Károly Bezdek
Antoine Deza
Yinyu Ye
Editors

Discrete Geometry and Optimization

FIELDS

Springer

Fields Institute Communications

VOLUME 69

The Fields Institute for Research in Mathematical Sciences

The Fields Institute is a centre for research in the mathematical sciences, located in Toronto, Canada. The Institutes mission is to advance global mathematical activity in the areas of research, education and innovation. The Fields Institute is supported by the Ontario Ministry of Training, Colleges and Universities, the Natural Sciences and Engineering Research Council of Canada, and seven Principal Sponsoring Universities in Ontario (Carleton, McMaster, Ottawa, Toronto, Waterloo, Western and York), as well as by a growing list of Affiliate Universities in Canada, the U.S. and Europe, and several commercial and industrial partners.

For further volumes:
http://www.springer.com/series/10503

Károly Bezdek • Antoine Deza • Yinyu Ye
Editors

Discrete Geometry and Optimization

The Fields Institute for Research
in the Mathematical Sciences

Springer

Editors
Károly Bezdek
Department of Mathematics
   & Statistics
University of Calgary
Calgary, AB, Canada

Antoine Deza
Department of Computing and Software
McMaster University
Hamilton, ON, Canada

Yinyu Ye
Department of Management Science
   and Engineering
Stanford University
Stanford, CA, USA

Cover illustration: Drawing of J.C. Fields by Ken Yeomans

Printed on acid-free paper

Kevin Cheung, Jun-Koo Park, Alexey Glazyrin, Elissa Ross, Zachary Voller, Megan Owen, Anthony Nixon, Walter Whiteley, Vincent Pilaud, Lorenz Klaus, Csaba Tóth, Canek Peláez, Tamon Stephen, Bernd Schulze, Monique Laurent, David Avis, Franz Rendl, David Bremner, Jan Foniok, Matthias Köppe, Edward Kim, Michel Deza, Itamar Halevy, István Szalkai

Oleg Musin, Frank Vallentin, Kim-Chuan Toh, Robert Connelly, Jesús De Loera, Antoine Deza, Károly Bezdek, Joseph Mitchell, Thomas Rehn, Katrin Herr

Ting Kei Pong, Pablo Parrilo, Levent Tunçel, Hayato Waki, Jon Lee, Michael Todd, Kurt Anstreicher, Nicolas Gillis, Gabor Pataki, Miguel Anjos, Lorenz Klaus, Vincent Pilaud, Kim-Chuan Toh, István Szalkai, Javier Peña

György Dósa, Adrian Lewis, Robert Freund, Marcel De Carli Silva, Antoine Deza, Károly Bezdek, Itamar Halevy, Gili Deza

Preface

Optimization has long been a source of both inspiration and applications for geometers, and conversely, discrete and convex geometry have provided the foundations for many optimization techniques, leading to a rich interplay between these subjects. The purpose of the Workshop on Discrete Geometry, the Conference on Discrete Geometry and Optimization, and the Workshop on Optimization, held in September 2011 at the Fields Institute, Toronto, was to further stimulate the interaction between geometers and optimizers. This volume reflects the fruitful interplay between these areas.

We would like to thank the contributors for their high-quality papers, as well as the referees for their thorough reviews. We are grateful to the Fields Institute and the National Science Foundation for the generous funding provided for the Thematic Program on Discrete Geometry and Applications. We wish to thank Jesús De Loera and Joseph Mitchell for co-organizing the events related to discrete geometry and optimization. It is a pleasure to acknowledge the excellent support provided by the Fields Institute; in particular, we would like to offer special thanks to Edward Bierstone, Alison Conway, Claire Dunlop, Matheus Grasselli, Debbie Iscoe, Matthias Neufang, and Carl Riehm.

Calgary, AB, Canada	Károly Bezdek
Hamilton, ON, Canada	Antoine Deza
Stanford, CA, USA	Yinyu Ye

Contents

Discrete Geometry in Minkowski Spaces

Javier Alonso, Horst Martini, and Margarita Spirova

Abstract   In this expository article we reflect some recent results from the geometry of real Banach (or normed) spaces in the spirit of Discrete Geometry. It is our goal to demonstrate that still today surprisingly elementary questions (at least from the Euclidean point of view) present parts of the research front in this field. We will discuss incidence theorems, some covering and location problems, group-theoretical aspects, and regularity of figures in normed planes. In addition, some single research problems are presented, partially even referring to gauges.

Key words Birkhoff orthogonality • Circumcenters • Covering discs • Group theory • James orthogonality • Minimal enclosing ball • Minsum hyperplanes • Minsum hyperspheres • Orthocentricity • Regularity of figures • Universal covers

Subject Classifications: 46B20, 52A10, 52A21, 52B12, 52C15, 52C17, 52C20, 90B85

1   Introduction

The geometry of finite-dimensional real Banach spaces goes back to H. Minkowski (see [40]), who was motivated by non-isotropic situations (in the sense of direction dependence) in Physics and by the Geometry of Numbers to extend questions from

J. Alonso
Departamento de Matemáticas, Universidad de Extremadura,
Adva. Elvas s/n, 06006 Badajoz, Spain
e-mail: jalonso@unex.es

H. Martini (✉) • M. Spirova
Faculty of Mathematics, Technical University of Chemnitz, D-09107 Chemnitz, Germany
e-mail: martini@mathematik.tu-chemnitz.de; margarita.spirova@mathematik.tu-chemnitz.de

Euclidean Geometry to linear spaces with convex distance functions or norms. An even earlier contribution was given by B. Riemann who, in [43] (see also [23] for the history) already mentioned the l_4-norm, as norm of tangent spaces in the sense of Finsler geometry. Today this field is usually called *Minkowski Geometry* (but should not be mixed up with Lorentz Geometry, also called Minkowskian Space-Time Geometry). Over the twentieth century, this field was permanently enriched by many contributions, also from neighbouring disciplines, such as general Banach Space Theory, Finsler Geometry, Geometric Convexity, Differential Geometry, and some other areas. But it remained independent of abstract Banach Space Theory. It is very interesting to see that, in contrast to the latter, also questions which (from the Euclidean point of view) are very elementary yield themes for the research front in Minkowski Geometry. Thus, simple notions like bisectors of segments, circumcenters, circumballs, minimal enclosing balls, angular bisectors, etc. are still interesting subjects of research on the geometry of normed spaces. It is our aim to demonstrate this phenomenon by various recent results. The analytical part of Minkowski Geometry was excellently summarized in the monograph [46] by A. C. Thompson. More recently, also researchers from Discrete and Computational Geometry, Optimization (in particular, from Location Science), Approximation Theory and further (more applied) disciplines contributed, but there is no monograph on this part of the field, although there is some need for a fundamental geometric theory (e.g., for continuing purely geometric investigations algorithmically, in the spirit of Computational Geometry). In the present article we will collect some recent (and mainly planar) results in this direction, basically referring to incidence theorems, covering problems, location problems, group-theoretical aspects, and norm-regularity of figures. The paper contains also various inspiring research problems. For background material we refer to [46] and the surveys [38] and [35].

We denote by X an *n-dimensional normed space*, i.e., an n-dimensional real Banach space with norm $\| \cdot \|$, whose *unit ball* $B := \{x \in X : \|x\| \leq 1\}$ is a convex body (i.e., a compact, convex set with non-empty interior) centered at the origin o. By conv, bd and int we denote *convex hull*, *boundary* and *interior*, respectively; bd B is called the *unit sphere* (for $n = 2$ unit circle) of X. A homothetical copy $x + \lambda \text{bd} B$ of bd B is said to be a *hypersphere* with center x and radius λ, denoted by $S(x, \lambda)$. In the case $n = 2$ it is called a *circle* denoted $C(x, \lambda)$. We say that X is *strictly convex* if bd B does not contain a non-trivial segment. Further on, X is *smooth* if B has precisely one supporting hyperplane at each $x \in \text{bd} B$, and X is *polyhedral* if B is the convex hull of a finite subset of X. A vector $x \in X$ is said to be *Birkhoff orthogonal* to $y \in X$, denoted by $x \perp_B y$, if $\|x + \alpha y\| \geq \|x\|$ for any $\alpha \in \mathbb{R}$, and x is *James orthogonal* to y if $\|x + y\| = \|x - y\|$. We refer to [3] for many properties of these orthogonality types. Clearly, Birkhoff orthogonality is in general not symmetric; but if it is, then, for $n = 2$, bd B is called a Radon curve and X a *Radon plane*. If X is an inner product space, then it is the *Euclidean space*.

2 Some Incidence Theorems

In this part we will demonstrate in which way generality might be lost when leaving the Euclidean plane to general normed planes. For example, *various theorems which are true in the Euclidean case for circles of arbitrary radii, hold in sufficiently large classes of normed planes only for circles of equal radii.*

We start with the basic "three-circles theorem", which is fundamental for many topics from Elementary Geometry (orthocentric point systems, inversive geometry and complex numbers, Feuerbach circles, etc.), Discrete Geometry (circle coverings), Descriptive Geometry (Theorem of Pohlke), Convexity (equilateral zonogons as cube shadows), configurations (Clifford's chain of theorems), and further topics; see the introduction of [29] for many references. It was extended in [5] to all strictly convex, smooth normed planes, and the following form of it (see [29]) can be proved without the smoothness assumption.

Theorem 1. *Let p_1, p_2, p_3 be distinct points of a strictly convex circle $C(x, \lambda)$, and let $C(x_i, \lambda), i = 1, 2, 3$, be three circles different from $C(x, \lambda)$ each of which contains two of the three points p_i. Then the intersection of the three circles $C(x_i, \lambda)$ is non-empty and consists of exactly one point p, called the C-orthocenter of the triangle $p_1 p_2 p_3$; see Fig. 1 (left side).*

We note that we need this theorem also below in our section on group-theoretical results. In the Euclidean subcase, p is the classical orthocenter, i.e., the intersection of the altitudes of the triangle formed by p_1, p_2, p_3 (each p_i playing the same role with respect to the remaining three points in this quadruple, therefore yielding an orthocentric system). Thus it makes sense to ask which type of orthogonality justifies, for strictly convex normed planes, the name "C-orthocenter" for p above. It turns out that *James orthogonality* yields the right answer, i.e., for any triangle formed by a triple from $\{p, p_1, p_2, p_3\}$ in Theorem 1, every vertex connects with the fourth point in a line which is James orthogonal to the opposite triangle side. In addition, we get for the configuration of Theorem 1 that the *Feuerbach or nine-point circle* from Euclidean geometry has, in general, only a *six-point analogue*
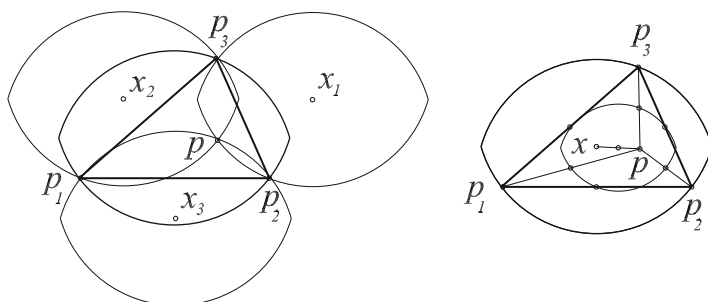
Fig. 1 The three-circles theorem (*left*) and the six-point circle (*right*)
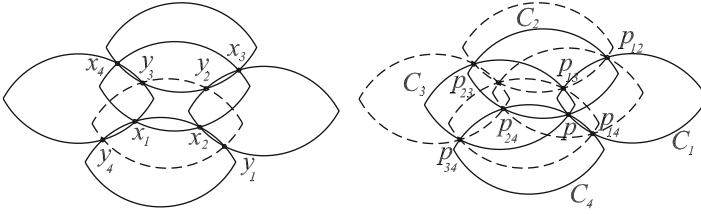
Fig. 2 Miquel's theorem (*left*) and Clifford's chain (*right*)

$C := C(\frac{1}{2}(x + p), \frac{1}{2}\lambda)$ passing through the three side midpoints and the three points $\frac{1}{2}(p + p_i)$. Also, C is a homothetical copy of the triangle's circumcircle, with center p and ratio $\frac{1}{2}$; see Fig. 1 (right side). More related results can be found in [5] and [29].

The famous *Theorem of Miquel* refers to $(8_3, 6_4)$-configurations of circles of possibly different sizes, meaning that they are formed by eight points and six circles such that each point lies on three circles and every circle passes through four points (see [8, p. 131]). Its appropriate extension (see [45] for this and further results) holds only for circles of equal radii.

Theorem 2. *Let $C(o, 1)$ be the unit circle of a strictly convex, smooth normed plane containing the four points x_1, x_2, x_3, x_4 in this order, and let C_i denote the four translates of $C(o, 1)$ (different to $C(o, 1)$, in each case) passing through pairs of neighbouring points from that quadruple. Then either there exists a proper translate of $C(o, 1)$ passing through the four points y_i, where $y_i = C_i \cap C_{i+1}, C_5 = C_1$, and $y_i \notin C$, or $y_i = x_i$ for $i = 1, 2, 3, 4$; see Fig. 2 (left side).*

It is well known that *Clifford's chain of theorems* can be transformed via inversive geometry to the following form for Euclidean circles (of possibly different radii; see [13, p. 262]): Let C_1, C_2, C_3 be coplanar circles all passing through p, and let C_i, C_j meet again in p_{ij}. Then p_{12}, p_{13}, p_{23} lie on a circle C_{123}; circles $C_{123}, C_{124}, C_{134}, C_{234}$ meet in p_{1234}; points $p_{1234}, \ldots, p_{2345}$ lie on a circle C_{12345}, and so on ad infinitum. Again, in normed planes it holds only for circles of equal radii; see [30].

Theorem 3. *Let $C_i = x_i + \lambda C(o, 1), i = 1, 2, 3, 4$, be four circles in a strictly convex normed plane passing through a point p. Let p_{ij} be the second intersection point of C_i and C_j. Then each triangle $p_{ij} p_{jk} p_{ki}$ with i, j, k pairwise different and from $\{1, 2, 3, 4\}$ has a circumcircle C_{ijk}. Furthermore, the four circles $C_{123}, C_{234}, C_{341}, C_{241}$ all pass through a point p_{1234}. Let $C_5 = x_5 + \lambda C(o, 1)$ be a fifth unit circle through p. Then the five points $p_{1234}, \ldots, p_{2345}$ lie on the circle C_{12345}, and so on ad infinitum; see Fig. 2 (right side).*

In contrast to the last two theorems, there are also Minkowskian analogues of classical theorems still referring to circles of arbitrary radii. An example is the Apollonius problem on circles touching three given ones in strictly convex, smooth normed planes; cf. [44].

We finish this section with a general observation on the overlap between Minkowski Geometry and fields like Foundations of Geometry (see again [6] and [8]) and Elementary Geometry (cf. [4] and [19]), since the Euclidean versions of the incidence theorems above belong to these fields. Checking the existing literature in Minkowski Geometry (see mainly [38, 46], and [35]), one will clearly observe that until now no "Minkowskian analogues" of many (in this sense certainly extendable) theorems from Euclidean Geometry are known. For example, there is no properly developed theory of triangles and simplices, of (metrically defined) classes of n-gons and polytopes, of circle and ball configurations, and of incidence and closure theorems referring to simply defined geometric figures in Minkowski spaces; see also the final remark of Sect. 5 below. The authors believe that fruitful research is possible in these directions.

3   Some Covering (and Related) Problems

In 1914, H. Lebesgue posed his famous *universal cover problem* for the Euclidean plane, namely to find a set of smallest possible area (called *universal cover*) having a congruent copy of any planar set of unit diameter as subset; see D15 and D16 in [14], § 11.4 in [11], and Chaps. V and VIII of [9] for many partial results, extensions, and applications. There are many variants of Lebesgue's (still unsettled) problem, e.g., by solving it using only restricted families of convex sets, like regular hexagons, balls and so on, as covers. But also in this restricted form, universal covers in normed planes and spaces were almost not investigated. For reasons described in the section group-theoretical aspects below, in general normed planes "congruent copy" above should be replaced by "translate", in the Euclidean case yielding the notion of *strong universal covers* instead. The paper [34] seems to be the only publication dealing explicitly with this problem in normed planes, although the Borsuk problem, considered in normed planes by several authors, has strong relations to universal covers there; see [16] and Chap. V in [9]. It is well-known that for a general norm $\| \cdot \|$ the relation $x \perp_B y$ does not imply $y \perp_B x$. But according to H. Busemann [12], for any given norm $\| \cdot \|$ there is a norm $\| \cdot \|_a$ with unit ball B_a (unique up to a factor) for which $y \perp_{B_a} x$ follows, for any pair with $x \perp_B y$. The norm $\| \cdot \|_a$ is called the *antinorm* of $\| \cdot \|$. (We note that bdB_a is nothing else than the *isoperimetrix* with respect to the original norm.) We say that a convex hexagon is *anti-regular* if it is affine regular and norm-equilateral with respect to $\| \cdot \|_a$. Moreover, since one can construct a Minkowskian angular measure regarding which a norm-equilateral triangle is also equiangular in the norm, we can define *regular* and also anti-regular triangles. The following theorems were proved in [34], and they refer to strong universal covers having triangular, hexagonal and 4-gonal shapes.

Theorem 4.   *In any normed plane X, an anti-regular hexagon circumscribed about a norm-circle of diameter 1 is a strong universal cover. Analogously, an anti-regular triangle circumscribed about a norm-circle of diameter 1 has the same property.*

It makes sense to call a parallelogram, which is norm-equilateral and has two neighbouring sides which are Birkhoff orthogonal, a *square* in this norm.

Theorem 5. *In any normed plane X, there is a square of side-length* 1 *being a strong universal cover, and X is Radon if and only if every such square is a strong universal cover.*

Further research in this direction should refer to universal covers belonging to larger classes of convex figures, and to higher dimensions. For this purpose, the related problems still open in the Euclidean subcase (see § 11.4 in [11] and Chaps. V and VIII in [9]) might be inspiring and should be suitably extended to normed planes and space.

At the end of this section we refer to the problem of *finding all possible locations of circumcircles, circumcenters and minimal enclosing discs of arbitrary triangles in general normed planes*. It is surprising that these (at first glance simple) location problems were completely solved only in the very recent papers [1] and [2]. The theorems derived there are too long and complicated to be presented here. We only mention that they are based on the complete classification of all possible intersection shapes that two homothetical norm circles can create (see [7, 16] and, for a refined presentation, [1]), and on some new notions, like norm-acuteness and norm-obtuseness of triangles. Besides many other surprising results in [1] and [2] it turns out that there are triangles having infinitely many minimal enclosing discs but a unique circumcircle, and that the completely described locus of all possible circumcenters is bounded also by parts of strangely occuring conics. It would be natural to extend the notions of circumcircles and minimal enclosing discs to more general given point sets and higher dimensions, also in the computational sense. This would yield deeper insights regarding unit distance graphs, Voronoi diagrams, and coresets in normed spaces (see, e.g., [31] for these notions and first steps in this direction).

4 Some Location Problems

Let X be an n-dimensional normed space ($n \geq 2$), and let $Y := \{y_1, \ldots, y_m\} \subset X$ be a finite set of points with corresponding weights $w_1, \ldots, w_m \in \mathbb{R}^+$. The famous Fermat-Torricelli problem (see [10], Chap. II) asks for the unique point in n-dimensional Euclidean space having the minimal sum of m weighted distances to the set Y. In an n-dimensional normed space X, the solution set $FT(Y) \subset X$ of this problem need not be (depending on the shape of the unit ball B) a singleton. Results on the geometric description of $FT(Y)$ are summarized in [39]. Among the many modifications and generalizations of this problem, the *minsum hyperplane problem* is particularly interesting: *Find the hyperplanes $H \subset X$ such that, among all hyperplanes of X, the weighted distance sum*

$$\sum_{i=1}^{m} w_i d(x_i, H), \quad d(x_i, H) := \min\{|\lambda| : (x + \lambda B) \cap H \neq \emptyset\},$$
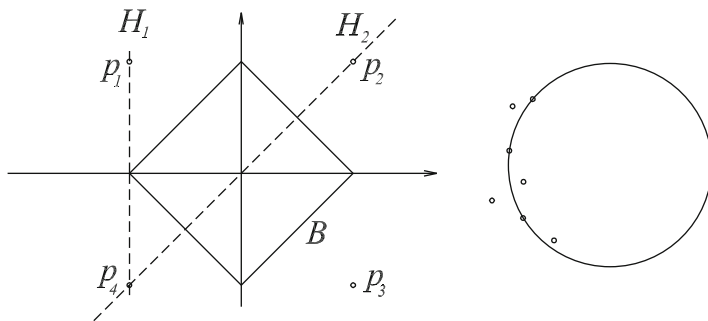
Fig. 3 Minsum lines (*left*) and a minsum circle (*right*)

is minimal. Such an optimal hyperplane is called a *minsum hyperplane* with respect to the weighted point set $Y \subset X$. Furthermore, a hyperplane H is said to be *pseudo-halving* with respect to Y if

$$\sum_{x_i \in H^+} w_i \leq \frac{W}{2} \quad \text{and} \quad \sum_{x_i \in H^-} w_i \leq \frac{W}{2},$$

where H^+, H^- are the two open halfspaces separated by H, and $\sum_{i=1}^m w_i = W$. In [26] and [25] the following theorem was proved, for which we exclude the trivial cases that Y is contained in a hyperplane of X.

Theorem 6. *For any weighted point set Y in an n-dimensional normed space $X, n \geq 2$, the following statements hold:*

1. *There exists a minsum hyperplane with respect to Y which is spanned by n affinely independent points from Y, and each minsum hyperplane is pseudo-halving with respect to Y.*
2. *The unit ball of X is smooth if and only if any minsum hyperplane with respect to Y is the affine hull of n affinely independent points from Y.*

In Fig. 3 (left side) we see that in the Manhattan norm also minsum hyperplanes exist which contain even no given point (the coordinate axes). Based on this theorem, the application of topological hyperplane sweep techniques for homogeneous hyperplane arrangements in $(n + 1)$-space yields $O(m^n)$ time (and $O(m)$ space) complexity to find all minsum hyperplanes spanned by n affinely independent given points. In certain cases (like for $n = 2$ and $n = 3$), this time complexity can be reduced a bit, using combinatorial results on halving lines and planes, and for polyhedral unit balls (with $2k$ vertices) even the time complexity $O(mk)$ is sufficient; see again [26], Chap. II of [10], and the references given there.

Considering hyperplanes like also points as "degenerate cases" of hyperspheres, one gets a natural extension of the minsum hyperplane (and the Fermat-Torricelli) problem. By $\mathscr{G}$ we denote the *set of all non-degenerate hyperspheres* of X.

For $a \in X$ we call $d(S, a) := \inf\{\|y - a\| : y \in S\}$ the *point-hypersphere distance*, and for Y like above we get the *minsum hypersphere problem: Find*

$$\min_{S \in \mathcal{G}} f(S) := \sum_{i=1}^{m} w_i \cdot d(S, y_i)$$

and the correspondingly optimal hyperspheres. This problem was already studied in Euclidean space (see [42] and the references there), and at first glance one might think that an incidence criterion analogous to Theorem 6 (1.) might hold; see Fig. 3 (right). As we shall see below, this is far from being true. For the following results we refer to [22]. Using the notion of d-segments (see [9, § 9]), one can prove that $d(S(x, r), a) = |\|x - a\| - r|$, and thus $d(S(x, r), a)$ is convex and piecewise linear in r. One can easily show that, for any normed space X, *no degenerate hypersphere with $r = 0$ can be optimal*, and *there always exists a set Y and a hyperplane H such that the objective value of H is superior to any $f(S), S \in \mathcal{G}$. If X is smooth, but not strictly convex, then there exists a finite set Y and a hyperplane such that the objective value of H is equal to that of a minsum hypersphere, and the same is true for polyhedral norms.* Nevertheless, for polyhedral norms the set $\mathcal{G}$ always contains a minsum hypersphere. Besides these existence criteria, the following position and incidence criteria can be fixed.

Theorem 7. *If $\mathcal{G}$ contains a minsum hypersphere for an arbitrary finite set $Y \subset X$, then there exists a least one minsum hyperpshere containing at least one point from Y. The intersection of any minsum hypersphere S and $\mathrm{conv}\, Y$ is non-empty, and we have $|S \cap \mathrm{conv} Y| \geq 2$ if the given set Y is inside or on S. If $|S \cap \mathrm{conv} Y|$ is finite, then $S \cap \mathrm{conv} Y \subset Y$.*

Defining the sets $J_{\pm}(S) := \{y_i \in Y : \|y_i - x\| \gtrless r\}$ and $J_0(S) := \{y_i \in Y : \|y_i - x\| = 0\}$ for $S = S(x, r)$, we get also

Proposition 1. *If $\mathcal{G}$ contains a minsum hypersphere $S = S(x, r)$, then*

$$\left| \sum_{y_i \in J_-(S)} w_i - \sum_{y_i \in J_+(S)} w_i \right| \leq \sum_{y_i \in J_0(S)} w_i.$$

As a next step, one should look for algorithmical approaches to minsum hyperspheres, and they also could be replaced by other approximating figures (like, e.g., by different types of minsum cylinders).

5   Group-Theoretical Results in Strictly Convex Normed Planes

The group of isometries of a (strictly convex) non-Euclidean normed plane consists of the semi-direct product of the translation group with a finite group of Euclidean rotations, or it is the dihedral group generated by two Euclidean line reflections.

This poorness motivated the authors of [32] to introduce and study left reflections for normed planes. In general, these are not isometries, but they are closely related to Birkhoff orthogonality: if every such reflection preserves Birkhoff orthogonality, then the respective plane is Radon! In general, these left reflections generate the group of affine transformations of determinant 1, but imposing additional properties for them and their products, or for the group generated by them, we get new characterizations of Radon planes or of the Euclidean plane (see, as continuation of [32], the paper [37]).

Definition 1. Given a line l in a strictly convex normed plane X, we define a transform $\Phi_i = (p \to p' : X \to X)$ to be a *left reflection in the line l* if

(i) $p' = p$ holds only for all $p \in l$,
(ii) $p' \neq p$ and $\text{aff}\{p, p'\} \perp_B l$ hold for all points not from l, and
(iii) the midpoint of the segment $[p, p']$ lies on l for all points p of X.

Thus left reflections are involutory, affine and have the property that all their fixed lines are Birkhoff orthogonal to the fixed-point line l, called their *axis*. Furthermore, the product of two of them in parallel lines is a translation, and all results presented in this section for this type of reflection analogously hold for the corresponding concept of "right reflections in lines for smooth normed planes"; see again [32]. Note that in smooth and strictly convex normed planes any left reflection is also a right reflection if and only if the plane is Radon, and if every left reflection preserves Birkhoff orthogonality, then the plane is again Radon. For the following results we refer to [37].

Theorem 8. *Let l_1, l_2, l_3 be three lines with common point p in a strictly convex, smooth normed plane. This plane is Euclidean if and only if there is a fourth line l_4 through p such that $\Phi_{l_3} \circ \Phi_{l_2} \circ \Phi_{l_1} = \Phi_{l_4}$.*

In this theorem the smoothness assumption can be replaced by the property that the strictly convex plane is Radon. This theorem also implies

Theorem 9. *A strictly convex normed plane, in addition being smooth or Radon, is Euclidean if and only if for any two lines l_1, l_2 intersecting in p and an arbitrary third line l_1' through p there exists a fourth line l_2' through p such that $\Phi_{l_2'} \circ \Phi_{l_1'} = \Phi_{l_2} \circ \Phi_{l_1}$.*

With the description of all isometries in normed planes at the beginning of this section we obtain also

Theorem 10. *In a strictly convex normed plane, every product of two left reflections is an isometry if and only if the plane is Euclidean.*

For the symmetry ψ_p with respect to an arbitrary point p of a strictly convex normed plane there exist two mutually Birkhoff orthogonal lines l_1, l_2 such that $\Phi_{l_2} \circ \Phi_{l_1} = \psi_p$. For $\mathscr{R}$ as family of all left reflections in a strictly convex normed plane we introduce the sets $\mathscr{R}^2 := \{\Phi_2 \circ \Phi_1 : \Phi_1, \Phi_2 \in \mathscr{R}, \Phi_1 \neq \Phi_2\}$ and $\mathscr{R}^3 := \{\Phi_3 \circ \Phi_2 \circ \Phi_1 : \Phi_1, \Phi_2, \Phi_3 \in \mathscr{R}\}$.

Theorem 11. *In a strictly convex normed plane the set $\mathscr{R}^* = \mathscr{R}^2 \cup \mathscr{R}^3$ is a group if and only if the plane is Euclidean. From this it follows that a strictly convex normed plane is Euclidean if and only if its left reflections generate a proper closed subgroup of the equi-affine group, and if and only if the image of any circle with respect to any left reflection is also a circle.*

Analogously we can prove that *a strictly convex normed plane is Euclidean if and only if every left reflection preserves James orthogonality*. For such a plane, we denote by P_p the pencil of all lines through a point p. The locus of all points that are images of some point $x \neq p$ with respect to the left reflections in lines from P_p is called the *cycle* of x with respect to P_p and denoted by $CY(p, x)$. For example, one can easily show that the cycle $CY(p, x)$ contains a line segment iff the unit circle has a corner point (belonging to more than one supporting line of B). The natural notion of cycle yields several interesting characterizations of large classes of normed planes.

Theorem 12. *Every two cycles $CY(p, x)$ and $CY(q, x)$ in a strictly convex normed plane, where p, q, x are collinear, have the unique point x in common if and only if the plane is smooth.*

And with respect to Radon planes we obtain

Theorem 13. *Any cycle $CY(p, x)$ in a strictly convex normed plane is symmetric with respect to p if and only if the plane is Radon. And for any non-degenerate line segment $[x, y]$ with midpoint p in a strictly convex normed plane, the cycles $CY(p, x)$ and $CY(p, y)$ coincide if and only if the plane is Radon.*

On the other hand, for the Euclidean subcase we get

Theorem 14. *Let $CY(p, x)$ be an arbitrary cycle in a smooth and strictly convex normed plane. Then, for any $y \in CY(p, x)$, the cycles $CY(p, x), CY(p, y)$ are identical if and only if the plane is Euclidean. And without the smoothness assumption, any cycle $CY(p, x)$ is a circle with center p if and only if the plane is Euclidean.*

We finish this section with some ideas on the concept of regularity of figures in normed planes. It is clear from above that notions like *rotation* and *angle* create difficulties in general normed planes, and that the notion of *symmetry group of a figure* is based on them. To our best knowledge, no *theory of regular figures in normed planes* is developed until now. In the monograph [46] we only find the known construction of convex *norm-regular hexagons* uniquely inscribed to a given norm circle (and with *norm-regular* meaning the combination of affine regularity and norm-equilaterality of these hexagons; see §4.1 in [46]). Simple norm examples show that already for triangles norm-equilaterality is not a satisfying notion for regularity. We present now one way to introduce such a notion for strictly convex normed planes; see [33]. The point p in Theorem 1 above was called the *C-orthocenter* of the triangle $p_1 p_2 p_3$ (with respect to James orthogonality) and satisfies $p = p_1 + p_2 + p_3 - 2x$; if the *C-orthocenter* p coincides with

the *circumcenter* x of the triangle (if this uniquely exists; see § 7.1 in [38]), we call $p_1 p_2 p_3$ *AG-regular* (due to the authors of [5]), and this is equivalent to $p_1 + p_2 + p_3 = 3x$ or to the property that the centroid, the circumcenter and the C-orthocenter of $p_1 p_2 p_3$ coincide. It is interesting to investigate how the notions of *AG-regularity* and *norm-equilaterality* behave to each other, and to characterize those strictly convex normed planes in which the incenter of any *AG*-regular triangle coincides with its circumcenter. Moreover, one even should clarify whether there exist *AG*-regular triangles; see [33].

Theorem 15. *Let X be a strictly convex normed plane with unit circle C. If p_1 lies on C, then there is exactly one triangle with vertex p_1 and circumcircle C which is AG-regular.*

It turns out that the construction of this (for given C and p_1 unique) *AG*-regular triangle is closely related to the classical construction of norm-regular hexagons mentioned above. Namely, every second vertex of such a hexagon, with starting vertex p_1, yields this unique *AG*-regular triangle (of course, its reflection at the center of C is of the same type, but does not contain p_1). This connection yields a new approach to affine regular, norm-equilateral hexagons and allows to define the *trigonal symmetry group* (containing the only symmetry operations) of an *AG*-regular triangle $p_1 p_2 p_3$ in a strictly convex normed plane, consisting of $\tau, \tau^2, \tau^3 = \iota$; here $\tau = \delta^2$, where δ maps the vertex set of the respective norm-regular hexagon cyclically onto itself, see again [33]. From this we obtain many useful applications of *AG*-regular figures. For example, *AG*-regular triangles allow constructions of *Dirichlet regions of lattices* and of *norm-regular* $\{3, 6\}$- *and* $\{6, 3\}$-*tessellations*, for arbitrary strictly convex normed planes; see again [33], but also [31].

AG-regularity presents only one possible type of regularity in Minkowski planes; there are various other ways to define regularity of polygons and polytopes in normed planes and spaces. To our best knowledge, almost nothing was systematically done in this direction.

6   Outlook: Jordan Curves and Gauges

We finish this paper by mentioning some single results showing again that the study of various *natural geometric research problems in normed planes* seems to be, in some sense, at the beginning only. For example, the *inscribed square problem* asks whether for every Jordan curve in the Euclidean plane there exists at least one inscribed square. We refer to Sect. 11 of [21] and to [20] for detailed discussions of this unsolved problem. Special cases have been positively answered, like that of *convex* Jordan curves or sufficiently *smooth* ones. A related result (see [41]) says that any Jordan curve contains the four vertices of some rhombus with two sides parallel to any given line in the Euclidean plane. It seems that there are (except for [36]) no extensions of such results to normed planes. We say that a parallelogram

uvst, embedded in a normed plane X, is a *rhombus* if $\|u - v\| = \|v - t\|$ (note that, unlike the Euclidean subcase, quadrilaterals with four equal side-lengths need not be rhombi). Moreover, a rhombus is said to be a *square* if also $\|u - s\| = \|v - t\|$ (cf. the different definition of *square* in our discussion on universal covers above). With these notions, the following was proved in [36].

Theorem 16. *For any convex Jordan curve in an arbitrary normed plane there exists at least one inscribed square.*

Further on, in a strictly convex normed plane no two different rhombi with the same prescribed direction of one diagonal can be inscribed in the same strictly convex Jordan curve (see again [36]). These results (and their proofs) might give a good starting point for solving other interesting subcases (analogous to existing Euclidean results; see again [14, Sect. 11]) of the inscribed square problem! Here the large variety of possible definitions of different types of 4-gons in normed planes seems to yield a rich field of geometrically interesting problems.

Another type of results refers to the more general *gauges* or *convex distance functions*, a concept which was introduced by H. Minkowski before that of norms! The unit circles of planes with convex distances functions are defined as those for normed planes, only central symmetry is not demanded (but the origin o is still an interior point). Convex distance functions have increasing importance, e.g. in Computational Geometry. As for normed planes (see [35]), it is natural to study quantities like the self-circumference of the unit ball for convex distance functions, called *normalizing figure B*. Since this self-circumference depends on the orientation of bd$B$, we use $L_o^-(B)$ and $L_o^+(B)$ for both these quantities, emphasizing also their dependence on the position of $o \in \text{int}B$. The following is known about the polygonal cases. B. Grünbaum [18] proved that $\min_{o \in \text{int}B} L_o^{\mp}(B) \leq 9$, and that equality holds for B a triangle. The global lower bound is given by $L^{\mp}(B) \geq 6$, with both cases of equality iff B is an affine regular hexagon (as for norms; see [17]). V. V. Makeev [24] observed that the self-perimeter is a strictly convex function with respect to the position of $o \in \text{int}B$; thus for affine regular k-gons P_k the lower bound is attained when o is their barycenter (and equals 9 when $k = 3$; see above). By continuity arguments, this yields $\inf_{\{P_k\}} L^{\mp}(P_k) = 6$, for all convex k-gons with $k \geq 6$. The remaining two cases $k = 4$ and $k = 5$ were settled in [27] and [28], respectively.

Theorem 17. *If P_4 is the normalizing quadrangle of a convex distance function in the plane, then the sharp lower bound*

$$L^{\mp}(P_4) \geq 3 \cdot (x_0 + 1) + x_0^{-1} = 7,7290\ldots$$

holds, where x_0 denotes the real root of the equation $x^3 - x - 1 = 0$. On the other hand, if P_5 is the normalizing pentagon of a convex distance function in the plane, then the sharp lower bound

$$L^{\mp}(P_5) \geq 3 \cdot (1 + x_0^{-1}) + 2x_0^{-2} = 6,7996\ldots$$

holds, where x_0 is the positive root of the equation $x^4 - x - 1 = 0$.

There are different ways to extend the notion of self-perimeter of polygonal unit circles to higher dimensions. For example, one can replace "perimeter" by "surface area", thus coming to isoperimetric problems in normed spaces (see Chap. 5 of [46]). On the other hand, "perimeter of polygons" can be replaced by "total edge-length of polytopes", measured in the norm induced by them. The following conjecture is due to K. J. Swanepoel (oral communication).

Conjecture. Let P, centred at the origin o, be the polyhedral unit ball of a normed 3-space. Then P has the minimum total edge-length $90 - 30\sqrt{5} \approx 22,92$, measured in the norm induced by itself, if and only if P is an affine image of a Platonic dodecahedron.

For partial results we refer to [15], where this minimum is found for the following classes of convex polyhedra: Platonic, Archimedean, and Catalan solids, further semiregular polyhedra and their duals, zonohedra and their duals. For the class of zonohedra the sharp lower bound is 24, and the equality cases are characterized.

References

1. Alonso, J., Martini, H., Spirova, M.: Minimal enclosing discs, circumcircles, and circumcenters in normed planes (Part I). Comput. Geom. **45**, 258–274 (2012)
2. Alonso, J., Martini, H., Spirova, M.: Minimal enclosing discs, circumcircles, and circumcenters in normed planes (Part II). Comput. Geom. **45**, 350–369 (2012)
3. Alonso, J., Martini, H., Wu, S.: On Birkhoff orthogonality and isosceles orthogonality in normed linear spaces. Aequ. Math. **83**, 153–189 (2012)
4. Altshiller-Court, N.: College Geometry. Barnes & Noble, New York (1952)
5. Asplund, E., Grünbaum, B.: On the geometry of Minkowski planes. L'Enseignement Math. (2) **6**, 299–306 (1960)
6. Bachmann, F.: Aufbau der Geometrie aus dem Spiegelungsbegriff. Springer, Berlin/Göttingen/Heidelberg (1959)
7. Banasiak, J.: Some contributions to the geometry of normed linear spaces. Math. Nachr. **139**, 175–184 (1988)
8. Benz, W.: Vorlesungen über die Geometrie der Algebren. Springer, Berlin/Heidelberg/New York (1973)
9. Boltyanski, V., Martini, H., Soltan, P.S.: Excursions into Combinatorial Geometry. Springer, Berlin/Heidelberg (1997)
10. Boltyanski, V., Martini, H., Soltan, V.: Geometric Methods and Optimization Problems. Kluwer, Dordrecht (1999)
11. Brass, P., Moser, W., Pach, J.: Research Problems in Discrete Geometry. Springer, New York (2005)
12. Busemann, H.: The isoperimetric problem in the Minkowski plane. Am. J. Math. **69**, 863–871 (1947)
13. Coxeter, H.S.M.: Introduction to Geometry, 2nd edn. Wiley, New York (1969)
14. Croft, H.T., Falconer, K.J., Guy, R.K.: Unsolved Problems in Geometry. Springer, Berlin/Heidelberg/New York (1991)

15. Düvelmeyer, N.: On the total length of edges of a polyhedral unit ball. Util. Math. **68**, 33–52 (2005)
16. Grünbaum, B.: Borsuk's partition conjecture in Minkowski planes. Bull. Res. Counc. Isr. Sect. F **7F**, 25–30 (1957/1958)
17. Grünbaum, B.: Self-circumference of convex sets. Colloq. Math. **13**, 55–57 (1964)
18. Grünbaum, B.: The perimeter of Minkowski unit discs. Colloq. Math. **15**, 135–139 (1966)
19. Johnson, R.A.: Advanced Euclidean Geometry. Dover, New York (1960)
20. Klee, V.: Some unsolved problems in plane geometry. Math. Mag. **52**, 131–145 (1979)
21. Klee, V., Wagon, S.: Unsolved Problems in Plane Geometry and Number Theory. MAA Dolciani Mathematical Expositions, No. 11. The Mathematical Association of America, Washington, DC (1991)
22. Körner, M.-C., Martini, H., Schöbel, A.: Minsum hyperspheres in normed spaces. Discret. Appl. Math. **160**, 2221–2233 (2012)
23. Laugwitz, D.: Bernhard Riemann 1826–1866. Wendepunkte in der Auffassung der Mathematik. Birkhäuser, Basel (1996)
24. Makeev, V.V.: On upper estimates for the perimeter of non-symmetric unit circles of Minkowski planes (in Russian). Zap. Nauchn. Semin. LOMI **299**, 262–266 (2003)
25. Martini, H., Schöbel, A.: Two characterizations of smooth norms. Geom. Dedicata **77**, 173–183 (1999)
26. Martini, H., Schöbel, A.: Median and center hyperplanes in Minkowski spaces – a unified approach. Discret. Math. **241**, 407–426 (2001)
27. Martini, H., Shcherba, A.: On the self-perimeter of quadrangles for gauges. Beitr. Algebra Geom. **52**, 191–203 (2011)
28. Martini, H., Shcherba, A.: On the self-perimeter of pentagonal gauges. Aequ. Math. **84**, 157–183 (2012)
29. Martini, H., Spirova, M.: The Feuerbach circle and orthocentricity in normed planes. L'Enseignement Math. **53**, 237–258 (2007)
30. Martini, H., Spirova, M.: Clifford's chain of theorems in strictly convex Minkowski planes. Publ. Math. Debr. **72**, 371–383 (2008)
31. Martini, H., Spirova, M.: On regular 4-coverings and their application for lattice coverings in normed planes. Discret. Math. **309**, 5158–5168 (2009)
32. Martini, H., Spirova, M.: Reflections in strictly convex Minkowski planes. Aequ. Math. **78**, 71–85 (2009)
33. Martini, H., Spirova, M.: Regular tessellations in normed planes. Symmetry **22**, 223–235 (2011)
34. Martini, H., Spirova, M.: On universal covers in normed planes. Adv. Geom. **13**, 41–50 (2013)
35. Martini, H., Swanepoel, K.J.: The geometry of Minkowski spaces – a survey. Part II. Expo. Math. **22**, 93–144 (2004)
36. Martini, H., Wu, S.: Minkowskian rhombi and squares inscribed in convex Jordan curves. Colloq. Math. **120**, 249–261 (2010)
37. Martini, H., Spirova, M., Strambach, K.: Geometric algebra in strictly convex Minkowski planes. Aequat. Math., p. 18 (to appear)
38. Martini, H., Swanepoel, K.J., Weiss, G.: The geometry of Minkowski spaces – a survey. Part I. Expo. Math. **19**, 97–142 (2001)
39. Martini, H., Swanepoel, K.J., Weiss, G.: The Fermat-Torricelli problem in normed planes and spaces. J. Optim. Theory Appl. **115**, 283–314 (2002)
40. Minkowski, H.: Sur les propriétés des nombres entiers qui sont dérivées de l'intuition de l'espace. Nouvelles Annales de Mathematiques, 3e Série **15** (1896); also in: Ges. Abh. 1, Band XII, pp. 271–277
41. Nielsen, M.J.: Rhombi inscribed in simple closed curves. Geom. Dedicata **54**, 245–254 (1995)
42. Nievergelt, Y.: Median spheres: theory, algorithms, applications. Numer. Math. **114**, 573–606 (2010)

43. Riemann, B.: Über die Hypothesen, welche der Geometrie zu Grunde liegen. Abh. Königl. Ges. Wiss. Göttingen **13**, 133–152 (1868)
44. Spirova, M.: Circle configurations in strictly convex normed planes. Adv. Geom. **10**, 631–646 (2010)
45. Spirova, M.: Miquel's theorem and inversions in normed planes. Monatsh. Math. **161**, 335–345 (2010)
46. Thompson, A.C.: Minkowski Geomery. Cambridge University Press, Cambridge (1996)

Engineering Branch-and-Cut Algorithms for the Equicut Problem

Miguel F. Anjos, Frauke Liers, Gregor Pardella, and Andreas Schmutzer

Abstract A minimum equicut of an edge-weighted graph is a partition of the nodes of the graph into two sets of equal size such that the sum of the weights of edges joining nodes in different partitions is minimum. We compare basic linear and semidefinite relaxations for the equicut problem, and find that linear bounds are competitive with the corresponding semidefinite ones but can be computed much faster. Motivated by an application of equicut in theoretical physics, we revisit an approach by Brunetta et al. and present an enhanced branch-and-cut algorithm. Our computational results suggest that the proposed branch-and-cut algorithm has a better performance than the algorithm of Brunetta et al. Further, it is able to solve to optimality in reasonable time several instances with more than 200 nodes from the physics application.

Key words Equicut • Maximum-Cut • Bisection • Graph partitioning • Linear programming • Semidefinite programming • Branch-and-cut

Subject Classifications: 90C57, 90C22, 90C05, 90C27

M.F. Anjos (✉)
Canada Research Chair in Discrete Nonlinear Optimization in Engineering, GERAD &
École Polytechnique de Montréal, CP 6079, succ. Centre-ville, Montreal, QC H3C 3A7, Canada
e-mail: anjos@stanfordalumni.org

F. Liers
Department Mathematik, Friedrich-Alexander-Universität Erlangen-Nürnberg,
Cauerstraße 11, 91058 Erlangen, Germany
e-mail: frauke.liers@math.uni-erlangen.de

G. Pardella • A. Schmutzer
Institut für Informatik, Universität zu Köln, Weyertal 121, 50931 Köln, Germany
e-mail: pardella@informatik.uni-koeln.de; schmutzer@informatik.uni-koeln.de

1 Introduction

The maximum cut problem on an edge-weighted graph $G = (V, E)$ is to find a partition of the set of nodes V into two shores (node sets) such that the weight of the cut (the sum of the weights of edges with endpoints in different shores) is maximum. It is a prominent NP-hard combinatorial optimization problem that has been studied intensively in the literature.

We consider the equicut problem which is the max-cut problem with the additional restriction that the sizes (number of nodes) of the shores must be equal. This work is motivated by an application in theoretical physics: equicuts can be used for the calculation of minimum-energy states, or ground states, for so-called Coulomb glasses. In a Coulomb glass, charges may be placed on the sites of a lattice. The number of charges is exactly half the number of sites. Randomly chosen local fields act on the charges. Since a quadratic function is used to represent the energy of a state as a graph G, the task is to determine an equicut in G.

Polyhedral investigations of the equicut polytope have been presented in [10, 11, 14]. Building upon this theoretical knowledge, an integer programming-based branch-and-cut approach was implemented in [8]. The bisection problem is the more general task of determining a cut in which the shore sizes are constrained (but not necessarily equal). Formulations using integer programming, semidefinite programming (SDP), a polyhedral study and computational results are presented in [2, 3, 29]. Another branch-and-bound algorithm using SDP formulations of the bisection problem that specifically accounts for the special case of equicut is given in [25]. A quadratic convex reformulation of the bisection problem is given in [6]. Further generalizations of the bisection problem with additional node capacities are studied in [17, 18, 22]. The maximum cut problem as well as other related problems have been investigated in great depth [15]. SDP-models for related graph partition problems have been introduced in [19, 24] and in the recent preprints [27, 32].

In this work, we first summarize important facts about the cut and the equicut polytopes and their relationships. The main part of this paper presents an algorithm engineering approach for an exact branch-and-cut algorithm for computing equicuts. Experimentally, we find that the bounds obtained from the SDP relaxation of [19] with several additional linear constraints are usually not stronger than those from solving the same linear relaxation without the SDP model. Furthermore, the computation of the SDP bounds often needed considerably more time. We then focus on the usage of integer linear programming (ILP) methods. We take the most important ingredients from the method of [8] and enrich it by target cuts, a variant of local cuts [9]. It turns out that within this separation routine, orthogonal projections on node-induced subgraphs together with zero-lifting of the separated inequalities works well in practice. For complete graphs, the projection and lifting approach is equivalent to the graph shrinking procedure from [23]. As target cut separation yields very strong bounds but can be time-consuming, we then design fast heuristics to separate those inequalities found by target cut separation. We finally show that our approach yields an effective method for determining optimal equicuts in graphs,

i.e., the computation times needed are about two orders of magnitude less than those reported in [8] on average. Furthermore we could reduce the average number of branching nodes by a factor of approximately 4; for some of the largest instances the computation time was three orders of magnitude faster than those reported in [8]; and instances that required more than 50 branching nodes are now solved at the root node. Hence we could solve some large instances that were not reported in [8] as well as instances of Coulomb glasses to optimality. We also solved instances with more than 200 nodes, doubling the size of the instances reported in [8].

2 The Equicut Problem

Let a graph $G = (V, E)$ with node set V, edge set E, and edge weights $c : E \to \mathbb{R}$ be given. We denote by n and m the cardinalities of V and E, respectively. The complete graph on n nodes is denoted by K_n. For $S \subseteq V$ a cut $\delta(S) \subseteq E$ is the set of edges with exactly one node in S. Its weight is the sum of the weights of edges in $\delta(S)$. We call S and $V \setminus S$ shores of $\delta(S)$. A cut $\delta(S)$ is an equicut if $\left\lfloor \frac{n}{2} \right\rfloor \leq |S| \leq \left\lceil \frac{n}{2} \right\rceil$. We associate with every cut $\delta(S)$ its characteristic vector $x(\delta(S)) = (x_1, \ldots, x_m) \in \{0, 1\}^m$ where $x_e = 1$ if $e \in \delta(S)$ and $x_e = 0$ otherwise. The cut polytope $\mathscr{C}(G)$ and the equicut polytope $\mathscr{Q}(G)$ of a graph G are the convex hulls of the characteristic vectors of all cuts and of all equicuts of G, respectively. The equicut problem consists in finding an equicut in G with minimum weight.

2.1 The Cut and the Equicut Polytope

In this section, we review known results about the equicut polytope [10, 11].

For a complete graph K_{2p}, the edge set induced by an equicut contains exactly p^2 many edges. Furthermore, every equicut in K_{2p+1} contains $p(p + 1)$ many edges. Thus, an equicut on a complete graph K_n has exactly $\left\lfloor \frac{n}{2} \right\rfloor \left\lceil \frac{n}{2} \right\rceil$ edges in the cut.

For complete graphs, the equicut polytope $\mathscr{Q}(K_n)$ can be defined as

$$\mathscr{Q}(K_n) = conv\left\{ x \in \mathscr{C}(K_n) \mid \sum_e x_e = \left\lfloor \frac{n}{2} \right\rfloor \left\lceil \frac{n}{2} \right\rceil \right\}. \tag{1}$$

Given a relaxation of the cut polytope we can use (1) to obtain a relaxation of the equicut problem as all inequalities remain valid.

We use the following classes of valid inequalities, some of which are well-known from the cut polytope [15]. For a more in-depth discussion see [10, 11] and [14].

Definition 1. Let $K_p = (V', E')$ be a complete subgraph (clique) of $G = (V, E)$. The clique inequality is given by

$$\sum_{e \in E'} x_e \leq \left\lfloor \frac{p}{2} \right\rfloor \left\lceil \frac{p}{2} \right\rceil. \tag{2}$$

For each cut $\delta(S)$ in K_p the switched clique inequality reads

$$\sum_{e \in E' \setminus \delta(S)} x_e - \sum_{e \in \delta(S)} x_e \leq \left\lfloor \frac{p}{2} \right\rfloor \left\lceil \frac{p}{2} \right\rceil - |\delta(S)|. \tag{3}$$

For the special case $p = 3$ the clique inequalities reduce to the triangle inequalities. For any triplet i, j, k such that $(i, j), (i, k), (j, k) \in E$, they take the following forms:

$$x_{ij} + x_{ik} + x_{jk} \leq 2 \tag{4}$$

$$x_{ij} - x_{ik} - x_{jk} \leq 0. \tag{5}$$

Barahona et al. [4] proved that the clique inequalities (2) are facets of the cut polytope $\mathscr{C}(K_p)$ iff p is odd. Hence the equicut polytope $\mathscr{Q}(K_n)$ is a face of the cut polytope and it is a facet iff n is odd [10].

As the number of cut edges in a cycle is even, for every cycle C with $|C| = p+1$ and $n = 2p$ the cycle inequalities (6) are valid for $\mathscr{Q}(K_n)$:

$$\sum_{e \in C} x_e \geq 2. \tag{6}$$

For a formal proof and facet inducing properties of these inequalities we refer to Conforti et al. [11].

If an edge $e = st$ is fixed to be in the cut, the resulting polytope is known as s-t-(equi-)cut polytope. Given a path $P = (s = v_0, v_1, \ldots, v_k = t)$ we consider the following s-t path inequalities that make sure that at least one edge on that path is cut:

$$\sum_{e \in P} x_e \geq 1. \tag{7}$$

It is easy to see that inequalities (7) are valid for the respective s-t-cut polytopes and they can be separated efficiently by shortest path computations. For further reasons to use these inequalities we refer to Conforti et al. [11].

For complete graphs K_{2p}, any equicut is also a p-regular subgraph of K_{2p} [8], i.e. each node has degree $|\delta(v)| = p$. Therefore, $\mathscr{Q}(K_{2p})$ is contained in $\mathscr{P}\mathscr{F}(K_{2p})$, which is the convex hull of all p-regular subgraphs of K_{2p}. Edmonds et al. [16] give the following complete description of $\mathscr{P}\mathscr{F}(K_{2p})$:

$$\sum_{e \in \delta(v)} x_e = p \tag{8}$$

$$\sum_{e \in W \times W} x_e + \sum_{e \in T} x_e \leq \frac{p|W| + |T| - 1}{2}, \tag{9}$$

where $v \in V$, $W \subseteq V$, $T \subseteq \delta(W)$ and $p|W| + |T|$ is odd. Inequalities (9) are called blossom inequalities.

We will work with complete graphs K_{2p} in the following as the inequalities (8) may be invalid for non-complete graphs.

2.2 Inequalities Outside the Template Paradigm

The classes of inequalities introduced earlier follow the template paradigm, i.e. inequalities within a class share a similar structure. A general procedure to generate valid inequalities for some polytope is given by separation of local cuts [1] or their variant called target cuts [9]. For the separation, the polytope in question and the point to be separated are projected into a low-dimensional space. A cutting plane separating the projected point from the projected polytope is generated by solving a small linear program. The size of the linear program is basically determined by the number of vertices of the projected polytope. Inequalities are then lifted to inequalities valid for the original problem.

2.3 Non-polyhedral Model

Given a weight matrix $C = \{c_{ij}\}$ for the edges $e = ij$ of a complete graph K_n the following SDP relaxation for the equicut problem was introduced by Frieze and Jerrum [19]:

$$\min \left\{ \frac{1}{4} trace(C(J_n - X)) \mid diag(X) = e, Xe = 0, X \succeq 0 \right\} . \qquad (10)$$

The notation $trace(A)$ refers to the sum of all elements of the main diagonal of a matrix A and the matrix J_n is the $n \times n$ matrix of all ones. In the positive semidefinite relaxation of the cut polytope [15], we restrict ourselves to positive-semidefinite matrices X ($X \succeq 0$) where all elements of the main diagonal are equal to one ($diag(X) = e$). This SDP relaxation is also referred to as the elliptope $\mathscr{E}_n$. A relaxation of the equicut problem is obtained by adding the constraint $Xe = 0$ as in [19] or in the form $e^T Xe = 0$ as in [25].

The variables $x_{ij}^{lp} \in \{0, 1\}$ from the LP formulation are in one-to-one correspondence with the elements $x_{ij}^{sdp} \in X$ of the positive-semidefinite matrix X via the transformation $x_{ij}^{sdp} = 1 - 2x_{ij}^{lp}$.

Our basic LP relaxation includes degree and triangle inequalities:

$$\min \left\{ \sum_e c_e x_e \mid \forall i, j, k \in V : (4) - (5), \forall v \in V : (8) \right\} . \qquad (11)$$

As we will see, the relaxation obtained by separating (switched) clique (3), cycle (6) and blossom (9) inequalities is already very good in practice. In fact we can solve most instances of the benchmark from Brunetta et al. [8] at the root node without branching. In some cases a better performance of our method can be obtained by forcing the algorithm to branch when there is too little improvement in the objective value (cf. tailing-off, Sect. 3.1). Target cut separation may improve the performance when applied before branching.

Table 1 LP and SDP bounds of root relaxations

Instance	$|V|$	Opt	All	Triangles		All but cliques	
				LP	SDP	LP	SDP
reti/2 × 11	22	11	11	10.24	10.34	10.24	10.34
tori/15 × 2	30	11	11	9.42	9.46	9.42	9.46
reti/16 × 2	32	10	9.44	6.27	6.27	6.27	6.27
tori/16 × 2	32	11	11	9.31	9.31	9.31	9.31
tori/18 × 2	36	13	13	9.58	9.58	9.58	9.58
misti/2 × 19 m	38	388	386.61	383.88	383.88	383.88	383.88
reti/19 × 2	38	10	8.94	6.77	6.78	6.77	6.78
reti/2 × 19	38	6	4.83	2.89	2.9	2.89	2.9
rand/q0.20	40	1,238	1,238	1,152	1,237.73	1,152	1,237.73
rand/q0.60	40	530	530	489.55	528.32	489.55	528.32
reti/20 × 2	40	10	9.78	5.87	5.87	5.87	5.87

In order to evaluate the respective bounds based on SDP and LP relaxations we first applied our branch-and-cut algorithm to the benchmark from Brunetta et al. [8]. For each instance we studied the relaxation that was obtained in the root, i.e. without branching.

In order to get reasonable computation times and memory usage we chose all instances with less than 10^5 inequalities in the final LP. There were 11 such instances. Since they were already solved at the root node, the respective SDP bounds are not significantly better than the LP bounds. In order to evaluate their strengths in practice nevertheless, we compare the LP bounds with those of the SDP when adding only a fraction of the inequalities separated by branch-and-cut. All SDP relaxations were computed using CSDP 6.1.1. [7]. The LPs were solved using version 12.1 of the CPLEX callable library [12]. We show in Table 1 the bounds that are obtained by solving the LP as well as the SDP when adding all separated inequalities (All), when adding only all triangle inequalities (Triangles) and when adding all inequalities but leaving out the (switched) clique inequalities (All but cliques). The generated inequalities are very strong in practice. It turns out that the triangle inequalities are important in practice as the corresponding bound is quite close to that given by all inequalities. This is known to hold for maximum cut as well, see e.g. [31].

Furthermore, the switched clique inequalities are most significant for improving the relaxation over the triangle bound as the remaining inequalities do not improve the bound any further. Similar results are obtained for most instances. Hence we conclude that using SDP relaxations does not significantly improve the root bounds compared to those obtained from the LP approach. Furthermore, the LP relaxations were obtained within minutes whereas some of the SDP relaxations needed several days of CPU time. We also tried to use a general maximum cut solver (BiqMac [31]) by dualizing the equicut constraint in the form $e^T X e = 0$. Therefore we introduced sufficiently large penalties for solutions that would use less than $\frac{n^2}{4}$ cut edges. More precisely we added the sum of the absolute values of edge weights to each edge

in the maximization problem. Hence optimal maximum cuts had to be equicuts and its values could be obtained by substracting a large constant. The solver sometimes was about two orders of magnitudes faster than our implementation. Unfortunately we observed that the precision of the solution values was limited to three digits. While this did not adversely affect the results for instances from [8], the solutions obtained by this approach for instances with fractional edge weights may be incorrect. Specifically for instances coming from the physics application, we could not obtain useful solutions even for the smallest instances. We thus chose to use the LP relaxation for our purposes.

Armbruster et al. [3] present relaxations using LP and SDP methods to solve a generalization of the equicut problem, i.e. the minimum bisection problem, where the number of nodes in each shore has to be larger than some parameter $F \leq \lfloor \frac{n}{2} \rfloor$. Their computational results on sparse instances suggest that SDP relaxations are superior to the corresponding LP relaxations. As equicut is a special case of the minimum bisection problem this appears to contradict the above observations. However, in contrast to the relaxations used in [3], our model is defined on complete graphs and includes constraints that are not valid for the minimum bisection polytope in general. Hence it is likely that our LP relaxation for the equicut problem is stronger than more general relaxations for the minimum bisection problem, and that it is especially well suited for dense instances such as those coming from the physics application.

3 Enhanced Branch-and-Cut Algorithm

Branch-and-cut is a framework often used for solving NP-hard combinatorial optimization problems exactly. Upper and lower bounds on the objective value are iteratively improved until optimality of a known solution can be proven. The size of the branching tree is kept small by using strong relaxations for the lower bounds and primal heuristics for the upper bounds. In this section, we describe cutting-plane separation and primal heuristics for our proposed branch-and-cut algorithm.

3.1 Cutting-Plane Separation

Given a (possibly fractional) vector $x^* \in \mathbb{R}^m, 0 \leq x^* \leq 1$, the separation problem asks for either an inequality violated by x^* or a proof that all inequalities valid for the polytope in question are satisfied. An algorithm that solves the separation problem for any fractional solution is called *exact*. In contrast *heuristic* separation algorithms find violated inequalities but if none is found, they cannot prove that there are no violated inequalities for the polytope. Separation algorithms are often defined for classes of inequalities such as those introduced in Sect. 2.

In general some classes of inequalities are more important than others or may become more important after several iterations. Brunetta et al. [8] use the relative change in the objective value from the previous iteration to decide which class of inequalities should be separated. Therefore they introduce certain threshold values for each class of inequalities. Hence they separate triangle inequalities if the relative change is above a certain value, then clique separation is applied if the relative change is above another value and then blossom inequalities and $(p + 1)$-cycle inequalities are separated if the relative change is above their respective thresholds. If the relative change is below any thresholds branching is applied. Because the thresholds are very sensitive to small changes in the separation routines, we decided to use just a single threshold to decide whether any separation is applied or we branch. We found that a factor $\alpha = 10^{-4}$ for the relative change of the objective value was a reasonable choice. Further in each iteration we stop separation whenever we found a certain number of violated inequalities. We found that 500 violated inequalities are a reasonable choice for most instances.

The degree inequalities (8) are all added from the beginning. In contrast to the heuristics for separating blossom inequalities in [8], we separate them exactly by the efficient algorithm [28].

Finally, target cut separation is applied whenever the other routines do not find any violated inequality. Target cuts were introduced in [9]. The key observation for target cuts is the following. For $k \leq m$, let π be some projection $\mathbb{R}^m \to \mathbb{R}^k$. Then the projection $\overline{P} = \pi(P) \subseteq \mathbb{R}^k$ of some polytope P in $\mathbb{R}^m$ is the convex hull of all points in $\mathbb{R}^k$ that can be extended to a point in P. Thus, $\overline{P}$ is the convex hull of all points $\pi(x_1), \ldots, \pi(x_r) \in \mathbb{R}^k$ such that $x_1, \ldots, x_r$ are the vertices of P. For $k \ll m$, many of the $\pi(x_i)$ are equal so that for small k, $\overline{P}$ can be dealt with efficiently. For details, we refer the reader to [9].

For the equicut polytope, we use orthogonal projections to (node-induced) subgraphs. More specifically, edges incident on nodes which are not in the considered subgraph are neglected. The projected polytope is then again a cut polytope. Given an inequality valid for $\overline{P}$, the corresponding inequality with coefficients set to zero for the neglected edges is then valid for P ('zero-lifting').

For the maximum cut problem, another projection is given through shrinking nodes to supernodes as introduced by Jünger et al. [23]. This projection is especially tailored for sparse graphs. For an edge $(s, t) \in E$ nodes s, t are replaced by a supernode v. Loops and multiple edges are deleted. Given a valid inequality $ax \leq b$ for $\overline{P}$ with complete graphs the corresponding lifted inequality $a'x \leq b'$ w.l.o.g. is defined as $a'_{st} = 0$, $a'_{sn} = a_{vn}$ and $a'_{tn} = 0$ for all nodes that are neighbours of s, t and v.

It is easy to see that for complete graphs the shrinking procedure is equivalent to the orthogonal projection that we use. Indeed, in the notation introduced above, a supernode v may be replaced by either s or t. Thus, when lifting an inequality the coefficients of variables x_{st} and x_{tn} are zero-lifted.

For the equicut problem, we can solve the target cut linear programs for subgraphs with up to 20 nodes within reasonable time. Therefore if we solve graphs with more than $n = 40$ nodes the subgraph induced by our projection always has

less than $\lceil \frac{n}{2} \rceil$ nodes. Thus, for instances of interesting sizes the projection of the equicut polytope is again a cut polytope without any restrictions on the size of the shores.

Next, we describe how we choose the nodes of the subgraphs used for projection. In order to separate a valid inequality the projected fractional solution needs to be infeasible. It is well known, that the cut polytope is a very symmetric object, i.e. the structure of inequalities valid with equality is the same at each vertex of the polytope. Furthermore the barycenter of the (equi)cut polytope is the vector $m = (0.5, 0.5, \ldots, 0.5)$ and the closer a fractional solution $\tilde{x}$ is to m, the less likely it is to be violated. Therefore given the fractional value $\tilde{x}_e$ of an edge e we assume that the value $b_e = |0.5 - \tilde{x}_e|$ correlates to the probability that the variable x_e contributes to a violated inequality, and we assume that these probabilities can be cumulated as $b_u = \sum_{e \in \delta(u)} b_e$ at node u. Unfortunately choosing the first k nodes with the largest values of b_u may as well result in a subgraph where the fractional solution $\tilde{x}$ is integral, hence given an integer feasible solution there is no violated inequality. Consequently the more fractional the solution $\tilde{x}$ is, the more likely it will yield a violated inequality. Further results on the projection we used as well as on the performance of target cut separation are described in Sect. 4.

3.2 Primal Heuristic

Given a fractional optimum solution $x^* \in \mathbb{R}^m$ of the current LP-relaxation, primal heuristics round x^* to a feasible solution that is hopefully better than the best one known to date. For the maximum cut problem, a primal heuristic works as follows [5]. A cut is given by a spanning tree where each edge is either a cut or a non-cut edge according to the corresponding value of the solution. The 'most decided' edges are used if the spanning tree in G is minimum with respect to the weights $w_e = \min(x_e^*, 1 - x_e^*)$ on the edges $e \in E$. The LP-values on the tree are then rounded appropriately, and the corresponding cut is returned. A minimum spanning tree may be computed by Prim's algorithm in $O(|E| + |V| \log |V|)$ [30].

For the equicut problem, a cut has to additionally satisfy the cardinality constraint on the shore size. We thus adapt the above greedy approach using Prim's algorithm. In each step, several trees are combined to larger trees until a spanning tree arises. Each of these trees induces a cut in the subgraph induced by its nodes, which we will call partial cuts.

Furthermore we have to make sure that in each iteration it is possible to combine these partial cuts to an equicut. We will call such a set of partial cuts compatible, and incompatible otherwise. Let a_i and b_i denote the size of the shores induced by the partial cut $\delta_i(S)$ and $d_i = |a_i - b_i|$ the absolute difference of the shore sizes. The subset-sum problem asks whether a subset $D' \subseteq \{d_i\}$ exists such that $\sum_{d_i \in D'} d_i = k$. Choosing $k = \frac{\sum d_i}{2}$, the partial cuts are compatible if the answer to the subset-sum problem is positive, and incompatible otherwise. In

general the subset-sum problem is NP-complete. Nevertheless, we use the well-known pseudo-polynomial algorithm due to Ibarra and Kim [21] for the knapsack problem which can be used to solve the subset-sum problem as well. As in our case the number of items and their weights are bounded above by $|V|$, the pseudo-polynomial algorithm yields an $O(|V|^2)$ algorithm.

Within Prim's algorithm, we make sure that whenever an edge is added to the spanning tree the partial cuts are compatible. This can be achieved by either skipping critical edges which lead to incompatible partial cuts or by repairing the partial cuts in such a way that the partial cuts become compatible. In either case we must never add edges that lead to a cycle. We also avoid using edges with weights w_e larger than a certain value r by temporarily removing them from G. From our experiments, $r = 0.25$ is a good choice.

Then according to Prim's algorithm all edges are iterated in order of increasing values w_e and edges which lead to cycles are skipped. Our method then uses three phases where edges are added and components of the graph are joined until a spanning tree is found. In the first phase we avoid repairing the partial cuts by skipping critical edges. In the second phase edges which lead to incompatible partial cuts are added. After adding an edge the partial cuts are then repaired greedily with respect to a minimum increase of the total edge weights w_e of the tree and such that no partial cut induces a shore of size greater than $\frac{|V|}{2}$. In the third phase we iteratively join the two largest components, which may occur from removing all edges with weights $w_e > r$. Again if a shore size exceeds $\frac{|V|}{2}$ nodes we have to repair that shore as in the second phase. Finally, we apply the Kernighan-Lin heuristic [26] to the solution to further improve the objective value.

We applied the above heuristic in our branch-and-cut algorithm. For all instances that were computed, whenever the root relaxation was strong enough to avoid branching, the optimal solution had been found by our primal heuristic.

4 Computational Results

In this section we evaluate our proposed branch-and-cut algorithm which is based upon a reimplementation of the algorithm presented in [8] using state-of-the-art tools. We used C++ and version 12.1 of CPLEX callable library as the branch-and-cut framework. For our experiments we used machines with Intel Xeon CPUs E5410 at 2.33 GHz. We use instances from [8] and instances from the physics application to evaluate the performance of our algorithm. The data for the instances we used and complete tables of computational results are available at [20].

In the determination of ground states in Coulomb glasses, we need to compute the minimum of the energy function:

$$H(q) = \sum_{i<j} p_{ij} q_i q_j + \sum_i c_i q_i. \tag{12}$$

The values $q_i \in \{-1, 1\}$ represent the positive or negative charges of sites i and are to be optimized. We are interested in charge-neutral systems, i.e. $\sum_i q_i = 0$. Those sites are located on a lattice and the values p_{ij} represent the pairwise interaction of sites i and j. We can use the same variable transformation as in Sect. 2.3 to obtain a quadratic unconstrained binary optimization (QUBO) problem. It was proved by de Simone [13] that QUBO is equivalent to the maximum cut problem. For this transformation the quadratic terms in the objective function are represented by edges in a graph. The linear terms are represented by edges connected to an artificial node s that is added to the graph. The Coulomb glass instances are generally defined on an even number of sites, hence the above transformation would yield a graph with an odd number of nodes. A complete graph K_{2n} is then obtained by adding another artificial node t. Further the constant term from the transformation to QUBO is represented by the weight of the edge st. Since we restrict to charge-neutral systems we have to find a minimum s-t equicut (cf. Sect. 2.1).

In order to improve the quality of the LP relaxation we used target cut separation. We found that a significant number of violated inequalities found by target cut separation were switched clique inequalities (3).

We use a greedy heuristic to separate switched clique inequalities that extends the algorithm described in [8] in a straight forward way. We start with the most violated triangle inequality which is also a switched clique inequality. Given a switched clique inequality for a clique $K_p = (V', E')$ and a cut $\delta(S)$ with $|S| \leq |\bar{S}|$, we iteratively compute switched clique inequalities for a clique K_{p+1} unless $p + 1$ exceeds a certain node limit k. Considering the switching operation we further improve the violation of the inequality in each iteration by iteratively switching pairs of nodes i and j if the violation of the switched clique inequality is increased.

In Table 2 we give the results for instances reported in [8] with our branch-and-cut algorithm including switched clique and target cut separation with different projections to subgraphs with 15 nodes. Further we give the number of subproblems reported in [8] as a reference.

The results suggest that target cut separation improves the LP relaxation for all projections and no choice of projection dominates the others in terms of the number of subproblems. Considering CPU times for most instances the overhead of target cut separation is moderate for the given size of projections but helps to reduce the number of subproblems. Comparing our results with those reported in [8] is very difficult since their reported computation times are for experiments carried out in the mid-1990s. Nevertheless we point out that our computation times are two orders of magnitude smaller on average, and more importantly, we need fewer subproblems. We conclude that our computation times outperform previous approaches based on LP relaxations. With respect to using SDP, the computation times reported in [25] for the branch-and-bound algorithm based on SDP relaxations seem to be better while our method needs fewer subproblems. In general LP relaxations can be solved much faster than SDP relaxations. Consequently the method described in [25] needs to solve fewer relaxations. Using equivalent SDP and LP relaxations we further

Table 2 Root bounds, number of subproblems and CPU time to solve instances reported in [8] without target cut separation (*N*), with random (*R*) and with greedy (*G*) projections. Number of branching nodes reported in [8] are given in columns (*B*). Results for instances with less than 40 nodes and instances that are not reported in [8] are omitted. Bounds are given as "*" if optimum was found in the root node

Instance	|V|	Opt	Root bound			No. of subs				Time (s)		
			N	R	G	N	R	G	B	N	R	G
rand/q0.90	40	63	*	*	*	1	1	1	1	4	5	5
rand/q0.80	40	199	*	*	*	1	1	1	1	73	86	93
rand/q0.30	40	1,056	1,052.01	*	*	9	1	1	5	161	226	256
rand/q0.20	40	1,238	1,235.16	*	*	7	1	1	3	131	174	205
rand/q0.20	40	1,238	1,235.16	*	*	7	1	1	3	131	174	205
rand/q0.10	40	1,425	1,420.38	*	*	9	1	1	7	99	173	169
rand/q0.00	40	1,606	*	*	*	1	1	1	1	44	48	54
rand/c0.90	50	122	*	*	*	1	1	1	7	6	7	7
rand/c0.80	50	368	359.71	*	*	5	1	1	11	478	1,150	1,212
rand/c0.70	50	603	585.54	*	*	3	1	1	1	1,203	2,864	3,159
rand/c0.30	50	1,658	1,510.82	*	*	3	1	1	9	815	1,611	1,554
rand/c0.10	50	2,226	2,090.9	*	*	3	1	1	7	674	986	870
rand/c0.00	50	2,520	2,510.78	*	*	7	1	1	5	420	552	549
rand/c2.90	52	123	*	*	*	1	1	1	7	8	10	10
rand/c4.90	54	160	*	*	*	1	1	1	17	32	36	40
rand/c6.90	56	177	*	*	*	1	1	1	17	35	38	45
rand/c8.90	58	226	216.04	216.55	216.54	3	2	2	n.a.	6,698	12,446	8,956
rand/s0.90	60	238	235.97	*	*	3	1	1	7	240	341	451
reti/5 × 8	40	18	*	*	*	1	1	1	7	2	2	3
reti/3 × 14	42	10	*	*	*	1	1	1	5	12	23	30
reti/5 × 10	50	22	*	*	*	1	1	1	3	10	23	34
reti/6 × 10	60	28	*	*	*	1	1	1	31	101	142	151
reti/7 × 10	70	23	*	*	*	1	1	1	n.a.	140	145	171
tori/21 × 2	42	9	*	*	*	1	1	1	3	3	18	24
tori/23 × 2	46	9	*	*	*	1	1	1	3	60	63	69
tori/4 × 12	48	24	*	*	*	1	1	1	5	4	18	24
tori/5 × 10	50	33	*	*	*	1	1	1	13	12	28	37
tori/10 × 6	60	42	41.81	*	*	3	1	1	3	97	123	100
tori/7 × 10	70	45	*	*	*	1	1	1	33	17	19	23
misti/10 × 4 m	40	436	*	*	*	1	1	1	1	3	13	28
misti/5 × 10 m	50	670	*	*	*	1	1	1	5	6	8	7
misti/13 × 4 m	52	721	*	*	*	1	1	1	5	21	25	26
misti/4 × 13 m	52	721	*	*	*	1	1	1	7	27	28	35
misti/9 × 6 m	54	792	*	*	*	1	1	1	n.a.	16	18	22
misti/10 × 6 m	60	954	*	*	*	1	1	1	9	22	38	46
misti/10 × 7 m	70	1,288	*	*	*	1	1	1	13	60	84	110
negative/q0.n.70	40	−298	*	*	*	1	1	1	1	29	29	35
negative/q0.n.50	40	−389	*	*	*	1	1	1	1	59	64	77
negative/q0.n.40	40	−450	*	*	*	1	1	1	1	9	10	11
negative/q0.n.00	40	−471	−474.02	*	*	3	1	1	1	192	254	298

(continued)

Table 2 (continued)

negative/c0.n.00	50	−829	−1,069.94	−954.81	*	3	2	1	5	2,591	4,717	4,027
negative/s0.n.80	60	−465	*	*	*	1	1	1	1	45	52	53
negative/o0.n.80	80	−690	−725.88	*	*	n.a.	1	1	n.a.	n.a.	40,287	37,562
real/ma.i	54	2	*	*	*	1	1	1	29	6	7	8
real/me.i	60	3	*	*	*	1	1	1	37	7	7	9
real/m6.i	70	7	*	*	*	1	1	1	55	215	251	242
real/mb.i	74	4	*	*	*	1	1	1	33	889	1,119	1,153
real/mc.i	74	6	*	*	*	1	1	1	53	92	129	149
real/md.i	80	4	*	*	*	1	1	1	57	762	781	884
real/mf.i	90	4	3.58	*	3.58	2	1	4	47	470	803	687
real/m1.i	100	4	*	*	*	1	1	1	101	153	171	196
real/m8.i	148	7	*	*	*	1	1	1	n.a.	543	632	713

observed that the bounds are very similar (cf. Sect. 2.3). Therefore we suspect that better computation times can be explained by stronger inequalities that were separated due to the different fractional points given by the interior-point method.

In Table 3 we give the bounds at the root node, the number of subproblems and the CPU time needed by our branch-and-cut algorithm with and without switched clique separation for instances from [8], including some instances that were not reported in [8]. Further we did not apply target cut separation. The results suggest that switched clique separation improves the bounds significantly.

Furthermore, as switched clique inequalities and target cut separation significantly improve the LP relaxations, we were able to solve larger instances than those reported in [8]. We illustrate this by presenting our results on Coulomb glass instances with up to 258 nodes. Table 4 gives the number of subproblems and CPU time required. For some instances that could not be solved, we report the gaps after 3 days of computation.

5 Conclusions

Our experimental results support the conclusion that the proposed branch-and-cut algorithm based on a linear relaxation with additional switched clique inequalities and target cut separation is able to efficiently solve medium-sized instances of equicut and larger instances of Coulomb glasses. This new algorithm thus contributes to the practical solution of equicut problems.

Most inequalities separated by target cut separation are hypermetric inequalities which is a very general class of inequalities [15]. It would be interesting to find heuristics to separate more specific hypermetric inequalities. Therefore target cut separation could be used to classify important inequalities. Further improving the performance of target cut separation would allow the use of larger subgraphs for the projection. It would also be interesting to improve the projections to find violated inequalities more efficiently.

Table 3 Root bounds, number of subproblems and CPU time to solve instances from [8] with and without switched clique (*SC*) separation. Results for instances with less than 52 nodes are omitted. Bounds are given as "*" if optimum was found in the root node. No target cut separation was applied

Instance	$|V|$	Opt	Root bound		No. of subs		Time (s)	
			No SC	SC	No SC	SC	No SC	SC
rand/c2.90	52	123	*	*	1	1	9	8
rand/c4.90	54	160	156.68	*	3	1	30	33
rand/c6.90	56	177	176.33	*	2	1	28	30
rand/c8.90	58	226	205.1	215.05	108	3	234	3, 200
rand/s0.90	60	238	229.45	*	8	1	72	276
reti/13 × 4	52	20	15.7	*	7	1	29	45
reti/6 × 10	60	28	26.43	*	6	1	80	101
reti/10 × 6	60	19	*	*	1	1	20	19
reti/7 × 10	70	23	21.58	*	15	1	252	198
tori/13 × 4	52	20	15.7	*	7	1	32	45
tori/6 × 10	60	35	*	*	1	1	7	7
tori/10 × 6	60	42	41.81	41.81	3	3	107	97
tori/7 × 10	70	45	*	*	1	1	18	17
tori/10 × 8t	80	43	42.99	42.99	2	2	450	429
misti/4 × 13 m	52	721	*	*	1	1	28	26
misti/13 × 4 m	52	721	*	*	1	1	25	21
misti/9 × 6 m	54	792	*	*	1	1	17	16
misti/10 × 6 m	60	954	*	*	1	1	22	23
misti/10 × 7 m	70	1, 288	*	*	1	1	63	58
negative/s0.n.80	60	-465	*	*	1	1	53	45
negative/tt0.n.80	70	-550	−579.08	−579.08	56	3	1, 534	22, 170
negative/o0.n.80	80	-690	−725.88	−725.88	86	3	5, 232	31, 124
real/ma.i	54	2	*	*	1	1	6	6
real/me.i	60	3	*	*	1	1	8	7
real/m6.i	70	7	*	*	1	1	216	198
real/mc.i	74	6	5.87	*	3	1	116	89
real/mb.i	74	4	3.26	*	6	1	243	752
real/md.i	80	4	3.46	3.71	3	2	347	508
real/mf.i	90	4	3.58	3.58	4	2	470	423
real/m1.i	100	4	*	*	1	1	163	140
real/m8.i	148	7	*	*	1	1	551	604

Table 4 Mean values for the relative gap at the root node, number of subproblems and computation times for large instances from the Coulomb glass model with two- and three-dimensional grid graphs. Each class of instances is characterized by the length of the grids L and its dimension. For each class of instances we present results averaged over five randomly generated instances with different random seeds to compute the values c_i of the local field. Due to our transformation the number of nodes is given as $|V| = Size + 2$

$|V|$	Size	Root gap (%)	No. of subs	Time (h)
66	8^2	0	1	0.01
66	4^3	0.02	1.4	0.01
102	10^2	0	1	0.12
146	12^2	0.14	1	1.7
198	14^2	0.01	1	24.81
218	6^3	0.9	14.6	58.86
258	16^2	4.06	n.a.	>72

References

1. Applegate, D., Bixby, R.E., Chvátal, V., Cook, W.J. et al.: TSP cuts which do not conform to the template paradigm. Comput. Comb. Optim. **2241**, 261–303 (2001)
2. Armbruster, M., Fügenschuh, M., Helmberg, C., Martin, A.: A comparative study of linear and semidefinite branch-and-cut methods for solving the minimum graph bisection problem. In: Integer Programming and Combinatorial Optimization, IPCO'08, Bertinoro, pp. 112–124 (2008)
3. Armbruster, M., Fügenschuh, M., Helmberg, C., Martin, A.: LP and SDP branch-and-cut algorithms for the minimum graph bisection problem: a computational comparison. Math. Program. Comput. **4**, 275–306 (2012)
4. Barahona, F., Grötschel, M., Mahjoub, A.R.: Facets of the bipartite subgraph polytope. Math. Oper. Res. **10**(2), 340–358 (1985)
5. Barahona, F., Grötschel, M., Jünger, M., Reinelt, G.: An application of combinatorial optimization to statistical physics and circuit layout design. Oper. Res. **36**(3), 493–513 (1988)
6. Billionnet, A., Elloumi, S., Plateau, M.C.: Quadratic convex reformulation: a computational study of the graph bisection problem. Technical report, Laboratoire CEDRIC (2005)
7. Borchers, B.: CSDP, A C library for semidefinite programming. Optim. Methods Softw. **11**(1–4), 613–623 (1999). Special Issue: Interior Point Methods
8. Brunetta, L., Conforti, M., Rinaldi, G.: A branch-and-cut algorithm for the equicut problem. Math. Program. B **78**(2), 243–263 (1997)
9. Buchheim, C., Liers, F., Oswald, M.: Local cuts revisited. Oper. Res. Lett. **36**(4), 430–433 (2008)
10. Conforti, M., Rao, M.R., Sassano, A.: The equipartition polytope. I: formulations, dimension and basic facets. Math. Program. A **49**, 49–70 (1990)
11. Conforti, M., Rao, M.R., Sassano, A.: The equipartition polytope. II: valid inequalities and facets. Math. Program. A **49**, 71–90 (1990)
12. CPLEX® Callable Library version 12.1 – C API Reference Manual. ftp://ftp.software.ibm.com/software/websphere/ilog/docs/optimization/cplex/refcallablelibrary.pdf (2009)
13. de Simone, C.: The cut polytope and the boolean quadric polytope. Discret. Math. **79**(1), 71–75 (1990)
14. de Souza, C.C., Laurent, M.: Some new classes of facets for the equicut polytope. Discret. Appl. Math. **62**(1–3), 167–191 (1995)

15. Deza, M.M., Laurent, M.: Geometry of Cuts and Metrics, 1st edn. Springer, New York (1997)
16. Edmonds, J., Johnson, E.L.: Matching: a well-solved class of integer linear programs. In: Guy, R. (ed.) Combinatorial Structures and Their Applications, pp. 89–92. Gordon and Breach, New York (1970)
17. Ferreira, C., Martin, A., de Souza, C., Weismantel, R., Wolsey, L.: Formulations and valid inequalities for the node capacitated graph partitioning problem. Math. Program. **74**, 247–266 (1996)
18. Ferreira, C.E., Martin, A., de Souza, C.C., Weismantel, R., Wolsey, L.A.: The node capacitated graph partitioning problem: a computational study. Math. Program. **81**, 229–256 (1998)
19. Frieze, A.M., Jerrum, M.: Improved approximation algorithms for max k-cut and max bisection. In: Proceedings of the 4th International IPCO Conference on Integer Programming and Combinatorial Optimization, Copenhagen, pp. 1–13. Springer, London (1995)
20. Anjos, M.F., Liers, F., Pardella, G., and Schmutzer, A.: Instances and computational results. (2012) http://cophy.informatik.uni-koeln.de/eng_eq_ref.html
21. Ibarra, O.H., Kim, C.E.: Fast approximation algorithms for the knapsack and sum of subset problems. J. ACM **22**, 463–468 (1975)
22. Johnson, E.L., Mehrotra, A., Nemhauser, G.L.: Min-cut clustering. Math. Program. **62**, 133–151 (1993)
23. Jünger, M., Reinelt, G., Rinaldi, G.: Lifting and separation procedures for the cut polytope. Technical report, IASI-CNR, R. 11–14 (2011)
24. Karisch, S.E., Rendl, F.: Semidefinite programming and graph equipartition. In: Pardalos, P.M., Wolkowicz, H. (eds.) Topics in Semidefinite and Interior-Point Methods, pp. 77–95. AMS, Providence (1998)
25. Karisch, S.E., Rendl, F., Clausen, J.: Solving graph bisection problems with semidefinite programming. INFORMS J. Comput. **12**(3), 177–191 (2000)
26. Kernighan, B., Lin, S.: An efficient heuristic procedure for partitioning graphs. Bell Syst. Tech. J. **49**, 291–307 (1970)
27. Klerk, E., Pasechnik, D., Sotirov, R., Dobre, C.: On semidefinite programming relaxations of maximum k-section. Math. Program. **136**(2):1–26 (2012)
28. Letchford, A.N., Reinelt, G., Theis, D.O.: A faster exact separation algorithm for blossom inequalities. In: IPCO'04, New York, pp. 196–205 (2004)
29. Mehrotra, A.: Cardinality constrained boolean quadratic polytope. Discret. Appl. Math. **79**(1–3), 137–154 (1997)
30. Prim, R.C.: Shortest connection networks and some generalizations. Bell Syst. Tech. J. **36**, 1389–1401 (1957)
31. Rendl, F., Rinaldi, G., Wiegele, A.: Solving max-cut to optimality by intersecting semidefinite and polyhedral relaxations. Math. Program. **121**(2), 307–355 (2010)
32. Sotirov, R.: An Efficient Semidefinite Programming Relaxation for the Graph Partition Problem. INFORMS J. Comput. (to appear)

An Approach to the Dodecahedral Conjecture Based on Bounds for Spherical Codes

Kurt M. Anstreicher

Abstract The dodecahedral conjecture states that in a packing of unit spheres in $\Re^3$, the Voronoi cell of minimum possible volume is a regular dodecahedron with inradius one. The conjecture was first stated by L. Fejes Tóth in 1943, and was finally proved by Hales and McLaughlin over 50 years later using techniques developed by Hales for his proof of the Kepler conjecture. In 1964, Fejes Tóth described an approach that would lead to a complete proof of the dodecahedral conjecture if a key inequality were established. We describe a connection between the key inequality required to complete Fejes Tóth's proof and bounds for spherical codes and show how recently developed strengthened bounds for spherical codes may make it possible to complete Fejes Tóth's proof.

Key words Dodecahedral conjecture • Kepler conjecture • Spherical codes • Delsarte bound • Semidefinite programming

Subject Classifications: 52C17, 90C22, 90C26

1   Introduction

The dodecahedral conjecture states that in a packing of unit spheres in $\Re^3$, the Voronoi (or Dirichlet) cell of minimum possible volume is a regular dodecahedron with inradius one. More precisely, let $\bar{x}_i$, $i = 1, \ldots, m$ be points in $\Re^3$ with $\|\bar{x}_i\| \geq 1$ for each i, and $\|\bar{x}_i - \bar{x}_j\| \geq 1$ for all $i \neq j$. Then the points $2\bar{x}_i$ can be taken to be the centers of m non-overlapping spheres of radius one which also do not overlap a

K.M. Anstreicher (✉)
Department of Management Sciences, University of Iowa, S210 PBB,
Iowa City, IA 52242, USA
e-mail: kurt-anstreicher@uiowa.edu

sphere of radius one centered at $x_0 = 0$. The Voronoi cell associated with x_0 is then

$$V(\bar{x}_1, \ldots, \bar{x}_m) = \{x : \bar{x}_i^T x \leq \|\bar{x}_i\|^2, \ i = 1, \ldots, m\}.$$

Let $D \subset \Re^3$ denote a regular dodecahedron of inradius one, and Vol$(\cdot)$ denote volume in $\Re^3$.

The Dodecahedral Conjecture [5, 6] *Let $\bar{x}_i \in \Re^3$, $i = 1, \ldots, m$ with $\|\bar{x}_i\| \geq 1$ for each i, and $\|\bar{x}_i - \bar{x}_j\| \geq 1$ for all $i \neq j$. Then* Vol$(V(\bar{x}_1, \ldots, \bar{x}_m)) \geq$ Vol(D).

The dodecahedral conjecture was stated by L. Fejes Tóth in 1943 [5]. Fejes Tóth's interest in the conjecture was to obtain a good upper bound on the maximal density of a sphere packing in $\Re^3$. In particular, the dodecahedral conjecture implies an upper bound of approximately 0.7545, compared to the maximal density of approximately 0.7405 asserted by the Kepler conjecture. Hales and McLaughlin [9] describe a complete proof of the dodecahedral conjecture based on techniques developed by Hales for his proof of the Kepler conjecture. The proof of [9] is believed to be correct, but is difficult to verify due to the many cases and extensive computations required.

Let $R_D = \sqrt{3} \tan 36° \approx 1.2584$ be the radius of a sphere that circumscribes D, and let $\mathscr{S}_D = \{x \in \Re^3 : \|x\| \leq R_D\}$. Fejes Tóth's 1943 paper contains a proof of the dodecahedral conjecture under the assumption that there are at most 12 i such that $\bar{x}_i \in \mathscr{S}_D$. In [6, pp. 296–298] Fejes Tóth restates the dodecahedral conjecture and describes an approach that would lead to a complete proof if a key inequality were established. The details of this approach are described in the next section. In Sect. 3 we describe a connection between the key inequality required to complete Fejes Tóth's proof and bounds for spherical codes. Using constraints from the well-known Delsarte bound for spherical codes, we are able to prove the key inequality for some but not all of the required possible cases. We then consider applying additional constraints from recently described semidefinite programming (SDP) bounds for spherical codes [2]. The use of the SDP constraints improves our bounds, but is not sufficient to eliminate more cases than were already eliminated using the linear programming constraints associated with the Delsarte bound.

In recent work, Hales [7] announced a proof of the "strong" dodecahedral conjecture, which is the original dodecahedral conjecture with surface area replacing volume throughout. The proof methodology of [7] also utilizes Fejes Tóth's key inequality, which is apparently the basis for a new computational proof of the Kepler conjecture in [8]. These recent developments suggest that continued efforts to provide a direct proof of the key inequality remain a very interesting topic for further research.

2 Fejes Tóth's Proof

In this section we describe the proof of the dodecahedral conjecture suggested in [6]. The first ingredient is a strengthened version of the result proved in [5].

Theorem 1 ([6, p. 265]). *Let* $\hat{x}_i$, $i = 1, \ldots, m$ *be points in* $\Re^3$ *with* $\|\hat{x}_i\| \geq 1$ *for each* i. *If* $m \leq 12$, *then* $\mathrm{Vol}(V(\hat{x}_1, \ldots, \hat{x}_m) \cap \mathscr{S}_D) \geq \mathrm{Vol}(D)$.

Note that in Theorem 1 it is *not* assumed that the points satisfy $\|\hat{x}_i - \hat{x}_j\| \geq 1$, $i \neq j$. Also, the assumption that $\|\hat{x}_i\| < R_D$ for each i could be added, since if $\|\hat{x}_i\| \geq R_D$ the constraint $\hat{x}_i^T x \leq \|\hat{x}_i\|^2$ in the definition of $V(\hat{x}_1, \ldots, \hat{x}_m)$ does not eliminate any points in $\mathscr{S}_D$.

The second important component of the argument suggested in [6] is a "point adjustment procedure" that facilitates the use of Theorem 1 when $m > 12$. For the Voronoi cell $V(\hat{x}_1, \ldots, \hat{x}_m)$, let $F_i(\hat{x}_1, \ldots, \hat{x}_m)$ be the face of $V(\hat{x}_1, \ldots, \hat{x}_m) \cap \mathscr{S}_D$ corresponding to the points with $\hat{x}_i^T x = \|\hat{x}_i\|^2$ (it is possible that $F_i(\hat{x}_1, \ldots, \hat{x}_m) = \emptyset$).

Point Adjustment Procedure

Step 0. Input $\bar{x}_i$, $1 \leq \|\bar{x}_i\| \leq R_D$, $i = 1, \ldots, m$ with $m > 12$ and $\|\bar{x}_i - \bar{x}_j\| \geq 1$, $i \neq j$. Let $\hat{x}_i = \bar{x}_i$, $i = 1, \ldots, m$.

Step 1. If $|\{i : 1 < \|\hat{x}_i\| < R_D\}| < 2$ then go to Step 3. Otherwise choose $j \neq k$ such that $1 < \|\hat{x}_j\| < R_D$, $1 < \|\hat{x}_k\| < R_D$, and the surface area of $F_j(\hat{x}_1, \ldots, \hat{x}_m)$ is less than or equal to that of $F_k(\hat{x}_1, \ldots, \hat{x}_m)$.

Step 2. Let $\delta = \min\{R_D - \|\hat{x}_j\|, \|\hat{x}_k\| - 1\}$, and

$$\hat{x}_j \leftarrow (\|\hat{x}_j\| + \delta)\frac{\hat{x}_j}{\|\hat{x}_j\|}, \quad \hat{x}_k \leftarrow (\|\hat{x}_k\| - \delta)\frac{\hat{x}_k}{\|\hat{x}_k\|}.$$

Go to Step 1.

Step 3. Output $\hat{x}_i$, $i = 1, \ldots, m$.

As pointed out in [6], $R_D < \sqrt{2}$ implies that the area of $F_i(\lambda_1 x_1, \ldots, \lambda_m x_m)$ is monotone decreasing in λ_i. It follows that the adjustment in Step 2 leaves $\sum_{i=1}^{m} \|\hat{x}_i\|$ unchanged, while $\mathrm{Vol}(V(\hat{x}_1, \ldots, \hat{x}_m) \cap \mathscr{S}_D)$ is nonincreasing.[1] Note that the adjustment in Step 2 is executed at most $m - 1$ times, since each adjustment decreases $|\{i : 1 < \|\hat{x}_i\| < R_D\}|$ by at least 1. Then Theorem 1 can be applied if the $\hat{x}_i$ output by the procedure have at most 12 i with $\|\hat{x}_i\| < R_D$. (Note that the output points $\hat{x}_i$ will generally *not* satisfy $\|\hat{x}_i - \hat{x}_j\| \geq 1$, $i \neq j$, but this assumption is not required in Theorem 1.) This will be the case if the input points $\bar{x}_i$ satisfy

$$\sum_{i=1}^{m} \|\bar{x}_i\| \geq 12 + (m - 12)R_D. \tag{1}$$

To see this, note that there is at most one $\hat{x}_j$ with $1 < \|\hat{x}_j\| < R_D$, so if $|i : \|\hat{x}_i\| = 1| \leq 11$ there is nothing to show. Assume on the other hand that $\|\hat{x}_i\| = 1$, $i = 1, \ldots, 12$. Then (1) and the fact that $\|\hat{x}_i\| \leq R_D$ for each i together imply

---

[1] Fejes Tóth does not explicitly consider the possibility that the two faces $F_j(\hat{x}_1, \ldots, \hat{x}_m)$ and $F_k(\hat{x}_1, \ldots, \hat{x}_m)$ intersect. However in this case it is easy to see that the increase in volume that results from increasing $\hat{x}_j$ is even less than if the faces do not intersect.

$$\sum_{i=1}^{m} \|\hat{x}_i\| = \sum_{i=1}^{m} \|\bar{x}_i\| \geq 12 + (m - 12)R_D$$

$$12 + \sum_{i=13}^{m} \|\hat{x}_i\| \geq 12 + (m - 12)R_D$$

$$(m - 12)R_D \geq \sum_{i=13}^{m} \|\hat{x}_i\| \geq (m - 12)R_D,$$

implying that $\|\hat{x}_i\| = R_D$ for $i = 13, \ldots, m$.

A complete proof of the dodecahedral conjecture thus requires only a proof that (1) holds for any $\bar{x}_i$, $i = 1, \ldots, m$ with $1 \leq \|\bar{x}_i\| \leq R_D$ for each i, and $\|x_i - x_j\| \geq 1$ for all $i \neq j$. Unfortunately Fejes Tóth was unable to prove (1), even though all evidence suggests that (1) actually holds with R_D replaced by the larger constant $7/\sqrt{27} \approx 1.347$ [6].[2] Remarkably, the key inequality also appears in Hales' [7] recent paper that describes a proof of the strong dodecahedral conjecture. In [7] the key inequality is labeled as the inequality L_{12}, and is written with the value of $R_D \approx 1.2584$ rounded up to $h_0 := 1.26$. (It is also assumed that the initial points satisfy $1 \leq \|\bar{x}_i\| \leq h_0$.) A computational proof of the inequality L_{12} is apparently the basis for a new proof of the Kepler conjecture in [8].

3   Applying Bounds for Spherical Codes

We now describe an approach to proving the key inequality (1) based on bounds for spherical codes. A set $\mathscr{C} = \{x_i\}_{i=1}^{m} \subset \mathfrak{R}^3$ is called a spherical z-code if $\|x_i\| = 1$ for each i, and $x_i^T x_j \leq z$ for all $i \neq j$. For example, a packing of unit spheres that all touch (or "kiss") a unit sphere centered at the origin generates a spherical $1/2$-code.

To begin we establish that if R is sufficiently small and $\{\bar{x}_i\}_{i=1}^{m}$ are points with $1 \leq \|\bar{x}_i\| \leq R$ for each i and $\|\bar{x}_i - \bar{x}_j\| \geq 1$ for all $i \neq j$, then the normalized points $x_i = \bar{x}/\|\bar{x}_i\|$ form a z-code for a suitable value of z.

Lemma 1. *Suppose that* $1 \leq \|\bar{x}_i\| \leq R$, $i = 1, \ldots, m$, *where* $1 \leq R \leq \frac{1+\sqrt{5}}{2}$ *and* $\|\bar{x}_i - \bar{x}_j\| \geq 1$ *for all* $i \neq j$. *Let* $x_i = \bar{x}_i/\|\bar{x}_i\|$, $i = 1, \ldots, m$. *Then* $x_i^T x_j \leq 1 - \frac{1}{2R^2}$ *for all* $i \neq j$.

Proof. The case $R = 1$ is trivial. For $R > 1$ and $i \neq j$, consider the problem

---

[2] Note that (1) implies that for $m = 13$, if $\|\bar{x}_i\| = 1$ for $i = 1, \ldots, 12$, then $\|\bar{x}_{13}\| \geq R_D$. It has been incorrectly stated that the latter implication is the "missing ingredient" in Fejes Tóth's proof. In fact the stronger statement (1) is exactly what is required.

$$\max \|\lambda_i x_i - \lambda_j x_j\|^2$$
$$\text{s.t. } 1 \le \lambda_i \le R, \ 1 \le \lambda_j \le R. \tag{2}$$

The objective in (2) is convex, so the solution lies at an extreme point of the feasible region. Letting $s := x_i^T x_j$, the value of the objective at the extreme points is:

$$\lambda_i^2 + \lambda_j^2 - 2s\lambda_i\lambda_j = \begin{cases} 2(1-s) & \text{if } \lambda_i = \lambda_j = 1, \\ 1 + R^2 - 2Rs & \text{if } \lambda_i = 1, \ \lambda_j = R \text{ or } \lambda_i = R, \ \lambda_j = 1, \\ 2R^2(1-s) & \text{if } \lambda_i = \lambda_j = R. \end{cases}$$

Obviously the maximum cannot occur at $\lambda_i = \lambda_j = 1$. Note that the solution value in (2) is at least one, from the assumption that $\|\bar{x}_i - \bar{x}_j\| \ge 1$. It follows that if the max occurs at $\lambda_i = 1$, $\lambda_j = R$ (or equivalently $\lambda_i = R$, $\lambda_j = 1$) then $1 + R^2 - 2Rs \ge 1$, which is equivalent to $s \le \frac{R}{2}$. Similarly, if the max occurs at $\lambda_i = \lambda_j = R$, then $2R^2(1-s) \ge 1$, which is equivalent to $s \le 1 - \frac{1}{2R^2}$. Finally it is easy to verify that $\frac{R}{2} \le 1 - \frac{1}{2R^2}$ for $1 \le R \le \frac{1+\sqrt{5}}{2}$. □

Next, for $x_i \ne x_j$ with $\|x_i\| = \|x_j\| = 1$, consider the problem

$$\min \lambda_i + \lambda_j$$
$$\text{s.t. } \|\lambda_i x_i - \lambda_j x_j\| \ge 1 \tag{3}$$
$$1 \le \lambda_i \le R, \ 1 \le \lambda_j \le R.$$

Note that by Lemma 1 and its proof, if $R \le \frac{1+\sqrt{5}}{2}$ then (3) is feasible if and only if $x_i^T x_j \le 1 - \frac{1}{2R^2}$. The next result gives a complete characterization of the solution value in (3).

Theorem 2. *Let* $1 \le R \le \frac{1+\sqrt{5}}{2}$, $\|x_i\| = \|x_j\| = 1$ *and* $s = x_i^T x_j \le 1 - \frac{1}{2R^2}$. *Then the solution value in problem (3) is* $\lambda_i^* + \lambda_j^* = f(s, R)$, *where*

$$f(s, R) = \begin{cases} 2 & \text{if } s \le \frac{1}{2}, \\ 1 + 2s & \text{if } \frac{1}{2} \le s \le \frac{R}{2}, \\ R(1+s) + \sqrt{1 - R^2(1-s^2)} & \text{if } \frac{R}{2} \le s \le 1 - \frac{1}{2R^2}. \end{cases}$$

Proof. The case of $s \le 0.5$ is trivial, so assume that $s > 0.5$ and the objective in (3) attains a value $\lambda_i + \lambda_j = c$, where $2 < c \le 2R$. Since the constraint $\|\lambda_i x_i - \lambda_j x_j\| \ge 1$ is equivalent to $(\lambda_i + \lambda_j)^2 \ge 1 + 2\lambda_i\lambda_j(1 + s)$, this implies that we must have

$$c^2 \ge 1 + 2(1 + s)\lambda_i\lambda_j .$$

To find the minimum possible value of c we are thus led to consider the problem

$$\min \; \lambda_i \lambda_j$$
$$\text{s.t.} \; \lambda_i + \lambda_j = c, \tag{4}$$
$$1 \le \lambda_i \le R, \; 1 \le \lambda_i \le R.$$

The objective in (4) can be written in the form $\lambda_i(c - \lambda_i)$, which is a concave function, so the solution of (4) must occur at an extreme point of the feasible region. There are two possibilities for the form of such an extreme point, depending on the value of c.

Case 1: $c \le 1 + R$. In this case the extreme points of (4) have (λ_i, λ_j) equal to $(1, c - 1)$ and $(c - 1, 1)$, both of which have $\lambda_i \lambda_j = c - 1$. To find the minimum possible value of $c = \lambda_i + \lambda_j$ in (3), we must find the minimum $c \ge 2$ such that

$$c^2 \ge 1 + 2(1 + s)(c - 1),$$

which is easily determined to be $c = 1 + 2s$. Since by assumption $c \le 1 + R$, this solution applies whenever $1 + 2s \le 1 + R$, or $s \le \frac{R}{2}$.

Case 2: $c \ge 1 + R$. In this case the extreme points of (4) have (λ_i, λ_j) equal to $(R, c - R)$ and $(c - R, R)$, both of which have $\lambda_i \lambda_j = R(c - R)$. To find the minimum possible value of $c = \lambda_i + \lambda_j$ in (3), we must find the minimum $c \ge 2$ such that

$$c^2 \ge 1 + 2(1 + s)R(c - R),$$

which is easily determined to be $c = R(1 + s) + \sqrt{1 - R^2(1 - s^2)}$. Since by assumption $c \ge 1 + R$, this solution applies whenever $R(1 + s) + \sqrt{1 - R^2(1 - s^2)} \ge 1 + R$, which is equivalent to $s \ge \frac{R}{2}$. $\qquad\square$

In Fig. 1 we plot $f(s, R_D)$ for $\frac{1}{2} \le s \le 1 - \frac{1}{2R_D^2}$. It is evident from the figure, and is easy to prove, that $f(s, R_D)$ is concave in the interval $\frac{R_D}{2} \le s \le 1 - \frac{1}{2R_D^2}$.

Now assume that $m > 12$, $1 \le \|\bar{x}_i\| \le R_D$, $i = 1, \ldots m$, and $\|\bar{x}_i - \bar{x}_j\| \ge 1$ for all $i \ne j$. Let $\lambda_i = \|\bar{x}_i\|$ and $x_i = (1/\lambda_i)\bar{x}_i$, $i = 1, \ldots, m$. Our goal is to prove (1), which can be written as

$$\sum_{i=1}^{m} \lambda_i \ge 12 + (m - 12)R_D. \tag{5}$$

Define $N_i = |\{j \ne i : x_i^T x_j \ge 0.5\}|$ to be the number of "close neighbors" of x_i, $i = 1, \ldots, m$ and $\mathcal{N} = \{(i, j), \; i \ne j : x_i^T x_j \ge 0.5\}$. Note that $(i, j) \in \mathcal{N} \iff (j, i) \in \mathcal{N}$, and $|\mathcal{N}| = \sum_{i=1}^{m} N_i$. Moreover we have

$$\sum_{(i,j) \in \mathcal{N}} (\lambda_i + \lambda_j - 2) = \sum_{(i,j) \in \mathcal{N}} (\lambda_i - 1) + (\lambda_j - 1) = 2\sum_{i=1}^{m} N_i(\lambda_i - 1).$$
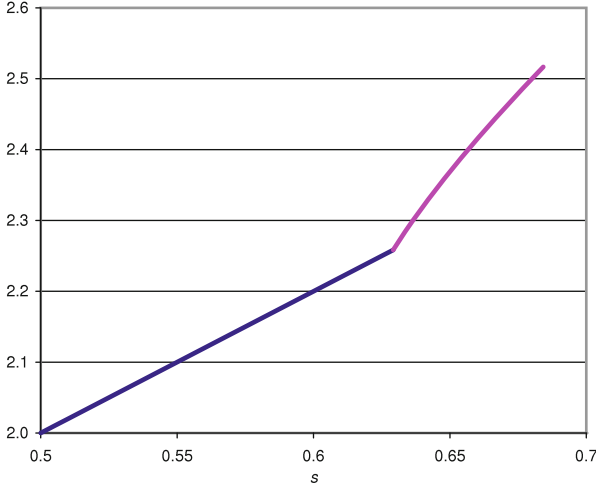
Fig. 1 Function $f(s, R_D)$ from Theorem 2

Applying Theorem 2, it follows that

$$2\sum_{i=1}^{m} N_i(\lambda_i - 1) \geq \sum_{(i,j)\in\mathcal{N}} [f(x_i^T x_j, R_D) - 2]$$

$$\sum_{i=1}^{m}(\lambda_i - 1) \geq \frac{1}{2N_{\max}} \sum_{(i,j)\in\mathcal{N}} [f(x_i^T x_j, R_D) - 2],$$

where $N_{\max} := \max\{N_i\}_{i=1}^{m}$. Thus to prove (5) it suffices to show that

$$\frac{1}{2N_{\max}} \sum_{(i,j)\in\mathcal{N}} [f(x_i^T x_j, R_D) - 2] \geq 12 + (m-12)R_D - m = (m-12)(R_D - 1). \quad (6)$$

To bound $N_{\max}$ we utilize the following result, which is a slight generalization of [1, Lemma 5].

Proposition 1. *Suppose that a spherical triangle with sides a,b,c has* $\cos c \leq z_c$, $0 \leq z_a \leq \cos a \leq \cos b \leq z_b < 1$, $z_c \geq z_a z_b$. *Let γ be the spherical angle between the sides a and b. Then*

$$\cos \gamma \leq \max\left\{ \frac{z_c - z_a^2}{1 - z_a^2}, \frac{z_c - z_a z_b}{\sqrt{(1 - z_a^2)(1 - z_b^2)}} \right\}.$$

Lemma 2. $N_{\max} \leq 6$. *Moreover, for $m = 13$, if $N_{\max} = 6$ then* (5) *holds.*

Proof. Applying Proposition 1 with $z_a = 0.5$, $z_b = z_c = 1 - 1/(2R_D^2) \approx 0.6843$, we obtain $\cos \gamma \leq 0.5791$, or $\gamma \geq 54.6°$. It follows immediately that $N_{\max} \leq 6$, since $7(54.6°) > 360°$. For $m = 13$, Theorem 2 implies that (5) immediately holds if $x_i^T x_j \geq R_D/2$ for any $i \neq j$. Assume alternatively that $x_i^T x_j \leq R_D/2$ for all $i \neq j$. Applying Proposition 1 with $z_a = 0.5$, $z_b = z_c = R_D/2 \approx 0.6292$, we obtain $\cos \gamma \leq 0.5056$, or $\gamma \geq 59.6288°$. Hence $N_{\max} = 6$ is still possible, so assume that $N_i = 6$ for some i. Reindexing the points $\{x_j\}_{j=1}^{13}$, we can assume that $i = 7$ and the points $\{x_j\}_{j=1}^6$, have $x_j^T x_{(j \bmod 6)+1} \leq R_D/2$, $j = 1, \ldots, 6$. However, the fact that $\gamma \geq 59.6288°$ in each spherical triangle with vertices $x_7, x_j, x_{(j \bmod 6)+1}$ also implies that $\gamma \leq 360° - 5(59.6288°) = 61.856°$. Since Proposition 1 with $z_a = 0.5$, $z_b = R_D/2$, $z_c = 0.6$ obtains $\gamma \geq 62.18°$, we can conclude that $x_j^T x_{(j \bmod 6)+1} \geq 0.6$, $j = 1, \ldots, 6$. Applying Theorem 2, we conclude that $\lambda_j + \lambda_{j+1} \geq 1 + 2(0.6) = 2.2$ for $j = 1, 3, 5$. It follows that

$$\sum_{i=1}^{13} \lambda_i \geq 7 + 3(2.2) = 13.6,$$

which implies (5).   □

With an upper bound for $N_{\max}$ determined, a lower bound for the left-hand side of (6) can be obtained using the Delsarte bound for spherical codes. Specifically, $\mathcal{C} = \{x_i\}_{i=1}^m$ is a spherical z-code in $\Re^3$, with $z = 1 - 1/(2R_D^2) \approx 0.6843$. We define the usual *distance distribution* of the code to be the function $\alpha(\cdot) : [-1, 1] \rightarrow \Re_+$ defined as

$$\alpha(s) = \frac{|\{(i, j) : x_i^T x_j = s\}|}{m}. \tag{7}$$

It is then easy to see that $\alpha(\cdot) \geq 0$, and

$$\sum_{-1 \leq s \leq z} \alpha(s) = m - 1. \tag{8}$$

Let $\Phi_k(\cdot)$, $k = 0, 1, \ldots$ denote the Gegenbauer, or ultraspherical, polynomials $\Phi_k(t) = P_k^{(0,0)}(t)$ where $P_k^{(s,s)}$ is a Jacobi polynomial. It can be shown [4], [3, Chaps. 9, 13] that

$$1 + \sum_{-1 \leq s \leq z} \alpha(s) \Phi_k(s) \geq 0, \quad k = 1, 2, \ldots. \tag{9}$$

From (8) and (9), using $k = 1, \ldots, d$, a bound on the left-hand side of (6) can be obtained via the semi-infinite linear programming problem
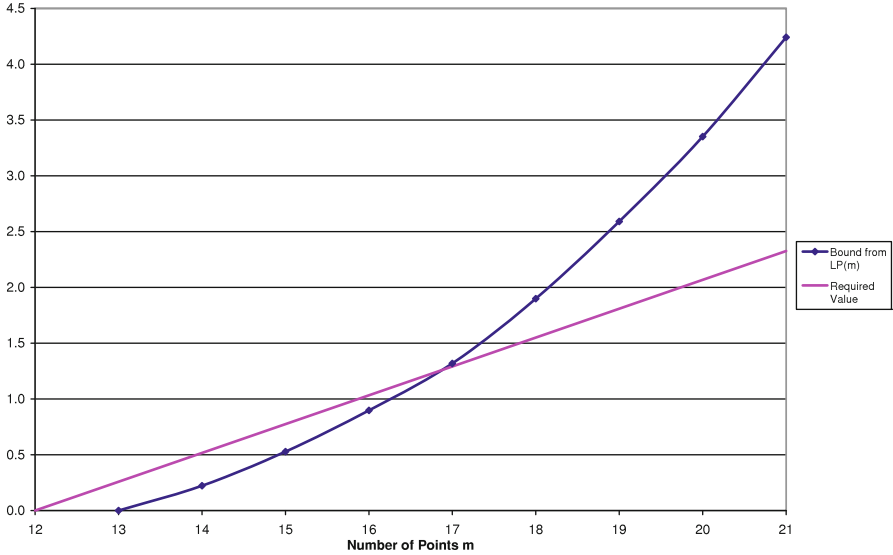
Fig. 2 LP bounds for inequality (6)

$$LP(m): \qquad \min \sum_{0.5 \le s \le z} m[f(s, R_D) - 2]\alpha(s)$$

$$\text{s.t.} \quad \sum_{s \in Z} \alpha(s)\Phi_k(s) \ge -1, \quad k = 1, \ldots, d,$$

$$\sum_{s \in Z} \alpha(s) = m - 1, \quad \alpha(s) \ge 0, \quad s \in Z,$$

where $Z := [-1, z]$. For $z = 1 - 1/(2R_D^2)$ the constraints of LP are feasible up to $m = 21$. (In other words, 21 is the Delsarte bound for the size of this spherical z-code. The maximum cardinality of a z-code for this value of z actually appears to be 20 [10].) Let $v_{LP}^*(m)$ denote the solution value in $LP(m)$. We obtain an approximate value of $v_{LP}^*(m)$ for $m = 13, \ldots, 21$ by numerically solving a discretized version of $LP(m)$ using $d = 16$, and values of $s \in Z$ incremented by 0.002.[3] In Fig. 2 we plot the lower bound $v_{LP}^*(m)/(2N_{max})$ for the left-hand side of (6) (using $N_{max} = 6$, except $N_{max} = 5$ for $m = 13$) and the required value $(m - 1)(R_D - 1)$ from the right-hand side of (6). The lower bound based on $v_{LP}^*(m)$ is sufficient to prove that (5) holds for $m \ge 17$.[4] The value $v_{LP}^*(13) = 0$ is a consequence of the well-known fact that the Delsarte bound for a $1/2$-code in $\Re^3$ is 13, despite the fact that the

---

[3] A rigorous lower bound for each $v_{LP}^*(m)$ can be obtained by solving the dual of the discretized problem and adjusting the dual solution to account for the discretization of s [3]. Alternatively a sum-of-squares formulation for the dual of $LP(m)$ could be used to solve the dual problem exactly.

[4] A referee has indicated that geometric arguments due to Marchal should also be able to establish that (5) holds for these cases, and possibly $m = 16$.

actual maximal size of such a code is 12. Indeed, this observation means that the approach based on LP(m) has no chance of establishing (5) for $m = 13$.

To prove (5) for $13 \leq m \leq 16$ requires stronger restrictions on the distance distribution than the constraints (9). The most attractive possibility appears to be the strengthened semidefinite programming constraints from [2]. In particular the constraints in [2] are sufficient to prove that the maximum cardinality of a 1/2-code in $\Re^3$ is 12, which is essential if one is to have any chance of proving (6) for $m = 13$. Applying the methodology of [2] results in a problem SDP(m) of the form

$$\text{SDP}(m) : \min \sum_{0.5 \leq s \leq z} m[f(s, R_D) - 2]\alpha(s)$$

$$\text{s.t. } 3\sum_{s \in Z} \alpha(s)S_k(s, s, 1) + \sum_{s,t,u \in Z} \alpha'(s, t, u)S_k(s, t, u) \succeq -S_k(1, 1, 1),$$

$$\sum_{s \in Z} \alpha(s)\Phi_k(s) \geq -1, \quad k = 1, \ldots, d$$

$$\sum_{s \in Z} \alpha(s) = m - 1, \quad \alpha(s) \geq 0, \quad s \in Z$$

$$\sum_{s,t,u \in Z} \alpha'(s, t, u) = (m - 1)(m - 2), \quad \alpha'(s, t, u) \geq 0, \quad s, t, u \in Z.$$

In SDP(m), $\alpha'(\cdot, \cdot, \cdot)$ is the three-point distance distribution

$$\alpha'(s, t, u) = \frac{|\{(i, j, k) : x_i^T x_j = s, x_i^T x_k = t, x_j^T x_k = u\}|}{m},$$

and $S_k(s, t, u)$ is a $(d + 1 - k) \times (d + 1 - k)$ symmetric matrix whose entries are symmetric polynomials of degree k in the variables (s, t, u); see [2] for details. (The notation $X \succeq Y$ means that $X - Y$ is positive semidefinite.) In Fig. 3 we show the bounds $v_{\text{SDP}}^*(m)/(2N_{\max})$ for the left-hand side of (6), as well as the required value $(m - 1)(R_D - 1)$ from the right-hand side of (6), for $13 \leq m \leq 16$.[5] For comparison we also give the previously described bounds based on $v_{\text{LP}}^*(m)$. As can be seen from the figure, the use of SDP(m) gives a substantial improvement over LP(m) for $m = 13$, but the magnitude of the difference appears to diminish as m increases, and the improved bound is unable to eliminate any more cases than were eliminated using LP(m).[6]

Although the use of SDP(m) is not sufficient to prove the key inequality (1) for all required m, there are several possible ways in which the approach based on SDP(m)

---

[5]The values of $v_{\text{SDP}}^*(m)$ are approximate, based on solving a discretization of SDP(m). It is possible to obtain rigorous bounds by applying a sum-of-squares formulation to the dual of SDP(m); see [2].

[6]As noted by a referee, it is possible that (6) is false even though (1) is true. We have not attempted to find a counter-example to (6) for the unresolved cases $13 \leq m \leq 16$.
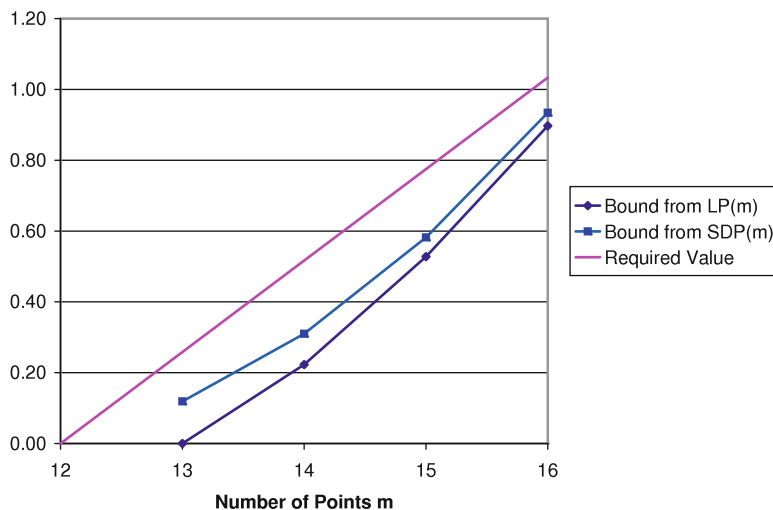
Fig. 3 LP and SDP bounds for inequality (6)

might be strengthened. In particular, since SDP(m) uses the three-point distance distribution, it should be possible to utilize a more elaborate version of Theorem 2 to give lower bounds on terms of the form $\lambda_i + \lambda_j + \lambda_k$. In addition, since the elements of the three-point distance distribution include the triangles in a Delaunay triangulation of the surface of the sphere, it might be possible to add valid constraints that can be derived for the Delaunay triangulation, as in [1]. The possibility that further strengthening of SDP(m) might suffice to establish (1) remains a very interesting topic for ongoing research, especially given the connection between (1) and the recent work of Hales [7,8] on the Kepler conjecture and related problems.

References

1. Antreicher, K.M.: The thirteen spheres: a new proof. Discret. Comput. Geom. **31**, 613–625 (2004)
2. Bachoc, C., Vallentin, F.: New upper bounds for kissing numbers from semidefinite programming. J. AMS **21**, 909–924 (2008)
3. Conway, J.H., Sloane, N.J.A.: Sphere Packings, Lattices and Groups, 3rd edn. Springer, New York (1999)
4. Delsarte, P., Goethals, J.-M., Seidel, J.J.: Spherical codes and designs. Geom. Dedicata **6**, 363–388 (1977)

5. Fejes Tóth, L.: Über die dichteste Kugellagerung. Math. Zeit. **48**, 676–684 (1943)
6. Fejes Tóth, L.: Regular Figures. Pergamon Press, New York (1964)
7. Hales, T.C.: The strong dodecahedral conjecture and Fejes Tóth's contact conjecture. Preprint, University of Pittsburgh. Available via arXiv. http://arxiv.org/abs/1110.0402v1 (2011)
8. Hales, T.C.: Dense Sphere Packings: A Blueprint for Formal Proofs. London Mathematical Society Lecture Note Series, vol. 400. Cambridge University Press, Cambridge/New York (2012)
9. Hales, T.C., McLaughlin, S.: A proof of the dodecahedral conjecture. J. AMS **23**, 299–344 (2010)
10. Sloane, N.J.A.: Spherical codes (packings). http://www.research.att.com/~njas/packings

On Minimal Tilings with Convex Cells Each Containing a Unit Ball

Károly Bezdek

Abstract We investigate the following problem that one can regard as a very close relative of the densest sphere packing problem. If the Euclidean 3-space is partitioned into convex cells each containing a unit ball, how should the shapes of the cells be designed to minimize the average edge curvature of the cells? In particular, we prove that the average edge curvature in question is always at least $13.8564\ldots$.

Key words Tiling • Convex cell • Unit sphere packing • Average edge curvature • Foam problem

Subject Classifications: 05B40, 05B45, 52B60, 52C17, 52C22

1 Introduction

We start with the following question: *if the Euclidean 3-space is partitioned into convex cells each containing a unit ball, how should the shapes of the cells be designed to minimize the average surface area of the cells?* In order to state the above question in more precise terms we proceed as follows. Let $\mathscr{T}$ be a tiling of the three-dimensional Euclidean space $\mathbb{E}^3$ into convex polyhedra $\mathbf{P}_i, i = 1, 2, \ldots$ each containing a unit ball say, $\mathbf{P}_i$ containing the closed three-dimensional ball $\mathbf{B}_i$ having radius 1 for $i = 1, 2, \ldots$. Also, we assume that there is a finite upper bound for the diameters of the convex cells in $\mathscr{T}$, i.e., $\sup\{\operatorname{diam}(\mathbf{P}_i) | i = 1, 2, \ldots\} < \infty$, where $\operatorname{diam}(\cdot)$ denotes the diameter of the corresponding set. In short, we say

K. Bezdek (✉)
Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada

Department of Mathematics, University of Pannonia, Veszprém, Hungary, Canada
e-mail: bezdek@math.ucalgary.ca

that $\mathscr{T}$ is a *normal tiling* of $\mathbb{E}^3$ with the underlying packing $\mathscr{P}$ of the unit balls $\mathbf{B}_i, i = 1, 2, \ldots$. Then we define the (lower) *average surface area* $\underline{s}(\mathscr{T})$ of the cells in $\mathscr{T}$ as follows:

$$\underline{s}(\mathscr{T}) := \liminf_{L \to \infty} \frac{\sum_{\{i \,|\, \mathbf{B}_i \subset \mathbf{C}_L\}} \text{sarea}(\mathbf{P}_i \cap \mathbf{C}_L)}{\text{card}\{i \,|\, \mathbf{B}_i \subset \mathbf{C}_L\}},$$

where $\mathbf{C}_L$ denotes the cube centered at the origin $\mathbf{o}$ with edges parallel to the coordinate axes of $\mathbb{E}^3$ and having edge length L. Furthermore, sarea($\cdot$) and card($\cdot$) denote the surface area and cardinality of the corresponding sets. (We note that it is rather straightforward to show that $\underline{s}(\mathscr{T})$ is independent from the choice of the coordinate system of $\mathbb{E}^3$.)

There is a very natural way to generate a large family of normal tilings. Namely, let $\mathscr{P}_R$ be an arbitrary packing of unit balls in $\mathbb{E}^3$ with the property that each closed ball of radius R in $\mathbb{E}^3$ contains the center of at least one unit ball in $\mathscr{P}_R$. Recall that the Voronoi cell of a unit ball in $\mathscr{P}_R$ is the set of points that are not farther away from the center of the given ball than from any other ball's center. It is well known that the Voronoi cells in question form a tiling of $\mathbb{E}^3$ (for more details see [17]). Furthermore, the Voronoi tiling obtained in this way is going to be a normal one because each Voronoi cell is contained in the closed ball of radius R concentric to the unit ball of the given Voronoi cell and therefore the diameter of each Voronoi cell is at most $2R$. Also, we recall here the *strong dodecahedral conjecture* of [3]: the surface area of every (bounded) Voronoi cell in a packing of unit balls is at least that of a regular dodecahedron of inradius 1, i.e., it is at least $16.6508\ldots$. After a sequence of partial results obtained in [3, 6], and [1] (proving the lower bounds $16.1433\ldots$, $16.1445\ldots$, and $16.1977\ldots$), just very recently, Hales [14] has announced a computer-assisted proof of the strong dodecahedral conjecture.

By adjusting Kertész's volume estimation technique [15] to the problem of estimating surface area and making the necessary modifications, we give a proof of the following inequality.

Theorem 1. *Let $\mathscr{T}$ be an arbitrary normal tiling of $\mathbb{E}^3$. Then the average surface area of the cells in $\mathscr{T}$ is always at least $\frac{24}{\sqrt{3}}$, i.e.,*

$$\underline{s}(\mathscr{T}) \geq \frac{24}{\sqrt{3}} = 13.8564\ldots.$$

Most likely the lower bound in Theorem 1 can be improved further; however, any such improvement would require additional new ideas. In particular, recall that in the face-centered cubic lattice packing of unit balls in $\mathbb{E}^3$, when each ball is touched by 12 others, the Voronoi cells of the unit balls are regular rhombic dodecahedra of inradius 1 and of surface area $12\sqrt{2}$ (for more details on the geometry involved see [8]). Thus, it is immediate to raise the following question: *prove or disprove that if $\mathscr{T}$ is an arbitrary normal tiling of $\mathbb{E}^3$, then*

$$\underline{s}(\mathscr{T}) \geq 12\sqrt{2} = 16.9705\ldots. \tag{1}$$

Let us mention that an affirmative answer to (1) (resp., a partially affirmative answer to (1) when only Voronoi tilings of unit ball packings are considered) would imply the Kepler conjecture. As is well known, the Kepler conjecture has been proved by Hales in a sequence of celebrated papers [9–13] concluding that the density of any unit ball packing in $\mathbb{E}^3$ is at most $\frac{\pi}{\sqrt{18}}$. Now, if $\underline{s}(\mathscr{T}) \geq 12\sqrt{2}$ were true for arbitrary normal tiling $\mathscr{T}$ of $\mathbb{E}^3$ with the unit balls of the cells forming the packing $\mathscr{P}$ in $\mathbb{E}^3$ (resp., if the inequality $\underline{s}(\mathscr{T}) \geq 12\sqrt{2}$ were true for Voronoi tilings $\mathscr{T}$ of unit ball packings $\mathscr{P}$ in $\mathbb{E}^3$), then based on the obvious inequalities

$$\sum_{\{i\,|\,\mathbf{B}_i \subset \mathbf{C}_L\}} \mathrm{vol}(\mathbf{P}_i \cap \mathbf{C}_L) \leq \mathrm{vol}(\mathbf{C}_L) \ \text{ and } \ \frac{1}{3}\mathrm{sarea}(\mathbf{P}_i \cap \mathbf{C}_L) \leq \mathrm{vol}(\mathbf{P}_i \cap \mathbf{C}_L),$$

(where $\mathrm{vol}(\cdot)$ denotes the volume of the corresponding set) we would get that the (upper) *density* $\overline{\delta}(\mathscr{P}) := \limsup_{L\to\infty} \dfrac{\frac{4\pi}{3}\mathrm{card}\{i\,|\,\mathbf{B}_i \subset \mathbf{C}_L\}}{\mathrm{vol}(\mathbf{C}_L)}$ of the packing $\mathscr{P}$ must satisfy the inequality

$$\begin{aligned} \overline{\delta}(\mathscr{P}) &\leq \limsup_{L\to\infty} \frac{\frac{4\pi}{3}\mathrm{card}\{i\,|\,\mathbf{B}_i \subset \mathbf{C}_L\}}{\sum_{\{i\,|\,\mathbf{B}_i \subset \mathbf{C}_L\}} \mathrm{vol}(\mathbf{P}_i \cap \mathbf{C}_L)} \\ &\leq \limsup_{L\to\infty} \frac{4\pi\,\mathrm{card}\{i\,|\,\mathbf{B}_i \subset \mathbf{C}_L\}}{\sum_{\{i\,|\,\mathbf{B}_i \subset \mathbf{C}_L\}} \mathrm{sarea}(\mathbf{P}_i \cap \mathbf{C}_L)} = \frac{4\pi}{\underline{s}(\mathscr{T})} \\ &\leq \frac{\pi}{\sqrt{18}} \end{aligned}$$

Thus, one could regard the affirmative version of (1), stated for Voronoi tilings of unit ball packings, as a *strong version of the Kepler conjecture*.

As an additional observation we mention that an affirmative answer to (1) would imply also the rhombic dodecahedral conjecture of [4]. According to that conjecture the surface area of any three-dimensional parallelohedron of inradius at least 1 (i.e., the surface area of any convex polyhedron containing a unit ball and having a family of translates tiling $\mathbb{E}^3$) is at least as large as $12\sqrt{2} = 16.9705\ldots$.

By taking a closer look of the proof of Theorem 1 we derive the following stronger version. In order to state it in a proper form we need to introduce some additional terminology. Let $\mathbf{P}$ be a convex polyhedron in $\mathbb{E}^3$ and let $E(\mathbf{P})$ denote the family of the edges of $\mathbf{P}$. Then $\mathrm{ecurv}(\mathbf{P}) := \sum_{e\in E(\mathbf{P})} L(e)\tan\frac{\alpha_e}{2}$ is called the *edge curvature* of $\mathbf{P}$, where $L(e)$ stands for the length of the edge $e \in E(\mathbf{P})$ and α_e denotes the angle between the outer normal vectors of the two faces of $\mathbf{P}$ meeting along the edge e. (For more insight on edge curvature we refer the interested reader to p. 287 in [8].) Now, if $\mathscr{T}$ is an arbitrary normal tiling of $\mathbb{E}^3$ with the underlying packing $\mathscr{P}$ of the unit balls $\mathbf{B}_i, i = 1, 2, \ldots$, then we define the (lower) *average edge curvature* $\underline{ec}(\mathscr{T})$ of the cells in $\mathscr{T}$ as follows:

$$\underline{ec}(\mathscr{T}) := \liminf_{L\to\infty} \frac{\sum_{\{i\,|\,\mathbf{B}_i \subset \mathbf{C}_L\}} \mathrm{ecurv}(\mathbf{P}_i \cap \mathbf{C}_L)}{\mathrm{card}\{i\,|\,\mathbf{B}_i \subset \mathbf{C}_L\}}.$$

Based on this we can state the following stronger version of Theorem 1.

Theorem 2. *Let $\mathscr{T}$ be an arbitrary normal tiling of $\mathbb{E}^3$. Then*

$$\underline{s}(\mathscr{T}) \geq \underline{ec}(\mathscr{T}) \geq \frac{24}{\sqrt{3}} = 13.8564\ldots .$$

Moreover, if $\mathscr{T}$ is a Voronoi tiling of a unit ball packing in $\mathbb{E}^3$, then

$$\underline{ec}(\mathscr{T}) \geq \frac{2\pi^2}{6\sqrt{6}\arcsin\left(\frac{1}{\sqrt{3}}\right) - \pi\sqrt{6}} = 14.6176\ldots .$$

Thus, we have the following stronger version of our original problem: *if the Euclidean 3-space is partitioned into convex cells each containing a unit ball, how should the shapes of the cells be designed to minimize the average edge curvature of the cells*?

Last but not least, it is very tempting to further relax the conditions in our original problem by replacing convex cells with cells that are measurable and have measurable boundaries and ask the following more general question: *if the Euclidean 3-space is partitioned into cells each containing a unit ball, how should the shapes of the cells be designed to minimize the average surface area of the cells*? One can regard this question as a *foam problem*, in particular, as a relative of Kelvin's foam problem (on partitioning $\mathbb{E}^3$ into unit volume cells with minimum average surface area) since foams are simply tilings of space that (under proper conditions) minimize surface area. Although foams are well studied (see the relevant sections of the highly elegant book [16] of Morgan), it is far not clear what would be a good candidate for the proper minimizer in the foam question just raised. As a last note we mention that Brakke (November, 2011, personal communication), by properly modifying the Williams foam, has just obtained a partition of the Euclidean 3-space into congruent cells each containing a unit ball and having surface area $16.95753 < 12\sqrt{2} = 16.9705\ldots .$

2 Proof of Theorem 1

First, we prove the following "compact" version of Theorem 1 (also because it might be of independent interest).

Theorem 3. *If the cube $\mathbf{C}$ is partitioned into the convex cells $\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_n$ each containing a unit ball in $\mathbb{E}^3$, then the sum of the surface areas of the n convex cells is at least $\frac{24}{\sqrt{3}}n$, i.e.,*

$$\sum_{i=1}^{n} \text{sarea}(\mathbf{Q}_i) \geq \frac{24}{\sqrt{3}}n .$$

Proof. It is well known that the Brunn–Minkowski inequality implies the following inequality:

$$\text{sarea}^2(\mathbf{Q}_i) \geq 3\text{vol}(\mathbf{Q}_i)\text{ecurv}(\mathbf{Q}_i) , \tag{2}$$

where $1 \leq i \leq n$. (For a proof we refer the interested reader to p. 287 in [8].) In what follows it is more proper to use the inner dihedral angles $\beta_e := \pi - \alpha_e$ and the relevant formula for the edge curvature:

$$\text{ecurv}(\mathbf{Q}_i) = \sum_{e \in E(\mathbf{Q}_i)} L(e) \cot \frac{\beta_e}{2}. \tag{3}$$

As, by assumption, $\mathbf{Q}_i$ contains a unit ball therefore

$$\text{vol}(\mathbf{Q}_i) \geq \frac{1}{3}\text{sarea}(\mathbf{Q}_i). \tag{4}$$

Hence, (2), (3), and (4) imply in a straightforward way that

$$\text{sarea}(\mathbf{Q}_i) \geq \sum_{e \in E(\mathbf{Q}_i)} L(e) \cot \frac{\beta_e}{2} \tag{5}$$

holds for all $1 \leq i \leq n$.

Now, let $s \subset \mathbf{C}$ be a closed line segment along which exactly k members of the family $\{\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_n\}$ meet having inner dihedral angles $\beta_1, \beta_2, \ldots, \beta_k$. There are the following three possibilities:

(a) s is on an edge of the cube $\mathbf{C}$.
(b) s is in the relative interior either of a face of $\mathbf{C}$ or of a face of a convex cell in the family $\{\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_n\}$.
(c) s is in the interior of $\mathbf{C}$ and not in the relative interior of any face of any convex cell in the family $\{\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_n\}$.

In each of the above cases we can make the following easy observations:

(a) $\beta_1 + \beta_2 + \cdots + \beta_k = \frac{\pi}{2}$ with $k \geq 1$.
(b) $\beta_1 + \beta_2 + \cdots + \beta_k = \pi$ with $k \geq 2$.
(c) $\beta_1 + \beta_2 + \cdots + \beta_k = 2\pi$ with $k \geq 3$.

As $y = \cot x$ is convex and decreasing over the interval $0 < x \leq \frac{\pi}{2}$ therefore the following inequalities must hold:

(a) $\cot \frac{\beta_1}{2} + \cot \frac{\beta_2}{2} + \cdots + \cot \frac{\beta_k}{2} \geq k \cot \frac{\pi}{4k} \geq k.$
(b) $\cot \frac{\beta_1}{2} + \cot \frac{\beta_2}{2} + \cdots + \cot \frac{\beta_k}{2} \geq k \cot \frac{\pi}{2k} \geq k.$
(c) $\cot \frac{\beta_1}{2} + \cot \frac{\beta_2}{2} + \cdots + \cot \frac{\beta_k}{2} \geq k \cot \frac{\pi}{k} \geq \frac{1}{\sqrt{3}}k.$

In short, the following inequality holds in all three cases:

$$\cot \frac{\beta_1}{2} + \cot \frac{\beta_2}{2} + \cdots + \cot \frac{\beta_k}{2} \geq \frac{1}{\sqrt{3}}k. \tag{6}$$

Thus, by adding together the inequalities (5) for all $1 \leq i \leq n$ and using (6) we get that

$$\sum_{i=1}^{n} \text{sarea}(\mathbf{Q}_i) \geq \frac{1}{\sqrt{3}} \sum_{i=1}^{n} \sum_{e \in E(\mathbf{Q}_i)} L(e). \tag{7}$$

Finally, recall the elegant theorem of Besicovitch and Eggleston [2] claiming that the total edge length of any convex polyhedron containing a unit ball in $\mathbb{E}^3$ is always at least as large as the total edge length of a cube circumscribing a unit ball. This implies that

$$\sum_{e \in E(\mathbf{Q}_i)} L(e) \geq 24 \tag{8}$$

holds for all $1 \leq i \leq n$. Hence, (7) and (8) finish the proof of Theorem 3.   □

Second, we take a closer look at the given normal tiling $\mathscr{T}$ and, using the above proof of Theorem 3, we give a proof of Theorem 1. The details are as follows.

By assumption $D := \sup\{\mathrm{diam}(\mathbf{P}_i) | i = 1, 2, \ldots\} < \infty$. Thus, clearly each closed ball of radius D in $\mathbb{E}^3$ contains at least one of the convex polyhedra $\mathbf{P}_i, i = 1, 2, \ldots$ (forming the tiling $\mathscr{T}$ of $\mathbb{E}^3$). Now, let $\mathbf{C}_{L_N}, N = 1, 2, \ldots$ be an arbitrary sequence of cubes centered at the origin $\mathbf{o}$ with edges parallel to the coordinate axes of $\mathbb{E}^3$ and having edge length $L_N, N = 1, 2, \ldots$ with $\lim_{N \to \infty} L_N = \infty$. It follows that

$$0 < \liminf_{N \to \infty} \frac{\frac{4\pi}{3} \mathrm{card}\{i \,|\, \mathbf{B}_i \subset \mathbf{C}_{L_N}\}}{\mathrm{vol}(\mathbf{C}_{L_N})} \leq \limsup_{N \to \infty} \frac{\frac{4\pi}{3} \mathrm{card}\{i \,|\, \mathbf{B}_i \subset \mathbf{C}_{L_N}\}}{\mathrm{vol}(\mathbf{C}_{L_N})} < 1. \tag{9}$$

Note that clearly

$$\frac{\mathrm{card}\{i \,|\, \mathbf{P}_i \cap \mathrm{bd}\mathbf{C}_{L_N} \neq \emptyset\}}{\mathrm{card}\{i \,|\, \mathbf{B}_i \subset \mathbf{C}_{L_N}\}} \leq \frac{\left(\mathrm{vol}(\mathbf{C}_{L_N+2D}) - \mathrm{vol}(\mathbf{C}_{L_N-2D})\right)\mathrm{vol}(\mathbf{C}_{L_N})}{\mathrm{vol}(\mathbf{C}_{L_N})\frac{4\pi}{3}\mathrm{card}\{i \,|\, \mathbf{B}_i \subset \mathbf{C}_{L_N}\}}. \tag{10}$$

Moreover,

$$\lim_{N \to \infty} \frac{\mathrm{vol}(\mathbf{C}_{L_N+2D}) - \mathrm{vol}(\mathbf{C}_{L_N-2D})}{\mathrm{vol}(\mathbf{C}_{L_N})} = 0. \tag{11}$$

Thus, (9), (10), and (11) imply in a straightforward way that

$$\lim_{N \to \infty} \frac{\mathrm{card}\{i \,|\, \mathbf{P}_i \cap \mathrm{bd}\mathbf{C}_{L_N} \neq \emptyset\}}{\mathrm{card}\{i \,|\, \mathbf{B}_i \subset \mathbf{C}_{L_N}\}} = 0. \tag{12}$$

Moreover, (5) yields that

$$\mathrm{sarea}(\mathbf{P}_i) \geq \mathrm{ecurv}(\mathbf{P}_i) = \sum_{e \in E(\mathbf{P}_i)} L(e) \cot \frac{\beta_e}{2} \tag{13}$$

holds for all $i = 1, 2, \ldots$. As a next step, using

$$\mathrm{sarea}(\mathbf{P}_i) = \mathrm{sarea}\left(\mathrm{bd}(\mathbf{P}_i \cap \mathbf{C}_L) \setminus \mathrm{bd}\mathbf{C}_L\right) + \delta_i \tag{14}$$

and

$$\mathrm{ecurv}(\mathbf{P}_i) \geq \sum_{e \in E(\mathrm{bd}(\mathbf{P}_i \cap \mathbf{C}_L) \setminus \mathrm{bd}\mathbf{C}_L)} L(e) \cot \frac{\beta_e}{2} \tag{15}$$

(with bd$(\cdot)$ denoting the boundary of the corresponding set) we obtain the following from (13):

$$\text{sarea}\left(\text{bd}(\mathbf{P}_i \cap \mathbf{C}_L) \setminus \text{bd}\mathbf{C}_L\right) + \delta_i \geq \sum_{e \in E(\text{bd}(\mathbf{P}_i \cap \mathbf{C}_L) \setminus \text{bd}\mathbf{C}_L)} L(e) \cot \frac{\beta_e}{2}, \qquad (16)$$

where clearly $0 \leq \delta_i \leq \text{sarea}(\mathbf{P}_i)$. Hence, (16) combined with (6) yields

Corollary 1.

$$f(L) := \sum_{\{i \mid \text{int}\mathbf{P}_i \cap \mathbf{C}_L \neq \emptyset\}} \text{sarea}\left(\text{bd}(\mathbf{P}_i \cap \mathbf{C}_L) \setminus \text{bd}\mathbf{C}_L\right) + \sum_{\{i \mid \mathbf{P}_i \cap \text{bd}\mathbf{C}_L \neq \emptyset\}} \delta_i$$

$$\geq g(L) := \frac{1}{\sqrt{3}} \sum_{\{i \mid \text{int}\mathbf{P}_i \cap \mathbf{C}_L \neq \emptyset\}} \left(\sum_{e \in E(\text{bd}(\mathbf{P}_i \cap \mathbf{C}_L) \setminus \text{bd}\mathbf{C}_L)} L(e) \right)$$

Now, it is easy to see that

$$f(L) = \Delta(L) + \sum_{\{i \mid \mathbf{B}_i \subset \mathbf{C}_L\}} \text{sarea}(\mathbf{P}_i \cap \mathbf{C}_L), \qquad (17)$$

where $0 \leq \Delta(L) \leq 2\sum_{\{i \mid \mathbf{P}_i \cap \text{bd}\mathbf{C}_L \neq \emptyset\}} \text{sarea}(\mathbf{P}_i)$.

Moreover, (8) implies that

$$g(L) \geq -\overline{\Delta}(L) + \sum_{\{i \mid \mathbf{B}_i \subset \mathbf{C}_L\}} \frac{24}{\sqrt{3}}, \qquad (18)$$

where $0 \leq \overline{\Delta}(L) \leq \sum_{\{i \mid \mathbf{P}_i \cap \text{bd}\mathbf{C}_L \neq \emptyset\}} \sum_{e \in E(\mathbf{P}_i)} L(e)$.

Lemma 1.

$$A := \sup\{\text{sarea}(\mathbf{P}_i) \mid i = 1, 2, \dots\} < \infty$$

and

$$E := \sup\left\{ \sum_{e \in E(\mathbf{P}_i)} L(e) \mid i = 1, 2, \dots \right\} < \infty.$$

Proof. Recall that $D = \sup\{\text{diam}(\mathbf{P}_i) \mid i = 1, 2, \dots\} < \infty$. Hence, according to Jung's theorem [7] each $\mathbf{P}_i$ is contained in a closed ball of radius $\sqrt{\frac{3}{8}} D$ in $\mathbb{E}^3$. Thus, $A \leq \frac{3}{2}\pi D^2 < \infty$.

For a proof of the other claim recall that $\mathbf{P}_i$ contains the unit ball $\mathbf{B}_i$ centered at $\mathbf{o}_i$. If the number of faces of $\mathbf{P}_i$ is f_i, then $\mathbf{P}_i$ must have at least f_i neighbors (i.e., cells of $\mathscr{T}$ that have at least one point in common with $\mathbf{P}_i$) and as each neighbor is contained in the closed 3-dimensional ball of radius $2D$ centered at $\mathbf{o}_i$ therefore the

number of neighbors of $\mathbf{P}_i$ is at most $(2D)^3 - 1$ and so, $f_i \leq 8D^3 - 1$. (Here, we have used the fact that each neighbor contains a unit ball and therefore its volume is larger than $\frac{4\pi}{3}$.) Finally, Euler's formula implies that the number of edges of $\mathbf{P}_i$ is at most $3f_i - 6 \leq 24D^3 - 9$. Thus, $E \leq 24D^4 - 9D < \infty$ (because the length of any edge of $\mathbf{P}_i$ is at most D). $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Thus, Corollary 1, (17), (18), and Lemma 1 imply the following inequality in a straightforward way.

Corollary 2.

$$\frac{2A\,\mathrm{card}\{i\,|\mathbf{P}_i \cap \mathrm{bd}\mathbf{C}_L \neq \emptyset\} + \sum_{\{i\,|\mathbf{B}_i \subset \mathbf{C}_L\}} \mathrm{sarea}(\mathbf{P}_i \cap \mathbf{C}_L)}{\mathrm{card}\{i\,|\mathbf{B}_i \subset \mathbf{C}_L\}}$$

$$\geq \frac{-E\,\mathrm{card}\{i\,|\mathbf{P}_i \cap \mathrm{bd}\mathbf{C}_L \neq \emptyset\} + \sum_{\{i\,|\mathbf{B}_i \subset \mathbf{C}_L\}} \frac{24}{\sqrt{3}}}{\mathrm{card}\{i\,|\mathbf{B}_i \subset \mathbf{C}_L\}} \,.$$

Finally, Corollary 2 and (12) yield that

$$\liminf_{N \to \infty} \frac{\sum_{\{i\,|\mathbf{B}_i \subset \mathbf{C}_{L_N}\}} \mathrm{sarea}(\mathbf{P}_i \cap \mathbf{C}_{L_N})}{\mathrm{card}\{i\,|\mathbf{B}_i \subset \mathbf{C}_{L_N}\}} \geq \frac{24}{\sqrt{3}} \,,$$

finishing the proof of Theorem 1.

3   Proof of Theorem 2

If $\mathscr{T}$ is an arbitrary normal tiling of $\mathbb{E}^3$ with the underlying packing $\mathscr{P}$ of the unit balls $\mathbf{B}_i, i = 1, 2, \ldots$, then (5) implies

$$\mathrm{sarea}(\mathbf{P}_i \cap \mathbf{C}_L) \geq \mathrm{ecurv}(\mathbf{P}_i \cap \mathbf{C}_L)$$

and therefore $\underline{s}(\mathscr{T}) \geq \underline{ec}(\mathscr{T})$ follows in a straightforward way. So, we are left to show that $\underline{ec}(\mathscr{T}) \geq \frac{24}{\sqrt{3}}$. In order to achieve that, we take a closer look of the given normal tiling $\mathscr{T}$ and using some of the properly modified estimates of the proof of Theorem 1 we derive the inequality $\underline{ec}(\mathscr{T}) \geq \frac{24}{\sqrt{3}}$. The details are as follows.

We start with the following immediate analogue of (16):

$$\mathrm{ecurv}(\mathbf{P}_i) = \delta_i^* + \sum_{e \in E(\mathrm{bd}(\mathbf{P}_i \cap \mathbf{C}_L) \setminus \mathrm{bd}\mathbf{C}_L)} L(e) \cot \frac{\beta_e}{2} \,, \qquad (19)$$

where $0 \leq \delta_i^* \leq \mathrm{ecurv}(\mathbf{P}_i) \leq \mathrm{sarea}(\mathbf{P}_i)$. If

$$\mathrm{ecurv}\big(\mathrm{bd}(\mathbf{P}_i \cap \mathbf{C}_L) \setminus \mathrm{bd}\mathbf{C}_L\big) := \sum_{e \in E(\mathrm{bd}(\mathbf{P}_i \cap \mathbf{C}_L) \setminus \mathrm{bd}\mathbf{C}_L)} L(e) \cot \frac{\beta_e}{2} \,,$$

then (19) combined with (6) yields

Corollary 3.

$$f^*(L) := \sum_{\{i\,|\,\mathrm{int}\mathbf{P}_i \cap \mathbf{C}_L \neq \emptyset\}} \mathrm{ecurv}\big(\mathrm{bd}(\mathbf{P}_i \cap \mathbf{C}_L) \setminus \mathrm{bd}\mathbf{C}_L\big) + \sum_{\{i\,|\,\mathbf{P}_i \cap \mathrm{bd}\mathbf{C}_L \neq \emptyset\}} \delta_i^*$$

$$\geq g(L) = \frac{1}{\sqrt{3}} \sum_{\{i\,|\,\mathrm{int}\mathbf{P}_i \cap \mathbf{C}_L \neq \emptyset\}} \left(\sum_{e \in E(\mathrm{bd}(\mathbf{P}_i \cap \mathbf{C}_L) \setminus \mathrm{bd}\mathbf{C}_L)} L(e) \right)$$

Now, it is easy to see that

$$f^*(L) = \Delta^*(L) + \sum_{\{i\,|\,\mathbf{B}_i \subset \mathbf{C}_L\}} \mathrm{ecurv}(\mathbf{P}_i \cap \mathbf{C}_L), \tag{20}$$

where $0 \leq \Delta^*(L) \leq 2\sum_{\{i\,|\,\mathbf{P}_i \cap \mathrm{bd}\mathbf{C}_L \neq \emptyset\}} \mathrm{ecurv}(\mathbf{P}_i)$. Moreover, $g(L)$ must satisfy (18). Thus, Corollary 3, (20), (18), and Lemma 1 imply the following inequality in a straightforward way.

Corollary 4.

$$\frac{2A\,\mathrm{card}\{i\,|\,\mathbf{P}_i \cap \mathrm{bd}\mathbf{C}_L \neq \emptyset\} + \sum_{\{i\,|\,\mathbf{B}_i \subset \mathbf{C}_L\}} \mathrm{ecurv}(\mathbf{P}_i \cap \mathbf{C}_L)}{\mathrm{card}\{i\,|\,\mathbf{B}_i \subset \mathbf{C}_L\}}$$

$$\geq \frac{-E\,\mathrm{card}\{i\,|\,\mathbf{P}_i \cap \mathrm{bd}\mathbf{C}_L \neq \emptyset\} + \sum_{\{i\,|\,\mathbf{B}_i \subset \mathbf{C}_L\}} \frac{24}{\sqrt{3}}}{\mathrm{card}\{i\,|\,\mathbf{B}_i \subset \mathbf{C}_L\}}.$$

Hence, Corollary 4 and (12) yield that

$$\liminf_{N \to \infty} \frac{\sum_{\{i\,|\,\mathbf{B}_i \subset \mathbf{C}_{L_N}\}} \mathrm{ecurv}(\mathbf{P}_i \cap \mathbf{C}_{L_N})}{\mathrm{card}\{i\,|\,\mathbf{B}_i \subset \mathbf{C}_{L_N}\}} \geq \frac{24}{\sqrt{3}} \,,$$

finishing the proof of the inequality $\underline{ec}(\mathscr{T}) \geq \frac{24}{\sqrt{3}}$.

Finally, let $\mathscr{T}$ be the Voronoi tiling of a unit ball packing in $\mathbb{E}^3$ consisting of the Voronoi cells $\mathbf{P}_i, i = 1, 2, \dots$ (each containing a unit ball). First, recall the inequality

$$\mathrm{ecurv}(\mathbf{P}_i) > 2\pi\,\mathrm{mwidth}(\mathbf{P}_i), \tag{21}$$

where $i = 1, 2, \dots$ and $\mathrm{mwidth}(\cdot)$ denotes the mean width of the corresponding set. (For more details on this inequality see p. 287 in [8] as well as the relevant discussion on p. 392 in [5].) Second, recall that according to a recent result of the author [5] the inequality

$$\text{mwidth}(\mathbf{P}_i) \geq \frac{\pi}{6\sqrt{6}\arcsin\left(\frac{1}{\sqrt{3}}\right) - \pi\sqrt{6}} = 2.3264\ldots \tag{22}$$

holds for all $i = 1, 2, \ldots$. Thus, (21) and (22) yield

$$\text{ecurv}(\mathbf{P}_i) > \frac{2\pi^2}{6\sqrt{6}\arcsin\left(\frac{1}{\sqrt{3}}\right) - \pi\sqrt{6}} = 14.6176\ldots$$

from which it follows in a straightforward way that $\underline{ec}(\mathscr{T}) \geq 14.6176\ldots$, finishing the proof of Theorem 2.

References

1. Ambrus, G., Fodor, F.: A new lower bound on the surface area of a Voronoi polyhedron. Period. Math. Hungar. **53**(1–2), 45–58 (2006)
2. Besicovitch, A.S., Eggleston, H.G.: The total length of the edges of a polyhedron. Q. J. Math. Oxford Ser. **2**(8), 172–190 (1957)
3. Bezdek, K.: On a stronger form of Rogers's lemma and the minimum surface area of Voronoi cells in unit ball packings. J. Reine Angew. Math. **518**, 131–143 (2000)
4. Bezdek, K.: On rhombic dodecahedra. Contrib. Alg. Geom. **41**(2), 411–416 (2000)
5. Bezdek, K.: A lower bound for the mean width of Voronoi polyhedra of unit ball packings in E^3. Arch. Math. **74**(5), 392–400 (2000)
6. Bezdek, K., Daróczy-Kiss, E.: Finding the best face on a Voronoi polyhedron – the strong dodecahedral conjecture revisited. Monatsh. Math. **145**(3), 191–206 (2005)
7. Dekster, B.V.: An extension of Jung's theorem. Isr. J. Math. **50**(3), 169–180 (1985)
8. Fejes Tóth, L.: Regular Figures. Pergamon Press, New York (1964)
9. Hales, T.C.: A proof of the Kepler conjecture. Ann. Math. **162**(3), 1065–1185 (2005)
10. Hales, T.C., Ferguson, S.P.: A formulation of the Kepler conjecture. Discrete Comput. Geom. **36**(1), 21–69 (2006)
11. Hales, T.C.: Sphere packings III, extremal cases. Discrete Comput. Geom. **36**(1), 71–110 (2006)
12. Hales, T.C.: Sphere packings IV, detailed bounds. Discrete Comput. Geom. **36**(1), 111–166 (2006)
13. Hales, T.C.: Sphere packings VI, Tame graphs and linear programs. Discrete Comput. Geom. **36**(1), 205–265 (2006)
14. Hales, T.C.: Dense Sphere Packings. A Blueprint for Formal Proofs. Cambridge University Press, Cambridge (2012)
15. Kertész, G.: On totally separable packings of equal balls. Acta Math. Hungar. **51**(3–4), 363–364 (1988)
16. Morgan, F.: Geometric Measure Theory - A Beginner's Guide. Elsevier – Academic Press, Amsterdam (2009)
17. Rogers, C.A.: Packing and Covering. Cambridge University Press, Cambridge (1964)

On Volumes of Permutation Polytopes

Katherine Burggraf, Jesús De Loera, and Mohamed Omar

Abstract This paper focuses on determining the volumes of permutation polytopes associated to cyclic groups, dihedral groups, groups of automorphisms of tree graphs, and Frobenius groups. We do this through the use of triangulations and the calculation of Ehrhart polynomials. We also briefly discuss the theta body hierarchy of various permutation polytopes.

Key words Permutation polytopes • Birkhoff's polytope • Volumes • Ehrhart polynomials • Gale duality • Triangulations • Theta bodies

Subject Classifications: 52A38, 52B20, 52B35

1 Introduction

Volumes are fundamental geometric invariants of convex bodies that often carry algebraic and combinatorial data. The computation of volumes appears in many areas of mathematical literature, including in the context of order polytopes [32], the Chan-Robbins-Yuen polytope [11, 39], the convex hull of the positive root configuration $A_n^{(+)}$ [20], and polytopes arising from algebraic geometry [5, 34, 35]. A key example of such a convex body is the *Birkhoff polytope B_n*. It is defined as

K. Burggraf • J. De Loera (✉)
Department of Mathematics, University of California, One Shields Ave.,
Davis, CA 95616, USA
e-mail: kburggraf@ucdavis.edu; deloera@math.ucdavis.edu

M. Omar
Department of Mathematics, California Institute of Technology,
Mathematics 253-37, Pasadena, CA 91125, USA
e-mail: momar@caltech.edu

the convex hull of all $n \times n$ permutation matrices, or equivalently, as the convex hull of the natural permutation representation of the symmetric group S_n; see [4, 6, 7, 9, 10, 15, 18, 30, 34] and references therein for a summary of its known properties. Subpolytopes of B_n have been shown to have remarkably beautiful properties; see [1,6,8,10,11,19,29,31,39] and references therein. This is particularly true for permutation polytopes, those polytopes that arise by taking convex hulls of permutation representations of special subgroups of S_n with concrete sets of generators. Their geometry reflects their group-theoretic structure. Our focus in this paper is on determining the volumes, or normalized volumes, of permutation polytopes associated to cyclic groups, dihedral groups, groups of automorphisms of tree graphs, and Frobenius groups. We take two essential approaches: triangulating a polytope by simplices of equal volume, or obtaining the volume through the Ehrhart polynomial of the polytope. The Ehrhart polynomial of a polytope P, which we denote by $i(P, t)$, is the function that counts the number of integer points in the tth dilation of P, and it turns out, whose leading coefficient is the volume of the polytope. For more on these approaches, see Sect. 2. For known results on permutation polytopes, see [3, 8, 12, 14, 22, 24, 33] and references therein. For definitions pertaining to permutation polytopes, see [3, 22].

Before stating our results, we will clarify some terminology. The *normalized volume* of a d-dimensional polytope $P \subset \mathbb{R}^n$ with respect to an affine lattice $L \subset \mathbb{R}^n$ is the volume form that assigns a volume of one to the smallest d-dimensional simplices in $\mathbb{R}^n$ whose vertices are in L. The *volume* of P is its normalized volume in the lattice $\mathrm{aff}(P) \cap \mathbb{Z}^n$. We say P is *unimodular* with respect to L if it has a triangulation whose simplices are all unimodular; that is, the vertices of any simplex in the triangulation span the lattice L. For more on triangulations with respect to particular lattices used in this paper, see Sect. 2. In what follows, we identify the symmetric group S_n on $\{1, 2, \ldots, n\}$ through its representation by $n \times n$ permutation matrices; that is, for any $g \in S_n$, we identify g with the $n \times n$ matrix whose (i, j)-entry is one if $g(i) = j$ and 0 otherwise. We denote the identity by e throughout. We denote a subgroup G of S_n by $G \leq S_n$. Such a subgroup is called a *permutation group*. For any permutation group $G \leq S_n$, we refer to the polytope $P(G) := \mathrm{conv}\{g \mid g \in G\}$ as the *permutation polytope* associated to G.

Example 1. Let $G \leq S_4$ be the group consisting of the four permutations $\{e, (1\ 2), (3\ 4), (1\ 2)(3\ 4)\}$. Then $P(G)$ is the convex hull of the matrices

$$
\begin{pmatrix} 1\ 0\ 0\ 0 \\ 0\ 1\ 0\ 0 \\ 0\ 0\ 1\ 0 \\ 0\ 0\ 0\ 1 \end{pmatrix}, \quad
\begin{pmatrix} 0\ 1\ 0\ 0 \\ 1\ 0\ 0\ 0 \\ 0\ 0\ 1\ 0 \\ 0\ 0\ 0\ 1 \end{pmatrix}, \quad
\begin{pmatrix} 1\ 0\ 0\ 0 \\ 0\ 1\ 0\ 0 \\ 0\ 0\ 0\ 1 \\ 0\ 0\ 1\ 0 \end{pmatrix}, \quad
\begin{pmatrix} 0\ 1\ 0\ 0 \\ 1\ 0\ 0\ 0 \\ 0\ 0\ 0\ 1 \\ 0\ 0\ 1\ 0 \end{pmatrix}.
$$

This polytope is geometrically a square. Now let $H \leq S_4$ be the group consisting of the four permutations $\{e, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\}$. Then $P(H)$ is the convex hull of the matrices

$$
\begin{pmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \end{pmatrix}, \quad
\begin{pmatrix} 0\,1\,0\,0 \\ 1\,0\,0\,0 \\ 0\,0\,0\,1 \\ 0\,0\,1\,0 \end{pmatrix}, \quad
\begin{pmatrix} 0\,0\,1\,0 \\ 0\,0\,0\,1 \\ 1\,0\,0\,0 \\ 0\,1\,0\,0 \end{pmatrix}, \quad
\begin{pmatrix} 0\,0\,0\,1 \\ 0\,0\,1\,0 \\ 0\,1\,0\,0 \\ 1\,0\,0\,0 \end{pmatrix}.
$$

This polytope is geometrically a tetrahedron.

Note that Example 1 shows that the geometric structure of a permutation polytope depends on the presentation of the group that defines it (i.e., on the choice of generators). Both of the examples above are groups isomorphic to $(\mathbb{Z}/2\mathbb{Z})^2$, but their permutation polytopes are not even combinatorially isomorphic. We now provide an example of computing the Ehrhart polynomial of a permutation polytope to kick off our investigation.

Example 2. The *cyclic group* $C_n \leq S_n$ is the group generated by the permutation $(1\ 2\ \cdots\ n)$. We claim that the volume of $P(C_n)$ is $\frac{1}{(n-1)!}$ and its Ehrhart polynomial of $P(C_n)$ is $\binom{t+n-1}{n-1}$. Indeed, since the matrices associated to the elements of C_n have disjoint support, there is a lattice isomorphism between $P(C_n)$ and Δ_n, the standard $(n-1)$-simplex $\mathrm{conv}\{e_i \mid 1 \leq i \leq n\} \subseteq \mathbb{R}^n$. It is easy to see that the number of integer points in Δ_n is the number of nonnegative integer solutions to the diophantine equation $x_1 + x_2 + \cdots + x_n = t$, which is $\binom{t+n-1}{n-1}$. The volume follows.

We now introduce our first class of permutation polytopes to investigate. The *dihedral group* $D_n \leq S_n$ is the group generated by the permutations $r = (1\ 2\ \cdots\ n)$ and $f = (1\ n)(2\ n-1)\cdots(\lfloor \frac{n+1}{2} \rfloor\ \lceil \frac{n+1}{2} \rceil)$. In Sect. 3, we determine particular unimodular triangulations of $P(C_n)$ and $P(D_n)$ with respect to the lattices $\mathrm{aff}(P(C_n)) \cap \mathbb{Z}^{n\times n}$ and $\mathrm{aff}(P(D_n)) \cap \mathbb{Z}^{n\times n}$ respectively. This allows us to recover their volumes via their Ehrhart polynomials.

Theorem 1. *Let n be an integer, $n \geq 2$.*

1. *If n is odd, the volume of $P(D_n)$ is $\frac{n}{(2n-2)!}$. The Ehrhart polynomial of $P(D_n)$ is*

$$
\sum_{k=0}^{n-2} \binom{2n}{k+1}\binom{t-1}{k} + \sum_{k-1}^{2n-2} \left(\binom{2n}{k+1} - \binom{n}{k-n+1} \right) \binom{t-1}{k}.
$$

2. *If n is even, $n = 2m$, the volume of $P(D_n)$ is $\frac{n^2}{4\cdot(2n-3)!}$. The Ehrhart polynomial of $P(D_n)$ is*

$$
\sum_{k=0}^{m-2} \binom{2n}{k+1}\binom{t-1}{k} - \sum_{k=m-1}^{2m-2} \left(\binom{2n}{k+1} - 2\binom{2n-m}{k+1-m} \right) \binom{t-1}{k}
$$

$$
+ \sum_{k=2m-1}^{4m-3} \left(\binom{2n}{k+1} - 2\binom{2n-m}{k+1-m} + \binom{2n-2m}{k+1-2m} \right) \binom{t-1}{k}.
$$

In Sect. 4, we study *Frobenius polytopes* as introduced by Collins and Perkinson in [12]. These are permutation polytopes $P(G)$ where G is a *Frobenius group*. A group $G \leq S_n$ is Frobenius if it has a proper subgroup H such that for all $x \in G \setminus H$, $H \cap (xHx^{-1}) = \{e\}$. The special subgroup H is known as the *Frobenius complement* of G and is unique up to conjugation. Moreover, every Frobenius group $G \leq S_n$ has a special proper subgroup N of size n called the *Frobenius kernel* which consists of the identity and all elements of G that have no fixed points; see Chap. 16 of [2]. The Frobenius kernel and Frobenius complement have trivial intersection, and $G = NH$. The class of Frobenius groups includes semi-direct products of cyclic groups, some matrix groups over finite fields, the alternating group A_4, and many others. See [38] for more on Frobenius groups. We determine triangulations of Frobenius polytopes and a formula for their normalized volumes, in particular showing that the normalized volumes are completely characterized by the size of the Frobenius complement and the size of the Frobenius kernel.

Theorem 2. *Let $G \leq S_n$ be a Frobenius group with Frobenius complement H and Frobenius kernel N. The normalized volume of $P(G)$ in the sublattice of $\mathbb{Z}^{n \times n}$ spanned by its vertices is*

$$\frac{1}{(|H||N| - |H|)!} \sum_{\ell=0}^{\lfloor \frac{|H|(|N|-1)-1}{|N|} \rfloor} \binom{(|H|-\ell)|N|}{(|H|-\ell)|N| - |H| + 1} \binom{|H|-1}{\ell} (-1)^\ell.$$

In order to better understand these polytopes, we can approximate them with a sequence of special convex bodies. In the 1980s, L. Lovász approximated the stable set polyhedra from graph theory using a convex body called the *theta body*; see [26]. In [21], Gouveia et al. generalize Lovász's theta body for 0/1 polytopes (that is, polytopes whose vertices have coordinates of zero and one) to generate a sequence of semidefinite programming relaxations of the convex hull of the common zeroes of a set of real polynomials; see [26, 27] and Sect. 2 for more on this topic. We briefly study the theta body hierarchy of our permutation polytopes. For instance, we prove that convergence of the first iterate always occurs for Frobenius groups. This implies many structural results, such as the existence of reverse lexicographic unimodular triangulations. See [36] for more on this. The conditions of being two-level and TH_1 exact are equivalent for 0/1-varieties, as shown in Sect. 2. Our main result regarding TH_1-exactness for Frobenius groups is the following.

Proposition 1. *If $G \leq S_n$ is a Frobenius group, then $P(G)$ is two-level and hence G is TH_1-exact.*

In Sect. 5, we study miscellaneous permutation polytopes. We begin by developing a method for computing the Ehrhart polynomial of $P(G)$ when G is the automorphism group of a rooted binary tree on n vertices. This method relates the Ehrhart polynomials of permutation polytopes associated to direct products and wreath products of groups to the Ehrhart polynomials of the individual permutation polytopes themselves. A key theorem in this regard is the following. This method relates the Ehrhart polynomial $i(P(G), t)$ of $P(G)$ (in the variable t) to that of Ehrhart polynomials at smaller evaluations.

Theorem 3. *Let $G \leq S_n$, and $G \wr S_2$ be the wreath product of G with the symmetric group S_2. Then*

$$i(P(G \wr S_2), t) = \sum_{k=0}^{t} i^2(P(G), k) \cdot i^2(P(G), t-k)$$

for any integer $t \geq 2$.

We continue our study of miscellaneous permutation polytopes by studying $P(A_n)$, where $A_n \leq S_n$, the *alternating group* on $\{1, 2, \ldots, n\}$, consists of permutations with even signature. One of the main focuses in the literature is on determining the facets of $P(A_n)$. Cunningham and Wang [14], and independently Hood and Perkinson [24], proved that $P(A_n)$ has exponentially many facets in n, resolving a problem of Brualdi and Liu [8]. However, a full facet description is still not known. Moreover, no polynomial time algorithm in n is known for membership in $P(A_n)$. The difficulty of attaining a description of all facets of these polytopes is demonstrated by the following proposition, which shows that the first iterate of the theta body hierarchy for the polytopes $P(A_n)$ is almost never equal to $P(A_n)$ itself.

Proposition 2. *The polytope $P(A_n)$ is two-level, and hence A_n is TH_1-exact, if and only if $n \leq 4$. Moreover, for $n \geq 8$, $P(A_n)$ is at least $(\lfloor \frac{n}{4} \rfloor + 1)$-level.*

We conclude the paper with computations of volumes and Ehrhart polynomials of permutation polytopes for many subgroups of S_3, S_4, and S_5.

2   Preliminaries

Given a convex d-dimensional polytope $P \subset \mathbb{R}^n$, its *Ehrhart polynomial* $i(P, t)$ is the function that counts the number of points in $\mathbb{Z}^n \cap tP$, where $tP = \{tX \mid X \in P\}$ is the tth dilation of P. It is well known that the normalized volume of P is the leading coefficient of the polynomial $i(P, t)$ (this number is a multiple of the usual Euclidean volume of P when P is full-dimensional). In order to determine the Ehrhart polynomial or the normalized volume of a lattice polytope P, it is often useful to know something about the triangulations and subdivisions of P (see [17]): Let $V = \{v_1, v_2, \ldots, v_r\}$ denote the vertices of P in $\mathbb{R}^n$. A *subdivision* of P is a collection T of subsets of V, called *cells*, whose convex hulls form a polyhedral complex with support P. If each cell in T is a simplex, then T is a *triangulation* of P. We are interested on a special kind of triangulation. Let L be an affine lattice $L \subset \mathbb{R}^n$. A simplex with vertices $v_1, v_2, \ldots, v_m \in L$ is said to be *unimodular* in the lattice L if $\{v_m - v_1, v_{m-1} - v_1, \ldots, v_2 - v_1\}$ is a basis for the lattice L. A polytope whose vertices lie in L is said to have a *unimodular triangulation* in L if it has a triangulation in which all maximal dimensional simplices are unimodular in L. We will be interested in unimodularity with respect to two kinds of lattices. We say that a polytope P is P-unimodular if it has a unimodular triangulation in the lattice spanned by the vertices of P, and we will say that P is $\mathbb{Z}$-unimodular if it has a

unimodular triangulation in the lattice $\text{aff}(P) \cap \mathbb{Z}^n$. The following lemma shows that if a polytope has a $\mathbb{Z}$-unimodular triangulation and the number of faces of each dimension in the triangulation is known, then its Ehrhart polynomial, and hence its volume, can be determined immediately.

Lemma 1 (See Theorem 9.3.25 in [17]). *Let $P \subset \mathbb{R}^n$ be a lattice polytope. Assume that P has a $\mathbb{Z}$-unimodular triangulation with f_k faces of dimension k. Then the Ehrhart polynomial of P is*

$$i(P, t) = \sum_{k=0}^{n} \binom{t-1}{k} f_k.$$

In order to compute triangulations of our polytopes, we make repeated use of Gale duality. In what follows, let $P \subset \mathbb{R}^n$ be a d-dimensional polytope with r vertices $V = \{v_1, v_2, \ldots, v_r\}$. Let $V \in \mathbb{R}^{n \times r}$ be the matrix given by

$$\begin{pmatrix} v_1 & v_2 & \cdots & v_r \end{pmatrix}. \tag{1}$$

Let $\mathscr{G} \in \mathbb{R}^{(r-d-1) \times r}$ be a matrix whose rows form a basis for the space of linear dependences of the columns of (1). The *Gale dual* of P is the vector configuration $\{\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_r\}$ consisting of the columns of $\mathscr{G}$ (for those familiar with Gale duals, we do not affinize here because the vectors lie on a common subspace). Note that $\mathscr{G}$ is unique up to linear coordinate transformations. When constructing triangulations one can rely on using *regular triangulations*. We will compute them using Gale duality and chambers, but for completeness let us include here a different (more common) definition. Every vector $w = (w_1, \ldots, w_r)$ in $\mathbb{R}^n$ induces a subdivision of P as follows. Again let $V = \{v_1, v_2, \ldots, v_r\}$ be the vertices of P. Consider the polytope $Q_w = \text{conv}(\{(v_1, w_1), \ldots, (v_r, w_r)\})$. Generally Q_w is a polytope of dimension $dim(Q) + 1$. The *lower envelope* of Q_w is the collection of faces of the form $\{x \in Q_w | cx = c_0\}$ with Q_w contained in the halfspace $cx \leq c_0$ and the last coordinate c_{d+1} is negative. The lower envelope of Q_w is a polyhedral complex of dimension $dim(Q)$. We define T_w to be the subdivision of P whose cells are the projection of the cells of the lower envelope of Q_w. In other words, $\{v_{i_1}, v_{i_2}, \ldots, v_{i_k}\}$ is a cell of T_w if $\{(v_{i_1}, w_{i_1}), (v_{i_2}, w_{i_2}), \ldots, (v_{i_k}, w_{i_k})\}$ are the vertices of a face in the lower envelope of Q_w. We observe that for a generic choice of the vector w the subdivision T_w is in fact a triangulation of V. A subdivision of P is *regular* if it is of the form T_w for some vector w. An important example for our purposes are the *reverse lexicographic triangulations* which are obtained by taking the weight vector to be $w = (1, 1, 1, \ldots, 1)$ and when these heights do not give a triangulation one perturbs further with $(1, 1, \ldots, 1, 1, 0)$, $(1, 1, \ldots, 1, 0, 0)$, etc. (see [34] and Chap. 9 in [17] for more on weight vectors). We remark that given a regular subdivision T_w, we can find an explicit vector w of lifting heights inducing T_w by solving a linear programming problem. We will use the strong connection of regular triangulations to Gale duality soon.

The relationship between triangulations of a polytope and the structure of its Gale dual hinges on the *chamber complex* of $\mathscr{G}$. Denote by $\Sigma_{\mathscr{G}}$ the set of cones generated by all bases of $\mathscr{G}$, that is, all subsets of $\{\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_r\}$ that form bases for the column space of $\mathscr{G}$. If $\sigma \in \Sigma_{\mathscr{G}}$, let $\partial\sigma$ denote its boundary, and let $\partial\Sigma_{\mathscr{G}}$ be the union of the boundaries of all cones $\sigma \in \Sigma_{\mathscr{G}}$. The complement of $\partial\Sigma_{\mathscr{G}}$ inside the cone generated by $\{\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_r\}$ consists of open convex cones. The closure of such an open convex cone is called a *chamber*, and the *chamber complex* of $\mathscr{G}$ is the collection of all these chambers. The chamber complex of $\mathscr{G}$ and its relationship to triangulations of P is encapsulated in the following lemma.

Lemma 2 (See Theorems 5.4.5, 5.4.7, and 5.4.9 in [17]). *Let* $P \subset \mathbb{R}^n$ *be a d-dimensional polytope with vertex set* $V = \{v_1, v_2, \ldots, v_r\}$ *and Gale dual* $\{\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_r\}$. *Let τ be a chamber of the chamber complex of $\mathscr{G}$. Then*

$$\Delta = \bigcup \operatorname{conv}(V \setminus \{v_{j_1}, v_{j_2}, \ldots, v_{j_{r-d-1}}\}),$$

taken over all $\{v_{j_1}, v_{j_2}, \ldots, v_{j_{r-d-1}}\}$ *such that* $\tau \subseteq \operatorname{cone}\{\bar{v}_{j_1}, \bar{v}_{j_2}, \ldots, \bar{v}_{j_{r-d-1}}\}$ *(the cone generated by* $\{\bar{v}_{j_1}, \bar{v}_{j_2}, \ldots, \bar{v}_{j_{r-d-1}}\}$*) is a full-dimensional cone in the Gale dual, is a regular triangulation of P. Moreover, all regular triangulations of P arise in this way from some chamber τ.*

Unfortunately, the aforementioned triangulations given by the Gale dual may not be $\mathbb{Z}$- nor P-unimodular, so we still need methods to determine if a given polytope P has a $\mathbb{Z}$-unimodular or P-unimodular triangulation. One way to do this is through the use of Gröbner bases of toric ideals. Though this can be addressed in a more general setting, we will restrict ourselves to permutation polytopes arising from subgroups of a particular S_n. Let $G = \{g_1, g_2, \ldots, g_k\}$ be elements of such a subgroup, and as usual consider g_i as an $n \times n$ permutation matrix for each i. Let $\mathbb{C}[\mathbf{x}] = \mathbb{C}[x_{g_1}, x_{g_2}, \ldots, x_{g_k}]$ be the polynomial ring in k indeterminates indexed by the elements of G and let $\mathbb{C}[\mathbf{t}] := \mathbb{C}[t_{\ell m} : 1 \leq \ell, m \leq n]$. The algebra homomorphism induced by the map

$$\hat{\pi}_G : \mathbb{C}[\mathbf{x}] \to \mathbb{C}[\mathbf{t}], \ \hat{\pi}_G(x_{g_i}) = \prod_{1 \leq \ell, m \leq n} t_{\ell m}^{(g_i)_{\ell m}}, \ \ 1 \leq i \leq k$$

has as its kernel the ideal I_G. Given a monomial order $\prec$ on $\mathbb{C}[\mathbf{x}]$, the ideal I_G can determine a $P(G)$-unimodular triangulation of $P(G)$. Moreover, this triangulation is always regular. See [17, 34] for more on regular triangulations.

Lemma 3 (See Corollary 8.9 in [34] and Theorem 9.4.5 in [17]). *Let* $\operatorname{in}_{\prec}(I_G)$ *be the initial ideal of I_G with respect to the term order $\prec$. The support vectors of the generators of the radical of* $\operatorname{in}_{\prec}(I_G)$ *are the minimal non-faces of a regular triangulation of $P(G)$. Moreover,* $\operatorname{in}_{\prec}(I_G)$ *is square-free if and only if the corresponding triangulation $\Delta_{\prec}$ of $P(G)$ is $P(G)$-unimodular.*

By Lemma 3 and the theory of Gröbner bases, $P(G)$ will have a $P(G)$-unimodular triangulation if there is a term order $\prec$ on $\mathbb{C}[\mathbf{x}]$ such that the Gröbner basis of I_G is

generated by polynomials whose initial terms are square-free. This will be exploited in Sect. 4. For more on the relationship between toric ideals, Gröbner bases, and triangulations, see [34] and Chap. 9 in [17].

Recently, Gouveia, Parrilo, and Thomas [21] constructed a hierarchy of convex bodies, each given as the projection of the feasible region of a semidefinite program, approximating the convex hull of an arbitrary real variety. In Sect. 4, given a permutation polytope $P(G) \subset \mathbb{R}^{n \times n}$, we discuss the convergence of this hierarchy of relaxations $TH_1(I) \supseteq TH_2(I) \supseteq \cdots \supseteq P(G)$, which are known as theta bodies. Here, I is an ideal in $\mathbb{R}[x_{ij} : 1 \leq i, j \leq n]$ whose real variety is the vertex set of $P(G)$. This hierarchy of relaxations has the property that if $TH_k(I) = P(G)$ for some fixed k, linear optimization over $P(G)$ can be performed in polynomial time in n provided a certain algebraic oracle. Of particular interest then are polytopes for which $TH_1(I) = P(G)$, in which case we say G is TH_1-*exact*. The concept of TH_1-exactness was defined in [21] for general ideals I, but we focus on ideals whose zero-sets are vertices of permutation polytopes. TH_1-exact varieties are characterized polyhedrally in [21], and we again restrict this characterization to permutation polytopes.

Lemma 4 (See Theorem 4.2 in [21]). *The group G is TH_1-exact if and only if for every facet defining inequality $c \cdot x - \alpha \geq 0$ of $P(G)$, there is a plane $c \cdot x - \beta = 0$ parallel to $c \cdot x - \alpha = 0$ such that all vertices of $P(G)$ lie in $\{x \mid c \cdot x - \alpha = 0\} \cup \{x \mid c \cdot x - \beta = 0\}$.*

A polytope satisfying the facet property in Lemma 4 is called *compressed* or *two-level*. Using Lemma 4, we can use Gale duality to characterize groups whose permutation polytopes are TH_1-exact. We do so in the following lemma, which was proved independently by Gouveia, Parrilo, and Thomas, but we provide a self-contained proof.

Lemma 5. *Let $P(G) \subseteq \mathbb{R}^{n \times n}$ be a permutation polytope with vertex set $\{v_1, v_2, \ldots, v_r\}$ and Gale dual $\{\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_r\}$. Then G is TH_1-exact if and only if for every $J \subseteq \{1, 2, \ldots, r\}$ such that* conv$\{v_j \mid j \in J\}$ *is a facet of $P(G)$, $\sum_{j \notin J} \bar{v}_j = 0$.*

Proof. Throughout this proof, we use the equivalence of TH_1-exactness with the property of $P(G)$ being two-level, which was proved in Lemma 4. Let $J \subseteq \{1, 2, \ldots, r\}$ such that conv$\{v_j \mid j \in J\}$ is a facet of $P(G)$ with the defining inequality $c \cdot x - \alpha \geq 0$ valid on $P(G)$. Then

$$0 = (c, -\alpha) \begin{pmatrix} v_1 & v_2 & \cdots & v_r \\ 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \bar{v}_1 \\ \bar{v}_2 \\ \vdots \\ \bar{v}_r \end{pmatrix} = (c \cdot v_1 - \alpha, c \cdot v_2 - \alpha, \cdots, c \cdot v_r - \alpha) \begin{pmatrix} \bar{v}_1 \\ \bar{v}_2 \\ \vdots \\ \bar{v}_r \end{pmatrix}.$$

Since G is exact, $c \cdot v_j - \alpha$ can take at most two values. By construction, one of these values is zero. If the other value is β, then $c \cdot v_j - \alpha = \beta$ if and only if $j \notin J$. Thus $\sum_{j \notin J} \beta \bar{v}_j = 0$, which implies that $\sum_{j \notin J} \bar{v}_j = 0$ since $\beta \neq 0$.

For the converse, suppose that $\sum_{j \notin J} \bar{v}_j = 0$ for every J such that conv$\{v_j \mid j \in J\}$ is a facet. Fix such a J and assume that the facet inequality of $P(G)$ defining it is $c \cdot x - \alpha \geq 0$. Then, as we have done above,

$$0 = \sum_{j \notin J}(c \cdot v_j - \alpha)\bar{v}_j = \sum_{j \notin J}(c \cdot v_j)\bar{v}_j - \alpha \sum_{j \notin J} \bar{v}_j = \sum_{j \notin J}(c \cdot v_j)\bar{v}_j.$$

Suppose that there are at least two distinct values among $\{c \cdot v_j \mid j \notin J\}$ and that γ is the least value. Then

$$0 = \sum_{j \notin J}(c \cdot v_j - \gamma)\bar{v}_j$$

yields a positive dependence relation on $\{\bar{v}_j \mid j \notin J\}$ that does not use all the elements in the set. This contradicts the assumption that J induces a facet of $P(G)$. Thus $\{c \cdot v_j \mid j \notin J\}$ has only one element, and hence G is exact. $\square$

From Lemma 5 we see that if $P(G)$ is TH_1-exact, then for any facet F of $P(G)$, the vertices not on F lie on a common hyperplane. As a consequence, if $P(G)$ is TH_1-exact, then it contains a $P(G)$-unimodular triangulation. Moreover, as we will see, all simplices in this triangulation have the same volume.

3 Dihedral Groups

We dedicate this section to the proof of Theorem 1. The dimension of permutation polytopes arising from dihedral groups was computed in [33]. We omit the proof here, because it follows immediately from the proof of Lemma 7 where we compute the set of linear relations among the elements of D_n.

Lemma 6 (See Theorem 4.1 of [33]). *The dimension of the polytope $P(D_n)$ is $2n - 2$ if n is odd and $2n - 3$ if n is even.*

Lemma 6 indicates that Gale duality is very useful for determining the Ehrhart polynomial of $P(D_n)$, since the Gale dual lies in a space of dimension $|D_n| - \dim(P(D_n)) - 1$, which is one if n is odd and two if n is even.

Lemma 7. *If n is odd, the Gale dual of $P(D_n)$ is a vector configuration in $\mathbb{R}$ consisting of n copies of each of the vectors ± 1. If n is even, $n = 2m$, the Gale dual of $P(D_n)$ is the vector configuration in $\mathbb{R}^2$ consisting of m copies of each of the four vectors $[\pm 1, 0]^T, [0, \pm 1]^T$.*

Proof. Throughout this proof, let $\mathscr{G}$ be the matrix whose columns form the Gale dual of $P(D_n)$ with its columns indexed by

$$\{e, r, r^2, \ldots, r^{n-1}, f, fr, fr^2, \ldots, fr^{n-1}\}$$

in that order. The following linear relation holds for D_n:

$$e + r + r^2 + \ldots + r^{n-1} = f + fr + fr^2 + \ldots + fr^{n-1} = J_{n \times n}, \quad (2)$$

where $J_{n \times n}$ is the $n \times n$ matrix whose entries are all one. When n is odd, Lemma 6 implies that $P(D_n)$ is $2n - 2$ dimensional, so the Gale dual of $P(D_n)$ is one dimensional. Thus, Equation (2) implies that

$$\mathscr{G} = \begin{pmatrix} 1 & 1 & \cdots & 1 & -1 & -1 & \cdots & -1 \end{pmatrix},$$

with n copies of 1 and n copies of -1. When n is even, $n = 2m$, Lemma 6 implies that $P(D_n)$ is $2n - 3$ dimensional, so the Gale dual of $P(D_n)$ is two dimensional. We observe that the relation

$$\sum_{j=0}^{m-1} r^{2j+1} = \sum_{j=0}^{m-1} fr^{2j} \quad (3)$$

holds for D_n when n is even. The linear relations (2) and (3) are linearly independent, so we deduce that

$$\mathscr{G} = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 & -1 & \ldots & -1 \\ 0 & 1 & \ldots & 1 & -1 & 0 & \ldots & 0 \end{pmatrix}.$$

We conclude that the Gale dual is the vector configuration in $\mathbb{R}^2$ consisting of n copies of each of the four vectors $[\pm 1, 0]^T, [0, \pm 1]^T$. $\qquad \square$

We now compute the Ehrhart polynomial of $P(D_n)$ and proceed as follows. First, we show in Proposition 3 that $P(D_n)$ has a $P(D_n)$-unimodular regular triangulation Δ, and then show that this triangulation must be a $\mathbb{Z}$-unimodular triangulation (see Proposition 4). Now, the structure of the Gale dual of $P(D_n)$ shows us that, by Lemma 2, the number of faces in any dimension in any two regular triangulations of $P(D_n)$ is the same. Thus, we shall find the number of faces of every dimension in an arbitrary regular triangulation of $P(D_n)$, and hence the number of faces of every dimension in Δ, and then apply Lemma 1 to recover the Ehrhart polynomial.

Proposition 3. *The polytope $P(D_n)$ has a $P(D_n)$-unimodular regular triangulation.*

Proof. Let G be the graph with vertices $\{1, 2, \ldots, n\}$ and edges $i, i + 1$ for each $i \in \{1, 2, \ldots, n - 1\}$ along with an edge from n to 1. Let A_G be the adjacency matrix of G. Consider the polytope

$$P_G = \left\{ X \in [0, 1]^{n \times n} \ : \ A_G X = X A_G, \ \sum_{j=1}^{n} X_{ij} = 1 \ \forall i, \ \sum_{i=1}^{n} X_{ij} = 1 \ \forall j \right\}.$$

The integer points of P_G are permutations commuting with A_G, so they are precisely the automorphisms of G. Since the automorphism group of G is D_n and P_G is integral (see Theorem 2 of [37]), this implies that $P_G = P(D_n)$. But by Theorem 4.4 of [16], the vertex set of P_G is exact (i.e., it is the zero set of a TH_1-exact ideal), which by Theorem 2.4 of [36] and Theorem 4.2 of [21] implies that every reverse lexicographic triangulation of $P(D_n)$ is $P(D_n)$-unimodular. Since every point set has a reverse lexicographic triangulation, and it is known to be regular (see [34] and Chap. 9 in [17]), the result follows. $\square$

In order to establish that $P(D_n)$ is $\mathbb{Z}$-unimodular, we prove that the index of the lattice generated by its vertices in the lattice $\mathrm{aff}(P(D_n)) \cap \mathbb{Z}^{n \times n}$ is one.

Proposition 4. *The lattice generated by the vertices of $P(D_n)$ has index one in the lattice $\mathrm{aff}(D_n) \cap \mathbb{Z}^{n \times n}$.*

Proof. First, consider when n is odd. For simplicity, let D_n consist of the matrices $v_1, v_2, \ldots, v_{2n}$, where $v_{2i+1} = r^i$ and v_{2i+2} is the unique flip in D_n fixing $i + 1$, $0 \le i \le n-1$. It suffices to prove that if $X \in \mathbb{Z}^{n \times n}$ and X is an $\mathbb{R}$-linear combination of the matrices $\{v_{2n} - v_1, v_{2n-1} - v_1, \ldots, v_2 - v_1\}$, then X is a $\mathbb{Z}$-linear combination of these matrices. Assume then that $X = \sum_{j=2}^{2n} \alpha_j (v_1 - v_j) = \left(\sum_{j=2}^{2n} \alpha_j \right) v_1 - \sum_{j=2}^{2n} \alpha_j v_j$, with $\alpha_j \in \mathbb{R}$. Let $\alpha \in [0, 1)$ such that $\sum_{j=2}^{2n} \alpha_j + \alpha \in \mathbb{Z}$. Since e and v_{2i+2} are the only elements of D_n with the $(i + 1, i + 1)$-entry in their support, and since X has integer entries, we conclude that $\alpha_{2i+2} - \alpha \in \mathbb{Z}$ for all $0 \le i \le n - 1$. Moreover, for any i, there is a unique flip with the $(1, i + 1)$-entry in its support. Since r^i is the only rotation with the $(1, i+1)$-entry in its support, and again since X has integer entries, we deduce that $\alpha_{2i+1} + \alpha \in \mathbb{Z}$ for all $0 \le i \le n - 1$. Now recall from Equation (2) in the proof of Lemma 7 that $\sum_{i=0}^{n-1} v_{2i+1} - \sum_{i=0}^{n-1} v_{2i+2} = 0$, so we have that

$$X = \left(\sum_{j=2}^{2n} \alpha_j \right) v_1 - \sum_{j=2}^{2n} \alpha_j v_j - \alpha \left(\sum_{i=0}^{n-1} v_{2i+1} - \sum_{i=0}^{n-1} v_{2i+2} \right)$$

$$= \sum_{i=1}^{n-1} (\alpha_{2i+1} + \alpha)(v_1 - v_{2i+1}) + \sum_{i=0}^{n-1} (\alpha_{2i+2} - \alpha)(v_1 - v_{2i+2}),$$

and hence X is a $\mathbb{Z}$-linear combination of $\{v_{2n} - v_1, v_{2n-1} - v_1, \ldots, v_2 - v_1\}$. A similar analysis can be done in the case n is even, and we leave this as an exercise to the reader. $\square$

We now determine the number of faces of each dimension in a particular triangulation of $P(D_n)$. This together with Lemma 1 and Propositions 3 and 4 proves parts (2) and (3) of Theorem 1.

Proof (Theorem 1). First consider when n is odd. By Lemma 7, the Gale dual of $P(D_n)$ consists of the vectors $\{e_1^{(1)}, e_1^{(2)}, \ldots, e_1^{(n)}, -e_1^{(1)}, -e_1^{(2)}, \ldots, -e_1^{(n)}\}$ where the $e_1^{(i)}, -e_1^{(i)}$ are copies of the vectors $e_1, -e_1$ in $\mathbb{R}$ respectively, $1 \le i \le n$. The set

consisting of the vector e_1 is the only extreme ray in one chamber in the Gale dual, so by Lemma 2, $P(D_n)$ has a triangulation Δ with maximal dimensional simplices $\{\text{conv}\{G \backslash \{r^i\}\} | \ 1 \leq i \leq n\}$. The number of $(k+1)$-element subsets of G is $\binom{2n}{k+1}$. By Lemma 2, of these subsets, the ones that are not simplices in Δ are those that contain all of $\{e, r, r^2, \ldots, r^{n-1}\}$. There are precisely $\binom{2n-n}{k+1-n}$ such subsets, so we conclude that the number of k-dimensional faces f_k in Δ is

$$f_k = \binom{2n}{k+1} - \binom{n}{k+1-n}.$$

By the symmetry in the Gale dual, this is also the number of k-dimensional faces in a reverse lexicographic triangulation of $P(D_n)$, which is $P(D_n)$-unimodular by Proposition 3 and hence $\mathbb{Z}$-unimodular by Proposition 4. The Ehrhart polynomial follows from Lemma 1. Moreover, we see that $f_{2n-2} = \binom{2n}{2n-1} - \binom{n}{n-1} = n$, so the volume of $P(D_n)$ is $\frac{n}{(2n-2)!}$.

Now consider $P(D_n)$ when n is even, $n = 2m$. By Lemma 7, the Gale dual of $P(D_n)$ consists of the copies $\{e_1^{(i)}, e_2^{(i)}, -e_1^{(i)}, -e_2^{(i)} \ | \ 1 \leq i \leq m\}$ of $e_1, e_2, -e_1, -e_2$ respectively in $\mathbb{R}^2$. Consider the chamber of the Gale dual whose extreme rays are the vectors $\{e_1, e_2\}$. By Lemma 2, this chamber gives the regular triangulation Δ of $P(D_n)$ whose maximal dimensional simplices are $\{\text{conv}\{G \backslash \{r^{2i-1}, r^{2j}\}\} | \ 1 \leq i, j \leq n\}$. By a similar counting argument as in the odd case, we conclude that

$$f_k = \binom{2n}{k+1} - \binom{2}{1}\binom{2n-m}{k+1-m} + \binom{2n-2m}{k+1-2m}.$$

Again, since Lemma 1 implies that $P(D_n)$ has a $P(D_n)$-unimodular triangulation and hence by Proposition 3 a $\mathbb{Z}$-unimodular triangulation with the same face numbers, the Ehrhart polynomial follows by Lemma 1. Lastly, we see that the volume of $P(D_n)$ when n is even is f_{2n-3}, which is

$$\frac{1}{(2n-3)!}\left(\binom{2n}{2n-2} - 2\binom{2n-m}{2n-m-2} + \binom{2n-m}{2n-2m-2}\right) = \frac{n^2}{4 \cdot (2n-3)!}.$$

$\square$

4　Frobenius Groups

In this section, we discuss triangulations and normalized volumes of Frobenius polytopes, leading to a proof of Theorem 2. We also establish that all Frobenius groups are exact, hence proving Proposition 1. For the remainder of this section, we assume that $G \leq S_n$ is a Frobenius group. We let $N = \{u_1, u_2, \ldots, u_n\}$ be its Frobenius kernel ($n = |N|$), and we let $H = \{v_1, v_2, \ldots, v_h\}$ be its

Frobenius complement ($h = |H|$). We assume throughout that H is the set of coset representatives for N in G. We let $\mathscr{G}$ denote the matrix whose columns form the Gale dual of $P(G)$. Recall that $G = NH$ and $H \cap N = \{e\}$, and so G consists of the nh matrices

$$u_1 v_1, u_2 v_1, \ldots, u_n v_1, u_1 v_2, u_2 v_2, \ldots, u_n v_2, \ldots, u_1 v_h, u_2 v_h, \ldots, u_n v_h$$

and we index the columns of $\mathscr{G}$ by G in this order. The following lemmas are proven in [12].

Lemma 8 (See Proposition 4.2 in [12]). *If $G \leq S_n$ is Frobenius, then $\sum_{i=1}^{n} u_i v_j = J_{n \times n}$ for all j, $1 \leq j \leq h$, where $J_{n \times n}$ is the $n \times n$ matrix of all 1s.*

Lemma 9 (See Corollary 4.5 in [12]). *If $G \leq S_n$ is Frobenius, the dimension of $P(G)$ is $|G| - |H|$.*

Lemma 8 gives us the $|H| - 1$ linearly independent relations $\sum_{i=1}^{n} u_i v_1 = \sum_{i=1}^{n} u_i v_j$, $2 \leq j \leq h$. The dimension formula in Lemma 9 tells us that the $|H| - 1$ relations in Lemma 8 actually form a basis for the space of linear dependences of G. As a consequence, we get the Gale dual of $P(G)$.

Proposition 5. *The Gale dual of $P(G)$ is the vector configuration consisting of n copies $\{\mathbf{1}^{(1)}, \mathbf{1}^{(2)}, \ldots, \mathbf{1}^{(n)}\}$ of the all-ones vector $\mathbf{1}$ in $\mathbb{R}^{h-1}$, together with n copies $\{-e_i^{(1)}, -e_i^{(2)}, \ldots, -e_i^{(n)}\}$ in $\mathbb{R}^{h-1}$ of $-e_i$ for $1 \leq i \leq h - 1$, where e_i is the ith standard basis vector. In particular, the $u_i v_j$ column of the matrix $\mathscr{G}$ is the vector $\mathbf{1}$ if $j = 1$ and $-e_{j-1}$ otherwise.*

Proof. This follows directly from Lemmas 8 and 9.   □

Now consider the chamber in the Gale dual whose extreme rays are $\{-e_1, -e_2, \ldots, -e_{h-1}\}$. From Lemma 2, $P(G)$ has a corresponding regular triangulation Δ whose maximal dimensional simplices are

$$\Delta = \left\{ \operatorname{conv} \{ G \setminus \{ u_{i_1} v_2, u_{i_2} v_3, \ldots, u_{i_{h-1}} v_h \} \} \mid 1 \leq i_j \leq n \right\}. \tag{4}$$

Furthermore, from the structure of the Gale dual as given by Proposition 5, all regular triangulations of $P(G)$ (and hence all triangulations, since the Gale dual tells us all of them are regular) have the same number of k-dimensional faces for any k. Thus, if we can determine a $P(G)$-unimodular triangulation of $P(G)$ and count the number of faces of dimension k for each k in the triangulation Δ, we can prove Theorem 2. We proceed by showing that $P(G)$ has a $P(G)$-unimodular triangulation and then by determining the number of faces of given dimensions in Δ.

Proposition 6. *If G is Frobenius, then G has a $P(G)$-unimodular triangulation.*

Proof. Our proof appeals to toric algebra. Let $A \in \mathbb{R}^{n^2 \times |G|}$ be the matrix whose columns are the elements of G written as n^2-dimensional column vectors by reading rows left to right and top to bottom. We index the columns of A by the elements of

G as in $\mathscr{G}$. The toric ideal $I_G \subseteq \mathbb{C}[\mathbf{x}] = \mathbb{C}[x_{u_r v_s} : 1 \leq r \leq n, \ 1 \leq s \leq h]$ is the kernel of the homomorphism

$$\hat{\pi} : \mathbb{C}[\mathbf{x}] \to \mathbb{C}[\mathbf{t}], \ \ \hat{\pi}(x_{u_r v_s}) = \prod_{1 \leq \ell, m \leq n} t_{\ell m}^{(u_r v_s)_{\ell m}},$$

and by Lemma 4.1 of [34], $I_G = \langle x^u - x^v \mid A(u - v) = 0, \ u, v \in \mathbb{Z}^{|G|}\rangle$. By Lemmas 8 and 9, $\ker(A)$ has the basis $\{b_1, b_2, \ldots, b_k\}$ where $b_i = e_{u_1} + e_{u_2} + \cdots + e_{u_n} - e_{u_1 v_i} - e_{u_2 v_i} - \cdots - e_{u_n v_i}$ for each i. Now if $u - v \in \ker(A)$ is integral, then $u - v = \sum_{i=1}^{h} \lambda_i b_i$, where $\lambda_i \in \mathbb{Q}$ for each i. In fact, $\lambda_i \in \mathbb{Z}$ for each i since the $u_\ell v_i$ component of $u - v$ is $\pm \lambda_i$. We conclude by Corollary 4.4 of [34] that $I_G = \langle x_{H_1} - x_{H_\ell} : 2 \leq \ell \leq h\rangle$ where $x_{H_\ell} = \prod_{i=1}^{n} x_{u_i v_\ell}$ for each ℓ.

In fact, $\{x_{H_1} - x_{H_\ell} : 2 \leq \ell \leq h\}$ is a Gröbner basis for I_G with respect to the reverse lexicographic order $\prec$; here, $u_{r_1} v_{s_1}$ comes lexicographically before $u_{r_2} v_{s_2}$ if and only if $r_1 \leq r_2, s_1 \leq s_2$. To see this, we use Buchberger's criterion. For an introduction to this criterion, Buchberger's algorithm and details of terms to follow, particularly S-pairs, see [13]. The essential idea is that, to check whether a set of polynomials F is a Gröbner basis for the ideal they generate with respect to a term order, all that one needs to check is whether some very special binomials (the S-pairs), defined from the possible pairs of elements of F, reduce to zero modulo the division algorithm by F.

Consider any pair of polynomials $f_r = x_{H_1} - x_{H_r}, f_s = x_{H_1} - x_{H_s}$ in our generating set for I_G. With respect to $\prec$, we compute the S-pair $S(f_r, f_s)$ and see that

$$S(f_r, f_s) = \frac{x_{H_r} x_{H_s}}{-x_{H_r}}(x_{H_1} - x_{H_r}) - \frac{x_{H_r} x_{H_s}}{-x_{H_s}}(x_{H_1} - x_{H_s}) = x_{H_1} x_{H_r} - x_{H_1} x_{H_s}.$$

Now since $x_{H_1} x_{H_r} - x_{H_1} x_{H_s} = x_{H_1}(x_{H_1} - x_{H_s}) - x_{H_1}(x_{H_1} - x_{H_r})$, we see that $\overline{S(f_r, f_s)}^{f_r f_s} = 0$ (this denotes the residue of the S-pair modulo the ideal [13]). Since r, s were arbitrary, Buchberger's algorithm concludes that $\{x_{H_1} - x_{H_\ell} : 2 \leq \ell \leq h\}$ is a Gröbner basis for I_G. By Lemma 3, we conclude that $P(G)$ has a $P(G)$-unimodular triangulation. $\qquad\square$

We now proceed to prove Theorem 2.

Proof (Theorem 2). By Proposition 6, $P(G)$ has a $P(G)$-unimodular triangulation, and by Proposition 5 and Lemma 2, all triangulations of $P(G)$ have the same face numbers. Thus it suffices to determine the number of top dimensional faces in the triangulation Δ in (4) and apply Lemma 1. We more generally determine the number of k-dimensional faces f_k for each k. Any k-simplex in Δ must be a subset of some maximal dimensional simplex of Δ, and by Lemma 2, all maximal dimensional simplices in Δ do not contain $\{u_1 v_i, u_2 v_i, \ldots, u_n v_i\}$ as a subset for any $i \geq 2$. Conversely, if a $(k + 1)$-element subset of G does not contain $\{u_1 v_i, u_2 v_i, \ldots, u_n v_i\}$ as a subset for any $i \geq 2$, then there exists m_i for each $i \geq 2$ such that $u_{m_i} v_i$ is not in the given subset, and this $(k + 1)$-element subset is therefore a k-simplex that is

a face of the maximal dimensional simplex $\operatorname{conv}\{G\backslash\{u_{m_1}v_2, u_{m_2}v_3, \ldots, u_{m_{h-1}}v_h\}\}$. We conclude that a $(k+1)$-element subset of G is a k-simplex in Δ if and only if it does not contain $\{u_1v_i, u_2v_i, \ldots, u_nv_i\}$ as a subset for any $i \geq 2$. Thus, to determine f_k, we need to count the number of $(k+1)$-element subsets of G that do not contain $\{u_1v_i, u_2v_i, \ldots, u_nv_i\}$ as a subset for any $i \geq 2$.

Let us call a subset of the form $\{u_1v_i, u_2v_i, \ldots, u_nv_i\}$ a *complete copy*. There are $\binom{(h+1)n}{k+1}$ $(k+1)$-element subsets of G, and the number of such subsets that contain ℓ complete copies as subsets is $\binom{hn-\ell n}{k+1-\ell n}\binom{h-1}{\ell}$. Thus by inclusion-exclusion

$$f_k = \sum_{\ell \geq 0} \binom{(h-\ell)n}{k+1-\ell n}\binom{h-1}{\ell}(-1)^\ell.$$

Since each maximal dimensional simplex in Δ has volume $\frac{1}{\dim(P(G))!}$, the result follows. $\qquad\square$

We now establish that Frobenius groups are two-level. This relies on an important lemma in [12].

Lemma 10 (See Corollary 4.5 in [12]). *The complement of any set of $|H|$ elements of G, one chosen from each of the cosets of N, forms the set of vertices of a facet of $P(G)$. All facets of $P(G)$ arise this way.*

Proof (Theorem 1). Let $J \subseteq G$ be the set of vertices of a facet of G. Choose H to be the set of coset representatives of N. By Lemma 10, $J = G\backslash\{u_i, u_iv_1, u_iv_2, \ldots, u_iv_h\}$ for some fixed i. Now let $\mathbf{1}$ be the all ones vector in $\mathbb{R}^{h-1}$ and let e_i be the standard basis vectors. Then we have

$$\sum_{j \notin J} \bar{j} = \bar{u_i} + \overline{u_iv_1} + \overline{u_iv_2} + \cdots + \overline{u_iv_h} = \mathbf{1} - e_1 - e_2 - \cdots - e_k = 0.$$

Since J was arbitrary, we conclude by Lemma 5 that $P(G)$ is two-level and thus TH_1-exact. $\qquad\square$

5 Miscellaneous Permutation Polytopes

In this section, we study a potpourri of permutation polytopes. We begin by presenting a method for computing the Ehrhart polynomials of groups that arise as automorphism groups of finite rooted binary trees. Recall that a finite *rooted binary tree* is a simple finite graph with no cycles, all of whose vertices have degree at most 3, with a distinguished vertex called the *root* that has degree at most 2 (see Fig. 1 for an example). The crux of this method lies in Theorem 3. We first introduce some necessary group theoretic preliminaries. For any groups $G \leq S_m$, $H \leq S_n$, the *direct product* $G \times H \leq S_m \times S_m \leq S_{m+n}$ consists of elements

$\{(g, h) : g \in G, h \in H\}$ with product $(g_1, h_1) \cdot (g_2, h_2) = (g_1 g_2, h_1 h_2)$. By construction, the vertices of the permutation polytope of $G \times H$ are block matrices of the form $\{g \oplus h : g \in G, h \in H\}$. The *wreath product* of G by S_n, denoted $G \wr S_n$, is the group $\{(g, h) : g \in G^n, h \in S_n\}$ under the operation defined by

$$(g', h') \cdot (g, h) = ((g'_1, g'_2, \ldots, g'_n), h') \cdot ((g_1, g_2, \ldots, g_n), h)$$
$$:= ((g'_{h'(1)} g_1, g'_{h'(2)} g_2, \ldots, g'_{h'(n)} g_n), h'h).$$

Remark 1. The vertices of the permutation polytope $P(G \wr S_n)$ are the $mn \times mn$ matrices $\{g \otimes h : g \in G, h \in S_n\}$. Moreover, it is shown in [25] and [28] that $P(G \wr S_n)$ is affinely isomorphic to free join of $|S_n|$ copies of G, however this equivalence is not necessarily a lattice equivalence, and hence the Ehrhart polynomial of $P(G \wr S_n)$ can not necessarily be computed by known methods such as convolution (see [23]).

We now prove that automorphism groups of rooted binary trees are always composed of direct products and wreath products of groups.

Lemma 11. *Let G be the automorphism group of a rooted binary tree T. Then G can be written as a sequence of direct products of groups and wreath products by symmetric groups of order at most two.*

Proof. Label the vertices of T by the positive integers $\{1, 2, \ldots, n\}$ such that the root vertex is labeled 1. First assume the root of T has one child and without loss of generality assume its label is 2. Letting T_2 be the subtree of T rooted at 2, we have $\text{Aut}(T) = S_1 \times \text{Aut}(T_2)$. Now assume instead that the root has two children that are labeled 2 and 3 without loss of generality. Let T_2 be the subtree of T rooted at 2 and T_3 be the subtree of T rooted at 3. If T_2 and T_3 are not isomorphic, then $\text{Aut}(T) = S_1 \times (\text{Aut}(T_2) \times \text{Aut}(T_3))$. If T_2 and T_3 are isomorphic, then $\text{Aut}(T) = S_1 \times (\text{Aut}(T_2) \wr S_2)$. The result then follows inductively. □

The proof of Lemma 11 indicates that computing the Ehrhart polynomial of groups arising as automorphism groups of rooted binary trees requires repeated computation of Ehrhart polynomials of direct products and wreath products by symmetric groups of order two. Theorem 3 indicates how Ehrhart polynomials behave under wreath products by symmetric groups of order two, and we prove this theorem now.

Proof (Theorem 3). The vertices of the polytope $P(G \wr S_2)$ are precisely the matrices

$$\left\{ \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}, \begin{pmatrix} 0 & X_1 \\ X_2 & 0 \end{pmatrix} \right\}, \tag{5}$$

where the X_i are vertices of $P(G)$. Thus, $P(G \wr S_2)$ is lattice isomorphic to $(P(G) \times P(G)) \otimes (P(G) \times P(G))$, the free join of $P(G) \times P(G)$ with itself. The result then follows by Lemma 1.3 in [23]. □

Fig. 1 A rooted binary tree
with root 1
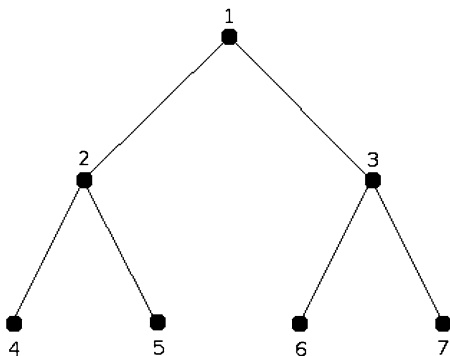
Theorem 3 gives us a method for computing Ehrhart polynomials and hence volumes of permutation polytopes from groups arising as automorphism groups of rooted binary trees. First, given a rooted binary tree T, we compute the automorphism group $\mathrm{Aut}(T)$ as a sequence of direct products and wreath products. Then we read the group $\mathrm{Aut}(T)$ from left to right. If we encounter a direct product, we compute the Ehrhart polynomials of the corresponding groups and take the product of the polynomials. If we encounter a wreath product, we apply Theorem 3. This produces the Ehrhart polynomial of the permutation polytope associated to the tree T.

Example 3. Consider the tree T shown in Fig. 1. Let T_2 and T_3 be the subtrees rooted at 2 and 3 respectively. Notice that $\mathrm{Aut}(T_2) = \mathrm{Aut}(T_3)$ because T is in fact unlabeled (we only place labels to illustrate how to compute the automorphism group). The automorphism group of T is therefore $S_1 \times [(\mathrm{Aut}(T_2)) \wr S_2]$. Thus, by Theorem 3, its Ehrhart polynomial is

$$i(P(Aut(T)), t)$$

$$= i(P(S_1), t) \cdot \left(\sum_{k=0}^{t} i^2(P(\mathrm{Aut}(T_2)), k) \cdot i^2(P(\mathrm{Aut}(T_2)), t - k) \right)$$

$$= 1 \cdot \left(\sum_{k=0}^{t} (k+1)^2 \cdot (t-k+1)^2 \right)$$

$$= \sum_{k=0}^{t} (k+1)^2 \cdot (t-k+1)^2$$

$$= \sum_{k=0}^{t} k^4 + (-2t) \sum_{k=0}^{t} k^3 + (t^2 - 2t - 2) \sum_{k=0}^{t} k^2 + (t^2 - 1) \sum_{k=0}^{t} k + (t+1)^2$$

$$= \frac{1}{30}(t+1)(t+2)(t^2 + 4t + 5).$$

By Theorem 1, this is precisely the Ehrhart polynomial of D_4, which we should expect, since $\mathrm{Aut}(T)$ is $S_1 \times D_4$ up to a relabeling of the generating set of D_4. Moreover, we conclude that the volume of $P(Aut(T))$ is $\frac{1}{30}$.

We can further prove that for any rooted T, $P(\mathrm{Aut}(T))$ has a $P(\mathrm{Aut}(T))$-unimodular regular triangulation.

Proposition 7. *If T is a rooted tree, then $P(Aut(T))$ has a $P(Aut(T))$-unimodular regular triangulation.*

Proof. Let T be any rooted tree, and let A_T be its adjacency matrix. Consider the polytope

$$P_T = \left\{ X \in [0,1]^{n \times n} \ : \ A_T X = X A_T, \ \sum_{j=1}^{n} X_{ij} = 1 \ \forall i, \ \sum_{i=1}^{n} X_{ij} = 1 \ \forall j \right\}.$$

The integer points of P_T are permutations commuting with A_T, so they are precisely the automorphisms in $\mathrm{Aut}(T)$. Since P_T is integral (see Theorem 2 of [37]), this implies that $P_T = P(\mathrm{Aut}(T))$. But by Theorem 4.4 of [16], P_T is exact, which by Theorem 2.4 of [36] and Theorem 4.2 of [21] implies that any reverse lexicographic triangulation of $P(\mathrm{Aut}(T))$ is $P(\mathrm{Aut}(T))$-unimodular. Since reverse lexicographic triangulations are regular, the result follows. □

We continue our study of miscellaneous permutation polytopes by looking at the alternating group, in particular proving Proposition 2. The following proposition shows the difficulty of dealing with general permutation polytopes.

Proof (Proposition 2). Since $P(A_2)$ and $P(A_3)$ have one and three vertices respectively, they are trivially two-level. Since A_4 is a Frobenius group, Proposition 1 implies that $P(A_4)$ is two-level. For $n \geq 5$, by choosing $(\sigma, t, h) = (e, 1, 2)$ as in Theorem 3 of [14], we deduce that $P(A_n)$ has the facet defining inequality $\ell(x) \leq n-2$, where $\ell(x) = \sum_{j=3}^{n} x_{j,j} + \sum_{j=3}^{n} x_{j,1} + \sum_{j=3}^{n} x_{1,j}$. Now $\ell(e) = n-2$, $\ell((1\ 2)(4\ 5)) = n - 4$, and $\ell((3\ 4\ 5)) = n - 5$, and hence $P(A_n)$ is not two-level for $n \geq 5$. To show that $P(A_n)$ is at least $(\lfloor \frac{n}{4} \rfloor + 1)$-level for $n \geq 8$, we evaluate ℓ on σ_i, where $\sigma_0 = e$ and $\sigma_k = (1\ 2)(3\ 4) \cdots (4k-1\ 4k)$ for $1 \leq k \leq \lfloor \frac{n}{4} \rfloor$. □

To conclude the paper, the following table lists the subgroups of S_3, S_4, and S_5 and some of their Ehrhart polynomials. Two groups stand out as incomplete, the alternating group A_5 and the general affine group of degree one over the field of five elements. The latter group is generated by taking the semidirect product of the additive and multiplicative groups of the field of five elements and is denoted by $GA(1,5)$.

Subgroups of S_3

Order	Generators	Dim	Ehrhart polynomial
1	$\langle e \rangle \cong \{e\}$	0	1
2	$\langle (1\,2) \rangle \cong C_2$	1	$t+1$
3	$\langle (1\,2\,3) \rangle \cong C_3$	2	$\frac{1}{2}t^2 + \frac{3}{2}t + 1$
6	$\langle (1\,2),(1\,3),(2\,3) \rangle \cong S_3$	4	$\frac{1}{8}t^4 + \frac{3}{4}t^3 + \frac{15}{8}t^2 + \frac{9}{4}t + 1$

Subgroups of S_4

Order	Group	Dim	Ehrhart polynomial
1	$\langle e \rangle \cong \{e\}$	0	1
2	$\langle (1\,2) \rangle \cong C_2$	1	$t+1$
2	$\langle (1\,2)(3\,4) \rangle \cong C_2$	1	$t+1$
3	$\langle (1\,2\,3) \rangle \cong C_3$	2	$\frac{1}{2}t^2 + \frac{3}{2}t + 1$
4	$\langle (1\,2),(3\,4) \rangle \cong C_2 \times C_2$	2	$t^2 + 2t + 1$
4	$\langle (1\,2)(3\,4),(1\,3)(2\,4) \rangle \cong C_2 \times C_2$	3	$\frac{1}{6}t^3 + t^2 + \frac{11}{6}t + 1$
4	$\langle (1\,2\,3\,4) \rangle \cong C_4$	3	$\frac{1}{6}t^3 + t^2 + \frac{11}{6}t + 1$
6	$\langle (1\,2),(1\,3),(2\,3) \rangle \cong S_3$	4	$\frac{1}{8}t^4 + \frac{3}{4}t^3 + \frac{15}{8}t^2 + \frac{9}{4}t + 1$
8	$\langle (1\,2\,3\,4),(1\,2)(3\,4) \rangle \cong D_4$	5	$\frac{1}{30}t^5 + \frac{1}{3}t^4 + \frac{4}{3}t^3 + \frac{8}{3}t^2 + \frac{79}{30}t + 1$
12	$\langle (1\,2\,3),(1\,2\,4),(1\,3\,4),(2\,3\,4) \rangle \cong A_4$	9	$\frac{1}{5670}t^9 + \frac{1}{504}t^8 + \frac{23}{1890}t^7 + \frac{1}{15}t^6 + \frac{173}{540}t^5 + \frac{9}{8}t^4 + \frac{29797}{11340}t^3 + \frac{1199}{315}t^2 + \frac{383}{126}t + 1$
24	$\langle (1\,2),(1\,3),(1\,4),(2\,3),(2\,4),(3\,4) \rangle \cong S_4$	9	$\frac{11}{11340}t^9 + \frac{11}{630}t^8 + \frac{19}{135}t^7\,\frac{1}{2}t^6 + \frac{1109}{540}t^5 + \frac{43}{10}t^4 + \frac{35117}{5670}t^3 + \frac{379}{63}t^2 + \frac{65}{18}t + 1$

Subgroups of S_5

Order	Generators	Dim	Ehrhart polynomial
1	$\langle e \rangle \cong \{e\}$	0	1
2	$\langle (1\,2) \rangle \cong C_2$	1	$t+1$

(continued)

Subgroups of S_5 (continued)

Order	Generators	Dim	Ehrhart polynomial
2	$\langle (1\,2)(3\,4)\rangle \cong C_2$	1	$t + 1$
3	$\langle (1\,2\,3)\rangle \cong C_3$	2	$\frac{1}{2}t^2 + \frac{3}{2}t + 1$
4	$\langle (1\,2), (3\,4)\rangle \cong C_2 \times C_2$	2	$t^2 + 2t + 1$
4	$\langle (1\,2)(3\,4), (1\,3)(2\,4)\rangle \cong C_2 \times C_2$	3	$\frac{1}{6}t^3 + t^2 + \frac{11}{6}t + 1$
4	$\langle (1\,2\,3\,4)\rangle \cong C_4$	3	$\frac{1}{6}t^3 + t^2 + \frac{11}{6}t + 1$
5	$\langle (1\,2\,3\,4\,5)\rangle \cong C_5$	4	$\frac{1}{24}t^4 + \frac{5}{12}t^3 + \frac{35}{24}t^2 + \frac{25}{12}t + 1$
6	$\langle (1\,2\,3)(4\,5)\rangle \cong C_6$	3	$\frac{1}{2}t^3 + 2t^2 + \frac{5}{2}t + 1$
6	$\langle (1\,2), (2\,3), (1\,3)\rangle \cong S_3$	4	$\frac{1}{8}t^4 + \frac{3}{4}t^3 + \frac{15}{8}t^2 + \frac{9}{4}t + 1$
6	$\langle (1\,2)(4\,5), (1\,3)(4\,5), (2\,3)(4\,5)\rangle \cong S_3$	5	$\frac{1}{40}t^5 + \frac{1}{8}t^4 + \frac{5}{8}t^3 + \frac{15}{8}t^2 + \frac{47}{20}t + 1$
8	$\langle (1\,2\,3\,4), (1\,2)(3\,4)\rangle \cong D_4$	5	$\frac{1}{30}t^5 + \frac{1}{3}t^4 + \frac{4}{3}t^3 + \frac{8}{3}t^2 + \frac{79}{30}t + 1$
10	$\langle (1\,2\,3\,4\,5), (2\,5)(3\,4)\rangle \cong D_5$	8	$\frac{1}{8064}t^8 + \frac{5}{2016}t^7 + \frac{5}{192}t^6 + \frac{25}{144}t^5 + \frac{95}{128}t^4 + \frac{575}{288}t^3 + \frac{6515}{2016}t^2 + \frac{475}{168}t + 1$
12	$\langle (1\,2\,3)(4\,5), (1\,2)(4\,5)\rangle \cong D_6$	5	$\frac{1}{8}t^5 + \frac{7}{8}t^4 + \frac{21}{8}t^3 + \frac{33}{8}t^2 + \frac{13}{4}t + 1$
12	$\langle (1\,2\,3), (1\,3\,4), (2\,3\,4)\rangle \cong A_4$	9	$\frac{1}{5670}t^9 + \frac{1}{504}t^8 + \frac{23}{1890}t^7 + \frac{1}{15}t^6 + \frac{173}{540}t^5 + \frac{9}{8}t^4 + \frac{29797}{11340}t^3 + \frac{1199}{315}t^2 + \frac{383}{126}t + 1$
20	$\langle (1\,2\,3\,4\,5), (1\,2\,4\,3)\rangle \cong GA(1,5)$	9	Too large to compute; volume$= \frac{19}{6538371840}$

$\langle(1\,2), (1\,3), (1\,4), (2\,3), (2\,4), (3\,4)\rangle \cong S_4$	24	9	$\frac{11}{11340}t^9 + \frac{11}{630}t^8 + \frac{19}{135}t^7 + \frac{2}{3}t^6 + \frac{1109}{540}t^5 + \frac{43}{10}t^4 + \frac{35117}{5670}t^3 + \frac{379}{63}t^2 + \frac{65}{18}t + 1$
$\langle(1\,2\,3), (1\,2\,4), (1\,2\,5), (1\,3\,4), (1\,3\,5), (1\,4\,5),$ $(2\,3\,4), (2\,3\,5), (2\,4\,5), (3\,4\,5)\rangle \cong A_5$	60	16	Too large to compute
$\langle(1\,2), (1\,3), (1\,4), (1\,5), (2\,3), (2\,4), (2\,5), (3\,4),$ $(3\,5), (4\,5)\rangle \cong S_5$	120	16	$\frac{188723}{836911595520}t^{16} + \frac{188723}{2092278988}t^{15} + \frac{1008757}{5977939968}t^{14} + \frac{112655}{57480192}t^{13} + \frac{72750523}{4598415360}t^{12} + \frac{984101}{10450944}t^{11} + \frac{125188639}{292626432}t^{10} + \frac{55426325}{36578304}t^9 + \frac{3541860299}{836075520}t^8 + \frac{196563587}{20901888}t^7 + \frac{3812839477}{229920768}t^6 + \frac{6441118435}{28740096}t^5 + \frac{438177965089}{17435658240}t^4 + \frac{3028287247}{145297152}t^3 + \frac{6229735}{494208}t^2 + \frac{725}{144}t + 1$

References

1. Ahmed, M.: Polytopes of magic labelings of graphs and the faces of the birkhoff polytope. Ann. Comb. **12**(3), 241–269 (2008)
2. Alperin, J.L., Bell, R.B.: Groups and Representations. Volume 162 of Graduate Texts in Mathematics. Springer, New York (1995)
3. Baumeister, B., Haase, C., Nill, B., Paffenholz, A.: On permutation polytopes. Adv. Math. **222**(2), 431–452 (2009)
4. Beck, M., Pixton, D.: The ehrhart polynomial of the birkhoff polytope. Discret. Comput. Geom. **30**(4), 623–637 (2003)
5. Bernstein, D.N., Kouchnirenko, A.G., Khovanskii, A.G.: Newton polyhedra. Uspenskhi Mat. Nauk. **31**, 201–202 (1976)
6. Brualdi, R.A.: Some applications of doubly stochastic matrices. Linear Algebra Appl. **107**, 77–100 (1988)
7. Brualdi, R.A., Gibson, P.M.: Convex polyhedra of double stochastic matrices. II. Graph of U_n. J. Comb. Theory B **22**(2), 175–198 (1977)
8. Brualdi, R.A., Liu, B.L.: The polytope of even doubly stochastic matrices. J. Comb. Theory A **57**(2), 243–253 (1991)
9. Canfield, E.R., McKay, B.D.: The asymptotic volume of the birkhoff polytope. Online J. Anal. Comb. (4), Art. 2, 4, (2009)
10. Chan, C.S., Robbins, D.P.: On the volume of the polytope of doubly stochastic matrices. Exp. Math. **8**(3), 291–300 (1999)
11. Chan, C.S., Robbins, D.P., Yuen, D.S.: On the volume of a certain polytope. Exp. Math. **9**(1), 91–99 (2000)
12. Collins, J., Perkinson, D.: Frobenius polytopes. Available at http://arxiv.org/abs/1102.0988 (2010)
13. Cox, D., Little, J., O'Shea, D.: Ideals, Varieties, and Algorithms. Undergraduate Texts in Mathematics, 3rd edn. Springer, New York (2007)
14. Cunningham, W., Wang, Y.: On the even permutation polytope. Linear Algebra Appl. **381**, 269–281 (2004)
15. De Loera, J.A., Liu, F., Yoshida, R.: A generating function for all semi-magic squares and the volume of the birkhoff polytope. J. Algebraic Comb. **30**(1), 113–139 (2009)
16. De Loera, J.A., Hillar, C.J., Malkin, P.N., Omar, M.: Recognizing graph theoretic properties with polynomial ideals. Electron. J. Comb. **17**(1), Research Paper 114, 26 (2010)
17. De Loera, J.A., Rambau, J., Santos, F.: Triangulations: Structures for Algorithms and Applications. Volume 25 of Algorithms and Computation in Mathematics. Springer, Berlin (2010)
18. Diaconis, P., Gangolli, A.: Rectangular arrays with filxed margins. In: Discrete Probability and Algorithms (Minneapolis, MN, 1993). Volume 72 of IMA Volumes in Mathematics and Its Applications, pp 15–41. Springer, New York (1995)
19. Diaconis, P.L., Wood, M.: Random doubly stochastic tridiagonal matrices. Random Struct. Algorithms (2012, to appear)

20. Gelfand, I.M., Graev, M.I., Postnikov, A.: Combinatorics of hypergeometric functions associated with positive roots. In: The Arnold-Gelfand Mathematical Seminars, pp. 205–221. Birkhäuser, Boston (1997)
21. Gouveia, J., Parrilo, P., Thomas, R.R.: Theta bodies for polynomial ideals. SIAM J. Optim. **20**(4), 2097–2118 (2010)
22. Guralnick, R.M., Perkinson, D.: Permutation polytopes and indecomposable elements in permutation groups. J. Comb. Theory A **113**(7), 1243–1256 (2006)
23. Henk, M., Tagami, M.: Lower bounds on the coefficients of ehrhart polynomials. Eur. J. Comb. **30**(1):70–83 (2009)
24. Hood, J., Perkinson, D.: Some facets of the polytope of even permutation matrices. Linear Algebra Appl. **381**, 237–244 (2004)
25. Joswig, M., Lutz, F.H.: One-point suspensions and wreath products of polytopes and spheres. J. Comb. Theory A **110**(2), 193–216 (2005)
26. Lovász, L.: Stable sets and polynomials. Discret. Math. **124**(1–3), 137–153 (1994). Graphs and combinatorics (Qawra, 1990)
27. Lovász, L., Schrijver., A.: Cones of matrices and set-functions and 0-1 optimization. SIAM J. Optim. **1**(2), 166–190 (1991)
28. McMahan, P.: Wreath product polytopes. Undergraduate thesis, Reed College (2003)
29. Mirsky, L.: Even doubly stochastic matrices. Math. Ann. **144**, 418–421 (1961)
30. Pak, I.: Four questions on Birkhoff polytope. Ann. Comb. **4**(1), 83–90 (2000)
31. Schreck, H., Tinhofer, G.: A note on certain subpolytopes of the assignment polytope associated with circulant graphs. Linear Algebra Appl. **111**, 125–134 (1988)
32. Stanley, R.P.: Two poset polytopes. Discrete Comput Geom **1**:9–23 (1986)
33. Steinkamp, H.: Convex polytopes and permutation matrices. Undergraduate thesis, Reed College (1999)
34. Sturmfels, B.: Gröbner Bases and Convex Polytopes. Volume 8 of University Lecture Series. American Mathematical Society, Providence (1996)
35. Sturmfels, B.: Solving Systems of Polynomial Equations. Volume 97 of CBMS Regional Conference Series in Mathematics. Conference Board of the Mathematical Sciences, Washington (2002)
36. Sullivant, S.: Compressed polytopes and statistical disclosure limitation. Tohoku Math. J. (2) **58**(3), 433–445 (2006)
37. Tinhofer, G.: Graph isomorphism and theorems of birkhoff type. Computing **36**(4), 285–300 (1986)
38. Wielandt, H.: Finite Permutation Groups. Translated from the German by R. Bercov. Academic, New York (1964)
39. Zeilberger, D.: Proof of a conjecture of Chan, Robbins, and Yuen. Electron. Trans. Numer. Anal. **9**, 147–148 (electronic) (1999). Orthogonal polynomials: numerical and symbolic algorithms (Leganés, 1998).

Monotone Paths in Planar Convex Subdivisions and Polytopes[*]

Adrian Dumitrescu, Günter Rote, and Csaba D. Tóth

Abstract Consider a connected subdivision of the plane into n convex faces where every vertex is incident to at most Δ edges. Then, starting from every vertex there is a path with at least $\Omega(\log_\Delta n)$ edges that is monotone in some direction. This bound is the best possible. Consider now a connected subdivision of the plane into n convex faces where exactly k faces are unbounded. Then, there is a path with at least $\Omega(\log(n/k)/\log\log(n/k))$ edges that is monotone in some direction. This bound is also the best possible. Our methods are constructive and lead to efficient algorithms for computing monotone paths of lengths specified above. In 3-space, we show that for every $n \geq 4$, there exists a polytope P with n vertices, bounded vertex degrees, and triangular faces such that every monotone path on the 1-skeleton of P has at most $O(\log^2 n)$ edges. We also construct a polytope Q with n vertices, and triangular faces, (with unbounded degree however), such that every monotone path on the 1-skeleton of Q has at most $O(\log n)$ edges.

A. Dumitrescu (✉)
Department of Computer Science, University of Wisconsin–Milwaukee, 3200 N.
Cramer Street, Milwaukee, WI 53201, USA
e-mail: dumitres@uwm.edu

G. Rote
Institut für Informatik, Freie Universität Berlin, Takustraße 9, 14195 Berlin, Germany
e-mail: rote@inf.fu-berlin.de

C.D. Tóth
Department of Mathematics and Statistics, University of Calgary, Room MS 432,
2500 University Drive NW, Calgary, AB T2N 1N4, Canada
e-mail: cdtoth@ucalgary.ca

1 Introduction

A *geometric graph* $G = (V, E)$ in Euclidean d-space is a set V of distinct points (vertices) in Euclidean d-space $\mathbb{R}^d$, and a set E of line segments (edges) between vertices such that no vertex lies in the relative interior of any edge. For our investigation, it is convenient to define an *extended geometric graph* $G = (V, E)$, where E may also contain rays, each emitted by a vertex, and lines (disjoint from vertices). A directed path p in an extended geometric graph G is *monotone* (resp., *weakly monotone*) if there exists a unit vector $\mathbf{u}$ such that the inner product $\mathbf{e} \cdot \mathbf{u}$ is positive (non-negative) for every directed edge $\mathbf{e}$ of p. In $\mathbb{R}^2$, in particular, the direction of a unit vector $\mathbf{u} = (\cos \theta, \sin \theta)$ is determined by the angle $\theta \in (-\pi, \pi]$. A directed path p is *x-monotone* (resp., *y-monotone*) if it is monotone in direction 0 (resp., $\frac{\pi}{2}$). The *size* (or *length*) of a path is the number of edges in the path, or equivalently, one plus the number of vertices on the path. Notice that any path (monotone or not) in an extended geometric graph contains at most two rays.

We study monotone paths in the 1-skeletons of polytopes and convex subdivisions. The *1-skeleton* $G(P)$ of a bounded polytope P in $\mathbb{R}^d$ is the geometric graph formed by the vertices and edges of P. Similarly, the *1-skeleton* $G(P)$ of an unbounded polytope P is the extended geometric graph formed by the vertices and edges of P.

A *convex subdivision* (for short, *subdivision*) of $\mathbb{R}^d$ is a set Π of (bounded or unbounded) convex polytopes (called *faces*) that tile $\mathbb{R}^d$. The 1-skeleton of a subdivision Π of the plane $\mathbb{R}^2$ is the extended geometric graph $G(\Pi)$ whose vertices are the points incident to three or more edges, and whose edges are the line segments, rays, and lines lying on the common boundary of two faces. To exclude some trivial cases, we always consider convex subdivisions whose 1-skeleton is connected, referred to as *connected subdivisions* for short. We are looking for long monotone paths in the 1-skeletons of polytopes and convex subdivisions of the plane. Our results are the following.

Theorem 1. *Let Π be a connected subdivision of the plane into n convex cells in which every vertex is incident to at most Δ edges. Then, for every vertex v, there is a weakly monotone path with at least $c \log_\Delta n$ edges starting from v, where $c > 0$ is an absolute constant. Apart from the constant c, this bound is the best possible.*

Theorem 2. *Let Π be a connected subdivision of the plane into n convex cells, k of which are unbounded with $n > k \geq 3$. Then $G(\Pi)$ contains a monotone path with at least $c \log \frac{n}{k} / \log \log \frac{n}{k}$ edges, where $c > 0$ is an absolute constant. Apart from the constant c, this bound is the best possible.*

We also consider long monotone paths in the 1-skeleton of a convex polytope in 3-space. We present two constructions, one with bounded vertex degrees and one with arbitrary degrees.

Theorem 3. *For every $n \geq 4$, there is a polytope P in $\mathbb{R}^3$ with n vertices, bounded vertex degrees, and triangular faces such that every monotone path in $G(P)$ has at most $O(\log^2 n)$ edges.*

Theorem 4. *For every $n \geq 4$, there is a polytope Q in $\mathbb{R}^3$ with n vertices and triangular faces such that every monotone path in $G(Q)$ has at most $O(\log n)$ edges.*

We do not know whether the bounds in Theorems 3 and 4 are asymptotically tight. The diameter of a bounded degree graph on n vertices is $\Omega(\log n)$. It follows that every monotone path connecting a diametral pair of vertices of a polytope with n vertices of bounded degree has $\Omega(\log n)$ edges. If the maximum vertex degree of the polytope is not bounded, then a lower bound of $\Omega(\log n / \log \log n)$ follows from the result of Chazelle et al. [4] (see below), applied to the dual graph of a plane projection, using reciprocal diagrams and the Maxwell-Cremona correspondence.

Related work. It is well known that the classical simplex algorithm in linear programming produces a monotone path on the 1-skeleton of a d-dimensional polytope of feasible solutions; it is called a *parametric simplex path*. According to the old *monotone Hirsch conjecture* [17], for any vector **u**, the 1-skeleton of every d-dimensional polytope with n facets contains a **u**-monotone path of at most $n - d$ edges from any vertex to a **u**-maximal vertex. For the weakly monotone version, counterexamples have already been found by Todd [15] in the 1980s. Recent counterexamples for this conjecture found by Santos [14] show that the monotone variant is also false. It is not known whether the Hirsch conjecture can be relaxed so that it holds when the length $n - d$ is replaced by some appropriate polynomial in d and n.

The *monotone upper bound problem* posed by Klee [8] asks for the maximum number of edges $M(d, n)$ of a monotone path on the 1-skeleton of a polytope with $n \geq d + 1$ facets in $\mathbb{R}^d$. An obvious upper bound for $M(d, n)$ (within 1) is the maximum number of vertices such a polytope can have, as given by the Upper Bound Theorem [9]. Pfeifle [10] and Pfeifle and Ziegler [11] have shown that $M(d, n)$ always attains this bound for $1 \leq d \leq 4$, but it does not always reach it for $d = 6$.

Balogh et al. [3] showed that there is a convex subdivision Π_n generated by n lines in the plane with $O(n^2)$ faces such that $G(\Pi_n)$ contains a monotone path of length $\Omega(n^2 / c^{\sqrt{n}})$ for some constant $c > 1$. In such a monotone path, many consecutive edges are collinear. Dumitrescu [5] proved that every simple arrangement of n lines admits a monotone path of length at least n in the convex subdivision generated by n lines. On the other hand, for every $n \geq 2$, there exists a line arrangement in which no monotone path is longer than $3n/2 + O(\log n)$. Again, consecutive collinear edges are allowed in such a path. If these are disallowed, the lower bound is unchanged but the upper bound drops to $4n/3 + O(\log n)$. As regarding algorithms

for monotone paths, Arkin et al. [1] devised a polynomial-time algorithm that computes an x-monotone path between two given points in the plane in the presence of polygonal obstacles, or reports that none exists.

A *monotone face sequence* in a convex subdivision Π is a sequence of faces such that there is a direction **u** such that any two consecutive faces, f_1 and f_2, are adjacent and a vector of direction **u** crosses their common boundary from f_1 to f_2. Chazelle et al. [4] showed that in a subdivision of the plane into n convex faces in which every face is adjacent to at most d other faces, there is a monotone face sequence of length

$$\Omega(\log_d n + \log n / \log \log n), \tag{1}$$

and this bound is tight. Moreover, a monotone face sequence of this length can even be achieved by faces stabbed by a line. The latter result was generalized to d-dimensions by Tóth [16]: for every subdivision of $\mathbb{R}^d$ into n convex faces, there is a line that stabs $\Omega((\log n / \log \log n)^{\frac{1}{d-1}})$ faces, and this bound is the best possible.

Motivation. We were led to the above result of Chazelle et al. [4] by the complexity analysis of an algorithm for partial matching between two finite planar point sets under translations [12]. There, one could show that a certain subdivision Q contained no monotone face sequence longer than some polynomial in the given parameters. If the lower bound $\Omega(\log_d n + \log n / \log \log n)$ could be strengthened to a polynomial bound of the form $\Omega(n^\alpha)$ for some $\alpha > 0$, this would have implied a polynomial bound on the number of faces of Q.

If P is the vertical projection of a piecewise linear convex terrain $\hat{P}$ in $\mathbb{R}^3$, one can apply a polarity transform to this terrain with respect to the paraboloid $z = x^2 + y^2$, yielding another piecewise linear convex terrain $\hat{P}^*$. The projection P^* of this terrain is a *reciprocal diagram* of P (cf. [2]): its graph is dual to the graph of P, in the sense that vertices of P^* correspond to faces of P and vice-versa. Each edge in P^* has a corresponding edge in P, and moreover, these two edges are perpendicular. (This last property distinguishes a reciprocal diagram from a general drawing of the dual graph of P.) A monotone path in P becomes a monotone face sequence in P^*.

Thus, for subdivisions P that are projections of convex terrains, the question about long monotone paths is completely answered by applying the tight bound of (1) for $Q = P^*$. However, for general subdivisions, the problems are not directly related, and in fact, the answers are different: We will see in Sect. 2 that one cannot add any other term that grows to infinity (such as the term $\Omega(\log n / \log \log n)$ in (1)) to the lower bound in Theorem 1.

Outline. We start with the proof of Theorem 1 in Sect. 2. We study convex subdivisions of simple polygons in Sect. 3.1. The tools developed there are instrumental in the proof of Theorem 2 in Sect. 3.2. The proofs of Theorems 3 and 4 regarding monotone paths on polytopes appear in Sects. 4 and 5, respectively. We conclude with some open problems in Sect. 6.

2 Proof of Theorem 1

Lower bound. The lower bound in Theorem 1 follows from the following lemma in a straightforward way (by counting, or inductively).

Lemma 1. *Let v be a vertex in a connected convex planar subdivision Π. Then $G(\Pi)$ contains a spanning tree rooted at v such that all paths starting at v are weakly monotone.*

Proof. For a generic direction **u**, we define the *rightmost path* $R(\mathbf{u})$ starting at v as follows; see Fig. 1(left): start at v and always follow the rightmost outgoing edge that is weakly monotone in direction **u** until we arrive at an unbounded ray.

Now we start rotating **u** clockwise. At some direction $\mathbf{u}'$, $R(\mathbf{u}')$ will be different from $R(\mathbf{u})$. At that point, $R(\mathbf{u})$ is still weakly monotone in direction $\mathbf{u}'$. Now, any vertex w (and any edge) in the region between $R(\mathbf{u})$ and $R(\mathbf{u}')$ can be reached by a weakly monotone path in direction $\mathbf{u}'$. Indeed, simply start at w and go monotonically in the direction $-\mathbf{u}'$ until reaching $R(\mathbf{u})$ or $R(\mathbf{u}')$. From there, follow $R(\mathbf{u})$ or $R(\mathbf{u}')$ to v. In this way, we can form a spanning tree of all vertices between $R(\mathbf{u})$ and $R(\mathbf{u}')$ with the desired properties.

Continuing the rotation in this way, we eventually reach all vertices and all infinite rays.

The subdivision in Fig. 2 shows that the lemma does not hold with (strictly) monotone paths. However, if there are no angles of 180°, the statement extends to strictly monotone paths.

Let v be a vertex in a connected convex planar subdivision Π. By Lemma 1, $G(\Pi)$ contains a spanning tree rooted a v such that all paths starting at v are weakly monotone. The maximum degree in $G(\Pi)$ is at most Δ. Hence the spanning tree contains a path of size $\Omega(\log_{\Delta} n)$ from v to some vertex of $G(\Pi)$.
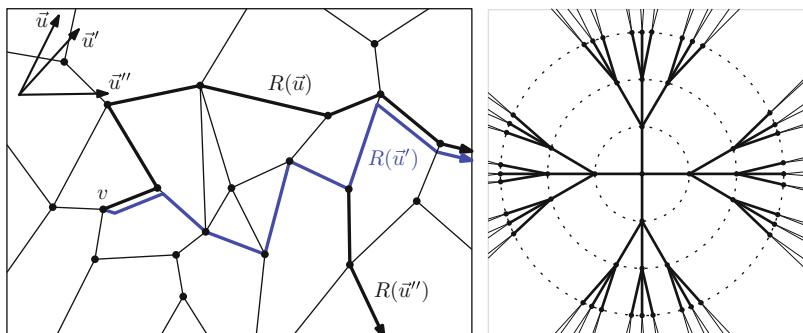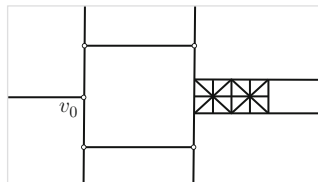

Fig. 1 *Left*: The rightmost path $R(\mathbf{u})$ starting from vertex v in direction **u**. *Right*: An balanced tree of maximum degree 4, which is the extended graph $G(\Pi)$ of a convex subdivision Π

Fig. 2 A convex subdivision
where only the five vertices
marked with *empty dots* can
be reached from v_0 along
(strictly) monotone paths

Upper bound. If the maximum degree Δ is $\Theta(n)$, Theorem 1 gives only a trivial statement. Dividing the plane into n convex sectors, by n rays starting from the origin shows that, indeed, there is no non-constant lower bound on the length of monotone paths in this case.

This construction generalizes for any $\Delta \in \mathbb{N}$, and $n \geq 3$. We construct a convex subdivision Π such that $G = G(\Pi)$ is a tree with maximum degree Δ and diameter $O(\log_\Delta n)$. See Fig. 1(right). We first construct the tree G, and then obtain a convex subdivision Π by attaching two rays to each leaf of G. Arrange the vertices of G on concentric circles of radii $0, 1, 2, \ldots, \lceil \log_\Delta n \rceil$. Place the root vertex v_0 at the center, and arrange Δ children of v_0 on the circle of radius 1 such that v_0 lies in the convex hull of its children. For $j = 1, 2, \ldots, \lfloor \log_\Delta n \rfloor$, to each vertex on a circle of radius j, attach $\Delta - 1$ children and place them on the circle or radius $j + 1$. Inductively, we can ensure that the cyclic order of the vertices on circle j corresponds to the cyclic order of the direction vectors of the edges that connect them to the previous circle $j - 1$. Note that the leaves of G are located on the largest circle, and we obtain a convex subdivision Π with $G = G(\Pi)$ by adding two infinite rays to each leaf of G.

3 Proof of Theorem 2

3.1 Monotone Paths in Simple Polygons

Monotone polygons. We start by introducing some notation for simple polygons in $\mathbb{R}^2$. A *polygonal domain* (for short, *polygon*) P is a closed set in the plane bounded by a piecewise linear simple closed curve. A polygon P is monotone if its boundary is the union of two paths, which are both monotone with respect to a vector $\mathbf{u}$. In particular, P is y-monotone if it is bounded by two y-monotone paths. A *convex subdivision* of a polygon P is a set $\Pi = \Pi(P)$ of pairwise disjoint open convex sets (called *faces*) such that the union of their closures is P. The faces in Π together with the complement $\bar{P} = \mathbb{R}^2 \setminus P$ (the *outer face*) form a (nonconvex) subdivision of the plane $\Pi \cup \{\bar{P}\}$. We also define a geometric graph $G(\Pi) = G(\Pi(P))$, where the *vertices* are the union of all vertices of P and the set of points incident to three or more faces in $\Pi \cup \{\bar{P}\}$; and the *edges* are the line segments lying on the common boundaries of two faces in $\Pi \cup \{\bar{P}\}$.

A simple but crucial observation is that for every vertex v of $G(\Pi)$ lying in the interior of P and every direction $\mathbf{u}$, there is an edge vw such that $\mathbf{u} \cdot \vec{vw} \geq 0$, otherwise the face incident to v in direction $\mathbf{u}$ would not be convex. This implies the following.

Observation 1. *Let P be a simple polygon with a convex subdivision $\Pi = \Pi(P)$, let v be a vertex of $G(\Pi)$ lying in the interior of P, and $\mathbf{u}$ be a unit vector. Then*

 (i) *There is a weakly $\mathbf{u}$-monotone directed path in $G(\Pi)$ from v to some vertex on the boundary of P;*
(ii) *If $\mathbf{u}$ is not orthogonal to any edge of $G(\Pi)$, then this path is $\mathbf{u}$-monotone.*

Proof. We can construct a required path as follows. Start with a single-vertex path $p = v$, and successively append to p new edges whose inner product with $\mathbf{u}$ is nonnegative. If $G(\Pi)$ has no edges orthogonal to $\mathbf{u}$, then any weakly $\mathbf{u}$-monotone path is necessarily (strictly) $\mathbf{u}$-monotone. □

Recall that a y-monotone polygon is bounded from the left and from the right by two y-monotone directed paths. The common start (resp., end) point of the two boundary paths is called the *bottom* (resp., *top*) vertex of P.

Observation 2. *Let $\Pi = \Pi(P)$ be a convex subdivision of a y-monotone polygon P with no horizontal edges. For every vertex v of $G(\Pi)$, there is a y-monotone path from the bottom vertex of P to the top vertex of P which is incident to v.*

Proof. The claim is immediate if v is on the boundary of P. Assume that v is in the interior of P. By Observation 1, $G(\Pi)$ contains monotone paths from v to the boundary of P in the two opposite directions $\pi/2$ and $-\pi/2$. The combination of these paths (together with parts of the boundary of P, if needed) gives the required y-monotone path through v from the bottom to the top vertex of P. □

Criterion for y-monotone polygons. To prove the lower bound in Theorem 2, we constructively build a monotone path of the required length for any given convex partition. In our recursive construction, we successively subdivide a y-monotone polygon P into smaller polygons along certain y-monotone paths in $G(\Pi(P))$. Lemma 2 below provides a criterion for producing y-monotone pieces. A vertex v in a geometric graph G is called y-*maximal* (resp., y-*minimal*) if all edges of G incident to v lie in a closed halfplane below (resp., above) the horizontal line passing through v. A vertex v is y-*extremal* if it is y-maximal or y-minimal. It is clear that the boundary of a y-monotone polygon has exactly two y-extremal vertices, namely its top vertex and its bottom vertex.

Lemma 2. *Let P be a simple polygon with a convex subdivision Π such that no edge in $G = G(\Pi)$ is horizontal. Let H be a subgraph of G that contains all edges and vertices of P. Then all bounded faces of H are y-monotone polygons if and only if all y-extremal vertices of H are convex vertices of P.*

Proof. Assume that all bounded faces of H are y-monotone polygons. Suppose that H has a y-extremal vertex v. We may assume without loss of generality that
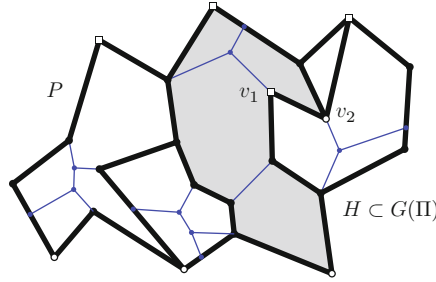
Fig. 3 The graph $G(\Pi)$ of a convex partition Π of a simple polygon P. A subgraph H in *bold* contains all edges of P. The y-maximal (resp., y-minimal) vertices of H are marked with *empty squares* (resp., *empty circles*). Vertices v_1 and v_2 are y-extremal in H, but not convex vertices of P (here they are not even vertices of P)

all edges of H incident to v are in the halfplane *above* v. Let f_v be the face of H incident to v that lies directly *below* v. Face f_v has a reflex interior angle at v, and v is neither the top nor the bottom vertex of f_v. Hence f_v is not y-monotone, and so it has to be an unbounded face of H. Since H contains all boundary edges of P, the face f_v is the unbounded face of G, as well. It follows that v is a convex vertex of P.

Assume that all y-extremal vertices of H are convex vertices of P. Consider a bounded face f of H. Let p_1 and p_2 be edge-disjoint directed paths on the boundary of f from a bottom (lowest) vertex of f to a top (highest) vertex of f (ties are broken arbitrarily), such that f lies on the right side of p_1 and on the left side of p_2. Suppose for contradiction that f is not y-monotone. We may assume without loss of generality that p_2 is not y-monotone, as in Fig. 3 (where f is the shaded face). The first (resp., last) edge of p_2 has a positive inner product with $(0, 1)$ by construction. There are two consecutive edges $\mathbf{e}_1$ and $\mathbf{e}_2$ in p_2 such that $\mathbf{e}_1 \cdot (0, 1) > 0 > \mathbf{e}_2 \cdot (0, 1)$, since otherwise p_2 would be y-monotone. Let v be the common vertex of these two edges. Since f is on the left side of both $\mathbf{e}_1$ and $\mathbf{e}_2$, these two edges are consecutive in the counterclockwise rotation order of the edges of H incident to v. Hence, H has no edge incident to v in the halfplane above v, and so v is y-extremal (y-maximal) in H. However, v is a reflex vertex of face f, hence it cannot be a convex vertex of P (specifically, v it is either interior to P, or a reflex vertex of P). This contradicts our initial assumption and completes the proof. □

Subdividing a polygon into y-monotone pieces. Our upper bound relies on the following two lemmas. In Lemma 3, we partition the bounded faces of a convex subdivision of the plane into monotone polygons. In Lemma 4, we subdivide a y-monotone polygon P into smaller y-monotone polygons which are not incident to both the top and the bottom vertex of P.

Lemma 3. *Let Π be a subdivision of the plane into n convex faces, $k \geq 3$ of which are unbounded. Then there is a subset $\Pi' \subset \Pi$ of at least $(n - k)/(k - 2)$ faces such that Π' is the convex subdivision of a monotone polygon.*

Proof. Let H denote the graph of vertices and bounded edges of $G = G(\Pi)$, without rays and lines. We proceed by induction on the number of 2-connected components, i.e., blocks of H. In the base case, H is 2-connected. Then the union of the closures of all bounded faces in Π forms a simple polygon P. Suppose that no edge of G is horizontal or vertical. Let V_0 be the set of y-extremal reflex vertices of P. If V_0 is empty, then P is a y-monotone polygon, and the $n - k$ bounded faces form a convex subdivision of P. If V_0 is nonempty, then we construct a y-monotone path $\gamma(v)$ for each $v \in V_0$, in an arbitrary order, as follows. If $v \in V_0$ is y-maximal (resp., y-minimal), then construct $\gamma(v)$ starting from v by successively appending edges in direction $\pi/2$ (resp., $-\pi/2$) until the path reaches another vertex on the boundary of P or a previously constructed path $\gamma(v')$, $v' \in V_0$. The paths $\gamma(v)$ subdivide P into $|V_0| + 1$ simple polygons, each of which is y-monotone by Lemma 2.

It remains to show that $|V_0| \leq k - 3$ after an appropriate rigid motion. Notice that every y-extremal reflex vertex of P is a y-extremal vertex of some unbounded face. An unbounded face cannot have both a top and a bottom vertex. Two of the unbounded faces, namely those containing rays in directions $(1, 0)$ and $(-1, 0)$, have neither a top nor a bottom vertex. This already implies $|V_0| \leq k - 2$. Let $\overrightarrow{e_0}$ be a ray edge of G emitted by an extremal vertex of the convex hull of P. Assume, by applying a reflection if necessary, that the unbounded face on the left of $\overrightarrow{e_0}$ is not a halfplane. Rotate the subdivision such that no edge in G is horizontal and edge $\overrightarrow{e_0}$ has the smallest positive slope. Now the unbounded face above $\overrightarrow{e_0}$ has a bottom vertex on the convex hull of P, which is not a reflex vertex of P. Therefore, we have $|V_0| \leq k - 3$.

Assume now that H has several 2-connected components. We distinguish two cases. Case 1: H is disconnected. Then G is also disconnected, and it is the disjoint union of G_1 and G_2, which are incident to k_1 and k_2 unbounded faces, respectively. Then there is at most one face (a parallel strip) incident to both G_1 and G_2. Hence $k_1 + k_2 \leq k + 1$, and so $(k_1 - 2) + (k_2 - 2) < k - 2$. Induction completes the proof as follows: Let G_1 and G_2 have $\bar{n}_1$ and $\bar{n}_2$ bounded faces, with $\bar{n}_1 + \bar{n}_2 = \bar{n} = n - k$. Then we have $\max(\frac{\bar{n}_1}{k_1-2}, \frac{\bar{n}_2}{k_2-2}) \geq \frac{\bar{n}}{k-2}$, and hence it is sufficient to find the required subdivision in G_1 or G_2. Case 2: H is connected but has a cut vertex v. Then G decomposes into subgraphs $G_1, \ldots, G_\ell$, for some $\ell \geq 2$, whose only common vertex is v. G_i can have a reflex angle incident to v; thus we augment it to a convex subdivision $\hat{G}_i$ by adding a ray incident to v. Denote by k_i the number of unbounded faces in $\hat{G}_i$ for $i = 1, \ldots, \ell$. G has ℓ unbounded faces between consecutive subgraphs around v, which are not faces in some $\hat{G}_i$. On the other hand, each $\hat{G}_i$ has two unbounded faces that are not faces of G. Hence $\sum_{i=1}^{\ell}(k_i - 2) = k - \ell \leq k - 2$. Induction completes the proof as above. $\square$

Lemma 4. *Let P be a y-monotone polygon with bottom vertex s, top vertex t, and a convex subdivision Π. Let m be the maximum size of a monotone path in $G = G(\Pi)$. Then at most two faces in Π are incident to both s and t; the remaining faces can be partitioned into at most $m + 1$ sets, each of which is the convex subdivision of a y-monotone polygon whose top or bottom vertex is not in $\{s, t\}$.*
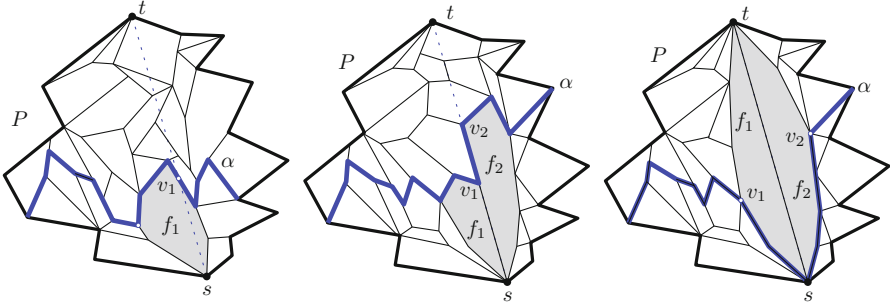
Fig. 4 A y-monotone polygon P with a *bottom* vertex s, a *top* vertex t, and a convex subdivision. *Left*: faces f_1 and f_2 are the same, but neither is incident to t. *Middle*: faces f_1 and f_2 are not incident to t. *Right*: faces f_1 and f_2 are both incident to t

Proof. If a face $f \in \Pi$ is incident to both s and t, then by convexity the closure of f contains the line segment st. Thus, at most two faces in Π are incident to both s and t. We first partition the remaining faces into two subsets, lying on opposite sides of a monotone path α such that the faces in each subset are incident to at most one of s and t. Then we further partition each of the two subsets to form y-monotone polygons. Let H_{st}^- and H_{st}^+ denote the closed halfplanes on the left and right of st, respectively. We distinguish two cases.

Case 1: no face in Π is incident to both s and t. We define two points, v_1 and v_2, in the relative interior of the segment st. Let $f_1 \in \Pi$ be a face incident to s whose closure contains an initial portion of st, and let the segment sv_1 be the intersection of the closure of f_1 with st. Refer to Fig. 4(left, middle). We may assume, by applying a reflection with respect to the y-axis if necessary, that the interior of f_1 intersects H_{st}^-. Since f_1 is not incident to t, there is an edge in G that contains v_1 and enters the interior of H_{st}^-. If there is some edge in G that contains v_1 and enters the interior of H_{st}^+ or if v_1 is on the boundary of polygon P, then let $v_2 = v_1$. Otherwise, all edges of G incident to v_1 lie in H_{st}^-, hence two consecutive edges are contained in st. Then v_1 is incident to a unique face $f_2 \in \Pi$ on the right of st. Since f_2 is not incident to both s and t, it has a vertex along st which is different from both s and t; let v_2 be an arbitrary such vertex. In summary: the segment $v_1 v_2 \subset st$ is covered by edges of G; an edge of G contains v_1 and enters the interior of H_{st}^-; and v_2 is either on the boundary of P or an edge of G contains v_2 and enters the interior of H_{st}^+.

Let **u** be a unit vector orthogonal to st. Slightly rotate **u**, if necessary, such that **u** is not parallel or orthogonal to any edge of G and $\mathbf{u} \cdot \overrightarrow{v_1 v_2} > 0$ if $v_1 \neq v_2$. By Observation 1, there is a $(-\mathbf{u})$-monotone path from v_1 to the boundary of P, and a **u**-monotone path from v_2 to the boundary of P. The union of these two paths and the segment $v_1 v_2$ forms a **u**-monotone path, denoted α, between two boundary points of P and passing through v_1 and v_2. Direct α such that its starting point is in H_{st}^-. Since the endpoints of α are on the boundary of P, every face in Π is either on left or on right side of α. By construction, every face incident to s is on the right of α, and every face incident to t is on the left of α.

By our assumption, α has at most m edges. Construct a path $\gamma(v)$ for every y-extremal interior vertex v of α, in an arbitrary order, as follows. If v is a y-maximal (resp., y-minimal) vertex in α, then construct y-monotone path $\gamma(v)$ starting from v by successively appending edges in direction $\pi/2$ (resp., $-\pi/2$) until the path reaches a vertex on α, the boundary of P or a previously constructed path $\gamma(v')$, $v' \neq v$. These paths $\gamma(v)$ together with α subdivide P into at most $(m - 1) + 2 = m + 1$ simple polygons, each of which is y-monotone by Lemma 2.

Case 2: some face in Π is incident to both s and t. Let $f_1 \in \Pi$ be a face incident to both s and t. Refer to Fig. 4(right). We may assume, by applying a reflection with respect to the y-axis if necessary, that the interior of f_1 intersects H_{st}^-. Let $\mathbf{u}$ be a unit vector orthogonal to st. Slightly rotate $\mathbf{u}$, if necessary, such that $\mathbf{u}$ is not parallel or orthogonal to any edge of G and $\mathbf{u} \cdot \overrightarrow{st} > 0$. Let v_1 be the $\mathbf{u}$-minimal vertex of f_1. We need to be more careful when defining v_2. If there is a face f_2 (possibly $f_2 = f_1$) incident to both s and t whose interior intersects H_{st}^+, then let v_2 be the $\mathbf{u}$-maximal vertex of f_2. Otherwise, if st is an edge of the face f_1 and G has vertices in the relative interior of st, then let v_2 be an arbitrary such vertex. Otherwise, st is an edge on the boundary of P, and then let $v_2 = s$.

Similarly to Case 1, construct a $(-\mathbf{u})$-monotone path from v_1 to the boundary of P, and a $\mathbf{u}$-monotone path from v_2 to the boundary of P. A $\mathbf{u}$-monotone path α is formed by the union of these two paths, a path from v_1 to s along the boundary of f_1, and a path from s to v_2 along the boundary of f_1 or f_2. Direct α such that its starting point is in H_{st}^-. Since the endpoints of α are on the boundary of P, every face in Π is either on the left or on the right side of α. By construction, every face incident to s other than f_1 and f_2 (if it exists) is on the right of α, and every face incident to t is on the left of α.

By our assumption, α has at most m edges. Construct a path $\gamma(v)$ for every y-extremal interior vertex of α that precedes v_1 or follows v_2 as in Case 1. If v_1 is an interior vertex of α, then let $\gamma(v_1)$ be the y-monotone path from v_1 to t along the boundary of f_1. Similarly, if v_2 is an interior vertex of α, then let $\gamma(v_2)$ be the y-monotone path from v_2 to t along the boundary of f_1 or f_2. Note that some of the interior vertices of α between v_1 and v_2 may be on the boundary of P. If exactly k interior vertices of α are on the boundary of P, then α subdivides P into at most $k + 2$ simple polygons. The paths γ (which are not defined for vertices on the boundary of P) further subdivide these polygons into at most $(m - 1) + 2 = m + 1$ simple polygons, each of which is y-monotone by Lemma 2.

By construction, one of these polygons is formed by the faces in Π incident to both s and t. Thus, the faces not incident to both s and t are partitioned into at most m sets, each of which is the convex subdivision of a y-monotone polygon whose top or bottom vertex is not in $\{s, t\}$. $\qquad\square$

In Lemma 4, we have partitioned almost all faces of Π into subsets that each form a y-monotone polygon. In the proof of the lower bound in Theorem 2, we will recurse on only one of these polygons.

Corollary 1. *Let P be a y-monotone polygon with bottom vertex s and top vertex t. Let Π be a subdivision of P into n convex faces. Let m be the maximum size of a monotone path in $G = G(\Pi)$. If $n \geq 3$, then there is a subset $\Pi' \subset \Pi$ of at least $n/(m+3)$ faces such that Π' is the convex subdivision of a y-monotone sub-polygon of P whose top or bottom vertex is not in $\{s, t\}$.*

3.2   Proof of Theorem 2

Lower bound proof. Let $3 \leq k \leq n$ and Π be a subdivision of the plane into n convex faces, where k faces are unbounded. By Lemma 3 there exists a subset $\Pi_0 \subset \Pi$ of $n_0 \geq (n-k)/(k-2)$ faces that form a convex subdivision of a monotone polygon P_0. We may assume, by applying a rotation if necessary, that no edge of $G(\Pi)$ is horizontal and P_0 is y-monotone. We can assume that $n_0 \geq 4$. It is enough to show that $G(\Pi_0)$ contains a monotone path of size at least $c_0 \log n_0 / \log \log n_0$, where $c_0 > 0$ is an absolute constant.

Let m denote the maximum size of a monotone path in $G(\Pi_0)$. We use Corollary 1 to construct a nested sequence $\Pi_0 \supset \Pi_1 \supset \cdots \supset \Pi_t$ such that $|\Pi_i| = n_i \geq n_0/(m+3)^i$ and Π_i is a convex subdivision of a y-monotone polygon P_i for $i = 1, 2, \ldots, t$. Moreover, the bottom or top vertex of P_i is different from that of P_{i-1} for $i = 1, 2 \ldots, t$. If $|\Pi_i| > 2$, then Π_{i+1} can be constructed from Π_i by Corollary 1. We may therefore assume that $n_t \in \{1, 2\}$.

Let β_t be an arbitrary y-monotone path in $G(\Pi_t)$ between the top and bottom vertex of P_t. For $i = t, t-1, \ldots, 1$, the path β_i can be extended to a y-monotone path β_{i-1} between the top and bottom vertex of $G(\Pi_{i-1})$ by Observation 1. Note that β_{i-1} is strictly longer than β_i, since at least one of two endpoints of β_i is not the top or bottom vertex of P_{i-1}. Therefore, β_0 is a y-monotone path with at least $t + 1$ edges, and $t \leq m - 1$ by the definition of m. We have

$$2 \geq n_t \geq \frac{n_0}{(m+3)^t} \geq \frac{n_0}{(m+3)^{m-1}}.$$

hence $n_0 \leq 2(m+3)^{m-1} = 2^{1+(m-1)\log(m+3)}$. This implies $m \geq c_0 \log n_0 / \log \log n_0$, for some absolute constant $c_0 > 0$, as required.

Upper bound construction. For every pair of integers $k, n \in \mathbb{N}$, where $3 \leq k < n$, we subdivide the plane into a set Π of $\Omega(n)$ convex cells, exactly k of which are unbounded, such that every monotone path in $G(\Pi)$ has $O(\log \frac{n}{k} / \log \log \frac{n}{k})$ edges. We first construct the unbounded faces. If $k = 3$, then let Q_1 be a triangle, and subdivide the exterior of Q_1 into 3 convex faces by 3 rays emitted from the vertices of Q_1. If $k \geq 4$, then subdivide the plane into k unbounded faces by a star graph with $\lfloor k/2 \rfloor$ leaves, $q_1, \ldots, q_{\lfloor k/2 \rfloor}$, and 2 or 3 rays emitted from each leaf (similarly to the subdivision in Fig. 1(right)). Then replace each vertex q_i, $1 \leq i \leq \lfloor k/2 \rfloor$ by a small triangle Q_i. Now it is enough to subdivide each triangle Q_i into at least $n_0 = 5n/k$ bounded faces such that every monotone path restricted to Q_i has
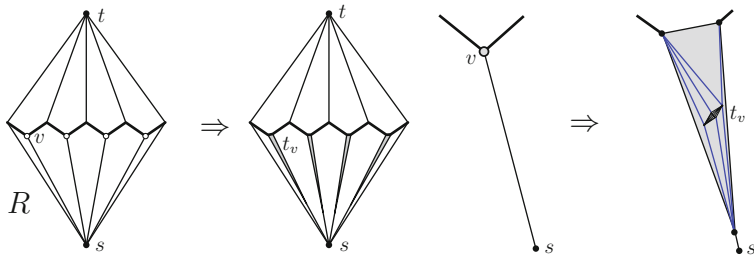
Fig. 5 *Left*: The subgraph R with $m = 4$. The vertices in Z are marked with *empty circles*. In one phase of our construction, the vertices in Z are replaced by long and skinny triangles. *Right*: A vertex $v \in Z$ is replaced by a skinny triangle that contains an affine copy of R, and the space between the triangle and R is triangulated

$O(\log n_0 / \log \log n_0)$ edges. Since a monotone path can visit at most two triangles Q_i, it has $O(\log n_0 / \log \log n_0)$ edges.

Let $m = 2 \cdot \lceil \log n_0 / \log \log n_0 \rceil$. The basic building block of our construction is a plane geometric graph R shown in Fig. 5(left). The outer face of R is a rhombus symmetric with respect to both the x- and the y-axes. Two opposite corners of the rhombus are connected by an x-monotone zig-zag path ξ of $2m$ edges. The edges of ξ have alternately negative and positive slopes, and so the $2m - 1$ interior vertices are alternately y-minimal and y-maximal. Denote by Z the y-minimal interior vertices of ξ, so $|Z| = m$. The y-minimal interior vertices of ξ are joined to the bottom vertex s of the rhombus, and the y-maximal ones to the top vertex t. The vertices s and t are sufficiently far below and respectively far above ξ such that all bounded faces of R are convex. The graph R contains a monotone path of size $2m$ in directions close to horizontal, but every monotone path has at most three edges in directions close to vertical. After an appropriate affine transformation, every monotone path has at most three edges in all directions except for those in a prescribed interval of length $\frac{\pi}{2m}$ in $[0, 2\pi)$.

We construct a subdivision Π of Q in m phases. For $i = 1, \dots, m$, we maintain a convex subdivision Π_i, a set Z_i of special vertices in $G(\Pi_i)$, and a special edge incident to each vertex in Z_i. For constructing Π_1, consider a small affine copy of R, such that any monotone path has at most three edges except for directions in the interval $(\frac{\pi}{2m}, \frac{2\pi}{2m})$. Place this copy of R in the interior of Q, and triangulate the space between R and Q arbitrarily to obtain the convex subdivision Π_1 of Q. For each $v \in Z_1 = Z$, let the special edge of v be vs, the edge joining v to the bottom vertex of R.

In phase $i = 2, \dots, m$, we construct Π_i from Π_{i-1} as follows. Replace each vertex $v \in Z_{i-1}$ by a long skinny triangle t_v along the special edge incident to v as in Fig. 5(right). In the interior of t_v, place a small affine copy of R near the midpoint of t_v, such that any monotone path has at most three edges in R except for directions in the interval $\left(\frac{(2i-1)\pi}{2m}, \frac{2i\pi}{2m} \right)$. Denote by r_v the outer boundary of this copy of R. Triangulate the space between r_v and t_v arbitrarily by using $O(1)$ edges; the edges of

the triangulation are almost parallel to the special edge sv if t_v is sufficiently skinny and r_v is sufficiently small. Now, any monotone path γ can enter the interior of *at most two* of the m skinny triangles t_v: to see this, note that the direction θ in which γ is monotone is close to perpendicular to at most one skinny triangle, say t_{v_0}. For any other $t_v \neq t_{v_0}$, γ must either enter or leave through the vertex s if it visits the interior. This means that at most two of these triangles can be visited. Moreover, if two of them are visited, then t_{v_0} cannot be entered.

Let Z_i be the union of the vertex sets Z from all affine copies of R created in phase i, and let the special edge of each vertex in Z_i be the edge connecting that vertex to the vertex s of the corresponding copy of R.

We show that $\Pi = \Pi_m$ has $\Omega(n_0)$ faces, and the longest monotone path in $G = G(\Pi)$ has size $O(m) = O(\log n_0 / \log \log n_0)$. Initially, we have $|Z_1| = m$ special vertices. Since $|Z_i| = m|Z_{i-1}|$ for $i = 2, \ldots, m$, it follows that $|Z_m| = m^m$. Note that for each special vertex in Z_m, there is an incident quadrilateral face in a copy of R which is not incident to any other vertex in Z_m. Hence Π has at least $|Z_m| = m^m = \Omega(n_0)$ faces.

For any $v \in Z_i$, $i = 1, 2 \ldots, m$, our recursive construction did not modify the edges of the triangle t_v and the rhombus r_v (only the interior edges of a copy of R inside r_v are modified in subsequent phases). Let $\mathcal{T}_i$ and $\mathcal{R}_i$ denote the set of triangles t_v and rhombi r_v, respectively, for all $v \in Z_i$. Note that a monotone path enters and exits the interior of a triangle or a rhombus in G at most once. Let γ be a path in $G(\Pi)$ that is monotone in some direction θ. Assume that $\theta \in \left[\frac{(j-1)\pi}{m}, \frac{j\pi}{m} \right)$, for some $j \in \{1, 2, \ldots, m\}$. For $i < j$, the path γ enters at most one triangle of $\mathcal{T}_i$. For $j = i$, it can visit m triangles of $\mathcal{T}_i$, that all lie in a common triangle $t \in \mathcal{T}_{i-1}$. However, as we have discussed above, γ can reach the rhombus $r_v \subset t_v$ in at most two triangles $t_v \in \mathcal{T}_i$. For $i > j$, the path γ enters at most two triangles of $\mathcal{T}_i$, at most one inside each rhombus in $\mathcal{R}_{i-1}$.

It follows that γ traverses $O(m)$ edges in at most one zig-zag path created in phase j, and it traverses $O(1)$ edges created in any of the other $m-1$ phases. Consequently, every monotone path in $G(\Pi)$ has at most $O(m)$ edges, as required. $\qquad\square$

We note that the construction uses only $O(m^2)$ distinct slopes of edges: at each level i, the copies of R are rotated into a fixed direction; this gives $O(m)$ slopes for the edges of R. R is embedded inside m distinctly oriented triangles t_v of the higher level; thus we need only m versions of the $O(1)$ edges filling the space between r_v and t_v, for another $O(m)$ slopes. This gives $O(m)$ slopes per level, and $O(m^2)$ slopes in total.

4 Proof of Theorem 3

We construct a polytope P with n vertices with no monotone path longer than $O(\log^2 n)$. The polytope will be built as a hierarchical system of *cells* Δ. The basic building block Δ is the polytope shown in Fig. 6. It is symmetric with respect to the
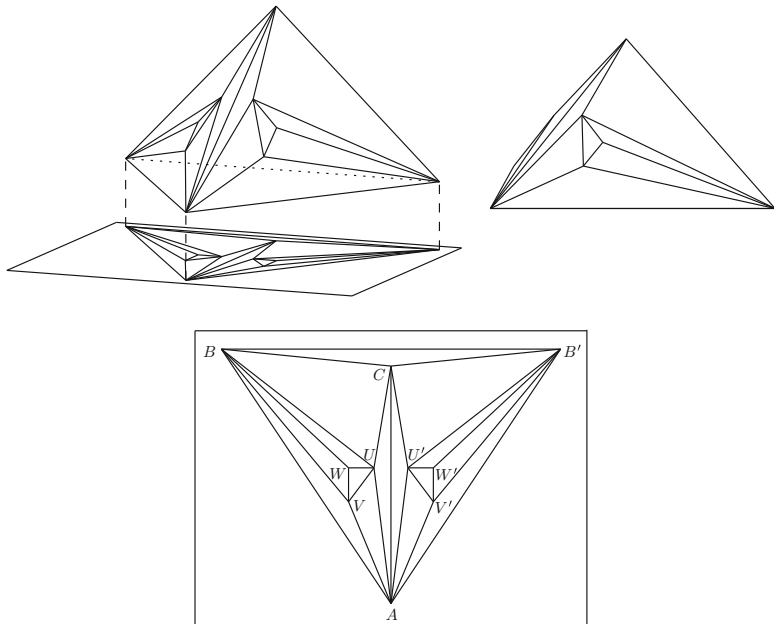
Fig. 6 The basic building block Δ. The *upper part* shows two spatial views, and the *lower part* shows the vertical projection

Table 1 The vertex coordinates of the basic polytope Δ

Point	(x,	y,	z)
A	(0,	0,	0)
B, B'	(∓ 1,	1.5,	0)
C	(0,	1.4,	1)
U, U'	(∓ 0.1,	0.8,	0.55)
V, V'	(∓ 0.25,	0.6,	0.25)
W, W'	(∓ 0.25,	0.8,	0.39)

yz-plane. The coordinates of its 10 vertices are given in Table 1. Its *base* is the outer triangle ABB' and it has two *inner triangles* UVW and $U'V'W'$ on which smaller (and transformed) copies of the cell Δ will be recursively placed, as sketched in Fig. 7.

The *characteristic region* $\chi(\Delta)$ of Δ is the set of directions c $=$ $(u, v, 1)$ for which there exists a monotone path that starts at a boundary vertex (A, B, or B'), visits at least two vertices of each triangle UVW and $U'V'W'$, and ends at the boundary. The characteristic region for the standard cell is shown in Fig. 8. It has been generated with the help of a computer by enumerating all simple paths from the boundary via UVW and $U'V'W'$ (in either order) back to the boundary according to the definition, and by generating the set of directions for which the path is monotone. This is a (possibly empty) polyhedral set. There are 106 such paths in total; and among these, 38 paths with a nonempty region of directions.
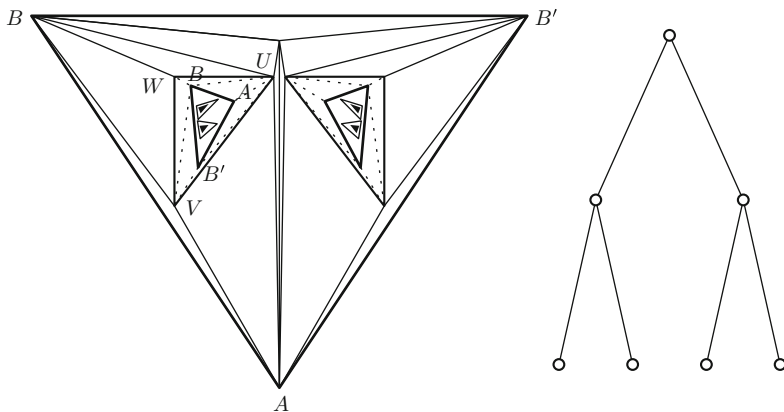
Fig. 7 A schematic drawing of two nested levels of cells (not to scale) and the corresponding binary tree
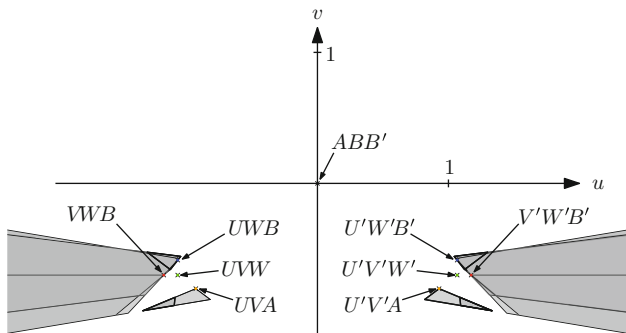

Fig. 8 The characteristic region $\chi(\Delta)$. It is symmetric and consists of two unbounded wedge-like regions and two triangles. For orientation, the normal vectors of the triangles ABB', UVW, and some other triangles are indicated. What is important is that UVW and $U'V'W'$ are outside $\chi(\Delta)$

In Fig. 8, all these regions are drawn on top of each other, and their union yields the characteristic region. Our restriction to directions $\mathbf{c} = (u, v, w)$ with $w = 1$ excludes the horizontal directions ($w = 0$). These directions appear as infinite rays in the wedges of the figure.

As announced we will recursively place smaller copies of Δ on the *inner triangles UVW and $U'V'W'$.* Before placing them there, we subject them to geometric transformations to make them fit together as a convex polytope, and to ensure certain properties of the characteristic regions. The structure of this construction will be a balanced binary tree of depth k, with $k + 1$ levels and $2^{k+2} - 1$ nodes. Each vertex of P will belong to only one copy of Δ. Thus, there will be $10(2^{k+2} - 1)$ vertices in total. Each copy of Δ (each node of the tree) will have its characteristic region.
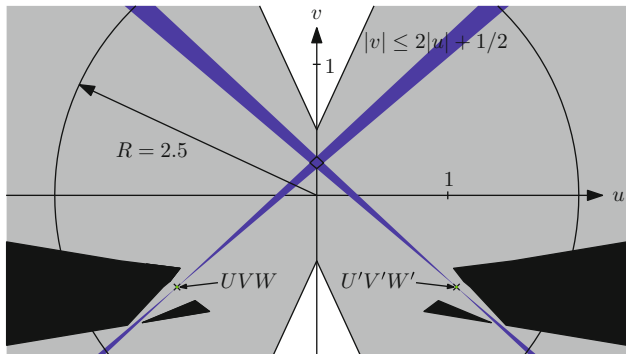
Fig. 9 The recursive construction in terms of characteristic regions

The root is the cell Δ in its original form. We describe how to place two transformed copies of a construction with k levels on Δ. Inductively, we maintain the following properties.

(i) No three characteristic regions intersect.
(ii) Two characteristic regions can only intersect in a disk of radius $R = 3$ around the origin.
(iii) The characteristic regions lie inside the region $|v| \le 2|u| + 1/2$.

For the basic building block itself (the induction basis $k = 0$), (i) and (ii) are vacuous, and (iii) can be checked by inspection, see Fig. 9.

Geometric Transformations. Table 2 lists some simple affine transformations that can be applied to some polytope P, and the effect that they have on the characteristic region χ of some cell Δ that belongs to P. Here, $\varepsilon > 0$ is a small constant, a, b, α are arbitrary constants.

Suppose we have a polytope P with $k - 1$ levels that fulfills the properties (i)–(iii). We first *squeeze* it to make the wedge $|v| \le 2|u| + 1/2$ narrow enough (see Fig. 10). Then we *turn* it, in order to be able to fit it between the two components of the characteristic region, as shown in Fig. 9. We also have to *tilt* it, to make the base parallel to the triangle UVW. In terms of uv-coordinates, this corresponds to translating the origin to the normal vector of UVW that is shown in Fig. 9. We want to place the tilted base slightly above the triangle UVW. Thus, before tilting it, we *scale* it to make it small enough to fit inside UVW, and we *squash* it in the z-direction. Squashing achieves two purposes.

(a) The polytope P is made very flat (almost parallel to its base). Thus, when P is *translated* close enough over UVW so that it does not see any face beyond UVW, the triangle UVW in turn does not see any vertex of P except the base. Thus the convex hull of the union will have the combined set faces of Δ and P, except UWV and the base of P, plus a *collar* or *ring* of six triangle

Table 2 The affine transformations to which the cells are subjected

Operation	Coordinate transformation	Effect on $\chi(\Delta)$
Squash	$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} x \\ y \\ \varepsilon \cdot z \end{pmatrix}$	$\begin{pmatrix} u \\ v \end{pmatrix} \mapsto \varepsilon \cdot \begin{pmatrix} u \\ v \end{pmatrix}$
Squeeze	$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} \varepsilon \cdot x \\ y \\ \varepsilon \cdot z \end{pmatrix}$	$\begin{pmatrix} u \\ v \end{pmatrix} \mapsto \begin{pmatrix} u \\ \varepsilon \cdot v \end{pmatrix}$
Scale	$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \varepsilon \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix}$	$\begin{pmatrix} u \\ v \end{pmatrix} \mapsto \begin{pmatrix} u \\ v \end{pmatrix}$
Turn	$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$	$\begin{pmatrix} u \\ v \end{pmatrix} \mapsto \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}$
Tilt	$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} x \\ y \\ z + ax + by \end{pmatrix}$	$\begin{pmatrix} u \\ v \end{pmatrix} \mapsto \begin{pmatrix} u \\ v \end{pmatrix} - \begin{pmatrix} a \\ b \end{pmatrix}$
Translate	$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} a \\ b \\ c \end{pmatrix}$	$\begin{pmatrix} u \\ v \end{pmatrix} \mapsto \begin{pmatrix} u \\ v \end{pmatrix}$

faces between UVW and the base of P. (The ring is shown with dotted edges in Fig. 7. The faces are guaranteed to be triangles except when the turning operation has created parallel edges between the base and UVW.)

(b) The second effect of squashing, in terms of characteristic regions, is to scale the uv-coordinates down. This brings any intersections between characteristic regions of P (which were initially within a radius of $R = 2.5$) as close as we want to the point representing the normal of UVW. We ensure that they are within distance $1/2$ of UVW.

A symmetric procedure is applied to another copy P' of P which is placed on $U'V'W'$.

Property (iii) can be checked visually from Fig. 9. Any intersections between two critical regions of P lie within distance $1/2$ of UVW the point, and thus within radius $R = 2.5$ of the origin. By construction there are no intersections between critical regions of P (they are confined in the squeezed wedge) and $\chi(\Delta)$. There can be intersections between critical regions of P and critical regions of P'. As Fig. 9 shows, these new intersections are confined to the intersections between the two corresponding squeezed wedges, which is a small quadrilateral area symmetric to the v-axis, close to the origin. So these new intersections are certainly within radius $R = 2.5$ about the origin, establishing property (ii). Finally, there cannot by any triple intersections, since the regions of P are disjoint when they are more than 0.5 away from UVW, and similarly for P'.

It is clear that the desired effects can be obtained by choosing the constants ε in the different steps small enough. One could work out particular values of ε (and α, a,

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} \varepsilon \cdot x \\ y \\ \varepsilon \cdot z \end{pmatrix}$$

squeeze

$$\begin{pmatrix} u \\ v \end{pmatrix} \mapsto \begin{pmatrix} u \\ \varepsilon \cdot v \end{pmatrix}$$

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \varepsilon \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

scale

turn

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} x \\ y \\ \varepsilon \cdot z \end{pmatrix}$$

squash

translate

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} x \\ y \\ z + ax + by \end{pmatrix}$$
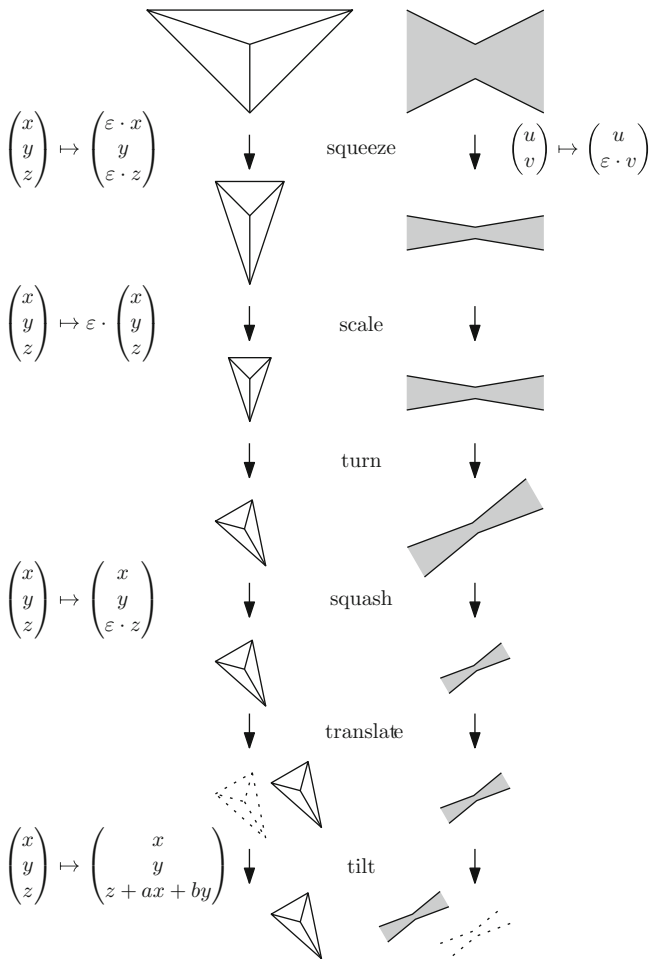
tilt

Fig. 10 Schematic drawing of the sequence of transformations on Δ and their effects on $\chi(\Delta)$

and b) and get an explicit affine transformation for placing P on top of UVW and $U'V'W'$.

Let us now prove that there is no long monotone path in P. We say that a cell Δ is *visited* if the path contains at least two vertices of the cell. These might be just two boundary vertices like A and B, or also vertices inside Δ. From the definition of characteristic regions, we can conclude the following. Let Δ be a cell with children Δ_1 and Δ_2. Then a path P that is monotone in direction c can visit *both Δ_1 and Δ_2* only if

(a) P starts or ends inside Δ, or
(b) c lies in the characteristic region of Δ.

We now analyze the set S of nodes that correspond to the visited cells. By definition, if a cell is visited, its parent is also visited. Thus, S is a subtree of T containing the root.

A node Δ in S can have both its children in S only if conditions (a) and (b) are fulfilled. Therefore, there can be at most $2k + 1$ nodes with two children: $2k - 1$ because of condition (a), plus 2 because of condition (b), since by property (i) there cannot be more than two cells whose characteristic set c lies in.

It follows that the tree S is the union of at most $2k + 2$ paths, each of length at most $k + 1$. So at most $2(k + 2)^2$ nodes are visited. The monotone path can have at most 12 vertices per visited node: 10 for the corresponding cell Δ, and at most one vertex from each child node. In total, the length is at most $24(k + 2)^2$.

The polytope P we have constructed has $n = (2^{k+1} - 1) \cdot 10 = 2^k \cdot 20 - 10$ vertices and $2n - 4 = 2^k \cdot 40 - 24$ triangular faces. To get an example P' with any given number of vertices, one can add (flat) pyramids on a subset of the faces. Between any two new vertices of P', a path must visit at least one "old" vertex of P. Hence the length of monotone paths can at most double. □

We can also analyze the size of the coordinates or the *spread* (ratio between the diameter and the smallest distance between two points) of this construction. The sequence of geometric transformations which are applied to place a small copy of the level-k construction on UVW can be composed into one fixed affine transformation (and its "mirror image", for $U'V'W'$). This transformation is iterated k times. It follows that the size of coordinates, or the spread, grows like c^k for some constant $c > 1$; in terms of n, this is a polynomial.

Remark. We note that P is even a *stacked* polytope: such a polytope can be generated from an initial tetrahedron by successively gluing tetrahedra on triangular faces. (Equivalently, the graph $G(P)$ can be generated by successively inserting new degree-3 vertices into triangular faces.) It is easy to check that the basic building block Δ is a stacked polytope. The only place where one has to be careful are the connecting edges between successive levels of P (the dashed edges of Fig. 7). This connecting polytope is the convex hull of two triangles that lie in parallel planes. It is not hard to see that this such a polytope a stacked polytope if and only if the edge directions of the two triangles, UVW and the small copy of ABB', do not alternate when going around the triangles. (Otherwise it has the combinatorial structure of an octahedron.) One can check that in our construction, this alternation does not happen: the main reason for this is that (contrary to the appearance in Fig. 7), the triangle ABB' is squeezed so much that the sides AB and AB' are almost parallel.

5 Proof of Theorem 4

The problem with the construction of the previous section is that it has to make the pieces very "flat" as the recursion proceeds. Therefore we could not control the entrance into sibling subtrees for the nodes in which the monotone path starts or ends.
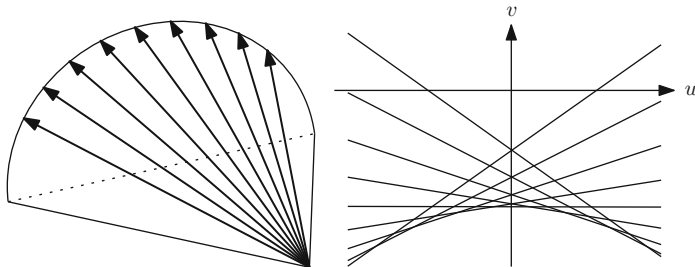
Fig. 11 $2^k + 1 = 9$ initial directions, and the characteristic lines of their normals (for $k = 3$)

In this section, we construct another polytope Q with a recursive structure. The "principal directions" of the parts will be chosen in advance, and they are not subject to transformations. We have to pay for this advantage by having a universal vertex A that is connected to all other vertices. In terms of a degree bound, our example is therefore as bad as it can be.

We start by picking $2^k + 1$ equally spaced *initial directions* on the cone K with the y-axis as the cone-axis and opening angle $45°$, all pointing upwards such that all angles made by these directions with the xy-plane are bounded by $45°$, see Fig. 11.

As in the previous section, we visualize directions by the (u, v)-component of the direction vectors $(u, v, 1)$. For an initial direction ℓ, the directions that are perpendicular to ℓ form a line in uv-space, called the *characteristic line χ_ℓ of ℓ*. The characteristic lines of the initial directions form an arrangement of lines in uv-space, as shown in the right part of Fig. 11. The perpendicular planes of the cone of initial directions are tangent to the polar cone K^* of K, which is in this case also a cone with opening angle $45°$. The intersections of these planes with the plane $z = 1$, which are the characteristic lines that we see in Fig. 11, are therefore tangent to the hyperbola which is the intersection of K^* with the plane $z = 1$. Since these lines are tangent to a common hyperbola, there are no triple intersections.

Every edge of our polytope Q (with one exception, the edge BB') will have a direction very close to one of the initial directions. It will therefore be difficult to traverse two such edges that are "almost parallel" in opposite directions. The following lemma quantifies this intuition.

Lemma 5. *Consider a line ℓ that makes an angle at most $\pi/4$ with the xy-plane, and two directed edges c_1 and c_2 which make an angle at most $\alpha \leq 1/10$ with ℓ but point in opposite directions. If a monotone path passes through both edges, then any direction $(u, v, 1)$ in which it is monotone lies inside the double-wedge of opening angle 2α and width 8α, centered about the characteristic line χ_ℓ of ℓ, as shown in Fig. 12.*

We call this region the *sharp-turn* region of ℓ.

Proof. This is an elementary calculation. We assume w.l.o.g. that ℓ has unit direction vector $(0, b, c)$. Since the angle with the horizontal plane is at most $45°$,

Fig. 12 The sharp-turn
region of two almost opposite
directions
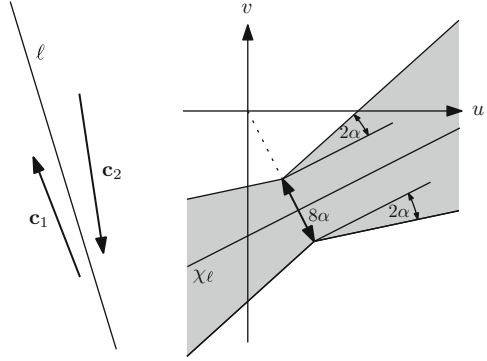
we have $c \leq b$, and since $b^2 + c^2 = 1$, we also get $b \geq 1/\sqrt{2}$, and

$$b - \alpha \geq 1/\sqrt{2} - 1/10 \geq 1/2. \tag{2}$$

The characteristic line χ_ℓ is then given by the solution of $0 \cdot u + b \cdot v + c \cdot 1 = 0$, that is, $v = -c/b$. (In Fig. 12 this would correspond to a rotated coordinate system where χ_ℓ becomes horizontal.)

We write the unit direction vector of c_1 as $(0, b, c) + (\Delta x, \Delta y, \Delta z)$, where

$$\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2} \leq 2 \sin \frac{\alpha}{2} \leq \alpha.$$

The normal plane to c_1 appears in the uv-plane as given by the equation

$$\Delta x \cdot u + (b + \Delta y) \cdot v + (c + \Delta z) \cdot 1 = 0,$$

which yields

$$v = -\frac{c + \Delta z}{b + \Delta y} - \frac{\Delta x}{b + \Delta y} \cdot u.$$

We claim that (i) the slope $\frac{-\Delta x}{b + \Delta y}$ of this line is bounded by $\tan(2\alpha)$ in absolute value and (ii) the intercept $-\frac{c + \Delta z}{b + \Delta y}$ deviates from $-c/b$ by at most 4α. This means that the line lies within the claimed sharp-turn region.

The normal plane to c_2 appears in the uv-plane as another line within this wedge, and the directions which are monotone in c_1 and c_2 must lie above one line and below the other, i.e., they are confined to some wedge within the sharp-turn region.

To finish the proof of the lemma, we verify the claims (i) and (ii).

(i) The absolute slope is bounded by $\left| \frac{-\Delta x}{b + \Delta y} \right| \leq \frac{|\Delta x|}{b - |\Delta y|} \leq \frac{\alpha}{b - \alpha} \leq \frac{\alpha}{1/2} = 2\alpha \leq \tan(2\alpha)$, by (2).
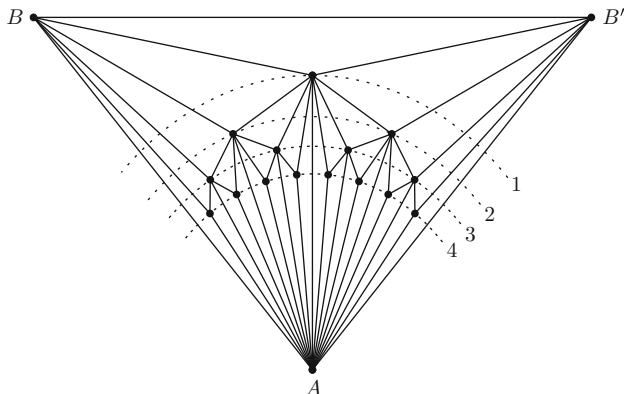
(ii) Similarly, the difference of intercepts is

Fig. 13 Schematic drawing of the polytope Q with $k = 4$ inner layers

$$\left| \frac{c}{b} - \frac{c + \Delta z}{b + \Delta y} \right| = \left| \frac{c\Delta y - b\Delta z}{b(b + \Delta y)} \right| \leq \frac{c}{b} \left| \frac{\Delta y}{b + \Delta y} \right| + \left| \frac{\Delta z}{b + \Delta y} \right|$$

$$\leq \left| \frac{\Delta y}{b + \Delta y} \right| + \left| \frac{\Delta z}{b + \Delta y} \right| = \frac{|\Delta y| + |\Delta z|}{|b + \Delta y|}$$

$$\leq \frac{|\Delta y| + |\Delta z|}{b - |\Delta y|} \leq \frac{\alpha + \alpha}{b - \alpha} \leq \frac{2\alpha}{1/2} = 4\alpha.$$

By choosing α small enough, we can therefore ensure that sharp-turn regions of all initial directions remain so close to their characteristic lines that no triple intersections of sharp-turn regions appear. This implies that a monotone path can make at most two such "sharp turns" as specified in Lemma 5.

The polytope Q has an outer triangle ABB' and k interior layers $1, 2, \ldots, k$. Layer i has 2^{i-1} vertices, and thus there are $2^k + 2$ vertices in total. A schematic drawing of the polytope with $k = 4$ inner layers is given in Fig. 13. From vertex A, there is an edge to each other vertex, and these edges have the chosen initial directions. We construct the polytope layer by layer. We start with an isosceles triangle ABB', and the vertices B and B' form layer 0. The vertices of layer i are drawn interspersed (in angular order) with the vertices of the previous layers. When a vertex is first generated, it has two edges to previous layers and an edge to A. Later it collects two edges to each successive layer.

So far we have described the combinatorial structure of Q. The geometry is already constrained insofar as each vertex lies on one of the initial direction rays through A. It remains to specify the distances from A. All vertices of a level i have the same distance d_i from A. When constructing layer i, we choose the distance d_i so much smaller than all previous distances that the new points appear extremely close to A. More precisely, we want that an edge UV between a vertex U on layer $j < i$ and a vertex V on layer i has an angle at most α with the edge UA.

By a continuity argument, such a value d_i always exists. For instance, it suffices to choose $d_i \leq \alpha d_{i-1}$. In addition, we have to ensure that the polytope has the right combinatorial structure, i.e., the convex hull edges are the edges shown in Fig. 13. By construction, all edges emanating from A lie on the convex hull. We need to make sure that the insertion of a layer does not destroy any previously inserted edge; the new layer will just add small triangular pyramids on top of some triangular faces. This can again be achieved by choosing d_i small enough, by a continuity argument. More explicitly, one can argue that is sufficient to set $d_i \leq 2d_{i-1}^2/d_0$: this ensures that the new vertex cannot even see any other vertex from previous layers in the plane projection (Fig. 13), without crossing edges incident to A. (In fact, it would not matter if edges were flipped; the only property we actually need is that no edge crosses between vertices of the same level.)

Now, consider a monotone path in the 1-skeleton of Q. If it goes through A, B or B', we cut it at these vertices, into at most four pieces. To show that any monotone path on the boundary of Q has length $O(k) = O(\log n)$, it is therefore sufficient to consider monotone paths that do not pass through A, B or B'. Each edge in the remaining graph connects vertices from different levels. An edge on the monotone path can either ascend (into a smaller level) or descend (into a larger level). When the path changes at a vertex U from ascending to descending, it must traverse two edges that are almost parallel to UA, in opposite orientations. As we have observed, this can only happen at most twice. $\square$

Remark. By the way in which it is constructed, Q, just as P in the previous section, is a *stacked* polytope. As for the size of coordinates and the spread, we need to fulfill the conditions $d_i \leq \alpha d_{i-1}$ and $d_i \leq 2d_{i-1}^2/d_0$. Setting $d_0 = 1$, and $d_i = \alpha^{2^{i-1}}$ for $i \geq 1$ would work. The angle α can be chosen as a rational function involving a polynomial in n. But the whole construction leads then to coordinates that are exponential in n (but require only polynomially many bits, in a suitable fixed-precision approximation). As we have mentioned, if we do not insist on Q having the given combinatorial structure, we can ignore the second condition, and setting $d_i = \alpha^i$ for $i \geq 0$ leads to coordinates that are polynomial in n. But then we can no longer guarantee that Q is a stacked polytope.

6 Conclusion

Open problems. The proofs of Lemmas 1–3 crucially depend on the planarity of the subdivisions. Extending Theorems 1 and 2 for convex subdivisions of $\mathbb{R}^d$, $d \geq 3$, remain as open problems.

It also left for future research to close the gaps between: (i) the upper bound $O(\log^2 n)$ in Theorem 3 and the trivial lower bound $\Omega(\log n)$ for n-vertex polytopes with bounded vertex degrees; (ii) the upper bound $O(\log n)$ in Theorem 4 and the lower bound $\Omega(\log n/\log\log n)$ for n-vertex polytopes with arbitrary vertex degrees.

We have shown (Theorem 2) that in any connected subdivision with n faces, $k \geq 3$ of which are unbounded, there exists a monotone path of length $\Omega(\log \frac{n}{k} / \log \log \frac{n}{k})$. Figure 2 shows that this lower bound does not hold for (strictly) monotone paths starting from an arbitrary vertex. Deciding whether there exists a *weakly* monotone path of length $\Omega(\log \frac{n}{k} / \log \log \frac{n}{k})$ starting from *every* vertex in a convex subdivision with k unbounded faces remains an open problem.

The authors thank János Pach for insistently asking the question to which Theorem 1 gives the answer. A recursive construction involving a hierarchy of squeezed zigzags, similar to the upper-bound construction for Theorem 2 in Sect. 3.2, has been suggested by Boris Bukh.

References

1. Arkin, E.M., Connelly, R., Mitchell, J.S.B.: On monotone paths among obstacles with applications to planning assemblies. In: Proceedings of the 5th Annual Symposium on Computational Geometry, Saarbrücken, ACM Press, pp. 334–343 (1989)
2. Aurenhammer, F.: A criterion for the affine equality of cell complexes in $\mathbb{R}^d$ and convex polyhedra in $\mathbb{R}^{d+1}$. Discret. Comput. Geom. **2**, 49–64 (1987)
3. Balogh, J., Regev, O., Smyth, C., Steiger, W., Szegedy, M.: Long monotone paths in line arrangements. Discret. Comput. Geom. **32**, 167–176 (2004)
4. Chazelle, B., Edelsbrunner, H., Guibas, L.J.: The complexity of cutting complexes. Discret. Comput. Geom. **4**, 139–181 (1989)
5. Dumitrescu, A.: On some monotone path problems in line arrangements. Comput. Geom. **32**(1), 13–25 (2005)
6. Dumitrescu, A., Tóth, Cs.D.: Monotone paths in planar convex subdivisions. In: Abstracts of the 21st Fall Workshop on Computational Geometry, New York (2011)
7. Dumitrescu, A., Rote, G., Tóth, Cs.D.: Monotone paths in planar convex subdivisions. In: Gudmundsson, J., Mestre, J., Viglas, T. (eds.) Computing and Combinatorics. Proceedings of the 18th Annual International Computing and Combinatorics Conference (COCOON 2012), Sydney. Lecture Notes in Computer Science, vol. 7434, pp. 240–251. Springer (2012)
8. Klee, V.: Heights of convex polytopes. J. Math. Anal. Appl. **11**, 176–190 (1965)
9. McMullen, P.: The maximum numbers of faces of a convex polytope. Mathematika **17** 179–184 (1971)
10. Pfeifle, J.: Long monotone paths on simple 4-polytopes. Isr. J. Math. **150**(1), 333–355 (2005)
11. Pfeifle, J., Ziegler, G.M.: On the monotone upper bound problem. Exp. Math. **13**, 1–11 (2004)
12. Rote, G.: Partial least-squares point matching under translations. In: Abstracts of the 26th European Workshop on Computational Geometry, Dortmund, pp. 249–251 (2010)
13. Rote, G.: Long monotone paths in convex subdivisions. In: Abstracts of the 27th European Workshop on Computational Geometry, Morschach, pp. 183–184 (2011)
14. Santos, F.: A counterexample to the Hirsch conjecture. Ann. Math. (2) **176**, 383–412 (2012). arXiv:1006.2814

15. Todd, M.J.: The monotonic bounded Hirsch conjecture is false for dimension at least 4. Math. Oper. Res. **5**(4), 599–601 (1980)
16. Tóth, Cs.D.: Stabbing numbers of convex subdivisions. Period. Math. Hung. **57**(2), 217–225 (2008)
17. Ziegler, G.M.: Lectures on Polytopes. GTM, vol. 152, pp. 83–93. Springer, New York (1994)

Complexity of the Positive Semidefinite Matrix Completion Problem with a Rank Constraint

Marianna E.-Nagy, Monique Laurent, and Antonios Varvitsiotis

Abstract We consider the decision problem asking whether a partial rational symmetric matrix with an all-ones diagonal can be completed to a full positive semidefinite matrix of rank at most k. We show that this problem is $\mathcal{NP}$-hard for any fixed integer $k \geq 2$. In other words, for $k \geq 2$, it is $\mathcal{NP}$-hard to test membership in the rank constrained elliptope $\mathcal{E}_k(G)$, defined by the set of all partial matrices with an all-ones diagonal and off-diagonal entries specified at the edges of G, that can be completed to a positive semidefinite matrix of rank at most k. Additionally, we show that deciding membership in the convex hull of $\mathcal{E}_k(G)$ is also $\mathcal{NP}$-hard for any fixed integer $k \geq 2$.

Key words Elliptope • Correlation matrix • psd matrix completion

Subject Classifications: 90C22, 68Q17, 05C62

1 Introduction

Geometric representations of graphs are widely studied within a broad range of mathematical areas, ranging from combinatorial matrix theory, linear algebra, discrete geometry, and combinatorial optimization. They arise typically when

M. E.-Nagy (✉) • A. Varvitsiotis
Centrum Wiskunde & Informatica (CWI), P.O. Box 94079,
1090 GB, Amsterdam, The Netherlands
e-mail: M.E.Nagy@cwi.nl; A.Varvitsiotis@cwi.nl

M. Laurent
Centrum Wiskunde & Informatica (CWI), P.O. Box 94079,
1090 GB, Amsterdam, The Netherlands

Tilburg University, Tilburg, The Netherlands
e-mail: M.Laurent@cwi.nl

labeling the nodes by vectors assumed to satisfy certain properties. For instance, one may require that the vectors labeling adjacent nodes are at distance 1, leading to unit distance graphs. Or one may require that the vectors labeling adjacent nodes are orthogonal, leading to orthogonal representations of graphs. One may furthermore ask, e.g., that nonadjacent nodes receive vector labels that are not orthogonal. Many other geometric properties of orthogonal labelings and other types of representations related, e.g., to Colin de Verdière type graph parameters, are of interest and have been investigated (see [8]). A basic question is to determine the smallest possible dimension of such vector representations. There is a vast literature, we refer in particular to the surveys [10, 11, 20] and further references therein for additional information.

In this note we revisit orthogonal representations of graphs, in the wider context of Gram representations of weighted graphs. We show some complexity results for the following notion of Gram dimension, which has been considered in [18, 19].

Definition 1. Given a graph $G = (V = [n], E)$ and $x \in \mathbb{R}^E$, a *Gram representation* of x in $\mathbb{R}^k$ consists of a set of unit vectors $v_1, \cdots, v_n \in \mathbb{R}^k$ such that

$$v_i^{\mathsf{T}} v_j = x_{ij} \quad \forall \{i, j\} \in E.$$

The *Gram dimension* of x, denoted as $\mathrm{gd}(G, x)$, is the smallest integer k for which x has such a Gram representation in $\mathbb{R}^k$ (assuming it has one in some space).

As we restrict our attention to Gram representations of $x \in \mathbb{R}^E$ by unit vectors, all coordinates of x should lie in the interval $[-1, 1]$, so that we can parametrize x as

$$x = \cos(\pi a), \quad \text{where } a \in [0, 1]^E.$$

In other words, the inequality $\mathrm{gd}(G, x) \le k$ means that (G, a) can be isometrically embedded into the spherical metric space $(\mathbf{S}^{k-1}, d_{\mathbf{S}})$, where $\mathbf{S}^{k-1}$ is the unit sphere in the Euclidean space $\mathbb{R}^k$ and $d_{\mathbf{S}}$ is the spherical distance:

$$d_{\mathbf{S}}(u, v) = \arccos(u^{\mathsf{T}} v)/\pi \quad \forall u, v \in \mathbf{S}^{k-1}.$$

Moreover, there are also tight connections with graph realizations in the Euclidean space (cf. [5, 6]); see Sect. 2.3 for a brief discussion and Sect. 3.2 for further results.

Determining the Gram dimension can also be reformulated in terms of finding low rank positive semidefinite matrix completions of partial matrices, as we now see. We use the following notation: $\mathscr{S}^n$ denotes the set of symmetric $n \times n$ matrices and $\mathscr{S}_+^n$ is the cone of positive semidefinite (psd) matrices in $\mathscr{S}^n$. The subset

$$\mathscr{E}_n = \{X \in \mathscr{S}_+^n : X_{ii} = 1 \ \forall i \in [n]\},$$

consisting of all positive semidefinite matrices with an all-ones diagonal (aka the correlations matrices), is known as the *elliptope*. Given a graph $G = ([n], E)$, π_E denotes the projection from $\mathscr{S}^n$ onto the subspace $\mathbb{R}^E$ indexed by the edges of G. Then, the projection $\mathscr{E}(G) = \pi_E(\mathscr{E}_n)$ is known as the *elliptope* of the graph G. Given an integer $k \ge 1$, define the *rank constrained elliptope*

$$\mathscr{E}_{n,k} = \{X \in \mathscr{E}_n : \mathrm{rank}(X) \le k\},$$

and, for any graph G, its projection $\mathscr{E}_k(G) = \pi_E(\mathscr{E}_{n,k})$. Then the points x in the elliptope $\mathscr{E}(G)$ correspond precisely to those vectors $x \in \mathbb{R}^E$ that admit a Gram representation by unit vectors. Moreover, $x \in \mathscr{E}_k(G)$ precisely when it has a Gram representation by unit vectors in $\mathbb{R}^k$; that is:

$$x \in \mathscr{E}_k(G) \iff \mathrm{gd}(G, x) \leq k.$$

The elements of $\mathscr{E}(G)$ can be seen as the G-partial symmetric matrices, i.e., the partial matrices whose entries are specified at the off-diagonal positions corresponding to edges of G and whose diagonal entries are all equal to 1, that can be completed to a positive semidefinite matrix. Hence the problem of deciding membership in $\mathscr{E}(G)$ can be reformulated as the problem of testing whether a given G-partial matrix can be completed to a correlation matrix. Moreover, for fixed $k \geq 1$, the membership problem in $\mathscr{E}_k(G)$ is the problem of deciding whether a given G-partial matrix has a psd completion of rank at most k. Using the notion of Gram dimension this can be equivalently formalized as:

> *Given a graph $G = (V, E)$ and $x \in \mathbb{Q}^E$, decide whether $\mathrm{gd}(G, x) \leq k$.*

A first main result of this paper is to prove that this problem is $\mathcal{NP}$-hard for any fixed $k \geq 2$ (cf. Theorems 3 and 4). Additionally, we consider the problem of testing membership in the convex hull of the rank constrained elliptope:

> *Given a graph $G = (V, E)$ and $x \in \mathbb{Q}^E$, decide whether $x \in \mathrm{conv}\,\mathscr{E}_k(G)$.*

The study of this problem is motivated by the relevance of the convex set $\mathrm{conv}\,\mathscr{E}_k(G)$ to the maximum cut problem and to the rank constrained Grothendieck problem. Indeed, for $k = 1$, $\mathrm{conv}\,\mathscr{E}_1(G)$ coincides with the cut polytope of G and it is well known that linear optimization over the cut polytope is $\mathcal{NP}$-hard [12]. For any $k \geq 2$, the worst case ratio of optimizing a linear function over the elliptope $\mathscr{E}(G)$ versus the rank constrained elliptope $\mathscr{E}_k(G)$ (equivalently, versus the convex hull $\mathrm{conv}\,\mathscr{E}_k(G)$) is known as the *rank k Grothendieck constant* of the graph G (see [7] for results and further references). It is believed that linear optimization over $\mathrm{conv}\,\mathscr{E}_k(G)$ is also hard for any fixed k (cf., e.g., the quote of Lovász [20, p. 61]). We show that the strong membership problem in $\mathrm{conv}\,\mathscr{E}_k(G)$ is $\mathcal{NP}$-hard, thus providing some evidence of hardness of optimization (cf. Theorem 6).

Contents. In Sect. 2 we present some background geometrical facts about cut and metric polytopes, about elliptopes, and about Euclidean graph realizations. In Sect. 3 we show $\mathcal{NP}$-hardness of the membership problem in $\mathscr{E}_k(G)$ for any fixed $k \geq 2$; we use two different reductions depending whether $k = 2$ or $k \geq 3$. In Sect. 4 we show $\mathcal{NP}$-hardness of the membership problem in the convex hull of $\mathscr{E}_k(G)$ for any fixed $k \geq 2$. In Sect. 2.3 we discuss links to complexity results for Euclidean graph realizations, and in Sect. 5 we conclude with some open questions.

Notation. Throughout $K_n = ([n], E_n)$ is the complete graph on n nodes; C_n denotes the circuit of length n, with node set $[n]$ and with edges the pairs $\{i, i + 1\}$ for $i \in [n]$ (indices taken modulo n), and its set of edges is again denoted as C_n for simplicity. Given a graph $G = (V, E)$, its *suspension graph* ∇G is the new graph obtained from G by adding a new node, called the *apex* node and often denoted as 0, which is adjacent to all the nodes of G. A *minor* of G is any graph which can be obtained from G by iteratively deleting edges or nodes and contracting edges. For a vector $a \in \mathbb{R}^n$ and $S \subseteq [n]$ we set $a(S) = \sum_{i \in S} a_i$.

2 Preliminaries

We recall here some basic geometric facts about metric and cut polyhedra, about elliptopes, and about Euclidean graph realizations.

2.1 Metric and Cut Polytopes

First we give the definition of the *metric polytope* $\mathrm{MET}(G)$ of a graph $G = (V, E)$. As a motivation recall the following basic 3D geometric result: Given a matrix $X \in \mathscr{S}_3$ with an all ones-diagonal parametrized as before by $x_{ij} = \cos(\pi a_{ij})$ where $a_{ij} \in [0, 1]$, then $X \succeq 0$ if and only if the a_{ij}'s satisfy the following *triangle inequalities*:

$$a_{ij} \leq a_{ik} + a_{jk}, \ a_{ij} + a_{ik} + a_{jk} \leq 2 \tag{1}$$

for distinct $i, j, k \in \{1, 2, 3\}$. (See e.g. [4]). The elliptope $\mathscr{E}_3$ (or rather, its bijective image $\mathscr{E}(K_3)$) is illustrated in Fig. 1.

The metric polytope of the complete graph $K_n = ([n], E_n)$ is the polyhedron in $\mathbb{R}^{E_n}$ defined by the above $4\binom{n}{3}$ triangle inequalities (1). More generally, the metric polytope of a graph $G = ([n], E)$ is the polyhedron $\mathrm{MET}(G)$ in $\mathbb{R}^E$, which is defined by the following linear inequalities (in the variable $a \in \mathbb{R}^E$):

$$0 \leq a_e \leq 1 \ \ \forall e \in E, \tag{2}$$

$$a(F) - a(C \setminus F) \leq |F| - 1 \tag{3}$$

for all circuits C of G and for all odd cardinality subsets $F \subseteq C$.

As is well known, the inequality (2) defines a facet of $\mathrm{MET}(G)$ if and only if the edge e does not belong to a triangle of G, while (3) defines a facet of $\mathrm{MET}(G)$ if and only if the circuit C has no chord (i.e., two non-consecutive nodes on C are not adjacent in G). In particular, for $G = K_n$, $\mathrm{MET}(K_n)$ is defined by the triangle inequalities (1), obtained by considering only the inequalities (3) where C is

a circuit of length 3. Moreover, $\text{MET}(G)$ coincides with the projection of $\text{MET}(K_n)$ onto the subspace $\mathbb{R}^E$ indexed by the edge set of G. (See [9] for details.)

A main motivation for studying the metric polytope is that it gives a tractable linear relaxation of the cut polytope. Recall that the rank 1 matrices in the elliptope $\mathscr{E}_n$ are of the form uu^T for all $u \in \{\pm 1\}^n$. They are sometimes called the *cut matrices* since they correspond to the cuts of the complete graph K_n. The *cut polytope* $\text{CUT}(G)$ is defined as the projection onto $\mathbb{R}^E$ of the convex hull of the cut matrices:

$$\text{CUT}(G) = \pi_E(\text{conv}(\mathscr{E}_{n,1})). \tag{4}$$

It is always true that $\text{CUT}(G) \subseteq \text{MET}(G)$, and equality holds if and only if G has no K_5 minor [3]. Linear optimization over the cut polytope models the maximum cut problem, well known to be $\mathscr{N}\mathscr{P}$-hard [12], and testing membership in the cut polytope $\text{CUT}(K_n)$ or, equivalently, in the convex hull of the rank constrained elliptope $\mathscr{E}_{n,1}$, is an $\mathscr{N}\mathscr{P}$-complete problem [2].

2.2 Elliptopes

From the above discussion about the elliptope $\mathscr{E}_3$ and the metric polytope, we can derive the following necessary condition for membership in the elliptope $\mathscr{E}(G)$ of a graph G, which turns out to be sufficient when G has no K_4 minor.

Proposition 1 ([16]). *For any graph $G = (V, E)$,*

$$\mathscr{E}(G) \subseteq \left\{ x \in [-1, 1]^E : \frac{1}{\pi} \arccos x \in \text{MET}(G) \right\}.$$

Moreover, equality holds if and only if G has no K_4 minor.

This result permits, in particular, to characterize membership in the elliptope $\mathscr{E}(C_n)$ of a circuit.

Corollary 1 ([4]). *Consider a vector $x = \cos(\pi a) \in \mathbb{R}^{C_n}$ with $a \in [0, 1]^{C_n}$. Then, $x \in \mathscr{E}(C_n)$ if and only if a satisfies the linear inequalities*

$$a(F) - a(C_n \setminus F) \leq |F| - 1 \quad \forall F \subseteq C_n \text{ with } |F| \text{ odd.} \tag{5}$$

We also recall the following result of [18] which characterizes membership in the rank constrained elliptope $\mathscr{E}_k(C_n)$ of a circuit in the case $k = 2$; see Lemma 4 for an extension to arbitrary graphs.

Lemma 1 ([18]). *For $x \in [-1, 1]^{C_n}$, $x \in \mathscr{E}_2(C_n)$ if and only if there exists $\epsilon \in \{\pm 1\}^{C_n}$ such that $\epsilon^\mathsf{T} \arccos x \in 2\pi\mathbb{Z}$.*
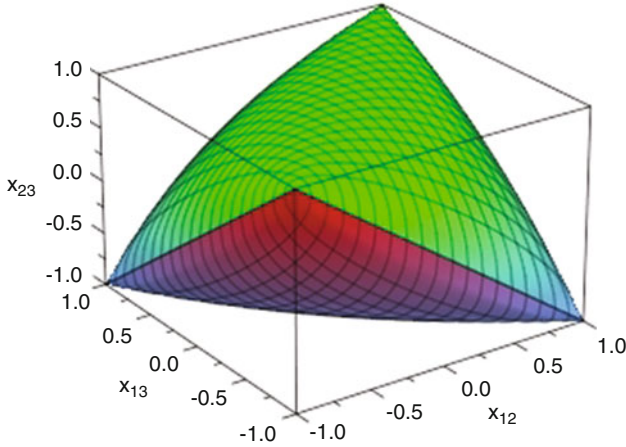
Fig. 1 The elliptope $\mathscr{E}(C_3)$

We conclude with some observations about the elliptope $\mathscr{E}(C_n)$ of a circuit. Figure 1 shows the elliptope $\mathscr{E}(C_3)$. Points x on the boundary of $\mathscr{E}(C_3)$ have $gd(C_3, x) = 2$ except $gd(C_3, x) = 1$ at the four corners (corresponding to the four cuts of K_3), while points in the interior of $\mathscr{E}(C_3)$ have $gd(C_3, x) = 3$.

Now let $n \geq 4$. Let $x = \cos(\pi a) \in \mathscr{E}(C_n)$, thus $a \in [0, 1]^{C_n}$ satisfies the inequalities (5). It is known that $gd(C_n, x) \leq 3$ (see [18], or derive it directly by triangulating C_n and applying Lemma 5 below). Moreover, x lies in the interior of $\mathscr{E}(C_n)$ if and only if x has a positive definite completion or, equivalently, a lies in the interior of the metric polytope $\mathrm{MET}(C_n)$.

If x lies on the boundary of $\mathscr{E}(C_n)$ then, either (i) $a_e \in \{0, 1\}$ for some edge e of C_n, or (ii) a satisfies an inequality (5) at equality. In case (i), $gd(C_n, x)$ can be equal to 1 (x is a cut), 2, or 3. In case (ii), by Lemma 1, $gd(C_n, x) \leq 2$ since $a(F) - a(C_n \setminus F) = |F| - 1 \in 2\mathbb{Z}$ for some $F \subseteq C_n$. If x is in the interior of $\mathscr{E}(C_n)$ then $gd(C_n, x) \in \{2, 3\}$.

As an illustration, for $n = 4$, consider the vectors $x_1 = (0, 0, 0, 1)^\mathsf{T}$, $x_2 = (0, \sqrt{3}/2, \sqrt{3}/2, \sqrt{3}/2)^\mathsf{T}$, $x_3 = (0, 0, 0, 0)^\mathsf{T}$ and $x_4 = (0, 0, 0, 1/2)^\mathsf{T} \in \mathbb{R}^{C_4}$. Then both x_1 and x_2 lie on the boundary of $\mathscr{E}(C_4)$ with $gd(C_4, x_1) = 3$ and $gd(C_4, x_2) = 2$, and both x_3 and x_4 lie in the interior of $\mathscr{E}(C_4)$ with $gd(C_4, x_3) = 2$ and $gd(C_4, x_4) = 3$.

2.3 Euclidean Graph Realizations

In this section we recall some basic facts about Euclidean graph realizations.

Definition 2. Given a graph $G = ([n], E)$ and $d \in \mathbb{R}_+^E$, a Euclidean (distance) representation of d in $\mathbb{R}^k$ consists of a set of vectors $p_1, \ldots, p_n \in \mathbb{R}^k$ such that

$$\|p_i - p_j\|^2 = d_{ij} \quad \forall \{i, j\} \in E.$$

Then, $\mathrm{ed}(G, d)$ denotes the smallest integer $k \geq 1$ for which d has a Euclidean representation in $\mathbb{R}^k$ (assuming d has a Euclidean representation in some space).

Then the problem of interest is to decide whether a given vector $d \in \mathbb{Q}_+^E$ admits a Euclidean representation in $\mathbb{R}^k$. Formally, for fixed $k \geq 1$, we consider the following problem:

> Given a graph $G = (V, E)$ and $d \in \mathbb{Q}_+^E$, decide whether $\mathrm{ed}(G, d) \leq k$.

This problem has been extensively studied (e.g. in [5, 6]) and its complexity is well understood. In particular, using a reduction from the 3SAT problem, Saxe [25] shows the following complexity result.

Theorem 1 ([25]). *For any fixed $k \geq 1$, deciding whether $\mathrm{ed}(G, d) \leq k$ is $\mathcal{NP}$-hard, already when restricted to weights $d \in \{1, 2\}^E$.*

We now recall a well known connection between Euclidean and Gram realizations. Given a graph $G = (V, E)$ and its suspension graph ∇G, consider the one-to-one map $\phi : \mathbb{R}^{V \cup E} \mapsto \mathbb{R}^{E(\nabla G)}$, which maps $x \in \mathbb{R}^{V \cup E}$ to $\varphi(x) = d \in \mathbb{R}^{E(\nabla G)}$ defined by

$$d_{0i} = x_{ii} \ (i \in [n]), \quad d_{ij} = x_{ii} + x_{jj} - 2x_{ij} \ (\{i, j\} \in E). \tag{6}$$

Then the vectors $u_1, \ldots, u_n \in \mathbb{R}^k$ form a Gram representation of x if and only if the vectors $u_0 = 0, u_1, \ldots, u_n$ form a Euclidean representation of $d = \varphi(x)$ in $\mathbb{R}^k$. This implies the following:

Lemma 2. *Let $G = (V, E)$ be a graph and $x \in \mathscr{E}(G)$. Then,*

$$\mathrm{gd}(G, x) = \mathrm{ed}(\nabla G, \varphi(x)).$$

As we will see in the next section, this connection will enable us to recover the above result of Saxe for the case $k \geq 3$ from results about the Gram dimension (cf. Corollary 2).

3 Testing Membership in $\mathscr{E}_k(G)$

In this section we discuss the complexity of testing membership in the rank constrained elliptope $\mathscr{E}_k(G)$. Specifically, for fixed $k \geq 1$ we consider the following problem:

> Given a graph $G = (V, E)$ and $x \in \mathbb{Q}^E$, decide whether $\mathrm{gd}(G, x) \leq k$.
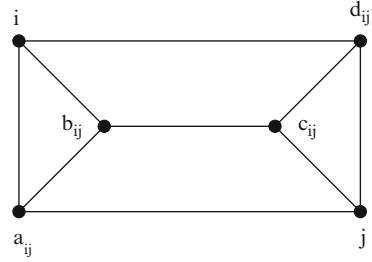
In the language of matrix completions this corresponds to deciding whether a rational G-partial matrix has a psd completion of rank at most k.

For $k = 1$, $x \in \mathscr{E}_1(G)$ if and only if $x \in \{\pm 1\}^E$ corresponds to a cut of G, and it is an easy exercise that this can be decided in polynomial time. In this section we show that the problem is $\mathscr{N}\mathscr{P}$-hard for any $k \geq 2$. It turns out that we have to use different reductions for the cases $k \geq 3$ and $k = 2$.

3.1   The Case $k \geq 3$

First we consider the problem of testing membership in $\mathscr{E}_k(G)$ when $k \geq 3$. We show that this is an $\mathscr{N}\mathscr{P}$-hard problem, already when $G = \nabla^{k-3} H$ is the suspension of a planar graph H and $x = 0$, the vector with zero entries at all edges.

The key idea is to relate the parameter $gd(G, \mathbf{0})$ to the chromatic number $\chi(G)$ (the minimum number of colors needed to color the nodes of G in such a way that adjacent nodes receive distinct colors). It is easy to check that

$$gd(G, \mathbf{0}) \leq \chi(G), \tag{7}$$

with equality if $\chi(G) \leq 2$ (i.e., if G is a bipartite graph). For $k \geq 3$ the inequality (7) can be strict. This is the case, e.g., for orthogonality graphs of Kochen-Specker sets (see [15]).

However, Peeters [22, Theorem 3.1] gives a polynomial time reduction of the problem of deciding 3-colorability of a graph to that of deciding $gd(G, \mathbf{0}) \leq 3$. Namely, given a graph G, he constructs (in polynomial time) a new graph G' having the property that

$$\chi(G) \leq 3 \iff \chi(G') \leq 3 \iff gd(G', \mathbf{0}) \leq 3. \tag{8}$$

The graph G' is obtained from G by adding for each pair of distinct nodes $i, j \in V$ the gadget graph H_{ij} shown in Fig. 2. Moreover, using a more involved construction, Peeters [21] constructs (in polynomial time) from any graph G a new *planar* graph

G' satisfying (8). As the problem of deciding whether a given planar graph is 3-colorable is $\mathscr{N}\mathscr{P}$-complete (see [26]) we have the following result.

Theorem 2 ([21]). *It is $\mathscr{N}\mathscr{P}$-hard to decide whether* $\mathrm{gd}(G, \mathbf{0}) \leq 3$, *already for the class of planar graphs.*

This hardness result can be extended to any fixed $k \geq 3$ using the suspension operation on graphs. The *suspension* graph $\nabla^p G$ is obtained from G by adding p new nodes that are pairwise adjacent and that are adjacent to all the nodes of G. It is an easy observation that

$$\mathrm{gd}(\nabla^p G, \mathbf{0}) = \mathrm{gd}(G, \mathbf{0}) + p. \tag{9}$$

Theorem 2 combined with equation (9) implies:

Theorem 3. *Fix $k \geq 3$. It is $\mathscr{N}\mathscr{P}$-hard to decide whether* $\mathrm{gd}(G, \mathbf{0}) \leq k$, *already for graphs of the form $G = \nabla^{k-3} H$ where H is a planar graph.*

As an application we can recover the complexity result of Saxe from Theorem 1 for the case $k \geq 3$.

Corollary 2. *For fixed $k \geq 3$, it is an $\mathscr{N}\mathscr{P}$–hard problem to decide whether* $\mathrm{ed}(G, d) \leq k$, *already when $G = \nabla^{k-2} H$ with H planar and d is $\{1, 2\}$-valued (more precisely, all edges adjacent to a given apex node have weight 1 and all other edges have weight 2).*

Proof. This follows directly from Lemma 2 combined with Theorem 3: By Lemma 2, $\mathrm{gd}(\nabla^{k-3} H, \mathbf{0}) = \mathrm{ed}(\nabla^{k-2} H, \varphi(\mathbf{0}))$ and observe that the image $d = \varphi(\mathbf{0})$ of the zero vector under the map φ from (6) satisfies: $d_{0i} = 1$ and $d_{ij} = 2$ for all nodes i, j of $\nabla^{k-3} H$. □

3.2 The Case $k = 2$

In this section we show $\mathscr{N}\mathscr{P}$-hardness of testing membership in $\mathscr{E}_2(G)$. Our strategy to show this result is as follows: Given a graph $G = (V, E)$ with edge weights $d \in \mathbb{R}_+^E$, define the new edge weights $x = \cos(d) \in \mathbb{R}^E$. We show a close relationship between the two problems of testing whether $\mathrm{ed}(G, d) \leq 1$, and whether $\mathrm{gd}(G, x) \leq 2$ (or, equivalently, $x \in \mathscr{E}_2(G)$). More precisely, we show that each of these two properties can be characterized in terms of the existence of a ± 1-signing of the edges of G satisfying a suitable 'flow conservation' type property; moreover, both are equivalent when the edge weights d are small enough.

As a motivation, let us consider first the case when $G = C_n$ is a circuit of length n. Say, weight d_i (resp., $x_i = \cos d_i$) is assigned to the edge $(i, i + 1)$ for $i \in [n]$ (where indices are taken modulo n). Then the following property holds:

$$\mathrm{ed}(C_n, d) \leq 1 \iff \exists \epsilon \in \{\pm 1\}^n \text{ such that } \epsilon^\mathsf{T} d = 0. \tag{10}$$

This is the key fact used by Saxe [25] for showing $\mathcal{N}\mathcal{P}$-hardness of the problem of testing $\mathrm{ed}(C_n, d) \leq 1$ by reducing the Partition problem for $d = (d_1, \cdots, d_n) \in \mathbb{Z}_+^n$ to it. Lemma 1 shows the analogous result for the Gram dimension:

$$\mathrm{gd}(C_n, \cos d) \leq 2 \iff \exists \epsilon \in \{\pm 1\}^n \text{ such that } \epsilon^\mathsf{T} d \in 2\pi\mathbb{Z}. \tag{11}$$

We now observe that these two characterizations extend for an arbitrary graph G. To formulate the result we need to fix an (arbitrary) orientation $\tilde{G}$ of G. Let $P = (u_0, u_1, \cdots, u_{k-1}, u_k)$ be a walk in G, i.e., $\{u_i, u_{i+1}\} \in E$ for all $0 \leq i \leq k - 1$. Recall that in a walk repetition of vertices is allowed; the walk P is said to be *closed* when $u_0 = u_k$. For $\epsilon \in \{\pm 1\}^E$, we define the following weighted sum along the edges of P:

$$\phi_{d,\epsilon}(P) = \sum_{i=0}^{k-1} d_{u_i, u_{i+1}} \epsilon_{u_i u_{i+1}} \eta_i, \tag{12}$$

setting $\eta_i = 1$ if the edge $\{u_i, u_{i+1}\}$ is oriented in $\tilde{G}$ from u_i to u_{i+1} and $\eta_i = -1$ otherwise.

Lemma 3. *Consider a graph $G = (V, E)$ with edge weights $d \in \mathbb{R}_+^E$ and fix an orientation $\tilde{G}$ of G. The following assertions are equivalent.*

(i) $\mathrm{ed}(G, d) \leq 1$.
(ii) *There exists an edge-signing $\epsilon \in \{\pm 1\}^E$ for which the function $\phi_{d,\epsilon}$ from (12) satisfies: $\phi_{d,\epsilon}(C) = 0$ for all closed walks C of G (equivalently, for all circuits of G).*

Proof. Assume that (i) holds. Let $f : V \to \mathbb{R}$ satisfying $|f(u) - f(v)| = d_{uv}$ for all $\{u, v\} \in E$. If the edge $\{u, v\}$ is oriented from u to v in $\tilde{G}$, let $\epsilon_{uv} \in \{\pm 1\}$ such that $f(v) - f(u) = d_{uv}\epsilon_{uv}$. This defines an edge-signing $\epsilon \in \{\pm 1\}^E$; we claim that (ii) holds for this edge-signing. For this, pick a circuit $C = (u_0, u_1, \cdots, u_k = u_0)$ in G. By construction of the edge-signing, the term $\epsilon_{u_i u_{i+1}} d_{u_i u_{i+1}} \eta_i$ is equal to $f(u_{i+1}) - f(u_i)$ for all $0 \leq i \leq k - 1$, where indices are taken modulo k. This implies that $\phi_{d,\epsilon}(C) = \sum_{i=0}^{k-1} f(u_{i+1}) - f(u_i) = 0$ and thus (ii) holds. Conversely, assume (ii) holds. We may assume that G is connected (else apply the following to each connected component). Fix an arbitrary node $u_0 \in V$. We define the function $f : V \to \mathbb{R}$ by setting $f(u_0) = 0$ and, for $u \in V \setminus \{u_0\}$, $f(u) = \phi_{d,\epsilon}(P)$ where P is any walk from u_0 to u. It is easy to verify that since (ii) holds this definition does not depend on the choice of P. We claim that f is a Euclidean embedding of (G, d) into $\mathbb{R}$. For this, pick an edge $\{u, v\} \in E$; say, it is oriented from u to v in $\tilde{G}$. Pick a walk P from u_0 to u, so that $Q = (P, v)$ is a walk from u_0 to v. Then, $f(u) = \phi_{d,\epsilon}(P)$, $f(v) = \phi_{d,\epsilon}(Q) = \phi_{d,\epsilon}(P) + d_{uv}\epsilon_{uv} = f(u) + d_{uv}\epsilon_{uv}$, which implies that $|f(v) - f(u)| = d_{uv}$. $\square$

Next we prove the analogous result for the spherical setting.

Lemma 4. *Consider a graph $G = (V, E)$ with edge weights $d \in \mathbb{R}_+^E$ and fix an orientation $\tilde{G}$ of G. The following assertions are equivalent.*

(i) $\mathrm{gd}(G, \cos d) \leq 2$.
(ii) *There exists an edge-signing $\epsilon \in \{\pm 1\}^E$ for which the function $\phi_{d,\epsilon}$ from (12) satisfies: $\phi_{d,\epsilon}(C) \in 2\pi\mathbb{Z}$ for all closed walks C of G (equivalently, for all circuits of G).*

Proof. Assume (i) holds. Then, there exists a labeling of the nodes $u \in V$ by unit vectors $g(u) = (\cos f(u), \sin f(u))$ where $f(u) \in [0, 2\pi]$ such that for any edge $\{u, v\} \in E$, we have $\cos d_{uv} = g(u)^\mathsf{T} g(v) = \cos(f(u) - f(v))$. If $\{u, v\}$ is oriented from u to v, define $\epsilon_{uv} \in \{\pm 1\}$ such that $f(v) - f(u) - \epsilon_{uv} d_{uv} \in 2\pi\mathbb{Z}$. This defines an edge-signing $\epsilon \in \{\pm 1\}^E$ which satisfies (ii) (same argument as in the proof of Lemma 3).

Conversely, assume (ii) holds. Analogously to the proof of Lemma 3, fix a node $u_0 \in V$ and consider the unit vectors $g(u_0) = (1, 0)$ and $g(u) = (\cos(\phi_{d,\epsilon}(P_u)), \sin(\phi_{d,\epsilon}(P_u)))$, where P_u is a walk from $u_0 \in V$ to $u \in V \setminus \{u_0\}$; one can verify that these vectors form a Gram realization of $(G, \cos d)$. $\qquad\square$

Corollary 3. *Consider a graph $G = (V, E)$ with edge weights $d \in \mathbb{R}_+^E$ satisfying $d(E) < 2\pi$. Then, $\mathrm{ed}(G, d) \leq 1$ if and only if $\mathrm{gd}(G, \cos d) \leq 2$.*

Proof. Note that if C is a circuit of G, then $\phi_{d,\epsilon}(C) \in 2\pi\mathbb{Z}$ implies $\phi_{d,\epsilon}(C) = 0$, since $|\phi_{d,\epsilon}(C)| \leq d(E) < 2\pi$. The result now follows directly by applying Lemmas 3 and 4. $\qquad\square$

We can now show $\mathcal{NP}$-hardness of testing membership in the rank constrained elliptope $\mathscr{E}_2(G)$. For this we use the result of Theorem 1 for the case $k = 1$: Given edge weights $d \in \{1, 2\}^E$, it is $\mathcal{NP}$-hard to decide whether $\mathrm{ed}(G, d) \leq 1$.

Theorem 4. *Given a graph $G = (V, E)$ and rational edge weights $x \in \mathbb{Q}^E$, it is $\mathcal{NP}$-hard to decide whether $x \in \mathscr{E}_2(G)$ or, equivalently, $\mathrm{gd}(G, x) \leq 2$.*

Proof. Fix edge weights $d \in \{1, 2\}^E$. We reduce the problem of testing whether $\mathrm{ed}(G, d) \leq 1$ to the problem of testing whether $\mathrm{gd}(G, \cos(\alpha d)) \leq 2$, for some appropriately chosen parameter $\alpha > 0$.

For this, define the angle $\alpha > 0$ by

$$\cos\alpha = \frac{16d(E)^2 - 1}{16d(E)^2 + 1} \in \mathbb{Q}, \quad \sin\alpha = \frac{8d(E)}{16d(E)^2 + 1} \in \mathbb{Q}.$$

As $d_e \in \{1, 2\}$ it follows that $\cos(\alpha d_e) \in \{\cos\alpha, \cos(2\alpha) = 2\cos^2\alpha - 1\}$ is rational valued for all edges $e \in E$. Moreover, the size of $\cos(\alpha d_e)$ is polynomial in the size of the instance (G, d) and thus the reduction can be carried out in polynomial time. Additionally, since $\sin\alpha < 1/(2d(E)) \leq 1/2 \leq \sin 1$, it follows that $\alpha < 1$ and thus $\alpha < 2\sin\alpha$. In turn, this implies that $\alpha < 1/d(E)$ and thus $\alpha d(E) < 2\pi$. Lastly, applying Corollary 3 we have that $\mathrm{gd}(G, \cos(\alpha d)) \leq 2$ is equivalent to $\mathrm{ed}(G, \alpha d) \leq 1$ and thus to $\mathrm{ed}(G, d) \leq 1$. $\qquad\square$

We conclude with a remark about the complexity of the Gram dimension of weighted circuits.

Remark 1. Consider the case when $G = C_n$ is a circuit and the edge weights $d \in \mathbb{Z}_+^{C_n}$ are integer valued. Relation (10) shows that $\mathrm{ed}(C_n, d) \leq 1$ if and only if the sequence $d = (d_1, \cdots, d_n)$ can be partitioned, thus showing $\mathcal{NP}$-hardness of the problem of testing $\mathrm{ed}(C_n, d) \leq 1$.

As in the proof of Theorem 4 let us choose α such that $\cos \alpha, \sin \alpha \in \mathbb{Q}$ and $\alpha < 1/(\sum_{i=1}^n d_i)$; then $\cos(t\alpha) \in \mathbb{Q}$ for all $t \in \mathbb{Z}$. The analogous relation (11) holds, which shows that $\mathrm{gd}(C_n, \cos(\alpha d)) \leq 2$ if and only if the sequence $d = (d_1, \cdots, d_n)$ can be partitioned. However, it is not clear how to use this fact in order to show $\mathcal{NP}$-hardness of the problem of testing $\mathrm{gd}(C_n, x) \leq 2$. Indeed, although all $\cos(\alpha d_i)$ are rational valued, the difficulty is that it is not clear how to compute $\cos(\alpha d_i)$ in time polynomial in the bit size of d_i (while it can be shown to be polynomial in d_i).

Finally we point out the following link to the construction of Aspnes et al. [1, §IV]. Consider the edge weights $x = \cos(\alpha d) \in \mathbb{R}^{C_n}$ for the circuit C_n and $y = \varphi(x)$ for its suspension ∇C_n, which is the wheel graph W_{n+1}. Thus $y_{0i} = 1$ and $y_{i,i+1} = 2 - 2\cos(\alpha d_i) = 4\sin^2(\alpha d_i/2)$ for all $i \in [n]$. Taking square roots we find the edge weights used in [1] to claim $\mathcal{NP}$-hardness of realizing weighted wheels (that have the property of admitting unique (up to congruence) realizations in the plane). As explained in the proof of Theorem 4, if we suitably choose α we can make sure that all $\sin(\alpha d_i/2)$ be rational valued, while [1] uses real numbers. However, it is not clear how to control their bit sizes, and thus how to deduce $\mathcal{NP}$-hardness.

4 Testing Membership in conv $\mathscr{E}_k(G)$

In the previous section we showed that testing membership in the rank constrained elliptope $\mathscr{E}_k(G)$ is an $\mathcal{NP}$-hard problem for any fixed $k \geq 2$. A related question is to determine the complexity of optimizing a linear objective function over $\mathscr{E}_k(G)$ or, equivalently, over its convex hull conv $\mathscr{E}_k(G)$. This question has been raised, in particular, by Lovász [20, p. 61] and more recently in [7], and we will come back to it in Sect. 5. In turn, this is related to the problem of testing membership in the convex hull conv $\mathscr{E}_k(G)$ which we address in this section. Specifically, for any fixed $k \geq 1$ we consider the following problem:

> Given a graph $G = (V, E)$ and $x \in \mathbb{Q}^E$, decide whether $x \in$ conv $\mathscr{E}_k(G)$.

For $k = 1$, conv $\mathscr{E}_1(G)$ coincides with the cut polytope of G, for which the membership problem is $\mathcal{NP}$-complete [2]. In this section we will show that this problem is $\mathcal{NP}$-hard for any fixed $k \geq 2$. The key fact to prove hardness is to

Fig. 3 The graph $\hat{C}_5$
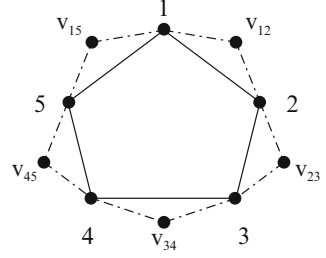
consider the membership problem in conv $\mathscr{E}_k(G)$ for extreme points of the elliptope $\mathscr{E}(G)$.

For a convex set K recall that a point $x \in K$ is an *extreme point* of K if $x = \lambda y + (1 - \lambda)z$ where $0 < \lambda < 1$ and $y, z \in K$ implies that $x = y = z$. The set of extreme points of K is denoted by ext K. Clearly, for $x \in \text{ext}\,\mathscr{E}(G)$,

$$x \in \text{conv}\,\mathscr{E}_k(G) \Longleftrightarrow x \in \mathscr{E}_k(G). \tag{13}$$

Our strategy for showing hardness of membership in conv $\mathscr{E}_k(G)$ is as follows: Given a graph $G = (V, E)$ and a rational vector $x \in \mathscr{E}(G)$, we construct (in polynomial time) a new graph $\hat{G} = (\hat{V}, \hat{E})$ (containing G as a subgraph) and a new rational vector $\hat{x} \in \mathbb{Q}^{\hat{E}}$ (extending x) satisfying the following properties:

$$\hat{x} \in \text{ext}\,\mathscr{E}(\hat{G}), \tag{14}$$

$$x \in \mathscr{E}_k(G) \Longleftrightarrow \hat{x} \in \mathscr{E}_k(\hat{G}). \tag{15}$$

Combining these two conditions with (13), we deduce:

$$x \in \mathscr{E}_k(G) \Longleftrightarrow \hat{x} \in \mathscr{E}_k(\hat{G}) \Longleftrightarrow \hat{x} \in \text{conv}\,\mathscr{E}_k(\hat{G}). \tag{16}$$

Given $G = (V, E)$, the construction of the new graph $\hat{G} = (\hat{V}, \hat{E})$ is as follows: For each edge $\{i, j\}$ of G, we add a new node v_{ij}, adjacent to the two nodes i and j. Let C_{ij} denote the clique on $\{i, j, v_{ij}\}$ and set $\hat{V} = V \cup \{v_{ij} : \{i, j\} \in E\}$. Then $\hat{G}$ has node set $\hat{V}$ and its edge set is the union of all the cliques C_{ij} for $\{i, j\} \in E$. As an illustration Fig. 3 shows the graph $\widehat{C_5}$.

Given $x \in \mathbb{Q}^E$, the construction of the new vector $\hat{x} \in \mathbb{Q}^{\hat{E}}$ is as follows: For each edge $\{i, j\} \in E$,

$$\hat{x}_{ij} = x_{ij}, \tag{17}$$

$$\hat{x}_{i,v_{ij}} = 4/5, \ \hat{x}_{j,v_{ij}} = 3/5 \qquad \text{if } x_{ij} = 0, \tag{18}$$

$$\hat{x}_{i,v_{ij}} = x_{ij}, \ \hat{x}_{j,v_{ij}} = 2x_{ij}^2 - 1 \ \text{if } x_{ij} \neq 0. \tag{19}$$

We will use the following result characterizing the extreme points of the elliptope $\mathscr{E}_3$.

Theorem 5 ([13]). *A matrix $X = (x_{ij}) \in \mathscr{E}_3$ is an extreme point of $\mathscr{E}_3$ if either* rank$(X) = 1$, *or* rank$(X) = 2$ *and* $|x_{ij}| < 1$ *for all* $i \neq j \in \{1, 2, 3\}$.

We also need the following well known (and easy to check) result permitting to construct points in the elliptope of clique sums of graphs.

Lemma 5. *Given two graphs $G_l = (V_l, E_l)$ ($l = 1, 2$), where $V_1 \cap V_2$ is a clique in both G_1, G_2, the graph $G = (V_1 \cup V_2, E_1 \cup E_2)$ is called their* clique sum. *Given $x_l \in \mathbb{R}^{E_l}$ ($l = 1, 2$) such that $(x_1)_{ij} = (x_2)_{ij}$ for $i, j \in V_1 \cap V_2$, let $x = (x_{ij}) \in \mathbb{R}^E$ be their common extension, defined as $x_{ij} = (x_l)_{ij}$ if $i, j \in V_l$. Then, for any integer $k \geq 1$,*

$$x \in \mathscr{E}_k(G) \iff x_1 \in \mathscr{E}_k(G_1) \text{ and } x_2 \in \mathscr{E}_k(G_2).$$

We can now show that our construction for $\hat{x}$ satisfies the two properties (14) and (15).

Lemma 6. *Given a graph $G = (V, E)$ and $x \in \mathbb{Q}^E$, let $\hat{G} = (\hat{V}, \hat{E})$ be defined as above and let $\hat{x} \in \mathbb{Q}^{\hat{E}}$ be defined by (17)–(19). For fixed $k \geq 2$ we have that $x \in \mathscr{E}_k(G)$ if and only if $\hat{x} \in \mathscr{E}_k(\hat{G})$ and $\hat{x} \in \text{ext}\,\mathscr{E}(\hat{G})$.*

Proof. Sufficiency is clear so it remains to prove necessity. Applying Theorem 5, we find that the two matrices

$$\begin{pmatrix} 1 & 0 & 3/5 \\ 0 & 1 & 4/5 \\ 3/5 & 4/5 & 1 \end{pmatrix}, \begin{pmatrix} 1 & x_{ij} & x_{ij} \\ x_{ij} & 1 & 2x_{ij}^2 - 1 \\ x_{ij} & 2x_{ij}^2 - 1 & 1 \end{pmatrix} \text{ where } x_{ij} \in [-1, 1] \setminus \{0\}, \quad (20)$$

are extreme points of $\mathscr{E}_3$. Therefore, for each edge $\{i, j\} \in E$, the restriction $\hat{x}_{C_{ij}}$ of $\hat{x}$ to the clique C_{ij} is an extreme point of $\mathscr{E}(C_{ij})$. By construction, $\hat{G}$ is obtained as the clique sum of G with the cliques C_{ij}. As both matrices in (20) have rank at most 2 and as $k \geq 2$, Lemma 5 implies that $\hat{x} \in \mathscr{E}_k(\hat{G})$.

Finally, we show that $\hat{x}$ is an extreme point of $\mathscr{E}(\hat{G})$. Assume that $\hat{x} = \sum_{k=1}^m \lambda_k \hat{x}_k$ where $\lambda_k > 0$, $\sum_{k=1}^m \lambda_k = 1$ and $\hat{x}_k \in \mathscr{E}(\hat{G})$. For any $\{i, j\} \in E$, taking the projection onto the clique C_{ij} and using the fact that $\hat{x}_{C_{ij}} \in \text{ext}\,\mathscr{E}(C_{ij})$ we deduce that $(\hat{x}_k)_{C_{ij}} = \hat{x}_{C_{ij}}$ for all $k \in [m]$. As the cliques $\{C_{ij} : \{i, j\} \in E\}$ cover the graph $\hat{G}$ it follows that $\hat{x} = \hat{x}_k$ for all $k \in [m]$. □

Combining these results we arrive at the main result of this section.

Theorem 6. *For any fixed $k \geq 2$, given a graph $G = (V, E)$ and rational edge weights $x \in \mathbb{Q}^E$, it is $\mathscr{N}\mathscr{P}$-hard to decide whether $x \in \text{conv}\,\mathscr{E}_k(G)$.*

Proof. We show that the problem is hard already when the input is restricted to extreme points of $\mathscr{E}(G)$. By relation (13), for such points, testing membership in conv $\mathscr{E}_k(G)$ is equivalent to testing membership in $\mathscr{E}_k(G)$.

In Theorems 3 and 4 we established that for any fixed $k \geq 2$ testing membership in $\mathscr{E}_k(G)$ is $\mathscr{N}\mathscr{P}$-hard. Using Lemma 6, testing membership in $\mathscr{E}_k(G)$ reduces to

testing membership in $\text{conv}\mathscr{E}_k(\hat{G})$ for extreme points of $\mathscr{E}_k(\hat{G})$. As the reduction described in Lemma 6 can be carried out in polynomial time, the latter problem is $\mathscr{N}\mathscr{P}$-hard. □

5 Concluding Remarks

In this note we have shown $\mathscr{N}\mathscr{P}$-hardness of the membership problem in the rank constrained elliptope $\mathscr{E}_k(G)$ and in its convex hull $\text{conv}\,\mathscr{E}_k(G)$, for any fixed $k \geq 2$. As mentioned earlier, it would be interesting to settle the complexity status of linear optimization over $\text{conv}\,\mathscr{E}_k(G)$. The case $k = 1$ is settled: Then $\text{conv}\,\mathscr{E}_1(G)$ is the cut polytope and both the membership problem and the linear optimization problem are $\mathscr{N}\mathscr{P}$-complete. For $k \geq 2$, the convex set $\text{conv}\,\mathscr{E}_k(G)$ is in general non-polyhedral. Hence the right question to ask is about the complexity of the *weak* optimization problem. It follows from general results about the ellipsoid method (see, e.g., [14] for details) that the weak optimization problem and the weak membership problems for $\text{conv}\,\mathscr{E}_k(G)$ have the same complexity status. Although we could prove that the (strong) membership problem in $\text{conv}\,\mathscr{E}_k(G)$ is $\mathscr{N}\mathscr{P}$-hard, we do not know whether this is also the case for the *weak* membership problem.

A second question of interest is whether the problems belong to $\mathscr{N}\mathscr{P}$. Indeed it is not clear how to find *succinct* certificates for membership in $\mathscr{E}(G)$ or in $\mathscr{E}_k(G)$. For one thing, even if the given partial matrix x is rational valued and is completable to a psd matrix, it is not known whether it admits a *rational* completion. (A positive result has been shown in [17] in the case of chordal graphs, and for graphs with minimum fill-in 1). In a more general setting, it is not known whether the problem of testing feasibility of a semidefinite program belongs to $\mathscr{N}\mathscr{P}$. On the positive side it is known that this problem belongs to $\mathscr{N}\mathscr{P}$ if and only if it belongs to co-$\mathscr{N}\mathscr{P}$ [24] and that it can be solved in polynomial time when fixing the dimension or the number of constraints [23].

References

1. Aspnes, J., Eren, T., Goldenberg, D.K., Morse, A.S., Whiteley W., Yang Y.R., Anderson B.D.O., Belhumeur, P.N.: A theory of network localization. IEEE Trans. Mobile Comput. **5**(12), 1663–1678 (2006)
2. Avis, D., Deza, M.: The cut cone, L1 embeddability, complexity, and multicommodity flows. Networks **21**(6), 595–617 (1991)
3. Barahona, F., Mahjoub, A.: On the cut polytope. Math. Program. **36**, 157–173 (1986)
4. Barrett, W.W., Johnson, C.R., Tarazaga, P.: The real positive definite completion problem: cycle completability. Mem. Am. Math. Soc. **584**, 69pp. (1996)

5. Belk, M.: Realizability of graphs in three dimensions. Disc. Comput. Geom. **37**, 139–162 (2007)
6. Belk, M., Connelly, R.: Realizability of graphs. Disc. Comput. Geom. **37**, 125–137 (2007)
7. Briët, J., de Oliveira Filho, F., Vallentin F.: Grothendieck inequalities for semidefinite programs with rank constraint (2011, Preprint). Available at arXiv:1011.1754v1
8. Colin de Verdière, Y.: Multiplicities of eigenvalues and tree-width of graphs. J. Comb. Theory B **74**(2), 121–146 (1998)
9. Deza, M.M., Laurent, M.: Geometry of Cuts and Metrics. Springer, New York (1997)
10. Fallat, S.M., Hogben, L.: The minimum rank of symmetric matrices described by a graph: a survey. Linear Algebra Appl. **426**, 558–582 (2007)
11. Fallat, S.M., Hogben, L.: Variants on the minimum rank problem: a survey II (2011, Preprint). Available at arXiv:1102.5142v1
12. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman, San Francisco (1979)
13. Grone, R., Pierce, S., Watkins, W.: Extremal correlation matrices. Linear Algebra Appl. **134**, 63–70 (1990)
14. Grötschel, M., Lovász, L., Schrijver, A.: Geometric Algorithms and Combinatorial Optimization. Springer, Berlin/New York (1988)
15. Haynes, G., Park, C., Schaeffer, A., Webster, J., Mitchell, L.M.: Orthogonal vector coloring. Electron. J. Comb. **17**, 1–18 (2010)
16. Laurent, M.: The real positive semidefinite completion problem for series-parallel graphs. Linear Algebra Appl. **252**(1–3), 347–366 (1997)
17. Laurent, M.: Polynomial instances of the positive semidefinite and Euclidean distance matrix completion problems. SIAM J. Matrix Anal. A **22**, 874–894 (2000)
18. Laurent, M., Varvitsiotis, A.: The Gram dimension of a graph. In: Proceedings of the 2nd International Symposium on Combinatorial Optimization, Athens. LNCS, vol. 7422, pp. 356–367 (2012)
19. Laurent, M., Varvitsiotis, A.: A new graph parameter related to bounded rank positive semidefinite matrix completions Mathematical Programming, Series A. (16 February 2013)
20. Lovász, L.: Geometric Representations of Graphs. Lecture Notes (2001). http://www.cs.elte.hu/~lovasz/geomrep.pdf
21. Peeters, R.: Ranks and structure of graphs. PhD thesis, Tilburg University (1995)
22. Peeters, R.: Orthogonal representations over finite fields and the chromatic number of graphs. Combinatorica **16**(3), 417–431 (1996)
23. Porkolab, L., Khachiyan, L.: On the complexity of semidefinite programs. J. Glob. Optim. **10**, 351–365 (1997)
24. Ramana, M.: An exact duality theory for semidefinite programming and its complexity implications. Math. Program. **77**(1), 129–162 (1997)
25. Saxe, J.B.: Embeddability of weighted graphs in k-space is strongly NP-hard. In: Proceedings of 17-th Allerton Conference on Communication Control Computing, Monticello, IL, pp. 480–489 (1979)
26. Stockmeyer, L.: Planar 3-colorability is polynomial complete. SIGACT News **5**(3), 19–25 (1973)

The Strong Dodecahedral Conjecture and Fejes Tóth's Conjecture on Sphere Packings with Kissing Number Twelve

Thomas C. Hales

Abstract This article sketches the proofs of two theorems about sphere packings in Euclidean 3-space. The first is K. Bezdek's strong dodecahedral conjecture: the surface area of every bounded Voronoi cell in a packing of balls of radius 1 is at least that of a regular dodecahedron of inradius 1. The second theorem is L. Fejes Tóth's conjecture on sphere packings with kissing number twelve, which asserts that in 3-space, any packing of congruent balls such that each ball is touched by 12 others consists of hexagonal layers. Both proofs are computer assisted. Complete proofs of these theorems appear in Hales TC (Dense sphere packings: a blueprint for formal proofs. London mathematical society lecture note series, vol 400. Cambridge University Press, Cambridge/New York, 2012; A proof of Fejes Tóth's conjecture on sphere packings with kissing number twelve. arXiv:1209.6043, 2012).

Key words Sphere packings • Discrete geometry • Voronoi cell

Subject Classifications: 52C17

1 The Strong Hexagonal Conjecture

To describe methods, we begin with a proof of the following elementary computer-assisted theorem in $\mathbb{R}^2$.

Theorem 1. *The perimeter of any bounded Voronoi cell of a packing of congruent balls of radius* 1 *in* $\mathbb{R}^2$ *is at least* $4\sqrt{3}$*, the perimeter of a regular hexagon with inradius* 1*.*

T.C. Hales (✉)
Department of Mathematics, University of Pittsburgh,
301 Thackeray Hall, Pittsburgh, PA 15260, USA
e-mail: hales@pitt.edu

Fig. 1 The truncated
boundary of the Voronoi cell
(*shaded*) is partitioned into
three types of pieces,
indicated with a thick curve
marked ℓ
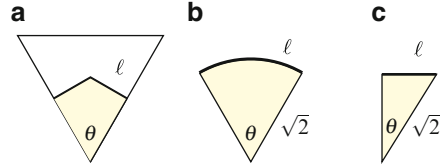
If we adopt the convention that the perimeter of an unbounded Voronoi cell is
infinite, then the boundedness hypothesis can be dropped from the theorem.

Proof. Fix a bounded Voronoi cell in a packing of congruent balls of radius 1 and fix
a coordinate system with the center of the Voronoi cell at the origin. The intersection
of the Voronoi cell with a disk of radius $\sqrt{2}$ at the origin is a convex disk whose
boundary C consists of circular arcs and straight line segments. The length of C is
no greater than the original perimeter of the Voronoi cell. It suffices to show that the
length of C is at least $4\sqrt{3}$.

The boundary C is determined by the set of centers V of balls at distance less
than $\sqrt{8}$ from the origin, excluding the ball centered at the origin.

The following piecewise linear function arises in the proof of the strong
dodecahedral conjecture in three-dimensions. We make repeated use of it.

$$L(h) = \begin{cases} (h_0 - h)/(h_0 - 1), & \text{if } h \leq h_0, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $h_0 = 1.26$.

Let $\mathbf{u}_1, \mathbf{u}_2 \in V$ be distinct points such that $T = \{\mathbf{0}, \mathbf{u}_1, \mathbf{u}_2\}$ has circumradius
less than $\sqrt{2}$. Let $\ell(\mathbf{u}_1, \mathbf{u}_2)$ be the length of the part of C contained in the convex
hull of T, and let θ be the angle at $\mathbf{0}$ between $\mathbf{u}_1$ and $\mathbf{u}_2$ (see Fig. 1a). The following
inequality has been verified by computer using interval arithmetic.

$$\ell(\mathbf{u}_1, \mathbf{u}_2) - b\theta(\mathbf{u}_1, \mathbf{u}_2) - cL(\|\mathbf{u}_1\|/2) - cL(\|\mathbf{u}_2\|/2) \geq 0, \tag{2}$$

where $b = 4/3$ and $c = \sqrt{3}/3 - 2\pi/9 \approx -0.12$. Equality holds when T is an
equilateral triangle with side 2.

Remark 1. Let $\mathbf{u}_1(t)$ and $\mathbf{u}_2(t)$ be points such that

$$\|\mathbf{u}_1(t)\| = \|\mathbf{u}_2(t)\| = 2, \quad \|\mathbf{u}_1(t) - \mathbf{u}_2(t)\| = t,$$

that is a triangle at the origin with sides $2, 2, t$. Along this curve with parameter t,
the constants b and c are the unique choice of constants that give the left-hand side
of (2) a local minimum with value 0 at $t = 2$.

The entire boundary C can be partitioned into finitely many (a) pieces lying in
convex hulls of triangles T, (b) arcs of circles of radius $\sqrt{2}$, and (c) linear segments
from $\mathbf{u}/2$, where $\mathbf{u} \in V$, to a point on the circle of radius $\sqrt{2}$ centered at the origin
(Fig. 1).

We extend the inequality (2) to boundary arcs of type (b) in the form

$$\ell - b\theta \geq 0 \tag{3}$$

where ℓ is the length of the circular arc and θ is the subtended angle. This inequality is obvious, because $\ell = \sqrt{2}\,\theta$, and $\sqrt{2} > b = 4/3$. We extend the inequality (2) to segments of type (c) in the form

$$\ell(\mathbf{v}) - b\theta(\mathbf{v}) - cL(h) \geq 0, \tag{4}$$

where $h = \|\mathbf{v}\|/2$, θ is the subtended angle at the origin, and $\ell = \sqrt{2 - h^2}$ is the length of the segment. When $h \leq h_0$, inequality (4) is a consequence of inequality (2), because the segment can be reflected through a mirror to form the two segments in (2). When $h \geq h_0$, the term $L(h)$ is zero. In this case, basic calculus gives the inequality.

Let ℓ_C be the length of C. We sum the inequalities over the boundary pieces of C of types (a), (b), and (c), using inequalities (2), (3), and (4):

$$\ell_C - b(2\pi) - 2c \sum_{\mathbf{v} \in V} L(\|\mathbf{v}\|/2) \geq 0. \tag{5}$$

The function L is zero on $\{h \mid h \geq h_0\}$. We drop such terms from the sum. Lemma 1 and Inequality (5) give

$$\ell_C \geq 2\pi b + 12c = 4\sqrt{3}.$$

This proves the theorem. □

The following lemma is used in the proof

Lemma 1. *Let V be a set of points contained in a closed annulus at the origin of inner radius* 2 *and outer radius* $2h_0$. *Assume that the mutual separation of points in V is at least* 2. *Then*

$$\sum_{\mathbf{v} \in V} L(\|\mathbf{v}\|/2) \leq 6.$$

Equality is obtained uniquely when V is the set of extreme points of a regular hexagon of circumradius 2.

Proof. In case card$(V) \leq 6$, by the inequality $L(\|\mathbf{v}\|/2) \leq 1$, it is clear that the sum is at most 6, with equality uniquely obtained for the regular hexagon of circumradius 2. An easy estimate shows that the angles at the origin between $\mathbf{u}, \mathbf{v} \in V$ is greater than $\pi/4$, so that card$(V) \leq 7$. We may therefore assume without loss of generality that card$(V) = 7$.

We index the seven points $\mathbf{v}_i$ by $i \in \mathbb{Z}/7\mathbb{Z}$ in their natural cyclic order around the annulus. Let θ_i be the angle subtended at the origin between $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$. Let

$$\alpha_i = \arccos(\|\mathbf{v}\|/4) - \pi/6.$$

We have the following inequality

$$\theta_i \geq \alpha_i + \alpha_{i+1}, \quad i \in \mathbb{Z}/7\mathbb{Z}, \tag{6}$$

which is proved by basic calculus: it follows from the intermediate value theorem and from an explicit analytic formula for the terms in the inequality [8, Lemma 6.107]. Further, we have

$$\alpha_i - 0.16 \, L(\|\mathbf{v}_i\|/2) - 0.32 \geq 0 \tag{7}$$

which is also proved by basic calculus: by a second derivative test the left-hand side is concave as a function of $\|\mathbf{v}_i\|$ so that the inequality holds if it holds at the endpoints $\|\mathbf{v}_i\| = 2, 2.52$, which is easily checked.

Summing θ_i over i we get

$$2\pi = \sum_i \theta_i \geq 2 \sum_i \alpha_i \geq 2(0.16) \sum_i L(\|\mathbf{v}_i\|/2) + 14(0.32).$$

Computing constants, we get

$$6 > \sum_i L(\|\mathbf{v}_i\|/2). \qquad \square$$

Remark 2. The proof can be organized in a way that carries over directly from two dimensions to three. In the first step (Lemma 1), $\sum L(\cdot)$ is shown to be at most the kissing number (which is 6 in dimension 2 and 12 in dimension 3). In the second step, the estimate of the boundary of the truncated Voronoi cell is transformed into an estimate of $\sum L(\cdot)$. The second step can be broken into two smaller steps: (a) use a simplex whose circumradius is less than $\sqrt{2}$ to design an inequality with a local minimum at the desired solution of to the Voronoi cell problem; (b) extend the inequality from part (a) so that it holds on a full geometric partition of the boundary of the truncated Voronoi cell. In a final short step, sum all the inequalities to obtain the desired result.

2   Marchal Cells

In this section, we give details of the partition of the boundary C of the truncated Voronoi cell. The partition is based on the partition of Euclidean space into Marchal cells [12].

Figure 2 shows a packing V of cardinality five. We use the constant $\sqrt{2}$ to partition the plane into levels $0, \ldots, 3$. Every point has level ≥ 0. For every $\mathbf{v} \in V$, we form a closed disk of radius $\sqrt{2}$. A point at level ≥ 1 is a point that lies inside some such disk. We form a closed rhombus of side $\sqrt{2}$ for every pair of distinct
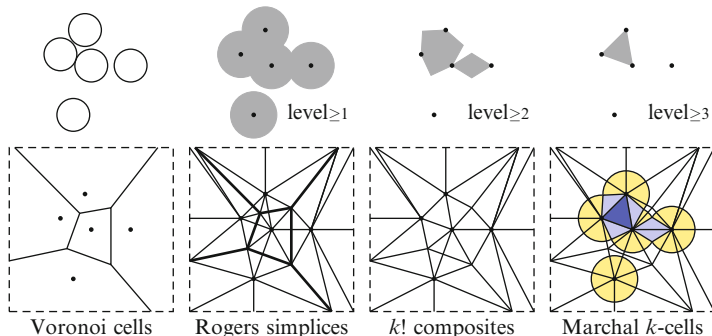
Fig. 2 Partitions of the plane (Image source [8])

Fig. 3 Convex hulls used to construct level sets in three dimensions

card(S)	X_S	conv($S \cup X_S$)
1	sphere	ball
2	circle	bi-cone
3	pair of points	bi-pyramid
4	∅	simplex

points in V whose separation is less than $\sqrt{8}$. By construction, the two points in V are opposite vertices of the rhombus. A point of level ≥ 2 is a point that lies some rhombus. We form a closed triangle for every triple of distinct points in V whose circumradius is less than $\sqrt{2}$. A point of level 3 is a point that lies inside some such triangle. No point has level ≥ 4. A point of level k is a point of level $\geq k$ that does not have level $\geq k + 1$.

The points of a given level can be further partitioned using the Rogers partition of the plane into simplices [15]. For each $k = 0, \dots, 3$, Rogers simplices that meet the set of level k can be naturally grouped into collections of $k!$ simplices. If P is the union of the $k!$ simplices, then the set of points of level k in P is called a *Marchal k-cell P_k*.

The construction can be generalized to three or more dimensions, again using the parameter $\sqrt{2}$. In n dimensions, the levels extend from 0 to $n + 1$ in an analogous manner. Let $S \subset V$ be a set of cardinality $k + 1$ whose circumradius is less than $\sqrt{2}$. The shapes used to define level sets are the convex hulls of

$$S \cup X_S$$

where X_S is the set of points at equidistance $\sqrt{2}$ from every point of S. The set X_S is sphere of dimension $n - \text{card}(S)$. When $n = 3$, the shapes are balls of radius $\sqrt{2}$, bicones, bipyramids, and tetrahedra (Fig. 3). Again in higher dimensions, the Rogers simplices can be grouped into collections of $k!$ simplices, giving Marchal k-cells P_k, at each level k.

Remark 3. Marchal introduced cells to show that the Kepler conjecture in three-dimensions can be reduced to an inequality of the form

$$\sum_{\mathbf{v} \in V} M(\|\mathbf{v}\|) \leq 12,$$

where M is a certain quartic polynomial, and V is a finite packing contained in a closed annulus of inner radius 2 and outer radius $\sqrt{8}$.

Remark 4. The book [8] strengthens Marchal's argument to reduce the Kepler conjecture to the inequality

$$\sum_{\mathbf{v} \in V} L(\|\mathbf{v}\|/2) \leq 12, \qquad (L_{12})$$

where L is the function defined above, and V is a packing in the closed annulus with inner radius 2 and outer radius $2h_0$. (In adapting this inequality from dimension two to dimension three, the two-dimensional kissing number 6 is replaced with the three-dimensional kissing number 12.) The book also gives a computer-assisted proof of the inequality (L_{12}), to obtain a new proof of the Kepler conjecture. This article shows how to deduce the strong dodecahedral conjecture and Fejes Tóth's conjecture on packings with kissing number twelve from (L_{12}).

Remark 5. An old conjecture by L. Fejes Tóth [5, p. 178] asserts that the minimum of

$$\sum_{\mathbf{v} \in V} \|\mathbf{v}\|, \qquad (8)$$

is $24 + 14/\sqrt{27} \approx 26.69$ as V runs over packings of cardinality 13 contained in a closed annulus with inner radius 2 and outer radius $\sqrt{8}$. The inequality (L_{12}) gives the best known result:

$$\sum_{\mathbf{v} \in V} \|\mathbf{v}\| \geq 24 + 2h_0 = 26.52. \qquad (9)$$

The inequality (L_{12}) also gives upper bounds for the Tammes problem when $\mathrm{card}(V) = 13, 14, 15$, but these upper bounds are weaker than those known by semi-definite programming [1].

3 Strong Dodecahedral Conjecture

This section sketches a proof of the strong dodecahedral conjecture [2]:

Theorem 2. *The surface area of every bounded Voronoi cell in a packing of balls of radius* 1 *is at least the surface area of a regular dodecahedron of inradius* 1. *Equality is obtained exactly when the bounded Voronoi cell is itself a regular dodecahedron.*

Remark 6. Fejes Tóth's classical dodecahedral conjecture [3] is the same statement, replacing *surface area* with *volume*. The classical dodecahedral conjecture was proved by McLaughlin [11]. To deduce the volume statement from the surface area statement, it is enough to use the volume formula $Bh/3$ for a tetrahedron, where B is its base area (the face of a Voronoi cell), and $h \geq 1$ is its height.

Proof. We pick coordinates so that the center of a chosen bounded Voronoi cell is at the origin. As in the two-dimensional case, we may intersect the Voronoi cell with a closed ball of radius $\sqrt{2}$. The boundary C after truncation is no greater than before. Let V be the set of centers of the packing in the annulus with inner radius 2 and outer radius $\sqrt{8}$.

There is a partition $C \cap P_k$ of C associated with the set of Marchal k-cells P_k associated with Rogers simplices at the origin. Write

$$\text{area}(C) = \sum_{P_k} \text{area}(C \cap P_k),$$

for the areas of the various contributions. Write $\text{sol}(P_k)$ for the solid angle of the Marchal cell at the origin, and write $\text{dih}(P_k, \mathbf{v})$ for the dihedral angle of a Marchal cell P_k along the edge through the line through $\{\mathbf{0}, \mathbf{v}\}$.

As a reference cell, we let $P_{D,4}$ be a Marchal 4-cell of the packing giving the regular dodecahedron of inradius 1. There exist constants a_D and $b_D > 0$ such that

$$\text{area}(C \cap P_k) + 3a_D \text{sol}(P_k) + 3b_D \sum_{\mathbf{v} \in P_k \cap V} L(\|\mathbf{v}\|/2) \text{dih}(P_k, \mathbf{v}) \geq 0, \qquad (10)$$

for all k-cells P_k and for all V. The constants $a_D \approx -0.581$ and $b_D \approx 0.0232$ are uniquely determined if we insist that equality is attained when $P_k = P_{D,4}$. This inequality has been proved by computer by interval arithmetic.

In more detail, the constants a_D and b_D are determined by a 1-dimensional family $P_4(t)$ of tetrahedra with sides $2, 2, 2, t, t, t$, for $t \in \mathbb{R}$ where the three edges of length 2 meet at the origin. When $t = t_D \approx 2.1029$ (the separation of balls in the arrangement giving the regular dodecahedron), $P_4(t_D)$ is congruent to $P_{D,4}$. Forcing the equality to be exact and the derivative to vanish when $t = t_D$, we obtain two linear equations in two unknowns that determine a_D and b_D.

If we sum (10) over all cells, the solid angles sum to 4π, dihedral angles sum to 2π, and L sums to at most 12 by Inequality (L_{12}), giving

$$\text{area}(C) = \sum_{P_k} \text{area}(C \cap P_k) \geq -3a_D 4\pi - 3b_D (2\pi)(12).$$

By the choice of a_D and b_D, equality is obtained for the boundary C_D of the regular dodecahedron,

$$\text{area}(C_D) = -3a_D 4\pi - 3b_D (2\pi)(12).$$

Hence area$(C) \geq$ area(C_D). This is the desired conclusion. (The circumradius of the regular dodecahedron is less than $\sqrt{2}$ so that C_D is both the truncated and untruncated boundary.) Tracing through the case of equality, inequality (10) is an equality exactly when the cell has measure zero or is congruent to $P_{D,4}$.

4 Fejes Tóth's Conjecture on Packings with Kissing Number Twelve

L. Fejes Tóth conjectured the following result in 1969 [4, 6].

Theorem 3. *In 3-space any packing of equal ball such that each ball is touched by 12 others consists of hexagonal layers.*

The proof of this theorem is much longer than the proof of the strong dodeca-hedral conjecture. This section describes the proof strategy. The details are found in [9].

It is enough to prove that the contact pattern of every ball is the hexagonal-close packing (HCP) or face-centered cubic (FCC) kissing arrangement, because these can only be extended in hexagonal layers. In fact, the HCP piece has a preferred plane of symmetry. Once a single HCP piece occurs, the preferred plane must be filled with HCP pieces. A plane forces another hexagonal layer above it and another hexagonal layer below it, leading to a packing of hexagonal layers. If no HCP piece occurs, the packing is the face-centered cubic packing, which also consists of hexagonal layers.

Lemma 2. *Let V be a packing in which every ball touches 12 others. Then for all distinct $\mathbf{u}, \mathbf{v} \in V$, either $\|\mathbf{u} - \mathbf{v}\| = 2$ or $\|\mathbf{u} - \mathbf{v}\| \geq 2h_0$.*

Proof. Let $\mathbf{u}_1, \ldots, \mathbf{u}_{12}$ be the 12 kissing points around $\mathbf{u}$. Assume that $\mathbf{v} \neq \mathbf{u}_i, \mathbf{u}$. By Inequality (L_{12}),

$$L(h(\mathbf{u}, \mathbf{v})) + 12 = L(h(\mathbf{u}, \mathbf{v})) + \sum_{i=1}^{12} L(h(\mathbf{u}, \mathbf{u}_i)) \leq 12.$$

This implies that $L(h(\mathbf{u}, \mathbf{v})) \leq 0$, so $\|\mathbf{u} - \mathbf{v}\| \geq 2h_0$.   □

4.1 Graph Classification Problems

Definition 1. Let S^2 be the sphere of radius 2, centered at $\mathbf{0}$. Let $\mathscr{V}$ be the set of packings $V \subset \mathbb{R}^3$ such that

1. card$(V) = 12$,
2. $V \subset S^2$,
3. $\|\mathbf{u} - \mathbf{v}\| \in \{0, 2\} \cup [2.52, 4]$ for all $\mathbf{u}, \mathbf{v} \in V$.

For each $V \in \mathcal{V}$, let E_{ctc} be the contact graph on vertex set V; that is, the set of $\{\mathbf{u}, \mathbf{v}\} \subset V$ such that $\|\mathbf{u} - \mathbf{v}\| = 2$.

Fejes Tóth's conjecture follows from the Inequality (L_{12}), together with a proof that the classification of graphs (V, E_{ctc}) with $V \in \mathcal{V}$ up to isomorphism contains exactly two graphs: the FCC contact graph and the HCP contact graph.

We formulate Inequality (L_{12}) as a graph classification problem as well. The inequality holds trivially for a finite packing of cardinality at most 12. For a contradiction, we may assume that V belongs to the set of finite packings of cardinality at least 13, contained in a closed annulus of radii $[2, 2h_0]$ and that violate the inequality:

$$\sum_{\mathbf{v} \in V} L(\|v\|/2) > 12.$$

Let E_{std} be the set of edges $\{\mathbf{u}, \mathbf{v}\} \subset V$ such that $2 \leq \|\mathbf{u} - \mathbf{v}\| \leq 2h_0$. The graph classification problem equivalent to (L_{12}) is that the set of graphs (V, E_{std}), with V from this set of counterexamples, is empty.

In summary, the proof of Fejes Tóth's conjecture consists of two graph classification problems: one for the contact graphs (V, E_{ctc}) involving vertex sets of cardinality 12 and one for the graphs (V, E_{std}) involving vertex sets of cardinality at least 13 for the Inequality (L_{12}). The proofs of these two classification results differ in detail, but the high-level structure is the same in both cases. The graphs are first represented combinatorially as hypermaps. (A hypermap is a finite set D together with three permutations e, n, f of D that satisfy $enf = I$, the trivial permutation.) A computer program classifies the hypermaps satisfying given combinatorial properties obtained from the constraints imposed on the graphs (V, E). Linear programs eliminate the extraneous cases; namely, those hypermaps that exist combinatorially but that do not admit a geometric realization. Finally, the inequalities used in the linear programs are proved by computer.

4.2 Hypermap Classification by Computer

The computer program that classifies hypermaps has been the subject of a exhaustive computer code formal verification project by G. Bauer and T. Nipkow [13]. The original scope of the project was the set of graphs from the 1998 proof of the Kepler conjecture [10], but in 2010, Nipkow extended this work to include the classification of hypermaps needed for the L_{12} inequality.[1]

---

[1]There are about 25,000 graphs that arise in the L_{12} classification and only 8 graphs that arise in the contact graph. Because of the vast difference in complexity of these two classification problems, our discussion will focus on the L_{12} classification.

4.3 Linear Programs

As mentioned above, linear programs eliminate the extraneous cases. The technology related to the linear programming has been significantly improved in the years following the proof of the Kepler conjecture. The thesis of S. Obua implemented the formal verification of linear programming proof certificates and used this to eliminate about 92% of the graphs that appear in the original proof of the Kepler conjecture [14]. More recent work by Solvyev has optimized the formal verification of linear programs to such a degree that the speed of the formal verification of a linear program rivals the speed of the unverified execution of a linear program [16]. Work in progress by Solovyev intends to make a full formal verification of all linear programs needed to prove Inequality (L_{12}).

The linear programs are generated in GLPK from an AMPL model that is independent of the hypermap. An OCaml program generates a separate AMPL data file for each linear program. When a single linear program fails to eliminate a hypermap, branch and bound methods are used to iteratively subdivide the domain into smaller pieces until linear programs are obtained that eliminate the hypermap. The process that was used to obtain a system of linear programming inequalities that works uniformly on all hypermaps was fully automated [7]. In brief synopsis, when a linear program fails to eliminate a hypermap, two models of corresponding metric graph are compared, one based purely on the linear programming estimates of lengths and angles, and a second nonlinear model based on nonlinear relations between lengths and angles. A comparison of models is used to determine inadequacies in the linearization. This data is fed to a Mathematica program based on various heuristics to construct a candidate nonlinear inequality. The inequality is then shipped to the nonlinear optimization package CFSQP for extensive nonrigorous testing. From there, a formal specification of the inequality is generated in the proof assistant HOL Light. The formal specification is exported to program that uses interval arithmetic to verify inequalities by computer; and finally, the AMPL model is automatically updated with the new inequality. This process works remarkably well in practice to develop a small set of inequalities[2] that uniformly eliminate all undesired hypermaps.

4.4 The Classification of Contact Graphs

As mentioned in an earlier footnote, there are about 25,000 graphs that arise in the L_{12} classification and only 8 graphs that arise in the contact graph classification. For the eight graphs, it was not necessary to follow the lengthy linear programming procedure described in the previous subsection. This final subsection sketches a much simpler procedure to eliminate the unwanted cases. (Two of the eight possibilities are the HCP and FCC, and the other six cases are unwanted.)

---

[2]About 500 inequalities occur.
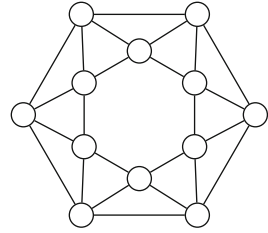
Five of the six are eliminated with linear programming inequalities. The linear programs are based on the following simple inequalities:

1. The angles around each node sum to 2π.
2. The angle of each triangle in the contact graph equals $\arccos(1/3) \approx 1.23096$.
3. The opposite angles of each rhombus are equal.
4. Each angle of every rhombus is between 1.6292 and 2.16672.

The final case is the graph shown in Fig. 4. It is eliminated with the following observations. The perimeter of a spherical hexagon with sides $\pi/3$ is 2π. However, the hexagons in the graph are spherically convex, and 2π is a strict upper bound on the perimeter of a spherically convex hexagon. Thus, this case does not admit a geometric realization as a contact graph. Fejes Tóth's conjecture on sphere packings with kissing number twelve ensues. □

References

1. Bachoc, C., Vallentin, F.: New upper bounds for kissing numbers from semidefinite programming. J. Am. Math. Soc. **21**, 909–924 (2008)
2. Bezdek, K.: On a stronger form of Rogers' lemma and the minimum surface area of Voronoi cells in unit ball packings. J. Reine Angew. Math. **518**, 131–143 (2000)
3. Fejes Tóth, L.: über die dichteste Kugellagerung. Mathematische Zeitschrift **48**, 676–684 (1943)
4. Fejes Tóth, L.: Remarks on a theorem of R.M. Robinson. Studia Scientiarum Hungarica **4**, 441–445 (1969)
5. Fejes Tóth, L.: Lagerungen in der Ebene auf der Kugel und im Raum, 2nd edn. Springer, Berlin/New York (1972)
6. Fejes Tóth, L.: Research problems. Periodica Mathematica Hungarica **29**, 89–91 (1989)
7. Hales, TC.: Linear programs for the Kepler conjecture. In: Fukuda, K., van der Hoeven, J., Joswig, M., Takayama, N. (eds.) Mathematical Software – ICMS 2010, Springer, Berlin/New York (2010)
8. Hales, T.C.: Dense sphere packings: a blueprint for formal proofs. London Mathematical Society Lecture Note Series, vol. 400. Cambridge University Press, Cambridge/New York (2012)
9. Hales, T.C.: A proof of Fejes Tóth's conjecture on sphere packings with kissing number twelve. arXiv:1209.6043 (2012)

10. Hales, TC., Ferguson, S.P.: The Kepler conjecture. Discret. Comput. Geom. **36**(1), 1–269 (2006)
11. Hales, T.C., McLaughlin, S.: A proof of the dodecahedral conjecture. J. AMS **23**, 299–344 (2010). http://arxiv.org/abs/math/9811079
12. Marchal, C.: Study of the Kepler's conjecture: the problem of the closest packing. Math. Z. **267** (3–4), 737–765 (2011)
13. Nipkow, T., Bauer, G., Schultz, P.: Flyspeck I: tame graphs. In: Furbach, U., Shankar, N. (eds.) International Joint Conference on Automated Reasoning, Seattle. Lecture Notes in Computer Science, vol. 4130, pp. 21–35. Springer (2006)
14. Obua, S.: Flyspeck II: the basic linear programs. Ph.D. thesis, Technische Universität München (2008)
15. Rogers, C.A.: The packing of equal spheres. J. Lond. Math. Soc. **3/8**, 609–620 (1958)
16. Solovyev, A., Hales, T.C.: Efficient formal verification of bounds of linear programs. LNCS, vol. 6824, pp. 123–132. Springer, Berlin (2011)

Solving Nuclear Norm Regularized and Semidefinite Matrix Least Squares Problems with Linear Equality Constraints

Kaifeng Jiang, Defeng Sun, and Kim-Chuan Toh

Abstract We introduce a partial proximal point algorithm for solving nuclear norm regularized and semidefinite matrix least squares problems with linear equality constraints. For the inner subproblems, we show that the positive definiteness of the generalized Hessian of the objective function for the inner subproblems is equivalent to the constraint nondegeneracy of the corresponding primal problem, which is a key property for applying a semismooth Newton-CG method to solve the inner subproblems efficiently. Numerical experiments on large scale matrix least squares problems arising from low rank matrix approximation, as well as regularized kernel estimation and Euclidean distance matrix completion problems in molecular conformation, show that our algorithm is efficient and robust.

K. Jiang
Department of Mathematics, National University of Singapore, Block S17,
10 Lower Kent Ridge Road, Singapore 119076, Singapore
e-mail: kaifengjiang@nus.edu.sg

D. Sun
Department of Mathematics and Risk Management Institute, National University
of Singapore, Blk S17, 10 Lower Kent Ridge Road, Singapore 119076, Singapore
e-mail: matsundf@nus.edu.sg

K.-C. Toh (✉)
Department of Mathematics, National University of Singapore, Blk S17, 10 Lower
Kent Ridge Road, Singapore 119076, Singapore

Singapore-MIT Alliance, 4 Engineering Drive 3, Singapore 117576, Singapore
e-mail: mattohkc@nus.edu.sg

1 Introduction

Let $\Re^{p \times q}$ be the space of all $p \times q$ matrices equipped with the standard trace inner product $\langle X, Y \rangle = \text{Tr}(X^T Y)$ and its induced Frobenius norm $\| \cdot \|$. Without loss of generality, we assume $p \leq q$ throughout this paper. For a given $X \in \Re^{p \times q}$, its nuclear norm $\|X\|_*$ is defined as the sum of all its singular values and its operator norm $\|X\|_2$ is the largest singular value. Let $\mathscr{S}^n$ be the space of all $n \times n$ symmetric matrices and $\mathscr{S}^n_+$ be the cone of symmetric positive semidefinite matrices. We use the notation $X \succeq 0$ to denote that X is a symmetric positive semidefinite matrix.

In this paper, we consider the following nuclear norm regularized matrix least squares problem with linear equality constraints:

$$\min_{X \in \Re^{p \times q}} \left\{ \frac{1}{2} \|\mathscr{A}(X) - b\|^2 + \rho \|X\|_* + \langle C, X \rangle \ : \ \mathscr{B}(X) = d \right\}, \qquad (1)$$

where $\mathscr{A} : \Re^{p \times q} \rightarrow \Re^m$ and $\mathscr{B} : \Re^{p \times q} \rightarrow \Re^s$ are given linear maps, $C \in \Re^{p \times q}, b \in \Re^m, d \in \Re^s$, and ρ is a given positive parameter. Note that the nuclear norm has been a very popular regularizer which favors a low rank solution of (1) [4, 9, 10, 25]. The problem (1) arises in many applications when one needs to find a low rank approximation of a given matrix while preserving certain desired structures. In many data analysis problems, the collected empirical data, which are usually messy and incomplete, typically do not have the specified structure or the desired low rank. It is important to find the nearest low rank approximation of the given matrix while maintaining the underlying structure of the original system. For example, in statistics, the regression matrix for the multiple regression model with a constant term has a column of all ones, and this column should not be perturbed during the low rank approximation.

When $C = 0$ and either $\mathscr{A}$ or $\mathscr{B}$ is absent in (1), the problem (1) includes the well studied matrix completion problem if either $\mathscr{B}$ or $\mathscr{A}$ is the projection onto the set of observed matrix entries. We should mention that many specialized first-order algorithms have been designed for various variants of the matrix completion problem; see for example [3, 20, 23, 29]. But as far as we are aware of, no papers have specifically discussed the nuclear norm regularized matrix least squares problem with additional structural constraints. The problem can of course be solved by several general first-order methods such as [6–8, 11, 14]. However, our numerical experiments show that these first-order methods may not achieve a satisfactory level of accuracy within a reasonable time.

In this paper, we design a partial proximal point algorithm (PPA) proposed by Ha [17] for solving (1), in which only some of the variables appear in the quadratic proximal term. Given a sequence of parameters σ_k such that

$$0 < \sigma_k \uparrow \sigma_\infty \leq +\infty, \qquad (2)$$

and an initial point $X^0 \in \Re^{p \times q}$, the partial PPA for solving (1) generates a sequence $\{(u^k, X^k)\} \subseteq \Re^m \times \Re^{p \times q}$ via the following scheme:

$$(u^{k+1}, X^{k+1}) \approx \arg\min \left\{ f_\rho(u, X) + \frac{1}{2\sigma_k} \|X - X^k\|^2 \, : \, \mathscr{A}(X) + u = b, \, \mathscr{B}(X) = d \right\}, \tag{3}$$

where $f_\rho(u, X) := \frac{1}{2}\|u\|^2 + \rho\|X\|_* + \langle C, X \rangle$. A key issue in the partial PPA which we must address is how to solve the partially regularized problem (3) efficiently. In our algorithm, we solve (3) via its dual, which is an unconstrained concave maximization problem whose objective function is continuously differentiable but not twice continuously differentiable. Because of the latter property, standard Newton's method cannot be used to solve the inner subproblem. However, we can show that the objective function is strongly semismooth due to the strong semismoothness of the soft thresholding operator [19, Theorem 2.1]. Thus we can apply the semismooth Newton method of Qi and Sun [24] to solve the inner subproblem. Based on the classical augmented Lagrangian framework of Rockafellar [27, 28], recently Zhao, Sun and Toh [32] proposed a Newton-CG augmented Lagrangian (SDPNAL) method for solving SDP problems, in which the inner subproblems are solved by using an inexact semismooth Newton-CG method. Their numerical results on a variety of large scale SDP problems demonstrated that the SDPNAL method is very efficient. This strongly motivated us to use a semismooth Newton-CG (SSNCG) method to solve the inner subproblems for achieving fast convergence. For our case, the global and fast local convergence of the SSNCG method is established under a constraint nondegeneracy condition, together with the strong semismoothness property of the soft thresholding operator.

The partial PPA which we will develop for solving (1) can easily be modified to solve the following semidefinite matrix least squares problem:

$$\min_{X \in \mathscr{S}^n} \left\{ \frac{1}{2}\|\mathscr{A}(X) - b\|^2 + \langle C, X \rangle \, : \, \mathscr{B}(X) = d, \, X \succeq 0 \right\}, \tag{4}$$

where $\mathscr{A} : \mathscr{S}^n \to \Re^m$ and $\mathscr{B} : \mathscr{S}^n \to \Re^s$ are given linear maps, $b \in \Re^m, d \in \Re^s$, and $C \in \mathscr{S}^n$. Thus in this paper, we also design a partial PPA to solve (4).

For the partial PPA (with SSNCG method for solving the inner subproblems) we have designed and implemented numerical experiments on large scale matrix least squares problems arising from low rank matrix approximation, as well as regularized kernel estimation and Euclidean distance matrix completion problems in molecular conformation, show that our algorithm is efficient and robust.

The remaining parts of this paper are organized as follows. In Sect. 2, we present some preliminaries about semismooth functions. In Sect. 3, we describe how to use the partial PPA to solve (1) and introduce a SSNCG method for solving the inner subproblems. The convergence analysis of our proposed algorithm is also established. In Sect. 4, we briefly explain how the SSNCG partial PPA for solving (1) can be modified to solve (4). In Sect. 5, we report the numerical performance of our algorithm for solving the various classes of problems mentioned in the last paragraph. We conclude the paper in Sect. 6.

2 Preliminaries

In this section, we give a brief introduction on some basic concepts such as the B-subdifferential and Clarke's generalized Jacobian of the soft-thresholding operator. These concepts and properties will be critical for us to develop a SSNCG method for solving the inner subproblems in our partial PPA.

Let $F : \Re^m \longrightarrow \Re^l$ be a locally Lipschitz function. By Rademacher's theorem, F is Fréchet differentiable almost everywhere. Let D_F denote the set of points where F is differentiable. The B-subdifferential of F at $x \in \Re^m$ is defined by

$$\partial_B F(x) := \{V : V = \lim_{k \to \infty} F'(x^k), \ x^k \longrightarrow x, \ x^k \in D_F\},$$

where $F'(x)$ denotes the Jacobian of F at $x \in D_F$. Then Clarke's [5] generalized Jacobian of F at $x \in \Re^m$ is defined as the convex hull of $\partial_B F(x)$, i.e., $\partial F(x) = \text{conv}\{\partial_B F(x)\}$.

Let $Y \in \Re^{p \times q}$ admit the following singular value decomposition (SVD):

$$Y = U[\Sigma \ 0]V^T, \tag{5}$$

where $U \in \Re^{p \times p}$ and $V \in \Re^{q \times q}$ are orthogonal matrices, $\Sigma = \text{Diag}(\sigma_1, \cdots, \sigma_p)$ is the diagonal matrix of singular values of Y, with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$. Define $g_\rho : \Re \to \Re$ by

$$g_\rho(t) := (t - \rho)_+ - (-t - \rho)_+. \tag{6}$$

For each parameter $\rho > 0$, the soft-thresholding operator $D_\rho : \Re^{p \times q} \to \Re^{p \times q}$ is defined by

$$D_\rho(Y) = U[\Sigma_\rho \ 0]V^T, \tag{7}$$

where $\Sigma_\rho = \text{Diag}(g_\rho(\sigma_1), \ldots, g_\rho(\sigma_p))$. From [19, Thorem 2.1], we know that $D_\rho(\cdot)$ is strongly semismooth everywhere in $\Re^{p \times q}$. Decompose $V \in \Re^{q \times q}$ into the form $V = [V_1 \ V_2]$, where $V_1 \in R^{q \times p}$ and $V_2 \in \Re^{q \times (q-p)}$. Let the orthogonal matrix $Q \in \Re^{(p+q) \times (p+q)}$ be defined by

$$Q := \frac{1}{\sqrt{2}} \begin{bmatrix} U & U & 0 \\ V_1 & -V_1 & \sqrt{2}V_2 \end{bmatrix}, \tag{8}$$

and $\Xi : \Re^{p \times q} \to \mathscr{S}^{p+q}$ be defined by

$$\Xi(Y) := \begin{bmatrix} 0 & Y \\ Y^T & 0 \end{bmatrix}, \ Y \in \Re^{p \times q}. \tag{9}$$

Then, by Golub and van Loan [15, Sect. 8.6], we know that the symmetric matrix $\Xi(Y)$ has the following spectral decomposition:

$$\Xi(Y) \; = \; Q \begin{bmatrix} \Sigma & 0 & 0 \\ 0 & -\Sigma & 0 \\ 0 & 0 & 0 \end{bmatrix} Q^T, \tag{10}$$

i.e., the eigenvalues of $\Xi(Y)$ are $\pm \sigma_i, i = 1, \ldots, p$, and 0 of multiplicity $q - p$. For any $W = P\mathrm{Diag}(\lambda_1, \cdots, \lambda_{p+q})P^T \in \mathscr{S}^{p+q}$, define $G_\rho : \mathscr{S}^{p+q} \to \mathscr{S}^{p+q}$ by

$$G_\rho(W) := P\mathrm{Diag}(g_\rho(\lambda_1), \cdots, g_\rho(\lambda_{p+q}))P^T = (W - \rho I)_+ - (-W - \rho I)_+,$$

where $(\cdot)_+$ denotes the projection onto the cone of positive semidefinite matrices. By direct calculations, we have

$$\Psi(Y) \; := \; G_\rho(\Xi(Y)) \; = \; Q \begin{bmatrix} \Sigma_\rho & 0 & 0 \\ 0 & -\Sigma_\rho & 0 \\ 0 & 0 & 0 \end{bmatrix} Q^T = \begin{bmatrix} 0 & D_\rho(Y) \\ D_\rho(Y)^T & 0 \end{bmatrix}. \tag{11}$$

Note that (11) provides an easy way for us to calculate the derivative (if it exists) of D_ρ at Y, as we shall see in Proposition 1. For later discussion, we define the following three index sets:

$$\alpha := \{1, \ldots, p\}, \quad \gamma := \{p + 1, \ldots, 2p\}, \quad \beta := \{2p + 1, \ldots, p + q\}. \tag{12}$$

For any $\lambda = (\lambda_1, \ldots, \lambda_{p+q})^T \in \Re^{p+q}$ and $\lambda_i \neq \pm\rho, i = 1, \ldots, p + q$, we denote by Ω the $(p + q) \times (p + q)$ first divided difference symmetric matrix of $g_\rho(\cdot)$ at λ [2] whose (i, j)th entry is given by

$$\Omega_{ij} = \begin{cases} \dfrac{g_\rho(\lambda_i) - g_\rho(\lambda_j)}{\lambda_i - \lambda_j} & \text{if } \lambda_i \neq \lambda_j, \\ g_\rho'(\lambda_i) & \text{if } \lambda_i = \lambda_j. \end{cases}$$

Proposition 1. *Let $Y \in \Re^{p\times q}$ admit the SVD in (5). If $\sigma_i \neq \rho, i = 1, \ldots, p$, then $D_\rho(\cdot)$ is differentiable at Y, and for any $H \in \Re^{p\times q}$, we have*

$$D_\rho'(Y)H = U\left[\left(\Omega_{\alpha\alpha} \circ \left(\frac{H_1 + H_1^T}{2}\right) + \Omega_{\alpha\gamma} \circ \left(\frac{H_1 - H_1^T}{2}\right)\right) V_1^T + \left(\Omega_{\alpha\beta} \circ H_2\right) V_2^T\right], \tag{13}$$

where $H_1 = U^T H V_1$ and $H_2 = U^T H V_2$.

Proof. For any $\lambda = (\lambda_1, \ldots, \lambda_{p+q})^T \in \Re^{p+q}$, let $\lambda_i = \sigma_i$ for $i \in \alpha$, $\lambda_i = -\sigma_{i-p}$ for $i \in \gamma$, and $\lambda_i = 0$ for $i \in \beta$. Since $\sigma_i \neq \rho, i = 1, \ldots, p$, from (10) and (11) we can obtain the first divided difference matrix for $g_\rho(\cdot)$ at λ:

$$\Omega = \begin{pmatrix} \Omega_{\alpha\alpha} & \Omega_{\alpha\gamma} & \Omega_{\alpha\beta} \\ \Omega_{\alpha\gamma}^T & \Omega_{\gamma\gamma} & \Omega_{\gamma\beta} \\ \Omega_{\alpha\beta}^T & \Omega_{\gamma\beta}^T & \Omega_{\beta\beta} \end{pmatrix}. \tag{14}$$

Since $g_\rho(\cdot)$ is an odd function, we have the following results:

$$\Omega_{\gamma\gamma} = \Omega_{\alpha\alpha}, \quad \Omega_{\alpha\gamma} = \Omega_{\alpha\gamma}^T, \quad \Omega_{\gamma\beta} = \Omega_{\alpha\beta}, \quad \Omega_{\beta\beta} = 0.$$

Now, by a result of Löwner [21], we have from (11) that for any $H \in \Re^{p \times q}$,

$$\Psi'(Y)H = G'_\rho(\Xi(Y))\Xi(H) = Q[\Omega \circ (Q^T \Xi(H)Q)]Q^T.$$

Since

$$Q^T \Xi(H)Q = \frac{1}{2} \begin{bmatrix} H_1 + H_1^T & H_1^T - H_1 & \sqrt{2}H_2 \\ H_1 - H_1^T & -(H_1 + H_1^T) & \sqrt{2}H_2 \\ \sqrt{2}H_2^T & \sqrt{2}H_2^T & 0 \end{bmatrix}, \tag{15}$$

by simple algebraic calculations, we have that

$$\Psi'(Y)H = Q[\Omega \circ (Q^T \Xi(H)Q)]Q^T = \begin{bmatrix} 0 & M_{12} \\ M_{12}^T & 0 \end{bmatrix}, \tag{16}$$

where $M_{12} = U\left[\left(\Omega_{\alpha\alpha} \circ \left(\frac{H_1 + H_1^T}{2}\right) + \Omega_{\alpha\gamma} \circ \left(\frac{H_1 - H_1^T}{2}\right)\right) V_1^T + \left(\Omega_{\alpha\beta} \circ H_2\right) V_2^T\right]$. Since

$$\Psi'(Y)H = \begin{bmatrix} 0 & D'_\rho(Y)H \\ (D'_\rho(Y)H)^T & 0 \end{bmatrix},$$

we have from (16) that $D'_\rho(Y)H = M_{12}$

Next, we give a characterization of the generalized Jacobian of $D_\rho(\cdot)$, which was presented in [31, Lemma 2.3.6 and Proposition 2.3.7]. For any $\lambda = (\lambda_1, \ldots, \lambda_{p+q})^T \in \Re^{p+q}$, let $\lambda_i = \sigma_i$ for $i \in \alpha$, $\lambda_i = -\sigma_{i-p}$ for $i \in \gamma$, and $\lambda_i = 0$ for $i \in \beta$. For each threshold parameter $\rho > 0$, we decompose the index set α into the following three subindex sets:

$$\alpha_1 := \{i \mid \sigma_i(Y) > \rho\}, \ \alpha_2 := \{i \mid \sigma_i(Y) = \rho\}, \ \alpha_3 := \{i \mid \sigma_i(Y) < \rho\}. \tag{17}$$

Let Γ denote the following $(p+q) \times (p+q)$ symmetric matrix

$$\Gamma = \begin{pmatrix} \Gamma_{\alpha\alpha} & \Gamma_{\alpha\gamma} & \Gamma_{\alpha\beta} \\ \Gamma_{\alpha\gamma}^T & \Gamma_{\gamma\gamma} & \Gamma_{\gamma\beta} \\ \Gamma_{\alpha\beta}^T & \Gamma_{\gamma\beta}^T & \Gamma_{\beta\beta} \end{pmatrix}, \tag{18}$$

whose (i, j)th entry is given by

$$
\Gamma_{ij} = \begin{cases}
\dfrac{g_\rho(\lambda_i) - g_\rho(\lambda_j)}{\lambda_i - \lambda_j} & \text{if } \lambda_i \neq \lambda_j, \\
1 & \text{if } \lambda_i = \lambda_j \text{ and } |\lambda_i| > \rho, \\
\in \partial g_\rho(\lambda_i) = [0,1] & \text{if } \lambda_i = \lambda_j \text{ and } |\lambda_i| = \rho, \\
0 & \text{if } \lambda_i = \lambda_j \text{ and } |\lambda_i| < \rho.
\end{cases}
\tag{19}
$$

Proposition 2. *Let $Y \in \Re^{p \times q}$ admit the SVD in (5). Then, for any $\mathscr{V} \in \partial_B \Psi(Y)$ and any $H \in \Re^{p \times q}$, we have*

$$
\mathscr{V}(H) = Q(\Gamma \circ (Q^T \Xi(H) Q)) Q^T.
\tag{20}
$$

Moreover, for any $\mathscr{W} \in \partial_B D_\rho(Y)$, we have

$$
\mathscr{W}(H) = U\left[\left(\Gamma_{\alpha\alpha} \circ \left(\frac{H_1 + H_1^T}{2}\right) + \Gamma_{\alpha\gamma} \circ \left(\frac{H_1 - H_1^T}{2}\right) \right) V_1^T + \left(\Gamma_{\alpha\beta} \circ H_2\right) V_2^T \right],
\tag{21}
$$

where $H_1 = U^T H V_1$, $H_2 = U^T H V_2$, and

$$
\Gamma_{\alpha\alpha} = \begin{pmatrix} \tau_{\alpha_1\alpha_1} & \tau_{\alpha_1\alpha_2} & \tau_{\alpha_1\alpha_3} \\ \tau_{\alpha_1\alpha_2}^T & \nu_{\alpha_2\alpha_2} & 0 \\ \tau_{\alpha_1\alpha_3}^T & 0 & 0 \end{pmatrix}, \quad
\begin{aligned}
& \tau_{ij} = 1, \text{ for } i \in \alpha_1, \ j \in \alpha_1 \cup \alpha_2, \\
& \tau_{ij} = \frac{\sigma_i - \rho}{\sigma_i - \sigma_j}, \text{ for } i \in \alpha_1, \ j \in \alpha_3, \\
& \nu_{ij} = \nu_{ji} \in [0,1], \text{ for } i,j \in \alpha_2,
\end{aligned}
\tag{22}
$$

$$
\Gamma_{\alpha\gamma} = \begin{pmatrix} \omega_{\alpha_1\alpha_1} & \omega_{\alpha_1\alpha_2} & \omega_{\alpha_1\alpha_3} \\ \omega_{\alpha_1\alpha_2}^T & 0 & 0 \\ \omega_{\alpha_1\alpha_3}^T & 0 & 0 \end{pmatrix}, \quad \omega_{ij} := \frac{(\sigma_i - \rho)_+ + (\sigma_j - \rho)_+}{\sigma_i + \sigma_j}, \text{ for } i \in \alpha_1, \ j \in \alpha,
\tag{23}
$$

$$
\Gamma_{\alpha\beta} = \begin{pmatrix} \mu_{\alpha_1\bar{\beta}} \\ 0 \end{pmatrix}, \quad \bar{\beta} = \beta - 2p = \{1, \ldots, q - p\}, \quad \mu_{ij} = \frac{\sigma_i - \rho}{\sigma_i}, \text{ for } i \in \alpha_1, \ j \in \bar{\beta}.
\tag{24}
$$

Proof. See [31, Lemma 2.3.6 and Proposition 2.3.7].

Let the operator $\mathscr{W}^0 : \Re^{p \times q} \to \Re^{p \times q}$ be defined by

$$
\mathscr{W}^0(H) = U\left[\left(\Gamma_{\alpha\alpha}^0 \circ \left(\frac{H_1 + H_1^T}{2}\right) + \Gamma_{\alpha\gamma} \circ \left(\frac{H_1 - H_1^T}{2}\right) \right) V_1^T + \left(\Gamma_{\alpha\beta} \circ H_2\right) V_2^T \right],
\tag{25}
$$

where $\Gamma_{\alpha\alpha}^0$ is of the form (22) with $(\Gamma_{\alpha\alpha}^0)_{\alpha_2\alpha_2} = 0$. Then we have that $\mathscr{W}^0$ is an element in $\partial_B D_\rho(Y)$.

3 A Partial Proximal Point Algorithm for Matrix Least Squares Problems

In this section, we will show how to use the partial proximal point algorithm (PPA) to solve the problem (1).

It is easy to see that (1) can be rewritten as follows:

$$\min_{u\in\Re^m, X\in\Re^{p\times q}} \left\{ f_\rho(u, X) := \frac{1}{2}\|u\|^2 + \rho\|X\|_* + \langle C, X\rangle \ : \ \mathscr{A}(X) + u = b, \mathscr{B}(X) = d \right\}.$$

(26)

Note that the objective function $f_\rho(u, X)$ is strongly convex in u for all $X \in \Re^{p\times q}$. Let $l(u, X; \zeta, \xi) : \Re^m \times \Re^{p\times q} \times \Re^m \times \Re^s \to \Re$ be the Lagrangian function for (26):

$$l(u, X; \zeta, \xi) := f_\rho(u, X) + \langle \zeta, b - \mathscr{A}(X) - u\rangle + \langle \xi, d - \mathscr{B}(X)\rangle. \quad (27)$$

Then the essential objective function in (26) is

$$f(u, X) := \sup_{\zeta\in\Re^m, \xi\in\Re^s} l(u, X; \zeta, \xi) = \begin{cases} f_\rho(u, X) \text{ if } (u, X) \in \mathscr{F}_P, \\ +\infty \quad \text{ if } (u, X) \notin \mathscr{F}_P, \end{cases} \quad (28)$$

where $\mathscr{F}_P = \{(u, X) \in \Re^m \times \Re^{p\times q} \mid \mathscr{A}(X) + u = b, \mathscr{B}(X) = d\}$ is the feasible set of (26). The dual problem of (26) is given by:

$$\max \left\{ g_\rho(\zeta, \xi) \ : \ \mathscr{A}^*(\zeta) + \mathscr{B}^*(\xi) + Z = C, \|Z\|_2 \le \rho, \zeta \in \Re^m, \xi \in \Re^s, Z \in \Re^{p\times q} \right\},$$

(29)

where $g_\rho(\zeta, \xi) := -\frac{1}{2}\|\zeta\|^2 + \langle b, \zeta\rangle + \langle d, \xi\rangle$. Since $f(u, X)$ is strongly convex in u for all $X \in \Re^{p\times q}$, we apply the partial PPA proposed by Ha [17] to the maximal monotone operator $\mathscr{T}_f = \partial f$, in which only the variable X appears in the quadratic proximal term. Let $\Pi : \Re^m \times \Re^{p\times q} \to \Re^m \times \Re^{p\times q}$ be the orthogonal projector of $\Re^m \times \Re^{p\times q}$ onto $\{0\} \times \Re^{p\times q}$, i.e., $\Pi(u, X) = (0, X)$ and $P_\sigma := (\Pi + \sigma\mathscr{T}_f)^{-1}\Pi$ for a given positive parameter σ. From [19, Proposition 3.1], we know that the operator P_σ is single-valued. Given a starting point $(u^0, X^0) \in \Re^m \times \Re^{p\times q}$, the partial PPA for solving problem (26) can be expressed as follows:

$$(u^{k+1}, X^{k+1}) \approx P_{\sigma_k}(u^k, X^k) := \underset{u\in\Re^m, X\in\Re^{p\times q}}{\mathrm{argmin}} \left\{ f(u, X) + \frac{1}{2\sigma_k}\|X - X^k\|^2 \right\}, \quad (30)$$

where the sequence $\{\sigma_k\}$ satisfies (2). Note that for the standard PPA, the map Π in P_σ is replaced by the identity map.

Next we compute the partial quadratic regularization of f in (30), which plays a key role in the study of the partial PPA for solving (26). For a given parameter $\sigma > 0$, the partial quadratic regularization of f in (28) associated with σ is given by

$$F_\sigma(X) = \min_{u \in \Re^m, Y \in \Re^{p \times q}} \left\{ f(u, Y) + \frac{1}{2\sigma} \|Y - X\|^2 \right\}. \tag{31}$$

From [19, Sect. 3], we have that

$$F_\sigma(X) = \sup_{\zeta \in \Re^m, \, \xi \in \Re^s} \theta_\sigma(\zeta, \xi; X),$$

where

$$\theta_\sigma(\zeta, \xi; X) := -\frac{1}{2}\|\zeta\|^2 + \langle b, \zeta \rangle + \langle d, \xi \rangle + \frac{1}{2\sigma}\|X\|^2 - \frac{1}{2\sigma}\|D_{\rho\sigma}(W(\zeta, \xi; X))\|^2 \tag{32}$$

and $W(\zeta, \xi; X) = X - \sigma(C - \mathscr{A}^*\zeta - \mathscr{B}^*\xi)$. By the saddle point theorem [26, Theorem 28.3], we have that for any

$$(\zeta(X), \xi(X)) \in \underset{\zeta \in \Re^m, \, \xi \in \Re^s}{\text{argsup}} \; \theta_\sigma(\zeta, \xi; X),$$

the point $(\zeta(X), D_{\rho\sigma}(W(\zeta(X), \xi(X); X)))$ is the unique solution to (31).

Now we formally present the partial PPA for solving (26).

Algorithm 1

Given a tolerance $\varepsilon > 0$, $(u^0, X^0) \in \Re^m \times \Re^{p \times q}$, $\sigma_0 > 0$. Set $k = 0$. Iterate:

Step 1.   Compute an approximate maximizer

$$(\zeta^{k+1}, \xi^{k+1}) \approx \underset{\zeta \in \Re^m, \xi \in \Re^s}{\text{argsup}} \; \theta_{\sigma_k}(\zeta, \xi; X^k), \tag{33}$$

   where $\theta_{\sigma_k}(\zeta, \xi; X^k)$ is defined in (32).

Step 2.   Compute $W^{k+1} := W(\zeta^{k+1}, \xi^{k+1}; X^k)$. Set

$$u^{k+1} = \zeta^{k+1}, \; X^{k+1} = D_{\rho\sigma_k}(W^{k+1}), \quad Z^{k+1} = \frac{1}{\sigma_k}(D_{\rho\sigma_k}(W^{k+1}) - W^{k+1}).$$

Step 3.   If $\|(X^k - X^{k+1})/\sigma_k\| \le \varepsilon$; stop; else; update σ_k; end.

Suppose that $(\bar{\zeta}(X^k), \bar{\xi}(X^k))$ is an optimal solution of the inner subproblem (33) for each X^k. In order to terminate (33) in the above partial PPA, we introduce the following stopping criteria:

$$\sup \theta_k(\zeta, \xi) - \theta_k(\zeta^{k+1}, \xi^{k+1}) \leq \frac{\varepsilon_k^2}{4\sigma_k}, \tag{34a}$$

$$\|\zeta^{k+1} - \bar{\zeta}(X^k)\|^2 \leq \frac{1}{2}\varepsilon_k^2, \quad \varepsilon_k > 0, \sum_{k=0}^{\infty} \varepsilon_k < \infty, \tag{34b}$$

$$\sup \theta_k(\zeta, \xi) - \theta_k(\zeta^{k+1}, \xi^{k+1}) \leq \frac{\delta_k^2}{2\sigma_k}\|X^{k+1} - X^k\|^2, \tag{34c}$$

$$\|\zeta^{k+1} - \bar{\zeta}(X^k)\|^2 \leq \delta_k^2\|\zeta^{k+1} - \zeta^k\|^2, \quad \delta_k > 0, \sum_{k=0}^{\infty} \delta_k < \infty, \tag{34d}$$

$$\|\nabla\theta_k(\zeta^{k+1}, \xi^{k+1})\| \leq \frac{\delta_k'}{\sigma_k}\|X^{k+1} - X^k\|, \quad 0 \leq \delta_k' \to 0. \tag{34e}$$

In [19], it has been shown that under mild assumptions, the sequence $\{(u^k, X^k)\}$ generated by the partial PPA under criterion (34a) and (34b) converges to an optimal solution $(\bar{u}, \overline{X})$ of (26), and $\{(\zeta^k, \xi^k)\}$ is asymptotically maximizing for problem (29). If, in addition, (34c) and (34d) are also satisfied and $\mathscr{T}_f^{-1}$ is Lipschitz continuous at the origin, then $\{(u^k, X^k)\}$ locally converges to the unique optimal solution $(\bar{u}, \overline{X})$ of (26) at least at a linear rate which tends to zero as $\sigma_k \to +\infty$. For details on the convergence analysis, we refer the reader to [19, Theorems 3.1 and 3.2].

3.1 A Semismooth Newton-CG Method for Solving Unconstrained Inner Subproblems

In this subsection, we introduce a semismooth Newton-CG (SSNCG) method for solving the unconstrained inner subproblem (33), which is the most expensive step in each PPA iteration. For later convenience, we let

$$\hat{\mathscr{A}} = \begin{pmatrix} \mathscr{A} \\ \mathscr{B} \end{pmatrix}, \ \hat{b} = (b; d) \in \mathfrak{R}^{m+s}, \ T = \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix} \in \mathfrak{R}^{(m+s)\times(m+s)}, \ y = (\zeta; \xi) \in \mathfrak{R}^{m+s}. \tag{35}$$

For the convergence analysis, we assume that the following Slater condition holds:

$$\begin{cases} \mathscr{B} : \mathfrak{R}^{p\times q} \to \mathfrak{R}^s \text{ is onto,} \\ \exists X_0 \in \mathfrak{R}^{p\times q} \text{ such that } \mathscr{B}(X_0) = d. \end{cases} \tag{36}$$

In our proposed partial PPA, for some fixed $X \in \Re^{p \times q}$ and $\sigma > 0$, we need to solve an inner subproblem of the following form:

$$\min_{y \in \Re^{m+s}} \left\{ \varphi(y) := \frac{1}{2} \langle y, Ty \rangle + \frac{1}{2\sigma} \| D_{\rho\sigma}(W(y; X)) \|^2 - \langle \hat{b}, y \rangle \right\}, \qquad (37)$$

where $W(y; X) = X - \sigma(C - \hat{\mathscr{A}}^* y)$ and $\hat{\mathscr{A}}^* = (\mathscr{A}^*, \mathscr{B}^*)$ is the adjoint of $\hat{\mathscr{A}}$. The optimality condition for (37) is given by

$$\nabla \varphi(y) = Ty + \hat{\mathscr{A}} D_{\rho\sigma}(W(y; X)) - \hat{b} = 0. \qquad (38)$$

Since the soft-thresholding operator $D_{\rho\sigma}(\cdot)$ is Lipschitz continuous with modulus 1 [19, 20], the mapping $\nabla \varphi(y)$ is Lipschitz continuous on $\Re^{m+s}$. Thus for any $y \in \Re^{m+s}$, the generalized Hessian of $\varphi(y)$ is well defined and it is given by

$$\partial^2 \varphi(y) := \partial(\nabla \varphi)(y), \qquad (39)$$

where $\partial(\nabla \varphi)(y)$ is the Clarke's generalized Jacobian of $\nabla \varphi$ at y [5]. However, it is hard to express $\partial^2 \varphi(y)$ exactly, so we define the following alternative for $\partial^2 \varphi(y)$,

$$\hat{\partial}^2 \varphi(y) := T + \sigma \hat{\mathscr{A}} \, \partial D_{\rho\sigma}(W(y; X)) \hat{\mathscr{A}}^*. \qquad (40)$$

From [5, p. 75], we have for any $h \in \Re^{m+s}$,

$$\partial^2 \varphi(y) h \subseteq \hat{\partial}^2 \varphi(y) h, \qquad (41)$$

which implies that if all elements in $\hat{\partial}^2 \varphi(y)$ are positive definite, so are those in $\partial^2 \varphi(y)$.

Since the soft-thresholding operator $D_{\rho\sigma}(\cdot)$ is strongly semismooth, $\nabla \varphi(\cdot)$ is also strongly semismooth. We can solve the nonlinear equation (38) by using a SSNCG method for which the direction r at an iterate y is computed from the following linear system of equations:

$$\underbrace{(T + \sigma \hat{\mathscr{A}} \mathscr{W} \hat{\mathscr{A}}^*)}_{\mathscr{V}} r = -\nabla \varphi(y), \qquad (42)$$

where $\mathscr{W}$ is any element in $\partial D_{\rho\sigma}(W(y; X))$. Note that if $s = 0$, i.e., the constraint $\mathscr{B}X = d$ is absent, then $\mathscr{V}$ is always positive definite due to the fact that all the elements in $\partial D_{\rho\sigma}(\cdot)$ are positive semidefinite [19, Proposition 2.1] and $T = I_m$.

Define the operator $\mathscr{W}_y^0 : \Re^{p \times q} \to \Re^{p \times q}$ as in (25). To implement the above SSNCG method, we need to choose an explicit element $\mathscr{W}$ in $\partial D_{\rho\sigma}(W(y; X))$, which we take to be $\mathscr{W}_y^0$. With this specific choice, then the coefficient matrix in (42) is given by

$$\mathscr{V}_y^0 = T + \sigma \hat{\mathscr{A}} \mathscr{W}_y^0 \hat{\mathscr{A}}^* \in \hat{\partial}^2 \varphi(y). \qquad (43)$$

Next, we shall study a certain constraint nondegeneracy condition and its connection to the positive definiteness of $\mathscr{V}_y \in \hat{\partial}^2 \varphi(y)$. Suppose that the Slater condition (36) holds and $\overline{y} = (\overline{\zeta}; \overline{\xi}) \in \Re^{m+s}$ is the optimal solution to problem (37). Let $W(\overline{y}; X) = X - \sigma(C - \mathscr{A}^* \overline{y})$ and $\overline{X} = D_{\rho\sigma}(W(\overline{y}; X))$. Let $W(\overline{y}; X)$ admit the SVD as in (5). For the given threshold value $\rho\sigma$, we decompose the index set $\alpha = \{1, \dots, p\}$ into the following three subindex sets:

$$\alpha_1 := \{i \mid \sigma_i(W) > \rho\sigma\}, \ \alpha_2 := \{i \mid \sigma_i(W) = \rho\sigma\}, \ \alpha_3 := \{i \mid \sigma_i(W) < \rho\sigma\}.$$

The constraint nondegeneracy condition is said to hold at $\overline{X}$ [19] if

$$\mathscr{B}(\mathscr{T}(\overline{X})) = \Re^s, \tag{44}$$

where the subspace $\mathscr{T}(\overline{X})$ of $\Re^{p \times q}$ is defined as

$$\mathscr{T}(\overline{X}) := \left\{ H \in \Re^{p \times q} \mid [U_{\alpha_2} \ U_{\alpha_3}]^T H [V_{\alpha_2} \ V_{\alpha_3} \ V_2] = 0 \right\}, \tag{45}$$

and its orthogonal complement is given by

$$\mathscr{T}^\perp(\overline{X}) = \left\{ H \in \Re^{p \times q} \mid U_{\alpha_1}^T H = 0, \ \ HV_{\alpha_1} = 0 \right\}. \tag{46}$$

The following lemma will be needed to analyze the connection between the constraint nondegeneracy condition at $\overline{X}$ and the positive definiteness of the elements of $\hat{\partial}^2 \varphi(\overline{y})$.

Lemma 1. *Let $W(\overline{y}; X)$ admit the SVD as in (5). For any $\mathscr{W} \in \partial D_{\rho\sigma}(W(\overline{y}; X))$ and $H \in \Re^{p \times q}$ such that $\mathscr{W} H = 0$, it holds that*

$$H \in \mathscr{T}^\perp(\overline{X}). \tag{47}$$

Proof. Let $\mathscr{W} \in \partial D_{\rho\sigma}(W(\overline{y}; X))$ and $H \in \Re^{p \times q}$ be such that $\mathscr{W} H = 0$. Then we have

$$0 = \langle H, \mathscr{W} H \rangle = \frac{1}{2} \langle \Xi(H), \Xi(\mathscr{W} H) \rangle = \frac{1}{2} \langle \Xi(H), Q(\Gamma \circ (Q^T \Xi(H) Q)) Q^T \rangle$$

$$= \frac{1}{2} \langle Q^T \Xi(H) Q, \Gamma \circ (Q^T \Xi(H) Q) \rangle = \frac{1}{2} \langle \tilde{H}, \Gamma \circ \tilde{H} \rangle,$$

where $\Gamma \in \mathscr{S}^{p+q}$ is defined as in (18) and $\tilde{H} = Q^T \Xi(H) Q$. Let $H_1 = U^T H V_1, H_2 = U^T H V_2, H_1^s = (H_1 + H_1^T)/2$ and $H_1^a = (H_1 - H_1^T)/2$. From (15) and (21), we have

$$0 = \frac{1}{2} \langle \tilde{H}, \Gamma \circ \tilde{H} \rangle = \sum_{i \in \alpha} \sum_{j \in \alpha} \Gamma_{ij} (H_1^s)_{ij}^2 + \sum_{i \in \alpha} \sum_{j \in \gamma} \Gamma_{ij} (H_1^a)_{ij}^2 + \sum_{i \in \alpha} \sum_{j \in \beta} \Gamma_{ij} (H_2)_{ij}^2.$$

Since $\Gamma_{ij} \in [0, 1]$ for all $i, j = 1, \ldots, p + q$, it follows that

$$\sum_{i \in \alpha} \sum_{j \in \alpha} \Gamma_{ij} (H_1^s)_{ij}^2 = 0, \quad \sum_{i \in \alpha} \sum_{j \in \gamma} \Gamma_{ij} (H_1^a)_{ij}^2 = 0, \quad \sum_{i \in \alpha} \sum_{j \in \beta} \Gamma_{ij} (H_2)_{ij}^2 = 0.$$

Then from (22), (23) and (24), we have that

$$(H_1^s)_{\alpha_1 \alpha} = 0, \ (H_1^s)_{\alpha \alpha_1} = 0, \ (H_1^a)_{\alpha_1 \alpha} = 0, \ (H_1^a)_{\alpha \alpha_1} = 0, \ (H_2)_{\alpha_1 \bar{\beta}} = 0,$$

where $\bar{\beta} = \{1, \ldots, q - p\}$. Since $H_1 = H_1^s + H_1^a$, we have that $(H_1)_{\alpha_1 \alpha} = 0$ and $(H_1)_{\alpha \alpha_1} = 0$. From $H_1 = [U_{\alpha_1} \ U_{\alpha_2} \ U_{\alpha_3}]^T H [V_{\alpha_1} \ V_{\alpha_2} \ V_{\alpha_3}]$ and $H_2 = [U_{\alpha_1} \ U_{\alpha_2} \ U_{\alpha_3}]^T H V_2$, we obtain that

$$U_{\alpha_1}^T H V_1 = 0, \quad U_{\alpha_1}^T H V_2 = 0, \quad U^T H V_{\alpha_1} = 0.$$

Since both U and $V = [V_1 \ V_2]$ are orthogonal matrices, we have $U_{\alpha_1}^T H = 0, H V_{\alpha_1} = 0$, which means that $H \in \mathscr{T}^\perp(\overline{X})$.

Proposition 3. *Suppose that the Slater condition (36) is satisfied. Let $\overline{y}$ be the optimal solution to problem (37), $W(\overline{y}; X) = X - \sigma(C - \hat{\mathscr{A}}^* \overline{y})$ admit the SVD as in (5), and $\overline{X} = D_{\rho\sigma}(W(\overline{y}; X))$. Then the following conditions are equivalent:*

(a) The constraint nondegeneracy condition (44) holds at $\overline{X}$.
(b) Every $\mathscr{V}_{\overline{y}} \in \hat{\partial}^2 \varphi(\overline{y})$ is symmetric and positive definite.
(c) $\mathscr{V}_{\overline{y}}^0 \in \hat{\partial}^2 \varphi(\overline{y})$ is symmetric and positive definite.

Proof. "(a) $\Rightarrow$ (b)". Let $\mathscr{V}_{\overline{y}}$ be an arbitrary element in $\hat{\partial}^2 \varphi(\overline{y})$. Then there exists an element $\mathscr{W}_{\overline{y}} \in \partial D_{\rho\sigma}(W(\overline{y}; X))$ such that

$$\mathscr{V}_{\overline{y}} = T + \sigma \hat{\mathscr{A}} \mathscr{W}_{\overline{y}} \hat{\mathscr{A}}^* = T + \sigma \begin{bmatrix} \mathscr{A} \mathscr{W}_{\overline{y}} \mathscr{A}^* & \mathscr{A} \mathscr{W}_{\overline{y}} \mathscr{B}^* \\ \mathscr{B} \mathscr{W}_{\overline{y}} \mathscr{A}^* & \mathscr{B} \mathscr{W}_{\overline{y}} \mathscr{B}^* \end{bmatrix}. \tag{48}$$

Since $\mathscr{W}_{\overline{y}}$ is self-adjoint and positive semidefinite [19, Proposition 2.1], we have that $\mathscr{V}_{\overline{y}}$ is self-adjoint and positive semidefinite. From (48) we obtain that $\mathscr{V}_{\overline{y}}$ is positive definite if only if $\mathscr{B} \mathscr{W}_{\overline{y}} \mathscr{B}^*$ is positive definite. Hence, it enough to show the positive definiteness of $\mathscr{B} \mathscr{W}_{\overline{y}} \mathscr{B}^*$. Let $h \in \Re^s$ be such that $\mathscr{B} \mathscr{W}_{\overline{y}} \mathscr{B}^* h = 0$. Then we have

$$0 = \langle h, \mathscr{B} \mathscr{W}_{\overline{y}} \mathscr{B}^* h \rangle = \langle \mathscr{B}^* h, \mathscr{W}_{\overline{y}} \mathscr{B}^* h \rangle \geq \langle \mathscr{W}_{\overline{y}} \mathscr{B}^* h, \mathscr{W}_{\overline{y}} \mathscr{B}^* h \rangle,$$

where the last inequality follows from [19, Proposition 2.1], which implies that $\mathscr{W}_{\overline{y}}(\mathscr{B}^* h) = 0$. From Lemma 1, we have $\mathscr{B}^* h \in \mathscr{T}(\overline{X})^\perp$. Since the constraint nondegeneracy condition holds at $\overline{X}$, there exists a $Y \in \mathscr{T}(\overline{X})$ such that $\mathscr{B} Y = h$. Then, we have

$$\langle h, h \rangle = \langle h, \mathscr{B} Y \rangle = \langle \mathscr{B}^* h, Y \rangle = 0.$$

Thus $h = 0$, which implies that $\mathscr{B} \mathscr{W}_{\overline{y}} \mathscr{B}^*$ is positive definite. Hence, $\mathscr{V}_{\overline{y}}$ is positive definite.

"(b) $\Rightarrow$ (c)". This is obviously true since $\mathscr{V}_{\overline{y}}^0 \in \hat{\partial}^2 \varphi(\overline{y})$.

"(c) $\Rightarrow$ (a)". Suppose that the constraint nondegeneracy condition (44) does not hold at $\overline{X}$. Then there exists a non-zero $h \in [\mathscr{B} \mathscr{T}(\overline{X})]^{\perp}$. And we have

$$0 = \langle h, \mathscr{B} Y \rangle = \langle H, Y \rangle \quad \forall\, Y \in \mathscr{T}(\overline{X}),$$

where $H = \mathscr{B}^* h$, which implies that $H \in \mathscr{T}(\overline{X})^{\perp}$. From (46), we have $U_{\alpha_1}^T H = 0$ and $H V_{\alpha_1} = 0$. Then it follows that

$$U_{\alpha_1}^T H V = U_{\alpha_1}^T H [V_1\ V_2] = 0 \quad \text{and} \quad U^T H V_{\alpha_1} = 0. \tag{49}$$

Let $H_1 = U^T H V_1$ and $H_2 = U^T H V_2$. We have from (49) that

$$(H_1)_{\alpha_1 \alpha} = 0, \quad (H_1)_{\alpha \alpha_1} = 0, \quad \text{and} \ (H_2)_{\alpha_1 \bar{\beta}} = 0,$$

where $\bar{\beta} = \{1, \ldots, q - p\}$, from which we can further have that

$$(H_1^s)_{\alpha_1 \alpha} = 0, \quad (H_1^s)_{\alpha \alpha_1} = 0, \quad (H_1^a)_{\alpha_1 \alpha} = 0, \quad \text{and} \ (H_1^a)_{\alpha \alpha_1} = 0,$$

where $H_1^s = (H_1 + H_1^T)/2$ and $H_1^a = (H_1 - H_1^T)/2$. Then we have

$$\Gamma_{\alpha \alpha}^0 \circ (H_1^s) = 0, \quad \Gamma_{\alpha \gamma} \circ (H_1^a) = 0, \quad \text{and} \ \Gamma_{\alpha \beta} \circ H_2 = 0.$$

From the definition of $\mathscr{W}_{\overline{y}}^0$ in (25), it follows that $\mathscr{W}_{\overline{y}}^0(H) = 0$, and hence

$$\langle h, \mathscr{B} \mathscr{W}_{\overline{y}}^0 \mathscr{B}^* h \rangle = \langle H, \mathscr{W}_{\overline{y}}^0(H) \rangle = 0. \tag{50}$$

Since $\mathscr{V}_{\overline{y}}^0$ is positive definite, it follows from (48) that $\mathscr{B} \mathscr{W}_{\overline{y}}^0 \mathscr{B}^*$ is also positive definite. Then (50) implies that $h = 0$, which contradicts the assumption that $h \neq 0$. Hence, we have that (a) holds. $\qquad \square$

Now we present the steps of the SSNCG algorithm for solving (37).

In Algorithm 2, since $\mathscr{V}_t$ is always positive semidefinite, the matrix $\mathscr{V}_t + \varepsilon_t I$ is positive definite as long as $\nabla \varphi(y^t) \neq 0$. From [32, Lemma 3.1], we know that the generated search direction r^t is always a descent direction.

To analyze the global convergence of Algorithm 2, we assume that $\nabla \varphi(y^t) \neq 0$ for any $t \geq 0$. The global convergence and the rate of local convergence of Algorithm 2 can be derived similarly as in [32].

Theorem 1. *Suppose that the Slater condition (36) holds. Then Algorithm 2 is well defined and any accumulation point $\overline{y}$ of $\{y^t\}$ generated by Algorithm 2 is an optimal solution to the inner subproblem (37).*

Proof. See [32, Theorem 3.4]. $\qquad \square$

Algorithm 2 A semismooth Newton-CG method

Given $y^0 \in \Re^{m+s}$, $\eta \in (0,1)$, $\tau \in (0,1]$, $\tau_1, \tau_2 \in (0,1)$, and $c \in (0,1/2)$, $\delta \in (0,1)$. Set $t = 0$. Iterate:

Step 1.   Compute $\eta_t := \min\{\eta, \|\nabla\varphi(y^t)\|^{1+\tau}\}$. Apply the CG method to find an approximation solution r^t to

$$(\mathscr{V}_t + \varepsilon_t I)\, r = -\nabla\varphi(y^t), \qquad (51)$$

where $\mathscr{V}_t \in \hat{\partial}^2\varphi(y^t)$ is defined in (43) and $\varepsilon_t = \min\{\tau_2, \tau_1\|\nabla\varphi(y^t)\|\}$, so that r^t satisfies the following condition:

$$\|(\mathscr{V}_t + \varepsilon_t I)r^t + \nabla\varphi(y^t)\| \le \eta_t. \qquad (52)$$

Step 2.   Set $\alpha_t = \delta^{m_t}$, where m_t is the first nonnegative integer m for which

$$\varphi(y^t + \delta^m r^t) \;\le\; \varphi(y^t) + c\delta^m \langle r^t, \nabla\varphi(y^t)\rangle.$$

Step 3.   Set $y^{t+1} = y^t + \alpha_t r^t$.

Theorem 2. *Suppose that the Slater condition (36) holds. Let $\overline{y}$ be an accumulation point of the infinite sequence $\{y^t\}$ generated by Algorithm 2 for solving the inner subproblem (37). Suppose also that at each step $t \ge 0$, the inexact direction r^t satisfies the accuracy condition in (52). Assume that the constraint nondegeneracy condition (44) holds at $\overline{X} := D_{\rho\sigma}(W(\overline{y}; X))$. Then the whole sequence $\{y^t\}$ converges to $\overline{y}$ and*

$$\|y^{t+1} - \overline{y}\| = O(\|y^t - \overline{y}\|^{1+\tau}). \qquad (53)$$

Proof. See [32, Theorem 3.5].

4   Semidefinite Matrix Least Squares Problems

In this section, we show that the partial PPA developed for solving (26) can easily be adapted for solving the semidefinite matrix least squares problem (4). It is easy to see that (4) can be rewritten as follows:

$$\min_{u\in\Re^m, X\in\mathscr{S}^n} \left\{\frac{1}{2}\|u\|^2 + \langle C, X\rangle \;:\; \mathscr{A}(X) + u = b, \; \mathscr{B}(X) = d, \; X \succeq 0\right\}. \quad (54)$$

The dual problem of (54) is given by:

$$\max_{\zeta\in\Re^m, \xi\in\Re^s, Z\in\mathscr{S}^n} \left\{-\frac{1}{2}\|\zeta\|^2 + \langle b, \zeta\rangle + \langle d, \xi\rangle \;:\; \mathscr{A}^*(\zeta) + \mathscr{B}^*(\xi) + Z = C, \; Z \succeq 0\right\}.$$

$$(55)$$

For some fixed $X \in \mathscr{S}^n$ and $\sigma > 0$, the partial quadratic regularization of problem (54) is given by:

$$\min_{u \in \Re^m, Y \in \mathscr{S}^n} \left\{ \frac{1}{2} \|u\|^2 + \langle C, Y \rangle + \frac{1}{2\sigma} \|Y - X\|^2 \; : \; \mathscr{A}(Y) + u = b, \; \mathscr{B}(Y) = d, \; Y \succeq 0 \right\},$$
(56)

and the Lagrangian dual problem of (56) is given by

$$\max_{\zeta \in \Re^m, \xi \in \Re^s} \theta_\sigma(\zeta, \xi; X) := \inf_{u \in \Re^m, Y \succeq 0} L_\sigma(u, Y; \zeta, \xi, X),$$
(57)

where

$$
\begin{aligned}
&L_\sigma(u, Y; \zeta, \xi, X) \\
&= \frac{1}{2} \|u\|^2 + \langle C, Y \rangle + \frac{1}{2\sigma} \|Y - X\|^2 + \langle \zeta, b - \mathscr{A}(Y) - u \rangle + \langle \xi, d - \mathscr{B}(Y) \rangle \\
&= \frac{1}{2} \|u\|^2 - \langle \zeta, u \rangle + \langle b, \zeta \rangle + \langle d, \xi \rangle + \frac{1}{2\sigma} \|Y - W(\zeta, \xi; X)\|^2 \\
&\quad + \frac{1}{2\sigma} (\|X\|^2 - \|W(\zeta, \xi; X)\|^2),
\end{aligned}
$$

where $W(\zeta, \xi; X) = X - \sigma(C - \mathscr{A}^*\zeta - \mathscr{B}^*\xi)$. By minimizing $L_\sigma(u, Y; \zeta, \xi, X)$ over $Y \succeq 0$, we have

$$\theta_\sigma(\zeta, \xi; X) = -\frac{1}{2} \|\zeta\|^2 + \langle b, \zeta \rangle + \langle d, \xi \rangle + \frac{1}{2\sigma} \|X\|^2 - \frac{1}{2\sigma} \|\Pi_{\mathscr{S}^n_+}(W(\zeta, \xi; X))\|^2,$$
(58)

where $\Pi_{\mathscr{S}^n_+}(\cdot)$ is the metric projector of $\mathscr{S}^n$ onto $\mathscr{S}^n_+$. The problem (57) is an unconstrained continuously differentiable convex optimization problem, and it can be efficiently solved by the SSNCG method developed in [32]. The SSNCG method for solving (57) is analogous to Algorithm 2 where for some fixed $X \in \mathscr{S}^n$ and $\sigma > 0$, the function φ is now given by

$$\varphi(y) = \frac{1}{2} \langle y, Ty \rangle + \frac{1}{2\sigma} \|\Pi_{\mathscr{S}^n_+}(W(y; X))\|^2 - \langle \hat{b}, y \rangle$$

and the operator $\mathscr{V}_t$ in (51) is replaced by

$$\mathscr{V}_t = T + \sigma \hat{\mathscr{A}} \Pi'_{\mathscr{S}^n_+}(W(y^t; X)) \hat{\mathscr{A}}^*,$$
(59)

where $\Pi'_{\mathscr{S}^n_+}(W(y^t; X))$ denotes an element of $\partial \Pi_{\mathscr{S}^n_+}(W(y^t; X))$.

The fast local convergence of the SSNCG method for solving (57) can be established in a similar fashion as Theorem 2 where the positive definiteness of the element $\mathscr{V}_{\bar{y}}$ defined in (59) at the optimal solution $\bar{y}$ is again equivalent to a constraint nondegeneracy condition similar to (44) at $\overline{X} := \Pi_{\mathscr{S}^n_+}(W(\bar{y}; X))$.

5 Numerical Results

In this section, we report some numerical results to demonstrate the efficiency of our SSNCG partial PPA. We implemented our algorithm in MATLAB 2012a (version 7.14), and the numerical experiments are run in MATLAB under a Windows 7 64-bit system on an Intel Xeon 4 Cores 3.20 GHz CPU with 12 GB memory.

In our numerical implementation, we use the alternating direction method of multipliers (ADMM) [8] to generate a reasonably good starting point for our SSNCG partial PPA. The augmented Lagrangian function for (29) corresponding to the linear equality constraints is defined as:

$$L_\sigma(y, Z; X) = -\frac{1}{2}\langle y, Ty \rangle + \langle \hat{b}, y \rangle + \langle X, C - \hat{\mathscr{A}}^* y - Z \rangle - \frac{\sigma}{2}\|C - \hat{\mathscr{A}}^* y - Z\|^2,$$

where $X \in \Re^{p \times q}$ and $\sigma > 0$. Given a starting point $(X^0; y^0; Z^0)$, the ADMM generates new iterates by the following procedure:

$$y^{k+1} := \text{argmax}_{y \in \Re^{m+s}} L_\sigma(y, Z^k; X^k), \tag{60}$$

$$Z^{k+1} := \text{argmax}_{\|Z\|_2 \le \rho} L_\sigma(y^{k+1}, Z; X^k) = \frac{1}{\sigma}(D_{\rho\sigma}(W^{k+1}) - W^{k+1}), \tag{61}$$

$$X^{k+1} := D_{\rho\sigma}(W^{k+1}), \tag{62}$$

where $W^{k+1} = X^k - \sigma(C - \hat{\mathscr{A}}^* y^{k+1})$. Note that the iterate y^{k+1} in (60) can be computed by solving the following linear system of equations:

$$(T + \sigma\hat{\mathscr{A}}\hat{\mathscr{A}}^*)y = \hat{b} - \hat{\mathscr{A}}(X^k) + \sigma\hat{\mathscr{A}}(C - Z^k).$$

We measure the infeasibilities and optimality for the primal problem (26) and the dual problem (29) as follows:

$$R_P = \frac{\|\hat{b} - (\zeta; 0) - \hat{\mathscr{A}}(X)\|}{1 + \|\hat{b}\|},$$

$$R_D = \frac{\|C - \hat{\mathscr{A}}^* y - Z\|}{1 + \|\hat{\mathscr{A}}^*\|},$$

$$\text{relgap} = \frac{f_\rho(\zeta, X) - g_\rho(\zeta, \xi)}{1 + |f_\rho(\zeta, X)| + |g_\rho(\zeta, \xi)|},$$

where $y = (\zeta; \xi)$, $Z = (D_{\rho\sigma}(W) - W)/\sigma$ with $W = X - \sigma(C - \hat{\mathscr{A}}^* y)$, and $f_\rho(\zeta, X)$ and $g_\rho(\zeta, \xi)$ are the objective functions of the primal and dual problems, respectively. The infeasibility of the condition $\|Z\|_2 \le \rho$ is not checked since

it is satisfied up to machine precision throughout the algorithm. In our numerical experiments, we stop the partial PPA when

$$\max\{R_P, R_D\} \leq \texttt{Tol}, \tag{63}$$

where Tol is a pre-specified accuracy tolerance. We choose the initial $\sigma_0 = 1$ and update it as $\sigma_{k+1} = \min(2\sigma_k, 10^8)$ if $R_D^{k+1}/R_D^k > 0.5$; otherwise $\sigma_{k+1} = \sigma_k$. Unless otherwise specified, we set the parameter ρ in (1) to be $\rho = 10^{-3}\|\mathscr{A}^*b\|_2$ and $\texttt{Tol} = 10^{-6}$ as the default.

In solving the subproblem (37), the SSNCG algorithm is stopped when $R_P = \|\nabla\varphi(y)\|/(1 + \|\hat{b}\|) \leq 0.2R_D$ or the number of Newton iterations exceeds 50. We cap the number of CG steps for solving each Newton system of linear equations at 600 and stop the CG solver when $\|(\mathscr{V} + \varepsilon I)r + \nabla\varphi(y)\| \leq \min(0.05, 0.1\|\nabla\varphi(y)\|)$.

The reason we used the ADMM instead of other first order methods to generate a starting point is based on our belief and experience that the ADMM is perhaps the most efficient first order method for solving the problems which we are interested in. Since this belief may be challenged without strong numerical evidence to support it, as suggested by one of the referees, we also tested the primal-dual splitting method by Condat [7] for solving (1). To apply Condat's method in [7], we rewrite (1) in the following form:

$$\min_X \ F(X) + G(X) + H(\mathscr{B}(X)), \tag{64}$$

where $F(X) = \frac{1}{2}\|\mathscr{A}(X) - b\|^2 + \langle C, X\rangle$ with $\nabla F(X) = \mathscr{A}^*(\mathscr{A}(X) - b) + C$, $G(X) = \rho\|X\|_*$, and $H(\cdot)$ is the indicator function over the singleton set $\{d\}$. Note that the optimality condition for (64) is given by:

$$\text{find } (X, y) \text{ such that } \begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \nabla F(X) + \partial G(X) + \mathscr{B}^*y \\ -\mathscr{B}(X) + \partial H^*(y) \end{pmatrix} \tag{65}$$

where H^* is the conjugate function of H with $H^*(y) = \langle d, y\rangle$. Let L_F be a Lipschitz constant for ∇F, which in our case can be set to $L_F = \lambda_{\max}(\mathscr{A}^*\mathscr{A})$. Given parameters $\lambda > 0, \beta > 0$ such that $\lambda^{-1} \geq L_F/2 + \beta\lambda_{\max}(\mathscr{B}\mathscr{B}^*)$, a sequence of positive numbers $\{\mu_k\}$, and a starting point $(X^0, y^0) \in \Re^{p\times q} \times \Re^s$, Condat's primal-dual splitting method generates new iterates as follows:

$$\tilde{X}^{k+1} = \text{prox}_{\lambda G}(X^k - \lambda\nabla F(X^k) - \lambda\mathscr{B}^*(y^k)) = D_{\rho\lambda}(W^k), \tag{66}$$

$$\tilde{y}^{k+1} = \text{prox}_{\beta H^*}(y^k + \beta\mathscr{B}(2\tilde{X}^{k+1} - X^k)) = y^k + \beta(\mathscr{B}(2\tilde{X}^{k+1} - X^k) - d), \tag{67}$$

$$(X^{k+1}, y^{k+1}) = \mu_k(\tilde{X}^{k+1}, \tilde{y}^{k+1}) + (1 - \mu_k)(X^k, y^k), \tag{68}$$

where $W^k = X^k - \lambda\nabla F(X^k) - \lambda\mathscr{B}^*(y^k)$ and $\text{prox}_{\lambda G}$ denotes the proximal point mapping associated with λG defined by $\text{prox}_{\lambda G}(X) = \text{argmin}_{Y\in\Re^{p\times q}}\lambda G(Y) +$

$\frac{1}{2}\|Y - X\|^2$. We can estimate how close the iterate $(\tilde{X}^{k+1}, \tilde{y}^{k+1})$ is to optimality by noting that $\text{dist}(0, -\mathscr{B}(\tilde{X}^{k+1}) + \partial H^*(\tilde{y}^{k+1})) = \|d - \mathscr{B}(\tilde{X}^{k+1})\|$, and

$$\text{dist}(0, \nabla F(\tilde{X}^{k+1}) + \partial G(\tilde{X}^{k+1}) + \mathscr{B}^* \tilde{y}^{k+1}) \le \|\frac{1}{\lambda}(W^k - \tilde{X}^{k+1}) + \mathscr{B}^* \tilde{y}^{k+1} + \nabla F(\tilde{X}^{k+1})\|,$$

based on the fact that $W^k - \tilde{X}^{k+1} \in \partial(\lambda G)(\tilde{X}^{k+1})$. By letting $Z^{k+1} = \frac{1}{\lambda}(\tilde{X}^{k+1} - W^k), u^{k+1} = \zeta^{k+1} = b - \mathscr{A}(\tilde{X}^{k+1})$ and $\xi^{k+1} = -\tilde{y}^{k+1}$, then

$$\|[\mathscr{A}\tilde{X}^{k+1} + u^{k+1} - b; d - \mathscr{B}(\tilde{X}^{k+1})]\| = \|d - \mathscr{B}(\tilde{X}^{k+1})\|$$

$$\|\frac{1}{\lambda}(W^k - \tilde{X}^{k+1}) + \mathscr{B}^* \tilde{y}^{k+1} + \nabla F(\tilde{X}^{k+1})\| = \|C - (\mathscr{A}^* \zeta^{k+1} + \mathscr{B}^* \xi^{k+1} + Z^{k+1})\|,$$

which correspond to the residuals of the primal and dual equality constraints in (26) and (29), respectively. Therefore, we can stop the primal-dual splitting method when (63) is satisfied. In our numerical experiments, we also stop the method when the number of iterations exceeds 2,000.

5.1   Example 1

We consider the low rank matrix approximation problem in which certain specified entries of the matrix are fixed. In [16], Golub, Hoffman and Stewart derived an explicit formula for finding the nearest lower-rank approximation of the target matrix while certain specified columns of the matrix are fixed. In our numerical experiments, we assume that only partial information of the original matrix is available and the specified fixed entries can be in any random position of the original matrix. For each triplet (p, q, r), we first generate a random matrix $M \in \Re^{p \times q}$ by setting $M = M_1 M_2^T$ where $M_1 \in \Re^{p \times r}$, $M_2 \in \Re^{q \times r}$ have i.i.d. Gaussian entries. Then we sample a subset $\mathscr{E}$ of m entries of M uniformly at random, and generate a random matrix $N_{\mathscr{E}} \in \Re^{p \times q}$ with sparsity pattern $\mathscr{E}$ and i.i.d Gaussian entries. Then we assume that the observed data is given by $\tilde{M}_{\mathscr{E}} = M_{\mathscr{E}} + \tau N_{\mathscr{E}} \|M_{\mathscr{E}}\| / \|N_{\mathscr{E}}\|$, where τ is the noise factor. The minimization problem which we solve can be stated as follows:

$$\min_{X \in \Re^{p \times q}} \left\{ \frac{1}{2}\|X_{\mathscr{E}} - \tilde{M}_{\mathscr{E}}\|_F^2 + \rho \|X\|_* : X_{i_t, j_t} = M_{i_t, j_t}, 1 \le t \le k \right\}, \qquad (69)$$

where $(i_1, j_1), \ldots, (i_k, j_k)$ are distinct pairs. In our numerical experiments, we set $k = \lceil 10^{-3} pq \rceil$, which is the number of prescribed entries selected uniformly at random, and the noise level $\tau = 0, 0.1$.

For each triplet $(p, q, r), m$ and τ, we generate five random instances. In Table 1, we report the average number of the following quantities: number of sampled entries (m); number of outer iterations (it); total number of inner iterations (itsub); number of CG steps taken to solve each linear system in (51) (cg); infeasibilities (R_p, R_D);

Table 1 Numerical performance of the partial PPA on (69). In the table, $dr = r(p + q - r)$ is the degree of freedom in an $p \times q$ matrix of rank r

$p/q/\tau$	r	$m; m/dr$	it; itsub|cg	$R_p \mid R_D$	relgap	MSE	#sv	Time
1,000/1,000/0.0	10	199,104; 10	10.2 | 30.4 | 18.8	2.20e−7 | 8.05e−7	−1.57e−4	1.32e−3	10	1:41
	50	487,314; 5	10.0 | 16.0 | 15.3	1.26e−7 | 7.07e−7	−3.94e−5	1.64e−3	50	1:04
	100	949,927; 5	9.0 | 10.2 | 12.4	1.23e−7 | 4.38e−7	−8.96e−6	1.48e−3	100	43
1,500/1,500/0.0	10	299,272; 10	13.0 | 32.4 | 16.9	1.95e−7 | 1.92e−7	−1.70e−5	1.30e−3	10	5:33
	50	738,078; 5	11.0 | 18.8 | 15.2	1.38e−7 | 6.46e−7	−3.54e−5	1.68e−3	50	3:45
	100	1,449,450; 5	10.0 | 15.2 | 18.7	1.27e−7 | 3.42e−7	−1.07e−5	1.60e−3	100	3:43
1,000/1,000/0.1	10	199,104; 10	20.0 | 46.4 | 10.5	4.97e−7 | 7.95e−7	−8.79e−5	7.75e−2	10	2:11
	50	487,314; 5	23.0 | 44.4 | 11.0	1.67e−7 | 5.18e−7	−9.04e−6	9.64e−2	50	2:42
	100	949,927; 5	21.0 | 25.2 | 9.7	3.68e−7 | 6.09e−7	−3.01e−7	9.66e−2	100	1:54
1,500/1,500/0.1	10	299,272; 10	21.4 | 48.6 | 10.7	1.17e−7 | 8.67e−7	−1.46e−4	7.50e−2	10	7:09
	50	738,078; 5	22.8 | 44.8 | 11.7	1.59e−7 | 7.07e−7	−2.20e−5	9.72e−2	50	8:09
	100	1,449,450; 5	23.0 | 39.4 | 11.4	7.33e−7 | 6.63e−7	−2.95e−6	9.62e−2	100	7:57
100/10,000/0.0	10	504,310; 5	14.2 | 23.4 | 13.4	5.73e−7 | 5.39e−7	−2.07e−5	1.66e−3	10	38
100/50,000/0.0	10	2,504,270; 5	14.4 | 21.6 | 16.3	4.05e−7 | 6.44e−7	−2.58e−5	1.67e−3	10	4:10
100/100,000/0.0	10	5,004,144; 5	14.4 | 22.4 | 18.4	2.70e−7 | 4.68e−7	−1.95e−5	1.67e−3	10	9:09
100/10,000/0.1	10	504,310; 5	22.0 | 34.2 | 5.7	3.26e−7 | 6.49e−7	−4.07e−5	8.95e−2	10	31
100/50,000/0.1	10	2,504,270; 5	22.2 | 33.0 | 5.7	2.73e−7 | 7.22e−7	−4.57e−5	8.94e−2	10	3:29
100/100,000/0.1	10	5,004,144; 5	21.8 | 31.0 | 5.9	5.31e−7 | 7.04e−7	−4.49e−5	8.93e−2	10	6:47
200/10,000/0.0	10	510,046; 5	13.4 | 31.4 | 18.6	4.66e−7 | 2.24e−7	−1.79e−5	1.76e−3	10	2:23
200/50,000/0.0	10	2,509,504; 5	13.6 | 31.2 | 22.3	4.23e−7 | 2.11e−7	−1.63e−5	1.78e−3	10	14:43
200/100,000/0.0	10	5,010,168; 5	13.8 | 29.2 | 22.8	2.30e−7 | 5.18e−7	−1.73e−5	1.78e−3	10	28:45
500/10,000/0.0	10	524,511; 5	16.8 | 50.4 | 28.3	3.40e−7 | 2.75e−7	−2.00e−5	1.78e−3	10	17:21
200/10,000/0.1	10	510,046; 5	19.0 | 36.0 | 7.6	5.40e−7 | 9.12e−7	−1.15e−4	8.78e−2	10	1:37
200/50,000/0.1	10	2,509,504; 5	19.0 | 34.0 | 8.0	4.68e−7 | 9.24e−7	−1.17e−4	8.67e−2	10	9:29
200/100,000/0.1	10	5,010,168; 5	19.6 | 35.4 | 8.1	3.46e−7 | 7.85e−7	−9.87e−5	8.65e−2	10	20:25
500/10,000/0.1	10	524,511; 5	24.6 | 51.6 | 11.5	2.91e−7 | 7.56e−7	−1.62e−4	9.34e−2	10	11:44

relative duality gap (relgap); relative mean square error MSE $:= \|X - M\|/\|M\|$; numerical rank of X (#sv); and the CPU time taken. Here we report the numerical rank of X defined as follows:

$$\#\mathrm{sv}(X) := \max\{k \ : \ \sigma_k(X) \geq \max\{10^{-8}, \tau\}\sigma_1(X)\}. \tag{70}$$

In this example, we compare the performance of the SSNCG partial PPA with the ADMM method [8] which is applied to the dual problem (29) and the primal-dual splitting method [7]. We use the same stopping criterion (63) for ADMM. Table 2 reports the average results of ADMM for each instance of five runs. We may observe from the tables that our algorithm is overall more efficient than ADMM for solving (69). For the problem where p is moderate but q is large, e.g., $p = 100$, $q = 100{,}000$ and $\tau = 0$, it takes the partial PPA less than half of the time needed by ADMM to achieve the tolerance of 10^{-6} while the MSE is reasonably small. Table 3 reports the average results of the primal-dual splitting method for each instance of five runs. We set the parameters $\lambda = 1.5$ and $\beta = 1/6$ after some tuning for good performance. The parameter μ_k is set to be 1 for all k as this choice gives the best performance. We can observe from the tables that the primal-dual splitting method is much slower than ADMM for solving (69). For $p = q = 1{,}500$, $\tau = 0$ and $r = 10$, the primal-dual splitting method cannot achieve the tolerance of 10^{-6} within 1 h and the obtained solution is of very high rank with relatively large MSE.

5.2 Example 2

In the Euclidean metric embedding problem, we are given an incomplete, possibly noisy, dissimilarity matrix $B \in \mathscr{S}^n$ with $\mathrm{Diag}(B) = 0$ and sparsity pattern specified by the set of indices $\mathscr{E} = \{(i, j) \mid B_{ij} \neq 0, 1 \leq i < j \leq n\}$. The goal is to find an Euclidean distance matrix (EDM) [1] that is nearest to B. If the measure of nearness is in the Frobenius norm, then the mathematical formulation of the problem is as follows:

$$\min\left\{\frac{1}{2}\sum_{(i,j)\in\mathscr{E}} W_{ij}(D_{ij} - B_{ij})^2 + \frac{\rho}{2n}\langle E, D\rangle \ : \ D \text{ is an EDM}\right\}, \tag{71}$$

where $W_{ij} > 0, (i, j) \in \mathscr{E}$, are given weights, $E \in \mathscr{S}^n$ is the matrix of all ones and $\rho > 0$ is a regularization parameter. Here we add the term $\frac{\rho}{2n}\langle E, D\rangle$ to encourage a sparse solution. Recall that a standard characterization [1] of an EDM D is that $D = \mathrm{Diag}(X)e^T + e\,\mathrm{Diag}(X)^T - 2X$ for some $X \succeq 0$ with $Xe = 0$, where $e \in \mathfrak{R}^n$ is the vector of all ones. Thus the problem (71) can be rewritten as:

$$\min\left\{\frac{1}{2}\sum_{(i,j)\in\mathscr{E}} W_{ij}(\langle A_{ij}, X\rangle - B_{ij})^2 + \rho\langle I, X\rangle \ : \ \langle E, X\rangle = 0, \ X \succeq 0\right\}, \tag{72}$$

Table 2 Numerical performance of the ADMM method on (69)

$p/q/\tau$	r	m; m/dr	it.	$R_p \mid R_D \mid$ relgap			MSE	#sv	Time
1,000/1,000/0.0	10	199,104; 10	360.6	9.91e − 7	2.36e − 7	1.03e − 5	1.32e − 3	10	4:01
	50	487,314; 5	184.6	9.75e − 7	3.24e − 7	3.06e − 6	1.64e − 3	50	2:29
	100	949,927; 5	147.8	9.68e − 7	3.05e − 7	1.57e − 6	1.48e − 3	100	2:24
1,500/1,500/0.0	10	299,272; 10	464.2	9.92e − 7	2.59e − 7	1.31e − 5	1.30e − 3	10	16:11
	50	738,078; 5	330.2	9.63e − 7	4.69e − 7	8.56e − 6	1.68e − 3	50	12:43
	100	1,449,450; 5	169.8	9.83e − 7	4.19e − 7	1.88e − 6	1.60e − 3	100	7:22
1,000/1,000/0.1	10	199,104; 10	183.6	9.85e − 7	5.94e − 7	−5.02e − 5	7.75e − 2	10	2:02
	50	487,314; 5	99.0	9.17e − 7	5.87e − 7	−1.79e − 6	9.64e − 2	50	1:22
	100	949,927; 5	72.4	9.79e − 7	4.98e − 7	1.22e − 7	9.66e − 2	100	1:15
1,500/1,500/0.1	10	299,272; 10	198.6	9.69e − 7	1.77e − 7	−1.93e − 5	7.50e − 2	10	6:56
	50	738,078; 5	93.0	9.61e − 7	2.34e − 7	−2.88e − 6	9.72e − 2	50	3:38
	100	1,449,450; 5	74.0	9.47e − 7	4.36e − 7	−2.11e − 7	9.62e − 2	100	3:20
100/10,000/0.0	10	504,310; 5	276.6	9.78e − 7	3.48e − 7	4.31e − 6	1.66e − 3	10	1:48
100/100,000/0.0	10	5,004,144; 5	254.6	9.85e − 7	3.30e − 7	3.90e − 6	1.67e − 3	10	24:04
500/10,000/0.0	10	524,511; 5	679.2	9.91e − 7	2.90e − 7	1.20e − 5	1.78e − 3	10	26:16
100/10,000/0.1	10	504,310; 5	66.2	9.58e − 7	5.27e − 7	−1.71e − 5	8.95e − 2	10	27
100/100,000/0.1	10	5,004,144; 5	63.4	9.41e − 7	4.28e − 7	−1.34e − 5	8.93e − 2	10	6:05
500/10,000/0.1	10	524,511; 5	329.2	9.91e − 7	3.62e − 7	−5.33e − 5	9.33e − 2	10	13:08

Table 3 Numerical performance of the primal-dual splitting method on (69)

$p/q/\tau$	r	$m; m/dr$	it.	R_p	R_D	relgap	MSE	#sv	Time
1,000/1,000/0.0	10	199,104; 10	2,000.0	2.10e−5	7.09e−5	−2.31e−3	2.57e−3	11	19:11
	50	487,314; 5	806.4	9.83e−7	5.30e−8	−1.19e−7	1.64e−3	50	8:03
1,500/1,500/0.0	10	299,272; 10	2,000.0	1.73e−5	1.79e−3	−5.26e−1	1.81e−1	302	1:04:46
	50	738,078; 5	1,379.6	9.83e−7	6.92e−8	−2.55e−7	1.68e−3	50	45:24
1,000/1,000/0.1	10	199,104; 10	2,000.0	9.54e−6	3.64e−4	−7.78e−2	8.76e−2	10	19:09
	50	487,314; 5	2,000.0	2.19e−8	1.44e−6	−2.79e−5	9.64e−2	50	20:01
1,500/1,500/0.1	10	299,272; 10	2,000.0	1.28e−5	1.41e−3	−3.71e−1	2.34e−1	10	1:04:47
	50	738,078; 5	2,000.0	1.38e−6	1.19e−4	−5.97e−3	9.84e−2	50	1:05:26
100/10,000/0.0	10	504,310; 5	1,292.8	9.93e−7	2.27e−8	−1.82e−8	1.66e−3	10	4:03
100/100,000/0.0	10	5,004,144; 5	1,328.2	9.99e−7	2.34e−8	−1.71e−8	1.67e−3	10	1:21:59
100/10,000/0.1	10	504,310; 5	1,706.4	2.37e−8	9.98e−7	−4.91e−5	8.95e−2	10	5:29
100/100,000/0.1	10	5,004,144; 5	1,711.0	2.49e−8	9.96e−7	−4.94e−5	8.93e−2	10	1:46:43

where $A_{ij} = (e_i - e_j)(e_i - e_j)^T$ and e_i is the i-th standard unit vector in $\Re^n$. Note that under the condition $X \succeq 0$, the constraint $Xe = 0$ is equivalent to $\langle E, X \rangle = 0$. It is interesting to note that desiring sparsity in the EDM D leads to the regularization term $\rho \langle I, X \rangle$, which is a proxy for desiring a low-rank X.

Let $m = |\mathcal{E}|$. The linear maps $\mathcal{A} : \mathcal{S}^n \to \Re^m$ and $\mathcal{B} : \mathcal{S}^n \to \Re$ for the problem (72) are given as follows:

$$(\mathcal{A}(X))_{(i,j)} = \langle \sqrt{W_{ij}} A_{ij}, X \rangle, \ \forall \ (i,j) \in \mathcal{E}, \quad \mathcal{B}(X) = \langle E, X \rangle.$$

Note that the components of a vector in $\Re^m$ are enumerated based on the elements in $\mathcal{E}$. And the operator $\mathcal{V}_t$ in (59) is given as follows:

$$\mathcal{V}_t = \sigma \left(\frac{1}{\sigma} T + \begin{bmatrix} \mathcal{A} \\ \mathcal{B} \end{bmatrix} \Pi'_{\mathcal{S}^n_+} (W(y^t; X)) \begin{bmatrix} \mathcal{A}^* & \mathcal{B}^* \end{bmatrix} \right). \tag{73}$$

For the EDM problem (72), the condition number of $\mathcal{V}_t$ can be quite large and it is important to find a good preconditioner for $\mathcal{V}_t$ so that the CG method can have a reasonable convergence speed when solving the linear system of equations associated with $\mathcal{V}_t$. Let $\mathbf{A}, \mathbf{B}$ and $\mathbf{S}$ be the matrix representations of $\mathcal{A}, \mathcal{B}$ and $\Pi'_{\mathcal{S}^n_+} (W(y^t; X))$ with respect to the standard basis of $\mathcal{S}^n$ and $\Re^m$, respectively. Let $\mathbf{h} \in \Re^{n(n+1)/2}$ and $\Gamma = \{(i,j) \mid 1 \leq i \leq j \leq n\}$. Suppose $\mathrm{Diag}(\mathbf{h})$ is a positive definite diagonal approximation of $\mathbf{S}$. (In our implementation, we choose $\mathbf{h}$ to be the approximate diagonal of $\Pi'_{\mathcal{S}^n_+} (W(y^t; X))$ considered in [13].) Let $H \in \mathcal{S}^n$ be the matrix such that $H_{ij} = H_{ji} = \mathbf{h}_{(i,j)}$ for all $(i,j) \in \Gamma$. We consider the following approximation of $\frac{1}{\sigma} \mathcal{V}_t$:

$$\mathcal{M} = \begin{bmatrix} M & q \\ q^T & \alpha \end{bmatrix}, \tag{74}$$

where $q = \mathcal{A}(H) \in \Re^m, \alpha = \langle E, H \rangle$, and

$$M = \frac{1}{\sigma} I_m + \mathbf{A} \mathrm{Diag}(\mathbf{h}) \mathbf{A}^T \in \Re^{m \times m}. \tag{75}$$

Note that the rows and columns of M are enumerated based on the elements of $\mathcal{E}$. We have that

$$M_{(i,j),(s,t)} = \begin{cases} 0 & \text{if } i \neq s, j \neq t, \\ \sqrt{W_{ij} W_{st}} \, H_{ss} & \text{if } i = s, j \neq t, \\ \sqrt{W_{ij} W_{st}} \, H_{tt} & \text{if } i \neq s, j = t, \\ 1/\sigma + \sqrt{W_{ij} W_{st}} (H_{ss} + H_{tt} + 2H_{st}) & \text{if } i = s, j = t. \end{cases}$$

Let $\bar{\mathbf{h}}_{(i,j)} = W_{ij} H_{ij}$ for all $(i,j) \in \mathcal{E}$. Then we know that M has the following structure

$$M = \mathbf{D} + JJ^T, \tag{76}$$

where $\mathbf{D} = \frac{1}{\sigma} I_m + 2\mathrm{Diag}(\bar{\mathbf{h}})$, and $J \in \Re^{m \times n}$ is the weighted arc-node incidence matrix for which the entry at the (s, t)-th row and k-th column is given by

$$J_{(s,t),k} = \begin{cases} \sqrt{W_{st} H_{ss}} & \text{if } k = s, \\ \sqrt{W_{st} H_{tt}} & \text{if } k = t, \\ 0 & \text{otherwise.} \end{cases}$$

To use $\mathcal{M}$ as a preconditioner for $\mathcal{V}_t$, we need the inverse of $\mathcal{M}$, which is given in the following expression:

$$\mathcal{M}^{-1} = \begin{bmatrix} S^{-1} & -\alpha^{-1} S^{-1} q \\ -\alpha^{-1} q^T S^{-1} & \alpha^{-1} + \alpha^{-2} q^T S^{-1} q \end{bmatrix}, \tag{77}$$

where

$$S = M - \alpha^{-1} q q^T = \mathbf{D} + \underbrace{[J, \, q]}_{\hat{J}} \begin{bmatrix} I_n & 0 \\ 0 & -\alpha^{-1} \end{bmatrix} \begin{bmatrix} J^T \\ q^T \end{bmatrix}.$$

By using the Sherman-Morrison-Woodbury formula [15], we have that

$$S^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1} \hat{J} (\Lambda + \hat{J}^T \mathbf{D}^{-1} \hat{J})^{-1} \hat{J}^T \mathbf{D}^{-1}, \tag{78}$$

where $\Lambda = [I_n, 0; 0, -\alpha]$. Here we assume that $\Lambda + \hat{J}^T \mathbf{D}^{-1} \hat{J}$ is nonsingular; otherwise we may consider the following block diagonal approximation of $\frac{1}{\sigma} \mathcal{V}_t$:

$$\mathcal{M}^d = \begin{bmatrix} M & 0 \\ 0 & \alpha \end{bmatrix}, \tag{79}$$

where the inverse of M can also be computed via the Sherman-Morrison-Woodbury formula.

The Euclidean metric problem arises in many applications. For the regularized kernel estimation (RKE) problem in statistics [22], we are given a set of n objects and dissimilarity measures d_{ij} for certain object pairs $(i, j) \in \mathcal{E}$. The goal is to estimate a positive semidefinite kernel matrix $X \in \mathcal{S}_+^n$ such that the fitted squared distances induced by X between the objects satisfy the following condition:

$$X_{ii} + X_{jj} - 2X_{ij} = \langle A_{ij}, X \rangle \approx d_{ij}^2 \quad \forall \, (i, j) \in \mathcal{E}.$$

Formally, one version of the RKE problem proposed in [22] is the SDP problem (72).

In our numerical experiments, the data d_{ij} are normalized to be in the interval [0, 1], and $\mathcal{E} = \{(i, j) : 1 \le i < j \le 630\}$. We set $W_{ij} = 1$ for all $(i, j) \in \mathcal{E}$.

Fig. 1 A 3D representation
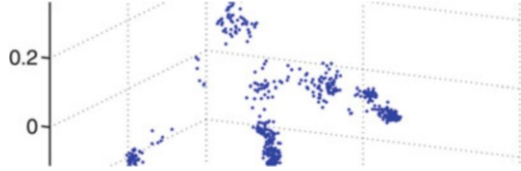of the sequence space for 630
proteins

Table 4 Numerical results on the RKE problem arising from protein clustering

problem	n	$m + s$	ρ	it.|itsub|cg	$R_p \mid R_D \mid$ relgap	#sv	Time
RKE630-$\mathscr{M}$	630	198,136	$5.07e-1$	6 | 36| 10.3	$2.67e-7 \mid 2.26e-8$ $\mid -1.54e-6$	388	36
PDB25-$\mathscr{M}$	1,898	1,646,031	$1.84e+0$	23 | 68 | 22.3	$1.89e-7 \mid 6.28e-7$ $\mid -1.31e-6$	1,371	23:43
PDB25	1,898	1,646,031	$1.84e+0$	23 | 80 | 76.7	$9.60e-7 \mid 6.32e-7$ $\mid -1.21e-6$	1,371	59:56

In [22], due to the prohibitive computational load encountered by the standard
interior-point solver (such as SDPT3 or SeDuMi) used to solve (72), a subset of 280
globin proteins were selected from the entire set of 630 proteins for the numerical
results reported in [22]. For each of the selected proteins, 55 dissimilarities were
randomly selected out of the total of 279. Here we are able to consider the entire set
of 630 proteins and the dissimilarities among all the pairs of proteins.

As mentioned in [22], the RKE methodology can provide an efficient way to
represent each protein sequence by a feature vector in an appropriate coordinate
system using the pairwise dissimilarity between protein sequences. Specifically,
we project the computed solution $\overline{X}$ onto a 3D space corresponding to the largest
three eigenvalues. Figure 1 displays a 3D representation of the sequence space
for 630 proteins. There are at least 4 classes visually identifiable in the data set
of 630 proteins, which is consistent with the observations in [22]. The numerical
results for solving (72) are reported in Table 4, where #sv is the number of positive
eigenvalues of $\overline{X}$. For the computed solution $\overline{X}$, we have $\langle \overline{X}, E \rangle = 4.46 \times 10^{-13}$
and $\langle \overline{X}, I \rangle = 1.85 \times 10^2$.

We also conducted numerical experiments on a much larger protein data set to
evaluate the performance of our algorithm. We used the PDB_SELECT 25 data
set [18], a representative subset of the Protein Data Bank database, which contains
1,898 protein chains. The numerical results for the PDB_SELECT 25 data set are
reported in Table 4. For the computed solution $\overline{X}$, we have $\langle \overline{X}, E \rangle = -1.94 \times 10^{-14}$
and $\langle \overline{X}, I \rangle = 8.76 \times 10^2$. To appreciate the usefulness of the preconditioner
$\mathscr{M}$, we also report the numerical results for the PDB25 problem without using
the preconditioner $\mathscr{M}$ in the last row of Table 4. We can observe that the CG
solver without using the preconditioner $\mathscr{M}$ requires substantially more CG steps
for computing the inexact Newton direction from the linear system (51).

5.3 Example 3: Molecular Conformation Problems

The molecular conformation problem for a molecule with n atoms is the problem of determining the positions $x_1, \ldots, x_n$ of the atoms, given estimated inter-atomic distances d_{ij} between some pairs of atoms. The estimated distances could be information derived from covalent bond lengths or measured from nuclear magnetic resonance (NMR) experiments. Let $\mathscr{E}$ be the set of pairs of indices (i, j) $(i < j)$ for which estimates on the distances $\|x_i - x_j\|$ are available. The molecular conformation problem can mathematically be formulated as follows:

$$\min \left\{ \frac{1}{2} \sum_{(i,j) \in \mathscr{E}} W_{ij} (\|x_i - x_j\|^2 - d_{ij}^2)^2 - \frac{\rho}{2n} \sum_{i,j=1}^{n} \|x_i - x_j\|^2 \mid \sum_{i=1}^{n} x_i = 0 \right\}, \quad (80)$$

where $W \in \mathscr{S}^n$ is a weight matrix with positive entries. The second term in the objective function is added to maximize the pairwise separations between the atoms. The equality constraint in (80) is included to set the center of mass of the molecule at the origin. The inclusion of weights is useful for differentiating data that are derived from different sources, and hence of different reliability. For example, distance data which are derived from covalent bond lengths are usually much more accurate than data which are derived from the NMR experiments.

Let $X = [x_1, \ldots, x_n] \in \Re^{3 \times n}$ and $A_{ij} = e_{ij} e_{ij}^T$, where $e_{ij} = e_i - e_j$. Then we have $x_i - x_j = X e_{ij}$ and hence $\|x_i - x_j\|^2 = \langle X^T X, A_{ij} \rangle$. Let $Y = X^T X$. The constraint $\sum_{i=1}^{n} x_i = 0$ can equivalently be replaced by $\langle E, Y \rangle = 0$. Note that under the latter constraint, it is easy to see that $\sum_{i,j=1}^{n} \|x_i - x_j\|^2 = 2n \langle I, Y \rangle$. By relaxing the non-convex constraint $Y = X^T X$ to $Y \succeq 0$ in (80), we get the following SDP problem:

$$\min \left\{ \frac{1}{2} \sum_{(i,j) \in \mathscr{E}} W_{ij} (\langle A_{ij}, Y \rangle - d_{ij}^2)^2 - \rho \langle I, Y \rangle \mid \langle E, Y \rangle = 0, Y \succeq 0 \right\}. \quad (81)$$

Observe that under the constraints $\langle E, Y \rangle = 0$ and $Y \succeq 0$, we have $\langle I, Y \rangle = \frac{1}{2n} \langle E, D \rangle$, where $D = \mathrm{Diag}(Y) e^T + e \, \mathrm{Diag}(Y)^T - 2Y$ and $e \in \Re^n$ is a vector of all ones. Thus (81) seeks an Euclidean distance matrix D which is encouraged to have as many nonzero entries as possible. Note that in the maximum variance unfolding problem [30], one also consider a problem that has exactly the same form as (81).

In this example, we focus on molecular conformation problems with noisy and sparse distance data. In our numerical experiments, we set $W_{ij} = 1/d_{ij}^2$ for all $(i, j) \in \mathscr{E}$ and $\rho = 8 \times 10^{-4} \|\mathscr{A}^* b\|_2$. For each molecule, we generated the partial inter-atomic distance matrix as follows. If the distance between two atoms is less than 6 Å (1 Å $= 10^{-8}$ cm), which is nearly the maximal distance that the NMR experiment can measure between two atoms, the distance is chosen; otherwise no distance information about the pair is known. Since not all the distances below 6 Å are known from NMR experiments, we randomly choose 30 % of all the distance

Table 5 Numerical results on molecular conformation problems with 30 % distances $\leq$6 Å, which are corrupted by 20 % normal noise

Molecule	$n; m + s$	it.| itsub | cg	R_p | R_D | relgap	RMSD	Time
1GM2	166; 1,119	23 | 81 | 58.3	$3.02e-7$ | $7.46e-7$ | $-4.92e-4$	1.15	7
1PBM	388; 3,145	29 | 113 | 89.5	$6.25e-7$ | $5.93e-7$ | $-2.55e-4$	1.20	52
1PTQ	402; 2,182	28 | 108 | 155.8	$1.04e-7$ | $3.42e-7$ | $-1.84e-4$	1.51	1:24
1CTF	487; 2,630	28 | 135 | 182.8	$3.11e-7$ | $3.65e-7$ | $-2.88e-4$	1.99	2:49
1AU6	506; 4,767	24 | 105 | 168.7	$2.88e-7$ | $6.88e-7$ | $-1.02e-3$	0.96	2:24
1HOE	558; 3,083	26 | 125 | 176.5	$9.73e-7$ | $6.70e-7$ | $-4.01e-4$	1.20	3:34
1PHT	814; 5,239	31 | 169 | 204.8	$2.90e-7$ | $9.36e-7$ | $-5.21e-4$	2.65	13:05
1POA	914; 5,045	31 | 192 | 224.4	$3.29e-7$ | $4.08e-7$ | $-4.41e-4$	2.57	20:40
1AX8	1,003; 5,563	30 | 184 | 207.3	$2.67e-7$ | $7.67e-7$ | $-6.70e-4$	2.05	22:49

Table 6 Numerical results on molecular conformation problems with 30 % distances $\leq$6 Å, which are corrupted by 20 % uniform noise

Molecule	$n; m + s$	it.| itsub | cg	R_p | R_D | relgap	RMSD	Time
1GM2	166; 1,119	24 | 88 | 63.3	$2.14e-7$ | $5.19e-7$ | $-1.51e-4$	0.91	8
1PBM	388; 3,145	24 | 105 | 106.6	$1.33e-7$ | $8.78e-7$ | $-3.13e-4$	1.02	57
1PTQ	402; 2,182	26 | 112 | 145.1	$1.03e-7$ | $8.20e-7$ | $-4.32e-4$	1.31	1:22
1CTF	487; 2,630	28 | 131 | 179.2	$3.37e-7$ | $3.18e-7$ | $-2.38e-4$	1.73	2:43
1AU6	506; 4,767	24 | 131 | 162.7	$3.32e-7$ | $5.27e-7$ | $-5.14e-4$	0.79	2:54
1HOE	558; 3,083	26 | 124 | 184.6	$9.59e-8$ | $6.92e-7$ | $-3.94e-4$	1.15	3:43
1PHT	814; 5,239	31 | 180 | 205.8	$6.22e-7$ | $5.13e-7$ | $-2.37e-4$	2.47	13:57
1POA	914; 5,045	30 | 185 | 209.5	$6.80e-7$ | $8.48e-7$ | $-8.18e-4$	2.25	18:43
1AX8	1,003; 5,563	30 | 175 | 214.4	$6.99e-7$ | $6.65e-7$ | $-5.07e-4$	1.84	22:13

below 6 Å in our experiment. For realistic molecular conformation problems, in which the exact distances are not known and only the lower bounds $\underline{d}_{ij}$ and upper bounds $\bar{d}_{ij}$ on distances are provided, we use the mean $d_{ij} = (\bar{d}_{ij} + \underline{d}_{ij})/2$ as the estimated distances. After selecting 30 % of inter-atomic distances, we add certain amount of normal noise or uniform noise to the distances to generate the lower and upper bounds. Suppose that $\hat{d}_{ij}$ is the exact distance between atom i and atom j, we set

$$\underline{d}_{ij} = \max(1, (1 - |\underline{\varepsilon}_{ij}|)\hat{d}_{ij}), \quad \bar{d}_{ij} = (1 + |\bar{\varepsilon}_{ij}|)\hat{d}_{ij}.$$

Let τ be a given noise level. In the normal noise model, $\underline{\varepsilon}_{ij}, \bar{\varepsilon}_{ij} \sim \mathcal{N}(0, \pi\tau^2/2)$ are independent normal random variables. In the uniform noise model, $\underline{\varepsilon}_{ij}, \bar{\varepsilon}_{ij}$ are independent uniform random variables in the interval $[0, 2\tau]$. We said that the distances are corrupted by 20 % noise if $\tau = 0.2$.

In Tables 5 and 6, we report the numerical results on molecular conformation problems under the normal noise model and uniform noise model, respectively, where the root mean square deviation (RMSD) is used to measure the accuracy of the estimated positions. The RMSD is defined by the following formula:

$$\text{RMSD} := \frac{1}{\sqrt{n}} \Big(\sum_{i=1}^{n} \| x_i - \hat{x}_i \|^2 \Big)^{1/2}, \tag{82}$$

where x_i is the estimated position and $\hat{x}_i$ is the actual position. We can observe from the tables that the estimated atomic positions via a simple projection of Y onto the 3-dimensional space are fairly accurate with RMSD $\approx 2\,\overset{\circ}{\text{A}}$.

6 Conclusion

In this paper, we introduced a partial PPA for solving nuclear norm regularized and semidefinite matrix least squares problems with linear equality constraints. The inner subproblems are solved inexactly by a semismooth Newton-CG method, whose convergence analysis is established under a constraint nondegeneracy condition, together with the strong semi-smoothness property of the soft-thresholding operator and the metric projector $\Pi_{\mathscr{S}_+^n}$. Numerical experiments conducted on nuclear norm regularized matrix least squares problems, regularized kernel estimation problems and molecular conformation problems demonstrated that our algorithm is efficient and robust. In [12], the graph visualization problem with applications in social network data analysis can be modeled as an optimization problem which is analogous to (80). It will be very interesting to investigate how to adapt our SSNCG partial PPA algorithm to solve large scale social network graph visualization problems.

References

1. Alfakih, A.Y., Khandani, A., Wolkowicz, H.: Solving Euclidean distance matrix completion problems via semidefinite programming. Comput. Optim. Appl. **12**, 13–30 (1999)
2. Bhatia, R.: Matrix Analysis. Springer, NewYork (1997)
3. Cai, J.F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. Optim. **20**, 1956–1982 (2010)
4. Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Found. Comput. Math. **9**, 717–772 (2009)
5. Clarke, F.: Optimization and Nonsmooth Analysis. Wiley, New York (1983)
6. Combettes, P.L., Pesquet, J.-C.: Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators. http://arxiv.org/abs/1107.0081v2 (2011)
7. Condat, L.: A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms (2011, preprint)
8. Eckstein, J., Bertsekas, D.: On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**, 293–318 (1992)
9. Fazel, M.: Matrix rank minimization with applications. Ph.D. thesis, Stanford University (2002)
10. Fazel, M., Hindi, H., Boyd, S.: A rank minimization heuristic with application to minimum order system approximation. In: Proceedings of the American Control Conference, Arlington (2001)

11. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximations. Comput. Math. Appl. **2**, 17–40 (1976)
12. Gansner, E.R., Hu, Y., North, S.: A maxent-stress model for graph layout. In: Proceedings of IEEE Pacific Visualization Symposium, Songdo (2012)
13. Gao, Y., Sun, D.F.: Calibrating least squares semidefinite programming with equality and inequality constraints. SIAM J. Matrix Anal. Appl. **31**, 1432–1457 (2009)
14. Glowinski, R., Marroco, A.: Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéares. Revue Francaise d'Automatique, Informatique et Recherche Opérationelle **9(R-2)**, 41–76 (1975)
15. Golub, G.H., van Loan, C.F.: Matrix Computations, 3rd edn. The Johns Hopkins University Press, Baltimore (1996)
16. Golub, G.H., Hoffman, A., Stewart, G.W.: A generalization of the Eckart-Young-Mirsky matrix approximation theorem. Linear Algebra Appl. **88**, 317–327 (1987)
17. Ha, C.: A generalization of the proximal point algorithm. SIAM J. Control Optim. **28**, 503–512 (1990)
18. Hou, J., Jun, S.R., Zhang, C., Kim, S.H.: Global mapping of the protein structure space and application in structure-based inference of protein function. Proc. Natl. Acad. Sci. USA **102**, 3651–3656 (2005)
19. Jiang, K.F., Sun, D.F., Toh, K.C.: A partial proximal point algorithm for nuclear norm regularized matrix least squares problems with polyhedral constraints (2012, preprint)
20. Liu, Y.J., Sun, D.F., Toh, K.C.: An implementable proximal point algorithmic framework for nuclear norm minimization. Math. Program. **133**, 399–436 (2012)
21. Löwner, K.: Über monotone matrixfunktionen. Math. Z. **38**, 177–216 (1934)
22. Lu, F., Keleş, S., Wright, S., Wahba, G.: Framework for kernel regularization with application to protein clustering. Proc. Natl. Acad. Sci. USA **102**, 12332–12337 (2005)
23. Ma, S., Goldfarb, D., Chen, L.: Fixed point and Bregman iterative methods for matrix rank minimization. Math. Program. **128**, 321–353 (2011)
24. Qi, L., Sun, J.: A nonsmooth version of Newton's method. Math. Program. **58**, 353–367 (1993)
25. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. **52**, 471–501 (2010)
26. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
27. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. **14**, 877–898 (1976)
28. Rockafellar, R.T.: Augmented Lagrangains and applications of the proximal point algorithm in convex programming. Math. Oper. Res. **1**, 97–116 (1976)
29. Toh, K.C., Yun, S.W.: An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. Pac. J. Optim. **6**, 615–640 (2010)
30. Weinberger, K.Q., Sha, F., Zhu, Q., Saul, L.K.: Graph Laplacian regularization for large-scale semidefinite programming. In: Schoelkopf, B., Platt, J., Hofmann, T. (eds.) Advances in Neural Information Processing Systems, pp. 1489–1496. MIT, Cambridge (2007)
31. Yang, Z.: A study on nonsymmetric matrix-valued functions. Master thesis, Department of Mathematics, National University of Singapore (2009)
32. Zhao, X.Y., Sun, D.F., Toh, K.C.: A Newton-CG augmented Lagrangian method for semidefinite programming. SIAM J. Optim. **20**, 1737–1765 (2010)

Techniques for Submodular Maximization

Jon Lee

Abstract Maximization of a submodular function is a central problem in the algorithmic theory of combinatorial optimization. On the one hand, it has the feel of a clean and stylized problem, amenable to mathematical analysis, while on the other hand, it comfortably contains several rather different problems which are independently of interest from both theoretical and applied points of view. There have been successful analyses from the point of view of theoretical computer science, specifically approximation algorithms, and from an operations research viewpoint, specifically novel branch-and-bound methods have proven to be effective on broad subclasses of problems. To some extent, both of these points of view have validated what some practitioners have known all along: Local-search methods are very effective for many of these problems.

1   Introduction

Submodular functions have been a central subject in the development of discrete optimization algorithms for over 40 years. Just as convex functions have a central place in continuous optimization, submodular functions, as the discrete analog of convex functions, play an extremely important role in the development and

J. Lee (✉)
IOE Department, University of Michigan, 1205 Beal Avenue,
Ann Arbor, MI 48109-2117, USA
e-mail: jonxlee@umich.edu

progress of techniques for discrete optimization. Examples of submodular functions and their optimization abound, with direct applications in very practical areas like electrical engineering, statics, environmental (and other spatial) monitoring, machine learning, and biodiversity conservation, but also in very useful and generally applicable integer-programming model classes like set covering and matching (with applications in, for example, airline crew scheduling, clustering and routing), as well as more mathematical areas like game theory and combinatorics (see, for example, [12, 13, 24, 31, 39, 44, 48]).

A real-valued function f defined on subsets of a finite ground set E is *submodular* if

$$f(X \cup Y) + f(X \cap Y) \le f(X) + f(Y), \quad \forall X, Y \subseteq E.$$

With this definition, we can readily see how such functions appear to be a kind of discrete analog of convex continuous functions. Equivalently, f is submodular if

$$f(Y \cup \{e\}) - f(Y) \le f(X \cup \{e\}) - f(X), \quad \forall X \subseteq Y \subseteq E.$$

From this point of view, such functions are those that model *non-increasing marginal value*. As these two concepts are of central importance for optimization associated with continuous functions, it seems clear that submodularity has the potential to be a key concept for optimization in the discrete setting. Indeed, submodularity has proven to be a unifying concept for distinguishing between tractable and intractable discrete-optimization problem classes, and for pointing the way toward useful solution methodologies (see, for example, [13, 17, 31]). Still, the concept is not a panacea—we do not yet have practical, general-purpose algorithms and software for submodular function optimization.

In recent years, there has been an enormous amount of research on algorithms for optimizing submodular functions. First, there was the seminal work establishing the polynomial-time complexity of *minimizing* a submodular function (see [16], recognized for a Fulkerson Prize in 1982), capitalizing on the polynomial-time solvability of linear programming via the ellipsoid method. Next, theoretically-efficient, combinatorial algorithms were discovered (see [20, 47], recognized for a Fulkerson Prize in 2003).

The problem of *maximizing* a submodular function is very different. First of all, it has long been know that this problem is NP-Hard. So, practically speaking, the available approaches are heuristics (aimed at large-scale instances), approximations algorithms (aimed at gaining a theoretical understanding of what heuristics can hope to achieve in the worst case), and practical exact methods (aimed at exact optimal solution of moderate-sized instances). On the side of heuristics, there are of course available greedy and local-search methods. Indeed, practitioners have long known these methods to be useful for particular subclasses of submodular maximization problems. More recently, these heuristics have gained some theoretical credence, as approximation guarantees based on these simple search methods have been established (see, for example, [33–36, 50]). Still, the

search neighborhoods, while of constant size for a given level of performance, grows much too quickly to have practical value. Then there are general-purpose integer-programming methods available for maximizing a submodular function (see, for example, [41]), but these methods are largely untested and do not appear to be very powerful. Finally, there are specific submodular maximization problems that have been successfully attacked—meaning, we are able to solve moderate-to-large-sized instances to provable optimality. These include: (i) max-cut problem (see [45, 46]), and (ii) maximum-entropy sampling problems (see, for example, [30]).

In the remainder of Sect. 1, we describe several examples of broad classes of submodular functions relevant to discrete optimization. In Sect. 2, we give a very brief overview of the most important results concerning minimization of submodular functions. In Sect. 3, we briefly survey some results on approximation algorithms for constrained submodular maximization. In Sect. 4, we describe the known mathematical-optimization techniques for maximization of submodular functions. We touch on techniques for general submodular functions and also cut functions, and then we give a more detailed overview of such methods for the less-known entropy function. Finally, in Sect. 5, we make some brief concluding remarks.

1.1 Cuts

A key example of a submodular function comes from graphs. They can be undirected graphs, directed graphs and even hypergraphs; but in the interest of familiarity, yet still broad applicability, we will assume that G is a directed graph with vertex set E and edge set $\mathscr{E}$, non-negative capacities $c(e)$, for all edges $e \in \mathscr{E}$. We assume that G has a distinguished source vertex s and sink vertex t. For a subset S of vertices containing the source s but not containing the sink t, we let $\delta_+(S)$ denote the set of edges of G that leave S (i.e., those with tail in S and head not in S). Then we define

$$f(S) := \sum_{e \in \delta_+(S)} c(e), \quad \forall S \text{ such that } s \in S \subseteq E \setminus \{t\}.$$

That is, $f(S)$ is the sum of the edge capacities of the directed $s - t$ cut from S to $E \setminus S$. By the celebrated Max-Flow/Min-Cut Theorem, the minimum value of $f(S)$ is exactly equal to the value of a maximum $s - t$ flow in G. Algorithmically, this classical and important problem is well solved via linear programming or specialized and very efficient network-flow techniques.

On the other hand, if we instead seek to maximize f, this is a very difficult and classical NP-Hard problem with many applications. Usually, in the case of maximization, the graph is undirected and there is no distinguished source and sink. Then the function that we seek to maximize is simply

$$f(S) := \sum_{e \in \delta(S)} c(e), \quad \forall S \text{ such that } \emptyset \neq S \subsetneq E,$$

where $\delta(S)$ denotes the (undirected) edges with one end in S and the other end not in S. Some of the key works in this deeply-studied area include: see [5, 15, 25, 43, 45, 46]. In particular, a mathematical-programming based approach using a combination of semidefinite-programming and polyhedral relaxation methods seems to be the best method currently available (see [45, 46]).

1.2 Matroids

Matroids are fundamental objects in combinatorial optimization. Briefly, a *matroid* is defined on a nonempty finite *ground set* via its non-empty set $\mathscr{I}$ of *independent sets* satisfying:

(I1) $Y \subset X \in \mathscr{I} \Longrightarrow Y \in \mathscr{I}$;
(I2) $X, Y \in \mathscr{I}, |Y| > |X| \Longrightarrow \exists e \in Y \setminus X$ such that $X \cup \{e\} \in \mathscr{I}$.

The *rank function r* of M, defined by

$$r(X) := \max \{|Y| \ : \ Y \subseteq X, Y \in \mathscr{I}\},$$

is a submodular function.

With respect to the matroid M, we can define the *matroid polytope*

$$\mathscr{P} : \ \mathrm{conv}\{x(S) \ : \ S \in \mathscr{I}\},$$

where $x(S) \in \{0, 1\}^E$ denotes the incidence vector of S. The matroid polytope $\mathscr{P}$ is precisely the solution set of

 (i) $x(e) \geq 0, e \in E$;
(ii) $\sum_{e \in T} x(e) \leq r(T)$.

For any $\hat{x}$ satisfying (i), the *separation problem* of finding a set T for which (ii) is violated by $\hat{x}$ can easily be recast as a problem of minimizing a submodular function. In this way, via the ellipsoid method, we have a rather exotic algorithm to find a maximum-weight independent set in a matroid.

For $i = 1, 2$, let M_i, be a matroid on the common ground set E, having independent set $\mathscr{I}_i$ and rank function r_i. Then we have Edmonds' Matroid Intersection Duality Theorem:

$$\max \{|S| \ : \ S \in \mathscr{I}_1 \cap \mathscr{I}_1\} = \min_{S \subset E} \{r_1(S) + r_2(E \setminus S)\},$$

so the problem of finding the size of a maximum-cardinality set that is independent for a pair of matroids on a common ground set can be recast as that of minimizing a submodular function.

Coming back to polytopes, the intersection of the polytopes of two matroids on a common ground set has all of its extreme points $\{0, 1\}$-valued. So, by remarks above concerning a single matroid polytope, using submodular function minimization for

separating rank inequalities, we have we have a rather exotic algorithm to find a maximum-weight set that is independent in a pair of matroids on a common ground set.

1.3 Matchings

Another important submodular function, defined with respect to the edge set E of an undirected graph $G = (V, E)$ is as follows. Let

$$f(S) := -|S| + |\{v \in V \ : \ e \in \delta(v) \text{ for some } e \in S\}|, \quad \forall S \subseteq E.$$

That is, $f(S)$ is the number of vertices touched by edges of S minus the number of edges in S. It can be shown that if S maximizes $f(S)$, then $f(S)$ is the maximum number of edges in a matching of G. As maximum-cardinality matching is a well-solved problem (see, for example, [41]), we see that there are interesting and non-trivial well-solved examples of maximizing a submodular function.

1.4 Coverage

To define a coverage function, we assume that we have some finite universal set U. The ground set of a coverage function is a set E of labels of subsets of the universal set U. That is, every e in our ground set E is the label of a subset U_e of the universal set U. Then, the *coverage* $f(S)$ of a subset S of our ground set E is simply the number of elements of the universal set "covered" by the sets labeled from S. That is,

$$f(S) := |\cup_{e \in S} U_e|, \quad \forall S \subseteq E.$$

It is very easy to check that coverage functions are submodular. A typical goal in applications of maximizing such an f is to maximize the number of elements covered, subject to a constraint on the cost of the sets used to do the covering—that is, there is usually a cardinality constraint $|S| = k$ or a budget constraint $\sum_{e \in S} c(e) \leq \beta$. Indeed, many complicated and useful discrete optimization can often be thought of as coverage models with side constraints. So methods that can attack coverage models successfully can be imbedded in algorithms (e.g., via Dantzig-Wolfe decomposition or Lagrangian relaxation) to solve still broader problems.

1.5 Welfare

The Submodular Welfare Problem: Given m items and n players with monotone, non-negative, submodular utility functions w_i, we seek a partition of the items into disjoint sets $S_1, S_2, \ldots, S_n$ so as to maximize the total welfare $\sum_{i=1}^{n} w_i(S_i)$.

In fact, Lehmann et al. [37] demonstrated that the Submodular Welfare Problem, first studied implicitly by Fisher et al. [11], is a special case of maximizing a submodular function subject to a special side constraint (of "matroid type"). This latter problem and extensions to multiple side constraints have recently been the subject of intense study from the point of view of approximation algorithms (see, for example, [7, 33–36, 50]).

1.6 Entropy

The goal of maximum entropy sampling is to choose a most informative subset of s random variables from a set of n random variables, subject to side constraints. A typical side constraint might be a budget restriction, where we have a cost for observing each random variable. Other possibilities include logical constraints (e.g., multiple choice or precedence constraints). In many situations, we can assume that the random variables are Gaussian, or that they can be suitably transformed.

We assume that we have n Gaussian random variables

$$Y_j, \quad j \in E := \{1, 2, \ldots, n\}.$$

Our goal is to choose the 'most informative' subset S from E, having s elements, possibly subject to additional constraints

$$\sum_{j \in S} a_{ij} \le b_i, \quad i \in M := \{1, 2, \ldots, m\}.$$

We let Y_S denote the set of random variables indexed by S, and we let ϕ_S denote the joint density function of Y_S. Our measure of information, which we seek to maximize, is the *Boltzmann-Shannon entropy*

$$h(S) := -\mathbf{E}(\ln \phi_S(Y_S)).$$

We assume that the random variables have a joint Gaussian distribution, and we let C denote the covariance matrix for Y_E. Then, letting $C[S, T]$ denote the submatrix of C with rows indexed by S and columns indexed by T, we have that $C[S, S]$ is the covariance matrix of Y_S. It turns out that in the Gaussian case, the entropy $h(S)$ is just an increasing linear function of

$$H(S) := \ln \det C[S, S],$$

where det denotes determinant. So, in what follows, we refer to $H(S)$ as the entropy associated with Y_S, and it is this quantity that we seek to maximize, referring to the maximum value as H^*. It has long been know that the entropy function is submodular (see, for example, [21]), and this submodular maximization problem is know to be NP-Hard (see [22]).

The term entropy was coined by R. Clausius. Boltzmann developed the concept mathematically when he built the foundations of statistical mechanics. Shannon popularized entropy in the field of information theory. Shewry and Wynn introduced the maximum-entropy sampling problem (without side constraints), in the context of the optimal design of spatial sampling networks (see [8, 9, 18, 49, 51]).

The maximum entropy sampling problem takes on considerable importance in experimental situations in which we seek to gain information concerning potential observations at a large number of locations while observing only a few. One such situation involves configuring or reconfiguring a network of spatially disbursed environmental monitoring stations. For example, the US National Atmospheric Deposition Program/National Trends Network (NADP/NTN) is a nationwide network of precipitation monitoring stations (see [38]). Currently there are approximately 240 active stations, mostly in the continental US, with data from some of these dating back to 1978. Data is collected on the chemistry of precipitation for monitoring geographical and temporal trends. Precipitation at each station is collected weekly and analyzed for pH, hydrogen, sulfate, nitrate, ammonium, chloride, calcium, magnesium, potassium, and sodium. As a means of evaluating existing networks and assessing their configuration and possible reconfiguration, we can ask which set of stations, having some pre-specified number, provides the most information. We can formalize this by focusing on a single chemical and deriving a covariance matrix. The covariance matrix is estimated in part from historical observations, and in part by interpolation using available data with an appropriate model (see [18], for example). Aggregating weekly data over months and applying a logarithmic transformation has been found to be valuable toward meeting the assumption that the underlying random variables have a joint Gaussian distribution. Wu and Zidek [51] carried out a detailed analysis of part of the NADP/NTN using the maximum-entropy framework.

In this setting, budget constraints are quite natural, and we can also incorporate other types of logical constraints to meet potential historical or political concerns.

The (constrained) maximum-entropy sampling problem has been the subject of intense investigation from the computational side (see [1–4, 6, 19, 22, 27–30, 32]). Now, moderate-sized instances can routinely be solved to exact optimality.

1.7 Other

There are many other application areas for optimizing submodular functions. Some of the more developed ones that have significant current activity are: Machine Learning [24], Biodiversity Conservation [48], and Statics [44]. In particular, there are many applications for submodular-function optimization in machine learning (e.g., active learning, clustering, influence maximization, informative path planning, ranking, sparse experimental design, and structure learning), which in turn has many concrete applications, so the potential for significant impact is quite large.

2   Minimization

General submodular-function *minimization*, where the function is given by a black
box, is a celebrated problem in discrete optimization. The problem of minimizing
a submodular function f on the 0/1 vertices of a unit cube (the 0/1 vertices pick
out subsets of the ground set) can be modeled as one of minimizing the so-called
"Lovász extension" $\hat{f}$ on the entire unit cube. The Lovász extension is a piecewise-
linear *convex* function whose optimum is attained at a vertex of the unit cube.

Specifically, for any non-negative $\hat{x}$ in the unit cube $[0, 1]^n$, we can easily and
uniquely write

$$\hat{x} = \sum_{i=1}^{m} \lambda_i v^i,$$

where the λ_i are positive, and the v^i are 0/1-vectors satisfying $v^1 \geq v^2 \geq \cdots \geq v^m$.
Then we simply define

$$\hat{f}(x) := \sum_{i=1}^{m} \lambda_i f(v^i).$$

Properties of $\hat{f}$ including its convexity enable us to apply the ellipsoid method
to minimize it on the unit cube. This approach yielded the first polynomial-time
algorithm for the minimization of general submodular functions. Details can be
found in [17,41]. However, due to the reliance on the ellipsoid method, the resulting
algorithm is woefully impractical.

Still, the fact that, via the ellipsoid method, submodular-function minimization
was proved to be of polynomial-time complexity, gave hope for discovering
practical algorithms. Eventually this hope led to ground-breaking, polynomial-time,
combinatorial algorithms for submodular-function minimization (see [20,47]). But,
from a practical viewpoint, the situation has not greatly improved (see [14, 23] for
some progress). To this date, there is no known, practical and reliable algorithm for
minimizing a black-box submodular function. Still, the existence of very-efficient
algorithms for special cases (e.g., minimum cuts in directed graphs with non-
negative capacities) suggests that there is the possibility for practical success on
much more general classes of submodular functions.

3   Approximation Algorithms for Maximization

Besides general integer-programming techniques for submodular-function maxi-
mization, in the final section of their book, Nemhauser and Wolsey champion
approximation algorithms for submodular-function maximization (see [10, 41],
pp. 711–712). Indeed, their early work with Fisher on approximation algorithms for
submodular-function maximization was possibly the earliest work on approximation

algorithms (see [11, 40]). Recently, there has been a flurry of activity on approximation algorithms for constrained submodular-function maximization (see, for example, [33–36, 50]). A constant number of side constraints either of knapsack or matroid type can be handled. On the positive side, it has been shown that in polynomial time, we can achieve very strong approximation guarantees for submodular-function maximization, even in the presence of side constraints, using very simple local-search mechanisms. This is great theory, but the approximation guarantees are not good enough for the algorithms to be viewed as truly practical. Still, these simple local-search methods can serve as the springboard to more powerful, practical methods (e.g., compare this with the relationship between simple k-opt edge swapping and the more practically powerful Lin-Kernighan heuristic for the TSP).

4 Mathematical Optimization for Maximization

4.1 General Techniques

We have already pointed out how the maximum-cardinality matching problem is an important special case of submodular-function maximization (see Sect. 1.3). On the other side of the spectrum, we have a general-purpose integer-programming approach to submodular-function maximization (see [41], pp. 710–711).

Specifically, we can assume that $f(\emptyset) = 0$, by subtracting the constant $f(\emptyset)$ from $f(S)$, $\forall S$. Let $E := \{1, 2, \ldots, n\}$. We can let $f^*(S) := f(S) - \sum_{j=1}^{n} [f(E \setminus \{j\}) - f(E)]$, $\forall S$. Again, we have just shifted f to get f^*, but in this way the equivalent submodular function f^* is non-decreasing. The submodular maximization problem can then be recast as the following mixed-integer linear program:

$$\max \eta$$
$$\eta - \sum_{j \in E \setminus S} \left[f^*(S \cup \{j\}) - f^*(S) \right] x_j$$
$$+ \sum_{j=1}^{n} [f(E \setminus \{j\}) - f(E)] x_j \leq f^*(S), \quad \forall S \subseteq E \, ;$$
$$\eta \in \mathbb{R}, \ x_j \in \{0, 1\}^n, \quad j = 1, 2, \ldots, n.$$

To our knowledge, an algorithm based on such a formulation has never been implemented. Indeed, such an algorithm may not be practical at all due to the reliance on an exponential number of (what appear to be weak) inequalities with no obvious efficient technique for separating them.

4.2 Max-Cut

The Max-Cut problem (see Sect. 1.1) is one of the shining examples of success
for integer-programming methods for submodular-function maximization. (see
[5, 15, 25, 43, 45, 46]). Some of the useful methodology uses classical *linear* integer-
programming techniques, but some of the exciting and newer work models the
Max-Cut problem as an integer program with quadratic functions in 0/1 variables.
The natural analysis leads to an important role for semidefinite-programming
relaxations and spectral methods. Interestingly, some of the most useful bound-
ing techniques for maximum-entropy sampling, another important success story
for submodular-function maximization, also makes use of spectral methods (see
Sect. 4.3).

4.3 Max-Entropy Sampling

The (constrained) maximum-entropy sampling problem is a beautiful example of
a broadly-applicable, NP-Hard, submodular maximization problem (see Sect. 1.6)
that does not have any obvious linear or quadratic formulation as an integer
program. This is in sharp contrast to many integer programming problems defined
on graphs, where edge variables can be related in a quadratic relationship with
vertex variables (e.g., cut, packing, assignment and routing problems). Still, there
has been enormous progress on this problem over the last 20 years, and a wide
variety of optimization techniques have been fruitfully employed, and often in a non-
straightforward manner: e.g., branch-and-bound, Lagrangian relaxation, convex
relaxation, dynamic programming, integer-linear programming, spectral methods,
etc. (see [1–4, 6, 19, 22, 27–30, 32]). In particular, branching, in the context of
branch-and-bound, is done is a subtle manner: Excluding a random variable from
observation is easy, as we just strike out the corresponding row and column
of the covariance matrix, but for forcing a random variable to be observed, we
compute a *conditional* covariance matrix. There are many different upper-bounding
techniques, which can be utilized within branch-and-bound.

A technique that can improve any upper-bound method is to exploit the matrix
identity

$$\ln \det C[S, S] = \ln \det C + \ln \det C^{-1}[E \setminus S, E \setminus S].$$

This identity says that the entropy of any set S of random variables to be observed
can be thought of as the sum of the entropy of the entire set (of random variables
under consideration) and the "inverse entropy" of the set of random variables that
we do not observe. Then bounding techniques applied to choosing $n - s$ variables
with respect to the "inverse entropy" translate to new bounds for choosing s random
variables with respect to the ordinary entropy.

Some of the best upper-bounding methodologies for entropy are based on eigenvalues of matrices derived from the covariance matrix (see [6, 19, 22, 27–30, 32]). A *mask* is a (symmetric) $X \succeq 0$ having $\mathrm{diag}(X) = e$ (i.e., a point in the "elliptope"). Anstreicher and Lee established

Theorem 1 ([1]).

$$H^* \leq \xi_{C,s}(X) := \sum_{l=1}^{s} \ln \left(\lambda_l \left(C \circ X \right) \right).$$

The proof is based on "Oppenheim's Inequality" (see [42]). Combinatorial (see [19]) and continuous (see [1, 6]) local-optimization techniques have been developed to attempt to find good masks for this *masked spectral bound*.

For constrained problems, it is vital that upper-bounding methods exploit the constraints, otherwise computed bounds will be too weak to be effective. For ease of exposition, we will demonstrate how to exploit side constraints to improve the spectral bound above, when the mask X is fixed as the all-ones matrix,

Theorem 2 ([28]).

$$H^* \leq \min_{\pi \in \mathbb{R}_+^m} v(\pi)$$

where

$$v(\pi) := \left\{ \sum_{l=1}^{s} \ln \lambda_l \left(D^\pi \ C \ D^\pi \right) + \sum_{i \in M} \pi_i b_i \right\},$$

and D^π is the diagonal matrix having

$$D_{jj}^\pi := \exp \left\{ -\frac{1}{2} \sum_{i \in M} \pi_i a_{ij} \right\}$$

Moreover, v_π is convex (in π), and v_π is analytic when

$$\lambda_s \left(D^\pi \ C \ D^\pi \right) > \lambda_{s+1} \left(D^\pi \ C \ D^\pi \right).$$

Next, let x^l be the eigenvector (of unit Euclidean norm) associated with λ_l. Define the *continuous solution* $\tilde{x} \in \mathbb{R}^E$ by $\tilde{x}_j := \sum_{l=1}^{s} \left(x_j^l \right)^2$, for $j \in E$. Define $\gamma \in \mathbb{R}^M$ by $\gamma_i := b_i - \sum_{j \in N} a_{ij} \tilde{x}_j$.

Theorem 3 ([28]). *If $\lambda_s > \lambda_{s+1}$, then γ is the gradient of f at π.*

These theorems give all of the ingredients for an effective method for quickly finding a good $\pi \in \mathbb{R}_+^m$.

Other useful bounding approaches use continuous relaxation (see [2, 3]). It is straightforward to check that the following formulation is a continuous relaxation of the constrained maximum entropy sampling problem.

$$\max f(x) := \ln \det\left(\text{Diag}(x_j^{p_j})\, C\, \text{Diag}(x_j^{p_j}) + \text{Diag}(d_j^{x_j} - d_j x_j^{p_j}) \right)$$

$$\text{subject to } \sum_{j \in E} a_{ij} x_j \leq b_i, \forall i;$$

$$\sum_{j \in E} x_j = s;$$

$$0 \leq x_j \leq 1, \forall j,$$

where the constants $d_j > 0$ and $p_j \geq 1$ satisfy $d_j \leq \exp(p_j - \sqrt{p_j})$, and $\text{Diag}(d_j) - C[E, E] \succeq 0$.

Moreover, the relaxation is tractable, due to the following result.

Theorem 4 ([3]). *Assume that* $D \succeq C$, $p_j \geq 1$, $0 < d_j \leq \exp(p_j - \sqrt{p_j})$. *Then* f *is concave for* $0 < x \leq e$.

There are a few other techniques that have been exploited for upper bounding; for example, dynamic programming, matching and integer-linear programming (see [32]).

Although the state-of-the-art is that moderate-sized instances of maximum-entropy sampling can be routinely solved to optimality, there is considerable room for greater success on large-scale instances.

5   Conclusion

We have seen remarkable progress in the state-of-the-art for submodular function optimization in the last couple of decades. In particular, on the NP-Hard maximization side, we have new approximation algorithms, with provable performance guarantees, based mostly on local search. On the mathematical–programming side, we now have good algorithms, for special cases like max-cut and maximum entropy sampling, capable of handling ever larger instances. In many cases where we can calculate exact optima, we find that these same solutions are found by simple heuristics. For example, in the case of entropy, Le and Zidek have said:

> "Therefore, the finding of [22] is encouraging in that the greedy/swap algorithm described above often produced the exact optimum, where the latter is computable." (see [26, p. 201]).

Therefore, our conclusion is that when possible, branch-and-bound methods should be used to calculate exact optima for submodular maximization problems. When not

possible, one can use greedy and local-search algorithms with a reasonable degree of comfort. As Bob Marley crooned:

> "Don't worry about a thing, oh no! 'Cause every little thing gonna be all right!" (Bob Marley & The Wailers, "Three little birds", 1977).

References

1. Anstreicher, K.M., Lee, J.: A masked spectral bound for maximum-entropy sampling. In: MODA 7 – Advances in Model-Oriented Design and Analysis. Contributions to Statistics, pp. 1–12. Springer, Berlin (2004)
2. Anstreicher, K.M., Fampa, M., Lee, J., Williams, J.D.: Continuous relaxations for constrained maximum-entropy sampling. In: Integer Programming and Combinatorial Optimization, Vancouver. Lecture Notes in Computer Science, vol. 1084, pp. 234–248. Springer, Berlin (1996)
3. Anstreicher, K.M., Fampa, M., Lee, J., Williams, J.D.: Using continuous nonlinear relaxations to solve constrained maximum-entropy sampling problems. Math. Program. A **85**, 221–240 (1999)
4. Anstreicher, K.M., Fampa, M., Lee, J., Williams, J.D.: Maximum-entropy remote sampling. Discret. Appl. Math. **108**, 211–226 (2001)
5. Barahona, F., Grötschel, M., Jünger, M., Reinelt, G.: An application of combinatorial optimization to statistical physics and circuit layout design. Oper. Res. **36**(3), 493–513 (1988)
6. Burer, S., Lee, J.: Solving maximum-entropy sampling problems using factored masks. Math. Program. **109**(2–3), 263–281 (2007)
7. Calinescu, G., Chekuri, C., Pál, M., Vondrák, J.: Maximizing a submodular set function subject to a matroid constraint. In: Proceedings of 12th IPCO, Ithaca, pp. 182–196 (2007)
8. Caselton, W.F., Zidek, J.V.: Optimal monitoring network designs. Stat. Probab. Lett. **2**, 223–227 (1984)
9. Caselton, W.F., Kan, L., Zidek, J.V.: Quality data networks that minimize entropy. In: Walden, A.T., Guttorp, P. (eds.) Statistics in the Environmental and Earth Sciences, pp. 10–38. Arnold, London (1992)
10. D'Ambrosio, C., Lee, J., Wächter, A.: An algorithmic framework for MINLP with separable non-convexity. In: Leyffer, S., Lee, J. (eds.) Mixed Integer Nonlinear Programming. The IMA Volumes in Mathematics and Its Applications, vol. 154, pp.315–347. Springer, New York (2012)
11. Fisher, M.L., Nemhauser, G.L., Wolsey, L.A.: An analysis of approximations for maximizing submodular set functions II. Math. Program. Study **8**, 73–87 (1978)
12. Frank, A.: Applications of submodular functions. In: Walker, K. (ed.) Surveys in Combinatorics. London Mathematical Society, Lecture Note Series, vol. 187, pp.850–136. Cambridge University Press, Cambridge (1993)
13. Fujishige, S.: Submodular functions and optimization. Annals of Discrete Mathematics, vol. 58, 2nd edn. Elsevier, Amsterdam (2005)
14. Fujishige, S., Isotani, S.: A submodular function minimization algorithm based on the minimum-norm base. Pac. J. Optim. **7**, 3–17 (2011)
15. Goemans, M.X., Williamson, D.P.: Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. J. ACM **42**(6), 1115–1145 (1995)
16. Grötschel, M., Lovász, L., Schrijver, A.: The ellipsoid method and its consequence in combinatorial optimization. Combinatorica **1**(2), 169–197 (1981)

17. Grötschel, M., Lovász, L., Schrijver, A.: Geometric Algorithms and Combinatorial Optimization. Springer, Berlin (1993)

18. Guttorp, P., Le, N.D., Sampson, P.D., Zidek, J.V.: Using Entropy in the redesign of an environmental monitoring network. In: Patil, G.P., Rao, C.R. (eds.) Multivariate Environmental Statistics, pp. 175–202. North Holland, New York (1993)

19. Hoffman, A.J., Lee, J., Williams, J.D.: New upper bounds for maximum-entropy sampling. In: Atkinson, A.C., Hackl, P., Müller, W.G. (eds.) MODA 6 – Advances in Model-Oriented Design and Analysis. Contributions to Statistics, pp. 143–153. Springer, Berlin (2001)

20. Iwata, S., Fleischer, L., Fujishige, S.: A combinatorial, strongly polynomial-time algorithm for minimizing submodular functions. In: Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing, Portland, pp. 97–106 (electronic). ACM, New York (2000)

21. Kelmans, A.K., Kimelfeld, B.N.: Multiplicative submodularity of a matrix?s principal minor as a function of the set of its rows and some combinatorial applications. Discret. Math. **44**(1), 113–116 (1980)

22. Ko, C.-W., Lee, J., Queyranne, M.: An exact algorithm for maximum entropy sampling. Oper. Res. **43**(4), 684–691 (1995)

23. Krause, A.: SFO: a toolbox for submodular function optimization. J. Mach. Learn. Res. **11**, 1141–1144 (2010)

24. Krause, A., Guestrin, C.: Beyond convexity: submodularity in machine learning (2013). http://submodularity.org/

25. Laurent, M.: Semidefinite relaxations for max-cut. In: The Sharpest Cut. MPS/SIAM Series on Optimization, pp. 257–290. SIAM, Philadelphia (2004)

26. Le, N.D., Zidek, J.V.: Statistical analysis of environmental space-time processes. Springer Series in Statistics. Springer, New York (2006)

27. Lee, J.: Discussion on: 'A state-space-model approach to optimal spatial sampling design based on entropy'. Environ. Ecol. Stat. **5**, 45–46 (1998)

28. Lee, J.: Constrained maximum-entropy sampling. Oper. Res. **46**, 655–664 (1998)

29. Lee, J.: Semidefinite programming in experimental design. In: Wolkowicz, H., Saigal, R., Vandenberghe, L. (eds.) Handbook of Semidefinite Programming. International Series in Operations Research and Management Science, vol. 27. Kluwer, Boston (2000)

30. Lee, J.: Maximum entropy sampling. In: El-Shaarawi, A.H., Piegorsch, W.W. (eds.) Encyclopedia of Environmetrics, 2nd edn., vol. 3, pp. 1570–1574. Wiley, Chichester (2012)

31. Lee, J.: A First Course in Combinatorial Optimization. Cambridge University Press, Cambridge/New York (2004)

32. Lee, J., Williams, J.: A linear integer programming bound for maximum-entropy sampling. Math. Program. B **94**, 247–256 (2003)

33. Lee, J., Sviridenko, M., Vondrák, J.: Submodular maximization over multiple matroids via generalized exchange properties. In: Dinur, I., Jansen, K., Naor, S., Rolim, J.D.P. (eds.) Proceedings of APPROX 2009, 12th International Workshop, Berkeley, 21–23 Aug 2009. Lecture Notes in Computer Science, vol. 5687, pp. 244–257. Springer (2009)

34. Lee, J., Mirrokni, V.S., Nagarajan, V., Sviridenko, M.: Non-monotone submodular maximization under matroid and knapsack constraints. In: Proceedings of the 41st Annual ACM Symposium on Theory of Computing STOC 2009, Bethesda, pp. 323–332 (2009)

35. Lee, J., Sviridenko, M., Vondrák, J.: Approximate maximization of a submodular function over multiple matroids via generalized exchange properties. Math. Oper. Res. **35**(4), 795–806 (2010)

36. Lee, J., Mirrokni, V.S., Nagarajan, V., Sviridenko, M.: Maximizing non-monotone submodular functions under matroid and knapsack constraints. SIAM J. Discret. Math. **23**(4), 2053–2078 (2010)

37. Lehmann, B., Lehmann, D.J., Nisan, N.: Combinatorial auctions with decreasing marginal utilities (journal version). Games Econ. Behav. **55**, 270–296 (2006)

38. NADP/NTN website (2013). http://nadp.sws.uiuc.edu

39. Narayanan, H.: Submodular functions and electrical networks. Annals of Discrete Mathematics, vol. 54. North-Holland, Amsterdam (1997)
40. Nemhauser, G.L., Wolsey, L.A., Fisher, M.: An analysis of approximations for maximizing submodular set functions. I. Math. Program. **14**(3), 265–294 (1978)
41. Nemhauser, G.L., Wolsey, L.A., Laurence A.: Integer and Combinatorial Optimization. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley, New York (1988)
42. Oppenheim, A.: Inequalities connected with definite Hermitian forms. J. Lond. Math. Soc. **5**, 114–119 (1930)
43. Poljak, S., Rendl, F.: Solving the max-cut problem using eigenvalues. Special volume on partitioning and decomposition in combinatorial optimization. Discret. Appl. Math. **62**(1–3), 249–278 (1995)
44. Recski, A.: Matroid theory and its applications in electric network theory and in statics. Algorithms and Combinatorics, vol. 6. Springer, Berlin (1989)
45. Rendl, F., Rinaldi, G., Wiegele, A.: Solving max-cut to optimality by intersecting semidefinite and polyhedral relaxations. Math. Program. A **121** (2), 307–335 (2010)
46. Rendl, F., Rinaldi, G., Wiegele, A.: Biq mac solver – binary quadratic and max cut solver (2013). http://biqmac.uni-klu.ac.at/
47. Schrijver, A.: A combinatorial algorithm minimizing submodular functions in strongly polynomial time. J. Comb. Theory B **80**(2), 346–355 (2000)
48. Semple, C.: Submodular functions and biodiversity conservation (2010). http://www-theory.phys.utas.edu.au/phylomania2010/website/2010/slides/slides/Semplehobart10.pdf
49. Shewry, M.C., Wynn, H.P.: Maximum entropy sampling. J. Appl. Stat. **14**, 165–170 (1987)
50. Sviridenko, M.: A note on maximizing a submodular set function subject to knapsack constraint. Oper. Res. Lett. **32**, 41–43 (2004)
51. Wu, S., Zidek, J.V.: An entropy based review of selected NADP/NTN network sites. Atmos. Environ. **26A**, 2089–2103 (1992)

A Further Generalization of the Colourful Carathéodory Theorem

Frédéric Meunier and Antoine Deza

Abstract Given $d+1$ sets, or colours, $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{d+1}$ of points in $\mathbb{R}^d$, a *colourful set* is a set $S \subseteq \bigcup_i \mathbf{S}_i$ such that $|S \cap \mathbf{S}_i| \leq 1$ for $i = 1, \ldots, d+1$. The convex hull of a colourful set S is called a *colourful simplex*. Bárány's colourful Carathéodory theorem asserts that if the origin $\mathbf{0}$ is contained in the convex hull of $\mathbf{S}_i$ for $i = 1, \ldots, d+1$, then there exists a colourful simplex containing $\mathbf{0}$. The sufficient condition for the existence of a colourful simplex containing $\mathbf{0}$ was generalized to $\mathbf{0}$ being contained in the convex hull of $\mathbf{S}_i \cup \mathbf{S}_j$ for $1 \leq i < j \leq d+1$ by Arocha et al. and by Holmsen et al. We further generalize the sufficient condition and obtain new colourful Carathéodory theorems. We also give an algorithm to find a colourful simplex containing $\mathbf{0}$ under the generalized condition. In the plane an alternative, and more general, proof using graphs is given. In addition, we observe that any condition implying the existence of a colourful simplex containing $\mathbf{0}$ actually implies the existence of $\min_i |\mathbf{S}_i|$ such simplices.

Key words Colourful Carathéodory theorem • Colourful simplicial depth • Discrete geometry

Subject Classifications: 52C45, 52A35

F. Meunier
Université Paris Est, CERMICS, 6–8 Avenue Blaise Pascal, Cité Descartes,
77455 Marne-la-Vallée Cedex 2, France
e-mail: frederic.meunier@cermics.enpc.fr

A. Deza (✉)
Advanced Optimization Laboratory, Department of Computing and Software, McMaster
University, 1280 Main Street West, Hamilton, ON L8S 4K1, Canada

Equipe Combinatoire et Optimisation, Université Pierre et Marie Curie, Paris, France
e-mail: deza@mcmaster.ca

1   Colourful Carathéodory Theorems

Given $d + 1$ sets, or colours, $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{d+1}$ of points in $\mathbb{R}^d$, we call a set of points drawn from the $\mathbf{S}_i$'s *colourful* if it contains at most one point from each $\mathbf{S}_i$. A *colourful simplex* is the convex hull of a colourful set S, and a colourful set of d points which misses $\mathbf{S}_i$ is called an $\hat{i}$-*transversal*. The colourful Carathéodory Theorem 1 by Bárány provides a sufficient condition for the existence of a colourful simplex containing the origin $\mathbf{0}$.

Theorem 1 ([2]). *Let $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{d+1}$ be finite sets of points in $\mathbb{R}^d$ such that $\mathbf{0} \in \mathrm{conv}(\mathbf{S}_i)$ for $i = 1 \ldots d+1$. Then there exists a set $S \subseteq \bigcup_i \mathbf{S}_i$ such that $|S \cap \mathbf{S}_i| = 1$ for $i = 1, \ldots, d + 1$ and $\mathbf{0} \in \mathrm{conv}(S)$.*

Theorem 1 was generalized by Arocha et al. [1] and by Holmsen et al. [11] providing a more general sufficient condition for the existence of a colourful simplex containing the origin $\mathbf{0}$.

Theorem 2 ([1, 11]). *Let $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{d+1}$ be finite sets of points in $\mathbb{R}^d$ such that $\mathbf{0} \in \mathrm{conv}(\mathbf{S}_i \cup \mathbf{S}_j)$ for $1 \le i < j \le d + 1$. Then there exists a set $S \subseteq \bigcup_i \mathbf{S}_i$ such that $|S \cap \mathbf{S}_i| = 1$ for $i = 1, \ldots, d + 1$ and $\mathbf{0} \in \mathrm{conv}(S)$.*

We further generalize the sufficient condition for the existence of a colourful simplex containing the origin. Moreover, the proof, given in Sect. 2.1, provides an alternative and geometric proof for Theorem 2. Let $\overrightarrow{x_k \mathbf{0}}$ denote the ray originating from x_k towards $\mathbf{0}$.
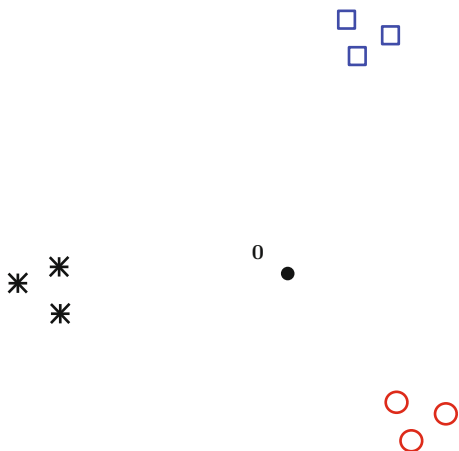
Theorem 3. *Let $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{d+1}$ be finite sets of points in $\mathbb{R}^d$. Assume that, for each $1 \le i < j \le d + 1$, there exists $k \notin \{i, j\}$ such that, for all $x_k \in \mathbf{S}_k$, the convex hull of $\mathbf{S}_i \cup \mathbf{S}_j$ intersects the ray $\overrightarrow{x_k \mathbf{0}}$ in a point distinct from x_k. Then there exists a set $S \subseteq \bigcup_i \mathbf{S}_i$ such that $|S \cap \mathbf{S}_i| = 1$ for $i = 1, \ldots, d + 1$ and $\mathbf{0} \in \mathrm{conv}(S)$.*

Under the general position assumption, Theorem 3 can be derived from the slightly stronger Theorem 4 where $H^+(T_i)$ denotes, for any $\hat{i}$-transversal T_i, the open half-space defined by $\mathrm{aff}(T_i)$ and containing $\mathbf{0}$.

Theorem 4. *Let $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{d+1}$ be finite sets of points in $\mathbb{R}^d$ such that the points in $\bigcup_i \mathbf{S}_i \cup \{\mathbf{0}\}$ are distinct and in general position. Assume that, for any $i \ne j$, $(\mathbf{S}_i \cup \mathbf{S}_j) \cap H^+(T_j) \ne \emptyset$ for any $\hat{j}$-transversal T_j. Then there exists a set $S \subseteq \bigcup_i \mathbf{S}_i$ such that $|S \cap \mathbf{S}_i| = 1$ for $i = 1, \ldots, d + 1$ and $\mathbf{0} \in \mathrm{conv}(S)$.*

Note that, as the conditions of Theorems 1 and 2, but unlike the one of Theorem 4, the condition of Theorem 3 is computationally easy to check. Indeed, testing whether a ray intersects the convex hull of a finite number of points amounts to solve a linear optimization feasibility problem which is polynomial-time solvable.

Fig. 1 A set in dimension 2 satisfying the condition of Theorem 3 but not the one of Theorem 2

In the plane and assuming general position, Theorem 3 can be generalized to Theorem 5. The proofs of Theorems 3, 4, and 5 are given in Sect. 2.

Theorem 5. *Let $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3$ be finite sets of points in $\mathbb{R}^2$ such that the points in $\mathbf{S}_1 \cup \mathbf{S}_2 \cup \mathbf{S}_3 \cup \{\mathbf{0}\}$ are distinct and in general position. Assume that, for pairwise distinct $i, j, k \in \{1, 2, 3\}$, the convex hull of $\mathbf{S}_i \cup \mathbf{S}_j$ intersects the line $\mathrm{aff}(x_k, \mathbf{0})$ for all $x_k \in \mathbf{S}_k$. Then there exists a set $S \subseteq \mathbf{S}_1 \cup \mathbf{S}_2 \cup \mathbf{S}_3$ such that $|S \cap \mathbf{S}_i| = 1$ for $i = 1, 2, 3$ and $\mathbf{0} \in \mathrm{conv}(S)$.*

Figures 1 and 2 illustrate sets satisfying the condition of Theorem 3 but not the ones of Theorems 1 and 2. Let $S_d^{\triangle}$ denote the d-dimensional configuration where the points in $\mathbf{S}_i$ are clustered around the ith vertex of a simplex containing $\mathbf{0}$, see Fig. 1 for an illustration of $S_2^{\triangle}$. While all the $(d+1)^{d+1}$ colourful simplices of this configuration contain $\mathbf{0}$, $S_{d \geq 3}^{\triangle}$ does not satisfy the conditions of Theorems 1, 2, or 3, but satisfies the one of Theorem 4. While the set given in Fig. 3 satisfies the condition of Theorem 4, it does not satisfy the condition of Theorem 3 for $i = \square$ and $j = \bullet$. Figure 4 illustrates a set satisfying the condition of Theorem 5 but not the one of Theorem 4.

One can check that Theorem 4 is still valid if the general position assumption is replaced by: *there is at least one transversal T such that $\mathbf{0} \notin \mathrm{aff}(T)$ and such that the points of T are affinely independent.* However, we are not aware of an obvious way to handle, via Theorem 4, configurations where all points and the origin lie in the same hyperplane. Note that Theorem 3 can be applied to such degenerate configurations. See Sect. 2.2 for a proof of Theorem 3 and a configuration which illustrates the gap between Theorem 3 and its general position version, and justifies the specific treatment for the degenerate cases.

Fig. 2 A degenerate set in dimension 3 satisfying the condition of Theorem 3 but not the one of Theorem 2
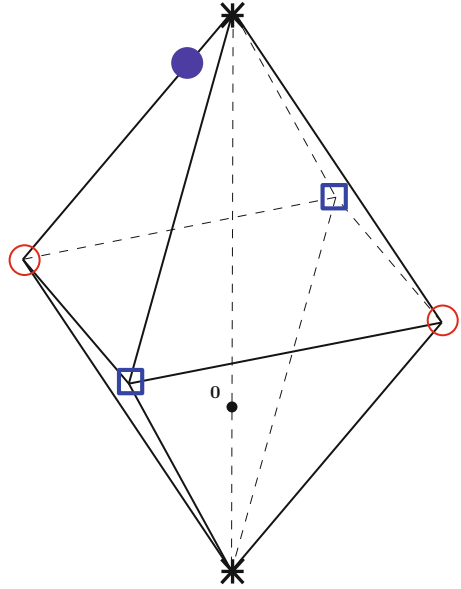

Fig. 3 A degenerate set in dimension 3 satisfying, up to a slight perturbation, the condition of Theorem 4 but not the one of Theorem 3
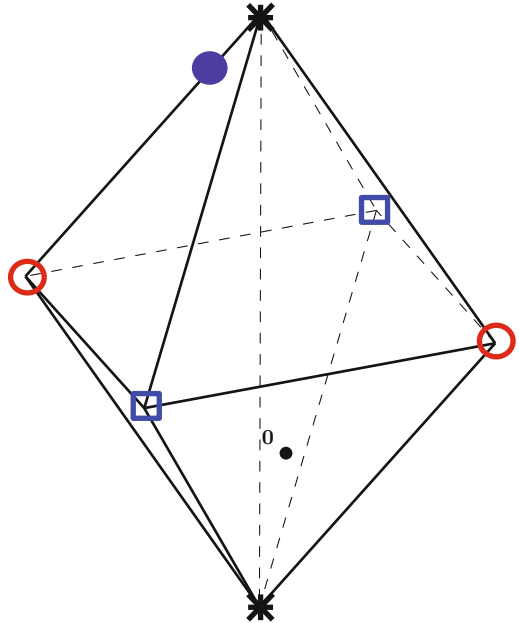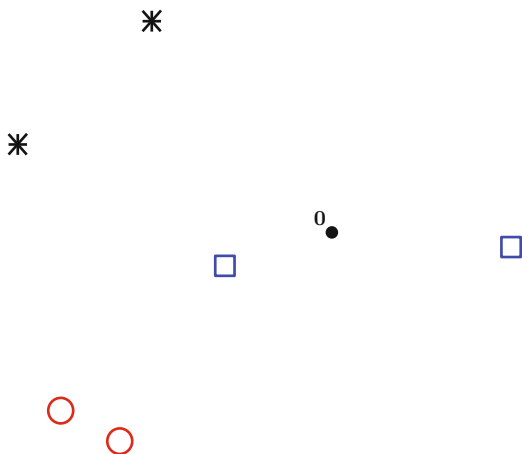
Fig. 4 A set in dimension 2
satisfying the condition of
Theorem 5 but not the one of
Theorem 4

2 Proofs

2.1 Proof of Theorem 4

We recall that a k-simplex σ is the convex hull of $(k+1)$ affinely independent points.
An *abstract simplicial complex* is a family $\mathscr{F}$ of subsets of a finite ground set such
that whenever $F \in \mathscr{F}$ and $G \subseteq F$, then $G \in \mathscr{F}$. These subsets are called *abstract*
simplices. The *dimension* of an abstract simplex is its cardinality minus one. The
dimension of a simplicial complex is the dimension of largest simplices. A *pure*
abstract simplicial complex is a simplicial complex whose maximal simplices have
all the same dimension. A *combinatorial d-pseudomanifold* $\mathscr{M}$ is a pure abstract d-
dimensional simplicial complex such that any abstract $(d-1)$-simplex is contained
in exactly two abstract d-simplices.

Consider a ray $\mathbf{r}$ originating from $\mathbf{0}$ and intersecting at least one colourful $(d-1)$-
simplex. Under the general position assumption for points in $\bigcup_i \mathbf{S}_i \cup \{\mathbf{0}\}$, one can
choose $\mathbf{r}$ such that it intersects the interior of the colourful $(d-1)$-simplex, and
that no two colourful simplices have the same intersection with $\mathbf{r}$. Let σ be the *first*
colourful $(d-1)$-simplex intersected by $\mathbf{r}$. Note that, given $\mathbf{r}$, σ is uniquely defined.
Without loss of generality, we can assume that the vertices of σ form the $\widehat{d+1}$-
transversal $\{v_1, \dots, v_d\}$.

Setting $j = d+1$, and $T_{d+1} = \{v_1, \dots, v_d\}$ in Theorem 4 gives $(\mathbf{S}_i \cup \mathbf{S}_{d+1}) \cap H^+(T_{d+1}) \neq \emptyset$. In other words, there is, for each i, a point either in
$\mathbf{S}_{d+1} \cap H^+(T_{d+1})$ or in $(\mathbf{S}_i \setminus \{v_i\}) \cap H^+(T_{d+1})$. Assume first that for one i
the corresponding point belongs to $\mathbf{S}_{d+1}$, and name it v'_{d+1}. Then $\mathbf{r}$ intersects the
boundary of $\mathrm{conv}(v_1, \dots, v_d, v'_{d+1})$ in only one point as otherwise $\mathbf{r}$ would intersect
another colourful $(d-1)$-simplex before intersecting σ. Indeed, $\mathbf{r}$ leaves $H^+(T_{d+1})$
after intersecting σ. Thus, $\mathbf{r}$ intersects $\mathrm{conv}(v_1, \dots, v_d, v'_{d+1})$ in exactly one point;

that is, $\mathbf{0} \in \mathrm{conv}(v_1, \ldots, v_d, v'_{d+1})$. Therefore, we can assume that for each i there is a point $v'_i \neq v_i$ in $\mathbf{S}_i \cap H^+(T_{d+1})$, and consider the $\widehat{d+1}$-transversal $T' = \{v'_1, \ldots, v'_d\}$ and the associated colourful $(d-1)$-simplex $\sigma' = \mathrm{conv}(v'_1, \ldots, v'_d)$. Let $\mathscr{M}$ be the abstract simplicial complex defined by

$$\mathscr{M} = \{F \cup F' : F \subseteq V(\sigma), \ F' \subseteq V(\sigma') \text{ and } c(F) \cap c(F') = \emptyset\}$$

where $V(\sigma)$ denotes the vertex set of σ and $c(x) = i$ for $x \in \mathbf{S}_i$. The simplicial complex $\mathscr{M}$ is a combinatorial $(d-1)$-pseudomanifold. Note that $V(\sigma)$ and $V(\sigma')$ are abstract simplices of $\mathscr{M}$. Let M be the collection of the convex hulls of the abstract simplices of $\mathscr{M}$. Note that the vertices of all maximal simplices of M form $\widehat{d+1}$-transversals and that $\mathscr{M}$ is not necessarily a simplicial complex in the geometric meaning as some pairs of geometric $(d-1)$-simplices might have intersecting interiors.

We recall that for any generic ray originating from $\mathbf{0}$, the parity of the number of times its intersects M is the same. We remark that this number can not be even as, otherwise, we would have a colourful $(d-1)$-simplex closer to $\mathbf{0}$ than σ on $\mathbf{r}$ since, M being contained in the closure of $H^+(T_{d+1})$, when $\mathbf{r}$ intersects σ, it is the last intersection. Thus, the number of times $\mathbf{r}$ intersects M is odd, and actually equal to 1. Take now any point $v \in \mathbf{S}_{d+1}$ and consider the ray originating from $\mathbf{0}$ towards the direction opposite to v. This ray intersects M in a colourful $(d-1)$-simplex τ; that is, $\mathbf{0} \in \mathrm{conv}(\tau \cup \{v\})$. $\square$

One can check that the proof of Theorem 4 still works if there is at least one transversal T such that $\mathbf{0} \notin \mathrm{aff}(T)$ and such that the points of T are affinely independent. Indeed, in that case, we can always choose a ray $\mathbf{r}$ such that, for any pair (T, T') of transversals, $\mathbf{r} \cap \mathrm{aff}(T) = \mathbf{r} \cap \mathrm{aff}(T')$ if and only if $\mathrm{aff}(T) = \mathrm{aff}(T')$.

Remark 1. The topological argument that the parity of the number of times a ray originating from $\mathbf{0}$ intersects M depends only on the respective positions of $\mathbf{0}$ and M can be replaced by Proposition 1 as used in the description of the algorithm given in Sect. 3.3. In other words, we get a geometric proof of Theorem 4.

Assuming $\bigcup_i \mathbf{S}_i$ lies on the sphere $\mathbb{S}^{d-1}$, the $\hat{i}$-transversals generate full dimensional colourful cones pointed at $\mathbf{0}$. We say that a transversal *covers* a point if the point is contained in the associated cone. Colourful simplices containing $\mathbf{0}$ are generated whenever the antipode of a point of colour i is covered by an $\hat{i}$-transversal. In particular, one can consider combinatorial *octahedra* generated by pairs of disjoint $\hat{i}$-transversals, and rely on the fact that every octahedron Ω either covers all of $\mathbb{S}^{d-1}$ with colourful cones, or every point $x \in \mathbb{S}^{d-1}$ that is covered by colourful cones from Ω is covered by at least two distinct such cones, see for example the *Octahedron Lemma* of [3]. One of the key argument in the proof of Theorem 4 can be reformulated as: either the pair of $\widehat{d+1}$-transversals (T, T') forms a octahedron covering $\mathbb{S}^{d-1}$, or $\mathbf{0}$ belongs to a colourful simplex having $\mathrm{conv}(T)$ as a facet.

2.2 Proof of Theorem 3

Consider a configuration satisfying the conditions of Theorem 3 and with $\bigcup_i \mathbf{S}_i \cup \{\mathbf{0}\}$ distinct and in general position. Consider $i \neq j$ and a $\hat{j}$-transversal T_j, then there is $x_k \in \mathbf{S}_k \cap T_j$ such that the $\overrightarrow{x_k \mathbf{0}}$ intersect the convex hull of $\mathbf{S}_i \cup \mathbf{S}_j$ in a point in $H^+(T_j)$, and therefore at least one point of $\mathbf{S}_i \cup \mathbf{S}_j$ belongs to $H^+(T_j)$.

Let consider degenerate configurations and let a denote the maximum cardinality of an affinely independent colourful set whose affine hull does not contain $\mathbf{0}$.

If $a = d$, there is at least one transversal T such that $\mathbf{0} \notin \text{aff}(T)$ and such that the points of T are affinely independent. Therefore we can use the stronger version of Theorem 4 relying on the existence of such a transversal T.

Assume that $a < d$. We can choose a ray $\mathbf{r}$ such that the non-empty intersections with $\text{aff}(A)$ for all colourful sets A of cardinality a are distinct. Let A^0 be an affinely independent colourful set of cardinality a such that $\text{aff}(A^0)$ is the first intersected by $\mathbf{r}$. Without loss of generality, let $A^0 = \{v_1, \ldots, v_a\}$ with $v_s \in \mathbf{S}_s$. Note that $\mathbf{S}_{a+1} \cup \ldots \cup \mathbf{S}_{d+1} \subset \text{aff}(A^0 \cup \{\mathbf{0}\})$ as otherwise $\mathbf{0} \notin \text{aff}(A^0 \cup \{v_j\})$ for $v_j \in \mathbf{S}_j$ with $j > a$ which contradicts the maximality of a. If there is a colourful simplex containing $\mathbf{0}$, we are done. Therefore, we can assume that, in $\text{aff}(A^0 \cup \{\mathbf{0}\})$, we have an open half-space defined by $\text{aff}(A^0)$ containing $\mathbf{0}$ but not $\mathbf{S}_{a+1} \cup \ldots \cup \mathbf{S}_{d+1}$, and will derive a contradiction.

Let $B_0 = \{a+1, \ldots, d+1\}$. We remark that, for all $i, j \in B_0$ with $i \neq j$, the k, such that $\text{conv}(\mathbf{S}_i \cup \mathbf{S}_j)$ intersects $\overrightarrow{x_k \mathbf{0}}$ in a point distinct from x_k, satisfies $k \in B_0$ since $\mathbf{S}_i \cup \mathbf{S}_j$ are separated from $\mathbf{0}$ by $\text{aff}(A^0)$ in $\text{aff}(A^0 \cup \{\mathbf{0}\})$; and therefore we have $|B_0| \geq 3$. We can define the following set map:

$$\mathscr{F}(B) = \begin{cases} \{k : \exists (i, j) \in B \times B, i \neq j, \forall x_k \in \mathbf{S}_k, \text{conv}(\mathbf{S}_i \cup \mathbf{S}_j) \cap \overrightarrow{x_k \mathbf{0}} \setminus \{x_k\} \neq \emptyset\} & \text{if } |B| \geq 2 \\ \emptyset & \text{otherwise.} \end{cases}$$

We have $\mathscr{F}(B) \subseteq \mathscr{F}(B')$ if $B \subseteq B'$. Let $B_\ell = \mathscr{F}(B_{\ell-1})$ for $\ell = 1, 2, \ldots$ As remarked above $B_1 \subseteq B_0$ and, by induction, $B_\ell \subseteq B_{\ell-1}$ for $\ell \geq 1$. Thus, the sequence (B_ℓ) converges towards a set B^* satisfying $\mathscr{F}(B^*) = B^*$. Finally, note that, by induction, $|B_\ell| \geq 3$: The base case holds as $|B_0| \geq 3$, and a pair $i, j \in B_\ell$ with $i \neq j$ yields a $k \in B_{\ell+1}$, then i, k yields an additional k' in $B_{\ell+1}$, which in turn, with k, yields a third element in $B_{\ell+1}$; and thus $|B^*| \geq 3$.

For any $v \in \bigcup_{k \in B^*} \mathbf{S}_k$, the ray $\overrightarrow{v\mathbf{0}}$ intersects the convex hull of $\bigcup_{k \in B^*} \mathbf{S}_k$ in a point distinct from v since $\mathscr{F}(B^*) = B^*$. It contradicts the fact that $\text{aff}(A^0)$ separates $\mathbf{0}$ from $\mathbf{S}_{a+1} \cup \ldots \cup \mathbf{S}_{d+1}$ in $\text{aff}(A^0 \cup \{\mathbf{0}\})$ by the following argument. There exists at least one facet of $\text{conv}(\bigcup_{k \in B^*} \mathbf{S}_k)$ whose supporting hyperplane separates $\mathbf{0}$ from $\text{conv}(\bigcup_{k \in B^*} \mathbf{S}_k)$ and, for a vertex v of this facet, we have $\text{conv}(\bigcup_{k \in B^*} \mathbf{S}_k) \cap \overrightarrow{v\mathbf{0}} = \{v\}$, which is impossible. $\qquad\square$

The gap between Theorem 3 and its general position version is illustrated by the following example in $\mathbb{R}^3$ where $\bigcup_i \mathbf{S}_i \cup \{\mathbf{0}\}$ lie in the same plane. *Let $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3, \mathbf{S}_4$ be finite sets of points in $\mathbb{R}^2$. Assume that, for each $1 \leq i < j \leq d+1$, there*

exists $k \notin \{i, j\}$ such that, for all $x_k \in \mathbf{S}_k$, the convex hull of $\mathbf{S}_i \cup \mathbf{S}_j$ intersects the ray $\overrightarrow{x_k \mathbf{0}}$ in a point distinct from x_k. Then there exists a set $S \subseteq \bigcup_i \mathbf{S}_i$ such that $|S \cap \mathbf{S}_i| = 1$ for $i = 1, \ldots, d+1$ and $\mathbf{0} \in \text{conv}(S)$. This property cannot be obtained by simply applying Theorem 3 with $d = 2$ since its conditions might not be satisfied by $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3$. Indeed, k may be equal to 4 for some $i \neq j$. This property can neither be obtained by a compactness argument since it would require to find sequences $(\mathbf{S}_i^j)_{j=1,\ldots,\infty}$ of generic point sets converging to $\mathbf{S}_i$ while satisfying the condition of Theorem 3. The case when each $\mathbf{S}_i$ is reduced to one point s_i shows that such a sequence may fail to exist as the condition implies that $s_1^j, s_2^j, s_3^j, s_4^j$ and $\mathbf{0}$ lie in a common plane. This might explain why we could not avoid a Tarsky-type fixed point argument.

2.3   Proof of Theorem 5

We present a proof of Theorem 5 for the planar case providing an alternative and possibly more combinatorial proof of Theorem 4 in the plane. Consider the graph $G = (V, E)$ with $V = \mathbf{S}_1 \cup \mathbf{S}_2 \cup \mathbf{S}_3$ and where a pair of nodes are adjacent if and only if they have different colours. We get a directed graph $D = (V, A)$ by orienting the edges of G such that $\mathbf{0}$ is always on the right side of any arc, i.e. on the right side of the line extending it, with the induced orientation. Since $\text{conv}(\mathbf{S}_i \cup \mathbf{S}_j) \cap \text{aff}(x_k, \mathbf{0}) \neq \emptyset$ with i, j, k pairwise distinct and $x_k \in \mathbf{S}_k$, we have $\deg^+(v) \geq 1$ and $\deg^-(v) \geq 1$ for all $v \in V$. It implies that there exists at least one circuit in D, and we consider the shortest circuit C. We first show that the length of C is at most 4 since any circuit of length 5 or more has necessarily a chord. Indeed, take a vertex v, there is a vertex u on the circuit at distance 2 or 3 having a colour distinct from the colour of v, and thus the arc (u, v) or (v, u) exists in D. Therefore, the length of C must be 3 or 4. If the length is 3, we are done as the 3 vertices of C form a colourful triangle containing $\mathbf{0}$. If the length is 4, the circuit C is 2-coloured as otherwise we could again find a chord. Consider such a 2-coloured circuit C of length 4 and take any generic ray originating from $\mathbf{0}$. We recall that given an oriented closed curve $\mathscr{C}$ in the plane, with k_+, respectively k_-, denoting the number of times a generic ray intersects $\mathscr{C}$ while entering by the right, respectively left, side, the quantity $k_+ - k_-$ does not depend on the ray. Considering the realization of C as a curve $\mathscr{C}$, we have $k_- = 0$ by definition of the orientation of the arcs. Since we can choose a ray intersecting C at least once, k_+ remains constant and non-zero. Take now a vertex w of the missing colour, and take the ray originating from $\mathbf{0}$ in the opposite direction. This ray intersects an arc of C since $k_+ \neq 0$, and the endpoints of the arc together with w form a colourful triangle containing $\mathbf{0}$. $\square$

Remark 2. The fact that a directed graph missing a source or a sink has always a circuit is a key argument, and it is not clear to us how the planar proof could be extended or adapted to dimension 3 or more.

3 Related Results and an Algorithm

3.1 Given One, Find Another One

Bárány and Onn [4] raised the following algorithmic question: Given sets $\mathbf{S}_i$ containing $\mathbf{0}$ in their convex hulls, finding a colourful simplex containing $\mathbf{0}$ in its convex hull. This question, called *colourful feasibility problem*, belongs to the *Total Function Nondeterministic Polynomial* (TFNP) class, i.e. problems whose decision version has always a *yes* answer. The geometric algorithms introduced by Bárány [2] and Bárány and Onn [4] and other methods to tackle the colourful feasibility problem, such as multi-update modifications, are studied and benchmarked in [7]. The complexity of this challenging problem, i.e. whether it is polynomial-time solvable or not, is still an open question. However, there are strong indications that no TFNP-complete problem exists, see [12]. The following Proposition 1, which is similar in flavour to Theorem 1, may indicate an inherent hardness result for this relative of the colourful feasibility problem. Indeed, the algorithmic problem associated to Proposition 1 belongs to the *Polynomial Parity Argument* (PPA) class defined by Papadimitriou [12] for which complete problems are known to exist. In addition, the proof of Proposition 1 is a key ingredient of the algorithm finding a colourful simplex under the condition of Theorem 4.

Proposition 1. *Given $d+1$ sets, or colours, $\mathbf{S}_1^*, \mathbf{S}_2^*, \ldots, \mathbf{S}_{d+1}^*$ of points in $\mathbb{R}^d$ with $|\mathbf{S}_i^*| = 2$ for $i = 1, \ldots, d+1$, if there is a colourful simplex containing $\mathbf{0}$, then there is another colourful simplex containing $\mathbf{0}$.*

Proof. Without loss of generality we assume that the points in $\bigcup_i \mathbf{S}_i^* \cup \{\mathbf{0}\}$ are distinct and in general position. Consider the graph G whose nodes consist of some subsets $\bigcup_i \mathbf{S}_i^*$ partitioned into three types: (i) N_1 made of subsets v_1 of cardinality $d+2$ with $\mathbf{0} \in \text{conv}(v_1)$, $|v_1 \cap \mathbf{S}_i^*| = 1$ for $i = 1, \ldots, d$, and $|v_1 \cap \mathbf{S}_{d+1}^*| = 2$; (ii) N_2 made of subsets v_2 of cardinality $d+1$ with $\mathbf{0} \in \text{conv}(v_2)$, $|v_2 \cap \mathbf{S}_i^*| = 1$ for $i = 1, \ldots, d$ except for exactly one i, and $|v_2 \cap \mathbf{S}_{d+1}^*| = 2$; and (iii) N_3 made of subsets v_3 of cardinality $d+1$ with $\mathbf{0} \in \text{conv}(v_3)$ and $|v_3 \cap \mathbf{S}_i^*| = 1$ for $i = 1, \ldots, d+1$. The adjacency between the nodes of G is defined as follows. There is no edge between nodes of type v_2 and v_3. The nodes v_1 and v_2, respectively v_1 and v_3, are adjacent if and only if $v_2 \subseteq v_1$, respectively $v_3 \subseteq v_1$.

We show that G is a collection of node-disjoint paths and cycles by checking the degree of N_1, N_2, and N_3 nodes. First consider a N_1 node v_1. We recall that, under the general position assumption, there are exactly two $d+1$-subsets χ and χ' of v_1 containing $\mathbf{0}$ in their convex hull. This fact can be expressed as, using the simplex method terminology, there is a unique leaving variable in a pivot step of the simplex method assuming non-degeneracy. Both χ and χ' intersect $\mathbf{S}_i^*$ for $i = 1, \ldots, d$ in at least one point except maybe for one i. Thus, χ and χ' are N_2 or N_3 nodes, hence the degree of a N_1 node is 2. Consider now a N_2 node v_2, there is a $i_0 \neq d+1$ such that $|v_2 \cap \mathbf{S}_{i_0}^*| = 0$. The node v_2 is contained in exactly two N_1 nodes, each

of them obtained by adding one of the points in $\mathbf{S}^*_{i_0}$. Hence the degree of a N_2 node is 2. Finally, consider a N_3 node, it is contained in exactly one N_1 node obtained by adding the missing point of $\mathbf{S}^*_{d+1}$. Hence, the degree of a N_3 node is 1. The graph G is thus a collection of node disjoint paths and cycles.

Therefore, the existence of a colourful simplex containing $\mathbf{0}$ provides a N_3 node, and following the path in G until reaching the other endpoint provides another node of degree 1, i.e. a N_3 node corresponding to a distinct colourful simplex containing the origin $\mathbf{0}$.   □

Proposition 1 raises the following problem, which we call *Second covering colourful simplex*: Given $d+1$ sets, or colours, $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{d+1}$ of points in $\mathbb{R}^d$ with $|\mathbf{S}_i| \geq 2$ for $i = 1, \ldots, d+1$, and a colourful set $S \subseteq \bigcup_i \mathbf{S}_i$ containing $\mathbf{0}$ in its convex hull, find another such set. The key property used in the proof of Proposition 1 is the fact that the existence of one odd degree node in a graph implies the existence of another one. In other words, the proof of Proposition 1 shows that *Second covering colourful simplex* belongs to the PPA class, which forms precisely the problems in TFNP for which the existence is proven through this parity argument. Other examples of PPA problems include *Brouwer*, *Borsuk-Ulam*, *Second Hamiltonian circuit*, *Nash*, or *Room partitioning* [9]. The PPA class has a nonempty subclass of PPA-complete problems for which the existence of a polynomial algorithm would imply the existence of a polynomial algorithm for any problem in PPA, see Grigni [10]. We do not know whether *Second covering colourful simplex* is PPA-complete, but it is certainly a challenging question related to the complexity of colourful feasibility problem.

Note that Proposition 1 can also be proven by a degree argument on the map embedding the join of the $\mathbf{S}^*_i$ in $\mathbb{R}^d$, or using the Octahedron Lemma [3].

3.2   Minimum Number of Colourful Simplices Containing 0

As a corollary of Proposition 1, any condition implying the existence of a colourful simplex containing $\mathbf{0}$ actually implies the existence of $\min_i |\mathbf{S}_i|$ such simplices.

Corollary 1. *Given $d+1$ sets, or colours, $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{d+1}$ of points in $\mathbb{R}^d$, if there is a colourful simplex containing $\mathbf{0}$, then there are at least $\min_i |\mathbf{S}_i|$ colourful simplices containing the origin $\mathbf{0}$.*

Proof. Let $I = \min_i |\mathbf{S}_i|$ and $\mathbf{S}_i = \{v^1_i, v^2_i, \ldots\}$, and assume without loss of generality that the given colourful simplex containing $\mathbf{0}$ in its convex hull is $\text{conv}(v^1_1, v^1_2, \ldots, v^1_{d+1})$. Applying Proposition 1 $(I-1)$ times with $\mathbf{S}^*_i = \{v^1_i, v^k_i\}$ we obtain an additional distinct colourful simplex containing $\mathbf{0}$ for each $k = 2, \ldots, I$. □

We recall that a covering octahedron Ω is a pair of disjoint $\hat{i}$-transversals such that Ω covers all of $\mathbb{S}^{d-1}$ with colourful cones.

Corollary 2. *Given* $d + 1$ *sets, or colours,* $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{d+1}$ *of points in* $\mathbb{R}^d$, *if there is a covering octahedron, then there are at least* $\min_{i \neq j} |\mathbf{S}_i \cup \mathbf{S}_j| - 2$ *colourful simplices containing the origin* $\mathbf{0}$.

Proof. Denote $\mathbf{S}_i = \{v_i^1, v_i^2, \ldots\}$ and $I = \min_i |\mathbf{S}_i|$. Without loss of generality, we can assume that the octahedron covering $\mathbb{S}^{d-1}$ is formed by a pair (T_1, T_2) of $\widehat{d + 1}$-transversals such that $T_1 = \{v_1^1, v_2^1, \ldots, v_d^1\}$ covers a point in $\mathbf{S}_{d+1}$. Applying Proposition 1 $(I - 2)$ times with $\mathbf{S}_i^* = \{v_i^1, v_i^k\}$ we obtain, for each $k = 3, \ldots, I$, an additional distinct colourful simplex containing $\mathbf{0}$ distinct from the $|\mathbf{S}_{d+1}|$ colourful simplices containing $\mathbf{0}$ given by the assumption that (T_1, T_2) covers $\mathbb{S}^{d-1}$. $\qquad\square$

The minimum number $\mu(d)$ of colourful simplices containing $\mathbf{0}$ for sets satisfying the condition of Theorem 1, the general position assumption, and $|\mathbf{S}_i| \geq d + 1$ for all i was investigated in [3, 6, 8, 13]. While it is conjectured that $\mu(d) = d^2 + 1$ for all $d \geq 1$, the best current upper and lower bounds are $d^2 + 1 \geq \mu(d) \geq \left\lceil \frac{(d+1)^2}{2} \right\rceil$. In addition, we have $\mu(3) = 10$ and $\mu(d)$ even for odd d. For sets satisfying $|\mathbf{S}_i| \geq d + 1$ for all i, one can consider the analogous quantities $\mu^\diamond(d)$, respectively $\mu^\circ(d)$, defined as the minimum number of colourful simplices containing $\mathbf{0}$ for sets satisfying the condition of Theorem 2 and the general position assumption, respectively Theorem 4. Since $\mu^\circ(d) \leq \mu^\diamond(d)$ and, as noted in [5], $\mu^\diamond(d) = d + 1$, Theorem 4 and Corollary 1 imply that $\mu^\circ(d) = d + 1$ for $d \geq 2$. As the condition $\bigcap_i \text{conv}(\mathbf{S}_i) \neq \emptyset$ implies the existence of at least one octahedron covering $\mathbb{S}^{d-1}$, Corollary 2 generalizes the inequality $\mu(d) \geq 2d$ showed in [6].

3.3 Algorithm to Find a Colourful Simplex

We present an algorithm based on the proof of Proposition 1 finding a colourful simplex containing $\mathbf{0}$ for sets satisfying the conditions of Theorem 4, and, therefore, for sets satisfying the condition of Theorem 3 and the general position assumption. Note that the algorithm also finds a colourful simplex under the condition of Theorem 2.

Algorithm

Take any colourful $(d - 1)$-simplex σ whose vertices form, without loss of generality, a $\widehat{d + 1}$-transversal $T = \{v_1, \ldots, v_d\}$, and a ray $\mathbf{r}$ intersecting σ in its interior. Let $H^+(T)$ be the open half-space delimited by $\text{aff}(T)$ and containing $\mathbf{0}$. Check if there is a colourful d-simplex having σ as a facet and either containing $\mathbf{0}$ or having a facet τ intersecting $\mathbf{r}$ before σ. If there is none, we obtain d new vertices forming a $\widehat{d + 1}$-transversal $T' = \{v_1', \ldots, v_d'\}$ in $H^+(T)$, see Sect. 2.1. Take a point $x \notin \bigcup_i \mathbf{S}_i$ in $\text{aff}(\mathbf{r})$ such that $\mathbf{0} \in \text{conv}(v_1, \ldots, v_d, x)$, and choose

any point $v'_{d+1} \in \mathbf{S}_{d+1}$. We can use Proposition 1 and its constructive proof with $\mathbf{S}_i^* = \{v_i, v'_i\}$ for $i = 1, \ldots, d$ and $\mathbf{S}_{d+1}^* = \{x, v'_{d+1}\}$ to obtain a new colourful simplex containing $\mathbf{0}$ with at least one vertex in T'. If v'_{d+1} is a vertex of the new simplex, we do have a colourful simplex containing $\mathbf{0}$. Otherwise, the facet of the simplex not containing x is a colourful $(d-1)$-simplex τ intersecting $\mathbf{r}$ before σ since aff(T) forms the boundary of $H^+(T)$.

Given a colourful $(d-1)$-simplex σ intersecting $\mathbf{r}$, the proposed algorithm finds either a colourful simplex containing $\mathbf{0}$, or a colourful $(d-1)$-simplex τ intersecting $\mathbf{r}$ before σ. Since there is a finite number of colourful $(d-1)$-simplices, the algorithm eventually finds a colourful simplex containing $\mathbf{0}$. While non-proven to be polynomial, pivot-based algorithms, such as the Bárány-Onn ones or our algorithm, are typically efficient in practice.

References

1. Arocha, J.L., Bárány, I., Bracho, J., Fabila, R., Montejano, L.: Very colorful theorems. Discret. Comput. Geom. **42**, 142–154 (2009)
2. Bárány, I.: A generalization of Carathéodory's theorem. Discret. Math. **40**, 141–152 (1982)
3. Bárány, I., Matoušek, J.: Quadratically many colorful simplices. SIAM J. Discret. Math. **21**, 191–198 (2007)
4. Bárány, I., Onn, S.: Colourful linear programming and its relatives. Math. Oper. Res. **22**, 550–567 (1997)
5. Custard, G., Deza, A., Stephen, T., Xie, F.: Small octahedral systems. In: Proceedings of the 23rd Canadian Conference on Computational Geometry (CCCG'11), Toronto (2011)
6. Deza, A., Huang, S., Stephen, T., Terlaky, T.: Colourful simplicial depth. Discret. Comput. Geom. **35**, 597–604 (2006)
7. Deza, A., Huang, S., Stephen, T., Terlaky, T.: The colourful feasibility problem. Discret. Appl. Math. **156**, 2166–2177 (2008)
8. Deza, A., Stephen, T., Xie, F.: More colourful simplices. Discret. Comput. Geom. **45**, 272–278 (2011)
9. Edmonds, J., Sanità, L.: On finding another room-partitioning of the vertices. Electron. Notes Discret. Math. **36**, 1257–1264 (2010)
10. Grigni, M.: A Sperner lemma complete for PPA. Inf. Process. Lett. **77**, 255–259 (2001)
11. Holmsen, A.F., Pach, J., Tverberg, H.: Points surrounding the origin. Combinatorica **28**, 633–644 (2008)
12. Papadimitriou, C.: On the complexity of the parity argument and other inefficient proofs of existence. J. Comput. Syst. Sci. **48**, 498–532 (1994)
13. Stephen, T., Thomas, H.: A quadratic lower bound for colourful simplicial depth. J. Comb. Optim. **16**, 324–327 (2008)

Expected Crossing Numbers

Bojan Mohar and Tamon Stephen

Abstract The expected value for the weighted crossing number of a randomly weighted graph is studied. We focus on the case where $G = K_n$ and the edge-weights are independent random variables that are uniformly distributed on $[0, 1]$. The first non-trivial case is K_5. We compute this via an unexpectedly involved calculation, and consider bounds for larger values of n. A variation of the Crossing Lemma for expectations is proved.

Key words Graph • Crossing number • Weighted crossing number • Crossing lemma

Subject Classifications: 05C10, 60C05

1   Introduction

The *crossing number* of a graph is the minimum number of internal intersections of edges in a drawing of the graph on the plane. Computing the crossing number, even for complete graphs, is a surprisingly challenging problem and an active area of research [13, 14, 16].

B. Mohar
CRC in Graph Theory, Tier 1, Department of Mathematics, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada
e-mail: mohar@sfu.ca

T. Stephen (✉)
Department of Mathematics, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada
e-mail: tamon@sfu.ca

The notion of the weighted crossing number, when the edges have weights and each crossing counts as the product of the corresponding weights, has been used in various situations, since it mimics the possibility of having many edges in parallel. In this paper, we study the expected value of the weighted crossing number of the complete graph K_n on n vertices, where the weights of edges are independent random variables. We consider first the situation where the weights are i.i.d. variables with the uniform distribution on $[0, 1]$. The first non-trivial case is K_5; we show through an involved calculation that the expected value is $\frac{35,921}{1,108,800} \approx 0.032396$.

We then use a discrete distribution to show that the first two moments of the distribution on the edges are not sufficient to determine the expectation. If the weights are binary (Bernoulli) random variables, we get the problem of crossing numbers of random graphs. In [15], these are shown to have the $\Theta(n^4)$ asymptotics of the usual crossing number $\mathrm{cr}(K_n)$ of complete graphs. We give three proofs that this also true for the uniform distribution and compare the bounds obtained from each. The first proof is by comparison with the binary case, the second by using a similar recurrence as used for the usual crossing number of complete graphs and the third by proving and applying a variation of the Crossing Lemma for expectations.

2 Preliminaries

Given a graph $G = (V, E)$, we denote its crossing number by $\mathrm{cr}(G)$. This is the minimum over all drawings of G in the Euclidean plane $\mathbb{R}^2$ of the number of crossings of edges in the drawing. All drawings are assumed to have simple polygonal arcs representing the edges of the graph, and it is assumed that each pair of edges involves at most one intersection of their representing arcs. Here and in the remainder of the paper, we consider only internal intersections of edges. Formally, a *crossing* in a drawing $\mathscr{D}$ is an unordered pair $\{e, f\}$ of edges whose arcs in $\mathscr{D}$ intersect each other internally. We let $\mathbb{X}(\mathscr{D})$ denote the set of all crossings and set $\mathrm{cr}(\mathscr{D}) = |\mathbb{X}(\mathscr{D})|$.

Given non-negative weights $w : E \to \mathbb{R}_+$ on the edges of G, we define the *crossing weight* of a drawing $\mathscr{D}$ of G as:

$$\mathrm{cr}(\mathscr{D}, w) = \sum_{\{e, f\} \in \mathbb{X}(\mathscr{D})} w(e)w(f).$$

We define the *weighted crossing number* of a weighted graph G as:

$$\mathrm{cr}(G, w) = \min_{\mathscr{D}} \mathrm{cr}(\mathscr{D}, w). \tag{1}$$

For a fixed graph, the function $\mathrm{cr}(G, \cdot)$ is also called the *crossing function* for G. We take the domain of $\mathrm{cr}(G, \cdot)$ to be $\mathbb{R}_+^E$. We remark that $\mathrm{cr}(G, 0) = 0$ and $\mathrm{cr}(G, w) \geq 0$. We also note that $\mathrm{cr}(G, \cdot) \equiv \mathbf{0}$ if and only if $\mathrm{cr}(G) = 0$. The function $\mathrm{cr}(G, \cdot)$

is piecewise quadratic in w, and the chambers defined by these pieces correspond to (groups of) optimal drawings for the contained weightings; the forms in the chambers are neither convex nor concave. If $\mathbf{1} \in \mathbb{R}_+^E$ is the constant all-1 function, then $\mathrm{cr}(G) = \mathrm{cr}(G, \mathbf{1})$.

The crossing function of any n-vertex graph is just a specialization of the crossing function $\mathrm{cr}(K_n, w)$ of the complete graph K_n, where we put weight 0 for the non-edges in the graph. In this sense the crossing functions of complete graphs contain information about crossing numbers of all graphs. This universality property was the main goal to introduce this notion in [7, 8] and to propose its study.

Note that we allow the edges to be represented by any (polygonal) line, they need not be straight lines. As in the unweighted case, minimal drawings can be obtained without using double crossings (pairs of edges that cross more than once). The related questions of the *rectilinear crossing number*, the *pairwise crossing number* and the *odd crossing number* are also interesting and well-studied, see for example [10]. For the rectilinear crossing number, the edges are required to be straight lines. The pairwise crossing number counts the number of pairs of crossing edges, so a pair of edges that intersects several times counts only once. The odd crossing number counts the number of pairs of edges crossing an odd number of times. We can consider weighted versions of all four types of crossing numbers.

Clearly the odd crossing number is at most the pairwise crossing number. Also, the pairwise crossing number is at most the crossing number (with equality in all known cases), and the crossing number is at most the rectilinear crossing number (with strict inequality in some cases [4]). These relations remain for the weighted versions. For the computations performed in Sect. 3, these four numbers will be the same.

3 Computation of the Uniform Expected Crossing Number for K_5

We begin by considering the expected crossing number of the complete graph K_n for some small values of n. We take the weights on the edges to be independently identically distributed random variables, with uniform distributions on the interval $[0, 1]$. Let us denote the expected value of $\mathrm{cr}(K_n, w)$ under this distribution as $\mathrm{Eu}(n)$.

For $n \leq 4$, the graph can be drawn without crossings, so $\mathrm{Eu}(n) = 0 = \mathrm{cr}(K_n)$. For $n \geq 5$, we have $0 < \mathrm{Eu}(n) < \mathrm{cr}(K_n)$. In this section, we compute $\mathrm{Eu}(5)$ directly from the definition of expectation. Our somewhat cumbersome case analysis can also be viewed as determination of the piecewise quadratic chambers for the crossing function of K_5. We conclude that:

Theorem 1. *We have* $\mathrm{Eu}(5) = \frac{35,921}{1,108,800}$.

Proof. We will label the edges of K_5 as in Fig. 1. We will denote the random weight assigned to the ith edge by $\mathbf{X}_i$, $i = 1, \ldots, 10$. We note that $\mathrm{cr}(K_5) = 1$ and by

Fig. 1 Edge labelling of K_5
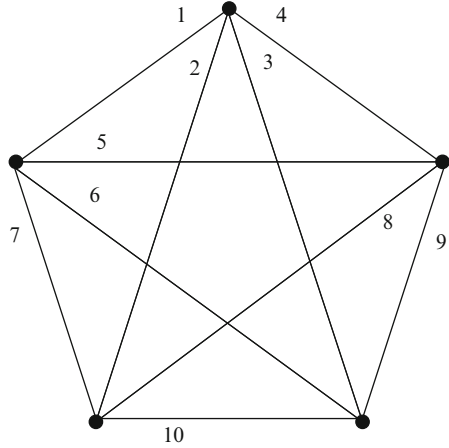
symmetry, for any two non-adjacent edges, K_5 can be drawn so that those two edges are the single pair of crossing edges. Hence:

$$\mathrm{Eu}(5) = \mathbb{E}[\min(\mathbf{X}_1\mathbf{X}_8, \mathbf{X}_1\mathbf{X}_9, \mathbf{X}_1\mathbf{X}_{10}, \mathbf{X}_2\mathbf{X}_5, \mathbf{X}_2\mathbf{X}_6, \mathbf{X}_2\mathbf{X}_9, \mathbf{X}_3\mathbf{X}_5, \mathbf{X}_3\mathbf{X}_7,$$
$$\mathbf{X}_3\mathbf{X}_8, \mathbf{X}_4\mathbf{X}_6, \mathbf{X}_4\mathbf{X}_8, \mathbf{X}_4\mathbf{X}_{10}, \mathbf{X}_5\mathbf{X}_{10}, \mathbf{X}_6\mathbf{X}_8, \mathbf{X}_7\mathbf{X}_9)]$$

We abbreviate the quantity inside the expectation as $m(\mathbf{X})$.

This is a problem in order statistics, see for instance [3]. The direct way to obtain $\mathrm{Eu}(5)$ is to evaluate:

$$\int_0^1 \int_0^1 \ldots \int_0^1 m(\mathbf{x}) \mathbf{d}x_1 \ldots \mathbf{d}x_9 \mathbf{d}x_{10} \tag{2}$$

where $m(\mathbf{x})$ is the function of $\mathbf{x} \in \mathbb{R}^{10}$ corresponding to the random variables of $m(\mathbf{X})$. To do this we break (2) into 10! terms based on the increasing order of the variables, i.e. we compute (2) via the sum:

$$\sum_{\sigma \in S_{10}} \int_0^1 \int_0^{x_{\sigma(10)}} \int_0^{x_{\sigma(9)}} \ldots \int_0^{x_{\sigma(2)}} m(\mathbf{x}) \mathbf{d}x_{\sigma(1)} \ldots \mathbf{d}x_{\sigma(9)} \mathbf{d}x_{\sigma(10)} \tag{3}$$

Here the permutations $\sigma \in S_{10}$ index the possible orderings of the random variables $\mathbf{X}$. This sum has 10! terms, but they can be grouped into a manageable number of cases. To begin, we note that by reordering the vertices, we can assume that $\mathbf{X}_1$ takes the smallest value, and, using the labelling of Fig. 1, $\mathbf{X}_2 \leq \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5, \mathbf{X}_6, \mathbf{X}_7$ and $\mathbf{X}_3 \leq \mathbf{X}_4$. This corresponds to a labelling of K_5 based on $\mathbf{X}$, breaking ties arbitrarily. Actually, we may assume that the weights $\mathbf{X}_i$, $1 \leq i \leq 10$, are pairwise different, since the set on which an equality occurs is of measure zero. Thus, each case with the above assumptions corresponds to 120 terms in (3).

With these assumptions, the minimum of the 15 pairs of random variables in $m(\mathbf{X})$ must be attained at one of $\mathbf{X}_1\mathbf{X}_8, \mathbf{X}_1\mathbf{X}_9, \mathbf{X}_1\mathbf{X}_{10}, \mathbf{X}_2\mathbf{X}_5, \mathbf{X}_2\mathbf{X}_6, \mathbf{X}_3\mathbf{X}_7$ since $\mathbf{X}_1\mathbf{X}_9 \leq \mathbf{X}_2\mathbf{X}_9, \mathbf{X}_7\mathbf{X}_9$; $\mathbf{X}_2\mathbf{X}_5 \leq \mathbf{X}_3\mathbf{X}_5$; etc. We note that these six terms are symmetric in the variables $\mathbf{X}_8, \mathbf{X}_9, \mathbf{X}_{10}$, and also in $\mathbf{X}_5, \mathbf{X}_6$. Thus we will also take $\mathbf{X}_8 = \min(\mathbf{X}_8, \mathbf{X}_9, \mathbf{X}_{10})$ and $\mathbf{X}_5 = \min(\mathbf{X}_5, \mathbf{X}_6)$, and treat the remaining cases by symmetry. Combined with our assumptions on $\mathbf{X}_1, \mathbf{X}_2$ and $\mathbf{X}_3$ we break the 10! terms of (3) into groups of 720 terms based on symmetry; this leaves us with 5,040 terms up to these symmetries. It also allows us to simplify our integrand further to $\min(\mathbf{X}_1\mathbf{X}_8, \mathbf{X}_2\mathbf{X}_5, \mathbf{X}_3\mathbf{X}_7)$.

We now divide into cases based on the relative orderings of some of the remaining variables. We remark that, depending on the order of the variables, the integrand may simplify further – for instance if the two smallest variables are $\mathbf{X}_1$ and $\mathbf{X}_8$, the minimum of the three terms will always be $\mathbf{X}_1\mathbf{X}_8$. We organize the cases by how the integrand simplifies.

Case 1: Orderings which ensure $\mathbf{X}_1\mathbf{X}_8 = \min(\mathbf{X}_1\mathbf{X}_8, \mathbf{X}_2\mathbf{X}_5, \mathbf{X}_3\mathbf{X}_7)$.

In these cases, the computation is relatively simple: the integral depends only on which position $\mathbf{X}_8$ occupies in the order of the $\mathbf{X}_i$'s. It can be anywhere from the second to fifth smallest. Suppose it is the second smallest, i.e. that the order of the variables is:

$$\mathbf{X}_1 \leq \mathbf{X}_8 \leq \mathbf{X}_{i_3} \leq \mathbf{X}_{i_4} \leq \mathbf{X}_{i_5} \leq \mathbf{X}_{i_6} \leq \mathbf{X}_{i_7} \leq \mathbf{X}_{i_8} \leq \mathbf{X}_{i_9} \leq \mathbf{X}_{i_{10}}.$$

Then we compute:

$$\int_0^1 \int_0^{x_{i_{10}}} \int_0^{x_{i_9}} \cdots \int_0^{x_{i_4}} \int_0^{x_{i_3}} \int_0^{x_8} x_1 x_8 \mathbf{d}x_1 \mathbf{d}x_8 \mathbf{d}x_{i_3} \dots \mathbf{d}x_{i_8} \mathbf{d}x_{i_9} \mathbf{d}x_{i_{10}}$$

$$= \int_0^1 \int_0^{x_{i_{10}}} \int_0^{x_{i_9}} \cdots \int_0^{x_{i_4}} \int_0^{x_{i_3}} \frac{x_8^3}{2} \mathbf{d}x_8 \mathbf{d}x_{i_3} \dots \mathbf{d}x_{i_8} \mathbf{d}x_{i_9} \mathbf{d}x_{i_{10}}$$

$$= \int_0^1 \int_0^{x_{i_{10}}} \int_0^{x_{i_9}} \cdots \int_0^{x_{i_4}} \frac{x_{i_3}^4}{2 \cdot 4} \mathbf{d}x_{i_3} \dots \mathbf{d}x_{i_8} \mathbf{d}x_{i_9} \mathbf{d}x_{i_{10}}$$

$$\cdots$$

$$= \int_0^1 \frac{x_{i_{10}}^{11}}{2 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 \cdot 11} \mathbf{d}x_{i_{10}} = \frac{3}{12!}.$$

A similar calculation shows that if $\mathbf{X}_8$ is ith smallest variable, the integral for a fixed ordering of the remaining variables will be $\frac{i+1}{12!}$.

Now observe that there are $\frac{8!}{24} = 1,680$ ways of ordering the variables with $\mathbf{X}_1$ as the smallest variable, $\mathbf{X}_8$ as the second smallest; $\mathbf{X}_2 \leq \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5, \mathbf{X}_6, \mathbf{X}_7$; $\mathbf{X}_3 \leq \mathbf{X}_4$; and $\mathbf{X}_5 \leq \mathbf{X}_6$. We remark that our symmetry assumptions guarantee that either $\mathbf{X}_2$ or $\mathbf{X}_8$ is the second smallest variable, so in the remainder of the analysis $\mathbf{X}_2$ will always be the second smallest variable.

Thus if $\mathbf{X}_8$ is the third smallest variable, we have fixed the order of the first three variables, and the remaining variables can be ordered in $7!/4 = 1{,}260$ ways, accounting for the facts that $\mathbf{X}_3 \leq \mathbf{X}_4$ and $\mathbf{X}_5 \leq \mathbf{X}_6$.

If $\mathbf{X}_8$ is the fourth smallest variable, we have two possible choices for the third smallest: $\mathbf{X}_3$ and $\mathbf{X}_7$. In the former case we have $6!/2 = 360$ possible orderings of the remaining variables (accounting for $\mathbf{X}_5 \leq \mathbf{X}_6$), while for the latter case we have $6!/4 = 180$ possible orderings.

Finally, under the assumptions of Case 1, $\mathbf{X}_8$ can be the fifth smallest variable only if the third and fourth variables are $\mathbf{X}_3$ and $\mathbf{X}_4$, respectively. There are $5!/2$ orderings of the remaining variables compatible with this. We remark that in this case, we can never have $\mathbf{X}_5 \leq \mathbf{X}_8$, or $\mathbf{X}_3, \mathbf{X}_7 \leq \mathbf{X}_8$ since then it may be the case that $\mathbf{X}_1 \mathbf{X}_8$ is not minimal, depending on the values chosen.

This already covers the majority of the cases, $3{,}540$ of the $5{,}040$. Thus the terms in (3) corresponding to these orderings of the variables have total weight per symmetry class of:

$$1{,}680 \cdot \frac{3}{12!} + 1{,}260 \cdot \frac{4}{12!} + 540 \cdot \frac{5}{12!} + 60 \cdot \frac{6}{12!} = \frac{13{,}140}{12!}.$$

The remaining cases are similar and are included in the Appendix.

As noted in Sect. 4.3, the computed value of $\mathrm{Eu}(5)$ is used in a lower bound for $\mathrm{Eu}(n)$ for $n \geq 5$.

It is straightforward to verify the first few digits of this number by simulating the ten uniform random variables. A short MATLAB program sampled the ten variables 10^{12} times and computed the minimum; the computed number agreed with our calculation to the seventh decimal place. While this number arises in a relatively simple way, we do not know of it arising in other places.

Unfortunately, it would be much harder to use such a simulation to get approximate values of $\mathrm{Eu}(6)$ or $\mathrm{Eu}(7)$. The proof method used above for K_5 does not generalize to K_6 or K_7 either. To simulate $\mathrm{Eu}(6)$ we would need to catalogue the minimal ways of drawing K_6, i.e. drawings $\mathscr{D}$ for which $\mathbb{X}(\mathscr{D})$ is inclusion-wise minimal.

Remark 1. We can get a simple lower bound by observing that each edge has weight at least $1-p$ with probability at least p; thus the expectation is at least $p^{10}(1-p)^2$ for any $p \in [0, 1]$. This expression is maximized at $p = \frac{5}{6}$, giving the bound $\mathrm{Eu}(5) \geq \frac{5^{10}}{6^{12}} \approx 0.0045$, which is weak, but can be generalized.

Remark 2. Gelasio Salazar notes that the same approach can be used to compute the expected crossing number of $K_{3,3}$ with uniform $[0, 1]$ random variables on the edges. This turns out to be less involved than the K_5 case due to the symmetries between the edges; using the same method we arrived at $\mathrm{Eu}(K_{3,3}) = \frac{23}{1{,}155} \approx 0.019913$.

4   Moments and Bounds for Eu(n)

In this section, we consider the expected crossing number under some discrete distributions. By doing this in the case where the random variable takes one of two values, we show that the first two moments of the distribution do not determine the expected crossing number, see Sect. 4.1.

When the two values are 0 and 1, the random variables define a random graph. We note that the expected crossing number for the uniform distribution cannot be much less than the expected crossing number of the random graph (Sect. 4.2). This fact gives us an asymptotic lower bound for Eu(n). In fact, such a bound can be obtained more directly via a recurrence as used for the usual crossing number of complete graphs (Sect. 4.3), or by proving and applying a variation of the Crossing Lemma for expectations (Sect. 4.4).

Finally, we discuss upper bounds briefly in Sect. 4.6.

4.1   *Moments*

Consider for $0 \le t < u$, the discrete distribution where edges have weight t or u with probability $\frac{1}{2}$. Let Edisc(n, t, u) be the expected weighted crossing number of K_n with the distribution for given t, u; if the parameter u is omitted we will assume it is $1 - t$. Then it is easy to see that

$$\mathrm{cr}(K_5, w) = \begin{cases} t^2 & \text{if there is a pair of non-adjacent edges of weight } t \\ u^2 & \text{if all edges have weight } u \\ tu & \text{otherwise.} \end{cases}$$

All 2^{10} possible assignments of t's and u's to the edges are equally likely. There is only one way for all edges to have weight u. Otherwise, if we do not have two non-adjacent edges of weight t, we must either have all edges of weight t incident with a single vertex, or three edges forming a triangle. In the former situation, we may have one edge (10 assignments), two edges (30 assignments), three edges (20 assignments) or four edges (5 assignments). For the triangles, we have ten more assignments. The remaining 948 assignments of t's and u's to the edges have a pair of non-adjacent edges of weight t. Therefore, Edisc$(5, t, u) = \frac{1}{1,024}(948t^2 + 75tu + u^2)$, which simplifies to Edisc$(5, t) = \frac{1}{1,024}(874t^2 + 73t + 1)$ when $u = 1 - t$.

The mean and variance of the considered discrete distribution are $\frac{u+t}{2}$ and $\frac{(u-t)^2}{4}$, respectively. If we take $u = 1 - t$, then the mean is $\frac{1}{2}$, matching the mean of the uniform distribution, while the variance is $\frac{(1-2t)^2}{4}$. Since the variance of the uniform case is $\frac{1}{12}$, by choosing $t = \frac{3-\sqrt{3}}{6}$, we get a distribution that matches the uniform distribution in its first two moments. However the above calculation shows that

$$\text{Edisc}\left(5, \frac{3-\sqrt{3}}{6}\right) = \frac{1{,}973 - 947\sqrt{3}}{6{,}144} \approx 0.05416 > \text{Eu}(5).$$

Thus the first two moments of the input distribution on the edges are not sufficient to determine the expected crossing number. We believe that a constant number of higher moments is not sufficient either. Perhaps, up to $\binom{n}{2}$ moments are required.

4.2 Random Graphs

If we take $t = 0$ and $u = 1$ above we have a random graph on n vertices where the edges are chosen randomly and independently with probability $p = \frac{1}{2}$. We can also consider more general $p > 0$, and denote the random graph on n vertices with edge probability p as $G(n, p)$.

The expected crossing number of $G(n, p)$ is studied in [15], focusing its behaviour as n increases, with p perhaps a function of n. They find several interesting results, which are unexpectedly different for pairwise crossing numbers, crossing numbers and rectilinear crossing numbers. The differences may be the availability of proof techniques rather than in the numbers themselves.

We note that given a lower bound ρ for $\mathbb{E}[\text{cr}(G(n, \frac{1}{2}))]$ we get a lower bound of $\frac{\rho}{4}$ for $\text{Eu}(n)$ by rounding down the uniform $[0, 1]$ variables to 0 or $\frac{1}{2}$ respectively. In particular, it is shown in [15] that $\mathbb{E}[\text{cr}(G(n, \frac{1}{2}))] = \Theta(n^4)$, so the same holds true of $\text{Eu}(n)$. (Since $\text{cr}(K_n)$ is $O(n^4)$ and an upper bound for $\text{Eu}(n)$, we need only show the lower bound.) In the following section we give two direct proofs of this fact by adapting standard arguments used for crossing numbers to work for the expectations.

4.3 Lower Bounds via a Recurrence

We recall that we denote the crossing weight of a given drawing $\mathscr{D}$ of a graph weighted by w as $\text{cr}(\mathscr{D}, w)$, and the weighted crossing number of G weighted by w (i.e. the minimum over all drawings) by $\text{cr}(G, w)$.

Given a drawing $\mathscr{D}$ of K_n with weights w, we can consider the induced drawings of copies of $K_n - v \approx K_{n-1}$ obtained by removing each vertex $v \in V = V(K_n)$ from K_n in turn. Then

$$\sum_{v \in V} \text{cr}(\mathscr{D}|_{K_n-v}, w|_{K_n-v}) = (n-4)\,\text{cr}(\mathscr{D}, w) \tag{4}$$

since each pair of disjoint edges $ij, i'j'$ of K_n appear in all but four of the terms on the left side of (4).

Now consider K_n for $n > 4$ with a fixed weighting w. There is some optimal drawing $\mathscr{D}^*$ of K_n such that $\operatorname{cr}(K_n, w) = \operatorname{cr}(\mathscr{D}^*, w)$. Now:

$$\operatorname{cr}(K_n, w) = \operatorname{cr}(\mathscr{D}^*, w) = \frac{1}{n-4} \sum_{v \in V} \operatorname{cr}(\mathscr{D}^*|_{K_n - v}, w|_{K_n - v})$$

$$\geq \frac{1}{n-4} \sum_{v \in V} \min_{\mathscr{D}} \operatorname{cr}(\mathscr{D}|_{K_n - v}, w|_{K_n - v}) = \frac{1}{n-4} \sum_{v \in V} \operatorname{cr}(K_n - v, w|_{K_n - v}).$$

If the weights in w are i.i.d. random variables, we can take expectations on both sides to get $\operatorname{Eu}(n) \geq \frac{n}{n-4} \operatorname{Eu}(n-1)$. Applying this inequality recursively, we find for $n \geq 6$ that $\operatorname{Eu}(n) \geq \frac{1}{5} \binom{n}{4} \operatorname{Eu}(5)$.

4.4 Lower Bounds via the Crossing Lemma

The following result, known as the Crossing Lemma, was proved independently by Ajtai et al. [2] and Leighton [5]. The version given below (with the specific constant $1,024/31,827 > 0.032$) is due to Pach et al. [11].

Theorem 2 (Crossing Lemma). *Let G be a graph of order n with $m \geq \frac{103}{16} n$ edges. Then*

$$\operatorname{cr}(G) \geq \frac{1,024}{31,827} \frac{m^3}{n^2}.$$

Let π be a probability distribution with expectation $\mathbb{E}(\pi) = \mu$. We define the *complementary probability distribution* π^* by setting $\pi^*(\mu + x) = \pi(\mu - x)$. For the purpose of the following argument, let us assume that our probability distribution is symmetric, i.e., $\pi = \pi^*$. Then, given a random weight function w, the *complementary weight function w^**, defined as $w^*(e) = 2\mu - w(e)$, has the same distribution as w. Let us define w' to be either w or w^*, so that $w'(e) \geq \mu$ holds for at least half of the edges $e \in E(G)$. Finally, let w_1 be defined as $w_1(e) = 0$ if $w'(e) < \mu$, and $w_1(e) = 1$ if $w'(e) \geq \mu$. Since $\operatorname{cr}(G, w) + \operatorname{cr}(G, w^*) \geq \operatorname{cr}(G, w') \geq \mu^2 \operatorname{cr}(G, w_1)$, the following holds:

$$\mathbb{E}(\operatorname{cr}(G, w)) = \frac{1}{2} \mathbb{E}(\operatorname{cr}(G, w) + \operatorname{cr}(G, w^*)) \geq \frac{1}{2} \mathbb{E}(\operatorname{cr}(G, w'))$$

$$\geq \frac{\mu^2}{2} \mathbb{E}(\operatorname{cr}(G, w_1)) \geq \frac{\mu^2}{2} \cdot \frac{1,024}{31,827} \frac{(m/2)^3}{n^2} = \frac{64\mu^2}{31,827} \frac{m^3}{n^2}.$$

This gives a version of the crossing lemma for expectations. With a little more care we can improve the above bound and also get rid of the symmetry condition. In order to do this, we replace the mean by the *median*, i.e. the largest number ν such that $Prob[w(e) \geq \nu] \geq \frac{1}{2}$.

Theorem 3 (Crossing Lemma for Expectations). *Let G be a graph of order n with $m \geq \frac{103}{16(1-4^{-1/3})} n$ edges. Suppose that each edge $e \in E(G)$ gets a random weight $w(e)$, where the weights of distinct edges are independent non-negative random variables (not necessarily i.d.) whose median is at least $v > 0$. Then*

$$\mathbb{E}(\mathrm{cr}(G, w)) \geq \frac{128 v^2}{31{,}827} \cdot \frac{m^3}{n^2}.$$

Proof. Given w, we introduce related weights w'' and w_2 in a similar (but not identical) way as above: we let $w''(e) = 0$ if $w(e) < v$, and $w''(e) = v$ if $w(e) \geq v$; we let $w_2(e) = w''(e)/v$ be the corresponding weight with values 0 and 1. Note that $Prob[w''(e) = v] \geq \frac{1}{2}$ and $Prob[w_2(e) = 1] \geq \frac{1}{2}$. Similarly as before, we have $\mathrm{cr}(G, w) \geq \mathrm{cr}(G, w'') = v^2 \mathrm{cr}(G, w_2)$.

Note that w determines a spanning subgraph $F_w \subseteq G$, whose edges are those edges of G for which $w_2(e) = 1$. The graph F_w is a random subgraph of G, and for each spanning subgraph F of G we let $Prob(F)$ be the probability that $F = F_w$. We will need a lower bound for the sum $\sum \mathrm{cr}(F) Prob(F)$ taken over all (spanning) $F \subseteq G$. To do this, let us define $F' \subseteq F$ as the spanning subgraph of G such that $e \in E(F')$ if $w(e) \geq v_e \geq v$, where v_e is the median of the random variable $w(e)$. The threshold case when $w(e) = v_e$ is to be considered so that $Prob[e \in E(F')] = \frac{1}{2}$. Then F' is also a random spanning subgraph of G and $Prob(F') = 2^{-m}$. Since the event that an edge e is in F' is contained in the event that $e \in E(F)$, we have for each F

$$Prob(F) = \sum_{F' \subseteq F} \alpha(F, F') Prob(F'),$$

where $\alpha(F, F') \geq 0$ is the probability that we have $F_w = F$ under the condition that F' is given. Clearly, $\sum_{F \supseteq F'} \alpha(F, F') = 1$ for every fixed F'. Since $\mathrm{cr}(F')$ is an increasing function, we have:

$$\sum_{F \subseteq G} \mathrm{cr}(F) Prob(F) = \sum_{F \subseteq G} \mathrm{cr}(F) \sum_{F' \subseteq F} \alpha(F, F') Prob(F')$$

$$\geq \sum_{F' \subseteq G} \mathrm{cr}(F') Prob(F') \sum_{F \supseteq F'} \alpha(F, F')$$

$$= \sum_{F' \subseteq G} \mathrm{cr}(F') Prob(F') = \sum_{F' \subseteq G} 2^{-m} \mathrm{cr}(F').$$

We will employ another notion:

$$\lambda(k, n) = \min\{\mathrm{cr}(F) : |V(F)| = n, |E(F)| = k\}.$$

By the Crossing Lemma, $\lambda(k, n) \geq \frac{1{,}024}{31{,}827} \frac{k^3}{n^2}$ if $k \geq \frac{103}{16} n$. Using the introduced quantities, we obtain the following estimate:

$$\mathbb{E}[cr(G, w)] = \int_{\mathbb{R}^E} cr(G, w)\mathbf{d}w(x) \geq \int_{\mathbb{R}^E} cr(G, w')\mathbf{d}w(x) = v^2 \int_{\mathbb{R}^E} cr(G, w_1)\mathbf{d}w(x)$$

$$\geq v^2 \sum_{F \subseteq G} cr(F) Prob(F) \geq v^2 \sum_{F' \subseteq G} 2^{-m} cr(F')$$

$$\geq v^2 \sum_{k=0}^{m} \sum_{F \subseteq G, |E(F)|=k} 2^{-m} cr(F)$$

$$\geq v^2 2^{-m} \sum_{k=0}^{m} \binom{m}{k} \lambda(k, n)$$

$$\geq \frac{1{,}024 v^2}{31{,}827 \cdot 2^m n^2} \sum_{k=\lceil 103n/16 \rceil}^{m} \binom{m}{k} k^3.$$

The integrals in the first row are with respect to the random weight function w. Note that we have $k^3 + (m-k)^3 \geq \frac{1}{4}m^3$ for $0 \leq k \leq m$, and that for $k < 103n/16$, we have $(m-k)^3 \geq (m - 103n/16)^3 \geq (m - (1 - 4^{-1/3})m)^3 = \frac{1}{4}m^3$. Thus,

$$\sum_{k=\lceil 103n/16 \rceil}^{m} \binom{m}{k} k^3 \geq \frac{1}{2} \sum_{k=0}^{m} \binom{m}{k} \frac{1}{4}m^3 = \frac{1}{8}2^m m^3.$$

The above inequalities imply:

$$\mathbb{E}[cr(G, w)] \geq \frac{1{,}024 v^2}{31{,}827 \cdot 2^m n^2} \frac{1}{8}2^m m^3 = \frac{128 v^2 \cdot m^3}{31{,}827 \cdot n^2}$$

which we were to prove. For uniform random variables on K_n this works out to $Eu(n) \geq \frac{4n(n-1)^3}{31{,}827}$.

4.5 Comparison of Lower Bounds

The lower bounds of Sects. 4.2, 4.3, and 4.4 all have the form $cn^4 + O(n^3)$. Computing the values of c obtained from the three proofs, we see that the Crossing Lemma for Expectations gives $c_{cl} = \frac{4}{31{,}827} \approx 0.000126$ while the recurrence gives $c_{rec} = \frac{Eu(5)}{120} \approx 0.000270$, and is tight for $n = 5$. The Crossing Lemma bound has the advantage of being applicable as long as there are sufficiently many edges and only requiring the distributions on the edges to be independent, non-negative and have mean v.

The constant c_{rg} found in [15] from the random graph approach is asymptotically $\frac{1}{1{,}920} \approx 0.000521$; for fixed n it is $\frac{1}{1{,}920L^2}$ where $L > \frac{\log n}{\log n - \log 2}$. In fact, it requires

some effort to produce an explicit lower bound from this technique for a given n as several lower order terms will need to be calculated; for small n the bound it yields will not be as strong as the others. The proof relies on counting induced copies of K_5 in the random graph, and works even for the pairwise crossing number. The constant $\frac{1}{1,920}$ may not be optimal.

For the rectilinear crossing number the constant can be further improved to $c_{rg'} = \frac{\gamma}{64}$ where γ is the limit of the rectilinear crossing number of K_n divided by $\binom{n}{2}^2$. At present γ is not known exactly, but [6] shows that $\gamma > \frac{1+\epsilon}{16}$ for a small $\epsilon \approx 0.00003$. This yields $c_{rg'} > 0.000976$, but again may not produce effective bounds for particular small n.

4.6 Upper Bounds

The best known upper bound for the regular crossing number of K_n is $\mathrm{cr}(K_n) \leq g(n) := \frac{1}{4} \lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor \lfloor \frac{n-2}{2} \rfloor \lfloor \frac{n-3}{2} \rfloor$. There is a drawing that meets this bounds but it is not known to be tight for $n > 12$ and is an ongoing research challenge, see for instance [1, 12]. (For the rectilinear crossing number the best known drawing has slightly more crossings.) By using this drawing we get that $\mathrm{Eu}(n) \leq g(n)$.

In fact, we can improve this to $\mathrm{Eu}(n) \leq \frac{g(n)}{4}$ since the expectation on given crossing pair of edges e_1 and e_2 from the drawing is simply the expectation of the product of a pair of independent uniform $[0, 1]$ variables, i.e. $\int_0^1 \int_0^1 xy \mathbf{d}x \mathbf{d}y = \frac{1}{4}$. This is far from tight even for $n = 5$, where the bound is $\frac{1}{4}$, about 7.7 times the actual value of $\mathrm{Eu}(5)$. On the other hand, comparison with the lower bounds show that this ratio will not get much worse as n increases.

Appendix

Here we compute the remaining cases from Sect. 3.

Case 2: Orderings which ensure $\min(\mathbf{X}_1\mathbf{X}_8, \mathbf{X}_2\mathbf{X}_5, \mathbf{X}_3\mathbf{X}_7)$ can be attained at both $\mathbf{X}_1\mathbf{X}_8$ and $\mathbf{X}_2\mathbf{X}_5$, but not $\mathbf{X}_3\mathbf{X}_7$.

In these cases, $\mathbf{X}_2$ and $\mathbf{X}_5$ are between $\mathbf{X}_1$ and $\mathbf{X}_8$. However, $\mathbf{X}_3$ and $\mathbf{X}_7$ are not both between $\mathbf{X}_2$ and $\mathbf{X}_5$. The integrand will be $m(\mathbf{X}) = \min(\mathbf{X}_1\mathbf{X}_8, \mathbf{X}_2\mathbf{X}_5)$, and the

two smallest variables are X_1 and X_2. We break into subcases based on the positions of X_5 and X_8. Only the simplest case is described in detail.

Subcase 2i: The four smallest variables are X_1, X_2, X_5 and X_8. Then we need to evaluate:

$$\int_0^1 \int_0^{x_{i_{10}}} \cdots \int_0^{x_{i_6}} \int_0^{x_{i_5}} \int_0^{x_8} \int_0^{x_5} \int_0^{x_2} \min(x_1 x_8, x_2 x_5) dx_1 dx_2 dx_5 dx_8 dx_{i_5} \ldots dx_{i_9} dx_{i_{10}}$$

$$= \int_0^1 \int_0^{x_{i_{10}}} \cdots \int_0^{x_{i_6}} \int_0^{x_{i_5}} \int_0^{x_8} \int_0^{x_5} \int_0^{\frac{x_2 x_5}{x_8}} x_1 x_8 dx_1 dx_2 dx_5 dx_8 dx_{i_5} \ldots dx_{i_9} dx_{i_{10}}$$

$$+ \int_0^1 \int_0^{x_{i_{10}}} \cdots \int_0^{x_{i_6}} \int_0^{x_{i_5}} \int_0^{x_8} \int_0^{x_5} \int_{\frac{x_2 x_5}{x_8}}^{x_2} x_2 x_5 dx_1 dx_2 dx_5 dx_8 dx_{i_5} \ldots dx_{i_9} dx_{i_{10}}$$

$$= \int_0^1 \int_0^{x_{i_{10}}} \cdots \int_0^{x_{i_6}} \int_0^{x_{i_5}} \int_0^{x_8} \int_0^{x_5} \left(\frac{x_2^2 x_5^2}{2x_8} + x_2^2 x_5 - \frac{x_2^2 x_5^2}{x_8} \right) dx_2 dx_5 dx_8 dx_{i_5} \ldots dx_{i_9} dx_{i_{10}}$$

$$= \int_0^1 \int_0^{x_{i_{10}}} \cdots \int_0^{x_{i_6}} \int_0^{x_{i_5}} \int_0^{x_8} \left(\frac{x_5^4}{3} - \frac{x_5^5}{6x_8} \right) dx_5 dx_8 dx_{i_5} \ldots dx_{i_9} dx_{i_{10}}$$

$$= \int_0^1 \int_0^{x_{i_{10}}} \cdots \int_0^{x_{i_6}} \int_0^{x_{i_5}} \left(\frac{x_8^5}{15} - \frac{x_8^5}{36} \right) dx_8 dx_{i_5} \ldots dx_{i_9} dx_{i_{10}}$$

$$= \int_0^1 \int_0^{x_{i_{10}}} \cdots \int_0^{x_{i_6}} \frac{7 x_{i_5}^6}{180} dx_{i_5} \ldots dx_{i_9} dx_{i_{10}} = \int_0^1 \frac{7 \cdot 5! \, x_{i_{10}}^{11}}{180 \cdot 11!} dx_{i_{10}} = \frac{14}{3 \cdot 12!}.$$

The number of orderings of the variables in Subcase 2i up to symmetries is $\frac{6!}{2} = 360$, since we require $X_3 \leq X_4$.

The integrals in the remaining cases are essentially similar, so we will simply list the initial sequence of integrands and then compute the number of orderings of the variables corresponding to each case.

Subcase 2ii: The five smallest variables are X_1, X_2, X_5, X_j and X_8. This produces the following integrands:

$$\min(x_1 x_8, x_2 x_5); \, x_2^2 x_5 - \frac{x_2^2 x_5^2}{2x_8}; \, \frac{x_5^4}{3} - \frac{x_5^5}{6x_8}; \, \frac{x_j^5}{15} - \frac{x_j^6}{36x_8}; \, \frac{x_8^6}{140}; \, \frac{x_{i_6}^7}{140 \cdot 7}; \ldots; \, \frac{6! x_{i_{10}}}{140 \cdot 11!}; \, \frac{36}{7 \cdot 12!}.$$

There are three possibilities for j: 3, 6 and 7. When $j = 3$ we have 5! cases, and when $j = 6$ or $j = 7$ we have $\frac{5!}{2}$ cases as we need to account for the fact that $X_3 \leq X_4$ in the remaining variables. This is the total of 240 orderings.

Subcase 2iii: The six smallest variables are X_1, X_2, X_5, X_j, X_k and X_8. The integrands remain as in the previous case up to $\frac{x_j^5}{15} - \frac{x_j^6}{36x_8}$. The next integrands will be $\frac{x_k^6}{90} - \frac{x_k^7}{36 \cdot 7 x_8}$ and $\frac{11 x_8^7}{7 \cdot 1,440}$, the remaining simple integrations give $\frac{11}{2 \cdot 12!}$.

There are seven possibilities for (j, k) in this case: $(3, 4), (3, 6), (3, 7), (6, 3)$, $(7, 3), (6, 7)$ and $(7, 6)$. The first five of these are each associated to $4!$ orderings of the remaining variables, while the last two are each associated to $\frac{4!}{2}$. This is a total of 144 orderings. *Subcase 2iv*: The seven smallest variables are $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_5$, $\mathbf{X}_j$, $\mathbf{X}_k$, $\mathbf{X}_l$ and $\mathbf{X}_8$. We proceed from the integrand $\frac{x_k^6}{90} - \frac{x_k^7}{36 \cdot 7 x_8}$ to $\frac{x_l^7}{7 \cdot 90} - \frac{x_l^8}{7 \cdot 8 \cdot 36 x_8}$, and then to $\frac{13 x_8^8}{7 \cdot 8 \cdot 9 \cdot 180}$. Continuing to the end, the integral is $\frac{52}{9 \cdot 12!}$. There are 12 possibilities for (j, k, l) arising from choosing three of $3, 4, 6, 7$ and requiring 4 to be preceded by 3. Each of these has $3!$ orderings of the remaining three variables, for a total of 72 orderings.

Subcase 2v: The eight smallest variables are $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_5, \mathbf{X}_j, \mathbf{X}_k, \mathbf{X}_l, \mathbf{X}_m$ and $\mathbf{X}_8$. We proceed from the integrand $\frac{x_l^7}{7 \cdot 90} - \frac{x_l^8}{7 \cdot 8 \cdot 36 x_8}$ to $\frac{x_m^8}{7 \cdot 8 \cdot 90} - \frac{x_m^9}{7 \cdot 8 \cdot 9 \cdot 36 x_8}$ and then to $\frac{x_8^9}{7 \cdot 8 \cdot 9 \cdot 120}$. Continuing to the end, the integral is $\frac{6}{12!}$.

Again there are 12 possibilities, as (j, k, l, m) are chosen from $3, 4, 6, 7$ with 3 preceding 4. There are two ways of arranging the remaining two variables, for a total of 24 orderings.

Subcase 2vi: The five smallest variables are $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_j, \mathbf{X}_5$ and $\mathbf{X}_8$. The sequence of integrands that we see is then:

$$\min(x_1 x_8, x_2 x_5)\ ;\ x_2^2 x_5 - \frac{x_2^2 x_5^2}{2 x_8}; \frac{x_j^3 x_5}{3} - \frac{x_j^3 x_5}{6 x_8}; \frac{x_5^5}{12} - \frac{x_5^6}{24 x_8}; \frac{x_8^6}{126}; \frac{x_{i_6}^7}{126 \cdot 7};$$

$$\cdots \frac{6! x_{i_{10}}}{126 \cdot 11!}; \frac{40}{7 \cdot 12!}.$$

In this case, j must be either 3 or 7. If it is 3, there are $5!$ ways of ordering the remaining variables, and if it is 7 there are $\frac{5!}{2}$ ways of ordering the remaining variables, for a total of 180 orderings.

Subcase 2vii: The six smallest variables are $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_j, \mathbf{X}_5, \mathbf{X}_k$ and $\mathbf{X}_8$. We then see the same integrands through $\frac{x_5^5}{12} - \frac{x_5^6}{24 x_8}$ followed by $\frac{x_k^6}{72} - \frac{x_k^7}{7 \cdot 24 x_8}$, $\frac{x_8^7}{7 \cdot 72} - \frac{x_8^7}{7 \cdot 8 \cdot 24}$. Following the remaining routine integrations, we get $\frac{25}{4 \cdot 12!}$.

The possibilities for (j, k) are similar to those of Subcase 2iii, but 6 cannot be used in the first position. This leaves $(3, 4), (3, 6), (3, 7), (7, 3)$ and $(7, 6)$. The first four cases correspond to $4!$ orderings of the remaining variables, while $(7, 6)$ corresponds to $\frac{4!}{2}$ as 3 must precede 4. This gives a total of 108 orderings.

Subcase 2viii: The seven smallest variables are $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_j, \mathbf{X}_5, \mathbf{X}_k, \mathbf{X}_l$ and $\mathbf{X}_8$. We proceed as in the previous subcase through $\frac{x_k^6}{72} - \frac{x_k^7}{7 \cdot 24 x_8}$, then to $\frac{x_l^7}{7 \cdot 72} - \frac{x_l^8}{7 \cdot 8 \cdot 24 x_8}$, and $\frac{x_8^8}{7 \cdot 8 \cdot 72} - \frac{x_8^8}{7 \cdot 8 \cdot 9 \cdot 24}$. The remaining integrations bring us to $\frac{20}{3 \cdot 12!}$.

In this subcase, we have (j, k, l) chosen from $3, 4, 6, 7$ with the conditions that 4 must be preceded by 3, and 6 may not appear in the first position. This second

condition removes 3 of the 12 orderings as compared to Subcase 2iv, leaving us with 9. There are always 3! orderings of the remaining variables, giving a total of 54 orderings for this case.

Subcase 2ix: The eight smallest variables are $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_j, \mathbf{X}_5, \mathbf{X}_k, \mathbf{X}_l, \mathbf{X}_m$ and $\mathbf{X}_8$. We proceed as in the previous subcase through $\frac{x_l^7}{7 \cdot 72} - \frac{x_l^8}{7 \cdot 8 \cdot 24 x_8}$. The next integrand is $\frac{x_m^8}{7 \cdot 8 \cdot 72} - \frac{x_m^9}{7 \cdot 8 \cdot 9 \cdot 24 x_8}$, followed by $\frac{7 x_8^9}{24 \cdot 30 \cdot 7 \cdot 8 \cdot 9}$, and eventually $\frac{7}{12!}$.

There are nine possible choices for (j, k, l, m) since we have the same conditions as in the previous subcase, with the remaining number assigned to m. There are two orders for the remaining two variables, giving a total of 18 orderings.

Subcase 2x: The six smallest variables are $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_j, \mathbf{X}_k, \mathbf{X}_5$ and $\mathbf{X}_8$. We proceed as in the previous subcase through the integrand $\frac{x_j^3 x_5}{3} - \frac{x_j^3 x_5^2}{6 x_8}$. Continuing, we see integrands $\frac{x_k^4 x_5}{12} - \frac{x_k^4 x_5^2}{24 x_8}, \frac{x_5^6}{60} - \frac{x_5^7}{120 x_8}$, and $\frac{x_8^7}{7 \cdot 60} - \frac{x_8^7}{16 \cdot 60}$ on our way to $\frac{27}{4 \cdot 12!}$.

In fact, this case requires $j = 3$ and $k = 4$, since we can't have either $\mathbf{X}_6$ or both of $\mathbf{X}_3$ and $\mathbf{X}_7$ precede $\mathbf{X}_5$. There are $4! = 24$ ways of ordering the remaining variables.

Subcase 2xi: The seven smallest variables are $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5, \mathbf{X}_l$ and $\mathbf{X}_8$. This matches the previous subcase through $\frac{x_5^6}{60} - \frac{x_5^7}{120 x_8}$; the next two integrands are $\frac{x_l^7}{7 \cdot 60} - \frac{x_l^8}{8 \cdot 120 x_8}$ and $\frac{x_8^8}{7 \cdot 8 \cdot 60} - \frac{x_8^8}{8 \cdot 9 \cdot 120}$. Continuing we arrive at $\frac{22}{3 \cdot 12!}$.

We must have $l = 6$ or $l = 7$. There are 3! ways of ordering the remaining variables for a total of 12 orderings.

Subcase 2xii: The eight smallest variables are $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5, \mathbf{X}_l, \mathbf{X}_m$ and $\mathbf{X}_8$. This matches the previous subcase through $\frac{x_l^7}{7 \cdot 60} - \frac{x_l^8}{8 \cdot 120 x_8}$. Next we have $\frac{x_m^8}{7 \cdot 8 \cdot 60} - \frac{x_m^9}{8 \cdot 9 \cdot 120 x_8}$. and $\frac{x_8^9}{7 \cdot 8 \cdot 9 \cdot 60} - \frac{x_8^9}{8 \cdot 9 \cdot 10 \cdot 120}$. Continuing we arrive at $\frac{39}{5 \cdot 12!}$.

We must have (l, m) equal to $(6, 7)$ or $(7, 6)$, and there are two ways of ordering the remaining two variables, for a total of four orderings.

This completes Case 2, which contains 1,240 possible orderings up to the symmetries. The terms in (3) corresponding to these orderings of the variables have total weight per symmetry class of: $360 \cdot \frac{14}{3 \cdot 12!} + \ldots + 4 \cdot \frac{39}{5 \cdot 12!} = \frac{235{,}797}{35 \cdot 12!}$.

Case 3: Orderings which ensure $\min(\mathbf{X}_1 \mathbf{X}_8, \mathbf{X}_2 \mathbf{X}_5, \mathbf{X}_3 \mathbf{X}_7)$ is attained at both $\mathbf{X}_1 \mathbf{X}_8$ and $\mathbf{X}_3 \mathbf{X}_7$, but not $\mathbf{X}_2 \mathbf{X}_5$.

Since $\mathbf{X}_2$ is the second smallest variable, these will occur only when $\mathbf{X}_3$, $\mathbf{X}_7 \leq \mathbf{X}_8$, but $\mathbf{X}_8 \leq \mathbf{X}_5$. Only the simplest case is described in detail.

Subcase 3i: The five smallest variables are $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_7$ and $\mathbf{X}_8$ or $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_7, \mathbf{X}_3$ and $\mathbf{X}_8$.

We proceed to evaluate:

$$\int_0^1 \int_0^{x_{i_{10}}} \cdots \int_0^{x_{i_6}} \int_0^{x_8} \int_0^{x_7} \int_0^{x_3} \int_0^{x_2} \min(x_1 x_8, x_3 x_7) \mathbf{d}x_1 \mathbf{d}x_2 \mathbf{d}x_3 \mathbf{d}x_7 \mathbf{d}x_8 \ldots \mathbf{d}x_{i_9} \mathbf{d}x_{i_{10}}$$

The inner integral is piecewise linear in x_1, with a single break point at $x_1 = \frac{x_3 x_7}{x_8}$, however $\frac{x_3 x_7}{x_8}$ may or may not be greater than x_2. We decompose the inner integral as:

$$\int_0^{\min(x_2, \frac{x_3 x_7}{x_8})} x_1 x_8 \mathbf{d}x_1 + \int_{\min(x_2, \frac{x_3 x_7}{x_8})}^{x_2} x_3 x_7 \mathbf{d}x_1$$

Evaluating this integral leaves us with the new inner integral:

$$\int_0^{x_3} \left(\frac{1}{2} \min(x_2, \frac{x_3 x_7}{x_8})^2 x_8 + x_2 x_3 x_7 - \min(x_2, \frac{x_3 x_7}{x_8}) x_3 x_7 \right) \mathbf{d}x_2$$

This again needs to be split, this time with breakpoint at $x_2 = \frac{x_3 x_7}{x_8}$:

$$\int_0^{\frac{x_3 x_7}{x_8}} \left(\frac{x_2^2 x_8}{2} + x_2 x_3 x_7 - x_2 x_3 x_7 \right) \mathbf{d}x_2 + \int_{\frac{x_3 x_7}{x_8}}^{x_3} \left(\frac{x_3^2 x_7^2}{2 x_8} + x_2 x_3 x_7 - \frac{x_3^2 x_7^2}{x_8} \right) \mathbf{d}x_2$$

Happily, we see some cancellation of terms, both before evaluating the integral and after. This yields:

$$\int_0^{x_7} \left(\frac{x_3^3 x_7^3}{6 x_8^2} + \frac{x_3^3 x_7}{2} - \frac{x_3^3 x_7^2}{2 x_8} \right) \mathbf{d}x_3$$

We proceed through the following integrands:

$$\frac{x_7^7}{24 x_8^2} + \frac{x_7^5}{8} - \frac{x_7^6}{8 x_8}; \quad \frac{11 x_8^6}{1,344}; \quad \ldots; \quad \frac{165}{28 \cdot 12!}$$

Accounting for the fact that $\mathbf{X}_5 \leq \mathbf{X}_6$, there are $\frac{5!}{2}$ orderings of the remaining variables. With the two orderings of $\mathbf{X}_3$ and $\mathbf{X}_7$ (which do not affect the computation of the integral), we have the total of 120 orderings corresponding to this subcase.

Subcase 3ii: The six smallest variables are $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_7, \mathbf{X}_4$ and $\mathbf{X}_8$, with $\mathbf{X}_3$ and $\mathbf{X}_7$ possibly switched.

The integrands then remain identical through $\frac{x_7^7}{24 x_8^2} + \frac{x_7^5}{8} - \frac{x_7^6}{8 x_8}$. These are followed by $\frac{x_4^8}{8 \cdot 24 x_8^2} + \frac{x_4^6}{6 \cdot 8} - \frac{x_4^7}{7 \cdot 8 x_8}$ and $\frac{x_8^7}{756}$. Subsequent integrations yield $\frac{20}{3 \cdot 12!}$. There are $\frac{4!}{2}$ orderings of the remaining variables, and $\mathbf{X}_3$ and $\mathbf{X}_7$ can be switched, giving a total of 24 orderings.

Subcase 3iii: The six smallest variables are $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_7$ and $\mathbf{X}_8$.

The integrands are identical to Subcase 3i until $\frac{x_3^3 x_7^3}{6 x_8^2} + \frac{x_3^3 x_7}{2} - \frac{x_3^3 x_7^2}{2 x_8}$. We proceed to:

$$\frac{x_4^4 x_7^3}{24 x_8^2} + \frac{x_4^4 x_7}{8} - \frac{x_4^4 x_7^2}{8 x_8}; \quad \frac{x_7^8}{120 x_8^2} + \frac{x_7^6}{40} - \frac{x_7^7}{40 x_8}; \quad \frac{83 x_8^7}{60,480}; \quad \cdots \quad \frac{83}{12 \cdot 12!}$$

There are $\frac{4!}{2} = 12$ orderings of the remaining variables.

This completes Case 3, which comprises 156 orderings of the variables. The terms in (3) corresponding to these orderings of the variables have total weight per symmetry class of $\frac{6,651}{7 \cdot 12!}$.

Case 4: Orderings in which $\min(\mathbf{X}_1 \mathbf{X}_8, \mathbf{X}_2 \mathbf{X}_5, \mathbf{X}_3 \mathbf{X}_7)$ is attained at all three terms.

This comprises a small number of orderings that feature a messy inner integral. We note that in all these cases the two smallest variables are $\mathbf{X}_1$ and $\mathbf{X}_2$, while the third smallest variable is either $\mathbf{X}_3$ or $\mathbf{X}_7$. Our integrand is symmetric in $\mathbf{X}_3$ and $\mathbf{X}_7$, so we will do the computation only with $\mathbf{X}_3$ as the smaller of the two variables. We will proceed to evaluate the three innermost integrals before breaking into subcases, assuming that the fourth smallest variable is $\mathbf{X}_j$:

$$\int_0^{x_j} \int_0^{x_3} \int_0^{x_2} \min(x_1 x_8, x_2 x_5, x_3 x_7) \mathbf{d}x_1 \mathbf{d}x_2 \mathbf{d}x_3$$

$$= \int_0^{x_j} \int_0^{x_3} \int_0^{\frac{\min(x_2 x_5, x_3 x_7)}{x_8}} x_1 x_8 \mathbf{d}x_1 + \int_{\frac{\min(x_2 x_5, x_3 x_7)}{x_8}}^{x_2} \min(x_2 x_5, x_3 x_7) \mathbf{d}x_1 \mathbf{d}x_2 \mathbf{d}x_3$$

$$= \int_0^{x_j} \int_0^{x_3} \left(\frac{\min(x_2 x_5, x_3 x_7)^2}{2 x_8} + x_2 \min(x_2 x_5, x_3 x_7) - \frac{\min(x_2 x_5, x_3 x_7)^2}{x_8} \right) \mathbf{d}x_2 \mathbf{d}x_3$$

$$= \int_0^{x_j} \int_0^{x_3} \left(x_2 \min(x_2 x_5, x_3 x_7) - \frac{\min(x_2 x_5, x_3 x_7)^2}{2 x_8} \right) \mathbf{d}x_2 \mathbf{d}x_3$$

$$= \int_0^{x_j} \int_0^{\frac{x_3 x_7}{x_5}} \left(x_2^2 x_5 - \frac{x_2^2 x_5^2}{2 x_8} \right) \mathbf{d}x_2 \mathbf{d}x_3 + \int_{\frac{x_3 x_7}{x_5}}^{x_3} \left(x_2 x_3 x_7 - \frac{x_3^2 x_7^2}{2 x_8} \right) \mathbf{d}x_2 \mathbf{d}x_3$$

$$= \int_0^{x_j} \left(\frac{x_3^3 x_7^3 x_5}{3 x_5^3} - \frac{x_3^3 x_7^3 x_5^2}{6 x_8 x_5^3} + \frac{x_3^3 x_7}{2} - \frac{x_3^3 x_7^2}{2 x_8} - \frac{x_3^3 x_7^3}{2 x_5^2} + \frac{x_3^3 x_7^3}{2 x_8 x_5} \right) \mathbf{d}x_3$$

$$= \int_0^{x_j} x_3^3 \left[\frac{x_7}{2} - \frac{x_7^2}{2 x_8} - \frac{x_7^3}{6 x_5^2} + \frac{x_7^3}{3 x_8 x_5} \right] \mathbf{d}x_3 = \frac{x_j^4}{4} \left[\frac{x_7}{2} - \frac{x_7^2}{2 x_8} - \frac{x_7^3}{6 x_5^2} + \frac{x_7^3}{3 x_8 x_5} \right].$$

We now proceed to the subcases which are based on the ordering of the remaining variables.

Subcase 4i: The six smallest variables are X_1, X_2, X_3, X_7, X_5 and X_8. Then $X_j = X_7$ and we need to evaluate:

$$\int_0^1 \int_0^{x_{i_{10}}} \int_0^{x_{i_9}} \int_0^{x_{i_8}} \int_0^{x_{i_7}} \int_0^{x_8} \int_0^{x_5} \left(\frac{x_7^5}{8} - \frac{x_7^6}{8x_8} - \frac{x_7^7}{24x_5^2} + \frac{x_7^7}{12x_8x_5} \right)$$

$$\times \, dx_7 dx_5 dx_8 dx_{i_7} dx_{i_8} dx_{i_9} dx_{i_{10}}$$

$$= \frac{1}{32} \int_0^1 \int_0^{x_{i_{10}}} \int_0^{x_{i_9}} \int_0^{x_{i_8}} \int_0^{x_{i_7}} \int_0^{x_8} \left(\frac{x_5^6}{2} - \frac{5x_5^7}{21x_8} \right) dx_5 dx_8 dx_{i_7} dx_{i_8} dx_{i_9} dx_{i_{10}}$$

$$= \frac{1}{768} \int_0^1 \int_0^{x_{i_{10}}} \int_0^{x_{i_9}} \int_0^{x_{i_8}} \int_0^{x_{i_7}} x_8^7 dx_8 dx_{i_7} dx_{i_8} dx_{i_9} dx_{i_{10}} = \frac{7!}{768 \cdot 12!} = \frac{105}{16 \cdot 12!}.$$

As noted previously, there is a second ordering corresponding to this subcase, in which X_3 and X_7 are reversed, and there are 4! orderings of the remaining variables, giving us 48 orderings in this case.

Subcase 4ii: The seven smallest variables are $X_1, X_2, X_3, X_7, X_5, X_k$ and X_8.

This calculation is quite similar to the previous one until it reaches the integral with integrand $\frac{1}{32}[\frac{x_5^6}{2} - \frac{5x_5^7}{21x_8}]$. Subsequent integrands are:

$$\frac{1}{32}\left[\frac{x_k^7}{14} - \frac{5x_k^8}{168x_8} \right]; \frac{1}{32}\left[\frac{x_8^8}{112} - \frac{5x_8^9}{1,512x_8} \right]; \text{ and } \frac{1}{1,792}\left[\frac{17x_8^8}{54} \right].$$

The remaining integrations bring us to $\frac{85}{12 \cdot 12!}$.

There are 3! orderings of the remaining variables, k may be either 4 or 6, and X_3 and X_7 can again be reversed, giving us a total of 24 orderings in this subcase.

Subcase 4iii: The seven smallest variables are $X_1, X_2, X_3, X_7, X_4, X_5$ and X_8.

The first integration is similar to the first integration in Subcase 4i, and we proceed from there, via the following integrands:

$$\frac{1}{4}\left[\frac{x_4^6}{12} - \frac{x_4^7}{14x_8} - \frac{x_4^8}{48x_5^2} + \frac{x_4^8}{24x_8x_5} \right]; \frac{1}{12,096}\left[29x_5^7 - \frac{13x_5^8}{x_8} \right]; \frac{157x_8^8}{870,912}.$$

The remaining integrations bring us to $\frac{785}{108 \cdot 12!}$.

There are 3! orderings of the remaining variables, and X_3 and X_7 can be again be switched, giving 12 orderings in this subcase.

Subcase 4iv: The seven smallest variables are $X_1, X_2, X_3, X_4, X_7, X_5$ and X_8.
Unlike the previous cases, we have $X_j = X_4$, so we restart with just the inner three integrals evaluated at the top of the section:

$$\int_0^1 \int_0^{x_{i_{10}}} \int_0^{x_{i_9}} \int_0^{x_{i_8}} \int_0^{x_8} \int_0^{x_5} \int_0^{x_7} \frac{x_4^4}{4} \left[\frac{x_7}{2} - \frac{x_7^2}{2x_8} - \frac{x_7^3}{6x_5^2} + \frac{x_7^3}{3x_8 x_5} \right]$$

$$\times \, \mathbf{d}x_4 \mathbf{d}x_7 \mathbf{d}x_5 \mathbf{d}x_8 \mathbf{d}x_{i_8} \mathbf{d}x_{i_9} \mathbf{d}x_{i_{10}}$$

$$= \frac{1}{20} \int_0^1 \int_0^{x_{i_{10}}} \int_0^{x_{i_9}} \int_0^{x_{i_8}} \int_0^{x_8} \int_0^{x_5} \left(\frac{x_7^6}{2} - \frac{x_7^7}{2x_8} - \frac{x_7^8}{6x_5^2} + \frac{x_7^8}{3x_8 x_5} \right)$$

$$\times \, \mathbf{d}x_7 \mathbf{d}x_5 \mathbf{d}x_8 \mathbf{d}x_{i_8} \mathbf{d}x_{i_9} \mathbf{d}x_{i_{10}}$$

$$= \frac{1}{540} \int_0^1 \int_0^{x_{i_{10}}} \int_0^{x_{i_9}} \int_0^{x_{i_8}} \int_0^{x_8} \left(\frac{10x_5^7}{7} - \frac{11x_5^8}{16x_8} \right) \mathbf{d}x_5 \mathbf{d}x_8 \mathbf{d}x_{i_8} \mathbf{d}x_{i_9} \mathbf{d}x_{i_{10}}$$

$$= \frac{1}{540} \int_0^1 \int_0^{x_{i_{10}}} \int_0^{x_{i_9}} \int_0^{x_{i_8}} \frac{103x_8^8}{1{,}008} \mathbf{d}x_8 \mathbf{d}x_{i_8} \mathbf{d}x_{i_9} \mathbf{d}x_{i_{10}} = \frac{103 \cdot 8!}{544{,}320 \cdot 12!} = \frac{206}{27 \cdot 12!}.$$

There are 3! orderings of the remaining variables, and $\mathbf{X}_3$ cannot be interchanged with $\mathbf{X}_7$ due to the interceding $\mathbf{X}_4$.

Subcase 4v: The eight smallest variables are $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_7, \mathbf{X}_5, \mathbf{X}_k, \mathbf{X}_l$ and $\mathbf{X}_8$.

This follows Subcase 4ii until we arrive at integrand $\frac{1}{32} \left[\frac{x_k^7}{14} - \frac{5x_k^8}{168x_8} \right]$. We continue through integrands

$$\frac{1}{32} \left[\frac{x_l^8}{112} - \frac{5x_l^9}{1{,}512x_8} \right] \text{ and } \frac{x_8^9}{96{,}768}$$

and eventually to $\frac{4}{15 \cdot 12!}$.

There are two orderings of the remaining variables, (k, l) can be $(4, 6)$ or $(6, 4)$ and $\mathbf{X}_3$ and $\mathbf{X}_7$ may be reversed, giving a total of eight orderings in this subcase.

Subcase 4vi: The eight smallest variables are $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_7, \mathbf{X}_4, \mathbf{X}_5, \mathbf{X}_6$ and $\mathbf{X}_8$.

This follows Subcase 4iii until we arrive at the integrand $\frac{1}{12{,}096} \left[29x_5^7 - \frac{13x_5^8}{x_8} \right]$. Two more integrations bring integrands:

$$\frac{1}{12{,}096} \left[\frac{29x_6^8}{8} - \frac{13x_6^9}{9x_8} \right] \text{ and } \frac{1}{12{,}096} \left[\frac{31x_8^9}{120} \right].$$

Continuing we get $\frac{31}{4 \cdot 9!}$.

There are two orderings of the remaining variables, and $\mathbf{X}_3$ and $\mathbf{X}_7$ can be reversed, giving a total of four orderings in this subcase.

Subcase 4vii: The eight smallest variables are $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_7, \mathbf{X}_5, \mathbf{X}_6$ and $\mathbf{X}_8$.

This follows Subcase 4iv until we arrive at the integrand $\frac{1}{540}\left[\frac{10x_5^7}{7} - \frac{11x_5^8}{16x_8}\right]$. Continuing, we see:

$$\frac{1}{540}\left[\frac{10x_6^8}{56} - \frac{11x_6^9}{144x_8}\right] \text{ and } \frac{1}{540}\left[\frac{123x_8^9}{10,080}\right]$$

on our way to $\frac{41}{5\cdot 12!}$. In this subcase, $\mathbf{X}_3$ and $\mathbf{X}_7$ cannot be interchanged, and there are two orderings of the remaining two variables.

This completes Case 4, which contains the remaining 104 possible orderings of the variables, and the terms in (3) corresponding to these orderings of the variables have total weight per symmetry class of $\frac{3,627}{5\cdot 12!}$.

Summing over the four cases, the contributions of $\frac{13,140}{12!}$, $\frac{235,797}{35\cdot 12!}$, $\frac{6,651}{7\cdot 12!}$ and $\frac{3,627}{5\cdot 12!}$ respectively, give a total of $\frac{107,763}{5\cdot 12!}$ summed over the 5,040 symmetry class representatives in (3). Multiplying by the 720 symmetries of the variables, we find that Eu(5) $= \frac{35,921}{1,108,800} = 0.032396\overline{284271}$.

References

1. Aichholzer, O., Aurenhammer, F., Krasser, H.: On the crossing number of complete graphs. Computing **76**, 165–176 (2006). doi:10.1007/s00607-005-0133-3
2. Ajtai, M., Chvátal, V., Newborn, M.M., Szemerédi, E.: Crossing-free subgraphs. In: Turgeon, J., Rosa, A., Sabidussi, G. (eds.) Theory and Practice of Combinatorics. North-Holland Mathematics Studies, vol. 60, pp. 9–12. North-Holland, Amsterdam (1982)
3. David, H.A., Nagaraja, H.N.: Order Statistics. Wiley Series in Probability and Statistics, 3rd edn. Wiley-Interscience, Hoboken (2003)
4. Guy, R.K.: Crossing numbers of graphs. In: Graph Theory and Applications. Proceedings of the Conference, Western Michigan University, Kalamazoo, 1972, pp. 111–124. Lecture Notes in Mathematics, vol. 303. Springer, Berlin (1972)
5. Leighton, F.T.: New lower bound techniques for VLSI. Math. Syst. Theory **17**, 47–70 (1984). doi:10.1007/BF01744433
6. Lovász, L., Vesztergombi, K., Wagner, U., Welzl, E.: Convex quadrilaterals and k-sets. In: Pach, J. (ed.) Towards a Theory of Geometric Graphs. Contemporary Mathematics, vol. 342, pp. 139–148. American Mathematical Society, Providence (2004)
7. Mohar, B.: Crossing numbers of graphs on the plane and on other surfaces. In: Abstracts of the 20th Workshop on Topological Graph Theory in Yokohama, 25–28 Nov 2008, Yokohama (2008)
8. Mohar, B.: Do we really understand the crossing numbers? In: Hliněný, P., et al. (eds.) Mathematical Foundations of Computer Science 2010: Proceedings of the 35th International Symposium, MFCS 2010, Brno, 23–27 Aug 2010. Lecture Notes in Computer Science, vol. 6281, pp. 38–41. Springer, Berlin (2010)
9. Mohar, B., Stephen, T.: Expected crossing numbers. In: Nešetřil, J., Györi, E., Sali, A. (eds.) Proceedings of the European Conference on Combinatorics, Graph Theory and Applications, Budapest. Electronic Notes in Discrete Mathematics, vol. 38, pp. 651–656. Elsevier, Oxford (2011)
10. Pach, J., Tóth, G.: Which crossing number is it anyway? J. Comb. Theory Ser. B **80**(2), 225–246 (2000)

11. Pach, J., Radoičić, R., Tardos, G., Tóth, G.: Improving the crossing lemma by finding more crossings in sparse graphs. Discret. Comput. Geom. **36**, 527–552 (2006). doi:10.1007/s00454-006-1264-9
12. Pan, S., Richter, R.B.: The crossing number of K_{11} is 100. J. Graph Theory **56**, 128–134 (2007). doi:10.1002/jgt.20249
13. Richter, R.B., Salazar, G.: Crossing numbers. In: Beineke, L.W., Wilson, R.J. (eds.) Topics in Topological Graph Theory. Encyclopedia of Mathematics and Its Applications, vol. 128, pp. 133–150. Cambridge University Press, Cambridge (2009)
14. Shahrokhi, F., Székely, L., Vrt'o, I.: Crossing numbers of graphs, lower bound techniques and algorithms: a survey. In: Tamassia, R., Tollis, I. (eds.) Graph Drawing. Lecture Notes in Computer Science, vol. 894, pp. 131–142. Springer, Berlin/New York (1995). http://dx.doi.org/10.1007/3-540-58950-3_364
15. Spencer, J., Tóth, G.: Crossing numbers of random graphs. Random Struct. Algorithms **21**(3–4), 347–358 (2002)
16. Vrt'o, I.: Crossing numbers of graphs: a bibliography (2010). Available at: http://www.ifi.savba.sk/~imrich

EL-Labelings and Canonical Spanning Trees for Subword Complexes

Vincent Pilaud and Christian Stump

Abstract We describe edge labelings of the increasing flip graph of a subword complex on a finite Coxeter group, and study applications thereof. On the one hand, we show that they provide canonical spanning trees of the facet-ridge graph of the subword complex, describe inductively these trees, and present their close relations to greedy facets. Searching these trees yields an efficient algorithm to generate all facets of the subword complex, which extends the greedy flip algorithm for pointed pseudotriangulations. On the other hand, when the increasing flip graph is a Hasse diagram, we show that the edge labeling is indeed an EL-labeling and derive further combinatorial properties of paths in the increasing flip graph. These results apply in particular to Cambrian lattices, in which case a similar EL-labeling was recently studied by M. Kallipoliti and H. Mühle.

Key words Subword complexes • Increasing flips • Spanning trees • EL-labelings • Möbius function • Enumeration algorithm

Subject Classifications: 20F55, 06A07, 05C05, 68R05

V. Pilaud (✉)
CNRS & Laboratoire d'Informatique (LIX) École Polytechnique,
91128 Palaiseau Cedex, France
e-mail: vincent.pilaud@lix.polytechnique.fr

C. Stump
Institut für Algebra, Zahlentheorie, Diskrete Mathematik, Universität Hannover,
Room B402, Welfengarten 1, 30167 Hannover, Germany
e-mail: stump@math.uni-hannover.de

1 Introduction

Subword complexes on Coxeter groups were defined and studied by A. Knutson and E. Miller in the context of Gröbner geometry of Schubert varieties [7, 8]. Type A spherical subword complexes can be visually interpreted using pseudoline arrangements on primitive sorting networks. These were studied by V. Pilaud and M. Pocchiola [10] as combinatorial models for pointed pseudotriangulations of planar point sets [18] and for multitriangulations of convex polygons [11]. These two families of geometric graphs extend in two different ways the family of triangulations of a convex polygon.

The greedy flip algorithm was initially designed to generate all pointed pseudo-triangulations of a given set of points or convex bodies in general position in the plane [3, 13]. It was then extended in [10] to generate all pseudoline arrangements supported by a given primitive sorting network. The key step in this algorithm is to construct a spanning tree of the flip graph on the combinatorial objects, which has to be sufficiently canonical to be visited in polynomial time per node and polynomial working space.

In the present paper, we study natural edge lexicographic labelings of the increasing flip graph of a subword complex on any finite Coxeter group. As a first line of applications of these EL-labelings, we obtain canonical spanning trees of the flip graph of any subword complex. We provide alternative descriptions of these trees based on their close relations to greedy facets, which are defined and studied in this paper. Moreover, searching these trees provides an efficient algorithm to generate all facets of the subword complex. For type A spherical subword complexes, the resulting algorithm is that of [10], although the presentation is quite different.

The second line of applications of the EL-labelings concerns combinatorial properties ensuing from EL-shellability [1,2]. Indeed, when the increasing flip graph is the Hasse diagram of the increasing flip poset, this poset is EL-shellable, and we can compute its Möbius function. These results extend recent work of M. Kallipoliti and H. Mühle [6] on EL-shellability of N. Reading's Cambrian lattices [14–17], which are, for finite Coxeter groups, increasing flip posets of specific subword complexes studied by C. Ceballos, J.-P. Labbé and C. Stump [4] and by the authors in [12].

2 Edge Labelings of Graphs and Posets

In [1], A. Björner introduced EL-labelings of partially ordered sets to study topological properties of their order complexes. These labelings are edge labelings of the Hasse diagrams of the posets with certain combinatorial properties. In this paper, we consider edge labelings of finite, acyclic, directed graphs which might differ from the Hasse diagrams of their transitive closures.

2.1 ER-Labelings of Graphs and Associated Spanning Trees

Let $G := (V, E)$ be a finite, acyclic, directed graph. For $u, v \in V$, we write $u \to v$ if there is an edge from u to v in G, and $u \dashrightarrow v$ if there is a *path* $u = x_1 \to x_2 \to \cdots \to x_{\ell+1} = v$ from u to v in G (this path has *length* ℓ). The *interval* $[u, v]$ in G is the set of vertices $w \in V$ such that $u \dashrightarrow w \dashrightarrow v$.

An *edge labeling* of G is a map $\lambda : E \to \mathbb{N}$. It induces a labeling $\lambda(p)$ of any path $p : x_1 \to x_2 \to \cdots \to x_\ell \to x_{\ell+1}$ given by $\lambda(p) := \lambda(x_1 \to x_2) \cdots \lambda(x_\ell \to x_{\ell+1})$. The path p is *λ-rising* (resp. *λ-falling*) if $\lambda(p)$ is strictly increasing (resp. weakly decreasing). The labeling λ is an *edge rising labeling* of G (or *ER-labeling* for short) if there is a unique λ-rising path p between any vertices $u, v \in V$ with $u \dashrightarrow v$.

Remark 1 (Spanning trees). Let $u, v \in V$, and $\lambda : E \to \mathbb{N}$ be an ER-labeling of G. Then the union of all λ-rising paths from u to any other vertex of the interval $[u, v]$ forms a spanning tree of $[u, v]$, rooted at and directed away from u. We call it the *λ-source tree* of $[u, v]$ and denote it by $\mathbf{T}_\lambda([u, v])$. Similarly, the union of all λ-rising paths from any vertex of the interval $[u, v]$ to v forms a spanning tree of $[u, v]$, rooted at and directed towards v. We call it the *λ-sink tree* of $[u, v]$ and denote it by $\mathbf{T}_\lambda^*([u, v])$. In particular, if G has a unique source and a unique sink, this provides two canonical spanning trees $\mathbf{T}_\lambda(G)$ and $\mathbf{T}_\lambda^*(G)$ for the graph G itself.

Example 1 (Cube). Consider the 1-skeleton $\square_d$ of the d-dimensional cube $[0, 1]^d$, directed from $\mathbf{0} := (0, \dots, 0)$ to $\mathbf{1} := (1, \dots, 1)$. Its vertices are the elements of $\{0, 1\}^d$ and its edges are the pairs of vertices which differ in a unique position. Note that $\varepsilon := (\varepsilon_1, \dots, \varepsilon_d) \dashrightarrow \varepsilon' := (\varepsilon'_1, \dots, \varepsilon'_d)$ if and only if $\varepsilon_k \le \varepsilon'_k$ for all $k \in [d]$.

For any edge $\varepsilon \to \varepsilon'$ of $\square_d$, let $\lambda(\varepsilon \to \varepsilon')$ denote the unique position in $[d]$ where ε and ε' differ. Then the map λ is an ER-labeling of $\square_d$. If $\varepsilon \in \{0, 1\}^d \smallsetminus \mathbf{0}$, then the father of ε in $\mathbf{T}_\lambda(\square_d)$ is obtained from ε by changing its last 1 into a 0. Similarly, if $\varepsilon \in \{0, 1\}^d \smallsetminus \mathbf{1}$, then the father of ε in $\mathbf{T}_\lambda^*(\square_d)$ is obtained from ε by changing its first 0 into a 1. See Fig. 1.
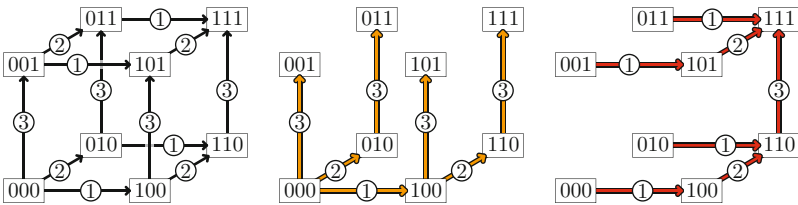
Fig. 1 An ER-labeling λ of the 1-skeleton $\square_3$ of the 3-cube, the λ-source tree $\mathbf{T}_\lambda(\square_3)$ and the λ-sink tree $\mathbf{T}_\lambda^*(\square_3)$

2.2   EL-Labelings of Graphs and Posets

Although ER-labelings of graphs are sufficient to produce canonical spanning trees, we need the following extension for further properties. The labeling $\lambda : E \rightarrow \mathbb{N}$ is an *edge lexicographic labeling* of G (or *EL-labeling* for short) if for any vertices $u, v \in V$ with $u \dashrightarrow v$,

 (i) There is a unique λ-rising path p from u to v, and
(ii) Its labeling $\lambda(p)$ is lexicographically first among the labelings $\lambda(p')$ of all paths p' from u to v.

For example, the ER-labeling of the 1-skeleton of the cube presented in Example 1 is in fact an EL-labeling.

   Remember now that one can associate a finite poset to a finite acyclic directed graph and vice versa. Namely,

 (i) The *transitive closure* of a finite acyclic directed graph $G = (V, E)$ is the finite poset $(V, \dashrightarrow)$;
(ii) The *Hasse diagram* of a finite poset P is the finite acyclic directed graph whose vertices are the elements of P and whose edges are the *cover relations* in P, i.e. $u \rightarrow v$ if $u <_P v$ and there is no $w \in P$ such that $u <_P w <_P v$.

The transitive closure of the Hasse diagram of P always coincides with P, but the Hasse diagram of the transitive closure of G might also be only a subgraph of G. An *EL-labeling* of the poset P is an EL-labeling of the Hasse diagram of P. If such a labeling exists, then the poset is called *EL-shellable*.

   As already mentioned, A. Björner [1] originally introduced EL-labelings of finite posets to study topological properties of their order complex. In particular, they provide a tool to compute the Möbius function of the poset. Recall that the *Möbius function* of the poset P is the map $\mu : P \times P \rightarrow \mathbb{Z}$ defined recursively by

$$\mu(u, v) := \begin{cases} 1 & \text{if } u = v, \\ -\sum_{u \leq_P w <_P v} \mu(u, w) & \text{if } u <_P v, \\ 0 & \text{otherwise.} \end{cases}$$

When the poset is EL-shellable, this function can be computed as follows.

Proposition 1 ([2, Proposition 5.7]). *Let λ be an EL-labeling of the poset P. For every $u, v \in P$ with $u \leq_P v$, we have*

$$\mu(u, v) = \text{even}_\lambda(u, v) - \text{odd}_\lambda(u, v),$$

where $\text{even}_\lambda(u, v)$ (resp. $\text{odd}_\lambda(u, v)$) denotes the number of even (resp. odd) length λ-falling paths from u to v in the Hasse diagram of P.

Example 2 (Cube). The directed 1-skeleton $\square_d$ of the d-dimensional cube $[0, 1]^d$ is the Hasse diagram of the boolean poset. The edge labeling λ of $\square_d$ of Example 1 is

thus an EL-labeling of the boolean poset. Moreover, for any two vertices $\varepsilon \dashrightarrow \varepsilon'$ of $\square_d$, there is a unique λ-falling path between ε and ε', whose length is the *Hamming distance* $\delta(\varepsilon, \varepsilon') := |\{k \in [d] \mid \varepsilon_k \neq \varepsilon_k'\}|$. The Möbius function is thus given by $\mu(\varepsilon, \varepsilon') = (-1)^{\delta(\varepsilon, \varepsilon')}$. In particular, $\mu(\mathbf{0}, \mathbf{1}) = (-1)^d$.

3 Subword Complexes on Coxeter Groups

3.1 Coxeter Systems

We recall some basic notions on Coxeter systems needed in this paper. More background material can be found in [5].

Let V be an n-dimensional Euclidean vector space. For $v \in V \smallsetminus 0$, we denote by s_v the reflection interchanging v and $-v$ while fixing pointwise the orthogonal hyperplane. We consider a *finite Coxeter group* W acting on V, i.e. a finite group generated by reflections. We assume without loss of generality that the intersection of all reflecting hyperplanes of W is reduced to 0.

A *root system* for W is a set Φ of vectors stable under the action of W and containing precisely two opposite vectors orthogonal to each reflection hyperplane of W. Fix a linear functional $f : V \to \mathbb{R}$ such that $f(\beta) \neq 0$ for all $\beta \in \Phi$. It splits the root system Φ into the set of *positive roots* $\Phi^+ := \{\beta \in \Phi \mid f(\beta) > 0\}$ and the set of *negative roots* $\Phi^- := -\Phi^+$. The *simple roots* are the roots which lie on the extremal rays of the cone generated by Φ^+. They form a basis Δ of the vector space V. The *simple reflections* $S := \{s_\alpha \mid \alpha \in \Delta\}$ generate the Coxeter group W. The pair (W, S) is a *finite Coxeter system*. For $s \in S$, we let $\alpha_s \in \Delta$ be the simple root orthogonal to the reflecting hyperplane of s.

The *length* of an element $w \in W$ is the length $\ell(w)$ of the smallest expression of w as a product of the generators in S. An expression $w = s_1 \cdots s_p$, with $s_1, \ldots, s_p \in S$, is *reduced* if $p = \ell(w)$. The length of w is also known to be the cardinality of the *inversion set* of w, defined as the set $\mathrm{inv}(w) := \Phi^+ \cap w(\Phi^-)$ of positive roots sent to negative roots by w^{-1}. Indeed, $\mathrm{inv}(w) = \{\alpha_{s_1}, s_1(\alpha_{s_2}), \ldots, s_1 \cdots s_{\ell-1}(\alpha_{s_\ell})\}$ for any reduced expression $w = s_1 \cdots s_\ell$ of w. The *(right) weak order* is the partial order on W defined by $u \leq w$ if there exists $v \in W$ with $uv = w$ and $\ell(u) + \ell(v) = \ell(w)$. In other words, $u \leq v$ if and only if $\mathrm{inv}(u) \subseteq \mathrm{inv}(v)$.

Example 3 (Type A—Symmetric groups). The symmetric group $\mathfrak{S}_{n+1}$, acting on the linear hyperplane $\mathbf{1}^\perp := \{x \in \mathbb{R}^{n+1} \mid \langle \mathbf{1} | x \rangle = 0\}$ by permutation of the coordinates, is the reflection group of *type A_n*. It is the group of isometries of the standard n-dimensional regular simplex $\mathrm{conv}\{e_1, \ldots, e_{n+1}\}$. Its reflections are the transpositions of $\mathfrak{S}_{n+1}$ and the set $\{e_i - e_j \mid i \neq j\}$ is a root system for A_n. We can choose the linear functional f such that the simple reflections are the adjacent transpositions $\tau_i := (i \ \ i+1)$, and the simple roots are the vectors $e_{i+1} - e_i$.

3.2 Subword Complexes

We consider a finite Coxeter system (W, S), a word $Q := q_1 q_2 \cdots q_m$ on the generators of S, and an element $\rho \in W$. A. Knutson and E. Miller [7] define the *subword complex* $\mathbf{SC}(Q, \rho)$ to be the simplicial complex of those subwords of Q whose complements contain a reduced expression for ρ as a subword. A vertex of $\mathbf{SC}(Q, \rho)$ is a position of a letter in Q. We denote by $[m] := \{1, 2, \ldots, m\}$ the set of positions in Q. A facet of $\mathbf{SC}(Q, \rho)$ is the complement of a set of positions which forms a reduced expression for ρ in Q. We denote by $\mathbf{F}(Q, \rho)$ the set of facets of $\mathbf{SC}(Q, \rho)$. We write $\rho \prec Q$ when Q contains a reduced expression of ρ, i.e. when $\mathbf{SC}(Q, \rho)$ is non-empty.

Example 4. Consider the type A Coxeter group $\mathfrak{S}_4$ generated by $\{\tau_1, \tau_2, \tau_3\}$. Let $Q^{\mathrm{ex}} := \tau_2 \tau_3 \tau_1 \tau_3 \tau_2 \tau_1 \tau_2 \tau_3 \tau_1$ and $\rho^{\mathrm{ex}} := [4, 1, 3, 2]$. The reduced expressions of ρ^{ex} are $\tau_2 \tau_3 \tau_2 \tau_1$, $\tau_3 \tau_2 \tau_3 \tau_1$, and $\tau_3 \tau_2 \tau_1 \tau_3$. Thus, the facets of the subword complex $\mathbf{SC}(Q^{\mathrm{ex}}, \rho^{\mathrm{ex}})$ are given by $\{1, 2, 3, 5, 6\}$, $\{1, 2, 3, 6, 7\}$, $\{1, 2, 3, 7, 9\}$, $\{1, 3, 4, 5, 6\}$, $\{1, 3, 4, 6, 7\}$, $\{1, 3, 4, 7, 9\}$, $\{2, 3, 5, 6, 8\}$, $\{2, 3, 6, 7, 8\}$, $\{2, 3, 7, 8, 9\}$, $\{3, 4, 5, 6, 8\}$, $\{3, 4, 6, 7, 8\}$, and $\{3, 4, 7, 8, 9\}$. Let $I^{\mathrm{ex}} := \{1, 3, 4, 7, 9\}$ and $J^{\mathrm{ex}} := \{3, 4, 7, 8, 9\}$ denote two facets of $\mathbf{SC}(Q^{\mathrm{ex}}, \rho^{\mathrm{ex}})$. We will use this example throughout this paper to illustrate further notions.

Example 5 (Type A—Primitive networks and pseudoline arrangements). For type A Coxeter systems, subword complexes can be visually interpreted using primitive networks. A *network* **N** is a collection of $n + 1$ horizontal lines (called *levels*, and labeled from bottom to top), together with m vertical segments (called *commutators*, and labeled from left to right) joining two different levels and such that no two of them have a common endpoint. We only consider *primitive* networks, where any commutator joins two consecutive levels. See Fig. 2 (left).

A *pseudoline* supported by the network **N** is an abscissa monotone path on **N**. A commutator of **N** is a *crossing* between two pseudolines if it is traversed by both pseudolines, and a *contact* if its endpoints are contained one in each pseudoline. A *pseudoline arrangement* Λ is a set of $n + 1$ pseudolines on **N**, any two of which have at most one crossing, possibly some contacts, and no other intersection. We label the pseudolines of Λ from bottom to top on the left of the network, and we define $\pi(\Lambda) \in \mathfrak{S}_{n+1}$ to be the permutation given by the order of these pseudolines on the right of the network. Note that the crossings of Λ correspond to the inversions of $\pi(\Lambda)$. See Fig. 2 (right).
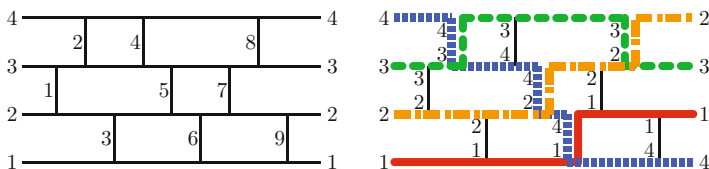

Fig. 2 The network $\mathbf{N}_{Q^{\mathrm{ex}}}$ (*left*) and the pseudoline arrangement $\Lambda_{I^{\mathrm{ex}}}$ for the facet $I^{\mathrm{ex}} = \{1, 3, 4, 7, 9\}$ of $\mathbf{SC}(Q^{\mathrm{ex}}, \rho^{\mathrm{ex}})$ (*right*)

Consider the type A Coxeter group $\mathfrak{S}_{n+1}$ generated by $S = \{\tau_i \mid i \in [n]\}$, where τ_i is the adjacent transposition $(i \quad i + 1)$. To a word $Q := q_1 q_2 \cdots q_m$ with m letters on S, we associate a primitive network $\mathbf{N_Q}$ with $n + 1$ levels and m commutators. If $q_j = \tau_p$, the jth commutator of $\mathbf{N_Q}$ is located between the pth and $(p + 1)$th levels of $\mathbf{N_Q}$. See Fig. 2 (left). For $\rho \in \mathfrak{S}_{n+1}$, a facet I of $\mathbf{SC}(Q, \rho)$ corresponds to a pseudoline arrangement Λ_I supported by $\mathbf{N_Q}$ and with $\pi(\Lambda_I) = \rho$. The positions of the contacts (resp. crossings) of Λ_I correspond to the positions of I (resp. of the complement of I). See Fig. 2 (right).

Example 6 (Combinatorial models for geometric graphs). As pointed out in [10], pseudoline arrangements on primitive networks give combinatorial models for the following families of geometric graphs (see Fig. 3):

  (i) Triangulations of convex polygons;
 (ii) Multitriangulations of convex polygons [11];
(iii) Pointed pseudotriangulations of points in general position in the plane [18];
(iv) Pseudotriangulations of disjoint convex bodies in the plane [13].

For example, consider a triangulation T of a convex $(n + 3)$-gon. Define the direction of a line of the plane to be the angle $\theta \in [0, \pi)$ of this line with the horizontal axis. Define also a bisector of a triangle $\triangle$ to be a line passing through a vertex of $\triangle$ and separating the other two vertices of $\triangle$. For any direction $\theta \in [0, \pi)$, each triangle of T has precisely one bisector in direction θ. We can thus order the $n + 1$ triangles of T according to the order π_θ of their bisectors in direction θ. The pseudoline arrangement associated to T is then given by the evolution of the order π_θ when the direction θ describes the interval $[0, \pi)$. A similar duality holds for the other three families of graphs, replacing triangles by the natural cells decomposing the geometric graph (stars for multitriangulations [11], or pseudotriangles for pseudotriangulations [18]), see Fig. 3. Details can be found in [10].
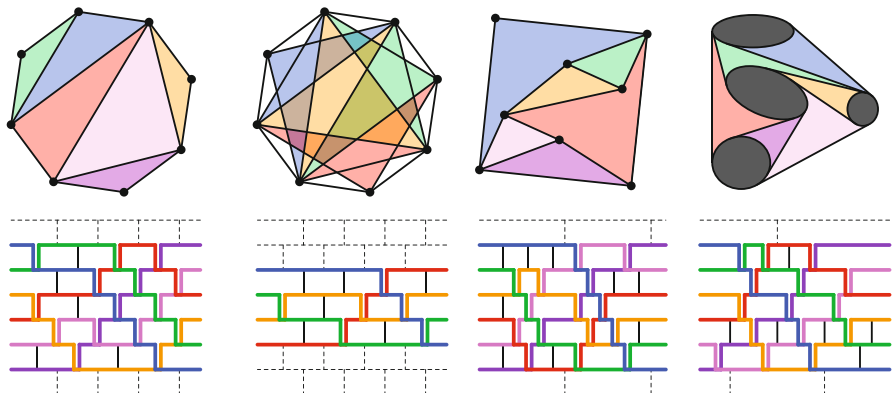

Fig. 3 Primitive sorting networks are combinatorial models for triangulations, multitriangulations, and pseudotriangulations of points or disjoint convex bodies

Remark 2. There is a natural reversal operation on subword complexes. Namely,

$$\mathbf{SC}(q_m \cdots q_1, \rho^{-1}) = \{\{m + 1 - i \mid i \in I\} \mid I \in \mathbf{SC}(q_1 \cdots q_m, \rho)\}.$$

We will use this operation to relate positive and negative labelings, facets and trees.

3.3 Inductive Structure

We denote by $Q_\vdash := q_2 \cdots q_m$ and $Q_\dashv := q_1 \cdots q_{m-1}$ the words on S obtained from $Q := q_1 \cdots q_m$ by deleting its first and last letters, respectively. We denote by $X^\rightarrow$ the right shift $\{x + 1 \mid x \in X\}$ of a subset X of $\mathbb{Z}$. For a collection $\mathbf{X}$ of subsets of $\mathbb{Z}$, we write $\mathbf{X}^\rightarrow$ for the set $\{X^\rightarrow \mid X \in \mathbf{X}\}$. Moreover, we denote by $\mathbf{X} \star z$ (or by $z \star \mathbf{X}$) the join $\{X \cup z \mid X \in \mathbf{X}\}$ of $\mathbf{X}$ with some $z \in \mathbb{Z}$. Remember that $\ell(\rho)$ denotes the length of ρ and that we write $\rho \prec Q$ when Q contains a reduced expression of ρ.

We can decompose inductively the facets of the subword complex $\mathbf{SC}(Q, \rho)$ depending on whether or not they contain the last letter of Q. Denoting by ε the empty word and by e the identity of W, we have $\mathbf{F}(\varepsilon, e) = \{\varnothing\}$ and $\mathbf{F}(\varepsilon, \rho) = \varnothing$ if $\rho \neq e$. Moreover, for a non-empty word Q on S, the set $\mathbf{F}(Q, \rho)$ is given by

(i) $\mathbf{F}(Q_\dashv, \rho q_m)$ if m appears in none of the facets of $\mathbf{SC}(Q, \rho)$ (i.e. if $\rho \not\prec Q_\dashv$);
(ii) $\mathbf{F}(Q_\dashv, \rho) \star m$ if m appears in all the facets of $\mathbf{SC}(Q, \rho)$ (i.e. if $\ell(\rho q_m) > \ell(\rho)$);
(iii) $\mathbf{F}(Q_\dashv, \rho q_m) \sqcup \left(\mathbf{F}(Q_\dashv, \rho) \star m\right)$ otherwise.

By reversal (see Remark 2), there is also a similar inductive decomposition of the facets of the subword complex $\mathbf{SC}(Q, \rho)$ depending on whether or not they contain the first letter of Q. Namely, for a non-empty word Q, the set $\mathbf{F}(Q, \rho)$ is given by

(i) $\mathbf{F}(Q_\vdash, q_1\rho)^\rightarrow$ if 1 appears in none of the facets of $\mathbf{SC}(Q, \rho)$ (i.e. if $\rho \not\prec Q_\vdash$);
(ii) $1 \star \mathbf{F}(Q_\vdash, \rho)^\rightarrow$ if 1 appears in all the facets of $\mathbf{SC}(Q, \rho)$ (i.e. if $\ell(q_1\rho) > \ell(\rho)$);
(iii) $\mathbf{F}(Q_\vdash, q_1\rho)^\rightarrow \sqcup \left(1 \star \mathbf{F}(Q_\vdash, \rho)^\rightarrow\right)$ otherwise.

Although we will only use these decompositions for the facets $\mathbf{F}(Q, \rho)$, they extend to the whole subword complex $\mathbf{SC}(Q, \rho)$ and are used to obtain the following result.

Theorem 1 ([7, Corollary 3.8]). *The subword complex* $\mathbf{SC}(Q, \rho)$ *is either a simplicial sphere or a simplicial ball.*

3.4 Flips and Roots

Let I be a facet of $\mathbf{SC}(Q, \rho)$ and i be a position in I. If there exists a facet J of $\mathbf{SC}(Q, \rho)$ and a position $j \in J$ such that $I \smallsetminus i = J \smallsetminus j$, we say that I and J are *adjacent* facets, that i is *flippable* in I, and that J is obtained from I by *flipping* i. Note that, if they exist, J and j are unique by Theorem 1. We say that the flip from I to J *flips out* i and *flips in* j.
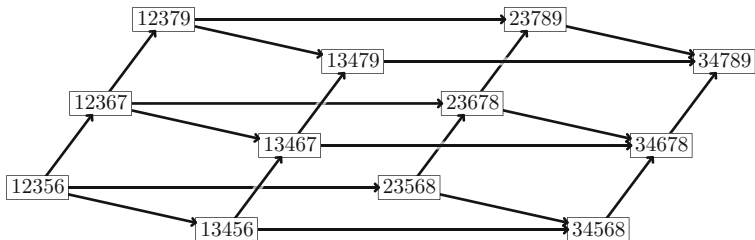
Fig. 4 The increasing flip graph $\mathbf{G}(Q^{ex}, \rho^{ex})$

We denote by $\mathbf{G}(Q, \rho)$ the graph of flips, whose vertices are the facets of $\mathbf{SC}(Q, \rho)$ and whose edges are pairs of adjacent facets. That is, $\mathbf{G}(Q, \rho)$ is the ridge graph of the simplicial complex $\mathbf{SC}(Q, \rho)$. This graph is connected according to Theorem 1.

This graph can be naturally oriented by the direction of the flips as follows. Let I and J be two adjacent facets of $\mathbf{SC}(Q, \rho)$ with $I \smallsetminus i = J \smallsetminus j$. We say that the flip from I to J is *increasing* if $i < j$. We consider the flip graph $\mathbf{G}(Q, \rho)$ oriented by increasing flips.

Example 7. Figure 4 represents the increasing flip graph $\mathbf{G}(Q^{ex}, \rho^{ex})$ for the subword complex $\mathbf{SC}(Q^{ex}, \rho^{ex})$ of Example 4. The facets of $\mathbf{SC}(Q^{ex}, \rho^{ex})$ appear in lexicographic order from left to right. Thus, all flips are increasing from left to right.

Remark 3. The increasing flip graph of $\mathbf{SC}(Q, \rho)$ was already considered by A. Knutson and E. Miller [7, Remark 4.5]. It carries various combinatorial informations about the subword complex $\mathbf{SC}(Q, \rho)$. In particular, since the lexicographic ordering of the facets of $\mathbf{SC}(Q, \rho)$ is a shelling order for $\mathbf{SC}(Q, \rho)$, the *h*-vector of the subword complex $\mathbf{SC}(Q, \rho)$ is the in-degree sequence of the increasing flip graph $\mathbf{G}(Q, \rho)$.

Throughout the paper, we consider flips as elementary operations on subword complexes. In practice, the necessary information to perform flips in a facet I of $\mathbf{SC}(Q, \rho)$ is encoded in its root function $\mathsf{r}(I, \cdot) : [m] \to \Phi$ defined by

$$\mathsf{r}(I, k) := \Pi Q_{[k-1] \smallsetminus I}(\alpha_{q_k}),$$

where ΠQ_X denotes the product of the reflections $q_x \in Q$ for $x \in X$. The *root configuration* of the facet I is the multiset $\mathsf{R}(I) := \{\!\{\mathsf{r}(I, i) \mid i \text{ flippable in } I\}\!\}$. The root function was introduced by C. Ceballos, J.-P. Labbé and C. Stump [4], and we extensively studied root configurations in [12] in the construction of brick polytopes for spherical subword complexes. The main properties of the root function are summarized in the following proposition, whose proof is similar to that of [4, Lemmas 3.3 and 3.6] or [12, Lemma 3.3].
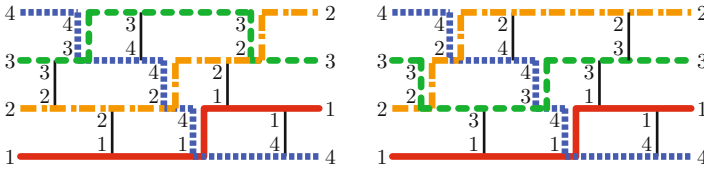
Fig. 5 The increasing flip from facet $I^{\mathrm{ex}} = \{1, 3, 4, 7, 9\}$ to facet $J^{\mathrm{ex}} = \{3, 4, 7, 8, 9\}$ of the subword complex $\mathbf{SC}(Q^{\mathrm{ex}}, \rho^{\mathrm{ex}})$, illustrated on the network $\mathbf{N}_{Q^{\mathrm{ex}}}$

Proposition 2. *Let I be any facet of the subword complex $\mathbf{SC}(Q, \rho)$.*

1. *The map $\mathsf{r}(I, \cdot) : i \mapsto \mathsf{r}(I, i)$ is a bijection from the complement of I to the inversion set of ρ.*
2. *The map $\mathsf{r}(I, \cdot)$ sends the flippable positions in I to $\{\pm \beta \mid \beta \in inv(\rho)\}$ and the unflippable ones to $\Phi^+ \smallsetminus inv(\rho)$.*
3. *If I and J are two adjacent facets of $\mathbf{SC}(Q, \rho)$ with $I \smallsetminus i = J \smallsetminus j$, the position j is the unique position in the complement of I for which $\mathsf{r}(I, j) \in \{\pm \mathsf{r}(I, i)\}$.*
4. *In the situation of (3), we have $\mathsf{r}(I, i) = \mathsf{r}(I, j) \in \Phi^+$ if $i < j$ (increasing flip), while $\mathsf{r}(I, i) = -\mathsf{r}(I, j) \in \Phi^-$ if $i > j$ (decreasing flip).*
5. *In the situation of (3), the map $\mathsf{r}(J, \cdot)$ is obtained from the map $\mathsf{r}(I, \cdot)$ by:*

$$\mathsf{r}(J, k) = \begin{cases} s_{\mathsf{r}(I, i)}(\mathsf{r}(I, k)) & \text{if } \min(i, j) < k \leq \max(i, j), \\ \mathsf{r}(I, k) & \text{otherwise.} \end{cases}$$

We call $\mathsf{r}(I, i) = -\mathsf{r}(J, j)$ the *direction* of the flip from the facet I to the facet J.

Example 8. In type A, roots and flips can easily be described using the primitive network interpretation presented in Example 5. Consider a word Q on the simple reflections $\{\tau_i \mid i \in [n]\}$, an element $\rho \in \mathfrak{S}_{n+1}$, and a facet I of $\mathbf{SC}(Q, \rho)$. For any $k \in [m]$, the root $\mathsf{r}(I, k)$ is the difference $e_t - e_b$ where t and b are the indices of the pseudolines of Λ_I which arrive respectively on the top and bottom endpoints of the kth commutator of $\mathbf{N}_Q$. A flip exchanges a contact between two pseudolines t and b of Λ_I with the unique crossing between t and b in Λ_I (when it exists). Such a flip is increasing if the contact lies before the crossing, i.e. if $t > b$. Figure 5 illustrates the properties of Proposition 2 on the subword complex $\mathbf{SC}(Q^{\mathrm{ex}}, \rho^{\mathrm{ex}})$ of Example 4.

3.5 Restriction of Subword Complexes to Parabolic Subgroups

In the proof of our main result, we will need to restrict subword complexes to dihedral parabolic subsystems. The following statement can essentially be found in [12, Proposition 3.7], we provide a proof here as well for the sake of completeness.

Proposition 3 ([12, Proposition 3.7]). *Let* $\mathbf{SC}(Q, \rho)$ *be a subword complex for a Coxeter system* (W, S) *acting on* V, *and let* $V' \subseteq V$ *be a subspace of* V. *The simplicial complex given by all facets* J *of* $\mathbf{SC}(Q, \rho)$ *reachable from a particular facet* I *by flips whose directions are contained in* V' *is isomorphic to a subword complex* $\mathbf{SC}(Q', \rho')$ *for the restriction of* (W, S) *to* V'. *The order of the letters is preserved by this isomorphism. In particular, the restriction of the increasing flip graph* $\mathbf{G}(Q, \rho)$ *to these reachable facets is isomorphic to the increasing flip graph* $\mathbf{G}(Q', \rho')$.

Proof. To prove this proposition, we explicitly construct the word Q' on S' and the element $\rho' \in W'$ where (W', S') is the restriction of (W, S) to the subspace V'.

First, the element ρ' only depends on ρ and on V': it is given by the projection of ρ onto W'. This is to say that ρ' is the unique element in W' whose inversion set is $\mathrm{inv}(\rho') = \mathrm{inv}(\rho) \cap V'$. To see that $\mathrm{inv}(\rho) \cap V'$ is again an inversion set, remember that a subset $\mathbf{I}$ of Φ^+ is an inversion set for an element in W if and only if for all $\alpha, \beta, \gamma \in \Phi^+$ such that $\gamma = a\alpha + b\beta$ for some $a, b \in \mathbb{R}_{\geq 0}$,

$$\alpha, \beta \in \mathbf{I} \implies \gamma \in \mathbf{I} \implies (\alpha \in \mathbf{I} \text{ or } \beta \in \mathbf{I}),$$

see e.g. [9]. Moreover, this property is preserved under intersection with linear subspaces.

We now construct the word Q' and the facet I' of $\mathbf{SC}(Q', \rho')$ corresponding to the particular facet I of $\mathbf{SC}(Q, \rho)$. For this, let $X := \{x_1, \ldots, x_p\}$ be the set of positions $k \in [m]$ such that $\mathsf{r}(I, k) \in V'$. The word Q' has p letters corresponding to the positions in X, and the facet I' contains precisely the positions $k \in [p]$ such that the position x_k is in I. To construct the word Q', we scan Q from left to right as follows. We initialize Q' to the empty word, and for each $1 \leq k \leq p$, we add a letter $\mathsf{q}'_k \in S'$ to Q' in such a way that $\mathsf{r}(I', k) = \mathsf{r}(I, x_k)$. To see that such a letter exists, we distinguish two cases. Assume first that $\mathsf{r}(I, x_k)$ is a positive root. Let $\mathbf{I}$ be the inversion set of $w := \Pi Q_{[x_k-1] \smallsetminus I}$ and $\mathbf{I}' = \mathbf{I} \cap V'$ be the inversion set of $w' := \Pi Q'_{[k-1] \smallsetminus I'}$. Then the set $\mathbf{I}' \cup \{\mathsf{r}(I, x_k)\}$ is again an inversion set (as the intersection of V' with the inversion set $\mathbf{I} \cup \{\mathsf{r}(I, x_k)\}$ of $w q_{x_k}$) which contains the inversion set $\mathscr{I}'$ of w' together with a unique additional root. Therefore, the corresponding element of W' can be written as $w'q'_k$ for some simple reflection $q'_k \in S'$. Assume now that $\mathsf{r}(I, x_k)$ is a negative root. Then $x_k \in I$, so that we can flip it with a position $x_{k'} < x_k$, and we can then argue on the resulting facet.

By the procedure described above, we eventually obtain the subword complex $\mathbf{SC}(Q', \rho')$ and its facet I' corresponding to the facet I. Finally observe that sequences of flips in $\mathbf{SC}(Q, \rho)$ starting at the facet I, and whose directions are contained in V', correspond bijectively to sequences of flips in $\mathbf{SC}(Q', \rho')$ starting at the facet I'. In particular, let J and J' be two facets reached from I and from I', respectively, by such a sequence. We then have that the root configuration of J' is exactly the root configuration of J intersected with V', and that the order in which the roots appear in the root configurations is preserved. This completes the proof.
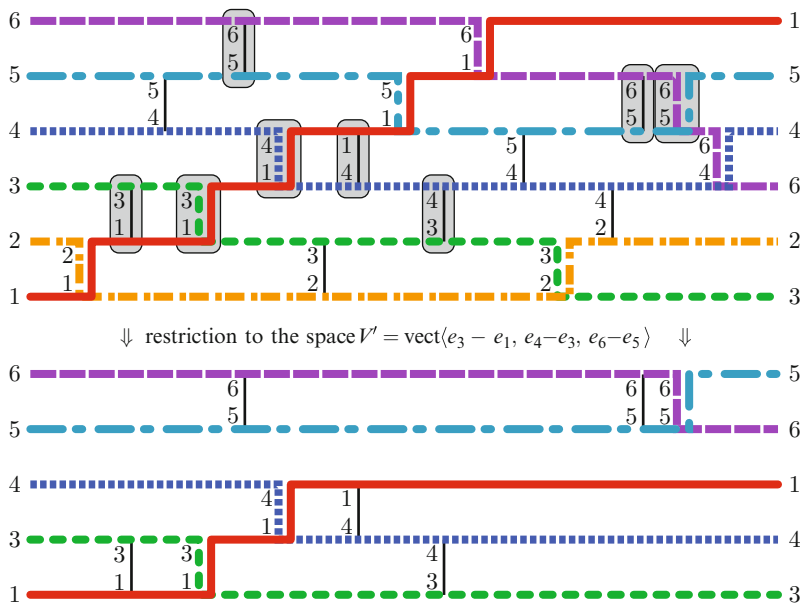
Fig. 6 Restricting subword complexes

Example 9. To illustrate different possible situations happening in this restriction, we consider the subword complex $\mathbf{SC}(Q, \rho)$ on the Coxeter group $A_5 = \mathfrak{S}_6$ generated by $S = \{\tau_1, \ldots, \tau_5\}$, the word $Q := \tau_1 \tau_2 \tau_4 \tau_2 \tau_5 \tau_3 \tau_1 \tau_3 \tau_4 \tau_2 \tau_5 \tau_3 \tau_1 \tau_2 \tau_4 \tau_4 \tau_3$ and the element $\rho := [3, 2, 6, 4, 5, 1] = \tau_1 \tau_2 \tau_3 \tau_4 \tau_5 \tau_1 \tau_4 \tau_3$. The sorting network corresponding to the subword complex $\mathbf{SC}(Q, \rho)$ and the pseudoline arrangement corresponding to the facet $I := \{2, 3, 5, 7, 8, 10, 12, 14, 15\}$ of $\mathbf{SC}(Q, \rho)$ are shown in Fig. 6 (top). Let V' be the subspace of V spanned by the roots $e_3 - e_1$, $e_4 - e_3$ and $e_6 - e_5$. Let $X = \{x_1, \ldots, x_8\} = \{2, 4, 5, 6, 8, 10, 15, 16\}$ denote the set of positions $k \in [17]$ for which $\mathsf{r}(I, k) \in V'$. These positions are circled in Fig. 6 (top).

We can now directly read off the subword complex $\mathbf{SC}(Q', \rho')$ corresponding to the restriction of $\mathbf{SC}(Q, \rho)$ to all facets reachable from I by flips with directions in V'. Namely, the restriction of (W, S) to V' is the Coxeter system (W', S') where W' is generated by $S' = \{\tau_1', \tau_2', \tau_3'\} = \{(1\ 3), (3\ 4), (5\ 6)\}$, and thus of type $A_2 \times A_1$. Moreover, we have $Q' = \tau_1' \tau_1' \tau_3' \tau_2' \tau_1' \tau_3' \tau_3'$, corresponding to the roots at positions in X, and $\rho' = \tau_1' \tau_2' \tau_3'$, with inversion set given by the positive roots corresponding to the roots at positions in $X \smallsetminus I$. Finally, the facet I' corresponding to I is given by $I' = \{1, 3, 5, 6, 7\}$. The sorting network corresponding to the restricted subword complex $\mathbf{SC}(Q', \rho')$ and the pseudoline arrangement corresponding to the facet I' of $\mathbf{SC}(Q', \rho')$ are shown in Fig. 6 (bottom).

As stated in Proposition 3, the map which sends a facet J of $\mathbf{SC}(Q', \rho')$ to the facet $\{x_j \mid j \in J\} \cup (I \smallsetminus X)$ of $\mathbf{SC}(Q, \rho)$ defines an isomorphism between the increasing flip graph $\mathbf{G}(Q', \rho')$ and the restriction of the increasing flip graph $\mathbf{G}(Q, \rho)$ to all facets reachable from I by flips with directions in V'.

4 EL-Labelings and Spanning Trees for the Subword Complex

4.1 EL-Labelings of the Increasing Flip Graph

We now define two natural edge labelings of the increasing flip graph $\mathbf{G}(Q, \rho)$.

Let I and J be two adjacent facets of $\mathbf{SC}(Q, \rho)$, with $I \smallsetminus i = J \smallsetminus j$ and $i < j$. We label the edge $I \to J$ of $\mathbf{G}(Q, \rho)$ with the positive edge label $\mathsf{p}(I \to J) := i$ and with the negative edge label $\mathsf{n}(I \to J) := j$. In other words, p labels the position flipped out while n labels the position flipped in during the flip $I \to J$. We call $\mathsf{p} : E(\mathbf{G}(Q, \rho)) \to [m]$ the *positive edge labeling* and $\mathsf{n} : E(\mathbf{G}(Q, \rho)) \to [m]$ the *negative edge labeling* of the increasing flip graph $\mathbf{G}(Q, \rho)$. The terms "positive" and "negative" emphasize the fact that the roots $\mathsf{r}(I, \mathsf{p}(I \to J))$ and $\mathsf{r}(J, \mathsf{n}(I \to J))$ are always positive and negative roots respectively.

The positive and negative edge labelings are reverse to one another (see Remark 2). Namely, $I \to J$ is an edge in the increasing flip graph $\mathbf{G}(\mathsf{q}_m \cdots \mathsf{q}_1, \rho^{-1})$ if and only if $J' := \{m + 1 - j \mid j \in J\} \to I' := \{m + 1 - i \mid i \in I\}$ is an edge in the increasing flip graph $\mathbf{G}(\mathsf{q}_1 \cdots \mathsf{q}_m, \rho)$, and in this case $\mathsf{n}(I \to J) = m + 1 - \mathsf{p}(J' \to I')$. However, we will work in parallel with both labelings, since we believe that certain results are simpler to present on the positive side while others are simpler on the negative side. We always provide proofs on the easier side and leave it to the reader to translate to the opposite side.

Example 10. Consider the subword complex $\mathbf{SC}(Q^{\mathrm{ex}}, \rho^{\mathrm{ex}})$ of Example 4. We have represented on Fig. 7 the positive and negative edge labelings p and n. Since we have represented the graph $\mathbf{G}(Q^{\mathrm{ex}}, \rho^{\mathrm{ex}})$ such that the flips are increasing from left to right, each edge has its positive label on the left and its negative label on the right.

The central result of this paper concerns the positive and negative edge labelings of the increasing flip graph.

Theorem 2. *The positive edge labeling p and the negative edge labeling n are both EL-labelings of the increasing flip graph.*
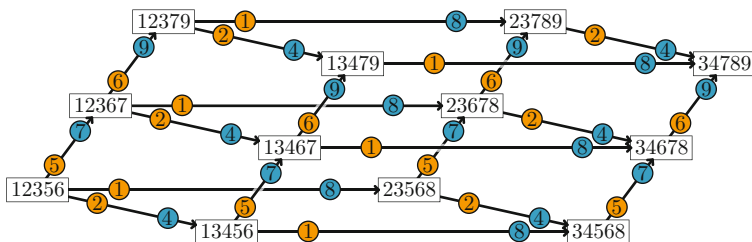

Fig. 7 The positive and negative edge labelings p and n of $\mathbf{G}(Q^{\mathrm{ex}}, \rho^{\mathrm{ex}})$. Each edge has its positive label on the left (*orange*) and its negative label on the right (*blue*)

For Cambrian lattices, whose Hasse diagrams were shown to be particular cases of increasing flip graphs in [12, Sect. 6], a similar result was recently obtained by M. Kallipoliti and H. Mühle in [6]. See Sect. 5.2.1 for details.

In Sects. 4.2–4.4, we present applications of Theorem 2 to the construction of canonical spanning trees and to the generation of the facets of the subword complex. Further combinatorial applications of this theorem are also discussed in Sect. 5. We prove Theorem 2 only for the positive edge labeling p, and leave it to the reader to translate the proof to the negative edge labeling n (through the reversal operation of Remark 2). Let I and J be two facets of $\mathbf{SC}(Q, \rho)$ such that $I \dashrightarrow J$. To show that p is indeed an EL-labeling, we have to show that (i) there is a p-rising path from I to J in $\mathbf{G}(Q, \rho)$ which is (ii) unique and (iii) lexicographically first among all paths from I to J in $\mathbf{G}(Q, \rho)$. We start with (ii) and (iii), which are direct consequences of the following proposition.

Proposition 4. *Let* $I_1 \to \cdots \to I_{\ell+1}$ *be a path of increasing flips, and define the labels* $\mathsf{p}_k := \mathsf{p}(I_k \to I_{k+1})$ *and* $\mathsf{n}_k := \mathsf{n}(I_k \to I_{k+1})$. *Then, for all* $k \in [\ell]$, *we have*

$$\min\{\mathsf{p}_k, \ldots, \mathsf{p}_\ell\} = \min(I_k \smallsetminus I_{\ell+1}) \quad and \quad \max\{\mathsf{n}_1, \ldots, \mathsf{n}_k\} = \max(I_{k+1} \smallsetminus I_1).$$

Moreover, the path is p*-rising if and only if* $\mathsf{p}_k = \min(I_k \smallsetminus I_{\ell+1})$ *for all* $k \in [\ell]$, *while the path is* n*-rising if and only if* $\mathsf{n}_k = \max(I_{k+1} \smallsetminus I_1)$ *for all* $k \in [\ell]$.

Proof. The position $\min\{\mathsf{p}_k, \ldots, \mathsf{p}_\ell\}$ is in $I_k \smallsetminus I_{\ell+1}$ since it is flipped out and never flipped in along the path from I_k to $I_{\ell+1}$ (because all flips are increasing). Moreover, $\min\{\mathsf{p}_k, \ldots, \mathsf{p}_\ell\}$ has to coincide with $\min(I_k \smallsetminus I_{\ell+1})$ otherwise this position would never be flipped out along the path.

This property immediately yields the characterization of p-rising paths. Indeed, if the path is p-rising, then we have $\mathsf{p}_k = \min(\mathsf{p}_k, \ldots, \mathsf{p}_\ell) = \min(I_k \smallsetminus I_{\ell+1})$ for all $k \in [\ell]$. Reciprocally, if $\mathsf{p}_k = \min(I_k \smallsetminus I_{\ell+1})$ for all $k \in [\ell]$, then we have $\mathsf{p}_k = \min(I_k \smallsetminus I_{\ell+1}) < \min(I_{k+1} \smallsetminus I_{\ell+1}) = \mathsf{p}_{k+1}$ so that the path is p-rising. The proof is similar for the negative edge labeling n. $\qquad\square$

We now need to prove the existence of a p-rising path from I to J. Before proving it in full generality, we prove its crucial part in the particular case of dihedral subword complexes.

Lemma 1. *Let* $\mathbf{SC}(Q, \rho)$ *be a subword complex for a dihedral reflection group* $W = I_2(m)$. *Let* I *and* K *be two of its facets such that there is a path* $I \to J \to K$ *from* I *to* K *in* $\mathbf{G}(Q, \rho)$ *with* $\mathsf{p}(I \to J) > \mathsf{p}(J \to K)$. *Then there is as well a* p*-rising path from* I *to* K *in* $\mathbf{G}(Q, \rho)$.

Proof. First, we remark that we construct a path only using letters in Q at positions not used in I (those positions corresponding to the reduced expression for ρ), together with the two positions $i := \mathsf{p}(I \to J)$ and $j := \mathsf{p}(J \to K)$. Observe here that both i and j are already contained in I.

We distinguish two cases: the roots $\mathsf{r}(I, i)$ and $\mathsf{r}(I, j)$ generate either a 1- or a 2-dimensional space. In the first case, we have $\mathsf{r}(I, i) = \mathsf{r}(I, j)$ and we can

directly flip position j in the facet I to obtain the facet K. In the second case, it is straightforward to check that we can perform a p-rising path from I to K, starting with position j, followed by position i, and finishing by a possibly empty p-rising sequence of flips.

We are now ready to prove Theorem 2. Restricting subword complexes to dihedral parabolic subgroups as presented in Sect. 3.5, we will reduce the general case to several applications of the dihedral situation treated in Lemma 1.

Proof (Proof of Theorem 2). Let I and J be two facets of $\mathbf{SC}(Q, \rho)$ related by a path $I = I_1 \to \cdots \to I_{\ell+1} = J$ of increasing flips. Let $\mathsf{p}_k := \mathsf{p}(I_k \to I_{k+1})$. Assume that this path is not p-rising, and let k be the smallest index such that $\mathsf{p}_k \neq \min\{\mathsf{p}_k, \ldots, \mathsf{p}_\ell\}$, and let $k' > k$ such that $\mathsf{p}_{k'} = \min\{\mathsf{p}_k, \ldots, \mathsf{p}_\ell\}$. We now prove that we can flip $\mathsf{p}_{k'}$ instead of $\mathsf{p}_{k'-1}$ in $I_{k'-1}$, and still obtain a path from I to J where $\mathsf{p}_{k'}$ is still smaller than all positive edge labels appearing after it. In Example 11, we illustrate this procedure on an explicit example.

Clearly $\mathsf{p}_{k'-1} > \mathsf{p}_{k'}$, and we have a p-falling sequence of two flips given by $I_{k'-1} \to I_{k'} \to I_{k'+1}$. Using Proposition 3, we can now see these two flips as well in a subword complex for the dihedral parabolic subsystem. For this, restrict (W, S) to the subspace V' spanned by the two roots $\mathsf{r}(I_{k'-1}, \mathsf{p}_{k'-1})$ and $\mathsf{r}(I_{k'-1}, \mathsf{p}_{k'}) = \mathsf{r}(I_{k'}, \mathsf{p}_{k'})$. This restricted subword complex corresponds to all facets of $\mathbf{SC}(Q, \rho)$ reachable from the particular facet $I_{k'-1}$ by flips whose directions are contained in V'. Applying Lemma 1, we can thus replace the subpath $I_{k'-1} \to I_{k'} \to I_{k'+1}$ by a p-rising path from $I_{k'-1}$ to $I_{k'+1}$ flipping first position $\mathsf{p}_{k'}$ and then a (possibly empty) sequence of positions larger than or equal to $\mathsf{p}_{k'-1}$.

Repeating this operation, we construct a path from I to J such that $\mathsf{p}_k = \min\{\mathsf{p}_k, \ldots, \mathsf{p}_\ell\}$. By this procedure, we obtain eventually a p-rising path from I to J. This path is unique and lexicographically first among all paths from I to J in $\mathbf{G}(Q, \rho)$ according to the characterization given in Proposition 4. This concludes the proof that p is an EL-labeling of $\mathbf{G}(Q, \rho)$.

The proof is similar for the negative edge labeling n (by the reversal operation in Remark 2).

Example 11. Consider the subword complex $\mathbf{SC}(Q^{\mathrm{ex}}, \rho^{\mathrm{ex}})$ of Example 4, whose labeled increasing flip graph is shown in Fig. 7, and the path

$$12356 \xrightarrow{-5} 12367 \xrightarrow{-6} 12379 \xrightarrow{-2} 13479 \xrightarrow{-1} 34789$$

in $\mathbf{G}(Q, \rho)$, where the numbers on the arrows are the positive edge labels. In the language of the proof of Theorem 2, we have $k = 1$, $k' = 4$, and therefore we replace the subpath $12379 \xrightarrow{-2} 13479 \xrightarrow{-1} 34789$ by the subpath $12379 \xrightarrow{-1} 23789 \xrightarrow{-2} 34789$, thus obtaining the path

$$12356 \xrightarrow{-5} 12367 \xrightarrow{-6} 12379 \xrightarrow{-1} 23789 \xrightarrow{-2} 34789.$$

Applying this operation again and again produces the sequence of paths given by

$$12356 \xrightarrow{-5} 12367 \xrightarrow{-1} 23678 \xrightarrow{-6} 23789 \xrightarrow{-2} 34789,$$

$$12356 \xrightarrow{-1} 23568 \xrightarrow{-5} 23678 \xrightarrow{-6} 23789 \xrightarrow{-2} 34789,$$

$$12356 \xrightarrow{-1} 23568 \xrightarrow{-5} 23678 \xrightarrow{-2} 34678 \xrightarrow{-6} 34789,$$

$$12356 \xrightarrow{-1} 23568 \xrightarrow{-2} 34568 \xrightarrow{-5} 34678 \xrightarrow{-6} 34789.$$

The resulting path is p-rising. In this example, all paths happen to have the same length. This does not hold in general, compare Fig. 15 on page 245, where the path

$$123 \xrightarrow{-2} 137 \xrightarrow{-3} 178 \xrightarrow{-1} 678 \xrightarrow{-7} 689$$

is, for example, replaced by the path

$$123 \xrightarrow{-2} 137 \xrightarrow{-1} 357 \xrightarrow{-3} 567 \xrightarrow{-5} 678 \xrightarrow{-7} 689.$$

In contrast to the rising paths, we can have none, one, or more than one p-falling and n-falling paths between two facets I and J of $\mathbf{SC}(Q, \rho)$. Even if we will not need it in the remainder of the paper, we observe in the next proposition that there are always as many p-falling paths as n-falling paths from I to J.

Remember that we say that a path $I_1 \to I_2 \to \cdots \to I_{\ell+1}$ *flips out* the multiset $\mathsf{P} := \{\!\{\mathsf{p}(I_k \to I_{k+1}) \mid k \in [\ell]\}\!\}$ and *flips in* the multiset $\mathsf{N} := \{\!\{\mathsf{n}(I_k \to I_{k+1}) \mid k \in [\ell]\}\!\}$. Observe that a p-falling (resp. n-falling) path is determined by the multiset P (resp. N) of positions that it flips out (resp. in).

Proposition 5. *Let I and J be two facets of $\mathbf{SC}(Q, \rho)$. Then there are as many p-falling paths as n-falling paths from I to J. More precisely, for any multisubsets P and N of $[m]$, there exists a p-falling path from I to J which flips out P and flips in N, if and only if there exists an n-falling path with the same property.*

Proof. Consider a p-falling path $I = I_1 \to \cdots \to I_{\ell+1} = J$. Define $\mathsf{p}_k := \mathsf{p}(I_k \to I_{k+1})$ and $\mathsf{n}_k := \mathsf{n}(I_k \to I_{k+1})$. We want to prove that there is as well an n-falling path which flips out $\mathsf{P} := \{\mathsf{p}_k \mid k \in [\ell]\}$ and flips in $\mathsf{N} := \{\mathsf{n}_k \mid k \in [\ell]\}$.

If the path $I = I_1 \to \cdots \to I_{\ell+1} = J$ happens to be n-falling, we are done. Otherwise, consider the first position k such that $\mathsf{n}_{k-1} < \mathsf{n}_k$. Since the path is p-falling, we thus have $\mathsf{p}_k < \mathsf{p}_{k-1} < \mathsf{n}_{k-1} < \mathsf{n}_k$. By Proposition 2(3), we know that $\mathsf{r}(I_{k-1}, \mathsf{p}_{k-1}) = \mathsf{r}(I_{k-1}, \mathsf{n}_{k-1})$ and $\mathsf{r}(I_k, \mathsf{p}_k) = \mathsf{r}(I_k, \mathsf{n}_k)$. According to Proposition 2(5) and to the previous inequalities, we therefore obtain

$$\mathsf{r}(I_{k-1}, \mathsf{p}_k) = \mathsf{r}(I_k, \mathsf{p}_k) = \mathsf{r}(I_k, \mathsf{n}_k) = \mathsf{r}(I_{k-1}, \mathsf{n}_k).$$

Thus, in the facet I_{k-1}, flipping out p_k flips in n_k. We denote by I_k' the facet of $\mathbf{SC}(Q, \rho)$ obtained by this flip. Using again Proposition 2(5) and the previous inequalities, we obtain that

$$\mathsf{r}(I_k', \mathsf{p}_{k-1}) = s_{\mathsf{r}(I_{k-1}, \mathsf{p}_k)}(\mathsf{r}(I_{k-1}, \mathsf{p}_{k-1})) = s_{\mathsf{r}(I_{k-1}, \mathsf{p}_k)}(\mathsf{r}(I_{k-1}, \mathsf{n}_{k-1})) = \mathsf{r}(I_k', \mathsf{n}_{k-1}).$$

Therefore, in the facet I_k', flipping out p_{k-1} flips in n_{k-1}. After these two flips, we thus obtain I_{k+1} (since we flipped out p_k and p_{k-1}, while we flipped in n_k and n_{k-1}). In other words, we can replace the subpath $I_{k-1} \to I_k \to I_{k+1}$ by the path $I_{k-1} \to I_k' \to I_{k+1}$ where we flip first p_k to n_k and then p_{k-1} to n_{k-1}. The new path still flips out P and flips in N, and the first k positions it flips in are in decreasing order. Repeating this transformation finally yields an n-falling path from I to J which still flips out P and flips in N. Observe that this path does not necessarily coincide with the p-falling path we started from.

Since a p-falling (resp. n-falling) path is determined by the set of positions it flips out (resp. in), we obtain a bijection between p-falling paths and n-falling paths from I to J. They are thus equinumerous. □

Remark 4. Observe that Proposition 5 can be deduced from the following observations in the situation of double root free subword complexes studied in Sect. 5. In this situation, the flip graph is the Hasse diagram of its transitive closure and the p- and n-labelings are both EL-labelings thereof. By Theorem 3, all p- and n-falling paths have the same length. Therefore, Proposition 1 implies that they are equinumerous. A similar topological construction in the situation of subword complexes having double roots is yet to be found.[1]

4.2 Greedy Facets

We now characterize the unique source and sink of the increasing flip graph $\mathbf{G}(Q, \rho)$.

Proposition 6. *The lexicographically smallest (resp. largest) facet of* $\mathbf{SC}(Q, \rho)$ *is the unique source (resp. sink) of* $\mathbf{G}(Q, \rho)$.

Proof. The lexicographically smallest facet is a source of $\mathbf{G}(Q, \rho)$ since none of its flips can be decreasing. We prove that this source is unique by induction on the word Q. Denote by $\mathsf{P}(Q_\dashv, \rho)$ (resp. $\mathsf{P}(Q_\dashv, \rho q_m)$) the lexicographically smallest facet of $\mathbf{SC}(Q_\dashv, \rho)$ (resp. $\mathbf{SC}(Q_\dashv, \rho q_m)$) and assume that it is the unique source of the flip graph $\mathbf{G}(Q_\dashv, \rho)$ (resp. $\mathbf{G}(Q_\dashv, \rho q_m)$). Consider a source P of $\mathbf{G}(Q, \rho)$. We distinguish two cases:

- If $\ell(\rho q_m) > \ell(\rho)$, then q_m cannot be the last reflection of a reduced expression for ρ. Thus $\mathbf{SC}(Q, \rho) = \mathbf{SC}(Q_\dashv, \rho) \star m$ and $\mathsf{P} = \mathsf{P}(Q_\dashv, \rho) \cup m$.
- Otherwise, $\ell(\rho q_m) < \ell(\rho)$. If m is in P, then

$$\mathsf{r}(\mathsf{P}, m) = \rho(\alpha_{q_m}) \in \Phi^- \cap \rho(\Phi^+).$$

Since $\Phi^- \cap \rho(\Phi^+) = -\mathrm{inv}(\rho)$, we obtain that m is flippable (by Proposition 2(3)) and its flip is decreasing (by Proposition 2(4)). This would contradict the assumption that P is a source of $\mathbf{G}(Q, \rho)$. Consequently, $m \notin \mathsf{P}$. Since the facets

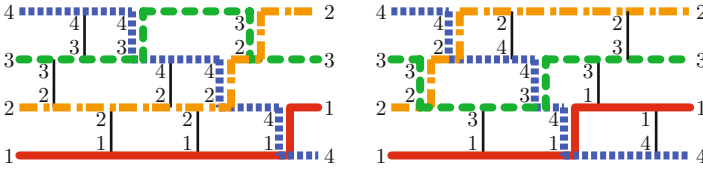---

[1]We thank an anonymous referee for raising this question.

Fig. 8 The positive and negative greedy facets of $\mathbf{SC}(Q^{ex}, \rho^{ex})$

of $\mathbf{SC}(Q, \rho)$ which do not contain m coincide with the facets of $\mathbf{SC}(Q_\dashv, \rho q_m)$, we obtain that $P = P(Q_\dashv, \rho q_m)$.

In both cases, we obtain that the source P is the lexicographically smallest facet of $\mathbf{SC}(Q, \rho)$. The proof is similar for the sink.

We call *positive* (resp. *negative*) *greedy facet* and denote by $P(Q, \rho)$ (resp. by $N(Q, \rho)$) the unique source (resp. sink) of the graph $\mathbf{G}(Q, \rho)$ of increasing flips. The term "positive" (resp. "negative") emphasizes that $P(Q, \rho)$ (resp. $N(Q, \rho)$) is the unique facet of $\mathbf{SC}(Q, \rho)$ whose root configuration is a subset of positive (resp. negative) roots, while the term "greedy" refers to the greedy properties of these facets underlined in Lemmas 2 and 3.

These greedy facets are reverse to one another (see Remark 2). Namely,

$$N(q_m \cdots q_1, \rho^{-1}) = \{m + 1 - p \mid p \in P(q_1 \cdots q_m, \rho)\}.$$

We still work with both in parallel to simplify the presentation in the next section.

Example 12. Consider the subword complex $\mathbf{SC}(Q^{ex}, \rho^{ex})$ presented in Example 4. Its positive and negative greedy facets are given by $P(Q^{ex}, \rho^{ex}) = \{1, 2, 3, 5, 6\}$ and $N(Q^{ex}, \rho^{ex}) = \{3, 4, 7, 8, 9\}$, respectively, see Fig. 8. They appear respectively as the leftmost and rightmost facets in Fig. 4.

The following two lemmas provide two (somehow inverse) greedy inductive procedures to construct the greedy facets $P(Q, \rho)$ and $N(Q, \rho)$. These lemmas are direct consequences of the definition of the greedy facets and of the induction formulas for the facets $\mathbf{F}(Q, \rho)$ presented in Sect. 3.3. Remember that we denote by $Q_\vdash := q_2 \cdots q_m$ and $Q_\dashv := q_1 \cdots q_{m-1}$ the words on S obtained from $Q := q_1 \cdots q_m$ by deleting its first and last letters respectively, and by $X^{\rightarrow} := \{x + 1 \mid x \in X\}$ the right shift of a subset $X \subset \mathbb{Z}$.

Lemma 2. *The greedy facets $P(Q, \rho)$ and $N(Q, \rho)$ can be constructed inductively from $P(\varepsilon, e) = N(\varepsilon, e) = \varnothing$ using the following formulas:*

$$P(Q, \rho) = \begin{cases} P(Q_\dashv, \rho) \cup m & \text{if m appears in all facets of } \mathbf{SC}(Q, \rho), \\ P(Q_\dashv, \rho q_m) & \text{otherwise.} \end{cases}$$

$$N(Q, \rho) = \begin{cases} 1 \cup N(Q_\vdash, \rho)^{\rightarrow} & \text{if 1 appears in all facets of } \mathbf{SC}(Q, \rho), \\ N(Q_\vdash, q_1\rho)^{\rightarrow} & \text{otherwise.} \end{cases}$$

Lemma 3. *The greedy facets* $\mathsf{P}(Q, \rho)$ *and* $\mathsf{N}(Q, \rho)$ *can be constructed inductively from* $\mathsf{P}(\varepsilon, e) = \mathsf{N}(\varepsilon, e) = \varnothing$ *using the following formulas:*

$$\mathsf{P}(Q, \rho) = \begin{cases} \mathsf{P}(Q_\vdash, q_1\rho)^\rightarrow & \text{if } 1 \text{ appears in none of the facets of } \mathbf{SC}(Q, \rho), \\ 1 \cup \mathsf{P}(Q_\vdash, \rho)^\rightarrow & \text{otherwise.} \end{cases}$$

$$\mathsf{N}(Q, \rho) = \begin{cases} \mathsf{N}(Q_\dashv, \rho q_m) & \text{if } m \text{ appears in none of the facets of } \mathbf{SC}(Q, \rho), \\ \mathsf{N}(Q_\dashv, \rho) \cup m & \text{otherwise.} \end{cases}$$

Lemmas 2 and 3 can be reformulated to obtain greedy sweep procedures on the word Q itself, avoiding the use of induction. Namely, the positive greedy facet $\mathsf{P}(Q, \rho)$ is obtained:

1. Either sweeping Q from right to left placing inversions as soon as possible,
2. Or sweeping Q from left to right placing non-inversions as long as possible.

The negative greedy facet is obtained similarly, reversing the directions of the sweeps.

We have seen in Theorem 2 that for any two facets $I, J \in \mathbf{F}(Q, \rho)$ such that $I \dashrightarrow J$, there is a p-rising (resp. n-rising) path from I to J. In particular, there is always a p-rising (resp. n-rising) path from $\mathsf{P}(Q, \rho)$ to $\mathsf{N}(Q, \rho)$. We will now show that there is also at least one p-falling (resp. n-falling) path from $\mathsf{P}(Q, \rho)$ to $\mathsf{N}(Q, \rho)$ if the subword complex $\mathbf{SC}(Q, \rho)$ is spherical.

Proposition 7. *For any spherical subword complex* $\mathbf{SC}(Q, \rho)$*, there is always a* p*-falling and an* n*-falling path from* $\mathsf{P}(Q, \rho)$ *to* $\mathsf{N}(Q, \rho)$*.*

Proof. Since the subword complex $\mathbf{SC}(Q, \rho)$ is spherical, recall that any position in any facet of $\mathbf{SC}(Q, \rho)$ is flippable. We will prove that starting from the positive greedy facet $\mathsf{P}(Q, \rho)$ and successively flipping all its positions in decreasing order yields the negative greedy facet $\mathsf{N}(Q, \rho)$, thus providing a p-falling path from $\mathsf{P}(Q, \rho)$ to $\mathsf{N}(Q, \rho)$.

Let $\ell := |Q| - \ell(\rho)$ denote the size of each facet of $\mathbf{SC}(Q, \rho)$. Let $\mathsf{p}_1 > \cdots > \mathsf{p}_\ell$ denote the positions of the positive greedy facet $\mathsf{P}(Q, \rho)$ in decreasing order. We consider the p-falling path $\mathsf{P}(Q, \rho) = I_1 \rightarrow \cdots \rightarrow I_{\ell+1}$ defined by $\mathsf{p}(I_k \rightarrow I_{k+1}) = \mathsf{p}_k$. We also set $\mathsf{n}_k := \mathsf{n}(I_k \rightarrow I_{k+1})$. By definition, we have $I_k = \{\mathsf{n}_1, \ldots, \mathsf{n}_{k-1}, \mathsf{p}_k, \ldots, \mathsf{p}_\ell\}$. We will prove that the root $\mathsf{r}(I_k, \mathsf{n}_j)$ is negative for any $j < k \in [\ell + 1]$. This implies in particular that $I_{\ell+1}$ is the negative greedy facet $\mathsf{N}(Q, \rho)$.

To see this, fix $j \in [\ell]$. For any $k \in [j + 1, \ell + 1]$, denote by x_k the position in the complement of I_k such that $\mathsf{r}(I_k, \mathsf{x}_k) = \pm\mathsf{r}(I_k, \mathsf{n}_j)$. We prove by induction on k that $\mathsf{p}_k < \mathsf{x}_k < \mathsf{n}_j$, and thus (by Proposition 2(4)) that the root $\mathsf{r}(I_k, \mathsf{n}_j) = -\mathsf{r}(I_k, \mathsf{x}_k)$ is negative for any $j < k \leq \ell + 1$. First, this is immediate for $k = j + 1$ since $\mathsf{x}_{j+1} = \mathsf{p}_j$ (because we just flipped out p_j to flip

in n_j in I_j) and $p_{j+1} < p_j < n_j$. Assume now that we proved that $p_k < x_k < n_j$ for a certain k. We distinguish two cases:

(i) If $n_k < n_j$, then $r(I_{k+1}, n_j) = r(I_k, n_j)$ by Proposition 2(5). Since this root is negative, Proposition 2(4) ensures that $x_{k+1} < n_j$. Moreover, if $x_{k+1} \leq p_{k+1}$, then we would have $x_{k+1} < p_k$, and thus Proposition 2(5) would give

$$r(I_k, x_{k+1}) = r(I_{k+1}, x_{k+1}) = -r(I_{k+1}, n_j) = -r(I_k, n_j).$$

By definition, this would imply that $x_k = x_{k+1} < p_k$, contradicting the induction hypothesis.

(ii) If $n_k > n_j$, then we have $p_k < x_k < n_j < n_k$. Therefore, Proposition 2(5) ensures that

$$r(I_{k+1}, x_k) = s_{r(I_k, p_k)}(r(I_k, x_k)) = -s_{r(I_k, p_k)}(r(I_k, n_j)) = -r(I_{k+1}, n_j).$$

By definition, this implies that $x_{k+1} = x_k$.

In both cases, we obtained that $p_{k+1} < x_{k+1} < n_j$, thus concluding our inductive argument.

The proof for the n-falling path is similar.

Note that this proposition fails if we drop the condition that $\mathbf{SC}(Q, \rho)$ is spherical, as illustrated in the subword complex $\mathbf{SC}(Q^{\text{ex}}, \rho^{\text{ex}})$ of Example 4. A smaller example is given by the subword complex $\mathbf{SC}(\tau_1 \tau_2 \tau_1 \tau_2, \tau_1 \tau_2)$.

4.3 Spanning Trees

As discussed in Remark 1, the edge labelings p and n automatically produce canonical spanning trees of any interval of the increasing flip graph $\mathbf{G}(Q, \rho)$. Since $\mathbf{G}(Q, \rho)$ has a unique source $\mathsf{P}(Q, \rho)$ and a unique sink $\mathsf{N}(Q, \rho)$, we obtain in particular four spanning trees of the graph $\mathbf{G}(Q, \rho)$ itself. The goal of this section is to give alternative descriptions of these four spanning trees.

We call respectively *positive source tree*, *positive sink tree*, *negative source tree*, and *negative sink tree*, and denote respectively by $\mathbf{P}(Q, \rho)$, $\mathbf{P}^*(Q, \rho)$, $\mathbf{N}(Q, \rho)$, and $\mathbf{N}^*(Q, \rho)$, the p-source, p-sink, n-source, and n-sink trees of $\mathbf{G}(Q, \rho)$. The tree $\mathbf{P}(Q, \rho)$ (resp. $\mathbf{N}(Q, \rho)$) is formed by all p-rising (resp. n-rising) paths from the positive greedy facet $\mathsf{P}(Q, \rho)$ to all the facets of $\mathbf{SC}(Q, \rho)$. Both $\mathbf{P}(Q, \rho)$ and $\mathbf{N}(Q, \rho)$ are rooted at and directed away from the positive greedy facet $\mathsf{P}(Q, \rho)$. The tree $\mathbf{P}^*(Q, \rho)$ (resp. $\mathbf{N}^*(Q, \rho)$) is formed by all p-rising (resp. n-rising) paths from all the facets of $\mathbf{SC}(Q, \rho)$ to the negative greedy facet $\mathsf{N}(Q, \rho)$. Both $\mathbf{P}^*(Q, \rho)$ and $\mathbf{N}^*(Q, \rho)$ are rooted at and directed towards the negative greedy facet $\mathsf{N}(Q, \rho)$.

The positive source and negative sink trees (resp. the positive sink and the negative source trees) are reverse to one another (see Remark 2). Namely, as we already observed, $I \to J$ is an edge in the increasing flip graph $\mathbf{G}(q_m \cdots q_1, \rho^{-1})$ if and only if $J' := \{m + 1 - j \mid j \in J\} \to I' := \{m + 1 - i \mid i \in I\}$ is an edge in the increasing flip graph $\mathbf{G}(q_1 \cdots q_m, \rho)$. Moreover, $I \to J$ belongs to $\mathbf{P}(q_m \cdots q_1, \rho^{-1})$ if and only if $J' \to I'$ belongs to $\mathbf{N}^*(q_1 \cdots q_m, \rho)$. Similarly, $I \to J$ belongs to $\mathbf{P}^*(q_m \cdots q_1, \rho^{-1})$ if and only if $J' \to I'$ belongs to $\mathbf{N}(q_1 \cdots q_m, \rho)$.

Example 13. Consider the subword complex $\mathbf{SC}(Q^{ex}, \rho^{ex})$ from Example 4. Figures 9–12 represent respectively the trees $\mathbf{P}(Q^{ex}, \rho^{ex})$, $\mathbf{P}^*(Q^{ex}, \rho^{ex})$, $\mathbf{N}(Q^{ex}, \rho^{ex})$, and $\mathbf{N}^*(Q^{ex}, \rho^{ex})$. Observe that these four canonical spanning trees of $\mathbf{G}(Q, \rho)$ are all different in general.

We now give a direct description of the father of a facet I in $\mathbf{P}^*(Q, \rho)$ and $\mathbf{N}(Q, \rho)$ in terms of $I \smallsetminus \mathbf{N}(Q, \rho)$ and $I \smallsetminus \mathbf{P}(Q, \rho)$.

Proposition 8. *Let I be a facet of $\mathbf{SC}(Q, \rho)$. If $I \neq \mathbf{N}(Q, \rho)$, then the father of I in $\mathbf{P}^*(Q, \rho)$ is obtained from I by flipping the smallest position in $I \smallsetminus \mathbf{N}(Q, \rho)$. Similarly, if $I \neq \mathbf{P}(Q, \rho)$, then the father of I in $\mathbf{N}(Q, \rho)$ is obtained from I by flipping the largest position in $I \smallsetminus \mathbf{P}(Q, \rho)$.*

Proof. Since the father of I in $\mathbf{P}^*(Q, \rho)$ (resp. in $\mathbf{N}(Q, \rho)$) is the facet next to I on the unique p-rising path towards $\mathbf{N}(Q, \rho)$ (resp. the facet previous to I on the unique n-rising path from $\mathbf{P}(Q, \rho)$), this is a direct consequence of Proposition 4. $\qquad\square$

We now focus on the positive source tree $\mathbf{P}(Q, \rho)$ and on the negative sink tree $\mathbf{N}^*(Q, \rho)$, and provide two different descriptions of them. The first is an inductive description of $\mathbf{P}(Q, \rho)$ and $\mathbf{N}^*(Q, \rho)$ (see Propositions 10 and 11). The second is a direct description of the father of a facet I in $\mathbf{P}(Q, \rho)$ and $\mathbf{N}^*(Q, \rho)$ in terms of greedy prefixes and suffixes of I (see Propositions 12 and 13). These descriptions mainly rely on the following property of the greedy facets.

Proposition 9. *If m is a flippable position of $\mathbf{N}(Q, \rho)$, then $\mathbf{N}(Q_\dashv, \rho q_m)$ is obtained from $\mathbf{N}(Q, \rho)$ by flipping m. Similarly, if 1 is a flippable position of $\mathbf{P}(Q, \rho)$, then $\mathbf{P}(Q_\vdash, q_1\rho)$ is obtained from $\mathbf{P}(Q, \rho)$ by flipping 1 and shifting to the left.*

Proof. Although the formulation is simpler for the negative greedy facets, the proof is simpler for the positive ones (due to the direction chosen in the definition of the root function). Assume that 1 is a flippable position of $\mathbf{P}(Q, \rho)$. Let $J \in \mathbf{F}(Q, \rho)$ and $j \in J$ be such that $\mathbf{P}(Q, \rho) \smallsetminus 1 = J \smallsetminus j$. Consider the facet $J^{\leftarrow}$ of $\mathbf{SC}(Q_\vdash, q_1\rho)$ obtained shifting J to the left. Proposition 2(5) enables us to compute the root function $\mathsf{r}(J, \cdot)$ for J, which in turn gives us the root function for $J^{\leftarrow}$:

$$\mathsf{r}(J^{\leftarrow}, k) = \begin{cases} \mathsf{r}(\mathbf{P}(Q, \rho), k + 1) & \text{if } 1 \leq k \leq j - 1, \\ q_1(\mathsf{r}(\mathbf{P}(Q, \rho), k + 1)) & \text{otherwise.} \end{cases}$$
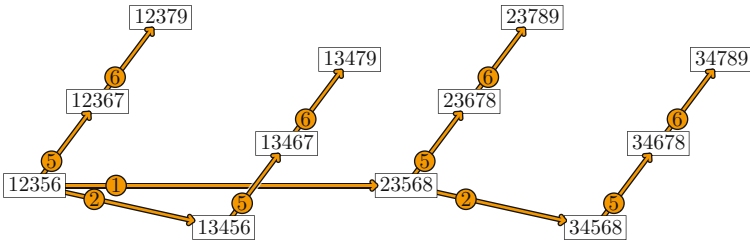
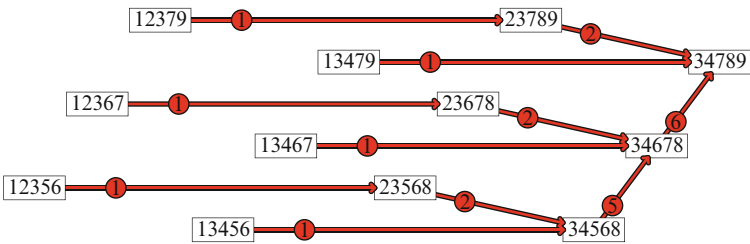**Fig. 9** The positive source tree $\mathbf{P}(Q^{ex}, \rho^{ex})$

Fig. 10 The positive sink tree $\mathbf{P}^*(Q^{ex}, \rho^{ex})$
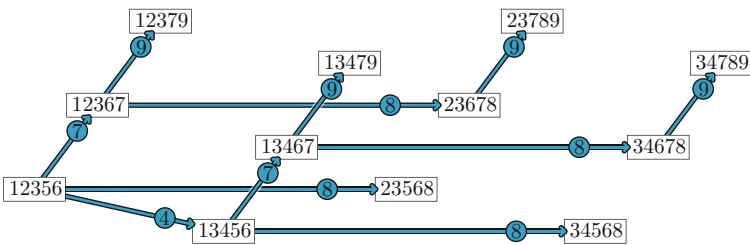
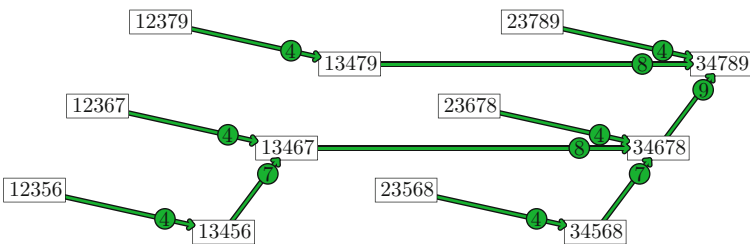**Fig. 11** The negative source tree $\mathbf{N}(Q^{ex}, \rho^{ex})$

Fig. 12 The negative sink tree $\mathbf{N}^*(Q^{ex}, \rho^{ex})$

Since all positions $i \in \mathsf{P}(\mathsf{Q}, \rho)$ such that $\mathsf{r}(\mathsf{P}(\mathsf{Q}, \rho), i) = \alpha_{q_1}$ are located before j, and since α_{q_1} is the only positive root sent to a negative root by the simple reflection q_1, all roots $\mathsf{r}(J^{\leftarrow}, k)$, for $k \in J^{\leftarrow}$, are positive. Consequently, $J^{\leftarrow} = \mathsf{P}(\mathsf{Q}_{\vdash}, q_1\rho)$.

We obtain the result for negative facets using the reversal operation of Remark 2.

Example 14. Consider the subword complex $\mathsf{SC}(\mathsf{Q}^{\mathrm{ex}}, \rho^{\mathrm{ex}})$ of Example 4. Since position 9 is flippable in $\mathsf{N}(\mathsf{Q}^{\mathrm{ex}}, \rho^{\mathrm{ex}}) = \{3, 4, 7, 8, 9\}$, we have $\mathsf{N}(\mathsf{Q}^{\mathrm{ex}}_{\dashv}, \rho^{\mathrm{ex}}\tau_1) = \{3, 4, 6, 7, 8\}$. Moreover, since position 1 is flippable in $\mathsf{P}(\mathsf{Q}^{\mathrm{ex}}, \rho^{\mathrm{ex}}) = \{1, 2, 3, 5, 6\}$, we have $\mathsf{P}(\mathsf{Q}^{\mathrm{ex}}_{\vdash}, \tau_2\rho^{\mathrm{ex}}) = \{2, 3, 5, 6, 8\}^{\leftarrow} = \{1, 2, 4, 5, 7\}$.

Using Proposition 9, we can describe inductively the two trees $\mathbf{P}(\mathsf{Q}, \rho)$ and $\mathbf{N}^*(\mathsf{Q}, \rho)$. The induction follows the induction formulas for the facets $\mathbf{F}(\mathsf{Q}, \rho)$ presented in Sect. 3.3. For a tree $\mathbf{T}$ whose vertices are subsets of $\mathbb{Z}$ and for an element $z \in \mathbb{Z}$, we denote by $\mathbf{T} \star z = z \star \mathbf{T}$ the tree with a vertex $X \cup z$ for each vertex X of $\mathbf{T}$ and an edge $X \cup z \to Y \cup z$ for each edge $X \to Y$ of $\mathbf{T}$. Similarly, we will denote by $\mathbf{T}^{\to}$ the tree with a vertex $X^{\to} := \{x + 1 \mid x \in X\}$ for each vertex X of $\mathbf{T}$ and an edge $X^{\to} \to Y^{\to}$ for each edge $X \to Y$ of $\mathbf{T}$.

We start with the inductive description of the negative sink tree $\mathbf{N}^*(\mathsf{Q}, \rho)$, which is based on the right induction formula. For the empty word ε, the tree $\mathbf{N}^*(\varepsilon, e)$ is formed by the unique facet $\varnothing$ of $\mathsf{SC}(\varepsilon, e)$, and the tree $\mathbf{N}^*(\varepsilon, \rho)$ is empty if $\rho \neq e$. Otherwise, $\mathbf{N}^*(\mathsf{Q}, \rho)$ is obtained as follows.

Proposition 10. *For a non-empty word* Q, *the tree* $\mathbf{N}^*(\mathsf{Q}, \rho)$ *equals*

1. $\mathbf{N}^*(\mathsf{Q}_{\dashv}, \rho q_m)$ *if m appears in none of the facets of* $\mathsf{SC}(\mathsf{Q}, \rho)$;
2. $\mathbf{N}^*(\mathsf{Q}_{\dashv}, \rho) \star m$ *if m appears in all the facets of* $\mathsf{SC}(\mathsf{Q}, \rho)$;
3. *The disjoint union of* $\mathbf{N}^*(\mathsf{Q}_{\dashv}, \rho q_m)$ *and* $\mathbf{N}^*(\mathsf{Q}_{\dashv}, \rho) \star m$, *with an additional edge from* $\mathsf{N}(\mathsf{Q}_{\dashv}, \rho q_m)$ *to* $\mathsf{N}(\mathsf{Q}, \rho) = \mathsf{N}(\mathsf{Q}_{\dashv}, \rho) \cup m$, *otherwise.*

Proof. Assume that m is contained in at least one and not all facets of $\mathsf{SC}(\mathsf{Q}, \rho)$. In other words, m is a flippable position of $\mathsf{N}(\mathsf{Q}, \rho)$. Let $I = I_1 \to \cdots \to I_{\ell+1} = \mathsf{N}(\mathsf{Q}, \rho)$ be any n-rising path from an arbitrary facet $I \in \mathbf{F}(\mathsf{Q}, \rho)$ to $\mathsf{N}(\mathsf{Q}, \rho)$. If the label m appears in this path, then it should clearly appear last. By Proposition 9, we have therefore $I_{\ell} = \mathsf{N}(\mathsf{Q}_{\dashv}, \rho q_m)$, and $I = I_1 \to \cdots \to I_{\ell} = \mathsf{N}(\mathsf{Q}_{\dashv}, \rho q_m)$ is also an n-rising path from I to $\mathsf{N}(\mathsf{Q}_{\dashv}, \rho q_m)$ in the increasing flip graph $\mathbf{G}(\mathsf{Q}_{\dashv}, \rho q_m)$. Otherwise, if the label m does not appear in the path, then m is contained in all facets of this path, and $(I \smallsetminus m) = (I_1 \smallsetminus m) \to \cdots \to (I_{\ell+1} \smallsetminus m) = \mathsf{N}(\mathsf{Q}_{\dashv}, \rho)$ is an n-rising path from $I \smallsetminus m$ to $\mathsf{N}(\mathsf{Q}_{\dashv}, \rho)$ in the increasing flip graph $\mathbf{G}(\mathsf{Q}_{\dashv}, \rho)$. This corresponds precisely to the description of (3). The proofs of (1) and (2) are similar and left to the reader. $\qed$

We now give the inductive description of the positive source tree $\mathbf{P}(\mathsf{Q}, \rho)$, which is based on the left induction formula. For the empty word ε, the tree $\mathbf{P}(\varepsilon, e)$ is formed by the unique facet $\varnothing$ of $\mathsf{SC}(\varepsilon, e)$, and the tree $\mathbf{P}(\varepsilon, \rho)$ is empty if $\rho \neq e$. Otherwise, $\mathbf{P}(\mathsf{Q}, \rho)$ is obtained as follows.

Proposition 11. *For a non-empty word* Q, *the tree* $\mathbf{P}(\mathsf{Q}, \rho)$ *equals*

(i) $\mathbf{P}(\mathsf{Q}_{\vdash}, q_1\rho)^{\to}$ *if 1 appears in none of the facets of* $\mathsf{SC}(\mathsf{Q}, \rho)$;
(ii) $1 \star \mathbf{P}(\mathsf{Q}_{\vdash}, \rho)^{\to}$ *if 1 appears in all the facets of* $\mathsf{SC}(\mathsf{Q}, \rho)$;

(iii) The disjoint union of $\mathbf{P}(\mathsf{Q}_\vdash, q_1\rho)^\to$ *and* $1 \star \mathbf{P}(\mathsf{Q}_\vdash, \rho)^\to$, *with an additional edge from* $\mathsf{P}(\mathsf{Q}, \rho) = 1 \cup \mathsf{P}(\mathsf{Q}_\vdash, \rho)^\to$ *to* $\mathsf{P}(\mathsf{Q}_\vdash, q_1\rho)^\to$, *otherwise.*

Proof. We can either translate the proof of Proposition 10, or directly apply to Proposition 10 the reversal operation of Remark 2. $\qquad \square$

Note that we do not have a similar inductive description for the positive sink and negative source trees $\mathbf{P}^*(\mathsf{Q}, \rho)$ and $\mathsf{N}(\mathsf{Q}, \rho)$. Let I_{max} denote the neighbor of $\mathsf{N}(\mathsf{Q}, \rho)$ in $\mathbf{G}(\mathsf{Q}, \rho)$ which maximizes $\mathsf{p}_{max} := \mathsf{p}(I_{max} \to \mathsf{N}(\mathsf{Q}, \rho))$. We can use position p_{max} to decompose the positive sink tree $\mathbf{P}^*(\mathsf{Q}, \rho)$ as the union of a spanning tree of the graph of increasing flips on its link $\{I \in \mathbf{SC}(\mathsf{Q}, \rho) \mid \mathsf{p}_{max} \in I\}$ with a spanning tree of the graph of increasing flips on its deletion $\{I \in \mathbf{SC}(\mathsf{Q}, \rho) \mid \mathsf{p}_{max} \notin I\}$, together with the edge $I_{max} \to \mathsf{N}(\mathsf{Q}, \rho)$. However, contrarily to the link of p_{max}, the deletion of p_{max} is not a subword complex in general. This is a serious limit to an inductive decomposition of the positive sink tree $\mathbf{P}^*(\mathsf{Q}, \rho)$. The same observation holds for the negative source tree $\mathsf{N}(\mathsf{Q}, \rho)$.

We now give a direct characterization of the father of a facet I of $\mathbf{SC}(\mathsf{Q}, \rho)$ in the negative sink tree $\mathsf{N}(\mathsf{Q}, \rho)$. This description can be understood in terms of the longest greedy prefix of I.

Proposition 12. *Let* $I \neq \mathsf{N}(\mathsf{Q}, \rho)$ *be a facet of* $\mathbf{SC}(\mathsf{Q}, \rho)$. *Define* $y = y(I)$ *to be the smallest position in* $[m]$ *such that*

$$I \cap [y] \neq \mathsf{N}(\mathsf{q}_1 \cdots \mathsf{q}_y, \mathit{\Pi}\mathsf{Q}_{[y] \smallsetminus I}),$$

and $x = x(I)$ *to be the smallest position in* I *such that* $\mathsf{r}(I, x) = \mathsf{r}(I, y)$. *Then the father of the facet* I *in the negative sink tree* $\mathsf{N}^*(\mathsf{Q}, \rho)$ *is obtained from* I *by flipping position* x.

Proof. Let $x(I)$ and $y(I)$ be the positions defined in the statement of the proposition. Denote by J the father of I in the negative sink tree $\mathsf{N}^*(\mathsf{Q}, \rho)$, and let $\bar{x}(I)$ and $\bar{y}(I)$ be such that $I \smallsetminus \bar{x}(I) = J \smallsetminus \bar{y}(I)$. We want to prove that $x(I) = \bar{x}(I)$ and $y(I) = \bar{y}(I)$ for any facet $I \neq \mathsf{N}(\mathsf{Q}, \rho)$ of $\mathbf{SC}(\mathsf{Q}, \rho)$.

We first prove that $y(I) = \bar{y}(I)$ for any facet I of $\mathbf{SC}(\mathsf{Q}, \rho)$ by induction on the negative sink tree. For this, set $y(\mathsf{N}(\mathsf{Q}, \rho)) = \bar{y}(\mathsf{N}(\mathsf{Q}, \rho)) = m + 1$. Consider an arbitrary facet $I \neq \mathsf{N}(\mathsf{Q}, \rho)$ and its father J in $\mathsf{N}^*(\mathsf{Q}, \rho)$. In particular, we have $I \smallsetminus \bar{x}(I) = J \smallsetminus \bar{y}(I)$ with $\bar{x}(I) < \bar{y}(I) < \bar{y}(J)$. The first inequality holds since the flip $I \to J$ is increasing, and the second holds since the unique path from I to $\mathsf{N}(\mathsf{Q}, \rho)$ in $\mathsf{N}^*(\mathsf{Q}, \rho)$ is n-rising. We want to prove that $y(I) = \bar{y}(I)$, assuming by induction that $y(J) = \bar{y}(J)$. First, since $\bar{y}(I) < \bar{y}(J) = y(J)$ and $\mathit{\Pi}\mathsf{Q}_{[\bar{y}(I)] \smallsetminus J} = \mathit{\Pi}\mathsf{Q}_{[\bar{y}(I)] \smallsetminus I}$, we observe that

$$\bar{y}(I) \in J \cap [\bar{y}(I)] = \mathsf{N}(\mathsf{q}_1 \cdots \mathsf{q}_{\bar{y}(I)}, \mathit{\Pi}\mathsf{Q}_{[\bar{y}(I)] \smallsetminus J}) = \mathsf{N}(\mathsf{q}_1 \cdots \mathsf{q}_{\bar{y}(I)}, \mathit{\Pi}\mathsf{Q}_{[\bar{y}(I)] \smallsetminus I}).$$

Since $\bar{y}(I) \notin I \cap [\bar{y}(I)]$, this implies that $y(I) \leq \bar{y}(I)$. Second, the negative greedy flip property of Proposition 9 ensures that

$$I \cap [\bar{y}(I) - 1] = \mathsf{N}(\mathsf{q}_1 \cdots \mathsf{q}_{\bar{y}(I)-1}, \mathit{\Pi}\mathsf{Q}_{[\bar{y}(I)-1] \smallsetminus I})$$

since it is obtained from $J \cap [\bar{y}(I)] = \mathsf{N}(\mathsf{q}_1 \cdots \mathsf{q}_{\bar{y}(I)}, \mathit{\Pi}\mathsf{Q}_{[\bar{y}(I)] \smallsetminus J})$ by flipping $\bar{y}(I)$. Thus, we obtain that $y(I) > \bar{y}(I) - 1$. This concludes the proof that $y(I) = \bar{y}(I)$.

Finally, since $I \smallsetminus \bar{x}(I) = J \smallsetminus \bar{y}(I) = J \smallsetminus y(I)$, we know that $\mathsf{r}(I, \bar{x}(I)) = \mathsf{r}(I, y(J))$ by Proposition 2(3). Moreover, it has to be the smallest position in I with this property since otherwise $y(J)$ would be smaller than $y(I)$.

Finally, we give a similar direct characterization of the father of a facet I of $\mathbf{SC}(\mathsf{Q}, \rho)$ in the positive source tree $\mathbf{P}(\mathsf{Q}, \rho)$. This description can be understood in terms of the longest greedy suffix of I.

Proposition 13. *Let $I \neq \mathbf{P}(\mathsf{Q}, \rho)$ be a facet of $\mathbf{SC}(\mathsf{Q}, \rho)$. Define $y = y(I)$ to be the largest position in $[m]$ such that*

$$\{i - y \mid i \in I \smallsetminus [y]\} \neq \mathbf{P}(\mathsf{q}_{y+1} \cdots \mathsf{q}_m, \mathit{\Pi}\mathsf{Q}_{[y+1,m] \smallsetminus I}),$$

and $x = x(I)$ to be the largest position in I such that $\mathsf{r}(I, x) = -\mathsf{r}(I, y)$. Then the father of the facet I in the positive sink tree $\mathbf{P}^*(\mathsf{Q}, \rho)$ is obtained from I by flipping position x.

Proof. We can either translate the proof of Proposition 12, or directly apply to Proposition 12 the reversal operation of Remark 2. □

4.4 Greedy Flip Algorithm

The initial motivation of this paper was to find efficient algorithms for the exhaustive generation of the set $\mathbf{F}(\mathsf{Q}, \rho)$ of facets of the subword complex $\mathbf{SC}(\mathsf{Q}, \rho)$. For the evaluation of the time and space complexity of the different enumeration algorithms, we consider as parameters the rank n of the Coxeter group W and the size m of the word Q. Neither of these two parameters can be considered to be constant a priori. For example, if we want to generate all triangulations of a convex $(n + 3)$-gon (see Example 6), we consider a subword complex with a group W of rank n and with a word Q of size $n(n + 3)/2$.

The properties of the subword complex described in Sects. 3.3 and 3.4 already provide two immediate enumeration algorithms. First, the inductive structure of $\mathbf{F}(\mathsf{Q}, \rho)$ yields an *inductive algorithm* whose running time per facet is polynomial. More precisely, since all subword complexes which appear in the different cases of the right induction formula of Sect. 3.3 are non-empty, and since the tests $\rho \not< \mathsf{Q}_{\dashv}$ and $\ell(\rho q_m) > \ell(\rho)$ can be performed in $O(mn)$ time, the running time per facet of this inductive algorithm is in $O(m^2 n)$.

The second option is an *exploration of the flip graph* $\mathbf{G}(\mathsf{Q}, \rho)$. This flip graph is connected by Theorem 1, and it has degree bounded by $m - \ell(\rho)$. We can thus generate $\mathbf{F}(\mathsf{Q}, \rho)$ exploring the flip graph, and we need $O(m - \ell(\rho))$ flips per facet for this exploration. By Proposition 2, we can perform flips in the subword complex $\mathbf{SC}(\mathsf{Q}, \rho)$ in $O(mn)$ time if we store and update the facets of $\mathbf{F}(\mathsf{Q}, \rho)$ together with their root functions (note that this storage requires $O(mn)$ space).

We thus obtain again a running time of $O(m^2 n)$ per facet. The problem of a naive exploration of the flip graph is that we need to store all facets of $\mathbf{F}(Q, \rho)$ during the algorithm, which may require an exponential working space. This happens for example if we want to generate the $\frac{1}{n+2}\binom{2n+2}{n+1}$ triangulations of a convex $(n+3)$-gon (see Example 6).

Using the canonical spanning trees constructed in this paper, we can bypass this difficulty: we avoid to store all visited facets while preserving the same running time. The *greedy flip algorithm* generates all facets of the subword complex $\mathbf{SC}(Q, \rho)$ by a depth first search procedure on one[2] of the four canonical spanning trees described in Sect. 4.3. The preorder traversal of the tree also provides an iterator on the facets of $\mathbf{SC}(Q, \rho)$. Given a facet $I \in \mathbf{F}(Q, \rho)$, we can indeed compute its next element in the preorder traversal of the spanning tree, provided we know its root function (plus the path from I to the root in the tree if we work with either $\mathbf{P}(Q, \rho)$ or $\mathbf{N}^*(Q, \rho)$). These data can be updated at each step of the algorithm, using Proposition 2 for the root function.

We now bound the time and space complexity of the greedy flip algorithm. First, its working space is in $O(mn)$ since we only need to remember during the algorithm the current facet, together with its root function (plus its path to the root in the tree if we work with either $\mathbf{P}(Q, \rho)$ or $\mathbf{N}^*(Q, \rho)$). Concerning running time, each facet needs at most m flips to generate all its children in the spanning tree. Since a flip can be performed in $O(mn)$ time (by Proposition 2), the running time per facet of the greedy flip algorithm is still in $O(m^2 n)$.

We have implemented the greedy flip algorithm using the mathematical software Sage [19] as part of a project[3] on implementing subword complexes. We have seen that these two algorithms for generating facets have the same theoretical complexity, namely $O(m^2 n)$ per facet. To compare their experimental running time, we have constructed the k-cluster complex of type A_n for increasing values of k and n. Its facets correspond to the k-triangulations of the $(n + 2k + 1)$-gon (see Example 6 and [4] for the definition of multicluster complexes in any finite type). The rank of the group is n, while the length of the word is $kn + \binom{n}{2}$. Figure 13 presents the running time per facet for both enumeration algorithms in two situations: on the left, k is fixed at 1 while n increases; on the right, n is fixed at 3 while k increases. The greedy flip algorithm is better than the inductive algorithm in the first situation, and worse in the second. We observe a similar behavior for the computation of k-cluster complexes of types B_n and D_n. In general, the inductive algorithm is experimentally faster when the Coxeter group is fixed, but slower when the size of the Coxeter group increases.

Remark 5. Our algorithm is similar to that of [3] for pointed triangulations and that of [10] for primitive sorting networks. More precisely, the algorithms of [3] and [10]

---

[2]As observed by M. Pocchiola, searching on the positive sink tree or on the negative source tree improves the working space of the algorithm. This issue is relevant for the enumeration of pseudotriangulations and will be discussed in a forthcoming paper of his.

[3]The ongoing work on this patch can be found at http://trac.sagemath.org/sage_trac/ticket/11010.
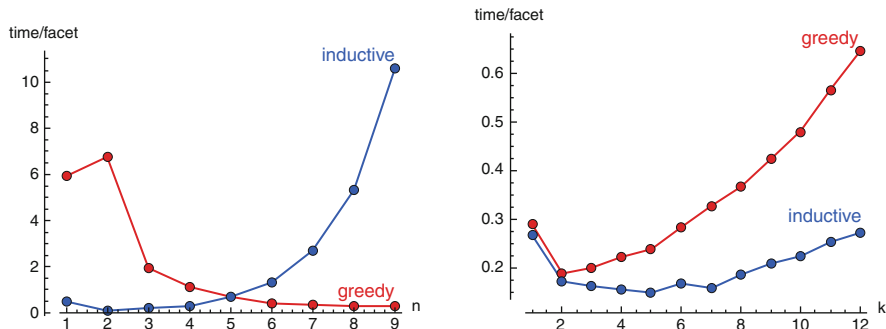
Fig. 13 Comparison of the running times of the inductive algorithm and the greedy flip algorithm to generate the k-cluster complex of type A_n. On the *left*, k is fixed at 1 while n increases; on the *right*, n is fixed at 3 while k increases. The time is presented in millisecond per facet

are both depth first search procedures on the positive source tree of particular subword complexes: subword complexes modeling pointed pseudotriangulations for [3] (see Example 6), and type A spherical subword complexes for [10].

5 Further Combinatorial Properties of the EL-Labelings

In this section, we discuss some implications of the EL-labelings of the increasing flip graph presented in Sect. 4.1. These results concern combinatorial properties of the *increasing flip poset* $\Gamma(Q, \rho)$, defined as the transitive closure of the increasing flip graph $\mathbf{G}(Q, \rho)$. The key requirement for the validity of these results is that the increasing flip graph $\mathbf{G}(Q, \rho)$ coincides with the Hasse diagram of the increasing flip poset $\Gamma(Q, \rho)$ (see the discussion in the beginning of Sect. 2.2). We first characterize and study the subword complexes which fulfill this property.

5.1 Double Root Free Subword Complexes

We say that the subword complex $\mathbf{SC}(Q, \rho)$ has a *double root* if there is a facet I in $\mathbf{SC}(Q, \rho)$ and two distinct positions $i \neq j \in [m]$ both flippable in I such that $\mathsf{r}(I, i) = \mathsf{r}(I, j)$. Otherwise, we say that the subword complex $\mathbf{SC}(Q, \rho)$ is *double root free*. In this section, we focus on double root free subword complexes due to the following characterization.

Proposition 14. *The subword complex $\mathbf{SC}(Q, \rho)$ is double root free if and only if its increasing flip graph $\mathbf{G}(Q, \rho)$ coincides with the Hasse diagram of its increasing flip poset $\Gamma(Q, \rho)$.*

Proof. Assume that $\mathbf{SC}(Q, \rho)$ has a double root. Let $i \neq j \in [m]$ be both flippable in I, and let $k \in [m] \smallsetminus I$ be such that $\mathsf{r}(I, i) = \mathsf{r}(I, j) = \pm\mathsf{r}(I, k)$ so that both i and j flip to k. Then the flip graph $\mathbf{G}(Q, \rho)$ contains a triangle formed by the facets I, $I \triangle \{i, k\}$, and $I \triangle \{j, k\}$ (where $A \triangle B := (A \cup B) \smallsetminus (A \cap B)$ denotes the *symmetric difference* of two sets A and B). Since a Hasse diagram cannot contain a triangle, the Hasse diagram of the increasing flip poset $\Gamma(Q, \rho)$ is only a strict subgraph of the increasing flip graph $\mathbf{G}(Q, \rho)$.

Assume reciprocally that the Hasse diagram of the increasing flip poset $\Gamma(Q, \rho)$ is a strict subgraph of the increasing flip graph $\mathbf{G}(Q, \rho)$. Let $I \to J$ be an oriented edge in $\mathbf{G}(Q, \rho)$ which is not an edge in the Hasse diagram of $\Gamma(Q, \rho)$. Let $i \in I$ and $j \in J$ be such that $I \smallsetminus i = J \smallsetminus j$ (thus $i < j$), and consider a path $I = I_1 \to \cdots \to I_{\ell+1} = J$ of increasing flips which prevents the edge $I \to J$ to be in the Hasse diagram of $\Gamma(Q, \rho)$ (in particular, $\ell > 1$). Let $\mathsf{p}_1 > \ldots > \mathsf{p}_\ell$ be the decreasing reordering of the set $\{\mathsf{p}(I_1 \to I_2), \ldots, \mathsf{p}(I_\ell \to I_{\ell+1})\}$ of positive edge labels along this path, and let $\mathsf{n}_1, \ldots, \mathsf{n}_\ell$ be the corresponding negative edge labels. That is to say, when we flip p_k out of a certain facet in this path, we obtain n_k in the next facet of the path. Since I and J differ only in positions i and j with $i < j$, and all flips are increasing, no position smaller than i can be flipped. Thus, we obtain that $\mathsf{p}_\ell = i$, and by a similar argument that $\mathsf{n}_1 = j$. Applying the same argument to the other positions that are flipped along the path, in increasing or in decreasing order, moreover gives

$$i = \mathsf{p}_\ell < \mathsf{n}_\ell = \mathsf{p}_{\ell-1} < \cdots < \mathsf{n}_2 = \mathsf{p}_1 < \mathsf{n}_1 = j.$$

Proposition 2 thus ensures that all roots $\mathsf{r}(I, \mathsf{p}_1), \ldots, \mathsf{r}(I, \mathsf{p}_\ell)$ coincide and are equal to $\mathsf{r}(I, \mathsf{n}_1)$, and that we moreover have $\mathsf{p}_k = \mathsf{p}(I_k \to I_{k+1})$ and $\mathsf{n}_k = \mathsf{n}(I_k \to I_{k+1})$. Since $\ell > 1$, this completes the proof. $\square$

The intervals in the increasing flip graph of a double root free subword complex have the following property. We will see in Remark 6 that this property, as well as its corollaries below, does not hold for subword complexes with double roots.

Proposition 15. *Let I and J be two facets of a double root free subword complex $\mathbf{SC}(Q, \rho)$. Then the intersection $I \cap J$ is contained in all facets of the interval $[I, J]$ in the increasing flip graph $\mathbf{G}(Q, \rho)$.*

We extract the crucial part of the proof of this proposition in the following lemma.

Lemma 4. *Let $I_0 \to I_1 \to \cdots \to I_{\ell+1}$ be a path in the increasing flip graph $\mathbf{G}(Q, \rho)$ with $\mathsf{p}_k := \mathsf{p}(I_k \to I_{k+1})$ and such that $\mathsf{p}_0 = \max\{\mathsf{p}_0, \ldots, \mathsf{p}_\ell\}$. Then, starting from I_0, it is possible to skip the first flip at position p_0, and directly successively flip positions $\mathsf{p}_1, \mathsf{p}_2, \ldots, \mathsf{p}_\ell$. If $I_0 = I_1' \to I_2' \to \cdots \to I_{\ell+1}'$ is the corresponding path for which $\mathsf{p}(I_k', I_{k+1}') = \mathsf{p}_k$ for all $k \in [\ell]$, we moreover have that $\mathsf{r}(I_k', p) = \mathsf{r}(I_k, p)$ for any position $p \leq \mathsf{p}_0$ and any $k \in [\ell + 1]$.*

Proof. The proof is based on the observation that flips are described using the root function, and that flipping out i and flipping in j only affects the roots located

between positions i and j, see Proposition 2. Remember that a position p in a facet I is increasingly flippable if and only if the root $r(I, p)$ is contained in the inversion set of ρ, compare Proposition 2(2) and (4).

We prove the statement by induction on k. Namely, we prove that

1. $r(I'_1, p) = r(I_1, p)$ for all positions $p \leq p_0$, and that
2. For any $k \in [\ell]$, if $r(I'_k, p) = r(I_k, p)$ for all positions $p \leq p_0$, then the position p_k is increasingly flippable in I'_k and $r(I'_{k+1}, p) = r(I_{k+1}, p)$ for all positions $p \leq p_0$.

To prove (1), observe that the flip $I_0 \to I_1$ does not affect roots located to the left of position p_0, so we have $r(I'_1, p) = r(I_0, p) = r(I_1, p)$ for any position $p \leq p_0$.

To prove (2), we assume that $r(I'_k, p) = r(I_k, p)$ for all positions $p \leq p_0$. In particular, $r(I'_k, p_k) = r(I_k, p_k)$ because $p_k \leq p_0$. Since p_k is increasingly flippable in I_k, this root is in the inversion set of ρ, and therefore, p_k is also increasingly flippable in I'_k. Here, we used twice Proposition 2(2) and (4). Define now $n_k := n(I_k \to I_{k+1})$ and $n'_k := n(I'_k \to I'_{k+1})$. If $n_k \leq p_0$, then

$$r(I'_k, p_k) = r(I_k, p_k) = r(I_k, n_k) = r(I'_k, n_k),$$

and thus $n_k = n'_k$. Here, we used twice Proposition 2(3). Similarly, if $n'_k \leq p_0$, then $n_k = n'_k$. We therefore obtain that either both n_k and n'_k are located to the right of p_0, or $n_k = n'_k$. In both cases, we know that $p_k < p \leq n_k$ if and only if $p_k < p \leq n'_k$ for any position $p \leq p_0$. Since $r(I'_k, p) = r(I_k, p)$, we thus obtain that $r(I'_{k+1}, p) = r(I_{k+1}, p)$ by Proposition 2(5).

Proof (Proof of Proposition 15). Let $I = I_0 \to I_1 \to \cdots \to I_\ell \to I_{\ell+1} = J$ be a path from I to J in the increasing flip graph $\mathbf{G}(Q, \rho)$. For $0 \leq k \leq \ell$, define $p_k := p(I_k \to I_{k+1})$ and $n_k := n(I_k \to I_{k+1})$. In other words, $p_k \in I_k$, $n_k \in I_{k+1}$ and $I_k \smallsetminus p_k = I_{k+1} \smallsetminus n_k$.

We assume by means of contradiction that there is a position in $I \cap J$ flipped out during the flip path which is flipped back later in the path. Up to shortening the path, we can assume without loss of generality that this position is flipped out during the first flip $I_0 \to I_1$ and flipped back in during the last flip $I_\ell \to I_{\ell+1}$, i.e. $p_0 = n_\ell$. We moreover assume that our path is a minimal length path which flips back in a position already flipped out.

Under these assumptions, we prove that

1. $p_0 = \max\{p_0, \ldots, p_\ell\}$,
2. Starting from facet I, we can successively flip positions $p_1, p_2, \ldots, p_{\ell-1}$ (just skipping the first and the last flips at positions p_0 and p_ℓ), and
3. The facet J' obtained after these flips has a double root at positions p_0 and p_ℓ.

To prove (1), assume that the index $m \in [0, \ell]$ such that $p_m = \max\{p_0, \ldots, p_\ell\}$ is different from 0. Note that $0 < m < \ell$ since $p_\ell < n_\ell = p_0$. Consider the path of flips

$$I = I_0 \to \cdots \to I_m = I'_m \to I'_{m+1} \to \cdots \to I'_\ell$$

defined by $\mathsf{p}(I_k, I_{k+1}) = \mathsf{p}_k$ for $k < m$ and $\mathsf{p}(I'_k, I'_{k+1}) = \mathsf{p}_{k+1}$ for $k \geq m$. In other words, starting from I, we flip positions $\mathsf{p}_0, \ldots, \mathsf{p}_{m-1}, \mathsf{p}_{m+1}, \ldots, \mathsf{p}_\ell$, skipping the flip at position p_m. According to Lemma 4, all flips in the path $I'_m \to I'_{m+1} \to \cdots \to I'_\ell$ are admissible since $\mathsf{p}_k \leq \mathsf{p}_m$ for all $k \geq m$, and we have

$$\mathsf{r}(I'_{\ell-1}, \mathsf{p}_\ell) = \mathsf{r}(I_\ell, \mathsf{p}_\ell) = \mathsf{r}(I_\ell, \mathsf{p}_0) = \mathsf{r}(I'_{\ell-1}, \mathsf{p}_0).$$

Therefore, we flip back position p_0 in facet I'_ℓ, thus contradicting the length minimality of the path $I = I_0 \to I_1 \to \cdots \to I_\ell \to I_{\ell+1} = J$.

We now prove (2) and (3). By (1), the path $I = I_0 \to I_1 \to \cdots \to I_\ell \to I_{\ell+1} = J$ satisfies the hypothesis of Lemma 4. We therefore obtain directly (2). Let J' denote the facet of $\mathbf{F}(Q, \rho)$ obtained after flipping successively $\mathsf{p}_1, \mathsf{p}_2, \ldots, \mathsf{p}_{\ell-1}$ starting from I. We moreover obtain

$$\mathsf{r}(J', \mathsf{p}_\ell) = \mathsf{r}(I_\ell, \mathsf{p}_\ell) = \mathsf{r}(I_\ell, \mathsf{p}_0) = \mathsf{r}(J', \mathsf{p}_0),$$

where the first and last equalities are ensured by Lemma 4, while the middle one holds by Proposition 2(3) since we flip position p_ℓ to position $\mathsf{n}_\ell = \mathsf{p}_0$ in facet I_ℓ. Since the facet J' contains both p_0 and p_ℓ, it has a double root, thus proving (3). $\square$

The following theorem is now a direct consequence of Proposition 15.

Theorem 3. *There is at most one p-falling (resp. n-falling) path between any two facets I and J of a double root free subword complex $\mathbf{SC}(Q, \rho)$. If it exists, its length is given by $|I \smallsetminus J| = |J \smallsetminus I|$.*

Proof. Let $I = I_1 \to \cdots \to I_{\ell+1} = J$ be a p-falling path from I to J in the increasing flip graph $\mathbf{G}(Q, \rho)$, and define $\mathsf{p}_k := \mathsf{p}(I_k \to I_{k+1})$ and $\mathsf{n}_k := \mathsf{n}(I_k \to I_{k+1})$. For $k < k'$, we then have $\mathsf{n}_k \neq \mathsf{p}_{k'}$ (because the flips are increasing and the path is p-falling) and $\mathsf{p}_k \neq \mathsf{n}_{k'}$ (otherwise, the position $\mathsf{p}_k = \mathsf{n}_{k'}$ would be flipped out and flipped back in during the path, thus contradicting Proposition 15). This implies that $\mathsf{p}_k \in I \smallsetminus J$ and $\mathsf{n}_k \in J \smallsetminus I$ for all $k \in [\ell]$. Therefore p_k is the kth largest position of $I \smallsetminus J$ and $\ell = |I \smallsetminus J| = |J \smallsetminus I|$. This uniquely determines the p-falling path from I to J. The proof is similar for the n-falling path (see also Proposition 5). $\square$

Corollary 1. *Let I and J be two facets of a double root free subword complex such that $I \dashrightarrow J$. The unique p-rising (resp. n-rising) path from I to J has maximal length among all path from I to J. Moreover, if there is a p-falling (resp. n-falling) path from I to J, it has minimal length.*

Proof. Consider a maximal length path from I to J. According to the proof of Theorem 2, we can modify this path to obtain the unique p-rising path from I to J. In the situation of a double root free subword complex, this procedure does not decrease the length of the path, since the first distinguished case in the proof of Lemma 1 cannot occur. This proves the result for the p-rising path. For the p-falling path, this follows directly from Theorem 3. The proof is similar for the negative edge labeling n. $\square$

Remark 6. Note that the conclusions of Proposition 15, Theorem 3, and Corollary 1 do indeed not hold if $\mathbf{SC}(Q, \rho)$ has double roots. Whenever one has a double root, one can reduce the situation to type A_1 with generator s for the word $Q = sss$ and the element $\rho = s$, using Proposition 3 (one might actually get that the word Q contains more than three letters, but the argument stays the same). In this case, the increasing flip graph $\mathbf{G}(Q, \rho)$ consists of the two paths

$$\{1, 2\} \ -2\!\rightarrow \{1, 3\} \ -1\!\rightarrow \{2, 3\} \quad \text{and} \quad \{1, 2\} \ -1\!\rightarrow \{2, 3\},$$

where the numbers on the edges are their positive edge labels. First, $\{1, 3\}$ lies in the interval $[\{1, 2\}, \{2, 3\}]$ of the increasing flip graph $\mathbf{G}(Q, \rho)$, but does not contain $\{1, 2\} \cap \{2, 3\} = \{2\}$, thus contradicting Proposition 15. Second, both paths are p-falling, contradicting the conclusions of Theorem 3. Third, the second path is p-rising and shorter than the first p-falling path, contradicting the conclusions of Corollary 1.

Corollary 2. *The Möbius function on the increasing flip poset $\Gamma(Q, \rho)$ of a double root free subword complex $\mathbf{SC}(Q, \rho)$ is given by*

$$\mu(I, J) = \begin{cases} (-1)^{|J \smallsetminus I|} & \textit{if there is a } \mathsf{p}\textit{-falling (resp. } \mathsf{n}\textit{-falling) path from } I \text{ to } J, \\ 0 & \textit{otherwise.} \end{cases}$$

Proof. This is a direct consequence of Propositions 1 and 5 and Theorem 3.

By this corollary, we can compute the Möbius function of an interval $[I, J]$ of the increasing flip poset as soon as we can decide whether or not there is a p-falling path from I to J. According to Proposition 7, there is always a p-falling path from the positive greedy facet to the negative greedy facet of a spherical subword complex. We therefore obtain the value of the Möbius function on the increasing flip poset $\Gamma(Q, \rho)$ of a spherical double root free subword complex.

Corollary 3. *In a spherical double root free subword complex $\mathbf{SC}(Q, \rho)$, we have*

$$\mu\big(\mathsf{P}(Q, \rho), \mathsf{N}(Q, \rho)\big) = (-1)^{|Q| - \ell(\rho)}.$$

Observe again that this result fails if we drop the condition that $\mathbf{SC}(Q, \rho)$ is spherical. The subword complex $\mathbf{SC}(Q^{\mathrm{ex}}, \rho^{\mathrm{ex}})$ of Example 4 and the subword complex $\mathbf{SC}(\tau_1 \tau_2 \tau_1 \tau_2, \tau_1 \tau_2)$ provide counter-examples.

5.2   Two Relevant Examples

We finish this section by two relevant families of examples of double root free subword complexes, to which the above results can be applied.

5.2.1 Cambrian Lattices

We start with recalling background on sortable elements in Coxeter groups and Cambrian lattices. Those were introduced by N. Reading in [14–17], originally to connect finite type cluster complexes to noncrossing partitions. Fix a Coxeter element c of W, and a reduced expression c of c. That is to say, c is a word on S where each simple reflection appears precisely once. For $w \in W$, we denote by w(c) the c-*sorting word* of w, i.e. the lexicographically first (as a sequence of positions) reduced subword of c^∞ for w. Moreover, this word can be written as w(c) $=$ $c_{K_1}c_{K_2}\cdots c_{K_p}$, where c_K denotes the subword of c only taking the simple reflections in $K \subset S$ into account. The element w is then called *c-sortable* if $K_1 \supseteq K_2 \supseteq \cdots \supseteq K_p$. Observe that the property of being c-sortable does not depend on the particular reduced expression c of the Coxeter element c. We denote by $\mathrm{SORT}_c(W)$ the set of c-sortable elements in W. The order induced by the weak order on W turns $\mathrm{SORT}_c(W)$ into a lattice, the *Cambrian lattice* for the Coxeter element c [17].

It was observed in [16, Remark 2.1] that Cambrian lattices are naturally equipped with a search-tree structure. The c-*sorting tree* $\mathbf{T}(c)$ has an edge between two c-sortable elements w and w' if the c-sorting word for w is obtained from the one for w' by deleting the last letter. See Example 15 and Fig. 14. Observe that the c-sorting tree really depends on the particular choice for the reduced expression c, and not only on the Coxeter element c.

In their recent work [6], M. Kallipoliti and H. Mühle define an EL-labeling of the Cambrian lattice $\mathrm{SORT}_c(W)$ as follows. They label a cover relation $w \to w'$ of $\mathrm{SORT}_c(W)$ by the first position within c^∞ which is used in the c-sorting word for w' but not in the c-sorting word for w. They observed in [6, Remark 3.5] that the spanning tree formed by all rising paths from the source e to any other c-sortable element coincides with the c-sorting tree mentioned above. See Example 15 and Fig. 14. They moreover use this EL-labeling to derive results on Möbius functions of Cambrian lattices [6, Theorems 4.1–4.3].

Example 15. Let $W = \mathfrak{S}_4$ and $c = \tau_1\tau_2\tau_3$. The c-sortable elements, the Hasse diagram of the Cambrian lattice, the EL-labeling of [6], and the c-sorting tree are represented in Fig. 14. We write *12.1* instead of $\tau_1\tau_2.\tau_1$ to simplify the picture (the dots mark the separation between the blocks c_{K_i}).

We now recall that Cambrian lattices can be seen as increasing flip posets. This interpretation was presented in [12, Sects. 6.3.2 and 6.4], based on previous connections between c-sortable elements and c-clusters [17], and between c-clusters and facets of the subword complex [4].

Let $w_\circ(c)$ denote the c-sorting word for the longest element $w_\circ \in W$. To simplify notations, we write $\mathbf{SC}(c)$ for the subword complex $\mathbf{SC}(cw_\circ(c), w_\circ)$. Similarly, we denote by $\mathbf{F}(c)$ its facets, by $\mathbf{G}(c)$ its increasing flip graph, by $\Gamma(c)$ its increasing flip poset, and by $\mathbf{P}(c)$ its positive source tree. Following [12, Sect. 5.1], we define a map $\kappa : W \to \mathbf{F}(c)$ by sending an element $w \in W$ to the unique facet $\kappa(w)$ whose root configuration $\mathsf{R}(\kappa(w))$ is contained in $w(\Phi^+)$. For the subword complex $\mathbf{SC}(c)$,
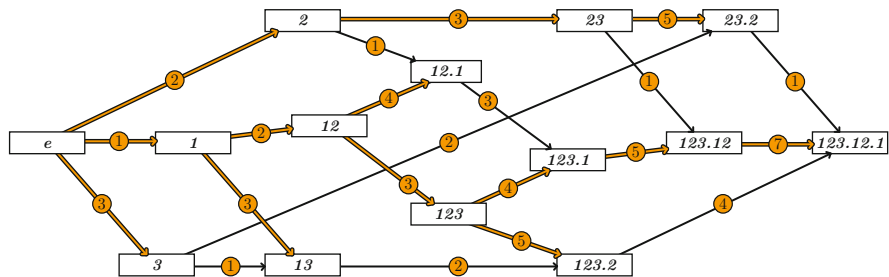
Fig. 14 The $(\tau_1\tau_2\tau_3)$-sorting tree is the spanning tree corresponding to an EL-labeling of the $(\tau_1\tau_2\tau_3)$-Cambrian lattice
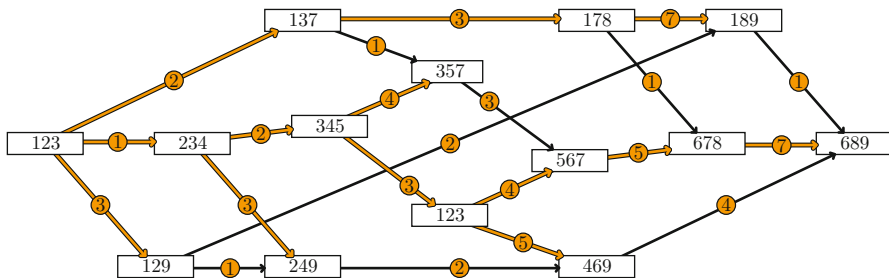

Fig. 15 The positive edge labeling p of $\mathbf{G}(\tau_1\tau_2\tau_3)$, and the positive source tree $\mathbf{P}(\tau_1\tau_2\tau_3)$

it turns out that the fibers of this map are intervals, and that their minimal elements are precisely the c-sortable elements. This gives the following proposition.

Proposition 16 ([12, Corollary 6.31]). *The map associating to a facet I the unique (weak order) minimal element in $\kappa^{-1}(I)$, is a poset isomorphism between the increasing flip poset and the Cambrian lattice.*

Through this isomorphism, we can transfer the results discussed in this paper to Cambrian lattices. We thus also obtain natural EL-labelings and spanning trees for Cambrian lattices.

Example 16. Let $W = \mathfrak{S}_4$ and $c = \tau_1\tau_2\tau_3$. The facets of $\mathbf{SC}(c)$, the Hasse diagram of $\Gamma(c)$, the positive edge labeling p of $\mathbf{G}(c)$, and the positive source tree $\mathbf{P}(c)$ are represented in Fig. 15. Compare to Fig. 14.

To finish, we want to observe that the positive edge labeling differs from the EL-labeling of [6] and that the positive source tree $\mathbf{P}(c)$ differs[4] from the c-sorting tree $\mathbf{T}(c)$. This is illustrated in the following (minimal) example.

---

[4]The contrary was stated in a previous version of this paper. We thank an anonymous referee for pointing out this mistake.
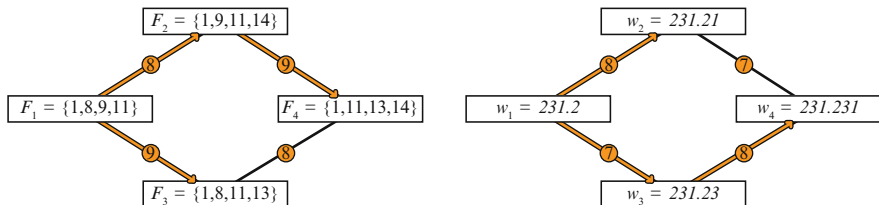
Fig. 16 The positive source tree **T**(c) differs from the c-sorting tree

Example 17 (Positive source tree $\neq$ Coxeter-sorting tree). Consider the Coxeter group $W = \mathfrak{S}_5$ and the Coxeter element $c = \tau_4\tau_2\tau_3\tau_1$. In this situation, the four facets of **SC**(c) given by

$$F_1 = \{1, 8, 9, 11\}, \quad F_2 = \{1, 9, 11, 14\}, \quad F_3 = \{1, 8, 11, 13\}, \quad F_4 = \{1, 11, 13, 14\},$$

are respectively sent by the isomorphism of Proposition 16 to the c-sortable elements

$$w_1 = \tau_2\tau_3\tau_1.\tau_2, \quad w_2 = \tau_2\tau_3\tau_1.\tau_2\tau_1, \quad w_3 = \tau_2\tau_3\tau_1.\tau_2\tau_3, \quad w_4 = \tau_2\tau_3\tau_1.\tau_2\tau_3\tau_1.$$

The facets F_1, F_2, F_3, F_4 (resp. the c-sortable elements w_1, w_2, w_3, w_4) form a square within the increasing flip poset (resp. within the Cambrian lattice). Figure 16 represents the two EL-labelings and their corresponding spanning trees restricted to these squares. The positive source tree **P**(c) contains all edges of this square except $F_3 \to F_4$, while the c-sorting tree **T**(c) contains all edges of this square except $w_2 \to w_4$.

5.2.2 Duplicated Words

Let $\rho := \rho_1 \cdots \rho_\zeta$ be a reduced expression of an element ρ of W. For $k \in [\zeta]$, we define a root $\alpha_k := \rho_1 \cdots \rho_{k-1}(\alpha_{\rho_k})$. Note that the roots $\alpha_1, \ldots, \alpha_\zeta$ are pairwise distinct and positive. They are the roots of the inversion set of ρ.

Let X be an arbitrary subset of $\chi := |X|$ positions of $[\zeta]$. We denote by Q^{dup} the word on S with $\zeta + \chi$ letters which is obtained by duplicating the letters of $\rho := \rho_1 \cdots \rho_\zeta$ at positions in X. To be more precise, define $k^\bullet := k + |X \cap [k-1]|$ for $k \in [\zeta]$. Observe that $[\zeta + \chi] = \{k^\bullet \mid k \in [\zeta]\} \sqcup \{x^\bullet + 1 \mid x \in X\}$. Then, we set $Q^{\mathrm{dup}} := q_1 \cdots q_{\zeta+\chi}$, where $q_{k^\bullet} := \rho_k$ for $k \in [\zeta]$ and $q_{x^\bullet+1} := \rho_x$ for $x \in X$. For $k \in [\zeta]$, the position $k^\bullet$ is the new position in Q^{dup} of the kth letter of ρ, and for $x \in X$, the position $x^\bullet + 1$ is the new position in Q^{dup} of the duplicated xth letter of ρ.

For any $x \in X$, the pair $\{x^\bullet, x^\bullet + 1\}$ of duplicated positions intersects any facet of **SC**(Q^{dup}, ρ), otherwise the expression would not be reduced. It follows that any

facet of $\mathbf{SC}(Q^{dup}, \rho)$ contains precisely one element of each pair $\{x^\bullet, x^\bullet + 1\}$ of duplicated positions and no other position. Therefore, the facets of $\mathbf{SC}(Q^{dup}, \rho)$ are precisely the sets $I_\varepsilon := \{x^\bullet + \varepsilon_x \mid x \in X\}$ where $\varepsilon := (\varepsilon_1, \ldots, \varepsilon_\chi) \in \{0, 1\}^X$. Moreover, the roots of the facet I_ε of $\mathbf{SC}(Q^{dup}, \rho)$ are given by $\mathsf{r}(I_\varepsilon, k^\bullet) = \alpha_k$ for $k \in [\zeta]$ and $\mathsf{r}(I_\varepsilon, x^\bullet + 1) = (-1)^{\varepsilon_x} \alpha_x$ for $x \in X$. Thus, the subword complex $\mathbf{SC}(Q^{dup}, \rho)$ is double root free, since the roots $\alpha_1, \ldots, \alpha_\zeta$ are pairwise distinct.

The subword complex $\mathbf{SC}(Q^{dup}, \rho)$ is the boundary complex of the χ-dimensional cross polytope. In particular, the graph of increasing flips $\mathbf{G}(Q^{dup}, \rho)$ is the directed 1-skeleton $\square_\chi$ of a χ-dimensional cube, and the increasing flip poset $\Gamma(Q^{dup}, \rho)$ is a boolean poset.

The positive greedy facet $\mathbf{P}(Q^{dup}, \rho)$ is the facet I_0, while the negative greedy facet $\mathbf{N}(Q^{dup}, \rho)$ is the facet I_1. The positive and negative edge labelings p and n of $\mathbf{SC}(Q^{dup}, \rho)$ are essentially the same as the edge labeling λ of $\square_\chi$ presented in Example 1. More precisely, for any edge $\varepsilon \to \varepsilon'$ of $\square_\chi$, we have

$$\psi \circ \lambda(\varepsilon \to \varepsilon') = \mathsf{p}(I_\varepsilon \to I_{\varepsilon'}) = \mathsf{n}(I_\varepsilon \to I_{\varepsilon'}) - 1,$$

where $\psi : [\chi] \to \{x^\bullet \mid x \in X\}$ is such that $\psi(1) < \psi(2) < \cdots < \psi(\chi)$. Since $\mathsf{p}(\cdot) = \mathsf{n}(\cdot) - 1$, the positive and negative source trees $\mathbf{P}(Q^{dup}, \rho)$ and $\mathbf{N}(Q^{dup}, \rho)$ coincide. Similarly the positive and negative sink trees $\mathbf{P}^*(Q^{dup}, \rho)$ and $\mathbf{N}^*(Q^{dup}, \rho)$ coincide as well. Moreover, the map $\varepsilon \mapsto I_\varepsilon$ defines a graph isomorphism from the λ-source tree of $\square_\chi$ to the source trees $\mathbf{P}(Q^{dup}, \rho) = \mathbf{N}^*(Q^{dup}, \rho)$, and from the λ-sink tree of $\square_\chi$ to the sink trees $\mathbf{P}^*(Q^{dup}, \rho) = \mathbf{N}^*(Q^{dup}, \rho)$. See Example 1 and Fig. 1.

Finally, the Möbius function on the increasing flip poset $\Gamma(Q^{dup}, \rho)$ is given by

$$\mu(I_\varepsilon, I_{\varepsilon'}) = \begin{cases} (-1)^{\delta(\varepsilon, \varepsilon')} & \text{if } \varepsilon \dashrightarrow \varepsilon', \\ 0 & \text{otherwise,} \end{cases}$$

where δ denotes the Hamming distance on the vertices of the cube. See Example 2.

References

 1. Björner, A.: Shellable and Cohen-Macaulay partially ordered sets. Trans. Am. Math. Soc. **260**(1), 159–183 (1980)
 2. Björner, A., Wachs, M.L.: Shellable nonpure complexes and posets. I. Trans. Am. Math. Soc. **348**(4), 1299–1327 (1996)
 3. Brönnimann, H., Kettner, L., Pocchiola, M., Snoeyink, J.: Counting and enumerating pointed pseudotriangulations with the greedy flip algorithm. SIAM J. Comput. **36**(3), 721–739 (electronic) (2006).
 4. Ceballos, C., Labbé, J.-P., Stump, C.: Subword complexes, cluster complexes, and generalized multi-associahedra. J. Algebraic Combin. 1–35 (2013). DOI 10.1007/s10801-013-0437-x
 5. Humphreys, J.E.: Reflection Groups and Coxeter Groups. Cambridge Studies in Advanced Mathematics, vol. 29. Cambridge University Press, Cambridge (1990)
 6. Kallipoliti, M., Mühle, H.: On the topology of the Cambrian semilattices (2012, Preprint). `arXiv:1206.6248`
 7. Knutson, A., Miller, E.: Subword complexes in Coxeter groups. Adv. Math. **184**(1), 161–176 (2004)
 8. Knutson, A., Miller, E.: Gröbner geometry of Schubert polynomials. Ann. Math. (2), **161**(3), 1245–1318 (2005)
 9. Papi, P.: A characterization of a special ordering in a root system. Proc. Am. Math. Soc. **120**(3), 661–665 (1994)
10. Pilaud, V., Pocchiola, M.: Multitriangulations, pseudotriangulations and primitive sorting networks. Discret. Comput. Geom. **48**(1), 142–191 (2012)
11. Pilaud, V., Santos, F.: Multitriangulations as complexes of star polygons. Discret. Comput. Geom. **41**(2), 284–317 (2009)
12. Pilaud, V., Stump, C.: Brick polytopes of spherical subword complexes: a new approach to generalized associahedra (2011, Preprint). `arXiv:1111.3349`
13. Pocchiola, M., Vegter, G.: Topologically sweeping visibility complexes via pseudotriangulations. Discret. Comput. Geom. **16**(4), 419–453 (1996)
14. Reading, N.: Lattice congruences of the weak order. Order **21**(4), 315–344 (2004/2005)
15. Reading, N.: Cambrian lattices. Adv. Math. **205**(2), 313–353 (2006)
16. Reading, N.: Clusters, Coxeter-sortable elements and noncrossing partitions. Trans. Am. Math. Soc. **359**(12), 5931–5958 (2007)
17. Reading, N.: Sortable elements and Cambrian lattices. Algebra Univers. **56**(3–4), 411–437 (2007)
18. Rote, G., Santos, F., Streinu, I.: Pseudo-triangulations—a survey. In: Goodman, J.E., Pach, J., Pollack, R. (eds.) Surveys on Discrete and Computational Geometry. Contemporary Mathematics, vol. 453, pp. 343–410. American Mathematical Society, Providence (2008)
19. Stein, W.A., et al.: Sage Mathematics Software (Version 4.8). The Sage Development Team. http://www.sagemath.org (2012)

Bandwidth, Vertex Separators, and Eigenvalue Optimization

Franz Rendl, Abdel Lisser, and Mauro Piacentini

Abstract A fundamental problem in numerical linear algebra consists in rearranging the rows and columns of a matrix in such a way that either the nonzero entries appear within a band of small width along the main diagonal, or such that the matrix has some block structure which is joined by only a few rows and columns. Such problems can be approached using graph partition techniques. From a practical point of view it is important that also large-scale instances can be dealt with. This rules out a direct application of the strong machinery for graph partition given by semidefinite optimization. We propose to use the weaker relaxations based on the Hoffman-Wielandt theorem, which lead to closed form bounds in terms of the Laplacian eigenvalues. We then try to improve these eigenvalue bounds by weight redistribution. This leads to nicely structured eigenvalue optimization problems. A similar approach has been used by Boyd, Diaconis and Xiao to increase the mixing rate of Markov chains. We use it to improve bounds on the bandwidth and the size of vertex separators in graphs. Moreover, the bounds can also be used to heuristically find good reorderings.

F. Rendl (✉)
Institut für Mathematik, Alpen-Adria Universität Klagenfurt,
Universitätsstraße 65-67, 9020 Klagenfurt, Austria
e-mail: franz.rendl@uni-klu.ac.at

A. Lisser
Laboratoire de Recherche en Informatique, Université Paris Sud,
PCRI, Batiment 650, Rue Noetzlin, 91190 Gif-sur-Yvette, France
e-mail: abdel.lisser@lri.fr

M. Piacentini
Sapienza – Università di Roma, Piazzale Aldo Moro 5, 00185 Roma, Italy
e-mail: piacentini@dis.uniroma1.it

1 Introduction

The investigation of structural properties of the nonzero pattern of matrices has important applications in many applied sciences. In numerical linear algebra for instance, the solution of a linear system may be speeded up in case that a reordering of the rows and columns of the system matrix is known such that all nonzeros after reordering are close to the main diagonal.

Alternatively a system matrix with block structure and only a few rows and columns linking the individual blocks leads to computational simplifications based on the Schur complement formula.

In this paper we investigate problems of this type. The zero pattern of a symmetric $n \times n$ matrix $M = (m_{ij})$ can be represented through a graph G having vertex set $N := \{1, \ldots, n\}$ and edges $[i, j]$ whenever $m_{ij} \neq 0$ for $i \neq j$. Thus $E(G) := \{[i, j] : m_{ij} \neq 0, \ i < j\}$. Since M is symmetric we take G as an undirected graph. We also ignore the entries on the main diagonal of M, hence G is also loopless.

If we relabel the vertices of G according to the permutation ϕ, then the bandwidth of G with respect to ϕ is defined to be

$$bw(\phi, G) := \max\{|\phi(i) - \phi(j)| : [i, j] \in E(G)\}.$$

The *bandwidth* of G is the minimum of $bw(\phi, G)$ over all permutations ϕ of N,

$$bw(G) := \min\{bw(\phi, G) : \phi \in \Pi_n\}.$$

The set of permutations of N is denoted by Π_n.

Determining $bw(G)$ is NP-hard, and it remains so even if G is restricted to the class of trees with maximal vertex degree equal three. In [6] several lower bounds for $bw(G)$ are investigated, based on the Laplacian spectrum of G. As an example, consider the matrix A from Table 1. Its bandwidth with respect to the given ordering is 10, attained at edge [1, 11]. After reordering according to the permutation p given in the second part of the table, the bandwidth is 4.

Another computational simplification for a linear system is possible in case that the system matrix has a vertex separator of small size. In graph theoretic terms a set $T \subseteq N$ is a *vertex separator* of G, if $G \setminus T$ decomposes into components of roughly equal size. Thus we are also interested in the question whether for a given number $k \geq 3$ there exists a subset $T \subseteq N$ such that the removal of T from G disconnects

Table 1 Matrix A and after reordering the rows and columns with respect to the permutation $p = (3\ 1\ 4\ 2\ 11\ 5\ 12\ 6\ 8\ 10\ 7\ 9)$. The zero entries are represented as dots

$$
A = \left(\begin{array}{ccccc|cccccc|cc}
. & 1 & 1 & . & . & . & . & . & . & . & . & 1 & . \\
1 & . & 1 & 1 & . & . & . & . & . & . & . & . & . \\
1 & 1 & . & . & . & . & . & . & . & . & . & 1 & . \\
. & 1 & . & . & 1 & . & . & . & . & . & . & . & 1 \\
. & . & . & 1 & . & . & . & . & . & . & . & . & 1 \\ \hline
. & . & . & . & . & . & 1 & . & 1 & . & . & . & 1 \\
. & . & . & . & . & 1 & . & 1 & 1 & . & . & . & . \\
. & . & . & . & . & . & 1 & . & 1 & 1 & . & 1 & . \\
. & . & . & . & . & 1 & 1 & 1 & . & 1 & . & . & . \\
. & . & . & . & . & . & 1 & 1 & 1 & . & . & . & . \\ \hline
1 & . & 1 & . & . & . & . & 1 & . & . & . & . & . \\
. & . & . & 1 & 1 & 1 & . & . & . & . & . & . & .
\end{array}\right)
\qquad
A_p = \left(\begin{array}{cccccccccccc}
. & 1 & . & 1 & 1 & . & . & . & . & . & . & . \\
1 & . & . & 1 & 1 & . & . & . & . & . & . & . \\
. & . & . & 1 & . & 1 & 1 & . & . & . & . & . \\
1 & 1 & 1 & . & . & . & . & . & . & . & . & . \\
1 & 1 & . & . & . & . & . & . & 1 & . & . & . \\
. & . & 1 & . & . & . & 1 & . & . & . & . & . \\
. & . & 1 & . & . & 1 & . & 1 & . & . & . & . \\
. & . & . & . & . & . & 1 & . & . & . & 1 & 1 \\
. & . & . & . & 1 & . & . & . & . & 1 & 1 & 1 \\
. & . & . & . & . & . & . & . & 1 & . & 1 & 1 \\
. & . & . & . & . & . & . & 1 & 1 & 1 & . & 1 \\
. & . & . & . & . & . & . & 1 & 1 & 1 & 1 & .
\end{array}\right)
$$

Table 2 The graph representing the nonzero entries of matrix A from Table 1
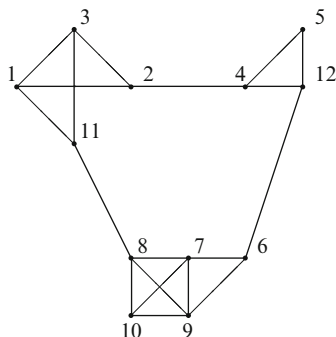
$G \setminus T$ into $k - 1$ components of roughly equal size. The graph in Table 2 has a vertex separator $T = \{11, 12\}$, disconnecting the graph into two components of five vertices.

Lipton and Tarjan [9] provide a polynomial time algorithm which determines a vertex separator T in a planar graph such that $|T| \leq \sqrt{6}\sqrt{n}$. In general, the vertex separator problem is NP-hard. A branch and bound based approach to solve the problem is described in [4]. In [6] the Laplacian spectrum of G is used again to get lower bounds on the size of the vertex separator.

It is the purpose of the present paper to explore extensions of the approach from [6] which are feasible for large scale problems. In Sect. 2 we provide a mathematical description of both the bandwidth and the separator problem. We use vertex partitions to approach both problems. These lead to relaxations over orthogonal vectors described in Sect. 3. A remarkable feature lies in the fact that these orthogonal relaxations have a closed form optimal solution, as already observed in [6]. We recall the relevant ideas in Sect. 4. It turns out that the orthogonal relaxation can also be formulated as a semidefinite program. Our goal here is

to avoid solving semidefinite programs to improve the relaxations as recently suggested in [10] or in [3]. The main new idea consists in applying the weight redistribution idea to the eigenvalue bounds derived in [6] to improve the resulting lower bounds, see Sect. 5. This idea was also exploited by Boyd, Diaconis and Zhao [2] in connection with the mixing rate of Markov chains. We close with Sect. 6 indicating some computational considerations.

Notation: The following notation will be used throughout. If G is a graph, we denote its vertex set by $V(G)$ which we assume to be $N = \{1, \ldots, n\}$. Its edge set is denoted by $E(G)$. We consider only undirected graphs. An edge $e \in E(G)$ connecting vertices i and j is denoted by $[i, j]$. For two disjoint subsets S, T of N,

$$\delta(S, T) := \{[u, v] : [u, v] \in E(G), \ u \in S, \ v \in T\}$$

denotes the set of edges of G joining vertices in S and T. Let $S = (S_1, \ldots, S_k)$ be a partition of N, meaning that the sets S_i are pairwise disjoint and their union is N. The set of edges with endpoints in distinct partition blocks $S_1, \ldots, S_{k-1}$ is therefore given by $\cup_{i<j<k}\delta(S_i, S_j)$. In a slight abuse of notation, we denote this edge set by

$$\delta(S) := \cup_{i<j<k}\delta(S_i, S_j).$$

Note that edges joining vertices in S_k are ignored.

The set of permutations of N is denoted by Π_n. If M is a square matrix, $\mathrm{diag}(M)$ denotes the vector containing the diagonal elements m_{ii} of M. Conversely, if m is a vector, then $\mathrm{Diag}(m)$ is the diagonal matrix having the entries of m on the main diagonal.

The columns of the identity matrix I_k are denoted $e_1, \ldots, e_k$. We set $E_{ij} := e_i e_j^T + e_j e_i^T$. The all-ones vector is denoted by e. The all-ones matrix is denoted by $J = ee^T$ or J_k to indicate its size.

2 Separators and Bandwidth Based on Vertex Partitions of Graphs

Let $m = (m_1, \ldots, m_k)^T$ be a vector of nonnegative integers such that $\sum_i m_i = n$. We denote by P_m the set of all partitions of N into k subsets S_i of cardinalities $|S_i| = m_i$,

$$P_m := \{(S_1, \ldots, S_k) : S_i \cap S_j = \emptyset \text{ for } i \neq j, \ \cup_i S_i = N, \ |S_i| = m_i\}. \quad (1)$$

We encode this partitioning using the $n \times k$ partition matrix $X = (x_{ij})$ where column j of X represents the incidence vector of S_j. Hence, $x_{ij} = 1$ if $i \in S_j$ and $x_{ij} = 0$ otherwise. Since each vertex appears in exactly one of the sets, we have

$Xe = e$. Recall that e denotes the all-ones vector of appropriate size. The cardinality constraints become $X^T e = m$. We collect these matrices in the set $\mathcal{M}_m$,

$$\mathcal{M}_m := \{X : X \text{ is } n \times k, \; x_{ij} \in \{0, 1\}, \; Xe = e, \; X^T e = m\} \tag{2}$$

which is clearly in bijective correspondence with P_m. Partition matrices can be characterized as follows.

Proposition 1. *Let* $m = (m_1, \ldots, m_k)^T$ *be a vector of nonnegative integers with* $\sum_i m_i = n$. *The matrix* $X \in \mathcal{M}_m$ *if and only if*

$$Xe = e, \; X^T e = m, \; X^T X = Diag(m), \; X \geq 0. \tag{3}$$

Proof. It is clear that any $X \in \mathcal{M}_m$ satisfies the conditions (3). Conversely, let X satisfy (3). Then $0 \leq x_{ij} \leq 1$, hence $x_{ij}^2 \leq x_{ij}$. We have $\mathrm{tr}(X^T X) = \sum_{ij} x_{ij}^2 = \sum_i m_i = n$ and $e^T Xe = \sum_{ij} x_{ij} = n$. Therefore $n = \sum_{ij} x_{ij}^2 \leq \sum_{ij} x_{ij} = n$, and we get equality throughout, thus $x_{ij}^2 = x_{ij}$, so $x_{ij} \in \{0, 1\}$. $\qquad\square$

Let us consider a partition $S = (S_1, \ldots, S_k) \in P_m$. In case that S_k is a vertex separator which disconnects G into the components $S_1, \ldots, S_{k-1}$, then there are no edges between S_i and S_j for any $i < j < k$. The set of edges with endpoints in distinct partition blocks $S_1, \ldots, S_{k-1}$ is given by

$$\delta(S) = \cup_{i < j < k} \delta(S_i, S_j).$$

Let us consider the minimum of $|\delta(S)|$ over all $S \in P_m$, which we denote by $cut(m)$,

$$cut(m) := \min\{|\delta(S)| : \; S \in P_m\}.$$

In case that for a given partition vector m, the graph G has a vertex separator of size m_k, separating G into $k - 1$ components of sizes $m_1, \ldots, m_{k-1}$, then there exists some $S \in P_m$ such that $\delta(S) = \emptyset$, thus $cut(m) = 0$. On the other hand, if no such separator exists then $cut(m) > 0$.

Example 1. As an example, consider the graph from Table 2. It is clear that $cut((m_1, m_2, 1)) > 0$, as the removal of a single vertex leaves the graph connected. On the other hand $cut((5, 5, 2)) = cut((6, 4, 2)) = 0$. Take $S_3 = \{11, 12\}$ in the first, and $S_3 = \{8, 12\}$ in the second case.

The optimal value $cut(m)$ can also be used to get information on the bandwidth $bw(G)$, see [6]. Let $m = (m_1, m_2, m_3)^T$. If $cut(m) > 0$ then $bw(G) \geq m_3 + 1$. To see this, let ϕ be an ordering of N such that $bw(G) = bw(\phi, G)$. Taking $S_1 = \{\phi^{-1}(1), \ldots, \phi^{-1}(m_1)\}$ and $S_2 = \{\phi^{-1}(m_1 + m_3 + 1), \ldots, \phi^{-1}(n)\}$, we have $|S_i| = m_i$. Since $cut(m) > 0$, there exists an edge joining $u \in S_1$ and $v \in S_2$. But $\phi(u) \in \{1, \ldots, m_1\}$ and $\phi(v) \in \{m_1 + m_3 + 1, \ldots n\}$, hence

$$bw(G) \geq |\phi(u) - \phi(v)| \geq (m_1 + m_3 + 1) - m_1 = m_3 + 1.$$

This idea was exploited by Juvan and Mohar [8] to get bounds on various labeling problems in graphs.

We are therefore interested to recognize whether for some given m we have $cut(m) = 0$ implying that S_k is a vertex separator. Alternatively, in the case $k = 3$ we conclude from $cut(m) > 0$ that the bandwidth is at least $m_3 + 1$.

3   Orthogonal Relaxation

We are now going to express $|\delta(S)|$ for $S \in P_m$ as a quadratic function in 0-1 variables using the partition matrices $X \in \mathscr{M}_m$. We define

$$B = \sum_{0 < i < j < k} E_{ij} = \begin{pmatrix} J_{k-1} - I_{k-1} & 0 \\ 0 & 0 \end{pmatrix}$$

of order k. Let G be a graph with adjacency matrix A. We denote by L the Laplacian matrix associated to A. It is defined as

$$L = \mathrm{Diag}(Ae) - A = \sum_{[i,j] \in E(G)} (e_i - e_j)(e_i - e_j)^T.$$

Proposition 2. *For $S \in P_m$ let $X \in \mathscr{M}_m$ be the associated partition matrix. Then*

$$|\delta(S)| = \frac{1}{2} tr AXBX^T = \frac{1}{2} tr(-L)XBX^T. \tag{4}$$

Proof. Let X be the partition matrix of $S \in P_m$. We use the column representation $X = (x_1, \ldots, x_k)$ of X and get $XBX^T = \sum_{0 < i < j < k} x_i x_j^T + x_j x_i^T$. Hence $Y = XBX^T$ is the adjacency matrix of the complete $(k-1)$-partite graph with partition $S_1, \ldots, S_k$. In particular, $y_{ij} = 1$ if and only if i, j are in distinct sets $S_1, \ldots, S_{k-1}$. Hence $y_{ij}a_{ij} = 1$ exactly if $[i, j] \in \delta(S)$. Therefore $\mathrm{tr} A(XBX^T) = \sum_{ij} a_{ij} y_{ij} = 2|\delta(S)|$, as each edge appears in the lower and the upper triangle of A. Note that outside the main diagonal we have $A = -L$. Since $\mathrm{diag}(XBX^T) = 0$, we can replace A by $-L$. □

Thus we get the following integer optimization problem to find $cut(m)$.

$$2cut(m) = \min\{\mathrm{tr}(-L)XBX^T : X \in \mathscr{M}_m\}.$$

This problem is NP-hard, so we are interested in tractable relaxations. We consider the following set $\mathscr{F}_m$, which contains all partition matrices $X \in \mathscr{M}_m$, see (3),

$$\mathscr{F}_m := \{X : Xe = e, \ X^T e = m, \ X^T X = \mathrm{Diag}(m)\}. \tag{5}$$

Note in particular that any $X \in \mathscr{F}_m$ with $X \geq 0$ will be a 0-1 matrix in view of Proposition 1.

It was pointed out in [13] that the set $\mathscr{F}_m$ can be parametrized using orthogonal unit vectors. We recall the details. Let V be a matrix with $V^T e = 0$, $V^T V = I_{n-1}$. The columns of V therefore form an orthonormal basis to the orthogonal complement of the all-ones vector $e \in \mathbb{R}^n$.

Given m as before, we define $\tilde{m} := (\sqrt{m_1}, \ldots, \sqrt{m_k})^T$ and let W be any matrix satisfying

$$W^T \tilde{m} = 0, \; W^T W = I_{k-1}.$$

In other words, the columns of W represent the orthogonal complement to $\tilde{m}$ in the form of an orthonormal basis. This implies in particular that

$$W W^T = I_k - \frac{1}{n} \tilde{m} \tilde{m}^T. \tag{6}$$

In a similar way we have $V V^T = I - \frac{1}{n} e e^T$. Here is now the parametrization of $\mathscr{F}_m$ in terms of orthogonal vectors. We use $\tilde{M} = \mathrm{Diag}(\tilde{m})$.

Lemma 1 ([13]). *In the notation above we have*

$$\mathscr{F}_m = \{\frac{1}{n} e m^T + V Z W^T \tilde{M} : \; Z^T Z = I_{k-1}\}.$$

We substitute this parametrization for $X \in \mathscr{F}_m$ into the objective function (4) and get the following.

Proposition 3. *For $X \in \mathscr{F}_m$ we have*

$$tr(L X B X^T) = tr(V^T L V) Z (W^T \tilde{M} B \tilde{M} W) Z^T, \tag{7}$$

where $X = \frac{1}{n} e m^T + V Z W^T \tilde{M}$ and $Z^T Z = I_{k-1}$.

Proof. This follows upon substitution and using the fact that $Le = 0$. $\qquad \square$

We summarize all this in the following main result of this section, which provides the lower bound f for $cut(m)$.

Theorem 1. *In the notation above we have $f \leq cut(m)$ where*

$$2f := \min\{tr(-L) X B X^T : X \in \mathscr{F}_m\}$$
$$= \min\{tr(-V^T L V) Z (W^T \tilde{M} B \tilde{M} W) Z^T : Z^T Z = I_{k-1}\}. \tag{8}$$

The key insight behind this theorem lies in the fact that optimizing a quadratic function over the set $\mathscr{F}_m$, which is described by the linear equations $Xe = e$, $X^T e = m$ and by the orthogonality condition $X^T X = \mathrm{Diag}(m)$ can equivalently be done by eliminating the equations, and optimizing over orthonormal vectors collected in Z.

The case $k = 3$ was already investigated in [6], leading to a closed form expression for f in terms of the eigenvalues of the two matrices involved. We will see in the next section that the general case $k > 3$ leads in a similar way to a closed form solution.

4   Eigenvalue Bounds Based on the Hoffman-Wielandt Theorem

The optimal value in (8) has a closed form solution, which can be derived from a theorem of John von Neumann [14], see also Hoffman and Wielandt [7]. We state this result in the following form.

Theorem 2. *Let A and B be symmetric matrices of order n and k with $k \leq n$. Then*

$$\min\{tr(AXBX^T) : \ X^T X = I_k\}$$

$$= \min\{\sum_{i=1}^{k} \lambda_i(B)\lambda_{\phi(i)}(A) : \ \phi : N \mapsto \{1,\ldots,k\} \ injection\}.$$

The minimum is attained for $X = (p_{\phi(1)},\ldots p_{\phi(k)})Q^T$, where $p_{\phi(i)}$ is a normalized eigenvector to $\lambda_{\phi(i)}(A)$ and $Q = (q_1,\ldots,q_k)$ contains the normalized eigenvectors q_i of $\lambda_i(B)$.

The result follows from the original papers [7, 14] by embedding the matrix B into an $n \times n$ matrix by adding zero rows and columns. See also [12] for an elementary proof and some historical notes. We therefore need to study the spectrum of $V^T L V$ and of

$$\tilde{B} := W^T \tilde{M} B \tilde{M} W.$$

We denote the eigenvalues of the Laplacian L by $\lambda_1 = 0 \leq \lambda_2(L) \leq \ldots \leq \lambda_n(L)$ with eigenvectors $v_1 = e, v_2, \ldots, v_n$ where $v_i \perp e$ for $i > 1$.

Lemma 2. *The eigenvalues of $V^T L V$ are $\lambda_2(L),\ldots,\lambda_n(L)$ with eigenvectors $V^T v_i$ for $i \geq 2$.*

Proof. Since $VV^T = I - \frac{1}{n}ee^T$ we have $V^T L V(V^T v_i) = V^T L(I - \frac{1}{n}ee^T)v_i = \lambda_i V^T v_i$. $\square$

To get a feeling for the spectrum of $\tilde{B}$ it is instructive to consider first the following special case, where the cardinalities of S_1 up to S_{k-1} are all equal to one another. Thus let $b := m_1 = \ldots = m_{k-1} > 0$ and $a := m_k = n - (k-1)b > 0$.

Lemma 3. *Let $k \geq 3$. For $b = m_1 = \ldots = m_{k-1} > 0$ and $a = m_k = n - (k-1)$ $b > 0$, the eigenvalues of $\tilde{B}$ are $-b$ (with multiplicity $k - 2$) and $\frac{k-2}{n}ab > 0$.*

Proof. The eigenvalues of $B = \begin{pmatrix} J_{k-1} - I_{k-1} & 0 \\ 0 & 0 \end{pmatrix}$ are $k - 2$ and 0 with multiplicity 1 and -1 with multiplicity $k - 2$. With $a = n - (k - 1)b$ and $m = (b, \ldots, b, a)^T$ we get $\tilde{M} B \tilde{M} = bB$, therefore $\tilde{B} = bW^T BW$. Therefore the eigenvalues of $\tilde{B}$ interlace those of bB. $\tilde{B}$ therefore has $-b$ with multiplicity $k - 3$ and x and y as eigenvalues. We can determine x and y from the equations

$$x + y - (k - 3)b = \operatorname{tr}\tilde{B} = -\frac{b^2}{n}(k - 1)(k - 2),$$

$$x^2 + y^2 + (k-3)b^2 = \operatorname{tr}\tilde{B}^2 = b^2(k-1)(k-2)\left[1 - \frac{2b}{n}(k-2) + \frac{b^2}{n^2}(k-1)(k-2)\right].$$

A somewhat tedious calculation shows that $x = -b, y = \frac{k-2}{n}ab$ solve the equations. $\square$

In the general case, the eigenvalues of $\tilde{B}$ are not available in closed form as before, but it can be shown that $k - 2$ eigenvalues are negative and exactly one is positive.

Lemma 4. *Let* $m = (m_1, \ldots, m_k)^T$ *be given with* $k \geq 3$ *and* $m_i > 0$ *and* $\sum m_i = n$. *Then* $\tilde{B}$ *has* $k - 2$ *negative eigenvalues and one positive eigenvalue.*

Proof. Let us introduce $\tilde{M} B \tilde{M} = \begin{pmatrix} B_0 & 0 \\ 0 & 0 \end{pmatrix}$. By Sylvester's inertia theorem, the eigenvalues of $J_{k-1} - I_{k-1}$ and B_0 have the same sign. Thus B_0 has $k - 2$ negative and one positive eigenvalue. Next we note that

$$\tilde{W} = \begin{pmatrix} \sqrt{m_2} & & & \\ -\sqrt{m_1} & \ddots & & \\ & \ddots & \sqrt{m_k} & \\ & & -\sqrt{m_{k-1}} & \end{pmatrix}$$

spans the orthogonal complement of $\tilde{m}$, hence the QR decomposition of $\tilde{W}$ can be assumed to satisfy $\tilde{W} = WR$. Let us write W in the form $W = \begin{pmatrix} W_0 \\ w_0 \end{pmatrix}$, where w_0 denotes the last row of W. Then $\tilde{B} = W^T (\tilde{M} B \tilde{M}) W = W_0^T B_0 W_0$. The first $k - 1$ rows of $\tilde{W}$ are linearly independent and equal to $W_0 R$, showing that W_0 must be invertible. This shows that $\tilde{B}$ and B_0 also have the same inertia. $\square$

Thus we have the following situation. The spectrum of $-V^T LV$ is given by

$$0 \geq -\lambda_2(L) \geq \ldots \geq -\lambda_n(L).$$

Let us denote the eigenvalues of $\tilde{B}$ by

$$\beta_1 \le \beta_2 \ldots \le \beta_{k-2} < 0 < \beta_{k-1}$$

in view of the previous lemma. This gives the following 'closed form' solution for f from (8) which we collect in the following theorem as the main conclusion from this section.

Theorem 3. *In the notation from above we have*

$$2f = \min\{tr(-L)XBX^T : X \in \mathscr{F}_m\} = -\lambda_2(L)\beta_1 - \ldots - \lambda_{k-1}(L)\beta_{k-2} - \lambda_n(L)\beta_{k-1}.$$

The value is attained for $X = \frac{1}{n}em^T + (v_2, \ldots, v_{k-1}, v_n)Q^T W^T \tilde{M}$ *in view of Theorem 2.*

Proof. The correctness of the bound is a consequence of Theorem 2 in combination with Lemmas 2 and 4. To see attainment, we recall from Proposition 3 that $X = \frac{1}{n}em^T + V^T ZW^T \tilde{M}$. Moreover, again using Lemmas 2 and 4 we have $Z = V(v_2, \ldots, v_{k-1}, v_n)Q^T$ with $Q^T \tilde{B} Q = \text{Diag}(\beta_1, \ldots, \beta_{k-1})$. Now $VV^T(v_2, \ldots, v_{k-1}, v_n)$
$= (v_2, \ldots, v_{k-1}, v_n)$ and the representation for X follows. $\square$

In the special case where $m_1 = \ldots = m_{k-1} = b$ and $m_k = a, m = (b, \ldots, b, a)$, we get a further simplification.

Corollary 1. *Let* $b = m_1 = \ldots = m_{k-1}$ *and* $a = m_k$. *Then*

$$2f = \min_{X \in F_m} tr(-LXBX^T) = b \sum_{i=2}^{k-1} \lambda_i(L) - \frac{k-2}{n}ab\lambda_n(L), \qquad (9)$$

This result is used in [6] with $k = 3$ to get the following lower bound on $bw(G)$,

$$bw(G) \ge n\frac{\lambda_2(L)}{\lambda_n(L)}.$$

Unfortunately, this bound is in general rather weak.

Example 2. We consider again the graph from Table 2. Setting $m = (5, 5, 2)$, we know from before that $cut(m) = 0$. The lower bound from the previous corollary is

$$5\lambda_2(L) - \frac{5}{6}\lambda_{12}(L) = -2.4597.$$

The relevant Laplacian eigenvalues are $\lambda_2(L) = 0.4543$ and $\lambda_{12}(L) = 5.6772$ respectively. Since the lower bound is less than zero, one can not draw any conclusion from it. Finally,

$$n\frac{\lambda_2(L)}{\lambda_n(L)} \approx 0.96,$$

and we only get the trivial lower bound $bw(G) \geq 1$.

Remark 1. In the case $a = 0$, i.e. partitioning into $k - 1$ blocks of equal size b, this result has already been investigated in [13]. The case $k = 3$ with arbitrary m_1 and m_2 is studied in [6]. The present theorem is in fact a straightforward generalization of the techniques from [6] to the case with $k > 3$.

We recall that f is most useful in case that $f > 0$, because this implies $cut(m) \geq f > 0$. In the case $f \leq 0$ we can not draw any conclusions about $cut(m)$ being positive or not. We note that f can be written in the form $f = f_1 + f_2$ where

$$2f_1 := -\lambda_2(L)\beta_1 - \ldots - \lambda_{k-1}(L)\beta_{k-2} > 0 \qquad (10)$$

and

$$2f_2 := -\lambda_n(L)\beta_{k-1} < 0, \qquad (11)$$

hence there is no immediate conclusion about the sign of f.

5 Weight Redistribution

We are now going to take a more general viewpoint. Instead of just looking at the graph G, where all its edges have weight $a_{ij} = 1 \; \forall [i, j] \in E(G)$ we now allow these values to vary, but under the condition that they remain nonnegative and their total sum equals $|E(G)|$. Thus let

$$\mathscr{S} := \{x \in IR^E : x \geq 0, \sum_{[i,j] \in E(G)} x_{[i,j]} = |E(G)|\}$$

denote the fundamental simplex on $E(G)$. We now consider the following family of Laplacian matrices

$$\mathscr{L} := \{L(x): \; L(x) = \sum_{[i,j] \in E} x_{[i,j]}(e_i - e_j)(e_i - e_j)^T : x \in S\}.$$

The lower bound f is now a function of x given as

$$f(x) = -\lambda_2(L(x))\beta_1 - \ldots - \lambda_{k-1}(L(x))\beta_{k-2} - \lambda_n(L(x))\beta_{k-1} = f_1(x) + f_2(x),$$

using (10) and (11) and $L(x) \in \mathscr{L}$.

The idea of weight redistribution has been used for instance by Boyd, Diaconis and Xiao [2] to optimize the mixing rate of Markov chains. Göhring, Helmberg and Reiss [5] investigate geometric properties of the eigenvector to the largest Laplacian eigenvalue in a similar setting.

In our context, we use $f(x)$ in the following way. In case that the initial bound $f = f(e) > 0$, we can draw conclusions about $bw(G)$. Otherwise, if $f(e) \leq 0$, we try to maximize $f(x)$ over $x \in S$. In case that we find $x \in S$ such that $f(x) > 0$ we can still conclude that for $L(x)$ we have $cut(m) > 0$, and therefore this also holds for the original Laplacian $L = L(e)$.

Thus we need to investigate

$$f^* := \max\{f(x) : x \in \mathscr{S}\}.$$

We first note that the ordering of the eigenvalues β_i shows that $f_1(x)$ is a concave function. Similarly, the largest eigenvalue is convex, hence $f_2(x)$ is also concave. We will in fact see that f^* is the optimal value of a linear semidefinite optimization problem. To see this, we need the following theorem of Anstreicher and Wolkowicz [1].

Theorem 4 ([1]). *Let A and B by symmetric matrices of order n. Then*

$$\min\{\langle A, XBX^T\rangle : X^T X = I\} = \max\{trS + trT : B \otimes A - S \otimes I - I \otimes T \succeq 0\}.$$

Remark 2. The matrices S and T are of order n and represent the Lagrange multipliers for the constraints $X^T X - I = 0, XX^T - I = 0$.

In our situation, the order k of B is much smaller than n so we need to adopt the statement of the theorem accordingly. For a proof we refer to Povh and Rendl [11], see also [12].

Theorem 5. *Let A and B by symmetric matrices of order n and k respectively and $k \leq n$. Then*

$$\min\{\langle A, XBX^T\rangle : X^T X = I_k\} = \max\{trS + trT : B \otimes A - S \otimes I - I \otimes T \succeq 0, -T \succeq 0\}.$$

Remark 3. The multiplier S of order k corresponds to the equation $X^T X = I_k$, and T of order n to the (redundant) constraint $I - XX^T \succeq 0$, therefore T also needs to be semidefinite.

Combining all these results we are now in a position to express f^* as the optimal value of a semidefinite program. This also constitutes the main insight of this section.

Theorem 6. *In the notation above we have*

$$\begin{aligned}
2f^* &= \max\{f(x) : x \in \mathscr{S}\} \\
&= \max\{trS + trT : \tilde{B} \otimes (-V^T L(x)V) - S \otimes I - I \otimes T \succeq 0, \\
&\quad -T \succeq 0, x \in \mathscr{S}\}.
\end{aligned}$$

The final problem is a semidefinite program in the matrix variable S of order $k - 1$, the semidefinite matrix variable T of order $n-1$, the nonnegative variables x and and

an additional semidefiniteness constraint on a matrix of order $(n-1) \times (k-1)$. We are not suggesting to actually determine f^* by solving this semidefinite program. In fact, our goal is a bit more modest. We are happy to show that given $f(e) \leq 0$, we can exhibit some $x \in \mathscr{S}$ such that $f(x) > 0$.

Example 3. We consider again the graph from Table 2 and set $m = (6, 5, 1)$. The initial lower bound from Theorem 3 is $f(e) = -0.121$ and does not allow us to draw any conclusions. After weight redistribution we get the improved bound $f^* = 0.566$, hence $cut(m) > 0$. The respective optimizers X_0 and X_1 are

$$X_0 = \begin{pmatrix} 0.9368 & -0.1322 & 0.1954 \\ 1.1261 & -0.0732 & -0.0529 \\ 0.9368 & -0.1322 & 0.1954 \\ 0.8773 & 0.0265 & 0.0962 \\ 0.8827 & -0.0656 & 0.1829 \\ -0.0815 & 0.6538 & 0.4276 \\ 0.0808 & 1.1943 & -0.2750 \\ -0.2651 & 0.5855 & 0.6796 \\ 0.0808 & 1.1943 & -0.2750 \\ -0.1854 & 1.0391 & 0.1463 \\ 0.8620 & 0.3667 & -0.2287 \\ 0.7486 & 0.3431 & -0.0917 \end{pmatrix}, X_1 = \begin{pmatrix} 0.0718 & 0.9827 & -0.0545 \\ -0.1905 & 0.9344 & 0.2561 \\ 0.0718 & 0.9827 & -0.0545 \\ 0.1644 & 0.8078 & 0.0278 \\ 0.1594 & 0.9825 & -0.1419 \\ 1.0591 & 0.2521 & -0.3112 \\ 1.0767 & -0.2715 & 0.1948 \\ 1.0748 & 0.2879 & -0.3627 \\ 1.0771 & -0.2816 & 0.2044 \\ 1.1184 & -0.2944 & 0.1760 \\ 0.1165 & 0.3308 & 0.5527 \\ 0.2005 & 0.2867 & 0.5128 \end{pmatrix}.$$

There is no separator of size 1, separating the graph into two pieces of roughly equal size. Ideally, the optimizers X_0 and X_1 should be close to 0-1 matrices, which they clearly are not. Note however, that rounding X_1 yields the partitions $S_1 = \{6, 7, 8, 9, 10\}$, $S_2 = \{1, 2, 3, 4, 5\}$ with separator $S_3 = \{11, 12\}$, while rounding X_0 does not even produce a partition.

In practical applications the graph G may be quite large. In this case it may be sufficient to work with the function $f(x)$ directly by exploiting the fact that $f(x)$ as an eigenvalue function can be evaluated, provided we are able to compute the largest eigenvalue $\lambda_n(L(x))$ and a few of the smallest eigenvalues of $L(x)$. The associated eigenvectors provide a subgradient to $f(x)$. Thus a more modest strategy, also suitable for very large scale problems, consists in using techniques from subgradient maximization of concave functions to achieve this goal.

6 Some Practical Considerations

We have just seen that the optimal weight redistribution could be determined by solving a single semidefinite optimization problem. For large problems however, say $n \geq 500$, this is far beyond the capabilities of ordinary desktop computers.

On the other hand it is possible to work directly with $f(x)$. First we compute the eigenvalue decomposition $Q^T \tilde{B} Q = \text{Diag}(\beta_1, \ldots, \beta_{k-1})$. These data will not change while maximizing $f(x)$. For $x \in \mathscr{S}$ we need to calculate the eigenvalues $\lambda_2(L(x)), \ldots, \lambda_{k-2}(L(x)), \lambda_n(L(x))$ together with the associated

eigenvectors $v_2, \ldots, v_{k-1}, v_n$. This can be done by iterative methods which only use subroutines calculating matrix times vector products, without actually needing the matrix explicitly. This gives the function value $f(x)$ and also the minimizer $X = \frac{1}{n}em^T + (v_2, \ldots, v_{k-1}, v_n)Q^T W^T \tilde{M} \in \mathscr{F}_m$ of

$$2f(x) = \min\{\text{tr}(-L(x))XBX^T : X \in \mathscr{F}_m\}.$$

The matrix X contains nontrivial information which can be used for rounding it to a partition matrix. Moreover, the eigenvectors can be used to compute a subgradient g of f at x. Hence a search direction based on g and perhaps some previous subgradients in the style of the bundle method will lead to a new trial point x^+ and one iterates. Finally, also the weight redistribution vector x contains interesting information. Which edges in the graph get high weight? What does this say about good partitions?

All these questions will be further investigated in a companion paper, where the theoretical results form the basis of several heuristic methods to get good partition matrices for finding reorderings of small bandwidth or for identifying small vertex separators.

References

1. Anstreicher, K.M., Wolkowicz, H.: On Lagrangian relaxation of quadratic matrix constraints. SIAM J. Matrix Anal. **22**, 41–55 (2000)
2. Boyd, S., Diaconis, P., Xiao, L.: Fastest mixing Markov chain on a graph. SIAM Rev. **46**(4), 667–689 (2004)
3. de Klerk, E., Nagy, M., Sotirov, R.: On semidefinite programming bounds for graph bandwidth. Technical report, Tilburg University, Netherlands (2011)
4. de Sousa, C., Balas, E.: The vertex separator problem: algorithms and computations. Math. Program. (A) **103**, 609–631 (2005)
5. Göhring, F., Helmberg, C., Reiss, S.: Graph realizations associated with minimizing the maximum eigenvalue of the Laplacian. Math. Program. (A) **131**, 95–111 (2012)
6. Helmberg, C., Mohar, B., Poljak, S., Rendl, F.: A spectral approach to bandwidth and separator problems in graphs. Linear Multilinear Algebra **39**, 73–90 (1995)
7. Hoffman, A.J., Wielandt, H.W.: The variation of the spectrum of a normal matrix. Duke Math. J. **20**, 37–39 (1953)
8. Juvan, M., Mohar, B.: Optimal linear labelings and eigenvalues of graphs. Discret. Appl. Math. **36**, 153–168 (1992)
9. Lipton, R.J., Tarjan, R.E.: A separator theorem for planar graphs. SIAM J. Appl. Math. **36**, 177–189 (1979)
10. Povh, J., Rendl, F.: A copositive programming approach to graph partitioning. SIAM J. Optim. **18**(1), 223–241 (2007)
11. Povh, J., Rendl, F.: Approximating non-convex quadratic programs by semidefinite and copositive programming. In: Neralic, L., Boljuncic, V., Soric, K. (eds.) Proceedings of the 11th International Conference on Operational Research, Pula, pp. 35–45. Croation Operations Research Society (2008)

12. Rendl, F.: Semidefinite relaxations for integer programming. In: Jünger, M., Liebling, Th.M., Naddef, D., Nemhauser, G.L., Pulleyblank, W.R., Reinelt, G., Rinaldi, G., Wolsey, L.A. (eds.) 50 Years of Integer Programming 1958–2008, pp. 687–726. Springer, Berlin/Heidelberg (2009)
13. Rendl, F., Wolkowicz, H.: A projection technique for partitioning the nodes of a graph. Ann. Oper. Res. **58**, 155–179 (1995)
14. von Neumann, J.: Some matrix inequalities and metrization of matrix space, pp. 205–219. Reprinted in: John von Neumann: Collected Works, vol. 4. MacMillan (1962/1937)

Exploiting Symmetries in Polyhedral Computations

Achill Schürmann

Abstract In this note we give a short overview on symmetry exploiting techniques in three different branches of polyhedral computations: The representation conversion problem, integer linear programming and lattice point counting. We describe some of the future challenges and sketch some directions of potential developments.

1 Introduction

Symmetric polyhedra such as the Platonic and Archimedean solids have not only fascinated mathematicians since time immemorial. They occur frequently in diverse contexts of art and science. Less known to a general audience, but of great importance to modern mathematics and its applications, are higher dimensional analogues of these familiar objects. One *standard description* is as a set of solutions to a system of linear inequalities

$$P = \{x \in \mathbb{R}^n : Ax \le b\},$$

where A is a real $m \times n$ matrix and $b \in \mathbb{R}^m$. A prominent example is the *n-cube* obtained by $2n$ inequalities $\pm x_i \le 1$. It has 2^n vertices (*extreme points*) with coordinates ± 1 and its *group of symmetries* is the *hyperoctahedral group* of order $2^n n!$.

A. Schürmann (✉)

Institute of Mathematics, University of Rostock, 18051 Rostock, Germany
e-mail: achill.schuermann@uni-rostock.de

Linear models, and therefore polyhedra, are used in a wide range of mathematical problems and in applications such as transportation logistics, machine scheduling, time tabling, air traffic flow management and portfolio planning. They are central objects in Mathematical Optimization (Mathematical Programming) and are for instance heavily used in Combinatorial Optimization. Frequently studied symmetric polyhedra have names like "Travelling Salesman", "Assignment", "Matching" and "Cut". For these and further examples we refer to [60] and the numerous references therein. Over the years a rich combinatorial and geometric theory of polyhedra has been developed (see [33, 69]). Symmetry itself is clearly a central topic in mathematics, and through the spread of computer algebra systems like [GAP] and [MAGMA], sophisticated tools from Computational Group Theory are widely used today (see [37]). Nevertheless, although many polyhedral problems are modeled with a high degree of symmetry, standard computational techniques for their solution do not take advantage of them. Even worse, often the used methods are known to work notoriously poorly on symmetric problems.

In this short survey we describe three main areas of polyhedral computations, in which the rich geometric structure of symmetric polyhedra can potentially be used for improved algorithms:

  I: Polyhedral representation conversion using symmetry
 II: Symmetric integer linear programming
III: Counting lattice points and exact volumes of symmetric polyhedra

There are multiple strong dependencies among the three topics and each one has its theoretical and algorithmic challenges as well as important applications. Before we take a closeup view on the three topics we give a brief introduction to the different types of polyhedral symmetries and how these can be determined and worked with.

What Are Polyhedral Symmetries?

The symmetries of a polyhedron can be of a purely combinatorial nature or they can also have a geometric manifestation as *affine symmetries*, that is, as affine maps of $\mathbb{R}^n$ preserving the polyhedron. Among these symmetries are the "more visually accessible" *isometries* which are composed of translations, rotations and reflections. All of the symmetries of the n-cube for example are part of its isometry group. There exists a representation as a *linear group* in $GL_n(\mathbb{R})$ and as a *finite orthogonal group* of isometries. However, if we perturb the defining inequalities a bit, all of these affine symmetries may be lost, while the new polyhedron is still *combinatorially equivalent* to a cube, sharing all of its *combinatorial symmetries*. These are defined as automorphisms of the *polyhedral face lattice* which encodes the combinatorial structure of a polyhedron. We refer to our survey [10] for further reading on these

different types of polyhedral symmetry groups. The study of combinatorial lattices and their automorphisms is itself an active research area (see [53]). The same is true for the study of possible isometry groups, respectively of *finite orthogonal groups* in $O_n(\mathbb{R})$. Their classification becomes in a way impractical for $n \geq 5$ (see [MO37136]), despite the classification of finite simple groups (see [18]). Even less is known about symmetry groups of polyhedra (see [57]). Here, an "implication phenomenon" occurs, which has not much been studied so far. For instance, if a 4-gon has an element of order 4 among its affine symmetries, the 4-gon has to be the affine image of a square (2-cube), with an affine symmetry group of order 8. These kind of implications clearly can potentially be exploited algorithmically, for example when detecting polyhedral symmetries.

It is important to note that the same abstract group can have different affine representations. We think that a key ingredient for future algorithmic improvements will be the use of geometric information coming with the affine representations of polyhedral symmetry groups. By a basic result in representation theory there is an invariant affine subspace $\mathcal{I}$ coming with each affine symmetry group. The polyhedron splits nicely into an invariant part $P \cap \mathcal{I}$ and symmetric slices orthogonal to it. These lie in fibers (pre-images) of the orthogonal projection onto $\mathcal{I}$. In a way, all of the symmetry is within these fibers.

Given a polyhedron with a group of symmetries, we say two vertices (or inequalities) are *equivalent*, if there exists a group element that maps one to the other. The set of vertices (and the set of *facets*/defining inequalities) splits into a number of *orbits* (disjoint sets of equivalent elements). For example, the n-cube has only one orbit of vertices and one orbit of facets. The same is true for all Platonic polyhedra and their higher dimensional analogues. In contrast, the Archimedean polyhedra like the soccer ball (truncated icosahedron) have more than one orbit of facets, but only one orbit of vertices. In all of these examples, their combinatorial symmetry group is equal to the group of affine symmetries. Its invariant affine subspace is a single point, the barycenter of the vertices.

In general, for a polyhedron P with a group of affine symmetries, the vertices of the polyhedron split into orbits $O_1, \ldots, O_l$ and the invariant part $P \cap \mathcal{I}$ is equal to the convex hull $\mathrm{conv}\{b_1, \ldots, b_l\}$, with $b_i = (\sum_{x \in O_i} x)/|O_i|$ being the barycenter of orbit O_i. This is due to two facts: The barycenter map, taking a point to the barycenter of its orbit, is an affine map. And second, the affine image of a convex hull of given points is equal to the convex hull of their affine images.

Thus working with the lower dimensional polyhedron $P \cap \mathcal{I}$ and its vertices gives us access to vertices of P. Orbits of integral points in P have barycenters at specific locations in $P \cap \mathcal{I}$. For instance, if the group acts transitive on the coordinates of $\mathbb{R}^n$ then orbits have barycenters at integral multiples of $\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$. For more general coordinate permutations the barycenters form a scaled copy of a standard lattice (see [36]).

How to Determine and Work with Symmetries?

If the symmetries of the polyhedron are not known, the first difficulty is their determination and how to represent them. In general we like to work with as many symmetries as possible. However, the combinatorial symmetries can usually not be found without having full knowledge about the vertex-facet incidences of the polyhedron (see [39]). In contrast, the group of affine symmetries can be determined from the vertices or defining inequalities alone, by finding the automorphism group of an edge colored graph. If P is given as the convex hull of its vertices $x_1, \ldots, x_k$, for instance, then the affine symmetry group can be obtained from the automorphism group of the complete graph with k vertices and edge labels $x_i^t Q^{-1} x_j$, where $Q = \sum_{i=1}^k x_i x_i^t$. For details and a proof we refer to [11]. For further methods to compute polyhedral symmetry groups we refer to [10]. Automorphism groups of graphs can be computed with software like [bliss] or [nauty]. Given a polyhedral description, the affine symmetries can conveniently be obtained directly with our software [SymPol], which by now can also be used through [polymake]. For instance, given a polyhedron with its description contained in input-file, simply call:

```
sympol --automorphisms-only input-file
```

If the symmetry group of a polyhedron (or parts of it) are given as a *permutation group*, we can use sophisticated tools from *Computational Group Theory*. Each element of the group is then viewed as a *permutation* of the index set $\{1, \ldots, m\}$ of the input, for instance of m defining inequalities. In practice, it is necessary to work with a small set of *group generators* if the group is large, and there are advanced heuristics to obtain such sets. Each face (and in particular each vertex) of a polyhedron is determined by a number of inequalities that are satisfied with equality; it can therefore be represented by a subset of $\{1, \ldots, m\}$. Given generators of a large permutation group and two subsets that represent faces, a typical computational bottleneck is to decide if both are in the same orbit. The fundamental data structures used for this in practice are *bases and strong generating sets* (BSGS, see [37, 64]). Based on them, backtrack searches can be used to perform essential tasks, such as deciding on (non-)equivalence, obtaining stabilizers or fusing and splitting orbits. An elaborate version is the *partition backtrack* introduced by Leon [47]. These backtracking methods work quite well in practice, although from a complexity point of view the mentioned problems are thought to be difficult (see [49]). Although computer algebra systems like [GAP] and [MAGMA] provide functions to work with permutation groups, for performance reasons it is often desirable to use problem specific code (see for example [41]). Nevertheless, all of these approaches, including [GAP] and [MAGMA], rely on efficient implementations of some partition backtrack. Therefore we have created a flexible C++-implementation [PermLib] of Leon's partition backtrack (see [55]) that can serve as a basis for the development of algorithms which combine tools from Computational Group Theory and Polyhedral Combinatorics. By now, [PermLib] has successfully been integrated into current versions of [SymPol, polymake] and [SCIP].

I. Representation Conversion

By a fundamental theorem in polyhedral combinatorics, the Farkas-Minkowski-Weyl theorem, every polyhedron has a second representation as the convex hull of finitely many *vertices* (extreme points) and, in the unbounded case, some *rays* (see [69]). Converting representations from inequalities to vertices (and rays), or vice versa, is a frequent task known as *representation conversion problem* (or *convex hull problem*). The importance of these conversions is due to the fact that some problems, like the maximization of a nonlinear convex function, are easy to solve in one presentation, but not in the other. Often, vertices represent objects that one would like to classify. These objects can be quite diverse, for instance perfect quadratic forms (see [25]) or the elementary flux modes in biochemical reaction systems (see [63]). Representation conversions are also often used to analyze polyhedra in Combinatorial Optimization (see [60]). So far there exists no efficient algorithm for finding all the vertices of a polyhedron. In fact, the existence of such an algorithm appears to be unlikely, as it is NP-complete for polyhedra that are possibly unbounded (see [42]). Nevertheless, several algorithms and implementations are widely used in practice (see for example [cdd] and [lrs] which are also available through [polymake]).

Quite often one is only interested in one representative for each orbit of vertices (or inequalities) in a representation conversion. For example, when maximizing a nonlinear convex function on a polyhedron, or when vertices and inequalities in one orbit correspond to equivalent objects of some sort. *Representation conversion up to symmetries* has been considered in different contexts, and depending on the problem, different techniques have been successful. The most successful approaches currently known are the *Incidence Decomposition Method* and the *Adjacency Decomposition Method* (see for instance [15, 20, 21, 25–27]). Both methods decompose the problem into a number of lower dimensional subproblems. They can be used recursively and can be parallelized (see [16, 22]). Loosely speaking, the Incidence Decomposition Method fixes an orbit of the input, whereas the Adjacency Decomposition Method fixes an orbit of the output and then lists all "neighboring" orbits. For details we refer to our survey [11]. We note that it is a priori not clear which method works best. We think best results can be achieved by a combination of different algorithms. All methods known so far do not use geometric insights and still rely on subproblem conversions that do not exploit available symmetry.

Our software [SymPol] and the experimental [GAP] package [Polyhedral] provide implementations of decomposition methods. These preliminary tools have already successfully been used in our own work [3, 24–27], but also by others: For instance, Kumar [44] obtains a classification of elliptic fibrations that was previously impossible. Jacques Martinet writes in [52] about the result in [25]: "It seems plainly impossible to classify 8-dimensional perfect lattices." [SymPol] can also be used to verify cumbersome calculations in proofs, like the edge-graph diameter analysis of the recently discovered, celebrated counterexamples to the Hirsch conjecture (see [56, 58]). For example, with the call

```
symbol --idm-adm-level 0 1 --adjacencies input-file
```

where `input-file` contains the 48 vertices of the 5-dimensional *Santos prismatoid* (see Table 1 in [58]), `SymPol` returns a text file with a description of the adjacency graph of facets up to symmetry. Using a visualization tool like [`Graphviz`], the produced textfile, say `adjacencies.dot`, can then easily be turned into an image with a command like

```
neato -Tpng -o adjacencies.png adjacencies.dot
```

From the obtained image (see figure) it is easily verified that the shortest path from facets 1–12 is of length 6, which is the key calculation in the proof of [58].
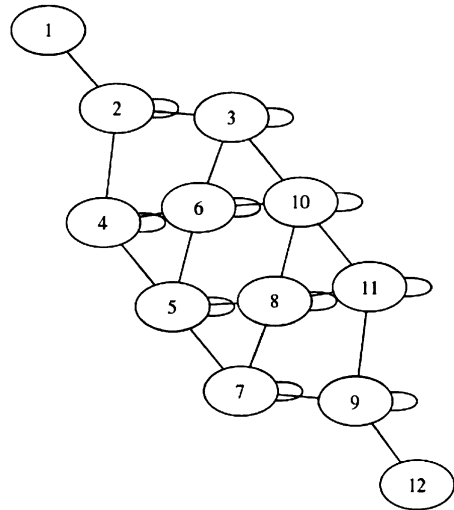
Let us make a remark on the increasing importance of mathematical software in general: As sophisticated computational tools become an increasingly important basis for high-level mathematical research, their creation also becomes an increasingly important service to the mathematical community. More and more mathematicians use computers in the creative process and to verify standard parts of difficult larger proofs (see e.g. [17, 34, 35]). Timothy Gowers [32] even guesses that at the end of the twenty-first century, computers will be better than humans in proving theorems. Although we would not go as far as Gowers, we are convinced that in the future, parts of proofs will routinely be performed by computers. With a symbiosis of human and computer reasoning we will see substantial advances in mathematical problems. In this way reliable mathematical software becomes an increasingly important part of mathematics itself.

Challenges

One of the most challenging polyhedral conversion problems arises in conjunction with lattice sphere packing problem, a classical problem in the Geometry of Numbers. Since its solution up to dimension 8 almost 80 years ago, it is still open in dimension $n \geq 9$, with the exception of dimension 24 (see [17, 52]). One way to approach this problem is via finding the vertices of a locally polyhedral object known as *Ryshkov polyhedron* (see [61] for details). The currently open 9-dimensional case leads to a challenging representation conversion problem of a 45-dimensional Ryshkov polyhedron. Main difficulties here come from faces that carry the symmetries of the exceptional Weyl group E_8. We think that this problem is a particular nice test case, as all finite rational matrix groups appear as stabilizers of faces in the Ryshkov polyhedron. So in a way, this challenging representation conversion problem gives a *universal test case* for any future algorithmic advances.

II. Integer Linear Programming

Linear programming is the task of maximizing (or minimizing) a linear function on a polyhedron given by linear inequalities. It serves as a fundamental basis for theory and computations in Mathematical Optimization (see [66]).

In *Integer linear programming* vectors to be optimized are restricted to integers (number of goods, etc.) or even to 0/1 entries (encoding a simple yes-no-decision). Integer linear programming is widely used in practical applications. In fact, "the vast majority of applications found in operations research and industrial engineering involve the use of discrete variables in problem formulation" (from a book review of [68]). In many of these problems the involved polyhedra have symmetries (see [60]). From a complexity point of view, integer programming is NP-complete (see [40]), whereas linear programming can be solved in polynomial time. For fixed dimension, polynomial time algorithms are known for integer linear programming (see [46]).

A linear programming problem max $c^t x$ with $x \in P$ is *invariant with respect to a linear symmetry group* $\Gamma \leq \mathrm{GL}_n(\mathbb{R})$, if the polyhedron P and the utility vector $c \in \mathbb{R}^n$ are preserved by it, that is, if $\Gamma P = P$ and $\Gamma c = c$. Any solution to the linear program, its orbit and the barycenter of its orbit lie in the same hyperplane orthogonal to the utility vector and therefore have the same utility value. Due to the convexity of polyhedra, the barycenter is also a feasible solution. As it lies in the *invariant linear subspace* $\mathcal{I}$ of Γ, the linear programming problem always has a solution attained within $\mathcal{I}$. Thus it is possible to solve the lower-dimensional linear program max $c^t x$ with $x \in P \cap \mathcal{I}$. Such symmetry reductions are often referred to as "dimension reduction" or "variable reduction". The symmetries of an integer linear program are more restrictive, as also $\mathbb{Z}^n$ has to be left invariant by the group Γ.

Exploiting symmetries in integer programming is much more difficult than in linear programming. In fact, symmetries are rather problematic, as standard

methods like *branch-and-bound* or *branch-and-cut* (see [59]) have to solve many equivalent sub-problems in such cases. In contrast to linear programming, it is not possible to simply consider the intersection with the invariant affine subspace, as integral solutions can lie outside. Nevertheless, in recent years it has been shown that it is possible to exploit symmetries in integer programming; see for example [14,30,38,48,50,54]. These specific methods fall into two main classes: They either modify the standard branching approach, using isomorphism tests or isomorphism free generation to avoid solving equivalent subproblems; or they use techniques to cut down the original symmetric problem to a less symmetric one, which contains at least one element of each orbit of solutions. For further reading we refer to the excellent survey [51].

As many real world applications can be modeled as (mixed) integer programming problems, a variety of professional software packages are available. Two of the leading ones, [`CPLEX`] and [`Gurobi`], have by now included some techniques to avoid or use symmetry. Unfortunately it is publicly not known what exactly is done.

None of the known methods uses the rich geometric properties of the involved symmetric polyhedra. Using the fact that solutions are "near" the linear invariant subspace, it is possible to do better. For the special case of a one dimensional invariant subspace, with the full symmetric group S_n acting transitively on the coordinates in $\mathbb{R}^n$, this is shown in [9]. We have highly promising results with a generalization to arbitrary symmetries in [36]: In particular for direct products of symmetric groups, we not only beat state-of-the-art professional solvers, but even solve a challenging, previously unsolved benchmark problem from [`MIPLIB`] (instance `toll-like`).

The main ingredient is the observation that any feasible integer linear programming problem with a non-trivial affine symmetry group contains certain *core-sets* of integral vectors that can be used as a kind of test-set, that is, if non of the points from the core-set is contained in the feasible region, no integral point is. Assume $\Gamma \leq \mathrm{GL}_n(\mathbb{R})$ is a linear representation of a given symmetry group and $\mathcal{I}$ denotes its invariant linear subspace containing the utility vector c. Then we say an integral point z in a *fiber* (pre-image) of the orthogonal projection onto $\mathcal{I}$ is in the *core-set of the fiber* if the convex hull of its orbit Γz does not contain any integral points aside those of the orbit itself. Then, by definition, representatives of each orbit Γz in the core-set can be used as a *test-set for feasibility of a fiber*.

Using the Flatness Theorem (from the Geometry of Numbers), it can be shown that core-sets are finite for irreducible groups. For direct products of symmetric groups acting on some of the coordinates, the test-set containing only representatives of orbits even reduces to a single point. Besides that not much is known so far about core-sets. Nevertheless, we think that they will serve as a powerful tool in the design of new algorithms for symmetric integer linear programs.

Challenges

Challenging examples of symmetric integer linear programs can be found in benchmark libraries like [`MIPLIB`]. These problems come from diverse contexts and have not been chosen to be particular symmetric. Nevertheless many symmetries can be found and exploiting them algorithmically, beating state-of-the-art commercial solvers, remains a challenging test case for future advances.

Some particular symmetric integer linear programming problems coming from difficult combinatorial problems in mathematics have been collected (and worked on) by Francois Margot at [`symlp`] (see for instance also [50] and [14]). As these problems have been intensively worked on, improving on the currently known results is certainly a hard problem. So this gives a very good benchmark for future improvements as well.

III. Lattice Point Counting and Exact Volumes

Often it is desirable to know how many integral solutions there are to a system of linear inequalities. Such problems occur frequently in Combinatorics (see [65]) but also in disciplines such as representation theory (Kostka and Littlewood-Richardson coefficients, see [8] and [43]), in statistics (contingency tables, see [23]), in voting theory (see [67] and [31]), and even in compiler optimization (see [`Graphite`]). We refer to [19] for an overview. Counting lattice points is moreover intimately related to integer linear programming (see [45]).

By a breakthrough result of Barvinok [4] in the 1990s, counting lattice points inside a rational polyhedron can be done in polynomial time for fixed dimension. His ideas are based on evaluating "short rational generating functions" and on constructing unimodular triangulations. His algorithm has been implemented in [`LattE`] and [`barvinok`]. The same applies to a slightly more general setting, in which one considers a one-parameter family of dilations λP, with P a rational polyhedron and λ an integer. By a theory initiated by Ehrhart in the 1960s (see [7, 29]), it is possible to obtain the number of integral points in the dilate λP by a quasi-polynomial in λ, with its degree equal to the dimension of P. A quasi-polynomial p is determined by a finite number of polynomials p_i, $i = 0, \ldots, k$, via the setting $p(\lambda) = p_i(\lambda)$ for all λ congruent to i mod k. In case of P being integral, the *Ehrhart quasi-polynomial* simply is a polynomial in λ. In general, the quasi-polynomial can also be computed in polynomial time by Barvinok methods (see [5]). Often, the main interest is only in the leading coefficient of the Ehrhart quasi-polynomial, which is the volume of P. Computing the volume itself is already a #P-hard problem (see [12, 28]).

Despite the fact that many counting problems have plenty of symmetries, they have not been exploited systematically so far. In other words, exploiting symmetry in lattice point counting, or more generally in Ehrhart theory, is a vastly open

subject. For volume computations the situation seems a bit better. For very special volume computations symmetry can be exploited (see [26]). However, there is still a huge potential for improved methods. Many of the difficulties originate from the fact that the "Barvinok methods" used to solve them rely on unimodular triangulations of polyhedral cones that usually do not inherit the symmetry of the polyhedron. New roads will have to be taken here.

In [62] it is shown that it is possible to exploit symmetries by using a decomposition into symmetric slices, together with a *weighted Ehrhart theory*. The theoretical background and first implementations for such a theory have just recently been developed (see [2, 13]). The polyhedral decomposition used in [62] is rather special: There is a linear invariant part and symmetric slices orthogonal to it, which are cross-products of regular simplices (simplotopes). A generalization to other decompositions is easily obtained, whenever there is a decomposition into an invariant part and slices orthogonal to it for which the Ehrhart quasi-polynomial is known. Note that the decomposition can easily be obtained in a automated way, as the invariant part is the intersection of the given polyhedron with the affine space fixed by its symmetry group.

For exploiting symmetry in corresponding volume computations, the integration of polynomials over a polyhedron is used. Using Brion-Lawrence-Varchenko theory, this can efficiently be done by integrating sums of powers of linear forms (see [1]). The new decomposition approach of [62] also allows to obtain exact volumes that have not been computable before. This is demonstrated on three well studied examples from Social Choice theory, which give the exact likelihood of certain election outcomes with four candidates that were previously known for three candidate elections only (see [31]).

Challenges

In Social Choice theory we face a large amount of challenging problems related to probability calculations of voting situations with four and more candidates. The only known results in the context of the "polyhedral model" (IAC hypothesis) appear to be those in [62], which are obtained by exploiting polyhedral symmetry as described above.

A challenging benchmark volume computation that several researchers previously have looked at is the volume of the Birkhoff polytope B_n (also known as *perfect matching polytope* of the complete bipartite graph $K_{n,n}$). The current known record is the volume of B_{10} due to Beck and Pixton [6], using a complex-analytic way to compute the Ehrhart polynomial. The computation of the volume of B_{11} would certainly be quite a computational achievement.

Conclusions

We expect that symmetry exploiting techniques for polyhedral computations can be vastly improved by using geometric properties that come with affine symmetries of polyhedra. Concentrating on improvements in polyhedral computations with affine symmetries is practically no restriction: If a polyhedron is given, either by linear inequalities or vertices and rays, the affine symmetries of the (potentially larger) combinatorial symmetry group are practically the only ones we can compute.

For polyhedral representation conversions we see potential in enhancing decomposition methods through the use of geometric information like fundamental domains, classical invariant theory and symmetric polyhedral decompositions. For integer linear programming we expect that a new class of algorithms based on the concept of core points will help to exploit symmetry on difficult symmetric integer linear programming problems. For exact volume computations and counting of lattice points, there is still a lot of potential for new ideas using symmetry. Overall, we think symmetry should be exploitable whenever it is available. For this goal to be reached there seem still quite some efforts necessary though.

References

1. Baldoni, V., Berline, N., De Loera, J.A., öppe, M.K, Vergne, M.: How to integrate a polynomial over a simplex. Math. Comp. **80**(273), 297–325 (2011)
2. Baldoni, V., Berline, N., De Loera, J.A., öppe, M.K., Vergne, M.: Computation of the highest coefficients of weighted Ehrhart quasi-polynomials of rational polyhedra. Found. Comput. Math. (2011, to appear). Preprint at http://arxiv.org/abs/1011.1602arxiv:1011.1602
3. Ballinger, B., Blekherman, G., Cohn, H., Giansiracusa, N., Kelly, E., Schürmann, A.: Experimental study of energy-minimizing point configurations on spheres. Experiment. Math. **18**, 257–283 (2009)
4. Barvinok, A.: A polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed. Math. Oper. Res. **19**, 769–779 (1994)
5. Barvinok, A.: Integer points in polyhedra. Eur. Math. Soc., Zrich, pp. viii+191 (2008). http://dx.doi.org/10.4171/052
6. Beck, M., Pixton, D.: The Ehrhart polynomial of the Birkhoff polytope. Discrete Comput. Geom. **30**, 623–637 (2003). See also the update in http://arxiv.org/abs/math/0305332arxiv:math/0305332
7. Beck, M., Robins, S., Computing the continuous discretely. Springer, New York (2007)
8. Berenstein, A.D., Zelevinsky, A.V.: Tensor product multiplicities, canonical bases and totally positive varieties. Invent. Math. **143**, 77–128 (2001)
9. Bödi, R., Herr, K., Joswig, M.: Algorithms for highly symmetric linear and integer programs. Math. Program. A (2012, to appear). Published online, http://dx.doi.org/10.1007/s10107-011-0487-6
10. Bremner, D., Dutour Sikirić, M., Pasechnik, D.V., Rehn, T., Schürmann, A.: Computing symmetry groups of polyhedra. Preprint at http://arxiv.org/abs/1210.0206arxiv:1210.0206
11. Bremner, D., Dutour Sikirić, M., Schürmann, A.: Polyhedral representation conversion up to symmetries. In: CRM Proceedings & Lecture Notes, Montreal, vol. 48, pp. 45–71. AMS, (2009)
12. Brightwell, G., Winkler, P.: Counting linear extensions. Order **8**, 225–242 (1991)

13. Bruns, W., Söger, C.: The computation of generalized Ehrhart series in Normaliz. Preprint at http://arxiv.org/abs/1211.5178arxiv:1211.5178
14. Bulutoglu, D.A., Margot, F.: Classification of orthogonal arrays by integer programming. J. Stat. Plan. Inference **138**, 654–666 (2008)
15. Christof, T., Reinelt, G.: Combinatorial optimization and small polytopes. Top **4**, 1–64 (1996)
16. Christof, T., Reinelt, G.: Decomposition and parallelization techniques for enumerating the facets of combinatorial polytopes. Internat. J. Comput. Geom. Appl. **11**, 423–437 (2001)
17. Cohn, H., Kumar, A.: Optimality and uniqueness of the Leech lattice among lattices. Ann. Math. **170**, 1003–1050 (2009)
18. Conway, J.H., Curtis, R.T., Norton, S.P., Parker, R.A., Wilson, R.A.: Atlas of finite groups. Oxford University Press, Oxford (1985)
19. De Loera, J.A.: The many aspects of counting lattice points in polytopes. Math. Semesterber. **52**, 175–195 (2005)
20. Deza, A., Fukuda, K., Mizutani, T., Vo, C.: On the face lattice of the metric polytope. In: Discrete and Computational Geometry. LNSC, vol. 2866, pp. 118–128. Springer, New York (2003)
21. Deza, A., Fukuda, K., Pasechnik, D., Sato, M.: On the skeleton of the metric polytope. In: Discrete and Computational Geometry. LNSC, vol. 2098, pp. 125–136. Springer, New York (2001)
22. Deza, A., Indik, G.: A counterexample to the dominating set conjecture. Optim. Lett. **1–2**, 163–169 (2007)
23. Diaconis, P., Sturmfels, B.: Algebraic algorithms for sampling from conditional distributions. Ann. Stat. **26**, 363–397 (1998)
24. Dutour Sikirić, M., Schürmann, A., Ellis, G.: On the integral homology of $PSL_4(\mathbb{Z})$ and other arithmetic groups. J. Number Theory **131**, 2368–2375 (2011)
25. Dutour Sikirić, M., Schürmann, A., Vallentin, F.: Classification of eight dimensional perfect forms. Electron. Res. Announc. Am. Math. Soc. **13**, 21–32 (2007)
26. Dutour Sikirić, M., Schürmann, A., Vallentin, F.: Complexity and algorithms for computing Voronoi cells of lattices. Math. Comp. **78**, 1713–1731 (2009)
27. Dutour Sikirić, M., Schürmann, A., Vallentin, F.: The contact polytope of the Leech lattice. Discret. Comput. Geom. **44**, 904–911 (2010)
28. Dyer, M.E., Frieze, A.M.: On the complexity of computing the volume of a polyhedron. SIAM J. Comput. **17**, 967–974 (1988)
29. Ehrhart, E.: Sur un problème de géométrie diophantienne linéaire. I. Polyèdres et réseaux. J. Reine Angew. Math. **226**, 1–29 (1967)
30. Friedman, E.J.: Fundamental domains for integer programs with symmetries. In: Combinatorial Optimization and Applications. LNSC, vol. 4616, pp. 146–153. Springer, New York (2007)
31. Gehrlein, W.V., Lepelley, D.: Voting paradoxes and group coherence Springer, Berlin/London (2011)
32. Gowers, W.T.: Rough structure and classification. In: Visions in Mathematics – Towards 2000, pp. 79–117. Birkhäuser, Basel (2000)
33. Grünbaum, B.: In: Kaibel, V., Klee, V., Ziegler, G.M. (eds.) Convex Polytopes, 2nd edn. Springer, London/New York (2003)
34. Hales, T.C.: A proof of the Kepler conjecture. Ann. Math. **162**, 1065–1185 (2005)
35. Hales, T.C.: The strong dodecahedral conjecture and fejes toth's conjecture on sphere packings with kissing number twelve. Preprint at http://arxiv.org/abs/1110.0402arxiv:1110.0402
36. Herr, K., Rehn, T., Schürmann, A.: Exploiting symmetry in integer convex optimization using core points. Preprint at http://arxiv.org/abs/1202.0435arxiv:1202.0435
37. Holt, D.F., Eick, B., O'Brien, E.A.: Handbook of computational group theory. Chapman & Hall/CRC, Boca Raton (2005)
38. Kaibel, V., Pfetsch, M.E.: Packing and partitioning orbitopes. Math. Program. A **114**, 1–36 (2008)
39. Kaibel, V., Schwartz, A.: On the complexity of polytope isomorphism problems. Graphs Combin. **19**, 215–230 (2003)

40. Karp, R.M.: Reducibility among combinatorial problems. In: Complexity of Computer Computations, pp. 85–103. Plenum, New York (1972)
41. Kaski, P., Östergård, P.R.J.: Classification algorithms for codes and designs. Springer, Berlin (2006)
42. Khachiyan, L., Boros, E., Borys, K., Elbassioni, K., Gurvich, V.: Generating all vertices of a polyhedron is hard. Discrete Comput. Geom. **39**, 174–190 (2008)
43. Knutson, A., Tao, T.: The honeycomb model of $GL_n(\mathbb{C})$ tensor products I. Proof of the saturation conjecture. J. Amer. Math. Soc. **12**, 1055–1090 (1999)
44. Kumar, A.: Elliptic fibrations on a generic Jacobian Kummer surface. J. Algebraic. Geom. (2012, to appear). Preprint at http://arxiv.org/abs/1105.1715arxiv:1105.1715
45. Lasserre, J.B.: Linear and integer programming vs linear integration and counting. Springer, New York/London (2009)
46. Lenstra Jr., H.W.: Integer programming with a fixed number of variables. Math. Oper. Res. **8**, 538–548 (1983)
47. Leon, J.S.: Permutation group algorithms based on partitions. I. Theory and algorithms. J. Symbolic Comput. **12**, 533–583 (1991)
48. Linderoth, J., Margot, F., Thain, G.: Improving bounds on the football pool problem via symmetry reduction and high-throughput computing. Inf. J. Comput. **21**, 445–457 (2009)
49. Luks, E.M.: Permutation groups and polynomial-time computation. In: Groups and Computation, pp. 139–175. AMS, Providence (1993)
50. Margot, F.: Exploiting orbits in symmetric ILP. Math. Program. B **98**, 3–21 (2003)
51. Margot, F.: Symmetry in integer linear programming. In: 50 Years of Integer Programming, pp. 647–686. Springer, Berlin/London (2010)
52. Martinet, J.: Perfect Lattices in Euclidean Spaces. Springer, Berlin/New York (2003)
53. McMullen, P., Schulte, E.: Abstract Regular Polytopes. Cambridge University Press, Cambridge/New York (2002)
54. Ostrowski, J., Linderoth, J., Rossi, F., Smriglio, S.: Orbital branching. Math. Program. A **126**, 147–178 (2011)
55. Rehn, T.: Fundamental permutation group algorithms for symmetry computation. Diploma Thesis in Computer Science, OvG University Magdeburg (2010)
56. Rehn, T., Schürmann, A.: C++ tools for exploiting polyhedral symmetries. In: Mathematical Software – ICMS 2010. LNCS, vol. 6327, pp. 295–298. Springer, Berlin (2010)
57. Robertson, S.A.: Polytopes and Symmetry. Cambridge University Press, Cambridge/New York (1984)
58. Santos, F.: A counterexample to the Hirsch conjecture. Ann. Math. **176**, 383–412 (2012)
59. Schrijver, A.: Theory of linear and integer programming. Wiley, Chichester/New York (1986)
60. Schrijver, A.: Combinatorial optimization. Polyhedra and efficiency, vols. A, B, C. Springer, Berlin (2003)
61. Schürmann, A.: Computational geometry of positive definite quadratic forms. AMS, Providence (2009)
62. Schürmann, A.: Exploiting polyhedral symmetries in social choice. Social Choice and Welfare (2012, to appear). Published online, http://dx.doi.org/10.1007/s00355-012-0667-1, preprint at http://arxiv.org/abs/1109.1545arxiv:1109.1545
63. Schuster, S., Hlgetag, C.: On elementary flux modes in biochemical reaction systems at steady state. J. Biol. Syst. **2**, 165–182 (1994)
64. Seress, Á.: Permutation group algorithms. Cambridge University Press, New York (2003)
65. Stanley, R.P.: Enumerative Combinatorics, vol. 1. Cambridge University Press, Cambridge/New York (1997)
66. Todd, M.J.: The many facets of linear programming. Math. Program. B **91**, 417–436 (2002)
67. Wilson, M.C., Pritchard, G.: Probability calculations under the IAC hypothesis. Math. Soc. Sci. **54**, 244–256 (2007)
68. Wolsey, L.A.: Integer Programming. Wiley, New York (1998)
69. Ziegler, G.M.: Lectures on Polytopes. Springer, New York (1997)

Software and Webpages

`barvinok`. *Counting lattice points*, http://www.kotnet.org/~skimo/barvinok/
`bliss`. *Graph and labeling automorphisms*, http://www.tcs.hut.fi/Software/bliss/
`cdd`. *Double Description*, http://www.ifor.math.ethz.ch/~fukuda/cdd_home/
`CPLEX`. *Mathematical programming technology*, http://www.ilog.com/products/cplex/
`GAP`. *Groups, Algorithms, Programming*, http://www.gap-system.org/
`Graphite`. *High-level memory optimizations*, http://gcc.gnu.org/wiki/Graphite
`Graphviz`. *Graph Visualization Software*, http://www.graphviz.org/
`Gurobi`. *High-end libraries for math programming*, http://www.gurobi.com/
`LattE`. *Lattice point count, volumes and integrals*, http://www.math.ucdavis.edu/latte/
`lrs`. *Lexicographic reverse search*, http://cgm.cs.mcgill.ca/~avis/C/lrs.html
`MAGMA`. *Computational Algebra System*, http://magma.maths.usyd.edu.au/magma/
`MO37136`. *Classification of finite groups of isometries*, http://mathoverflow.net/questions/37136/
    classification-of-finite-groups-of-isometries
`MIPLIB`. *Mixed Integer Problem Library*, http://miplib.zib.de/
`nauty`. *Graph and labeling automorphisms*, http://cs.anu.edu.au/~bdm/nauty/
`PermLib`. *Permutation groups library*, http://www.geometrie.uni-rostock.de/software/
`Polyhedral`. *A polyhedral GAP package*, http://www.liga.ens.fr/~dutour/Polyhedral/
`polymake`. *A framework for analyzing convex polytopes*, http://polymake.org/
`SCIP`. *Solving Constraint Integer Programs*, http://scip.zib.de/
`symlp`. *Symmetric IP problems*, http://wpweb2.tepper.cmu.edu/fmargot/lpsym.html
`SymPol`. *Symmetric polyhedra toolkit*, http://www.geometrie.uni-rostock.de/software/

Conditions for Correct Sensor Network Localization Using SDP Relaxation

Davood Shamsi, Nicole Taheri, Zhisu Zhu, and Yinyu Ye

Abstract A Semidefinite Programming (SDP) relaxation is an effective computational method to solve a Sensor Network Localization problem, which attempts to determine the locations of a group of sensors given the distances between some of them. In this paper, we analyze and determine new sufficient conditions and formulations that guarantee that the SDP relaxation is exact, i.e., gives the correct solution. These conditions can be useful for designing sensor networks and managing connectivities in practice. Our main contribution is threefold: First, we present the first non-asymptotic bound on the connectivity (or radio) range requirement of randomly distributed sensors in order to ensure the network is uniquely localizable with high probability. Determining this range is a key component in the design of sensor networks, and we provide a result that leads to a correct localization of each sensor, for any number of sensors. Second, we introduce a new class of graphs that can always be correctly localized by an SDP relaxation. Specifically, we show that adding a simple objective function to the SDP relaxation

D. Shamsi
Department of Management Science and Engineering, Huang Engineering Center,
Stanford University, 475 Via Ortega, Stanford, CA 94305, USA
e-mail: davood@stanford.edu

N. Taheri (✉)
IBM Research, Smarter Cities Technology Centre, Dublin, Ireland
e-mail: nicole.taheri@ie.ibm.com

Z. Zhu
IBM Corporation, 1001 East Hillsdale Boulevard, Suite 400, Foster City, CA 94404-1642, USA
e-mail: zhuzhisu@gmail.com

Y. Ye
Department of Management Science and Engineering, Huang Engineering Center 308,
School of Engineering, Stanford University, 475 Via Ortega, Stanford, CA 94305, USA. Research
supported in part by AFOSR Grant FA9550-12-1-0396.
e-mail: yinyu-ye@stanford.edu

model will ensure that the solution is correct when applied to a triangulation graph. Since triangulation graphs are very sparse, this is informationally efficient, requiring an almost minimal amount of distance information. Finally, we analyze a number of objective functions for the SDP relaxation to solve the localization problem for a general graph.

Key words Sensor network localization • Graph realization • Semidefinite programming

Subject Classifications: 90C22, 90C46, 90C90, 90B18

1   Introduction

Graph Realization is a commonly studied topic which attempts to map the nodes in a graph $G(V, E)$ to point locations in Euclidean space based on the non-negative weights of the edges in E; that is, the weight of each edge corresponds to the Euclidean distance between the incident points. There are a number of applications of the graph realization problem [9, 13, 16, 21, 25]. In this paper, we focus on the application to Sensor Network Localization (SNL).

A sensor network consists of a collection of *sensors* whose locations are unknown, and *anchors* whose locations are known. A common property of a sensor network is that each sensor detects others within a given connectivity (or radio) range and determines the distance from itself to these nearby sensors. Given this set of known distances, the goal is to determine the exact location of each sensor. The problem becomes a graph realization problem by forming the weighted undirected graph $G(V, E)$, where the node set V represents the sensors and each non-negative weighted edge in E represents a known distance between two sensors. A number of approaches have been proposed to solve this problem [17, 27–29].

The SNL problem has received a lot of attention recently because of the formulation of its relaxation as a Semidefinite Program (SDP) [2, 10, 11, 24, 30, 32, 33]. This formulation can find the exact locations of the sensors, given that the graph possesses certain properties.

Definition 1. A *correct localization*, or a *correct solution*, provides a set of points that is exactly equal to the sensor locations. That is, the solution not only solves a given formulation, but it provides the correct sensor locations in the desired dimension.

In this paper, we present a number of additional sufficient conditions that guarantee unique localizability (and hence a correct localization) of the SDP relaxation of the SNL problem. These conditions can be useful for designing sensor networks and managing connectivities in practice.

1.1 Background

We are given a graph $G(V, E \cup \bar{E})$ in a fixed dimension d, where the nodes, or points, of V are partitioned into two sets: the set $V_a = \{a_1, \ldots, a_m\}$ of m anchors (where $m \geq d + 1$) whose locations are known and the set $V_x = \{x_1, \ldots, x_n\}$ of n sensors, whose locations are unknown. The edge set also consists of two distinct sets: the set $E = \{(i, j) : i, j \in V_x\}$ of edges between sensors, and the set $\bar{E} = \{(k, j) : k \in V_a, j \in V_x\}$ of edges between an anchor and a sensor. Moreover, for each $(i, j) \in E$ (or $(k, j) \in \bar{E}$) the Euclidean distance between sensor i and sensor j (respectively, anchor k and sensor j) is known as d_{ij} (respectively $\bar{d}_{kj}$). The problem of finding the locations of the sensors can be formulated as finding points $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ that satisfy a set of quadratic equations:

$$\| x_i - x_j \|^2 = d_{ij}^2, \ \forall \ (i, j) \in E$$

$$\| a_k - x_j \|^2 = \bar{d}_{kj}^2, \ \forall \ (k, j) \in \bar{E}. \qquad \text{(SNL-norm)}$$

From this, a number of fundamental questions naturally arise: Is there a localization or realization of x_j's that solves this system? If there is a solution, is it unique? And is there a way to certify that a solution is unique? Is the network instance partially localizable, i.e., is the localization solution for a subset of the sensors unique? These questions were extensively studied in the graph rigidity and discrete geometry communities from a more combinatorial and theoretical prospective (see [15, 19, 20, 22] and references therein). However, the question of whether there is an efficient algorithm to numerically answer some of these questions remains open.

The SDP relaxation model (SNL-SDP) and corresponding method aim to answer these questions computationally (see [4, 6, 11, 18, 30, 34]). Let $e_i \in \mathbb{R}^n$ represent the ith column of the identity matrix in $\mathbb{R}^{n \times n}$, and define the symmetric matrices $A_{ij} := (\mathbf{0}; e_i - e_j)(\mathbf{0}; e_i - e_j)^T$ and $\bar{A}_{kj} := (a_k; -e_j)(a_k; -e_j)^T$, where $\mathbf{0} \in \mathbb{R}^d$ is the vector of all zeros. The SDP relaxation can be represented as:

$$
\begin{aligned}
\text{maximize } & 0 \\
\text{subject to } & Z_{(1:d,1:d)} = I_d \\
& A_{ij} \bullet Z = d_{ij}^2, \quad \forall (i, j) \in E \\
& \bar{A}_{kj} \bullet Z = \bar{d}_{kj}^2, \quad \forall (k, j) \in \bar{E} \\
& Z \succeq 0.
\end{aligned}
\qquad \text{(SNL-SDP)}
$$

Here, $Z_{(1:d,1:d)}$ represents the upper-left d-dimensional principle submatrix of Z, the matrix dot-product refers to the sum of element-wise products $A \bullet B = \sum_{ij} A_{ij} B_{ij}$, and $Z \succeq 0$ means that the symmetric variable matrix Z is positive semidefinite. Note that problem (SNL-SDP) is a convex semidefinite program and can be approximately solved in polynomial time by interior-point algorithms.

One can see that the solution matrix $Z \in \mathbb{R}^{(d+n) \times (d+n)}$ of (SNL-SDP) is a matrix that can be decomposed into submatrices,

$$Z = \begin{bmatrix} I & X \\ X^T & Y \end{bmatrix}.$$

The constraint $Z \succeq 0$ holds if and only if $Y \succeq X^T X$. If $Y = X^T X$, then the above formulation finds a matrix Z such that the columns of its submatrix $X = \begin{bmatrix} x_1 \ x_2 \ \cdots \ x_n \end{bmatrix}$ satisfy all quadratic equations in (SNL-norm).

Definition 2. A sensor network is *uniquely localizable* if there is a unique $X \in \mathbb{R}^{d \times n}$ whose columns satisfy (SNL-norm), and there is no $\bar{X} \in \mathbb{R}^{h \times n}$, for $h > d$, whose columns satisfy (SNL-norm) and $\bar{X} \neq (X; \mathbf{0})$. In other words, there is no nontrivial extension of $X \in \mathbb{R}^{d \times n}$ into higher dimension $h > d$ that also satisfies (SNL-norm) [30].

Note that the notion of unique localizability is stronger than the notion of global rigidity. A sensor network is globally rigid only if there is a unique $X \in \mathbb{R}^{d \times n}$ that satisfies (SNL-norm), but it may also have a solution in a higher dimension space, that is a nontrivial extension of $X \in \mathbb{R}^{d \times n}$, which satisfies (SNL-norm) [1, 30].

The following theorem was proved in [30]:

Theorem 1. *An SNL problem instance is uniquely localizable if and only if the maximum rank solution of its SDP relaxation* (SNL-SDP) *has rank d, or equivalently, every solution matrix Z of* (SNL-SDP) *satisfies $Y = X^T X$. Moreover, such a max-rank solution matrix can be computed approximately in polynomial time.*

The theorem asserts that the certification of a uniquely localizable network instance can be achieved by solving a convex optimization problem; the proof is constructive and produces a unique realization or localization solution for the original problem (SNL-norm).

The dual of the SDP relaxation (SNL-SDP)

$$\text{minimize} \ \ I_d \bullet V + \sum_{(i,j) \in E} y_{ij} d_{ij}^2 + \sum_{(k,j) \in \bar{E}} w_{kj} \bar{d}_{kj}^2$$

$$\text{subject to} \ \ \begin{pmatrix} V & 0 \\ 0 & 0 \end{pmatrix} + \sum_{(i,j) \in E} y_{ij} A_{ij} + \sum_{(k,j) \in \bar{E}} w_{kj} \bar{A}_{kj} \succeq 0 \qquad \text{(SDP-dual)}$$

is also useful, in that the solution to the dual tells us key properties about the primal. We define the dual slack matrix $U \in \mathbb{R}^{(d+n) \times (d+n)}$ as

$$U = \begin{pmatrix} V & 0 \\ 0 & 0 \end{pmatrix} + \sum_{(i,j) \in E} y_{ij} A_{ij} + \sum_{(k,j) \in \bar{E}} w_{kj} \bar{A}_{kj},$$

for $V \in \mathbb{R}^{d \times d}$. The dual slack matrix U is optimal if and only if it is feasible and meets the complementarity condition, $ZU = 0$. If complementarity holds, then rank(Z)+ rank(U) $\leq (d + n)$, and since rank(Z) $\geq d$, this means that rank(U) $\leq n$. Thus, if an optimal dual slack matrix has rank n, then every solution to (SNL-SDP) has rank d [30]. In fact, we have a stronger notion on localizability:

Definition 3. A sensor network is *strongly localizable* if there exists an optimal dual slack matrix with rank n.

Again, such a max-rank dual solution matrix can be computed approximately in polynomial time using SDP interior-point algorithms.

1.2 Our Contributions

In this paper, we present new conditions that guarantee unique localizability of the SDP relaxation of the problem, i.e., conditions that ensure the SDP will give the correct solution so that the sensor network can be localized in polynomial time. We also enhance the relaxation such that the new SDP relaxation will produce a correct solution in dimension d to satisfy (SNL-norm), even when the standard SDP relaxation (SNL-SDP) may not. More precisely, our result is twofold:

1. A very popular graph in the context of sensor network localization is the unit-disk graph, where any two sensor points (or a sensor point and an anchor point) are connected if and only if their Euclidean distance is less than a given connectivity radius r. It has been observed that when the radius (or radio range) increases, more sensors in the network can be correctly localized. There is an asymptotic analysis to explain this phenomenon when the sensor points are uniformly distributed in a unit-square [3]. In this paper, we present a *non-asymptotic* bound on the radius requirement of the points in order to ensure the network is uniquely localizable with high probability. Specifically, we decompose the area into sub-regions, which allows us to analyze whether the locations of points in each sub-region can be determined, as opposed to analyzing each point individually. We then determine the probability that the locations of all sensors can be determined, given a specified concentration of the sensors in a given area. This may have practical impact by providing guidance on communication power ranges that ensure the network is uniquely localizable.

2. The basic SDP localization model (SNL-SDP) is an SDP feasibility problem. An open question has been to determine whether adding a certain objective function to the basic model improves localizability of the problem; that is, if the SDP feasible region contains high-rank solutions, is the SDP optimal solution guaranteed to be unique and low-rank with a certain objective? We give an affirmative answer for a generic class of graphs, by identifying an objective function that will always result in a correct localization for this class of graphs. Our result may also have an influence on Compressed Sensing, which uses an

objective function to produce the sparsest solution. Based on this idea, we present numerical results by comparing several SDP objective functions to illustrate their effectiveness.

Moreover, although our theoretical analyses are based on exact distance measurements, similar extensions of our model (established in earlier SDP work) would be applicable to noisy distance data.

1.3 Paper Organization

The organization of this paper is as follows. First, Sect. 2 derives a lower bound for the connectivity radius in a sensor network that guarantees unique localizability with high probability. In Sect. 3, we prove that given a triangulation (i.e., a planar, chordal and convex) graph, if the sum of the distances between nodes that do *not* have an edge between them is maximized, then the graph will be strongly localizable. We use this idea, and test a number of heuristic objective functions on a large number of random sensor networks to determine how well each works in practice. Our results for these heuristics are presented in Sect. 4.

2 Bounding the Connectivity Radius

In this section, we consider the unit-disk graph model [7, 8, 14] for sensor networks, where the Euclidean distance between any two sensor points (or a sensor point and an anchor point) is known (i.e., the two points are connected) if and only if the distance between them is less than a given connectivity radius r. Assuming that the sensor points are randomly distributed in a region, we then establish a lower bound on radius r that guarantees unique localizability, with high probability, of the sensor network formed based on radius r. We do this by establishing a lower bound on radius r to ensure that the unit-disk graph is a $(d + 1)$-lateration graph, which is a sufficient condition for unique localizability.

Definition 4. For some $d, n \geq 1$, the graph $G(V, E)$ is a *(d+1)-lateration graph* if there exists a permutation of the points, $\{\pi(1), \pi(2), \ldots, \pi(n)\}$, such that the edges of the sub-graph $\pi(1), \ldots, \pi(d + 1)$ form a complete graph, and each successive point $\pi(j)$ for $j \geq d+2$ is connected to $d+1$ points in the set $\{\pi(1), \ldots, \pi(j-1)\}$. This permutation of the points, π, is called a $(d + 1)$-*lateration ordering*.

It is shown in [35] that if a sensor network graph contains a spanning $(d + 1)$-lateration graph and the points are in general position, then it is uniquely localizable. Zhu et al. [35] provide a rigorous proof, which is based on the intuitive concept that given $d + 1$ points in general position forming a complete graph, the locations of the points can be always be uniquely determined, and the location of any point connected to $d + 1$ points with known locations can also be determined.

Define $r(p)$ to be the smallest connectivity radius of the randomly distributed sensor points that ensures the network is uniquely localizable with probability at least p. To find a lower bound on $r(p)$, we can find a connectivity radius for which the unit-disk graph $G(V, E)$ will contain a spanning $(d + 1)$-lateration graph with at least probability p.

We approach the problem by considering a unit hypercube $\mathcal{H} = [0, 1]^d$, which contains all the sensor points. We then split the region $\mathcal{H}$ into a grid of M equal sub-hypercubes in dimension d, say $h_1, h_2, \ldots, h_M \subset \mathcal{H}$, where each sub-hypercube h_i will have a volume of $1/M$, and the length of each of its edges will be $\ell := 1/\sqrt[d]{M}$. Without loss of generality, we can assume $M = b^d$, where b is a positive integer and $b \geq 3$. Similarly, if the region considered is a hyper-rectangle in dimension d, we can assume $M = b_1 \cdot b_2 \cdots b_d$, where $b_i \geq 3$ for $i = 1, \ldots, d$ are positive integers. This partition will allow us to analyze the probability that the locations of sensors in a given region can be determined, as opposed to analyzing each individual point.

2.1 Ensuring a Clique in the Graph

Since a $(d + 1)$-lateration ordering on the points must begin with a $(d + 1)$-clique, we first find a lower bound on the radius r to ensure there exists at least one clique of $d + 1$ points in the graph.

Proposition 1. *Let $\mathcal{H}$ contain n points, and $r \geq \ell\sqrt{d} = \frac{\sqrt{d}}{\sqrt[d]{M}}$ and $M \leq \frac{n-1}{d}$ (or equivalently $r \geq \frac{\sqrt[d]{d}\sqrt{d}}{\sqrt[d]{n-1}}$). Then, there exists at least one clique of $d + 1$ points in the unit-disk graph $G(V, E)$.*

Proof. Note that $\frac{\sqrt{d}}{\sqrt[d]{M}}$ is the length of the diagonal of each sub-hypercube h_i. Thus, if r is lower-bounded by the given value, then every point in a sub-hypercube will be connected to any other point in the same sub-hypercube. Furthermore, since there are at most $\frac{n-1}{d}$ sub-hypercubes, by the pigeon-hole principle, at least one of them contains at least $(d + 1)$ points and they must form a clique of $d + 1$ points in the unit-disk graph with given radius r.

In what follows, we fix $n = d \cdot M + 1 = d \cdot b^d + 1$. We will initialize the spanning $(d+1)$-lateration graph construction by choosing r according to this lower bound, and let the points in the $(d + 1)$-clique be the first $d + 1$ points in the lateration ordering. Since these points are randomly distributed, they must be in general position with probability one. Thus, we may assume that these $d + 1$ points are anchors for the sensor network. This assumption is without loss of generality, because our bound on the radius r established in the following sections will be much greater than the bound specified in Proposition 1, simply because we need to ensure that not only does there exist a clique of $d + 1$ points, but also all sensor points in $\mathcal{H}$ form a spanning $(d + 1)$-lateration graph with a high probability.

2.2 Binomial Distribution Model

One way to let the sensor points be randomly distributed throughout the area of $\mathcal{H}$ is to let the points be binomially distributed throughout each sub-hypercube of $\mathcal{H}$. More specifically, the number of points, Y_i, placed in each sub-hypercube h_i, for $i = 1, \ldots, M$, will be independently and binomially generated according to $Y_i \sim B\left(n, \frac{1}{M}\right)$ with $n = d \cdot M + 1 = d \cdot b^d + 1$. Once Y_i is generated, we let these Y_i sensor points be arbitrarily placed in general position within sub-hypercube h_i.

Using this binomial distribution model, let $S_n = \sum_{i=1}^{M} Y_i$ denote the total number of points in the hypercube $\mathcal{H}$. Since the Y_i values are independently and identically distributed and all sub-hypercubes are equally sized, the total number of points will be more or less evenly distributed in the entire hypercube $\mathcal{H}$. Furthermore, by properties of the binomial distribution,

$$\mathrm{E}\left[S_n\right] = M \cdot \mathrm{E}\left[Y_1\right] = M\left(\frac{n}{M}\right) = n$$

$$\mathrm{Var}\left(S_n\right) = M \cdot \mathrm{Var}\left(Y_1\right) = M \cdot \left[\frac{n}{M}\left(1 - \frac{1}{M}\right)\right] = n\left(1 - \frac{1}{M}\right).$$

Thus, $\frac{S_n}{n} \to 1$ almost surely and the assumption of binomially distributed sensor points throughout each sub-hypercube is statistically equivalent to assuming a uniform distribution of n points throughout the whole region $\mathcal{H}$ when M is sufficiently large.

2.3 Connectivity Bound

We now form further conditions on the connectivity radius r to ensure that the unit-disk graph G contains a spanning $(d+1)$-lateration graph. We have assumed that the points are binomially distributed in each sub-hypercube, parametrized as $B\left(n, \frac{1}{M}\right)$. First, r must satisfy Proposition 1, since it ensures a $(d+1)$-clique in G. These points in the clique will represent the first $d+1$ points in the lateration ordering π of a spanning $(d+1)$-lateration graph (Definition 4).

We construct an improved bound on the probability of localizability through an ordering of the hypercubes, $h_i \in \mathcal{H}$, and hence an ordering on the points. For simplicity, we prove the following lemmas for the case of $d = 2$, and we refer to the sub-hypercubes as sub-squares. We also refer to $(d+1)$-lateration when $d = 2$ as *trilateration*. However, we note that the same analysis can be applied to hypercubes in higher dimensions, and our bound $r \geq 2\ell\sqrt{2}$ in Lemmas 1–3 is analogous to the bound $r \geq 2\ell\sqrt{d}$ in dimension d.

Lemma 1. *Assume that each sub-square in $\mathcal{H} \in \mathbb{R}^2$ has at least one point, and $r \geq 2\ell\sqrt{2}$. If the points of three sub-squares in the same row in three consecutive columns are in the trilateration ordering, then the points in all sub-squares in those*

Fig. 1 Conditions as described in Lemma 2 to ensure trilateration
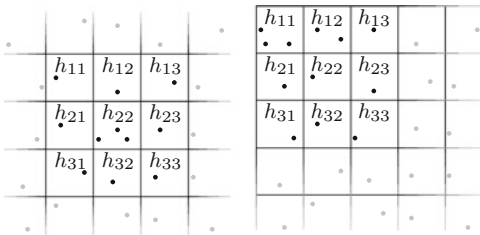
three columns are also in the ordering. Similarly, if the points in three consecutive sub-squares in the same column are in the trilateration ordering, then the points in all sub-squares in those three rows are also in the ordering.

Proof. First, note that the lower bound $r \geq 2\ell\sqrt{2}$ ensures that all points in a given sub-square are connected to all points in a neighboring sub-square, which share either an edge or a point within the given sub-square.

For ease of explanation, let (i, j) represent the sub-square in the i-th row and j-th column, and consider the case that all points in the first three sub-squares in the first row of the grid are already in the trilateration. Since all points in sub-square $(2, 2)$ are within the connectivity range of all three points in the first row, these points are in the trilateration. Then, all points in sub-square $(2, 1)$ (or $(2, 3)$) are within the connectivity range of at least three points in sub-square $(1, 2), (2, 2), (1, 1)$ (or $(1, 2), (2, 2), (1, 3)$), these points are also in the trilateration. Therefore, all points in the first three sub-squares of the second row are in the trilateration.

Similarly, all points in the third row of the grid in the first three columns are also in the trilateration. This pattern continues, until all points in the first three columns of the grid are in the trilateration.

A generalization of this shows that if there are three sub-squares in the same row and in consecutive columns with points in the trilateration, and each sub-square has at least one point, then all points in the corresponding columns are also in the trilateration.

An analogous result holds for three sub-squares in the same column and in consecutive rows.

Lemma 1 states that if there are three consecutive sub-squares in a row with points in the trilateration, then the trilateration ordering extends to all squares in the corresponding columns. This concept is used below in Lemma 2, which analyzes the cases depicted in Fig. 1.

Lemma 2. *Assume there is at least one point in each sub-square and $r \geq 2\ell\sqrt{2}$. Then the associated unit-disk graph contains a spanning trilateration graph if either:*

(a) There is a 3-clique in a non-corner sub-square
(b) There is a 3-clique in a corner sub-square and one of its neighbor squares has at least two points

Proof. Again, note that $r \geq 2\ell\sqrt{2}$ ensures all points in a given sub-square are connected to all points in neighboring sub-squares. We show that if either of the conditions of Lemma 2 are satisfied, then there exists a trilateration ordering on the points in the graph.

(a) Consider the example in the left grid of Fig. 1, where there is a 3-clique in the non-corner sub-square h_{22}. Let the points in this clique be the initial 3 points in the trilateration ordering. All points in the sub-squares $\{h_{11}, h_{12}, h_{13},$ $h_{21}, h_{23}, h_{31}, h_{32}, h_{33}\}$ are connected to this clique; let the points in these squares be next in the trilateration ordering.

   By Lemma 1, all points in the sub-squares in rows 1–3 are in the trilateration ordering. Since there are at least three columns in $\mathcal{H}$, the same argument applies for the columns, and inductively, there is a trilateration ordering on the points that spreads throughout the entire hyperspace $\mathcal{H}$.

(b) Now consider the right grid of Fig. 1, where there is a 3-clique in the corner sub-square h_{11}, and there are at least two points in a neighboring sub-square. Let the points in h_{11} be the first three points in the trilateration ordering. All points in the three sub-squares h_{12}, h_{21}, h_{22} are connected to the points in the clique and hence in the trilateration ordering. Next, let the points in sub-squares $h_{31}, h_{32}, h_{33}, h_{23}, h_{13}$ be the succeeding points in the ordering. With a similar argument as before using Lemma 1, we can construct a trilateration on the points in the graph, and all points are in the trilateration.

Therefore, if the conditions of Lemma 2 hold, the associated unit-disk graph contains a spanning trilateration graph.

The above lemma provides sufficient, but not necessary, conditions on a network for trilateration to exist, which implies unique localizability. Moreover, these are strict conditions for a sensor network, since the distribution of sensors in a network may not always ensure that there is one sensor in each sub-square. Thus, we extend these conditions to a more general case, and allow for the possibility of empty sub-squares. Clearly, too many empty sub-squares will result in a graph that is not uniquely localizable; also, if empty sub-squares exist, there must be restricting conditions to ensure the graph is not too sparse to ensure localizability. Thus, we establish additional properties of the graph that ensure a trilateration but allow for empty sub-squares.

Definition 5. Two neighboring sub-squares are called *adjacent* neighbors if they do not share any edges, but share a point; neighbors that share an edge are called *simple* neighbors. A sub-square is called *densely surrounded* if all its simple neighbors have at least two points and one of its simple neighbors has at least three points.

Lemma 3. *Assume every empty sub-square is densely surrounded and $r \geq 2\ell\sqrt{2}$. Then the associated unit-disk graph contains a spanning trilateration graph if there is a 3-clique in a non-corner sub-square.*
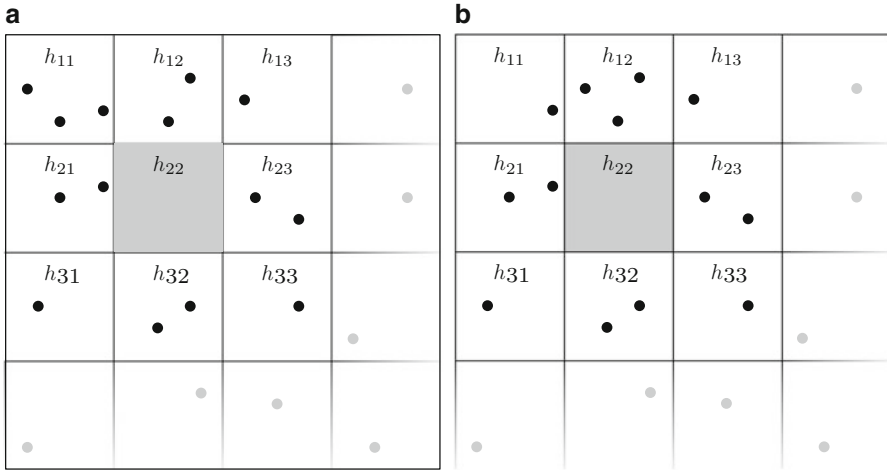
Fig. 2 Example grids for Lemma 3. (**a**) Example of grid that does not satisfy conditions of Lemma 3. (**b**) Example of grid that satisfies conditions of Lemma 3

Proof. Consider the grids in Fig. 2, which shows an example and counter example of the conditions in Lemma 3.

(a) First, consider the left grid of Fig. 2, which does not satisfy the condition of Lemma 3 because there is a clique in a corner sub-square. Notice that if a trilateration ordering starts with the points in sub-square h_{11}, it can continue to the points in sub-squares h_{12} and h_{21}, but will not spread to points in other sub-squares. That is, there is no trilateration ordering that starts with the points in h_{11} and extends to the points in the sub-squares $\{h_{13}, h_{23}, h_{31}, h_{32}, h_{33}\}$, because none of these sub-squares neighbor a subset of sub-squares, in the corresponding trilateration ordering, that contain at least three points combined. Thus, empty sub-squares must be densely surrounded to ensure a trilateration ordering on the points exists.

(b) Now, consider the right grid of Fig. 2, with a non-corner 3-clique, and a densely surrounded empty sub-square. This example shows the worst-case example of the condition in Lemma 3. The shaded sub-square h_{22} is empty and densely surrounded, and the 3-clique is along the edge of the area $\mathcal{H}$. We prove that a trilateration ordering exists on the points in this sample grid with a densely surrounded sub-square. This proves Lemma 3 holds in the worst-case; the proof that Lemma 3 holds in every case is a generalized extension of this.

Define the permutation on the points in Fig. 2 via the ordering on their sub-squares:

$$\Pi := \{h_{12}, h_{11}, h_{13}, h_{21}, h_{23}, h_{32}, h_{31}, h_{33}\}. \tag{1}$$

This permutation is a trilateration ordering on the points in the sample grid. Note that Π is a trilateration ordering only because the sub-squares $\{h_{21}, h_{23}\}$ contain at least three points combined; that is, a permutation containing the points in the sub-squares $\{h_{32}, h_{31}, h_{33}\}$ can only be a trilateration if $\{h_{21}, h_{23}\}$ together contain at least three points.

By Lemmas 1 and 2, if there are no other empty sub-squares in $\mathcal{H}$, then there is trilateration ordering on all the points in $\mathcal{H}$. However, if there *are* other densely surrounded empty sub-squares in $\mathcal{H}$, then by a similar construction as (1), there is still a trilateration ordering on all points in $\mathcal{H}$.

Therefore, if the condition of Lemma 3 holds, the associated graph contains a spanning trilateration graph, and hence is uniquely localizable in dimension 2.

We now use the fact that a sensor network containing a spanning trilateration is uniquely localizable [35] to establish a lower bound on the probability that the unit disk sensor network with radius $r \geq 2\ell\sqrt{2}$ is localizable. Define the two events:

$$C := \{\text{There are only 3-cliques in corner sub-squares}\},$$

$$\hat{C} := \{\text{There is a 3-clique in a non-corner sub-square}\}.$$

Then, the probability that a graph with such randomly distributed points is uniquely localizable will be

$$\text{P}\{\text{uniquely localizable}\} = \text{P}\left\{\text{uniquely localizable}|\hat{C}\right\}\text{P}\left\{\hat{C}\right\} +$$

$$\text{P}\{\text{uniquely localizable}|C\}\text{P}\{C\}$$

$$\geq \text{P}\left\{\text{uniquely localizable}|\hat{C}\right\}\text{P}\left\{\hat{C}\right\}.$$

Given that the total number of sub-squares is $M = b^2$ (for some integer $b \geq 3$), we introduce a parameter $\alpha := \sqrt{\frac{n}{M}}$ (or $\alpha := \sqrt[d]{\frac{n}{M}}$ for general d) such that $\ell = \alpha/\sqrt{n}$ is the edge-length of each sub-square and we can use the same connectivity radius lower bound as before, now in terms of α, $r(\alpha) \geq (2\alpha\sqrt{2})/\sqrt{n}$. The distribution of point number in each sub-square is binomial $B\left(n, \frac{1}{M}\right)$, and there are a total of $(M-4)$ non-corner sub-squares in $\mathcal{H}$. Thus, the probability that there is a 3-clique in a non-corner sub-square is

$$\text{P}\left\{\hat{C}\right\} = 1 - \left(\sum_{i=0}^{2}\binom{n}{i}\left(\frac{1}{M}\right)^i\left(1 - \frac{1}{M}\right)^{n-i}\right)^{M-4}.$$

Let k be the number of empty sub-squares. By Lemma 2,

$$\text{P}\left\{\text{uniquely localizable}|k = 0, \hat{C}\right\} = 1,$$

and if $p_0 = (1 - \frac{1}{M})^n$ is the probability that one specific sub-square is empty, we have

$$P\{k = i\} = \binom{M}{i} p_0^i (1 - p_0)^{M-i}.$$

Moreover, for any $i < M - 4$, we have

$$P\{\hat{C}|k = i\} \geq 1 - \left(\sum_{j=0}^{2} \binom{n}{j} \left(\frac{1}{M}\right)^j \left(1 - \frac{1}{M}\right)^{n-j}\right)^{M-4-i} := p_{\hat{C},i}.$$

From Lemma 3, we know

$$P\left\{\text{uniquely localizable}|k = i, \hat{C}\right\} \geq$$

$$P\left\{\text{empty sub-squares are densely surrounded}|k = i, \hat{C}\right\}.$$

The conditions of Lemma 3 require that empty sub-squares do not have empty simple neighbors; thus, we first find the probability that a sub-square does not have empty simple neighbors. Assume there are k empty sub-squares, say $s_1, s_2, \ldots, s_k$. Because of the independence assumption, these empty sub-squares are uniformly distributed.

Given the empty sub-square s_1, the probability that s_2 is not a simple neighbor of s_1 is at least $\left(1 - \frac{4}{M-1}\right)$; the probability that s_3 is not a simple neighbor of s_1 or s_2 is at least $(1 - 2 \cdot \frac{4}{M-2})$; and so on, so that the probability that no two empty sub-squares are neighbors is at least $\prod_{j=1}^{k-1}(1 - \frac{4j}{M-j})$. Moreover, the probability that an empty sub-square is densely surrounded, i.e., that all simple neighbors of an empty sub-square have at least two points and at least one of them has more than two points, is:

$$\hat{p} = P\{\text{All simple neighbors have at least two points}\} -$$

$$P\{\text{All simple neighbors have exactly two points}\}$$

$$= \left[1 - \sum_{j=0}^{1} \binom{n}{j} \left(\frac{1}{M}\right)^j \left(1 - \frac{1}{M}\right)^{n-j}\right]^4 - \left[\binom{n}{2} \left(\frac{1}{M}\right)^2 \left(1 - \frac{1}{M}\right)^{n-2}\right]^4.$$

Thus, the probability that all empty sub-squares are densely surrounded is

$$P\left\{\text{empty sub-squares are densely surrounded}| k = i, \hat{C}\right\} \geq \hat{p}^i \cdot \prod_{j=1}^{i-1}(1 - \frac{4j}{M-j}).$$

Note that the right hand side of the above equation is positive if $i < M/5$. Thus, we only consider grids with less than $u := \lfloor M/5 \rfloor - 1$ empty squares. Finally, we
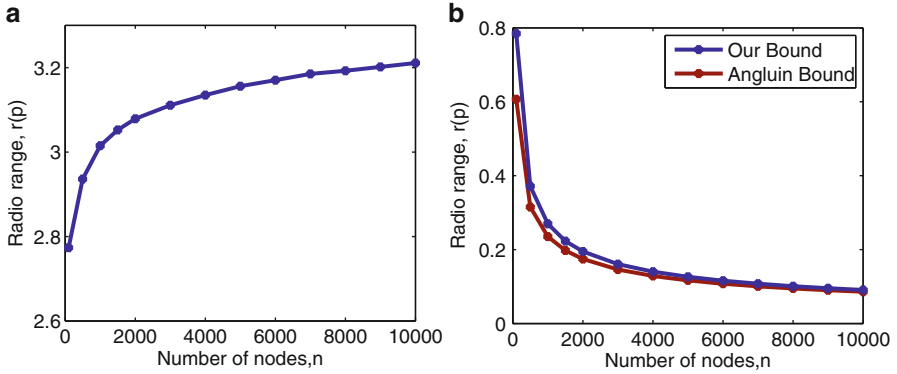
Fig. 3 Bound on the Connectivity Radius. (**a**) α vs. number of nodes. (**b**) r vs. number of nodes, compared to Angluin et al. bound

have the lower bound given by the following expression:

$$P\{\text{uniquely localizable}\}$$

$$\geq P\left\{\text{uniquely localizable}|\hat{C}\right\} P\left\{\hat{C}\right\}$$

$$= \sum_{i=0}^{u} P\{\text{uniquely localizable}|k=i, \hat{C}\} P\{\hat{C}|k=i\} P\{k=i\}$$

$$\geq \sum_{i=0}^{u} P\{\text{uniquely localizable}|k=i, \hat{C}\} P\{k=i\} p_{\hat{C},i}$$

$$\geq p_{\hat{C},0} P\{k=0\} + \sum_{i=1}^{u} p_{\hat{C},i} P\{k=i\}$$

$$\times P\left\{\text{empty sub-squares are densely surrounded}|k=i, \hat{C}\right\}$$

$$\geq p_{\hat{C},0} P\{k=0\} + \sum_{i=1}^{u} \hat{p}^{i} \cdot p_{\hat{C},i} P\{k=i\} \times \prod_{j=1}^{i-1}(1 - \frac{4j}{M-j}). \tag{2}$$

For different values of n (to be taken as the total number of sensor points), we can find values of M, and thus α (where α^2 can be viewed as the average number of sensor points in each sub-square), such that the right hand side of Equation (2) is at least 0.99. Figures 3a, b show α and r versus the number of points n such that the right hand side of Equation 2 is at least 0.99.

We also compare our connectivity bound against the bound of Angluin et al. [3] in Fig. 3. One can see that our bound and Angluin's are almost identical for any value of n. Thus, our result shows that the bound of Angluin et al. in (of $r > \frac{2\sqrt{2}\sqrt{\log n}}{\sqrt{n}}$ for $d = 2$) is true even when n is small, although it was initially proved to be an asymptotic bound when n is sufficiently large. Note that our bound, while not in an analytical form, is proved for any value of n.

We recently learned of another asymptotic bound that was independently developed by Javanmard and Montanari [23]. However, this bound is much weaker than ours and Angluin's.

Our connectivity result was proved for $\mathcal{H} = [0, 1]^2$, i.e., the unit square in dimension 2. The result can be extended to dimension $d > 2$. In summary, we have the following Theorem 2.

Theorem 2. *Let $\mathcal{H} \in [0, 1]^d$ be the unit hypercube in dimension d and be partitioned into a grid of $M = b^d$ equal sub-hypercubes, say $h_1, h_2, \ldots, h_M \subset \mathcal{H}$, where $\ell = 1/b$ is the edge length of each sub-hypercube. Let the number of sensor points in each sub-hypercube be independently and binomially generated according to $B\left(n, \frac{1}{M}\right)$ where $n = d \cdot M + 1$, and let one of the sub-hypercubes contain $d + 1$ anchors. Then, if the connectivity radius satisfies $r \geq 2\ell\sqrt{d}$, the probability that the sensor network is uniquely localizable is given by expression (2).*

Again, the parameter n of the binomial distribution can be viewed as the total number of sensor points in the region. We can also extend our result to another region $\mathcal{H}$ in dimension d into a grid of M equal sub-hypercubes in dimension d, say $h_1, h_2, \ldots, h_M \subset \mathcal{H}$, where each sub-hypercube h_i will have a volume of $1/M$, and the length of each of its edges will be $\ell := 1/\sqrt[d]{M}$. For example, we can assume $M = b_1 \cdot b_2 \cdots b_d$, where $b_i \geq 3$ for $i = 1, \ldots, d$ are positive integers.

3   Unique Localization of Triangulation Graph

The basic SDP localization model (SNL-SDP) is an SDP feasibility problem. When the network is not uniquely localizable, the max-rank of SDP feasible solutions is strictly greater than d. In practice, one may still be interested in finding a feasible SDP solution with rank d, representing one possible localization of points in $\mathbb{R}^d$. In this section, we show that adding an objective function that maximizes the sum of certain distances in a triangulation graph (in $\mathbb{R}^2$) will produce a rank-2 SDP solution. The result should be applicable to $d > 2$.

Definition 6. Consider a set of points $\mathcal{P} = \{p_1, p_2 \ldots p_n\} \in \mathbb{R}^2$. A *triangulation*, $\mathcal{T}_{\mathcal{P}}$, of the points in $\mathcal{P}$ is a subdivision of the convex hull of $\mathcal{P}$ into simplices (triangles) $\{p_i, p_j, p_k\}$, for some $i, j, k \in \{1, \ldots, n\}$, such that the edges of two simplices do not intersect or share a common face.

Definition 7. For a triangulation $\mathcal{T}_{\mathcal{P}}$, we define a *triangulation graph* $G_{\mathcal{T}_{\mathcal{P}}}(V, E)$ such that $V = \mathcal{P}$ and $(p_i, p_j) \in E$ if and only if (p_i, p_j) is an edge of a simplex in $\mathcal{T}_{\mathcal{P}}$. Note that triangles in a triangulation graph do not overlap, and triangles do not exist strictly inside other triangles.

Triangulation graphs and their properties have been studied in the literature [5, 12, 26]. Bruck et al. [12] showed that embedding a unit disk graph with local
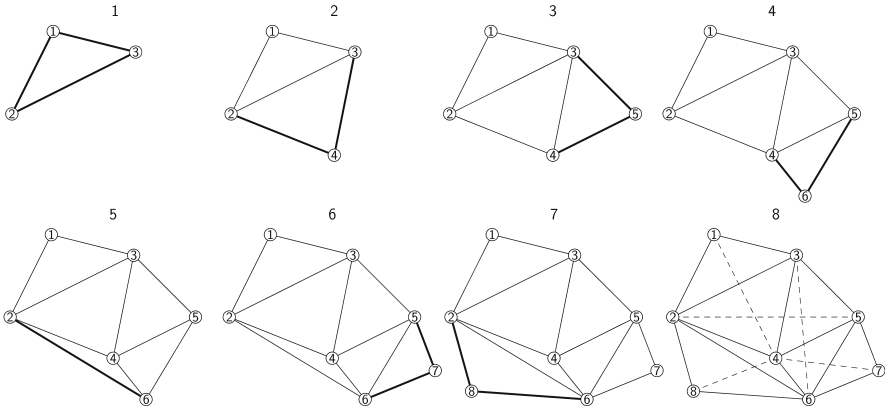
Fig. 4 Construction of triangulation graph

angle information (angles between points) is NP-hard, while the same problem on a triangulation graph is not. Araújo and Rodrigues [5] introduced an algorithm to construct a triangulation graph from a unit disk graph with $\mathcal{O}(n \log n)$ bit communications between points.

We formally decompose a triangulation $\mathcal{T}_\mathcal{P}$ into an initial clique K_3 and a set of actions $\mathcal{A} = \{a_1, a_2 \ldots a_m\}$, where an action a_i consists of adding a point and connecting it to either two adjacent points or two connected external points, where a point is called *external* if it is not strictly inside the convex hull of a cycle in the graph. This leads us to the following lemma, whose proof is omitted.

Lemma 4. *A triangulation can be constructed recursively by either adding an external point that connects to two adjacent points of a simplex (triangle) already in $\mathcal{T}_\mathcal{P}$ such that the new edges do not cross any existing edges (see Fig. 4, 1–4), or simply connecting two external points already in $\mathcal{T}_\mathcal{P}$ to form a triangle (see Fig. 4, 5).*

Proof. By induction on an external point; see Fig. 4. ∎

Step 8 of Fig. 4 shows the set of *virtual edges* in the sample triangulation graph. These virtual edges will be used to construct an appropriate objective function of the SDP relaxation for triangulation graphs.

Definition 8. In a triangulation graph, *adjacent triangles* are two triangles which share a common edge. A *virtual edge* exists between two points i and j when i and j belong to adjacent triangles, but $(i, j) \notin E$. The set of virtual edges between sensors is denoted E_v, and between sensors and anchors is denoted $\bar{E}_v$.

Consider adding an objective function to the SDP model (SNL-SDP) that maximizes the sum of the lengths of all virtual edges in a generic triangulation graph. The primal SDP relaxation, for $d = 2$, becomes:

$$\text{maximize} \quad \sum_{(k,j)\in\bar{E}_v} \bar{A}_{kj} \bullet Z + \sum_{(i,j)\in E_v} A_{ij} \bullet Z$$

$$\text{subject to} \quad Z_{(1:d,1:d)} = I_d$$

$$A_{ij} \bullet Z = d_{ij}^2, \forall(i,j) \in E \tag{3}$$

$$\bar{A}_{kj} \bullet Z = \bar{d}_{kj}^2, \forall(k,j) \in \bar{E}$$

$$Z \succeq 0$$

and the dual of (3) is:

$$\text{minimize} \quad I_d \bullet V + \sum_{(i,j)\in E} y_{ij} d_{ij}^2 + \sum_{(k,j)\in\bar{E}} w_{kj} \bar{d}_{kj}^2$$

$$\text{subject to} \quad U = \begin{pmatrix} V & 0 \\ 0 & 0 \end{pmatrix} + \sum_{(i,j)\in E} y_{ij} A_{ij} + \sum_{(k,j)\in\bar{E}} w_{kj} \bar{A}_{kj}$$

$$- \sum_{(k,j)\in\bar{E}_v} \bar{A}_{kj} - \sum_{(i,j)\in E_v} A_{ij} \tag{4}$$

$$U \succeq 0.$$

For a triangulation graph with at least three anchors, we can show that (4) is strictly feasible, i.e., there exists a feasible U with $U \succeq 0$ (see Proposition 4.1. in [31]). The primal SDP (3) also has a feasible point. As a result, the strong duality and complementarity condition hold for (3) and (4).

We derive the following exact-localization theorem.

Theorem 3. *Consider applying the SDP relaxation* (3) *to a generic triangulation graph with at least three anchors. Then, the rank of an optimal dual slack matrix of* (4) *is n and the rank of the optimal SDP solution of* (3) *is $d = 2$, so that the pair is strictly complementary and the SDP relaxation produces the correct localization.*

Proof. We use induction to show that the ranks of the optimal dual slack matrix U and primal SDP solution Z are n and $d = 2$, respectively. This implies that the strict complementarity conditions holds and (3) produces the correct localization, that is, the original true positions of the sensor points of the generic triangulation graph.

Assume the result is true for any triangulation graph with n points. It remains to be shown that this also holds for graphs with $n + 1$ points. It is clearly true for a single simplex when $n = 3$.

Let $X^n \in \mathbb{R}^{d \times n}$ be the correct locations of points, where the superindex n represents the number of points. By the induction assumption, the solution to (3) is $Z^n := \begin{pmatrix} I_d & X^n \\ (X^n)^T & (X^n)^T X^n \end{pmatrix}$. Moreover, the optimal dual slack matrix U^n satisfies $U^n \bullet Z^n = 0$ and has rank n; we can write the optimal dual slack matrix in terms of its submatrices $U^n = \begin{pmatrix} U_{11}^n & U_{12}^n \\ U_{21}^n & U_{22}^n \end{pmatrix}$, where $U_{11}^n \in \mathbb{R}^{d \times d}$ and $U_{22}^n \in \mathbb{R}^{n \times n}$. Note that $U_{22}^n \succ 0$, which follows from the fact that $\text{rank}(U_{22}^n) = n$ and $U^n \succeq 0$.

The complementarity condition $U^n \bullet Z^n = 0$ means the elements of U^n represent a stress on each edge such that the total force at all non-anchor points is zero (assuming, without loss of generality, a stress of -1 on all virtual edges).

Definition 9. Given a set of sensor locations $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$, let $G(V, E \cup \bar{E})$ be the corresponding graph. A matrix $U \in \mathbb{R}^{n \times n}$ is a *stress matrix* of the sensor network if it satisfies the constraints of (4) and $U \bullet (X^T X) = 0$. That is, each element of U represents a stress on the associated edge in $E \cup \bar{E}$ such that the total force on each non-anchor point is zero.

We decompose the triangulation graph into an initial simplex K_3, and actions $\mathcal{A} = \{a_1, a_2 \ldots a_m\}$. Without loss of generality, we assume the points in the first triangle are anchor points and let the last points added to the graph be x_{n+1}. For example, consider Fig. 4; let U^7 be the dual slack matrix on points 1–7 and assume the subgraph induced on the first seven points is uniquely localizable. When point 8 is added along with its incident edges, points $(2, 4, 6, 8)$ form a clique (when including the virtual edge between 4 and 8, which is unique when its length is maximized). Consider an SDP relaxation problem in dimension 4 that maximizes the length of the virtual edge between 4 and 8; this problem will have a unique optimal solution with rank 2 that determines the exact location of points $(2, 4, 6, 8)$, and an optimal dual slack matrix that forms a stress matrix for these four points.

Now consider the general case, where x_{n+1} is the last point added to the graph. A new triangle is created by adding x_{n+1}, its adjacent triangle and the virtual edge, which forms a 4-clique. Let Ω_0 be the corresponding positive-semidefinite stress matrix on the graph formed by x_{n+1}, the two points adjacent to x_{n+1} (say, g and h) and the point with which x_{n+1} has a virtual edge (say, k). We examine the case where g and h are sensors, however the case where at least one of them is an anchor is an easy extension. As before, the locations x_g, x_h, x_k, and x_{n+1} can be uniquely determined by solving an SDP relaxation, and Ω_0 is the optimal dual slack matrix that solves its dual problem:

$$\text{minimize} \quad y_{gk} d_{gk}^2 + y_{hk} d_{hk}^2 + y_{gh} d_{gh}^2 + y_{g,n+1} d_{g,n+1}^2 + y_{h,n+1} d_{h,n+1}^2 \quad (5)$$

$$\text{subject to} \quad U^4 \succeq 0,$$

where

U^4

$$= \begin{pmatrix} -1 + y_{gk} + y_{hk} & -y_{gk} & -y_{hk} & 1 \\ -y_{gk} & y_{gk} + y_{gh} + y_{g,n+1} & -y_{gh} & -y_{g,n+1} \\ -y_{hk} & -y_{gh} & y_{hk} + y_{gh} + y_{h,n+1} & -y_{h,n+1} \\ 1 & -y_{g,n+1} & -y_{h,n+1} & -1 + y_{g,n+1} + y_{h,n+1} \end{pmatrix}.$$

Assume $(y_{gk}, y_{hk}, y_{gh}, y_{g,n+1}, y_{h,n+1})$ is the optimal solution of this SDP, then

$$\Omega_0 = U^4.$$

It's easy to see that strong duality and complementarity condition hold for this SDP, and therefore

$$\sum_{i,j \in \{g,h,k,n+1\}} [\Omega_0]_{ij} (x_i^T x_j) = 0.$$

Note that $0 < (-1 + y_{g,n+1} + y_{h,n+1})$ because $\Omega_0 \succeq 0$, and consider the updated stress matrix

$$U_{22}^{n+1} := \begin{pmatrix} U_{22}^n & 0_{n \times 1} \\ 0_{1 \times n} & 0 \end{pmatrix} + \Omega,$$

where $\Omega \in \mathbb{R}^{(n+1) \times (n+1)}$ is the stress matrix of the new edges, that is, $\Omega_{([g,h,k,n+1],[g,h,k,n+1])} = \Omega_0$.

The new matrix U^{n+1} will be feasible for the dual, since Ω_0 is the solution of (5), and $\Omega \succeq 0$, $U^n \succeq 0$ implies that $U^{n+1} \succeq 0$.

Define

$$Z^{n+1} := \begin{pmatrix} Z^n & \begin{pmatrix} x_{n+1} \\ (X^n)^T x_{n+1} \end{pmatrix} \\ \begin{pmatrix} x_{n+1}^T & x_{n+1}^T X_n \end{pmatrix} & x_{n+1}^T x_{n+1} \end{pmatrix}$$

as the correct locations of the updated points. (Note that given this definition of Z^{n+1}, it does not immediately follow that $\mathrm{rank}(Z^{n+1}) = d$, since the added last row of Z^{n+1} can be linearly independent from the first n rows.) The sum of element-wise products of U^{n+1} and Z^{n+1} is

$$U^{n+1} \bullet Z^{n+1} = U^n \bullet Z^n + \sum_{(i,j)} [\Omega_0]_{ij} (x_i^T x_j) = 0.$$

Moreover, we can show that $U_{22}^{n+1} \succ 0$. Assume this is not true, i.e., assume that there is a vector $z \in \mathbb{R}^{n+1}$ such that

$$z^T U_{22}^{n+1} z = z^T \begin{bmatrix} U_{22}^n & 0 \\ 0 & 0 \end{bmatrix} z + z^T \Omega z = 0,$$

which holds if and only if $z^T \begin{bmatrix} U_{22}^n & 0 \\ 0 & 0 \end{bmatrix} z = 0$ and $z^T \Omega z = 0$. Since $U_{22}^n \succ 0$, this means that the first n elements of z are zero, i.e., $z_{(1:n)} = 0$. Thus,

$$z^T \Omega z = z_{n+1}^2 \Omega_{n+1} = z_{n+1}^2 (-1 + y_{g,n+1} + y_{h,n+1}) = 0$$

which implies $z_{n+1} = 0$. Thus, $z^T U_{22}^{n+1} z = 0$ if and only if $z = 0$, implying $U_{22}^{n+1} \succ 0$ and rank$(U^{n+1}) = n + 1$. Therefore, the rank of (Z^{n+1}) is d, and consequently from [30], Z^{n+1} is the unique solution to (3), so that the localization is correct and exact.

Theorem 3 implies that the strict complementarity condition holds when localizing a generic triangulation graph with the selected objective function. This result is interesting because, in general, it is difficult to prove strict complementarity for SDPs. How to compute a stress matrix (or optimal dual matrix) and determine whether the stress matrix has rank n are also important questions in rigidity theory for graph realization. Clearly, Theorem 3 is applicable to any graph that contains a generic triangulation graph as a spanning subgraph. In practice, the objective of the SDP relaxation may include all non-edges that are not specified in the given graph (rather than just virtual edges), which we experiment in the next section.

4   Heuristic Objective Function

Section 3 proves that adding a given objective function to (SNL-SDP) results in a correct localization for a certain class of graphs, whereas the formulation without an objective function may not.

Based on these findings, we tested a number of different SDP relaxation methods with different objective functions. For each method (i.e., each objective function), we ran the relaxation on a large number of random sensor networks and determined the success rate of each method. The following objective functions were tested to heuristically determine the best method.

1. (ZERO) Solve the formulation (SNL-SDP) (with no objective function). This can be viewed as a control simulation against which to compare other methods.
2. (MAX) Maximize the sum of all the 'non-edge' lengths by solving the formulation:

$$
\begin{aligned}
\text{maximize } & \sum_{(i,j)\notin E} d_{ij} + \sum_{(k,j)\notin \bar{E}} \bar{d}_{kj} \\
\text{subject to } & Z_{(1:d,1:d)} = I_d \\
& A_{ij} \bullet Z = d_{ij}^2, && \forall (i,j) \in E && \text{(SDP-MAX)} \\
& \bar{A}_{kj} \bullet Z = \bar{d}_{kj}^2, && \forall (k,j) \in \bar{E} \\
& Z \succeq 0.
\end{aligned}
$$

Table 1 Percent of networks correctly localized

		Method			
		ZERO	MAX	MIN	MAX-PT
Radio Range	0.15	0	0	0	0
	0.2	41	75	39	0
	0.25	87	95	88	0
	0.3	98	100	100	4
	0.35	100	100	100	7
	0.4	100	100	100	13

3. (MIN) Minimize the sum of all the 'non-edge' lengths by solving the formulation:

$$\begin{aligned}
\text{minimize} \quad & \sum_{(i,j)\notin E} d_{ij} + \sum_{(k,j)\notin \bar{E}} \bar{d}_{kj} \\
\text{subject to} \quad & Z_{(1:d,1:d)} = I_d \\
& A_{ij} \bullet Z = d_{ij}^2, \qquad \forall (i,j) \in E \qquad \text{(SDP-MIN)} \\
& \bar{A}_{kj} \bullet Z = \bar{d}_{kj}^2, \qquad \forall (k,j) \in \bar{E} \\
& Z \succeq 0.
\end{aligned}$$

4. (MAX-PT) Maximize the sum of the distances from each sensor location $x_i \in \mathbb{R}^d$ to a distant point, where x_i is set to the corresponding elements of the decision matrix Z. For example, for $\mathbf{1} \in \mathbb{R}^d$ the vector of all ones, we took the point $p := 1{,}000 \cdot \mathbf{1}$ and solved the formulation:

$$\begin{aligned}
\text{maximize} \quad & \sum_{i=1}^{n} \| p - x_i \|^2 \\
\text{subject to} \quad & Z_{(1:d,1:d)} = I_d \\
& A_{ij} \bullet Z = d_{ij}^2, \quad \forall (i,j) \in E \\
& \bar{A}_{kj} \bullet Z = \bar{d}_{kj}^2, \quad \forall (k,j) \in \bar{E} \\
& Z \succeq 0.
\end{aligned}$$

We constructed 200 uniformly distributed sensor networks and tested each method on the networks for a number of different radio ranges. Each randomly distributed sensor network has 100 points in a unit square (in dimension $d = 2$), and the distance between two points is known when they are within the given radio range. Table 1 shows the percent of sensor networks that were correctly localized using each method, for each radio range.

As can be seen from Table 1, maximizing the sum of the unknown distances out-performs the other three methods tested, and maximizing the sum of distances from a distant point does not produce good results. Moreover, the methods (ZERO), (MAX) and (MIN) all seemed to work very well when the radio range was at least 0.35. This radio range is much smaller than the lower bound given by (2), but has not been theoretically proved as a radio range that will lead to a correct localization.

References

1. Alfakih, A.Y.: On the universal rigidity of generic bar frameworks. Contrib. Discrete Math. **5**(3), 7–17 (2010)
2. Alfakih, A.Y., Khandani, A., Wolkowicz, H.: Solving euclidean distance matrix completion problems via semidefinite programming. Comput. Optim. Appl. **12**, 13–30 (1999)
3. Angluin, D., Aspnes, J., Chan, M., Fischer, M.J., Jiang, H., Peralta, R.: Stably computable properties of network graphs. In: Prasanna, V.K., Iyengar, S., Spirakis, P., Welsh, M. (eds.) Proceedings of the First IEEE International Conference Distributed Computing in Sensor Systems, DCOSS 2005, Marina del Rey, CA, USE, June/July, 2005. Lecture Notes in Computer Science, vol. 3560, pp. 63–74. Springer (2005)
4. Angluin, D., Aspnes, J., Diamadi, Z., Fischer, M.J., Peralta R.: Computation in networks of passively mobile finite-state sensors. Distrib. Comput. **18**(4), 235–253 (2006)
5. Araújo, F., Rodrigues, L.: Fast localized delaunay triangulation. In: Higashino, T. (ed.) Principles of Distributed Systems, vol. 3544, pp. 81–93. Springer, Berlin/Heidelberg (2005)
6. Aspnes, J., Eren, T., Goldenberg, D.K., Morse, A.S., Whiteley, W., Yang, Y.R., Anderson, B.D.O., Belhumeur, P.N.: A theory of network localization. IEEE Trans. Mob. Comput. **5**(12), 1663–1678, (2006)
7. Aspnes, J., Goldenberg, D., Yang, Y.R.: On the computational complexity of sensor network localization. In: First International Workshop on Algorithmic Aspects of Wireless Sensor Networks. Lecture Notes in Computer Science, vol. 3121 pp. 32–44. Springer (2004)
8. Badoiu, M., Demaine, E.D., Hajiaghayi, M., Indyk, P.: Low-dimensional embedding with extra information. Discrete Comput. Geom. **36**(4), 609–632 (2006)
9. Belk M., Connelly, R.: Realizability of graphs. In: Discrete and Computational Geometry, Springer-Verlag, vol. 37, pp. 7125–7137 (2007)
10. Biswas, P., Lian, T., Wang, T., Ye, Y.: Semidefinite programming based algorithms for sensor network localization. In: IPSN, ACM Transactions on Sensor Networks (TOSN), Berkeley, pp 46–54 (2004).
11. Biswas P., Ye, Y.: Semidefinite programming for ad hoc wireless network localization. In: IPSN, Berkeley, pp. 46–54 (2004)
12. Bruck, J., Gao, J., Jiang, A.: Localization and routing in sensor networks by local angle information. In: MobiHoc, the 6th ACM international symposium on Mobile ad hoc networking and computing, pp. 181–192. (2005)
13. Bulusu, N., Heidemann, J., Estrin, D.: Gps-less low-cost outdoor localization for very small devices. IEEE Pers. Commun. **7**(5), 28–34 (2000)
14. Clark, B.N., Colbourn, C.J., Johnson, D.S.: Unit disk graphs. Discrete Math. **86**(1–3), 165–177 (1991)
15. Connelly, R.: Generic global rigidity. Discrete Comput. Geom. **33**(4), 549–563 (2005)
16. Crippen G.M., Havel, T.F.: Distance geometry and molecular conformation. In: Chemometrics Series, Research Studies Press Ltd., Taunton, Somerset, England, volume 15, 1988.
17. Doherty, L., Pister, K.S.J., El Ghaoui, L.: Convex position estimation in wireless sensor networks. In: IEEE INFOCOM, Anchorge, vol. 3, pp. 1655–1663 (2001)
18. Eren, T., Goldenber, E.K., Whiteley, W., Yang, Y.R.: Rigidity, computation, and randomization in network localization. In: IEEE INFOCOM, Hong Kong, vol. 4, pp. 2673–2684 (2004)
19. Gortler, S.J., Healy, A.D., Thurston, D.P: Characterizing generic global rigidity. Am. J. Math. vol. 4, pp. 897–939 (2010)
20. Hendrickson, B.: Conditions for unique graph realizations. SIAM J. Comput. **21**(1), 65–84 (1992)
21. Gao, J., Bruck, J., Jiang, A.A.: Localization and routing in sensor networks by local angle information. ACM Trans. Sens. Networks **5**(1), 181–192 (2009)
22. Jackson B., Jordan, T: Connected rigidity matroids and unique realizations of graphs. J. Comb. Theory Ser. B **94**(1), 1–29 (2005)

23. Javanmard, A., Montanari, A.:Localization from incomplete noisy distance measurements. (2011). http://arxiv.org/abs/1103.1417v3
24. Krislock, N., Wolkowicz, H.: Explicit sensor network localization using semidefinite representations and clique reductions. SIAM J. Optim. **20**(5), 2679–2708 (2010)
25. Li, X.: Wireless Ad Hoc and Sensor Networks: Theory and Applications. Cambridge University Press, New York (2008)
26. Li, X.Y., Calinescu, G., Wan, P.J., Wang, Y.: Localized delaunay triangulation with application in ad hoc wireless networks. IEEE Transaction on Parallel and Distributed Systems, **14**(10), 1035–1047 (2003)
27. Priyantah, N.B., Balakrishnana, H., Demaine, E.D., Teller, S.: Mobile-assisted localization in wireless sensor networks. In: IEEE INFOCOM, Miami, vol. 1, pp. 172–183 (2005)
28. Recht, B., Fazel, M, Parrilo, P.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. **52**(3), 471–501 (2010)
29. Savvides, A., Han, C.C., Strivastava, M.B., Dynamic fine-grained localization in ad-hoc networks of sensors. In: MobiCom, Rome, pp. 166–179 (2001)
30. So, A.M.C., Ye, Y.: Theory of semidefinite programming for sensor network localization. In: Symposium on Discrete Algorithms, Mathematical Programming, Vancouver, Springer-verlag, **109**(2–3), 367–384 (2007)
31. So, A.M.C., Ye, Y.: A semidefinite programming approach to tensegrity theory and realizability of graphs. In: Symposium on Discrete Algorithms, Miami, pp. 766–775 (2006)
32. Tseng, P.: Second-order cone programming relaxation of sensor network localization. SIAM J. Optim. **18**(1), 156–185 (2007)
33. Wang, Z., Zheng, S., Ye, Y., Boyd, S.: Further relaxations of the semidefinite programming approach to sensor network localization. SIAM J. Optim. **19**, 655–673 (2008)
34. Yang, Z., Liu, Y., Li, X.Y.: Beyond trilateration: on the localizability of wireless ad-hoc networks. In: IEEE INFOCOM, Rio de Janeiro, pp. 2392–2400 (2009)
35. Zhu, Z., So, A.M.C., Ye,Y.: Universal rigidity: towards accurate and efficient localization of wireless networks. In: IEEE INFOCOM, San Diego (2010)

A Primal–Dual Smooth Perceptron–von Neumann Algorithm

Negar Soheili and Javier Peña

Abstract We propose an elementary algorithm for solving a system of linear inequalities $A^{\mathrm{T}} y > 0$ or its alternative $Ax = 0$, $x \geq 0$, $x \neq 0$. Our algorithm is a smooth version of the perceptron and von Neumann algorithms. Our algorithm retains the simplicity of these algorithms but has a significantly improved convergence rate. Our approach also extends to more general conic systems provided a suitable smoothing oracle is available.

Key words Perceptron algorithm • von Neumann algorithm • Condition number • Smoothing technique

Subject Classifications: 90C05, 90C25, 90C52

1 Introduction

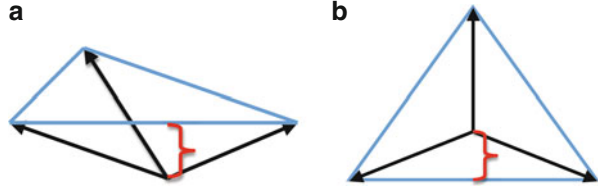We propose an elementary algorithm to solve the polyhedral feasibility problem

$$A^{\mathrm{T}} y > 0, \tag{1}$$

or its alternative

$$Ax = 0, \quad x \geq 0, \quad x \neq 0, \tag{2}$$

N. Soheili • J. Peña (✉)
Tepper School of Business, Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh, PA 15213, USA
e-mail: nsoheili@andrew.cmu.edu;jfp@andrew.cmu.edu

Fig. 1 Geometric
interpretation of $\rho(A)$. (**a**)
System (1) is feasible. (**b**)
System (2) is feasible

where A is a matrix in $\mathbb{R}^{m \times n}$. We refer to the systems (1) and (2) as *dual* and *primal*
feasibility problems, respectively. Feasibility problems like these are fundamental
in optimization, since every linear programming problem can be recast in the form
(1)–(2) via homogenization.

Our algorithm is based on a smooth version of both the perceptron and von
Neumann algorithms. The perceptron algorithm was introduced by Rosenblatt [17]
in 1958 for solving classification problems in machine learning. It is a simple greedy
algorithm that finds a solution to (1) when this system is feasible. Von Neumann
algorithm was privately communicated by von Neumann to Dantzig in the late
1940s, and later studied by Dantzig [6]. It is also a simple greedy algorithm that finds
an approximate solution to (2) when this system is feasible. Although the perceptron
and von Neumann algorithms have slow rate of convergence, the simplicity of their
iterations makes them attractive. As discussed in [18], these algorithms can be seen
as first-order methods to solve (1) and (2). The algorithm proposed herein is an
accelerated version of them, in the same spirit as the currently popular accelerated
first-order algorithms for convex optimization [11–13, 19].

The convergence rate of both the perceptron and von Neumann algorithms is
determined by the following parameter $\rho(A)$. Assume $A = \begin{bmatrix} a_1 \cdots a_n \end{bmatrix} \in \mathbb{R}^{m \times n}$ and
let $\| \cdot \|$ denote the Euclidean norm in $\mathbb{R}^m$. Let

$$\rho(A) := \left| \max_{\|y\|=1} \min_{j=1,\ldots,n} \frac{a_j^{\mathrm{T}} y}{\|a_j\|} \right|. \tag{3}$$

When (1) is feasible, $\rho(A)$ is precisely the *width* of the feasibility cone $\{y : A^{\mathrm{T}} y \geq 0\}$, as defined by Freund and Vera [10]. Furthermore, if the columns of A have all
Euclidean norm equal to one, then $\rho(A)$ is the Euclidean distance from the origin
to the boundary of the convex hull of $\{a_1, \ldots, a_n\}$. This geometric interpretation is
illustrated in Fig. 1 where the arrows depict the column vectors of A.

The parameter $\rho(A)$ is also a certain *radius of well-posedness* of the matrix A
(see [5, Theorem 1]). Cheung and Cucker [5] define $1/\rho(A)$ as a condition number
of the matrix A in relation to the problems (1) and (2). The quantity $1/\rho(A)$ can also
be seen as a special case of Renegar's condition number [15] for the systems (1) and
(2). (See [4, 5, 15, 16] for further details.)

Block [2] and Novikoff [14] showed that when (1) is feasible the perceptron
algorithm finds a solution to (1) after at most $1/\rho(A)^2$ iterations.

Given $\epsilon > 0$, we say that x is an ϵ-solution to (2) if $x \geq 0$, $\|x\|_1 = 1$ and
$\|Ax\| \leq \epsilon$. Under the assumption that the columns of A have Euclidean norm

one, Dantzig [6] showed that von Neumann algorithm finds an ϵ-solution to (2) in at most $\frac{1}{\epsilon^2}$ iterations when (2) is feasible. Epelman and Freund [9] showed that von Neumann algorithm either computes an ϵ-solution to (2) in $\mathcal{O}\left(\frac{1}{\rho(A)^2}\log\left(\frac{1}{\epsilon}\right)\right)$ iterations when (2) is feasible, or finds a solution to the alternative system (1) in $\mathcal{O}\left(\frac{1}{\rho(A)^2}\right)$ iterations if (1) is feasible.

The perceptron and von Neumann algorithms are not polynomial in the bit model of computation because the quantity $\rho(A)$ can be exponentially small in the bit length description of an input matrix A with rational entries. On the other hand, the quantity $\rho(A)$ makes sense for any A with real entries. The condition number $1/\rho(A)$, defined by Cheung and Cucker [5], is a natural parameter for studying the problems (1) and (2) in the real number model of computation introduced by Blum, Shub, and Smale [3].

There has been some recent research interest in the perceptron algorithm. Notably, Dunagan and Vempala [7, 8] proposed a randomized re-scaled version of the perceptron algorithm that with high probability terminates in $\mathcal{O}\left(m\log\left(\frac{1}{\rho(A)}\right)\right)$ re-scaling iterations, where each re-scaling iteration requires no more than $(32m)^2$ iterations of the perceptron algorithm. For an input matrix A with rational entries this algorithm is polynomial in the bit length description of A with high probability. More recently, Soheili and Peña [18] introduced a deterministic smooth perceptron algorithm that solves (1) in at most $\mathcal{O}\left(\frac{\sqrt{\log n}}{\rho(A)}\right)$ elementary iterations when (1) is feasible.

The main contribution in this paper is an *Iterated Smooth Perceptron–von Neumann Algorithm* (Algorithm ISPVN) that solves the pair of feasibility problems (1) and (2). Algorithm ISPVN relies on Nesterov's smoothing techniques [12, 13], and extends the smooth perceptron algorithm in [18]. For a given $A \in \mathbb{R}^{m\times n}$ with $\rho(A) > 0$, Algorithm ISPVN either finds an ϵ-solution to (2) in $\mathcal{O}\left(\frac{\sqrt{n}}{\rho(A)}\log\left(\frac{1}{\epsilon}\right)\right)$ elementary iterations, or finds a solution to the alternative system (1) in at most $\mathcal{O}\left(\frac{\sqrt{n}}{\rho(A)}\log\left(\frac{1}{\rho(A)}\right)\right)$ elementary iterations (see Theorem 1). Like the perceptron and von Neumann algorithms, the iterations in Algorithm ISPVN are *elementary* in the sense that they only involve simple computational steps. The iteration complexity of Algorithm ISPVN substantially improves the dependence on $\rho(A)$ of the iteration complexity $\mathcal{O}\left(\frac{1}{\rho(A)^2}\right)$ of the perceptron algorithm [2, 14], and the iteration complexity $\mathcal{O}\left(\frac{1}{\rho(A)^2}\log\left(\frac{1}{\epsilon}\right)\right)$ of von Neumann algorithm [9]. However, the new iteration bound incurs an extra factor of $\sqrt{n}$ whereas the former bounds depend solely on $\rho(A)$. In contrast to Soheili and Peña's smooth perceptron algorithm [18] that only applies to (1), our algorithm simultaneously handles both (1) and its alternative system (2). This comes at the expense of a weaker complexity bound in the case when (1) is feasible.

An extended version of Algorithm ISPVN applies to more general conic systems. More precisely, assume $\mathcal{K} \subseteq \mathbb{R}^n$ is a convex regular cone and $A \in \mathbb{R}^{m\times n}$. Then Algorithm ISPVN and its convergence properties extend in a natural fashion to the

conic system $A^{\mathrm{T}} y \in \mathrm{int}(\mathcal{K}^*)$, and its alternative $Ax = 0,\ x \in \mathcal{K},\ x \neq 0$ provided that a suitable smooth separation oracle for the cone $\mathcal{K}$ is available (see Theorem 2). An oracle of this kind is readily available for the main cones of interest in convex optimization, namely the non-negative orthant, the semidefinite cone, the second-order cone, and direct products of these types cones.

The paper is organized as follows. In Sect. 2 we briefly review the perceptron and von Neumann algorithms. In Sect. 3 we present our main contribution, namely a smooth perceptron–von Neumann algorithm. Section 4 presents a general version of smooth perceptron–von Neumann algorithm for conic systems. Section 5 presents the proof of a general version of our main theorem. Section 6 concludes the paper with some comments on potential directions for future work.

2   The Perceptron and von Neumann Algorithms

We next recall the perceptron and von Neumann algorithms, and observe that they can be seen as special cases of a general primal-dual template to solve the problems (1) and (2).

To solve (1) the perceptron algorithm starts with a trial point (usually zero). At each iteration, it updates the current trial point using the direction normal to one of the violated constraints (if any). This procedure is repeated until a solution to (1) is found.

For ease of notation we make the following assumption on the input matrix A throughout this section and Sect. 3.

Assumption 1. $A = \begin{bmatrix} a_1 \cdots a_n \end{bmatrix} \in \mathbb{R}^{m \times n}$ where $\|a_j\| = 1$ for $j = 1, \ldots, n$.

Here and throughout the entire paper $\| \cdot \|$ denotes the Euclidean norm.

Assumption 1 can be made without loss of generality since it only involves re-scaling the columns of A.

Perceptron Algorithm
```
begin
    y₀ := 0;
    for k = 0, 1, 2, . . .
        j := argmin aᵢᵀyₖ;
              i=1,...,n
        θₖ := 1/(k+1);
        y_{k+1} := (1 − θₖ)yₖ + θₖaⱼ;
        if Aᵀy_{k+1} > 0 then Halt // y_{k+1} is a solution to (1) fi
    end
end
```

We note that the above version of the perceptron algorithm includes a normalization of the iterates y_k and hence is slightly different from the description usually found in the literature, where the iteration is written $y_{k+1} = y_k + a_j$.

In contrast to the perceptron algorithm, von Neumann algorithm finds an approximate solution to the alternative system (2) and can be seen as a dual of the perceptron algorithm. Von Neumann algorithm starts with an initial point $x_0 \geq 0$ such that $\|x_0\|_1 = 1$, and iteratively generates a sequence $x_1, x_2, \ldots$ such that $x_k \geq 0$, $\|x_k\|_1 = 1$ and $\|Ax_k\| \to 0$, provided (2) is feasible. The point x_{k+1} is constructed as follows. First, identify the column a_j of A that forms the largest angle with Ax_k. Next, define x_{k+1} so that Ax_{k+1} is the point with smallest Euclidean norm along the segment joining Ax_k and a_j.

Let $e \in \mathbb{R}^n$ denote the n-dimensional vector of all ones and $e_i \in \mathbb{R}^n$ denote the unitary vector whose i-th entry is equal to one and all others are equal to zero.

Assume $\epsilon > 0$ is a given input.

Von Neumann Algorithm(ϵ)
begin
$\quad x_0 := \frac{e}{n}; y_0 := Ax_0;$
$\quad$ for $k = 0, 1, 2, \ldots$
$\quad\quad j := \underset{i=1,\ldots,n}{\operatorname{argmin}} a_i{}^{\mathrm{T}} y_k;$
$\quad\quad \theta_k := \underset{\lambda \in [0,1]}{\operatorname{argmin}} \{\|(1-\lambda)y_k + \lambda a_j\|\};$
$\quad\quad x_{k+1} := (1 - \theta_k)x_k + \theta_k e_j;$
$\quad\quad y_{k+1} := Ax_{k+1} = (1 - \theta_k)y_k + \theta_k a_j;$
$\quad\quad$ if $A^{\mathrm{T}} y_{k+1} > 0$ then Halt // y_{k+1} is a solution to (1) fi
$\quad\quad$ if $\|Ax_{k+1}\| < \epsilon$ then Halt // x_{k+1} is an ϵ-solution to (2) fi
$\quad$ end
end

The iterations in the above perceptron and von Neumann algorithms are similar. Each of these algorithms can be seen as a special case of the *Perceptron–von Neumann Template* below.

Let Δ_n denote the standard simplex in $\mathbb{R}^n$, that is, $\Delta_n := \{x \in \mathbb{R}^n : x \geq 0, \|x\|_1 = 1\}$. For $y \in \mathbb{R}^m$, let $x(y)$ denote any arbitrary point in the set $\operatorname{argmin}_{x \in \Delta_n} \langle A^{\mathrm{T}} y, x \rangle$. Observe that for a given $y \in \mathbb{R}^m$, we have $a_j{}^{\mathrm{T}} y = \min_i a_i{}^{\mathrm{T}} y$ if and only if $a_j = Ax(y)$ for $x(y) = e_j$.

Assume $\epsilon > 0$ is a given input.

Perceptron–von Neumann Template(ϵ)
begin
$\quad x_0 = \dfrac{e}{n}; y_0 := Ax_0;$
$\quad$ for $k = 0, 1, 2, \ldots$
$\quad\quad x_{k+1} := (1 - \theta_k)x_k + \theta_k x(y_k);$
$\quad\quad y_{k+1} := (1 - \theta_k)y_k + \theta_k Ax(y_k);$
$\quad\quad$ if $A^{\mathrm{T}} y_{k+1} > 0$ then Halt // y_{k+1} is a solution to (1) fi

```
        if ‖Axₖ₊₁‖ < ϵ then Halt //xₖ₊₁ is an ϵ-solution to (2) fi;
    end
end
```

Observe that the above perceptron–von Neumann template recovers the perceptron algorithm for $\theta_k := \frac{1}{k+1}$, and von Neumann algorithm for $\theta_k := \mathrm{argmin}_{\lambda \in [0,1]} \|(1-\lambda)\ y_k\ +\ \lambda A x(y_k)\|$ provided $x(y)$ is always chosen as one of the extreme points of the set $\mathrm{argmin}_{x \in \Delta_n} \langle A^{\mathrm{T}} y, x \rangle$.

3   Smooth Perceptron–von Neumann Algorithm

This section presents our main contribution, namely an *iterated smooth perceptron–von Neuman algorithm* (ISPVN) for solving both (1) and (2). Algorithm ISPVN relies on Nesterov's smoothing techniques [12, 13].

We start by considering the following *smooth* version of the map $y \mapsto x(y)$ defined above. Given $\bar{x} \in \Delta_n$ and $\mu > 0$ let $x_\mu : \mathbb{R}^m \to \Delta_n$ be defined as

$$x_\mu(y) := \underset{x \in \Delta_n}{\mathrm{argmin}} \left\{ \langle A^{\mathrm{T}} y, x \rangle + \frac{\mu}{2} \|x - \bar{x}\|^2 \right\}.$$

The minimizer $x_\mu(y)$ can be easily found by sorting the entries of $\bar{x} - \frac{1}{\mu} A^{\mathrm{T}} y$.

The following algorithm SPVN (Smooth Perceptron–von Neumann) is a smooth version of the Perceptron–von Neumann template. Assume $\bar{x} \in \Delta_n$ and $\delta > 0$ are given inputs.

Algorithm SPVN($\bar{x}, \delta$)
```
begin
    y₀ := Ax̄;  μ₀ := 2n;  x₀ := x_μ₀(y₀);
    for k = 0, 1, 2, . . .
        θₖ := 2/(k+3);
        yₖ₊₁ := (1 − θₖ)(yₖ + θₖAxₖ) + θₖ²Ax_μₖ(yₖ);
        μₖ₊₁ := (1 − θₖ)μₖ;
        xₖ₊₁ := (1 − θₖ)xₖ + θₖx_μₖ₊₁(yₖ₊₁);
        if Aᵀyₖ₊₁ > 0 then Halt // yₖ₊₁ is a solution to (1) fi
        if ‖Axₖ₊₁‖ ≤ δ then Return xₖ₊₁ fi;
    end
end
```

Algorithm SPVN is a slight modification of the smooth perceptron algorithm given in [18]. The main difference is that algorithm SPVN uses the Euclidean prox-function to smooth the map $x(y)$ instead of the entropy prox-function used in [18]. As Proposition 1 in Sect. 5 shows, Algorithm SPVN finds a solution to (1) in at most $\frac{2\sqrt{2n}}{\rho(A)} - 1$ iterations provided (1) is feasible and $\delta < \rho(A)$. On the other hand, when (2) is feasible, it can be shown that Algorithm SPVN halts after at most

$\mathcal{O}\left(\frac{1}{\sqrt{\delta}}\right)$ iterations with a δ-solution to (2). The following iterated version ISPVN of algorithm SPVN achieves a substantially better complexity when (2) is feasible.

Assume $\gamma > 1$ is a fixed constant and $\epsilon > 0$ is a given input.

Algorithm ISPVN(γ, ϵ)
begin
  $\tilde{x}_0 = \dfrac{e}{n}$;
  for $i = 0, 1, 2, \ldots$
    $\delta_i := \frac{\|A\tilde{x}_i\|}{\gamma}$;
    $\tilde{x}_{i+1} := \mathsf{SPVN}(\tilde{x}_i, \delta_i)$;
    if $\delta_i < \epsilon$ then Halt fi
  end
end

We are now ready to state our main result.

Theorem 1. *Assume $A \in \mathbb{R}^{m \times n}$ is such that $\rho(A) > 0$.*

(i) If the system (2) *is feasible then each call to* SPVN *in Algorithm* ISPVN *halts in at most*

$$\frac{2\sqrt{2n}\gamma}{\rho(A)} - 1$$

iterations.

  For any given $\epsilon > 0$ Algorithm ISPVN *finds an ϵ-solution to* (2) *in at most*

$$\frac{\log(1/\epsilon)}{\log(\gamma)}$$

outer iterations, that is, in at most $\left(\frac{2\sqrt{2n}\gamma}{\rho(A)} - 1\right) \cdot \left(\frac{\log(1/\epsilon)}{\log(\gamma)}\right) = \mathcal{O}\left(\frac{\sqrt{n}}{\rho(A)} \log\left(\frac{1}{\epsilon}\right)\right)$ elementary iterations.

(ii) If (1) *is feasible, then each call to* SPVN *in Algorithm* ISPVN *halts in at most*

$$\frac{2\sqrt{2n}}{\rho(A)} - 1$$

iterations.

  Algorithm ISPVN *finds either an ϵ-solution to* (2) *or a solution to* (1) *in at most*

$$\frac{\log(1/\rho(A))}{\log(\gamma)}$$

outer iterations, that is, in at most $\left(\frac{2\sqrt{2n}}{\rho(A)} - 1\right) \cdot \left(\frac{\log(1/\rho(A))}{\log(\gamma)}\right) = \mathcal{O}\left(\frac{\sqrt{n}}{\rho(A)} \log\left(\frac{1}{\rho(A)}\right)\right)$ elementary iterations.

Theorem 1 is a special case of the more general Theorem 2 presented in the next section.

4 Smooth Perceptron–von Neuman Algorithm for Conic Systems

Assume $\mathcal{K} \subseteq \mathbb{R}^n$ is a fixed *regular* convex cone, that is, $\mathcal{K}$ is closed, pointed and has non-empty interior. We next generalize Algorithms SPVN and ISPVN to the homogeneous conic system

$$A^{\mathrm{T}} y \in \operatorname{int}(\mathcal{K}^*), \tag{4}$$

and its alternative

$$Ax = 0, \ x \in \mathcal{K}, \ x \neq 0, \tag{5}$$

for a given matrix $A \in \mathbb{R}^{m \times n}$. We note that in contrast to Sects. 2 and 3, we do not assume that the columns of A are normalized.

We proceed by defining general versions of Δ_n, x_μ, and $\rho(A)$. Let $\mathbf{1} \in \operatorname{int}(\mathcal{K}^*)$ be fixed. Define the set $\Delta(\mathcal{K})$ as

$$\Delta(\mathcal{K}) := \{ x \in \mathbb{R}^n : \ x \in \mathcal{K}, \ \langle \mathbf{1}, x \rangle = 1 \}.$$

Given $\epsilon > 0$, we say that $x \in \Delta(\mathcal{K})$ is an ϵ-solution to (5) if $\|Ax\| \leq \epsilon$.

Observe that for $\mathcal{K} = \mathbb{R}^n_+$ and $\mathbf{1} = e = \begin{bmatrix} 1 \cdots 1 \end{bmatrix}^{\mathrm{T}}$, the set $\Delta(\mathcal{K})$ is precisely the standard simplex Δ_n. For the cone $\mathcal{K} = \mathbb{S}^n_+$ of symmetric positive semidefinite matrices in the space $\mathbb{S}^n$ of $n \times n$ symmetric matrices and $\mathbf{1} = I_n$, the set $\Delta(\mathcal{K})$ is $\{ X \in \mathbb{S}^n_+ : \operatorname{trace}(X) = 1 \}$ which is sometimes called the *spectraplex*. For the second-order cone $\mathbb{L}_n = \left\{ \begin{bmatrix} x_0 \\ \bar{x} \end{bmatrix} \in \mathbb{R}^n : \|\bar{x}\| \leq x_0 \right\}$ and $\mathbf{1} = \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix}$, the set $\Delta(\mathcal{K})$ is the lifted ball $\left\{ \begin{bmatrix} 1 \\ \bar{x} \end{bmatrix} \in \mathbb{R}^n : \|\bar{x}\| \leq 1 \right\}$.

Our extension of Algorithms SPVN and ISPVN to the conic systems (4) and (5) relies on the following key assumption.

Assumption 2. There is an available oracle that computes

$$\operatorname*{argmin}_{x \in \Delta(\mathcal{K})} \left\{ \frac{1}{2} \|x\|^2 - \langle g, x \rangle \right\} \tag{6}$$

for any given $g \in \mathbb{R}^n$.

Assumption 2 readily holds when $\mathcal{K} = \mathbb{R}^n_+$ and $\mathbf{1} = e = \begin{bmatrix} 1 \cdots 1 \end{bmatrix}^{\mathrm{T}}$. In this case $\Delta(\mathcal{K}) = \Delta_n$ and the solution to (6) is $x = (g - \theta e)^+$ where $\theta \in \mathbb{R}$ is such that $\|(g - \theta e)^+\|_1 = 1$. This θ can be obtained by sorting the values of g. Likewise, Assumption 2 holds when $\mathcal{K} = \mathbb{S}^n_+$ and $\mathbf{1} = I_n$. In this case the solution to (6) for $g \in \mathbb{S}^n$ is $x = U \operatorname{Diag}((\lambda(g) - \theta e)^+) U^{\mathrm{T}}$ where $g = U \operatorname{Diag}(\lambda(g)) U^{\mathrm{T}}$ is the spectral decomposition of g and $\theta \in \mathbb{R}$ is such that $\|(\lambda(g) - \theta e)^+\|_1 = 1$. This time the value of θ can be obtained by sorting the values of the vector of eigenvalues $\lambda(g)$.

Similarly, Assumption 2 also holds when $\mathcal{K} = \mathbb{L}_n$ and $\mathbf{1} = \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix}$. In this case the

solution to (6) for $g = \begin{bmatrix} g_0 \\ \bar{g} \end{bmatrix} \in \mathbb{R}^n$ is $x = (\lambda_1(g) - \theta)^+ \begin{bmatrix} 1 \\ \frac{\bar{g}}{\|\bar{g}\|} \end{bmatrix} + (\lambda_2(g) - \theta)^+ \begin{bmatrix} 1 \\ -\frac{\bar{g}}{\|\bar{g}\|} \end{bmatrix}$

where $g = \frac{1}{2}\lambda_1(g) \begin{bmatrix} 1 \\ \frac{\bar{g}}{\|\bar{g}\|} \end{bmatrix} + \frac{1}{2}\lambda_2(g) \begin{bmatrix} 1 \\ -\frac{\bar{g}}{\|\bar{g}\|} \end{bmatrix}$ is the Jordan algebra spectral

decomposition of g (see [1]), that is, $\lambda_1(g) = g_0 + \|\bar{g}\|$, $\lambda_2(g) = g_0 - \|\bar{g}\|$, and $\theta \in \mathbb{R}$ is such that $(\lambda_1(g) - \theta)^+ + (\lambda_2(g) - \theta)^+ = 1$. This value of θ is readily computable: $\theta = \begin{cases} \lambda_1(g) - 1 & \text{if } \lambda_1(g) \geq \lambda_2(g) + 1 \\ \frac{\lambda_1(g) + \lambda_2(g) - 1}{2} & \text{otherwise.} \end{cases}$

Proceeding in a similar fashion to the three cases above, it is easy to see that Assumption 2 also holds when $\mathcal{K}$ is a direct product of non-negative orthants, semidefinite cones, and second-order cones.

Given $\bar{x} \in \Delta(\mathcal{K})$ and $\mu > 0$ let $x_\mu : \mathbb{R}^m \to \Delta(\mathcal{K})$ be defined as

$$x_\mu(y) := \underset{x \in \Delta(\mathcal{K})}{\operatorname{argmin}} \left\{ \langle A^{\mathsf{T}} y, x \rangle + \frac{\mu}{2} \|x - \bar{x}\|^2 \right\}.$$

Observe that the mapping $x_\mu(\cdot)$ is computable by Assumption 2.

Assume M is a known upper bound on $\|A\|$. We are now ready to give the general versions of Algorithms SPVN and ISPVN.

Assume $\bar{x} \in \Delta(\mathcal{K})$ and $\delta > 0$ are given inputs.

Algorithm SPVNC$(\bar{x}, \delta)$
begin
    $y_0 := A\bar{x}$; $\mu_0 := 2M^2$; $x_0 := x_{\mu_0}(y_0)$;
    for $k = 0, 1, 2, \dots$
        $\theta_k := \frac{2}{k+3}$;
        $y_{k+1} := (1 - \theta_k)(y_k + \theta_k A x_k) + \theta_k^2 A x_{\mu_k}(y_k)$;
        $\mu_{k+1} := (1 - \theta_k)\mu_k$;
        $x_{k+1} := (1 - \theta_k)x_k + \theta_k x_{\mu_{k+1}}(y_{k+1})$;
        if $A^{\mathsf{T}} y_{k+1} > 0$ then Halt // y_{k+1} is a solution to (4) fi
        if $\|A x_{k+1}\| \leq \delta$ then Return x_{k+1} fi;
    end
end

Assume $\gamma > 1$ is a fixed constant and $\epsilon > 0$ is a given input.

Algorithm ISPVNC(γ, ϵ)
begin
    pick $\tilde{x}_0 \in \Delta(\mathcal{K})$;
    for $i = 0, 1, 2, \dots$
        $\delta_i := \frac{\|A\tilde{x}_i\|}{\gamma}$;
        $\tilde{x}_{i+1} := \text{SPVN}(\tilde{x}_i, \delta_i)$;

      if $\delta_i < \epsilon$ then Halt fi
   end
end

Let D denote the *diameter* of the set $\Delta(\mathcal{K})$, that is

$$D := \max_{u,v \in \Delta(\mathcal{K})} \|u - v\|. \tag{7}$$

Notice that D is well-defined and finite since the set $\Delta(\mathcal{K})$ is compact.

For a given $A \in \mathbb{R}^{m \times n}$ let

$$\rho(A) := \left| \max_{\|y\|=1} \min_{x \in \Delta(\mathcal{K})} \left\langle A^{\mathrm{T}} y, x \right\rangle \right|.$$

We have the following general version of Theorem 1.

Theorem 2. *Assume $A \in \mathbb{R}^{m \times n}$ is such that $\rho(A) > 0$.*

(i) If the system (5) is feasible then each call to SPVNC in Algorithm ISPVNC halts in at most

$$\frac{2MD\gamma}{\rho(A)} - 1 \tag{8}$$

iterations.

    For any given $\epsilon > 0$ Algorithm ISPVNC finds an ϵ-solution to (5) in at most

$$\frac{\log(\|A\tilde{x}_0\|/\epsilon)}{\log(\gamma)} \tag{9}$$

outer iterations, that is, in at most $\left(\frac{2MD\gamma}{\rho(A)} - 1 \right) \cdot \left(\frac{\log(\|A\tilde{x}_0\|/\epsilon)}{\log(\gamma)} \right) = \mathcal{O}\left(\frac{MD}{\rho(A)} \log\left(\frac{\|A\tilde{x}_0\|}{\epsilon} \right) \right)$ elementary iterations.

(ii) If (4) is feasible, then each call to SPVNC in Algorithm ISPVNC halts in at most

$$\frac{2MD}{\rho(A)} - 1 \tag{10}$$

iterations.

    Algorithm ISPVNC finds either an ϵ-solution to (5) or a solution to (4) in at most

$$\frac{\log(\|A\tilde{x}_0\|/\rho(A))}{\log(\gamma)} \tag{11}$$

outer iterations, that is, in at most $\left(\frac{2MD}{\rho(A)} - 1 \right) \cdot \left(\frac{\log(\|A\tilde{x}_0\|/\rho(A))}{\log(\gamma)} \right) = \mathcal{O}\left(\frac{MD}{\rho(A)} \log\left(\frac{\|A\tilde{x}_0\|}{\rho(A)} \right) \right)$ elementary iterations.

We conclude this section by showing that Theorem 1 follows from Theorem 2.

Proof (Proof of Theorem 1). Since the columns of A are normalized we have $\|A\| \leq \sqrt{n}$. Hence Algorithms SPVN and ISPVN are recovered as special cases of Algorithms SPVNC and ISPVNC respectively for $\mathcal{K} = \mathbb{R}_+^n$ and $\Delta(\mathcal{K}) = \Delta_n$. Next, observe that the diameter of Δ_n is $\sqrt{2}$. Furthermore, for the initial point $\tilde{x}_0 = \frac{e}{n}$ we have $\|A\tilde{x}_0\| \leq 1$ because $\|\tilde{x}_0\|_1 = 1$ and the columns of A are normalized. Therefore for $\mathcal{K} = \mathbb{R}_+^n$, $\Delta(\mathcal{K}) = \Delta_n$, $\tilde{x}_0 = \frac{e}{n}$, and A with normalized columns, we have $D = \sqrt{2}$, $M = \sqrt{n}$, and $\|A\tilde{x}_0\| \leq 1$. Consequently, in this case Theorem 2 reduces to Theorem 1. $\qquad\square$

5 Proof of Theorem 2

Proposition 1 and Proposition 3 below are the crux of the proof of Theorem 2. These propositions show that each call to SPVNC in Algorithm ISPVNC halts in $\mathcal{O}\left(\frac{MD}{\rho(A)}\right)$ iterations. The proofs of these propositions use ideas introduced by Soheili and Peña in [18]. Let $\varphi : \mathbb{R}^m \to \mathbb{R}$ be defined as

$$\varphi(y) := -\frac{1}{2}\|y\|^2 + \min_{x \in \Delta(\mathcal{K})} \langle A^{\mathsf{T}}y, x\rangle.$$

Observe that if $y \in \mathbb{R}^m$ is such that $\varphi(y) > 0$, then $A^{\mathsf{T}}y \in \operatorname{int}(\mathcal{K}^*)$.

Given $\bar{x} \in \Delta(\mathcal{K})$ and $\mu > 0$, consider the smooth approximation φ_μ of φ defined as follows:

$$\begin{aligned}
\varphi_\mu(y) &= -\frac{1}{2}\|y\|^2 + \min_{x \in \Delta(\mathcal{K})}\left\{\langle A^{\mathsf{T}}y, x\rangle + \frac{\mu}{2}\|x - \bar{x}\|^2\right\} \\
&= -\frac{1}{2}\|y\|^2 + \langle A^{\mathsf{T}}y, x_\mu(y)\rangle + \frac{\mu}{2}\|x_\mu(y) - \bar{x}\|^2.
\end{aligned} \tag{12}$$

We will rely on the following properties of the functions φ, φ_μ. Recall that D stands for the diameter of $\Delta(\mathcal{K})$ as defined in (7).

Lemma 1. *Assume $A \in \mathbb{R}^{m \times n}$ is given.*

(i) *For all $\mu > 0$*
$$0 \leq \varphi_\mu(y) - \varphi(y) \leq \frac{1}{2}\mu D^2.$$

(ii) *The iterates $x_k \in \Delta(\mathcal{K})$, $y_k \in \mathbb{R}^m$, $\mu_k \in \mathbb{R}$, $k = 0, 1, \dots$ generated by Algorithm SPVNC satisfy*
$$\frac{1}{2}\|Ax_k\|^2 \leq \varphi_{\mu_k}(y_k). \tag{13}$$

(iii) *If $\rho(A) > 0$ and the system (4) is feasible then for all $y \in \mathbb{R}^m$ and $x \in \Delta(\mathcal{K})$,*

$$\varphi(y) \leq \frac{1}{2}\rho(A)^2 \leq \frac{1}{2}\|Ax\|^2.$$

Lemma 1 is a straightforward extension of the ideas and proof techniques in [18]. To make our exposition self-contained we present a proof of it in Appendix 6.

Proposition 1. *Assume* $A \in \mathbb{R}^{m \times n}$ *is given and* $\rho(A) > 0$. *If the system* (4) *is feasible, Algorithm* **SPVNC** *halts in at most*

$$\frac{2MD}{\rho(A)} - 1$$

iterations.

Proof. It suffices to bound the number of iterations when $\delta < \rho(A)$ since otherwise the algorithm can only halt sooner. If indeed $\delta < \rho(A)$, Lemma 1(iii) implies that Algorithm **SPVNC** can only halt when a solution to (4) is found. In the algorithm $\mu_0 = 2M^2$ and $\mu_{k+1} = \frac{k+1}{k+3}\mu_k$, so

$$\mu_k = \frac{4M^2}{(k+1)(k+2)} < \frac{4M^2}{(k+1)^2}.$$

By Lemma 1(iii,ii,i) it follows that

$$\frac{1}{2}\rho(A)^2 \leq \frac{1}{2}\|Ax_k\|^2 \leq \varphi_{\mu_k}(y_k) \leq \varphi(y_k) + \frac{1}{2}\mu_k D^2 < \varphi(y_k) + \frac{2M^2 D^2}{(k+1)^2}.$$

Thus $\varphi(y_k) > 0$, and consequently y_k is a solution to (4), if $k \geq \frac{2MD}{\rho(A)} - 1$. $\square$

We will rely on the characterization (14) below of $\rho(A)$ when (5) is feasible. Figure 1(b) illustrates this characterization in the special case when $\mathcal{K}$ is the non-negative orthant. We note that this property is closely related to a characterization of Renegar's *distance to ill-posedness,* see [16, Theorem 3.5]. Related properties of $\rho(A)$ are also discussed in [5] and [4, Chap. 6].

Proposition 2. *Assume* $\rho(A) > 0$ *and the problem* (5) *is feasible. Then*

$$\rho(A) = \sup\{\delta : y \in \mathbb{R}^m, \ \|y\| \leq \delta \Rightarrow y \in A(\Delta(\mathcal{K}))\}. \tag{14}$$

Proof. We first show the inequality "$\geq$" in (14). To that end, suppose $\delta > 0$ is such that $\tilde{y} \in A(\Delta(\mathcal{K}))$ for any $\tilde{y} \in \mathbb{R}^m$ with $\|\tilde{y}\| \leq \delta$. Given an arbitrary $y \in \mathbb{R}^m$ with $\|y\| = 1$, put $\tilde{y} := -\delta y$. By our assumption on δ, there exists $\tilde{x} \in \Delta(\mathcal{K})$ such that $A\tilde{x} = \tilde{y}$. In addition, $\langle A\tilde{x}, y \rangle = \langle \tilde{y}, y \rangle = -\delta$. So $\min_{x \in \Delta(\mathcal{K})} \langle Ax, y \rangle \leq -\delta$. Since this holds for any arbitrary $y \in \mathbb{R}^m$ with $\|y\| = 1$, we have $\max_{\|y\|=1} \min_{x \in \Delta(\mathcal{K})} \langle Ax, y \rangle \leq -\delta$. Therefore, $\rho(A) \geq \delta$.

Next we show the inequality "$\leq$" in (14). To do so, it suffices to show that if $y \notin A(\Delta(\mathcal{K}))$ then $\rho(A) < \|y\|$. Observe that $A(\Delta(\mathcal{K}))$ is closed and convex because $\Delta(\mathcal{K})$ is compact and convex. Therefore, if $y \notin A(\Delta(\mathcal{K}))$ then there exists

a hyperplane that separates y and $A(\Delta(\mathcal{K}))$. More precisely, there exists $z \in \mathbb{R}^m$ with $\|z\| = 1$ such that

$$\langle z, y \rangle < \min_{x \in \Delta(\mathcal{K})} \langle z, Ax \rangle \leq \max_{\|y\|=1} \min_{x \in \Delta(\mathcal{K})} \langle y, Ax \rangle = -\rho(A).$$

Hence by Schwarz inequality,

$$\rho(A) < |\langle z, y \rangle| \leq \|z\| \|y\| = \|y\|. \qquad \square$$

Throughout the rest of this section let $S := \{x \in \Delta(\mathcal{K}) : Ax = 0\}$, and for $v \in \mathbb{R}^n$ let $\mathsf{dist}(v, S) := \min\{\|v - x\| : x \in S\}$.

Lemma 2. *Assume* (5) *is feasible and* $\rho(A) > 0$. *Then for all* $v \in \Delta(\mathcal{K})$

$$\mathsf{dist}(v, S) \leq \frac{\|Av\| D}{\rho(A)}. \qquad (15)$$

Proof. Given an arbitrary $v \in \Delta(\mathcal{K})$, the inequality (15) is clearly true if $v \in S$. Assume $v \in \Delta(\mathcal{K}) \setminus S$. Consider $y := -\frac{Av}{\|Av\|} \rho(A)$. By Proposition 2 there exists $u \in \Delta(\mathcal{K})$ such that $Au = y = -\frac{Av}{\|Av\|} \rho(A)$. Let $x = \lambda u + (1 - \lambda)v$ for $\lambda = \frac{\|Av\|}{\|Av\| + \rho(A)}$. Then $x \in S$ and

$$\|v - x\| = \lambda \|u - v\| \leq \lambda D = \frac{\|Av\| D}{\|Av\| + \rho(A)} \leq \frac{\|Av\| D}{\rho(A)}. \qquad \square$$

Proposition 3. *Assume* (5) *is feasible and* $\rho(A) > 0$. *If* $\bar{x} \in \Delta(\mathcal{K})$ *and* $\delta = \frac{\|A\bar{x}\|}{\gamma}$ *for some* $\gamma > 1$, *then Algorithm* **SPVNC** *with input* $(\bar{x}, \delta)$ *terminates in at most*

$$\frac{2MD\gamma}{\rho(A)} - 1 \qquad (16)$$

iterations.

Proof. By Lemma 1(ii), at iteration k of Algorithm **SPVN** we have

$$\begin{aligned}
\frac{1}{2} \|Ax_k\|^2 &\leq \varphi_{\mu_k}(y_k) \\
&\leq -\frac{\|y_k\|^2}{2} + \min_{x \in S} \left\{ \langle A^\mathsf{T} y_k, x \rangle + \frac{\mu_k}{2} \|x - \bar{x}\|^2 \right\} \\
&\leq \frac{\mu_k}{2} \min_{x \in S} \|x - \bar{x}\|^2 \\
&= \frac{\mu_k}{2} \mathsf{dist}(\bar{x}, S)^2.
\end{aligned} \qquad (17)$$

Thus by Lemma 2

$$\|Ax_k\| \leq \sqrt{\mu_k} \cdot \mathsf{dist}(\bar{x}, S) \leq \sqrt{\mu_k} \cdot \frac{D \|A\bar{x}\|}{\rho(A)} \leq \frac{2MD \|A\bar{x}\|}{(k+1)\rho(A)}.$$

So when $k \geq \frac{2MD\|A\tilde{x}\|}{\rho(A)\delta} - 1 = \frac{2MD\gamma}{\rho(A)} - 1$ we have $\|Ax_k\| \leq \delta$ and Algorithm SPVNC halts.   □

Proof (Proof of Theorem 2).

(i) The bound (8) readily follows from Proposition 3. For (9), observe that after N outer iterations algorithm ISPVNC yields $\tilde{x}_N \in \Delta(\mathcal{K})$ with

$$\|A\tilde{x}_N\| \leq \delta_{N-1} \leq \frac{\|A\tilde{x}_0\|}{\gamma^N}.$$

Thus, $\|A\tilde{x}_N\| \leq \delta_{N-1} < \epsilon$, and so algorithm ISPVNC halts, in at most $N = \frac{\log(\|A\tilde{x}_0\|/\epsilon)}{\log(\gamma)}$ outer iterations.

(ii) Proposition 1 readily yields the bound (10). Furthermore, the proof of Proposition 1 shows that algorithm ISPVNC halts with a solution to (4) when $\delta_{N-1} < \rho(A)$. By part (i) we know that $\delta_{N-1} \leq \frac{\|A\tilde{x}_0\|}{\gamma^N}$. Thus $\delta_{N-1} < \rho(A)$, and so algorithm ISPVNC halts with a solution to (4), in at most $N = \frac{\log(\|A\tilde{x}_0\|/\rho(A))}{\log(\gamma)}$ outer iterations. Note that Algorithm ISPVNC may halt with an ϵ-solution to (5) in fewer outer iterations if $\epsilon > \rho(A)$.   □

6   Conclusions and Future Work

We have proposed an iterated smooth perceptron-von Neumann algorithm (Algorithm ISPVN) for the alternative systems (1) and (2). Our algorithm retains the simplicity of the classical perceptron and von Neumann algorithms while improving their iteration complexity roughly from $\mathcal{O}(1/\rho(A)^2)$ to $\mathcal{O}(\sqrt{n}/\rho(A))$. A key ingredient of our approach is the smoothing of the mapping $x(y) = \operatorname{argmin}_{x \in \Delta_n} \langle A^\mathsf{T} y, x \rangle$ to $x_\mu(y) = \operatorname{argmin}_{x \in \Delta_n} \{ \langle A^\mathsf{T} y, x \rangle + \frac{\mu}{2} \|x - \bar{x}\|^2 \}$.

It is natural to ask whether our main results hold if a different prox-function is used to smooth the mapping $x(\cdot)$. In particular, the entropy function $\sum_{i=1}^n x_i \log x_i + \log n$ can be used in Algorithm SPVN in place of the Euclidean distance function $\frac{1}{2}\|x - \bar{x}\|^2$. In this case, a non-iterated version of Algorithm SPVN solves (1) in $\mathcal{O}(\sqrt{\log n}/\rho(A))$ iterations provided it is feasible, as shown by Soheili and Peña [18, Theorem 2.1]. We conjecture that the main factor $\mathcal{O}(\sqrt{n}/\rho(A))$ in Theorem 1 can be improved to $\mathcal{O}(\sqrt{\log n}/\rho(A))$ if the entropy function is suitably used. It is tempting to look for a proof of this conjecture by modifying Algorithm ISPVN and the proof of Theorem 1 in obvious ways. However, this attempt runs into a roadblock because it needs a bound as that in Lemma 2 but with the entropy-induced Bregman distance in place of the Euclidean distance. Such an analog of Lemma 2 does not hold.

It is also natural to ask whether Algorithm SPVN could be combined with the re-scaling phase proposed by Dunagan and Vempala [7, 8] to obtain a faster

probabilistic re-scaled perceptron algorithm. In particular, the bulk of the work of each outer iteration of Dunagan and Vempala's algorithm consists of up to $(32m)^2$ iterations in a perceptron phase, and up to $(32m)^2 \log m$ iterations in a perceptron improvement phase. We conjecture that the smoothing techniques used herein may reduce the number of iterations in both of these phases to something like $\mathcal{O}(m\sqrt{m})$ or $\mathcal{O}(m\sqrt{m} \log m)$.

Appendix

Proof of Lemma 1

(i) From the construction of φ and φ_μ it follows that

$$
\begin{aligned}
\varphi_\mu(y) &= -\frac{1}{2}\|y\|^2 + \left\langle A^\mathsf{T} y, x_\mu(y) \right\rangle + \tfrac{\mu}{2}\|x_\mu(y) - \bar{x}\|^2 \\
&\geq -\frac{1}{2}\|y\|^2 + \left\langle A^\mathsf{T} y, x_\mu(y) \right\rangle \\
&\geq -\frac{1}{2}\|y\|^2 + \min_{x \in \Delta(\mathcal{K})} \left\langle A^\mathsf{T} y, x \right\rangle \\
&= \varphi(y).
\end{aligned}
$$

In addition,

$$
\begin{aligned}
\varphi_\mu(y) &= -\frac{1}{2}\|y\|^2 + \min_{x \in \Delta(\mathcal{K})} \left\{ \left\langle A^\mathsf{T} y, x \right\rangle + \tfrac{\mu}{2}\|x - \bar{x}\|^2 \right\} \\
&\leq -\frac{1}{2}\|y\|^2 + \left\langle A^\mathsf{T} y, x(y) \right\rangle + \tfrac{\mu}{2}\|x(y) - \bar{x}\|^2 \\
&\leq \varphi(y) + \tfrac{1}{2}\mu D^2.
\end{aligned}
$$

(ii) We proceed by induction. For $k = 0$ we have:

$$
\begin{aligned}
\frac{1}{2}\|Ax_0\|^2 &= \frac{1}{2}\|A\bar{x}\|^2 + \left\langle A\bar{x}, A(x_0 - \bar{x}) \right\rangle + \frac{1}{2}\|A(x_0 - \bar{x})\|^2 \\
&\leq -\frac{1}{2}\|A\bar{x}\|^2 + \left\langle A^\mathsf{T} A\bar{x}, x_0 \right\rangle + \frac{1}{2}\|A\|^2 \|x_0 - \bar{x}\|^2 \\
&\leq -\frac{1}{2}\|y_0\|^2 + \left\langle A^\mathsf{T} y_0, x_{\mu_0}(y_0) \right\rangle + \frac{1}{2}\mu_0 \|x_{\mu_0}(y_0) - \bar{x}\|^2 \\
&= \varphi_{\mu_0}(y_0).
\end{aligned}
$$

Now we will show that if (13) holds for k then it also holds for $k + 1$. To ease notation, drop the index k and write y_+, x_+, μ_+ for y_{k+1}, x_{k+1}, μ_{k+1} respectively. Also, let $\hat{x} = (1 - \theta)x + \theta x_\mu(y)$ so that $y_+ = (1 - \theta)y + \theta A\hat{x}$. We have

$$
\begin{aligned}
\varphi_{\mu_+}(y_+) &= -\frac{\|y_+\|^2}{2} + \langle A^{\mathsf{T}} y_+, x_{\mu_+}(y_+) \rangle + \frac{\mu_+}{2} \|x_{\mu_+}(y_+) - \bar{x}\|^2 \\
&= -\frac{\|(1-\theta)y + \theta A\hat{x}\|^2}{2} \\
&\quad + (1-\theta)\left[\langle A^{\mathsf{T}} y, x_{\mu_+}(y_+) \rangle + \frac{\mu}{2} \|x_{\mu_+}(y_+) - \bar{x}\|^2 \right] \\
&\quad + \theta \langle A^{\mathsf{T}} A\hat{x}, x_{\mu_+}(y_+) \rangle \\
&\geq (1-\theta)\left[-\frac{\|y\|^2}{2} + \langle A^{\mathsf{T}} y, x_{\mu_+}(y_+) \rangle + \frac{\mu}{2} \|x_{\mu_+}(y_+) - \bar{x}\|^2 \right]_1 \\
&\quad + \theta \left[-\frac{\|A\hat{x}\|^2}{2} + \langle A^{\mathsf{T}} A\hat{x}, x_{\mu_+}(y_+) \rangle \right]_2 .
\end{aligned}
\tag{18}
$$

The last inequality follows from the concavity of the function $y \mapsto -\frac{\|y\|^2}{2}$. Using (12), we can estimate the expression in the first bracket in (18) as follows:

$$
\begin{aligned}
[.]_1 &= \varphi_\mu(y) + \langle A^{\mathsf{T}} y, x_{\mu_+}(y_+) - x_\mu(y) \rangle \\
&\quad + \frac{\mu}{2} \left(\|x_{\mu_+}(y_+) - \bar{x}\|^2 - \|x_\mu(y) - \bar{x}\|^2 \right) \\
&= \varphi_\mu(y) + \langle A^{\mathsf{T}} y + \mu(x_\mu(y) - \bar{x}), x_{\mu_+}(y_+) - x_\mu(y) \rangle \\
&\quad + \frac{\mu}{2} \|x_{\mu_+}(y_+) - x_\mu(y)\|^2 \\
&\geq \varphi_\mu(y) + \frac{\mu}{2} \|x_{\mu_+}(y_+) - x_\mu(y)\|^2 \\
&\geq \frac{1}{2} \|Ax\|^2 + \frac{\mu}{2} \|x_{\mu_+}(y_+) - x_\mu(y)\|^2 \\
&\geq \frac{1}{2} \|A\hat{x}\|^2 + \langle A^{\mathsf{T}} A\hat{x}, x - \hat{x} \rangle + \frac{\mu}{2} \|x_{\mu_+}(y_+) - x_\mu(y)\|^2 . \tag{19}
\end{aligned}
$$

The third step above follows from the optimality conditions for (12) at $x_\mu(y)$. The fourth step follows from the induction hypothesis (13).

The expression in the second bracket in (18) can be written as

$$
[.]_2 = \frac{1}{2} \|A\hat{x}\|^2 + \langle A^{\mathsf{T}} A\hat{x}, x_{\mu_+}(y_+) - \hat{x} \rangle .
\tag{20}
$$

Observe also that

$$\begin{aligned} x_+ - \hat{x} &= (1 - \theta)x + \theta x_{\mu_+}(y_+) - (1 - \theta)x - \theta x_\mu(y) \\ &= \theta(x_{\mu_+}(y_+) - x_\mu(y)). \end{aligned} \tag{21}$$

Plugging (19) and (20) into (18) we get

$$\begin{aligned} \varphi_{\mu_+}(y_+) &\geq (1 - \theta)\left[\tfrac{1}{2}\|A\hat{x}\|^2 + \langle A^{\mathrm{T}}A\hat{x}, x - \hat{x}\rangle + \tfrac{\mu}{2}\|x_{\mu_+}(y_+) - x_\mu(y)\|^2\right] \\ &\quad + \theta\left[\tfrac{1}{2}\|A\hat{x}\|^2 + \langle A^{\mathrm{T}}A\hat{x}, x_{\mu_+}(y_+) - \hat{x}\rangle\right] \\ &= \tfrac{1}{2}\|A\hat{x}\|^2 + \theta\langle A^{\mathrm{T}}A\hat{x}, x_{\mu_+}(y_+) - x_\mu(y)\rangle \\ &\quad + \tfrac{(1-\theta)\mu}{2}\|x_{\mu_+}(y_+) - x_\mu(y)\|^2 \\ &\geq \tfrac{1}{2}\|A\hat{x}\|^2 + \theta\langle A^{\mathrm{T}}A\hat{x}, x_{\mu_+}(y_+) - x_\mu(y)\rangle \\ &\quad + \tfrac{1}{2}\theta^2\|A\|^2\|x_{\mu_+}(y_+) - x_\mu(y)\|^2 \\ &\geq \tfrac{1}{2}\|A\hat{x}\|^2 + \langle A^{\mathrm{T}}A\hat{x}, x_+ - \hat{x}\rangle + \tfrac{1}{2}\|A(x_+ - \hat{x})\|^2 \\ &= \tfrac{1}{2}\|Ax_+\|^2. \end{aligned}$$

The second step above follows because $\hat{x} = (1 - \theta)x + \theta x_\mu(y)$. The third step follows because at iteration k we have $\frac{\theta^2\|A\|^2}{1-\theta} = \frac{4\|A\|^2}{(k+1)(k+3)} \leq \frac{4M^2}{(k+1)(k+2)} = \mu$. The fourth step follows from (21).

(iii) Since the mapping $v \mapsto \min_{x \in \Delta(\mathcal{K})} \langle A^{\mathrm{T}}v, x\rangle$ is positively homogeneous and (4) is feasible, it follows that

$$\begin{aligned} \varphi(y) &\leq \max_{v \in \mathbb{R}^m \setminus \{0\}} \varphi(v) \\ &= \max_{v \in \mathbb{R}^m \setminus \{0\}} \left\{-\tfrac{1}{2}\|v\|^2 + \|v\| \min_{x \in \Delta(\mathcal{K})} \left\langle A^{\mathrm{T}}\tfrac{v}{\|v\|}, x\right\rangle\right\} \\ &= \max_{t>0} \left\{-\tfrac{1}{2}t^2 + t\rho(A)\right\} \\ &= \tfrac{1}{2}\rho(A)^2. \end{aligned}$$

In addition, $\rho(A) = \min_{u \in \Delta(\mathcal{K})} \max_{\|y\|=1} \langle A^{\mathrm{T}}y, u\rangle \leq \max_{\|y\|=1} \langle A^{\mathrm{T}}y, x\rangle = \|Ax\|$ for any $x \in \Delta(\mathcal{K})$. □

References

1. Alizadeh, F., Goldfarb, D.: Second-order cone programming. Math. Program. **95**(1), 3–51 (2003)
2. Block, H.D.: The perceptron: a model for brain functioning. Rev. Mod. Phys. **34**, 123–135 (1962)

3. Blum, L., Cucker, F., Shub, M., Smale, S.: Complexity and Real Computation. Springer, New York (1998)
4. Burgisser, P., Cucker, F.: Condition. (Forthcoming).
5. Cheung, D., Cucker, F.: A new condition number for linear programming. Math. Program. **91**, 163–174 (2001)
6. Dantzig, G.B.: An ϵ-precise feasible solution to a linear program with a convexity constraint in $\frac{1}{\epsilon^2}$ iterations independent of problem size. Technical Report, Stanford University (1992)
7. Dunagan, J., Vempala, S.: A simple polynomial-time rescaling algorithm for solving linear programs. In: Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing, Chicago, ACM, pp. 315–320 (2004)
8. Dunagan, J., Vempala, S.: A simple polynomial-time rescaling algorithm for solving linear programs. Math. Program. **114**(1), 101–114 (2006)
9. Epelman, M., Freund, R.M.: Condition number complexity of an elementary algorithm for computing a reliable solution of a conic linear system. Math. Program. **88**, 451–485 (2000)
10. Freund, R., Vera, J.: Condition-based complexity of convex optimization in conic linear form via the ellipsoid algorithm. SIAM J. Optim. **10**, 155–176 (1999)
11. Nemirovski, A.: Prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with Lipschitz-continuous monotone operators and smooth convex-concave saddle point problems. SIAM J. Optim. **15**(1), 229–251 (2004)
12. Nesterov, Y.: A method for unconstrained convex minimization problem with rate of convergence $\mathcal{O}(1/k^2)$. Doklady AN SSSR (in russian). (English translation. Sov. Math. Dokl.) **269**, 543–547 (1983)
13. Nesterov, Y.: Excessive gap technique in nonsmooth convex minimization. SIAM J. Optim. **16**(1), 235–249 (2005)
14. Novikoff, A.B.J.: On convergence proofs on perceptrons. In: Proceedings of the Symposium on the Mathematical Theory of Automata, New York, vol. XII, pp. 615–622 (1962)
15. Renegar, J.: Incorporating condition measures into the complexity theory of linear programming. SIAM J. Optim. **5**, 506–524 (1995)
16. Renegar, J.: Linear programming, complexity theory and elementary functional analysis. Math. Program. **70**, 279–351 (1995)
17. Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain. Cornell Aeronautical Laboratory. Psychol. Rev. **65**(6), 386–408 (1958)
18. Soheili, N., Peña, J.: A smooth perceptron algorithm. SIAM J. Optim. **22**(2), 728-737 (2012)
19. Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization. SIAM J. Optim., University of Washington (2008, Unpublished Manuscript)

Selected Open Problems in Discrete Geometry and Optimization

Károly Bezdek, Antoine Deza, and Yinyu Ye

Abstract A list of questions and problems posed and discussed in September 2011 at the following consecutive events held at the Fields Institute, Toronto: Workshop on Discrete Geometry, Conference on Discrete Geometry and Optimization, and Workshop on Optimization. We hope these questions and problems will contribute to further stimulate the interaction between geometers and optimizers.

K. Bezdek
Department of Mathematics and Statistics, Center for Computational
and Discrete Geometry, University of Calgary, 612 Campus Place N.W.,
2500 University Drive NW, Calgary, AB, Canada T2N 1N4
e-mail: bezdek@math.ucalgary.ca

A. Deza (✉)
Advanced Optimization Laboratory, Department of Computing
and Software, McMaster University, 1280 Main Street West, Hamilton, ON, Canada L8S 4K1
e-mail: deza@mcmaster.ca

Y. Ye
Department of Management Science and Engineering,
Huang Engineering Center 308,
School of Engineering, Stanford University, Stanford,
475 Via Ortega, CA 94305, USA
e-mail: yinyu-ye@stanford.edu

1   Short Questions

1.1   Question by Imre Bárány, Endre Makai, Jr., Horst Martini, and Valeriu Soltan

The problem posed below can also be found in the article [27] (p. 469) written by the second and third of its posers.

Let $X \subset \mathbb{R}^d$. Following Klee, one calls $x', x'' \in X$ (where $x' \neq x''$) *antipodal* if there are different parallel supporting hyperplanes H', H'' of conv X such that $x' \in H'$ and $x'' \in H''$ (cf. [26], p. 420). Moreover, $X \subset \mathbb{R}^d$ is antipodal if for all $x', x'' \in X$ with $x' \neq x''$ we have that x' and x'' are antipodal. Answering a problem of Erdős and Klee, Danzer and Grünbaum [25] proved that for $X \subset \mathbb{R}^d$, if X is antipodal, then $|X| \leq 2^d$. This is sharp for the vertices of a parallelotope.

We pose a generalization of this theorem:

Question 1. *Suppose S is a set of segments in $\mathbb{R}^d$ such that for every $s', s'' \in S$ with $s' \neq s''$ there are different parallel supporting hyperplanes H', H'' of the convex hull conv$(\bigcup\{s : s \in S\})$, such that $s' \subset H'$ and $s'' \subset H''$. Then is it true that $|S| \leq 2^{d-1}$?*

Comments. If true, this would be sharp: an example would be the set of all edges of a parallelotope, parallel to a given edge.

Of course, more generally, we may consider a set S_k of k-simplices in $\mathbb{R}^d$, which satisfy the word-for-word analogue of the above property. Is it true, that then $|S_k| \leq 2^{d-k}$? If true, this would be sharp: an example would be simplices on all k-faces of a parallelotope, parallel to a given k-face. Here $1 \leq k \leq d - 2$. (Observe that for $k = d - 1$ the statement is evidently true.)

I. Talata, (Oral Communication, unpublished), proved the case $d = 3$ and $k = 1$.

1.2   Question by Károly Bezdek

A *plank* is a closed region of the d-dimensional Euclidean space $\mathbb{E}^d$ bounded by a pair of parallel hyperplanes. The *width* of a plank is the distance between its boundary hyperplanes.

Question 2. *Given a family of planks whose sum of widths is smaller than 2, what is the maximum volume of the part of the unit ball in $\mathbb{E}^d$ that can be covered by the planks?*

Comments. One might expect that the maximum volume in question is reached when the planks do not overlap and their union forms one plank concentric with the unit ball. Indeed, this is so in $\mathbb{E}^3$. For a somewhat stronger statement in $\mathbb{E}^3$ and its proof see Theorem 4.5.2 in [28]. The above question and the expected answer in two

dimensions are motivated by the well-known solution (attributed to A. Tarski, 1932) of a problem on covering the unit circle by a family of planks. (For more details we refer the interested reader to Chap. 4 of [28].)

Recall that $\mathbb{S}^d$ stands for the d-dimensional unit sphere in $(d+1)$-dimensional Euclidean space $\mathbb{E}^{d+1}, d \geq 2$. A *spherically convex body* is a closed, spherically convex subset $\mathbf{K}$ of $\mathbb{S}^d$ with interior points and lying in some closed hemisphere, thus, the intersection of $\mathbb{S}^d$ with a $(d+1)$-dimensional closed convex cone of $\mathbb{E}^{d+1}$ different from $\mathbb{E}^{d+1}$. The *inradius* $r(\mathbf{K})$ of $\mathbf{K}$ is the spherical radius of the largest spherical ball contained in $\mathbf{K}$.

Question 3. *Let the spherically convex bodies $\mathbf{K}_1, \ldots, \mathbf{K}_n$ cover the spherical ball $\mathbf{B}$ of radius $r(\mathbf{B}) < \frac{\pi}{2}$ in $\mathbb{S}^d, d \geq 2$. Then prove or disprove that $\sum_{i=1}^n r(\mathbf{K}_i) \geq r(\mathbf{B})$.*

Comments. R. Schneider and the author [29] have proved the following related result: If the spherically convex bodies $\mathbf{K}_1, \ldots, \mathbf{K}_n$ cover the spherical ball $\mathbf{B}$ of radius $r(\mathbf{B}) \geq \frac{\pi}{2}$ in $\mathbb{S}^d, d \geq 2$, then $\sum_{i=1}^n r(\mathbf{K}_i) \geq r(\mathbf{B})$. Furthermore, we note that the Euclidean analogue of the latter result has been proved by V. Kadets in [30] using an approach completely different from the one of [29].

1.3 Question by Peter Brass

Question 4. *Is it true that for each set of points in general position that results from an $\sqrt{n} \times \sqrt{n}$ grid in three-dimensional space by a small perturbation, every triangulation of the set consists of at most $O\left(n^{3/2}\right)$ simplices?*

Comments. Any set of n points in three-dimensional space that is in general position allows many different triangulations, and unlike in the two-dimensional situation, different numbers of simplices are possible. Any set of n points has a triangulation with $O(n)$ simplices, but it can have much larger triangulations, up to $O\left(n^2\right)$. But some point sets do not allow that large triangulations. I believe that for the perturbed-grid-square $O\left(n^{3/2}\right)$ is the maximum. For the perturbed-grid-cube I have a bound of $O\left(n^{5/3}\right)$, which is not sharp; the bound of $O\left(n^{3/2}\right)$ would be sharp for the perturbed-grid-squares.

1.4 Question by Antoine Deza

An arrangement $\mathscr{A}_{d,n}$ of n hyperplanes in dimension d is simple if any d hyperplanes intersect at a distinct point. The d-dimensional polyhedra defined by the hyperplanes of an arrangement $\mathscr{A}_{d,n}$ are called the cells of $\mathscr{A}_{d,n}$. The bounded facets of an unbounded cell are called external. Let $\Phi_{\mathscr{A}}(d,n)$ be the minimum number of external facets for any simple arrangement defined by n hyperplanes in dimension d.

Question 5. *We hypothesize that $\Phi_{\mathscr{A}}(d, n+1) \geq \Phi_{\mathscr{A}}(d, n) + \Phi_{\mathscr{A}}(d-1, n)$ for $n > d \geq 3$, and that the inequality is satisfied with equality for $d = 3$ and $n \geq 6$, i.e., $\Phi_{\mathscr{A}}(3, n) = n^2 - 3n + 4$ for $n \geq 6$.*

Comments. The hypothesized inequality holds for $n = d+1$ since $\Phi_{\mathscr{A}}(d, d+1) = d+1$ and $\Phi_{\mathscr{A}}(d, d+2) = d(d+1)$. For $d = 2$, we have $\Phi_{\mathscr{A}}(2, n) = 2(n-1)$ for $n \geq 4$ and, thus, $\Phi_{\mathscr{A}}(2, n+1) = \Phi_{\mathscr{A}}(2, n) + \Phi_{\mathscr{A}}(1, n)$ for $n \geq 4$ since $\Phi_{\mathscr{A}}(1, n) = 2$. The hypothesized inequality holds for all know values of $\Phi_{\mathscr{A}}(d, n)$ and is satisfied with equality for $(d, n) = (3, 6)$ and $(3, 7)$, see [31]. A strengthening of the lower bound of $\Phi_{\mathscr{A}}(3, n) \geq n(n-2)/3 + 2$ would improve the upper bound for the average diameter of a bounded cell of a simple arrangement of n hyperplanes in dimension 3. We refer to [32] for more details about the relation of the average diameter of a bounded cell of a simple arrangement of n hyperplanes in dimension d to $\Phi_{\mathscr{A}}(d, n)$, and to the Hirsch conjecture recently disproved by Santos [33].

1.5 Question by Gábor Fejes Tóth

The maximum volume of the intersection of a fixed ball in $\mathbb{E}^d$ and a variable simplex of given volume V is attained when the simplex is regular and concentric with the ball. This statement easily follows by Steiner symmetrization.

Question 6. *Show that the above statement holds true in spherical and hyperbolic space as well.*

Comments. Apart of the two-dimensional case the problem is open. If true, the statement has some important consequences. It implies that the simplex of maximum volume inscribed in a ball in S^d or H^d is regular, results proved by Böröczky [34] and Peyerimhoff [35], respectively. It also implies the conjecture that the simplex of minimum volume circumscribed a ball in S^d or H^d is regular. For the spherical case the statement implies the following: The volume of the part of S^d covered by $d + 2$ congruent balls attains its maximum if the centers of the balls lie in the vertices of a regular simplex.

1.6 Question by Włodzimierz Kuperberg

Question 7. *What is the minimum number $q(n)$ of cubes in $\mathbb{R}^n$ of edge length smaller than 1 whose union contains a unit cube?*

Comments. The smaller cubes in question are not assumed to be parallel (homothetic) to the covered unit cube, for in that case the corresponding minimum number would be exactly 2^n, since on one hand, a smaller homothetic cube contains at most one vertex of the unit cube, and on the other hand, 2^n smaller homothetic cubes suffice to cover the unit cube.

It is not difficult to prove that $q(2) = 3$ and that $q(n) \leq n + 1$ for every n, but the exact value of $q(n)$ has not yet been established for any $n \geq 3$.

1.7 Question by Jon Lee

Question 8. *For $n \geq 2$, the* Boolean Quadric Polytope $\mathscr{P}_n$ *is the convex hull in dimension $d = n(n + 1)/2$ of the 0/1 solutions to $x_i x_j = y_{ij}$ for all $i < j$ in $N := \{1, 2, \ldots, n\}$. Give a formula or good bounds for the d-dimensional volume of $\mathscr{P}_n$.*

Comments. The polytope $\mathscr{P}_n$ is contained in $\mathscr{Q}_n$, the solution set of the linear inequalities: $y_{ij} \leq x_i$, $y_{ij} \leq x_j$, $x_i + x_j \leq 1 + y_{ij}$, for all $i < j$ in N. In [36], we demonstrated that the d-dimensional volume of $\mathscr{Q}_n$ is $2^{2n-d} n!/(2n)!$. So this is an upper bound on the d-dimensional volume of $\mathscr{P}_n$. We would like to see a significant improvement in this upper bound and/or a non-trivial lower bound. There is quite a lot known about further linear inequalities satisfied by $\mathscr{P}_n$, so there are avenues to explore for trying to get a significant improvement in the upper bound.

1.8 Question by Horst Martini

Question 9. *Characterize geometrically those n-simplices in $\mathbb{E}^n$, $n \geq 3$, for which the incenter lies on the Euler line.*

Comments. It is well known that for any triangle T in $\mathbb{E}^2$ the *circumcenter C*, the *centroid S*, the *orthocenter O*, and the *center F of the nine-point circle* lie on one line—the *Euler line e* of T. It is also known that the *incenter I* of T lies on e if and only if T is isosceles, with e as axis of symmetry. No analogous characterization is known for n-simplices in $\mathbb{E}^n$ ($n \geq 3$) whose incenter I lies on their Euler line, which still is the affine hull of C and S; see the problem posed above. Only the following is known (see [37]): Let T be an n-dimensional *orthocentric* simplex ($n \geq 3$), i.e., the $n + 1$ altitudes of T still have a common point. Then C, S, and I of T are collinear if and only if T is *biregular*, which means: The vertex set of T can be partitioned into two disjoint subsets V_1, V_2 such that conv V_1, conv V_2 form regular simplices and all segments $[x, y]$, $x \in V_1$, $y \in V_2$, are of equal length. This directly generalizes the planar result and supports somewhat the "philosophy" that orthocentric n-simplices ($n \geq 3$) are the "true" higher dimensional analogues of triangles.

1.9 Question by Benjamin Matschke

Conjecture 1. (*A Multicolored Carathéodory Conjecture*).   Let $r \geq 2$ and $N \geq 1$ be integers. N can be assumed to be very large, that is, $N \geq N_0(r)$ for some $N_0(r)$. Suppose we are given $r(N + 1)$ points P_{ij} in $\mathbb{R}^N$ that are indexed by $1 \leq i \leq r$ and $1 \leq j \leq N + 1$. Assume that $0 \in \text{conv}\{P_{1j}, \ldots, P_{rj}\}$ for all $1 \leq j \leq N + 1$. Assume further that the index set $\{1, 2, \ldots, N + 1\}$ is partitioned as $C_1 \uplus \ldots \uplus C_m$ such that all color classes are small: $|C_k| \leq r - 1$ for all $1 \leq k \leq m$. Then there exist $k_1, \ldots, k_{N+1} \in \{1, \ldots, r\}$ such that $0 \in \text{conv}\{P_{k_1,1}, \ldots, P_{k_{N+1},N+1}\}$ and for any two distinct a, b in the same color class C_k we have $k_a \neq k_b$.

Comments. This is an—admittedly technical—multicolored version of Bárány's colored Carathéodory theorem (1982). If true this conjecture implies the new colored Tverberg theorem by Blagojević, Ziegler and me (2009), *also for non-primes r*. Hence, the conjecture is particularly interesting when r is not a prime and $r - 1$ divides N. The first interesting case is $r = 4$ and $N = 9$, which, if true, would imply the new colored Tverberg theorem in the smallest open non-prime case $r = 4$ and $d = 2$.

1.10 Question by Valeriu Soltan

Conjecture 2. If $K \subset \mathbb{R}^n$ is a compact convex set and $n_1, \ldots, n_s$ are positive integers with $n_1 + \cdots + n_s = n + 1$, then, for every point $z \in K$, non-empty faces $F_1, \ldots, F_s$ of K exist such that

$$z \in \text{conv}\,(F_1 \cup \cdots \cup F_s)$$

and

$$\dim F_i \leq n_i - 1 \ \text{ for all } \ i = 1, \ldots, s.$$

Comments. For convex polytopes K the conjecture holds true.

2   Comprehensive Research Problems

2.1 The Contact Number Problem of Unit Sphere Packings by Károly Bezdek

Let **B** be a ball in the d-dimensional Euclidean space $\mathbb{E}^d$. Then the contact graph of an arbitrary finite packing by non-overlapping translates of **B** in $\mathbb{E}^d$ is the (simple) graph whose vertices correspond to the packing elements and whose two vertices are

connected by an edge if and only if the corresponding two packing elements touch each other. One of the most basic questions on contact graphs is to find the maximum number of edges that a contact graph of n non-overlapping translates of the given Euclidean ball **B** can have in $\mathbb{E}^d$. Harborth [44] proved the following remarkable result on the contact graphs of congruent circular disk packings in $\mathbb{E}^2$. The maximum number of touching pairs in a packing of n congruent circular disks in $\mathbb{E}^2$ is precisely $\lfloor 3n - \sqrt{12n-3} \rfloor$. The analogue question in the hyperbolic plane has been studied by Bowen in [42]. We prefer to quote his result in the following geometric way: Consider circle packings in the hyperbolic plane, by finitely many congruent circles, which maximize the number of touching pairs for the given number of congruent circles. Then such a packing must have all of its centers located on the vertices of a triangulation of the hyperbolic plane by congruent equilateral triangles, provided the diameter D of the circles is such that an equilateral triangle in the hyperbolic plane of side length D has each of its angles equal to $\frac{2\pi}{N}$ for some $N > 6$.

Now, we are ready to phrase the *Contact Number Problem* of finite congruent sphere packings in $\mathbb{E}^3$. For a given positive integer $n \geq 2$ find the largest number $C(n)$ of touching pairs in a packing of n congruent balls in $\mathbb{E}^3$. One can regard this problem as a combinatorial relative of the long-standing Kepler conjecture on the densest unit sphere packings in $\mathbb{E}^3$, which has been recently proved by Hales [43]. It is natural to continue with the following question.

Problem 1. Find those positive integers n for which $C(n)$ can be achieved in a packing of n unit balls in $\mathbb{E}^3$ consisting of parallel layers of unit balls each being a subset of the densest infinite hexagonal layer of unit balls.

Harborth's result [44] implies in a straightforward way that if the maximum number of touching pairs in packings of n congruent circular disks in $\mathbb{E}^2$ is denoted by $c(n)$, then

$$\lim_{n \to +\infty} \frac{3n - c(n)}{\sqrt{n}} = \sqrt{12} = 3.464\ldots.$$

The author [39] has proved the following estimates in higher dimensions. The number of touching pairs in an arbitrary packing of $n > 1$ unit balls in $\mathbb{E}^d$, $d \geq 3$ is less than

$$\frac{1}{2}\tau_d\, n - \frac{1}{2^d}\delta_d^{-\frac{d-1}{d}}\, n^{\frac{d-1}{d}},$$

where τ_d stands for the kissing number of a unit ball in $\mathbb{E}^d$ (i.e., it denotes the maximum number of non-overlapping unit balls of $\mathbb{E}^d$ that can touch a given unit ball in $\mathbb{E}^d$) and δ_d denotes the largest possible density for (infinite) packings of unit balls in $\mathbb{E}^d$. Now, recall that on the one hand, according to the well-known theorem of Kabatiansky and Levenshtein [47] $\tau_d \leq 2^{0.401d(1+o(1))}$ and $\delta_d \leq 2^{-0.599d(1+o(1))}$ as $d \to +\infty$ on the other hand, $\tau_3 = 12$ (for the first complete proof see [48]) moreover, according to the recent breakthrough result of Hales [43] $\delta_3 = \frac{\pi}{\sqrt{18}}$. Thus, by combining the above results together we get that the number of touching pairs in an arbitrary packing of $n > 1$ unit balls in $\mathbb{E}^d$ is less than

$$\frac{1}{2}2^{0.401d(1+o(1))}\, n - \frac{1}{2}2^{-0.401(d-1)(1-o(1))}\, n^{\frac{d-1}{d}}$$

as $d \to +\infty$ and in particular, it is less than

$$6n - \frac{1}{8}\left(\frac{\pi}{\sqrt{18}}\right)^{-\frac{2}{3}} n^{\frac{2}{3}} = 6n - 0.152\ldots n^{\frac{2}{3}}$$

for $d = 3$. Next we report on a recent improvement on the latter estimate. In order, to state that theorem in a proper form we need to introduce a bit of additional terminology. If $\mathscr{P}$ is a packing of n unit balls in $\mathbb{E}^3$, then let $C(\mathscr{P})$ stand for the number of touching pairs in $\mathscr{P}$, that is, let $C(\mathscr{P})$ denote the number of edges of the contact graph of $\mathscr{P}$ and call it the *contact number* of $\mathscr{P}$. Moreover, let $C(n)$ be the largest $C(\mathscr{P})$ for packings $\mathscr{P}$ of n unit balls in $\mathbb{E}^3$. Finally, let us imagine that we generate packings of n unit balls in $\mathbb{E}^3$ in such a special way that each and every center of the n unit balls chosen, is a lattice point of the face-centered cubic lattice Λ_{fcc} with shortest non-zero lattice vector of length 2. Then let $C_{fcc}(n)$ denote the largest possible contact number of all packings of n unit balls obtained in this way. Before stating our main theorem we make the following comments. First, recall that according to [43] the lattice unit sphere packing generated by Λ_{fcc} gives the largest possible density for unit ball packings in $\mathbb{E}^3$, namely $\frac{\pi}{\sqrt{18}}$ with each ball touched by 12 others such that their centers form the vertices of a cuboctahedron. Second, it is easy to see that $C_{fcc}(2) = C(2) = 1, C_{fcc}(3) = C(3) = 3, C_{fcc}(4) = C(4) = 6$. Third, it is natural to conjecture that $C_{fcc}(9) = C(9) = 21$. Based on the trivial inequalities $C(n+1) \geq C(n)+3, C_{fcc}(n+1) \geq C_{fcc}(n)+3$ valid for all $n \geq 2$, it would follow that $C_{fcc}(5) = C(5) = 9, C_{fcc}(6) = C(6) = 12, C_{fcc}(7) = C(7) = 15$, and $C_{fcc}(8) = C(8) = 18$. In general, clearly $C(n) \geq C_{fcc}(n) \geq 3n - 6$. Furthermore, we note that $C(10) \geq 25, C(11) \geq 29$, and $C(12) \geq 33$. In order, to see that one should take the union $\mathbf{U}$ of two regular octahedra of edge length 2 in $\mathbb{E}^3$ such that they share a regular triangle face T in common and lie on opposite sides of it. If we take the unit balls centered at the nine vertices of $\mathbf{U}$, then there are exactly 21 touching pairs among them. Also, we note that along each side of T the dihedral angle of $\mathbf{U}$ is concave and in fact, it can be completed to 2π by adding twice the dihedral angle of a regular tetrahedron in $\mathbb{E}^3$. This means that along each side of T two triangular faces of $\mathbf{U}$ meet such that for their four vertices there exists precisely one point in $\mathbb{E}^3$ lying outside $\mathbf{U}$ and at distance 2 from each of the four vertices. Finally, if we take the 12 vertices of a cuboctahedron of edge length 2 in $\mathbb{E}^3$ along with its center of symmetry, then the 13 unit balls centered about them have 36 contacts implying that $C(13) \geq 36$. Whether in any of the inequalities $C(10) \geq 25, C(11) \geq 29, C(12) \geq 33$, and $C(13) \geq 36$ we have equality seems to be an open question. In connection with this problem we call the reader's attention to the very recent and highly elegant article of Hayes [45]. It gives an overview of the computational methods presented in the papers [38] and [46] that are based on exhaustive enumeration and elementary geometry. The main

results are: $C(9) = 21, C(10) = 25$ [38] and $C(11) = 29$ [46]. However, the status of the mathematical rigour of the approaches of [38] as well as [46] remains to be seen. For $C(n)$ in general, when n is an arbitrary positive integer, we have the following estimates proved in [40] and [41].

Theorem 1.

(i) $C(n) < 6n - 0.926n^{\frac{2}{3}}$ for all $n \geq 2$.

(ii) $C_{fcc}(n) < 6n - \frac{3\sqrt[3]{18\pi}}{\pi}n^{\frac{2}{3}} = 6n - 3.665\ldots n^{\frac{2}{3}}$ for all $n \geq 2$.

(iii) $6n - \sqrt[3]{486}n^{\frac{2}{3}} < C_{fcc}(n) \leq C(n)$ for all $n = \frac{k(2k^2+1)}{3}$ with $k \geq 2$.

As an immediate result we get

Corollary 1.

$$0.926 < \frac{6n - C(n)}{n^{\frac{2}{3}}} < \sqrt[3]{486} = 7.862\ldots$$

for all $n = \frac{k(2k^2+1)}{3}$ with $k \geq 2$.

The latter claim leads us to the following rather basic question.

Problem 2. Does the limit $\lim_{n \to +\infty} \frac{6n - C(n)}{n^{\frac{2}{3}}}$ exist?

The following was noted in [39]. Due to the Minkowski difference body method the family $\mathscr{P}_{\mathbf{K}} := \{\mathbf{t}_1 + \mathbf{K}, \mathbf{t}_2 + \mathbf{K}, \ldots, \mathbf{t}_n + \mathbf{K}\}$ of n translates of the convex body $\mathbf{K}$ in $\mathbb{E}^d$ is a packing if and only if the family $\mathscr{P}_{\mathbf{K_o}} := \{\mathbf{t}_1 + \mathbf{K_o}, \mathbf{t}_2 + \mathbf{K_o}, \ldots, \mathbf{t}_n + \mathbf{K_o}\}$ of n translates of the symmetric difference body $\mathbf{K_o} := \frac{1}{2}(\mathbf{K} + (-\mathbf{K}))$ of $\mathbf{K}$ is a packing in $\mathbb{E}^d$. Moreover, the number of touching pairs in the packing $\mathscr{P}_{\mathbf{K}}$ is equal to the number of touching pairs in the packing $\mathscr{P}_{\mathbf{K_o}}$. Thus, for this reason and for the reason that if $\mathbf{K}$ is a convex body of constant width in $\mathbb{E}^d$, then $\mathbf{K_o}$ is a ball of $\mathbb{E}^d$, Theorem 1 extends in a straightforward way to translative packings of convex bodies of constant width in $\mathbb{E}^3$.

2.2 On Gram and Euclidean Graph Realizations by Monique Laurent and Antonios Varvitsiotis

We present two open problems about the graph parameters $ed(G)$, $gd(G)$ and $\nu^=(G)$, which deal with some geometric realizations of graphs.

Problem 3. Determine the validity of the inequality:

$$ed(\nabla G) \leq ed(G) + 1, \tag{1}$$

relating the Euclidean dimension of a graph G and of its suspension ∇G.

Comments: Given a graph $G = ([n], E)$, its *Euclidean dimension* is the graph parameter $ed(G)$ which is defined as the smallest integer $k \geq 1$ such that, for every

family of vectors $p_1, \ldots, p_n$, there exists another family of vectors $q_1, \ldots, q_n \in \mathbb{R}^k$ satisfying

$$\|p_i - p_j\|_2 = \|q_i - q_j\|_2, \ \forall \{i, j\} \in E.$$

The suspension graph ∇G is obtained from G by adding to it a new node and making it adjacent to all the nodes of G.

The parameter $\mathrm{ed}(G)$ was studied in [49] where it is shown that for any fixed $k \geq 1$, the class of graphs satisfying $\mathrm{ed}(G) \leq k$ is closed under the operation of taking minors. That is, the Euclidean dimension does not increase if one deletes or contracts an edge e in G: $\mathrm{ed}(G \setminus e), \mathrm{ed}(G/e) \leq \mathrm{ed}(G)$. Then, the Graph Minor Theorem of Robertson and Seymour implies that, for any fixed $k \geq 1$, there exists a finite family of graphs $G_1, \ldots G_{t_k}$ having the property that $\mathrm{ed}(G) \leq k$ if and only if G does not have any minor isomorphic to any of $G_1, \ldots, G_{t_k}$. In other words, the graph property $\mathrm{ed}(G) \leq k$ can be characterized by finitely many minimal forbidden minors. In [49, 50] the full list of minimal forbidden minors is identified for $k \in \{1, 2, 3\}$. Specifically, K_{k+2} is the only minimal forbidden minor when $k \in \{1, 2\}$ and, for $\mathrm{ed}(G) \leq 3$, there are two minimal forbidden minors: K_5 and the octahedral graph $K_{2,2,2}$.

The following inequality is shown in [54], relating the Euclidean dimension of a graph and of its suspension:

$$\mathrm{ed}(\nabla G) \geq \mathrm{ed}(G) + 1. \tag{2}$$

Thus our first problem asks whether the converse inequality holds or, equivalently, whether it is true that

$$\mathrm{ed}(\nabla G) = \mathrm{ed}(G) + 1. \tag{3}$$

By combining results from [49] and [54] it follows that the answer is positive when $\mathrm{ed}(G) \leq 3$, i.e., when G is K_5 and $K_{2,2,2}$-minor free.

In a similar manner, the *Gram dimension* $\mathrm{gd}(G)$ is defined as the smallest integer $k \geq 1$ such that, for every family of vectors $p_1, \ldots, p_n$, there exists another family of vectors $q_1, \ldots, q_n \in \mathbb{R}^k$ satisfying

$$\|p_i\|_2 = \|q_i\|_2, \ \forall i \in [n], \ \text{ and } \ p_i^\mathsf{T} p_j = q_i^\mathsf{T} q_j, \ \forall \{i, j\} \in E.$$

This parameter was introduced in [53, 54] and its study is motivated by its connection with the low rank positive semidefinite matrix completion problem.

In [53, 54] it is shown that, for any fixed $k \geq 1$, the class of graphs satisfying $\mathrm{gd}(G) \leq k$ is closed under taking minors. Moreover, it is shown that K_{k+1} is the only minimal forbidden minor for $k \in \{1, 2, 3\}$ and that K_5 and $K_{2,2,2}$ are the only minimal forbidden minors for the graph property $\mathrm{gd}(G) \leq 4$. We also show the following equality, which relates the Gram dimension of a graph to the Euclidean dimension of its suspension:

$$\mathrm{gd}(G) = \mathrm{ed}(\nabla G). \tag{4}$$

Combining with (2), we obtain that $\text{gd}(G) \geq \text{ed}(G) + 1$ for any graph G. Therefore, Problem 1 is equivalent to the validity of the following equality:

$$\text{gd}(G) = \text{ed}(G) + 1. \qquad (5)$$

Problem 4. Determine the validity of the inequality

$$\text{gd}(G) \leq \nu^=(G), \qquad (6)$$

relating the Gram dimension $\text{gd}(G)$ and the van der Holst parameter $\nu^=(G)$.

Comments: Let $\mathscr{S}_+^n$ denote the cone of $n \times n$ positive semidefinite matrices. Given a graph $G = ([n], E)$ consider the cone

$$\mathscr{C}(G) = \{M \in \mathscr{S}_+^n : M_{ij} = 0 \text{ for } \{i, j\} \notin E \text{ and } i \neq j\}.$$

The parameter $\nu^=(G)$ is defined as the maximum corank of a matrix $M \in \mathscr{C}(G)$ satisfying the following nondegeneracy property:

$$\forall X \in \mathscr{S}^n \quad MX = 0, \ X_{ii} = 0 \ \forall i \in V, \ X_{ij} = 0 \ \forall \{i, j\} \in E \implies X = 0,$$

known as the *Strong Arnold Property*. This graph parameter was introduced in [52] and its study is motivated by its relation to the celebrated graph invariant $\mu(G)$ of Colin de Verdière [51].

In [52] is shown that, for any fixed $k \geq 1$, the class of graphs with $\nu^=(G) \leq k$ is closed under taking minors. Additionally, the full list of minimal forbidden minors was determined for $k \in \{1, 2, 3, 4\}$. Surprisingly, it turns out that the forbidden minors for the property $\nu^=(G) \leq k$ coincide with the forbidden minors for the property $\text{gd}(G) \leq k$, for each $k \in \{1, 2, 3, 4\}$.

This observation prompted the investigation of possible links between these two parameters. A first result in this direction was established in [53, 54] where it was shown that, for any graph G,

$$\text{gd}(G) \geq \nu^=(G). \qquad (7)$$

Our second problem asks for the validity of the converse inequality. In other words, is it true that the two graph parameters $\text{gd}(\cdot)$ and $\nu^=(\cdot)$ coincide? We know that the answer is positive, e.g., for the graphs with Gram dimension at most 4, and for chordal graphs.

2.3 Non-convex Optimization Approaches to Network Localization by Anthony Man-Cho So and Yinyu Ye

Determining the positions of a set of n points in Euclidean space based on knowledge of a subset of the $\binom{n}{2}$ pairwise distances is a fundamental geometric problem with numerous applications. For instance, in location-aware networks—which support a host of services such as emergency response [14], mobile advertising [18], and target tracking [23]—wireless nodes that are deployed in an area of interest must be able to localize themselves using distance measurements obtained from direct communications with their neighbors. Another example can be found in biochemistry, where the positions of atoms in a molecule—which provide important information about the properties and functions of the molecule—are typically determined from a set of geometric constraints that include a subset of the inter-atomic distances [6]. As the above examples suggest, in many applications of the localization problem, it is only meaningful to localize the points in an Euclidean space *of given dimension*, say in $\mathbb{R}^2$ or $\mathbb{R}^3$. Unfortunately, such a fixed-dimensional localization problem is intractable in general [17]. In fact, as shown in Biswas and Ye [5], the d-dimensional localization problem can be formulated as a rank-constrained semidefinite program (SDP), namely,

$$
\begin{aligned}
\text{find} \quad & Z \in \mathbb{R}^{n \times n} \\
\text{such that} \quad & \mathscr{E}(Z) = u, \\
& Z \succeq \mathbf{0}, \ \mathrm{rank}(Z) \leq d.
\end{aligned}
\tag{8}
$$

Here, the linear operator $Z \mapsto \mathscr{E}(Z) = (\mathrm{tr}(E_1 Z), \dots, \mathrm{tr}(E_m Z)) \in \mathbb{R}^m$ and vector $u \in \mathbb{R}^m$ are determined by the available distance measurements, $d \geq 1$ is the target dimension in which the input instance should be localized, and $Z \succeq \mathbf{0}$ means that Z is a symmetric positive semidefinite matrix. On the other hand, by dropping the non-convex constraint $\mathrm{rank}(Z) \leq d$, one immediately obtains an SDP relaxation of the fixed-dimensional localization problem. Such a relaxation and its variants have been extensively studied in recent years (see, e.g., [3,4,7,9,10,12,16,19,21,22,24]) and are very natural as far as polynomial-time solvability is concerned. Moreover, they have the added advantage that in many cases, localization accuracy guarantees can be established; see, e.g., [10,20–22,24]. However, standard interior-point algorithms for solving SDPs will always return the solution with the highest rank [21], which means that they are unlikely to deliver a feasible solution to the rank-constrained problem (8) in general. Thus, it is interesting to ask whether there are other efficient approaches for finding low-rank solutions to the SDP relaxation of (8).

In a recent work, Ji et al. [11] depart from the convex relaxation paradigm and develop a non-convex optimization approach for tackling Problem (8). Such an approach is motivated by ideas from low-rank matrix recovery—a topic that has received significant interest recently; see, e.g., the website [15] and the references therein. Specifically, for a given $p \in (0, 1]$, consider the following regularized version of Problem (8):

$$\Gamma^* = \text{minimize} \quad f_p(Z) = \sum_{i=1}^{n} \sigma_i(Z)^p$$
$$\text{subject to} \quad \mathscr{E}(Z) = u,$$
$$Z \succeq \mathbf{0}. \tag{9}$$

Here, $\sigma_i(Z)$ is the i-th singular value of Z. The value $(f_p(Z))^{1/p}$ is known as the *Schatten p-quasi-norm* of Z, and it is easy to verify that $f_1(Z) = \text{tr}(Z)$ and $f_p(Z) \to \text{rank}(Z)$ as $p \searrow 0$ for all $Z \succeq \mathbf{0}$. This suggests that the Schatten quasi-norms can be effective in finding a low-rank solution to Problem (9), especially when p is small. However, a fundamental challenge associated with Problem (9) is that the function $Z \mapsto f_p(Z)$ is non-convex when $p \in (0, 1)$. Indeed, the problem of minimizing the Schatten p-quasi-norm over a system of linear matrix inequalities is NP-hard for any fixed $p \in (0, 1)$; cf. [8]. To circumvent this difficulty, Ji et al. [11] design a potential reduction algorithm and show that it can approximate a first-order critical point of Problem (9) to any given accuracy in polynomial time. In other words, given an accuracy level $\epsilon > 0$, the algorithm will return a solution $\bar{Z}$ in polynomial time that is feasible for (9) and satisfies one of the following conditions:

(a) $\bar{Z}$ is an ϵ-*optimal solution*, i.e., $f_p(\bar{Z}) \leq \epsilon$.
(b) $\bar{Z}$ is an ϵ-*first-order critical point*, i.e., there exists a multiplier $\bar{y} \in \mathbb{R}^m$ such that

$$p\Lambda^{p-1} - \sum_{i=1}^{m} \bar{y}_i \left(U^T E_i U\right) \succeq \mathbf{0}$$

and

$$0 \leq \frac{\text{tr}\left(p\bar{Z}^p - \sum_{i=1}^{m} \bar{y}_i E_i \bar{Z}\right)}{f_p(\bar{Z})} \leq \epsilon,$$

where $\bar{Z} = U\Lambda U^T$ is the spectral decomposition of $\bar{Z}$ with $U \in \mathbb{R}^{n \times r}$, $\Lambda = \text{Diag}(\lambda_1, \ldots, \lambda_r) \in \mathbb{R}^{r \times r}$ and $r = \text{rank}(\bar{Z})$, and $\bar{Z}^p = U\Lambda^p U^T = U \text{Diag}(\lambda_1^p, \ldots, \lambda_r^p) U^T$.

Moreover, it is shown in [11] that if the input instance is universally rigid,[1] then the potential reduction algorithm can localize it in the required dimension, *even though the algorithm may only return a first-order critical point*. This indicates that the localizability guarantee of the potential reduction algorithm is at least as strong as that of the SDP relaxations in [4,5]. Computationally, it is observed that the potential reduction algorithm can localize some of the globally rigid[2] but not universally rigid

---

[1]A localization instance is said to be universally rigid if it has a unique (up to congruences) localization in *any* Euclidean space.

[2]A localization instance is said to be globally rigid in $\mathbb{R}^d$ if it has a unique (up to congruences) localization in $\mathbb{R}^d$.

input instances in the required dimension [11]. It is worth noting that by an earlier result of So and Ye [21], the SDP relaxation of Biswas and Ye [5] will necessarily fail to localize such instances in the required dimension. This phenomenon strongly motivates a deeper investigation of the approach proposed in [11].

Problem 5. In view of the above developments, it is clear that there are many directions for further investigation. One of the most immediate questions is to understand the power of the non-convex Schatten quasi-norm regularization in the context of localization. Specifically, can we characterize the class of input instances that can be localized in the required dimension by the potential reduction algorithm of Ji et al.? From the results in [11], it is clear that this class will be larger than that of universally rigid instances. However, it will certainly be smaller than that of globally rigid instances, since the problem of localizing an arbitrary globally rigid instance in the required dimension is intractable [2]. This also suggests that some new rigidity-theoretic notions may be waiting to be discovered.

Along the same direction, it will be interesting to study the rigidity-theoretic implications of Schatten quasi-norm regularization. A starting point could be to understand the rigidity-theoretic interpretations of the dual vector $\bar{y}$ in the definition of the first-order critical point. This is motivated by an earlier result of So and Ye [20], which states that each dual variable in the SDP relaxation of Biswas and Ye [5] corresponds to a stress on an edge of the input graph, and the optimality conditions of the SDP correspond to a certain equilibrium condition on the input graph. The work [20] has since motivated or been used to develop other rigidity-theoretic results (see, e.g., [1, 13]), and a natural question would be whether these results have counterparts in the Schatten quasi-norm regularization setting.

References

1. Alfakih, A.Y.: On bar frameworks, stress matrices and semidefinite programming. Math. Program. B **129**(1), 113–128 (2011)
2. Aspnes, J., Goldenberg, D., Yang, Y.R.: On the computational complexity of sensor network localization. In: Nikoletseas S., Rolim J.D.P. (eds.) Proceedings of the 1st International Workshop on Algorithmic Aspects of Wireless Sensor Networks (ALGOSENSORS 2004), Turku. Lecture Notes in Computer Science, vol. 3121, pp. 32–44. Springer (2004)
3. Biswas, P., Lian, T.-C., Wang, T.-C., Ye, Y.: Semidefinite programming based algorithms for sensor network localization. ACM Trans. Sens. Netw. **2**(2), 188–220 (2006)
4. Biswas, P., Toh, K.-C., Ye, Y.: A distributed SDP approach for large–scale noisy anchor–free graph realization with applications to molecular conformation. SIAM J. Sci. Comput. **30**(3), 1251–1277 (2008)
5. Biswas, P., Ye, Y.: Semidefinite programming for ad hoc wireless sensor network localization. In: Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks (IPSN 2004), Berkeley, pp. 46–54 (2004)
6. Crippen, G.M., Havel, T.F.: Distance Geometry and Molecular Conformation. Chemometrics Series, vol. 15. Research Studies Press, Taunton (1988)
7. Ding, Y., Krislock, N., Qian, J., Wolkowicz, H.: Sensor network localization, Euclidean matrix completions, and graph realization. Optim. Eng. **11**(1), 45–66 (2010)

8. Ge, D., Jiang, X., Ye, Y.: A note on the complexity of L_p minimization. Math. Program. B **129**(2), 285–299 (2011)

9. Gouveia, J., Pong, T.K.: Comparing SOS and SDP relaxations of sensor network localization. Comput. Optim. Appl. **52**(3), 609–627 (2012)

10. Javanmard, A., Montanari, A.: Localization from incomplete noisy distance measurements. Found. Comput. Math. (2012, Accepted for publication). doi: 10.1007/s10208-012-9129-5

11. Ji, S., Sze, K.-F., Zhou, Z., So, A.M.-C., Ye, Y.: Beyond convex relaxation: A polynomial–time non–convex optimization approach to network localization. In: To Appear in the Proceedings of the 32nd IEEE International Conference on Computer Communications (INFOCOM 2013), Torino (2013)

12. Kim, S., Kojima, M., Waki, H.: Exploiting sparsity in SDP relaxation for sensor network localization. SIAM J. Optim. **20**(1), 192–215 (2009)

13. Laurent, M., Varvitsiotis, A.: The Gram dimension of a graph. In: Mahjoub A.R., Markakis V., Milis I., Paschos V.T. (eds.) Proceedings of the 2nd International Symposium on Combinatorial Optimization (ISCO 2012), Athens. Lecture Notes in Computer Science, vol. 7422, pp. 356–367. Springer (2012)

14. Lorincz, K., Malan, D.J., Fulford-Jones, T.R.F., Nawoj, A., Clavel, A., Shnayder, V., Mainland, G., Welsh, M., Moulton., S.: Sensor networks for emergency response: Challenges and opportunities. IEEE Pervasive Comput. **3**(4), 16–23 (2004)

15. Low–Rank Matrix Recovery and Completion via Convex Optimization. http://perception.csl.illinois.edu/matrix-rank/references.html.

16. Pong, T.K., Tseng, P.: (Robust) edge–based semidefinite programming relaxation of sensor network localization. Math. Program. A **130**(2), 321–358 (2011)

17. Saxe, J.B.: Embeddability of weighted graphs in k–space is strongly NP–hard. In: Proceedings of the 17th Allerton Conference in Communication, Control, and Computing, Monticello, Illinois, pp. 480–489 (1979)

18. Sayed, A.H., Tarighat, A., Khajehnouri, N.: Network–based wireless location: Challenges faced in developing techniques for accurate wireless location information. IEEE Signal Process. Mag. **22**(4), 24–40 (2005)

19. Shames, I., Anderson, B.D.O., Fidan, B.: On the use of convex optimization in sensor network localization and synchronization. In: Proceedings of the 1st IFAC Workshop on Estimation and Control of Networked Systems, Don Orione Artigianelli, Italy, pp. 228–233 (2009)

20. So, A.M.-C., Ye, Y.: A semidefinite programming approach to tensegrity theory and realizability of graphs. In: Proceedings of the 17th Annual ACM–SIAM Symposium on Discrete Algorithms (SODA 2006), Miami, pp. 766–775 (2006)

21. So, A.M.-C., Ye, Y.: Theory of semidefinite programming for sensor network localization. Math. Program. B **109**(2), 367–384 (2007)

22. Sugihara, R., Gupta, R.K.: Sensor localization with deterministic accuracy guarantee. In: Proceedings of the 30th IEEE International Conference on Computer Communications (INFOCOM 2011), Shanghai, pp. 1772–1780 (2011)

23. Zhou, K., Roumeliotis, S.I.: Multirobot active target tracking with combinations of relative observations. IEEE Trans. Robot. **27**(4), 678–695 (2011)

24. Zhu, Z., So, A.M.-C., Ye, Y.: Universal rigidity and edge sparsification for sensor network localization. SIAM J. Optim. **20**(6), 3059–3081 (2010)

25. Danzer, L., Grünbaum, B.: Über zwei Probleme bezüglich konvexer Körper von P. Erdős und V. L. Klee, Math. Z. **79**, 95–99 (1962)

26. Grünbaum, B.: Convex Polytopes. Wiley-Interscience, London (1967)

27. Makai, E., Jr., H.Martini, On the number of antipodal or strictly antipodal pairs of points in finite subsets of $\mathbb{R}^d$. In: Gritzmann, P., Sturmfels B. (eds.) Appled Geometry and Discrete Mathematics, The V. Klee Festschrift. DIMACS Series in Discrete Mathematics and Theoretical Computer Sci., Vol. **4**, American Mathematical Society, Providence, RI, pp. 457–470 (1991)

28. Bezdek, K.: Classical Topics in Discrete Geometry. CMS Books in Mathematics, Springer, New York (2010)

29. Bezdek, K., Schneider, R.: Covering large balls with convex sets in spherical space. Beiträge Algebra Geom. **51**(1), 229–235 (2010)
30. Kadets, V.: Coverings by convex bodies and inscribed balls. Proc. Amer. Math. Soc. **133**(5), 1491–1495 (2005)
31. Deza, A., Moriyama, S., Miyata, H., Xie, F. : Hyperplane arrangements with large average diameter: A computational approach. Adv. Stud. Pure Math. **62**, 59–74 (2012)
32. Deza, A., Xie, F.: Hyperplane arrangements with large average diameter, American Mathematical Society series and Centre de Recherches Mathématiques **48**, 103–114 (2009)
33. Santos, F.: A counterexample to the hirsch conjecture. Ann. Math. **176**(1), 383–412 (2012)
34. Böröczky, K.: On an extremum property of the regular simplex in S^d, Intuitive Geometry (Siófok, 1985), 117–121, Colloq. Math. Soc. János Bolyai, **48**, North-Holland, Amsterdam, 1987; MR0910705.
35. Peyerimhoff, N.: Simplices of maximal volume or minimal total edge length in hyperbolic space. J. London Math. Soc. (2) **66**(3), 753–768 (2002). MR1934304
36. Ko, C-W., Lee, J., Steingrímsson, E.: The volume of relaxed Boolean-quadric and cut polytopes. Discret. Math. **163**(1–3), 293–298 (1997)
37. Edmonds, A.E., Hadjja, M., Martini, H.: Ortocentric simplices and biregularity. Results Math. **52**, 41–50 (2008)
38. Arkus, N., Manoharan, V.N., Brenner, M.P.: Deriving finite sphere packings. SIAM J. Discret. Math. **25**(4), 1860–1901 (2011)
39. Bezdek, K.: On the maximum number of touching pairs in a finite packing of translates of a convex body. J. Comb. Theory A **98**, 192–200 (2002)
40. Bezdek, K.: Contact numbers for congruent sphere packings in Euclidean 3-space. Discret. Comput. Geom. **48**(2), 298–309 (2012)
41. Bezdek, K., Reid, S.: On touching pairs, triplets, and quadruples in packings of congruent spheres, arXiv:1210.5756v1 [math.MG]1–19(2012)
42. Bowen, L.: Circle packing in the hyperbolic plane. Math. Phys. Electron. J. **6**, 1–10 (2000)
43. Hales, T. C.: A proof of the Kepler conjecture. Ann. Math. **162**(2–3), 1065–1185 (2005)
44. Harborth, H.: Lösung zu Problem 664A. Elem. Math. **29**, 14–15 (1974)
45. Hayes, B.: The science of sticky spheres. Am. Sci. **100**, 442–449 (2012)
46. Hoy, R.S., Harwayne-Gidansky, J., O'Hern, C.S.: Structure of finite sphere packings via exact enumeration: Implications for colloidal crystal nucleation. Phys. Rev. E **85** (2012). 051403
47. Kabatiansky, G.A., Levenshtein, V.I.: Bounds for packings on a sphere and in space. Problemy Peredachi Informatsii **14**, 3–25 (1978)
48. Schütte, K., van der Waerden, B.L.: Das Problem der dreizehn Kugeln. Math. Ann. **125**, 253–334 (1953)
49. Belk, M., Connelly, R.: Realizability of graphs. Discret. Comput. Geom. **37**, 125–137 (2007)
50. Belk. M.: Realizability of graphs in three dimensions. Discret. Comput. Geom. **37**, 139–162 (2007)
51. Colin de Verdière, Y.: Sur un nouvel invariant des graphes et un critère de planarité. J. Comb. Theory B **50**(1), 1121 (1990)
52. van der Holst, H.: Two tree-width-like graph invariants. Combinatorica **23**(4), 633–651 (2003)
53. Laurent, M., Varvitsiotis, A.: The Gram dimension of a graph. In: Mahjoub A.R., et al. (eds.) Proceedings of the 2nd International Symposium on Combinatorial Optimization (ISCO 2012), Athens. LCS, vol. 7422, pp. 356–367. Springer (2012)
54. Laurent M., Varvitsiotis, A.: A new graph parameter related to bounded rank positive semidefinite matrix completions. Preprint at arXiv:1204.0734 to appear in Mathematical Programming, Series A. (16 February 2013)