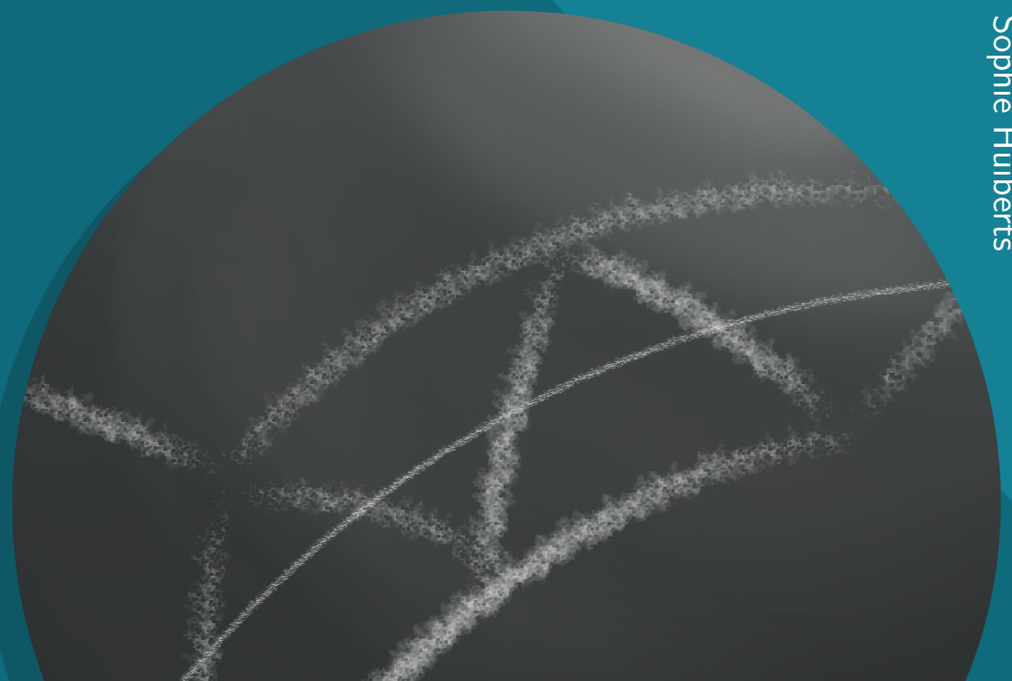
Geometric Aspects of Linear Programming

Sophie Huiberts



Geometric Aspects of Linear Programming

Shadow Paths, Central Paths, and a Cutting Plane Method

Sophie Huiberts

Copyright © 2022 by Sophie Huiberts
Cover artwork © 2022 by Tim Huiberts
Printed by Gildeprint

ISBN: 978-90-393-7462-7
doi:10.33540/1251

Geometric Aspects of Linear Programming:

Shadow Paths, Central Paths, and a Cutting Plane Method

Meetkundige Aspecten van Lineaire Programmering:

Schaduwpaden, Centrale Paden, en een Snijvlakmethode

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. H.R.B.M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op maandag 16 mei 2022 des middags te 2.15 uur

door

Sophie Huiberts

geboren op 24 augustus 1994 te Heerenveen

Promotor:
Prof. dr. G.L.M. Cornelissen
Copromotor:
Dr. D.N. Dadush
Beoordelingscommissie:
Prof. dr. R.H. Bisseling
Prof. dr. S. Pokutta
Prof. dr. H. Röglin
Dr. L. Sanità
Prof. dr. D.A. Spielman

Contents

Introduction

Before you lies a thesis about convex optimization. In optimization, we aim to find the "best" solution out of a set of options. Such tasks are often described by an *objective* as well as a number of *constraints* that must be satisfied. Examples include computing an efficient train schedule, digitally stabilizing video, or routing trucks to deliver goods at minimal cost.

Convex optimization deals with optimization problems whose set of *feasible solutions* is convex and whose objective is to minimize a convex function. The set of feasible solutions is convex when, if $x, y \in \mathbb{R}^n$ are solutions that satisfy all constraints and $0 \leq t \leq 1$ then $tx + (1-t)y$ also satisfies all constraints. An objective function f is convex if for $x, y \in \mathbb{R}^n$ and $0 \leq t \leq 1$ we have $f(tx+(1-t)y) \leq tf(x)+(1-t)f(y)$. Convex optimization captures a large class of natural problems and has a number of useful mathematical and computational properties.

In this thesis, we will focus on two classical convex optimization problems and study algorithms to solve them. The first is linear programming: the task of maximizing a linear function over a set of points described by linear inequalities. Chapters 2, 3, and 4 are all about questions that arise from the theory of linear programming and two popular algorithms for solving them. In Chapter 5 we shift our attention to the more general problem of minimizing a convex function when given access to a *separation oracle* for the set of feasible solutions. This introduction describes the new results of this thesis with a minimal number of references. Detailed background, history and discussion of prior work can be found in the individual chapters.

1.1 Linear Programming

In the first part of this thesis, the topic of interest will be one of the most basic optimization problems: linear programming. A linear program (LP) is any optimization problem that can be written in the following form:

$$\begin{aligned} &\text{maximize } c^\mathsf{T} x \\ &\text{subject to } Ax \leq b. \end{aligned} \tag{LP}$$

Here, the program data is known and consists of the matrix $A \in \mathbb{R}^{m \times n}$ and the vectors $b \in \mathbb{R}^m, c \in \mathbb{R}^n$. Our task is first to determine if there exists a *feasible solution*, i.e., a vector $x \in \mathbb{R}^n$ that satisfies all m linear inequalities. If this is the case, we need to find either such a feasible vector with maximal possible objective value or an infinite ray of feasible points that demonstrate the fact that no such maximizing point exists. Linear programs can model a large variety of optimization problems, and as such have been extensively used and studied in practice and in theory. There is a rich theory about linear programming, which can be found in various textbooks. An accessible introduction can be found in [137].

Among practical software available for solving general purpose linear programs, the key algorithms can be divided into two categories. The first category is that of simplex methods, which are the subject of Chapters 2 and 3. The second category is that of interior point methods, which are the subject of Chapter 4. We should note here that practical software typically uses floating-point arithmetic, while this thesis will use the real model of computation.

1.2 The Simplex Method

The simplex method is the oldest algorithm for linear programming, having been invented by Dantzig in 1947. The algorithm maintains a *basis*, a set $B \subseteq [m] := \{1, \ldots, m\}$ of rows such that A_B is invertible, such that $x_B = A_B^{-1} b_B$ is feasible for the linear program (LP). Here, the subscript means we restrict to the rows indexed by B. In every iteration of its main loop, it will replace one element of the basis B for another element of $[m]$ such that the new point x_B has better objective value. This repeats until a point with optimal objective value has been found. Geometrically, we can view this algorithm as navigating from vertex to vertex along the edges of the polyhedron of feasible points. One such path on a polyhedron is depicted in Figure 1.1.

Since a vertex can have neighbouring vertices, the simplex method needs a way to choose which basis element will be swapped out. Rules for making these choices are called pivot rules, and many different rules have been studied. A classic example is Dantzig's most negative reduced cost rule, which chooses the outgoing edge which maximizes the objective gain per unit of *slack*, the amount by which the constraint $A_i^\intercal x \leq b_i$ to be swapped out of the basis is no longer tight after the pivot step. Another well-known example is the steepest edge rule, as well as its approximate cousin, Harris' Devex rule, which chooses the edge whose angle to the objective is minimized.

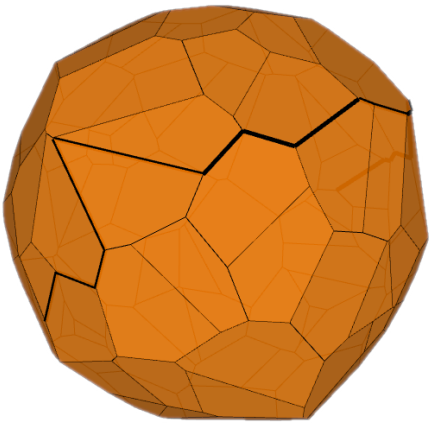


Figure 1.1: A simplex path on a convex polyhedron.

1.2.1 Analyzing the Simplex Method

The simplex method is implemented in most software packages for linear programming since it is very fast in practice. One aspect that contributes to its speed is that the simplex method typically visits only a small number of bases, and hence its main loop requires few iterations. To better understand why so few bases are visited, we want a theoretical analysis that supports this observation.

A major obstacle is that there are specially constructed inputs to the simplex method on which it visits at least a subexponentially large number of vertices of the feasible polyhedron. That means that the simplex method is slow in the worst case. Such worst-case inputs depend on the pivot rule, but are known for every major pivot rule that has been studied [6, 9, 72, 81, 85, 86, 94, 98, 113, 125, 125, 149]. As theoreticians, our task is thus to explain why difficult inputs do not appear in practice. This is the subject of Chapter 2. We study the performance of a simplex method under small perturbations of the input data by bounding the expected number of bases visited if a small amount of random noise is added to the description of the LP. The study of algorithms' performance under small perturbations is called *smoothed analysis* and was introduced by Spielman and Teng [179]. The simplex method is provably fast in expectation after the input has been perturbed. This is considered to imply that difficult inputs are very fragile, explaining why inputs from practice are relatively easy. In Chapter 2, we give the best known upper bound on the smoothed complexity of the simplex method, as summarized in the following theorem.

Theorem A (Theorem 2.5.12)**.** *There exists an explicit simplex method such that the following holds. Assume that* $m \geq n \geq 3$ *and that the entries of* $A \in \mathbb{R}^{m \times n}$ *and*

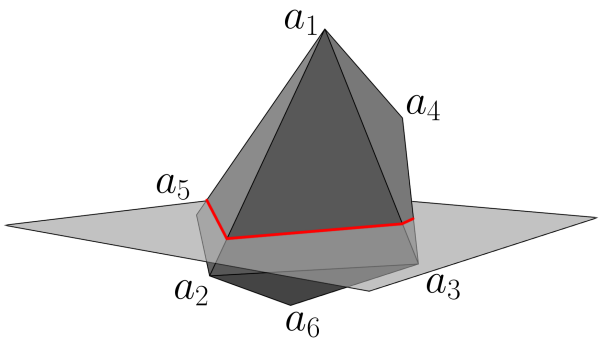


Figure 1.2: The convex hull of points $a_1, \ldots, a_6$ is intersected with a two-dimensional plane. The polygon at the intersection is marked in red.

$b \in \mathbb{R}^m$ are independent Gaussian distributed random variables with variance σ^2. If the rows of the expected matrix $\mathbb{E}[(A, b)] \in \mathbb{R}^{m\times(n+1)}$ each have norm at most 1*, then this algorithm solves* (LP) *using $O(n^2\sqrt{\log m}\,\sigma^{-2} + n^3 \log(m)^{3/2})$ pivot steps in expectation.*

The chapter is based on [50] and [51] and is an extension of my master's thesis [109], which proved a complexity bound of $O(n^2\sqrt{\log m}\,\sigma^{-2}+n^5 \log(m)^{3/2})$. The improved additive term is the result of a new algorithm which we call the symmetric random vertex algorithm. The analysis of the symmetric random vertex algorithm is based on the following key geometric estimate which is illustrated in Figure 1.2:

Theorem B (Theorem 2.4.1)**.** *Let $W \subseteq \mathbb{R}^n$ be a fixed two-dimensional subspace, $m \geq n \geq 3$ and let $A \in \mathbb{R}^{m\times n}$ be a matrix with rows $a_1, \ldots, a_m \in \mathbb{R}^n$, such that the entries of A are independent Gaussian random variables with variance σ^2 and such that $\|\mathbb{E}[a_i]\| \leq 1$ for every $i \in [m]$. Letting $Q(A) := \operatorname{conv}(a_1, \ldots, a_m)$ denote the convex hull of the row vectors, we find that the expected number of edges of the polygon $Q(A) \cap W$ is bounded as follows*:

$$\mathbb{E}[|\text{edges}(Q(A) \cap W)|] \leq O(n^2\sqrt{\log m}\,\sigma^{-2} + n^{2.5} \log m\,\sigma^{-1} + n^{2.5} \log(m)^{3/2}).$$

The above theorem is the main result in [109]. Bounding the complexity of the symmetric random vertex algorithm requires this same upper bound to hold when the rows $a_1, \ldots, a_m \in \mathbb{R}^n$ can be correlated in a limited way. For the sake of readability, we include the proof of Theorem 2.4.1 verbatim in Section 2.4 and point out the necessary minor adaptations in Section 2.5.

The same techniques used to prove the geometric estimate can be used to study a simpler question as well. Specifically, we study the expected number of vertices of a random polygon, where this polygon is the convex hull of points $a_1, \ldots, a_m \in \mathbb{R}^2$ that

are all independently Gaussian distributed with standard deviation σ and satisfying $\|\mathbb{E}[a_i]\| \leq 1$ for each $i \in [m]$. While the convex hull of m fixed points can have up to m vertices, we prove that the convex hull of our random points will have at most $O(\sigma^{-1} + \sqrt{\log m})$ points in expectation.

Theorem C (Theorem 2.3.1)**.** *For independently distributed points $a_1, \ldots a_m \in \mathbb{R}^2$, each with independent Gaussian distributed entries of variance σ^2 and $\|\mathbb{E}[a_i]\| \leq 1$ for all $i \in [m]$, the convex hull $\mathrm{conv}(a_1, \ldots, a_m)$ has $O(\sigma^{-1} + \sqrt{\log m})$ edges in expectation.*

This result is based on [51] and did not appear in [109]. Its proof can be found in Section 2.3 and can serve as a warm-up for the section following it. On a high level, the two-dimensional bound arises from the fact that every edge has length $\Omega(\sigma)$ in expectation, while the total perimeter is $O(1+\sigma\sqrt{\log m})$ in expectation. With a small trick, we can divide these two bounds to prove Theorem 2.3.1.

1.2.2 Diameter Bounds

So far we have thought about the lengths of paths from vertex to vertex on polyhedra generated according to the shadow vertex pivot rule. In Chapter 3, we set our eyes on a quantity that is more difficult to analyze: the shortest paths between vertices. Given a polyhedron, the smallest number k such that any two vertices are connected by a path of at most k edges is called the (combinatorial) diameter. The diameter of the cube in $\mathbb{R}^3$ for example is equal to 3, as illustrated in Figure 1.3. The polynomial Hirsch conjecture posits that any polyhedron in $\mathbb{R}^n$ with m facets can have combinatorial diameter at most a polynomial in n and m. This conjecture remains wide open to this day.

In Chapter 3, we study the combinatorial diameter for random polyhedra and prove high-probability upper and lower bounds. Specifically, we sample a matrix A whose rows come from a Poisson point process on the unit sphere and consider the convex hull $Q(A)$ of the rows as well as the polyhedron $P(A) = \{x \in \mathbb{R}^n : Ax \leq \vec{1}\}$. Assuming that we expect A to have more than $2^{\Omega(n)}$ rows, we find upper and lower bounds on the diameter of $Q(A)$ and $P(A)$ with high probability. Chapter 3 is based on [25] and proves the following result:

Theorem D. *Suppose that $n, m \in \mathbb{N}$ satisfy $n \geq 2$ and $m \geq 2^{\Omega(n)}$. Sample a matrix $A^\mathsf{T} := (a_1, \ldots, a_M) \in \mathbb{R}^{n\times M}$, where M is Poisson distributed with $\mathbb{E}[M] = m$, and $a_1, \ldots, a_M$ are sampled independently and uniformly from $\mathbb{S}^{n-1}$. Then, with probability at least $1 - m^{-n}$, we have that*

$$\Omega(nm^{\frac{1}{n-1}}) \leq \mathrm{diameter}(P(A)) \leq O(n^2 m^{\frac{1}{n-1}} + n^5 4^n).$$

$$\Omega(m^{\frac{1}{n-1}}) \leq \mathrm{diameter}(Q(A)) \leq O(nm^{\frac{1}{n-1}} + n^4 4^n).$$

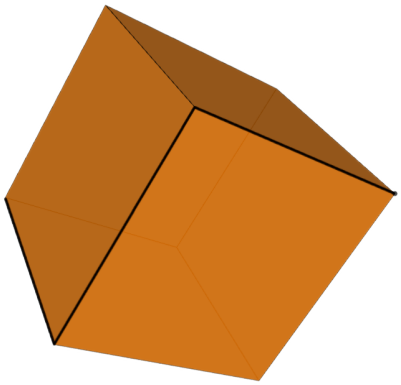


Figure 1.3: Any two vertices of the three-dimensional cube are connected by a path consisting of at most 3 edges.

The bounds we find are polynomial in m. The upper bound contains an additive term that is exponential in n, but this term does not depend on m. Compared to the worst-case diameter bound of $2^{n-3}m$ for a polyhedron $P(A)$ with m linear constraints, we find that our upper bound improves over this worst-case bound upper only when $m \geq \Omega(n^5 2^n)$. In the regime when $m \geq 2^{\Omega(n^2)}$, our upper and lower bounds are tight to within a factor n. Moreover, up to a constant factor the upper bound for $P(A)$ is equal to the length of an average shadow vertex method path induced by two random objective functions. This means that improving the upper bound would require using a different pivot rule to construct paths.

1.3 Interior Point Methods

Interior point methods are the other popular category of algorithms for solving linear programs. For these, the algorithm maintains a point inside the relative interior of the feasible set, and follows a curve inside the relative interior to find an optimal solution of the linear program.

In this section and in Chapter 4, we consider linear programs that are written in the following standard form:

$$\begin{aligned} &\text{maximize } c^{\mathsf{T}} x \\ &\text{subject to } Ax = b, \\ &\qquad x \geq \vec{0}. \end{aligned} \tag{LP'}$$

Assuming we are given a suitable interior point to start from, a standard interior point method can find an exact solution to (LP') in $O(\sqrt{n} \cdot L_{A,b,c})$ iterations, each iteration taking time polynomial in n and m [201]. Here, $L_{A,b,c}$ denotes the bit description length of the input data, assuming that $A \in \mathbb{Q}^{m \times n}$, $b \in \mathbb{Q}^m$, $c \in \mathbb{Q}^n$. After obtaining the initial point, e.g., through constructing an extended LP, these algorithms are *scaling-invariant*. To explain this term, let $D \in \mathbb{R}^{n \times n}$ be a diagonal matrix with strictly positive entries on the diagonal. Then we can consider the linear program

$$\begin{aligned} &\text{maximize } c^\mathsf{T} Dx \\ &\text{subject to } ADx = b, \\ &\qquad x \geq \vec{0}. \end{aligned}$$

We consider this alternative linear program to be equivalent under a diagonal rescaling, because its set of feasible solutions can be obtained by applying the linear transformation D^{-1} to the solutions of the first linear program. We say an interior point method is scaling invariant if, when applying it to these two equivalent linear programs starting at appropriate interior points x_0 and $D^{-1}x_0$, the internally maintained interior points are related by this same linear transformation.

In a separate line of research, people have found interior point methods whose iteration count depends only on n and L_A, the bit-complexity of the constraint matrix [129, 146, 147, 198]. So far, these algorithms have not been scaling-invariant, and whether a scaling-invariant algorithm could exist was an open question posed in [146].

Based on [52], Chapter 4 describes an interior point method that fills this gap, for it is both scaling invariant and has running time depending on n, m and L_A. This resolves the open question from [146] in the affirmative.

Theorem E (Theorem 4.3.16)**.** *There exists a scaling-invariant interior point method that finds an optimal solution to* (LP') *in*

$$O(n^{2.5} \log(n) \log(\bar{\chi}_A + n))$$

iterations, each iteration taking time polynomial in m and n.

In the above theorem statement, $\bar{\chi}_A$ is a condition number of the matrix A that equals

$$\bar{\chi}_A = \max\{\|B^{-1}A\| : B \text{ non-singular } m \times m\text{-submatrix of } A\}$$

and was already used in the context of interior point methods in [198]. The above theorem holds in the real model of computation where A, b, c have entries in $\mathbb{R}$. When $A \in \mathbb{Q}^{m \times n}$, a classic result states that $\bar{\chi}_A \leq 2^{O(L_A)}$, where L_A is the bit-complexity

of the matrix A. Moreover, if A is totally unimodular (every square non-singular submatrix has determinant 0, 1, or -1) then $\bar{\chi}_A \leq n$.

Since the algorithm is scaling-invariant, the bound can trivially be strengthened to $O(n^{2.5}\log(n)\log(\bar{\chi}_A^* + n))$, where $\bar{\chi}_A^* := \inf_D \bar{\chi}_{AD}$ is the infimum condition number over all strictly positive diagonal matrices D. The chapter also describes the first algorithm that can approximate $\bar{\chi}_A$ and can compute a diagonal matrix D such that $\bar{\chi}_{AD}$ is approximately equal to the infimum value $\bar{\chi}_A^*$.

Theorem F (Theorem 4.2.5)**.** *There is an $O(n^2m^2 + n^3)$ time algorithm that for any matrix $A \in \mathbb{R}^{m\times n}$ computes an estimate t of $\bar{\chi}_A$ such that*

$$t \leq \bar{\chi}_A \leq n(\bar{\chi}_A^*)^2 t$$

and a strictly positive diagonal matrix D such that

$$\bar{\chi}_A^* \leq \bar{\chi}_{AD} \leq n(\bar{\chi}_A^*)^3 .$$

In contrast to the above approximation result, it is known to be NP-hard to approximate $\bar{\chi}_A$ to within a factor $2^{\text{poly}(\text{rank}(A))}$, see [188].

1.4 The Separation Oracle Model

In the last chapter we look beyond linear programming and consider convex optimization in the *oracle model*. Suppose there is some convex compact set $K \subseteq \mathbb{R}^n$ and that we know $R \geq r > 0$ such that such that K contains some ball of radius r as a subset and such that K is contained inside the ball of radius R centered at the origin. Furthermore, we assume there is an L-Lipschitz convex function f with known $L > 0$. Our goal is to approximately minimize $f(x)$ over $x \in K$ when given access to suitable oracles for K and f. That means that K and f are not known in advance. Instead, we can make queries about K and f which will be answered.

For K, we have access to a *separation oracle*. This means that we can query any point $x \in \mathbb{R}^n$ and receive either the answer "$x \in K$" or we receive some $a \in \mathbb{R}^n, b \in \mathbb{R}$ such that $\langle a, x\rangle > b$ and such that $\langle a, y\rangle \leq b$ for every $y \in K$. In the latter case, the inequality defines a hyperplane that *separates* the queried point from the set K, providing a "reason" or certificate for why $x \notin K$. For f, we have access to a *first-order oracle*. We can query any point $x \in \mathbb{R}^n$ and receive the value $f(x)$ and a (sub)gradient $\nabla f(x)$.

Convex optimization in the separation oracle model is a fundamental topic in optimization, see, e.g., [103]. The model is quite general and covers a large variety of problems, including problems where linear programming formulations are known

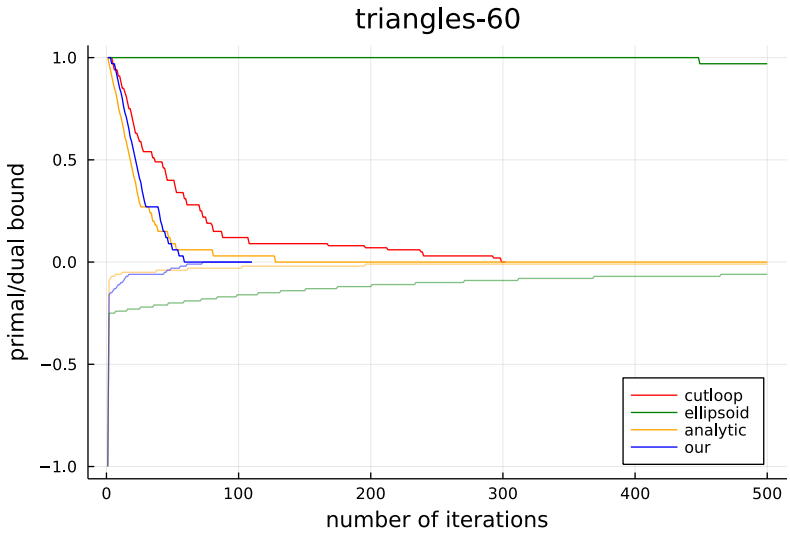


Figure 1.4: Progression of the upper (primal) and lower (dual) bounds of the different cutting plane algorithms for a maximum-cardinality matching problem consisting of a graph with 500 vertices and 60 randomly planted triangles.

but too large to write out explicitly, e.g., the minimum cost matching problem on general graphs.

One famous algorithm to solve convex optimization problems in the separation oracle model is called the *ellipsoid method*. It maintains an ellipsoid in $\mathbb{R}^n$ that contains the optimal solutions, and after querying a point from the interior of the ellipsoid, the algorithm will be able to find a new ellipsoid with smaller volume. The ellipsoid method will identify a point $x \in K$ satisfying $f(x) \leq \min_{z \in K} f(z) + \varepsilon$ after $O(n^2 \log(LR/(\varepsilon r)))$ queries to the oracles. This complexity depends only logarithmically on L, R, r and ε, but at the expense of the number of queries growing with the dimension n. The ellipsoid method is, generally speaking, too slow for most practical applications.

One popular approach to optimization in this model is the linear programming based cutting plane method. It maintains an LP consisting of all known constraints on the optimal solutions and queries an optimal solution to this linear program. If the queried point is not feasible or optimal, the newly obtained constraint is added to the linear program and the method repeats. This cutting plane method is often practical but does not have any theoretical guarantees on its convergence rate.

In Chapter 5 we introduce a new algorithm that both has a convergence guarantee and is competitive in experiments. The convergence guarantee is as follows:

Theorem G (Theorem 5.2.6)**.** *There exists an explicit algorithm which, given a separation oracle $\mathcal{A}$ for a convex body $K \subseteq \mathbb{R}^n$ satisfying $z + r\mathbb{B}_2^n \subseteq K \subseteq R\mathbb{B}_2^n$ and a first-order oracle for an L-Lipschitz convex function $f\colon \mathbb{R}^n \to \mathbb{R}$, finds a feasible solution $y \in K$ with $f(y) \geq \min_{x \in K} f(x) \geq f(y) - \varepsilon$ after*

$$O\left(\frac{R^2}{r^2} \cdot \frac{R^2 L^2}{\varepsilon^2}\right)$$

oracle queries.

This convergence rate depends polynomially on the problem parameters, but without a dependence on the dimension n of the problem.

We experimentally compare our algorithm to both methods named above as well as an analytic center cutting plane method. Comparing these algorithms on a testbed of combinatorial, semidefinite and machine learning problems, we find our algorithm to be competitive in terms of number of oracle queries. Figure 1.4 is a comparison of the algorithms on one of our combinatorial instances. The figure depicts for each algorithm the best lower (primal) and upper (dual) bound on the size of a maximum matching in a random graph. The vertical axis is scaled as a fraction of the gap closed between the optimal value and the initial bound, and the horizontal axis measures the number of oracle queries that have been made by the algorithms. Here, 'cutloop' denotes the LP-based cutting plane method, which only produces upper bounds, 'ellipsoid' denotes the ellipsoid method, 'analytic' stands for the analytic center cutting plane method and 'our' is the algorithm described in Chapter 5.

1.5 Notation

Here we define notation and basic concepts that will see use throughout this thesis. Individual chapters may have their own section for notation as well.

- For functions $f, g : \mathbb{R} \to \mathbb{R}$ we write
  - $f(x) = O(g(x))$ if there exist $C > 0$ and $x_0 \in \mathbb{R}$ such that $f(x) \leq Cg(x)$ for every $x \geq x_0$ in the domain of both functions,
  - $f(x) = \Omega(g(x))$ if there exist $C > 0$ and $x_0 \in \mathbb{R}$ such that $f(x) \geq Cg(x)$ for every $x \geq x_0$ in the domain of both functions,
  - $f(x) = \Theta(g(x))$ if both $f(x) = O(g(x))$ and $g(x) = \Omega(g(x))$
  - $f(x) = \tilde{O}(g(x))$ if there exists $c > 0$ such that $f(x) = O(g(x)\log(g(x))^c)$.

- Vector inequalities are defined coordinate-wise: $v \le w$ if and only if $v_i \le w_i$ for all $i \le n$.
- We abbreviate $[m] := \{1, \dots, m\}$ and $\binom{[m]}{n} = \{I \subseteq [m] \mid |I| = n\}$.
- for a set $K \subseteq \mathbb{R}^n$ and a scalar t, we write $tK := \{tx : x \in K\}$.
- For $a, b \in \mathbb{R}$ we have intervals $[a, b] = \{r \in \mathbb{R} : a \le r \le b\}$ and $(a, b) = \{r \in \mathbb{R} : a < r < b\}$.
- For $x > 0$, we define $\log x$ to be the logarithm base e of x.
- For a set A, we use the notation $\mathbb{1}[x \in A]$ to denote the indicator function of A, i.e., $\mathbb{1}[x \in A] = 1$ if $x \in A$ and 0 otherwise.
- Depending on the context, the inner product of x and y is written with the notation $x^\mathsf{T} y = \langle x, y \rangle = \sum_{i=1}^n x_i y_i$. We use the ℓ_2-norm $\|x\|_2 = \sqrt{x^\mathsf{T} x}$ and the ℓ_1-norm $\|x\|_1 = \sum_{i=1}^n |x_i|$. Any norm without subscript is the ℓ_2-norm.
- The standard basis of $\mathbb{R}^n$ consists of the vectors $e_i = (0, \dots, 0, 1, 0, \dots, 0), i \in [n]$ with the only non-zero entry being 1 on the i^{th} location. The all-ones vector is $\vec{1}$ and the all-zeroes vector is $\vec{0}$
- The unit sphere in $\mathbb{R}^n$ is denoted by $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$ and the unit ball is $\mathbb{B}_2^n = \{x \in \mathbb{R}^n : \|x\| \le 1\}$.
- We let $\mathbb{R}_{++}$ denote the set of strictly positive reals, and $\mathbb{R}_+$ the set of nonnegative reals.
- A set $V + p$ is an affine subspace if $V \subseteq \mathbb{R}^n$ is a linear subspace. We say its dimension $\dim(V + p)$ equals $\dim(V)$, the dimension of V as a vector space.
- If $S \subseteq \mathbb{R}^n$ then the affine hull $\mathrm{aff}(S)$ is the smallest affine subspace containing S. We say $\dim(S) = k$ if $\dim(\mathrm{aff}(S)) = k$.
- For any linear or affine subspace $V \subseteq \mathbb{R}^n$ the orthogonal projection onto V is denoted by π_V.
- When $V \subseteq \mathbb{R}^n$ is a linear subspace, its orthogonal complement is denoted $V^\perp = \{x \in \mathbb{R}^n : v^\mathsf{T} x = 0, \forall v \in V\}$. For $v \in \mathbb{R}^n$ we shorten $v^\perp := \mathrm{span}(v)^\perp$.
- We write $\mathrm{vol}_k(S)$ for the k-dimensional volume of S.
- For $A, B \subseteq \mathbb{R}^n$ we write the Minkowski sum $A + B = \{a + b : a \in A, b \in B\}$. For a vector $v \in \mathbb{R}^n$ we write $A + v = A + \{v\}$. For a set of scalars $S \subseteq \mathbb{R}$ we write $v \cdot S = \{sv : s \in S\}$.

- We say vectors $a_1, \ldots, a_k$ in $\mathbb{R}^n$ are *affinely independent* if there is no $(k-2)$-dimensional affine subspace containing all of $a_1, \ldots, a_k$. Algebraically, $a_1, \ldots, a_k$ are affinely independent if the system $\sum_{i \le k} \lambda_i a_i = \vec{0}, \sum_{i \le k} \lambda_i = 0$ has no non-trivial solution.
- For $A \in \mathbb{R}^{m \times n}$ a matrix and $B \subseteq [m]$ we write $A_B \in \mathbb{R}^{|B| \times n}$ for the submatrix of A consisting of the rows indexed in B, and for $b \in \mathbb{R}^m$ we write b_B for the restriction of b to the coordinates indexed in B.
- For a set X and function $f : X \to \mathbb{R}$, we write $\arg\min\{f(x) : x \in X\}$ to denote an arbitrary but fixed $x^* \in X$ such that $f(x^*) = \min_{x \in X} f(x)$.

Convexity A set $S \subseteq \mathbb{R}^n$ is convex if for all points $x, y \in S$ and $\lambda \in [0,1]$ we have $\lambda x + (1-\lambda)y \in S$. We write $\mathrm{conv}(S)$ to denote the convex hull of S, which is the intersection of all convex sets $T \supset S$. In an n-dimensional vector space, the convex hull equals

$$\mathrm{conv}(S) = \{\sum_{i=1}^{n+1} \lambda_i s_i : \lambda_1, \ldots, \lambda_{n+1} \ge 0, \sum_{i=1}^{n+1} \lambda_i = 1, s_1, \ldots, s_{n+1} \in S\}.$$

For $x, y \in \mathbb{R}^n$ the line segment between x and y is $[x, y] = \mathrm{conv}(\{x, y\})$ and we say it has length $\mathrm{length}([x, y]) = \|x - y\|$.

A *polyhedron* is a set that can be written $P = \{x \in \mathbb{R}^n : Ax \le b\}$ for $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$. We will only consider convex polyhedra. A *face* $F \subseteq P$ is a convex subset such that if $x, y \in P$ and $\lambda \in (0,1)$ satisfy $\lambda x + (1-\lambda)y \in F$, then $x, y \in F$. In particular, a set F is a face of the polyhedron P if and only if there exists $I \subseteq [m]$ such that F coincides with P intersected with $a_i^\intercal x = b_i, \forall i \in I$, where $a_i, i \in I$ denote rows of A. A zero-dimensional face is called a *vertex*, one-dimensional face is called an *edge*, and a $(\dim(P)-1)$-dimensional face is called a *facet*. We use the notation $\mathrm{vertices}(P)$ to denote the set of vertices of P and $\mathrm{edges}(P)$ for the set of edges of P.

For any polyhedron $P \subseteq \mathbb{R}^n$, a *path* is a sequence $v_1, v_2, \ldots, v_k \in P$ of vertices, such that each line segment $[v_i, v_{i+1}], i \in [k-1]$, is an edge of P. A path is *monotone* with respect to an inner product $\langle w, \cdot \rangle$ if $\langle w, v_{i+1} \rangle \ge \langle w, v_i \rangle$ for every $i \in [k-1]$. The *distance* between vertices $v_1, v_2 \in P$ is the minimum number k such that there exists a path $v_1', v_2', \ldots, v_{k+1}'$ with $v_1 = v_1'$ and $v_{k+1}' = v_2$. The *diameter* of P is the maximal distance between any two of its vertices.

When considering polyhedra, if A is a matrix with rows $a_1, \ldots, a_m \in \mathbb{R}^n$, we write $Q(A) = \mathrm{conv}(a_1, \ldots, a_m)$ and $P(A) = \{x \in \mathbb{R}^n : Ax \le \vec{1}\}$. Given a right-hand side $b \in \mathbb{R}^m$ we write $P(A, b) = \{x \in \mathbb{R}^n : Ax \le b\}$.

We will need the following classical comparison inequality for surface areas of convex sets (see for example [24, Chapter 7]).

Lemma 1.5.1 (Monotonicity of Surface Area)**.** *If $K_1 \subseteq K_2 \subseteq \mathbb{R}^n$ are compact full-dimensional convex sets, then $\mathrm{vol}_{n-1}(\partial K_1) \leq \mathrm{vol}_{n-1}(\partial K_2)$, where ∂ denotes the topological boundary of a body.*

Although this fact holds generally, its primary use in this thesis will be to argue that any compact polygon $K \subseteq \mathbb{R}^2$ has perimeter (the sum of the lengths of its edges) at most $2\pi \max_{x \in K} \|x\|$.

Random Variables

For a random variable $X \in \mathbb{R}$, we denote its expectation (mean) by $\mathbb{E}[X]$ and its variance by $\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2]$. For a random vector $Y \in \mathbb{R}^n$, we define its expectation (mean) $\mathbb{E}[Y] := (\mathbb{E}[Y_1], \ldots, \mathbb{E}[Y_n])$ and its variance (expected squared distance from the mean) $\mathrm{Var}(Y) := \mathbb{E}[\|Y - \mathbb{E}[Y]\|^2]$.

If μ is a probability density function, we write $x \sim \mu$ to denote that x is a random variable distributed with probability density μ. For an event $E \subseteq \Omega$ in a measure space, we write $E^c := \Omega \setminus E$ to denote its complement.

For jointly distributed $X \in \Omega_1, Y \in \Omega_2$, we will often minimize the expectation of X over instantiations $y \in A \subseteq \Omega_2$. For this, we use the notation

$$\min_{Y \in A} \mathbb{E}[X \mid Y] := \min_{y \in A} \mathbb{E}[X \mid Y = y].$$

Computational Model

We use the real model of computation, allowing basic arithmetic operations $+$, $-$, $\times$, $/$, comparisons, square root computations. Exact square root computations could be omitted by using approximate square roots; we assume exact computations for simplicity. In Chapter 2 we furthermore assume that we can sample from the Gaussian distribution.

Chapter 2

Smoothed Analysis of the Simplex Method

Explaining the excellent practical performance of the simplex method for linear programming has been a major topic of research for over 50 years. One of the most successful frameworks for understanding the simplex method was given by Spielman and Teng [179], who developed the notion of smoothed analysis. Starting from an arbitrary linear program with n variables and m constraints, Spielman and Teng analyzed the expected runtime over random perturbations of the LP, known as the smoothed LP, where variance σ^2 Gaussian noise is added to the LP data. In particular, they gave a two-stage shadow vertex simplex algorithm which uses an expected $\widetilde{O}(n^{55}m^{86}\sigma^{-30} + n^{70}m^{86})$ number of simplex pivots to solve the smoothed LP. Their analysis and runtime was substantially improved by Deshpande and Spielman [66] and later Vershynin [200]. The fastest current algorithm, due to Vershynin, solves the smoothed LP using an expected $O\big(\log(m)^2 \cdot \log\log m \cdot (n^3\sigma^{-4} + n^5\log(m)^2 + n^9\log(n)^4)\big)$ number of pivots, improving the dependence on n from polynomial to poly-logarithmic.

In this chapter we present a new shadow vertex based simplex algorithm which solves the smoothed LP using

$$O(n^2\sqrt{\log m}\,\sigma^{-2} + n^3\log(m)^{3/2})$$

pivot steps in expectation. The result uses a *shadow bound* from my master's thesis [109] and whose proof is included as Section 2.4 for the sake of readability. Furthermore, this chapter includes a new smoothed analysis of the number of edges of a two-dimensional polygon.

2.1 Introduction

The simplex method for linear programming is one of the most important algorithms of the 20th century. Invented by Dantzig in 1947 [58, 59], it remains to this day

This chapter is based on [50] and [51], joint works with Daniel Dadush. The shadow bound Theorem 2.4.1 and its application using the dimension-by-dimension algorithm also appeared in my master's thesis [109], whereas the symmetric random vertex algorithm and Theorem 2.3.1 are new.

one of the fastest methods for solving LPs in practice. The simplex method is not one algorithm however, but a class of LP algorithms, each differing in the choice of *pivot rule*. At a high level, the simplex method moves from vertex to vertex along edges of the feasible polyhedron, where the pivot rule decides which edges to cross, until an optimal vertex or unbounded ray is found. Important examples include Dantzig's most negative reduced cost [59], the Gass and Saaty parametric objective [88] and Goldfarb's steepest edge [93] method. We note that for solving LPs in the context of branch & bound and cutting plane methods for integer programming, where the successive LPs are "close together", the dual steepest edge method [82] is the *dominant* algorithm in practice [21, 22], due its observed ability to quickly re-optimize.

The continued success of the simplex method in practice is remarkable for two reasons. Firstly, there is no known polynomial time simplex method for LP. Indeed, there are exponential examples for almost every major pivot rule starting with constructions based on *deformed products* [6, 9, 94, 98, 113, 125, 149], such as the Klee-Minty cube [125], which defeat most classical pivot rules, and more recently based on Markov Decision Processes (MDP) [72, 81, 85, 86], which notably defeat randomized and history dependent pivot rules. Furthermore, for an LP with n variables and m constraints, the fastest provable (randomized) simplex method requires $2^{O(\sqrt{n \log(m-n)})}$ pivots [106, 117, 138], while the observed practical behavior is linear $O(n+m)$ [172]. Secondly, it remains the most popular way to solve LPs despite the tremendous progress for polynomial time methods [122], most notably, interior point methods [119, 132, 143, 160]. How can we explain the simplex method's excellent practical performance?

This question has fascinated researchers for decades. An immediate question is how to model instances in "practice", or at least instances where the simplex method should perform well? The research on this subject has, broadly speaking, followed three different lines: the analysis of average case LP models, where natural distributions of LPs are studied, the smoothed analysis of arbitrary LPs, where small random perturbations are added to the LP data, and work on structured LPs, such as totally unimodular systems and MDPs. We review the major results for the first two lines in the next section, as they are the most relevant to the present work, and defer additional discussion to the related work section. To formalize the model, we consider LPs in n variables and m constraints of the following form:

$$\begin{aligned} \max\ & c^{\mathsf{T}} x \\ & Ax \leq b. \end{aligned} \tag{2.1}$$

We denote the feasible polyhedron by $P = P(A, b) := \{x \in \mathbb{R}^n : Ax \leq b\}$. We now introduce relevant details for the simplex methods of interest to this work.

Parametric Simplex Algorithms While a variety of pivot rules have been studied, the most successfully analyzed in theory are the so-called parametric simplex methods, due to the useful geometric characterization of the paths they follow. The first such method, and the main one used in the context of smoothed analysis, is the parametric objective method of Gass and Saaty [88], dubbed the *shadow (vertex) simplex* method by Borgwardt [26]. Starting at a *known* vertex v of P maximizing an objective c', the parametric objective method computes the path corresponding to the sequence of maximizers for the objectives obtained by interpolating $c' \to c$. This path is well-defined under mild non-degeneracy assumptions. The name shadow vertex method is derived from the fact that the visited vertices are in correspondence with those on the projection of P onto $W := \text{span}(c, c')$, the 2D convex polygon known as the shadow of P on W (see figure 2.2 for an illustration). In particular, the number of vertices traversed by the method is bounded by the number of vertices of the projection, known as the *size of the shadow*.

An obvious problem, as with most simplex methods, is how to initialize the method at a feasible vertex if one exists. This is generally referred to as the Phase I problem, where Phase II then corresponds to finding an optimal solution. A common Phase I adds artificial variable(s) to make feasibility trivial and applies the simplex method to drive them to zero.

A more general method, popular in the context of average case analysis, is the self-dual parametric simplex method of Dantzig [60]. In this method, one *simultaneously* interpolates the objectives $c' \to c$ and right hand sides $b' \to b$ which has the effect of combining Phase I and II. Here c' and b' are chosen to induce a *known* initial maximizer. While the polyhedron is no longer fixed, the breakpoints in the path of maximizers (now a piecewise linear curve) can be computed via certain primal and dual pivots. This procedure was in fact generalized by Lemke [134] to solve linear complementarity problems. We note that the self dual method can roughly speaking be simulated in a higher dimensional space by adding an interpolation variable λ, i.e. $Ax \leq \lambda b + (1 - \lambda)b', 0 \leq \lambda \leq 1$, which has been the principal approach in smoothed analysis.

2.1.1 Prior Work

Here we present the main works in both average case and smoothed analysis which inform our main results, presented in the next section. A common theme in these works, which all study parametric simplex methods, is to first obtain a bound on the expected parametric path length, with respect to some distribution on interpolations and LPs, and then find a way to use the bounds algorithmically. This second step can be non-obvious, as it is often the case that one cannot directly find a starting vertex on the path in question. We now present the main random LP models that have

been studied, presenting path length bounds and algorithms. Lastly, as our results are in the smoothed analysis setting, we explain the high level strategies used to prove smoothed (shadow) path bounds.

Average case Models The first model, introduced in the seminal work of Borgwardt [26–28, 30], examined LPs of the form $\max c^\mathsf{T} x, Ax \leq \vec{1}$, possibly with $x \geq \vec{1}$ constraints (note that this model is always feasible at $\vec{0}$), where the rows of $A \in \mathbb{R}^{m \times n}$ are drawn i.i.d. from a rotationally symmetric distribution (RSD) and $c \in \mathbb{R}^n \setminus \{\vec{0}\}$ is fixed and non-zero. Borgwardt proved tight bounds on the expected shadow size of the feasible polyhedron when projected onto any fixed plane. For general RSD, he proved a sharp $\Theta(n^2 m^{1/(n-1)})$ [28, 30] bound, tight for rows drawn uniformly from the sphere, and for Gaussians a sharp $\Theta(n^{1.5}\sqrt{\log m})$ bound [28], though this last bound is only known when m is very large compared to n. On the algorithmic side, Borgwardt [27] gave a *dimension by dimension* (DD) algorithm which solves a linear program by traversing $n - 2$ different shadow vertex paths to iteratively solve the restrictions $\max c^\mathsf{T} x, Ax \leq \vec{1}, x_i = 0, i \in \{k+1, \ldots, n\}$, for $k \geq 2$, which are all of RSD type.

For the next class, Smale [173] analyzed the standard self dual method for LPs where A and (c, b) are chosen from independent RSD distributions, where Megiddo [141] gave the best known bound of $f(\min\{n, m\})$ iterations, for some exponentially large function f. Adler [1] and Haimovich [105] examined a much weaker model where the data is fixed, but where the signs of all the inequalities, including non-negativity constraints, are flipped uniformly at random. Using the combinatorics of hyperplane arrangements, they achieved a remarkable bound of $O(\min\{n, m\})$ for the expected length of parametric paths. These results were made algorithmic shortly thereafter [2,3,184], where it was shown that a lexicographic version of the parametric self dual simplex method[1] requires $\Theta(\min\{n, m\}^2)$ iterations, where tightness was established in [3]. While these results are impressive, a notable criticism of the symmetry model is that it results in infeasible LPs with overwhelming probability once m is a bit larger than n.

Smoothed LP Models The *smoothed* analysis framework, introduced in the breakthrough work of Spielman and Teng [179], helps explain the performance of algorithms whose worst-case examples are in essence *pathological*, i.e., which arise from very brittle structures in instance data. To get rid of these structures, the idea is to add a small amount of noise to the data, quantified by a parameter σ, where the general goal is then to prove an expected running time bound over any *smoothed instance* that

[1] These works use seemingly different algorithms, though they were shown to be equivalent to a lexicographic self-dual simplex method by Meggiddo [140].

scales inverse polynomially with σ. Beyond the simplex method, smoothed analysis has been successfully applied to many other algorithms such as interior point methods [177], Gaussian elimination [166], Lloyd's k-means algorithm [7], the 2-OPT heuristic for the TSP [79], and much more.

The smoothed LP model, introduced by [179], starts with any base LP

$$\max c^\mathsf{T} x,\ \bar{A}x \leq \bar{b}, \qquad \text{(Base LP)}$$

$\bar{A} \in \mathbb{R}^{m\times n}$, $\bar{b} \in \mathbb{R}^m$, $c \in \mathbb{R}^n \setminus \{\vec{0}\}$, where the rows of $(\bar{A}, \bar{b})$ are normalized to have ℓ_2 norm at most 1. From the base LP, we generate the smoothed LP by adding Gaussian perturbations to both the constraint matrix A and the right hand side b. Precisely, the data of the smoothed LP is

$$A = \bar{A} + \hat{A},\ b = \bar{b} + \hat{b},\ c,$$

and we wish to solve

$$\max c^\mathsf{T} x,\ Ax \leq b, \qquad \text{(Smoothed LP)}$$

where the perturbations $\hat{A}$,$\hat{b}$ have i.i.d. mean 0, variance σ^2 Gaussian entries. Note that the objective is not perturbed in this model, though we require that $c \neq \vec{0}$. An LP algorithm is said to have polynomial smoothed complexity if for any base LP data A, b, c as above, we have

$$\mathbb{E}_{\hat{A},\hat{b}}[T(A, b, c)] = \text{poly}(m, n, 1/\sigma), \qquad \text{(Smoothed Complexity)}$$

where $T(A, b, c)$ is the running time of the algorithm on a given smoothed instance. Crucially, this complexity measure allows for an inverse polynomial dependence on σ, the perturbation size. Focusing on the simplex method, T will measure the number of simplex pivots used by the algorithm as a proxy for the running time.

In previous works, the complexity of the algorithms is reduced in a black box manner to a shadow bound for *smoothed unit LPs*. In particular, a smoothed unit LP has a base system $\bar{A}x \leq \vec{1}$, where $\bar{A}$ has row norms at most 1, and smoothing is performed only to $\bar{A}$. Here the goal is to obtain a bound on the expected shadow size with respect to any fixed plane. Note that if $\bar{A}$ is the zero matrix, then this is exactly Borgwardt's Gaussian model, where he achieved the asymptotically tight bound of $\Theta(n^{1.5}\sqrt{\log m})$ when m is large compared to n [28]. For smoothed unit LPs, Spielman and Teng [179] gave the first bound of $O(n^3 m\sigma^{-6} + n^6 m \log(m)^3)$. Deshpande and Spielman [66] derived a bound of $O(nm^2 \log m\sigma^{-2} + n^2 m^2 \log(m)^2)$, substantially improving the dependence on σ while squaring the dependence on m. Lastly, Vershynin [200] achieved a bound of $O(n^3\sigma^{-4} + n^5 \log(m)^2)$, dramatically improving

the dependence on m to poly-logarithmic, though still with a worse dependence on σ than [66].

Before discussing the high level ideas for how these bounds are proved, we overview how they are used algorithmically. In this context, [179] and [200] provide two different reductions to the unit LP analysis, each via an interpolation method. Spielman and Teng first solve the smoothed LP with respect to an artificial "somewhat uniform" right hand side b', constructed to force a randomly chosen basis of A to yield a vertex of the artificial system. From here they use the shadow vertex method to compute a maximizer for the right hand side b', and continue via interpolation to derive an optimal solution for b. Here the analysis is quite challenging, since in both steps the LPs are not quite smoothed unit LPs and the used shadow planes correlate with the perturbations. To circumvent these issues, Vershynin uses a *random vertex* (RV) algorithm, which starts with $b' = \vec{1}$ (i.e. a unit LP) and adds a random additional set of n inequalities to the system to induce an "uncorrelated known vertex". From this random vertex, he proceeds similarly to Spielman and Teng, but now at every step the LP is of smoothed unit type and the used shadow planes are (almost) independent of the perturbations. In Vershynin's approach, the main hurdle was to give a simple shadow vertex algorithm to solve unit LPs, which correspond to the Phase I problem. An extremely simple method for this was in fact already given in the early work of Borgwardt [28], namely, the dimension by dimension (DD) algorithm. The application of the DD algorithm in the smoothed analysis context was however only discovered much later by Schnalzger [168]. As it is both simple and not widely known, we will describe the DD algorithm and its analysis in Section 2.5.

We note that beyond the above model, smoothed analysis techniques have been used to analyze the simplex method in other interesting settings. In [35], the successive shortest path algorithm for min-cost flow, which is a shadow vertex algorithm, was shown to be efficient when only the objective (i.e. edge costs) is perturbed. In [121], Kelner and Spielman used smoothed analysis techniques to give a "simplex like" algorithm which solves arbitrary LPs in polynomial time. Here they developed a technique to analyze the expected shadow size when only the right hand side of an LP is perturbed.

Shadow Bounds for Smoothed Unit LPs Let $a_1, \ldots, a_m \in \mathbb{R}^n, i \in [m]$, denote the rows of the constraint matrix of the smoothed unit LP $Ax \leq \vec{1}$. The goal is to bound from above the expected number of vertices in the projection of the feasible polyhedron $P(A) = \{x \in \mathbb{R}^n : Ax \leq \vec{1}\}$ onto a fixed 2D plane W. As noticed by Borgwardt, by a duality argument, this number of vertices is upper bounded by the number of edges in the *polar polygon* (see Figure 2.2 for an illustration). Letting $Q(A) := \text{conv}(a_1, \ldots, a_m)$, the convex hull of the rows, the polar polygon can be

expressed as $Q(A) \cap W$.

We survey the different approaches used in [66, 179, 200] to bound the number of edges of $Q(A) \cap W$. Let u_θ, $\theta \in [0, 2\pi]$, denote an angular parametrization of the unit circle in W, and let $r_\theta = u_\theta \cdot \mathbb{R}_{\geq 0}$ denote the corresponding ray. Spielman and Teng [179] bounded the probability that any two nearby rays r_θ and $r_{\theta+\varepsilon}$ intersect different edges of $Q(A) \cap W$ by a linear function of ε. Summing this probability over any fine enough discretization of the circle upper bounds the expected number of edges of $Q(A) \cap W$.[2] Their probability bound proceeds in two steps, first they estimate the probability that the Euclidean distance between the intersection of r_θ with its corresponding edge and the boundary of that edge is small (the *distance lemma*), and second they estimate the probability that angular distance is small compared to Euclidean distance (the *angle of incidence bound*). Vershynin [200] avoided the use of the angle of incidence bound by measuring the intersection probabilities with respect to the "best" of three different viewpoints, i.e. where the rays emanate from a well-chosen set of three equally spaced viewpoints as opposed to just the origin. This gave a much more efficient reduction to the distance lemma, and in particular allowed Vershynin to reduce the dependence on n from linear to poly-logarithmic. Deshpande and Spielman [66] bounded different probabilities to get their shadow bound. Namely, they bounded the probability that nearby objectives u_θ and $u_{\theta+\varepsilon}$ are maximized at different vertices of $Q(A) \cap W$. The corresponding discretized sum over the circle directly bounds the number of vertices of $Q(A) \cap W$, which is the same as the number of edges.

Complexity in two dimensions In two dimensions, the shadow size reduces to the complexity of the convex hull. The convex hull of random points on two dimensions was studied before by [57, 68, 168]. The best upper bound that can be found in the mentioned references is $O(\sqrt{\log m} + \sigma^{-1}\sqrt{\log m})$, asymptotically slightly worse than the bound $O(\sqrt{\log m} + \sigma^{-1})$ that we will prove in Theorem 2.3.1. The best available lower bound is when $\mathbb{E}[a_1], \ldots, \mathbb{E}[a_m]$ are equally spaced on the unit circle, where the convex hull has $\Omega(\min(m, \sqrt{\log m} + \frac{\sqrt[4]{\log m}}{\sqrt{\sigma}}))$ vertices in expectation. That result can be found in [68].

2.1.2 Results

While the original proof of Spielman and Teng has now been substantially simplified, the resulting analyses are still complex and the parameter improvements have not been uniform. In this chapter, we give a "best of all worlds" analysis, which is both much

[2] One must be a bit more careful when $Q(A) \cap W$ does not contain the origin, but the details are similar.

simpler and improves all prior parameter dependencies. The main novel contribution is a new algorithm, which is analyzed using a shadow bound from my master's thesis. Since our algorithmic result requires a minor adaptation, the proof of this shadow bound is included in full in this chapter.

Reference	Expected Number of Vertices	Model
[30]	$\Theta(n^2 m^{1/(n-1)})$	RSD
[28]	$\Theta(n^{3/2}\sqrt{\log m})$	Gaussian, m large
[179]	$O(n^3 m\sigma^{-6} + n^6 m \log(m)^3)$	Smooth
[66]	$O(nm^2 \log m\ \sigma^{-2} + n^2 m^2 \log(m)^2)$	Smooth
[200]	$O(n^3\sigma^{-4} + n^5 \log(m)^2)$	Smooth
[109]	$O(n^2\sqrt{\log m}\ \sigma^{-2} + n^{2.5} \log(m)^{3/2}(1+\sigma^{-1}))$	Smooth

Table 2.1: Shadow Bounds. Logarithmic factors are simplified. The "Gaussian, m large" lower bound applies in the smoothed model as well.

We note that the bounds below hold for $n \geq 3$. Recalling the models, the results in the Table 2.1 bound the expected number of vertices in the projection of a random polytope $Ax \leq 1$, $A \in \mathbb{R}^{m\times n}$, onto any fixed two-dimensional plane. The models differ in the class of distributions examined for A. In the RSD model, the rows of A are distributed i.i.d. according to an *arbitrary* rotationally symmetric distribution. In the Gaussian model, the rows of A are i.i.d. mean zero standard Gaussian vectors. Note that this is a special case of the RSD model. In the smoothed model, the rows of A are n-dimensional Gaussian random vectors with standard deviation σ centered at vectors of norm at most 1, i.e. the expected matrix $\mathbb{E}[A]$ has rows of ℓ_2 norm at most 1. The "m large" in the table indicates that that bound only holds for m large enough (compared to n). The Gaussian model is a special case of the smoothed analysis model, and hence the applicable lower bound of $\Omega(n^{3/2}\sqrt{\log m})$ also holds in the smoothed model under the same assumption of m being large enough.

No interesting lower bounds for the small σ regime are known for $n \geq 3$, though the results of [28, 67, 68] suggest that the correct lower bound might be lower than current upper bounds.

From the algorithmic perspective, we describe the two phase interpolation approach of Vershynin [200], which we instantiate using two different Phase I algorithms to solve unit LPs. As a warmup, we first describe Schnalzger's application of the dimension by dimension (DD) algorithm [168], as it yields the simplest known Phase I algorithm and is not widely known. Following this, we introduce a new,

Reference	Expected Number of Pivots	Model	Algorithm
[28, 168]	$n \cdot$ max shadow size	Multiple	DD + Int. LP
[28, 30, 108]	$O(n^{2.5} m^{1/(n-1)})$	RSD, m large	DD
[200]	$O\big(\log(m)^3 (n^3\sigma^{-4} + n^5 \log(m)^2 + n^9 \log(n)^4)\big)$	Smooth	RV + Int. LP
This chapter	$O(n^2\sqrt{\log m}\ \sigma^{-2} + n^3 \log(m)^{3/2})$	Smooth	Symmetric RV + Int. LP

Table 2.2: Running time bounds. Logarithmic factors are simplified.

symmetric variant of Vershynin's random vertex (RV) algorithm which induces an artificial (degenerate) random vertex by adding $2n - 2$ inequalities placed symmetrically around a randomly chosen objective. The symmetry condition ensures that this random vertex optimizes the chosen objective with probability 1. Vershynin's original approach added n random inequalities, which only induce the optimal vertex for the chosen objective if the noise is small. Via a more careful analysis of the RV algorithm combined with the additional guarantees ensured by our variant, we derive a substantially improved complexity estimate. Specifically, our Symmetric RV algorithm takes $O(n^2\sqrt{\log m}\ \sigma^{-2} + n^3 \log(m)^{3/2})$ pivot steps, which is faster than both the original RV algorithm and Borgwardt's dimension by dimension algorithm in all parameter regimes. We defer further discussion of this to Section 2.5.

2.1.3 Techniques: Improved Shadow Bound

We now give a detailed sketch of the proof of the used shadow bound. Proofs of all claims can be found in Section 2.4. The outline of the presentation is as follows. To begin, we explain our general edge counting strategy, where we depart from the previously discussed analyses. In particular, we adapt the approach of Kelner and Spielman (KS) [121], who analyzed a smoothing model where only the right-hand side is perturbed, to the present setting. Following this, we present a parametrized shadow bound, which applies to any class of perturbations for which the relevant parameters are bounded. The main motivation of the abstraction in the parametrized model is to clearly identify the relevant properties of the perturbations we need to obtain shadow bounds. Lastly, we give the high-level idea of how we estimate the relevant quantities in the KS approach within the parametrized model.

Edge Counting Strategy The goal is to compute a bound on the expected number of edges in the polygon $Q(A) \cap W$, where W is the two-dimensional shadow plane,

$Q(A) := \text{conv}(a_1, \ldots, a_m)$ and $a_1, \ldots, a_m \in \mathbb{R}^n$ are the smoothed constraints of a unit LP. Recall that this is an upper bound on the shadow size.

In [121], Kelner and Spielman developed a very elegant and useful alternative strategy to bound the expected number of edges, which can be applied to many distributions over 2D convex polygons. Whereas they analyzed the geometry of the primal shadow polygon, the projection of $P(A)$ onto W, we will instead work with the geometry of the polar polygon $Q(A) \cap W$. The analysis begins with the following elementary identity:

$$\mathbb{E}[\text{perimeter}(Q(A) \cap W)] = \mathbb{E}[\sum_{e \in \text{edges}(Q(A) \cap W)} \text{length}(e)] \,. \tag{2.2}$$

Starting from the above identity, the approach first derives a good upper bound on the perimeter and a lower bound on the right-hand side in terms of the number of edges and the minimum edge length. The bound on the number of edges is then derived as the ratio of the perimeter bound and the minimum edge length.

We focus first on the perimeter upper bound. Since $Q(A) \cap W$ is convex, any containing circle has larger perimeter. Furthermore, we clearly have $Q(A) \cap W \subseteq \pi_W(Q(A))$, where π_W is the orthogonal projection onto W. Combining these two observations, we derive the first useful inequalities:

$$\mathbb{E}[\text{perimeter}(Q(A) \cap W)] \leq \mathbb{E}[2\pi \max_{x \in Q(A) \cap W} \|x\|] \leq \mathbb{E}[2\pi \max_{i \in [m]} \|\pi_W(a_i)\|] \,. \tag{2.3}$$

To extract the expected number of edges from the right hand side of (2.2), we first note that every edge of $Q(A) \cap W$ is derived from a facet of $Q(A)$ intersected with W (see Figure 2.2 for an illustration). Assuming non-degeneracy, the possible facets of $Q(A)$ are $F_I := \text{conv}(a_i : i \in I)$, where $I \subseteq [m]$ is any subset of size n. Let E_I denote the event that F_I induces an edge of $Q(A) \cap W$, or more precisely, that F_I is a facet of $Q(A)$ and that $F_I \cap W \neq \emptyset$. From here, we get that

$$\begin{aligned}
\mathbb{E}[\sum_{e \in \text{edges}(Q(A) \cap W)} \text{length}(e)] &= \sum_{|I|=n} \mathbb{E}[\text{length}(F_I \cap W) \mid E_I] \Pr[E_I] \\
&\geq \min_{|I|=n} \mathbb{E}[\text{length}(F_I \cap W) \mid E_I] \cdot \sum_{|I|=n} \Pr[E_I] \\
&= \min_{|I|=n} \mathbb{E}[\text{length}(F_I \cap W) \mid E_I] \cdot \mathbb{E}[|\text{edges}(Q(A) \cap W)|].
\end{aligned} \tag{2.4}$$

Combining (2.2), (2.3), (2.4), we derive the following fundamental bound:

$$\mathbb{E}[|\text{edges}(Q(A) \cap W)|] \leq \frac{\mathbb{E}[2\pi \max_{i \in [m]} \|\pi_W(a_i)\|]}{\min_{|I|=n} \mathbb{E}[\text{length}(F_I \cap W) \mid E_I]} \,. \tag{2.5}$$

In the actual proof, we further restrict our attention to potential edges having probability $\Pr[E_I] \geq 2\binom{m}{n}^{-1}$ of appearing, which helps control how extreme the conditioning on E_I can be. Note that the edges appearing with probability smaller than $2\binom{m}{n}^{-1}$ contribute at most 2 to the expected number of edges, and hence can be ignored. Thus our task now directly reduces to showing that the maximum perturbation is not too large on average, an easy condition, while ensuring that the edges that are not too unlikely to appear are reasonably long on average, the more difficult condition.

We note that applying the KS approach already improves the situation with respect to the maximum perturbation size compared to earlier analyses, as [66, 179, 200] all require a bound to hold with high probability as opposed to on expectation. For this purpose, they enforced the condition $1/\sigma \geq \sqrt{n \log m}$ (for Gaussian perturbations), which we do not require here.

Bound for Parametrized Distributions We now present the parameters of the pertubation distributions we use to obtain bounds on the numerator and denominator of 2.5. We also discuss how these parameters behave for the Gaussian and Laplace distribution.

Let us assume that $a_1, \ldots, a_m \in \mathbb{R}^n$ are independently distributed. As before we assume that the centers $a_i := \mathbb{E}[a_i]$, $i \in [m]$, have norm at most 1. We denote the perturbations by $\hat{a}_i := a_i - \bar{a}_i$, $i \in [m]$. We will assume for simplicity of the presentation that all the perturbations $\hat{a}_1, \ldots, \hat{a}_m$ are i.i.d. according to a distribution with probability density μ (in general, they could each have a distinct distribution).

At a high-level, the main properties we require of the distribution are that it be smooth and that it have sufficiently strong tail bounds. We formalize these requirements via the following 4 parameters, where we let $X \sim \mu$ below:

1. μ is an L-log-Lipschitz probability density function, that is, $|\log \mu(x) - \log \mu(y)| \leq L\|x - y\|, \forall x, y \in \mathbb{R}^n$.

2. The variance of X, when restricted to any line $l \subseteq \mathbb{R}^n$, is at least τ^2.

3. The cutoff radius $R_{m,n} > 0$ is such that $\Pr[\|X\| \geq R_{m,n}] \leq \frac{1}{n\binom{m}{n}}$.

4. The m-th *deviation* r_m is such that, for all $\theta \in \mathbb{R}^n$, $\|\theta\| = 1$, and $X_1, \ldots, X_m$ i.i.d., we have $\mathbb{E}[\max_{i \in [m]} |\langle X_i, \theta \rangle|] \leq r_m$.

We refer the reader to subsection 2.4.1 for more formal definitions of these parameters. We note that these parameters naturally arise from the proof strategy and directly expose the relevant quantities for the shadow bound.

The first two parameters are smoothness related while the last two relate to tail bounds. Using these four parameters, we will derive appropriate bounds for the numerator and denominator in (2.5). Assuming the above parameter bounds for $\hat{a}_1, \ldots, \hat{a}_m$, the main "plug and play" bound on the expected shadow size is as follows (see Theorem 2.4.10):

$$\mathbb{E}[|\text{edges}(\text{conv}(a_1, \ldots, a_m) \cap W)|] = O(\frac{n^{1.5}L}{\tau}(1 + R_{m,n})(1 + r_n)) \ . \tag{2.6}$$

We can use this parametrized bound to prove the shadow bound for Gaussian and Laplace distributed noise. For the variance σ^2 Gaussian distribution in $\mathbb{R}^n$, it is direct to verify that $\tau = \sigma$ for any line (since every line restriction results in a 1D variance σ^2 Gaussian), and from standard tail bounds for the Gaussian distribution we get that $R_{m,n} = O(\sigma\sqrt{n \log m})$ and $r_m = O(\sigma\sqrt{\log m})$. The only parameter that cannot be bounded directly is the log-Lipschitz parameter L, since $\|x/\sigma\|^2/2$, the log of the Gaussian density, is quadratic. For Laplace distributed perturbations however, this last difficulty is completely avoided. Here a comparably sized Laplace perturbation (i.e. same expected norm) has density proportional to $e^{-(\sqrt{n}/\sigma)\|x\|}$ which is by definition log-Lipshitz with $L = \sqrt{n}/\sigma$. The other parameters are somewhat worse, it can be shown that $R_{m,n} = O(\sigma\sqrt{n} \log m)$, $r_m = O(\sigma \log m)$ and $\tau \geq \sigma/\sqrt{n}$, where in particular τ is a $\sqrt{n}$-factor smaller than the Gaussian. Thus, for Laplace perturbations the parametrized bound applies directly and yields a bound of $O(n^{2.5}\sigma^{-2})$ in the small σ regime.

To apply this analysis to the Gaussian setting, we start with the fact, noted in all prior analyses, that the Gaussian is locally smooth within any fixed radius. In particular, within radius $R_{m,n}$ of the mean, the Gaussian density is $O(\sqrt{n \log m}/\sigma)$-log-Lipschitz. As events that happen with probability $\ll \binom{m}{n}^{-1}$ have little effect on the expected shadow bound (recall that the shadow is always bounded by $\binom{m}{n}$), one can hope to condition on each perturbation living inside the $R_{m,n}$ radius ball. This is in fact the approach taken in the prior analyses [66, 179, 200]. This conditioning however does not ensure full log-Lipshitzness and causes problems for points near the boundary. Furthermore, the conditioning may also decrease line variances for lines near the boundary.

To understand why this is problematic, we note that the main role of the smoothness parameters L and τ is to ensure enough "wiggle-room" to guarantee that edges induced by any fixed basis are long on expectation. Using the above conditioning, it is clear that edges induced by facets whose perturbations occur close to the $R_{m,n}$ boundary must be dealt with carefully. To avoid such difficulties altogether, we leverage the local log-Lipshitzness of the Gaussian in a "smoother" way. Instead of conditioning, we simply replace the Gaussian density with a globally $O(\sqrt{n \log m}/\sigma)$-log-Lipshitz density which has statistical distance $\ll \binom{m}{n}^{-1}$ to the Gaussian (thus preserving the

shadow bound) and also yields nearly identical bounds for the other parameters. This distribution will consist of an appropriate gluing of a Gaussian and Laplace density, which we call the Laplace-Gaussian distribution (see Section 2.4.3 for details). Thus, by slightly modifying the distribution, we are able to use the parametrized model to obtain shadow bounds for Gaussian perturbations in a black box manner.

Bounding the Perimeter and Edge Length We now briefly describe how the perimeter and minimum edge length in (2.5) are bounded in the parametrized perturbation model to obtain (2.6). As this is the most technical part of the analysis, we refer the reader to the proofs in Section 2.4 and give only a very rough discussion here. As above, we will assume that the perturbations satisfy the bounds given by $L, \tau, R_{m,n}, r_m$.

For the perimeter bound, we immediately derive the bound

$$\mathbb{E}[\max_{i\in[m]} \|\pi_W(a_i)\|] \leq 1 + \mathbb{E}[\max_{i\in[m]} \|\pi_W(\hat{a}_i)\|] \leq 1 + 2r_m$$

by the triangle inequality. From here, we must bound the minimum expected edge length, which requires the majority of the work. For this task, we provide a clean analysis, which shares high-level similarities with the Spielman and Teng distance lemma, though our task is simpler. Firstly, we only need to show that an edge is large on average, whereas the distance lemma has the more difficult task of proving that an edge is unlikely to be small. Second, the conditioning is much milder. Namely, the distance lemma conditions a facet F_I on intersecting a specified ray r_θ, whereas we only condition F_I on intersecting W. This conditioning gives the edge much more "wiggle room", and is the main leverage we use to get the factor n improvement.

Let us fix $F := F_{[n]} = \mathrm{conv}(a_1, \ldots, a_n)$ as the potential facet of interest, under the assumption that $E := E_{[n]}$, i.e., that F induces an edge of $Q(A) \cap W$, has probability at least $2\binom{m}{n}^{-1}$. The analysis of the edge length conditioned on E proceeds as follows:

1. Show that if F induces an edge, then under this conditioning F has *small diameter* with good probability, namely its vertices are all at distance at most $O(1+R_{m,n})$ from each other (Lemma 2.4.17). This uses the tailbound defining $R_{m,n}$ and the fact that E occurs with non-trivial probability.

2. Condition on F being a facet of $Q(A)$ by fixing its containing affine hyperplane H (Lemma 2.4.19). This is standard and is achieved using a change of variables known as a Blaschke-Petkantschin formula (see Section 2.2 for details).

3. Let $l := H \cap W$ denote the line which intersects F to form an edge of $Q(A) \cap W$. Show that on average the longest chord of F parallel to l is long. We achieve

the bound $\Omega(\tau/\sqrt{n})$ (Lemma 2.4.27) using that the vertices of F restricted to lines parallel to l have variance at least τ^2.

4. Show that on average F is pierced by l through a chord that is not too much shorter than the longest one. Here we derive the final bound on the expected edge length of

$$\mathbb{E}[\text{length}(F \cap W) \mid E] = \Omega((\tau/\sqrt{n}) \cdot 1/(nL(1 + R_{m,n}))) \text{ (Lemma 2.4.26)}$$

using the fact that the distribution of the vertices is L-log-Lipschitz and that F has diameter $O(1 + R_{m,n})$.

This concludes the high-level discussion of the proof.

2.1.4 Related work

Structured Polytopes An important line of work has been to study LPs with good geometric or combinatorial properties. Much work has been done to analyze primal and dual network simplex algorithms for fundamental combinatorial problems on flow polyhedra such as bipartite matching [110], shortest path [70, 97], maximum flow [92, 95] and minimum cost flow [96, 156, 157]. Generalizing on the purely combinatorial setting, LPs where the constraint matrix $A \in \mathbb{Z}^{m \times n}$ is totally unimodular (TU), i.e. the determinant of any square submatrix of A is in $\{0, \pm 1\}$, were analyzed by Dyer and Frieze [75], who gave a random walk based simplex algorithm which requires $\text{poly}(n, m)$ pivots. Recently, an improved random walk approach was given by Eisenbrand and Vempala [78], which works in the more general setting where the subdeterminants are bounded in absolute value by Δ, who gave an $O(\text{poly}(n, \Delta))$ bound on the number of Phase II pivots (note that there is no dependence on m). Furthermore, randomized variants of the shadow vertex algorithm were analyzed in this setting by [36, 47], where in particular [47] gave an expected $O(n^5\Delta^2 \log(n\Delta))$ bound on the number of Phase I and II pivots. Another interesting class of structured polytopes comes from the LPs associated with MDP, where simplex rules such as Dantzig's most negative reduced cost correspond to variants of policy iteration. Ye [205] gave polynomial bounds for Dantzig's rule and Howard's policy iteration for MDPs with a fixed discount rate, and Ye and Post [159] showed that Dantzig's rule converges in strongly polynomial time for deterministic MDPs with variable discount rates.

2.1.5 Organization

Section 2.2 contains basic definitions and background material. We study the smoothed complexity of the convex hull in two dimensions in Section 2.3. Following a similar strategy, the proofs of the shadow bound are given in Section 2.4. The details regarding the two phase shadow vertex algorithm we use, which rely in an almost black box way on the shadow bound, are presented in Section 2.5.

2.2 Preliminaries

Gaussian distribution

Definition 2.2.1. The *Gaussian distribution* or *normal distribution* $N_n(\bar{a}, \sigma)$ in n variables with mean $\bar{a}$ and standard deviation σ has density $(2\pi)^{-n/2} e^{-\|x-\bar{a}\|^2/(2\sigma^2)}$. We abbreviate $N_n(\sigma) = N_n(\vec{0}, \sigma)$.

Important facts about the Gaussian distribution include:

- Given a k-dimensional affine subspace $W \subseteq \mathbb{R}^n$, if X is $N_n(\bar{a}, \sigma)$-distributed then both the orthogonal projection $\pi_W(X)$ and the restriction of X to W are $N_k(\pi_W(\bar{a}), \sigma)$-distributed in W.
- For $X \sim N_n(\bar{a}, \sigma)$ we have $\mathbb{E}[X] = \bar{a}$ and $\mathbb{E}[((X - \bar{a})^\mathsf{T}\theta)^2] = \sigma^2$ for all $\theta \in \mathbb{S}^{n-1}$.
- The expected squared distance to the mean is $\mathbb{E}[\|X - \bar{a}\|^2] = n\sigma^2$.
- The moment generating function of $X \sim N_1(0, \sigma)$ is $\mathbb{E}[e^{\lambda X}] = e^{\lambda^2\sigma^2/2}$, for all $\lambda \in \mathbb{R}$, and that of X^2 is $\mathbb{E}[e^{\lambda X^2}] = 1/\sqrt{1 - 2\lambda\sigma}$ for $\lambda < 1/(2\sigma)$.

We will need the following tail bound for Gaussian random variables. We include a proof for completeness.

Lemma 2.2.2 (Gaussian tail bounds)**.** *For $X \in \mathbb{R}^n$ distributed as $N_n(\vec{0}, \sigma)$, $t \geq 1$,*

$$\Pr[\|X\| \geq t\sigma\sqrt{n}] \leq e^{-(n/2)(t-1)^2}. \tag{2.7}$$

For $\theta \in \mathbb{S}^{n-1}$ and $t \geq 0$,

$$\Pr[|X^\mathsf{T}\theta| \geq t\sigma] \leq 2e^{-t^2/2}. \tag{2.8}$$

Proof. By homogeneity, we may without loss of generality assume that $\sigma = 1$.

Proof of (2.7)

$$\begin{aligned}
\Pr[\|X\| \geq \sqrt{n}t] &= \min_{\lambda\in(0,1/2)} \Pr[e^{\lambda\|X\|^2} \geq e^{\lambda t^2 n}] \\
&\leq \min_{\lambda\in(0,1/2)} \mathbb{E}[e^{\lambda\|X\|^2}]e^{-\lambda t^2 n} \qquad \text{(Markov's inequality)} \\
&= \min_{\lambda\in(0,1/2)} \left(\prod_{i=1}^{n} \mathbb{E}[e^{\lambda X_i^2}]\right) e^{-\lambda t^2 n} \qquad \text{(Independence of coefficients)} \\
&= \min_{\lambda\in(0,1/2)} \left(\frac{1}{1-2\lambda}\right)^{n/2} e^{-\lambda t^2 n} \\
&\leq e^{-(n/2)(t^2-2\log t-1)} \qquad \text{(setting } \lambda = \frac{1}{2}(1-1/t^2)\text{)} \\
&\leq e^{-(n/2)(t-1)^2} \qquad \text{(since } \log t \leq t-1 \text{ for } t \geq 1\text{).}
\end{aligned}$$

Proof of (2.8)

$$\begin{aligned}
\Pr[|X^\mathsf{T}\theta| \geq t] &= 2\Pr[X^\mathsf{T}\theta \geq t] \\
&\leq 2\min_{\lambda>0} \mathbb{E}[e^{\lambda X^\mathsf{T}\theta}]e^{-\lambda t} \\
&= 2\min_{\lambda>0} e^{\lambda^2/2-\lambda t} \leq 2e^{-t^2/2}\text{, setting } \lambda = t.
\end{aligned}$$

□

Laplace distribution Our shadow bounds will hold for a general class of distributions with bounds on certain parameters. We illustrate this for the n-dimensional Laplace distribution.

Definition 2.2.3. The *Laplace distribution* $L_n(\bar{a}, \sigma)$ or *exponential distribution* in $\mathbb{R}^n$ with mean vector $\bar{a}$ has probability density function

$$\frac{\sqrt{n}^n}{(n-1)!\sigma^n \mathrm{vol}_{n-1}(\mathbb{S}^{n-1})} e^{-\|x-\bar{a}\|\sqrt{n}/\sigma}.$$

We abbreviate $L_n(\sigma) = L_n(\vec{0}, \sigma)$. We have normalized the distribution to have expected norm $\sqrt{n}\sigma$. Additionally, the variance along any direction is $\sigma^2(1+\frac{1}{n})$.

The norm of a Laplace distributed random variable follows a *Gamma distribution*.

Definition 2.2.4. The Gamma distribution $\Gamma(\alpha, \beta)$, $\alpha \in \mathbb{N}, \beta \in \mathbb{R}$, on the non-negative real numbers has probability density $\frac{\beta^\alpha}{(\alpha-1)!}t^{\alpha-1}e^{-\beta t}$. The moment generating function of the Gamma distribution is $\mathbb{E}_{X\sim\Gamma(\alpha,\beta)}[e^{\lambda X}] = (1-\lambda/\beta)^{-\alpha}$ for $\lambda < \beta$.

One can generate a n-dimensional Laplace distribution $L_n(\sigma)$ as the product of an independent scalar and vector. The vector θ is sampled uniformly from the sphere $\mathbb{S}^{n-1}$. The scalar $s \sim \Gamma(n, \sqrt{n}/\sigma)$ is sampled from the Gamma distribution. The product $s\theta$ has a $L_n(\sigma)$-distribution.

We will need the following tail bound for Laplace distributed random variables. We include a proof for completeness.

Lemma 2.2.5 (Laplace tail bounds)**.** *For $X \in \mathbb{R}^n$, $n \geq 2$, distributed as $(\vec{0}, \sigma)$-Laplace and $t \geq 1$,*

$$\Pr[\|X\| \geq t\sigma\sqrt{n}] \leq e^{-n(t-\log t-1)} . \tag{2.9}$$

In particular, for $t \geq 2$,

$$\Pr[\|X\| \geq t\sigma\sqrt{n}] \leq e^{-nt/7} . \tag{2.10}$$

For $\theta \in \mathbb{S}^{n-1}$, $t \geq 0$,

$$\Pr[|X^{\mathsf{T}}\theta| \geq t\sigma] \leq \begin{cases} 2e^{-t^2/16} : 0 \leq t \leq 2\sqrt{n} \\ e^{-\sqrt{n}t/7} : t \geq 2\sqrt{n} \end{cases} . \tag{2.11}$$

Proof. By homogeneity, we may without loss of generality assume that $\sigma = 1$.

Proof of (2.9)

$$\begin{aligned} \Pr[\|X\| \geq \sqrt{n}t] &= \min_{\lambda \in (0,\sqrt{n})} \Pr[e^{\lambda\|X\|} \geq e^{\lambda\sqrt{n}t}] \\ &\leq \min_{\lambda \in (0,\sqrt{n})} \mathbb{E}[e^{\lambda\|X\|}]e^{-\lambda\sqrt{n}t} \qquad \text{(Markov's inequality)} \\ &\leq \min_{\lambda \in (0,\sqrt{n})} (1-\lambda/\sqrt{n})^{-n} e^{-\lambda\sqrt{n}t} \\ &= e^{-n(t-\log t-1)} \text{ , setting } \lambda = \sqrt{n}(1-1/t). \end{aligned}$$

For the case $t \geq 2$, the desired inequality follows from the fact that $t - \log t - 1 \geq t/7$ for $t \geq 2$, noting that $(t - \log t - 1)/t$ is an increasing function on $t \geq 1$.

Proof of (2.11) For $t \geq 2\sqrt{n}$, we directly apply equation (2.10):

$$\Pr[|X^{\mathsf{T}}\theta| \geq t\sigma] \leq \Pr[\|X\| \geq t\sigma] \leq e^{-\sqrt{n}t/7}.$$

For $t \leq 2\sqrt{n}$, express $X = s \cdot \omega$ for $s \sim \Gamma(n, \sqrt{n}/\sigma)$, $\omega \in \mathbb{S}^{n-1}$ uniformly sampled.

$$\begin{aligned} \Pr[|s\omega^{\mathsf{T}}\theta| \geq t\sigma] &\leq \Pr[|\omega^{\mathsf{T}}\theta| \geq t/(2\sqrt{n})] + \Pr[|s| \geq 2\sqrt{n}\sigma] \\ &\leq \Pr[|\omega^{\mathsf{T}}\theta| \geq t/(2\sqrt{n})] + e^{-n/4}. \end{aligned}$$

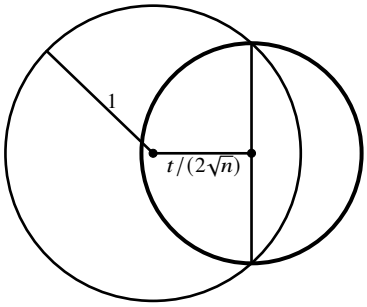


Figure 2.1: The small sphere has at least as much surface area as combined surface area of the enclosed sphere cap and the opposite cap together by the monotonicity of surface area (Lemma 1.5.1).

For the first term we follow [12, Lemma 2.2], where the second line is illustrated in Figure 2.1:

$$\begin{aligned}
\Pr[|\omega^{\mathsf{T}}\theta| \geq t/(2\sqrt{n})] &= \frac{\mathrm{vol}_{n-1}(\{\omega \in \mathbb{S}^{n-1} : |\omega^{\mathsf{T}}\theta| \geq t/(2\sqrt{n})\})}{\mathrm{vol}_{n-1}(\mathbb{S}^{n-1})} \\
&\leq \frac{\mathrm{vol}_{n-1}(\sqrt{1-\frac{t^2}{4n}}\mathbb{S}^{n-1})}{\mathrm{vol}_{n-1}(\mathbb{S}^{n-1})} \\
&= (1-\frac{t^2}{4n})^{(n-1)/2} \\
&\leq e^{-t^2(n-1)/(8n)} \leq e^{-t^2/16}.
\end{aligned}$$

The desired conclusion follows since $e^{-t^2/16} + e^{-n/4} \leq 2e^{-t^2/16}$ for $0 \leq t \leq 2\sqrt{n}$. □

Change of variables In Section 2.4 we make use of a change of variables that known as a Blaschke–Petkantschin formula and is standard in the study of convex hulls, see, e.g., [169].

Recall that a change of variables affects a probability distribution. Let the vector $y \in \mathbb{R}^n$ be a random variable with density μ. If $y = \varphi(x)$ and φ is invertible, then the induced density on x is

$$\mu(\varphi(x))\left|\det\left(\frac{\partial\varphi(x)}{\partial x}\right)\right|,$$

where $\left|\det\left(\frac{\partial\varphi(x)}{\partial x}\right)\right|$ is the Jacobian of φ. We describe a particular change of variables which has often been used for studying convex hulls, and, in particular, by Borgwardt [28] and Spielman and Teng [179] for deriving shadow bounds.

For affinely independent vectors $a_1, \ldots, a_n \in \mathbb{R}^n$ we have the coordinate transformation

$$(a_1, \ldots, a_n) \mapsto (\theta, t, b_1, \ldots, b_n),$$

where $\theta \in \mathbb{S}^{n-1}$ and $t \geq 0$ satisfy $\theta^\mathsf{T} a_i = t$ for every $i \in [n]$ and the vectors $b_1, \ldots, b_n \in \mathbb{R}^{n-1}$ parametrize the positions of $a_1, \ldots, a_n$ within the hyperplane $\{x \in \mathbb{R}^n \mid \theta^\mathsf{T} x = t\}$. We coordinatize the hyperplanes as follows:

Fix a reference vector $v \in \mathbb{S}^{n-1}$, and pick an isometric embedding $h : \mathbb{R}^{n-1} \to v^\perp$. For any unit vector $\theta \in \mathbb{S}^{n-1}$, define the map $R'_\theta : \mathbb{R}^n \to \mathbb{R}^n$ as the unique map that rotates v to θ along $\operatorname{span}(v, \theta)$ and fixes the orthogonal subspace $\operatorname{span}(v, \theta)^\perp$. We define $R_\theta = R'_\theta \circ h$. The change of variables from $\theta \in \mathbb{S}^{n-1}, t > 0, b_1, \ldots, b_n \in \mathbb{R}^{n-1}$ to $a_1, \ldots, a_n$ takes the form

$$(a_1, \ldots, a_n) = (R_\theta b_1 + t\theta, \ldots, R_\theta b_n + t\theta).$$

The change of variables as specified above is not uniquely defined when $a_1, \ldots, a_n$ are affinely dependent, when $t = 0$ or when $\theta = -v$.

Theorem 2.2.6. *Let $\theta \in \mathbb{S}^{n-1}$ be a unit vector, $t > 0$ and $b_1, \ldots, b_n \in \mathbb{R}^{n-1}$. Consider the map*

$$(\theta, t, b_1, \ldots, b_n) \mapsto (a_1, \ldots, a_n) = (R_\theta b_1 + t\theta, \ldots, R_\theta b_n + t\theta).$$

The Jacobian of this map equals

$$\left| \det \left(\frac{\partial \varphi(x)}{\partial x} \right) \right| = (n-1)! \operatorname{vol}_{n-1}(\operatorname{conv}(b_1, \ldots, b_n)).$$

2.2.1 Shadow vertex algorithm

We briefly introduce the shadow vertex algorithm. An alternative exposition about the shadow vertex algorithm can be found in [28].

Let $P(A, b) = \{x \in \mathbb{R}^n : Ax \leq b\}$ be a polyhedron, and let $a_1, \ldots, a_m \in \mathbb{R}^n$ correspond to the rows of A. We call a set $B \subseteq [m]$ a basis of $Ax \leq b$ if A_B is invertible. This implies that $|B| = n$. We say B is a feasible basis if $x_B = A_B^{-1} b_B$ satisfies $Ax_B \leq b$. Such a point x_B is always a vertex of $P(A, b)$. We say a feasible basis B is optimal for an objective $c \in \mathbb{R}^n$ if $c^\mathsf{T} A_B^{-1} \geq \vec{0}$, which happens if and only if $\max_{x \in P} c^\mathsf{T} x = c^\mathsf{T} x_B$.

The shadow vertex algorithm is a pivot rule for the simplex method. Given a feasible basis $B \subseteq [m]$, an objective $d \in \mathbb{R}^n$ for which B is optimal, and an objective function $c \in \mathbb{R}^n$ to optimize, where c and d are linearly independent, the shadow vertex algorithm (Algorithm 1) specifies which pivot steps to take to reach an optimal basis for c. We note that there are many possible choices for starting objective d.

Algorithm 1 Shadow vertex algorithm for non-degenerate polyhedron and shadow.

Input: $P(A,b) = \{x \in \mathbb{R}^n : Ax \leq b\}$, $c, d \in \mathbb{R}^n$, feasible basis $B \subseteq [m]$ optimal for d.

Output: Return optimal basis $B \subseteq [m]$ for c or *unbounded*.

$\lambda_0 \leftarrow 0$.
$i \leftarrow 0$.
loop
  $i \leftarrow i + 1$.
  $\lambda_i :=$ maximum $\lambda \leq 1$ such that $c_\lambda^\intercal A_B^{-1} \geq \vec{0}$.
  if $\lambda_i = 1$ **then return** B.
  $k := k \in B$ such that $(c_{\lambda_i}^\intercal A_B^{-1})_k = 0$.
  $x_B := A_B^{-1} b_B$.
  $s_i :=$ supremum $s > 0$ such that $A(x_B - sA_B^{-1}e_k) \leq b$.
  if $s_i = \infty$ **then return** *unbounded*.
  $j := j \in [m] - B$ such that $a_j{}^\intercal (x_B - s_i A_B^{-1} e_k) = b_j$.
  $B \leftarrow B \cup \{j\} \setminus \{k\}$.

We parametrize $c_\lambda := (1-\lambda)d + \lambda c$ and start at $\lambda = 0$. The shadow vertex rule increases λ until there are $j \neq k \in [m]$ such that a new feasible basis $B \cup \{j\} \setminus \{k\}$ is optimal for c_λ, and repeats with increased λ and new basis B until $\lambda = 1$.

The index $k \in B$ is such that the coordinate for k in $c_\lambda^\intercal A_B^{-1}$ first lowers to 0, and $j \notin B$ is such that $B \cup \{j\} \setminus \{k\}$ is a feasible basis: we follow the edge $A_B^{-1} b_B - A_B^{-1} e_k \mathbb{R}_+$ until we hit the first constraint $a_j^\intercal x \leq b_j$, and then replace k by j to get the new basis $B \cup \{j\} \setminus \{k\}$.

Changing the current basis from B to $B \cup \{j\} \setminus \{k\}$ is called a pivot step. As soon as $\lambda = 1$ we have $c_\lambda = c$, at which moment the current basis is optimal for our objective c. If at some point no choice of j exists, then an unbounded ray has been found.

Definition 2.2.7. We say that the system $Ax \leq b$ is *non-degenerate* if $m \geq n$, any $B \in \binom{[m]}{n}$ is a basis, and every vertex of the corresponding polyhedron $P(A,b)$ is tight at exactly n linearly independent inequalities. When the description $Ax \leq b$ is clear, we say that $P = P(A,b)$ is non-degenerate to mean that its describing system is.

Definition 2.2.8. We say that the shadow of a polyhedron P on a two-dimensional linear subspace W is *non-degenerate* if $\dim(\pi_W(P)) = 2$ and for every face F of P such that $\pi_W(F)$ is a face of $\pi_W(P)$ and $\dim(\pi_W(F)) \leq 1$, we have that $\dim(\pi_W(F)) = \dim(F)$.

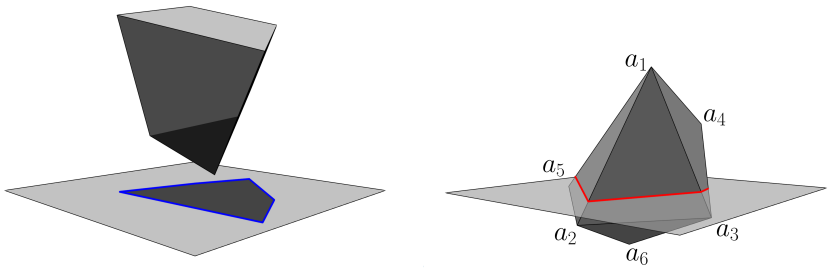


Figure 2.2: On the left, a polytope and its shadow. On the right, the corresponding polar polytope intersected with the plane. There are as many edges marked blue as there are edges marked red.

If both the polyhedron and the shadow are non-degenerate, each pivot step can be performed in $O(mn)$ time (see the pseudo-code for Algorithm 1). Under the distribution models we examine, degeneracy occurs with probability 0.

The shadow vertex rule is called as such because the visited vertices are in correspondence with vertices on the relative boundary of the orthogonal projection $\pi_W(P)$ of P onto $W = \text{span}(d, c)$, where we denote $\pi_W(P)$ as the shadow of P on W. See the left half of Figure 2.2. We call the total number of vertices of the projection the *shadow size*, and it is the key geometric estimate in our analysis of the simplex method.

Lemma 2.2.9. *For a polyhedron P having a non-degenerate shadow on W, the vertices of P optimizing objectives in $W \setminus \{\vec{0}\}$ are in one-to-one correspondence with the vertices of $\pi_W(P)$ under the map π_W.*

Proof. For every vertex $y \in \pi_W(P)$ of the shadow, there is some $\theta \in W - \{\vec{0}\}$ that is uniquely optimized by y over $\pi_W(P)$, i.e., y is optimal for θ, and $\theta^\intercal y > \theta^\intercal y'$ for all $y' \in \pi_W(P), y' \neq y$. Let $F \subseteq P$ be the face of optimizers of θ. It must satisfy $\pi_W(F) = y$, and by non-degeneracy of the shadow, F is a vertex.

For a vertex $x \in P$ optimizing $\theta \in W - \{\vec{0}\}$, its projection $\pi_W(x)$ optimizes θ over the projection $\pi_W(P)$, and hence it lies on the relative boundary of the shadow. If $\pi_W(x)$ is a vertex of $\pi_W(P)$, we are done, so suppose not. Then $\pi_W(x)$ must lie in the relative interior of an edge e of $\pi_W(P)$. We can lift e to $F = \pi_W^{-1}(e) \cap P$, the largest face of P satisfying $\pi_W(F) = e$. As the shadow is non-degenerate, F is an edge of P. Since $\dim(F) = 1$ and $\pi_W(x)$ is in the relative interior of e, we see that x is in the relative interior of F, contradicting our assumption that x is a vertex of P.

From the above arguments, we see that the map π_W yields a bijection from the set of vertices of P optimizing some objective in $W - \{\vec{0}\}$ and the set of vertices of

$\pi_W(P)$. The lemma thus follows. □

We will consider non-degenerate polyhedra of the form $P(A) = \{x \in \mathbb{R}^n : Ax \leq \vec{1}\}$, in which case $\vec{0}$ is always contained in the polyhedron. The problem thus has a known feasible solution. We will look at the geometry of shadow paths on such polyhedra from a *polar perspective*. For any non-degenerate polyhedron $P(A)$, we look at the polar polytope, defined as the convex hull $Q(A) := \text{conv}(a_1, \ldots, a_m)$ of the rows of A. For any index-set $I \in \binom{[m]}{n}$, if the (unique) solution x_I to the equations

$$a_i{}^{\mathsf{T}} x = 1 \qquad \forall i \in I$$

is a vertex of the original polyhedron $P(A)$, then the set $\text{conv}(a_i : i \in I)$ forms a facet of the polytope $Q(A)$. Conversely, if $\text{conv}(a_i : i \in I)$ induces a facet of $Q' := \text{conv}(\vec{0}, a_1, \ldots, a_m)$ (note the inclusion of $\vec{0}$), then x_I is a vertex of $P(A)$. The addition of $\vec{0}$ to the polar of $P(A)$ allows us to detect unboundedness. Precisely, the facets of the extended polar Q' containing $\vec{0}$ are in one to one correspondence with unbounded edges of $P(A)$. Moreover, $P(A)$ is bounded, i.e. a polytope, if and only if $\vec{0}$ is in the interior of $Q(A)$. In this case $Q(A) = Q'$, and hence every facet of $Q(A)$ is associated to a vertex of $P(A)$.

In the polar perspective, a pivot step moves from one facet of Q' to a neighboring facet. The shadow vertex algorithm moves the objective c_λ along the line segment $[d, c]$ and keeps track of which facet of Q' is intersected by the ray $c_\lambda \mathbb{R}^+$. If we move to a facet of Q' containing $\vec{0}$, we may conclude that the LP with objective c is in fact unbounded. Since we can only visit such facets at the end of a shadow path, we will be able to control the length of shadow paths using only the geometry of $Q(A)$, which will help simplify our analyses. The main bound on the size of the shadow we will use is given in the following lemma.

Lemma 2.2.10. *Let $P(A)$ be a non-degenerate polyhedron with a non-degenerate shadow on W. Then*

$$|\text{vertices}(\pi_W(P(A)))| \leq |\text{edges}(Q(A) \cap W)|.$$

Proof. Let x' be a vertex of $\pi_W(P)$. To prove the statement, we will associate x' with a unique edge of $Q(A) \cap W$. By non-degeneracy of the shadow and Lemma 2.2.9, there exists a unique vertex x of $P(A)$ such that $\pi_W(x) = x'$. By non-degeneracy of $P(A)$, there exists a unique basis $I \in \binom{[m]}{n}$ such that $x = x_I := A_I^{-1}\vec{1}_I$. By virtue of x' being a vertex of $\pi_W(P(A))$, since $\dim(\pi_W(P(A))) = 2$ there must an objective $\theta \in W - \{\vec{0}\}$ that is uniquely maximized by x' on $\pi_W(P(A))$. Consequently, by construction x also uniquely maximizes θ over $P(A)$. We now show that $\theta^{\mathsf{T}} A_I^{-1} > \vec{0}$. Firstly, since x_I maximizes θ, by non-degeneracy of $P(A)$, we must have that $\theta^{\mathsf{T}} A_I^{-1} \geq \vec{0}$. Now

assume for the sake of contradiction that $(\theta^\mathsf{T} A_I^{-1})_i = 0$ for some $i \in [n]$. In this case, by non-degeneracy of $P(A)$, we can pivot such that i leaves the basis and find an edge of $P(A)$ that is optimal for θ, and thus that also projects to x'. This violates non-degeneracy of the shadow, and so we must have $\theta^\mathsf{T} A_I^{-1} > \vec{0}$ as claimed.

The condition that $\theta^\mathsf{T} A_I^{-1} > \vec{0}$ is equivalent to the existence of $\gamma_i > 0$, for $i \in I$, and $\alpha > 0$, such that

$$\sum_{i \in I} \gamma_i a_i = \theta, \qquad \sum_{i \in I} \gamma_i = \alpha,$$

which is equivalent to

$$\sum_{i \in I} (\gamma_i / \alpha) a_i = \alpha^{-1} \theta, \qquad \sum_{i \in I} \gamma_i / \alpha = 1.$$

The ray $\theta \mathbb{R}_+$ intersects $\mathrm{conv}(a_i)_{i \in I}$. Since the inequalities are strict, there is a neighborhood around θ in which every vector θ' induces a ray intersecting $\mathrm{conv}(a_i)_{i \in I}$. In particular, we get that $\theta \mathbb{R}_+$ intersects the relative interior of $\mathrm{conv}(a_i)_{i \in I}$.

Recalling that x_I is a vertex of a non-degenerate polytope $P(A)$ and that $b = \vec{1}$, we have that ${a_i}^\mathsf{T} x_I \leq 1$ for all $i \leq m$, where equality holds iff $i \in I$. Therefore the points in $\mathrm{conv}(a_i)_{i \in I}$ are the maximizers of x_I, when maximizing over $Q(A)$. This means that $\mathrm{conv}(a_i)_{i \in I}$ forms a facet of $Q(A)$. As $\mathrm{conv}(a_i)_{i \in I}$ is intersected by $\theta \mathbb{R}_+$ in its relative interior, the intersection $\mathrm{conv}(a_i)_{i \in I} \cap W$ can not be zero-dimensional. If $\mathrm{conv}(a_i)_{i \in I} \cap W$ were two-dimensional, the origin would be contained in the affine hull of $\mathrm{conv}(a_i)_{i \in I}$ and hence $0 = \vec{0}^\mathsf{T} x_I = 1$. Thus $\mathrm{conv}(a_i)_{i \in I} \cap W$ must be one-dimensional and hence an edge of $Q(A) \cap W$.

For a feasible basis I, the facet $\mathrm{conv}(a_i)_{i \in I}$ uniquely determines x_I and hence x'. If two facets $\mathrm{conv}(a_i)_{i \in I}$ and $\mathrm{conv}(a_j)_{j \in J}$ of Q, where I and J are feasible bases, induce the same edge e of $Q \cap W$, then e passes through the relative interior of both facets by the argument in the previous paragraph. If two faces intersect in their relative interior, they must be equal. For suppose not. Then we could, without loss of generality, find some $x \in \mathrm{conv}(a_i)_{i \in I}$ and y in the relative interior of $\mathrm{conv}(a_i)_{i \in I} \cap \mathrm{conv}(a_j)_{j \in J}$. Then there is some $0 < \lambda < 1$ such that $\lambda x + (1 - \lambda) y \in \mathrm{conv}(a_j)_{j \in J}$, so $x \in \mathrm{conv}(a_j)_{j \in J}$. Since x was arbitrary, $\mathrm{conv}(a_j)_{j \in I} = \mathrm{conv}(a_j)_{j \in J}$. By non-degeneracy of $P(A)$, equality of $\mathrm{conv}(a_i)_{i \in I}$ and $\mathrm{conv}(a_j)_{j \in J}$ implies that $I = J$. We thus conclude that the above mapping from vertices of $\pi_W(P(A))$ to edges of $Q \cap W$ is injective. □

The number of pivot steps taken in a shadow path is bounded from above by the number of edges in the intersection $Q(A) \cap \mathrm{span}(d, c)$. Hence it suffices that we prove an upper bound on this geometric quantity. The following theorem summarizes the properties we will use of the shadow vertex algorithm.

Theorem 2.2.11. *Let $P(A, b)$ denote a non-degenerate polyhedron. Let $c, d \in \mathbb{R}^n$ denote two objectives inducing a non-degenerate shadow and let $W = \mathrm{span}(d, c)$. Given*

a feasible basis $I \in \binom{[m]}{n}$ for $Ax \leq b$ which is optimal for d, Algorithm 1 (*shadow vertex*) *finds a feasible basis $J \in \binom{[m]}{n}$ optimal for c or declares unboundedness in a number of pivot steps bounded by $|\text{vertices}(\pi_W(P(A)))|$, where π_W is the orthogonal projection onto W. In particular, when $b = \vec{1}$, the number of pivots is at most*

$$|\text{edges}(Q(A) \cap W)| \,.$$

Proof. We first establish that λ_i and s_i are well-defined. We know that the set of $\lambda \leq 1$ such that $c_\lambda^\intercal A_B^{-1} \geq \vec{0}$ is non-empty, because λ_{i-1} satisfies this property. The set of $s > 0$ such that $A(x_B - sA_B^{-1}e_k) \leq b$ is non-empty because if $A(x_B - sA_B^{-1}e_k) \not\leq b$ for all $s > 0$ then the vertex x_B would be tight at more than d inequalities, contradicting non-degeneracy of $P(A)$.

To show correctness, we prove that, in every iteration of the loop, the basis B is feasible, that the algorithm terminates, and that the output is correct. For this we first prove that in every iteration x_B is a shadow vertex.

If $\lambda_i \neq 1$, there must be some index k such that $(c_{\lambda_i}^\intercal A_B^{-1})_k = 0$. Since each of the points $(x_B - sA_B^{-1}e_k)$ with $s \leq s_i$ is feasible, if the index j exists then $B \cup \{j\} - \{k\}$ is a feasible basis. We know that c_{λ_i} is a non-negative combination of rows of $A_{B-\{k\}}$, so it is also a non-negative combination of rows of $A_{B\cup\{j\}-\{k\}}$. Thus $B \cup \{j\} - \{k\}$ is optimal for c_{λ_i}, and hence $x_{B\cup\{j\}-\{k\}}$ is a shadow vertex by Lemma 2.2.9, using the non-degeneracy of the shadow.

From our non-degeneracy conditions it follows that k is unique. Suppose it were not. The set $P(A) \cap \{x : A_{B-\{k,k'\}}x = b_{B-\{k,k'\}}\}$ would be a two-dimensional face of $P(A)$ optimizing the objective c_{λ_i}, and hence this face would project to an at most one-dimensional face of $\pi_W(P(A))$. This contradicts non-degeneracy of the shadow, so k is unique.

Now we prove that the algorithm has the runtime that was claimed. We do this by proving that no shadow vertex is visited by the algorithm twice.

Since at the start of a loop $c_{\lambda_i}^\intercal A_B^{-1} \geq \vec{0}$, we know from maximality of λ_{i+1} that $\lambda_{i+1} \geq \lambda_i$. Suppose that $\lambda_{i+1} = \lambda_i < 1$ for $i \geq 1$. This implies that

$$\max\{\lambda : c_\lambda^\intercal A_B^{-1} \geq \vec{0}\} = \max\{\lambda : c_\lambda^\intercal A_{B\cup\{j\}-\{k\}}^{-1} \geq \vec{0}\}.$$

Either one of the two sets is a singleton set or we find linearly independent objectives $\theta, \theta' \in W - \{\vec{0}\}$ that are both optimized by both x_B and $x_{B\cup\{j\}-\{k\}}$. One of the two sets being a singleton set contradicts non-degeneracy of the shadow, because otherwise one of the shadow vertices $x_B, x_{B\cup\{j\}-\{k\}}$ would not have an objective in W that it uniquely optimizes, contradicting $\pi_W(x_B)$ and $\pi_W(x_{B\cup\{j\}-\{k\}})$ being vertices of $\pi_W(P(A))$. If linearly independent objectives $\theta, \theta' \in W - \{\vec{0}\}$ are both optimized by distinct vertices x_B and $x_{B\cup\{j\}-\{k\}}$, the two vertices must project to the

same point in W, contradicting non-degeneracy of the shadow. Hence $\lambda_{i+1} \neq \lambda_i$ and $\lambda_{i+1} > \lambda_i$.

Since for any $\lambda > \lambda_{i+1}$, $c_\lambda^\intercal A_B^{-1} \not\geq \vec{0}$, no shadow vertex can be visited twice. This implies no basis is visited twice, and hence the algorithm terminates.

Every pivot step taken by the algorithm starts at a shadow vertex, and no two pivot steps start at the same vertex. The number of pivot steps is hence bounded by the number of shadow vertices. Using non-degeneracy of the shadow, this is bounded by $|\text{vertices}(\pi_W(P(A)))|$. By Lemma 2.2.10, if $b = \vec{1}$ then the number of pivot steps is bounded by $|\text{edges}(Q(A) \cap W)|$.

Now we show that the output is correct. Suppose that the algorithm returns a basis B. From the above we know that B is feasible. The choice of λ_i is such that $c_{\lambda_i}^\intercal A_B^{-1} \geq \vec{0}$, so when $\lambda_i = 1$ the basis B is indeed optimal for $c_1 = c$.

Now suppose the algorithm returns *unbounded*, so the ray $x_B + (-A_B^{-1}e_k)\mathbb{R}_+$ is feasible. We want to show that $c(-A_B^{-1}e_k) > 0$. Since $c_\lambda = d + \lambda(c - d)$ and λ_i is $\max\{\lambda : c_\lambda A_B^{-1} \geq 0\}$, we have $(c - d)^\intercal A_B^{-1} e_k < 0$ from uniqueness of k, for otherwise λ_i would not be maximal. Since $\lambda_i < 1$ we get

$$c(-A_B^{-1}e_k) = (c_{\lambda_i} + (1 - \lambda_i)(c - d))(-A_B^{-1}e_k) = (1 - \lambda_i)(c - d)(-A_B^{-1}e_k) > 0,$$

so the algorithm correctly returns *unbounded*. □

2.3 Smoothed Complexity of the Convex Hull

The convex hull of m points in $\mathbb{R}^2$ is a polygon and can have up to m vertices. If each point is randomly distributed, the expected number of vertices can be smaller. In this chapter we prove that if points $a_1, \ldots, a_m \in \mathbb{R}^2$, each of norm at most 1, are perturbed with independent Gaussian random noise of standard deviation σ, then the convex hull of the perturbed points has $O(\sigma^{-1} + \sqrt{\log m})$ vertices in expectation. The general proof strategy will be used and extended in the next chapter as well.

Theorem 2.3.1. *For independently distributed points $a_1, \ldots a_m \in \mathbb{R}^2$, each with independent Gaussian distributed entries of variance σ^2 and $\|\mathbb{E}[a_i]\| \leq 1$ for all $i \in [m]$, the convex hull $Q := \text{conv}(a_1, \ldots, a_m)$ has $O(\sigma^{-1} + \sqrt{\log m})$ edges in expectation.*

To prove the above theorem, we first need a small lemma.

Lemma 2.3.2. *Let $X \in \mathbb{R}$ be a random variable with $\mathbb{E}[X] = \mu$ and $\text{Var}(X) = \tau^2$. Then X satisfies*

$$\frac{\mathbb{E}\left[X^2\right]}{\mathbb{E}\left[|X|\right]} \geq (|\mu| + \tau)/2.$$

Proof. By definition one has $\mathbb{E}\left[X^2\right] = \mu^2 + \tau^2$. We will show that $\mathbb{E}\left[|X|\right] \leq |\mu| + \tau$ so that we can use the fact that $\mu^2 + \tau^2 \geq 2|\mu|\tau$ to derive that $\mu^2 + \tau^2 \geq (|\mu| + \tau)^2/2$. It then follows that $\mathbb{E}\left[X^2\right]/\mathbb{E}\left[|X|\right] \geq (|\mu| + \tau)/2$.

The expected absolute value $\mathbb{E}[|X|]$ satisfies

$$\mathbb{E}\left[|X|\right] \leq |\mu| + \mathbb{E}\left[|X - \mu|\right] \leq |\mu| + \mathbb{E}\left[(X - \mu)^2\right]^{1/2}$$

by Cauchy-Schwarz, hence $\mathbb{E}\left[|X|\right] \leq |\mu| + \tau$. □

Proof of Theorem 2.3.1. We will prove that, on average, the edges of Q are long and the perimeter of Q is small. This is sufficient to bound the expected number of edges.

For $i, j \in [m], i \neq j$, let $E_{i,j}$ denote the event that a_i and a_j are the end points of an edge of Q. By linearity of expectation we have the following equality:

$$\mathbb{E}[\text{perimeter}(Q)] = \sum_{1 \leq i < j \leq m} \mathbb{E}[\|a_i - a_j\| \mid E_{i,j}] \Pr[E_{i,j}].$$

We lower bound the right-hand side by taking the minimum over all conditional expectations and get

$$\sum_{1 \leq i < j \leq m} \mathbb{E}[\|a_i - a_j\| \mid E_{i,j}] \Pr[E_{i,j}] \geq \min_{k \neq l} \mathbb{E}[\|a_k - a_l\| \mid E_{k,l}] \sum_{1 \leq i < j \leq m} \Pr[E_{i,j}].$$

Dividing on both sides, we can estimate the expected number of edges

$$\mathbb{E}[|\text{edges}(Q)|] = \sum_{1 \leq i < j \leq m} \Pr[E_{i,j}] \leq \frac{\mathbb{E}[\text{perimeter}(Q)]}{\min_{k \neq l} \mathbb{E}[\|a_k - a_l\| \mid E_{k,l}]}. \tag{2.12}$$

We are left to bound the numerator and denominator on the right-hand side. For the first, we observe that Q is convex and thus has perimeter at most that of any containing disc. This yields the bound

$$\mathbb{E}[\text{perimeter}(Q)] \leq \mathbb{E}[2\pi \max_i \|a_i\|] \leq 2\pi(1 + 6\sigma\sqrt{\log m}), \tag{2.13}$$

using the traingle inequality and standard Gaussian tail bounds.

We are left to lower bound the denominator. Fix $k = 1, l = 2$ without loss of generality and write $E = E_{1,2}$. The quantity of interest is

$$\mathbb{E}[\|a_1 - a_2\| \mid E] = \frac{\int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \|a_1 - a_2\| \Pr[E] \mu_1(a_1) \mu_2(a_2) \, \mathrm{d}a_1 \, \mathrm{d}a_2}{\int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \Pr[E] \mu_1(a_1) \mu_2(a_2) \, \mathrm{d}a_1 \, \mathrm{d}a_2} \tag{2.14}$$

where μ_i is the probability density of a_i and the probability of $E = E_{1,2}(a_1, \ldots, a_n)$ is taken over the randomness in $a_3, a_4, \ldots, a_m$. To get control on the event E, we

perform a change of coordinates from $a_1, a_2 \in \mathbb{R}^2$ to $t \in [0, \infty], \theta \in \mathbb{S}^1, h_1, h_2 \in \mathbb{R}$ satisfying

$$a_1 = t\theta + R_\theta(h_1)$$
$$a_2 = t\theta + R_\theta(h_2)$$

where $R_\theta : \mathbb{R} \to \theta^\perp$ is the isometric linear embedding of $\mathbb{R}$ into the linear subspace orthogonal to θ with $R_\theta(1)$ having positive first coordinate. This transformation is uniquely defined and continuous whenever a_1 and a_2 are linearly independent and θ has non-zero first coordinate, which happens with probability 1. The Jacobian of this transformation is $|h_1 - h_2|$ and we can rewrite the above fraction (2.14) as

$$\frac{\int_0^\infty \int_{\mathbb{S}^1} \int_{-\infty}^\infty \int_{-\infty}^\infty |h_1 - h_2|^2 \Pr[E] \mu_1(t\theta + R_\theta(h_1)) \mu_2(t\theta + R_\theta(h_2)) \, \mathrm{d}h_1 \, \mathrm{d}h_2 \, \mathrm{d}\theta \, \mathrm{d}t}{\int_0^\infty \int_{\mathbb{S}^1} \int_{-\infty}^\infty \int_{-\infty}^\infty |h_1 - h_2| \Pr[E] \mu_1(t\theta + R_\theta(h_1)) \mu_2(t\theta + R_\theta(h_2)) \, \mathrm{d}h_1 \, \mathrm{d}h_2 \, \mathrm{d}\theta \, \mathrm{d}t}.$$

The event E is equivalent to asking that either $\theta^\mathsf{T} a_i \le t$ for all $i = 3, 4, \ldots, m$ or $\theta^\mathsf{T} a_i \ge t$ for all $i = 3, 4, \ldots, m$. This makes E a function of only $a_3, \ldots, a_n$ and θ and t, i.e. its value does not depend on h_1, h_2.

Now, we use that $\frac{\int g(p)h(p) \, \mathrm{d}p}{\int g(p) \, \mathrm{d}p} \ge \inf_p h(p)$ for any positive integrable g, h and find

$$\begin{aligned}
\mathbb{E}[\|a_1 - a_2\| \mid E] &\ge \inf_{t,\theta} \frac{\int_{-\infty}^\infty \int_{-\infty}^\infty |h_1 - h_2|^2 \mu_1(t\theta + R_\theta(h_1)) \mu_2(t\theta + R_\theta(h_2)) \, \mathrm{d}h_1 \, \mathrm{d}h_2}{\int_{-\infty}^\infty \int_{-\infty}^\infty |h_1 - h_2| \mu_1(t\theta + R_\theta(h_1)) \mu_2(t\theta + R_\theta(h_2)) \, \mathrm{d}h_1 \, \mathrm{d}h_2} \\
&= \inf_{t,\theta} \frac{\int_{-\infty}^\infty z^2 \left(\int_{-\infty}^\infty \mu_1(R_\theta(h_1)) \mu_2(R_\theta(h_1 - z)) \, \mathrm{d}h_1 \right) \mathrm{d}z}{\int_{-\infty}^\infty |z| \left(\int_{-\infty}^\infty \mu_1(R_\theta(h_1)) \mu_2(R_\theta(h_1 - z)) \, \mathrm{d}h_1 \right) \mathrm{d}z},
\end{aligned}$$

substituting $z = h_1 - h_2$ and simplifying. For fixed t, θ, we can reinterpret the last fraction as $\mathbb{E}[Z^2]/\mathbb{E}[|Z|]$ for Z a random variable with probability density proportional to

$$\int_{-\infty}^\infty \mu_1(R_\theta(h_1)) \mu_2(R_\theta(h_1 - z)) \, \mathrm{d}h_1.$$

This is the same probability density as that of the difference of two independent Gaussian random variables each of variance σ^2, which means that Z has variance $2\sigma^2$. If we apply Lemma 2.3.2 to Z, we deduce $\mathbb{E}[\|a_1 - a_2\| \mid E] \ge \sigma/\sqrt{2}$. We conclude that the expected total number of edges is bounded from above by

$$\mathbb{E}[\text{edges}(Q)] \le 2\pi \frac{1 + 6\sigma\sqrt{\log m}}{\sigma/\sqrt{2}} \le 9\sigma^{-1} + 54\sqrt{\log m}. \qquad \square$$

2.4 Shadow Bounds

In this section, we derive our new and improved shadow bounds for Laplace and Gaussian distributed perturbations. We achieve these results by first proving a shadow bound for parametrized distributions as described in the next section, and then specializing to the case of Laplace and Gaussian perturbations. The bounds we obtain are described below.

Theorem 2.4.1. *Let $W \subseteq \mathbb{R}^n$ be a fixed two-dimensional subspace, $m \geq n \geq 3$ and let $A \in \mathbb{R}^{m \times n}$ be a matrix with rows $a_1, \ldots, a_m \in \mathbb{R}^n$, such that the entries of A are independent Gaussian random variables with variance σ^2 and such that $\|\mathbb{E}[a_i]\| \leq 1$ for every $i = 1, \ldots, m$. Writing $Q(A) := \operatorname{conv}(a_1, \ldots, a_m)$ for the convex hull of the row vectors, we find that the expected number of edges is bounded by*

$$\mathbb{E}[|\text{edges}(Q(A) \cap W)|] \leq \mathcal{D}_g(n, m, \sigma)$$

where $\mathcal{D}_g(n, m, \sigma)$ is defined as

$$\mathcal{D}_g(n, m, \sigma) = O(n^2\sqrt{\log m}\ \sigma^{-2} + n^{2.5} \log m\ \sigma^{-1} + n^{2.5} \log(m)^{1.5}).$$

Our bound applies more generally for distributions satisfying certain parameters. We illustrate this with a shadow bound for perturbations distributed according to the Laplace distribution. This will serve as a good warm-up exercise for the slightly more involved analysis of the Gaussian distribution.

Theorem 2.4.2. *Let $W \subseteq \mathbb{R}^n$ be a fixed two-dimensional subspace, $m \geq n \geq 3$ and let $A \in \mathbb{R}^{m \times n}$ be a matrix with rows $a_1, \ldots, a_m \in \mathbb{R}^n$, such that the entries of A are independent Laplace random variables with parameter σ and such that $\|\mathbb{E}[a_i]\| \leq 1$ for every $i = [m]$. Writing $Q(A) := \operatorname{conv}(a_1, \ldots, a_m)$ for the convex hull of the row vectors, we find that the expected number of edges is bounded by*

$$\mathbb{E}[|\text{edges}(Q(A) \cap W)|] \leq O(n^{2.5}\sigma^{-2} + n^3 \log m\ \sigma^{-1} + n^3 \log(m)^2).$$

The proofs of Theorems 2.4.1 and 2.4.2 are given in respectively subsections 2.4.3 and 2.4.2.

2.4.1 Parametrized Shadow Bound

In this section, we prove a shadow bound theorem for any noise distribution that has non-trivial bounds on certain parameters. The parameters we will use are defined below.

Distribution parameters

Definition 2.4.3. A distribution with density μ on $\mathbb{R}^n$ is *L-log-Lipschitz* if for all $x, y \in \mathbb{R}^n$ we have $|\log(\mu(x)) - \log(\mu(y))| \leq L\|x - y\|$. Equivalently, μ is L-log-Lipschitz if $\mu(x)/\mu(y) \leq \exp(L\|x - y\|)$ for all $x, y \in \mathbb{R}^n$.

Definition 2.4.4. Given a probability distribution with density μ on $\mathbb{R}^n$, we define the *line variance* τ^2 as the infimum of the variances when restricted to any fixed line $l \subseteq \mathbb{R}^n$:

$$\tau^2 = \inf_{\text{line } l \subseteq \mathbb{R}^n} \operatorname{Var}(X \sim \mu \mid X \in l).$$

Both the log-Lipschitz constant and the minimal line variance relate to how "spread out" the probability mass is. The log-Lipschitzness of a random variable gives a lower bound on the line variance, which we prove in Lemma 2.4.7.

Definition 2.4.5. Given a distribution with probability density μ on $\mathbb{R}^n$ with expectation $\mathbb{E}_{X\sim\mu}[X] = y$ we define the m^{th}*-th deviation* r_m to be the smallest number such that for any unit vector $\theta \in \mathbb{R}^n$,

$$\int_{r_m}^{\infty} \Pr_{X\sim\mu}[|(X - y)^\mathsf{T}\theta| \geq t] \, \mathrm{d}t \leq r_m/m.$$

Note that as r_m increases to ∞, the left-hand side goes to 0 and the right-hand side goes to ∞. We see that there must exist a number satisfying this inequality, so r_m is well-defined.

The n-th deviation will allow us to give bounds on the expected maximum size $\mathbb{E}[\max_{i\leq m}|x_i{}^\mathsf{T}\theta|]$ of m separate perturbations in a given direction θ. We formalize this in Lemma 2.4.8.

Definition 2.4.6. Given a distribution with probability density μ on $\mathbb{R}^n$ with expectation $\mathbb{E}_{x\sim\mu}[x] = y$, we define, for all $1 > p > 0$, the *cutoff radius* $R(p)$ as the smallest number satisfying

$$\Pr_{x\sim\mu}[\|x - y\| \geq R(p)] \leq p.$$

The cutoff radius of interest is $R_{m,n} := R(\frac{1}{n\binom{m}{n}})$. The cutoff radius tells us how concentrated the probability mass of the random variable is, while the log-Lipschitzness tells us how spread out the probability mass is. These quantities cannot both be arbitrarily good (small) at the same time. We formalize this notion in Lemma 2.4.9.

Lemma 2.4.7. *If a distribution with probability density μ is L-log-Lipschitz, then its line variance satisfies $\tau \geq 1/(\sqrt{e}L)$.*

Proof. Let $v + w\mathbb{R}$ be a line and assume that $\mathbb{E}[x \mid x \in v + w\mathbb{R}] = v$ and $\|w\| = 1$. We show that with probability at least $1/e$, x has distance at least $1/L$ from v. Conditioning on $x \in v + w\mathbb{R}$, the induced probability mass is proportional to $\mu(x)$. We can bound the fraction of the induced probability mass that is far away from the expectation by the following calculation:

$$\begin{aligned}\int_{-\infty}^{\infty} \mu(v+\gamma w)\,\mathrm{d}\gamma &= \int_{-\infty}^{0} \mu(v+\gamma w)\,\mathrm{d}\gamma + \int_{0}^{\infty} \mu(v+\gamma w)\,\mathrm{d}\gamma \\ &= \int_{-\infty}^{-1/L} \mu(v+(\gamma+1/L)w)\,\mathrm{d}\gamma + \int_{1/L}^{\infty} \mu(v+(\gamma-1/L)w)\,\mathrm{d}\gamma \\ &\leq e\int_{-\infty}^{-1/L} \mu(v+\gamma w)\,\mathrm{d}\gamma + e\int_{1/L}^{\infty} \mu(v+\gamma w)\,\mathrm{d}\gamma.\end{aligned}$$

The integral on the first line exists because it is the integral of a continuous non-negative function, and, if the integral were infinite, then the integral along every parallel line would be infinite by log-Lipschitzness, contradicting the fact that μ has integral 1 over $\mathbb{R}^n$.

Hence,

$$\Pr[\|x-v\| \geq 1/L \mid x \in v+w\mathbb{R}] = \frac{\int_{-\infty}^{-1/L} \mu(v+\gamma w)\,\mathrm{d}\gamma + \int_{1/L}^{\infty} \mu(v+\gamma w)\,\mathrm{d}\gamma}{\int_{-\infty}^{\infty} \mu(v+\gamma w)\,\mathrm{d}\gamma} \geq 1/e,$$

and we can lower bound the variance

$$\mathrm{Var}(x \mid x \in v + w\mathbb{R}) \geq \frac{1}{e}(1/L)^2.$$

Since the line $v + w\mathbb{R}$ was arbitrary, it follows that $\tau \geq 1/(\sqrt{e}L)$. □

Lemma 2.4.8. *If $x_1, \ldots, x_m$ are each distributed with mean $\vec{0}$ and m-th deviation at most r_m, then for any $\theta \in \mathbb{S}^{n-1}$,*

$$\mathbb{E}[\max_{i\in[m]} |\theta^{\mathsf{T}} x_i|] \leq 2r_m.$$

Proof. We rewrite the expectation as

$$\mathbb{E}[\max_{i\in[m]} |\theta^{\mathsf{T}} x_i|] = \int_0^{\infty} \Pr[\max_{i\in[m]} |\theta^{\mathsf{T}} x_i| \geq t]\,\mathrm{d}t.$$

We separately bound the integral up to r_m and from r_m to ∞. Since a probability is at most 1 we have

$$\int_0^{r_m} \Pr[\max_{i\in[m]} |\theta^{\mathsf{T}} x_i| \geq t]\,\mathrm{d}t \leq r_m,$$

and by definition of the n-th deviation and the union bound:

$$\int_{r_m}^{\infty} \Pr[\max_{i\in[m]} |\theta^{\mathsf{T}} x_i| \geq t] \, \mathrm{d}t \leq \sum_{i\in[m]} \int_{r_m}^{\infty} \Pr[|\theta^{\mathsf{T}} x_i| \geq t] \leq r_m.$$

Together these estimates yield the desired inequality,

$$\mathbb{E}[\max_{i\leq m} |\theta^{\mathsf{T}} x_i|] \leq 2r_m. \qquad \square$$

Lemma 2.4.9. *For a n-dimensional distribution with probability density μ, where $n \geq 3$, with parameters L, R as described above, we have $LR(1/2) \geq n/3$.*

Proof. Let $\bar{R} := R(1/2)$. If $L\bar{R} \geq n$, we are already done, so we may assume that $L\bar{R} < n$. Also, without loss of generality, we may assume that μ has mean $\vec{0}$. For $\alpha > 1$ to be chosen later we know

$$\begin{aligned} 1 &\geq \int_{\alpha\bar{R}\mathbb{B}_2^n} \mu(x) \, \mathrm{d}x \\ &= \alpha^n \int_{\bar{R}\mathbb{B}_2^n} \mu(\alpha x) \, \mathrm{d}x \\ &\geq \alpha^n e^{-(\alpha-1)L\bar{R}} \int_{\bar{R}\mathbb{B}_2^n} \mu(x) \, \mathrm{d}x \\ &= \frac{\alpha^n}{2} e^{-(\alpha-1)L\bar{R}}. \end{aligned}$$

Taking logarithms, we find

$$0 \geq n\log(\alpha) - (\alpha-1)L\bar{R} - \log(2).$$

We choose $\alpha = \frac{n}{L\bar{R}} > 1$ and look at the resulting inequality:

$$0 \geq n\log(\frac{n}{L\bar{R}}) - n + L\bar{R} - \log(2).$$

For $n \geq 3$, this inequality can only hold if $L\bar{R} \geq n/3$, as needed. $\square$

Proving a shadow bound for parametrized distributions

The main result of this subsection is the following parametrized shadow bound.

Theorem 2.4.10 (Parametrized Shadow Bound)**.** *Let $a_1, \ldots, a_m \in \mathbb{R}^n$, where $m \geq n \geq 3$, be independently distributed according to L-log-Lipschitz distributions with centers of norm at most 1, line variances at least τ^2, cutoff radii at most $R_{m,n}$ and m-th deviations at most r_m. For any fixed two-dimensional linear subspace $W \subseteq \mathbb{R}^n$, the expected number of edges satisfies*

$$\mathbb{E}[|\text{edges}(Q(A) \cap W)|] \leq O(\frac{n^{1.5}L}{\tau}(1 + R_{m,n})(1 + r_m)).$$

The proof is given at the end of the subsection. It will be derived from the sequence of lemmas given below. We refer the reader to section 2.1.3 for a high-level overview of the proof.

In the rest of the subsection, $a_1, \ldots, a_m \in \mathbb{R}^n$, where $m \geq n \geq 3$, will be as in Theorem 2.4.10. We use $Q(A) := \text{conv}(a_1, \ldots, a_m)$ to denote the convex hull of the rows $a_1, \ldots, a_m$ of the constraint matrix A and W to denote the two-dimensional shadow plane.

The following non-degeneracy conditions on $a_1, \ldots, a_m$ will hold with probability 1, because $a_1, \ldots, a_m$ are independently distributed with continuous distributions.

1. Every $n+1$ vectors from $a_1, \ldots, a_m$ are affinely independent. Thus, every facet of $Q(A)$ is the convex hull of exactly n vectors from $a_1, \ldots, a_m$.

2. Any n distinct vectors $a_{i_1}, \ldots, a_{i_n}$, $i_1, \ldots, i_n \in [m]$, have a unique hyperplane through them. This hyperplane intersects W in a one-dimensional line, does not contain the origin $\vec{0}$, and its unit normal vector pointing away from the origin is not $-e_1$. Note that the last two conditions imply that the coordinate transformation on $a_{i_1}, \ldots, a_{i_n}$ is uniquely defined with e_1 as reference vector.

3. For every edge $e \subseteq Q(A) \cap W$ there is a unique facet F of $Q(A)$ such that $e = F \cap W$.

In what follows we will always assume the above conditions hold.

For our first lemma, in which we bound the number of edges in terms of two different expected lengths, we make a distinction between possible edges with high probability of appearing versus edges with low probability of appearing. The sets with probability at most $2\binom{m}{n}^{-1}$ to form an edge, together contribute at most 2 to the expected number of edges, as there are only $\binom{m}{n}$ bases.

For a basis with probability at least $2\binom{m}{n}^{-1}$ of forming an edge, we can safely condition on it forming an edge without forcing very unlikely events to happen. Because of this, we will later be able to condition on the vertices not being too far apart.

Definition 2.4.11. For $I \in \binom{[m]}{n}$, let E_I denote the event that $\mathrm{conv}(a_i : i \in I) \cap W$ forms an edge of $Q(A) \cap W$.

Definition 2.4.12. We define the set $B \subseteq \binom{[m]}{n}$ to be the set of those $I \subseteq [m]$ satisfying $|I| = n$ and $\Pr[E_I] \geq 2\binom{m}{n}^{-1}$.

The next lemma is inspired by Theorem 3.2 of [121].

Lemma 2.4.13. *The expected number of edges in $Q(A) \cap W$ satisfies*

$$\mathbb{E}[|\mathrm{edges}(Q(A) \cap W)|] \leq 2 + \frac{\mathbb{E}[\mathrm{perimeter}(Q(A) \cap W)]}{\min_{I \in B} \mathbb{E}[\mathrm{length}(\mathrm{conv}(a_i : i \in I) \cap W) \mid E_I]}.$$

Proof. We give a lower bound on the perimeter of the intersection $Q(A) \cap W$ in terms of the number of edges. By our non-degeneracy assumption, every edge can be uniquely represented as $\mathrm{conv}(a_i : i \in I) \cap W$, for $I \in \binom{[m]}{n}$. From this we derive the first equality, and we continue from that:

$$\begin{aligned}
\mathbb{E}[\mathrm{perimeter}(Q(A) \cap W)] &= \sum_{I \in \binom{[m]}{n}} \mathbb{E}[\mathrm{length}(\mathrm{conv}(a_i : i \in I) \cap W) \mid E_I] \Pr[E_I] \\
&\geq \sum_{I \in B} \mathbb{E}[\mathrm{length}(\mathrm{conv}(a_i : i \in I) \cap W) \mid E_I] \Pr[E_I] \\
&\geq \min_{I \in B} \mathbb{E}[\mathrm{length}(\mathrm{conv}(a_i : i \in I) \cap W) \mid E_I] \sum_{J \in B} \Pr[E_J].
\end{aligned}$$

The first line holds because whenever E_I holds, $\mathrm{conv}(a_i : i \in I) \cap W$ is an edge of $Q(A) \cap W$, and every edge of $Q(A) \cap W$ is formed by exactly one face F_J, by the non-degeneracy conditions we have assumed. By construction of B and linearity of expectation, $\sum_{J \in B} \Pr[E_J] \geq \sum_{J \in \binom{[m]}{n}} \Pr[E_J] - 2 = \mathbb{E}[|\mathrm{edges}(Q(A) \cap W)|] - 2$. By dividing on both sides of the inequality, we can now conclude

$$\mathbb{E}[|\mathrm{edges}(Q(A) \cap W)|] \leq 2 + \frac{\mathbb{E}[\mathrm{perimeter}(Q(A) \cap W)]}{\min_{I \in B} \mathbb{E}[\mathrm{length}(\mathrm{conv}(a_i : i \in I) \cap W) \mid E_I]}. \quad \square$$

Given the above, we may now restrict our task to proving an upper bound on the expected perimeter and a lower bound on the minimum expected edge length, which will be the focus on the remainder of the subsection.

The perimeter is bounded using a standard convexity argument. A convex shape has perimeter no more than that of any circle containing it. We exploit the fact that all centers have norm at most 1 and the expected perturbation sizes are not too big along any fixed axis.

Lemma 2.4.14. *The expected perimeter of $Q(A) \cap W$ is bounded by*

$$\mathbb{E}[\mathrm{perimeter}(Q(A) \cap W)] \leq 2\pi(1 + 4r_m),$$

where r_m is the m-deviation bound for $a_1, \ldots, a_m$.

Proof. By convexity, the perimeter is bounded from above by 2π times the norm of the maximum norm point. Let $\hat{a}_i := a_i - \mathbb{E}[a_i]$ denote the perturbation of a_i from the center of its distribution, recalling that $\|\mathbb{E}[a_i]\| \leq 1$ by assumption. We can now derive the bound

$$\begin{aligned}
\mathbb{E}[\text{perimeter}(Q(A) \cap W)] &\leq 2\pi\mathbb{E}[\max_{x \in Q(A) \cap W} \|x\|] \\
&= 2\pi\mathbb{E}[\max_{x \in Q(A) \cap W} \|\pi_W(x)\|] \\
&\leq 2\pi\mathbb{E}[\max_{x \in Q(A)} \|\pi_W(x)\|] \\
&= 2\pi\mathbb{E}[\max_{i \in [m]} \|\pi_W(a_i)\|] \\
&\leq 2\pi \left(1 + \mathbb{E}[\max_{i \leq m} \|\pi_W(\hat{a}_i)\|]\right),
\end{aligned}$$

where the last inequality follows since $a_1, \ldots, a_m$ have centers of norm at most 1. Pick an orthogonal basis v_1, v_2 of W. By the triangle inequality the expected perturbation size satisfies

$$\mathbb{E}[\max_{i \leq m} \|\pi_W(\hat{a}_i)\|] \leq \sum_{j \in \{1,2\}} \mathbb{E}[\max_{i \leq n} |{v_j}^{\mathsf{T}} \hat{a}_i|].$$

Each of the two expectations satisfies, by Lemma 2.4.8, $\mathbb{E}[\max_{i \leq m} |{v_j}^{\mathsf{T}} \hat{a}_i|] \leq 2r_m$, thereby concluding the proof. □

The rest of this subsection will be devoted to finding a suitable lower bound on the denominator $\mathbb{E}[\text{length}(\text{conv}(a_i : i \in I) \cap W) \mid E_I]$ uniformly over all choices of $I \in B$. Without loss of generality we assume that $I = [n]$ and write $E := E_{[n]}$.

Definition 2.4.15 (Containing hyperplane)**.** Define $H = \text{aff}(a_1, \ldots, a_n) = t\theta + \theta^{\perp}$, where $\theta \in \mathbb{S}^{n-1}$, $t > 0$ to be the hyperplane containing $a_1, \ldots, a_n$. Define $l = H \cap W$. From our non-degeneracy conditions we know that l is a line. Express $l = p + \omega \cdot \mathbb{R}$, where $\omega \in \mathbb{S}^{n-1}$ and $p \in \omega^{\perp}$.

To lower bound the length $\mathbb{E}[\text{length}(\text{conv}(a_1, \ldots, a_n) \cap W) \mid E]$ we will need the pairwise distances between the different a_i's for $i \in [n]$ to be small along $\omega^{\perp}$. This will allow us to get "wiggle room" around each vertex of $\text{conv}(a_1, \ldots, a_n)$ that is proportional to the size of the facet.

Definition 2.4.16 (Bounded diameter event)**.** We define the event D to hold exactly when $\|\pi_{\omega^{\perp}}(a_i) - \pi_{\omega^{\perp}}(a_j)\| \leq 2 + 2R_{m,n}$ for all $i, j \in [n]$.

We will condition on the event D. This will not change the expected length by much, because the probability that D does not occur is small compared to the probability of E by our assumption that $\Pr[E] \geq \frac{2}{n\binom{m}{n}}$.

Lemma 2.4.17. *The expected edge length satisfies*

$$\mathbb{E}[\text{length}(\text{conv}(a_1, \ldots, a_n) \cap W) \mid E] \geq \mathbb{E}[\text{length}(\text{conv}(a_1, \ldots, a_n) \cap W) \mid D, E]/2.$$

Proof. Let the vector $\hat{a}_i$ denote the perturbation $a_i - \mathbb{E}[a_i]$. Since distances can only decrease when projecting, the complementary event D^c satisfies

$$\begin{aligned}\Pr[D^c] &= \Pr[\max_{i,j \leq n} \|\pi_{\omega^\perp}(a_i - a_j)\| \geq 2 + 2R_{m,n}] \\ &\leq \Pr[\max_{i,j \leq n} \|a_i - a_j\| \geq 2 + 2R_{m,n}],\end{aligned}$$

by the triangle inequality and the bound of 1 on the norms of the centers, the line above is at most

$$\begin{aligned}&\leq \Pr[\max_{i \leq n} \|a_i\| \geq 1 + R_{m,n}] \\ &\leq \Pr[\max_{i \leq n} \|\hat{a}_i\| \geq R_{m,n}] \\ &\leq \binom{m}{n}^{-1}.\end{aligned}$$

By our assumption that $[n] \in B$, we know that $\Pr[E] \geq 2\binom{m}{n}^{-1}$. In particular, it follows that $\Pr[E \cap D] \geq \Pr[E] - \Pr[D^c] \geq \Pr[E]/2$. Thus, we may conclude that

$$\mathbb{E}[\text{length}(\text{conv}(a_1, \ldots, a_n) \cap W) \mid E] \geq \mathbb{E}[\text{length}(\text{conv}(a_1, \ldots, a_n) \cap W) \mid D, E]/2.$$

□

For the rest of this section, we use a change of variables on $a_1, \ldots, a_n$. The non-degeneracy conditions we have assumed at the start of this section make the change of variables well-defined.

Recall the change of variables mapping $(a_1, \ldots, a_n) \mapsto (\theta, t, b_1, \ldots, b_n)$ for $\theta \in \mathbb{S}^{n-1}, t > 0, b_1, \ldots, b_n \in \mathbb{R}^{n-1}$ from Theorem 2.2.6. We abbreviate $\bar{\mu}_i(\theta, t, b_i) = \mu_i(R_\theta(b_i) + t\theta)$ and we write $\bar{\mu}_i(b_i)$ when the values of θ, t are clear.

By Theorem 2.2.6 we know that for any fixed values of θ, t the vectors $b_1, \ldots, b_n$ have joint probability density proportional to

$$\text{vol}_{n-1}(\text{conv}(b_1, \ldots, b_n)) \prod_{i=1}^{n} \bar{\mu}_i(b_i) \,. \tag{2.15}$$

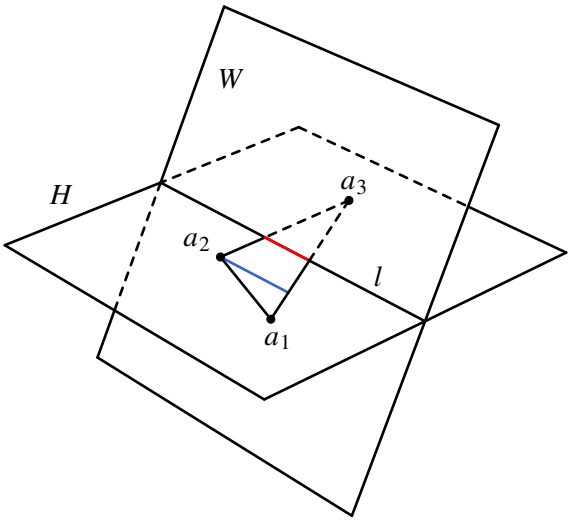


Figure 2.3: $a_1, \ldots, a_n$ are conditioned for $\mathrm{conv}(a_1, \ldots, a_n)$ to intersect W and lie in H. The red line corresponds to induced edge. The blue line represents the longest chord parallel to ℓ.

We assumed that $a_1, \ldots, a_n$ are affinely independent, so $b_1, \ldots, b_n$ are affinely independent as well.

In the next lemma, we condition on the hyperplane $H = t\theta + \theta^\perp$ and from then on we restrict our attention to what happens inside H. Conditioned on $a_1, \ldots, a_n$ lying in H, the set $\mathrm{conv}(a_1, \ldots, a_n)$ is a facet of $Q(A)$ if and only if all of $a_{n+1}, \ldots, a_m$ lie on one side of H. This means that the shape of $\mathrm{conv}(a_1, \ldots, a_n)$ in H does not influence the event that it forms a facet, so in studying this convex hull we can then ignore $a_{n+1}, \ldots, a_m$.

We identify the hyperplane H with $\mathbb{R}^{n-1}$ and define $\bar{l} = \bar{p} + \bar{\omega} \cdot \mathbb{R} \subseteq \mathbb{R}^{n-1}$ corresponding to $l = p + \omega \cdot \mathbb{R}$ by $\bar{p} = R_\theta^{-1}(p - t\theta)$, $\bar{\omega} = R_\theta^{-1}(\omega)$. We define $\bar{E}$ as the event that $\mathrm{conv}(b_1, \ldots, b_n) \cap \bar{l} \neq \emptyset$. Notice that E holds if and only if $\bar{E}$ and $\mathrm{conv}(a_1, \ldots, a_n)$ is a facet of $Q(A)$. See Figure 2.3.

We will condition on the shape of the projected simplex.

Definition 2.4.18 (Projected shape)**.** We define the projected shift variable by $x := x_\omega(b_1) = \pi_{\bar{\omega}^\perp}(b_1)$ and shape variable $S := S_\omega(b_1, \ldots, b_n)$ by

$$S_\omega(b_1, \ldots, b_n) = (\vec{0}, \pi_{\bar{\omega}^\perp}(b_2) - x, \ldots, \pi_{\bar{\omega}^\perp}(b_n) - x) \,.$$

We index $S = (s_1, \ldots, s_n)$, so $s_i \in \bar{\omega}^\perp$ is the i-th vector in S, and furthermore define the diameter function $\mathrm{diam}(S) = \max_{i,j \in [n]} \|s_i - s_j\|$. We will condition on the shape

being in the set of allowed shapes

$$\mathcal{S} := \{(s_1, \ldots, s_n) \in (\bar{\omega}^{\perp})^n : s_1 = \vec{0}, \operatorname{diam}(S) \leq 2 + 2R_{m,n}, \operatorname{rank}(s_2, \ldots, s_n) = n - 2\}.$$

Observe that $S \in \mathcal{S}$ if and only if the event D holds. To justify the rank condition on $s_2, \ldots, s_n$, note that by our non-degeneracy conditions, we have that $b_1, \ldots, b_n$ are affinely independent. In particular, they do not all lie in an $(n-2)$-dimensional affine subspace. This means that $s_1, \ldots, s_n$ do not all lie in a $(n-3)$-dimensional affine subspace, from which it follows that $\operatorname{rank}(s_2, \ldots, s_n) = n - 2$ (recalling that $s_1 = \vec{0}$).

Lemma 2.4.19. *Let $\theta \in \mathbb{S}^{n-1}, t > 0, b_1, \ldots, b_n \in \mathbb{R}^{n-1}$ denote the change of variables of $a_1, \ldots, a_n \in \mathbb{R}^n$ as discussed above. Then, the expected length satisfies*

$$\mathbb{E}[\operatorname{length}(\operatorname{conv}(a_1, \ldots, a_n) \cap W) \mid D, E] \geq \inf_{\theta, t, S \in \mathcal{S}} \mathbb{E}[\operatorname{length}(\operatorname{conv}(b_1, \ldots, b_n) \cap \bar{l}) \mid \theta, t, S, \bar{E}].$$

Proof. To derive the desired inequality, we first understand the effect of conditioning on E. Let E_0 denote the event that $F := \operatorname{conv}(a_1, \ldots, a_n)$ induces a facet of $Q(A)$. Note that E is equivalent to $E_0 \cap \bar{E}$, where $\bar{E}$ is as above. We now perform the change of variables from $a_1, \ldots, a_n \in \mathbb{R}^n$ to $\theta \in \mathbb{S}^{n-1}, t \in \mathbb{R}_+, b_1, \ldots, b_n \in \mathbb{R}^{n-1}$. The set F is a facet of $Q(A)$ if and only if $\theta^{\mathsf{T}} a_{n+i} \leq t$ for all $i \in [m-n]$ or $\theta^{\mathsf{T}} a_{n+i} \geq t$ for all $i \in [m-n]$. Given this, we see that

$$\begin{aligned}
&\mathbb{E}[\operatorname{length}(\operatorname{conv}(a_1, \ldots, a_n) \cap W) \mid D, E] \\
&= \mathbb{E}[\operatorname{length}(\operatorname{conv}(b_1, \ldots, b_n) \cap \bar{l}) \mid D, E_0, \bar{E}] \\
&= \frac{\mathbb{E}[\ \mathbb{1}[E_0] \cdot \operatorname{length}(\operatorname{conv}(b_1, \ldots, b_n) \cap \bar{l}) \mid D, \bar{E}]}{\Pr[E_0 \mid D, \bar{E}]} \\
&= \frac{\mathbb{E}_{\theta,t}[\ \mathbb{E}[\mathbb{1}[E_0] \cdot \operatorname{length}(\operatorname{conv}(b_1, \ldots, b_n) \cap \bar{l}) \mid \theta, t, D, \bar{E}]\ ]}{\mathbb{E}_{\theta,t}[\ \Pr[E_0 \mid \theta, t, D, \bar{E}]\ ]}
\end{aligned} \tag{2.16}$$

Since $a_1, \ldots, a_m$ are independent, conditioned on θ, t, the random vectors $b_1, \ldots, b_n$ are independent of $\theta^{\mathsf{T}} a_{n+1}, \ldots, \theta^{\mathsf{T}} a_m$. Since the events D and $\bar{E}$ only depend on $b_1, \ldots, b_n$, continuing from (2.16), we get that

$$\begin{aligned}
&\frac{\mathbb{E}_{\theta,t}[\ \mathbb{E}[\mathbb{1}[E_0] \cdot \operatorname{length}(\operatorname{conv}(b_1, \ldots, b_n) \cap \bar{l}) \mid \theta, t, D, \bar{E}]\ ]}{\mathbb{E}_{\theta,t}[\ \Pr[E_0 \mid \theta, t, D, \bar{E}]\ ]} \\
&= \frac{\mathbb{E}_{\theta,t}[\Pr[E_0 \mid \theta, t] \cdot \mathbb{E}[\operatorname{length}(\operatorname{conv}(b_1, \ldots, b_n) \cap \bar{l}) \mid \theta, t, D, \bar{E}]\ ]}{\mathbb{E}_{\theta,t}[\ \Pr[E_0 \mid \theta, t]]} \\
&\geq \inf_{\theta \in \mathbb{S}^{n-1}, t > 0} \mathbb{E}[\operatorname{length}(\operatorname{conv}(b_1, \ldots, b_n) \cap \bar{l}) \mid \theta, t, D, \bar{E}].
\end{aligned}$$

The last inequality uses the general fact that if f, g are functions then

$$\frac{\int f(x)g(x)\,\mathrm{d}x}{\int f(x)\,\mathrm{d}x} \geq \inf g(x)$$

when f is non-negative and has finite integral.

Lastly, since the event D is equivalent to $S := S_\omega(b_1, \ldots, b_n) \in \mathcal{S}$ as in Definition 2.4.18, we have that

$$\mathbb{E}[\text{length}(\text{conv}(a_1, \ldots, a_n) \cap W) \mid D, E] \geq \inf_{\theta, t, S \in \mathcal{S}} \mathbb{E}[\text{length}(\text{conv}(b_1, \ldots, b_n) \cap \bar{l}) \mid \theta, t, S, \bar{E}]. \qquad \square$$

Definition 2.4.20 (Kernel combination)**.** For $S \in \mathcal{S}$, define the combination $z := z(S)$ to be the unique (up to sign) $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$ satisfying

$$\sum_{i=1}^n z_i s_i = \vec{0},\ \sum_{i=1}^n z_i = 0,\ \|z\|_1 = 1.$$

To justify the above definition, it suffices to show that the system of equations

$$\sum_{i=1}^n z_i s_i = \vec{0}, \sum_{i=1}^n z_i = 0 \tag{2.17}$$

has a one-dimensional solution space. Since $s_1, \ldots, s_n$ live in a $(n-2)$ dimensional space, the solution space has dimension at least 1 by dimension counting. Next, note that z is a solution to (2.17) if and only if $z_1 = -\sum_{i=2}^n z_i$ and

$$\sum_{i=2}^n z_i s_i = \vec{0} \tag{2.18}$$

(since $s_1 = \vec{0}$). Thus, the solution space of (2.17) and (2.18) have the same dimension. Given our assumption that $\text{rank}(s_2, \ldots, s_n) = n-2$, it follows that the solution space of (2.18) is one-dimensional, exactly what is needed for the kernel combination $z(S)$ to be unique up to sign.

Observe that for $S := S_\omega(b_1, \ldots, b_n)$, z satisfies $\pi_{\bar{\omega}^\perp}(\sum_{i=1}^n z_i b_i) = \vec{0}$.

The vector z provides us with a unit to measure lengths in "convex combination space". We make this formal with the next definition:

Definition 2.4.21 (Chord combinations)**.** We define the set of convex combinations of the shape $S = (s_1, \ldots, s_n) \in \mathcal{S}$ that equal $q \in \bar{\omega}^\perp$ by

$$C_S(q) := \{(\lambda_1, \ldots, \lambda_n) \geq \vec{0} \ :\ \sum_{i=1}^n \lambda_i = 1,\ \sum_{i=1}^n \lambda_i s_i = q\} \subseteq \mathbb{R}^n.$$

When S is clear we drop the subscript.

Observe that $C(q)$ is a line segment of the form $C(q) = \lambda_q + z \cdot [0, d_q]$. We write $\|C(q)\|_1$ for the ℓ_1-diameter of $C(q)$. Since $C(q)$ is a line segment, $\|C(q)\|_1 = d_q$. We prove two basic properties of $\|C(q)\|_1$ as a function of q.

Lemma 2.4.22 (Properties of chord combinations)**.** *Let* $y := y(S) = \sum_{i=1}^n |z_i| s_i$*, with* $z := z(S)$ *as in Definition* 2.4.21*. Then the following holds*:

- $\|C(q)\|_1$ *is a concave function for* $q \in \mathrm{conv}(S)$.
- $\max_{q \in \mathrm{conv}(S)} \|C(q)\|_1 = \|C(y)\|_1 = 2$.

Proof. For the first claim, take $x, y \in \mathrm{conv}(S)$. Let $\alpha \in C(x)$ and $\beta \in C(y)$. Then we see that, for all $\gamma \in [0, 1]$,

$$\gamma\alpha + (1-\gamma)\beta \geq 0, \quad \sum_{i=1}^n \gamma\alpha_i + (1-\gamma)\beta_i = 1, \quad \sum_{i=1}^n (\gamma\alpha_i + (1-\gamma)\beta_i) s_i = \gamma x + (1-\gamma) y,$$

from which we derive that

$$\gamma C(x) + (1-\gamma) C(y) \subseteq C(\gamma x + (1-\gamma) y),$$

and hence

$$\|C(\gamma x + (1-\gamma)y)\|_1 \geq \|\gamma C(x) + (1-\gamma) C(y)\|_1 = \gamma \|C(x)\|_1 + (1-\gamma)\|C(y)\|_1.$$

For the second claim, we look at the combination $y := \sum_{i=1}^n |z_i| s_i \in \mathrm{conv}(S)$. For all $\gamma \in [-1, 1]$, we have $\sum_{i=1}^n (|z_i| + \gamma z_i) s_i = y$, $\sum_{i=1}^n |z_i| + \gamma z_i = \|z\|_1 = 1$ and $|z_i| + \gamma z_i \geq 0$, $\forall i \in [n]$. Hence, $\|C(y)\|_1 \geq 2$. Now suppose there is some y' with $\|C(y')\|_1 > 2$. That means there is some convex combination $\lambda = (\lambda_1, \ldots, \lambda_n) \geq \vec{0}$, $\|\lambda\|_1 = 1$, with $\sum_{i=1}^n \lambda_i s_i = y'$ such that coordinatewise $\lambda + z > \vec{0}$ and $\lambda - z > \vec{0}$. Let $I \cup J$ be a partition of $[n]$ such that $z_i \geq 0$ for $i \in I$ and $z_j < 0$ for $j \in J$. We know that $\sum_{i=1}^n z_i = 0$, so $\sum_{i \in I} z_i = -\sum_{j \in J} z_j$. This makes

$$1 = \|z\|_1 = \sum_{i \in I} z_i - \sum_{j \in J} z_i = 2 \sum_{i \in I} z_i,$$

so $\sum_{i \in I} z_i = 1/2$. The combination λ satisfies

$$\sum_{i \in I} \lambda_i > \sum_{i \in I} z_i = 1/2, \quad \sum_{j \in J} \lambda_j > \sum_{j \in J} -z_j = 1/2,$$

from which we conclude $\|\lambda\|_1 > 1$. As this is a contradiction, we must have $\max_{q \in \mathrm{conv}(S)} \|C(q)\|_1 = 2$. □

The ℓ_1-diameter $\|C(q)\|_1$ specified by $q \in \operatorname{conv}(S(b_1, \ldots, b_n))$ directly relates to the length of the chord $(q + x + \bar{\omega} \cdot \mathbb{R}) \cap \operatorname{conv}(b_1, \ldots, b_n)$, which projects to $q + x$ under $\pi_{\bar{\omega}^\perp}$. Specifically, $\|C(q)\|_1$ measures how long the chord is compared to the longest chord through the simplex. The exact relation is given below.

Lemma 2.4.23. *Let* $(h_1, \ldots, h_n) = (\bar{\omega}^\mathsf{T} b_1, \ldots, \bar{\omega}^\mathsf{T} b_n)$, $(s_1, \ldots, s_n) = S(b_1, \ldots, b_n)$, $x = \pi_{\bar{\omega}^\perp}(b_1)$. *For any* $q \in \operatorname{conv}(S)$ *the following equality holds*:

$$\operatorname{length}((x + q + \bar{\omega} \cdot \mathbb{R}) \cap \operatorname{conv}(b_1, \ldots, b_n)) = \|C(q)\|_1 \cdot |\sum_{i=1}^n z_i h_i|.$$

Proof. By construction there is a convex combination $\lambda_1, \ldots, \lambda_n \geq 0$, $\sum_{i=1}^n \lambda_i = 1$ satisfying $\sum_{i=1}^n \lambda_i s_i = q$ such that $C(q) = [\lambda, \lambda + \|C(q)\|_1 z]$ and hence

$$(x + q + \bar{\omega} \cdot \mathbb{R}) \cap \operatorname{conv}(b_1, \ldots, b_n) = [\sum_{i=1}^n \lambda_i b_i, \sum_{i=1}^n (\lambda_i + \|C(q)\|_1 z_i) b_i].$$

From this we deduce

$$\begin{aligned} \operatorname{length}((x + q + \bar{\omega} \cdot \mathbb{R}) \cap \operatorname{conv}(b_1, \ldots, b_n)) &= \left\| \sum_{i=1}^n (\lambda_i + \|C(q)\|_1 z_i) b_i - \sum_{i=1}^n \lambda_i b_i \right\| \\ &= \left\| \sum_{i=1}^n \|C(q)\|_1 z_i b_i \right\| \\ &= \|C(q)\|_1 \cdot |\sum_{i=1}^n z_i h_i|. \end{aligned}$$

The third equality follows from the definition of $z_1, \ldots, z_n$: as $\pi_{\bar{\omega}^\perp}(\sum_{i=1}^n z_i b_i) = \vec{0}$, we must have $\|\sum_{i=1}^n z_i b_i\| = \|\sum_{i=1}^n z_i h_i \bar{\omega}\| = |\sum_{i=1}^n z_i h_i|$. □

We can view the terms in the above product as follows: the length of the longest chord of $\operatorname{conv}(b_1, \ldots, b_n)$ parallel to $\bar{l}$ is $2|\sum_{i=1}^n z_i h_i|$, and the ratio of the length of the chord $\operatorname{conv}(b_1, \ldots, b_n) \cap \bar{l}$ to the length of the longest chord parallel to $\bar{l}$ equals $\|C(q)\|_1/2$. This follows from Lemma 2.4.22 since $\|C(q)\|_1$ achieves a maximum value of 2 at $q = y$. As discussed in the high-level description, we will bound the expected values of these two quantities separately.

The term $|\sum_{i=1}^n z_i h_i|$ can also be used to simplify the volume term in the probability density of $b_1, \ldots, b_n$ after we condition on the shape S. We prove this in the next lemma.

Lemma 2.4.24. *For fixed* $\theta \in \mathbb{S}^{n-1}, t > 0, S \in \mathcal{S}$, *define* $x \in \bar{\omega}^\perp, h_1, \ldots, h_n \in \mathbb{R}$ *conditioned on* θ, t, S *to have joint probability density function proportional to*

$$|\sum_{i=1}^n z_i h_i| \cdot \prod_{i=1}^n \bar{\mu}_i(x + s_i + h_i \bar{\omega}),$$

where $z := z(S)$ is as in Definition 2.4.20. *Then for $b_1, \dots, b_n \in \mathbb{R}^{n-1}$ distributed as in Lemma* 2.4.19, *conditioned on θ, t and the shape $S = (s_1, \dots, s_n)$, where $s_1 = \vec{0}$, we have equivalence of the distributions*

$$(b_1, \dots, b_n) \mid \theta, t, S \equiv (x + s_1 + h_1\bar{\omega}, \dots, x + s_n + h_n\bar{\omega}) \mid \theta, t, S.$$

Proof. The variables $b_1, \dots, b_n$ conditioned on θ, t have density proportional to

$$\mathrm{vol}_{n-1}(\mathrm{conv}(b_1, \dots, b_n)) \prod_{i=1}^{n} \bar{\mu}_i(b_i).$$

We make a change of variables from $b_1, \dots, b_n$ to $x, s_2, \dots, s_n \in \bar{\omega}^{\perp}, h_1, \dots, h_n \in \mathbb{R}$, defined by

$$(b_1, \dots, b_n) = (x + h_1\bar{\omega}, x + s_2 + h_n\bar{\omega}, \dots, x + s_n + h_n\bar{\omega}).$$

Recall that any invertible linear transformation has constant Jacobian. We observe that

$$\mathrm{vol}_{n-1}(\mathrm{conv}(b_1, \dots, b_n)) = \int_{\mathrm{conv}(S)} \mathrm{length}((x + q + \bar{\omega} \cdot \mathbb{R}) \cap \mathrm{conv}(b_1, \dots, b_n))\, \mathrm{d}q.$$

By Lemma 2.4.23 we find

$$\mathrm{vol}_{n-1}(\mathrm{conv}(b_1, \dots, b_n)) = |\sum_{i=1}^{n} z_i h_i| \int_{\mathrm{conv}(S)} \|C(q)\|_1\, \mathrm{d}q.$$

The integral of $\|C(q)\|_1$ over $\mathrm{conv}(S)$ is independent of $x, h_1, \dots, h_n$. Thus, for fixed $\theta \in \mathbb{S}^{n-1}, t > 0, S \in \mathcal{S}$, the random variables $x, h_1, \dots, h_n$ have joint probability density proportional to

$$|\sum_{i=1}^{n} z_i h_i| \cdot \prod_{i=1}^{n} \bar{\mu}_i(x + s_i + h_i\bar{\omega}). \qquad \square$$

Recall that $\bar{l} = \bar{p} + \bar{\omega} \cdot \mathbb{R}$. The event $\bar{E}$ that $\mathrm{conv}(b_1, \dots, b_n) \cap \bar{l} \neq \emptyset$ occurs if and only if $\bar{p} \in x + \mathrm{conv}(S)$, hence if and only if $\bar{p} - x \in \mathrm{conv}(S)$.

Lemma 2.4.25. *For fixed $\theta \in \mathbb{S}^{n-1}, t > 0, S \in \mathcal{S}$, let $b_1, \dots, b_n \in \mathbb{R}^{n-1}, h_1, \dots, h_n \in \mathbb{R}$, $x \in \omega^{\perp}$ be random variables distributed as in Lemma* 2.4.24. *Define $q := \bar{p} - x$. Then, the expected edge length satisfies*

$$\mathbb{E}[\mathrm{length}(\mathrm{conv}(b_1, \dots, b_n) \cap \bar{l}) \mid \theta, t, S, \bar{E}] \geq \mathbb{E}[\|C(q)\|_1 \mid \theta, t, S, \bar{E}] \cdot \inf_{x \in \bar{\omega}^{\perp}} \mathbb{E}[|\sum_{i=1}^{n} z_i h_i| \mid \theta, t, S, x].$$

Proof. We start with the assertion of Lemma 2.4.23:

$$\text{length}((x + q + \bar{\omega} \cdot \mathbb{R}) \cap \text{conv}(b_1, \ldots, b_n)) = \|C(q)\|_1 \cdot |\sum_{i=1}^{n} z_i h_i|.$$

We take expectation on both sides to derive the equality

$$\mathbb{E}[\text{length}(\text{conv}(b_1, \ldots, b_n) \cap \bar{l}) \mid \theta, t, S, \bar{E}] = \mathbb{E}[\|C(q)\|_1 \cdot |\sum_{i=1}^{n} z_i h_i| \mid \theta, t, S, \bar{E}].$$

Since $\|C(q)\|_1$ and $|\sum_{i=1}^{n} z_i h_i|$ do not share any of their variables, we separate the two expectations:

$$\begin{aligned}
\mathbb{E}[\|C(q)\|_1 \cdot |\sum_{i=1}^{n} z_i h_i| \mid \theta, t, S, \bar{E}] &= \mathbb{E}_{x, h_1, \ldots, h_n}[\|C(q)\|_1 \cdot |\sum_{i=1}^{n} z_i h_i| \mid \theta, t, S, \bar{E}] \\
&= \mathbb{E}_x[\|C(q)\|_1 \mathbb{E}_{h_1, \ldots, h_n}[|\sum_{i=1}^{n} z_i h_i| \mid \theta, t, S, x] \mid \theta, t, S, \bar{E}] \\
&\geq \mathbb{E}_x[\|C(q)\|_1 \mid \theta, t, S, \bar{E}] \inf_{x \in \bar{\omega}^{\perp}} \mathbb{E}_{h_1, \ldots, h_n}[|\sum_{i=1}^{n} z_i h_i| \mid \theta, t, S, x]. \quad \square
\end{aligned}$$

We will first bound the expected ℓ_1-diameter of $C(q)$, where $q = \bar{p} - x$, which depends on where $\bar{p} - x$ intersects the projected simplex $\text{conv}(S)$: where this quantity tends to get smaller as we approach the boundary of $\text{conv}(S)$. We recall that $\bar{E}$ occurs if and only if $q \in \text{conv}(S)$.

Lemma 2.4.26 (Chord combination bound)**.** *Let $\theta \in \mathbb{S}^{n-1}, t > 0$ and $S \in \mathcal{S}$ be fixed. Let $q = \bar{p} - x$ be distributed as in Lemma* 2.4.25*. Then, the expected ℓ_1-diameter of $C(q)$ satisfies*

$$\mathbb{E}[\|C(q)\|_1 \mid \theta, t, S, \bar{E}] \geq \frac{e^{-2}}{nL(1 + R_{m,n})}$$

Proof. To get a lower bound on the expected value of $\|C(q)\|_1$, we will use the concavity of $\|C(q)\|_1$ over $\text{conv}(S) = \text{conv}(s_1, \ldots, s_n)$ and that $\max_{q \in \text{conv}(S)} \|C(q)\|_1 = 2$. These facts are proven in Lemma 2.4.22. We show that shifting the projected simplex does not change the probability density too much (using log-Lipschitzness), and use the properties of $\|C(q)\|_1$ mentioned above.

Let $\hat{\mu}$ denote the probability density of q conditioned on $\theta, t, S, \bar{E}$. Note that $\hat{\mu}$ is supported on $\text{conv}(S)$ and has density proportional to

$$\int \cdots \int |\sum_{i=1}^{n} z_i h_i| \prod_{i=1}^{n} \bar{\mu}_i(\bar{p} - q + s_i + h_i \bar{\omega}) \, \mathrm{d}h_1 \cdots \mathrm{d}h_n.$$

We claim that $\hat{\mu}$ is nL-log-Lipschitz. To see this, note that since $\bar{\mu}_1, \ldots, \bar{\mu}_n$ are L-log-Lipschitz, for $v, v' \in \operatorname{conv}(S)$ we have that

$$\begin{aligned}
&\int \cdots \int |\sum_{i=1}^{n} z_i h_i| \prod_{i=1}^{n} \bar{\mu}_i(\bar{p} - v + s_i + h_i\bar{\omega}) \, \mathrm{d}h_1 \cdots \mathrm{d}h_n \\
&\leq \int \cdots \int |\sum_{i=1}^{n} z_i h_i| \prod_{i=1}^{n} e^{L\|v'-v\|} \bar{\mu}_i(\bar{p} - v' + s_i + h_i\bar{\omega}) \, \mathrm{d}h_1 \cdots \mathrm{d}h_n \\
&= e^{nL\|v'-v\|} \int \cdots \int |\sum_{i=1}^{n} z_i h_i| \prod_{i=1}^{n} \bar{\mu}_i(\bar{p} - v' + s_i + h_i\bar{\omega}) \, \mathrm{d}h_1 \cdots \mathrm{d}h_n,
\end{aligned}$$

as needed.

Let $\alpha \in (0, 1)$ be a scale factor to be chosen later, and let $y = y(S) \in \operatorname{conv}(S)$ be as in Lemma 2.4.22. Now we can write

$$\begin{aligned}
\mathbb{E}[\|C(q)\| \mid \theta, t, S, \bar{E}] &= \int_{\operatorname{conv}(S)} \|C(q)\|_1 \hat{\mu}(q) \, \mathrm{d}q \\
&\geq \int_{\alpha\operatorname{conv}(S)+(1-\alpha)y} \|C(q)\|_1 \hat{\mu}(q) \, \mathrm{d}q, \qquad (2.19)
\end{aligned}$$

because the integrand is non-negative. By concavity of $\|C(q)\|_1$ we have the lower bound $\|C(\alpha q + (1-\alpha)y)\| \geq 2(1-\alpha)$ for all $q \in \operatorname{conv}(S)$. Therefore, (2.19) is lower bounded by

$$\begin{aligned}
&\geq \int_{\alpha\operatorname{conv}(S)+(1-\alpha)y} 2(1-\alpha)\hat{\mu}(q) \, \mathrm{d}q \\
&= 2\alpha^n(1-\alpha) \int_{\operatorname{conv}(S)} \hat{\mu}(\alpha q + (1-\alpha)y) \, \mathrm{d}q \\
&\geq 2\alpha^n(1-\alpha) e^{-\max_{q\in\operatorname{conv}(S)}(1-\alpha)\|q-y\|\cdot nL} \int_{\operatorname{conv}(S)} \hat{\mu}(q) \, \mathrm{d}q, \\
&= 2\alpha^n(1-\alpha) e^{-\max_{i\in[n]}(1-\alpha)\|s_i-y\|\cdot nL}, \qquad (2.20)
\end{aligned}$$

where we used a change of variables in the first equality, the nL-log-Lipschitzness of $\hat{\mu}$ in the second inequality, and the convexity of the ℓ_2 norm in the last equality. Using the diameter bound of $2 + 2R_{m,n}$ for $\operatorname{conv}(S)$, (2.20) is lower bounded by

$$\geq 2\alpha^n(1-\alpha) e^{-(1-\alpha)nL(2+2R_{m,n})}. \qquad (2.21)$$

Setting $\alpha = 1 - \frac{1}{nL(2+2R_{m,n})} \geq 1 - 1/n$ (by Lemma 2.4.9) gives a lower bound for (2.21) of

$$\geq e^{-2} \frac{1}{nL(1+R_{m,n})}. \qquad \square$$

Recall that we have now fixed the position x and shape S of the projected simplex. The randomness we have left is in the positions $h_1, \dots, h_n$ of $b_1, \dots, b_n$ along lines parallel to the vector $\bar{\omega}$. As θ and t are also fixed, restricting b_i to lie on a line is the same as restricting a_i to lie on a line.

Lemma 2.4.27 (Height of simplex bound)**.** *Let $\theta \in \mathbb{S}^{n-1}, t \geq 0$, $S \in \mathcal{S}, x \in \bar{\omega}^{\perp}$ be fixed and let $z := z(S)$ be as in Definition* 2.4.20*. Then for $h_1, \dots, h_n \in \mathbb{R}$ distributed as in Lemma* 2.4.25*, the expected inner product satisfies*

$$\inf_{x \in \bar{\omega}^{\perp}} \mathbb{E}[|\sum_{i=1}^{n} z_i h_i| \mid \theta, t, S, x] \geq \tau/(2\sqrt{n}).$$

Proof. For fixed θ, t, S, x, let $g_1, \dots, g_n \in \mathbb{R}$ be independent random variables with respective probability densities $\tilde{\mu}_1, \dots, \tilde{\mu}_n$, where $\tilde{\mu}_i, i \in [n]$, is defined by

$$\tilde{\mu}_i(g_i) := \bar{\mu}(x + s_i + g_i\bar{\omega}) = \mu(R_\theta(x + s_i + g_i\bar{\omega}) + t\theta) \, .$$

Note that, by assumption, the variables $g_1, \dots, g_n$ each have variance at least τ^2. We recall from Lemma 2.4.24 that the joint probability density of $h_1, \dots, h_n$ is proportional to $|\sum_{i=1}^n z_i h_i| \prod_{i=1}^n \tilde{\mu}_i(h_i)$. Thus, we may rewrite the above expectation as

$$\begin{aligned}\mathbb{E}[|\sum_{i=1}^{n} z_i h_i| \mid \theta, t, S, x] &= \frac{\int \cdots \int_{\mathbb{R}} |\sum_{i=1}^n z_i h_i|^2 \prod_{i=1}^n \tilde{\mu}_i(h_i) \, \mathrm{d}h_1 \cdots \mathrm{d}h_n}{\int \cdots \int_{\mathbb{R}} |\sum_{i=1}^n z_i h_i| \prod_{i=1}^n \tilde{\mu}_i(h_i) \, \mathrm{d}h_1 \cdots \mathrm{d}h_n} \\ &= \frac{\mathbb{E}[|\sum_{i=1}^n z_i g_i|^2]}{\mathbb{E}[|\sum_{i=1}^n z_i g_i|]} \, ,\end{aligned}$$

where $g_1, \dots, g_n$ are distributed independently with densities $\tilde{\mu}_1, \dots, \tilde{\mu}_n$. By the additivity of variance for independent random variables, we see that

$$\mathrm{Var}(\sum_{i=1}^{n} z_i g_i) = \sum_{i=1}^{n} z_i^2 \mathrm{Var}(g_i) \geq \tau^2 \|z\|^2 \geq \tau^2 \|z\|_1^2/n = \tau^2/n.$$

We reach the desired conclusion by applying Lemma 2.3.2:

$$\frac{\mathbb{E}[|\sum_{i=1}^n z_i g_i|^2]}{\mathbb{E}[|\sum_{i=1}^n z_i g_i|]} \geq \frac{|\mathbb{E}[\sum_{i=1}^n z_i g_i]| + \sqrt{\mathrm{Var}(\sum_{i=1}^n z_i g_i)}}{2} \geq \tau/(2\sqrt{n}). \qquad \square$$

Using the bounds from the preceding lemmas, the proof of our main theorem is now given below.

Proof of Theorem 2.4.10 (Parametrized Shadow Bound). By Lemma 2.4.13, we derive the shadow bound by combining an upper bound on $\mathbb{E}[\text{perimeter}(Q(A) \cap W)]$ and a uniform lower bound on $\mathbb{E}[\text{length}(\text{conv}(a_i : i \in I) \cap W) \mid E_I]$ for all $I \in B$. For the perimeter upper bound, by Lemma 2.4.14 we have that

$$\mathbb{E}[\text{perimeter}(Q(A) \cap W)] \leq 2\pi(1 + 4r_m). \tag{2.22}$$

For the edge length bound, we assume w.l.o.g. as above that $I = [n]$. Combining prior lemmas, we have that

$$\begin{aligned}
&\mathbb{E}[\text{length}(\text{conv}(a_1, \ldots, a_n) \cap W) \mid E] \\
&\geq \frac{1}{2} \cdot \mathbb{E}[\text{length}(\text{conv}(a_1, \ldots, a_n) \cap W) \mid D, E] \quad (\text{ Lemma 2.4.17 }) \\
&\geq \frac{1}{2} \cdot \inf_{\substack{\theta \in \mathbb{S}^{n-1} \\ t>0}} \mathbb{E}[\text{length}(\text{conv}(b_1, \ldots, b_n) \cap \bar{l}) \mid \theta, t, S \in \mathcal{S}, \bar{E}] \quad (\text{ Lemma 2.4.19 }) \\
&\geq \frac{1}{2} \cdot \inf_{\substack{\theta \in \mathbb{S}^{n-1} \\ t>0, S \in \mathcal{S}}} \left(\mathbb{E}[\|C(\bar{p} - x)\|_1 \mid \theta, t, S, \bar{E}] \cdot \inf_{x \in \bar{\omega}^\perp} \mathbb{E}[|\sum_{i=1}^{n} z_i h_i| \mid \theta, t, S, x] \right) \\
&\qquad\qquad (\text{ Lemma 2.4.25 }) \\
&\geq \frac{1}{2} \cdot \frac{e^{-2}}{nL(1 + R_{m,n})} \cdot \frac{\tau}{2\sqrt{n}} \quad (\text{ Lemma 2.4.26 and Lemma 2.4.27 }) .
\end{aligned} \tag{2.23}$$

The theorem now follows by taking the ratio of (2.22) and (2.23). □

2.4.2 Shadow bound for Laplace perturbations

Theorem 2.4.10 is most naturally used to prove shadow bounds or distributions where all parameters are bounded, which we illustrate here for Laplace-distributed perturbations. The Laplace distribution is defined in section 2.2. To achieve the shadow bound, we use the abstract shadow bound as a black box, and we bound the necessary parameters of the Laplace distribution below.

Lemma 2.4.28. *For $m \geq n \geq 3$, the Laplace distribution $L_n(\bar{a}, \sigma)$, satisfies the following properties*:

1. *The density is $\sqrt{n}/\sigma$-log-Lipschitz.*
2. *Its cutoff radius satisfies $R_{m,n} \leq 14\sigma\sqrt{n}\log m$.*
3. *The n-th deviation satisfies $r_m \leq 7\sigma \log m$.*
4. *The variance after restricting to any line satisfies $\tau \geq \sigma/\sqrt{ne}$.*

Proof. By shift invariance of the parameters, we may assume w.l.o.g. that $\bar{a} = \vec{0}$. Let X be distributed as $L_n(\vec{0}, \sigma)$ for use below.

1 The density of the Laplace distribution is proportional to $e^{-\|x\|\sqrt{n}/\sigma}$, for $x \in \mathbb{R}^n$, and thus the logarithm of the density differs an additive constant from $-\|x\|\sqrt{n}/\sigma$, which is clearly $\sqrt{n}/\sigma$-Lipschitz.

2 The second property follows from Lemma 2.2.5:

$$\begin{aligned}\Pr[\|X\| \geq 14\sigma\sqrt{n}\log m] &\leq e^{-2n\log m} = m^{-2n} \\ &\leq \frac{1}{n\binom{m}{n}}.\end{aligned}$$

3 Again from Lemma 2.2.5. If $7\log m \geq 2\sqrt{n}$, we get that

$$\begin{aligned}\int_{7\sigma\log m}^{\infty} \Pr[|X^{\mathsf{T}}\theta| \geq t]\,\mathrm{d}t &\leq \int_{7\sigma\log m}^{\infty} e^{-\sqrt{n}t/(7\sigma)}\,\mathrm{d}t \\ &= \frac{7\sigma}{\sqrt{n}} m^{-\sqrt{n}\log m} \leq \frac{7\sigma\log m}{m}.\end{aligned}$$

If $7\log m \leq 2\sqrt{n}$, then

$$\begin{aligned}\int_{7\sigma\log m}^{\infty} \Pr[|X^{\mathsf{T}}\theta| \geq t]\,\mathrm{d}t &= \int_{7\sigma\log m}^{2\sigma\sqrt{n}} \Pr[|X^{\mathsf{T}}\theta| \geq t]\,\mathrm{d}t + \int_{2\sigma\sqrt{n}}^{\infty} \Pr[|X^{\mathsf{T}}\theta| \geq t]\,\mathrm{d}t \\ &\leq \int_{7\sigma\log m}^{2\sigma\sqrt{n}} 2e^{-t^2/(16\sigma^2)}\,\mathrm{d}t + \int_{2\sigma\sqrt{n}}^{\infty} e^{-\sqrt{n}t/(7\sigma)}\,\mathrm{d}t \\ &\leq 4\sigma\sqrt{n}e^{-(7\log m)^2/16} + \frac{7\sigma}{\sqrt{n}}e^{-2d/7} \\ &\leq 4\sigma\sqrt{n}/m^3 + 7\sigma/(\sqrt{n}m^{\sqrt{n}}) \leq \frac{7\sigma\log m}{m}.\end{aligned}$$

4 This follows from the $\sqrt{n}/\sigma$-log-Lipschitzness and Lemma 2.4.7. □

Proof of Theorem 2.4.2 (Shadow bound for Laplace perturbations). We get the desired result by plugging in the bounds from Lemma 2.4.28 for $L, R_{m,n}, r_m$ and τ into the upper bound $O((n^{1.5}L/\tau)(1 + R_{m,n})(1 + r_m))$ from Theorem 2.4.10. □

2.4.3 Shadow bound for Gaussian perturbations

In this subsection, we prove our shadow bound for Gaussian perturbations.

The Gaussian distribution is not log-Lipschitz, so we can not directly apply Theorem 2.4.10. We will define a *smoothed out* version of the Gaussian distribution to remedy this problem, which we call the Laplace-Gaussian distribution. The Laplace-Gaussian distribution, defined below, matches the Gaussian distribution in every meaningful parameter, while also being log-Lipschitz. We will first bound the shadow size for Laplace-Gaussian perturbations, and then show that the expected number of edges of $Q(A) \cap W$ for Gaussian perturbations is at most 1 larger.

Definition 2.4.29. We define a random variable $X \in \mathbb{R}^n$ to be (σ, r)-*Laplace-Gaussian distributed with mean* $\bar{a}$, or $X \sim LG_n(\bar{a}, \sigma, r)$, if its density is proportional to $f_{(\bar{a},\sigma,r)} : \mathbb{R}^n \to \mathbb{R}_+$ given by

$$f_{(\bar{a},\sigma,r)}(x) = \begin{cases} e^{-\|x-\bar{a}\|^2/(2\sigma^2)} & \text{if } \|x-\bar{a}\| \le r\sigma \\ e^{-\|x-\bar{a}\|r/\sigma + r^2/2} & \text{if } \|x-\bar{a}\| \ge r\sigma. \end{cases}$$

Note that at $\|x - \bar{a}\| = r\sigma$, both cases give the density $e^{-r^2/2}$, and hence $f_{(\bar{a},\sigma,r)}$ is well-defined and continuous on $\mathbb{R}^n$. For distributions with mean $\vec{0}$, we abbreviate $f_{(\sigma,r)} := f_{(\vec{0},\sigma,r)}$ and $LG_n(\sigma, r) := LG_n(\vec{0}, \sigma, r)$.

Just like for the shadow size bound for Laplace perturbations, we need strong enough tail bounds. We state these tail bounds here, and defer their proofs till the end of the section.

Lemma 2.4.30 (Laplace-Gaussian tail bounds)**.** *Let $X \in \mathbb{R}^n$ be (σ, r)-Laplace-Gaussian distributed with mean $\vec{0}$, where $r := c\sqrt{n \log m}$, $c \ge 4$. Then for $t \ge r$,*

$$\Pr[\|X\| \ge \sigma t] \le e^{-(1/4)rt}. \tag{2.24}$$

For $\theta \in \mathbb{S}^{n-1}$, $t \ge 0$,

$$\Pr[|X^\mathsf{T}\theta| \ge \sigma t] \le \begin{cases} e^{-(1/4)rt} & : t \ge r \\ 3e^{-t^2/4} & : 0 \le t \le r. \end{cases} \tag{2.25}$$

Lemma 2.4.31. *For $m \ge n \ge 3$, the $(\sigma, 4\sqrt{n \log m})$-Laplace-Gaussian distribution in $\mathbb{R}^n$ with mean $\bar{a}$ satisfies the following properties*:

1. *The density is $4\sigma^{-1}\sqrt{n \log m}$-log-Lipschitz.*
2. *Its cutoff radius satisfies $R_{m,n} \le 4\sigma\sqrt{n \log m}$.*

3. *The n-th deviation is* $r_m \leq 4\sigma\sqrt{\log m}$.

4. *The variance after restricting to any line satisfies* $\tau \geq \sigma/4$.

Proof. As before, by shift invariance, we may assume w.l.o.g that $\bar{a} = \vec{0}$. Let $X \sim LG_n(\sigma, 4\sqrt{n \log m})$ and let $r := 4\sqrt{n \log m}$.

1 The gradient of the function $\log(f_{(\sigma,r)}(x))$ has norm bounded by $4\sigma^{-1}\sqrt{n \log m}$ wherever it is defined, which by continuity implies $f_{(\sigma,r)}$ is $4\sigma^{-1}\sqrt{n \log m}$-log-Lipschitz.

2 Applying the tail bound from Lemma 2.4.30, we get that

$$\Pr[\|X\| \geq 4\sigma\sqrt{n \log m}] \leq e^{-4n \log m} \leq \frac{1}{n\binom{m}{n}}.$$

3 Again using Lemma 2.4.30,

$$\begin{aligned}
\int_{4\sigma\sqrt{\log m}}^{\infty} \Pr[|X^\mathsf{T}\theta| \geq t] \, \mathrm{d}t &= \int_{4\sigma\sqrt{\log m}}^{r\sigma} \Pr[|X^\mathsf{T}\theta| \geq t] \, \mathrm{d}t + \int_{r\sigma}^{\infty} \Pr[|X^\mathsf{T}\theta| \geq t] \, \mathrm{d}t \\
&\leq \int_{4\sigma\sqrt{\log m}}^{r\sigma} 3e^{-t^2/(4\sigma^2)} \, \mathrm{d}t + \int_{r\sigma}^{\infty} e^{-\sqrt{n \log m}\, t/\sigma} \, \mathrm{d}t \\
&\leq 4\sigma\sqrt{n \log m}(3m^{-4}) + \frac{\sigma}{\sqrt{n \log m}} m^{-4n} \\
&\leq 4\sigma\sqrt{\log m}/m.
\end{aligned}$$

4 For the line variance, by rotational symmetry, we may without loss of generality assume that $l := (y, 0) + e_n\mathbb{R}$, where $y \in \mathbb{R}^{n-1}$, and so $(y, 0)$ is the point on l closest to the origin. Since $f_{(\sigma,r)}((y, \lambda)) = f_{(\sigma,r)}((y, -\lambda))$ for every $\lambda \in \mathbb{R}$, the expectation $\mathbb{E}[X \mid X \in l] = (y, 0)$. Thus, $\mathrm{Var}(X \mid X \in l) = \mathbb{E}[X_n^2 \mid X \in l]$.

Let $\bar{l} = (y, 0) + [-\sigma, \sigma] \cdot e_n$. Since $|X_n|$ is larger on $l \setminus \bar{l}$ than on $\bar{l}$, we clearly have $\mathbb{E}[X_n^2 \mid X \in l] \geq \mathbb{E}[X_n^2 \mid X \in \bar{l}]$, so it suffices to lower bound the latter quantity.

For each y with $\|y\| \leq \sigma r$ we have for all $\lambda \in [-\sigma, \sigma]$ the inequality

$$1 \geq \frac{f_{(\sigma,r)}((y, \lambda))}{f_{(\sigma,r)}((y, 0))} \geq \frac{e^{-\|(y,\lambda)\|^2/(2\sigma^2)}}{e^{-\|(y,0)\|^2/(2\sigma^2)}} = e^{-\lambda^2/(2\sigma^2)} \geq e^{-1/2}. \tag{2.26}$$

Given the above, we have that

$$\begin{aligned}\mathbb{E}[X_n^2 \mid X \in \bar{l}] &\geq (\sigma^2/4)\Pr[|X_n| \geq \sigma/2 \mid X \in \bar{l}] \\ &= (\sigma^2/4)\frac{\int_{\sigma/2}^{\sigma} f_{(\sigma,r)}((y,t))\,\mathrm{d}t}{\int_0^{\sigma} f_{(\sigma,r)}((y,t))\,\mathrm{d}t} \\ &\geq (\sigma^2/4)\frac{\int_{\sigma/2}^{\sigma} f_{(\sigma,r)}((y,0))e^{-1/2}\,\mathrm{d}t}{\int_0^{\sigma} f_{(\sigma,r)}((y,0))\,\mathrm{d}t} \quad (\text{ by (2.26) }) \\ &= (\sigma^2/4)(e^{-1/2}/2) \geq \sigma^2/16 \ , \text{as needed .}\end{aligned} \tag{2.27}$$

For y with $\|y\| \geq \sigma r$, $\lambda \in [-\sigma, \sigma]$, we similarly have

$$\begin{aligned}\|(y,\lambda)\| &= \sqrt{\|y\|^2 + \lambda^2} \\ &\leq \|y\| + \frac{\lambda^2}{2\|y\|} \leq \|y\| + \frac{\lambda^2}{2r\sigma}.\end{aligned}$$

In particular, we get that

$$1 \geq \frac{f_{(\sigma,r)}((y,\lambda))}{f_{(\sigma,r)}((y,0))} = \frac{e^{-\|(y,\lambda)\|(r/\sigma)}}{e^{-\|(y,0)\|(r/\sigma)}} \geq e^{-\lambda^2/(2\sigma^2)} \geq e^{-1/2}\,. \tag{2.28}$$

The desired lower bound now follows by combining (2.27), (2.28). □

Given any unperturbed unit LP given by $c, \bar{a}_1, \ldots, \bar{a}_m$, we denote by $\mathbb{E}_{N_n(\sigma)}$ the expectation when its vertices are perturbed with noise distributed according to the Gaussian distribution of standard deviation σ and we write $\mathbb{E}_{LG_n(\sigma,r)}$ for the expectation when its vertices are perturbed by (σ, r)-Laplace-Gaussian noise. The same applies for $\Pr_{N_n(\sigma)}$ and $\Pr_{LG_n(\sigma,r)}$.

In the next lemma we prove that, for $r := 4\sqrt{n \log m}$, the expected number of edges for Gaussian distributed perturbations is not much bigger than the expected number for Laplace-Gaussian perturbations. We use the strong tail bounds we have on the two distributions along with the knowledge that restricted to a ball of radius $r\sigma$ the probability densities are equal. Recall that we use $\hat{a}_i$ to denote the perturbation $a_i - \mathbb{E}[a_i]$.

Lemma 2.4.32. *For $n \geq 3$, the number of edges in $Q(A) \cap W$ satisfies*

$$\begin{aligned}&\mathbb{E}_{N_n(\sigma)}[|\text{edges}(Q(A) \cap W)|] \\ &\qquad \leq 1 + \mathbb{E}_{LG_n(\sigma, 4\sqrt{n \log m})}[|\text{edges}(Q(A) \cap W)|].\end{aligned}$$

Proof. Let us abbreviate the edge count $C(A, W) := |\text{edges}(Q(A) \cap W)|$ and let $r := 4\sqrt{n \log m}$. We make use of the fact that $N_n(\sigma)$ and $LG_n(\sigma, r)$ are equal when restricted to distance at most σr from their centers.

$$\mathbb{E}_{N(\sigma)}[C(A, W)] \tag{2.29}$$
$$= \Pr_{N_n(\sigma)}[\exists i \in [m] \ \|\hat{a}_i\| > \sigma r]\mathbb{E}_{N_n(\sigma)}[C(A, W) \mid \exists i \in [m] \ \|\hat{a}_i\| > \sigma r]$$
$$+ \Pr_{N_n(\sigma)}[\forall i \in [m] \ \|\hat{a}_i\| \leq \sigma r]\mathbb{E}_{N_n(\sigma)}[C(A, W) \mid \forall i \in [m] \ \|\hat{a}_i\| \leq \sigma r]. \tag{2.30}$$

By Lemma 2.4.30, the first probability is at most $m^{-4n} \leq m^{-n}/4$, so we upper bound the first number of edges by $\binom{m}{n}$ making a total contribution of less than $1/4$. Now we use the fact that within radius $4\sigma\sqrt{n \log m}$ we have equality of densities between $N_n(\sigma)$ and $LG_n(\sigma, r)$. Continuing from (2.30),

$$\leq 1/4 + \mathbb{E}_{N_n(\sigma)}[C(A, W) \mid \forall i \in [m] \ \|\hat{a}_i\| \leq \sigma r]$$
$$= 1/4 + \mathbb{E}_{LG_n(\sigma,r)}[C(A, W) \mid \forall i \in [m] \ \|\hat{a}_i\| \leq \sigma r]$$
$$\leq 1/4 + \mathbb{E}_{LG_n(\sigma,r)}[C(A, W)]/\Pr_{LG_n(\sigma,r)}[\forall i \in [m] \ \|\hat{a}_i\| \leq \sigma r]. \tag{2.31}$$

The inequality above is true by non-negativity of the number of edges. Next we lower bound the denominator and continue (2.31),

$$\leq 1/4 + \mathbb{E}_{LG_n(\sigma,r)}[C(A, W)]/(1 - m^{-n}/4)$$
$$\leq 1/4 + (1 + m^{-n}/2)\mathbb{E}_{LG_n(\sigma,r)}[C(A, W)]. \tag{2.32}$$

The last inequality we deduce from the fact that $(1 - \varepsilon)(1 + 2\varepsilon) = 1 + \varepsilon - 2\varepsilon^2$, which is bigger than 1 for $0 < \varepsilon < 1/2$. Again using the trivial upper bound of $\binom{m}{n}$ edges, we arrive at our desired conclusion that

$$\mathbb{E}_{N_n(\sigma)}[C(A, W)] \leq 1 + \mathbb{E}_{LG_n(\sigma,r)}[C(A, W)]. \qquad \square$$

We now have all the ingredients to prove our bound on the expected number of edges for Gaussian perturbations.

Proof of Theorem 2.4.1 (Shadow bound for Gaussian perturbations).
By Lemma 2.4.32, we know that

$$\mathbb{E}_{N_n(\sigma)}[|\text{edges}(Q(A) \cap W)|]$$
$$\leq 1 + \mathbb{E}_{LG_n(\sigma, 4\sqrt{n \log m})}[|\text{edges}(Q(A) \cap W)|].$$

We now derive the shadow bound for Laplace-Gaussian perturbations by combining the parameter bounds in Lemma 2.4.31 with the parameterized shadow bound in Theorem 2.4.10. $\square$

We now prove the tail bounds for Laplace-Gaussian distributions. Recall that we set $r := c\sqrt{n \log m}$ with $c \geq 4$.

Proof of Lemma 2.4.30 (Tail bound for Laplace-Gaussian distribution). By homogeneity, we may w.l.o.g. assume that $\sigma = 1$. Define auxiliary random variables $Y \in \mathbb{R}^n$ distributed as $(\vec{0}, 1/(c\sqrt{\log m}))$-Laplace and $Z \in \mathbb{R}^n$ be distributed as $N_n(\vec{0}, 1)$.

Since X has density proportional to $f_{(1,r)}(x)$, which equals $e^{-\|x\|^2/2}$ for $\|x\| \leq r$ and $e^{-r\|x\|+r^2/2}$ for $\|x\| \geq r$, we immediately see that

$$\begin{aligned} Z \mid \|Z\| \leq r \;&\equiv\; X \mid \|X\| \leq r \\ Y \mid \|Y\| \geq r \;&\equiv\; X \mid \|X\| \geq r \end{aligned} \tag{2.33}$$

Proof of (2.24) By the above, for any $t \geq r$, we have that

$$\Pr[\|X\| \geq t] = \Pr[\|Y\| \geq t] \cdot \frac{\Pr[\|X\| \geq r]}{\Pr[\|Y\| \geq r]} \,. \tag{2.34}$$

For the first term, by the Laplace tail bound (2.9), we get that

$$\Pr[\|Y\| \geq t] \leq e^{-rt - n\log(\frac{c\sqrt{\log mt}}{\sqrt{n}}) - n} \,. \tag{2.35}$$

For the second term,

$$\begin{aligned} \frac{\Pr[\|X\| \geq r]}{\Pr[\|Y\| \geq r]} &= e^{r^2/2} \frac{\int_{\mathbb{R}^n} e^{-r\|x\|}\,\mathrm{d}x}{\int_{\mathbb{R}^n} f_{(\sigma,r)}(x)\,\mathrm{d}x} \leq e^{r^2/2} \frac{\int_{\mathbb{R}^n} e^{-r\|x\|}\,\mathrm{d}x}{\int_{\mathbb{R}^n} e^{-\|x\|^2/2}\,\mathrm{d}x} \\ &\leq e^{r^2/2} \frac{r^{-n} n! \mathrm{vol}_n(\mathbb{B}_2^n)}{\sqrt{2\pi}^n} \leq e^{(nc^2\log m)/2} (\frac{\sqrt{e}}{c\sqrt{\log m}})^n \\ &\leq e^{(nc^2\log m)/2} \,, \end{aligned} \tag{2.36}$$

where we have used the upper bound $\mathrm{vol}_n(\mathbb{B}_2^n) \leq (2\pi e/n)^{n/2}$, $r = c\sqrt{n\log m}$ and $c \geq \sqrt{e}$. Combining (2.35), (2.36) and that $t \geq r$, $c \geq 4$, we get

$$\begin{aligned} \Pr[\|X\| \geq t] &\leq e^{-rt - n\log(\frac{c\sqrt{\log mt}}{\sqrt{n}}) - n} \cdot e^{(nc^2\log m)/2} \\ &\leq e^{-rt/2 - n\log(\frac{c\sqrt{\log mt}}{\sqrt{n}}) - n} = e^{-n(\frac{rt}{2n} - \log(\frac{rt}{n}) - 1)} \\ &\leq e^{-n(\frac{rt}{4n})} = e^{-rt/4}, \end{aligned} \tag{2.37}$$

where the last inequality follows from $x/2 - \log(x) - 1 \geq x/4$, for $x \geq rt/n \geq c^2 \geq 16$.

Proof of (2.25) For $t \geq r$, using the bound (2.24), we get

$$\Pr[|X^{\mathsf{T}}\theta| \geq t] \leq \Pr[\|X\| \geq t] \leq e^{-c\sqrt{n \log mt}/4} . \tag{2.38}$$

For $t \leq r$, we see that

$$\begin{aligned} \Pr[|X^{\mathsf{T}}\theta| \geq t] &\leq \Pr[|X^{\mathsf{T}}\theta| \geq t, \|X\| \leq r] + \Pr[\|X\| \geq r] \\ &\leq \Pr[|X^{\mathsf{T}}\theta| \geq t, \|X\| \leq r] + e^{-r^2/4} . \end{aligned} \tag{2.39}$$

By the identity (2.33), for the first term, using the Gaussian tail bound (2.8), we have that

$$\begin{aligned} \Pr[|X^{\mathsf{T}}\theta| \geq t, \|X\| \leq r] &= \Pr[|Z^{\mathsf{T}}\theta| \geq t, \|Z\| \leq r] \cdot \frac{\Pr[\|X\| \leq r]}{\Pr[\|Z\| \leq r]} \\ &= \Pr[|Z^{\mathsf{T}}\theta| \geq t, \|Z\| \leq r] \cdot \frac{\int_{\mathbb{R}^n} e^{-\|x\|^2/2}\, \mathrm{d}x}{\int_{\mathbb{R}^n} f_{(1,r)}(x)\, \mathrm{d}x} \\ &\leq \Pr[|Z^{\mathsf{T}}\theta| \geq t] \leq 2e^{-t^2/2} . \end{aligned} \tag{2.40}$$

The desired inequality (2.25) now follows directly by combining (2.38), (2.39), (2.40), noting that $2e^{-t^2/2} + e^{-r^2/4} \leq 3e^{-t^2/4}$ for $0 \leq t \leq r$. □

2.5 Analyzing Simplex Algorithms

In this section, we describe how to use the shadow bound to bound the complexity of a complete shadow vertex simplex based algorithm. We will follow the two-stage interpolation strategy given by Vershynin in [200], and describe its usage with the dimension-by-dimension algorithm [28] as well as a faster variant of the Random Vertex algorithm from [200].

We will say that (Smoothed LP) is unbounded if the system $c^{\mathsf{T}}x > 0, Ax \leq \vec{0}$ is feasible, and bounded if this system is infeasible. Note that, under this definition, (Smoothed LP) can be simultaneously unbounded and infeasible under this definition. In that case, we are satisfied if an algorithm reports that the program is unbounded. If (Smoothed LP) is both bounded and feasible, then it has an optimal feasible solution.

For the execution of the algorithms as stated, we assume the non-degeneracy conditions listed in Theorem 2.2.11. That is, we assume both the feasible polyhedron and shadows to be non-degenerate. These conditions hold with probability 1.

2.5.1 Two-Phase Interpolation Method

We wish to apply the shadow vertex method to LP's with $b = \vec{1}$. To reduce solving any LP to this case, first define the Phase I Unit LP:

$$\begin{aligned} &\max\ c^\mathsf{T} x && \text{(Unit LP)} \\ &\quad Ax \leq \vec{1} \end{aligned}$$

and the Phase II interpolation LP with parametric objective for $\theta \in (-\pi/2, \pi/2)$:

$$\begin{aligned} &\max\ \cos(\theta) c^\mathsf{T} x + \sin(\theta)\lambda && \text{(Int. LP)} \\ &\quad Ax + (\vec{1} - b)\lambda \leq \vec{1} \\ &\quad 0 \leq \lambda \leq 1. \end{aligned}$$

The above form of interpolation was first introduced in the context of smoothed analysis by Vershynin [200].

Let us assume for the moment that (Smoothed LP) is bounded and feasible (i.e., has an optimal solution). Since boundedness is a property of A and not b, note that this implies that (Unit LP) is also bounded (and clearly always feasible).

To understand the Phase II interpolation LP, the key observation is that for θ sufficiently close to $-\pi/2$, the maximizer will be the optimal solution to (Unit LP), i.e., will satisfy $\lambda = 0$, and for θ sufficiently close to $\pi/2$ the maximizer will be the optimal solution to (Smoothed LP), i.e., will satisfy $\lambda = 1$. Thus given an optimal solution to the Phase I unit LP one can initialize a run of shadow vertex starting at θ just above $-\pi/2$, moving towards $\pi/2$ until the optimal solution to (Smoothed LP) is found. The corresponding shadow plane is generated by $(c, 0)$ and $(\vec{0}, 1)$ (associating λ with the last coordinate), and as usual the size of the shadow bounds the number of pivots.

If (Smoothed LP) is unbounded (i.e., the system $c^\mathsf{T} x > 0, Ax \leq \vec{0}$ is feasible), this will be detected during Phase I as (Unit LP) is also unbounded. If (Smoothed LP) is infeasible but bounded, then the shadow vertex run will terminate at a vertex having $\lambda < 1$. Thus, all cases can be detected by the two-phase procedure (see [200, Proposition 4.1] for a formal proof).

We bound the number of pivot steps taken to solve (Int. LP) given a solution to (Unit LP), and after that we describe how to solve (Unit LP).

Consider polyhedron $P' = \{(x, \lambda) \in \mathbb{R}^{n+1} : Ax + (\vec{1} - b)\lambda \leq \vec{1}\}$, the slab $H = \{(x, \lambda) \in \mathbb{R}^{n+1} : 0 \leq \lambda \leq 1\}$ and let $W = \text{span}(c, e^{n+1})$. In this notation, $P' \cap H$ is the feasible set of (Int. LP) and W is the shadow plane of (Int. LP). We know that it suffices to bound the number of vertices of $\pi_W(P' \cap H)$ of (Int. LP), which we do by relating it to $\pi_W(P')$.

The constraints defining P' are of smoothed unit type. Namely, the rows of $(A, \vec{1} - b)$ are variance σ^2 Gaussians centered at means of norm at most 2. We derive this from the triangle inequality. Thus, we know $\pi_W(P')$ has at most $\mathcal{D}_g(n+1, m, \sigma/2)$ expected vertices, where we denote by $\mathcal{D}_g(\cdot, \cdot, \cdot)$ the upper bound from Theorem 2.4.1. We divided σ by 2 because to apply the shadow bound we first scale down the data such that the expected rows have norm at most 1.

Since the shadow plane contains the normal vector $(\vec{0}, 1)$ to the inequalities $0 \leq \lambda \leq 1$, these constraints intersect the shadow plane W at right angles. It follows that $\pi_W(P' \cap H) = \pi_W(P') \cap H$. Adding 2 constraints to a 2D polyhedron can add at most 2 new edges, hence the constraints on λ can add at most 4 new vertices. By combining these observations, we directly derive the following lemma of Vershynin [200].

Lemma 2.5.1. *If* (Unit LP) *is unbounded, then* (Smooth LP) *is unbounded. If* (Unit LP) *is bounded, then given an optimal solution to* (Unit LP) *one can solve* (Smoothed LP) *using at most an expected* $\mathcal{D}_g(n+1, m, \sigma/2) + 4$ *shadow vertex pivots over* (*Int. LP*).

Given the above, our main task is now to solve (Unit LP), i.e., either to find an optimal solution or to determine unboundedness. The simplest algorithm is Borgwardt's dimension-by-dimension (DD) algorithm, which was first used in the context of smoothed analysis by Schnalzger [168].

2.5.2 DD algorithm

As outlined in the introduction, the DD algorithm solves (Unit LP) by iteratively solving the restrictions:

$$\begin{aligned} \max\ & (c^k)^\mathsf{T} x && \text{(Unit LP}_k\text{)} \\ & Ax \leq \vec{1} \\ & x_i = 0,\ \forall i \in \{k+1, \ldots, n\}, \end{aligned}$$

where $k \in [n]$ and $c^k := (c_1, \ldots, c_k, 0, \ldots, 0)$. The main idea here is that the solution of (Unit LP$_k$), $k \in [n-1]$, is generically on an edge of the shadow of (Unit LP$_{k+1}$) on the span of c^k and e_{k+1}, which is sufficient to initialize the shadow simplex path in the next step. We note that Borgwardt's algorithm can be applied to any LP with a known feasible point as long as appropriate non-degeneracy conditions hold (which occur with probability 1 for smoothed LPs). To avoid degeneracy, we will assume that $c^k \neq \vec{0}$ for all $k \in [n]$, which can always be achieved by permuting the coordinates. Note that (Unit LP$_1$) can be trivially solved, as the feasible region is an interval whose endpoints are easy to compute.

Theorem 2.5.2 ([27]). *Let W_k, $k \in \{2, \ldots, n\}$, denote the shadow of* (Unit LP_k) *on the span of c^{k-1} and e_k. Then, if each* (Unit LP_k) *and shadow W_k is non-degenerate, for $k \in \{2, \ldots, n\}$, the DD algorithm solves* (Unit LP) *using at most $\sum_{k=2}^{n} |\mathrm{vertices}(W_k)|$ number of pivots.*

Using the shadow bound of Theorem 2.4.1 for $n \geq 3$ and the $O(1/\sigma + \sqrt{\log m})$ bound for $n = 2$ (Chapter 2) we immediately derive the following corollary.

Corollary 2.5.3. *The smoothed* (Unit LP) *can be solved by the DD algorithm using an expected $\sum_{k=2}^{n} \mathcal{D}_g(k, m, \sigma) = O(n^3 \sqrt{\log m}\, \sigma^{-2} + n^{3.5}\sigma^{-1} \log m + n^{3.5} \log(m)^{1.5})$ number of shadow vertex pivots.*

Combining Lemma 2.5.1 and Corollary 2.5.3 we find the conclusion of this subsection:

Theorem 2.5.4. (Smoothed LP) *can be solved by a two-phase shadow simplex method using an expected number of pivots of $O(n^3 \sqrt{\log m}\, \sigma^{-2} + n^{3.5}\sigma^{-1} \log m + n^{3.5} \log(m)^{1.5})$.*

2.5.3 Random vertex method

Another procedure for solving (Unit LP) can be found in [200]. In that reference, the approach for initializing the shadow simplex method on (Unit LP) is to add a random smoothed system of n linear constraints to its description. These constraints are meant to induce a *known* random vertex v and corresponding maximizing objective d which are effectively uncorrelated with the original system. Starting at this vertex v, we then follow the shadow path induced by rotating d towards c. The main difficulty with this approach is to guarantee that the randomly generated system:

(i) adds a vertex

(ii) which is optimized at d, and

(iii) does not cut off the optimal solution or all unbounded rays.

Fortunately, each of these conditions is easily checkable, and hence if they fail (which will occur with constant probability), the process can be attempted again.

One restriction imposed by this approach is that the perturbation size needs to be rather small, namely

$$\sigma \leq \sigma_1 := \frac{c_1}{\max\{\sqrt{n \log m}, n^{1.5} \log m\}}$$

in [200] for some $c_1 > 0$. A more careful analysis can relax the restriction to

$$\sigma \leq \sigma_2 := \frac{c_2}{\max\{\sqrt{n \log m}, \sqrt{n} \log n\}}$$

for some $c_2 > 0$. This restriction is necessary due to the fact that we wish to predict the effect of smoothing the added constraints. In particular, the smoothing operation should not negate (i), (ii), or (iii). Recall that one can always artificially decrease σ by scaling down the matrix A as this does not change the structure of (Unit LP). The assumption on σ is thus without loss of generality. When stating running time bounds however, this restriction will be reflected by a larger additive term that does not depend on σ.

We adapt the Random Vertex algorithm to make (ii) guaranteed to hold, allowing us to relax the constraint on the perturbation size to

$$\sigma \leq \bar{\sigma} := \frac{1}{36\sqrt{n \log m}}. \tag{2.41}$$

Instead of adding n constraints, each with their own perturbation, we add $n - 1$ pairs of constraints with mirrored perturbations. This forces the desired objective to be maximized at the random vertex whenever this vertex exists.

Our algorithm is printed as Algorithm 2. We begin with some preliminary remarks. First, the goal of defining V is to create a new artificial LP, (Unit LP') $\max c^\mathsf{T} x, Ax \leq \vec{1}, Vx \leq \vec{1}$, such that x_0 is a vertex of the corresponding system which maximizes d. On line 9 and 10, the algorithm checks if x_0 is feasible and whether it is not the optimizer of c on (Unit LP'). Having passed these checks, (Unit LP') is solved via shadow vertex initialized at vertex x_0 with objective d. An unbounded solution to (Unit LP') is always an unbounded solution to (Unit LP). Lastly, it is checked on line 13 whether the bounded solution (if it exists) to (Unit LP') is a solution to (Unit LP). Correctness of the algorithm's output is thus straightforward. We do have to make sure that every step of the algorithm can be executed as described.

Lemma 2.5.5. *In* (Unit LP') *as defined on lines* 3-11 *of Algorithm* 2*, with probability* 1*,* x_0 *is well-defined, and, when entering the shadow simplex routine, the point* x_0 *is a shadow vertex and the edge defined by* B_0 *is a shadow edge on* (Unit LP')*. Moreover,* x_0 *is the only degenerate vertex.*

Proof. Without loss of generality, we assume $R = I_{n\times n}$. With probability 1, the coefficients $\lambda_1, \ldots, \lambda_n$ exist and are uniquely defined.

We now show that x_0 is well-defined. Let x_0^+ be the solution to the following system of n equalities

$$v_1^{+\mathsf{T}} x_0^+ = 1, v_2^{+\mathsf{T}} x_0^+ = 1, \ldots, v_{n-1}^{+\mathsf{T}} x_0^+ = 1, \qquad 8e^{n\mathsf{T}} x_0^+ = 2. \tag{2.42}$$

This system of equations almost surely has a single solution. We claim that $Vx_0^+ = \vec{1}$. By writing $v_i^- = 8e^n - v_i^+$, we find that $v_i^{-\mathsf{T}} x_0^+ = 1$ for all $i \in [n-1]$. Therefore, $x_0 = x_0^+$ is indeed well-defined.

Algorithm 2 Symmetric Random Vertex algorithm

Input: $c \in \mathbb{R}^n \setminus \{\vec{0}\}$, $A \in \mathbb{R}^{m \times n}$, A is standard deviation $\sigma \leq \bar{\sigma}$ Gaussian with rows having centers of norm at most 1.

Output: Decide whether (Unit LP) $\max c^\mathsf{T} x, Ax \leq \vec{1}$ is unbounded or return an optimal solution.

1: If some row of A has norm greater than 2, solve $\max c^\mathsf{T} x$, st. $Ax \leq \vec{1}$ using any simplex method that takes at most $\binom{m}{n}$ pivot steps.
2: **loop**
3: Let $l = 1/6\sqrt{\log n}$.
4: Sample a rotation matrix $R \in O(n)$ uniformly at random.
5: Sample $g_1, \ldots, g_{n-1} \sim N(\vec{0}, \sigma^2 I)$ independently.
6: Set $v_i^+ = R(4e^n + le^i + g_i)$, $v_i^- = R(4e^n - le^i - g_i)$ for all $i \in [n-1]$.
7: Put $V = (v_1^+, v_1^-, v_2^+, \ldots, v_{n-1}^+, v_{n-1}^-)^\mathsf{T}$, $d = Re^n$.
8: Find x_0 such that $Vx_0 = \vec{1}$.
9: If not $Ax_0 < \vec{1}$, restart the loop.
10: Solve $\sum_{i=1}^{n-1} \lambda_i R(le^i + g_i) = c + \lambda_n d$. If $\lambda_n + \sum_{i=1}^{n-1} 4|\lambda_i| \leq 0$, restart the loop. (This corresponds to x_0 being optimal for c.)
11: Follow the shadow path from d to c on

$$\begin{aligned} &\max c^\mathsf{T} x \\ &Ax \leq \vec{1} \\ &Vx \leq \vec{1}, \end{aligned} \qquad \text{(Unit LP')}$$

starting from the vertex x_0. For the first pivot, follow the edge which is tight at the constraints in $B_0 = (v_1^{\mathrm{sign}(\lambda_1)}, \ldots, v_{n-1}^{\mathrm{sign}(\lambda_{n-1})})$. All other pivot steps are as in Algorithm 1.
12: If (Unit LP') is unbounded, return "unbounded".
13: If (Unit LP') is bounded and the optimal vertex x^* satisfies $Vx^* < \vec{1}$, return x^* as the optimal solution to (Unit LP).
14: Otherwise, restart the loop.

By definition, upon entering the shadow simplex routine, x_0 satisfies $Ax < \vec{1}$, $Vx \leq \vec{1}$ and is thus a vertex.

For all $t > 0$, define x_t to be the solution to $8e^{n\mathsf{T}}x_t = 2 - t, B_0 x_t = \vec{1}$. For any $v_i^s \notin B_0$, we have $v_i^{s\mathsf{T}} x_t = 1 - t < 1$. As the x_t lie on a line and $Ax_0 < \vec{1}$, there exists some $\varepsilon > 0$ such that x_t is feasible for all $t \leq \varepsilon$. Hence the constraints in B_0 define an edge of the feasible set.

The point x_0 is tight at the inequalities $Vx \leq \vec{1}$, and $\frac{1}{8n-8}\sum_{i=1}^{n-1}(v_i^+ + v_i^-) = d$ certifies that d lies in the normal cone at x_0, so we know that x_0 is optimal for objective d and thus a shadow vertex.

Assume that x_0 is not optimal for objective c. One outgoing edge of x_0 is tight at the inequalities $v_i^{s\mathsf{T}} x \leq 1$ for all $v_i^s \in B_0$ and that edge is on the shadow path exactly if the cone spanned by B_0 intersects $\text{cone}(c, d)$ outside $\{0\}$. This intersection is exactly the ray spanned by

$$\begin{aligned}\sum_{i=1}^{n-1} |\lambda_i| v_i^{\text{sign}(\lambda_i)} &= \sum_{i=1}^{n-1} \lambda_i R(le^i + g_i) + 4|\lambda_i| d \\ &= c + \lambda_n d + \sum_{i=1}^{n-1} 4|\lambda_i| d,\end{aligned}$$

and we know that $\lambda_n + \sum_{i=1}^{n-1} 4|\lambda_i| > 0$ as otherwise we could rewrite the above equation to certify that $c \in \text{cone}(V\lambda : \lambda \in \mathbb{R}_+^n)$, which would contradict x_0 not being optimal for objective c. We conclude that $\sum_{i=1}^{n-1} |\lambda_i| v_i^{\text{sign}(\lambda_i)}$ is a non-negative linear combination of c, d and hence our description of the first shadow vertex pivot step is correct.

Lastly, we show that any vertex other than x_0 is tight at exactly n independently distributed constraint vectors. Fix any basis B such that there exists an $i \in [n-1]$ with $v_i^+, v_i^- \in B$ and which does not define the vertex x_0. Let x_B be such that $a^{\mathsf{T}} x_B = 1$ for all $a \in B$. There exists some $j \in [n-1]$ such that both $v_j^+, v_j^- \notin B$, for otherwise we would have $x_B = x_0$. We show that, almost surely, $v_j^{+\mathsf{T}} x_B > 1$ or $v_j^{-\mathsf{T}} x_B > 1$, which implies that x_B is almost surely not feasible. We know that $v_i^{+\mathsf{T}} x_B = v_i^{-\mathsf{T}} x_B = 1$, and hence $4d^{\mathsf{T}} x_B = 1$. It follows that $v_j^{+\mathsf{T}} x_B = 2 - v_j^{-\mathsf{T}} x_B$, The only way to have both $v_j^{+\mathsf{T}} x_B \leq 1$ and $v_j^{-\mathsf{T}} x_B \leq 1$ would be if $v_j^{+\mathsf{T}} x_B = 1$. However, x_B and v_j^+ are independently distributed and v_j^+ has a continuous probability distribution, so x_B is a vertex with probability 0. □

To bound the expected running time of Algorithm 2, we bound the expected number of pivot steps per iteration of the loop, and the expected number of iterations of the loop.

First, we bound the expected shadow size in a single iteration. Because the constraint vectors v_i^+, v_i^- are not independently distributed for any $i \in [n-1]$, we are unable to apply Theorem 2.4.1 in a completely black-box way. As we show below, in this new setting, the proof of Theorem 2.4.1 still goes through essentially without modification.

In the rest of this section, we abbreviate

$$\operatorname{conv}(A, V) := \operatorname{conv}(a_1, \ldots, a_m, v_1^+, \ldots, v_{n-1}^+, v_1^-, \ldots, v_{n-1}^-).$$

Lemma 2.5.6. *Let A have independent standard deviation σ Gaussian rows with centers of norm at most* 1 *and let V be sampled, independently from A, as in lines* 4-7 *of Algorithm* 2 *with $l \leq 1$. The shadow size $\mathbb{E}[|\text{edges}(\operatorname{conv}(A, V) \cap \operatorname{span}(c, d))|]$ is bounded by $\mathcal{D}_g(n, m + 2n - 2, \min(\sigma, \bar{\sigma})/5) + 1$.*

Proof. We fix the choice of R. The distribution of constraint vectors is now independent of the two-dimensional plane by the following formula:

$$\mathbb{E}[|\text{edges}(\operatorname{conv}(A, V) \cap \operatorname{span}(c, d))|] \leq \max_R \mathbb{E}[|\text{edges}(\operatorname{conv}(A, V) \cap \operatorname{span}(c, d))| \mid R].$$

The rows of A have centers of norm at most 1 and the rows of V have centers of norm at most $4 + l \leq 5$. After an appropriate rescaling, we can assume all $n + 2d - 2$ constraints have expectations of norm at most 1 and standard deviation $\sigma \leq \bar{\sigma}/5$.

To get the desired bound, we bound the number of edges other than the one induced by x_0, $W \cap \{y \in \mathbb{R}^n : y^\mathsf{T} x_0 = 1\}$, which yields the +1 in the final bound. The proof is essentially identical to that of Theorem 2.4.1, i.e. we bound the ratio of the expected perimeter divided by the minimum expected edge of the polar polygon. We sketch the key points below. Firstly, notice that the perimeter bound in Lemma 2.4.14 does not require independence of the perturbations, so it still holds. For the minimum edge length, we restrict to the bases B as in Lemma 2.4.13 (which also does not require independence) after removing those which induce x_0 as a vertex (it has already been counted). By Lemma 2.5.5, the remaining bases in B contain at most one of each pair $\{v_i^-, v_i^+\}$, $i \in [n-1]$, since bases containing two such vectors correspond to an edge different from the one induced by x_0 with probability 0. In particular, every basis we need to consider consists of only independent random vectors.

From here, the only remaining detail for the bound to go through is to to check that the conclusion of Lemma 2.4.19 still holds, i.e., that the position of vectors within their containing hyperplane does not affect the probability that these vectors form a facet of the convex hull. Without loss of generality, we consider the vectors $a_1, \ldots, a_i$, $v_1^+, \ldots, v_j^+$ with $i + j = n$. Define $\theta \in \mathbb{S}^{n-1}, t \geq 0$ by $\theta^\mathsf{T} a_k = t$ for all $k \in [i]$, $\theta^\mathsf{T} v_k^+ = t$ for all $k \in [j]$. The set $\operatorname{conv}(a_1, \ldots, a_i, v_1^+, \ldots, v_j^+)$ is a facet of the convex hull of the constraint vectors when either (1) $\theta^\mathsf{T} a_k < t$ for all $k > i$, $\theta^\mathsf{T} v_k^- < t$ for all $k \in [j]$

and $\theta^\mathsf{T} v_k^\pm < t$ for all $k > j$ or (2) when $\theta^\mathsf{T} a_k > t$ for all $k > i$, $\theta^\mathsf{T} v_k^- > t$ for all $k \in [j]$ and $\theta^\mathsf{T} v_k^\pm > t$ for all $k > j$. The only one of these properties that is not independent of $a_1, \ldots, a_i, v_1^+, \ldots, v_j^+$ is whether $\theta^\mathsf{T} v_k^- < t$ or $\theta^\mathsf{T} v_k^- > t$ for $k \in [j]$, but we know that $\theta^\mathsf{T} v_k^- = 8\theta^\mathsf{T} d - \theta^\mathsf{T} v_k^+ = 8\theta^\mathsf{T} d - t$ for all $k \in [j]$, and so the value $\theta^\mathsf{T} v_k^-$ does not depend on the positions of $a_1, \ldots, a_i, v_1^+, \ldots, v_j^+$ within their containing hyperplane. Hence the proof of Theorem 2.4.1 still goes through and we conclude that the expected number of edges is bounded by $\mathcal{D}_g(n, m + 2n - 2, \min(\sigma, \bar{\sigma})/5) + 1$. □

All that is left, is to show that the success probability of each loop is lower bounded by a constant.

Definition 2.5.7. For a matrix $M \in \mathbb{R}^{n \times n}$, we define its operator norm by

$$\|M\| = \max_{x \in \mathbb{R}^n \setminus \{\vec{0}\}} \frac{\|Mx\|}{\|x\|}$$

and its maximum and minimum singular values by

$$s_{\max}(M) = \|M\|, \qquad s_{\min}(M) = \min_{x \in \mathbb{R}^n \setminus \{\vec{0}\}} \frac{\|Mx\|}{\|x\|}.$$

Using the Gaussian tailbound (2.7) together with a $1/2$-net on the sphere we immediately obtain the following tail bound for the operator norm of random Gaussian matrices.

Lemma 2.5.8. *For a random $n \times n$ matrix G with independent standard normal entries, one has*

$$\Pr[\|G\| > 2t\sqrt{n}] \leq 8^n e^{-n(t-1)^2/2}.$$

Proof. Let $N \subseteq \mathbb{S}^{n-1}$ be a $1/2$-net of minimal size, which has size at most 8^n, see e.g., [136], page 314. For each $v \in N$ we observe that $Gv \in \mathbb{R}^n$ is distributed as $N_n(\vec{0}, \sigma)$. By (2.7) and the union bound, we find that $\|Gv\| > t\sqrt{n}$ with probability at most $8^n e^{-n(t-1)^2/2}$.

Now let $w \in \mathbb{S}^{n-1}$ satisfy $\|Gw\| = \|G\|$ and pick $v \in N$ such that $\|v - w\| \leq 1/2$. We get

$$\|G\| = \|Gw\| \leq \|Gv\| + \|G\|\|w - v\| \leq t\sqrt{n} + \|G\|/2$$

with probability at least $1 - 8^n e^{-n(t-1)^2/2}$. The result follows after rearranging. □

Lemma 2.5.9. *Let $A \in \mathbb{R}^{m \times n}$ have rows of norm at most 2 and $\sigma \leq \frac{1}{6\sqrt{n}}$. For x_0 sampled as in lines 4-8 of Algorithm 2, with probability at least 0.98, the point x_0 satisfies $Ax_0 < \vec{1}$.*

Proof. Without loss of generality, we assume $R = I_{n\times n}$. We claim that, with sufficient probability, $\|x_0 - e^n/4\| < 1/4$. Together with the triangle inequality and the assumption that $\|a_i\| \le 2$ for all $i \in [n]$, this suffices to show $Ax_0 < \vec{1}$.

Elementary calculations show that $x_0 - e^n/4$ satisfies ${e^n}^\mathsf{T}(x_0 - e^n/4) = 0$ and, for every $i \in [n-1]$, $(le^i + g_i)^\mathsf{T}(x_0 - e^n/4) = -{g_i}^\mathsf{T}e^n/4$. Let G be the matrix with rows consisting of the first $n-1$ entries of each of $g_1, \ldots, g_{n-1}$, and g be the vector consisting of the n^{th} entries of $g_1, \ldots, g_{n-1}$. From the above equalites we derive

$$\begin{pmatrix} lI_{n-1} + G & g \\ 0^\mathsf{T} & 1 \end{pmatrix}(x_0 - e^n/4) = \frac{1}{4}\begin{pmatrix} -g \\ 0 \end{pmatrix}$$
$$\begin{pmatrix} lI_{n-1} + G & 0 \\ 0^\mathsf{T} & 1 \end{pmatrix}(x_0 - e^n/4) = \frac{1}{4}\begin{pmatrix} -g \\ 0 \end{pmatrix}$$
$$x_0 - e^n/4 = \frac{1}{4}\begin{pmatrix} -\left(lI_{n-1} + G\right)^{-1} g \\ 0 \end{pmatrix}.$$

Note that the matrix is almost surely invertible. We abbreviate $M = lI_{n-1} + G$ and bound $\|x_0 - e^n/4\| \le \|M^{-1}\|\|g\|/4$. Using that $\sigma \le \frac{l}{6\sqrt{n}}$, we apply (2.7) to get $\|g\| \le l/2$ with probability at least 0.99.

The operator norm of the inverse matrix satisfies $\|M^{-1}\| = \frac{1}{s_{\min}(lI+G)}$, and by the triangle inequality we derive

$$s_{\min}(lI + G) \ge s_{\min}(lI) - s_{\max}(G) = l - s_{\max}(G).$$

By Lemma 2.5.8, we have $\|G\| \le 3\sqrt{n}\sigma \le l/2$ with probability at least 0.99. Putting the pieces together, we conclude that

$$\frac{1}{4}\|M^{-1}\|\|g\| \le \frac{1}{4} \cdot \frac{1}{l - l/2} \cdot \frac{l}{2} \le 1/4.$$

We take the union bound over the two bad events and thus conclude that ${a_i}^\mathsf{T}x_0 \le \|a_i\|\|x_0\| < 1$ for all $i \in [m]$ with probability at least 0.98. □

Lastly, we need to prove that the conditionals on lines 10, 12 and 13 of Algorithm 2 succeeds with sufficient probability.

Lemma 2.5.10 (Adapted from [200])**.** *Let $l \le 1/6\sqrt{\log n}$ and $\sigma \le 1/8\sqrt{n \log n}$. For fixed A and V sampled as in lines* 4-7 *of Algorithm* 2*, let x^* be the optimal solution to* (Unit LP') *if it exists. With probability at least* 0.24*,* (Unit LP) *being unbounded implies that* (Unit LP') *is unbounded and* (Unit LP) *being bounded implies $Vx^* < \vec{1}$.*

Proof. Let x be the maximizer of (Unit LP) if it exists, or otherwise a generator for an unbounded ray in (Unit LP), and let $\omega = x/\|x\|$. We aim to prove that $V\omega < \vec{0}$ with

probability at least 0.24 over the randomness in V, which is sufficient for the lemma to hold.

We fix A, and hence ω as well. We decompose

$$v_i^{+\mathsf{T}}\omega = 4d^{\mathsf{T}}\omega + (lRe^i)^{\mathsf{T}}\omega + (Rg_i)^{\mathsf{T}}\omega, \tag{2.43}$$

for all $i \in [n-1]$ and similarly for v_i^-, and we will bound the different terms separately.

The inner product $d^{\mathsf{T}}\omega$ has probability density proportional to $\sqrt{1-t^2}^{n-3}$, as it is the one-dimensional marginal distribution over the sphere $\mathbb{S}^{n-1}$ (see e.g., [80], equation 1.26). which can differ over the interval $[-\sqrt{\frac{2}{n-1}}, \sqrt{\frac{2}{n-1}}]$ by at most a factor $1/e$. We lower bound the probability that $d^{\mathsf{T}}\omega$ is far from being positive:

$$\begin{aligned}
\Pr[d^{\mathsf{T}}\omega < -\frac{1}{4}\sqrt{\frac{2}{n-1}}] &= \frac{1}{2}\Pr[d^{\mathsf{T}}\omega < -\frac{1}{4}\sqrt{\frac{2}{n-1}} \mid d^{\mathsf{T}}\omega \le 0] \\
&\ge \frac{1}{2}\Pr[d^{\mathsf{T}}\omega < -\frac{1}{4}\sqrt{\frac{2}{n-1}} \mid d^{\mathsf{T}}\omega \in [-\sqrt{\frac{2}{n-1}}, 0]] \\
&\ge \frac{1}{2}\cdot\frac{\frac{3}{4e}}{\frac{3}{4e}+\frac{1}{4}} \\
&\ge 0.26.
\end{aligned}$$

Hence, for d a randomly chosen unit vector independent of ω, we have $4d^{\mathsf{T}}\omega < -\sqrt{\frac{2}{n-1}}$ with probability at least 0.26. Now we will give an upper bound on the second and third terms in (2.43) with sufficient probability.

By the same measure concentration argument as in the proof of (2.11) we know that $\Pr[|(e^i)^{\mathsf{T}}R^{\mathsf{T}}\omega| > t/\sqrt{n-1}] \le e^{-t^2/2}$. We apply the above statement with $t = 3\sqrt{\log n}$ and find that

$$|(le^i)^{\mathsf{T}}R^{\mathsf{T}}\omega| < tl/\sqrt{n-1} \le 1/2\sqrt{n-1}$$

with probability at least $1 - \frac{0.01}{n}$.

For the last part, fix $R = I$ without loss of generality. The inner product ${g_i}^{\mathsf{T}}\omega$ is $N(0,\sigma^2)$ distributed, hence $\Pr[|{g_i}^{\mathsf{T}}\omega| < 4\sigma\sqrt{\log n}] \ge 1 - \frac{0.01}{n}$ by standard Gaussian tail bounds. Recall that $4\sigma\sqrt{\log n} \le 1/2\sqrt{n-1}$.

Putting it all together, we take the union bound over the three terms in (2.43) and all v_i^+, v_i^- with $i \in [n-1]$ and find that $v_i^{+\mathsf{T}}\omega < 0$ and $v_i^{-\mathsf{T}}\omega < 0$ for all $i \in [n-1]$ with probability at least $0.26 - (n-1)\frac{0.01}{n} - (n-1)\frac{0.01}{n} \ge 0.24$. □

Theorem 2.5.11. *For $\sigma \le \bar{\sigma}$, Algorithm 2 solves* (Unit LP) *in at most an expected $6 + 5\mathcal{D}_g(n, m+2n-2, \sigma/5)$ number of shadow vertex pivots.*

Proof. Let $a_1, \ldots, a_m \in \mathbb{R}^n$ denote the rows of A, where we recall that the centers $\bar{a}_i := \mathbb{E}[a_i]$, $i \in [m]$, have norm at most 1. We let L denote the event that the rows of $a_1, \ldots, a_m$ all have norm at most 2, and L^c denotes the complement of L.

Pivots from line 1 Noting that each a_i, $i \in [m]$, is a variance σ^2 Gaussian and $1/\sigma \geq 5\sqrt{n \log m}$, by Lemma 2.2.2 (Gaussian concentration), we have that

$$\begin{aligned}\Pr[L^c] &= \Pr[\exists i \in [m] : \|a_i\| \geq 2] \leq m \Pr[\|a_1 - \bar{a}_1\| \geq 1] \\ &\leq m \Pr[\|a_1 - \bar{a}_1\| \geq 5\sqrt{n \log m}\sigma] \leq e^{-(n/2)(5\sqrt{\log m}-1)^2} \leq m^{-n} .\end{aligned}$$

Therefore, the simplex run on line 1 is executed with probability at most m^{-n} incurring at most $m^{-n}\binom{m}{n} \leq 1$ pivots on expectation.

Pivots from the main loop Let $V_1, V_2, \ldots$ be independent samples of V as described in lines 3-7 of Algorithm 2. Define the random variable $N = N(A, V_i : i \in \mathbb{N}) \geq 0$ as the number of iterations of the main loop if Algorithm 2 were run on input A, c and the value of V in iteration i equals V_i. Note that $N = 0$ exactly if L^c. Note that the value of V_i unique specifies the value of d_i. Define the event F_i that the checks on lines 9 and 10 would pass on data V_i. Lastly, let $C(A, V_i)$ denote the number of pivot steps that an iteration of the main loop would perform on the data A, V_i. In particular, $C(A, V_i) > 0$ exactly when L and F_i.

The total number of pivot steps is given by the expectation

$$\begin{aligned}\mathbb{E}[\sum_{k=1}^{N} C(A, V_k)] &= \mathbb{E}[\sum_{k=1}^{\infty} C(A, V_k)\mathbb{1}[N \geq k]] \\ &= \sum_{k=1}^{\infty} \mathbb{E}[C(A, V_k)\mathbb{1}[N \geq k]].\end{aligned}$$

For any k, the event $N \geq k$ depends solely on $V_1, \ldots, V_{k-1}$, hence we get

$$\begin{aligned}\sum_{k=1}^{\infty} \mathbb{E}[C(A, V_k)\mathbb{1}[N \geq k]] &= \sum_{k=1}^{\infty} \mathbb{E}_{A,V_k}[C(A, V_k)\mathbb{E}_{V_1,\ldots,V_{k-1}}[\mathbb{1}[N \geq k \mid A]] \\ &= \sum_{k=1}^{\infty} \mathbb{E}[C(A, V_k) \Pr[N \geq k \mid A]] \\ &= \sum_{k=1}^{\infty} \mathbb{E}[C(A, V_k) \Pr[N > 1 \mid A]^{k-1}],\end{aligned}$$

where the last line follows from the observation that the separate trials are independent when A is fixed. When A is such that L^c holds, then $\Pr[N > 1 \mid A] = 0$. Now we

appeal to Lemma 2.5.9 and Lemma 2.5.10. The first shows that the Algorithm 2 does not restart on line 9 with probability at least 0.98 and the second shows that the algorithm does not restart on lines 10 and 14 with probability at least 0.24. By the union bound, this implies that $\Pr[N > 1|A] \leq 1 - 0.22$ for any A such that L holds. Hence we get

$$\sum_{k=1}^{\infty} \mathbb{E}_{A,V_k}[C(A,V_k)\Pr[N > 1 \mid A]^{k-1}] \leq \sum_{k=1}^{\infty} \mathbb{E}[C(A,V_k)(1-0.22)^{k-1}] = \frac{1}{0.22}\mathbb{E}[C(A,V_1)].$$

The number of pivot steps $C(A,V_1)$ is nonzero exactly when L and F_1 hold, and is always bounded by the shadow size according to Theorem 2.2.11. We bound this quantity using Lemma 2.5.6 and get

$$\begin{aligned}\frac{1}{0.22}\mathbb{E}[C(A,V_1)] &\leq 5\mathbb{E}[\mathbb{1}F_1 \cap L|\text{edges}(\text{conv}(A,V_1) \cap \text{span}(c,d_1))|] \\ &\leq 5\mathbb{E}[|\text{edges}(\text{conv}(A,V_1) \cap \text{span}(c,d_1))|] \\ &\leq 5\mathcal{D}_g(n, m+2n-2, \min(\sigma,\bar{\sigma})/5) + 5.\end{aligned}$$

Final Bound Combining the results from the above paragraphs, we get that the total expected number of simplex pivots in Algorithm 2 is bounded by:

$$\Pr[L^c]\binom{m}{n} + \mathbb{E}[\sum_{k=1}^{N} C(A,V_k)] \leq 6 + 5\mathcal{D}_g(n, m+2n-2, \sigma/5)\ ,$$

as needed. □

This finishes up the analysis of the symmetric RV algorithm.

Theorem 2.5.12. (Smoothed LP) *can be solved by a two-phase shadow simplex method using an expected number of pivots of* $O(n^2\sqrt{\log m}\,\sigma^{-2} + n^3\log(m)^{1.5})$.

Proof. Combining Lemma 2.5.1 and Theorem 2.5.11, the expected number of simplex pivots is bounded by

$$10 + \mathcal{D}_g(n+1, m, \sigma/2) + 5\mathcal{D}_g(n, m+2n-2, \min\{\sigma,\bar{\sigma}\}/5)\ ,$$

where $\bar{\sigma}$ is as defined in (2.41). Noting that $1/\bar{\sigma} = O(\sqrt{n\log m})$, by the smoothed Gaussian shadow bound (Theorem 2.4.1), the above is bounded by

$$O(\mathcal{D}_g(n,m,\sigma) + \mathcal{D}_g(n,m,(\sqrt{n\log m})^{-1})) = O(n^2\sqrt{\log m}\sigma^{-2} + n^3\log(m)^{1.5})\ ,$$

as needed. □

Chapter 3

Asymptotic Bounds on the Combinatorial Diameter of Random Polytopes

The combinatorial diameter $\mathrm{diam}(P)$ of a polytope P is the maximum shortest path distance between any pair of vertices. In this chapter, we provide upper and lower bounds on the combinatorial diameter of a random "spherical" polytope, which is tight to within one factor of dimension when the number of inequalities is large compared to the dimension. More precisely, for an n-dimensional polytope P defined by the intersection of m i.i.d. half-spaces whose normals are chosen uniformly from the sphere, we show that $\mathrm{diam}(P)$ is $\Omega(nm^{\frac{1}{n-1}})$ and $O(n^2m^{\frac{1}{n-1}} + n^5 4^n)$ with high probability when $m \geq 2^{\Omega(n)}$.

For the upper bound, we first prove that the number of vertices in any fixed two-dimensional projection sharply concentrates around its expectation when m is large, where we rely on the $\Theta(n^2m^{\frac{1}{n-1}})$ bound on the expectation due to Borgwardt [29]. To obtain the diameter upper bound, we stitch these "shadows paths" together over a suitable net using worst-case diameter bounds to connect vertices to the nearest shadow. For the lower bound, we first reduce to lower bounding the diameter of the dual polytope P°, corresponding to a random convex hull, by showing the relation $\mathrm{diam}(P) \geq (n-1)(\mathrm{diam}(P^\circ) - 2)$. We then prove that the shortest path between any "nearly" antipodal pair vertices of P° has length $\Omega(m^{\frac{1}{n-1}})$.

3.1 Introduction

When does a polyhedron have small (combinatorial) diameter? This question has fascinated mathematicians, operation researchers and computer scientists for more than half a century. In a letter to Dantzig in 1957, motivated by the study of the simplex method for linear programming, Hirsch conjectured that any n-dimensional polytope with m facets has diameter at most $m - n$. While recently disproved by Santos [167] (for unbounded polyhedra, counter-examples were already given by

This chapter is based on [25], a joint work with Gilles Bonnet, Daniel Dadush, Uri Grupel and Galyna Livshyts.

Klee and Walkup [126]), the question of whether the diameter is bounded from above by a polynomial in n and m, known as the *polynomial Hirsch conjecture*, remains wide open. In fact, the current counter-examples violate the conjectured $m - n$ bound by at most 25 percent.

The best known general upper bounds on the combinatorial diameter of polyhedra are the $2^{n-3}m$ bound by Barnette and Larman [14, 15, 130], which is exponential in n and linear in m, and the *quasi-polynomial* $m^{\log_2 n+1}$ bound by Kalai and Kleitman [118]. The Kalai-Kleitman bound was recently improved to $(m-n)^{\log_2 n}$ by Todd [186] and $(m-n)^{\log_2 O(n/\log n)}$ by Sukegawa [181]. Similar diameter bounds have been established for graphs induced by certain classes of simplicial complexes, which vastly generalize 1-skeleta of polyhedra. In particular, Eisenbrand et al. [77] proved both Barnette-Larman and Kalai-Kleitman bounds for so-called connected-layer families (see Theorem 3.3.15), and Labbé et al. [128] extended the Barnette-Larman bound to pure, normal, pseudo-manifolds without boundary.

Moving beyond the worst-case bounds, one may ask for which families of polyhedra does the Hirsch conjecture hold, or more optimistically, are there families for which we can significantly beat the Hirsch conjecture? Many interesting classes induced by combinatorial optimization problems are known to satisfy the Hirsch conjecture, including the class of polytopes with vertices in $\{0,1\}^n$ [150], Leontief substitution systems [102], transportation polyhedra and their duals [11, 32, 34], as well as the fractional stable-set and perfect matching polytopes [144, 165].

Relatedly, there has been progress on obtaining diameter bounds for classes of "well-conditioned" polyhedra. If P is a polytope defined by an integral constraint matrix $A \in \mathbb{Z}^{m\times n}$ with all square submatrices having determinant of absolute value at most Δ, then diameter bounds polynomial in m, n and Δ have been obtained [23, 47, 75, 151]. The best current bound is $O(n^3\Delta^2 \log(\Delta))$, due to Dadush and Hähnle [47]. Extending on the result of Naddef [150], strong diameter bounds have been proved for polytopes with vertices in $\{0, 1, \ldots, k\}^n$ [64, 69, 127]. In particular, [127] proved that the diameter is at most nk, which was improved to $nk - \lceil n/2 \rceil$ for $k \geq 2$ [64] and to $nk - \lceil 2n/3 \rceil - (k-2)$ for $k \geq 4$ [69].

3.1.1 Diameter of Random Polytopes

With a view of beating the Hirsch bound, the main focus of this chapter will be to analyze the diameter of random polytopes, which one may think of as well-conditioned on "average". Coming both from the average case and smoothed analysis literature that was mentioned in Chapter 2, there is tantilizing evidence that important classes of random polytopes may have very small diameters.

In the average-case context, Borgwardt [28, 29] proved that for $P(A) := \{x \in \mathbb{R}^n : Ax \leq \vec{1}\}$, $A \in \mathbb{R}^{m\times n}$ where the rows of A are drawn from any rotational symmetric

distribution (RSD), that the expected number of edges in any fixed two-dimensional projection of $P(A)$ – the so-called *shadow bound* – is $O(n^2 m^{\frac{1}{n-1}})$. Borgwardt also showed that this bound is tight up to constant factors when the rows of A are drawn uniformly from the sphere, that is, the expected shadow size is $\Theta(n^2 m^{\frac{1}{n-1}})$. In the smoothed analysis context, A has the form $\bar{A} + \sigma G$, where $\bar{A}$ is a fixed matrix with rows of ℓ_2 norm at most 1 and G has i.i.d. standard normally distributed entries and $\sigma > 0$. Bounds on the expected size of the shadow in this context were first studied by Spielman and Teng [179], later improved by [50, 200], where the best current bound is $O(n^2\sqrt{\log m}/\sigma^2)$ when $\sigma \leq \frac{1}{\sqrt{n \log m}}$, as seen in Chapter 2.

From the perspective of short paths, these results imply that if one samples objectives v, w uniformly from the sphere, then there is a path between the vertices maximizing v and w in $P(A)$ of expected length $O(n^2 m^{\frac{1}{n-1}})$ in the RSD model, and expected length $O(n^2\sqrt{\log m}/\sigma^2)$ in the smoothed model. That is, "most pairs" of vertices (with respect to the distribution in the last sentence), are linked by short expected length paths. Note that both of these bounds scale either sublinearly or logarithmically in m, which is far better than $m - n$. While these bounds provide evidence, they do not directly upper bound the diameter, since this would need to work for all pairs of vertices rather than most pairs.

A natural question is thus whether the shadow bound is close to the true diameter. In this chapter, we show that this is indeed the case, in the setting where the rows of A are drawn uniformly from the sphere and when m is (exponentially) large compared to n. More formally, our main result is as follows:

Theorem 3.1.1. *Suppose that $n, m \in \mathbb{N}$ satisfy $n \geq 2$ and $m \geq 2^{\Omega(n)}$. Let $A^\mathsf{T} := (a_1, \ldots, a_M) \in \mathbb{R}^{n \times M}$, where M is Poisson distributed with $\mathbb{E}[M] = m$, and $a_1, \ldots, a_M$ are sampled independently and uniformly from $\mathbb{S}^{n-1}$. Then, letting $P(A) := \{x \in \mathbb{R}^n : Ax \leq \vec{1}\}$, with probability at least $1 - m^{-n}$, we have that*

$$\Omega(nm^{\frac{1}{n-1}}) \leq \operatorname{diam}(P(A)) \leq O(n^2 m^{\frac{1}{n-1}} + n^5 4^n).$$

In the above, we note that the number of constraints M is chosen according to a Poisson distribution with expectation m. This is only for technical convenience (it ensures useful independence properties, see Proposition 3.2.5), and with small modifications, our arguments also work in the case where $M := m$ deterministically. Also, since the constraints are chosen from the sphere, M is almost surely equal to the number of facets of $P(A)$ above (i.e., there are no redundant inequalities).

From the bounds, we see that $\operatorname{diam}(P(A)) \leq O(n^2 m^{\frac{1}{n-1}})$ with high probability as long as $m \geq 2^{\Omega(n^2)}$. This shows that the shadow bound is indeed close to an upper bound for the expected diameter when m is sufficiently large. Furthermore, the shadow bound is tight to within one factor of dimension in this regime.

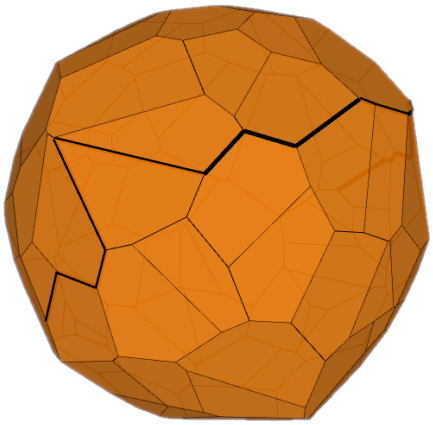


Figure 3.1: A shortest path between a diameter-achieving pair of vertices, on a random spherical polytope with 100 constraints.

We note that the upper bound is already non-trivial when $m \geq \Omega(n^5 4^n)$, since then $O(n^2 m^{\frac{1}{n-1}} + n^5 4^n) \leq m - n$.

While our bounds are only interesting when m is exponential, the bounds are nearly tight asymptotically, and as far as we are aware, they represent the first non-trivial improvements over worst-case upper bounds for a natural class of polytopes defined by random halfspaces.

Our work naturally leaves two interesting open problems. The first is whether the shadow bound upper bounds the diameter when m is polynomial in n. The second is to close the factor n gap between upper and lower bound in the large m regime.

3.1.2 Prior work

Lower bounds on the diameter of $P(A)$, $A^\mathsf{T} = (a_1, \ldots, a_m) \in \mathbb{R}^{n \times m}$, were studied by Borgwardt and Huhn [31]. They examined the case where each row of A is sampled from a RSD with radial distribution

$$\Pr_a[\|a\|_2 \leq r] = \frac{\int_0^r (1-t^2)^\beta t^{n-1} dt}{\int_0^1 (1-t^2)^\beta t^{n-1} dt},$$

for $r \in [0,1]$, $\beta \in (-1, \infty)$. Restricting their results to the case $\beta \to -1$, corresponding to the uniform distribution on the sphere (where the bound is easier to state), they show that

$$\mathbb{E}[\mathrm{diam}(P(A))] \geq \Omega(m^{\frac{1}{n} + \frac{1}{n(n-1)^2}}).$$

We improve their lower bound to $\Omega(nm^{1/(n-1)})$ when $m \geq 2^{\Omega(n)}$, noting that $m^{1/(n-1)} = O(1)$ for $m = 2^{O(n)}$.

In terms of upper bounds, the diameter of a *random convex hull of points*, instead of a random intersection of halfspaces, has been implicitly studied. Given a matrix $A^\mathsf{T} = (a_1, \ldots, a_m) \in \mathbb{R}^{n \times m}$, let us define

$$Q(A) := \mathrm{conv}(\{a_1, \ldots, a_m\}) \tag{3.1}$$

to be the convex hull of the rows of A. When the rows of A are sampled uniformly from the unit ball $\mathbb{B}_2^n$, the question of when the diameter of $Q(A)$ is exactly 1 (i.e., every pair of distinct vertices is connected by an edge) was studied by Bárány and Füredi [13]. They proved that with probability approaching 1, $\mathrm{diam}(Q(A)) = 1$ if $m \leq 1.125^n$ and $\mathrm{diam}(Q(A)) > 1$ if $m \geq 1.4^n$.

In dimension 3, letting $a_1, \ldots, a_M \in \mathbb{S}^2$ be chosen independently and uniformly from the 2-sphere, where M is Poisson distributed with $\mathbb{E}[M] = m$, Glisse, Lazard, Michel and Pouget [89] proved that with high probability the maximum number of edges in any two-dimensional projection of $Q(A)$ is $\Theta(\sqrt{m})$. This in particular proves that the combinatorial diameter is at most $O(\sqrt{m})$ with high probability.

It is important to note that the geometry of $P(A)$ and $Q(A)$ are strongly related. Indeed, as long as $m = \Omega(n)$ and the rows of A are drawn from a symmetric distribution, $P(A)$ and $Q(A)$ are polars of each other. To be precise, it always holds that

$$Q(A)^\circ := \{x \in \mathbb{R}^n : \langle x, y \rangle \leq 1, \forall y \in Q(A)\} = P(A)$$

and if $\vec{0} \in Q(A)$ then also

$$P(A)^\circ := \{x \in \mathbb{R}^n : \langle x, y \rangle \leq 1, \forall y \in P(A)\} = Q(A)$$

As a direct consequence of Wendel's theorem [169, Theorem 8.2.1], $\vec{0} \in Q(A)$ happens with probability approaching 1 when $m \geq cn$ for any fixed $c > 2$. In general $P(A)^\circ = \mathrm{conv}(A \cup \{\vec{0}\})$ holds.

As we will see, our proof of Theorem 3.1.1 will in fact imply similarly tight diameter bounds for $\mathrm{diam}(Q(A))$ as for $\mathrm{diam}(P(A))$, yielding analogues and generalizations of the above results, when $A^\mathsf{T} = (a_1, \ldots, a_M) \in \mathbb{R}^{n \times M}$ and M is Poisson with $\mathbb{E}[M] = m$. More precisely, we will show that for $m \geq 2^{\Omega(n)}$, with high probability

$$\Omega(m^{\frac{1}{n-1}}) \leq \mathrm{diam}(Q(A)) \leq O(nm^{\frac{1}{n-1}} + n^5 4^n).$$

In essence, for m large enough, our bounds for $\mathrm{diam}(Q(A))$ are a factor $\Theta(n)$ smaller than our bounds for $\mathrm{diam}(P(A))$. This relation will be explained in Section 3.4.

3.1.3 Proof Overview

In this section, we give the high level overview of our approach for both the upper and lower bound in Theorem 3.1.1.

The Upper Bound

In this overview, we will say that an event holds with high probability if it holds with probability $1 - m^{-\Omega(n)}$. To prove the upper bound on the diameter of $P(A)$, we proceed as follows. For simplicity, we will only describe the high level strategy for achieving a $O(n^2 m^{\frac{1}{n-1}} + 2^{O(n)})$ bound. To begin, we first show that the vertices of $P(A)$ maximizing objectives in a suitable net N of the sphere $\mathbb{S}^{n-1}$, are all connected to the vertex maximizing e^1, with a path of length $O(n^2 m^{\frac{1}{n-1}} + 2^{O(n)})$ with high probability. Second, we will show that with high probability, for all $v \in \mathbb{S}^{n-1}$, there is a path between the vertex of $P(A)$ maximizing v and the corresponding maximizer of closest objective $v' \in N$ of length at most $2^{O(n)} \log m$. Since every vertex of $P(A)$ maximizes some objective in $\mathbb{S}^{n-1}$, by stitching at most 4 paths together, we get that the diameter of $P(A)$ is at most $O(n^2 m^{\frac{1}{n-1}} + 2^{O(n)} \log m) = O(n^2 m^{\frac{1}{n-1}} + 2^{O(n)})$ with high probability.

We only explain the strategy for the first part, as the second part follows easily from the same techniques. The key estimate here is the sharp $\Theta(n^2 m^{\frac{1}{n-1}})$ bound on the expected number of vertices in a fixed two-dimensional projection due to Borgwardt [28, 29], the so-called *shadow bound*, which allows one to bound the expected length of paths between vertices maximizing any two fixed objectives (see Section 3.3 for a more detailed discussion). We first strengthen this result by proving that the size of the shadow sharply concentrates around its expectation when m is large (Theorem 3.3.4), allowing us to apply a union bound on a suitable net of shadows, each corresponding to a two-dimensional plane spanned by e^1 and some element of N above. To obtain such concentration, we show that the shadow decomposes into a sum of nearly independent "local shadows", corresponding to the vertices maximizing a small slice of the objectives in the plane, allowing us to apply concentration results on sums of nearly independent random variables.

Independence via Density We now explain the local independence structure in more detail. For this purpose, we examine the smallest $\varepsilon > 0$ such that rows of A are ε-dense on $\mathbb{S}^{n-1}$, that is, such that every point in $\mathbb{S}^{n-1}$ is at distance at most ε from some row of A. Using standard estimates on the measure of spherical caps and the union bound, one can show with high probability that $\varepsilon := \Theta((\log m/m)^{1/m})$ and that any spherical cap of radius $t\varepsilon$ contains at most $O(t^{n-1} \log m)$ rows of A for any fixed $t \geq 1$ (see Lemma 3.2.3 and Corollary 3.2.7).

We derive local independence from the fact that the vertex v of $P(A)$ maximizing a unit norm objective w is defined by constraints $a \in A$ which are distance at most 2ε from w (see Lemma 3.3.10 for a more general statement). This locality implies that the number of vertices in a projection of $P(A)$ onto a two-dimensional subspace $W \ni w$ maximizing objectives at distance ε from w (i.e., the slice of objectives) depends only on the constraints in A at distance at most $O(\varepsilon)$ from w. In particular, the number of relevant constraints for all objectives at distance ε from w is at most $2^{O(n)} \log m$ by the estimate in the last paragraph. By the independence properties of Poisson processes (see Proposition 3.2.5), one can in fact conclude that this local part of the shadow on W is independent of the constraints in A at distance more than $O(\varepsilon)$ from w.

Given the above, we decompose the shadow onto W into $k = O(1/\varepsilon)$ pieces, by placing k equally spaced objectives $w_0, \ldots, w_{k-1}, w_k = w_0$ on $\mathbb{S}^{n-1} \cap W$, so that $\|w_i - w_{i+1}\|_2 \leq \varepsilon$, $0 \leq i \leq k-1$, and defining $K_i \geq 0$, $0 \leq i \leq k-1$, to be the number of vertices maximizing objectives in $[w_i, w_{i+1}]$. This subdivision partitions the set of shadow vertices, so Borgwardt's bound applies to the expected sum: $\mathbb{E}[\sum_{i=0}^{k-1} K_i] = O(n^2 m^{1/(n-1)})$. Furthermore, as argued above, each K_i is (essentially) independent of all K_j's with $|i - j \mod k| = \Omega(1)$. This allows us to apply a Bernstein-type concentration bound for sums of nearly-independent bounded random variables to $\sum_{i=0}^{k-1} K_i$ (see Lemma 3.2.8).

Unfortunately, the worst-case upper bounds we have for each K_i, $0 \leq i \leq k-1$, are rather weak. Namely, we only know that in the worst-case, K_i is bounded by the total number of vertices induced by constraints relevant to the interval $[w_i, w_{i+1}]$, where $\|w_i - w_{i+1}\| \leq \varepsilon$. As mentioned above, the number of relevant constraints is $2^{O(n)} \log m$ and hence the number of vertices is at most $(2^{O(n)} \log m)^n$. With these estimates, we can show high probability concentration of the shadow size around its mean when $m \geq 2^{\Omega(n^3)}$. One important technical aspect ignored above is that both the independence properties and the worst-case upper bounds on each K_i crucially relies only on conditioning A to be "locally" ε-dense around $[w_i, w_{i+1}]$ (see Definition 3.3.11 and Lemma 3.3.14 for more details).

Abstract Diameter Bounds to the Rescue To allow tight concentration of the diameter to occur for $m = 2^{\Omega(n^2)}$, we adapt the above strategy by successively following shortest paths instead of the shadow path on W. More precisely, between the maximizer v_i of w_i and v_{i+1} of w_{i+1}, $0 \leq i \leq k-1$, we follow the shortest path from v_i to v_{i+1} in the subgraph induced by the vertices v of $P(A)$ satisfying $\langle v, w_{i+1} \rangle \geq \langle v_i, w_{i+1} \rangle$. We now let K_i, $0 \leq i \leq k-1$, denote the length of the corresponding shortest path. For such local paths, one can apply the abstract Barnette–Larman style bound of [77] to obtain much better worst-case bounds. Namely, we can

show $K_i \leq 2^{O(n)} \log m$, $0 \leq i \leq k-1$, instead of $(2^{O(n)} \log m)^n$ (see Lemma 3.3.16). Crucially, the exact same independence and locality properties hold for these paths as for the shadow paths, due to the generality of our main locality lemma (Lemma 3.3.10). Furthermore, as these paths are only shorter than the corresponding shadow paths, their expected sum is again upper bounded by Borgwardt's bound. With the improved worst-case bounds, our concentration estimates are sufficient to show that all paths indexed by planes in the net N have length $O(n^2 m^{\frac{1}{n-1}} + 2^{O(n)})$ with high probability.

The Lower Bound

For the lower bound, we first reduce to lower bounding the diameter of the polar polytope $P(A)^\circ = Q(A)$, where we show that $\operatorname{diam}(P(A)) \geq (n-1)(\operatorname{diam}(Q(A))-2)$ (see Lemma 3.4.1). This relation holds as long as $P(A)$ is a simple polytope containing the origin in its interior (which holds with probability $1 - 2^{-\Omega(m)}$). To prove it, we show that given any path between vertices v_1, v_2 of $P(A)$ of length D, respectively incident to distinct facets F_1, F_2 of $P(A)$, one can extract a facet path, where adjacent facets share an $(n-2)$-dimensional intersection (i.e., a ridge), of length at most $D/(n-1)+2$. Such facet paths exactly correspond to paths between vertices in $Q(A)$, yielding the desired lower bound.

For $m \geq 2^{\Omega(n)}$, proving that $\operatorname{diam}(P(A)) \geq \Omega(nm^{1/(n-1)})$ reduces to showing that $\operatorname{diam}(Q(A)) \geq m^{1/(n-1)}$ with high probability. For the $Q(A)$ lower bound, we examine the length of paths between vertices of $Q(A)$ maximizing antipodal objectives, e.g., $-e^1$ and e^1. From here, one can easily derive an $\Omega((m/\log m)^{\frac{1}{n-1}})$ lower bound on the length of such a path, by showing that every edge of $Q(A)$ has length $\varepsilon := \Theta((\log m/m)^{\frac{1}{n-1}})$ and that the vertices in consideration are at distance $\Omega(1)$. This is a straightforward consequence of $Q(A)$ being tightly sandwiched by Euclidean balls, namely $(1-\varepsilon^2/2)\mathbb{B}_2^n \subseteq Q(A) \subseteq \mathbb{B}_2^n$ (Lemma 3.3.6) with high probability. This sandwiching property is itself a consequence of the rows of A being ε-dense on $\mathbb{S}^{n-1}$, as mentioned in the previous section.

Removing the extraneous logarithmic factor (which makes the multiplicative gap between our lower and upper bound go to infinity as $m \to \infty$), requires a much more involved argument as we cannot rely on a worst-case upper bound on the length of edges. Instead, we first associate any antipodal path above to a continuous curve on the sphere from $-e^1$ to e^1 (Lemma 3.4.6), corresponding to objectives maximized by vertices along the path. From here, we decompose any such curve into $\Omega(m^{\frac{1}{n-1}})$ segments whose endpoints are at distance $\Theta(m^{-1/(n-1)})$ on the sphere. Finally, by appropriately bucketing the breakpoints (Lemma 3.4.7) and applying a careful union bound, we show that for any such curve, an $\Omega(1)$ fraction of the segments induce at least 1 edge on the corresponding path with overwhelming probability

(Theorem 3.4.2). For further details on the lower bound, including how we discretize the set of curves, we refer the reader to Section 3.4.

3.1.4 Organization

In Section 3.2, we introduce some basic notation as well as background materials on Poisson processes, the measure of spherical caps, and concentration inequalities for independent random variables. In Section 3.3, we prove the upper bound. Halfway into that section, we also prove Theorem 3.3.4, a tail bound on the shadow size that is of independent interest. We prove the lower bound in Section 3.4.

3.2 Preliminaries

For notational simplicity in the remainder of this chapter, it will be convenient to treat A as a subset of $\mathbb{S}^{n-1}$ instead of a matrix. For $A \subseteq \mathbb{S}^{n-1}$, we will slightly abuse notation and let $P(A) := \{x \in \mathbb{R}^n : \langle x, a\rangle \leq 1, \forall a \in A\}$ and $Q(A) := \mathrm{conv}(A)$.

3.2.1 Cap Volumes

For a subset $C \subseteq \mathbb{S}^{n-1}$, we write $\sigma(C) := \sigma_{n-1}(C)$ to denote the measure of C with respect to the uniform measure on $\mathbb{S}^{n-1}$. In particular, $\sigma(\mathbb{S}^{n-1}) = 1$.

Definition 3.2.1. For $w \in \mathbb{S}^{n-1}$ and $r \geq 0$, we denote the spherical cap of radius r centered at w by $C(w, r) = \{x \in \mathbb{S}^{n-1} : \|w - x\| \leq r\}$.

We say $A \subseteq \mathbb{S}^{n-1}$ is ε-dense in the sphere for $\varepsilon > 0$ if for every $w \in \mathbb{S}^{n-1}$ there exists $a \in A$ such that $a \in C(w, \varepsilon)$.

We will need relatively tight estimates on the measure of spherical caps. The following lemma gives useful upper and lower bounds on the ratio of cap volumes.

Lemma 3.2.2. *For any $s, \varepsilon > 0$ and $v \in \mathbb{S}^{n-1}$ we have*

$$\frac{\sigma(C(v, (1+s)\varepsilon))}{(1+s)^{n-1}} \leq \sigma(C(v, \varepsilon)) \leq \frac{\sigma(C(v, (1-s)\varepsilon))}{(1-s)^{n-1}},$$

assuming for the first inequality that $(1+s)\varepsilon \leq 2$ and for the second that $s < 1$ and $\varepsilon \leq 2$.

Proof. First we write the area of the cap as the following integral, for any $r \in [0, 2]$

$$\sigma(C(v, r)) = c_{n-1} \int_0^{r^2/2} \sqrt{2t - t^2}^{n-3} \, \mathrm{d}t,$$

where $c_{n-1} := \mathrm{vol}_{n-2}(\mathbb{S}^{n-2})/\mathrm{vol}_{n-1}(\mathbb{S}^{n-1})$. Note that $\sqrt{2t-t^2}$ is the radius of the slice $\mathbb{S}^{n-1} \cap \{x \in \mathbb{S}^{n-1} : \langle x, v\rangle = 1-t\} = (1-t)v + \sqrt{2t-t^2}(S^{n-1} \cap v^{\perp})$. The scaling of the volume of the central slice by $\sqrt{2t-t^2}^{n-3}$ instead of $\sqrt{2t-t^2}^{n-2}$ is to account for the curvature of the sphere. With this integral in our toolbox, we can prove our desired inequalities. We start with the first one, assuming that $(1+s)^2r^2/2 \leq 2$ so that we only take square roots of positive numbers.

$$\begin{aligned}
\sigma(C(v,(1+s)\varepsilon)) &= c_{n-1}\int_0^{(1+s)^2r^2/2} \sqrt{2t-t^2}^{n-3}\,\mathrm{dt} \\
&= c_{n-1}(1+s)^2\int_0^{r^2/2} \sqrt{2(1+s)^2u-(1+s)^4u^2}^{n-3}\,\mathrm{du} \\
&\leq c_{n-1}(1+s)^2\int_0^{r^2/2} \sqrt{2(1+s)^2u-(1+s)^2u^2}^{n-3}\,\mathrm{du} \\
&= (1+s)^{n-1}c_{n-1}\int_0^{r^2/2} \sqrt{2u-u^2}^{n-3}\,\mathrm{du} \\
&= (1+s)^{n-1}\sigma(C(v,\varepsilon)).
\end{aligned}$$

The second inequality is proven in a similar fashion, assuming that $1-s>0$:

$$\begin{aligned}
\sigma(C(v,(1-s)\varepsilon)) &= c_{n-1}\int_0^{(1-s)^2r^2/2} \sqrt{2t-t^2}^{n-3}\,\mathrm{dt} \\
&= c_{n-1}(1-s)^2\int_0^{r^2/2} \sqrt{2(1-s)^2t-(1-s)^4t^2}^{n-3}\,\mathrm{dt} \\
&\geq c_{n-1}(1-s)^2\int_0^{r^2/2} \sqrt{2(1-s)^2t-(1-s)^2t^2}^{n-3}\,\mathrm{dt} \\
&= (1-s)^{n-1}c_{n-1}\int_0^{r^2/2} \sqrt{2t-t^2}^{n-3}\,\mathrm{dt} \\
&= (1-s)^{n-1}\sigma(C(v,\varepsilon)).
\end{aligned}$$

□

We now give absolute estimates on cap volume measure that can be found in [33]. We note that [33] parametrize spherical caps with respect to the distance of their defining halfspace to the origin. The following lemma is derived using the fact that the cap $C(v,\varepsilon)$, $\varepsilon \in [0,\sqrt{2}]$, $v \in \mathbb{S}^{n-1}$, is induced by intersecting $\mathbb{S}^{n-1}$ with the halfspace $\langle v, x\rangle \geq 1-\varepsilon^2/2$, whose distance to the origin is exactly $1-\varepsilon^2/2$.

Lemma 3.2.3. [33, *Lemma* 2.1] *For $n \geq 2$, $\varepsilon \in [0,\sqrt{2}]$, $v \in \mathbb{S}^{n-1}$, the following estimates holds*:

- *If $\varepsilon \in [\sqrt{2(1-\frac{2}{\sqrt{n}})}, \sqrt{2}]$, then $\sigma(C(v,\varepsilon)) \in [1/12, 1/2]$.*

- *If* $\varepsilon \in [0, \sqrt{2(1-\frac{2}{\sqrt{n}})}]$, *then*

$$\frac{1}{6(1-\varepsilon^2/2)\sqrt{n}}(\varepsilon\sqrt{1-\varepsilon^2/4})^{n-1} \leq \sigma(C(v,\varepsilon)) \leq \frac{1}{2(1-\varepsilon^2/2)\sqrt{n}}(\varepsilon\sqrt{1-\varepsilon^2/4})^{n-1}.$$

3.2.2 Poisson Processes

The Poisson distribution $\mathrm{Pois}(\lambda)$ with parameter $\lambda \geq 0$ has probability mass function $f(x,\lambda) := e^{-\lambda}\frac{\lambda^x}{x!}$, $x \in \mathbb{Z}_+$. We note that $\mathrm{Pois}(0)$ is the random variable taking value 0 with probability 1. Recall that $\mathbb{E}[\mathrm{Pois}(\lambda)] = \lambda$. We will rely on the following standard tail-estimate (see [38, Theorem 1]):

Lemma 3.2.4. *Let* $X \sim \mathrm{Pois}(\lambda)$. *Then for* $x \geq 0$, *we have that*

$$\max\{\Pr[X \geq \lambda + x], \Pr[X \leq \lambda - x]\} \leq e^{-\frac{x^2}{2(\lambda+x)}}. \tag{3.2}$$

We define a random subset A to be distributed as $\mathrm{Pois}(\mathbb{S}^{n-1}, \lambda)$, $\lambda \geq 0$, if $A = \{a_1, \ldots, a_M\}$, where $|A| = M \sim \mathrm{Pois}(\lambda)$ and $a_1, \ldots, a_M$ are uniformly and independently distributed on $\mathbb{S}^{n-1}$. Note that $\mathbb{E}[|A|] = \lambda$. In standard terminology, A is called a homogeneous Poisson point process on $\mathbb{S}^{n-1}$ with intensity $\lambda > 0$.

A basic fact about such a Poisson process is that the number of samples landing in disjoint subsets are independent Poisson random variables. This property is known as "complete randomness", see, e.g., [56].

Proposition 3.2.5. *Let* $A \sim \mathrm{Pois}(\mathbb{S}^{n-1}, \lambda)$. *Let* $C_1, \ldots, C_k \subseteq \mathbb{S}^{n-1}$ *be pairwise disjoint measurable sets. Then, the random variables* $|A \cap C_i|$, $i \in [k]$, *are independent and* $|A \cap C_i| \sim \mathrm{Pois}(\lambda\sigma(C_i))$, $i \in [k]$.

3.2.3 Density Estimates

In this section, we give bounds on the fineness of the net induced by a Poisson distributed subset of $\mathbb{S}^{n-1}$. Roughly speaking, if A is $\mathrm{Pois}(\mathbb{S}^{n-1}, m)$ distributed then A will be $\Theta((\log m/m)^{1/(n-1)})$-dense, see Definition 3.2.1. While this estimate is standard in the stochastic geometry, it is not so easy to find a reference giving quantitative probabilistic bounds, as more attention has been given to establishing exact asymptotics as $m \to \infty$ (see [163]). We provide a simple proof of this fact here, together with the probabilistic estimates that we will need.

Lemma 3.2.6. *For* $m \geq n \geq 2$ *and* $0 < p < m^{-n}$, *have* $\varepsilon = \varepsilon(m,n,p) > 0$ *satisfy* $\sigma(C(v,\varepsilon)) = 3e\log(1/p)/m < 1/12$. *Then, for* $A \sim \mathrm{Pois}(\mathbb{S}^{n-1}, m)$,

$$\Pr[\exists v \in \mathbb{S}^{n-1} : \ C(v,\varepsilon) \cap A = \emptyset] \leq p$$

and for every $t \geq 1$,

$$\Pr[\exists v \in \mathbb{S}^{n-1} : \ |C(v, t\varepsilon) \cap A| \geq 45 \log(1/p) t^{n-1}] \leq p.$$

Proof. Let $N \subseteq \mathbb{S}^{n-1}$ denote the centers of a maximal packing of spherical caps of radius $\varepsilon/(2n)$. By maximality, N is ε/n-dense, i.e., an ε/n net. Comparing volumes, by Lemma 3.2.2, we see that

$$1 \geq |N|\sigma(C(v, \varepsilon/(2n)) \geq |N|(2n)^{-(n-1)}\sigma(C(v, \varepsilon)),$$

so $|N| \leq (2n)^{n-1}/\sigma(C(v, \varepsilon)) \leq (2n)^{n-1}m$. By way of a net argument, using that $|C(v, (1 - 1/n)\varepsilon) \cap A| \sim \text{Pois}(m\sigma(C(v, (1 - 1/n)\varepsilon)), \forall v \in \mathbb{S}^{n-1}$, we analyze our first probability

$$\begin{aligned}
\Pr[\exists v \in \mathbb{S}^{n-1} : \ C(v, \varepsilon) \cap A = \emptyset] &\leq \Pr[\exists v \in N : \ C(v, (1 - 1/n)\varepsilon) \cap A = \emptyset] \\
&\leq |N| \max_{v \in N} \Pr[C(v, (1 - 1/n)\varepsilon) \cap A = \emptyset] \\
&\leq (2n)^{n-1} m e^{-m\sigma(C(v,(1-1/n)\varepsilon))} \\
&\leq (2n)^{n-1} m e^{-(1-1/n)^{n-1} m\sigma(C(v,\varepsilon))} \\
&\leq (2n)^{n-1} m e^{-3\log(1/p)} \leq p.
\end{aligned}$$

We now prove the second estimate. Using the cap size estimate from Lemma 3.2.2, we have $m\sigma(C(v, (1+1/n)t\varepsilon)) \leq (1+1/n)^{n-1} t^{n-1} m\sigma(C(v, \varepsilon)) \leq 3e^2 t^{n-1} \log(1/p)$. Write $\lambda := 3e^2 t^{n-1} \log(1/p)$. By a similar net argument as above, we see that

$$\begin{aligned}
\Pr[\exists v \in \mathbb{S}^{n-1} : \ |C(v, t\varepsilon) \cap A| \geq 2\lambda] &\leq |N| \max_{v \in N} \Pr[|C(v, (1 + 1/n)t\varepsilon) \cap A| \geq 2\lambda] \\
&\leq |N| \Pr_{X \sim \text{Pois}(\lambda)}[X \geq 2\lambda] \\
&\leq |N| e^{-\left(2\lambda - m\sigma(C(v,(1+1/n)t\varepsilon))\right)^2/4\lambda} \\
&\quad \text{(by the Poisson tailbound, Lemma 3.2.4)} \\
&\leq |N| e^{-\frac{\lambda}{4}} \leq (2n)^{n-1} m e^{-3\log(1/p)} \leq p.
\end{aligned}$$

The proof is complete when we observe that $2\lambda \leq 45 t^{n-1} \log(1/p)$. □

We now give effective bounds on the density estimate ε above. Note that taking the $(n-1)^{th}$ root of the bounds for ε^{n-1} below yields $\varepsilon = \Theta((\log m/m)^{1/(n-1)})$ for $m = n^{\Omega(1)}$ and $p = 1/m^{-n}$. The stated bounds follow directly from the cap measure estimates in Lemma 3.2.3.

Corollary 3.2.7. *Let $\varepsilon > 0$ be as in Lemma 3.2.6, i.e., satisfying $\sigma(C(v,\varepsilon)) = 3e\log(1/p)/m \le 1/12$. Then $\varepsilon \in [0, \sqrt{2(1-\frac{2}{\sqrt{n}})}]$,*

$$\varepsilon^{n-1} \ge 12e\log(1/p)/m$$

and

$$\left(\varepsilon/\sqrt{2}\right)^{n-1} \le \left(\varepsilon\sqrt{1-\varepsilon^2/4}\right)^{n-1} \le 18\sqrt{n}\log(1/p)/m.$$

Proof. The claim $\varepsilon \in [0, \sqrt{2(1-\frac{2}{\sqrt{n}})}]$ follows by Lemma 3.2.3 part 1 and our assumption that $\sigma(C(v,\varepsilon)) \le 1/12$. The lower bound on ε^{n-1} follows from the upper bound from Lemma 3.2.3 part 2

$$\frac{3e\log(1/p)}{m} = \sigma(C(v,\varepsilon)) \le \frac{1}{2(1-\varepsilon^2/2)\sqrt{n}}(\varepsilon\sqrt{1-\varepsilon^2/4})^{n-1} \le \frac{\varepsilon^{n-1}}{4},$$

where the last inequality follows since $\varepsilon \in [0, \sqrt{2(1-\frac{2}{\sqrt{n}})}]$. For the upper bound on ε, we rely on the corresponding estimate in Lemma 3.2.3 part 2:

$$\frac{3e\log(1/p)}{m} = \sigma(C(v,\varepsilon)) \ge \frac{(\varepsilon\sqrt{1-\varepsilon^2/4})^{n-1}}{6(1-\varepsilon^2/2)\sqrt{n}} \ge \frac{(\varepsilon\sqrt{1-\varepsilon^2/4})^{n-1}}{6\sqrt{n}} \ge \frac{(\varepsilon/\sqrt{2})^{n-1}}{6\sqrt{n}},$$

where the last inequality follows from $\varepsilon \in [0, \sqrt{2}]$. The desired inequalities now follow by rearranging. □

3.2.4 Concentration for Nearly-Independent Random Variables

We will use the following variant on Bernstein's inequality that is a direct consequence of [112, Theorem 2.3], which proves a more general result using the fractional chromatic number of the dependency graph.

Lemma 3.2.8. *Suppose that $Y_1, \dots, Y_k$ are random variables taking values in $[0, M]$ and $\mathrm{Var}(Y_i) \le \sigma^2$ for each $i \in [k]$. Assume furthermore that there exists a partition $I_1 \cup I_2 \cup \cdots \cup I_q = \{Y_1, \dots, Y_k\}$ such that the random variables in any one set I_j are mutually independent. Then for any $t \ge 0$ we get*

$$\Pr\left[\left|\sum_{i=1}^{k} Y_i - \mathbb{E}[\sum_{i=1}^{k} Y_i]\right| \ge t\right] \le 2\exp\left(\frac{-8t^2}{25q(k\sigma^2 + Mt/3)}\right)$$

When we use the above lemma, we will bound the variance of the random variables using the following inequality:

Lemma 3.2.9. *Let $Y \in [0, M]$ be a random variable and $\mathbb{E}[Y] = \mu$. Then* $\mathrm{Var}(Y) \leq \mu(M - \mu)$.

Proof. The inequality follows from $\mathrm{Var}(Y) = \mathbb{E}[Y^2] - \mu^2 \leq M\mathbb{E}[Y] - \mu^2 = \mu(M - \mu)$, where we have used that $Y^2 \leq MY$ for $Y \in [0, M]$. □

3.3 Shadow size and upper bounding the diameter

In the first part of this section, we prove a concentration result on the number of *shadow vertices* of $P(A)$. This addresses an open problem from [28]. In the second part, we use the resulting tools to prove Theorem 3.3.5, our high-probability upper bound on the diameter of $P(A)$.

We start by defining a useful set of paths for which we know their expected lengths.

Definition 3.3.1. Let $P \subseteq \mathbb{R}^n$ be a polyhedron and $W \subseteq \mathbb{R}^n$ be a two-dimensional linear subspace. We denote by $\mathcal{S}(P, W)$ the set of *shadow vertices*: the vertices of P that maximize a non-zero objective function $\langle w, \cdot \rangle$ with $w \in W$.

From standard polyhedral theory, we get a characterization of shadow vertices:

Lemma 3.3.2. *Let $P(A)$ be a polyhedron given by $A \subseteq \mathbb{R}^n$ and $w \in \mathbb{R}^n \setminus \{\vec{0}\}$. A vertex $v \in P(A)$ maximizes $\langle w, \cdot \rangle$ if and only if $w\mathbb{R}_+ \cap \mathrm{conv}\{a \in A : \langle a, v \rangle = 1\} \neq \emptyset$.*

Hence for $W \subseteq \mathbb{R}^n$ a two-dimensional linear subspace, a vertex $v \in P(A)$ is a shadow vertex $v \in \mathcal{S}(P(A), W)$ if and only if $\mathrm{conv}\{a \in A : \langle a, v \rangle = 1\} \cap W \setminus \{\vec{0}\} \neq \emptyset$.

The set of shadow vertices for a fixed plane W induces a connected subgraph in the graph consisting of vertices and edges of P, and so any two shadow vertices are connected by a path of length at most $|\mathcal{S}(P, W)|$. As such, for nonzero $w_1, w_2 \in W$, we might speak of a *shadow path* from w_1 to w_2 to denote a path from a maximizer of $\langle w_1, \cdot \rangle$ to a maximizer of $\langle w_2, \cdot \rangle$ that stays inside $\mathcal{S}(P, W)$ and is monotonous with respect to $\langle w_2, \cdot \rangle$. The shadow path was studied by Borgwardt:

Theorem 3.3.3 ([28, 29])**.** *Let $m \geq n$ and fix a two-dimensional linear subspace $W \subseteq \mathbb{R}^n$. Pick any probability distribution on $\mathbb{R}^n$ that is invariant under rotations and let the entries of $A \subseteq \mathbb{R}^n$, $|A| = m$, be independently sampled from this distribution. Then, almost surely, for any linearly independent $w_1, w_2 \in W$ there is a unique shadow path from w_1 to w_2. Moreover, the vertices in $\mathcal{S}(P(A), W)$ are in one-to-one correspondence to the vertices of $\pi_W(P(A))$, the orthogonal projection of $P(A)$ onto W. The expected length of the shadow path from w_1 to w_2 is at most*

$$\mathbb{E}[|\mathcal{S}(P(A), W)|] = O(n^2 m^{\frac{1}{n-1}}).$$

This upper bound is tight up to constant factors for the uniform distribution on $\mathbb{S}^{n-1}$.

We prove a tail bound for the shadow size when $A \sim \text{Pois}(\mathbb{S}^{n-1}, m)$. This result answers a question of Borgwardt in the asymptotic regime, regarding whether bounds on higher moments of the shadow size can be given. To obtain such concentration, we show that the shadow decomposes into a sum of nearly independent "local shadows", using that A will be ε-dense per Lemma 3.2.6, allowing us to apply standard concentration results for sums of nearly independent random variables.

Theorem 3.3.4 (Shadow Size Concentration)**.** *Let* $e^{\frac{-m}{18\sqrt{n}(76\sqrt{2})^{n-1}}} < p < m^{-2n}$ *and let*

$$t_p := \max\left(\sqrt{O(Un^2 m^{\frac{1}{n-1}}\log(1/p))}, O(U\log(1/p))\right)$$

for $U := O(n2^{n^2}(\log(1/p))^n)$. *If* $A \sim \text{Pois}(\mathbb{S}^{n-1}, m)$ *then the shadow size satisfies*

$$\Pr\left[\left||\mathcal{S}(P(A), W)| - \mathbb{E}[|\mathcal{S}(P(A), W)|]\right| > t_p\right] \leq 4p.$$

In the second part of this section, we extend the resulting tools to obtain our upper bound on the diameter.

Theorem 3.3.5 (Diameter Upper Bound)**.** *Let* $e^{\frac{-m}{18\sqrt{n}(76\sqrt{2})^{n-1}}} < p < m^{-2n}$. *If* $A = \{a_1, \ldots, a_M\} \in \mathbb{S}^{n-1}$, *where* M *is Poisson with* $\mathbb{E}[M] = m$, *and* $a_1, \ldots, a_M$ *are uniformly and independently distributed in* $\mathbb{S}^{n-1}$. *Then, we have that*

$$\Pr[\text{diam}(P(A)) > O(n^2 m^{\frac{1}{n-1}} + n4^n \log(1/p)^2)] \leq O(\sqrt{p}).$$

3.3.1 Only 'nearby' constraints are relevant

We will start by showing that, with very high probability, constraints that are 'far away' from a given point on the sphere will not have any impact on the local shape of paths. That will result in a degree of independence between different parts of the sphere, which will be essential in getting concentration bounds on key quantities.

Lemma 3.3.6. *If* $A \subseteq \mathbb{S}^{n-1}$ *is* ε*-dense for* $\varepsilon \in [0, \sqrt{2})$ *then*

$$\mathbb{B}_2^n \subseteq P(A) \subseteq \left(1 - \frac{\varepsilon^2}{2}\right)^{-1} \mathbb{B}_2^n.$$

Proof. The first inclusion follows immediately from the construction of $P(A)$. We now show the second inclusion. Taking $x \in P(A) \setminus \{\vec{0}\}$, we must show that $\|x\| \leq (1-\varepsilon^2/2)^{-1}$. For this purpose, choose $a \in A$ such that $\|a - x/\|x\|\| \leq \varepsilon$, which exists by our assumption that A is ε-dense. Since $\varepsilon^2 \geq \|a-x/\|x\|\|^2 = 2(1-\langle a, x/\|x\|\rangle)$, we have that $\langle a, x/\|x\|\rangle \geq 1-\varepsilon^2/2$. Since $x \in P(A)$, we have $1 \geq \langle a, x\rangle \geq (1-\varepsilon^2/2)\|x\|$, and the bound follows by rearranging. □

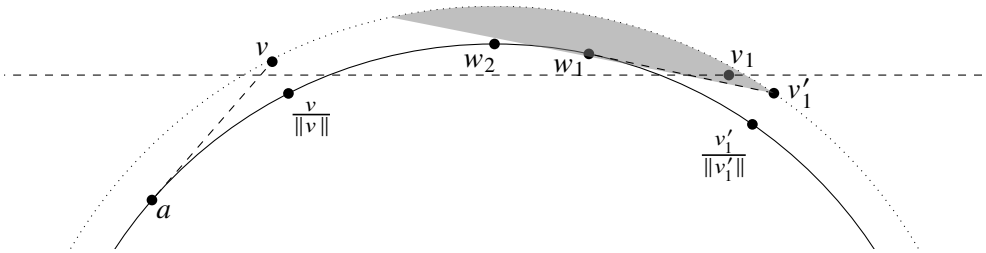


Figure 3.2: Illustration of the proof of Lemma 3.3.8. The inner (resp. outer dotted) curve represents part of the sphere $\mathbb{S}^{n-1}$ (resp. $(1-\varepsilon^2/2)^{-1}\mathbb{S}^{n-1}$). The horizontal dashed line represents the hyperplane $\{x \in \mathbb{R}^d : \langle x, w_2\rangle = \langle v_1, w_2\rangle\}$. The two oblique dashed line segments represent parts of the hyperplanes tangent to the unit sphere at the points a and w_1. The grey area represents the set B.

Lemma 3.3.7. *If $w \in \mathbb{S}^{n-1}$, $\alpha < 1$, $\|v\| \leq (1-\alpha)^{-1}$ and $\langle v, w\rangle \geq 1$ then we get $\|v/\|v\| - w\|^2 \leq 2\alpha$.*

Proof. We have $1 \leq \langle v, w\rangle = \|v\| \cdot \langle v/\|v\|, w\rangle \leq (1-\alpha)^{-1}\langle v/\|v\|, w\rangle$. Hence $1 - \|v/\|v\| - w\|^2/2 = \langle v/\|v\|, w\rangle \geq 1-\alpha$, which exactly implies that $\|v/\|v\| - w\|^2 \leq 2\alpha$ as required. □

We will use the above lemmas to prove the main technical estimate of this subsection: if $A \subseteq \mathbb{S}^{n-1}$ is ε-dense and $w_1, w_2 \in \mathbb{S}^{n-1}$ satisfy $\|w_1 - w_2\| \leq 2\varepsilon/n$ then any vertex on any path on $P(A)$ starting at a maximizer of $\langle w_1, \cdot\rangle$ that is non-decreasing with respect to $\langle w_2, \cdot\rangle$ can only be tight at constraints $\langle a, x\rangle = 1$ induced by $a \in A \cap C(w_2, (2+2/n)\varepsilon)$. All other constraints are strictly satisfied by every vertex on such a monotone path.

Lemma 3.3.8. *Let $\varepsilon \in [0,1]$ and assume that $w_1, w_2 \in \mathbb{S}^{n-1}$ satisfy $\|w_1 - w_2\| \leq (1-\varepsilon^2/2)$. Let $v_1, v \in \mathbb{R}^n$ satisfy $\langle w_1, v_1\rangle \geq 1$ and $\langle w_2, v\rangle \geq \langle w_2, v_1\rangle$, and assume $\|v_1\|, \|v\| \leq (1-\varepsilon^2/2)^{-1}$. Last, let $a \in \mathbb{S}^{n-1}$ satisfy $\langle a, v\rangle \geq 1$. Then we have $\|w_2 - a\| \leq 2\varepsilon + \|w_1 - w_2\|$.*

Proof. By Lemma 3.3.7, since $\langle w_1, v_1\rangle, \langle a, v\rangle \geq 1$ and $\|w_1\| = \|a\| = 1$, we get that $\|w_1 - v_1/\|v_1\|\|, \|a - v/\|v\|\| \leq \varepsilon$.

If $w_1 = w_2$, then by assumption $\langle v, w_2\rangle \geq \langle v_1, w_2\rangle = \langle v_1, w_1\rangle \geq 1$. Thus, Lemma 3.3.7 implies that $\|w_2 - v/\|v\|\| \leq \varepsilon$. By the triangle inequality, we conclude that $\|w_2 - a\| \leq \|w_2 - v/\|v\|\| + \|v/\|v\| - a\| \leq 2\varepsilon$, as needed.

Now assume that $w_1 \neq w_2$. To prove the lemma, we show that it suffices to find

a point v_1' such that the following two inequalities hold:

$$\left\| \frac{v}{\|v\|} - w_2 \right\| \leq \left\| \frac{v_1'}{\|v_1'\|} - w_2 \right\|, \qquad \left\| \frac{v_1'}{\|v_1'\|} - w_1 \right\| \leq \varepsilon. \tag{3.3}$$

Indeed, given v_1' as above, the triangle inequality and the first inequality of (3.3) imply that

$$\begin{aligned} \|w_2 - a\| &\leq \left\| w_2 - \frac{v}{\|v\|} \right\| + \left\| \frac{v}{\|v\|} - a \right\| \\ &\leq \left\| w_2 - \frac{v_1'}{\|v_1'\|} \right\| + \left\| \frac{v}{\|v\|} - a \right\| \\ &\leq \|w_2 - w_1\| + \left\| w_1 - \frac{v_1'}{\|v_1'\|} \right\| + \left\| \frac{v}{\|v\|} - a \right\|. \end{aligned}$$

From here, by the second inequality of (3.3) and $\|a - v/\|v\|\| \leq \varepsilon$, we get that

$$\|w_2 - a\| \leq \|w_2 - w_1\| + \varepsilon + \varepsilon,$$

which is the claim of the lemma. To construct v_1', let

$$B := \left\{ x \in \mathbb{R}^d : \langle w_1, x \rangle \geq 1,\ \|x\| \leq \left(1 - \frac{\varepsilon^2}{2}\right)^{-1} \right\}.$$

and define v_1' to be the minimizer of $\langle w_2, \cdot \rangle$ in B. Since $w_1 \neq w_2$, it is direct to verify the v_1' is uniquely defined and satisfies $\|v_1'\| = (1 - \frac{\varepsilon^2}{2})^{-1}$.

From Lemma 3.3.7 we have that any point $x \in B$ satisfies $\|x/\|x\| - w_1\| \leq \varepsilon$, and in particular this is true for v_1', making the second inequality of (3.3) hold. Note that $v_1 \in B$ as well. It remains to show the first inequality of (3.3). For this, we claim that

$$\langle w_2, v \rangle \geq \max\{0, \langle w_2, v_1' \rangle\},$$

By assumption, recall that $\langle w_2, v \rangle \geq \langle w_2, v_1 \rangle$. The first inequality now follows since $\langle w_2, v_1 \rangle \geq \langle w_1, v_1 \rangle - \|w_1 - w_2\|\|v_1\| \geq 1 - \|w_1 - w_2\|(1 - \varepsilon^2/2)^{-1} \geq 0$, by our assumption on $\|w_1 - w_2\|$. The second inequality now follows from $\langle w_2, v_1 \rangle \geq \langle w_2, v_1' \rangle$, which holds since $v_1 \in B$ and v_1' minimizes w_2 over B.

Using that $\|v\| \leq (1 - \varepsilon^2/2)^{-1} = \|v_1'\|$, we conclude that

$$\langle w_2, \frac{v}{\|v\|} \rangle \geq \langle w_2, \frac{v}{\|v'\|} \rangle \geq \langle w_2, \frac{v_1'}{\|v_1'\|} \rangle,$$

where the first inequality uses $\langle w_2, v \rangle \geq 0$. The first inequality of (3.3) now follows from the fact that $u \in \mathbb{S}^{n-1} \mapsto \|u - w_2\|$ is a decreasing function of $\langle u, w_2 \rangle$, and thus the proof is complete. □

To round out this subsection, we prove that the conclusion of Lemma 3.3.8 holds whenever $v, v_1 \in P(A)$ and A is ε-dense in a neighbourhood around w_2.

Definition 3.3.9. Given sets $A, C \subseteq \mathbb{S}^{n-1}$ and $\varepsilon > 0$, we say that A is ε-dense for C if for every $c \in C$ there exists $a \in A$ such that $\|a - c\| \leq \varepsilon$.

Lemma 3.3.10. *Let $A \subseteq \mathbb{S}^{n-1}$ be compact and ε-dense for $C(w_2, 4\varepsilon)$, $\varepsilon > 0$. Let $v_1, v \in P(A)$ and $w_1, w_2 \in \mathbb{S}^{n-1}$ satisfying $\langle w_1, v_1\rangle \geq 1$, $\langle w_2, v\rangle \geq \langle w_2, v_1\rangle$ and $\|w_1 - w_2\| \leq \varepsilon$. Now let $a \in \mathbb{S}^{n-1}$ satisfy $\langle a, v\rangle \geq 1$. Then we have $\|v_1\|, \|v\| \leq (1 - \varepsilon^2/2)^{-1}$ and $\|w_2 - a\| \leq 2\varepsilon + \|w_1 - w_2\|$.*

Proof. First, observe that if $\varepsilon \geq 1$ then the conclusion is trivially satisfied since $\|w_2 - a\| \leq 2 \leq 2\varepsilon + \|w_1 - w_2\|$. From now on, assume $\varepsilon < 1$.

Let $\delta = \|w_1 - w_2\|$, and let $A \subseteq A' \subseteq \mathbb{S}^{n-1}$ be ε-dense, such that $A' \cap C(w_2, 2\varepsilon + \delta) \subseteq A$. One valid choice is to take any ε-net $N \subseteq \mathbb{S}^{n-1}$ and define the set A' as $A' := A \cup (N \setminus C(w_2, 2\varepsilon + \delta))$. Then any $x \in C(w_2, 4\varepsilon)$ has an $a \in A \subseteq A'$ with $\|a - x\| \leq \varepsilon$ and any $y \notin C(w_2, 4\varepsilon)$ has some $b \in N$ with $\|y - b\| \leq b$ and $b \notin C(w_2, 3\varepsilon)$. Moreover we have $(N \setminus C(w_2, 3\varepsilon)) \cap C(w_2, (2 + 2/n)\varepsilon) = \emptyset$ so this choice of A' satisfies our requirements.

If $v, v_1 \in P(A')$, then $\|v\|, \|v_1\| \leq (1-\varepsilon^2/2)^{-1}$ by Lemma 3.3.6 and we can apply Lemma 3.3.8 to the set A' and vectors w_1, w_2, v, v_1 and a to conclude $\|w_2 - a\| \leq 2\varepsilon + \|w_1 - w_2\|$ as required.

We now prove that both the case $v_1 \notin P(A')$ and the case $v_1 \in P(A'), v \notin P(A')$ lead to contradiction. First, observe that given w_1 and w_2, the set of pairs (v_1, v) that satisfy $\langle w_1, v_1\rangle \geq 1, \langle w_2, v\rangle \geq \langle w_2, v_1\rangle$ and $\|v_1\|, \|v\| \leq (1 - \varepsilon^2/2)^{-1}$ is a closed convex set and contains (w_1, w_1).

If $v_1 \notin P(A')$, let (x, y) be the convex combination of (v_1, v_1) and (w_1, w_1) such that $x = y \in P(A')$ and there exists $a' \in A' \setminus A$ such that $\langle a', x\rangle = 1$. Such a' will exist because A' is compact.

Otherwise we have $v \notin P(A')$ and let (x, y) be a convex combination of (v_1, v) and (w_1, w_1) such that $x, y \in P(A')$ and there exists $a' \in A' \setminus A$ such that $\langle a', x\rangle = 1$. Such a' will exist because A' is compact.

Either way, apply Lemma 3.3.8 to A', w_1, w_2, x, y and a' to find that $\|w_2 - a'\| \leq 2\varepsilon + \|w_1 - w_2\|$. This contradicts the earlier claim that $a' \in A' \setminus A$. From this contradiction we conclude that $v, v_1 \in P(A')$, which finishes the proof. □

Note also the contrapositive of the above statement: for w_1, w_2, v_1, v, A satisfying the conditions above, we have for $a \in \mathbb{S}^{n-1}$ that $\|w_2 - a\| > 2\varepsilon + \|w_1 - w_2\|$ implies $\langle a, v\rangle < 1$.

3.3.2 Locality, independence, and concentration

With an eye to Lemma 3.3.10, this subsection is concerned with proving concentration for sums of random variables that behave nicely when A is dense in given neighbourhoods. The specific random variables that we will use this for are the paths between the maximizers of nearby objective vectors $w_1, w_2 \in \mathbb{S}^{n-1}$.

Definition 3.3.11. Given m, n, p, let $\varepsilon = \varepsilon(m, n, p) > 0$ be as in Lemma 3.2.6 and $A \subseteq \mathbb{R}^n$ be a random finite set. For $x, y \in \mathbb{S}^{n-1}$ define the event $E_{x,y}$ as:

- A is ε-dense for $C(x, \|x - y\| + 4\varepsilon)$, and
- for every $z \in [x, y]$ we have

$$\left| A \cap C(\frac{z}{\|z\|}, (2 + 2/n)\varepsilon) \right| \leq 45e2^n \log(1/p)$$

A random variable K is called (x, y)-local if $E_{x,y}$ implies that K is a function of $A \cap C(x, 5\varepsilon + \|x - y\|)$.

In particular, we will use that if K is (x, y)-local then $K\mathbb{1}[E_{x,y}]$ is a function of $A \cap C(x, 5\varepsilon + \|x - y\|)$.

To help prove that certain paths are local random variables, we will use the following lemma.

Lemma 3.3.12. *Let $w_1, w_2 \in \mathbb{S}^{n-1}$, and have $w_1 = v_1, v_2, \ldots, v_{k+1} = w_2$ be equally spaced on a shortest geodesic segment on $\mathbb{S}^{n-1}$ connecting w_1 and w_2. Then for every $i \in [k]$ we have $\|w_1 - w_2\|/k \leq \|v_i - v_{i+1}\| \leq \pi\|w_1 - w_2\|/k$.*

Proof. By the triangle inequality, we have $\|w_1 - w_2\| \leq \sum_{i=1}^{k} \|v_i - v_{i+1}\|$. Since each of the line segments $[v_i, v_{i+1}]$ has identical length, this gives us the first inequality $k\|v_i - v_{i+1}\| \geq \|w_1 - w_2\|$.

Furthermore, we know that the geodesic segment connecting w_1 and w_2 has length at most $\pi\|w_1 - w_2\|$. From this we get $\sum_{i=1}^{k} \|v_i - v_{i+1}\| \leq \pi\|w_1 - w_2\|$ and hence $\pi\|w_1 - w_2\| \geq k\|v_i - v_{i+1}\|$. □

Many paths on $P(A)$ turn out to be such local random variables. One example are short segments of the shadow paths from Theorem 3.3.3.

Lemma 3.3.13. *Let $w_1, w_2 \in \mathbb{S}^{n-1}$ satisfy $\|w_1 - w_2\| \leq \varepsilon$. Then the length of the shadow path on $P(A)$ from w_1 to w_2 is a (w_1, w_2)-local random variable. Assuming that $\|w_1 - w_2\| \leq \varepsilon$, the event E_{w_1,w_2} implies that this path has length at most $2n(45e2^n \log(1/p))^n$.*

Proof. Let us first assume that $\|w_1 - w_2\| \leq 2\varepsilon/n$. Consider the points $v_1, v \in P(A \cap C(w_2, 5\varepsilon))$ such that $\langle w_1, v_1\rangle \geq 1$ and $\langle w_2, v\rangle \geq \langle w_2, v_1\rangle$. By Lemma 3.3.10, assuming E_{w_1,w_2}, any such points have bounded norm. Hence, we can take v_1 to be a vertex maximizing $\langle w_1, \cdot\rangle$ and $v \in P(A \cap C(w_2, 5\varepsilon))$ be any vertex on the shadow path from w_1 to w_2.

Again by Lemma 3.3.10, assuming E_{w_1,w_2}, every $a \in A$ such that $\langle a, v\rangle = 1$ satisfies $a \in C(w_2, (2 + 2/n)\varepsilon)$, meaning that $v, v_1 \in P(A)$ as well.

Now Lemma 3.3.2 implies that if E_{w_1,w_2} then any vertex of $P(A \cap C(w_2, 5\varepsilon)$ on its shadow path from w_1 to w_2 is a shadow vertex of $P(A)$ on the shadow path from w_1 to w_2. Hence the shadow path on $P(A)$ from w_1 to w_2 is a (w_1, w_2)-local random variable.

The upper bound follows because every vertex on the shadow path is visited at most once and, assuming E_{w_1,w_2}, almost surely every vertex on the shadow path is induced by n constraints out of $A \cap C(w_2, (2 + 2/n)\varepsilon)$. The total number of subsets of size n of $A \cap C(w_2, (2 + 2/n)\varepsilon)$ is at most $|A \cap C(w_2, (2 + 2/n)\varepsilon)|^n \leq (45e2^n \log(1/p))^n$ by E_{w_1,w_2}.

To extend the conclusion to the case when $2\varepsilon/n < \|w_1 - w_2\| \leq \varepsilon$, pick $w_1 = v_1, v_2, \ldots, v_{2n+1} = w_2$ evenly spaced on the shortest geodesic segment connecting w_1 and w_2. For every $k \in [2n]$, by Lemma 3.3.12 the shadow path from v_k to v_{k+1} satisfies $\|v_k - v_{k+1}\| \leq 2\varepsilon/n$ and is thus a (v_k, v_{k+1})-local random variable and $E_{v_k,v_{k+1}}$ implies that this shadow path has length at most $(45e2^n \log(1/p))^n$ when $E_{v_k,v_{k+1}}$.

Now observe that the shadow path from w_1 to w_2 is obtained by concatenating the shadow paths from v_k to v_{k+1} for $k \in [n]$. Since E_{w_1,w_2} implies $E_{v_k,v_{k+1}}$ for every $k \in [2n]$, each of the shadow paths from v_k to v_{k+1} is a (w_1, w_2)-local random variable. Hence the shadow path from w_1 to w_2 is a (w_1, w_2)-local random variable and has length at most $2n(45e2^n \log(1/p))^n$. □

Lemma 3.3.14. *Let $0 < p < m^{-2n}$ and let $\varepsilon = \varepsilon(m, n, p) < 1/76$ be as in Lemma 3.2.6 and let $k \geq 2\pi/\varepsilon$ be the smallest number divisible by 76. Let $W \subseteq \mathbb{R}^n$ be a fixed 2D linear subspace and let $w_1, \ldots, w_k, w_{k+1} = w_1 \in W \cap \mathbb{S}^{n-1}$ be equally spaced around the circle. Assume for every $i \in [k]$ that $K_i \geq 0$ is a (w_i, w_{i+1})-local random variable and there exists $U \leq m^n$ such that $K_i \leq U$ whenever $E_{w_i,w_{i+1}}$. Furthermore assume that $\mathbb{E}[\sum_{i=1}^k K_i] \leq O(n^2 m^{\frac{1}{n-1}})$. Then*

$$\Pr\Big[\Big|\sum_{i\in[k]} K_i - \mathbb{E}\Big[\sum_{i\in[k]} K_i\Big]\Big| \geq t_p\Big] \leq 4p$$

for $t_p = \max\left(\sqrt{O(Un^2 m^{\frac{1}{n-1}} \log(1/p))}, O(U \log(1/p))\right)$.

Proof. Let F denote the event that E_{v_1,v_2} holds for every $v_1, v_2 \in \mathbb{S}^{n-1}$. By Lemma 3.2.6 we have $\Pr[F] \geq 1 - 2p$.

Our first observation is that $\Pr[\sum_{i=1}^{k} K_i = \sum_{i=1}^{k} K_i \mathbb{1}[E_i]] \geq \Pr[F] \geq 1 - p$. Since both sums only take values in the interval $[0, km^n]$, it follows that

$$\left|\mathbb{E}[\sum_{i=1}^{k} K_i] - \mathbb{E}[\sum_{i=1}^{k} K_i \mathbb{1}[E_i]]\right| \leq 2km^n p \leq 1.$$

From the above statements we deduce that

$$\begin{aligned}
&\Pr\left[\left\|\sum_{i=1}^{k} K_i - \mathbb{E}\left[\sum_{i=1}^{k} K_i\right]\right\| > t_p\right] \\
&\qquad \leq \Pr\left[\left\|\sum_{i=1}^{k} K_i - \mathbb{E}\left[\sum_{i=1}^{k} K_i \mathbb{1}[E_i]\right]\right\| > t_p - 1\right] \\
&\qquad \leq \Pr[F^c] + \Pr\left[F \wedge \left\|\sum_{i=1}^{k} K_i - \mathbb{E}\left[\sum_{i=1}^{k} K_i \mathbb{1}[E_i]\right]\right\| > t_p - 1\right] \\
&\qquad \leq 2p + \Pr\left[F \wedge \left\|\sum_{i=1}^{k} K_i \mathbb{1}[E_i] - \mathbb{E}\left[\sum_{i=1}^{k} K_i \mathbb{1}[E_i]\right]\right\| > t_p - 1\right] \\
&\qquad \leq 2p + \Pr\left[\left\|\sum_{i=1}^{k} K_i \mathbb{1}[E_i] - \mathbb{E}\left[\sum_{i=1}^{k} K_i \mathbb{1}[E_i]\right]\right\| > t_p - 1\right].
\end{aligned}$$

In the above, recall that F^c denotes the complement of F. We will now upper bound the last term.

For $j \in [76]$ define $I_j = \{i \in [k] \mid i \equiv j \mod 76\}$, forming a partition $I_1 \cup \cdots \cup I_{76} = [k]$. Observe that $w_1, \ldots, w_k$ are placed on a unit circle and every $[w_i, w_{i+1}]$ is an edge of $\mathrm{conv}(w_1, \ldots, w_k)$, hence $\sum_{i \in [k]} \|w_i - w_{i+1}\| \leq 2\pi$. Since $k \geq 2\pi/\varepsilon$ that gives us $\|w_i - w_{i+1}\| \leq \varepsilon$ for every $i \in [k]$. Next, from $\varepsilon \leq 1/76$ we know that $k \leq 2\pi/\varepsilon + 76 \leq 8/\varepsilon$. Since $k \geq 4$ we have $\sum_{i \in [k]} \|w_i - w_{i+1}\| \geq 4$ and hence $\|w_i - w_{i+1}\| \geq 4/k \geq \varepsilon/2$ for every $i \in [k]$. Last, we use that $\|w_i - w_{i+76}\| \leq \sum_{j=i}^{i+75} \|w_j - w_{j+1}\| \leq 76\varepsilon \leq 1$ to deduce

$$\|w_i - w_{i+76}\| \geq \frac{1}{\pi} \sum_{k=i}^{i+75} \|w_k - w_{k+1}\| \geq \frac{76}{\pi} \cdot \varepsilon/2 > 12\varepsilon.$$

This lets us conclude that if $i, i' \in I_j$ are distinct then $\|w_i - w_{i'}\| > 12\varepsilon$. In particular, for any $j \in [76]$ the random variables $K_i \mathbb{1}[E_i]$ for $i \in I_j$ are mutually independent since they are functions of A intersected with disjoints subsets of $\mathbb{S}^{n-1}$ due to being local random variables.

For any $i \in [k]$, the random variable $K_i \mathbb{1}[E_i] \in [0, U]$ has variance at most

$$\mathbb{E}[K_i \mathbb{1}[E_i]] \cdot U \leq \frac{O(n^2 m^{\frac{1}{n-1}})}{k} \cdot U$$

by Lemma 3.2.9.

We apply Lemma 3.2.8 to the random variables $K_i \mathbb{1}[E_i]$ for $i \in [k]$ and obtain

$$\Pr\left[\left|\sum_{i=1}^{k} K_i \mathbb{1}[E_i] - \mathbb{E}\left[\sum_{i=1}^{k} K_i \mathbb{1}[E_i]\right]\right| > t_p - 1\right] \leq 2\exp\left(\frac{-8(t_p-1)^2}{1900(UO(n^2 m^{\frac{1}{n-1}}) + (t_p-1)U)}\right).$$

By filling in t_p, we find that the right-hand side of the above inequality is at most $2p$.

Putting the bounds together we get our desired inequality

$$\Pr\Big[\sum_{i\in[k]} K_i \geq \mathbb{E}\Big[\sum_{i\in[k]} K_i\Big] + t\Big] \leq 4p. \qquad \square$$

3.3.3 Concentration of the shadow size around its mean

To illustrate the use of the above technical result, we show in this subsection that $|\mathcal{S}(P(A), W)|$ is concentrated around its mean when $m > 2^{O(n^3)}$.

Recall that by Theorem 3.3.3 we have $\mathbb{E}[|\mathcal{S}(P(A), W)|] = \Theta(n^2 m^{\frac{1}{n-1}})$.

Theorem 3.3.4 (Shadow Size Concentration)**.** *Let* $e^{\frac{-m}{18\sqrt{n}(76\sqrt{2})^{n-1}}} < p < m^{-2n}$ *and let*

$$t_p := \max\left(\sqrt{O(Un^2 m^{\frac{1}{n-1}} \log(1/p))}, O(U\log(1/p))\right)$$

for $U := O(n2^{n^2}(\log(1/p))^n)$*. If* $A \sim \mathrm{Pois}(\mathbb{S}^{n-1}, m)$ *then the shadow size satisfies*

$$\Pr\Big[\Big||\mathcal{S}(P(A), W)| - \mathbb{E}[|\mathcal{S}(P(A), W)|]\Big| > t_p\Big] \leq 4p.$$

Proof. From Corollary 3.2.7, we know that $\varepsilon^{n-1} \leq \frac{1}{76^{n-1}}$. As such, the lower bound on p implies that $\varepsilon(m, n, p) < 1/76$.

Let $w_1, \ldots, w_k$ be as in Lemma 3.3.14 and let K_i denote the number of edges on the shadow path from w_i to w_{i+1}. By Lemma 3.3.13, each K_i is a (w_i, w_{i+1})-local random variable which satisfies $K_i \leq 2n(45e2^n \log(1/p))^n$ when $E_{w_i, w_{i+1}}$.

By Theorem 3.3.3 we get $\sum_{i\in[k]} K_i = |\mathcal{S}(P(A), W)|$ almost surely, hence we have $\mathbb{E}[\sum_{i\in[k]} K_i] \leq O(n^2 m^{\frac{1}{n-1}})$. We apply Lemma 3.3.14 to $\sum_{i\in[k]} K_i$ to conclude

$$\Pr\Big[\Big||\mathcal{S}(P(A), W)| - \mathbb{E}[|\mathcal{S}(P(A), W)|]\Big| > t\Big] = \Pr\Big[\Big|\sum_{i\in[k]} K_i - \mathbb{E}[\sum_{i\in[k]} K_i]\Big| > t\Big] \leq 4p.$$

$\square$

3.3.4 Upper bound on the diameter

In this section we prove our high probability upper bound on $\operatorname{diam}(P(A))$. We start by proving that for fixed W the vertices in $\mathcal{S}(P(A), W)$ are connected by short paths, where we aim for an error term smaller than that of Theorem 3.3.4. We require the following abstract diameter bound from [77]. We will only need the Barnette–Larman style bound.

Theorem 3.3.15. *Let $G = (V, E)$ be a connected graph, where the vertices V of G are subsets of $[k]$ of cardinality n and the edges E of G are such that for each $u, v \in V$ there exists a path connecting u and v whose intermediate vertices all contain $u \cap v$.*

Then the following upper bounds on the diameter of G hold:

$$2^{n-1} \cdot k - 1 \quad (\textit{Barnette–Larman}), \qquad k^{1+\log n} - 1 \quad (\textit{Kalai–Kleitman}).$$

To confirm that the above theorem indeed gives variants of the Barnette–Larman and Kalai–Kleitman bounds, let $A = \{a_1, ..., a_m\} \subseteq \mathbb{S}^{n-1}$ be in general position. For a vertex $x \in P(A)$, we denote $A_x = \{a \in A : \langle a, x\rangle = 1\}$. Consider the following sets

$$V = \{A_x : x \text{ is a vertex of } P(A)\},$$
$$E = \{\{A_x, A_y\} : [x, y] \text{ is an edge of } P(A)\}.$$

The graph $G = (V, E)$ almost surely satisfies the assumptions of Theorem 3.3.15 which therefore shows that the combinatorial diameter of $P(A)$ is less than $\min(2^{n-1} \cdot M - 1, M^{1+\log n} - 1)$. Up to a constant factor difference, these bounds correspond to the same bounds described in the introduction. Instead of applying the bound to the full graph however, we will use it to bound the length of local paths.

Lemma 3.3.16. *Let $w_1, w_2 \in \mathbb{S}^{n-1}$ satisfy $\|w_1 - w_2\| \leq \varepsilon$, where $\varepsilon = \varepsilon(m, n, p)$ is as in Lemma* 3.2.6. *Furthermore, let K denote the maximum over all $w \in [w_1, w_2]$ of the length of the shortest path from a maximizer $v_w \in P(A)$ of $\langle w, \cdot\rangle$ to the maximizer of $\langle w_2, \cdot\rangle$ of which every vertex $v \in P(A)$ on the path satisfies $\langle w_2, v\rangle \geq \langle w_2, v_w\rangle$. Then K is a (w_1, w_2)-local random variable and E_{w_1,w_2} implies that K is at most $45en4^n \log(1/p)$.*

Proof. We start by assuming $\|w_1 - w_2\| \leq 2\varepsilon/n$. Let $w \in [w_1, w_2]$ and let $v_w \in P(A)$ be a vertex maximizing $\langle w, \cdot\rangle$. By Lemma 3.3.10, assuming E_{w_1,w_2}, for every vertex $v \in P(A)$ satisfying $\langle w_2, v\rangle \geq \langle w_2, v_w\rangle$ and every $a \in A$ such that $\langle a, v\rangle \geq 1$ we have $a \in A \cap C(w_2, (2 + 2/n)\varepsilon)$.

First, this implies that if E_{w_1,w_2} and if $v \in \mathbb{R}^n$ is satisfies $\langle w_2, v\rangle \geq \langle w_2, v_w\rangle$ then we need only inspect $A \cap C(w_2, (2 + 2/n)\varepsilon)$ to decide if v is a vertex of $P(A)$. From this we conclude that if E_{w_1,w_2} then the shortest path described in the lemma

statement can be computed knowing only $A \cap C(w_2, (2+2/n)\varepsilon)$. This implies that the path length is a (w_1, w_2)-local random variable.

Second, assuming E_{w_1,w_2} we consider the sets

$$\widehat{V} = \{v \in P(A) : v \text{ is a vertex and } \langle w_2, v\rangle \geq \langle w_2, v_1\rangle\},$$

$$\widehat{A} = \{a \in A : \text{ there exist } v \in \widehat{V} \text{ such that } \langle a, v\rangle = 1\} \subseteq A \cap C\left(w_2, \left(2+\frac{2}{n}\right)\varepsilon\right).$$

The last inclusion follows directly from Lemma 3.3.10.

Recall the notation $A_v = \{a \in A : \langle a, v\rangle = 1\}$ for vertices $v \in P(A)$. We will apply Theorem 3.3.15 to the graph

$$V = \{A_v : v \in \widehat{V}\} \simeq \widehat{V},$$
$$E = \{\{A_{v_1}, A_{v_2}\} : v_1, v_2 \in \widehat{V},\ [v_1, v_2] \text{ is an edge of } P(A)\}.$$

We need to check that the assumptions of Theorem 3.3.15 are met. First we note that almost surely $P(A)$ is a simple polytope and thus the vertices of the graph (V, E) are subsets of A of cardinality n. Consider two vertices $A_v = \{a_{i_1}, \ldots, a_{i_n}\}, A_{v'} = \{a_{i'_1}, \ldots, a_{i'_n}\} \in V$. Observe that the set

$$F = \{x \in P(A) : \langle x, a\rangle = 1\ \forall a \in A_v \cap A_{v'}\}$$

is the minimum face of $P(A)$ containing both v and v'. We can build paths $v_0 = v, v_1, \ldots, v_k$ and $v'_0 = v', v'_1, \ldots, v'_{k'}$ satisfying the following monotonicity properties

$$\langle w_2, v\rangle = \langle w_2, v_0\rangle \leq \langle w_2, v_1\rangle \leq \cdots \leq \langle w_2, v_k\rangle = \arg\max\{\langle w_2, x\rangle : x \in F\},$$
$$\langle w_2, v'\rangle = \langle w_2, v'_0\rangle \leq \langle w_2, v'_1\rangle \leq \cdots \leq \langle w_2, v'_{k'}\rangle = \arg\max\{\langle w_2, x\rangle : x \in F\}.$$

Moreover one can assume that $v_k = v'_{k'}$ by potentially completing the paths moving along the edges of $\arg\max\{\langle w_2, x\rangle : x \in F\}$ (in the case this face contains more than one vertex). By construction all vertices v_i and v'_i belong to $\widehat{V}$. Stitching the two paths and adopting the dual point of view we found a path $A_v = A_{v_0}, \ldots, A_{v_k} = A_{v'_{k'}}, \ldots A_{v'_0} = A_{v'}$ whose vertices contain the intersection $A_v \cap A_{v'}$.

We can thus apply Theorem 3.3.15 and conclude that there is a path in the graph (V, E) from A_{v_1} to A_{v_2} of length at most $2^{n-1} \cdot |A \cap C(w_2, (2+2/n)\varepsilon)|$. It follows that $K \leq 2^{n-1} \cdot |A \cap C(w_2, (2+2/n)\varepsilon)|$.

To extend the conclusion to the case when $2\varepsilon/n < \|w_1 - w_2\| \leq \varepsilon$, we do the same as in the proof of Lemma 3.3.13. □

Theorem 3.3.17. *Let* $0 < p < m^{-2n}$ *and let*

$$t_p = \max\left(\sqrt{O(Un^2 m^{\frac{1}{n-1}} \log(1/p))}, O(U \log(1/p))\right)$$

for $U = O(n4^n \log(1/p))$. If $W \subseteq \mathbb{R}^n$ is a fixed 2D linear subspace and $A \sim \text{Pois}(\mathbb{S}^{n-1}, m)$, the largest distance T between any two shadow vertices satisfies

$$\Pr[T \geq O(n^2 m^{\frac{1}{n-1}}) + t_p] \leq 4p$$

Proof. Let $w_1, \ldots, w_k$ be as in Lemma 3.3.14 and let K_i denote the maximum over all $w \in [w_i, w_{i+1}]$ of the length of the shortest path from a shadow vertex v_w maximizing $\langle w, \cdot \rangle$ to a vertex maximizing $\langle w_{i+1}, \cdot \rangle$ such that every vertex v on this path satisfies $\langle w_{i+1}, v \rangle \geq \langle w_{i+1}, v_w \rangle$.

From Lemma 3.3.16 we know that K_i is a (w_i, w_{i+1})-local random variable and $K_i \leq 45en4^n \log(1/p)$ whenever $E_{w_i, w_{i+1}}$.

Now recall Theorem 3.3.3. Observe that $T \leq \sum_{i \in [k]} K_i$ almost surely by concatenating the above-mentioned paths, and note that $\sum_{i \in [k]} K_i \leq \mathcal{S}(P(A), W)$ holds almost surely, which implies $\mathbb{E}[\sum_{i \in [k]} K_i] = O(n^2 m^{\frac{1}{n-1}})$. We apply Lemma 3.3.14 to $\sum_{i \in [k]} K_i$ and get the desired result. □

Theorem 3.3.5 (Diameter Upper Bound)**.** *Let $e^{\frac{-m}{18\sqrt{n}(76\sqrt{2})^{n-1}}} < p < m^{-2n}$. If $A = \{a_1, \ldots, a_M\} \in \mathbb{S}^{n-1}$, where M is Poisson with $\mathbb{E}[M] = m$, and $a_1, \ldots, a_M$ are uniformly and independently distributed in $\mathbb{S}^{n-1}$. Then, we have that*

$$\Pr[\text{diam}(P(A)) > O(n^2 m^{\frac{1}{n-1}} + n4^n \log(1/p)^2)] \leq O(\sqrt{p}).$$

Proof. From Corollary 3.2.7, we know that $\varepsilon^{n-1} \leq \frac{1}{76^{n-1}}$. As such, the lower bound on p implies that $\varepsilon(m, n, p) < 1/76$.

Let $N \subseteq \mathbb{S}^{n-1}$ be a fixed minimal ε-net. Consider the following statements:

- For every $n \in N$, any two vertices in $\mathcal{S}(P(A), \text{span}(e^1, n))$ are connected by a path of length at most $O(n^2 m^{\frac{1}{n-1}}) + t$, where t is defined in Theorem 3.3.17.
- A is ε-dense.
- For any $x \in \mathbb{S}^{n-1}$ we have $|A \cap C(x, (2 + 2/n)\varepsilon)| \leq 45e2^n \log(1/p)$.

For given $n \in N$, the first event holds with probability at least $1-4p$ by Theorem 3.3.17. The net N has $|N| \leq (4/\varepsilon)^n$ points, which is at most $4^n \cdot m$ by Corollary 3.2.7. By the union bound the first statement holds for all $n \in N$ simultaneously with probability at least $1 - \sqrt{p}$. From Lemma 3.2.6 we know that the second statement holds with probability at least $1 - p$ and the third statement holds with probability at least $1 - p$. We conclude that all three statements hold simultaneously with probability at least $1 - O(\sqrt{p})$.

We will show that the above conditions imply the bound on the combinatorial diameter of $P(A)$.

Observe that we only need to show an upper bound for all $w \in \mathbb{S}^{n-1}$ on the length of a path connecting any vertex maximizing $\langle w, \cdot \rangle$ to a vertex maximizing $\langle e^1, \cdot \rangle$. The combinatorial diameter of $P(A)$ is at most twice that upper bound.

Let $w \in \mathbb{S}^{n-1}$ and pick $n \in N$ such that $\|w - n\| \leq \varepsilon$. By the first statement, there is a path from the vertex maximizing $\langle n, \cdot \rangle$ to the vertex maximizing $\langle e^1, \cdot \rangle$ of length $O(n^2 m^{\frac{1}{n-1}}) + t$.

By the second two statements, E_{w_1,w_2} is satisfied for every $w_1, w_2 \in \mathbb{S}^{n-1}$. We conclude from Lemma 3.3.16 that there is a path from any vertex maximizing $\langle w, \cdot \rangle$ to the vertex maximizing $\langle n, \cdot \rangle$ of length $45en4^n \log(1/p)$.

Therefore, when all three statements hold the combinatorial diameter of $P(A)$ is at most $O(n^2 m^{\frac{1}{n-1}}) + t_p + 45en4^n \log(1/p)$ Now we fill in t_p and obtain an upper bound of

$$O(n^2 m^{\frac{1}{n-1}} + n4^n \log(1/p)^2). \qquad \square$$

3.4 Lower Bounding the Diameter

To begin, we first reduce to lower bounding the diameter of the polar polytope P°, corresponding to a convex hull of m uniform points on $\mathbb{S}^{n-1}$, via the following simple lemma.

Lemma 3.4.1 (Diameter Relation)**.** *For $n \geq 2$, let $P \subseteq \mathbb{R}^n$ be a simple bounded polytope containing the origin in its interior and denote its (simplicial) polar polytope by $Q = P^\circ := \{x \in \mathbb{R}^n : \langle x, y \rangle \leq 1, \forall y \in P\}$. Then,* $\operatorname{diam}(P) \geq (n-1)(\operatorname{diam}(Q) - 2)$.

We then associate any "antipodal" path to a continuous curve on the sphere corresponding to objectives maximized by vertices along the path. From here, we decompose any such curve into $\Omega(m^{\frac{1}{n-1}})$ segments whose endpoints are at distance $\Theta(m^{-1/(n-1)})$ on the sphere. Finally, we apply a suitable union bound, to show that for any such curve, an $\Omega(1)$ fraction of the segments induce at least 1 edge on the corresponding path.

Theorem 3.4.2 (Lower Bound for $Q(A)$)**.** *There exist positive constants $c_2 < 1$ and $c_3 > 1$ independent of $n \geq 3$ and m such that the following holds. Let $A = \{a_1, \ldots, a_M\} \in \mathbb{S}^{n-1}$, where M is Poisson with $\mathbb{E}[M] = m$, and $a_1, \ldots, a_M$ are uniformly and independently distributed in $\mathbb{S}^{n-1}$. Then, with probability at least $1 - e^{-c_3^{n-1} m^{1/(n-1)}}$, the combinatorial diameter of $Q(A)$ is at least $c_2 m^{1/(n-1)}$.*

3.4.1 Relating the diameter of $Q(A)$ and $P(A)$

Proof of Lemma 3.4.1. If $\text{diam}(Q) \leq 1$, the statement is trivial, so we may assume that $\text{diam}(Q) \geq 2$. Let $a_1, a_2 \in Q$ be vertices of Q at distance $\text{diam}(Q) \geq 2$. Since P is bounded, note that $\vec{0}$ is in the interior of Q and hence $a_1, a_2 \neq \vec{0}$. We must show that there exists a path from a_1 to a_2 of length $L \geq 2$ such $\text{diam}(P) \geq (n-1)(L-2)$.

Let $F_i := \{x \in P : \langle a_i, x\rangle = 1\}$, $i \in [2]$, the corresponding facets of P. Pick the two vertices $v_1 \in F_1$, $v_2 \in F_2$ whose distance in P is minimized. Let $v_1 := w_0, \ldots, w_D := v_2$ be a shortest path from v_1 to v_2 in P. Here $w_0, \ldots, w_D$ are all vertices of P, and $[w_i, w_{i+1}]$, $0 \leq i \leq D-1$, are edges of P. By definition, $D \leq \text{diam}(P)$.

To complete the proof, we will extract a walk from a_1 to a_2 in Q from the path $w_0, \ldots, w_D$ of length at most $D/(n-1)+2$. Let $Q_i := Q \cap \{x \in \mathbb{R}^n : \langle x, w_i\rangle = 1\}$, $0 \leq i \leq D$, denote the facet of Q induced by w_i. By our assumption that P is simple, each Q_i, $i \in [D]$, is a $(n-1)$-dimensional simplex, and hence there exists $S_i \subseteq \text{vertices}(Q)$, $|S_i| = n$, such that $Q_i := \text{conv}(a : a \in S_i)$. In particular, the combinatorial diameter of each Q_i, $0 \leq i \leq D$, is 1. That is, every distinct pair of vertices of Q_i induces an edge of Q_i, and hence an edge of Q.

By the above discussion, note that if $a_1, a_2 \in S_0$, then a_1, a_2 are adjacent in Q. Since we assume that the distance between a_1, a_2 is at least 2, we conclude that $a_2 \notin S_0$, and hence that $D \geq 1$. Furthermore, since we assume that v_1, v_2 are at minimum distance in P subject to $v_1 \in F_1, v_2 \in F_2$, we conclude that $a_1 \in S_0 \setminus \cup_{j=1}^{D} S_j$ and $a_2 \in S_L \setminus \cup_{j=0}^{D-1} S_j$, since otherwise we could shortcut the path.

We now define a walk $a_1 = u_0, \ldots, u_L = a_2$, for some $L \geq 2$, from a_1 to a_2 in Q as follows. Letting $l_0 = 0$ and $S_{D+1} := \emptyset$, for $i \geq 1$ inductively define $l_i := \max\{j \geq l_{i-1} : \cap_{r=l_{i-1}}^{j} S_r \neq \emptyset\}$ and let $L = \min\{i \geq 1 : l_i = D\} + 1$. For $1 \leq i \leq L-1$, choose u_i from $\cap_{r=l_{i-1}}^{l_i} S_r$ arbitrarily. To relate the length of the walk to D, we will need the following claim.

Claim 3.4.3. *For any interval $I \subseteq \{0, \ldots, D\}$, $|\bigcap_{i \in I} S_i| \geq n - |I| + 1$.*

Proof. First note that $|S_j \cap S_{j+1}| = n-1 = |S_j| - 1$ for all $0 \leq j \leq D-1$, since P is simple and $S_j \cap S_{j+1}$ indexes the tight constraints of an edge of P. In particular, $|S_j \setminus S_{j+1}| = 1, 0 \leq j \leq D-1$. Thus, for an interval $I = \{c, c+1, \ldots, d\} \subseteq \{0, \ldots, D\}$, we see that $|\cap_{i=c}^{d} S_i| \geq |\cap_{i=c}^{d-1} S_i| - |S_{d-1} \setminus S_d| = |\cap_{i=c}^{d-1} S_i| - 1 \geq |S_c| - (d-c) = n + 1 - |I|$. □

Applying the claim to the interval $I = \{l_{i-1}, \ldots, l_i + 1\}$, $1 \leq i \leq L-1$, we see that

$$\cap_{r=l_{i-1}}^{l_i+1} S_r = \emptyset$$

implies that either $l_i = D$ or that $|I| \geq n+1 \Leftrightarrow l_i - l_{i-1} \geq n-1$. In particular, $l_i - l_{i-1} \geq n-1$ for $0 \leq i \leq L-2$ and $l_{L-1} - l_{L-2} \geq 1$ (since $l_{L-1} = D$ and $l_{L-2} < D$).

Let us now verify that $a_1 = u_0, u_1, \ldots, u_L = a_2$ induces a walk in Q. Here, we must check that $[u_i, u_{i+1}]$, $0 \leq i \leq L-1$, is an edge of Q. By construction u_i, u_{i+1} are both vertices of the simplex Q_{l_i}. Furthermore, $u_i \neq u_{i+1}$, since either $u_i = a_1 \neq u_{i+1}$ or $u_{i+1} = a_2 \neq u_i$ or $u_{i+1} \in S_{l_i+1}$ and $u_i \notin S_{l_i+1}$. Thus, $[u_i, u_{i+1}]$ is indeed an edge of Q_i and thus of Q, as explained previously. Note by our assumption that a_1 and a_2, we indeed have $2 \leq \operatorname{diam}(Q) \leq L$.

We can now compare the diameters of P and Q as follows:

$$\begin{aligned}\operatorname{diam}(P) \geq D = l_{L-1} - l_0 &= \sum_{i=1}^{L-1}(l_i - l_{i-1}) \\ &\geq \sum_{i=1}^{L-2}(n-1) = (n-1)(L-2) \geq (n-1)(\operatorname{diam}(Q)-2),\end{aligned}$$

as needed. □

3.4.2 Lower Bounding the Diameter of $Q(A)$

For a discrete set $N \subseteq S^{n-1}$, a point $x_0 \in N$ and a positive number $\varepsilon > 0$ we denote by

$$\begin{aligned}X_k &:= X_k(N, x_0, \varepsilon) \\ &= \{\mathbf{x} \in N^k : x_i \neq x_j \text{ and } 6\varepsilon \leq \|x_i - x_{i+1}\| \leq 8\varepsilon \text{ for any } 0 \leq i < j \leq k\}\end{aligned}$$

the set of all sequences of k distinct points in N with jumps of length between 6ε and 8ε (including an extra initial jump between x_0 and x_1).

Lemma 3.4.4. *Let $\varepsilon > 0$. If $N \subseteq S^{n-1}$ is a maximal ε-separated set, then*

$$|X_k| \leq (17^{n-1})^k$$

Proof. For any $x \in N$ we find an upper bound for the number of points $y \in N$ such that $6\varepsilon \leq \|x - y\| \leq 8\varepsilon$. Recall that $C(x, r)$ denotes the closed spherical cap centered at x with radius $r > 0$. Since N is ε-separated, for any different points $y_1, y_2 \in N$ we have

$$\operatorname{int}(C(y_1, \varepsilon/2)) \cap C(y_2, \varepsilon/2) = \emptyset.$$

Taking a union of spherical caps centered at all points inside the annulus, we obtain a subset of the inflated annulus

$$C(x, 17\varepsilon/2) \setminus \operatorname{int}(C(x, 11\varepsilon/2)).$$

Since the caps $C(y, \varepsilon/2)$, $y \in N$, have pairwise disjoint interiors, the volume of their union is the sum of the volumes. Hence, the maximal number of points in the annulus is bounded by

$$\frac{\sigma(C(x, 17\varepsilon/2)) - \sigma(C(x, 11\varepsilon/2))}{\sigma(C(x, \varepsilon/2))} \leq \frac{\sigma(C(x, 17\varepsilon/2))}{\sigma(C(x, \varepsilon/2))}.$$

Using Lemma 3.2.2 we have

$$|\{y \in N : \ 6\varepsilon \leq \|x - y\| \leq 8\varepsilon\}| \leq \frac{(17/2)^{n-1}}{(1/2)^{n-1}} = 17^{n-1}.$$

Thus, the overall number of paths in X_k is bounded by

$$|X_k| \leq 17^{k(n-1)}. \qquad \square$$

Lemma 3.4.5. *Let $f\colon [0,1] \to S^{n-1}$ be a continuous function. Let $\varepsilon > 0$ and $N \subseteq S^{n-1}$ be a minimal ε-net, such that $f(0) \in N$. There exist $k \in \mathbb{N}_0$, $0 \leq t_0 < t_1 < \cdots < t_k \leq 1$ and $x_0, \ldots, x_k \in N$ such that*

1. *$\|f(t_i) - x_i\| \leq \varepsilon$ for any $i \in \{0, \ldots, k\}$,*
2. *$\|f(t) - x_i\| \geq \varepsilon$ for any $i \in \{0, \ldots, k\}$ and $t > t_i$,*
3. *$(x_1, \ldots, x_k) \in X_k(N, x_0, \varepsilon)$,*
4. *$\|x_k - f(1)\| < 7\varepsilon$.*

Proof. We build the desired couple of sequences (x_i) and (t_i) by induction. We start by taking $x_0 = f(0)$ and

$$t_0 = \sup\{t \geq 0 : \|f(t) - x_0\| \leq \varepsilon\}.$$

Note that with these choices, we have a couple of (very short) sequences for which 1-3 are fulfilled.

Assume that $x_0, \ldots, x_\ell$ and $0 \leq t_0 < \ldots < t_\ell \leq 1$ are sequences for which 1-3 hold true.

If $\|x_\ell - f(1)\| < 7\varepsilon$ then we may take $k = \ell$, and we are done.

Assume otherwise, and define

$$t' = \min\{t \in [t_\ell, 1] : \exists x_{\ell+1} \in N \text{ with } \|f(t) - x_{\ell+1}\| \leq \varepsilon \text{ and } \|x_{\ell+1} - x_\ell\| \geq 6\varepsilon\},$$

Since 4 is not fulfilled, the set is non-empty (it contains 1) and t' is well defined. We take $x_{\ell+1}$ as it appears in the definition of t'. Set

$$t_{\ell+1} = \sup\{t \in [0,1] : \|f(t) - x_{\ell+1}\| \leq \varepsilon\}.$$

By 2, for any $i \leq \ell$

$$\|f(t_{\ell+1}) - x_i\| > \varepsilon,$$

hence $x_i \neq x_{\ell+1}$. Combining this with the definition of $t_{\ell+1}$ and $x_{\ell+1}$ we only need to show that $\|x_\ell - x_{\ell+1}\| \leq 8\varepsilon$ in order to get that $0 \leq t_0 < \ldots < t_\ell < t_{\ell+1} \leq 1$ and $x_0, \ldots, x_{\ell+1}$ fulfill 1-3.

By the minimality of t', for any $s \in (t_\ell, t')$ we have $\|x_\ell - f(s)\| \leq 7\varepsilon$, otherwise there would be $x' \in N$ such that $\|x' - f(s)\| \leq \varepsilon$ but $\|x_\ell - x'\| \geq 6\varepsilon$, hence $t' \leq s$ in contradiction to the definition of s. Hence

$$\|x_\ell - x_{\ell+1}\| \leq \|x_\ell - f(s)\| + \|f(s) - f(t')\| + \|f(t') - x_{\ell+1}\| \leq 7\varepsilon + \|f(s) - f(t')\| + \varepsilon.$$

This holds for all $s \in (t_\ell, t')$. By continuity of f we may take $s \nearrow t'$ and have $\|f(s) - f(t')\| \to 0$. Thus $\|x_\ell - x_{\ell+1}\| \leq 8\varepsilon$.

Since N is finite and the points $x_0, \ldots, x_\ell$ are distinct the process must end after at most after $|N|$ steps. □

Lemma 3.4.6. *Let $A \subseteq S^{n-1}$ be a finite subset of the sphere. Let $[a_0, a_1]$, $[a_1, a_2]$, $\ldots$, $[a_{\ell-1}, a_\ell]$ be a path along the edges of $Q(A)$. There exists a continuous function $f\colon [0,1] \to S^{n-1}$ and $0 = s_0 < s_1 < \cdots < s_{\ell+1} = 1$ such that $f(0) = a_0$, $f(1) = a_\ell$, and for any $i \in \{0, 1, \ldots, \ell\}$ and any $t \in [s_i, s_{i+1}]$,*

$$a_i \in \arg\min_{a \in A}(\|f(t) - a\|).$$

Proof. First we consider the case where the path consist of a single edge, i.e. $\ell = 1$. Consider a point $x \in S^{n-1}$ and a real $r > 0$ such that the cap $C(x, r)$ contains a_0 and a_1 on its boundary and no point of A in its interior. A possible choice is given by the circumscribed cap of any facet of $Q(A)$ which contains $[a_0, a_1]$ as an edge. Now we set f such that it interpolates a_0, x and a_1 by two geodesic segments,

$$f(t) = \frac{\tilde{f}(t)}{\|\tilde{f}(t)\|}, \qquad \tilde{f}(t) = \begin{cases} (1-2t)a_0 + 2tx, & t \in [0, \frac{1}{2}], \\ (2-2t)x + (2t-1)a_1, & t \in [\frac{1}{2}, 1]. \end{cases}$$

By construction, for any $t \in [0, \frac{1}{2}]$ (resp. $t \in [\frac{1}{2}, 1]$), the cap $C(f(t), \|f(t) - a_0\|)$ (resp. $C(f(t), \|f(t) - a_1\|)$) is a subset of $C(x, r)$. Thus it contains a_0 (resp. a_1) on its boundary and no point of A in its interior. This implies that $f(0) = a_0$, $f(1) = a_1$, and

$$a_0 \in \arg\min_{a \in A}(\|f(t) - a\|), \quad t \in [0, \frac{1}{2}],$$
$$a_1 \in \arg\min_{a \in A}(\|f(t) - a\|), \quad t \in [\frac{1}{2}, 1].$$

This yields the proof in the case $\ell = 1$ (with $s_0 = 0 < s_1 = \frac{1}{2} < s_{1+1} = 1$). The general case follows by concatenating and renormalizing the functions corresponding to each edge. □

Lemma 3.4.7. *Let $A \subseteq \mathbb{S}^{n-1}$ be a finite subset of the sphere, containing two points $a_+, a_- \in A$ such that $\|a_+ - a_-\| \geq 1$. Let $\varepsilon > 0$ and N be a minimal ε-net, such that $a_+ \in N$. Set $x_0 = a_+$ and $k_0 = \lceil 1/8\varepsilon \rceil - 1$. It holds that*

$$\operatorname{diam}(Q(A)) \geq \min_{\substack{k \geq k_0 \\ \mathbf{x} \in X_k(N, x_0, \varepsilon)}} \sum_{0 \leq i \leq k-1} \mathbb{1}[C(x_i, \varepsilon/2) \cap A \neq \emptyset] \mathbb{1}[C(x_{i+1}, \varepsilon/2) \cap A \neq \emptyset].$$

Proof. The diameter of $Q(A)$ is at least the combinatorial distance between a_+ and a_-, i.e., the minimal number of edges required to form a path between these two vertices. Note that this minimum is realized for a path without loops. Let $[a_0, a_1]$, $[a_1, a_2], \ldots, [a_{\ell-1}, a_\ell]$ be such a path. Here we denote $a_0 = a_+ = x_0$ and $a_\ell = a_-$.

Consider a function f and a sequence $0 = s_0 < s_1 < \cdots < s_{\ell+1} = 1$ as in Lemma 3.4.6, and consider $k \in \mathbb{N}_0$, $0 \leq t_0 < t_1 < \cdots < t_k \leq 1$ and $x_0, \ldots, x_k \in N$ as in Lemma 3.4.5. We set $j(0) \leq j(1) \leq \cdots \leq j(k)$ such that $t_i \in [s_{j(i)}, s_{j(i)+1}]$. In particular, with this notation set up we have

$$\|x_i - x_{i+1}\| \geq 6\varepsilon, \qquad i \in \{0, \ldots, k-1\}, \tag{3.4}$$

$$\|a_{j(i)} - f(t_i)\| = \min_{a \in A} \|a - f(t_i)\|, \qquad i \in \{0, \ldots, k\}, \tag{3.5}$$

and

$$\|x_i - f(t_i)\| \leq \varepsilon, \qquad i \in \{0, \ldots, k\}. \tag{3.6}$$

From (3.6) we get $C(f(t_i), 3\varepsilon/2) \supset C(x_i, \varepsilon/2)$. Hence, if $C(x_i, \varepsilon/2) \cap A \neq \emptyset$, we have that $\|a_{j(i)} - f(t_i)\| \leq 3\varepsilon/2$ because of (3.5). Therefore if, for some $i \in \{0, \ldots, k-1\}$, both caps $C(x_i, \varepsilon/2)$ and $C(x_{i+1}, \varepsilon/2)$ contain points of A, then

$$\begin{aligned} \|a_{j(i)} - a_{j(i+1)}\| &\geq \|x_i - x_{i+1}\| - \|x_i - f(t_i)\| - \|f(t_i) - a_{j(i)}\| \\ &\qquad - \|a_{j(i+1)} - f(t_{i+1})\| - \|f(t_{i+1}) - x_{i+1}\| \\ &\geq 6\varepsilon - \varepsilon - 3\varepsilon/2 - 3\varepsilon/2 - \varepsilon = \varepsilon > 0 \end{aligned}$$

and we get $a_{j(i+1)} \neq a_{j(i)}$ which implies that $j(i) < j(i')$ for any $i' > i$. This shows that if

$$i, i' \in I = \{i : C(x_i, \varepsilon/2) \cap A \neq \emptyset \text{ and } C(x_{i+1}, \varepsilon/2) \cap A \neq \emptyset\} \subseteq \{0, 1, \ldots, k-1\},$$

with $i \neq i'$, then $a_{j(i)}$ and $a_{j(i')}$ are distinct vertices of the path. Therefore

$$\ell \geq |I| = \sum_{0 \leq i \leq k-1} \mathbb{1}[C(x_i, \varepsilon/2) \cap A \neq \emptyset] \mathbb{1}[C(x_{i+1}, \varepsilon/2) \cap A \neq \emptyset].$$

Also, we note that from

$$\begin{aligned}\|a_+ - a_-\| &\le \|a_+ - x_0\| + \sum_{1\le i\le k} \|x_i - x_{i-1}\| + \|x_k - a_-\| \\ &< \varepsilon + k \times 8\varepsilon + 7\varepsilon = 8(k+1)\varepsilon\end{aligned}$$

we have $k \ge k_0$, and therefore

$$(x_0, \dots, x_k) \in \cup_{k\ge k_0} X_k(N, x_0, \varepsilon). \qquad \square$$

Theorem 3.4.2 (Lower Bound for $Q(A)$)**.** *There exist positive constants $c_2 < 1$ and $c_3 > 1$ independent of $n \ge 3$ and m such that the following holds. Let $A = \{a_1, \dots, a_M\} \in \mathbb{S}^{n-1}$, where M is Poisson with $\mathbb{E}[M] = m$, and $a_1, \dots, a_M$ are uniformly and independently distributed in $\mathbb{S}^{n-1}$. Then, with probability at least $1 - e^{-c_3^{n-1} m^{1/(n-1)}}$, the combinatorial diameter of $Q(A)$ is at least $c_2 m^{1/(n-1)}$.*

Proof. Without loss of generality $m \ge (1/c_2)^{n-1}$ since otherwise the statement of the theorem is trivial.

In this proof the constants $1 < c_3 < c_4 < c_5 < c_6 < c_2^{-1}$ are large enough constants, independent from n and m.

We set $\varepsilon = c_6 m^{-1/(n-1)}$, and want to apply Lemma 3.4.7. Let N be an ε-net, obtained from a maximal ε-separated set, such that it contains a point a_+ from the set A. For independence properties needed later we take a_+ randomly and uniformly from the set A. With probability $1 - e^{-m/2}$ we have that A intersects the halfsphere $\{u \in \mathbb{S}^{n-1} : \langle a_+, u\rangle \le 0\}$. In which case there exists a point $a_- \in A$ such that $\|a_+ - a_-\| \ge \sqrt{2} \ge 1$. Therefore we can apply Lemma 3.4.7 with $x_0 = a_+$. Combined with the union bound, we get

$$\Pr\left(\operatorname{diam}(Q(A)) \le c_2 m^{1/(n-1)}\right) \le e^{-m/2} + \sum_{\substack{k\ge k_0 \\ \mathbf{x}\in X_k(N,x_0,\varepsilon)}} \Pr\left(\sum_{0\le i\le k-1} B_i \le c_2 m^{1/(n-1)}\right),$$

where

$$k_0 = \lceil 1/8\varepsilon \rceil + 1 \ge 1/8\varepsilon = m^{1/(n-1)}/8c_6,$$

and the summands in the probability are Bernoulli random variables

$$B_i = \mathbb{1}[C(x_i, \varepsilon/2) \cap A \ne \emptyset]\mathbb{1}[C(x_{i+1}, \varepsilon/2) \cap A \ne \emptyset].$$

For $1 \le i \le k-1$, they are identically distributed, with failure probability

$$\begin{aligned}\Pr(B_i = 0) &\le 2\Pr(C(x_i, \varepsilon/2) \cap A = 0) = 2\exp\left(-m\sigma(C(x_i, \varepsilon/2))\right) \\ &\le 2\exp\left(-m\,(\varepsilon/4)^{n-1}\right) = 2\exp\left(-\left(\frac{c_6}{4}\right)^{n-1}\right) =: 1 - p.\end{aligned}$$

In the above, note that we used Lemma 3.2.2 to lower bound the volume of the cap $\sigma(C(x_i, \varepsilon/2)) \geq (\varepsilon/4)^{n-1}\sigma(C(x_i, 2))$. Since N forms a maximal ε-separated set and the x_i are distinct, the caps $C(x_i, \varepsilon/2)$ are disjoint and therefore the random variables $B_1, B_3, B_5, \ldots$ are independent. Next we exploit this independence. Let $k \geq k_0$, and set $K = \lfloor k/2 \rfloor$. Note that $K \geq 1/16\varepsilon = m^{1/(n-1)}/16c_6$. Assuming that $c_2 \leq 1/32c_6$, we have

$$\Pr\left(\sum_{0\leq i\leq k-1} B_i \leq c_2 m^{1/(n-1)}\right) \leq \Pr\left(\sum_{1\leq i\leq K} B_{2i-1} \leq \frac{K}{2}\right) = \sum_{1\leq i\leq \lfloor K/2\rfloor} \binom{K}{i} p^i (1-p)^{K-i}.$$

Now we bound p by 1, $(1-p)^{K-i}$ by $(1-p)^{K/2}$ and $\sum \binom{K}{i}$ by 2^K, which provides us the bound

$$\begin{aligned}\Pr\left(\sum_{0\leq i\leq k-1} B_i \leq c_2 m^{-1/(n-1)}\right) &\leq (2(1-p)^{1/2})^K \\ &= \left(e^{\left(-\frac{1}{2}\left(\frac{c_6}{4}\right)^{n-1}+\frac{3}{2}\ln 2\right)}\right)^K \leq \left(e^{\left(-c_5^{n-1}\right)}\right)^K.\end{aligned}$$

Thus, with the bound $|X_k| \leq (17^{n-1})^k$ from Lemma 3.4.4, and the fact that $K \geq k/2$, we get

$$\begin{aligned}\Pr\left(\operatorname{diam}(Q(A)) \leq c_2 m^{-1/(n-1)}\right) &\leq e^{-m/2} + \sum_{k\geq k_0}\left(e^{\left(-\frac{1}{2}(c_5)^{n-1}+(n-1)\ln 17\right)}\right)^k \\ &\leq e^{-m/2} + \sum_{k\geq k_0}(e^{-(c_4)^{n-1}})^k \\ &= e^{-m/2} + \frac{e^{-k_0 c_4^{n-1}}}{1-e^{-(c_4)^{n-1}}} \\ &\leq e^{-m/2} + \frac{e^{-\frac{m^{1/(n-1)}}{8c_6}c_4^{n-1}}}{1-e^{-c_4^{n-1}}} \\ &\leq e^{-c_3^{n-1}m^{1/(n-1)}}.\end{aligned}$$

□

Chapter 4

A Scaling-Invariant Algorithm for Linear Programming whose Running Time Depends Only on the Constraint Matrix

Following the breakthrough work of Tardos [183] in the bit-complexity model, Vavasis and Ye [198] gave the first exact algorithm for linear programming in the real model of computation with running time depending only on the constraint matrix. For solving a linear program $\max c^{\mathsf{T}}x,\ Ax = b,\ x \geq \vec{0},\ A \in \mathbb{R}^{m\times n}$, Vavasis and Ye developed a primal-dual interior point method using a *'layered least squares'* (LLS) step, and showed that $O(n^{3.5}\log(\bar{\chi}_A + n))$ iterations suffice to solve the LP exactly, where $\bar{\chi}_A$ is a condition measure controlling the size of solutions to linear systems related to A.

Monteiro and Tsuchiya [146], noting that the central path is invariant under rescalings of the columns of A and c, asked whether there exists an LP algorithm depending instead on the measure $\bar{\chi}^*_A$, defined as the minimum $\bar{\chi}_{AD}$ value achievable by a column rescaling AD of A, and gave strong evidence that this should be the case. We resolve this open question affirmatively.

Our first main contribution is an $O(m^2n^2 + n^3)$ time algorithm which works on the linear matroid of A to compute a nearly optimal diagonal rescaling D satisfying $\bar{\chi}_{AD} \leq n(\bar{\chi}^*_A)^3$. This algorithm also allows us to approximate the value of $\bar{\chi}_A$ up to a factor $n(\bar{\chi}^*_A)^2$. This result is in (surprising) contrast to that of Tunçel [188], who showed NP-hardness for approximating $\bar{\chi}_A$ to within $2^{\mathrm{poly}(\mathrm{rank}(A))}$. The key insight for our algorithm is to work with ratios g_i/g_j of circuits of A — i.e., minimal linear dependencies $Ag = \vec{0}$ — which allow us to approximate the value of $\bar{\chi}^*_A$ by a maximum geometric mean cycle computation in what we call the *'circuit ratio digraph'* of A.

While this resolves Monteiro and Tsuchiya's question by appropriate preprocessing, it falls short of providing either a truly scaling invariant algorithm or an improvement upon the base LLS analysis. In this vein, as our second main contribution we develop a *scaling invariant* LLS algorithm, which uses and dynamically maintains improving estimates of the circuit ratio digraph, together with a refined potential function based analysis for LLS algorithms in general. With this analysis, we derive

This chapter is based on [52], a joint work with Daniel Dadush, Bento Natura, and László A Végh.

an improved $O(n^{2.5} \log n \log(\bar{\chi}_A^* + n))$ iteration bound for optimally solving the LP using our algorithm. The same argument also yields a factor $n/\log n$ improvement on the iteration complexity bound of the original Vavasis-Ye algorithm.

4.1 Introduction

The linear programming problem in primal-dual form is to solve

$$\begin{array}{rr} \min c^\mathsf{T} x & \max y^\mathsf{T} b \\ Ax = b & A^\mathsf{T} y + s = c \\ x \geq \vec{0}, & s \geq \vec{0}, \end{array} \tag{4.1}$$

where $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) = m \leq n$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ are given in the input, and $x, s \in \mathbb{R}^n$, $y \in \mathbb{R}^m$ are the variables. We consider the program in x to be the primal problem and the program in (y, s) to be the dual problem.

Khachiyan [122] used the ellipsoid method to give the first polynomial time LP algorithm in the bit-complexity model, that is, polynomial in the bit description length of (A, b, c). An outstanding open question is the existence of a *strongly polynomial* algorithm for LP, listed by Smale as one of the most prominent mathematical challenges for the 21st century [174]. Such an algorithm amounts to solving LP using $\text{poly}(n, m)$ basic arithmetic operations in the real model of computation.[1] Known strongly polynomially solvable LP problems classes include: feasibility for two variable per inequality systems [139], the minimum-cost circulation problem [182], the maximum generalized flow problem [155, 199], and discounted Markov decision problems [203, 205].

Towards this goal, the principal line of attack has been to develop LP algorithms whose running time is bounded in terms of natural *condition measures*. Such condition measures attempt to finely measure the "intrinsic complexity" of the LP. An important line of work in this area has been to parametrize LPs by the "niceness" of their solutions (e.g. the depth of the most interior point), where relevant examples include the Goffin measure [91] for conic systems and Renegar's distance to ill-posedness for general LPs [161, 162], and bounded ratios between the nonzero entries in basic feasible solutions [43, 123].

Parametrizing by the constraint matrix A second line of research, and the main focus of this chapter, focuses on the complexity of the constraint matrix A. The first

[1] In the bit-complexity model, a further requirement is that the algorithm must be in PSPACE.

breakthrough in this area was given by Tardos [183], who showed that if A has integer entries and all square submatrices of A have determinant at most Δ in absolute value, then (4.1) can be solved in $\text{poly}(n, m, \log \Delta)$ arithmetic operations, independent of the encoding length of the vectors b and c. This is achieved by finding the exact solutions to $O(nm)$ rounded LPs derived from the original LP, with the right hand side vector and cost function being integers of absolute value bounded in terms of n and Δ. From m such rounded problem instances, one can infer, via proximity results, that a constraint $x_i = 0$ must be valid for every optimal solution. The process continues by induction until the optimal primal face is identified.

Path-following methods and the Vavasis–Ye algorithm In a seminal work, Vavasis and Ye [198] introduced a new type of interior-point method that optimally solves (4.1) within $O(n^{3.5} \log(\bar{\chi}_A + n))$ iterations, where the condition number $\bar{\chi}_A$ controls the size of solutions to certain linear systems related to the kernel of A (see Section 4.2 for the formal definition).

Before detailing the Vavasis–Ye (henceforth VY) algorithm, we recall the basics of path following interior-point methods. If both the primal and dual problems in (4.1) are strictly feasible, the *central path* for (4.1) is the curve $((x(\mu), y(\mu), s(\mu)) : \mu > 0)$ defined by

$$\begin{aligned} x(\mu)_i s(\mu)_i &= \mu, \quad \forall i \in [n] \\ Ax(\mu) &= b,\ x(\mu) > \vec{0}, \\ A^\mathsf{T} y(\mu) + s(\mu) &= c,\ s(\mu) > \vec{0}, \end{aligned} \tag{CP}$$

which converges to complementary optimal primal and dual solutions (x^*, y^*, s^*) as $\mu \to 0$, recalling that the optimality gap at time μ is exactly $x(\mu)^\mathsf{T} s(\mu) = n\mu$. We thus refer to μ as the normalized duality gap. Methods that "follow the path" generate iterates that stay in a certain neighborhood around it while trying to achieve rapid multiplicative progress w.r.t. to μ, where given (x, y, s) 'close' to the path, we define the effective μ as $\mu(x, y, s) = \sum_{i=1}^n x_i s_i / n$. Given a target parameter μ' and starting point close to the path at parameter μ, standard path following methods [100] can compute a point at parameter below μ' in at most $O(\sqrt{n} \log(\mu/\mu'))$ iterations, and hence the quantity $\log(\mu/\mu')$ can be usefully interpreted as the length of the corresponding segment of the central path.

Crossover events and layered least squares steps At a very high level, Vavasis and Ye show that the central path can be decomposed into at most $\binom{n}{2}$ short but curved segments, possibly joined by long (apriori unbounded) but very straight segments. At the end of each curved segment, they show that a new ordering relation $x_i(\mu) > x_j(\mu)$—called a *'crossover event'*—is implicitly learned, where this relation did not

hold at the start of the segment, but will hold at every point from the end of the segment onwards. These $\binom{n}{2}$ relations give a combinatorial way to measure progress along the central path. In contrast to Tardos's algorithm, where the main progress is setting variables to zero explicitly, the variables participating in crossover events cannot be identified, only their existence is shown.

At a technical level, the VY-algorithm is a variant of the Mizuno–Todd–Ye [145] predictor-corrector method (MTY P-C). In predictor-corrector methods, corrector steps bring an iterate closer to the path, i.e., improve centrality, and predictor steps "shoot down" the path, i.e., reduce μ without losing too much centrality. Vavasis and Ye's main algorithmic innovation was the introduction of a new predictor step, called the *'layered least squares'* (LLS) step, which crucially allowed them to cross each aforementioned "straight" segment of the central path in *a single step*, recalling that these straight segments may be arbitrarily long. To traverse the short and curved segments of the path, the standard predictor step, known as *affine scaling* (AS), in fact suffices.

To compute the LLS direction, the variables are decomposed into 'layers' $J_1 \cup J_2 \cup \ldots \cup J_p = [n]$. The goal of such a decomposition is to eventually learn a refinement of the optimal partition of the variables $B^* \cup N^* = [n]$, where $B^* := \{i \in [n] : x_i^* > 0\}$ and $N^* := \{i \in [n] : s_i^* > 0\}$ for the limit optimal solution (x^*, y^*, s^*).

The primal affine scaling direction can be equivalently described by solving a weighted least squares problem in $\mathrm{Ker}(A)$, with respect to a weighting defined according to the current iterate. The primal LLS direction is obtained by solving a series of weighted least squares problems, starting with focusing only on the final layer J_p. This solution is gradually extended to the higher layers (i.e., layers with lower indices). The dual directions have analogous interpretations, with the solutions on the layers obtained in the opposite direction, starting with J_1. If we use the two-level layering $J_1 = B^*$, $J_2 = N^*$, and are sufficiently close to the limit (x^*, y^*, s^*) of the central path, then the LLS step reaches an exact optimal solution in a single step. We note that standard AS steps generically never find an exact optimal solution, and thus some form of "LLS rounding" in the final iteration is always necessary to achieve finite termination with an exact optimal solution.

Of course, guessing B^* and N^* correctly is just as hard as solving (4.1). Still, if we work with a "good" layerings, these will reveal new information about the "optimal order" of the variables, where B^* is placed on higher layers than N^*. The crossover events correspond to swapping two wrongly ordered variables into the correct ordering. Namely, a variable $i \in B^*$ and $j \in N^*$ are currently ordered on the same layer, or j is in a higher layer than i. After the crossover event, i will always be placed on a higher layer than j.

Computing good layerings and the $\bar{\chi}_A$ condition measure Given the above discussion, the obvious question is how to come up with "good" layerings? The philosophy behind LLS can be stated as saying that if modifying a set of variables x_I barely affects the variables in $x_{[n]\setminus I}$ (recalling that movement is constrained to $\Delta x \in \mathrm{Ker}(A)$), then one should optimize over x_I without regard to the effect on $x_{[n]\setminus I}$; hence x_I should be placed on lower layers.

VY's strategy for computing such layerings was to directly use the size of the coordinates of the current iterate x (where (x, y, s) is a point near the central path). In particular, assuming $x_1 \geq x_2 \geq \ldots \geq x_n$, the layering $J_1 \cup J_2 \cup \ldots \cup J_p = [n]$ corresponds to consecutive intervals constructed in decreasing order of x_i values. The break between J_i and J_{i+1} occurs if the gap $x_r/x_{r+1} > g$, where r is the rightmost element of J_i and $g > 0$ is a threshold parameter. Thus, the expectation is that if $x_i > gx_j$, then a small multiplicative change to x_j, subject to moving in $\mathrm{Ker}(A)$, should induce a small multiplicative change to x_i. By proximity to the central path, the dual ordering is reversed as mentioned above.

The threshold g for which this was justified in VY was a function of the $\bar{\chi}_A$ condition measure. We now provide a convenient definition, which immediately yields this justification (see Proposition 4.2.4). Letting $W = \mathrm{Ker}(A)$ and $\pi_I(W) = \{x_I : x \in W\}$, we define $\bar{\chi}_A := \bar{\chi}_W$ as the minimum number $M \geq 1$ such that for any $\emptyset \neq I \subseteq [n]$ and $z \in \pi_I(W)$, there exists $y \in W$ with $y_I = z$ and $\|y\| \leq M\|z\|$. Thus, a change of ε in variables in I can be lifted to a change of at most $\bar{\chi}_A\varepsilon$ in variables in $[n] \setminus I$. Crucially, $\bar{\chi}$ is a "self-dual" quantity. That is, $\bar{\chi}_W = \bar{\chi}_{W^\perp}$, where $W^\perp = \mathrm{range}(A^\mathsf{T})$ is the movement subspace for the dual problem, justifying the reversed layering for the dual (see Section 4.2 for more details).

The question of scale invariance and $\bar{\chi}^*_A$ While the VY layering procedure is powerful, its properties are somewhat mismatched with those of the central path. In particular, variable ordering information has *no intrinsic meaning* on the central path, as the path itself is *scaling invariant*. Namely, the central path point $(x(\mu), y(\mu), s(\mu))$ w.r.t. the problem instance (A, b, c) is in bijective correspondence with the central path point $(D^{-1}x(\mu), Dy(\mu), Ds(\mu)))$ w.r.t. the problem instance (AD, Dc, b) for any positive diagonal matrix D. The standard path following algorithms are also scaling invariant in this sense.

This lead Monteiro and Tsuchiya [146] to ask whether a scaling invariant LLS algorithm exists. They noted that any such algorithm would then depend on the potentially much smaller parameter

$$\bar{\chi}^*_A := \inf_D \bar{\chi}_{AD}\,, \tag{4.2}$$

where the infimum is taken over the set of $n \times n$ positive diagonal matrices. Thus,

Monteiro and Tsuchiya's question can be rephrased as to whether there exists an exact LP algorithm with running time $\text{poly}(n, m, \log \bar{\chi}_A^*)$.

Substantial progress on this question was made in the followup works [129, 147]. The paper [147] showed that the number of iterations of the MTY predictor-corrector algorithm [145] can get from $\mu_0 > 0$ to $\eta > 0$ on the central path in

$$O\left(n^{3.5} \log \bar{\chi}_A^* + \min\{n^2 \log\log(\mu^0/\eta), \log(\mu^0/\eta)\}\right)$$

iterations. This is attained by showing that the standard AS steps are reasonably close to the LLS steps. This proximity can be used to show that the AS steps can traverse the "curved" parts of the central path in the same iteration complexity bound as the VY algorithm. Moreover, on the "straight" parts of the path, the rate of progress amplifies geometrically, thus attaining a log log convergence on these parts. Subsequently, [129] developed an affine invariant *trust region* step, which traverses the full path in $O(n^{3.5} \log(\bar{\chi}_A^* + n))$ iterations. However, the running time of each iteration is weakly polynomial in b and c. The question of developing an LP algorithm with complexity bound $\text{poly}(n, m, \log \bar{\chi}_A^*)$ thus remained open.

A related open problem to the above is whether it is possible to compute a near-optimal rescaling D for program (4.2)? This would give an alternate pathway to the desired LP algorithm by simply preprocessing the matrix A. The related question of approximating $\bar{\chi}_A$ was already studied by Tunçel [188], who showed NP-hardness for approximating $\bar{\chi}_A$ to within a $2^{\text{poly}(\text{rank}(A))}$ factor. Taken at face value, this may seem to suggest that approximating the rescaling D should be hard.

A further open question is whether Vavasis and Ye's cross-over analysis can be improved. Ye in [204] showed that the iteration complexity can be reduced to $O(n^{2.5} \log(\bar{\chi}_A + n))$ for feasibility problems and further to $O(n^{1.5} \log(\bar{\chi}_A + n))$ for homogeneous systems, though the $O(n^{3.5} \log(\bar{\chi}_A + n))$ bound for optimization has not been improved since [198].

4.1.1 Our contributions

In this work, we resolve all of the above questions in the affirmative. We detail our contributions below.

1. Finding an approximately optimal rescaling. As our first contribution, we give an $O(m^2n^2 + n^3)$ time algorithm that works on the linear matroid of A to compute a diagonal rescaling matrix D which achieves $\bar{\chi}_{AD} \leq n(\bar{\chi}_A^*)^3$, given any $m \times n$ matrix A. Furthermore, this same algorithm allows us to approximate $\bar{\chi}_A$ to within a factor $n(\bar{\chi}_A^*)^2$. The algorithm bypasses Tunçel's hardness result by allowing the approximation factor to depend on A itself, namely on $\bar{\chi}_A^*$. This gives a simple first answer to Monteiro and Tsuchiya's question: by applying the Vavasis-Ye algorithm

directly on the preprocessed A matrix, we may solve any LP with constraint matrix A using $O(n^{3.5}\log(\bar{\chi}_A^* + n))$ iterations. Note that the approximation factor $n(\bar{\chi}_A^*)^2$ increases the runtime only by a constant factor.

To achieve this result, we work with the circuits of A, where a circuit $C \subseteq [n]$ corresponds to an inclusion-wise minimal set of linearly dependent columns. With each circuit, we can associate a vector $g^C \in \mathrm{Ker}(A)$ with $\mathrm{supp}(g^C) = C$ that is unique up to scaling. By the *'circuit ratio'* κ_{ij} associated with the pair of nodes (i, j), we mean the largest ratio $|g_j^C/g_i^C|$ taken over every circuit C of A such that $i, j \in C$. As our first observation, we show that the maximum of all circuit ratios, which we call the *'circuit imbalance measure'*, in fact characterizes $\bar{\chi}_A$ up to a factor n. This measure was first studied by Vavasis [197], who showed that it lower bounds $\bar{\chi}_A$, though, as far as we are aware, our upper bound is new. The circuit ratios of each pair (i, j) induce a weighted directed graph we call the *'circuit ratio digraph'* of A. From here, our main result is that $\bar{\chi}_A^*$ is up to a factor n equal to the maximum geometric mean cycle in the circuit ratio digraph. Our algorithm populates the circuit ratio digraph with approximations of the κ_{ij} ratios for each $i, j \in [n]$ using standard techniques from matroid theory, and then computes a rescaling by solving the dual of the maximum geometric mean ratio cycle on the *'approximate circuit ratio digraph'*.

2. Scaling invariant LLS algorithm. While the above yields an LP algorithm with $\mathrm{poly}(n, m, \log \bar{\chi}_A^*)$ running time, it does not satisfactorily address Monteiro and Tsuchiya's question for a scaling invariant algorithm. As our second contribution, we use the circuit ratio digraph directly to give a natural scaling invariant LLS layering algorithm together with a scaling invariant crossover analysis.

At a conceptual level, we show that the circuit ratios give a scale invariant way to measure whether '$x_i > x_j$' and enable a natural layering algorithm. Assume for now that the circuit imbalance value κ_{ij} is known for every pair (i, j). Given the circuit ratio graph induced by the κ_{ij}'s and given a primal point x near the path, our layering algorithm can be described as follows. We first rescale the variables so that x becomes the all ones vector, which rescales κ_{ij} to $\kappa_{ij}x_i/x_j$. We then restrict the graph to its edges of length $\kappa_{ij}x_i/x_j \geq 1/\mathrm{poly}(n)$—the *long edges* of the (rescaled) circuit ratio graph—and let the layering $J_1 \cup J_2 \cup \ldots \cup J_p$ be a topological ordering of its strongly connected components (SCC) with edges going from left to right. Intuitively, variables that "affect each other" should be in the same layer, which motivates the SCC definition.

We note that our layering algorithm does not in fact have access to the true circuit ratios κ_{ij}, as these are NP-hard to compute. Getting a good enough initial estimate for our purposes however is easy: we let $\hat{\kappa}_{ij}$ be the ratio corresponding to an *arbitrary* circuit containing i and j. This already turns out to be within a factor $(\bar{\chi}_A^*)^2$ from the true value κ_{ij}, which we recall is the maximum over all such circuits. Our layering

algorithm in fact learns better circuit ratio estimates if the "lifting costs" of our SCC layering, i.e., how much it costs to lift changes from lower layer variables to higher layers (as in the definition of $\bar{\chi}_A$), are larger than we expected them to be based on the previous estimates.

We develop a scaling-invariant analogue of cross-over events as follows. Before the crossover event, $\text{poly}(n)(\bar{\chi}_A^*)^n > \kappa_{ij} x_i/x_j$, and after the crossover event, $\text{poly}(n)(\bar{\chi}_A^*)^n < \kappa_{ij} x_i/x_j$ for all further central path points. Our analysis relies on $\bar{\chi}_A^*$ in only a minimalistic way, and does not require an estimate on the value of $\bar{\chi}_A^*$. Namely, it is only used to show that if $i, j \in J_q$, for a layer $q \in [p]$, then the rescaled circuit ratio $\kappa_{ij} x_i/x_j$ is in the range $(\text{poly}(n)\bar{\chi}_A^*)^{O(\pm|J_q|)}$. The argument to show this crucially utilizes the maximum geometric mean cycle characterization. Furthermore, unlike prior analyses [146, 198], our definition of a "good" layering (i.e., 'balanced' layerings, see Section 4.3.5), is completely independent of $\bar{\chi}_A^*$.

3. Improved potential analysis. As our third contribution, we improve the Vavasis–Ye crossover analysis using a new and simple potential function based approach. When applied to our new LLS algorithm, we derive an $O(n^{2.5} \log n \log(\bar{\chi}_A^* + n))$ iteration bound for path following, improving the polynomial term by an $\Omega(n/\log n)$ factor compared to the VY analysis.

Our potential function can be seen as a fine-grained version of the crossover events as described above. In case of such a crossover event, it is guaranteed that in every subsequent iteration, i is in a layer before j. We analyze less radical changes instead: an "event" parametrized by τ means that i and j are currently together on a layer of size $\leq \tau$, and after the event, i is on a layer before j, or if they are together on the same layer, then this layer must have size $\geq 2\tau$. For every LLS step, we can find a parameter τ such that an event of this type happens concurrently for at least $\tau - 1$ pairs within the next $O(\sqrt{n\tau} \log(\bar{\chi}_A^* + n))$ iterations,

Our improved analysis is also applicable to the original VY-algorithm. Let us now comment on the relation between the VY-algorithm and our new algorithm. The VY-algorithm starts a new layer once $x_{\pi(i)} > g x_{\pi(i+1)}$ between two consecutive variables where the permutation π is a non-increasing order of the x_i variables, and $g = \text{poly}(n)\bar{\chi}$. Setting the initial 'estimates' $\hat{\kappa}_{ij} = \bar{\chi}$ for a suitable polynomial, our algorithm runs the same way as the VY algorithm. Using these estimates, the layering procedure becomes much simpler: there is no need to verify 'balancedness' as in our algorithm.

However, using estimates $\hat{\kappa}_{ij} = \bar{\chi}$ has drawbacks. Most importantly, it does not give a lower bound on the true circuit ratio κ_{ij}—to the contrary, g will be an upper bound. In effect, this causes VY's layers to be "much larger" than ours, and for this reason, the connection to $\bar{\chi}_A^*$ is lost. Nevertheless, our potential function analysis can still be adapted to the VY-algorithm to obtain the same $\Omega(n/\log n)$ improvement on

the iteration complexity bound; see Section 4.4.1 for more details.

4.1.2 Related work

Since the seminal works of Karmarkar [119] and Renegar [160], there has been a tremendous amount of work on speeding up and improving interior-point methods. In contrast to the present work, the focus of these works has mostly been to improve complexity of *approximately solving* LPs. Progress has taken many forms, such as the development of novel barrier methods, such as Vaidya's volumetric barrier [190] and the recent entropic barrier of Bubeck and Eldan [37] and the weighted log-barrier of Lee and Sidford [132], together with new path following techniques, such as the predictor-corrector framework [143, 145], as well as advances in fast linear system solving [131, 178]. There has been substantial progress in improving IPM by amortizing the cost of the iterative updates, and working with approximate computations, see e.g. [160, 190] for classical results. Recently, Cohen, Lee and Song [45] developed a new inverse maintenance scheme to get a randomized $\tilde{O}(n^{\omega}\log(1/\varepsilon))$-time algorithm for ε-approximate LP, which was derandomized by van den Brand [192]; here $\omega \approx 2.37$ is the matrix multiplication exponent. A very recent result by van den Brand et al. [195] obtained a randomized $\tilde{O}((nm + m^3)\log(1/\varepsilon))$ algorithm. For special classes of LP such as network flow and matching problems, even faster algorithms have been obtained using, among other techniques, fast Laplacian solvers, see e.g. [55, 135, 193, 194]. Given the progress above, we believe it to be an interesting problem to understand to what extent these new numerical techniques can be applied to speed up LLS computations, though we expect that such computations will require very high precision. We note that no attempt has been made in the present work to optimize the complexity of the linear algebra.

Subsequent to this work, [53] extended Tardos's framework to the real model of computation, showing that $\mathrm{poly}(n, m, \log \bar{\chi}_A)$ running time can be achieved using approximate solvers in a black box manner. Combined with [192], one obtains a deterministic $O(mn^{\omega+1}\log(n)\log(\bar{\chi}_A + n))$ LP algorithm; using the initial rescaling subroutine from this chapter, the dependence can be improved to $\bar{\chi}_A^*$; together with the preprocessing time, this amounts to $O(m^2n^2 + mn^{\omega+1}\log(n)\log(\bar{\chi}_A^* + n))$. A weaker extension of Tardos's framework to the real model of computation was previously given by Ho and Tunçel [107].

With regard to LLS algorithms, the original VY-algorithm required explicit knowledge of $\bar{\chi}_A$ to implement their layering algorithm. The paper [142] showed that this could be avoided by computing all LLS steps associated with n candidate partitions and picking the best one. In particular, they showed that all such LLS steps can be computed in $O(m^2n)$ time. In [146], an alternate approach was presented to compute an LLS partition directly from the coefficients of the AS step. We note that these

methods crucially rely on the variable ordering, and hence are not scaling invariant. Kitahara and Tsuchiya [124], gave a 2-layer LLS step which achieves a running time depending only on $\bar{\chi}_A^*$ and right-hand side b, but with no dependence on the objective, assuming the primal feasible region is bounded.

A series of papers have studied the central path from a differential geometry perspective. Monteiro and Tsuchiya [148] showed that a curvature integral of the central path, first introduced by Sonnevend, Stoer, and Zhao [176], is in fact upper bounded by $O(n^{3.5}\log(\bar{\chi}_A^*+n))$. This has been extended to SDP and symmetric cone programming [116], and also studied in the context of information geometry [115].

Circuits have appeared in several papers on linear and integer optimization (see [63] and references within). The idea of using circuits within the context of LP algorithms also appears in [62]. They develop an augmentation framework for LP (as well ILP) and show that a simplex-like algorithm which takes steps according to the "best circuit" direction achieves linear convergence, though these steps are hard to compute.

Our algorithm makes progress towards strongly polynomial solvability of LP, by improving the dependence $\text{poly}(n, m, \log \bar{\chi}_A)$ to $\text{poly}(n, m, \log \bar{\chi}_A^*)$. However, in a remarkable recent paper, Allamigeon et al. [5] have shown, using tools from tropical geometry, that path-following methods for the standard logarithmic barrier *cannot* be strongly polynomial. In particular, they give a parametrized family of instances, where, for sufficiently large parameter values, any sequence of iterations following the central path must be of exponential length—thus, $\bar{\chi}_A^*$ will be doubly exponential.

4.1.3 Organization

The rest of the chapter is organized as follows. We conclude this section by introducing some notation. Section 4.2 discusses our results on the circuit imbalance measure. It starts with Section 4.2.1 on the necessary background on the condition measures $\bar{\chi}_A$ and $\bar{\chi}_A^*$. Section 4.2.2 introduces the circuit imbalance measure, and formulates and explains all main results of Section 4.2. The proofs are given in the rest of the sections: basic properties in Section 4.2.3, the min-max characterization in Section 4.2.4, the circuit finding algorithm in Section 4.2.5, the algorithms for approximating $\bar{\chi}_A^*$ and $\bar{\chi}_A$ in Section 4.2.6.

In Section 4.3, we develop our scaling invariant interior-point method. Interior-point preliminaries are given in Section 4.3.1. Section 4.3.2 introduces the affine scaling and layered-least-squares directions, and proves some basic properties. Section 4.3.3 provides a detailed overview of the high level ideas and a roadmap to the analysis. Section 4.3.4 further develops the theory of LLS directions and introduces partition lifting scores. Section 4.3.5 gives our scaling invariant layering procedure, and our overall algorithm is given in Section 4.3.6.

In Section 4.4, we give the potential function proof for the improved iteration bound, relying on technical lemmas. The full proof of these lemmas is deferred to Section 4.6; however, Section 4.4 provides the high-level ideas to each proof. Section 4.4.1 shows that our argument also leads to a factor $\Omega(n/\log n)$ improvement in the iteration complexity bound of the VY-algorithm.

In Section 4.5, we prove the technical properties of our LLS step, including its proximity to AS and step length estimates. Finally, in Section 4.7, we discuss the initialization of our interior-point method.

Besides reading linearly, we suggest two other possible strategies for navigating this chapter. Readers mainly interested in the circuit imbalance measure and its approximation may focus only on Section 4.2; this part can be understood without any familiarity with interior point methods. Other readers, who wish to mainly focus on our interior point algorithm may read Section 4.2 only up to Section 4.2.2; this includes all concepts and statements necessary for the algorithm.

4.1.4 Notation

Here we specify some general notation that we will use in this chapter. See also Section 1.5.

For a vector $x \in \mathbb{R}^n$, we let $\mathrm{Diag}(x) \in \mathbb{R}^{n\times n}$ denote the diagonal matrix with x on the diagonal. We let $\mathbf{D}_n$ denote the set of all positive $n \times n$ diagonal matrices, dropping the subscript if the dimension is clear. For $x, y \in \mathbb{R}^n$, we use the notation $xy \in \mathbb{R}^n$ to denote $xy = \mathrm{Diag}(x)y = (x_i y_i)_{i\in[n]}$. The inner product of the two vectors is denoted as $x^\mathsf{T} y$. For $p \in \mathbb{Q}$, we also use the notation x^p to denote the vector $(x_i^p)_{i\in[n]}$. Similarly, for $x, y \in \mathbb{R}^n$, we let x/y denote the vector $(x_i/y_i)_{i\in[n]}$. We denote the support of a vector $x \in \mathbb{R}^n$ by $\mathrm{supp}(x) = \{i \in [n] : x_i \neq 0\}$.

For an index subset $I \subseteq [n]$, we use $\pi_I : \mathbb{R}^n \to \mathbb{R}^I$ for the coordinate projection. That is, $\pi_I(x) = x_I$, and for a subset $S \subseteq \mathbb{R}^n$, $\pi_I(S) = \{x_I : x \in S\}$. We let $\mathbb{R}^n_I = \{x \in \mathbb{R}^n : x_{[n]\setminus I} = \vec{0}\}$.

For a matrix $B \in \mathbb{R}^{n\times k}$, $I \subseteq [n]$ and $J \subseteq [k]$ we let $B_{I,J}$ denote the submatrix of B restricted to the set of rows in I and columns in J. We also use $B_{I,\bullet} = B_{I,[k]}$ and $B_J = B_{\bullet,J} = B_{[n],J}$. We let $B^\dagger \in \mathbb{R}^{k\times n}$ denote the pseudo-inverse of B.

We let $\mathrm{Ker}(A)$ denote the kernel of the matrix $A \subseteq \mathbb{R}^{m\times n}$. Throughout, we assume that the matrix A in (4.1) has full row rank, and that $n \geq 3$.

Subspace formulation Throughout the chapter, we let $W = \mathrm{Ker}(A) \subseteq \mathbb{R}^n$ denote the kernel of the matrix A. Using this notation, (4.1) can be written in the form

$$\begin{array}{ll} \min c^\mathsf{T} x & \max d^\mathsf{T}(c-s) \\ x \in W + d & s \in W^\perp + c \\ x \geq \vec{0}, & s \geq \vec{0}, \end{array} \tag{4.3}$$

where $d \in \mathbb{R}^n$ satisfies $Ad = b$.

4.2 Finding an approximately optimal rescaling

4.2.1 The condition number $\bar{\chi}$

The condition number $\bar{\chi}_A$ is defined as

$$\begin{aligned} \bar{\chi}_A &= \sup\left\{ \|A^\mathsf{T}\left(ADA^\mathsf{T}\right)^{-1} AD\| \,:\, D \in \mathbf{D} \right\} \\ &= \sup\left\{ \frac{\|A^\mathsf{T} y\|}{\|p\|} \,:\, y \text{ minimizes } \|D^{1/2}(A^\mathsf{T} y - p)\| \text{ for some } \vec{0} \neq p \in \mathbb{R}^n, D \in \mathbf{D} \right\}. \end{aligned} \tag{4.4}$$

This condition number was first studied by Dikin [71], Stewart [180], and Todd [185], among others, and plays a key role in the analysis of the Vavasis–Ye interior point method [198]. There is an extensive literature on the properties and applications of $\bar{\chi}_A$, as well as its relations to other condition numbers. We refer the reader to the papers [107, 146, 198] for further results and references.

It is important to note that $\bar{\chi}_A$ only depends on the subspace $W = \mathrm{Ker}(A)$. Hence, we can also write $\bar{\chi}_W$ for a subspace $W \subseteq \mathbb{R}^n$, defined to be equal to $\bar{\chi}_A$ for some matrix $A \in \mathbb{R}^{k\times n}$ with $W = \mathrm{Ker}(A)$. We will use the notations $\bar{\chi}_A$ and $\bar{\chi}_W$ interchangeably.

The next lemma summarizes some important known properties of $\bar{\chi}_A$.

Proposition 4.2.1. *Let $A \in \mathbb{R}^{m\times n}$ with full row rank and $W = \mathrm{Ker}(A)$.*

(i) *If the entries of A are all integers, then $\bar{\chi}_A$ is bounded by $2^{O(L_A)}$, where L_A is the input bit length of A.*

(ii) $\bar{\chi}_A = \max\{\|B^{-1}A\| : B$ *non-singular* $m \times m$*-submatrix of* $A\}$.

(iii) *Let the columns of $B \in \mathbb{R}^{n\times(n-m)}$ form an orthonormal basis of W. Then*

$$\bar{\chi}_W = \max\left\{ \|BB^\dagger_{I,\bullet}\| : \emptyset \neq I \subseteq [n] \right\}.$$

(iv) $\bar{\chi}_W = \bar{\chi}_{W^\perp}$.

Proof. Part (i) was proved in [198, Lemma 24]. For part (ii), see [187, Theorem 1] and [198, Lemma 3]. In part (iii), the direction $\geq$ was proved in [180], and the direction $\leq$ in [154]. The duality statement (iv) was shown in [101]. □

In Proposition 4.3.8, we will also give another proof of (iv). We now define the lifting map, a key operation in this chapter, and explain its connection to $\bar{\chi}_A$.

Definition 4.2.2. Let us define the *lifting map* $L_I^W : \pi_I(W) \to W$ by

$$L_I^W(p) = \arg\min \{\|z\| : z_I = p, z \in W\}.$$

Note that L_I^W is the unique linear map from $\pi_I(W)$ to W such that $\left(L_I^W(p)\right)_I = p$ and $L_I^W(p)$ is orthogonal to $W \cap \mathbb{R}^n_{[n]\setminus I}$.

Lemma 4.2.3. *Let $W \subseteq \mathbb{R}^n$ be an $(n-m)$-dimensional linear subspace. Let the columns of $B \in \mathbb{R}^{n\times(n-m)}$ denote an orthonormal basis of W. Then, viewing L_I^W as a matrix in $\mathbb{R}^{n\times|I|}$,*

$$L_I^W = BB^\dagger_{I,\bullet}.$$

Proof. If $p \in \pi_I(W)$, then $p = B_{I,\bullet}y$ for some $y \in \mathbb{R}^{n-m}$. A property of the pseudo-inverse is that $B^\dagger_{I,\bullet}B_{I,\bullet}$ is the orthogonal projection onto the orthogonal complement of the kernel of $B_{I,\bullet}$, from which it follows that $B^\dagger_{I,\bullet}p = \arg\min_{p=B_{I,\bullet}y}\|y\|$. This solution satisfies $\pi_I(BB^\dagger_{I,\bullet}p) = p$ and $BB^\dagger_{I,\bullet}p \in W$. Since the columns of B form an orthonormal basis of W, we have $\|BB^\dagger_{I,\bullet}p\| = \|B^\dagger_{I,\bullet}p\|$. Consequently, $BB^\dagger_{I,\bullet}p$ is the minimum-norm point with the above properties. □

The above lemma and Proposition 4.2.1(iii) yield the following characterization. This will be the most suitable characterization of $\bar{\chi}_W$ for our purposes.

Proposition 4.2.4. *For a linear subspace $W \subseteq \mathbb{R}^n$,*

$$\bar{\chi}_W = \max\left\{\|L_I^W\| : I \subseteq [n], I \neq \emptyset\right\}.$$

The following notation will be convenient for our algorithm. For a subspace $W \subseteq \mathbb{R}^n$ and an index set $I \subseteq [n]$, if $\pi_I(W) \neq \{\vec{0}\}$ then we define the *lifting score*

$$\ell^W(I) := \sqrt{\|L_I^W\|^2 - 1}. \tag{4.5}$$

Otherwise, we define $\ell^W(I) = 0$. This means that for any $z \in \pi_I(W)$ and $x = L_I^W(z)$, $\|x_{[n]\setminus I}\| \leq \ell^W(I)\|z\|$.

The condition number $\bar{\chi}^*_A$ For every $D \in \mathbf{D}$, we can consider the condition number $\bar{\chi}_{WD} = \bar{\chi}_{AD^{-1}}$. We let

$$\bar{\chi}^*_W = \bar{\chi}^*_A = \inf\{\bar{\chi}_{WD} : D \in \mathbf{D}\}$$

denote the best possible value of $\bar{\chi}$ that can be attained by rescaling the coordinates of W. The main result of this section is the following theorem.

Theorem 4.2.5 (Proof on p. 139)**.** *There is an $O(n^2m^2 + n^3)$ time algorithm that for any matrix $A \in \mathbb{R}^{m\times n}$ computes an estimate ξ of $\bar{\chi}_W$ such that*

$$\xi \le \bar{\chi}_W \le n(\bar{\chi}^*_W)^2\xi$$

and a $D \in \mathbf{D}$ such that

$$\bar{\chi}^*_W \le \bar{\chi}_{WD} \le n(\bar{\chi}^*_W)^3 .$$

4.2.2 The circuit imbalance measure

The key tool in proving Theorem 4.2.5 is to study a more combinatorial condition number, the *circuit imbalance measure* which turns out to give a good proxy to $\bar{\chi}_A$.

Definition 4.2.6. For a linear subspace $W \subseteq \mathbb{R}^n$ and a matrix A such that $W = \mathrm{Ker}(A)$, a circuit is an inclusion-wise minimal dependent set of columns of A. Equivalently, a circuit is a set $C \subseteq [n]$ such that $W \cap \mathbb{R}^n_C$ is one-dimensional and that no strict subset of C has this property. The set of circuits of W is denoted $\mathcal{C}_W$.

Note that circuits defined above are the same as the circuits in the linear matroid associated with A. Every circuit $C \in \mathcal{C}_W$ can be associated with a vector $g^C \in W$ such that $\mathrm{supp}(g^C) = C$; this vector is unique up to scalar multiplication.

Definition 4.2.7. For a circuit $C \in \mathcal{C}_W$ and $i, j \in C$, we let

$$\kappa^W_{ij}(C) = \frac{\left|g^C_j\right|}{\left|g^C_i\right|} . \tag{4.6}$$

Note that since g^C is unique up to scalar multiplication, this is independent of the choice of g^C. For any $i, j \in [n]$, we define the *circuit ratio* as the maximum of $\kappa^W_{ij}(C)$ over all choices of the circuit C:

$$\kappa^W_{ij} = \max\left\{\kappa^W_{ij}(C) : C \in \mathcal{C}_W, i, j \in C\right\} . \tag{4.7}$$

By convention we set $\kappa_{ij}^W = 0$ if there is no circuit supporting i and j. Further, we define the *circuit imbalance measure* as

$$\kappa_W = \max\left\{\kappa_{ij}^W : i, j \in [n]\right\} .$$

Minimizing over all coordinate rescalings, we define

$$\kappa_W^* = \min\{\kappa_{WD} : D \in \mathbf{D}\} .$$

We omit the index W whenever it is clear from context. Further, for a vector $d \in \mathbb{R}^n_{++}$, we write $\kappa_{ij}^d = \kappa_{ij}^{\mathrm{Diag}(d)W}$ and $\kappa^d = \kappa_W^d = \kappa_{\mathrm{Diag}(d)W}$.

We want to remark that a priori it is not clear that κ_W^* is well-defined. Theorem 4.2.13 will show that the minimum of $\{\kappa_{WD} : D \in \mathbf{D}\}$ is indeed attained.

We next formulate the main statements on the circuit imbalance measure; proofs will be given in the subsequent subsections. Crucially, we show that the circuit imbalance κ_W is a good proxy to the condition number $\bar{\chi}_W$. The lower bound was already proven in [197], and the upper bound is from [53].

Theorem 4.2.8 (Proof on p. 130)**.** *For a linear subspace $W \subseteq \mathbb{R}^n$,*

$$\sqrt{1 + (\kappa_W)^2} \leq \bar{\chi}_W \leq n\kappa_W .$$

We now overview some basic properties of κ_W. Proposition 4.2.4 asserts that $\bar{\chi}_W$ is the maximum $\ell_2 \to \ell_2$ operator norm of the mappings L_I^W over $I \subseteq [n]$. In [53], it was shown that κ_W is in contrast the maximum $\ell_1 \to \ell_\infty$ operator norm of the same mappings; this easily implies the upper bound $\bar{\chi}_W \leq n\kappa_W$.

Proposition 4.2.9 ([53])**.** *For a linear subspace $W \subseteq \mathbb{R}^n$,*

$$\kappa_W = \max\left\{\frac{\|L_I^W(p)\|_\infty}{\|p\|_1} : I \subseteq [n], I \neq \emptyset, p \in \pi_I(W) \setminus \{\vec{0}\}\right\} .$$

Similarly to $\bar{\chi}_W$, κ_W is self-dual; moreover, this holds for all individual κ_{ij}^W values.

Lemma 4.2.10 (Proof on p. 131)**.** *For any subspace $W \subseteq \mathbb{R}^n$ and $i, j \in [n]$, $\kappa_{ij}^W = \kappa_{ji}^{W^\perp}$.*

The next lemma provides a subroutine that efficienctly yields upper bounds on $\ell^W(I)$ or lower bounds on some circuit imbalance values. Recall the definition of the lifting score $\ell^W(I)$ from (4.5).

Lemma 4.2.11 (Proof on p. 131)**.** *There exists a subroutine* VERIFY-LIFT(W, I, θ) *that, given a linear subspace $W \subseteq \mathbb{R}^n$, an index set $I \subseteq [n]$, and a threshold $\theta \in \mathbb{R}_{++}$, either returns the answer* 'pass'*, verifying $\ell^W(I) \le \theta$, or returns the answer* 'fail'*, and a pair $i \in I, j \in [n] \setminus I$ such that $\theta/n \le \kappa_{ij}^W$. The running time can be bounded as $O(n(n-m)^2)$.*

The proof of the above statements are given in Section 4.2.3.

A min-max theorem We next provide a combinatorial min-max characterization of κ_W^*. Consider the *circuit ratio digraph* $G = ([n], E)$ on the node set $[n]$ where $(i, j) \in E$ if $\kappa_{ij} > 0$, that is, there exists a circuit $C \in \mathcal{C}$ with $i, j \in C$. We will refer to $\kappa_{ij} = \kappa_{ij}^W$ as the weight of the edge (i, j). (Note that $(i, j) \in E$ if and only if $(j, i) \in E$, but the weight of these two edges can be different.)

Definition 4.2.12. Let H be a *cycle* in G: a sequence of indices $i_1, i_2, \ldots, i_k, i_{k+1} = i_1$. We use $|H| = k$ to denote the length of the cycle. (In our terminology, 'cycles' always refer to objects in G, whereas 'circuits' refer to the minimum supports in $\mathrm{Ker}(A)$.) We use the notation $\kappa(H) = \kappa_W(H) = \prod_{j=1}^{k} \kappa_{i_j i_{j+1}}^W$. For a vector $d \in \mathbb{R}_{++}^n$, we denote $\kappa_W^d(H) = \kappa_{\mathrm{Diag}(d)W}(H)$.

A simple but important observation is that such a rescaling does not change the value associated with the cycle, that is,

$$\kappa_W^d(H) = \kappa_W(H) \quad \forall d \in \mathbb{R}_{++}^n \quad \text{for any cycle } H \text{ in } G\,. \tag{4.8}$$

Theorem 4.2.13 (Proof on p. 132)**.** *For a subspace $W \subseteq \mathbb{R}^n$, we have*

$$\kappa_W^* = \min_{d > \vec{0}} \kappa_W^d = \max\left\{ \kappa_W(H)^{1/|H|} : \ H \text{ is a cycle in } G \right\}\,.$$

The proof relies on the following formulation:

$$\begin{aligned} \kappa_W^* = \quad & \min t \\ & \kappa_{ij} d_j / d_i \le t \quad \forall (i,j) \in E \\ & d > \vec{0}. \end{aligned}$$

Taking logarithms, we can rewrite this problem as

$$\begin{aligned} & \min s \\ & \log \kappa_{ij} + z_j - z_i \le s \quad \forall (i,j) \in E \\ & z \in \mathbb{R}^n. \end{aligned}$$

This is the dual of the minimum-mean cycle problem with weights $\log \kappa_{ij}$, and can be solved in polynomial time (see e.g. [4, Theorem 5.8]).

Whereas this formulation verifies Theorem 4.2.13, it does not give a polynomial-time algorithm to compute κ^*_W. The caveat is that the values κ^W_{ij} are typically not available; in fact, approximating them up to a factor $2^{O(m)}$ is NP-hard, as follows from the work of Tunçel [188].

Nevertheless, the following corollary of Theorem 4.2.13 shows that any arbitrary circuit containing i and j yields a $(\kappa^*)^2$ approximation to κ_{ij}.

Corollary 4.2.14 (Proof on p. 133)**.** *Let us be given a linear subspace $W \subseteq \mathbb{R}^n$ and $i, j \in [n]$, $i \neq j$, and a circuit $C \in \mathcal{C}_W$ with $i, j \in C$. Let $g \in W$ be the corresponding vector with* $\operatorname{supp}(g) = C$*. Then,*

$$\frac{\kappa^W_{ij}}{\left(\kappa^*_W\right)^2} \leq \frac{|g_j|}{|g_i|} \leq \kappa^W_{ij}\,.$$

The above statements are shown in Section 4.2.4. In Section 4.2.5, we use techniques from matroid theory and linear algebra to efficiently identify a circuit for any pair of variables that are contained in the same circuit. A matroid is non-separable if the circuit hypergraph is connected; precise definitions and background will be described in Section 4.2.5.

Theorem 4.2.15 (Proof on p. 136)**.** *Given $A \in \mathbb{R}^{m\times n}$, there exists an $O(n^2m^2)$ time algorithm* Find-Circuits(A) *that obtains a decomposition of $\mathcal{M}(A)$ to a direct sum of non-separable linear matroids, and returns a family $\hat{\mathcal{C}}$ of circuits such that if i and j are in the same non-separable component, then there exists a circuit in $\hat{\mathcal{C}}$ containing both i and j. Further, for each $i \neq j$ in the same component, the algorithm returns a value $\hat{\kappa}_{ij}$ as the the maximum of $|g_j/g_i|$ such that $g \in W$,* $\operatorname{supp}(g) = C$ *for some $C \in \hat{\mathcal{C}}$ containing i and j. For these values, $\hat{\kappa}_{ij} \leq \kappa_{ij} \leq (\kappa^*)^2\hat{\kappa}_{ij}$.*

Finally, in Section 4.2.6, we combine the above results to prove Theorem 4.2.5 on approximating $\bar{\chi}^*_W$ and κ^*_W.

Section 4.2.5 contains an interesting additional statement, namely that the logarithms of the circuit ratios satisfy the triangle inequality. This will also be useful in the analysis of the LLS algorithm. The proof uses similar arguments as the proof of Theorem 4.2.15.

Lemma 4.2.16 (Proof on p. 137)**.**

(i) *For any distinct i, j, k in the same connected component of $\mathcal{C}_W$, and any g^C with $i, j \in C$, $C \in \mathcal{C}_W$, there exist circuits $C_1, C_2 \in \mathcal{C}_W$, $i, k \in C_1$, $j, k \in C_2$ such that $|g^C_j/g^C_i| = |g^{C_2}_j/g^{C_2}_k| \cdot |g^{C_1}_k/g^{C_1}_i|$.*

(ii) *For any distinct i, j, k in the same connected component of $\mathcal{C}_W$, $\kappa_{ij} \leq \kappa_{ik} \cdot \kappa_{kj}$.*

4.2.3 Basic properties of κ_W

Theorem 4.2.8 (Repetition)**.** *For a linear subspace $W \subseteq \mathbb{R}^n$,*

$$\sqrt{1 + (\kappa_W)^2} \leq \bar{\chi}_W \leq n\kappa_W .$$

Proof. For the first inequality, let $C \in \mathcal{C}_W$ be the circuit and $i \neq j \in C$ such that $|g_j/g_i| = \kappa_W$ for the corresponding solution $g = g^C$. Let us use the characterization of $\bar{\chi}_W$ in Proposition 4.2.4. Let $I = ([n] \setminus C) \cup \{i\}$, and $p = g_i e^i$, that is, the vector with $p_i = g_i$ and $p_k = 0$ for $k \neq i$. Then, the unique vector $z \in W$ such that $z_I = p$ is $z = g$. Therefore,

$$\bar{\chi}_W \geq \min_{z \in W, z_I = p} \frac{\|z\|}{\|p\|} = \frac{\|g\|}{|g_i|} \geq \frac{\sqrt{|g_i|^2 + |g_j|^2}}{|g_i|} = \sqrt{1 + \kappa_W^2} .$$

The second inequality is immediate from Proposition 4.2.4 and Proposition 4.2.9, and the inequalities between ℓ_1, ℓ_2, and ℓ_∞ norms. The proof of the slightly weaker $\bar{\chi}_W \leq \sqrt{1 + (n\kappa_W)^2}$ follows from Lemma 4.2.11. □

The next lemma will be needed to prove Lemma 4.2.11 and also to analyze the LLS algorithm. Let us say that the vector $y \in \mathbb{R}^n$ is *sign-consistent* with $x \in \mathbb{R}^n$ if $x_i y_i \geq 0$ for all $i \in [n]$ and $x_i = 0$ implies $y_i = 0$ for all $i \in [n]$.

Lemma 4.2.17. *For $i \in I \subseteq [n]$ with $e^i_I \in \pi_I(W)$, let $z = L^W_I(e^i_I)$. Then for any $j \in \operatorname{supp}(z)$ we have $\kappa^W_{ij} \geq |z_j|$.*

Proof. We consider the cone $F \subseteq W$ of vectors sign-consistent with z. The faces of F are bounded by inequalities of the form $z_k y_k \geq 0$ or $y_k = 0$. The edges (rays) of F are of the form $\{\alpha g : \alpha \geq 0\}$ with $\operatorname{supp}(g) \in \mathcal{C}_W$. It is easy to see from the Minkowski-Weyl theorem that z can be written as

$$z = \sum_{k=1}^{h} g^k,$$

where $h \leq n$, $C_1, C_2, \ldots, C_h \in \mathcal{C}_W$ are circuits, and the vectors $g^1, g^2, \ldots, g^h \in W$ are sign-consistent with z and $\operatorname{supp}(g^k) = C_k$ for all $k \in [h]$. Note that $i \in C_k$ for all $k \in [h]$, as otherwise, $z' = z - g^k$ would also satisfy $z'_I = e^i_I$, but $\|z'\| < \|z\|$ due to g^k being sign-consistent with z, a contradiction to the definition of z.

At least one $k \in [h]$ contributes at least as much to $|z_j| = \frac{\sum_{k=1}^h |g^k_j|}{\sum_{k=1}^h g^k_i}$ as the average. Hence we find $\kappa^W_{ij} \geq |g^k_j/g^k_i| \geq |z_j|$. □

Lemma 4.2.11 (Repetition)**.** *There exists a subroutine* VERIFY-LIFT(W, I, θ) *that, given a linear subspace* $W \subseteq \mathbb{R}^n$*, an index set* $I \subseteq [n]$*, and a threshold* $\theta \in \mathbb{R}_{++}$*, either returns the answer* 'pass'*, verifying* $\ell^W(I) \leq \theta$*, or returns the answer* 'fail'*, and a pair* $i \in I, j \in [n] \setminus I$ *such that* $\theta/n \leq \kappa^W_{ij}$*. The running time can be bounded as* $O(n(n-m)^2)$*.*

Proof. Take any minimal $I' \subseteq I$ such that $\dim(\pi_{I'}(W)) = \dim(\pi_I(W))$. Then we know that $\pi_{I'}(W) = \mathbb{R}^{I'}$ and for $p \in \pi_I(W)$ we can compute $L^W_I(p) = L^W_{I'}(p_{I'})$. Let $B \in \mathbb{R}^{([n]\setminus I)\times I'}$ be the matrix sending any $q \in \pi_{I'}(W)$ to the corresponding vector $(L^W_{I'}(q))_{[n]\setminus I}$. The column B_i can be computed as $(L^W_{I'}(e^i_{I'}))_{[n]\setminus I}$ for $e^i_{I'} \in \mathbb{R}^{I'}$. We have $\|L^W_I(p)\|^2 = \|p\|^2 + \|(L^W_{I'}(p_{I'}))_{[n]\setminus I}\|^2 \leq \|p\|^2 + \|B\|^2\|p_{I'}\|^2$ for any $p \in \pi_I(W)$, and so $\ell^W(I) = \sqrt{\|L^W_I\|^2 - 1} \leq \|B\|$. We upper bound the operator norm by the Frobenius norm as $\|B\| \leq \|B\|_F = \sqrt{\sum_{ji} B^2_{ji}} \leq n \max_{ji} |B_{ji}|$. By Lemma 4.2.17 it follows that $|B_{ji}| = |(L^W_{I'}(e^i))_j| \leq \kappa^W_{ij}$. The algorithm returns the answer 'pass' if $n \max_{ji} |B_{ji}| \leq \theta$ and 'fail' otherwise.

To implement the algorithm, we first need to select a minimal $I' \subseteq I$ such that $\dim(\pi_{I'}(W)) = \dim(\pi_I(W))$. This can be found by computing $M \in \mathbb{R}^{(n-m)\times n}$ such that $\mathrm{range}(M) = W$, and selecting a maximal number of linearly independent columns of M_I. Then, we compute the matrix $B \in \mathbb{R}^{([n]\setminus I)\times I'}$ that implements the transformation $[L^W_{I'}]_{[n]\setminus I} : \pi_{I'}(W) \to \pi_{[n]\setminus I}(W)$. The algorithm returns the pair (i, j) corresponding to the entry maximizing $|B_{ji}|$. The running time analysis will be given in the proof of Lemma 4.3.15, together with an amortized analysis of a sequence of calls to the subroutine. □

Remark 4.2.18. We note that the algorithm VERIFY-LIFT does not need to compute the circuit as in Lemma 4.2.17. The following observation will be important in the analysis: the algorithm returns the answer 'fail' even if $\ell^W(I) \leq \theta < n|B_{ji}|$.

We now prove the duality property of the circuit imbalances.

Lemma 4.2.10 (Repetition)**.** *For any subspace* $W \subseteq \mathbb{R}^n$ *and* $i, j \in [n]$*,* $\kappa^W_{ij} = \kappa^{W^\perp}_{ji}$*.*

Proof. Choose a circuit $C \in \mathcal{C}_W$ and corresponding circuit solution $g := g^C \in W \cap \mathbb{R}^n_C$ such that $\kappa_{ij} = \kappa_{ij}(C) = |g_j/g_i|$. We will construct a circuit solution in $W^\perp$ that certifies $\kappa^{W^\perp}_{ji} \geq \kappa^W_{ij}$.

Define $h \in \mathbb{R}^C$ by $h_i = g_j, h_j = -g_i$ and $h_k = 0$ for all $k \in C \setminus \{i, j\}$. Then, h is orthogonal to g_C by construction, and hence $h \in (\pi_C(W \cap \mathbb{R}^n_C))^\perp = \pi_C(W^\perp)$. Furthermore, we have $\mathrm{supp}(h) \in \mathcal{C}_{\pi_C(W^\perp)}$ since $h \in \mathbb{R}^C$ is a support minimal vector orthogonal to g^C.

Take any vector $\bar{h} \in W^\perp$ satisfying $\bar{h}_C = h$ that is support minimal subject to these constraints. We claim that $\mathrm{supp}(\bar{h}) \in \mathcal{C}_{W^\perp}$. Assume not, then there exists a non-zero $v \in W^\perp$ with $\mathrm{supp}(v) \subseteq \mathrm{supp}(\bar{h})$. Since $\mathrm{supp}(\pi_C(v)) \subseteq \mathrm{supp}(\pi_C(\bar{h})) = \mathrm{supp}(h)$, we must have either $v_C = \vec{0}$ or $v_C = sh$ for $s \neq 0$. If $v_C = \vec{0}$, then $\bar{h} - \alpha v$ is also in $W^\perp$ satisfying $\pi_C(\bar{h}_C - \alpha v) = h$ for all $\alpha \in \mathbb{R}$, and since $v \neq \vec{0}$ we can choose α such that $\bar{h} - \alpha v$ has smaller support than $\bar{h}$, a contradiction. If $s \neq 0$ then $v/s \in W^\perp$ satisfies $\pi_C(v/s) = h$ and has smaller support than $\bar{h}$, again a contradiction.

By the above construction, we have

$$\kappa_{ji}^{W^\perp} \geq \left|\frac{\bar{h}_i}{\bar{h}_j}\right| = \left|\frac{h_i}{h_j}\right| = \left|\frac{g_j}{g_i}\right| = \kappa_{ij}^{W} \ .$$

By swapping the role of W and $W^\perp$ and i and j, we obtain $\kappa_{ij}^{W} \geq \kappa_{ji}^{W^\perp}$. The statement follows. □

4.2.4 Proving the min-max theorem on κ_W^*

The proof of the characterization of κ_W^* follows.

Theorem 4.2.13 (Repetition)**.** *For a subspace $W \subseteq \mathbb{R}^n$, we have*

$$\kappa_W^* = \min_{d > \vec{0}} \kappa_W^d = \max\left\{\kappa_W(H)^{1/|H|} : \ H \text{ is a cycle in } G\right\} \ .$$

Proof. For the direction $\kappa_W(H)^{1/|H|} \leq \kappa_W^*$ we use (4.8). Let $d > \vec{0}$ be a scaling and H a cycle. We have $\kappa_{ij}^d \leq \kappa_W^d$ for every $i, j \in [n]$, and hence $\kappa_W(H) = \kappa_W^d(H) \leq (\kappa_W^d)^{|H|}$. Since this inequality holds for every $d > \vec{0}$, it follows that $\kappa_W(H) \leq (\kappa_W^*)^{|H|}$.

For the reverse direction, consider the following optimization problem.

$$\begin{aligned} &\min t \\ &\kappa_{ij} d_j/d_i \leq t \quad \forall (i,j) \in E \\ &d > \vec{0}. \end{aligned} \tag{4.9}$$

For any feasible solution (d, t) and $\lambda > 0$, we get another feasible solution $(\lambda d, t)$ with the same objective value. As such, we can strengthen the condition $d > \vec{0}$ to $d \geq 1$ without changing the objective value. This makes it clear that the optimum value is achieved by a feasible solution.

Any rescaling $d > \vec{0}$ provides a feasible solution with objective value κ^d, which means that the optimal value t^* of (4.9) is $t^* = \kappa^*$. Moreover, with the variable

substitution $z_i = \log d_i$, $s = \log t$, (4.9) can be written as a linear program:

$$\begin{aligned} &\min s \\ &\log \kappa_{ij} + z_j - z_i \le s \quad \forall (i,j) \in E \\ &z \in \mathbb{R}^n. \end{aligned} \tag{4.10}$$

This is the dual of a minimum-mean cycle problem with respect to the cost function $\log(\kappa_{ij})$. Therefore, an optimal solution corresponds to the cycle maximizing $\sum_{ij\in H} \log \kappa_{ij}/|H|$, or in other words, maximizing $\kappa(H)^{1/|H|}$. □

The following example shows that $\kappa^* \le \bar{\chi}^*$ can be arbitrarily big.

Example 4.2.19. Take $W = \operatorname{span}((0,1,1,M),(1,0,M,1))$, where $M > 0$. Then $\{2,3,4\}$ and $\{1,3,4\}$ are circuits with $\kappa^W_{34}(\{2,3,4\}) = M$ and $\kappa^W_{43}(\{1,3,4\}) = M$. Hence, by Theorem 4.2.13, we see that $\kappa^* \ge M$.

Corollary 4.2.14 (Repetition)**.** *Let us be given a linear subspace $W \subseteq \mathbb{R}^n$ and $i, j \in [n]$, $i \ne j$, and a circuit $C \in \mathcal{C}_W$ with $i, j \in C$. Let $g \in W$ be the corresponding vector with* $\operatorname{supp}(g) = C$*. Then,*

$$\frac{\kappa^W_{ij}}{\left(\kappa^*_W\right)^2} \le \frac{|g_j|}{|g_i|} \le \kappa^W_{ij}.$$

Proof. The second inequality follows by definition. For the first inequality, note that the same circuit C yields $|g_i/g_j| \le \kappa^W_{ji}(C) \le \kappa^W_{ji}$. Therefore, $|g_j/g_i| \ge 1/\kappa^W_{ji}$.

From Theorem 4.2.13 we see that $\kappa^W_{ij}\kappa^W_{ji} \le (\kappa^*_W)^2$, giving $1/\kappa^W_{ji} \ge \kappa^W_{ij}/(\kappa^*_W)^2$, completing the proof. □

4.2.5 Finding circuits: a detour in matroid theory

We next prove Theorem 4.2.15, showing how to efficiently obtain a family $\hat{\mathcal{C}} \subseteq \mathcal{C}_W$ such that for any $i, j \in [n]$, $\hat{\mathcal{C}}$ includes a circuit containing both i and j, provided there exists such a circuit.

We need some simple concepts and results from matroid theory. We refer the reader to [170, Chapter 39] or [83, Chapter 5] for definitions and background. Let $\mathcal{M} = ([n], \mathcal{I})$ be a matroid on ground set $[n]$ with independent sets $\mathcal{I} \subseteq 2^{[n]}$. The rank $\operatorname{rk}(S)$ of a set $S \subseteq [n]$ is the maximum size of an independent set contained in S. The maximal independent sets are called *bases*. All bases have the same cardinality $\operatorname{rk}([n])$.

For the matrix $A \in \mathbb{R}^{m\times n}$, we work with the linear matroid $\mathcal{M}(A) = ([n], \mathcal{I}(A))$, where a subset $I \subseteq [n]$ is independent if the columns $\{A_i : i \in I\}$ are linearly independent. Note that $\operatorname{rk}([n]) = m$ under the assumption that A has full row rank.

The *circuits* of the matroid are the inclusion-wise minimal non-independent sets. Let $I \in \mathcal{I}$ be an independent set, and $i \in [n] \setminus I$ such that $I \cup \{i\} \notin \mathcal{I}$. Then, there exists a unique circuit $C(I,i) \subseteq I \cup \{i\}$ that is called the *fundamental circuit* of i with respect to I. Note that $i \in C(I,i)$.

The matroid $\mathcal{M}$ is *separable*, if the ground set $[n]$ can be partitioned to two nonempty subsets $[n] = S \cup T$ such that $I \in \mathcal{I}$ if and only if $I \cap S, I \cap T \in \mathcal{I}$. In this case, the matroid is the direct sum of its restrictions to S and T. In particular, every circuit is fully contained in S or in T.

For the linear matroid $\mathcal{M}(A)$, separability means that $\mathrm{Ker}(A) = \mathrm{Ker}(A_S) \oplus \mathrm{Ker}(A_T)$. In this case, solving (4.1) can be decomposed into two subproblems, restricted to the columns in A_S and in A_T, and $\kappa_A = \max\{\kappa_{A_S}, \kappa_{A_T}\}$.

Hence, we can focus on *non-separable* matroids. The following characterization is well-known, see e.g. [83, Theorems 5.2.5, 5.2.7–5.2.9]. For a hypergraph $H = ([n], \mathcal{E})$, we define the underlying graph $H_G = ([n], E)$ such that $(i,j) \in E$ if there is a hyperedge $S \in \mathcal{E}$ with $i, j \in S$. That is, we add a clique corresponding to each hyperedge. The hypergraph is called *connected* if the underlying graph $G = ([n], E)$ is connected.

Proposition 4.2.20. *For a matroid $\mathcal{M} = ([n], \mathcal{I})$, the following are equivalent*:

(i) *$\mathcal{M}$ is non-separable.*

(ii) *The hypergraph of the circuits is connected.*

(iii) *For any base B of $\mathcal{M}$, the hypergraph formed by the fundamental circuits $\mathcal{C}^B = \{C(B,i) : i \in [n] \setminus B\}$ is connected.*

(iv) *For any $i, j \in [n]$, there exists a circuit containing i and j.*

Proof. The implications (i) $\Leftrightarrow$ (ii), (iii) $\Rightarrow$ (ii), and (iv) $\Rightarrow$ (ii) are immediate from the definitions.

For the implication (ii) $\Rightarrow$ (iii), assume for a contradiction that the hypergraph of the fundamental circuits with respect to B is not connected. This means that we can partition $[n] = S \cup T$ such that for each $i \in S$, $C(B,i) \subseteq S$, and for each $i \in T$, $C(B,i) \subseteq T$. Consequently, $\mathrm{rk}(S) = |B \cap S|$, $\mathrm{rk}(T) = |B \cap T|$, and therefore $\mathrm{rk}([n]) = \mathrm{rk}(S) + \mathrm{rk}(T)$. It is easy to see that this property is equivalent to separability to S and T; see e.g. [83, Theorem 5.2.7] for a proof.

Finally, for the implication (ii) $\Rightarrow$ (iv), consider the undirected graph $([n], E)$ where $(i,j) \in E$ if there is a circuit containing both i and j. This graph is transitive according to [83, Theorem 5.2.5]: if $(i,j), (j,k) \in E$, then also $(i,k) \in E$. Consequently, whenever $([n], E)$ is connected, it must be a complete directed graph. □

We now give a different proof of (iii) $\Rightarrow$ (iv) that will be convenient for our algorithmic purposes. First, we need a simple lemma that is commonly used in matroid optimization, see e.g. [83, Lemma 13.1.11] or [170, Theorem 39.13].

Lemma 4.2.21. *Let I be an independent set of a matroid $\mathcal{M} = ([n], \mathcal{I})$, and $U = \{u_1, u_2, \dots, u_\ell\} \subseteq I$, $V = \{v_1, v_2, \dots, v_\ell\} \subseteq [n] \setminus I$ such that $I \cup \{v_i\}$ is dependent for each $i \in [\ell]$. Further, assume that for each $t \in [\ell]$, $u_t \in C(I, v_t)$ and $u_t \notin C(I, v_h)$ for all $h < t$. Then, $(I \setminus U) \cup V \in \mathcal{I}$.*

We give a sketch of the proof. First, we note that for each $t \in [\ell]$, $u_t \in C(I, v_t)$ means that exchanging v_t for u_t maintains independence. The statement follows by induction on ℓ: we consider the independent set $I' = (I \setminus \{u_\ell\}) \cup \{v_\ell\}$. We can apply induction for I', $U' = \{u_1, u_2, \dots, u_{\ell-1}\}$, and $V' = \{v_1, v_2, \dots, v_{\ell-1}\}$, noting that the assumption guarantees that $C(I', v_t) = C(I, v_t)$ for all $t \in [\ell - 1]$. Based on this lemma, we show the following exchange property.

Lemma 4.2.22. *Let B be a basis of the matroid $\mathcal{M} = ([n], \mathcal{I})$, and let $U = \{u_1, u_2, \dots, u_\ell\} \subseteq B$, and $V = \{v_1, v_2, \dots, v_\ell, v_{\ell+1}\} \subseteq [n] \setminus B$. Assume $C(B, v_1) \cap U = \{u_1\}$, $C(B, v_{\ell+1}) \cap U = \{u_\ell\}$, and for each $2 \leq t \leq \ell$, $C(B, v_t) \cap U = \{u_{t-1}, u_t\}$. Then $(B \setminus U) \cup V$ contains a unique circuit C, and $V \subseteq C$.*

The situation described here corresponds to a minimal path in the hypergraph $\mathcal{C}^B$ of the fundamental circuits with respect to a basis B. The hyperedges $C(B, v_i)$ form a path from v_1 to $v_{\ell+1}$ such that no shortcut is possible (note that this is weaker than requiring a shortest path).

Proof of Lemma 4.2.22. Note that $S = (B \setminus U) \cup V \notin \mathcal{I}$ since $|S| > |B|$ and B is a basis. For any $i \in [\ell + 1]$, we can use Lemma 4.2.21 to show that

$$S \setminus \{v_i\} = (B \setminus U) \cup (V \setminus \{v_i\}) \in \mathcal{I},$$

and thus $S \setminus \{v_i\}$ is a basis. To see this, we apply Lemma 4.2.21 for the ordered sets $V' = \{v_1, \dots, v_{i-1}, v_{\ell+1}, v_\ell, \dots, v_{i+1}\}$ and $U' = \{u_1, \dots, u_{i-1}, u_\ell, u_{\ell-1}, \dots, u_i\}$.

Consequently, every circuit in S must contain the entire set V. The uniqueness of the circuit in S follows by the well-known circuit axiom asserting that if $C, C' \in \mathcal{C}$, $C \neq C'$ and $v \in C \cap C'$, then there exists a circuit $C'' \in \mathcal{C}$ such that $C'' \subseteq (C \cup C') \setminus \{v\}$, contradicting the claim that every circuit in S contains the entire set V. □

We are ready to describe the algorithm that will be used to obtain lower bounds on all κ_{ij} values.

Theorem 4.2.15 (Repetition)**.** *Given $A \in \mathbb{R}^{m \times n}$, there exists an $O(n^2 m^2)$ time algorithm* Find-Circuits(A) *that obtains a decomposition of $\mathcal{M}(A)$ to a direct sum of non-separable linear matroids, and returns a family $\hat{\mathcal{C}}$ of circuits such that if i and j*

are in the same non-separable component, then there exists a circuit in $\hat{\mathcal{C}}$ containing both i and j. Further, for each $i \neq j$ in the same component, the algorithm returns a value $\hat{\kappa}_{ij}$ as the the maximum of $|g_j/g_i|$ such that $g \in W$, $\mathrm{supp}(g) = C$ for some $C \in \hat{\mathcal{C}}$ containing i and j. For these values, $\hat{\kappa}_{ij} \leq \kappa_{ij} \leq (\kappa^*)^2 \hat{\kappa}_{ij}$.

Proof. Once we have found the set of circuits $\hat{\mathcal{C}}$, and computed $\hat{\kappa}_{ij}$ as in the statement, the inequalities $\hat{\kappa}_{ij} \leq \kappa_{ij} \leq (\kappa^*)^2 \hat{\kappa}_{ij}$ follow easily. The first inequality is by the definition of κ_{ij}, and the second inequality is from Corollary 4.2.14.

We now turn to the computation of $\hat{\mathcal{C}}$. We first obtain a basis $B \subseteq [n]$ of $\mathrm{Ker}(A)$ via Gauss-Jordan elimination in time $O(nm^2)$. Recall the assumption that A has full row-rank. Let us assume that $B = [m]$ is the set of first m indices. The elimination transforms it to the form $A = (I_m|H)$, where $H \in \mathbb{R}^{m\times(n-m)}$ corresponds to the non-basis elements. In this form, the fundamental circuit $C(B, i)$ is the support of the ith column of A together with i for every $m + 1 \leq i \leq n$. We let $\mathcal{C}^B$ denote the set of all these fundamental circuits.

We construct an undirected graph $G = (B, E)$ as follows. For each $i \in [n] \setminus B$, we add a clique between the nodes in $C(B, i) \setminus \{i\}$. This graph can be constructed in $O(nm^2)$ time.

The connected components of G correspond to the connected components of $\mathcal{C}^B$ restricted to B. Thus, due to the equivalence shown in Proposition 4.2.20 we can obtain the decomposition by identifying the connected components of G. For the rest of the proof, we assume that the entire hypergraph is connected; connectivity can be checked in $O(m^2)$ time.

We initialize $\hat{\mathcal{C}}$ as $\mathcal{C}^B$. We will then check all pairs $i, j \in [n]$, $i \neq j$. If no circuit $C \in \hat{\mathcal{C}}$ exists with $i, j \in C$, then we will add such a circuit to $\hat{\mathcal{C}}$ as follows.

Assume first $i, j \in [n] \setminus B$. We can find a shortest path in G between the sets $C(B, i) \setminus \{i\}$ and $C(B, j) \setminus \{j\}$ in time $O(m^2)$. This can be represented by the sequences of points $V = \{v_1, v_2, \ldots, v_{\ell+1}\} \subseteq [n] \setminus B$, $v_1 = i$, $v_{\ell+1} = j$, and $U = \{u_1, u_2, \ldots, u_\ell\} \subseteq B$ as in Lemma 4.2.22. According to the lemma, $S = (B \setminus U) \cup V$ contains a unique circuit C that contains all v_t's, including i and j.

We now show how this circuit can be identified in $O(m)$ time, along with the vector g^C. Let A_S be the submatrix corresponding to the columns in S. Since $g = g^C$ is unique up to scaling, we can set $g_{v_1} = 1$. Note that for each $t \in [\ell]$, the row of A_S corresponding to u_t contains only two nonzero entries: $A_{u_t v_t}$ and $A_{u_t v_{t+1}}$. Thus, the value $g_{v_1} = 1$ can be propagated to assigning unique values to $g_{v_2}, g_{v_3}, \ldots, g_{v_{\ell+1}}$. Once these values are set, there is a unique extension of g to the indices $t \in B \cap S$ in the basis. Thus, we have identified g as the unique element of $\mathrm{Ker}(A_S)$ up to scaling. The circuit C is obtained as $\mathrm{supp}(g)$. Clearly, the above procedure can be implemented in $O(m)$ time.

The argument easily extends to finding circuits for the case $\{i, j\} \cap B \neq \emptyset$. If

$i \in B$, then for any choice of $V = \{v_1, v_2, \ldots, v_{\ell+1}\}$ and $U = \{u_1, u_2, \ldots, u_\ell\}$ as in Lemma 4.2.22 such that $i \in C(B, v_1)$ and $i \notin C(B, v_t)$ for $t > 1$, the unique circuit in $(B \setminus U) \cup V$ also contains i. This follows from Lemma 4.2.21 by taking $V' = \{v_{\ell+1}, v_\ell, \ldots, v_1\}$ and $U' = \{u_\ell, \ldots, u_1, i\}$, which proves that

$$S \setminus \{i\} = (B \setminus U') \cup V' \in \mathcal{I}.$$

Similarly, if $j \in B$ with $j \in C(B, v_{\ell+1})$ and $j \notin C(B, v_t)$ for $t < \ell + 1$, taking $V'' = V$ and $U'' = \{u_1, \ldots, u_\ell, j\}$ gives $S \setminus \{j\} \in \mathcal{I}$.

The bottleneck for the running time is finding the shortest paths for the $n(n-1)$ pairs, in time $O(m^2)$ each. □

The triangle inequality An interesting additional fact about the circuit ratio graph is that the logarithm of the weights satisfy the triangle inequality. The proof uses similar arguments as the proof of Theorem 4.2.15 above.

Lemma 4.2.16 (Repetition)**.**

(i) *For any distinct i, j, k in the same connected component of $\mathcal{C}_W$, and any g^C with $i, j \in C$, $C \in \mathcal{C}_W$, there exist circuits $C_1, C_2 \in \mathcal{C}_W$, $i, k \in C_1$, $j, k \in C_2$ such that $|g^C_j / g^C_i| = |g^{C_2}_j / g^{C_2}_k| \cdot |g^{C_1}_k / g^{C_1}_i|$.*

(ii) *For any distinct i, j, k in the same connected component of $\mathcal{C}_W$, $\kappa_{ij} \leq \kappa_{ik} \cdot \kappa_{kj}$.*

Proof. Note that part *(ii)* immediately follows from part *(i)* when taking $C \in \mathcal{C}_W$ such that $\kappa_{ij}(C) = \kappa_{ij}$. We now prove part *(i)*.

Let $A \in \mathbb{R}^{m \times n}$ be a full-rank matrix with $W = \mathrm{Ker}(A)$. If $C = \{i, j\}$, then the columns A_i, A_j are linearly dependent. Writing $A_i = \lambda A_j$, we have $\lambda = -g^C_j / g^C_i$. Let h be any circuit solution with $i, k \in \mathrm{supp}(h)$, and hence $j \notin \mathrm{supp}(h)$. By assumption, the vector $h' = h - h_i e^i + \lambda h_i e^j$ will satisfy $Ah' = \vec{0}$ and have $i \notin \mathrm{supp}(h')$, $j, k \in \mathrm{supp}(h')$. We know that h' is a circuit solution, because any circuit $C' \subseteq \mathrm{supp}(h')$ could, by the above process in reverse, be used to produce a kernel solution with strictly smaller support than h, contradicting the assumption that h is a circuit solution. Now we have $|h'_j / h'_k| \cdot |h_k / h_i| = |h'_j / h_i| = |\lambda|$ by construction. Thus, h and h' are the circuit solutions we are looking for.

Now assume $C \neq \{i, j\}$. If $k \in C$, the statement is trivially true with $C = C_1 = C_2$, so assume $k \notin C$. Pick $l \in C$, $l \notin \{i, j\}$ and set $B = C \setminus \{l\}$. Assume without loss of generality that $B \subseteq [m]$ and apply row operations to A such that $A_{B,B} = I_{B \times B}$ is an identity submatrix and $A_{[n] \setminus B, B} = 0$. Then the column A_l has support given by B, for otherwise g^C could not be in the kernel. The given circuit solution satisfies $g^C_t = -A_{t,l} g^C_l$ for all $t \in B$, and in particular $g^C_j / g^C_i = A_{j,l} / A_{i,l}$.

Take any circuit solution $h \in \mathrm{Ker}(A)$ such that $l, k \in \mathrm{supp}(h)$ and such that $C \cup \mathrm{supp}(h)$ is inclusion-wise minimal. Such a vectors exists by Proposition 4.2.20(iv). Now let $J = \mathrm{supp}(h) \setminus C$. Because $A_{[n]\setminus B,C} = 0$ and $Ah = \vec{0}$, we must have $\vec{0} \neq h_J \in \mathrm{Ker}(A_{[n]\setminus B,J})$. We show that we can uniquely lift any vector $x \in \mathrm{Ker}(A_{B,C\cup\{k\}})$ to a vector $x' \in \mathrm{Ker}(A_{C\cup J})$ with $x'_{C\cup k} = x$. Since this lift will send circuit solutions to circuit solutions by uniqueness, it suffices to find our desired circuits as solutions to the smaller linear system.

We first prove $\dim(\mathrm{Ker}(A_{[n]\setminus B,J})) = 1$. Suppose $\dim(\mathrm{Ker}(A_{[n]\setminus B,J})) \geq 2$, then $|J| \geq 2$ and there would exist some vector $y \in \mathrm{Ker}(A_{[n]\setminus B,J})$ linearly independent from h_J with $k \in \mathrm{supp}(y)$. This vector could be uniquely lifted to a vector $\bar{y} \in \mathrm{Ker}(A)$, and we could then find a linear combination $h + \alpha\bar{y}$ such that $\mathrm{supp}(h + \alpha\bar{y}) \subsetneq C \cup J$ but $l, k \in \mathrm{supp}(h + \alpha\bar{y})$. The existence of such a vector contradicts the minimality of $C \cup \mathrm{supp}(h)$. As such, we know that $\dim(\mathrm{Ker}(A_{[n]\setminus B,J})) = 1$.

This linear relation between any two entries in J for any vector in $\mathrm{Ker}(A_{[n]\setminus B,J})$ implies that we can apply row operations to A such that $A_{B,J}$ has non-zero entries only in the column $A_{B,\{k\}}$. Note that these row operations leave A_C unchanged because $A_{[n]\setminus B,C} = 0$. From this, we can see that any element in $\mathrm{Ker}(A_{B,C\cup\{k\}})$ can be uniquely lifted to an element in $\mathrm{Ker}(A_{C\cup J})$. Hence we can focus on $\mathrm{Ker}(A_{B,C\cup\{k\}})$. If $A_{i,k} = A_{j,k} = 0$, then any $x \in \mathrm{Ker}(A_{B,C\cup\{k\}})$ satisfies $x_i + A_{i,l}x_l = x_j + A_{j,l}x_l = 0$ and, in particular, any circuit $l, k \in \bar{C} \subseteq C \cup \{k\}$ contains $\{i, j\} \subseteq \bar{C}$ and fulfills $|g^C_j / g^C_i| = |A_{j,l}/A_{i,l}| = |g^{\bar{C}}_j / g^{\bar{C}}_i| = |g^{\bar{C}}_j / g^{\bar{C}}_k || g^{\bar{C}}_k / g^{\bar{C}}_i|$. Choosing $C_1 = C_2 = \bar{C}$ concludes the case.

Otherwise we know that $A_{i,k} \neq 0$ or $A_{j,k} \neq 0$, meaning that $\mathrm{Ker}(A_{\{i,j\},\{i,j,l,k\}})$ contains at least one circuit solution with k in its support. Observe that any circuit in $\mathrm{Ker}(A_{\{i,j\},\{i,j,l,k\}})$ can be lifted uniquely to an element in $\mathrm{Ker}(A_{B,C\cup\{k\}})$ since $A_{B,B}$ is an identity matrix and we can set the entries of $B \setminus \{i, j\}$ individually to satisfy the equalities. Note that this lifted vector is a circuit as well, again by uniqueness of the lift. Hence we may restrict our attention to the matrix $A_{\{i,j\},\{i,j,l,k\}}$. If the columns $A_{\{i,j\},k}, A_{\{i,j\},l}$ are linearly dependent, then any circuit solution to $A_{\{i,j\},\{i,j,l\}}x = 0, x_l \neq 0$, such as $g^C_{\{i,j,l\}}$, is easily transformed into a circuit solution to $A_{\{i,j\},\{i,j,k\}}x = \vec{0}, x_k \neq 0$ and we are done.

If $A_{\{i,j\},k}, A_{\{i,j\},l}$ are independent, we can write $A_{\{i,j\},\{i,j,l,k\}} = \left(\begin{smallmatrix} 1 & 0 & a & c \\ 0 & 1 & b & d \end{smallmatrix}\right)$, where $g^C_j / g^C_i = b/a$. For $\alpha = ad - bc$, which is non-zero since $\alpha = \det(\left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)) \neq 0$ by the independence assumption, we can check that $(\alpha, 0, -d, b)$ and $(0, \alpha, c, -a)$ are the circuits we are looking for. □

4.2.6 Approximating $\bar{\chi}$ and $\bar{\chi}^*$

Equipped with Theorem 4.2.13 and Theorem 4.2.15, we are ready to prove Theorem 4.2.5. Recall that we defined $\kappa_{ij}^d := \kappa_{ij}^{\mathrm{Diag}(d)W} = \kappa_{ij} d_j/d_i$ when $d > \vec{0}$. We can similarly define $\hat{\kappa}_{ij}^d := \hat{\kappa}_{ij} d_j/d_i$, and $\hat{\kappa}_{ij}^d$ approximates κ_{ij}^d just as in Theorem 4.2.15.

Theorem 4.2.5 (Repetition)**.** *There is an $O(n^2m^2 + n^3)$ time algorithm that for any matrix $A \in \mathbb{R}^{m\times n}$ computes an estimate ξ of $\bar{\chi}_W$ such that*

$$\xi \le \bar{\chi}_W \le n(\bar{\chi}_W^*)^2 \xi$$

and a $D \in \mathbf{D}$ such that

$$\bar{\chi}_W^* \le \bar{\chi}_{WD} \le n(\bar{\chi}_W^*)^3 .$$

Proof. Let us run the algorithm FINDING-CIRCUITS(A) described in Theorem 4.2.15 to obtain the values $\hat{\kappa}_{ij}$ such that $\hat{\kappa}_{ij} \le \kappa_{ij} \le (\kappa_W^*)^2 \hat{\kappa}_{ij}$. We let $G = ([n], E)$ be the circuit ratio digraph, that is, $(i, j) \in E$ if $\kappa_{ij} > 0$.

To show the first statement on approximating $\bar{\chi}$, we simply set $\xi = \max_{(i,j)\in E} \hat{\kappa}_{ij}$. Then,

$$\xi \le \kappa_W \le \bar{\chi}_W \le n\kappa_W \le n(\kappa_W^*)^2 \xi \le n(\bar{\chi}_W^*)^2 \xi$$

follows by Theorem 4.2.8.

For the second statement on finding a nearly optimal rescaling for $\bar{\chi}_W^*$, we consider the following optimization problem, which is an approximate version of (4.9) from Theorem 4.2.13.

$$\begin{aligned} &\min t \\ \hat{\kappa}_{ij} d_j/d_i \le t \quad &\forall (i, j) \in E \\ d > \vec{0}. & \end{aligned} \tag{4.11}$$

Let $\hat{d}$ be an optimal solution to (4.11) with value $\hat{t}$. We will prove that $\kappa^{\hat{d}} \le (\kappa_W^*)^3$.

First, observe that $\kappa_{ij}^{\hat{d}} = \kappa_{ij}\hat{d}_j/\hat{d}_i \le (\kappa_W^*)^2 \hat{\kappa}_{ij}\hat{d}_j/\hat{d}_i \le (\kappa_W^*)^2 \hat{t}$ for any $(i, j) \in E$. Now, let $d^* > \vec{0}$ be such that $\kappa^{d^*} = \kappa_W^*$. The vector d^* is a feasible solution to (4.11), and so $\hat{t} \le \max_{i\ne j} \hat{\kappa}_{ij} d_j^*/d_i^* \le \max_{i\ne j} \kappa_{ij} d_j^*/d_i^* = \kappa^{d^*}$. Hence we find that $\hat{d}$ gives a rescaling with

$$\bar{\chi}_{W\hat{D}} \le n\kappa^{\hat{d}} \le n(\kappa_W^*)^3 \le n(\bar{\chi}_W)^3 ,$$

where we again used Theorem 4.2.8.

We can obtain the optimal value $\hat{t}$ of (4.11) by solving the corresponding maximum-mean cycle problem (see Theorem 4.2.13). It is easy to develop a multiplicative version of the standard dynamic programming algorithm of the classical

minimum-mean cycle problem (see e.g. [4, Theorem 5.8]) that allows finding the optimum to (4.11) directly, in the same $O(n^3)$ time.

It is left to find the labels $d_i > 0$, $i \in [n]$ such that $\hat{\kappa}_{ij} d_j / d_i \leq \hat{t}$ for all $(i, j) \in E$. We define the following weighted directed graph. We associate the weight $w_{ij} = \log \hat{t} - \log \hat{\kappa}_{ij}$ with every $(i, j) \in E$, and add an extra source vertex r with edges (r, i) of weight $w_{ri} = 0$ for all $i \in [n]$.

By the choice of $\hat{t}$, this graph does not contain any negative weight directed cycles. We can compute the shortest paths from r to all nodes in $O(n^3)$ using the Bellman-Ford algorithm; let σ_i be the shortest path label for i. We then set $d_i = \exp(\sigma_i)$. One can avoid computing logarithms by using a multiplicative variant of the Bellman-Ford algorithm instead.

The running time of the whole algorithm will be bounded by $O(n^2 m^2 + n^3)$. The running time is dominated by the $O(n^2 m^2)$ complexity of FINDING-CIRCUITS(A) and the $O(n^3)$ complexity of solving the minimum-mean cycle problem and shortest path computation. □

4.3 A scaling-invariant layered least squares interior-point algorithm

4.3.1 Preliminaries on interior-point methods

In this section, we introduce the standard definitions, concepts and results from the interior-point literature that will be required for our algorithm. We consider an LP problem in the form (4.1), or equivalently, in the subspace form (4.3) for $W = \mathrm{Ker}(A)$. We let

$$\mathcal{P}^{++} = \{x \in \mathbb{R}^n : Ax = b, x > \vec{0}\}, \quad \mathcal{D}^{++} = \{(y, s) \in \mathbb{R}^{m+n} : A^\mathsf{T} y + s = c, s > \vec{0}\}.$$

Recall the *central path* defined in (CP), with $w(\mu) = (x(\mu), y(\mu), s(\mu))$ denoting the central path point corresponding to $\mu > 0$. We let $w^* = (x^*, y^*, s^*)$ denote the primal and dual optimal solutions to (4.1) that correspond to the limit of the central path for $\mu \to 0$.

For a point $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, the *normalized duality gap* is $\mu(w) = x^\mathsf{T} s / n$.

The *ℓ_2-neighborhood of the central path with opening $\beta > 0$* is the set

$$\mathcal{N}(\beta) = \left\{ w \in \mathcal{P}^{++} \times \mathcal{D}^{++} : \left\| \frac{xs}{\mu(w)} - \vec{1} \right\| \leq \beta \right\}$$

Throughout the chapter, we will assume β is chosen from $(0, 1/4]$; in Algorithm 4 we use the value $\beta = 1/8$. The following proposition gives a bound on the distance

between w and $w(\mu)$ if $w \in \mathcal{N}(\beta)$. See e.g. [100, Lemma 5.4], [146, Proposition 2.1].

Proposition 4.3.1. *Let $w = (x, y, s) \in \mathcal{N}(\beta)$ for $\beta \in (0, 1/4]$ and $\mu = \mu(w)$, and consider the central path point $w(\mu) = (x(\mu), y(\mu), s(\mu))$. For each $i \in [n]$,*

$$\frac{x_i}{1+2\beta} \leq \frac{1-2\beta}{1-\beta} \cdot x_i \leq x_i(\mu) \leq \frac{x_i}{1-\beta}, \quad \text{and}$$
$$\frac{s_i}{1+2\beta} \leq \frac{1-2\beta}{1-\beta} \cdot s_i \leq s_i(\mu) \leq \frac{s_i}{1-\beta}.$$

We will often use the following proposition which is immediate from definiton of β.

Proposition 4.3.2. *Let $w = (x, y, s) \in \mathcal{N}(\beta)$ for $\beta \in (0, 1/4]$, and $\mu = \mu(w)$. Then for each $i \in [n]$*

$$(1-\beta)\sqrt{\mu} \leq \sqrt{s_i x_i} \leq (1+\beta)\sqrt{\mu}\,.$$

Proof. By definition of $\mathcal{N}(\beta)$ we have for all $i \in [n]$ that $|\frac{x_i s_i}{\mu} - 1| \leq \|\frac{xs}{\mu} - \vec{1}\| \leq \beta$ and so $(1-\beta)\mu \leq x_i s_i \leq (1+\beta)\mu$. Taking roots gives the results. □

A key property of the central path is *"near monotonicity"*, formulated in the following lemma, see [198, Lemma 16].

Lemma 4.3.3. *Let $w = (x, y, s)$ be a central path point for μ and $w' = (x', y', s')$ be a central path point for $\mu' \leq \mu$. Then $\|x'/x + s'/s\|_\infty \leq n$. Further, for the optimal solution $w^* = (x^*, y^*, s^*)$ corresponding to the central path limit $\mu \to 0$, we have $\|x^*/x\|_1 + \|s^*/s\|_1 = n$.*

Proof. We show that $\|x'/x\|_1 + \|s'/s\|_1 \leq 2n$ for any feasible primal x' and dual (y', s') such that $(x')^\mathsf{T} s' \leq x^\mathsf{T} s = n\mu$; this implies the first statement with the weaker bound $2n$. For the stronger bound $\|x'/x + s'/s\|_\infty \leq n$, see the proof of [198, Lemma 16]. Since $x - x' \in W$ and $s - s' \in W^\perp$, we have $(x - x')^\mathsf{T}(s - s') = \vec{0}$. This can be rewritten as $x^\mathsf{T} s' + (x')^\mathsf{T} s = x^\mathsf{T} s + (x')^\mathsf{T} s'$. By our assumption on x' and s', the right hand side is bounded by $2n\mu$. Dividing by μ, and noting that $x_i s_i = \mu$ for all $i \in [n]$, we obtain

$$\left\|\frac{x'}{x}\right\|_1 + \left\|\frac{s'}{s}\right\|_1 = \sum_{i=1}^{n} \frac{x_i'}{x_i} + \frac{s_i'}{s_i} \leq 2n\,.$$

The second statement follows by using this to central path points (x', y', s') with parameter μ', and taking the limit $\mu' \to 0$. □

4.3.2 The affine scaling and layered-least-squares steps

Given $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, the search directions commonly used in interior-point methods are obtained as the solution $(\Delta x, \Delta y, \Delta s)$ to the following linear system for some $\sigma \in [0, 1]$.

$$A\Delta x = \vec{0} \tag{4.12}$$

$$A^{\mathsf{T}}\Delta y + \Delta s = \vec{0} \tag{4.13}$$

$$s\Delta x + x\Delta s = \sigma\mu\vec{1} - xs \tag{4.14}$$

Predictor-corrector methods, such as the Mizuno-Todd-Ye Predictor-Corrector (MTY P-C) algorithm [145], alternate between two types of steps. In *predictor steps*, we use $\sigma = 0$. This direction is also called the *affine scaling direction*, and will be denoted as $\Delta w^{\mathrm{a}} = (\Delta x^{\mathrm{a}}, \Delta y^{\mathrm{a}}, \Delta s^{\mathrm{a}})$ throughout. In *corrector steps*, we use $\sigma = 1$. This gives the *centrality direction*, denoted as $\Delta w^{\mathrm{c}} = (\Delta x^{\mathrm{c}}, \Delta y^{\mathrm{c}}, \Delta s^{\mathrm{c}})$.

In the predictor steps, we make progress along the central path. Given the search direction on the current iterate $w = (x, y, s) \in \mathcal{N}(\beta)$, the step-length is chosen maximal such that we remain in $\mathcal{N}(2\beta)$, i.e.

$$\alpha^{\mathrm{a}} := \sup\{\alpha \in [0, 1] \ : \forall \alpha' \in [0, \alpha] : w + \alpha'\Delta w^{\mathrm{a}} \in \mathcal{N}(2\beta)\}.$$

Thus, we obtain a point $w^+ = w + \alpha^{\mathrm{a}}\Delta w^{\mathrm{a}} \in \mathcal{N}(2\beta)$. The corrector step finds a next iterate $w^c = w^{\mathrm{a}} + \Delta w^{\mathrm{c}}$, where Δw^{c} is the centrality direction computed at w^{a}. The next proposition summarizes well-known properties, see e.g. [202, Section 4.5.1].

Proposition 4.3.4. *Let $w = (x, y, s) \in \mathcal{N}(\beta)$ for $\beta \in (0, 1/4]$.*

(i) *For the affine scaling step, we have $\mu(w^+) = (1 - \alpha)\mu(w)$.*

(ii) *The affine scaling step-length is*

$$\alpha^{\mathrm{a}} \geq \max\left\{\frac{\beta}{\sqrt{n}}, 1 - \frac{\|\Delta x^{\mathrm{a}}\Delta s^{\mathrm{a}}\|}{\beta\mu(w)}\right\}.$$

(iii) *For $w^+ \in \mathcal{N}(2\beta)$, and $w^{\mathrm{c}} = w^+ + \Delta w^{\mathrm{c}}$, we have $\mu(w^{\mathrm{c}}) = \mu(w^+)$ and $w^{\mathrm{c}} \in \mathcal{N}(\beta)$.*

(iv) *After a sequence of $O(\sqrt{n}t)$ predictor and corrector steps, we obtain an iterate $w' = (x', y', s') \in \mathcal{N}(\beta)$ such that $\mu(w') \leq \mu(w)/2^t$.*

Minimum norm viewpoint and residuals For any point $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ we define

$$\delta = \delta(w) = s^{1/2} x^{-1/2} \in \mathbb{R}^n. \tag{4.15}$$

With this notation, we can write (4.14) for $\sigma = 0$ in the form

$$\delta \Delta x + \delta^{-1} \Delta s = -s^{1/2} x^{1/2}. \tag{4.16}$$

Note that for a point $w(\mu) = (x(\mu), y(\mu), s(\mu))$ on the central path, we have $\delta_i(w(\mu)) = s_i(\mu)/\sqrt{\mu} = \sqrt{\mu}/x_i(\mu)$ for all $i \in [n]$. From Proposition 4.3.1, we see that if $w \in \mathcal{N}(\beta)$, and $\mu = \mu(w)$, then for each $i \in [n]$,

$$\sqrt{1 - 2\beta} \cdot \delta_i(w(\mu)) \leq \delta_i(w) \leq \frac{1}{\sqrt{1 - 2\beta}} \cdot \delta_i(w(\mu)). \tag{4.17}$$

The matrix $\mathrm{Diag}(\delta(w))$ will be often used for rescaling in the algorithm. That is, for the current iterate $w = (x, y, s)$ in the interior-point method, we will perform projections in the space $\mathrm{Diag}(\delta(w))W$. To simplify notation, for $\delta = \delta(w)$, we use L_I^{δ} and κ_{ij}^{δ} as shorthands for $L_I^{\mathrm{Diag}(\delta)W}$ and $\kappa_{ij}^{\mathrm{Diag}(\delta)W}$. The subspace $W = \mathrm{Ker}(A)$ will be fixed throughout.

It is easy to see from the optimality conditions that the components of the affine scaling direction $\Delta w^{\mathrm{a}} = (\Delta x^{\mathrm{a}}, \Delta y^{\mathrm{a}}, \Delta s^{\mathrm{a}})$ are the optimal solutions of the following minimum-norm problems.

$$\begin{aligned} \Delta x^{\mathrm{a}} &= \arg\min_{\Delta x \in \mathbb{R}^n} \{\|\delta(x + \Delta x)\|^2 : A\Delta x = \vec{0}\} \\ (\Delta y^{\mathrm{a}}, \Delta s^{\mathrm{a}}) &= \arg\min_{(\Delta y, \Delta s) \in \mathbb{R}^m \times \mathbb{R}^n} \{\|\delta^{-1}(s + \Delta s)\|^2 : A^{\mathsf{T}} \Delta y + \Delta s = \vec{0}\} \end{aligned} \tag{4.18}$$

Following [147], for a search direction $\Delta w = (\Delta x, \Delta y, \Delta s)$, we define the *residuals* as

$$Rx := \frac{\delta(x + \Delta x)}{\sqrt{\mu}}, \qquad Rs := \frac{\delta^{-1}(s + \Delta s)}{\sqrt{\mu}}. \tag{4.19}$$

We let Rx^{a} and Rs^{a} denote the residuals for the affine scaling direction Δw^{a}. Hence, the primal affine scaling direction Δx^{a} is the one that minimizes the ℓ_2-norm of the primal residual Rx^{a}, and the dual affine scaling direction $(\Delta y^{\mathrm{a}}, \Delta s^{\mathrm{a}})$ minimizes the ℓ_2-norm of the dual residual Rs^{a}. The next lemma summarizes simple properties of the residuals, see [147].

Lemma 4.3.5. *For $\beta \in (0, 1/4]$ such that $w = (x, y, s) \in \mathcal{N}(\beta)$ and the affine scaling direction $\Delta w = (\Delta x^{\mathrm{a}}, \Delta y^{\mathrm{a}}, \Delta s^{\mathrm{a}})$, we have*

(i)

$$Rx^{\mathrm{a}}Rs^{\mathrm{a}} = \frac{\Delta x^{\mathrm{a}}\Delta s^{\mathrm{a}}}{\mu}, \quad Rx^{\mathrm{a}} + Rs^{\mathrm{a}} = \frac{x^{1/2}s^{1/2}}{\sqrt{\mu}}, \tag{4.20}$$

(ii)

$$\|Rx^{\mathrm{a}}\|^2 + \|Rs^{\mathrm{a}}\|^2 = n,$$

(iii) *We have* $\|Rx^{\mathrm{a}}\|, \|Rs^{\mathrm{a}}\| \le \sqrt{n}$, *and for each* $i \in [n]$, $\max\{Rx_i^{\mathrm{a}}, Rs_i^{\mathrm{a}}\} \ge \frac{1}{2}(1-\beta)$.

(iv)

$$Rx^{\mathrm{a}} = -\frac{1}{\sqrt{\mu}}\delta^{-1}\Delta s^{\mathrm{a}}, \quad Rs^{\mathrm{a}} = -\frac{1}{\sqrt{\mu}}\delta\Delta x^{\mathrm{a}}.$$

Proof. Parts (i) and (iv) are immediate from the definitions and from (4.12)-(4.14) and (4.16). In part (ii), we use part (i) and $(Rx^{\mathrm{a}})^{\mathsf{T}}Rs^{\mathrm{a}} = 0$. In part, (iii), the first statement follows by part (ii), and the second statement follows from (i) and Proposition 4.3.2. □

For a subset $I \subseteq [n]$, we define

$$\varepsilon_I^{\mathrm{a}}(w) := \max_{i\in I} \min\{|Rx_i^{\mathrm{a}}|, |Rs_i^{\mathrm{a}}|\}, \quad \text{and} \quad \varepsilon^{\mathrm{a}}(w) := \varepsilon_{[n]}^{\mathrm{a}}(w). \tag{4.21}$$

The next claim shows that for the affine scaling direction, a small $\varepsilon(w)$ yields a long step; see [147, Lemma 2.5].

Lemma 4.3.6. *Let* $w = (x, y, s) \in \mathcal{N}(\beta)$ *for* $\beta \in (0, 1/4]$. *Then for the affine scaling step, we have*

$$\frac{\mu(w + \alpha^{\mathrm{a}}\Delta w^{\mathrm{a}})}{\mu(w)} \le \min\left\{1 - \frac{\beta}{\sqrt{n}}, \frac{\sqrt{n}\varepsilon^{\mathrm{a}}(w)}{\beta}\right\}.$$

Proof. Let $\varepsilon := \varepsilon^{\mathrm{a}}(w)$. From Lemma 4.3.5(i), we get $\|\Delta x^{\mathrm{a}}\Delta s^{\mathrm{a}}\|/\mu = \|Rx^{\mathrm{a}}Rs^{\mathrm{a}}\|$. We can bound $\|Rx^{\mathrm{a}}Rs^{\mathrm{a}}\| \le \varepsilon(\|Rx^{\mathrm{a}}\| + \|Rs^{\mathrm{a}}\|) \le \varepsilon\sqrt{n}$, where the latter inequality follows by Lemma 4.3.5(iii). From Proposition 4.3.4(ii), we get $\alpha^{\mathrm{a}} \ge \max\{\beta/\sqrt{n}, 1-\sqrt{n}\varepsilon/\beta\}$. The claim follows by part (i) of the same proposition. □

The layered-least-squares direction

Let $\mathcal{J} = (J_1, J_2, \ldots, J_p)$ be an *ordered partition* of $[n]$.[2] For $k \in [p]$, we use the notations $J_{<k} := J_1 \cup \ldots \cup J_{k-1}$, $J_{>k} := J_{k+1} \cup \ldots \cup J_p$, and similarly $J_{\le k}$ and $J_{\ge k}$.

[2] In contrast to how ordered partitions were defined in [147], we use the term *ordered* only to the p-tuple $(J_1, \ldots, J_p)$, which is to be viewed independently of δ.

We will also refer to the sets J_k as *layers*, and $\mathcal{J}$ as a *layering*. Layers with lower indices will be referred to as 'higher' layers.

Given $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, and the layering $\mathcal{J}$, the *layered-least-squares (LLS) direction* is defined as follows. For the primal direction, we proceed backwards, with $k = p, p-1, \ldots, 1$. Assume the components on the lower layers $\Delta x^{\text{ll}}_{J_{>k}}$ have already been determined. We define the components in J_k as the coordinate projection $\Delta x^{\text{ll}}_{J_k} = \pi_{J_k}(X_k)$, where the affine subspace X_k is defined as the set of minimizers

$$X_k := \operatorname*{arg\,min}_{\Delta x \in \mathbb{R}^n} \left\{ \left\| \delta_{J_k}(x_{J_k} + \Delta x_{J_k}) \right\|^2 : A\Delta x = \vec{0}, \Delta x_{J_{>k}} = \Delta x^{\text{ll}}_{J_{>k}} \right\} . \tag{4.22}$$

The dual direction Δs^{ll} is determined in the forward order of the layers $k = 1, 2, \ldots, p$. Assume we have already fixed the components $\Delta s^{\text{ll}}_{J_{<k}}$ on the higher layers. Then, $\Delta s^{\text{ll}}_{J_k} = \pi_{J_k}(S_k)$ for

$$S_k = \operatorname*{arg\,min}_{\Delta s \in \mathbb{R}^n} \left\{ \left\| \delta^{-1}_{J_k}(s_{J_k} + \Delta s_{J_k}) \right\|^2 : \exists y \in \mathbb{R}^m, A^{\mathsf{T}}\Delta y + \Delta s = \vec{0}, \Delta s_{J_{<k}} = \Delta s^{\text{ll}}_{J_{<k}} \right\} . \tag{4.23}$$

The component Δy^{ll} is obtained as the optimal Δy for the final layer $k = p$. We use the notation Rx^{ll} and $\varepsilon^{\text{ll}}(w)$ analogously to the affine scaling direction. This search direction was first introduced in [198].

The affine scaling direction is a special case for the single element partition. In this case, the definitions (4.22) and (4.23) coincide with those in (4.18).

4.3.3 Overview of ideas and techniques

A key technique in the analysis of layered least-squares algorithms [129, 146, 198] is to argue about variables that have 'converged'. According to Proposition 4.3.1 and Lemma 4.3.3, for any iterate $w = (x, y, s) \in \mathcal{N}(\beta)$ and the limit optimal solution $w^* = (x^*, y^*, s^*)$, the bounds $x^*_i \le O(n)x_i$ and $s^*_i \le O(n)s_i$ hold. We informally say that x_i (or s_i) has *converged*, if $x_i \le O(n)x^*_i$ ($s_i \le O(n)s^*_i$) hold for the current iterate or for any earlier iterate. Thus, the value of x_i (or s_i) remains within a multiplicative factor $O(n^2)$ for the rest of the algorithm. Note that if $\mu > \mu'$ and x_i has converged at μ, then $\frac{s_i(\mu')/s_i(\mu)}{\mu'/\mu} \in \left[\frac{1}{O(n^2)}, O(n^2)\right]$; thus, s_i keeps "shooting down" with the central path parameter.

Converged variables in the affine scaling algorithm Let us start by showing that at any point of the algorithm, at least one primal or dual variable has converged.

Suppose for simplicity that our current iterate is exactly on the central path, i.e., that $xs = \mu\vec{1}$. This assumption will be maintained throughout this overview. In this case, the residuals can be simply written as $Rx^{\mathrm{a}} = (x + \Delta x^{\mathrm{a}})/x$, $Rs^{\mathrm{a}} = (s + \Delta s^{\mathrm{a}})/s$. Recall from (4.18) that the affine scaling direction corresponds to minimizing the residuals Rx^{a} and Rs^{a}. From this choice, we see that

$$\left\|\frac{x^*}{x}\right\| \geq \left\|\frac{x + \Delta x^{\mathrm{a}}}{x}\right\|, \quad \left\|\frac{s^*}{s}\right\| \geq \left\|\frac{s + \Delta s^{\mathrm{a}}}{s}\right\|. \tag{4.24}$$

We have $\|Rx^{\mathrm{a}}\|^2 + \|Rs^{\mathrm{a}}\|^2 = n$ by Lemma 4.3.5(ii). Let us assume $\|Rx^{\mathrm{a}}\|^2 \geq n/2$; thus, there exists a $i \in [n]$ such that $x_i^* \geq x_i/\sqrt{2}$. In other words, just by looking at the residuals, we get the guarantee that a primal or a dual variable has already converged. Based on the value of the residuals, we can guarantee this to be a primal or a dual variable, but cannot identify which particular x_i or s_i this might be.

For $\|Rx^{\mathrm{a}}\|^2 \geq n/2$, a primal variable has already converged *before* performing the predictor and corrector steps. We now show that even if $\|Rx^{\mathrm{a}}\|$ is small, a primal variable will have converged *after* a single iteration. From (4.24), we see that there is an index i with $x_i^*/x_i \geq \|Rx^{\mathrm{a}}\|/\sqrt{n}$. Furthermore, Proposition 4.3.4(ii) and Lemma 4.3.5 imply that $1 - \alpha \leq \|Rx^{\mathrm{a}}\| \cdot \|Rs^{\mathrm{a}}\|/\beta \leq \sqrt{n}\|Rx^{\mathrm{a}}\|/\beta$, since $\|Rs^{\mathrm{a}}\| \leq \sqrt{n}$. The predictor step moves to $x^+ := x + \alpha\Delta x^{\mathrm{a}} = (1 - \alpha)x + \alpha(x + \Delta x^{\mathrm{a}})$. Hence, $x^+ \leq \left(\frac{\sqrt{n}\|Rx^{\mathrm{a}}\|}{\beta} + \|Rx^{\mathrm{a}}\|\right) x$. Putting the two inequalities together, we learn that $x_i^+ \leq O(n)x_i^*$ for some $i \in [n]$. Since $w^+ = (x^+, y^+, s^+) \in \mathcal{N}(2\beta)$, Proposition 4.3.1 implies that x_i will have converged after this iteration. An analogous argument proves that some s_j will also have converged after the iteration. We again emphasize that the argument only shows the existence of converged variables but we cannot identify them in general.

Measuring combinatorial progress Tying the above together, we find that after a single affine scaling step, at least one primal variable x_i and at least one dual variable s_j has converged. This means that for any $\mu' < \mu$, $\frac{x_i(\mu')/x_j(\mu')}{x_i(\mu)/x_j(\mu)} \in \left[\frac{\mu}{O(n^4)\mu'}, \frac{O(n^4)\mu}{\mu'}\right]$; thus, the ratio of these variables keeps asymptotically increasing. The x_i/x_j ratios serve as the main progress measure in the Vavasis–Ye algorithm. If x_i/x_j is between $1/(\mathrm{poly}(n)\bar{\chi})$ and $\mathrm{poly}(n)\bar{\chi}$ before the affine scaling step for the pair of converged variables x_i and s_j, then after $\mathrm{poly}(n)\log\bar{\chi}$ iterations, the x_i/x_j ratio must leave this interval and never return. Thus, we obtain a 'crossover-event' that cannot again occur for the same pair of variables. In the affine scaling algorithm, there is no guarantee that x_i/x_j falls in such a bounded interval for the converging variables x_i and s_j; in particular, we may obtain the same pairs of converged variables after each step.

The main purpose of layered-least-squares methods is to proactively force that in every certain number of iterations, some 'bounded' x_i/x_j ratios become 'large' and remain so for the rest of the algorithm.

In our approach, the first main insight is to focus on the scaling invariant quantities $\kappa_{ij}^W x_i/x_j$ instead. For simplicity's sake, we first present the algorithm with the assumption that all values κ_{ij}^W are known. We will then explain how this assumption can be removed by using gradually improving estimates on the values.

The combinatorial progress will be observed in the *'long edge graph'*. For a primal-dual feasible point $w = (x, y, x)$ and $\sigma = 1/O(n^6)$, this is defined as $G_{w,\sigma} = ([n], E_{w,\sigma})$ with edges (i, j) such that $\kappa_{ij}^W x_i/x_j \geq \sigma$. Observe that for any $i, j \in [n]$, at least one of (i, j) and (j, i) are long edges: this follows since for any circuit C with $i, j \in C$, we get lower bounds $|g_j^C/g_i^C| \leq \kappa_{ij}^W$ and $|g_i^C/g_j^C| \leq \kappa_{ji}^W$.

Intuitively, our algorithm will enforce the following two types of events. The analysis in Section 4.4 is based on a potential function analysis capturing roughly the same progress.

- For an iterate w and a value $\mu > 0$, we have $i, j \in [n]$ in a strongly connected component in $G_{w,\sigma}$ of size $\leq \tau$, and for any iterate w' with $\mu(w') > \mu$, if i, j are in a strongly connected component of $G_{w',\sigma}$ then this component has size $\geq 2\tau$.
- For an iterate w and a value $\mu > 0$, we have $(i, j) \notin E_{w,\sigma}$, and for any iterate w' with $\mu(w') > \mu$ we have $(i, j) \in E_{w',\sigma}$.

At most $O(n^2 \log n)$ such events can happen overall, so if we can prove that on average an event will happen every $O(\sqrt{n}\log(\bar{\chi}_A^* + n))$ iterations or the algorithm terminates, then we have the desired convergence bound of $O(n^{2.5}\log(n)\log(\bar{\chi}_A^* + n))$ iterations.

Converged variables cause combinatorial progress We now show that combinatorial progress as above must happen in the affine scaling step in the case when the graph $G_{w,\sigma}$ is strongly connected. As noted above, for the pair of converged variables x_i and s_j after the affine scaling step, x_i/x_j, and thus $\kappa_{ij}^W x_i/x_j$, will asymptotically increase by a factor 2 in every $O(\sqrt{n})$ iterations.

By the strong connectivity assumption, there is a directed path in the long edge graph from i to j of length at most $n-1$. Each edge has length at least σ, and by the cycle characterization (Theorem 4.2.13) we know that $(\kappa_{ji}^W x_j/x_i) \cdot \sigma^{n-1} \leq (\kappa_W^*)^n$. As such, $\kappa_{ji}^W x_j/x_i \leq (\kappa_W^*)^n/\sigma^{n-1}$. Since $\kappa_{ij}^W \kappa_{ji}^W \leq (\kappa_W^*)^2$ by the same theorem, we obtain the lower bound $\kappa_{ij}^W x_i/x_j \geq \sigma^{n-1}(\kappa_W^*)^{-n+2}$.

This means that after $O(\sqrt{n}\log((\kappa_W^*/\sigma)^n)) = O(n^{1.5}\log(\kappa_W^* + n))$ affine scaling steps, the weight of the edge (i, j) will be more than $(\kappa_W^*/\sigma)^{4n}$. There can never

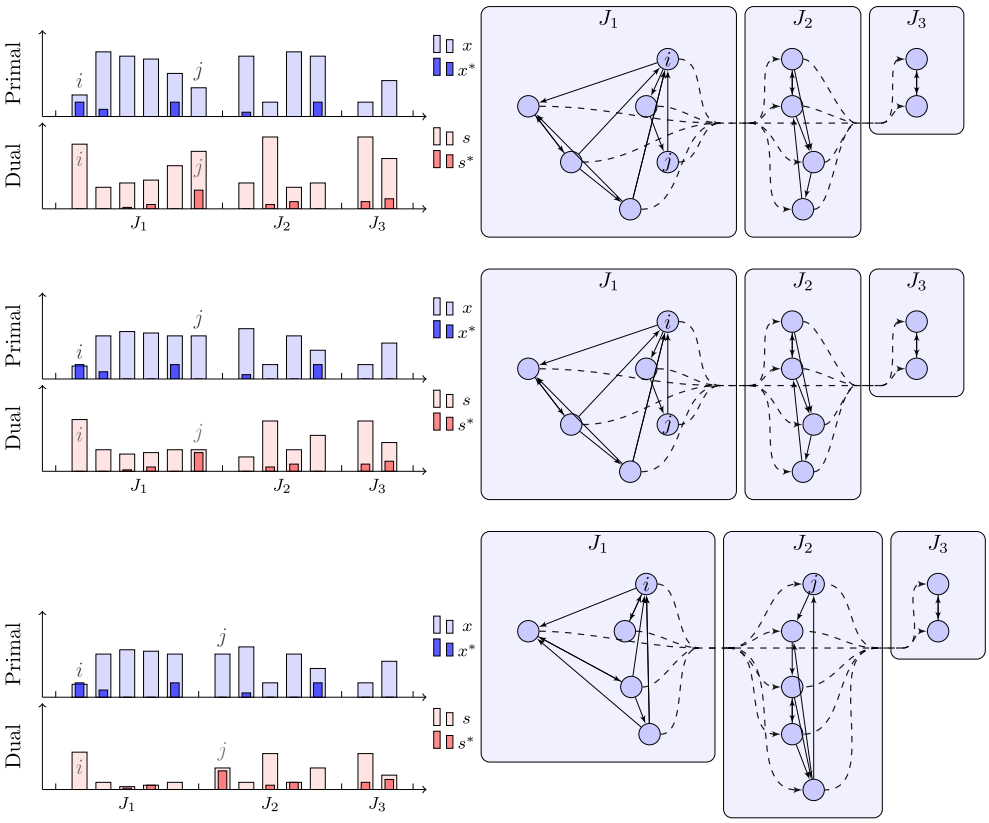


Figure 4.1: Top-down we have a chart of primal/dual variables and the *estimated* subgraph of the circuit ratio digraph (Definition 4.3.11) for three different iterations: 1) All variables are far away from their optimal values. 2) On J_1 there is a primal variable (i) and dual variable (j) that have converged, i.e. x_i is close to x_i^* and s_i is close to s_i^*. 3) j moves to layer J_2 due to a change in the underlying subgraph of the circuit ratio digraph.

again be a length n or shorter path from j to i in the long edge graph, for otherwise the resulting cycle would violate Theorem 4.2.13. Moreover, by the triangle inequality (Lemma 4.2.16), any other $k \neq i, j$ will have either (i, k) or (k, j) of length at least $(\kappa_W^*/\sigma)^{2n}$, similarly causing a pair of variables to never again be in the same connected component. As such, we took $O(n^{1.5} \log(\kappa_W^* + n))$ affine scaling steps and in that time at least $n - 1$ combinatorial progress events have occured.

The layered least squares step Similarly to the Vavasis–Ye algorithm [198] and subsequent literature, our algorithm is a predictor-corrector method using *layered*

least squares (LLS) steps as in Section 4.3.2 for certain predictor iterations. Our algorithm (Algorithm 4) uses LLS steps only sometimes, and most steps are the simpler affine scaling steps; but for simplicity of this overview, we can assume every predictor iteration uses an LLS step.

We define the ordered partition $\mathcal{J} = (J_1, J_2, \ldots, J_p)$ corresponding to the strongly connected components in topological ordering. Recalling that either (i, j) or (j, i) is a long edge for every pair $i, j \in [n]$, this order is unique and such that there is a complete directed graph of long edges from every J_k to $J_{k'}$ for $1 \leq k < k' \leq p$.

The first important property of the LLS step is that it is very close to the affine scaling step. In Section 4.3.4, we introduce the partition lifting cost $\ell^W(\mathcal{J}) = \max_{2 \leq k \leq p} \ell^W(J_{\geq k})$ as the cost of lifting from lower to higher layers; we let $\ell^{1/x}(\mathcal{J})$ be a shorthand for $\ell^{\mathrm{Diag}(1/x)W}(\mathcal{J})$. Note that this same rescaling is used for the affine scaling step in (4.18), since $\delta = \sqrt{\mu}/x$ if w is on the central path. In Lemma 4.3.10(ii), we show that for a small partition lifting cost, the LLS residuals will remain near the affine scaling residuals. Namely,

$$\|Rx^{\mathrm{ll}} - Rx^{\mathrm{a}}\|, \|Rs^{\mathrm{ll}} - Rs^{\mathrm{a}}\| \leq 6n^{3/2}\ell^{1/x}(\mathcal{J}) \,.$$

Recall that the LLS residuals can be written as $Rx^{\mathrm{ll}} = (x + \Delta x^{\mathrm{ll}})/x$, $Rs^{\mathrm{ll}} = (s + \Delta s^{\mathrm{ll}})/s$ for a point on the central path. For $\mathcal{J}$ defined as above, Lemma 4.2.11 yields $\ell^{1/x}(\mathcal{J}) \leq n \max_{i \in J_{>k}, j \in J_{\leq k}, k \in [p]} \kappa^W_{ij} x_i/x_j$. This will be sufficiently small as this maximum is taken over 'short' edges (not in $E_{w,\sigma}$).

A second, crucial property of the LLS step is that it "splits" our LP into p separate LPs that have "negligible" interaction. Namely, the direction $(\Delta x^{\mathrm{ll}}_{J_k}, \Delta s^{\mathrm{ll}}_{J_k})$ will be very close to the affine scaling step obtained in the problem restricted to the subspace $W_{\mathcal{J},k} = \{x_{J_k} : x \in W, x_{J_{>k}} = \vec{0}\}$ (Lemma 4.3.10(i))

Since each component J_k is strongly connected in the long edge graph $G_{w,\sigma}$, if there is at least one primal x_i and dual s_j in J_k that have converged after the LLS step, we can use the above argument to show combinatorial progress regarding the $\kappa^W_{ij} x_i/x_j$ value (Lemma 4.4.3).

Exploiting the proximity between LLS and affine scaling steps, Lemma 4.3.10(iv) gives a lower bound on the step size $\alpha \geq 1 - \frac{3\sqrt{n}}{\beta} \max_{i \in [n]} \min\{|Rx^{\mathrm{ll}}_i|, |Rs^{\mathrm{ll}}_i|\}$. Let J_k be the component where $\min\{\|Rx^{\mathrm{ll}}_{J_k}\|, \|Rs^{\mathrm{ll}}_{J_k}\|\}$ is the largest. Hence, the step size α can be lower bounded in terms of $\min\{\|Rx^{\mathrm{ll}}_{J_k}\|, \|Rs^{\mathrm{ll}}_{J_k}\|\}$.

The analysis now distinguishes two cases. Let $w^+ = w + \alpha\Delta s^{\mathrm{ll}}$ be the point obtained by the predictor LLS step. If the corresponding partition lifting cost $\ell^{1/x^+}(\mathcal{J})$ is still small, then a similar argument that has shown the convergence of primal and dual variables in the affine scaling step will imply that after the LLS step, at least one x_i and one s_j will have converged for $i, j \in J_k$. Thus, in this case we obtain the combinatorial progress (Lemma 4.4.4).

The remaining case is when $\ell^{1/x^+}(\mathcal{J})$ becomes large. In Lemma 4.4.5, we show that in this case a new edge will enter the long edge graph, corresponding to the second combinatorial event listed previously. Intuitively, in this case one layer "crashes" into another.

Refined estimates on circuit imbalances In the above overview, we assumed the circuit imbalance values κ_{ij}^W are given, and thus the graph $G_{w,\sigma}$ is available. Whereas these quantities are difficult to compute, we can naturally work with lower estimates. For each $i, j \in [n]$ that are contained in a circuit together, we start with the lower bound $\hat{\kappa}_{ij}^W = |g_j^C / g_i^C|$ obtained for an arbitrary circuit C with $i, j \in C$. We use the graph $\hat{G}_{w,\sigma} = ([n], \hat{E}_{w,\sigma})$ corresponding to these estimates. Clearly, $\hat{E}_{w,\sigma} \subseteq E_{w,\sigma}$, but some long edges may be missing. We determine the partition $\mathcal{J}$ of the strongly connected components of $\hat{G}_{w,\sigma}$ and estimate the partition lifting cost $\ell^{1/x}(\mathcal{J})$. If this is below the desired bound, the argument works correctly. Otherwise, we can identify a pair i, j responsible for this failure. Namely, we find a circuit C with $i, j \in C$ such that $\hat{\kappa}_{ij}^W < |g_j^C / g_i^C|$. In this case, we update our estimate, and recompute the partition; this is described in Algorithm 3. At each LLS step, the number of updates is bounded by n, since every update leads to a decrease in the number of partition classes. This finishes the overview of the algorithm.

4.3.4 A linear system viewpoint of layered least squares

We now continue with the detailed exposition of our algorithm. We present an equivalent definition of the LLS step introduced in Section 4.3.2, generalizing the linear system (4.13)–(4.14). We use the subspace notation. With this notation, (4.13)–(4.14) for the affine scaling direction can be written as

$$s\Delta x^{\mathrm{a}} + x\Delta s^{\mathrm{a}} = -xs\,, \quad \Delta x^{\mathrm{a}} \in W\,, \quad \text{and} \quad \Delta s^{\mathrm{a}} \in W^{\perp}\,, \tag{4.25}$$

which is further equivalent to $\delta\Delta x^{\mathrm{a}} + \delta^{-1}\Delta s^{\mathrm{a}} = -x^{1/2}s^{1/2}$.

Given the layering $\mathcal{J}$ and $w = (x, y, s)$, for each $k \in [p]$ we define the subspaces

$$W_{\mathcal{J},k} := \{x_{J_k} : x \in W, x_{J_{>k}} = \vec{0}\} \quad \text{and} \quad W^{\perp}_{\mathcal{J},k} := \{x_{J_k} : x \in W^{\perp}, x_{J_{<k}} = \vec{0}\}\,.$$

It is easy to see that these two subspaces are orthogonal complements. Our next goal is to show that, analogously to (4.25), the primal LLS step Δx^{ll} is obtained as the unique solution to the linear system

$$\delta\Delta x^{\mathrm{ll}} + \delta^{-1}\Delta s = -x^{1/2}s^{1/2}\,, \quad \Delta x^{\mathrm{ll}} \in W\,, \quad \text{and} \quad \Delta s \in W^{\perp}_{\mathcal{J},1} \oplus \cdots \oplus W^{\perp}_{\mathcal{J},p}\,, \tag{4.26}$$

and the dual LLS step Δs^{ll} is the unique solution to

$$\delta\Delta x + \delta^{-1}\Delta s^{\mathrm{ll}} = -x^{1/2}s^{1/2}\,, \quad \Delta x \in W_{\mathcal{J},1} \oplus \cdots \oplus W_{\mathcal{J},p}\,, \quad \text{and} \quad \Delta s^{\mathrm{ll}} \in W^{\perp}\,. \tag{4.27}$$

It is important to note that Δs in (4.26) may be different from Δs^{ll}, and Δx in (4.27) may be different from Δx^{ll}. In fact, $\Delta s^{\text{ll}} = \Delta s$ and $\Delta x^{\text{ll}} = \Delta x$ can only be the case for the affine scaling step.

The following lemma proves that the above linear systems are indeed uniquely solved by the LLS step.

Lemma 4.3.7. *For $t \in \mathbb{R}^n$, $W \subseteq \mathbb{R}^n$, $\delta \in \mathbb{R}^n_{++}$, and $\mathcal{J} = (J_1, J_2, \ldots, J_p)$, let $w = \text{LLS}^{W,\delta}_{\mathcal{J}}(t)$ be defined by*

$$\delta w + \delta^{-1} v = \delta t, \qquad w \in W, \qquad v \in W^{\perp}_{\mathcal{J},1} \oplus \cdots \oplus W^{\perp}_{\mathcal{J},p}.$$

Then $\text{LLS}^{W,\delta}_{\mathcal{J}}(t)$ is well-defined and

$$\|\delta_{J_k}(t_{J_k} - w_{J_k})\| = \min\{\|\delta_{J_k}(t_{J_k} - z_{J_k})\| : z \in W, z_{J_{>k}} = w_{J_{>k}}\}$$

for every $k \in [p]$.

In the notation of the above lemma, for ordered partitions $\mathcal{J} = (J_1, J_2, \ldots, J_p)$, $\bar{\mathcal{J}} = (J_p, J_{p-1}, \ldots, J_1)$, and $(x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ with $\delta = s^{1/2}x^{-1/2}$, we have $\Delta x^{\text{ll}} = \text{LLS}^{W,\delta}_{\mathcal{J}}(-x)$ and $\Delta s^{\text{ll}} = \text{LLS}^{W^{\perp},\delta^{-1}}_{\bar{\mathcal{J}}}(-s)$.

Proof of Lemma 4.3.7. We first prove the equality $W \cap (W^{\perp}_{\mathcal{J},1} \oplus \cdots \oplus W^{\perp}_{\mathcal{J},p}) = \{\vec{0}\}$, and by a similar argument we have $W^{\perp} \cap (W_{\mathcal{J},1} \oplus \cdots \oplus W_{\mathcal{J},p}) = \{\vec{0}\}$. By duality, this last equality tells us that

$$(W^{\perp} \cap (W_{\mathcal{J},1} \oplus \cdots \oplus W_{\mathcal{J},p}))^{\perp} = W + (W^{\perp}_{\mathcal{J},1} \oplus \cdots \oplus W^{\perp}_{\mathcal{J},p}) = \mathbb{R}^n.$$

Thus, the linear decomposition defining $\text{LLS}^{W,\delta}_{\mathcal{J}}(t)$ has a solution and its solution is unique.

Suppose $y \in W \cap (W^{\perp}_{\mathcal{J},1} \oplus \cdots \oplus W^{\perp}_{\mathcal{J},p})$. We prove $y_{J_k} = \vec{0}$ by induction on k, starting at $k = p$. The induction hypothesis is that $y_{J_{>k}} = \vec{0}$, which is an empty requirement when $k = p$. The hypothesis $y_{J_{>k}} = \vec{0}$ together with the assumption $y \in W$ is equivalent to $y \in W \cap \mathbb{R}^n_{J_{\leq k}}$, and implies $y_{J_k} \in \pi_{J_k}(W \cap \mathbb{R}^n_{J_{\leq k}}) := W_{\mathcal{J},k}$. Since we also have $y_{J_k} \in W^{\perp}_{\mathcal{J},k}$ by assumption, which is the orthogonal complement of $W_{\mathcal{J},k}$, we must have $y_{J_k} = \vec{0}$. Hence, by induction $y = \vec{0}$. This finishes the proof that $\text{LLS}^{W,\delta}_{\mathcal{J}}(t)$ is well-defined.

Next we prove that w minimizes $\min\{\|\delta_{J_k}(t_{J_k} - z_{J_k})\| : z \in W, z_{J_{>k}} = w_{J_{>k}}\}$. The optimality condition is for $\delta_{J_k}(t_{J_k} - z_{J_k})$ to be orthogonal to $\delta_{J_k} u$ for any $u \in W_{\mathcal{J},k}$. By the LLS equation, we have $\delta_{J_k}(t_{J_k} - w_{J_k}) = \delta^{-1}_{J_k} v_{J_k}$, where $v_{J_k} \in W^{\perp}_{\mathcal{J},k}$. Noting then that $\langle \delta_{J_k} u, \delta^{-1}_{J_k} v \rangle = \langle u_{J_k}, v_{J_k} \rangle = 0$ for $u \in W_{\mathcal{J},k}$, the optimality condition follows immediately. □

With these tools, we can prove that the lifting costs are self-dual. This explains the reverse order in the dual vs primal LLS step and justifies our attention on the lifting cost in a self-dual algorithm. The next proposition generalizes the result of [101].

Proposition 4.3.8 (Proof on p. 166)**.** *For a linear subspace $W \subseteq \mathbb{R}^n$ and index set $I \subseteq [n]$ with $J = [n] \setminus I$,*

$$\|L_I^W\| \leq \max\{1, \|L_J^{W^\perp}\|\}.$$

In particular, $\ell^W(I) = \ell^{W^\perp}(J)$.

We defer the proof to Section 4.5. Note that this proposition also implies Proposition 4.2.1(iv).

Partition lifting scores

A key insight is that if the layering $\mathcal{J}$ is "well-separated", then we indeed have $x\Delta s^{\text{ll}} + s\Delta x^{\text{ll}} \approx -xs$, that is, the LLS direction is close to the affine scaling direction. This will be shown in Lemma 4.3.10. The notion of "well-separatedness" can be formalized as follows. Recall the definition of the lifting score (4.5). The lifting score of the layering $\mathcal{J} = (J_1, J_2, \ldots, J_p)$ of $[n]$ with respect to W is defined as

$$\ell^W(\mathcal{J}) := \max_{2 \leq k \leq p} \ell^W(J_{\geq k})\,.$$

For $\delta \in \mathbb{R}^n_{++}$, we use $\ell^{W,\delta}(I) := \ell^{\mathrm{Diag}(\delta)W}(I)$ and $\ell^{W,\delta}(\mathcal{J}) := \ell^{\mathrm{Diag}(\delta)W}(\mathcal{J})$. When the context is clear, we omit W and write $\ell^\delta(I) := \ell^{W,\delta}(I)$ and $\ell^\delta(\mathcal{J}) := \ell^{W,\delta}(\mathcal{J})$.

The following important duality claim asserts that the lifting score of a layering equals the lifting score of the reverse layering in the orthogonal complement subspace. It is an immediate consequence of Proposition 4.3.8.

Lemma 4.3.9. *Let $W \subseteq \mathbb{R}^n$ be a linear subspace, $\delta \in \mathbb{R}^n_{++}$. For an ordered partition $\mathcal{J} = (J_1, J_2, \ldots, J_p)$, let $\bar{\mathcal{J}} = (J_p, J_{p-1}, \ldots, J_1)$ denote the reverse ordered partition. Then, we have*

$$\ell^{W,\delta}(\mathcal{J}) = \ell^{W^\perp,\delta^{-1}}(\bar{\mathcal{J}}).$$

Proof. Let $U = \mathrm{Diag}(\delta)W$. Note that $U^\perp = \mathrm{Diag}(\delta^{-1})W^\perp$. Then by Proposition 4.3.8, for $2 \leq k \leq p$, we have that

$$\ell^{W,\delta}(J_{\geq k}) = \ell^U(J_{\geq k}) = \ell^{U^\perp}(J_{\leq k-1}) = \ell^{U^\perp}(\bar{J}_{\geq p-k+2}) = \ell^{W^\perp,\delta^{-1}}(\bar{J}_{\geq p-k+2}).$$

In particular, $\ell^{W,\delta}(\mathcal{J}) = \ell^{W^\perp,\delta^{-1}}(\bar{\mathcal{J}})$, as needed. □

The next lemma summarizes key properties of the LLS steps, assuming the partition has a small lifting score. We show that if $\ell^\delta(\mathcal{J})$ is sufficiently small, then on the one hand, the LLS step will be very close to the affine scaling step, and on the other hand, on each layer $k \in [p]$, it will be very close to the affine scaling step restricted to this layer for the subspace $W_{\mathcal{J},k}$. The proof is deferred to Section 4.5.

Lemma 4.3.10 (Proof on p. 171)**.** *Let $w = (x, y, s) \in \mathcal{N}(\beta)$ for $\beta \in (0, 1/4]$, let $\mu = \mu(w)$ and $\delta = \delta(w)$. Let $\mathcal{J} = (J_1, \ldots, J_p)$ be a layering with $\ell^\delta(\mathcal{J}) \le \beta/(32n^2)$, and let $\Delta w^{\mathrm{ll}} = (\Delta x^{\mathrm{ll}}, \Delta y^{\mathrm{ll}}, \Delta s^{\mathrm{ll}})$ denote the LLS direction for the layering $\mathcal{J}$. Then the following properties hold.*

(i) *We have*

$$\|\delta_{J_k}\Delta x^{\mathrm{ll}}_{J_k} + \delta^{-1}_{J_k}\Delta s^{\mathrm{ll}}_{J_k} + x^{1/2}_{J_k}s^{1/2}_{J_k}\| \le 6n\ell^\delta(\mathcal{J})\sqrt{\mu}\,, \quad \forall k \in [p], \text{ and} \tag{4.28}$$

$$\|\delta\Delta x^{\mathrm{ll}} + \delta^{-1}\Delta s^{\mathrm{ll}} + x^{1/2}s^{1/2}\| \le 6n^{3/2}\ell^\delta(\mathcal{J})\sqrt{\mu}\,. \tag{4.29}$$

(ii) *For the affine scaling direction $\Delta w^{\mathrm{a}} = (\Delta x^{\mathrm{a}}, \Delta y^{\mathrm{a}}, \Delta s^{\mathrm{a}})$,*

$$\|Rx^{\mathrm{ll}} - Rx^{\mathrm{a}}\|, \|Rs^{\mathrm{ll}} - Rs^{\mathrm{a}}\| \le 6n^{3/2}\ell^\delta(\mathcal{J})\,.$$

(iii) *For the residuals of the LLS steps we have $\|Rx^{\mathrm{ll}}\|, \|Rs^{\mathrm{ll}}\| \le \sqrt{2n}$. For each $i \in [n]$, $\max\{|Rx^{\mathrm{ll}}_i|, |Rs^{\mathrm{ll}}_i|\} \ge \frac{1}{2} - \frac{3}{4}\beta$.*

(iv) *Let $\varepsilon^{\mathrm{ll}}(w) = \max_{i\in[n]} \min\{|Rx^{\mathrm{ll}}_i|, |Rs^{\mathrm{ll}}_i|\}$, and define the step length as*

$$\alpha := \sup\{\alpha' \in [0,1] : \forall \bar\alpha \in [0,\alpha'] : w + \bar\alpha\Delta w^{\mathrm{ll}} \in \mathcal{N}(2\beta)\}\,.$$

We obtain the following bounds on the progress in the LLS step:

$$\mu(w + \alpha\Delta w^{\mathrm{ll}}) = (1-\alpha)\mu\,, \quad \text{and}$$

$$\alpha \ge 1 - \frac{3\sqrt{n}\varepsilon^{\mathrm{ll}}(w)}{\beta}.$$

(v) *We have $\varepsilon^{\mathrm{ll}}(w) = 0$ if and only if $\alpha = 1$. These are further equivalent to $w + \Delta w^{\mathrm{ll}} = (x + \Delta x^{\mathrm{ll}}, y + \Delta y^{\mathrm{ll}}, s + \Delta s^{\mathrm{ll}})$ being an optimal solution to* (4.1)*.*

4.3.5 The layering procedure

Our algorithm performs LLS steps on a layering with a low lifting score. A further requirement is that within each layer, the circuit imbalances κ^δ_{ij} defined in (4.7) are suitably bounded. The rescaling here is with respect to $\delta = \delta(w)$ for the current iterate $w = (x, y, s)$. To define the precise requirement on the layering, we first introduce an auxiliary graph. Throughout we use the parameter

$$\gamma := \frac{\beta}{2^{10}n^5}\,. \tag{4.30}$$

The auxiliary graph For a vector $\delta \in \mathbb{R}^n_{++}$ and $\sigma > 0$, we define the directed graph $G_{\delta,\sigma} = ([n], E_{\delta,\sigma})$ such that $(i,j) \in E_{\delta,\sigma}$ if $\kappa^\delta_{ij} \geq \sigma$. This is a subgraph of the *circuit ratio digraph* studied in Section 4.2, including only the edges where the circuit ratio is at least the threshold σ. Note that we do not have direct access to this graph, as we cannot efficiently compute the values κ^δ_{ij}.

At the beginning of the entire algorithm, we run the subroutine FIND-CIRCUITS(A) as in Theorem 4.2.15, where $W = \mathrm{Ker}(A)$. We assume the matroid $\mathcal{M}(A)$ is non-separable. For a separable matroid, we can solve the subproblems of our LP on the components separately. Thus, for each $i \neq j$, $i, j \in [n]$, we obtain an estimate $\hat{\kappa}_{ij} \leq \kappa_{ij}$. These estimates will be gradually improved throughout the algorithm.

Note that $\kappa^\delta_{ij} = \kappa_{ij}\delta_j/\delta_i$ and $\hat{\kappa}^\delta_{ij} = \hat{\kappa}_{ij}\delta_j/\delta_i$. If $\hat{\kappa}^\delta_{ij} \geq \sigma$, then we are guaranteed $(i,j) \in E_{\delta,\sigma}$.

Definition 4.3.11. Define $\hat{G}_{\delta,\sigma} = ([n], \hat{E}_{\delta,\sigma})$ to be the directed graph with edges (i,j) such that $\hat{\kappa}^\delta_{ij} \geq \sigma$; clearly, $\hat{G}_{\delta,\sigma}$ is a subgraph of $G_{\delta,\sigma}$.

Lemma 4.3.12. *Let $\delta \in \mathbb{R}^n_{++}$. For every $i \neq j$, $i, j \in [n]$, $\hat{\kappa}^\delta_{ij} \cdot \hat{\kappa}^\delta_{ji} \geq 1$. Consequently, for any $0 < \sigma \leq 1$, at least one of $(i,j) \in \hat{E}_{\delta,\sigma}$ or $(j,i) \in \hat{E}_{\delta,\sigma}$.*

Proof. We show that this property holds at the initialization. Since the estimates can only increase, it remains true throughout the algorithm. Recall the definition of $\hat{\kappa}_{ij}$ from Theorem 4.2.15. This is defined as the maximum of $|g_j/g_i|$ such that $g \in W$, $\mathrm{supp}(g) = C$ for some $C \in \hat{\mathcal{C}}$ containing i and j. For the same vector g, we get $\hat{\kappa}_{ji} \geq |g_i/g_j|$. Consequently, $\hat{\kappa}_{ij} \cdot \hat{\kappa}_{ji} \geq 1$, and also $\hat{\kappa}^\delta_{ij} \cdot \hat{\kappa}^\delta_{ji} \geq 1$. The second claim follows by the assumption $\sigma \leq 1$. □

Balanced layerings We are ready to define the requirements on the layering in the algorithm. In the algorithm, $\delta = \delta(w)$ will correspond to the scaling of the current iterate $w = (x, y, s)$.

Definition 4.3.13. Let $\delta \in \mathbb{R}^n_{++}$. The layering $\mathcal{J} = (J_1, J_2, \ldots, J_p)$ of $[n]$ is *δ-balanced* if

(i) $\ell^\delta(\mathcal{J}) \leq \gamma$, and

(ii) J_k is strongly connected in $G_{\delta,\gamma/n}$ for all $k \in [p]$.

The following lemma shows that within each layer, the κ^δ_{ij} values are within a bounded range. This will play an important role in our potential analysis.

Lemma 4.3.14. *Let $0 < \sigma < 1$ and $t > 0$, and $i, j \in [n]$, $i \neq j$.*

(i) *If the graph $G_{\delta,\sigma}$ contains a directed path of at most $t-1$ edges from j to i, then*

$$\kappa_{ij}^{\delta} < \left(\frac{\kappa^*}{\sigma}\right)^t .$$

(ii) *If $G_{\delta,\sigma}$ contains a directed path of at most $t-1$ edges from i to j, then*

$$\kappa_{ij}^{\delta} > \left(\frac{\sigma}{\kappa^*}\right)^t .$$

Proof. For part (i), let $j = i_1, i_2, \ldots, i_h = i$ be a path in $G_{\delta,\sigma}$ in J from j to i with $h \leq t$. That is, $\kappa^{\delta}_{i_\ell i_{\ell+1}} \geq \sigma$ for each $\ell \in [h]$. Theorem 4.2.13 yields

$$(\bar{\kappa}^*)^t \geq \kappa_{ij}^{\delta} \cdot \sigma^{h-1} > \kappa_{ij}^{\delta} \cdot \sigma^t ,$$

since $h \leq t$ and $\sigma < 1$. Part (ii) follows using part (i) for j and i, and that $\kappa_{ij}^{\delta} \cdot \kappa_{ji}^{\delta} \geq 1$ according to Lemma 4.3.12. □

Description of the layering subroutine Consider an iterate $w = (x, y, s) \in \mathcal{N}(\beta)$ of the algorithm with $\delta = \delta(w)$, The subroutine LAYERING$(\delta, \hat{\kappa})$, described in Algorithm 3, constructs a δ-balanced layering. We recall that the approximated auxilliary graph $\hat{G}_{\delta,\gamma/n}$ with respect to $\hat{\kappa}$ is as in Definition 4.3.11

Algorithm 3 LAYERING$(\delta, \hat{\kappa})$

Input: $\delta \in \mathbb{R}^n_{++}$ and $\hat{\kappa} \in \mathbb{R}^E_{++}$.
Output: δ-balanced layering $\mathcal{J} = (J_1, \ldots, J_p)$ and updated values $\hat{\kappa} \in \mathbb{R}^E_{++}$.

Compute the strongly connected components $C_1, C_2, \ldots, C_\ell$ of $\hat{G}_{\delta,\gamma/n}$, listed in the ordering imposed by $\hat{G}_{\delta,\gamma/n}$
$\bar{E} \leftarrow \hat{E}_{\delta,\gamma/n}$
for $k = 2, \ldots, \ell$ **do**
  Call VERIFY-LIFT$(\mathrm{Diag}(\delta)W, C_{\geq k}, \gamma)$ that answers *'pass'* or *'fail'*
  if the answer is *'fail'* **then**
    Let $i \in C_{\geq k}$, $j \in C_{<k}$, and t be the output of VERIFY-LIFT such that $\gamma/n \leq t \leq \kappa_{ij}^{\delta}$
    $\hat{\kappa}_{ij} \leftarrow t\delta_i/\delta_j$
    $\bar{E} \leftarrow \bar{E} \cup \{(i, j)\}$
Compute strongly connected components $J_1, J_2, \ldots, J_p$ of $([n], \bar{E})$, listed in the ordering imposed by $\hat{G}_{\delta,\gamma/n}$ **return** $\mathcal{J} = (J_1, J_2, \ldots, J_p), \hat{\kappa}$.

We now give an overview of the subroutine LAYERING$(\delta, \hat{\kappa})$. We start by computing the strongly connected components (SCCs) of the directed graph $\hat{G}_{\delta,\gamma/n}$.

The edges of this graph are obtained using the current estimates $\hat{\kappa}_{ij}^{\delta}$. According to Lemma 4.3.12, we have $(i,j) \in \hat{E}_{\delta,\gamma/n}$ or $(j,i) \in \hat{E}_{\delta,\gamma/n}$ for every $i,j \in [n]$, $i \neq j$. Hence, there is a linear ordering of the components $C_1, C_2, \ldots, C_\ell$ such that $(u,v) \in \hat{E}_{\delta,\gamma/n}$ whenever $u \in C_i$, $v \in C_j$, and $i < j$. We call this the ordering imposed by $\hat{G}_{\delta,\gamma/n}$.

For each $k = 2, \ldots, \ell$, we use the subroutine VERIFY-LIFT$(\mathrm{Diag}(\delta)W, C_{\geq k}, \gamma)$ described in Lemma 4.2.11. If the subroutine returns 'pass', then we conclude $\ell^{\delta}(C_{\geq k}) \leq \gamma$, and proceed to the next layer. If the answer is 'fail', then the subroutine returns as certificates $i \in C_{\geq k}$, $j \in C_{<k}$, and t such that $\gamma/n \leq t \leq \kappa_{ij}^{\delta}$. In this case, we update $\hat{\kappa}_{ij}^{\delta}$ to the higher value t. We add (i,j) to an edge set $\bar{E}$; this edge set was initialized to contain $\hat{E}_{\delta,\gamma/n}$. After adding (i,j), all components C_ℓ between those containing i and j will be merged into a single strongly connected component. To see this, recall that if $i' \in C_\ell$ and $j' \in C_{\ell'}$ for $\ell < \ell'$, then $(i',j') \in \hat{E}_{\delta,\gamma/n}$ according to Lemma 4.3.12.

Finally, we compute the strongly connected components of $([n], \bar{E})$. We let $J_1, J_2, \ldots, J_p$ denote their unique acyclic order, and return these layers.

Lemma 4.3.15. *The subroutine* LAYERING$(\delta, \hat{\kappa})$ *returns a δ-balanced layering in $O(nm^2 + n^2)$ time.*

The difficult part of the proof is showing the running time bound. We note that the weaker bound $O(n^2m^2)$ can be obtained by a simpler argument.

Proof. We first verify that the output layering is indeed δ-balanced. For property (i) of Definition 4.3.13, note that each J_q component is the union of some of the C_k's. In particular, for every $q \in [p]$, the set $J_{\geq q} = C_{\geq k}$ for some $k \in [\ell]$. Assume now $\ell^{\delta}(C_{\geq k}) > \gamma$. At step k of the main cycle, the subroutine VERIFY-LIFT returned the answer 'fail', and a new edge $(i,j) \in E$ was added with $i \in C_{\geq k}$, $j \in C_{<k}$. Note that we already had $(j,i) \in \hat{E}_{\delta,\gamma/n}$, since $j \in C_r$ for some $r < k$, and $i \in C_{r'}$ for $r' \geq k$. This contradicts the choice of $J_{\geq q}$ as a maximal strongly connected component in $([n], E)$.

Property (ii) follows since all new edges added to E have $\kappa_{ij} \geq \gamma/n$. Therefore, $([n], E)$ is a subgraph of $G_{\delta,\gamma/n}$.

Let us now turn to the computational cost. The initial strongly-connected components can be obtained in time $O(n^2)$, and the same bound holds for the computation of the final components. (The latter can be also done in linear time, exploiting the special structure that the components C_i have a complete linear ordering.)

The second computational bottleneck is the subroutine VERIFY-LIFT. We assume a matrix $M \in \mathbb{R}^{n \times (n-m)}$ is computed at the very beginning such that $\mathrm{range}(M) = W$. We first explain how to implement one call to VERIFY-LIFT in $O(n(n-m)^2)$ time.

We then sketch how to amortize the work across the different calls to VERIFY-LIFT, using the nested structure of the layering, to implement the whole procedure in $O(n(n-m)^2)$ time. To turn this into $O(nm^2)$, we recall that the layering procedure is the same for W and $W^\perp$ due to duality (Proposition 4.3.8). Since $\dim(W^\perp) = m$, applying this subroutine on $W^\perp$ instead of W achieves the same result in time $O(nm^2)$.

We now explain the implementation of VERIFY-LIFT, where we are given as input $C \subseteq [n]$ and the basis matrix $M \in \mathbb{R}^{n\times(n-m)}$ as above with $\mathrm{range}(M) = W$. Clearly, the running time is dominated by the computation of the set $I \subseteq C$ and the matrix $B \in \mathbb{R}^{([n]\setminus C)\times|I|}$ satisfying $L_C^W(x)_{[n]\setminus C} = Bx_I$, for $x \in \pi_C(W)$. We explain how to compute I and B from M using column operations (note that these preserve the range). The valid choices for $I \subseteq C$ are in correspondence with maximal sets of linear independent rows of $M_{C,\bullet}$, noting then that $|I| = r$ where $r := \mathrm{rk}(M_{C,\bullet})$. Let $D_1 = [n-m-r]$ and $D_2 = [n-m] \setminus [n-m-r]$. By applying columns operations to M, we can compute $I \subseteq C$ such that $M_{I,D_2} = I_r$ ($r \times r$ identity) and $M_{C,D_1} = 0$. This requires $O(n(n-m)|C|)$ time using Gaussian elimination. At this point, note that $\pi_C(W) = \mathrm{range}(M_{C,D_2})$, $\pi_I(W) = \mathbb{R}^I$ and $\mathrm{range}(M_{\bullet,D_1}) = W \cap \mathbb{R}^n_{[n]\setminus C}$. To compute B, we must transform the columns of $M_{\bullet,D_2}$ into minimum norm lifts of $e^i \in \pi_I(W)$ into W, for all $i \in I$. For this purpose, it suffices to make the columns of $M_{[n]\setminus C,D_2}$ orthogonal to the range of $M_{[n]\setminus C,D_1}$. Applying Gram-Schmidt orthogonalization, this requires $O((n-|C|)(n-m)(n-m-r))$ time. From here, the desired matrix $B = M_{[n]\setminus C,D_2}$. Thus, the total running time of VERIFY-LIFT is $O(n(n-m)|C| + (n-|C|)(n-m)(n-m-r)) = O(n(n-m)^2)$.

We now sketch how to amortize the work of all the calls of VERIFY-LIFT during the layering algorithm, to achieve a total $O(n(n-m)^2)$ running time. Let $C_1, \ldots, C_\ell$ denote the candidate SCC layering. Our task is to compute the matrices $B_k, 2 \le k \le \ell$, needed in the calls to VERIFY-LIFT on $W, C_{\ge k}, 2 \le k \le \ell$, in total $O(n(n-m)^2)$ time. We achieve this in three steps working with the basis matrix M as above. Firstly, by applying column operations to M, we compute sets $I_k \subseteq C_k$ and $D_k = [|I_{\le k}|]\setminus[|I_{<k}|]$, $k \in [\ell]$, such that $M_{I_k,D_k} = I_{r_k}$, where $r_k = |I_k|$, and $M_{C_{\ge k},D_{<k}} = 0, 2 \le k \le \ell$. Note that this enforces $\sum_{k=1}^{\ell} r_k = (n-m)$. This computation requires $O(n(n-m)^2)$ time using Gaussian elimination. This computation achieves $\mathrm{range}(M_{C_k,D_k}) = \pi_{C_k}(W \cap \mathbb{R}^n_{C_{\le k}})$, $\mathrm{range}(M_{C_{\ge k},D_{\ge k}}) = \pi_{C_{\ge k}}(W)$ and $\mathrm{range}(M_{\bullet,D_{\le k}}) = W \cap \mathbb{R}^n_{C_{\le k}}$, for all $k \in [\ell]$.

From here, we block orthogonalize M, such that the columns of $M_{\bullet,D_k}$ are orthogonal to the range of $M_{\bullet,D_{<k}}$, $2 \le k \le \ell$. Applying an appropriately adapted Gram-Schmidt orthogonalization, this requires $O(n(n-m)^2)$ time. Note that this operation maintains $M_{I_k,D_k} = I_{r_k}$, $k \in [\ell]$, since $M_{C_{\ge k},D_{<k}} = 0$. At this point, for $k \in [\ell]$ the columns of $M_{\bullet,D_k}$ are in correspondence with minimum norm lifts of $e^i \in \pi_{D_{\ge k}(W)}$ into W, for all $i \in I_k$. Note that to compute the matrix B_k we need the lifts of $e^i \in \pi_{D_{\ge k}(W)}$, for all $i \in I_{\ge k}$ instead of just $i \in I_k$.

We now compute the matrices $B_\ell, \ldots, B_2$ in this order via the following iterative procedure. Let k denote the iteration counter, which decrements from ℓ to 2. For $k = \ell$ (first iteration), we let $B_\ell = M_{C_{<\ell}, D_\ell}$ and decrement k. For $k < \ell$, we eliminate the entries of $M_{I_k, D_{>k}}$ by using the columns of $M_{\bullet, D_k}$. We then let $B_k = M_{C_{<k}, D_{\geq k}}$ and decrement k. To justify correctness, one need only notice that at the end of iteration k, we maintain the orthogonality of $M_{\bullet, D_{\geq k}}$ to the range of $M_{\bullet, D_{<k}}$ and that $M_{I_{\geq k}, D_{\geq k}} = I_{|I_{\geq k}|}$ is the appropriate identity. The cost of this procedure is the same as a full run of Gaussian elimination and thus is bounded by $O(n(n-m)^2)$. The calls to VERIFY-LIFT during the layering procedure can thus be executed in $O(n(n-m)^2))$ amortized time as claimed. □

4.3.6 The overall algorithm

Algorithm 4 LP-SOLVE(A, b, c, w^0)

Input: $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, and an initial feasible solution $w^0 = (x^0, y^0, s^0) \in \mathcal{N}(1/8)$ to (4.1).

Output: Optimal solution $w^* = (x^*, y^*, s^*)$ to (4.1). Call FIND-CIRCUITS(A) to obtain the lower bounds $\hat{\kappa}_{ij}$ for each $i, j \in [n], i \neq j$ $k \leftarrow 0, \alpha \leftarrow 0$

while $\mu(w^k) \neq 0$ **do**
  Compute affine scaling direction $\Delta w^{\mathrm{a}} = (\Delta x^{\mathrm{a}}, \Delta y^{\mathrm{a}}, \Delta s^{\mathrm{a}})$ for w^k ▷ Predictor
  if $\varepsilon^{\mathrm{a}}(w) < 10n^{3/2}\gamma$ **then** ▷ Recall $\varepsilon^{\mathrm{a}}(w)$ defined in (4.21)
    $\delta \leftarrow (s^k)^{1/2}(x^k)^{-1/2}$
    $(\mathcal{J}, \hat{\kappa}) \leftarrow$LAYERING$(\delta, \hat{\kappa})$
    Compute Layered Least Squares direction $\Delta w^{\mathrm{ll}} = (\Delta x^{\mathrm{ll}}, \Delta y^{\mathrm{ll}}, \Delta s^{\mathrm{ll}})$ for the layering $\mathcal{J}$ and w
    $\Delta w \leftarrow \Delta w^{\mathrm{ll}}$
  else
    $\Delta w \leftarrow \Delta w^{\mathrm{a}}$
  $\alpha \leftarrow \sup\{\alpha' \in [0,1] : \forall \bar{\alpha} \in [0, \alpha'] : w + \bar{\alpha}\Delta w \in \mathcal{N}(1/4)\}$
  $w' \leftarrow w^k + \alpha \Delta w$
  Compute centrality direction $\Delta w^{\mathrm{c}} = (\Delta x^{\mathrm{c}}, \Delta y^{\mathrm{c}}, \Delta s^{\mathrm{c}})$ for w' ▷ Corrector
  $w^{k+1} \leftarrow w' + \Delta w^{\mathrm{c}}$
  $k \leftarrow k + 1$
return $w^k = (x^k, y^k, s^k)$.

Algorithm 4 presents the overall algorithm LP-SOLVE(A, b, c, w^0). We assume that an initial feasible solution $w^0 = (x^0, y^0, s^0) \in \mathcal{N}(\beta)$ is given. We address this in Section 4.7, by adapting the extended system used in [198]. We note that this subroutine requires an upper bound on $\bar{\chi}^*$. Since computing $\bar{\chi}^*$ is hard, we can

implement it by a doubling search on $\log \bar{\chi}^*$, as explained in Section 4.7. Other than for initialization, the algorithm does not require an estimate on $\bar{\chi}^*$.

The algorithm starts with the subroutine FIND-CIRCUITS(A) as in Theorem 4.2.15. The iterations are similar to the MTY Predictor-Corrector algorithm [145]. The main difference is that certain affine scaling steps are replaced by LLS steps. In every predictor step, we compute the affine scaling direction, and consider the quantity $\varepsilon^{\mathrm{a}}(w) = \max_{i\in[n]} \min\{|Rx_i^{\mathrm{a}}|, |Rs_i^{\mathrm{a}}|\}$. If this is above the threshold $10n^{3/2}\gamma$, then we perform the affine scaling step. However, in case $\varepsilon^{\mathrm{a}}(w) < 10n^{3/2}\gamma$, we use the LLS direction instead. In each such iteration, we call the subroutine LAYERING$(\delta, \hat{\kappa})$ (Algorithm 3) to compute the layers, and we compute the LLS step for this layering.

Another important difference is that the algorithm does not require a final rounding step. It terminates with the exact optimal solution w^* once a predictor step is able to perform a full step with $\alpha = 1$.

Theorem 4.3.16. *For given $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, and an initial feasible solution $w^0 = (x^0, y^0, s^0) \in \mathcal{N}(1/8)$, Algorithm* 4 *finds an optimal solution to* (4.1) *in $O(n^{2.5} \log n \log(\bar{\chi}_A^* + n))$ iterations.*

Remark 4.3.17. Whereas using LLS steps enables us to give a strong bound on the total number of iterations, finding LLS directions has a significant computational overhead as compared to finding affine scaling directions. The layering $\mathcal{J}$ can be computed in time $O(nm^2)$ (Lemma 4.3.15), and the LLS steps also require $O(nm^2)$ time, see [142, 198]. This is in contrast to the computational cost $O(n^\omega)$ of an affine scaling direction. Here $\omega < 2.373$ is the matrix multiplication constant [196].

We now sketch a possible approach to amortize the computational cost of the LLS steps over the sequence of affine scaling steps. It was shown in [147] that for the MTY P-C algorithm, the "bad" scenario between two crossover events amounts to a series of affine scaling steps where the progress in μ increases exponentially from every iteration to the next. This corresponds to the term $O(\min\{n^2 \log\log(\mu_0/\eta), \log(\mu_0/\eta)\})$ in their running time analysis. Roughly speaking, such a sequence of affine scaling steps indicates that an LLS step is necessary.

Hence, we could observe these accelerating sequences of affine scaling steps, and perform an LLS step after we see a sequence of length $O(\log n)$. The progress made by these affine scaling steps offsets the cost of computing the LLS direction.

4.4 The potential function and the overall analysis

Let $\mu > 0$ and $\delta(\mu) = s(\mu)^{1/2}x(\mu)^{-1/2} = \sqrt{\mu}/x(\mu) = s(\mu)/\sqrt{\mu}$ correspond to the point on the central path and recall the definition of γ in (4.30). For $i, j \in [n]$, $i \neq j$,

we define

$$\varrho^{\mu}(i,j) := \frac{\log \kappa_{ij}^{\delta(\mu)}}{\log\left(4n\kappa_A^*/\gamma\right)},$$

and the main potentials in the algorithm as

$$\Psi^{\mu}(i,j) := \max\left\{1, \min\left\{2n, \inf_{0<\mu'<\mu} \varrho^{\mu'}(i,j)\right\}\right\} \quad \text{and} \quad \Psi(\mu) := \sum_{i,j\in[n], i\neq j} \log \Psi^{\mu}(i,j).$$

The quantity $\Psi^{\mu}(i,j)$ is motivated by the bounds in Lemma 4.3.14. This statement together with (4.17) imply Lemma 4.4.1.

Lemma 4.4.1. *Let $w = (x,y,s) \in \mathcal{N}(\beta)$ for $\beta \in (0,1/4]$, let $\mu = \mu(w)$, and $\delta = \delta(w)$. Let $i,j \in [n]$, $i \neq j$.*

1. *If $G_{\delta,\gamma/(4n)}$ contains a path from j to i of at most $t-1$ edges, then $\varrho^{\mu}(i,j) < t$.*
2. *If $G_{\delta,\gamma/(4n)}$ contains a path from i to j of at most $t-1$ edges, then $\varrho^{\mu}(i,j) > -t$.*
3. *If $\Psi^{\mu}(i,j) \geq t$, then i and j cannot be together on a layer of size at most t, and j cannot be on a layer preceding the layer containing i in any $\delta(w')$-balanced layering, where $w' = (x',y',s') \in \mathcal{N}(\beta)$ with $\mu(w') < \mu$.*

Our potentials $\Psi^{\mu}(i,j)$ can be seen as fine-grained analogues of the crossover events analyzed in [146, 147, 198]. Roughly speaking, a crossover event corresponds to $\Psi^{\mu}(i,j)$ increasing above n, meaning that i and j cannot be contained in the same layer after the normalized duality gap decreases below μ.

In what follows, we formulate four important lemmas crucial for the proof of Theorem 4.3.16. For the lemmas, we only highlight some key ideas here, and defer the full proofs to Section 4.6.

For a triple $w \in \mathcal{N}(\beta)$, Δw^{ll} refers to the LLS direction found in the algorithm, and Rx^{ll} and Rs^{ll} denote the residuals as in (4.19). For a subset $I \subseteq [n]$ recall the definition

$$\varepsilon_I^{\mathrm{ll}}(w) := \max_{i\in I} \min\{|Rx_i^{\mathrm{ll}}|, |Rs_i^{\mathrm{ll}}|\}.$$

We introduce another important quantity ξ for the analysis:

$$\xi_I^{\mathrm{ll}}(w) := \min\{\|Rx_I^{\mathrm{ll}}\|, \|Rs_I^{\mathrm{ll}}\|\}$$

for a subset $I \subseteq [n]$. For a layering $\mathcal{J} = (J_1, J_2, \ldots, J_p)$, we let

$$\xi_{\mathcal{J}}^{\mathrm{ll}}(w) = \max_{k\in[p]} \xi_{J_k}^{\mathrm{ll}}(w).$$

The key idea of the analysis is to extract information about the optimal solution $w^* = (x^*, y^*, s^*)$ from the LLS direction. The first main lemma shows that if $\|Rx^{\text{ll}}_{J_q}\|$ is large on some layer J_q, then for at least one index $i \in J_q$, $x^*_i/x_i \geq 1/\text{poly}(n)$, i.e., the variable x_i has "converged". The analogous statement holds on the dual side for $\|Rs^{\text{ll}}_{J_q}\|$ and an index $j \in J_q$.

Lemma 4.4.2 (Proof on p. 174)**.** *Let $w = (x, y, s) \in \mathcal{N}(\beta)$ for $\beta \in (0, 1/8]$ and let $w^* = (x^*, y^*, s^*)$ be the optimal solution corresponding to $\mu^* = 0$ on the central path. Let further $\mathcal{J} = (J_1, \ldots, J_p)$ be a $\delta(w)$-balanced layering (Definition 4.3.13), and let $\Delta w^{\text{ll}} = (\Delta x^{\text{ll}}, \Delta y^{\text{ll}}, \Delta s^{\text{ll}})$ be the corresponding LLS direction. Then the following statement holds for every $q \in [p]$:*

(i) *There exists $i \in J_q$ such that*

$$x^*_i \geq \frac{2x_i}{3\sqrt{n}} \cdot (\|Rx^{\text{ll}}_{J_q}\| - 2\gamma n)\,. \tag{4.31}$$

(ii) *There exists $j \in J_q$ such that*

$$s^*_j \geq \frac{2s_j}{3\sqrt{n}} \cdot (\|Rs^{\text{ll}}_{J_q}\| - 2\gamma n)\,. \tag{4.32}$$

We outline the main idea of the proof of part (i); part (ii) follows analogously using the duality of the lifting scores (Lemma 4.3.9). On layer q, the LLS step minimizes $\|\delta_{J_q}(x_{J_q} + \Delta x_{J_q})\|$, subject to $\Delta x_{J_{>q}} = \Delta x^{\text{ll}}_{J_{>q}}$ and subject to existence of $\Delta x_{J_{<q}}$ such that $\Delta x \in W$. By making use of $\ell^{\delta(w)}(J_{>q}) \leq \gamma$ due to $\delta(w)$-balancedness, we can show the existence of a point $z \in W + x^*$ such that $\|\delta_{J_q}(z_{J_q} - x^*_{J_q})\|$ is small, and $z_{J_{>q}} = x_{J_{>q}} + \Delta x^{\text{ll}}_{J_{>q}}$. By the choice of $\Delta x^{\text{ll}}_{J_q}$, we have $\|\delta_{J_q} z_{J_q}\| \geq \|\delta_{J_q}(x_{J_q} + \Delta x^{\text{ll}}_{J_q})\| = \sqrt{\mu}\|Rx^{\text{ll}}_{J_q}\|$. Therefore, $\|\delta_{J_q} x^*_{J_q}/\sqrt{\mu}\|$ cannot be much smaller than $\|Rx^{\text{ll}}_{J_q}\|$. Noting that $\delta_{J_q} x^*_{J_q}/\sqrt{\mu} \approx x^*_{J_q}/x_{J_q}$, we obtain a lower bound on x^*_i/x_i for some $i \in J_q$.

We emphasize that the lemma only shows the existence of such indices i and j, but does not provide an efficient algorithm to identify them. It is also useful to note that for any $i \in [n]$, $\max\{|Rx^{\text{ll}}_i|, |Rs^{\text{ll}}_i|\} \geq \frac{1}{2} - \frac{3}{4}\beta$ according to Lemma 4.3.10(iii). Thus, for each $q \in [p]$, we obtain a positive lower bound in either case (i) or case (ii).

The next lemma allows us to argue that the potential function $\Psi^{\cdot}(\cdot, \cdot)$ increases for multiple pairs of variables, if we have strong lower bounds on both x^*_i and s^*_j for some $i, j \in [n]$, along with a lower and upper bound on $\varrho^{\mu}(i, j)$.

Lemma 4.4.3 (Proof on p. 175)**.** *Let $w = (x, y, s) \in \mathcal{N}(2\beta)$ for $\beta \in (0, 1/8]$, let $\mu = \mu(w)$ and $\delta = \delta(w)$. Let $i, j \in [n]$ and $2 \leq \tau \leq n$ such that for the optimal solution $w^* = (x^*, y^*, s^*)$, we have $x^*_i \geq \beta x_i/(2^{10}n^{5.5})$ and $s^*_j \geq \beta s_j/(2^{10}n^{5.5})$, and*

assume $\varrho^{\mu}(i,j) \geq -\tau$. *After* $O(\beta^{-1}\sqrt{n}\tau\log(\bar{\chi}^* + n))$ *further iterations the duality gap* μ' *fulfills* $\Psi^{\mu'}(i,j) \geq 2\tau$, *and for every* $\ell \in [n] \setminus \{i,j\}$, *either* $\Psi^{\mu'}(i,\ell) \geq 2\tau$, *or* $\Psi^{\mu'}(\ell,j) \geq 2\tau$.

We note that i and j as in the lemma are necessarily different, since $i = j$ would imply $0 = x_i^* s_i^* \geq \beta^2\mu/(2^{20}n^{11}) > 0$.

Let us illustrate the idea of the proof of $\Psi^{\mu'}(i,j) \geq 2\tau$. For i and j as in the lemma, and for a central path element $w' = w(\mu')$ for $\mu' < \mu$, we have $x_i' \geq x_i^*/n \geq \beta x_i/(2^{10}n^{6.5})$ and $s_j' \geq s_j^*/n \geq \beta s_j/(2^{10}n^{6.5})$ by the near-monotonicity of the central path (Lemma 4.3.3). Note that

$$\kappa_{ij}^{\delta'} = \kappa_{ij} \cdot \frac{\delta_j'}{\delta_i'} = \kappa_{ij} \cdot \frac{x_i' s_j'}{\mu'} \geq \kappa_{ij} \cdot \frac{\beta^2 x_i s_j}{2^{20}n^{13}\mu'} \geq \frac{\beta^2(1-\beta)^2}{2^{20}n^{13}} \cdot \kappa_{ij}^{\delta} \cdot \frac{\mu}{\mu'},$$

where the last inequality uses Proposition 4.3.2. Consequently, as μ' sufficiently decreases, $\kappa_{ij}^{\delta'}$ will become much larger than κ_{ij}^{δ}. The claim on $\ell \in [n] \setminus \{i,j\}$ can be shown by using the triangle inequality $\kappa_{ik} \cdot \kappa_{kj} \geq \kappa_{ij}$ shown in Lemma 4.2.16.

Assume now $\xi_{J_q}^{\mathrm{ll}}(w) \geq 4\gamma n$ for some $q \in [p]$ in the LLS step. Then, Lemma 4.4.2 guarantees the existence of $i,j \in J_q$ such that $x_i^*/x_i, s_j^*/s_j \geq \frac{4}{3\sqrt{n}}\gamma n > \beta/(2^{10}n^{5.5})$. Further, Lemma 4.4.1 gives $\varrho^{\mu}(i,j) \geq -|J_q|$. Hence, Lemma 4.4.3 is applicable for i and j with $\tau = |J_q|$.

The overall potential argument in the proof of Theorem 4.3.16 uses Lemma 4.4.3 in three cases: $\xi_{\mathcal{J}}^{\mathrm{ll}}(w) \geq 4\gamma n$ (Lemma 4.4.2 is applicable as above); $\xi_{\mathcal{J}}^{\mathrm{ll}}(w) < 4\gamma n$ and $\ell^{\delta^+}(\mathcal{J}) \leq 4\gamma n$ (Lemma 4.4.4); and $\xi_{\mathcal{J}}^{\mathrm{ll}}(w) < 4\gamma n$ and $\ell^{\delta^+}(\mathcal{J}) > 4\gamma n$ (Lemma 4.4.5). Here, δ^+ refers to the value of δ after the LLS step. Note that $\delta^+ > 0$ is well-defined, unless the algorithm terminated with an optimal solution.

To prove these lemmas, we need to study how the layers "move" during the LLS step. We let $\boldsymbol{B} = \{t \in [n] : |Rs_t^{\mathrm{ll}}| < 4\gamma n\}$ and $\boldsymbol{N} = \{t \in [n] : |Rx_t^{\mathrm{ll}}| < 4\gamma n\}$. The assumption $\xi_{\mathcal{J}}^{\mathrm{ll}}(w) < 4\gamma n$ means that for each layer J_k, either $J_k \subseteq \boldsymbol{B}$ or $J_k \subseteq \boldsymbol{N}$; we accordingly refer to $\boldsymbol{B}$-layers and $\boldsymbol{N}$-layers.

Lemma 4.4.4 (Proof on p. 178)**.** *Let* $w = (x,y,s) \in \mathcal{N}(\beta)$ *for* $\beta \in (0, 1/8]$, *and let* $\mathcal{J} = (J_1, \ldots, J_p)$ *be a* $\delta(w)$*-balanced partition. Assume that* $\xi_{\mathcal{J}}^{\mathrm{ll}}(w) < 4\gamma n$, *and let* $w^+ = (x^+, y^+, s^+) \in \mathcal{N}(2\beta)$ *be the next iterate obtained by the LLS step with* $\mu^+ = \mu(w^+)$ *and assume* $\mu^+ > 0$. *Let* $q \in [p]$ *such that* $\xi_{\mathcal{J}}^{\mathrm{ll}}(w) = \xi_{J_q}^{\mathrm{ll}}(w)$. *If* $\ell^{\delta^+}(\mathcal{J}) \leq 4\gamma n$, *then there exist* $i,j \in J_q$ *such that* $x_i^* \geq \beta x_i^+/(16n^{3/2})$ *and* $s_j^* \geq \beta s_j^+/(16n^{3/2})$. *Further, for any* $\ell, \ell' \in J_q$, *we have* $\varrho^{\mu^+}(\ell,\ell') \geq -|J_q|$.

For the proof sketch, without loss of generality, let $\xi_{\mathcal{J}}^{\mathrm{ll}} = \xi_{J_q}^{\mathrm{ll}} = \|Rx_{J_q}^{\mathrm{ll}}\|$, that is, J_q is an $\boldsymbol{N}$-layer. The case $\xi_{J_q}^{\mathrm{ll}} = \|Rs_{J_q}^{\mathrm{ll}}\|$ can be treated analogously. Since the

residuals $\|Rx^{\mathrm{ll}}_{J_q}\|$ and $\|Rs^{\mathrm{ll}}_{J_q}\|$ cannot be both small, Lemma 4.4.2 readily provides a $j \in J_q$ such that $s^*_j/s_j \geq 1/(6\sqrt{n})$. Using Lemma 4.3.3 and Proposition 4.3.1, $s^*_j/s^+_j = s^*_j/s_j \cdot s_j/s^+_j > (1-\beta)/(6(1+4\beta)n^{3/2}) > \beta/(16n^{3/2})$.

The key ideas of showing the existence of an $i \in J_q$ such that $x^*_i \geq x^+_i/(16n^{3/2})$ are the following. With $\approx$, $\lessapprox$ and $\gtrapprox$, we write equalities and inequalities that hold up to small polynomial factors. First, we show that *(i)* $\|\delta_{J_q} x^+_{J_q}\| \lessapprox \mu^+/\sqrt{\mu}$, and then, that *(ii)* $\|\delta_{J_q} x^*_{J_q}\| \gtrapprox \mu^+/\sqrt{\mu}$.

If we can show *(i)* and *(ii)* as above, we obtain that $\|\delta_{J_q} x^*_{J_q}\| \gtrapprox \|\delta_{J_q} x^+_{J_q}\|$, and thus, $x^*_i \gtrapprox x^+_i$ for some $i \in J_q$.

Let us now sketch the first step. By the assumption $J_q \subseteq \boldsymbol{N}$, one can show $x^+_{J_q}/x_{J_q} \approx \mu^+/\mu$, and therefore

$$\|\delta_{J_q} x^+_{J_q}\| \approx \frac{\mu^+}{\mu}\|\delta_{J_q} x_{J_q}\| \approx \frac{\mu^+}{\mu}\sqrt{\mu} = \frac{\mu^+}{\sqrt{\mu}}\,.$$

The second part of the proof, namely, lower bounding $\|\delta_{J_q} x^*_{J_q}\|$, is more difficult. Here, we only sketch it for the special case when $J_q = [n]$. That is, we have a single layer only; in particular, the LLS step is the same as the affine scaling step $\Delta x^{\mathrm{ll}} = \Delta x^{\mathrm{a}}$. The general case of multiple layers follows by making use of Lemma 4.3.10, i.e. exploiting that for a sufficiently small $\ell^\delta(\mathcal{J})$, the LLS step is close to the affine scaling step.

Hence, assume that $\Delta x^{\mathrm{ll}} = \Delta x^{\mathrm{a}}$. Using the equivalent definition of the affine scaling step (4.18) as a minimum-norm point, we have $\|\delta x^*\| \geq \|\delta(x + \Delta x^{\mathrm{ll}})\| = \sqrt{\mu}\|Rx^{\mathrm{ll}}\| = \sqrt{\mu}\xi^{\mathrm{ll}}_{\mathcal{J}}$. From Lemma 4.3.6, $\mu^+/\mu \leq \sqrt{n}\varepsilon^{\mathrm{a}}(w)/\beta \leq \sqrt{n}\xi^{\mathrm{ll}}_{\mathcal{J}}/\beta$. Thus, we see that $\|\delta x^*\| \geq \beta\mu^+/(\sqrt{n\mu})$.

The final statement on lower bounding $\varrho^{\mu^+}(\ell,\ell') \geq -|J_q|$ for any $\ell,\ell' \in J_q$ follows by showing that $\delta^+_\ell/\delta^+_{\ell'}$ remains close to $\delta_\ell/\delta_{\ell'}$, and hence the values of $\kappa^{\mu^+}(\ell,\ell')$ and $\kappa^{\mu}(\ell,\ell')$ are sufficiently close for indices on the same layer (Lemma 4.6.1).

Lemma 4.4.5 (Proof on p. 181)**.** *Let $w = (x,y,s) \in \mathcal{N}(\beta)$ for $\beta \in (0,1/8]$, and let $\mathcal{J} = (J_1,\ldots,J_p)$ be a $\delta(w)$-balanced partition. Assume that $\xi^{\mathrm{ll}}_{\mathcal{J}}(w) < 4\gamma n$, and let $w^+ = (x^+,y^+,s^+) \in \mathcal{N}(2\beta)$ be the next iterate obtained by the LLS step with $\mu^+ = \mu(w^+)$ and assume $\mu^+ > 0$. If $\ell^{\delta^+}(\mathcal{J}) > 4\gamma n$, then there exist two layers J_q and J_r and $i \in J_q$ and $j \in J_r$ such that $x^*_i \geq x^+_i/(8n^{3/2})$, and $s^*_j \geq s^+_j/(8n^{3/2})$. Further, $\varrho^{\mu^+}(i,j) \geq -|J_q \cup J_r|$, and for all $\ell,\ell' \in J_q \cup J_r$, $\ell \neq \ell'$ we have $\Psi^{\mu}(\ell,\ell') \leq |J_q \cup J_r|$.*

Consider now any $\ell \in J_k \subseteq \boldsymbol{B}$. Then, since Rx^{ll}_ℓ is very close to 1, $x^+_\ell \approx x_\ell$; on the other hand s^+_ℓ will "shoot down" close to the small value $Rs^{\mathrm{ll}}_\ell \cdot s_\ell$. Conversely, for $\ell \in J_k \subseteq \boldsymbol{N}$, $s^+_\ell \approx s_\ell$, and x^+_ℓ will "shoot down" to a small value.

The key step of the analysis is showing that the increase in $\ell^{\delta^+}(\mathcal{J})$ can be attributed to an $\boldsymbol{N}$-layer J_r "crashing into" a $\boldsymbol{B}$-layer J_q. That is, we show the existence of an

edge $(i', j') \in E_{\delta^+,\gamma/(4n)}$ for $i' \in J_q$ and $j' \in J_r$, where $r < q$ and $J_q \subseteq \boldsymbol{B}$, $J_r \subseteq \boldsymbol{N}$. This can be achieved by analyzing the matrix B used in the subroutine VERIFY-LIFT.

For the layers J_q and J_r, we can use Lemma 4.4.2 to show that there exists an $i \in J_q$ where x_i^*/x_i is lower bounded, and there exists a $j \in J_r$ where s_j^*/s_j is lower bounded. The lower bound on $\varrho^{\mu^+}(i, j)$ and the upper bounds on the $\Psi^{\mu}(\ell, \ell')$ values can be shown by tracking the changes between the $\kappa^{\delta}(\ell, \ell')$ and $\kappa^{\delta^+}(\ell, \ell')$ values, and applying Lemma 4.4.1 both at w and at w^+.

Proof of Theorem 4.3.16. We analyze the overall potential function $\Psi(\mu)$. By definition, $0 \le \Psi(\mu) \le n(n-1)(\log_2 n + 1)$, and if $\mu' < \mu$ then $\Psi(\mu') \ge \Psi(\mu)$. By the *iteration at* μ we mean the iteration where the normalized duality gap of the current iterate is μ.

If $\mu^+ = 0$ at the end of an iteration, the algorithm terminates with an optimal solution. Recall from Lemma 4.3.10(v) that this happens if and only if $\varepsilon^{\mathrm{ll}}(w) = 0$ at a certain iteration.

From now on, assume that $\mu^+ > 0$. We distinguish three cases at each iteration. These cases are well-defined even at iterations where affine scaling steps are used. At such iterations, $\xi^{\mathrm{ll}}_{\mathcal{J}}(w)$ still refers to the LLS residuals, even if these have not been computed by the algorithm.

(i) $\xi^{\mathrm{ll}}_{\mathcal{J}}(w) \ge 4\gamma n$;

(ii) $\xi^{\mathrm{ll}}_{\mathcal{J}}(w) < 4\gamma n$ and $\ell^{\delta^+}(\mathcal{J}) \le 4\gamma n$; and

(iii) $\xi^{\mathrm{ll}}_{\mathcal{J}}(w) < 4\gamma n$ and $\ell^{\delta^+}(\mathcal{J}) > 4\gamma n$.

Recall that the algorithm uses an LLS direction instead of the affine scaling direction whenever $\varepsilon^{\mathrm{a}}(w) < 10n^{3/2}\gamma$. Consider now the case when an affine scaling direction is used, that is, $\varepsilon^{\mathrm{a}}(w) \ge 10n^{3/2}\gamma$. According to Lemma 4.3.10(ii), $\|Rx^{\mathrm{ll}} - Rx^{\mathrm{a}}\|, \|Rs^{\mathrm{ll}} - Rs^{\mathrm{a}}\| \le 6n^{3/2}\gamma$. This implies that $\xi^{\mathrm{ll}}_{\mathcal{J}}(w) \ge 4n^{3/2}\gamma \ge 4n\gamma$. Therefore, in cases (ii) and (iii), an LLS step will be performed.

Starting with any given iteration, in each case we will identify a set $J \subseteq [n]$ of indices with $|J| > 1$, and start a *phase* of $O(\sqrt{n}|J| \log(\bar{\chi}^* + n))$ iterations (that can be either affine scaling or LLS steps). In each phase, we will guarantee that Ψ increases by at least $|J| - 1$. As we can partition the union of all iterations into disjoint phases, this yields the bound $O(n^{2.5} \log n \log(\bar{\chi}^* + n))$ on the total number of iterations.

We now consider each of the cases. We always let μ denote the normalized duality gap at the current iteration, and we let $q \in [p]$ be the layer such that $\xi^{\mathrm{ll}}_{\mathcal{J}}(w) = \xi^{\mathrm{ll}}_{J_q}(w)$.

Case (i): $\xi_{\mathcal{J}}^{\mathrm{ll}}(w) \geq 4\gamma n$. Lemma 4.4.2 guarantees the existence of $x_i, s_j \in J_q$ such that $x_i^*/x_i, s_j^*/s_j \geq 4\gamma n/(3\sqrt{n}) > 1/(2^{10}n^{5.5})$. Further, according to Lemma 4.4.1, $\varrho^{\mu}(i,j) \geq -|J_q|$. Thus, Lemma 4.4.3 is applicable for $J = J_q$. The phase starting at μ comprises $O(\sqrt{n}|J_q|\log(\bar{\chi}^* + n))$ iterations, after which we get a normalized duality gap μ' such that $\Psi^{\mu'}(i,j) \geq 2|J_q|$, and for each $\ell \in [n] \setminus \{i,j\}$, either $\Psi^{\mu'}(i,\ell) \geq 2|J_q|$, or $\Psi^{\mu'}(\ell,j) \geq 2|J_q|$.

We can take advantage of these bounds for indices $\ell \in J_q$. Again by Lemma 4.4.1, for any $\ell, \ell' \in J_q$, we have $\Psi^{\mu}(\ell,\ell') \leq \varrho^{\mu}(\ell,\ell') \leq |J_q|$. Thus, there are at least $|J_q| - 1$ pairs of indices (ℓ,ℓ') for which $\Psi^{\mu}(\ell,\ell')$ increases by at least $|J_q|$ between iterations at μ and μ'.

We note that this analysis works regardless whether an LLS step or an affine scaling step was performed in the iteration at μ.

Case (ii): $\xi_{\mathcal{J}}^{\mathrm{ll}}(w) < 4\gamma n$ and $\ell^{\delta^+}(\mathcal{J}) \leq 4\gamma n$. As explained above, in this case we perform an LLS step in the iteration at μ, and we let w^+ denote the iterate obtained by the LLS step. For $J = J_q$, Lemma 4.4.4 guarantees the existence of $i, j \in J_q$ such that $x_i^*/x_i^+, s_j^*/s_j^+ > \beta/(16n^{3/2})$, and further, $\varrho^{\mu^+}(i,j) > -|J_q|$. We can therefore apply Lemma 4.4.3. The phase starting at μ includes the LLS step leading to μ^+ (and the subsequent centering step), and the additional $O(\sqrt{n}|J_q|\log(\bar{\chi}^* + n))$ iterations (β is a fixed constant in Algorithm 4) as in Lemma 4.4.3. As in Case I, we get the desired potential increase compared to the potentials at μ in layer J_q.

Case (iii): $\xi_{\mathcal{J}}^{\mathrm{ll}}(w) < 4\gamma n$ and $\ell^{\delta^+}(\mathcal{J}) > 4\gamma n$. Again, the iteration at μ will use an LLS step. We apply Lemma 4.4.5, and set $J = J_q \cup J_r$ as in the lemma. The argument is the same as in Case II, using that Lemma 4.4.5 explicitly states that $\Psi^{\mu}(\ell,\ell') \leq |J|$ for any $\ell, \ell' \in J$, $\ell \neq \ell'$. □

4.4.1 The iteration complexity bound for the Vavasis-Ye algorithm

We now show that the potential analysis described above also gives an improved bound $O(n^{2.5}\log n \log(\bar{\chi}_A + n))$ for the original VY algorithm [198].

We recall the VY layering step. Order the variables via π such that $\delta_{\pi(1)} \leq \delta_{\pi(2)} \leq \ldots \leq \delta_{\pi(n)}$. The layers will be consecutive sets in the ordering; a new layer starts with $\pi(i+1)$ each time $\delta_{\pi(i+1)} > g\delta_{\pi(i)}$, for a parameter $g = \mathrm{poly}(n)\bar{\chi}$.

As outlined in the Introduction, the VY algorithm can be seen as a special implementation of our algorithm by setting $\hat{\kappa}_{ij} = g\gamma/n$. With these edge weights, we have that $\hat{\kappa}_{ij}^{\delta} \geq \gamma/n$ precisely if $g\delta_j \geq \delta_i$. Note that, for simplicity, in the Introduction we used $gx_i \geq x_j$ instead. These quantities are almost the same when in close proximity to the central path.

With these edge weights, it is easy to see that our LAYERING($\delta, \hat{\kappa}$) subroutine finds the exact same components as VY. Moreover, the layers will be the initial strongly connected components C_i of $G_{\delta,\gamma/n}$: due to the choice of g, this partition is automatically δ-balanced. There is no need to call VERIFY-LIFT.

The essential difference compared to our algorithm is that the values $\hat{\kappa}_{ij} = g\gamma/n$ are not lower bounds on κ_{ij} as we require, but upper bounds instead. This is convenient to simplify the construction of the layering. On the negative side, the strongly connected components of $\hat{G}_{\delta,\gamma/n}$ may not anymore be strongly connected in $G_{\delta,\gamma/n}$. Hence, we cannot use Lemma 4.4.1, and consequently, Lemma 4.4.3 does not hold.

Still, the $\hat{\kappa}_{ij}$ bounds are overestimating κ_{ij} by at most a factor $\text{poly}(n)\bar{\chi}$. Therefore, the strongly connected components of $\hat{G}_{\delta,n/\gamma}$ are strongly connected in $G_{\delta,\sigma}$ for some $\sigma = 1/(\text{poly}(n)\bar{\chi})$.

Hence, the entire argument described in this section is applicable to the VY algorithm, with a different potential function defined with $\bar{\chi}$ instead of $\bar{\chi}^*$. This is the reason why the iteration bound in Lemma 4.4.3, and therefore in Theorem 4.3.16, also changes to $\bar{\chi}$ dependency.

It is worth noting that due to the overestimation of the κ_{ij} values, the VY algorithm uses a coarser layering than our algorithm. Our algorithm splits up the VY layers into smaller parts so that $\ell^\delta(\mathcal{J})$ remains small, but within each part, the gaps between the variables are bounded as a function of $\bar{\chi}^*_A$ instead of $\bar{\chi}_A$.

4.5 Properties of the layered least square step

This section is dedicated to the proofs of Proposition 4.3.8 on the duality of lifting scores and Lemma 4.3.10 on properties of LLS steps.

Proposition 4.3.8 (Repetition)**.** *For a linear subspace $W \subseteq \mathbb{R}^n$ and index set $I \subseteq [n]$ with $J = [n] \setminus I$,*

$$\|L_I^W\| \leq \max\{1, \|L_J^{W^\perp}\|\}.$$

In particular, $\ell^W(I) = \ell^{W^\perp}(J)$.

Proof. We first treat the case where $\pi_I(W) = \{\vec{0}\}$ or $\pi_J(W^\perp) = \{\vec{0}\}$. If $\pi_I(W) = \{\vec{0}\}$ then $\|L_I^W\| = \ell^W(I) = 0$. Furthermore, in this case $\mathbb{R}^I = \pi_I(W)^\perp = \pi_I(W^\perp \cap \mathbb{R}^n_I)$, and thus $\pi_{\mathbb{R}^n_J}(W^\perp) \subseteq W^\perp$. In particular, $\|L_J^W\| \leq 1$ and $\ell^{W^\perp}(J) = 0$. Symmetrically, if $\pi_J(W^\perp) = \{\vec{0}\}$ then $\|L_J^{W^\perp}\| = \ell^{W^\perp}(J) = 0$, $\|L_I^W\| \leq 1$ and $\ell^W(I) = 0$.

We now restrict our attention to the case where both $\pi_I(W), \pi_J(W^\perp) \neq \{\vec{0}\}$. Under this assumption, we show that $\|L_I^W\| = \|L_J^{W^\perp}\|$ and thus that $\ell^W(I) = \ell^{W^\perp}(J)$. Note that by non-emptyness, we clearly have that $\|L_I^W\|, \|L_J^{W^\perp}\| \geq 1$.

We formulate a more general claim. Let $\{\vec{0}\} \neq U, V \subseteq \mathbb{R}^n$ be linear subspaces such that $U + V = \mathbb{R}^n$ and $U \cap V = \{\vec{0}\}$. Note that for the orthogonal complements in $\mathbb{R}^n$, we also have $\{\vec{0}\} \neq U^\perp, V^\perp$, $U^\perp + V^\perp = \mathbb{R}^n$ and $U^\perp \cap V^\perp = \{\vec{0}\}$.

Claim 4.5.1. *Let $\{\vec{0}\} \neq U, V \subseteq \mathbb{R}^n$ be linear subspaces such that $U + V = \mathbb{R}^n$ and $U \cap V = \{\vec{0}\}$. Thus, for $z \in \mathbb{R}^n$, there are unique decompositions $z = u + v$ with $u \in U$, $v \in V$ and $z = u' + v'$ with $u' \in U^\perp$ and $v' \in V^\perp$. Let $T : \mathbb{R}^n \to V$ be the map sending $Tz = v$. Let $T' : \mathbb{R}^n \to V^\perp$ be the map sending $T'z = v'$. Then, $\|T\| = \|T'\|$.*

Proof. To prove the statement, we claim that it suffices to show that if $\|T\| > 1$ then $\|T'\| \geq \|T\|$. To prove sufficiency, note that by symmetry, we also get that if $\|T'\| > 1$ then $\|T\| \geq \|T'\|$.Note that $V, V^\perp \neq \{\vec{0}\}$ by assumption, and $Tz = z$ for $z \in V$, $T'z = z$ for $z \in V^\perp$. Thus, we always have $\|T\|, \|T'\| \geq 1$, and therefore the equality $\|T\| = \|T'\|$ must hold in all cases. We now assume $\|T\| > 1$ and show $\|T'\| \geq \|T\|$.

Representing T as an $n \times n$ matrix, we write $T = \sum_{i=1}^{k} \sigma_i v_i u_i^\mathsf{T}$ using a singular value decomposition with $\sigma_1 \geq \cdots \geq \sigma_k > 0$. As such, $v_1, \ldots, v_k$ is an orthonormal basis of V, since the range$(T) = V$, and $u_1, \ldots, u_k$ is an orthonormal basis of $U^\perp$, since $\mathrm{Ker}(T) = U$, noting that we have restricted to the singular vectors associated with positive singular values. By assumption, we have that $\|T\| = \|Tu_1\| = \sigma_1 > 1$.

The proof is complete by showing that

$$\|T'(v_1 - u_1/\sigma_1)\| \geq \sigma_1 \|v_1 - u_1/\sigma_1\|, \tag{4.33}$$

and that $\|v_1 - u_1/\sigma_1\| > 0$, since then the vector $v_1 - u_1/\sigma_1$ will certify that $\|T'\| \geq \sigma_1$.

The map T is a linear projection with $T^2 = T$. Hence ${u_i}^\mathsf{T} v_i = \sigma_i^{-1}$ and ${u_i}^\mathsf{T} v_j = 0$ for all $i \neq j$.

We show that $v_1 - \sigma_1^{-1} u_1$ can be decomposed as $v_1 - \sigma_1 u_1 + (\sigma_1 - \sigma_1^{-1}) u_1$ such that $v_1 - \sigma_1 u_1 \in V^\perp$ and $(\sigma_1 - \sigma_1^{-1}) u_1 \in U^\perp$. Therefore, $T'(v_1 - \sigma_1^{-1} u_1) = v_1 - \sigma_1 u_1$.

The containment $(\sigma_1 - \sigma_1^{-1}) u_1 \in U^\perp$ is immediate. To show $v_1 - \sigma_1 u_1 \in V^\perp$, we need that $v_1 - \sigma_1 {u_1}^\mathsf{T} v_i = 0$ for all $i \in [k]$. For $i \geq 2$, this is true since ${u_i}^\mathsf{T} v_j = 0$ and ${v_i}^\mathsf{T} v_j = 0$. For $i = 1$, we have $v_1 - \sigma_1 {u_1}^\mathsf{T} v_1 = 0$ since $\|v_1\| = 1$ and ${u_1}^\mathsf{T} v_1 = \sigma_1^{-1}$. Consequently, $T'(v_1 - \sigma_1^{-1} u_1) = v_1 - \sigma_1 u_1$.

We compute $\left\|v_1 - \sigma_1^{-1} u_1\right\| = \sqrt{1 - \sigma_1^{-2}} > 0$, since $\sigma_1 > 1$, and $\|v_1 - \sigma_1 u_1\| = \sqrt{\sigma_1^2 - 1}$. This verifies (4.33), and thus $\|T'\| \geq \sigma_1 = \|T\|$. □

To prove the lemma, we define $\mathcal{J} = (J, I)$, $U = W_{\mathcal{J},1}^\perp \oplus W_{\mathcal{J},2}^\perp$ and $V = W$ and let $T : \mathbb{R}^n \to V$ and $T' : \mathbb{R}^n \to V^\perp$ be as in Claim 4.5.1. By assumption, $\{\vec{0}\} \neq \pi_I(W) \Rightarrow \{\vec{0}\} \neq V$ and $\{\vec{0}\} \neq \pi_J(W^\perp) = W_{\mathcal{J},1}^\perp \Rightarrow \{\vec{0}\} \neq U$. Applying Lemma 4.3.7, U, V satisfy the conditions of Claim 4.5.1 and $T = \mathrm{LLS}_{\mathcal{J}}^{W,1}$. In particular, $\|T'\| = \|T\|$. Using the fact that $U^\perp = W_{\mathcal{J},1} \oplus W_{\mathcal{J},2}$ and $V^\perp = W^\perp$, we

similarly get that $T' = \mathcal{LLS}^{W^\perp,1}_{\bar{\mathcal{J}}}$, where $\bar{\mathcal{J}} = (I, J)$. By (4.22) we have, for any $t \in \pi_{\mathbb{R}^n_I}(W)$, that $Tt = \mathrm{LLS}^{W,1}_{\mathcal{J}}(t) = L^W_I(t_I)$. Thus $\|T\| \geq \|L^W_I\| \geq 1$.

To finish the proof of the lemma from the claim, we show that $\|T\| \leq \|L^W_I\|$. By a symmetric argument we get $\|T'\| = \|L^{W^\perp}_J\|$.

If $x \in \mathbb{R}^n_J$, then $Tx \in W \cap \mathbb{R}^n_J$ because any $s \in W^\perp_{\mathcal{J},2}, t \in \pi_I(W)$ with $s + t = \vec{0}$ must have $s = t = \vec{0}$ since $W^\perp_{\mathcal{J},2}$ is orthogonal to $\pi_I(W)$. But $W \cap \mathbb{R}^n_J$ and $W^\perp_{\mathcal{J},1}$ are orthogonal, so $\|Tx\| \leq \|x\|$ because $x = Tx + (x - Tx)$ is an orthogonal decomposition.

If $y \in \mathbb{R}^n_I$, then $y_J = \vec{0}$ and hence $(Ty)_J = (Ty - y)_J$. Since $(Ty - y)_J \in W^\perp_{\mathcal{J},1} = \pi_J(W \cap \mathbb{R}^n_J)^\perp$, we see that $Ty \in (W \cap \mathbb{R}^n_J)^\perp$. As such, for any $x \in \mathbb{R}^n_J, y \in \mathbb{R}^n_I$, we see that $x \perp y$ and $Tx \perp Ty$. For $x, y \neq \vec{0}$, we thus have that

$$\frac{\|T(x+y)\|^2}{\|x+y\|^2} = \frac{\|T(x)\|^2 + \|T(y)\|^2}{\|x\|^2 + \|y^2\|} \leq \max\left\{\frac{\|T(x)\|^2}{\|x\|^2}, \frac{\|T(y)\|^2}{\|y\|^2}\right\} \leq \max\left\{1, \frac{\|T(y)\|^2}{\|y\|^2}\right\}.$$

Since $\|L^W_I\| \geq 1$, we must have that $\|Tt\|/\|t\|$ is maximized by some $t \in \mathbb{R}^n_I$. From $\mathrm{Ker}(T) = U$ it is clear that $\|Tt\|/\|t\|$ is maximized by some $t \in U^\perp$. Now, $U^\perp \cap \mathbb{R}^n_I = \pi_{\mathbb{R}^n_I}(W)$, so any t maximizing $\|Tt\|/\|t\|$ satisfies $Tt = L^W_I(t_I)$. Therefore, $\|L^W_I\| \geq \|T\|$. □

Our next goal is to show Lemma 4.3.10: for a layering with small enough $\ell^\delta(\mathcal{J})$, the LLS step approximately satisfies (4.14), that is, $\delta\Delta x^{\mathrm{ll}} + \delta^{-1}\Delta s^{\mathrm{ll}} \approx -x^{1/2}s^{1/2}$. This also enables us to derive bounds on the norm of the residuals and on the step-length. We start by proving a few auxiliary technical claims. The next simple lemma allows us to take advantage of low lifting scores in the layering.

Lemma 4.5.2. *Let $u, v \in \mathbb{R}^n$ be two vectors such that $u - v \in W$. Let $I \subseteq [n]$, and $\delta \in \mathbb{R}^n_{++}$. Then there exists a vector $u' \in W + u$ satisfying $u'_I = v_I$ and*

$$\|\delta_{[n]\setminus I}(u'_{[n]\setminus I} - u_{[n]\setminus I})\| \leq \ell^\delta(I)\|\delta_I(u_I - v_I)\|.$$

Proof. We let

$$u' := u + \delta^{-1}L^\delta_I(\delta_I(v_I - u_I)).$$

The claim follows by the definition of the lifting score $\ell^\delta(I)$. □

The next lemma will be the key tool to prove Lemma 4.3.10. It is helpful to recall the characterization of the LLS step in Section 4.3.4.

Lemma 4.5.3. *Let $w = (x, y, s) \in \mathcal{N}(\beta)$ for $\beta \in (0, 1/4]$, let $\mu = \mu(w)$ and $\delta = \delta(w)$. Let $\mathcal{J} = (J_1, \ldots, J_p)$ be a $\delta(w)$-balanced layering, and let $\Delta w^{\mathrm{ll}} = (\Delta x^{\mathrm{ll}}, \Delta y^{\mathrm{ll}}, \Delta s^{\mathrm{ll}})$ denote the corresponding LLS direction. Let $\Delta x \in \bigoplus_{k=1}^{p} W_{\mathcal{J},k}$ and $\Delta s \in \bigoplus_{k=1}^{p} W_{\mathcal{J},k}^{\perp}$ as in* (4.26) *and* (4.27)*, that is*

$$\delta \Delta x^{\mathrm{ll}} + \delta^{-1} \Delta s + x^{1/2} s^{1/2} = \vec{0}\,, \tag{4.34}$$

$$\delta \Delta x + \delta^{-1} \Delta s^{\mathrm{ll}} + x^{1/2} s^{1/2} = \vec{0}. \tag{4.35}$$

Then, there exist vectors $\Delta \bar{x} \in \bigoplus_{k=1}^{p} W_{\mathcal{J},k}$ and $\Delta \bar{s} \in \bigoplus_{k=1}^{p} W_{\mathcal{J},k}^{\perp}$ such that

$$\|\delta_{J_k}(\Delta \bar{x}_{J_k} - \Delta x^{\mathrm{ll}}_{J_k})\| \le 2n\ell^{\delta}(\mathcal{J})\sqrt{\mu} \quad \forall k \in [p] \quad \textit{and} \tag{4.36}$$

$$\|\delta^{-1}_{J_k}(\Delta \bar{s}_{J_k} - \Delta s^{\mathrm{ll}}_{J_k})\| \le 2n\ell^{\delta}(\mathcal{J})\sqrt{\mu} \quad \forall k \in [p]\,. \tag{4.37}$$

Proof. Throughout, we use the shorthand notation $\lambda = \ell^{\delta}(\mathcal{J})$. We construct $\Delta \bar{x}$; one can obtain $\Delta \bar{s}$, using that the reverse layering has lifting score λ in $W^{\perp}\,\mathrm{Diag}(\delta^{-1})$ according to Lemma 4.3.9.

We proceed by induction, constructing $\Delta \bar{x}_{J_k} \in W_{\mathcal{J},k}$ for $k = p, p-1, \ldots, 1$. This will be given as $\Delta \bar{x}_{J_k} = \Delta x^{(k)}_{J_k}$ for a vector $\Delta x^{(k)} \in W$ such that $\Delta x^{(k)}_{J_{>k}} = \vec{0}$. We prove the inductive hypothesis

$$\left\| \delta_{J_{\le k}} \left(\Delta x^{(k)}_{J_{\le k}} - \Delta x^{\mathrm{ll}}_{J_{\le k}} \right) \right\| \le 2\lambda\sqrt{\mu} \sum_{q=k+1}^{p} \sqrt{|J_q|}\,. \tag{4.38}$$

Note that (4.36) follows by restricting the norm on the LHS to J_k and since the sum on the RHS is $\le n$.

For $k = p$, the RHS is 0. We simply set $\Delta x^{(p)} = \Delta x^{\mathrm{ll}}$, that is, $\Delta \bar{x}_{J_p} = \Delta x^{\mathrm{ll}}_{J_p}$, trivially satisfying the hypothesis. Consider now $k < p$, and assume that we have a $\Delta \bar{x}_{J_{k+1}} = \Delta x^{(k+1)}_{J_{k+1}}$ satisfying (4.38) for $k+1$. From (4.34) and the induction hypothesis, we get that

$$\begin{aligned}
&\|\delta_{J_{k+1}} \Delta \bar{x}_{J_{k+1}} + \delta^{-1}_{J_{k+1}} \Delta s_{J_{k+1}}\| \le \|x^{1/2}_{J_{k+1}} s^{1/2}_{J_{k+1}}\| + \|\delta_{J_{k+1}}(\Delta \bar{x}_{J_{k+1}} - \Delta x^{\mathrm{ll}}_{J_{k+1}})\| \\
&\le \|x^{1/2}_{J_{k+1}} s^{1/2}_{J_{k+1}}\| + 2\lambda\sqrt{\mu} \sum_{q=k+2}^{p} \sqrt{|J_q|} \le \sqrt{1+\beta}\sqrt{\mu|J_{k+1}|} + 2n\lambda\sqrt{\mu} < 2\sqrt{\mu|J_{k+1}|}\,,
\end{aligned}$$

using also that $w \in \mathcal{N}(\beta)$, Proposition 4.3.2, and the assumptions $\beta \le 1/4$, $\lambda \le \beta/(32n^2)$. Note that $\Delta \bar{x}_{J_{k+1}} \in W_{\mathcal{J},k}$ and $\Delta s_{J_{k+1}} \in W^{\perp}_{\mathcal{J},k}$ are orthogonal vectors. The above inequality therefore implies

$$\|\delta_{J_{k+1}} \Delta \bar{x}_{J_{k+1}}\| \le 2\sqrt{\mu|J_{k+1}|}\,.$$

Let us now use Lemma 4.5.2 to obtain $\Delta x^{(k)}$ for $u = \Delta x^{(k+1)}$, $v = \vec{0}$, and $I = J_{>k}$. That is, we get $\Delta x^{(k)}_{J_{>k}} = \vec{0}$, $\Delta x^{(k)} \in W$, and

$$\begin{aligned}\|\delta_{J_{\le k}}(\Delta x^{(k)}_{J_{\le k}} - \Delta x^{(k+1)}_{J_{\le k}})\| &\le \lambda \|\delta_{J_{>k}} \Delta x^{(k+1)}_{J_{>k}}\| \\ &= \lambda \|\delta_{J_{k+1}} \Delta \bar{x}_{J_{k+1}}\| \le 2\lambda \sqrt{\mu |J_{k+1}|}\,.\end{aligned}$$

By the triangle inequality and the induction hypothesis (4.38) for $k+1$,

$$\begin{aligned}\|\delta_{J_{\le k}}(\Delta x^{(k)}_{J_{\le k}} - \Delta x^{\text{ll}}_{J_{\le k}})\| &\le \|\delta_{J_{\le k}}(\Delta x^{(k)}_{J_{\le k}} - \Delta x^{(k+1)}_{J_{\le k}})\| + \|\delta_{J_{\le k}}(\Delta x^{(k+1)}_{J_{\le k}} - \Delta x^{\text{ll}}_{J_{\le k}})\| \\ &\le 2\lambda \sqrt{\mu |J_{k+1}|} + 2\lambda \sum_{q=k+2}^{p} \sqrt{\mu |J_q|},\end{aligned}$$

yielding the induction hypothesis for k. □

Lemma 4.3.10 (Repetition)**.** *Let $w = (x, y, s) \in \mathcal{N}(\beta)$ for $\beta \in (0, 1/4]$, let $\mu = \mu(w)$ and $\delta = \delta(w)$. Let $\mathcal{J} = (J_1, \ldots, J_p)$ be a layering with $\ell^\delta(\mathcal{J}) \le \beta/(32n^2)$, and let $\Delta w^{\text{ll}} = (\Delta x^{\text{ll}}, \Delta y^{\text{ll}}, \Delta s^{\text{ll}})$ denote the LLS direction for the layering $\mathcal{J}$. Then the following properties hold.*

(i) *We have*

$$\|\delta_{J_k} \Delta x^{\text{ll}}_{J_k} + \delta^{-1}_{J_k} \Delta s^{\text{ll}}_{J_k} + x^{1/2}_{J_k} s^{1/2}_{J_k}\| \le 6n\ell^\delta(\mathcal{J})\sqrt{\mu}\,, \quad \forall k \in [p], \text{ and} \tag{4.28}$$

$$\|\delta \Delta x^{\text{ll}} + \delta^{-1} \Delta s^{\text{ll}} + x^{1/2} s^{1/2}\| \le 6n^{3/2}\ell^\delta(\mathcal{J})\sqrt{\mu}\,. \tag{4.29}$$

(ii) *For the affine scaling direction $\Delta w^{\text{a}} = (\Delta x^{\text{a}}, \Delta y^{\text{a}}, \Delta s^{\text{a}})$,*

$$\|Rx^{\text{ll}} - Rx^{\text{a}}\|, \|Rs^{\text{ll}} - Rs^{\text{a}}\| \le 6n^{3/2}\ell^\delta(\mathcal{J})\,.$$

(iii) *For the residuals of the LLS steps we have $\|Rx^{\text{ll}}\|, \|Rs^{\text{ll}}\| \le \sqrt{2n}$. For each $i \in [n]$, $\max\{|Rx^{\text{ll}}_i|, |Rs^{\text{ll}}_i|\} \ge \frac{1}{2} - \frac{3}{4}\beta$.*

(iv) *Let $\varepsilon^{\text{ll}}(w) = \max_{i\in[n]} \min\{|Rx^{\text{ll}}_i|, |Rs^{\text{ll}}_i|\}$, and define the step length as*

$$\alpha := \sup\{\alpha' \in [0, 1] : \forall \bar{\alpha} \in [0, \alpha'] : w + \bar{\alpha}\Delta w^{\text{ll}} \in \mathcal{N}(2\beta)\}\,.$$

We obtain the following bounds on the progress in the LLS step:

$$\mu(w + \alpha\Delta w^{\text{ll}}) = (1 - \alpha)\mu\,, \quad \text{and}$$

$$\alpha \ge 1 - \frac{3\sqrt{n}\varepsilon^{\text{ll}}(w)}{\beta}.$$

(v) *We have* $\varepsilon^{\mathrm{ll}}(w) = 0$ *if and only if* $\alpha = 1$. *These are further equivalent to* $w + \Delta w^{\mathrm{ll}} = (x + \Delta x^{\mathrm{ll}}, y + \Delta y^{\mathrm{ll}}, s + \Delta s^{\mathrm{ll}})$ *being an optimal solution to* (4.1).

Proof. Again, we use $\lambda = \ell^{\delta}(\mathcal{J})$.

Part (i)**.** Clearly, (4.28) implies (4.29). To show (4.28), we use Lemma 4.5.3 to obtain $\Delta\bar{x}$ and $\Delta\bar{s}$ as in (4.36) and (4.37). We will also use $\Delta x \in \bigoplus_{k=1}^{p} W_{\mathcal{J},k}$ and $\Delta s \in \bigoplus_{k=1}^{p} W_{\mathcal{J},k}^{\perp}$ as in (4.34) and (4.35).

Select any layer $k \in [p]$. From (4.34), we get that

$$\|\delta_{J_k}\Delta\bar{x}_{J_k} + \delta_{J_k}^{-1}\Delta s_{J_k} + x_{J_k}^{1/2}s_{J_k}^{1/2}\| = \|\delta_{J_k}(\Delta\bar{x}_{J_k} - \Delta x_{J_k}^{\mathrm{ll}})\| \le 2n\lambda\sqrt{\mu}\,. \tag{4.39}$$

Similarly, from (4.35), we see that

$$\|\delta_{J_k}^{-1}\Delta\bar{s}_{J_k} + \delta_{J_k}\Delta x_{J_k} + x_{J_k}^{1/2}s_{J_k}^{1/2}\| = \|\delta_{J_k}^{-1}(\Delta\bar{s}_{J_k} - \Delta s_{J_k}^{\mathrm{ll}})\| \le 2n\lambda\sqrt{\mu}\,.$$

From the above inequalities, we see that

$$\|\delta_{J_k}(\Delta\bar{x}_{J_k} - \Delta x_{J_k}) + \delta_{J_k}^{-1}(\Delta s_{J_k} - \Delta\bar{s}_{J_k})\| \le 4n\lambda\sqrt{\mu}\,.$$

Since $\delta_{J_k}(\Delta\bar{x}_{J_k} - \Delta x_{J_k})$ and $\delta_{J_k}^{-1}(\Delta s_{J_k} - \Delta\bar{s}_{J_k})$ are orthogonal vectors, we have

$$\|\delta_{J_k}(\Delta\bar{x}_{J_k} - \Delta x_{J_k})\|,\ \|\delta_{J_k}^{-1}(\Delta s_{J_k} - \Delta\bar{s}_{J_k})\| \le 4n\lambda\sqrt{\mu}\,.$$

Together with (4.36), this yields $\|\delta_{J_k}(\Delta x_{J_k}^{\mathrm{ll}} - \Delta x_{J_k})\| \le 6n\lambda\sqrt{\mu}$. Combined with (4.27), we get

$$\|\delta_{J_k}\Delta x_{J_k}^{\mathrm{ll}} + \delta_{J_k}^{-1}\Delta s_{J_k}^{\mathrm{ll}} + x_{J_k}^{1/2}s_{J_k}^{1/2}\| = \|\delta_{J_k}(\Delta x_{J_k}^{\mathrm{ll}} - \Delta x_{J_k})\| \le 6n\lambda\sqrt{\mu}\,,$$

thus, (4.28) follows.

Part (ii)**.** Recall from Lemma 4.3.5(i) that $\sqrt{\mu}Rx^{\mathrm{a}} + \sqrt{\mu}Rs^{\mathrm{a}} = x^{1/2}s^{1/2}$. From part (i), we can similarly see that

$$\|\sqrt{\mu}Rx^{\mathrm{ll}} + \sqrt{\mu}Rs^{\mathrm{ll}} - x^{1/2}s^{1/2}\| \le 6n^{3/2}\lambda\sqrt{\mu}\,.$$

From these, we get

$$\|(Rx^{\mathrm{ll}} - Rx^{\mathrm{a}}) + (Rs^{\mathrm{ll}} - Rs^{\mathrm{a}})\| \le 6n^{3/2}\lambda\,.$$

The claim follows since $Rx^{\mathrm{ll}} - Rx^{\mathrm{a}} \in \mathrm{Diag}(\delta)W$ and $Rs^{\mathrm{ll}} - Rs^{\mathrm{a}} \in \mathrm{Diag}(\delta^{-1})W^{\perp}$ are orthogonal vectors.

Part (iii)**.** Both bounds follow from the previous part and Lemma 4.3.5(iii), using the assumption $\ell^{\delta}(\mathcal{J}) \le \beta/(32n^2)$.

Part (iv)**.** Let $w^+ = w + \alpha\Delta w^{\text{ll}}$. We need to find the largest value $\alpha > 0$ such that $w^+ \in \mathcal{N}(2\beta)$. To begin, we first show that the normalized duality gap $\mu(w^+) = (1-\alpha)\mu$ for any $\alpha \in \mathbb{R}$. For this purpose, we use the decomposition:

$$(x+\alpha\Delta x^{\text{ll}})(s+\alpha\Delta s^{\text{ll}}) = (1-\alpha)xs + \alpha(x+\Delta x^{\text{ll}})(s+\Delta s^{\text{ll}}) - \alpha(1-\alpha)\Delta x^{\text{ll}}\Delta s^{\text{ll}}. \quad (4.40)$$

Recall from Part (i) that there exists $\Delta x \in \bigoplus_{k=1}^{p} W_{\mathcal{J},k}$ and $\Delta s \in \bigoplus_{k=1}^{p} W_{\mathcal{J},k}^{\perp}$ as in (4.34) and (4.35) such that $\delta\Delta x^{\text{ll}} + \delta^{-1}\Delta s = -\delta x$ and $\delta\Delta x + \delta^{-1}\Delta s^{\text{ll}} = -\delta^{-1}s$. In particular, $x+\Delta x^{\text{ll}} = \delta^{-2}\Delta s$ and $s+\Delta s^{\text{ll}} = \delta^2\Delta x$. Noting that $\Delta x^{\text{ll}} \perp \Delta s^{\text{ll}}$ and $\Delta x \perp \Delta s$, taking the average of the coordinates on both sides of (4.40), we get that

$$\begin{aligned}\mu(w+\alpha\Delta w^{\text{ll}}) &= (1-\alpha)\mu(w) + \alpha\langle x+\Delta x^{\text{ll}}, s+\Delta s^{\text{ll}}\rangle/n - \alpha(1-\alpha)\langle\Delta x^{\text{ll}}, \Delta s^{\text{ll}}\rangle/n \\ &= (1-\alpha)\mu(w) + \alpha\langle\delta^{-2}\Delta s, \delta^2\Delta x\rangle/n \\ &= (1-\alpha)\mu(w), \end{aligned} \quad (4.41)$$

as needed.

Let $\varepsilon := \varepsilon^{\text{ll}}(w)$. To obtain the desired lower bound on the step-length, given (4.41) it suffices to show that for all $0 \le \alpha < 1 - \frac{3\sqrt{n}\varepsilon}{\beta}$ that

$$\left\| \frac{(x+\alpha\Delta x^{\text{ll}})(s+\alpha\Delta s^{\text{ll}})}{(1-\alpha)\mu} - \vec{1} \right\| \le 2\beta\,. \quad (4.42)$$

We will need a bound on the product of the LLS residuals:

$$\begin{aligned}\left\| Rx^{\text{ll}}Rs^{\text{ll}} - \frac{1}{\mu}\Delta x^{\text{ll}}\Delta s^{\text{ll}} \right\| &= \left\| \frac{x^{1/2}s^{1/2}}{\sqrt{\mu}} \cdot \frac{\delta\Delta x^{\text{ll}} + \delta^{-1}\Delta s^{\text{ll}} + x^{1/2}s^{1/2}}{\sqrt{\mu}} \right\| \\ &\le 6(1+2\beta)n^{3/2}\lambda \le \frac{\beta}{4}\,, \end{aligned} \quad (4.43)$$

using Proposition 4.3.1, part (i), and the assumptions $\lambda \le \beta/(32n^2)$, $\beta \le 1/4$. Another useful bound will be

$$\begin{aligned}\|Rx^{\text{ll}}Rs^{\text{ll}}\|^2 &= \sum_{i\in[n]} \left|Rx_i^{\text{ll}}\right|^2 \left|Rs_i^{\text{ll}}\right|^2 \le \varepsilon^2 \sum_{i\in[n]} \max\left\{ \left|Rx_i^{\text{ll}}\right|^2, \left|Rs_i^{\text{ll}}\right|^2 \right\} \\ &\le \varepsilon^2(\|Rx^{\text{ll}}\|^2 + \|Rs^{\text{ll}}\|^2) \le 2n\varepsilon^2\,. \end{aligned} \quad (4.44)$$

The last inequality uses part (ii). We are ready to get a bound as in (4.42).

$$\begin{aligned}
&\left\|\frac{(x+\alpha\Delta x^{\mathrm{ll}})(s+\alpha\Delta s^{\mathrm{ll}})}{(1-\alpha)\mu}-\vec{1}\right\| \\
&\leq \beta+\left\|\frac{\alpha}{(1-\alpha)\mu}(x+\Delta x^{\mathrm{ll}})(s+\Delta s^{\mathrm{ll}})+\frac{\alpha}{\mu}(xs+x\Delta s^{\mathrm{ll}}+s\Delta x^{\mathrm{ll}})\right\| \\
&\leq \beta+\frac{\alpha}{1-\alpha}\|Rx^{\mathrm{ll}}Rs^{\mathrm{ll}}\|+\alpha\left\|Rx^{\mathrm{ll}}Rs^{\mathrm{ll}}-\frac{1}{\mu}\Delta x^{\mathrm{ll}}\Delta s^{\mathrm{ll}}\right\| \\
&\leq \beta+\frac{\sqrt{2n}\varepsilon}{1-\alpha}+\frac{\beta}{4}\leq\frac{5}{4}\beta+\frac{\sqrt{2n}\varepsilon}{1-\alpha}\,.
\end{aligned}$$

This value is $\leq 2\beta$ whenever $2\sqrt{n}\varepsilon/(1-\alpha)\leq(3/4)\beta \Leftarrow \alpha<1-\frac{3\sqrt{n}\varepsilon}{\beta}$, as needed.

Part (v)**.** From part (iv), it is immediate that $\varepsilon^{\mathrm{ll}}(w)=0$ implies $\alpha=1$. If $\alpha=1$, we have that $w+\Delta w^{\mathrm{ll}}$ is the limit of (strictly) feasible solutions to (4.1) and thus is also a feasible solution. Optimality of $w+\Delta w^{\mathrm{ll}}$ now follows from Part (iv), since $\alpha=1$ implies $\mu(w+\Delta w^{\mathrm{ll}})=0$. The remaining implication is that if $w+\Delta w^{\mathrm{ll}}$ is optimal, then $\varepsilon^{\mathrm{ll}}(w)=0$. Recall that $Rx_i^{\mathrm{ll}}=\delta_i(x_i+\Delta x_i^{\mathrm{ll}})/\sqrt{\mu}$ and $Rs_i^{\mathrm{ll}}=\delta_i^{-1}(s_i+\Delta s_i^{\mathrm{ll}})/\sqrt{\mu}$. The optimality of $w+\Delta w^{\mathrm{ll}}$ means that for each $i\in[n]$, either $x_i+\Delta x_i^{\mathrm{ll}}=\vec{0}$ or $s_i+\Delta s_i^{\mathrm{ll}}=\vec{0}$. Therefore, $\varepsilon^{\mathrm{ll}}(w)=0$. □

4.6 Proofs of the main lemmas for the potential analysis

Lemma 4.4.2 (Repetition)**.** *Let $w=(x,y,s)\in\mathcal{N}(\beta)$ for $\beta\in(0,1/8]$ and let $w^*=(x^*,y^*,s^*)$ be the optimal solution corresponding to $\mu^*=0$ on the central path. Let further $\mathcal{J}=(J_1,\ldots,J_p)$ be a $\delta(w)$-balanced layering (Definition* 4.3.13*), and let $\Delta w^{\mathrm{ll}}=(\Delta x^{\mathrm{ll}},\Delta y^{\mathrm{ll}},\Delta s^{\mathrm{ll}})$ be the corresponding LLS direction. Then the following statement holds for every $q\in[p]$:*

(i) *There exists $i\in J_q$ such that*

$$x_i^*\geq\frac{2x_i}{3\sqrt{n}}\cdot(\|Rx_{J_q}^{\mathrm{ll}}\|-2\gamma n)\,. \tag{4.31}$$

(ii) *There exists $j\in J_q$ such that*

$$s_j^*\geq\frac{2s_j}{3\sqrt{n}}\cdot(\|Rs_{J_q}^{\mathrm{ll}}\|-2\gamma n)\,. \tag{4.32}$$

Proof of Lemma 4.4.2. We prove part (i); part (ii) follows analogously when using Lemma 4.3.9. Let z be a vector fulfilling the statement of Lemma 4.5.2 for $u = x^*$, $v = x + \Delta x^{\mathrm{ll}}$, and $I = J_{>q}$. Then $z \in W + d$, $z_{J_{>q}} = x_{J_{>q}} + \Delta x^{\mathrm{ll}}_{J_{>q}}$ and by $\ell^\delta(\mathcal{J}) \leq \gamma$

$$\left\|\delta_{J_{\leq q}}(x^*_{J_{\leq q}} - z_{J_{\leq q}})\right\| \leq \gamma \left\|\delta_{J_{>q}}\left(x^*_{J_{>q}} - (x_{J_{>q}} + \Delta x^{\mathrm{ll}}_{J_{>q}})\right)\right\|.$$

Restricting to the components in J_q, and dividing by $\sqrt{\mu}$, we get

$$\left\|\frac{\delta_{J_q}(x^*_{J_q} - z_{J_q})}{\sqrt{\mu}}\right\| \leq \gamma \left\|\frac{\delta_{J_{>q}}\left(x^*_{J_{>q}} - (x_{J_{>q}} + \Delta x^{\mathrm{ll}}_{J_{>q}})\right)}{\sqrt{\mu}}\right\| \leq \gamma \left\|\frac{\delta_{J_{>q}} x^*_{J_{>q}}}{\sqrt{\mu}}\right\| + \gamma \|R x^{\mathrm{ll}}_{J_{>q}}\|. \tag{4.45}$$

Since $w \in \mathcal{N}(\beta)$, from Proposition 4.3.1 and (4.17) we see that for $i \in [n]$

$$\frac{\delta_i}{\sqrt{\mu}} \leq \frac{1}{\sqrt{1-2\beta}} \cdot \frac{\delta_i(w(\mu))}{\sqrt{\mu}} = \frac{1}{\sqrt{1-2\beta}} \cdot \frac{1}{x_i(\mu)},$$

and therefore

$$\left\|\frac{\delta_{J_{>q}} x^*_{J_{>q}}}{\sqrt{\mu}}\right\| \leq \frac{1}{\sqrt{1-2\beta}} \left\|x(\mu)^{-1}_{J_{>q}} x^*_{J_{>q}}\right\| \leq \frac{1}{\sqrt{1-2\beta}} \cdot \left\|x(\mu)^{-1}_{J_{>q}} x^*_{J_{>q}}\right\|_1 \leq \frac{n}{\sqrt{1-2\beta}},$$

where the last inequality follows by Lemma 4.3.3.

Using the above bounds with (4.45), along with $\|R x^{\mathrm{ll}}_{J_{\geq q}}\| \leq \|R x^{\mathrm{ll}}\| \leq \sqrt{2n}$ from Lemma 4.3.10(iii), we get

$$\left\|\frac{\delta_{J_q} z_{J_q}}{\sqrt{\mu}}\right\| \leq \left\|\frac{\delta_{J_q} x^*_{J_q}}{\sqrt{\mu}}\right\| + \frac{\gamma n}{\sqrt{1-2\beta}} + \gamma\sqrt{2n} \leq \left\|\frac{\delta_{J_q} x^*_{J_q}}{\sqrt{\mu}}\right\| + 2\gamma n,$$

using that $\beta \leq 1/8$ and $n \geq 3$. Note that z is a feasible solution to the least-squared problem which is optimally solved by $x^{\mathrm{ll}}_{J_q}$ for layer J_q and so

$$\|R x^{\mathrm{ll}}_{J_q}\| \leq \left\|\frac{\delta_{J_q} z_{J_q}}{\sqrt{\mu}}\right\|.$$

It follows that

$$\left\|\frac{\delta_{J_q} x^*_{J_q}}{\sqrt{\mu}}\right\| \geq \|R x^{\mathrm{ll}}_{J_q}\| - 2\gamma n.$$

Let us pick $i = \arg\max_{t \in J_q} |\delta_t x^*_t|$. Using Proposition 4.3.2,

$$\frac{x^*_i}{x_i} \geq \frac{1}{1+\beta} \cdot \frac{\delta_i x^*_i}{\sqrt{\mu}} \geq \frac{\|R x^{\mathrm{ll}}_{J_q}\| - 2\gamma n}{(1+\beta)\sqrt{n}} \geq \frac{2}{3\sqrt{n}} \cdot (\|R x^{\mathrm{ll}}_{J_q}\| - 2\gamma n),$$

completing the proof. □

Lemma 4.4.3 (Repetition)**.** *Let $w = (x, y, s) \in \mathcal{N}(2\beta)$ for $\beta \in (0, 1/8]$, let $\mu = \mu(w)$ and $\delta = \delta(w)$. Let $i, j \in [n]$ and $2 \le \tau \le n$ such that for the optimal solution $w^* = (x^*, y^*, s^*)$, we have $x_i^* \ge \beta x_i/(2^{10}n^{5.5})$ and $s_j^* \ge \beta s_j/(2^{10}n^{5.5})$, and assume $\varrho^\mu(i, j) \ge -\tau$. After $O(\beta^{-1}\sqrt{n}\tau \log(\bar{\chi}^* + n))$ further iterations the duality gap μ' fulfills $\Psi^{\mu'}(i, j) \ge 2\tau$, and for every $\ell \in [n] \setminus \{i, j\}$, either $\Psi^{\mu'}(i, \ell) \ge 2\tau$, or $\Psi^{\mu'}(\ell, j) \ge 2\tau$.*

Proof of Lemma 4.4.3. Let us select a value μ' such that

$$\log \mu - \log \mu' \ge 5\tau \log\left(\frac{4n\kappa^*}{\gamma}\right) + 31 \log n + 44 - 4 \log \beta\,.$$

The normalized duality gap decreases to such value within $O(\beta^{-1}\sqrt{n}\tau \cdot \log(\bar{\chi}^* + n))$ iterations, recalling that $\log(\bar{\chi}^* + n) = \Theta(\log(\kappa^* + n))$. The step-lengths for the affine scaling and LLS steps are stated in Proposition 4.3.4 and Lemma 4.3.10(iv).Whenever the algorithm chooses an LLS step, $\varepsilon^{\mathrm{a}}(w) < 10n^{3/2}\gamma$. Thus, the progress in μ will be at least as much (in fact, much better) than the $(1 - \beta/\sqrt{n})$ guarantee we use for the affine scaling step in Proposition 4.3.4.

Let $w' = (x', y', s')$ be the central path element corresponding to μ', and let $\delta' = \delta(w')$. From now on we use the shorthand notation

$$\Theta := \log\left(\frac{4n\kappa^*}{\gamma}\right).$$

We first show that

$$\Theta\varrho^{\mu'}(i, j) \ge 4\Theta\tau + 18 \log n + 22 \log 2 - 2 \log \beta \tag{4.46}$$

for μ', and therefore, $\Theta\Psi^{\mu'}(i, j) \ge \min(2\Theta n, 4\Theta\tau + 18 \log n + 22 \log 2 - 2 \log \beta) \ge 2\Theta\tau$ as $\tau \le n$. Recalling the definition $\kappa_{ij}^\delta = \kappa_{ij}\delta_j/\delta_i$, we see that according to Proposition 4.3.2,

$$\kappa_{ij}^\delta \le \frac{\kappa_{ij}}{(1-\beta)^2} \cdot \frac{x_i s_j}{\mu}, \quad \text{and} \quad \kappa_{ij}^{\delta'} = \kappa_{ij} \cdot \frac{x_i' s_j'}{\mu'}\,.$$

Thus,

$$\begin{aligned}
&\Theta\varrho^{\mu'}(i, j)\\
&\ge \Theta\varrho^{\mu}(i, j) + \log \mu - \log \mu' + 2\log(1-\beta) + \log x_i' - \log x_i + \log s_j' - \log s_j\\
&\ge \Theta\varrho^{\mu}(i, j) + 5\Theta\tau + 31 \log n + 44 - 4 \log \beta + 2\log(1-\beta) + \log(x_i'/x_i) + \log(s_j'/s_j).
\end{aligned}$$

Using the near-monotonicity of the central path (Lemma 4.3.3), we have $x'_i \geq x^*_i/n$ and $s'_j \geq s^*_j/n$. Together with our assumptions $x^*_i \geq \beta x_i/(2^{10}n^{5.5})$ and $s^*_i \geq \beta s_i/(2^{10}n^{5.5})$, we see that

$$\log(x'_i/x_i) + \log(s'_j/s_j) \geq -13\log n - 20\log 2 + 2\log\beta\,.$$

Using the assumption $\varrho^{\mu}(i,j) > -\tau$ of the lemma, we can establish (4.46) as $\beta < 1/8$.

Next, consider any $\ell \in [n]\setminus\{i,j\}$. From the triangle inequality Lemma 4.2.16(ii) it follows that $\kappa^{\delta'}_{ij} \leq \kappa^{\delta'}_{i\ell}\cdot\kappa^{\delta'}_{\ell j}$, which gives $\varrho^{\mu'}(i,\ell) + \varrho^{\mu'}(\ell,j) \geq \varrho^{\mu'}(i,j)$. We therefore get

$$\max\{\Theta\varrho^{\mu'}(i,\ell), \Theta\varrho^{\mu'}(\ell,j)\} \geq \frac{1}{2}\Theta\varrho^{\mu'}(i,j) \overset{(4.46)}{\geq} 2\Theta\tau + 9\log n + 11\log 2 - \log\beta.$$

We next show that if $\Theta\varrho^{\mu'}(i,\ell) \geq 2\Theta\tau + 9\log n + 11\log 2 - \log\beta$, then $\Psi^{\mu'}(i,\ell) \geq 2\tau$. The case $\Theta\varrho^{\mu'}(\ell,j) \geq 2\Theta\tau + 9\log n + 11\log 2 - \log\beta$ follows analogously.

Consider any $0 < \bar{\mu} < \mu'$ with the corresponding central path point $\bar{w} = (\bar{x},\bar{y},\bar{s})$. The proof is complete by showing $\Theta\varrho^{\bar{\mu}}(i,\ell) \geq \Theta\varrho^{\mu'}(i,\ell) - 9\log n - 11\log 2 + \log\beta$. Recall that for central path elements, we have $\kappa^{\delta'}_{ij} = \kappa_{ij}x'_i/x'_j$, and $\kappa^{\bar{\delta}}_{ij} = \kappa_{ij}\bar{x}_i/\bar{x}_j$. Therefore

$$\Theta\varrho^{\bar{\mu}}(i,j) = \Theta\varrho^{\mu'}(i,j) + \log\bar{x}_i - \log x'_i - \log\bar{x}_j + \log x'_j\,.$$

Using Proposition 4.3.1, Lemma 4.3.3 and the assumption $x^*_i \geq \beta x_i/(2^{10}n^{5.5})$, we have $\bar{x}_j \leq nx'_j$ and

$$\bar{x}_i \geq \frac{x^*_i}{n} \geq \frac{\beta x_i}{2^{10}n^{6.5}} \geq \frac{\beta(1-\beta)x'_i}{2^{10}n^{7.5}} \geq \frac{\beta x'_i}{2^{11}n^{7.5}}\,.$$

Using these bounds, we get

$$\Theta\varrho^{\bar{\mu}}(i,j) \geq \Theta\varrho^{\mu'}(i,j) - 9\log n - 11\log 2 + \log\beta,$$

completing the proof. □

It remains to prove Lemma 4.4.4 and Lemma 4.4.5, addressing the more difficult case $\xi^{\mathrm{II}}_{\mathcal{J}} < 4\gamma n$. It is useful to decompose the variables into two sets. We let

$$\boldsymbol{B} := \{t \in [n] : |Rs^{\mathrm{II}}_t| < 4\gamma n\}, \quad \text{and} \quad \boldsymbol{N} := \{t \in [n] : |Rx^{\mathrm{II}}_t| < 4\gamma n\}\,. \tag{4.47}$$

The assumption $\xi^{\mathrm{II}}_{\mathcal{J}} < 4\gamma n$ implies that for every layer J_k, either $J_k \subseteq \boldsymbol{B}$ or $J_k \subseteq \boldsymbol{N}$. The next two lemmas describe the relations between δ and δ^+.

Lemma 4.6.1. *Let $w \in \mathcal{N}(\beta)$ for $\beta \in (0, 1/8]$, and assume $\ell^\delta(\mathcal{J}) \le \gamma$ and $\varepsilon^{\mathrm{ll}}(w) < 4\gamma n$. For the next iterate $w^+ = (x^+, y^+, s^+) \in \mathcal{N}(2\beta)$, we have*

(i) *For $i \in \boldsymbol{B}$,*

$$\frac{1}{2} \cdot \sqrt{\frac{\mu^+}{\mu}} \le \frac{\delta_i^+}{\delta_i} \le 2 \cdot \sqrt{\frac{\mu^+}{\mu}} \quad \text{and} \quad \delta_i^{-1} s_i^+ \le \frac{3\mu^+}{\sqrt{\mu}} .$$

(ii) *For $i \in \boldsymbol{N}$,*

$$\frac{1}{2} \cdot \sqrt{\frac{\mu}{\mu^+}} \le \frac{\delta_i^+}{\delta_i} \le 2 \cdot \sqrt{\frac{\mu}{\mu^+}} \quad \text{and} \quad \delta_i x_i^+ \le \frac{3\mu^+}{\sqrt{\mu}} .$$

(iii) *If $i, j \in \boldsymbol{B}$ or $i, j \in \boldsymbol{N}$, then*

$$\frac{1}{4} \le \frac{\kappa_{ij}^\delta}{\kappa_{ij}^{\delta^+}} = \frac{\delta_i^+ \delta_j}{\delta_i \delta_j^+} \le 4 .$$

(iv) *If $i \in \boldsymbol{N}$ and $j \in \boldsymbol{B}$, then*

$$\frac{\kappa_{ij}^\delta}{\kappa_{ij}^{\delta^+}} \ge 4n^{3.5} .$$

Proof. **Part (i).** By Lemma 4.3.10(i), we see that

$$\begin{aligned} \|\delta_B \Delta x_B^{\mathrm{ll}}\|_\infty &\le \|\delta_B \Delta x_B^{\mathrm{ll}} + \delta_B^{-1} \Delta s_B^{\mathrm{ll}} + x_B^{1/2} s_B^{1/2}\|_\infty + \|\delta_B^{-1}(\Delta s_B^{\mathrm{ll}} + s_B)\|_\infty \\ &= \|\delta_B \Delta x_B^{\mathrm{ll}} + \delta_B^{-1} \Delta s_B^{\mathrm{ll}} + x_B^{1/2} s_B^{1/2}\|_\infty + \sqrt{\mu}\|R s_B^{\mathrm{ll}}\|_\infty \\ &\le \sqrt{\mu}\left(6n\ell^\delta(\mathcal{J}) + 4n\gamma\right) \le 10n\gamma\sqrt{\mu} \le \sqrt{\mu}/64 , \end{aligned}$$

by the assumption on $\ell^\delta(\mathcal{J})$ and the definition of $\boldsymbol{B}$.

By construction of the LLS step, $|x_i^+ - x_i| = \alpha^+ |\Delta x^{\mathrm{ll}}| \le |\Delta x^{\mathrm{ll}}|$, recalling that $0 \le \alpha^+ \le 1$. Using the bound derived above, for $i \in \boldsymbol{B}$ we get

$$\left|\frac{x_i^+}{x_i} - 1\right| \le \left|\frac{\Delta x_i^{\mathrm{ll}}}{x_i}\right| = \frac{|\delta_i \Delta x_i^{\mathrm{ll}}|}{\delta_i x_i} \le \frac{\sqrt{\mu}}{64\delta_i x_i} \le \frac{1}{32} ,$$

where the last inequality follows from Proposition 4.3.2. As

$$\frac{\delta_i^+}{\delta_i} = \sqrt{\frac{x_i^+ s_i^+}{x_i s_i}} \cdot \frac{x_i}{x_i^+} \quad \text{and} \quad \frac{1 - 2\beta}{1 + \beta} \frac{\sqrt{\mu^+}}{\sqrt{\mu}} \le \sqrt{\frac{x_i^+ s_i^+}{x_i s_i}} \le \frac{1 + 2\beta}{1 - \beta} \frac{\sqrt{\mu^+}}{\sqrt{\mu}}$$

by Proposition 4.3.2 the claimed bounds follow with $\beta \le 1/8$.

To get the upper bound on $\delta_i^{-1} s_i^+$, again with Proposition 4.3.2

$$\delta_i^{-1} s_i^+ = \frac{\delta_i^+}{\delta_i \delta_i^+} s_i^+ = \frac{\delta_i^+}{\delta_i} \cdot \sqrt{x_i^+ s_i^+} \le 2\sqrt{\frac{\mu^+}{\mu}} \cdot (1 + 2\beta)\sqrt{\mu^+} \le \frac{3\mu^+}{\sqrt{\mu}} .$$

Part (ii)**.** Analogously to (i).

Part (iii)**.** Immediate from parts (i) and (ii).

Part (iv)**.** Follows by parts (i) and (ii), and by the lower bound on $\sqrt{\mu/\mu^+}$ obtained from Lemma 4.3.10(iv) as follows

$$\frac{\kappa_{ij}^{\delta}}{\kappa_{ij}^{\delta^+}} = \frac{\delta_i^+\delta_j}{\delta_i\delta_j^+} \geq \frac{\mu}{4\mu^+} = \frac{1}{4(1-\alpha^+)} \geq \frac{\beta}{12\sqrt{n}\varepsilon^{\mathrm{ll}}(w)} \geq 4n^{3.5}. \qquad \square$$

Lemma 4.4.4 (Repetition)**.** *Let $w = (x, y, s) \in \mathcal{N}(\beta)$ for $\beta \in (0, 1/8]$, and let $\mathcal{J} = (J_1, \ldots, J_p)$ be a $\delta(w)$-balanced partition. Assume that $\xi_{\mathcal{J}}^{\mathrm{ll}}(w) < 4\gamma n$, and let $w^+ = (x^+, y^+, s^+) \in \mathcal{N}(2\beta)$ be the next iterate obtained by the LLS step with $\mu^+ = \mu(w^+)$ and assume $\mu^+ > 0$. Let $q \in [p]$ such that $\xi_{\mathcal{J}}^{\mathrm{ll}}(w) = \xi_{J_q}^{\mathrm{ll}}(w)$. If $\ell^{\delta^+}(\mathcal{J}) \leq 4\gamma n$, then there exist $i, j \in J_q$ such that $x_i^* \geq \beta x_i^+/(16n^{3/2})$ and $s_j^* \geq \beta s_j^+/(16n^{3/2})$. Further, for any $\ell, \ell' \in J_q$, we have $\varrho^{\mu^+}(\ell, \ell') \geq -|J_q|$.*

Proof of Lemma 4.4.4. Without loss of generality, let $\xi_{\mathcal{J}}^{\mathrm{ll}} = \xi_{J_q}^{\mathrm{ll}} = \|Rx_{J_q}^{\mathrm{ll}}\|$ for a layer q with $J_q \subseteq \boldsymbol{N}$. The case $\xi_{J_q}^{\mathrm{ll}} = \|Rs_{J_q}^{\mathrm{ll}}\|$ and $J_q \subseteq \boldsymbol{B}$ can be treated analogously.

By Lemma 4.3.10(iii), $\|Rs_{J_q}^{\mathrm{ll}}\| \geq \frac{1}{2} - \frac{3}{4}\beta > \frac{1}{4} + 2n\gamma$, and therefore Lemma 4.4.2 provides a $j \in J_q$ such that $s_j^*/s_j \geq 1/(6\sqrt{n})$. Using Lemma 4.3.3 and Proposition 4.3.1 we find that $s_j^+/s_j \leq 2n$ and so $s_j^*/s_j^+ = s_j^*/s_j \cdot s_j/s_j^+ \geq 1/(12n^{3/2}) > 1/(16n^{3/2})$.

The final statement $\varrho^{\mu^+}(\ell, \ell') \geq -|J_q|$ for any $\ell, \ell' \in J_q$ is also straightforward. From Lemma 4.6.1(iii) and the strong connectivity of J_q in $G_{\delta,\gamma/n}$, we obtain that J_q is strongly connected in $G_{\delta^+,\gamma/(4n)}$. Hence, $\varrho^{\mu^+}(\ell, \ell') \geq -|J_q|$ follows by Lemma 4.4.1.

The rest of the proof is dedicated to showing the existence of an $i \in J_q$ such that $x_i^* \geq \beta x_i^+/(16n^{3/2})$. For this purpose, we will prove following claim.

Claim 4.6.2. $\|\delta_{J_q} x_{J_q}^*\| \geq \frac{\beta\mu^+}{8\sqrt{n\mu}}$.

In order to prove Claim 4.6.2, we define

$$z := (\delta^+)^{-1} L_{J_{>q}}^{\delta^+} \left(\delta_{J_{>q}}^+ (x_{J_{>q}}^* - x_{J_{>q}}^+)\right) \quad \text{and } w := x^* - x^+ - z\,,$$

as in Lemma 4.5.2. By construction, $w \in W$ and $w_{J_{>q}} = \vec{0}$. Thus, $w_{J_q} \in W_{\mathcal{J},q}$ as defined in Section 4.3.4.

Using the triangle inequality, we get

$$\|\delta_{J_q} x_{J_q}^*\| \geq \|\delta_{J_q}(x_{J_q}^+ + w_{J_q})\| - \|\delta_{J_q} z_{J_q}\|\,. \tag{4.48}$$

We bound the two terms separately, starting with an upper bound on $\|\delta_{J_q} z_{J_q}\|$. Since $\ell^{\delta^+}(\mathcal{J}) \le 4\gamma n$, we have with Lemma 4.5.2 that

$$\begin{aligned}
\left\|\delta^+_{J_q} z_{J_q}\right\| &\le \ell^{\delta^+}(\mathcal{J}) \left\|\delta^+_{J_{>q}} \left(x^*_{J_{>q}} - x^+_{J_{>q}}\right)\right\| \\
&\le 4n\gamma \left\|\delta^+_{J_{>q}} \left(x^*_{J_{>q}} - x^+_{J_{>q}}\right)\right\| \\
&= 4n\gamma \left\|\delta^+_{J_{>q}} x^+_{J_{>q}} \left(\frac{x^*_{J_{>q}}}{x^+_{J_{>q}}} - \vec{1}\right)\right\| \\
&\le 4n\gamma \left(\|\delta^+ x^+\|_\infty \cdot \left\|\frac{x^*}{x^+}\right\|_1 + \sqrt{n\mu^+}\right) \\
&\le 4n\gamma \left(\frac{3}{2}\sqrt{\mu^+} \cdot \frac{4}{3}n + \sqrt{n\mu^+}\right) \\
&\le 16n^2\sqrt{\mu^+}\gamma,
\end{aligned} \tag{4.49}$$

where the penultimate inequality follows by Proposition 4.3.1, Proposition 4.3.2 and Lemma 4.3.3. We can use this and Lemma 4.6.1(ii) to obtain

$$\|\delta_{J_q} z_{J_q}\| \le \|\delta_{J_q}/\delta^+_{J_q}\|_\infty \cdot \|\delta^+_{J_q} z_{J_q}\| \le \frac{32n^2\gamma\mu^+}{\sqrt{\mu}} \le \frac{\beta\mu^+}{32n^3\sqrt{\mu}}\,, \tag{4.50}$$

using the definition of γ.

The first RHS term in (4.48) will be bounded as follows.

Claim 4.6.3. $\|\delta_{J_q}(x^+_{J_q} + w_{J_q})\| \ge \frac{1}{2}\sqrt{\mu}\xi^{\mathrm{ll}}_{\mathcal{J}}$.

Proof. [Claim 4.6.3] We recall the characterization (4.26) of the LLS step $\Delta x^{\mathrm{ll}} \in W$. Namely, there exists $\Delta s \in W^\perp_{\mathcal{J},1} \oplus \cdots \oplus W^\perp_{\mathcal{J},q}$ that is the unique solution to $\delta^{-1}\Delta s + \delta\Delta x^{\mathrm{ll}} = -\delta x$. From the above, note that

$$\|\delta^{-1}_{J_q}\Delta s_{J_q}\| = \|\delta_{J_q}(x_{J_q} + \Delta x^{\mathrm{ll}}_{J_q})\| = \sqrt{\mu}\|Rx^{\mathrm{ll}}_{J_q}\| = \sqrt{\mu}\xi^{\mathrm{ll}}_{\mathcal{J}}\,.$$

From the Cauchy-Schwarz inequality,

$$\begin{aligned}
\|\delta^{-1}_{J_q}\Delta s_{J_q}\| \cdot \|\delta_{J_q}(x^+_{J_q} + w_{J_q})\| &\ge \left|\left\langle \delta^{-1}_{J_q}\Delta s_{J_q}, \delta_{J_q}(x^+_{J_q} + w_{J_q})\right\rangle\right| \\
&= \left|\left\langle \delta^{-1}_{J_q}\Delta s_{J_q}, \delta_{J_q} x^+_{J_q}\right\rangle\right|\,.
\end{aligned} \tag{4.51}$$

Here, we used that $\Delta s_{J_q} \in W^\perp_{\mathcal{J},q}$ and $w_{J_q} \in W_{\mathcal{J},q}$. Note that

$$x^+ = x + \alpha\Delta x^{\mathrm{ll}} = x + \Delta x^{\mathrm{ll}} - (1-\alpha)\Delta x^{\mathrm{ll}} = -\delta^{-2}\Delta s - (1-\alpha)\Delta x^{\mathrm{ll}}\,.$$

Therefore,

$$\left|\left\langle \delta_{J_q}^{-1}\Delta s_{J_q}, \delta_{J_q} x^+_{J_q}\right\rangle\right| = \left|\left\langle \delta_{J_q}^{-1}\Delta s_{J_q}, -\delta_{J_q}^{-1}\Delta s_{J_q} - (1-\alpha)\delta_{J_q}\Delta x^{\mathrm{ll}}_{J_q}\right\rangle\right|$$
$$\geq \|\delta_{J_q}^{-1}\Delta s_{J_q}\|^2 - (1-\alpha)\left|\left\langle \delta_{J_q}^{-1}\Delta s_{J_q}, \delta_{J_q}\Delta x^{\mathrm{ll}}_{J_q}\right\rangle\right|.$$

By Lemma 4.5.3, there exists $\Delta\bar{x} \in W_{\mathcal{J},1}\oplus\cdots\oplus W_{\mathcal{J},p}$ such that $\|\delta_{J_q}(\Delta x^{\mathrm{ll}}_{J_q} - \Delta\bar{x}_{J_q})\| \leq 2n\ell^\delta(\mathcal{J})\sqrt{\mu}$. Therefore, using the orthogonality of Δs_{J_q} and $\Delta\bar{x}_{J_q}$, we get that

$$\left|\left\langle \delta_{J_q}^{-1}\Delta s_{J_q}, \delta_{J_q}\Delta x^{\mathrm{ll}}_{J_q}\right\rangle\right| = \left|\left\langle \delta_{J_q}^{-1}\Delta s_{J_q}, \delta_{J_q}(\Delta x^{\mathrm{ll}}_{J_q} - \Delta\bar{x}^{\mathrm{ll}}_{J_q})\right\rangle\right| \leq 2n\ell^\delta(\mathcal{J})\sqrt{\mu}\cdot\|\delta_{J_q}^{-1}\Delta s_{J_q}\|.$$

From the above inequalities, we see that

$$\|\delta_{J_q}(x^+_{J_q} + w_{J_q})\| \geq \|\delta_{J_q}^{-1}\Delta s_{J_q}\| - 2(1-\alpha)n\ell^\delta(\mathcal{J})\sqrt{\mu} = \sqrt{\mu}\xi^{\mathrm{ll}}_{\mathcal{J}} - 2(1-\alpha)n\ell^\delta(\mathcal{J})\sqrt{\mu}.$$

It remains to show $(1-\alpha)n\ell^\delta(\mathcal{J}) \leq \xi^{\mathrm{ll}}_{\mathcal{J}}/4$. From Lemma 4.3.10(iv), we obtain

$$(1-\alpha)n\ell^\delta(\mathcal{J}) \leq 3n^{3/2}\ell^\delta(\mathcal{J})\xi^{\mathrm{ll}}_{\mathcal{J}}\beta^{-1},$$

using $\xi^{\mathrm{ll}}_{\mathcal{J}} \geq \varepsilon^{\mathrm{ll}}$. The claim now follows by the assumption $\ell^\delta(\mathcal{J}) \leq \gamma$, and the choice of γ. ⌟

Proof. [Claim 4.6.2] Using Lemma 4.3.10(iv),

$$\mu^+ \leq \frac{3\sqrt{n}\xi^{\mathrm{ll}}_{\mathcal{J}}\mu}{\beta},$$

implying $\|\delta_{J_q}(x^+_{J_q} + w_{J_q})\| \geq \beta\mu^+/(6\sqrt{n\mu})$ by Claim 4.6.3. Now the claim follows using (4.48) and (4.50). ⌟

By Lemma 4.6.1(ii), we see that

$$\|\delta_{J_q}x^+_{J_q}\| \leq \sqrt{n}\|\delta_{J_q}x^+_{J_q}\|_\infty \leq \frac{3\sqrt{n}\mu^+}{\sqrt{\mu}}.$$

Thus, the lemma follows immediately from Claim 4.6.2: for at least one $i \in J_q$, we must have

$$\frac{x^*_i}{x_i} \geq \frac{\|\delta_{J_q}x^*_{J_q}\|}{\|\delta_{J_q}x^+_{J_q}\|} \geq \frac{\beta}{24n} \geq \frac{\beta}{16n^{3/2}}.$$ □

Lemma 4.4.5 (Repetition)**.** *Let $w = (x, y, s) \in \mathcal{N}(\beta)$ for $\beta \in (0, 1/8]$, and let $\mathcal{J} = (J_1, \dots, J_p)$ be a $\delta(w)$-balanced partition. Assume that $\xi^{\mathrm{ll}}_{\mathcal{J}}(w) < 4\gamma n$, and let $w^+ = (x^+, y^+, s^+) \in \mathcal{N}(2\beta)$ be the next iterate obtained by the LLS step with $\mu^+ = \mu(w^+)$ and assume $\mu^+ > 0$. If $\ell^{\delta^+}(\mathcal{J}) > 4\gamma n$, then there exist two layers J_q and J_r and $i \in J_q$ and $j \in J_r$ such that $x_i^* \geq x_i^+/(8n^{3/2})$, and $s_j^* \geq s_j^+/(8n^{3/2})$. Further, $\varrho^{\mu^+}(i, j) \geq -|J_q \cup J_r|$, and for all $\ell, \ell' \in J_q \cup J_r$, $\ell \neq \ell'$ we have $\Psi^{\mu}(\ell, \ell') \leq |J_q \cup J_r|$.*

Proof of Lemma 4.4.5. Recall the sets $\boldsymbol{B}$ and $\boldsymbol{N}$ defined in (4.47). The key is to show the existence of an edge

$$(i', j') \in E_{\delta^+, \gamma/(4n)} \quad \text{such that} \quad i' \in J_q \subseteq \boldsymbol{B}, \quad j' \in J_r \subseteq \boldsymbol{N}, \quad r < q\,. \tag{4.52}$$

Before proving the existence of such i' and j', we show how the rest of the statements follow. The existence of $i \in J_q$ and $j \in J_r$ such that $x_i^* \geq x_i^+/(8n^{3/2})$ and $s_j^* \geq s_j^+/(8n^{3/2})$ follow immediately from Lemma 4.4.2. The other statements are that $\varrho^{\mu^+}(i, j) \geq -|J_q \cup J_r|$, and for each $\ell, \ell' \in J_q \cup J_r$, $\ell \neq \ell'$, $\Psi^{\mu}(\ell, \ell') \leq |J_q \cup J_r|$. According to Lemma 4.4.1, the latter is true (even with the stronger bound $\max\{|J_q|, |J_r|\}$) whenever $\ell, \ell' \in J_q$, or $\ell, \ell' \in J_r$, or if $\ell \in J_q$ and $\ell' \in J_r$. It is left to show the lower bound on $\varrho^{\mu^+}(i, j)$ and $\Psi^{\mu}(\ell, \ell') \leq |J_q \cup J_r|$ for $\ell' \in J_q$ and $\ell \in J_r$.

From Lemma 4.6.1(iii), we have that if $\ell, \ell' \in J_q \subseteq \boldsymbol{B}$ or $\ell, \ell' \in J_r \subseteq \boldsymbol{N}$, then $\kappa^{\delta}_{\ell\ell'}/4 \leq \kappa^{\delta^+}_{\ell\ell'}$. Hence, the strong connectivity of J_r and J_q in $G_{\delta,\gamma}$ implies the strong connectivity of these sets in $G_{\delta^+,\gamma/(4n)}$. Together with the edge (i', j'), we see that every $\ell' \in J_q$ can reach every $\ell \in J_r$ on a directed path of length $\leq |J_q \cup J_r| - 1$ in $G_{\delta^+,\gamma/(4n)}$. Applying Lemma 4.4.1 for this setting, we obtain $\Psi^{\mu}(\ell, \ell') \leq \varrho^{\mu^+}(\ell, \ell') \leq |J_q \cup J_r|$ for all such pairs, and also $\varrho^{\mu^+}(i, j) \geq -|J_q \cup J_r|$.

The rest of the proof is dedicated to showing the existence of i' and j' as in (4.52). We let $k \in [p]$ such that $\ell^{\delta^+}(J_{\geq k}) = \ell^{\delta^+}(\mathcal{J}) > 4n\gamma$. To simplify the notation, we let $I = J_{\geq k}$.

When the subroutine LAYERING$(\delta, \hat{\kappa})$ was constructing $\mathcal{J}$, the subroutine VERIFY-LIFT$(\mathrm{Diag}(\delta)W, I, \gamma)$ was called for the set $I = J_{\geq k}$, with the answer 'pass'. Besides $\ell^{\delta}(I) \leq \gamma$, this guaranteed the stronger property that $\max_{ji} |B_{ji}| \leq \gamma$ for the matrix B implementing the lift (see Remark 4.2.18).

Let us recall how this matrix B was obtained. The subroutine starts by finding a minimal $I' \subseteq I$ such that $\dim(\pi_{I'}(W)) = \dim(\pi_I(W))$. Recall that $\pi_{I'}(W) \cong \mathbb{R}^{I'}$ and $L^{\delta}_I(p) = L^{\delta}_{I'}(p_{I'})$ for every $p \in \pi_I(\mathrm{Diag}(\delta)W)$.

Consider the optimal lifting $L^{\delta}_I : \pi_I(\mathrm{Diag}(\delta)W) \to \mathrm{Diag}(\delta)W$. We defined $B \in \mathbb{R}^{([n]\setminus I)\times I'}$ as the matrix sending any $q \in \pi_{I'}(\mathrm{Diag}(\delta)W)$ to the corresponding vector $[L^{\delta}_{I'}(q)]_{[n]\setminus I}$. The column B_i can be computed as $[L^{\delta}_{I'}(e^i)]_{[n]\setminus I}$ for $e^i \in \mathbb{R}^{I'}$.

We consider the transformation

$$\bar{B} := \mathrm{Diag}(\delta^+\delta^{-1}) B \,\mathrm{Diag}\left((\delta^+_{I'})^{-1}\delta_{I'}\right).$$

This maps $\pi_{I'}(\mathrm{Diag}(\delta^+)W) \to \pi_{[n]\setminus I}(\mathrm{Diag}(\delta^+)W)$.

Let $z \in \pi_I(\mathrm{Diag}(\delta^+)W)$ be the singular vector corresponding to the maximum singular value of $L_I^{\delta^+}$, namely, $\|[L_I^{\delta^+}(z)]_{[n]\setminus I}\| > 4n\gamma\|z\|$. Let us normalize z such that $\|z_{I'}\| = 1$. Thus,

$$\|[L_{I'}^{\delta^+}(z_{I'})]_{[n]\setminus I}\| > 4n\gamma\,.$$

Let us now apply $\bar{B}$ to $z_{I'} \in \pi_{I'}(\mathrm{Diag}(\delta^+)W)$. Since $L_I^{\delta^+}$ is the minimum-norm lift operator, we see that

$$\|\bar{B}z_{I'}\| \geq \|[L_{I'}^{\delta^+}(z_{I'})]_{n\setminus I}\| > 4n\gamma\,.$$

We can upper bound the operator norm using the Frobenius norm $\|\bar{B}\| \leq \|\bar{B}\|_F = \sqrt{\sum_{ji} \bar{B}_{ji}{}^2} \leq n \max_{ji} |\bar{B}_{ji}|$, and therefore

$$\max_{ji} |\bar{B}_{ji}| > 4\gamma\,.$$

Let us fix $i' \in I'$ and $j' \in [n] \setminus I$ as the indices giving the maximum value of $\bar{B}$. Note that $\bar{B}_{j'i'} = B_{j'i'}\delta^+_{j'}\delta_{i'}/(\delta^+_{i'}\delta_{j'})$.

Let us now use Lemma 4.2.17 for the pair i', j', the matrix B and the subspace $\mathrm{Diag}(\delta)W$. Noting that $B_{j'i'} = [L_{I'}^{\delta}(e^{i'})]_{j'}$, we obtain $\kappa^{\delta}_{i'j'} \geq |B_{j'i'}|$. Now,

$$\kappa^{\delta^+}_{i'j'} = \kappa^{\delta}_{i'j'} \cdot \frac{\delta^+_{j'}\delta_{i'}}{\delta^+_{i'}\delta_{j'}} \geq |B_{j'i'}| \cdot \frac{\delta^+_{j'}\delta_{i'}}{\delta^+_{i'}\delta_{j'}} = |\bar{B}_{j'i'}| > 4\gamma\,. \tag{4.53}$$

The next claim finishes the proof.

Claim 4.6.4. *For i' and j' selected as above,* (4.52) *holds.*

Proof. $(i', j') \in E_{\delta^+,\gamma/(4n)}$ holds by (4.53). From the above, we have

$$|B_{j'i'}| > 4\gamma \cdot \frac{\delta^+_{i'}\delta_{j'}}{\delta_{i'}\delta^+_{j'}}\,.$$

According to Remark 4.2.18, $|B_{j'i'}| \leq \gamma$ follows since VERIFY-LIFT$(\mathrm{Diag}(\delta)W, I, \gamma)$ returned with 'pass'. We thus have

$$\frac{\delta^+_{i'}\delta_{j'}}{\delta_{i'}\delta^+_{j'}} < \frac{1}{4}.$$

Lemma 4.6.1 excludes the scenarios $i', j' \in N$, $i', j' \in B$, and $i' \in N$, $j' \in B$, leaving $i' \in B$ and $j' \in N$ as the only possibility. Therefore, $i' \in J_q \subseteq B$ and $j' \in J_r \subseteq N$. We have $r < q$ since $i \in I = J_{\geq k}$ and $j \in [n] \setminus I = J_{<k}$. □

4.7 Initialization

Our main algorithm (Algorithm 4 in Section 4.3.6), requires an initial solution $w^0 = (x^0, y^0, s^0) \in \mathcal{N}(\beta)$. In this section, we remove this assumption by adapting the initialization method of [198] to our setting.

We use the "big-M method", a standard initialization approach for path-following interior point methods that introduces an auxiliary system whose optimal solutions map back to the optimal solutions of the original system. The primal-dual system we consider is

$$\begin{array}{rlrl} \min & c^\mathsf{T}x + M\vec{1}^\mathsf{T}\underline{x} & \max & y^\mathsf{T}b + 2M\vec{1}^\mathsf{T}z \\ & Ax - A\underline{x} = b & & A^\mathsf{T}y + z + s = c \\ & x + \bar{x} = 2Me & & z + \bar{s} = \vec{0} \\ & x, \bar{x}, \underline{x} \geq \vec{0} & & -A^\mathsf{T}y + \underline{s} = Me \\ & & & s, \bar{s}, \underline{s} \geq \vec{0}. \end{array} \tag{Init-LP}$$

The constraint matrix used in this system is

$$\hat{A} = \begin{pmatrix} A & -A & 0 \\ I & 0 & I \end{pmatrix}$$

The next lemma asserts that the $\bar{\chi}$ condition number of $\hat{A}$ is not much bigger than that of A of the original system (4.1).

Lemma 4.7.1 ([198, Lemma 23])**.** $\bar{\chi}_{\hat{A}} \leq 3\sqrt{2}(\bar{\chi}_A + 1)$.

We extend this bound for $\bar{\chi}^*$.

Lemma 4.7.2. $\bar{\chi}^*_{\hat{A}} \leq 3\sqrt{2}(\bar{\chi}^*_A + 1)$.

Proof. Let $D \in \mathbf{D}_n$ and let $\hat{D} \in \mathbf{D}_{3n}$ the matrix consisting of three copies of D, i.e.

$$\hat{D} = \begin{pmatrix} D & 0 & 0 \\ 0 & D & 0 \\ 0 & 0 & D \end{pmatrix}.$$

Then

$$\hat{A}\hat{D} = \begin{pmatrix} AD & -AD & 0 \\ D & 0 & D \end{pmatrix}$$

Row-scaling does not change $\bar{\chi}$ as the kernel of the matrix remains unchanged. Thus, we can rescale the last n rows of $\hat{A}\hat{D}$, to the identity matrix, i.e. multiplying by (I, D^{-1}) from the left hand side. We observe that

$$\bar{\chi}_{\hat{A}\hat{D}} = \bar{\chi}\left(\begin{pmatrix} AD & -AD & 0 \\ I & 0 & I \end{pmatrix}\right) \leq 3\sqrt{2}(\bar{\chi}_{AD} + 1)$$

where the inequality follows from Lemma 4.7.1. The lemma now readily follows as

$$\bar{\chi}^*_{\hat{A}} = \inf\{\bar{\chi}_{\hat{A}\hat{D}} : D \in \mathbf{D}_{3n}\} \leq \inf\{3\sqrt{2}(\bar{\chi}_{AD} + 1) : D \in \mathbf{D}_n\} = 3\sqrt{2}(\bar{\chi}^*_A + 1). \quad \square$$

We show next that the optimal solutions of the original system are preserved for sufficiently large M. We let d be the min-norm solution to $Ax = b$, i.e., $d = A^\mathsf{T}(AA^\mathsf{T})^{-1}b$.

Proposition 4.7.3. *Assume both primal and dual of* (4.1) *are feasible, and* $M > \max\{(\bar{\chi}_A + 1)\|c\|, \bar{\chi}_A\|d\|\}$. *Every optimal solution* (x, y, s) *to* (4.1)*, can be extended to an optimal solution* $(x, \underline{x}, \bar{x}, y, s, \underline{s}, \bar{s})$ *to* (Init-LP); *and conversely, from every optimal solution* $(x, \underline{x}, \bar{x}, y, z, s, \underline{s}, \bar{s})$ *to* (Init-LP)*, we obtain an optimal solution* (x, y, s) *by deleting the auxiliary variables.*

Proof. If system (4.1) is feasible, it admits a basic optimal solution (x^*, y^*, s^*) with basis B such that $A_B x^*_B = b, x^* \geq \vec{0}, A_B^\mathsf{T} y^* = c$ and $A^\mathsf{T} y^* \leq c$. Using Proposition 4.2.1(ii) we see that

$$\|x^*_B\| = \|A_B^{-1}b\| = \|A_B^{-1}Ad\| \leq \bar{\chi}_A\|d\| < M\,, \tag{4.54}$$

and using that $\|A\| = \|A^\mathsf{T}\|$ we observe

$$\|A^\mathsf{T}y^*\| = \|A^\mathsf{T}A_B^{-\mathsf{T}}c\| \leq \|A^\mathsf{T}A_B^{-\mathsf{T}}\|\|c\| = \|A_B^{-1}A\|\|c\| \leq \bar{\chi}_A\|c\| < M. \tag{4.55}$$

We can extend this solution to a solution of system (Init-LP) via setting $\bar{x}^* = 2Me - x^*, \underline{x}^* = \vec{0}, z^* = \bar{s}^* = \vec{0}$ and $\underline{s}^* = Me + A^\mathsf{T}y^*$. Observe that $\bar{x}^* > \vec{0}$ and $\underline{s}^* > \vec{0}$ by (4.54) and (4.55). Furthermore observe that by complementary slackness this extended solution for (Init-LP) is an optimal solution. The property that $\underline{s}^* > \vec{0}$ immediately tells us that $\underline{x}$ vanishes for all optimal solutions of (Init-LP) and thus all optimal solutions of (4.1) coincide with the optimal solutions of (Init-LP), with the auxiliary variables removed. $\square$

The next lemma is from [146, Lemma 4.4]. Recall that $w = (x, y, s) \in \mathcal{N}(\beta)$ if $\|xs/\mu(w) - \vec{1}\| \leq \beta$.

Lemma 4.7.4. *Let* $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$*, and let* $\nu > 0$*. Assume that* $\|xs/\nu - \vec{1}\| \leq \tau$*. Then* $(1 - \tau/\sqrt{n})\nu \leq \mu(w) \leq (1 + \tau/\sqrt{n})\nu$ *and* $w \in \mathcal{N}(\tau/(1 - \tau))$.

The new system has the advantage that we can easily initialize the system with a feasible solution in close proximity to central path:

Proposition 4.7.5. *We can initialize system* (Init-LP) *close to the central path with initial solution* $w^0 = (x^0, y^0, s^0) \in \mathcal{N}(1/8)$ *and parameter* $\mu(w^0) \approx M^2$ *if* $M > 15 \max\{(\bar{\chi}_A + 1)\|c\|, \bar{\chi}_A\|d\|\}$.

Proof. The initialization follows along the lines of [198, Section 10]. We let d as above, and set

$$\begin{aligned} \bar{x}^0 &= Me, x^0 = Me, \underline{x}^0 = Me - d \\ y^0 &= \vec{0}, z^0 = -Me \\ \bar{s}^0 &= Me, s^0 = Me + c, \underline{s}^0 = Me. \end{aligned}$$

This is a feasible primal-dual solution to system (Init-LP) with parameter

$$\mu^0 = (3n)^{-1}(x^{0\mathsf{T}} s^0 + \underline{x}^{0\mathsf{T}} \underline{s}^0 + \bar{x}^{0\mathsf{T}} \bar{s}^0) = (3n)^{-1}(3nM^2 + Mc^{\mathsf{T}}\vec{1} - Md^{\mathsf{T}}\vec{1}) \approx M^2 .$$

We see that

$$\left\| \frac{1}{M^2} \begin{pmatrix} \bar{x}^0 \bar{s}^0 \\ x^0 s^0 \\ \underline{x}^0 \underline{s}^0 \end{pmatrix} - e \right\|^2 = M^{-2}\|c\|^2 + M^{-2}\|d\|^2 \leq \frac{1}{9^2 \bar{\chi}_A^2} \leq \frac{1}{9^2}.$$

With Lemma 4.7.4 we conclude that $w^0 = (x^0, y^0, s^0) \in \mathcal{N}\left(\frac{1/9}{1-1/9}\right) = \mathcal{N}(1/8)$. □

Detecting infeasibility To use the extended system (Init-LP), we still need to assume that both the primal and dual programs in (4.1) are feasible. For arbitrary instances, we first need to check if this is the case, or conclude that the primal or the dual (or both) are infeasible.

This can be done by employing a two-phase method. The first phase decides feasibility by running (Init-LP) with data $(A, b, \vec{0})$ and $M > \bar{\chi}_A\|d\|$. The objective value of the optimal primal-dual pair is 0 if and only if (4.1) has a feasible solution. If the optimal primal/dual solution $(x^*, \underline{x}^*, \bar{x}^*, y^*, s^*, \underline{s}^*, \bar{s}^*)$ has positive objective value, we can extract an infeasibility certificate in the following way.

By the characterization of $\bar{\chi}_A$ as in Proposition 4.2.1(ii), there exists an optimal solution x' with $\bar{x}' > \vec{0}$, and so by strong duality, $\bar{s}^* = \vec{0}$. From the dual, we conclude that $z = \vec{0}$, and therefore $A^{\mathsf{T}} y^* \leq A^{\mathsf{T}} y^* + s^* + z = c = \vec{0}$. On the other hand, by assumption the objective value of the dual is positive, and so $y^{\mathsf{T}} b + 2M\vec{1}^{\mathsf{T}} z > 0$.

Feasibility of the dual of (4.1) can be decided by running (Init-LP) on data $(A, \vec{0}, c)$ and $M > (\bar{\chi}_A + 1)\|c\|$ with the same argumentation: Either the objective value of

the dual is 0 and therefore the dual optimal solution $(y^*, \underline{s}^*, s^*, \bar{s}^*)$ corresponds to a feasible dual solution of (4.1) or the objective value is negative and we extract a dual infeasibility certificate in the following way: By assumption $c^\mathsf{T} x \leq c^\mathsf{T} x + M\vec{1}^\mathsf{T}\underline{x} < 0$. Furthermore, there exists a basic optimal solution to the dual of (Init-LP) with $\underline{s} > \vec{0}$ and therefore $\underline{x}^* = \vec{0}$ for the optimal primal solution $(\underline{x}^*, x^*, \bar{x}^*)$. So, we have $Ax^* = b = \vec{0}$, together with $c^\mathsf{T} x < 0$ yielding the certificate.

Finding the right value of M While Algorithm 4 does not require any estimate on $\bar{\chi}^*$ or $\bar{\chi}$, the initialization needs to set $M \geq \max\{(\bar{\chi}_A + 1)\|c\|, \bar{\chi}_A\|d\|\}$ as in Proposition 4.7.3.

A straightforward guessing approach (attributed to J. Renegar in [198]) starts with a constant guess, say $\bar{\chi}_A = 100$, constructs the extended system, and runs the algorithm. In case the optimal solution to the extended system does not map to an optimal solution of (4.1), we restart with $\bar{\chi}_A = 100^2$ and try again; we continue squaring the guess until an optimal solution is found.

This would still require a series of $\log\log \bar{\chi}_A$ guesses, and thus, result in a dependence on $\bar{\chi}_A$ in the running time. However, if we initially rescale our system using the near-optimal rescaling Theorem 4.2.5, the we can turn the dependence from $\bar{\chi}_A$ to $\bar{\chi}_A^*$. The overall iteration complexity remains $O(n^{2.5} \log n \log(\bar{\chi}_A^* + n))$, since the running time for the final guess on $\bar{\chi}_A^*$ dominates the total running time of all previous computations due to the repeated squaring.

An alternative approach, that does not rescale the system, is to use Theorem 4.2.5 to approximate $\bar{\chi}_A$. In this case we repeatedly square a guess of $\bar{\chi}_A^*$ instead of $\bar{\chi}_A$ which takes $\mathcal{O}(\log\log \bar{\chi}_A^*)$ iterations until our guess corresponds to a valid upper bound for $\bar{\chi}_A$.

Note that either guessing technique can handle bad guesses gracefully. For the first phase, if neither a feasible solution to (4.1) is returned nor a Farkas' certificate can be extracted, we have proof that the guess was too low by the above paragraph. Similarly, in phase two, when feasibility was decided in the affirmative for primal and dual, an optimal solution to (Init-LP) that corresponds to an infeasible solution to (4.1) serves as a certificate that another squaring of the guess is necessary.

Chapter 5

A Simple Method for Convex Optimization in the Oracle Model

We give a simple and natural method for computing approximately optimal solutions for minimizing a convex function f over a convex set K given by a separation oracle. Our method utilizes the Frank–Wolfe algorithm over the cone of valid inequalities of K and subgradients of f. Under the assumption that f is L-Lipschitz and that K contains a ball of radius r and is contained inside the origin centered ball of radius R, using $O(\frac{(RL)^2}{\varepsilon^2} \cdot \frac{R^2}{r^2})$ iterations and calls to the oracle, our main method outputs a point $x \in K$ satisfying $f(x) \leq \varepsilon + \min_{z \in K} f(z)$.

Our algorithm is easy to implement, and we believe it can serve as a useful alternative to existing cutting plane methods. As evidence towards this, we show that it compares favorably in terms of iteration counts to the standard LP based cutting plane method and the analytic center cutting plane method, on a testbed of combinatorial, semidefinite and machine learning instances.

5.1 Introduction

We consider the problem of minimizing a convex function $f : \mathbb{R}^n \to \mathbb{R}$ over a compact convex set $K \subseteq \mathbb{R}^n$. We assume that K contains an (unknown) Euclidean ball of radius $r > 0$ and is contained inside the origin centered ball of radius $R > 0$, and that f is L-Lipschitz. We have first-order access to f that yields $f(x)$ and a subgradient of f at x for any given x. Moreover, we only have access to K through a separation oracle (SO), which, given a point $x \in \mathbb{R}^n$, either asserts that $x \in K$ or returns a linear constraint valid for K but violated by x.

Convex optimization in the SO model is one of the fundamental settings in optimization. The model is relevant for a wide variety of implicit optimization problems, where an explicit description of the defining inequalities for K is either too large to store or not fully known. The SO model was first introduced in [152]

This chapter is based on [48], a joint work with Daniel Dadush, Christopher Hojny, and Stefan Weltge.

where it was shown that an additive ε-approximate solution can be obtained using $O(n \log(LR/(\varepsilon r)))$ queries via the center of gravity method and $O(n^2 \log(LR/(\varepsilon r)))$ queries via the ellipsoid method. This latter result was used by Khachiyan [122] to give the first polynomial time method for linear programming. The study of oracle-type models was greatly extended in the classic book of Grötschel, Lovász, and Schrijver [103], where many applications to combinatorial optimization were provided. Further progress on the SO model was given by Vaidya [191], who showed that the $O(n \log(LR/(\varepsilon r)))$ oracle complexity can be efficiently achieved using the so-called volumetric barrier as a potential function, where the best current running time for such methods was given very recently [114, 133].

From the practical perspective, two of the most popular methods in the SO model are the standard linear programming (LP) based cutting plane method, independently discovered by Kelley [120], Goldstein-Cheney [41] as well as Gomory [99] (in the integer programming context), and the analytic center cutting plane method [175] (ACCPM).

The LP based cutting plane method, which we henceforth dub the standard cut loop, proceeds as follows: starting with finitely many linear underestimators of f and linear constraints valid for K, in each iteration it solves a linear program that minimizes the lower envelope of f subject to the current linear relaxation of K. The resulting point x is then used to query f and the SO to obtain a new underestimator for f and a new constraint valid for K. Note that if f is a linear function, it repeatedly minimizes f over linear relaxations of K. While it is typically fast in practice, it can be unstable, and no general quantitative convergence guarantees are known for the standard cut loop.

To link to integer programming, in that context K is the convex hull of integer points of some polytope P and the objective is often linear, and the method is initialized with a linear description of P. A crucial difference there is that the separator SO is generally only efficient when queried at vertices of the current relaxation.

ACCPM is a barrier based method, in which the next query point is the minimizer of the barrier for the current inequalities in the system. ACCPM is in general a more stable method with provable complexity guarantees. Interestingly, while variants of ACCPM achieving $O(n \log(1/\varepsilon)^2)$ queries exist, achieved by judiciously dropping constraints [8], the more practical variants achieve only converge in $O(n/\varepsilon^2)$ queries [153].

In this chapter, we describe a new method for convex optimization in the SO model that computes an additive ε-approximate solution within $O(R^4L^2/r^2\varepsilon^2)$ iterations. Our algorithm is easy to implement, and we believe it can serve as a useful alternative to existing methods. In our experimental results, we show that it compares favorably in terms of iteration counts to the standard cut loop and the analytic center cutting plane

method, on a testbed of combinatorial, semidefinite and machine learning instances.

Before explaining our approach, we review the relevant work in related models. To begin, there has been a tremendous amount of work in the context of first-order methods [16,18], where the goal is to minimize a possibly complicated function, given by a gradient oracle, over a *simple domain* K (e.g., the simplex, cube, ℓ_2 ball). These methods tend to have cheap iterations and to achieve $\text{poly}(1/\varepsilon)$ convergence rates. They are often superior in practice when the requisite accuracy is low or moderate, e.g., within 1% of optimal. For these methods, often variants of (sub-)gradient descent, it is generally assumed that computing (Euclidean) projections onto K as well as linear optimization over K are easy. If one only assumes access to a linear optimization (LO) oracle on K, K can become more interesting (e.g., the shortest-path or spanning-tree polytope). In this context, one of the most popular methods is the so-called *Frank–Wolfe* algorithm [84] (see [111] for a modern treatment), which iteratively computes a convex combination of vertices of K to obtain an approximate minimizer of a smooth convex function.

In the context of combinatorial optimization, there has been a considerable line of work on solving (implicit) packing and covering problems using the so-called multiplicative weights update (MWU) framework [87, 158, 171]. In this framework, one must be able to implement an MWU oracle, which in essence computes optimal solutions for the target problem after the "difficult" constraints have been aggregated according to the current weights. This framework has been applied for getting fast $(1 \pm \varepsilon)$-approximate solutions to multi-commodity flow [87, 171], packing spanning trees [40], the Held–Karp approximation for TSP [39], and more, where the MWU oracle computes shortest paths, minimum cost spanning trees, minimum cuts respectively in a sequence of weighted graphs. The MWU oracle is in general just a special type of LO oracle, which can often be interpreted as a SO that returns a maximally violated constraint. While certainly related to the SO model, it is not entirely clear how to adapt MWU to work with a general SO, in particular in settings unrelated to packing and covering.

A final line of work, which directly inspires our work, has examined simple iterative methods for computing a point in the interior of a cone Σ that directly apply in the SO model. The application of simple iterative methods for solving conic feasibility problems can be traced to Von Neumann in 1948 (see [61]), and a variant of this method, the perceptron algorithm [164] is still very popular today. Von Neumann's algorithm computes a convex combination of the defining inequalities of the cone, scaled to be of unit length, of nearly minimal Euclidean norm. The separation oracle is called to find an inequality violated by the current convex combination, and this inequality is then used to make the current convex combination shorter, in an analogous way to Frank–Wolfe. This method is guaranteed to find a point in the cone

in $O(1/\varrho^2)$ iterations, where ϱ is the so-called width of Σ (the radius of the largest ball contained in Σ centered at a point of norm 1). Starting in 2004, polynomial time variants of this and related methods (i.e., achieving $\log 1/\varrho$ dependence) have been found [19, 42, 73], which iteratively "rescale" the norm to speed up the convergence. These rescaled variants can also be applied in the oracle setting [17, 44, 54] with appropriate adaptations. The main shortcoming of existing conic approaches is that they are currently not well-adapted for solving optimization problems rather than feasibility problems.

Our approach. In this work, we build upon von Neumann's approach and utilize the Frank–Wolfe algorithm over the cone of valid inequalities of K as well as the subgradients of f in a way that yields a clean, simple, and flexible framework for solving general convex optimization problems in the SO model. For simpler explanation, let us assume that $f(x) = \langle c, x\rangle$ is a linear function and that we know an upper bound UB on the minimum of f over K. Given some linear inequalities $\langle a_i, x\rangle \le b_i$ that are valid for all $x \in K$, our goal is to find convex combinations p of the *homogenized* points (c, UB) and (a_i, b_i) that are "close" to the origin. Note that if $p = \vec{0}$, the fact that K is full-dimensional implies that (c, UB) appears with a nonzero coefficient and hence $(-c, -\mathrm{UB})$ is a nonnegative combination of the points (a_i, b_i), which in turn shows that UB is equal to the minimum of f over K. In view of this, we will consider a potential $\Phi\colon \mathbb{R}^{n+1} \to \mathbb{R}_+$ with the property that if $\Phi(p)$ is sufficiently small, then the convex combination will yield an explicit certificate that UB is close to the minimum of f over K.

Given a certain convex combination p, note that the gradient of Φ at p provides information about whether moving towards one of the known points will (significantly) decrease $\Phi(p)$. However, if no such known point exists, it turns out that the "dehomogenization" of the gradient (a scaling of its projection onto the first n coordinates) is a natural point $x \in \mathbb{R}^n$ to query the SO with. In fact, if $x \in K$, it will have improved objective value with respect to f. Otherwise, the SO will provide a linear inequality such that moving towards its homogenization decreases $\Phi(p)$.

In this work, we will show that the above paradigm immediately yields a rigorous algorithm for various natural choices of Φ and scalings of inequalities. We will also see that general convex functions can be directly handled in the same manner by simply replacing (c, UB) with all subgradient cuts of f learned throughout the iterations. The same applies to pure feasibility problems for which we set $f = \vec{0}$. The convergence analysis of our algorithm is simple and based on standard estimates for the Frank–Wolfe algorithm.

Besides its conceptual simplicity and distiction to existing methods for convex optimization in the SO model, we also regard it as a practical alternative. In fact,

in terms of iterations, our vanilla implementation in `Julia` [49] performs similarly and often even better than the standard cut loop and the analytic center cutting plane method evaluated on a testbed of oracle-based linear optimization problems for matching problems, semidefinite relaxations of the maximum cut problem, and LPBoost. Moreover, the flexibility of our framework leaves several degrees of freedom to obtain optimized implementations that outperform our naive implementation.

5.2 Algorithm

Recall that we are given first-order access to a convex function $f\colon \mathbb{R}^n \to \mathbb{R}$ that we want to minimize over a convex body $K \subseteq \mathbb{R}^n$. In the case where f is not differentiable, with a slight abuse of notation we interpret $\nabla f(x)$ to be *any* subgradient of f at x. We can access K by a separation oracle that, given a point $x \in \mathbb{R}^n$, either asserts that $x \in K$ or returns a point $(a, b) \in \mathcal{A} \subseteq \mathbb{R}^{n+1}$ with $\langle a, x\rangle > b$ such that $\langle a, y\rangle \leq b$ holds for all $y \in K$. Here, $\langle \cdot, \cdot \rangle$ denotes the standard scalar product and we assume that all points in $\mathcal{A}$ correspond to linear constraints valid for K. To state our algorithm, let $\|\cdot\|$ denote any norm on $\mathbb{R}^{n+1}$ and $\|\cdot\|_*$ its dual norm. Moreover, let $\Phi\colon \mathbb{R}^{n+1} \to \mathbb{R}_+$ be any strictly convex and differentiable function with $\min_{x\in\mathbb{R}^{n+1}} \Phi(x) = \Phi(\vec{0}) = 0$. Our method is given in Algorithm 5, in which we denote the number of iterations by T for later reference. However, T does not need to be specified in advance, and the algorithm may be stopped at any time, e.g., when a solution or bound of desired accuracy has been found.

In line 5, $\nabla\Phi(p_t)[1 : n]$ denotes the first n components of $\nabla\Phi(p_t)$, and $\nabla\Phi(p_t)[n + 1]$ denotes the last component of $\nabla\Phi(p_t)$. The sets A_t and G_t denote the already known/separated inequalities and objective gradients during iteration t.

Lemma 5.2.1. *When $x_t \in \mathbb{R}^n$ is computed in iteration t of Algorithm 5, it is well-defined and we have $\langle c, x_t\rangle \leq d$ for every $(c, d) \in A_t \cup G_t$.*

Proof. Since p_t minimizes Φ over $\operatorname{conv}(A_t \cup G_t)$, for every $q \in \operatorname{conv}(A_t \cup G_t)$ we have $\langle \nabla\Phi(p_t), q - p_t\rangle \geq 0$. If $p_t \neq \vec{0}$ then from strict convexity of Φ and $\min_{x\in\mathbb{R}^{n+1}} \Phi(x) = \Phi(\vec{0}) = 0$ we get

$$\langle \nabla\Phi(p_t), q\rangle \geq \langle \nabla\Phi(p_t), p_t\rangle > 0. \tag{5.1}$$

First, apply this inequality to $q = (\vec{0}, 1)/\|(\vec{0}, 1)\|_* \in A_t$ and conclude $\nabla\Phi(p_t)[n + 1] > 0$. This makes sure that x_t can be computed. Second, we apply (5.1) to $q = (c, d) \in A_t \cup G_t$ and find that $d - \langle c, x_t\rangle = \frac{1}{\nabla\Phi(p_t)[n+1]}\langle \nabla\Phi(p_t), (c, d)\rangle > 0$, thus x_t satisfies $\langle c, x_t\rangle \leq d$ for all $(c, d) \in A_t \cup G_t$. □

Algorithm 5

1: UB $\leftarrow \infty$, $A_1 \leftarrow \{(\vec{0}, 1)/\|(\vec{0}, 1)\|_*\}$, $G_1 \leftarrow \emptyset$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: $\quad p_t \leftarrow \arg\min\{\Phi(p) : p \in \mathrm{conv}(A_t \cup G_t)\}$
4: $\quad$ **if** $p_t = \vec{0}$ **then return** UB.
5: $\quad x_t \leftarrow -\nabla\Phi(p_t)[1:n]/\nabla\Phi(p_t)[n+1]$
6: $\quad$ **if** $x_t \in K$ **then**
7: $\quad\quad$ UB $\leftarrow \min\{\mathrm{UB}, f(x_t)\}$
8: $\quad\quad A_{t+1} \leftarrow A_t$.
9: $\quad\quad G_{t+1} \leftarrow G_t \cup \{(\nabla f(x_t), \langle \nabla f(x_t), x_t\rangle)\}$
10: $\quad$ **else**
11: $\quad\quad$ get $(a, b) \in \mathcal{A}$, with $\langle a, x_t\rangle > b$ and $\|(a, b)\|_* = 1$
12: $\quad\quad A_{t+1} \leftarrow A_t \cup \{(a, b)\}$.
13: $\quad\quad G_{t+1} \leftarrow G_t$.
14: **return** UB.

Note that, for the sake of presentation, in line 3 we require p_t to be the convex combination of minimum Φ-value. However, it is usually not necessary to compute such a minimum. The same convergence rates can be obtained if, in every iteration, p_t is a suitable convex combination of p_{t-1} and some $(c, d) \in A_t \cup G_t$ with $\langle \nabla\Phi(p_{t-1}), (c, d)\rangle < 0$. If the last coordinate of p_{t-1}, as discussed in the above proof, is not positive, then such an update can be made towards $(\vec{0}, 1)/\|(\vec{0}, 1)\|_* \in A_t$. Any such update will significantly decrease $\Phi(p_t)$, and the computation in line 3 is guaranteed to make at least that much progress. This shows that simple updates of p_t, which may be more preferable in practice, still suffice to achieve the claimed convergence rates.

Definition 5.2.2. A continuously differentiable function $g : \mathbb{R}^k \to \mathbb{R}$ is called β-smooth with respect to a norm $\|\cdot\|$ if the gradient ∇g is β-Lipschitz, that is

$$\|\nabla g(x) - \nabla g(y)\| \leq \beta\|x - y\|$$

holds for all $x, y \in \mathbb{R}^k$.

Lemma 5.2.3. *Suppose that Φ is 1-smooth with respect to $\|\cdot\|_*$ and that*

$$\|(\nabla f(x), \langle \nabla f(x), x\rangle)\|_* \leq 1$$

for every $x \in K$. Then for every $t = 1, \ldots, T$, Algorithm 5 satisfies $\Phi(p_t) \leq \frac{8}{t+1}$.

Proof. Let $\gamma_t = \frac{2}{t+1}$. If we added a constraint v in iteration t, i.e., if $A_{t+1} \cup G_{t+1} = \{v\} \cup A_t \cup G_t$, then we are guaranteed that $(1-\gamma_t)p_t + \gamma_t v \in \operatorname{conv}(A_{t+1} \cup G_{t+1})$ and that $\langle \nabla\Phi(p_t), v\rangle < 0$. Therefore we get, by 1-smoothness, by construction, and by convexity, that

$$\begin{aligned}\Phi(p_{t+1}) - \Phi(p_t) &\leq \langle \nabla\Phi(p_t), (p_{t+1} - p_t)\rangle + \frac{1}{2}\|p_{t+1} - p_t\|^2 \\ &\leq \gamma_t \langle \nabla\Phi(p_t), (v - p_t)\rangle + 2\gamma_t^2 \\ &\leq \gamma_t \langle \nabla\Phi(p_t), (\vec{0} - p_t)\rangle + 2\gamma_t^2 \\ &\leq -\gamma_t \Phi(p_t) + 2\gamma_t^2\end{aligned}$$

From this, we can derive that $\Phi(p_{t+1}) \leq (1-\gamma_t)\Phi(p_t) + 2\gamma_t^2$. Furthermore, by 1-smoothness we know that $\Phi(p_1) \leq 1/2$. By induction it follows that $\Phi(p_t) \leq \frac{8}{t+1}$. □

The proof of the above lemma is in line with standard proofs for the analysis of Frank–Wolfe algorithms, see, e.g., Theorem 1 in [111].

The following lemma yields conditions under which a small value of $\Phi(p_t)$ implies that UB is close to the minimum of f over K. Note in particular that it proves that if $\|p_t\| = 0$ then UB = OPT, where OPT := $\min_{x\in K} f(x)$ is the optimal value of the optimization problem.

Lemma 5.2.4. *Assume that* $\|(x,-1)\| \leq 2$ *holds for every* $x \in K$*, and there exist* $z \in K$ *and* $\alpha \in (0,1]$ *such that* $\langle (a,b), (-z,1)\rangle \geq \alpha \|(-z,1)\| \|(a,b)\|_*$ *holds for every* $(a,b) \in \mathcal{A} \cup \{(\vec{0},1)\}$*. Moreover, assume that* $\|(\nabla f(x), \langle \nabla f(x), x\rangle)\|_* \leq 1$ *holds for every* $x \in K$*. If* $\|p_T\| \leq \alpha/2$ *in Algorithm 5, then the returned value satisfies* $\mathrm{UB} \geq \mathrm{OPT} \geq \mathrm{UB} - \frac{4\|p_T\|_*(1+\alpha)}{\alpha}$.

Proof. Let $x^* \in K$ minimize $f(x)$ over $x \in K$ and let $F \subseteq [T-1]$ be the set of iterations t (except the last one) in which $x_t \in K$. Now write the point p_T as a convex combination

$$p_T = \sum_{(a,b)\in A_T} \lambda_{(a,b)}(a,b) + \sum_{t\in F} \gamma_t (\nabla f(x_t), \langle \nabla f(x_t), x_t\rangle)$$

where $\lambda \geq \vec{0}, \gamma \geq \vec{0}$ and $\|(\lambda, \gamma)\|_1 = 1$. Then we have

$$\begin{aligned}
&\sum_{t\in F} \gamma_t (f(x_t) - f(x^*)) \leq \sum_{t\in F} \gamma_t \langle \nabla f(x_t), x_t - x^* \rangle \\
&\quad = \Big\langle \sum_{t\in F} \gamma_t (\nabla f(x_t), \langle \nabla f(x_t), x_t \rangle),\ (-x^*, 1) \Big\rangle \\
&\quad \leq \Big\langle \sum_{t\in F} \gamma_t (\nabla f(x_t), \langle \nabla f(x_t), x_t \rangle) + \sum_{(a,b)\in A_T} \lambda_{(a,b)} (a, b),\ (-x^*, 1) \Big\rangle \\
&\quad = \langle p_T, (-x^*, 1) \rangle \\
&\quad \leq \|p_T\|_* \cdot \|(-x^*, 1)\| \leq 2\|p_T\|_*.
\end{aligned}$$

Here, the inequalities respectively arise from convexity of f, that $x^* \in K$ satisfies $\langle (a, b), (-x^*, 1) \rangle \geq 0$ for every $(a, b) \in A_T$, and the Cauchy–Schwarz inequality. In particular, we find that $\min_{t\in F} f(x_t) - f(x^*) \leq \frac{2\|p_T\|_*}{\sum_{t\in F} \gamma_t}$ whenever $\sum_{t\in F} \gamma_t > 0$. To lower bound the denominator, we use the assumptions on z to derive the inequalities

$$\begin{aligned}
&\alpha \left(1 - \sum_{t\in F} \gamma_t\right) \|(-z, 1)\| = \alpha \|(-z, 1)\| \sum_{(a,b)\in A_T} \lambda_{(a,b)} \\
&\quad \leq \langle \sum_{(a,b)\in A_T} \lambda_{(a,b)} (a, b), (-z, 1) \rangle \quad (\text{ since } \|(a, b)\|_* = 1 \,) \\
&\quad = \langle p_T, (-z, 1) \rangle - \sum_{t\in F} \gamma_t \langle (\nabla f(x_t), \langle \nabla f(x_t), x_t \rangle),\ (-z, 1) \rangle \\
&\quad \leq \|p_T\|_* \cdot \|(-z, 1)\| + \sum_{t\in F} \gamma_t \|(\nabla f(x_t), \langle \nabla f(x_t), x_t \rangle)\|_* \cdot \|(-z, 1)\|.
\end{aligned}$$

Now observe that $\|(\nabla f(x_t), \langle \nabla f(x_t), x_t \rangle)\|_* \leq 1$ for every $t \in F$ and divide through by $\|(-z, 1)\|$ to find $\alpha(1 - \sum_{t\in F} \gamma_t) \leq \|p_T\|_* + \sum_{t\in F} \gamma_t$. Hence, if $\|p_T\|_* \leq \frac{\alpha}{2}$ then $\alpha/2 \leq (\alpha + 1) \sum_{t\in F} \gamma_t$. This lower bound on $\sum_{t\in F} \gamma_t$ suffices to prove the lemma. □

Combining the previous two lemmas, we obtain the following convergence rate of our algorithm:

Theorem 5.2.5. *Assume that $\beta > 0$ is such that $\Phi(x) \geq \beta\|x\|_*^2$ for all $x \in \mathbb{R}^{n+1}$. Under the assumptions of Lemmas* 5.2.3 *and* 5.2.4*, Algorithm* 5 *computes, for every $T \geq \frac{32}{\beta\alpha^2}$, a value* $\mathrm{UB} < \infty$ *satisfying*

$$\mathrm{UB} \geq \min_{x\in K} f(x) \geq \mathrm{UB} - \frac{16}{\sqrt{\beta(T+2)}} \cdot \frac{1+\alpha}{\alpha}.$$

Proof. After T iterations, we have $\beta\|p_T\|_*^2 \le \Phi(p_T) \le \frac{8}{T+2} \le \beta\alpha^2/4$ from using Lemma 5.2.3. Since then $\|p_T\|_* \le \frac{\sqrt{8}}{\sqrt{\beta(T+2)}} \le \alpha/2$, Lemma 5.2.4 tells us that $\mathrm{OPT} \ge \mathrm{UB} - \frac{16(1+\alpha)}{\sqrt{\beta(T+2)}\,\alpha}$. □

Let us now apply the previous findings to a concrete setting, in which we assume that the objective function f is L-Lipschitz, i.e., $|f(x) - f(y)| \le L\|x - y\|_2$ for all $x, y \in \mathbb{R}^n$.

Theorem 5.2.6. *Let $K \subseteq \mathbb{R}^n$ be a convex body satisfying $z + r\mathbb{B}_2^n \subseteq K \subseteq R\mathbb{B}_2^n$, given by a separation oracle $\mathcal{A}$, and let $f\colon \mathbb{R}^n \to \mathbb{R}$ be an L-Lipschitz convex function given by a subgradient oracle.*

Apply Algorithm 5 to the function $\frac{1}{RL}f$ using norm $\|(x,y)\| := \sqrt{2}\|(x/R, y)\|_2$ and potential $\Phi(a,b) := \frac{1}{4}\|(Ra,b)\|_2^2$. Then, for every $\varepsilon > 0$, after

$$T = O\left(\frac{R^2}{r^2} \cdot \frac{R^2L^2}{\varepsilon^2}\right)$$

iterations we have $\mathrm{UB} \ge \min_{x\in K} f(x) \ge \mathrm{UB} - \varepsilon$.

Proof. By replacing $f(x)$ by $f(Rx)/(RL)$, K by K/R, ε by $\varepsilon/(RL)$, r by r/R, we may assume that $R = L = 1$, that $r \in (0, 1]$. After this rescaling, note $\|(x,y)\| := \sqrt{2}\|(x,y)\|_2$ and $\Phi(a,b) := \frac{1}{4}\|(a,b)\|_2^2 = \frac{1}{2}\|(a,b)\|_*^2$. Crucially, note that Algorithm 5 is invariant under the above replacement.

We now claim that our choice of input satisfies the conditions of Theorem 5.2.5 with $\beta = 1/2$ and $\alpha = r/4$. Given the claim, Theorem 5.2.5 directly proves the result. To prove the claim, apart from verifying that the bounds on β and α hold, we must verify smoothness of Φ with respect to the dual norm, a bound of 2 on the norm of $(-x, 1)$ for $x \in K$, as well as a dual norm bound of 1 on $(\nabla f(x), \langle \nabla f(x), x\rangle)$ for $x \in K$.

The setting $\beta = 1/2$ is direct by definition of Φ. Since $\|\cdot\|_*$ is a Euclidean norm, it is immediate that Φ is 1-smooth with respect to $\|\cdot\|_*$. For each $x \in K$, using that $R = L = 1$, we may also verify that

$$\|(x,1)\| = \sqrt{2}\|(x,1)\|_2 = \sqrt{2}\sqrt{\|x\|_2^2 + 1} \le \sqrt{2}\sqrt{R^2+1} = 2,$$

and

$$\begin{aligned}\|(\nabla f(x), \langle \nabla f(x), x\rangle)\|_* &= \frac{1}{\sqrt{2}}\|(\nabla f(x), \langle \nabla f(x), x\rangle)\|_2 \\ &\le \frac{1}{\sqrt{2}}\sqrt{\|\nabla f(x)\|_2^2 + \|\nabla f(x)\|^2\|x\|^2} \\ &\le \frac{1}{\sqrt{2}}\sqrt{L^2 + L^2R^2} = 1.\end{aligned}$$

We now show the lower bound $\alpha \ge r/4$. Firstly, since $\|(-z,1)\|\|(\vec{0},1)\|_* = \|(-z,1)\|_2\|(\vec{0},1)\|_2 \le \sqrt{2}$, we see that $\langle(-z,1),(\vec{0},1)\rangle = 1 \ge \frac{1}{2}\|(-z,1)\|\|(\vec{0},1)\|_*$. Next, any (a,b) returned by the oracle is normalized so that $\|(a,b)\|_* = 1 \Leftrightarrow \|(a,b)\|_2 = \sqrt{2}$. Note then that $\|(-z,1)\|\|(a,b)\|_* \le 2$. From here, we observe that

$$\langle(a,b),(-z,1)\rangle = b - \langle a, z\rangle = b - \langle a, z + ra/\|a\|_2\rangle + r\|a\|_2 \ge r\|a\|_2,$$

since $z + ra/\|a\|_2 \in K$ by assumption. Furthermore, $b - \langle a, z\rangle \ge b - \|a\|_2\|z\|_2 \ge b - \|a\|_2$ and $0 \le b - \langle a, z\rangle \le b + \|a\|_2$. Thus, $b - \langle a, z\rangle \ge \max\{r\|a\|_2, b - \|a\|_2\}$. We now examine two cases. If $\|a\|_2 \ge 1/2$, then $b - \langle a, z\rangle \ge r/2 \ge r/4 \cdot \|(-z,1)\|\|(a,b)\|_*$. If $\|a\|_2 \le 1/2$, then $|b| \ge 1$ since $\|(a,b)\|_2^2 = 2$. Since $b + \|a\|_2 \ge 0 \Rightarrow b \ge 1$. This gives $b - \langle a, z\rangle \ge b - \|a\|_2 \ge 1/2 \ge r/2$. Thus, $\alpha \ge r/4$, as needed. □

5.3 Computational experiments

In this section, we provide a computational comparison of our method with the standard cut loop, the ellipsoid method, and the analytic center cutting plane method on a testbed of linear optimization instances. For comparison purposes, all four methods are embedded into a common cutting plane framework such that the same termination criteria apply.

Framework. Each method has access to a separation oracle that is equipped with a set of initial linear inequalities valid for K (such as bounds on variables), which are incorporated within each method in a straightforward way. For instance, we initialize our algorithm by adding these constraints to the set A_1. Moreover, for each instance, we will be given a finite upper bound UB and incorporate the linear inequality $f(x) \le \text{UB}$ in a similar way. This upper bound gets updated whenever a feasible solution of better objective value was found. Our framework collects all inequalities queried by the current method and computes the resulting lower bound on

the optimum value in every iteration. Each method is stopped whenever the difference of upper and lower bound is below 10^{-3}.

We will also inspect the possibility of a *smart* oracle that, regardless of whether a given point x is feasible, may still provide a valid inequality as well as a feasible solution (for instance, by modifying x in a simple way so that it becomes feasible). For some problems we consider, such an oracle is available and will be specified below.

Implementation. The framework has been implemented in `julia 1.6.2` [20] using `JuMP` [74] and `Gurobi 9.1.1` [104]. To guarantee a fair comparison, all four methods have been implemented in a straightforward fashion. We use the textbook implementation of the ellipsoid method, and Badenbroek's [10] implementation of the analytic center cutting plane method. Our method is implemented [49] in the spirit of Theorem 5.2.6, where p_t is computed using `Gurobi`.

Test sets. We use three problem classes in our experiments: linear programming formulations of the maximum-cardinality matching problem, semidefinite relaxations of the maximum cut problem, and LPBoost instances for classification problems.

For the maximum-cardinality matching problem, we consider the linear program

$$\max \Big\{ \sum_{e\in E} x_e : x \in [0,1]^E, \sum_{e\in\delta(v)} x_e \le 1 \text{ for all } v \in V, \sum_{e\in E[U]} x_e \le \tfrac{|U|-1}{2} \text{ for all } U \subseteq V \text{ with } |U| \text{ odd} \Big\},$$

due to Edmonds [76], where $G = (V, E)$ is a given undirected graph, $\delta(v)$ is the set of all edges incident to v, and $E[U]$ is the set of all edges with both endpoints in U. The latter constraints are handled within an oracle that computes an inequality minimizing $(|U| - 1)/2 - \sum_{e\in E[U]} x_e$, whereas the other inequalities are provided as initial constraints. For the above problem, the smart version of the oracle does not provide a feasible point since there is no obvious way of transforming a given point into a feasible one. However, the smart version always provides the minimizing inequality.

We consider 16 random instances with 500 nodes, generated as follows. For each $r \in \{30, 33, \ldots, 75\}$ we build an instance by sampling r triples of nodes $\{u, v, w\}$ and adding the edges of the induced triangles to the graph. We believe that these instances are interesting because the r triangles give rise to many constraints to be added by the oracle. Moreover, we selected all 13 instances from the Color02 symposium [46] with less than 300 edges.

Our second set of instances is based on the semidefinite relaxation of Goemans and Williamson [90] for the maximum cut problem

$$\max \Big\{ \sum_{\{v,w\}\in E} c(v,w)(1 - X_{v,w})/2 : X_{v,w} = X_{w,v} \text{ for all } v, w \in V, \\ X_{v,v} = 1 \text{ for all } v \in V, \\ X \text{ is positive semidefinite} \Big\},$$

where c are edge weights on the edges of (V, E). We add the constraints $X \in [-1, 1]^{V\times V}$ to the initial constraints and handle the semidefiniteness constraint by a separation oracle that, given X, computes an eigenvector h of X of minimum eigenvalue and returns the inequality $\langle hh^\intercal, X\rangle \geq 0$.

Within the smart version of the oracle, this constraint is returned regardless of the feasibility of X. If X is not feasible, the semidefinite matrix $\frac{1}{\lambda-1}X - \frac{\lambda}{\lambda-1}I$ is returned, where λ denotes the minimum eigenvalue and I the identity matrix. We generated 10 complete graphs with edge weights chosen uniformly at random in $[0, 1]$.

Our third set of instances arises from LPBoost [65], a classifier algorithm based on column generation. To solve the pricing problem in column generation, the following linear program is solved:

$$\max \Big\{ \gamma : (\gamma, \lambda) \in [-1, 1] \times [0, D]^n, \langle \vec{1}, \lambda\rangle = 1, \sum_{i=1}^{m} y_i h(x^i, \omega)\lambda_i \leq -\gamma \text{ for } \omega \in \Omega \Big\},$$

where Ω is a set of parameters, for $i \in [m]$, x^i is a data point labeled as $y_i = \pm 1$, $h(\cdot, \omega)$ is a classifier parameterized by $\omega \in \Omega$ that predicts the label of x^i as $h(x^i, \omega) \in \{-1, +1\}$, and $D > 0$ is a parameter. In our experiments, we restrict $h(\cdot, \omega)$ to be a decision tree of height 1, so-called tree stumps, and choose $D = \frac{5}{n}$. To separate a point (γ', λ'), we use `julia`'s `DecisionTree` module to compute a decision stump with score function λ' that weights the data points, whose corresponding inequality classifies (γ', λ') as feasible or not. A smart oracle always returns the computed inequality and decreases γ' until (γ', λ') becomes feasible according to the found decision stump.

We extracted all data sets from the UC Irvine Machine Learning Repository [189] that are labeled as multivariate, classification, ten-to-hundred attributes, hundred-to-thousand instances. Data sets with alpha-numeric values or too many missing values have been discarded.

Results. In what follows, we report on the number of iterations, i.e., oracle calls, each method needs to obtain a gap (upper bound minus lower bound) below 10^{-3}. We impose a limit of 500 iterations per instance. Since we are testing naive implementations of each method, we do not report on running time.

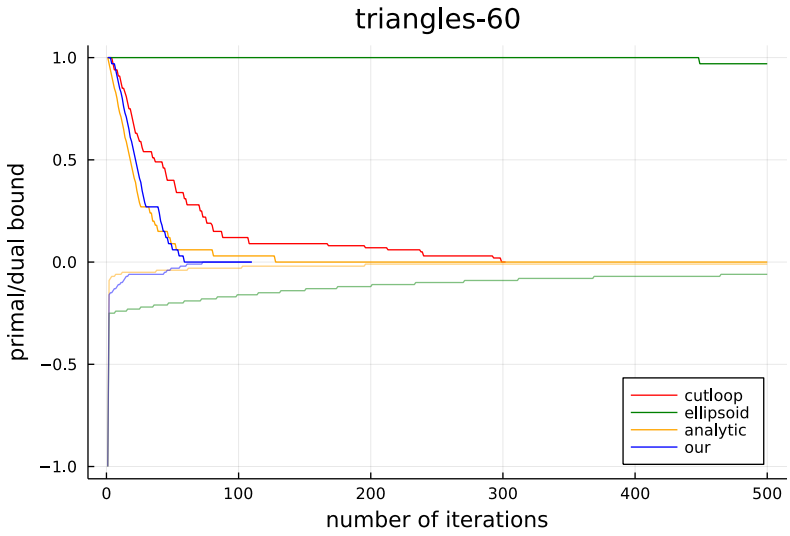


Figure 5.1: Typical primal/dual bounds for a random matching instance.

To get more insights on the primal and dual performance of the tested methods, we also report on their *primal and dual integrals*. Note that we are solving maximization problems in this section, as opposed to minimization problems in Section 5.2. That is, primal (dual) solutions provide lower (upper) bounds on OPT. If ℓ_i is the lower bound on the optimal objective value OPT in iteration i, the *primal integral* is $\sum_{i=1}^{500} \frac{\text{OPT}-\ell_i}{\text{OPT}-\ell_1}$. The *dual integral* is computed analogously. If an integral is small, this indicates quick progress in finding the correct value of the corresponding bound.

Table 5.1 summarizes our results without smart oracles, where all numbers are average values. Here, "matching" refers to the random instances and "matching02" to the instances from the Color02 symposium. The standard cut loop is referred to as "LP", the ellipsoid method as "ellipsoid", the analytic center method as "analytic", and Algorithm 5 as "our". Note that Table 5.1 does not report on the primal integral of "LP" since the standard cut loop is a dual method.

We see that the ellipsoid and analytic center methods are struggling with solving any instance within 500 iterations independent from the problem class. Our algorithm solves the instances of the matching and max-cut problem much faster than the standard cut loop. Only for LPBoost, the standard cut loop clearly dominates our algorithm. To better understand this behavior, the integrals reveal that our algorithm is better in improving the primal bound than the dual bound, with the only exception being LPBoost. The analytic center method, however, performs significantly worse than our algorithm in improving the primal bound. Regarding the dual bound, it

Table 5.1: Comparison of iterations and dual/primal integral without smart oracles.

	number of iterations			
instance	LP	ellipsoid	analytic	our
matching	175.44	500.00	500.00	99.81
matching02	283.77	460.77	491.69	47.15
maxcut	265.30	500.00	500.00	193.30
LPboost	91.94	489.06	479.12	278.06
	dual integral			
matching	48.34	473.02	22.13	21.10
matching02	257.76	339.67	194.26	21.64
maxcut	7.72	44.32	3.48	6.14
LPboost	3.15	13.62	20.65	53.15
	primal integral			
matching	—	52.12	9.29	4.40
matching02	—	23.41	5.91	2.13
maxcut	—	21.15	9.04	6.32
LPboost	—	459.97	100.71	64.08

performs better than our algorithm (with the exception of matching02). The ellipsoid method is much worse in improving the primal bound in comparison with the analytic center method and our algorithm. Regarding the dual bound, a similar trend can be observed with LPBoost being an exception.

In summary, the analytic center cutting plane method improves the dual bound more quickly than our algorithm. It can find a good primal solution early as the primal integral is small, however it fails to close the remaining gap within the iteration limit. Our algorithm is able to close the primal gap faster, with the trade-off of a slightly slower dual convergence. A typical plot of the of the relative primal and dual gaps is given in Figure 5.1.

In a second experiment, we investigate the effect of smart oracles. As Table 5.2 shows, there is no impact of smart oracles on the matching instances. For max-cut, our algorithm gets slightly slower and the other methods do not seem to be affected by smartness. For LPBoost, all methods benefit from a smart oracle with the biggest effect for analytic center and our algorithm. The reason for the positive effect for LPBoost might be in the particular structure of these instances: the objective just consists of γ and every truncated convex combination λ is feasible.

Table 5.2: Comparison of iterations and dual/primal integral with smart oracles.

	number of iterations			
instance	LP	ellipsoid	analytic	our
matching	175.44	500.00	500.00	99.81
matching02	283.77	460.77	491.69	47.15
maxcut	265.30	500.00	500.00	231.00
LPboost	86.94	346.38	88.00	127.00
	dual integral			
matching	48.34	473.02	22.13	21.10
matching02	257.76	339.67	194.26	21.64
maxcut	7.72	42.90	3.48	6.15
LPboost	3.04	13.50	5.54	5.46
	primal integral			
matching	—	52.12	9.29	4.40
matching02	—	23.41	5.91	2.13
maxcut	—	20.42	8.91	5.59
LPboost	—	25.41	6.83	6.95

Bibliography

[1] Ilan Adler. The expected number of pivots needed to solve parametric linear programs and the efficiency of the self-dual simplex method. Technical report, University of California, Dept. of IE and OR, 1983. URL: https://adler.ieor.berkeley.edu/ilans_pubs/lp_expected_self_dual_1983.pdf.

[2] Ilan Adler, Richard M. Karp, and Ron Shamir. A simplex variant solving an $m \times d$ linear program in $O(\min(m^2, d^2))$ expected number of pivot steps. *Journal of Complexity*, 3(4):372–387, 1987. doi:10.1016/0885-064X(87)90007-0.

[3] Ilan Adler and Nimrod Megiddo. A simplex algorithm whose average number of steps is bounded between two quadratic functions of the smaller dimension. *Journal of the ACM*, 32(4):871–895, 1985. doi:10.1145/4221.4222.

[4] Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice-Hall, Inc., 1993.

[5] Xavier Allamigeon, Pascal Benchimol, Stéphane Gaubert, and Michael Joswig. Log-barrier interior point methods are not strongly polynomial. *SIAM Journal on Applied Algebra and Geometry*, 2(1):140–178, 2018. doi:10.1137/17m1142132.

[6] Nina Amenta and Günter M. Ziegler. Deformed products and maximal shadows. *Contemporary Mathematics*, 223:57–90, 1998. doi:10.1090/conm/223/03132.

[7] David Arthur, Bodo Manthey, and Heiko Röglin. Smoothed analysis of the k-means method. *Journal of the ACM*, 58(5):Art. 19, 31, 2011. Preliminary version in FOCS '09. doi:10.1145/2027216.2027217.

[8] David S. Atkinson and Pravin M. Vaidya. A cutting plane algorithm for convex programming that uses analytic centers. *Mathematical Programming*, 69(1):1–43, 1995. doi:10.1007/bf01585551.

[9] David Avis and Vasek Chvátal. Notes on Bland's pivoting rule. In *Polyhedral Combinatorics*, volume 8 of *Mathematical Programming Studies*, pages 24–34. Springer, 1978. doi:10.1007/bfb0121192.

[10] Riley Badenbroek and Etienne de Klerk. An analytic center cutting plane method to determine complete positivity of a matrix. *INFORMS Journal on Computing*, 2021. doi:10.1287/ijoc.2021.1108.

[11] Michel L. Balinski. The Hirsch conjecture for dual transportation polyhedra. *Mathematics of Operations Research*, 9(4):629–633, 1984. doi:10.1287/moor.9.4.629.

[12] Keith Ball. An elementary introduction to modern convex geometry. *Flavors of Geometry*, 31:1–58, 1997. URL: http://library.msri.org/books/Book31/files/ball.pdf.

[13] Imre Bárány and Zoltán Füredi. On the shape of the convex hull of random points. *Probability Theory and Related Fields*, 77(2):231–240, 1988. doi:10.1007/bf00334039.

[14] David Barnette. Wv paths on 3-polytopes. *Journal of Combinatorial Theory*, 7(1):62–70, 1969. doi:10.1016/s0021-9800(69)80007-4.

[15] David Barnette. An upper bound for the diameter of a polytope. *Discrete Mathematics*, 10(1):9–13, 1974. doi:10.1016/0012-365X(74)90016-8.

[16] Amir Beck. *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. SIAM, 2017. doi:10.1137/1.9781611974997.

[17] Alexandre Belloni, Robert M. Freund, and Santosh Vempala. An efficient rescaled perceptron algorithm for conic systems. *Mathematics of Operations Research*, 34(3):621–641, 2009. doi:10.1287/moor.1090.0388.

[18] Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on Modern Convex Optimization*. MOS-SIAM Series on Optimization. SIAM, 2001. doi:10.1137/1.9780898718829.

[19] Ulrich Betke. Relaxation, new combinatorial and polynomial algorithms for the linear feasibility problem. *Discrete & Computational Geometry*, 32(3), 2004. doi:10.1007/s00454-004-2878-4.

[20] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017. doi:10.1137/141000671.

[21] Robert E. Bixby. A brief history of linear and mixed-integer programming computation. *Documenta Mathematica*, pages 107–121, 2012. URL: https://assets.gurobi.com/pdfs/a-brief-history-of-optimization.pdf.

[22] Robert E. Bixby, Mary Fenelon, Zonghao Gu, Ed Rothberg, and Roland Wunderling. MIP: theory and practice—closing the gap. In *System Modelling and Optimization*, pages 19–49, 2000. doi:10.1007/978-0-387-35514-6.

[23] Nicolas Bonifas, Marco Di Summa, Friedrich Eisenbrand, Nicolai Hähnle, and Martin Niemeier. On sub-determinants and the diameter of polyhedra. *Discrete & Computational Geometry*, 52(1):102–115, 2014. doi:10.1007/s00454-014-9601-x.

[24] Tommy Bonnesen and Werner Fenchel. *Theory of convex bodies*. BCS Associates, 1987. Translated from the German and edited by L. Boron, C. Christenson and B. Smith.

[25] Gilles Bonnet, Daniel Dadush, Uri Grupel, Sophie Huiberts, and Galyna Livshyts. Asymptotic bounds on the combinatorial diameter of random polytopes. 2021. arXiv:2112.13027.

[26] Karl Heinz Borgwardt. *Untersuchungen zur Asymptotik der mittleren Schrittzahl von Simplexverfahren in der linearen Optimierung*. PhD thesis, Universität Kaiserslautern, 1977.

[27] Karl Heinz Borgwardt. The average number of pivot steps required by the simplex-method is polynomial. *Zeitschrift für Operations Research*, 26(5):A157–A177, 1982. doi:10.1007/bf01917108.

[28] Karl Heinz Borgwardt. *The simplex method: A probabilistic analysis*, volume 1 of *Algorithms and Combinatorics: Study and Research Texts*. Springer, 1987. doi:10.1007/978-3-642-61578-8.

[29] Karl Heinz Borgwardt. Erratum: A sharp upper bound for the expected number of shadow vertices in LP-polyhedra under orthogonal projection on two-dimensional planes. *Mathematics of Operations Research*, 24(4):925–984, 1999. doi:10.1287/moor.24.4.925.

[30] Karl Heinz Borgwardt. A sharp upper bound for the expected number of shadow vertices in LP-polyhedra under orthogonal projection on two-dimensional planes. *Mathematics of Operations Research*, 24(3):544–603, 1999. A corrected version of Figure 1 appears in Volume 24(4), pp. 925–984. doi:10.1287/moor.24.3.544.

[31] Karl Heinz Borgwardt and Petra Huhn. A lower bound on the average number of pivot-steps for solving linear programs valid for all variants of the simplex-algorithm. *Mathematical Methods of Operations Research*, 49(2):175–210, 1999. doi:10.1007/s186-1999-8373-5.

[32] Steffen Borgwardt, Jesús A. De Loera, and Elisabeth Finhold. The diameters of network-flow polytopes satisfy the Hirsch conjecture. *Mathematical Programming*, 171(1-2):283–309, 2017. doi:10.1007/s10107-017-1176-x.

[33] Andreas Brieden, Peter Gritzmann, Ravindran Kannan, Victor Klee, László Lovász, and Miklós Simonovits. Deterministic and randomized polynomial-time approximation of radii. *Mathematika*, 48(1-2):63–105, 2001. doi:10.1112/s0025579300014364.

[34] Graham Brightwell, Jan Van den Heuvel, and Leen Stougie. A linear bound on the diameter of the transportation polytope. *Combinatorica*, 26(2):133–139, 2006. doi:10.1007/s00493-006-0010-5.

[35] Tobias Brunsch, Kamiel Cornelissen, Bodo Manthey, Heiko Röglin, and Clemens Rösner. Smoothed analysis of the successive shortest path algorithm. *SIAM Journal on Computing*, 44(6):1798–1819, 2015. Preliminary version in SODA '13. doi:10.1137/140989893.

[36] Tobias Brunsch, Anna Großwendt, and Heiko Röglin. Solving totally unimodular LPs with the shadow vertex algorithm. In *Proceedings of the 32nd Symposium on Theoretical Aspects of Computer Science*, pages 171–183, 2015. doi:10.4230/LIPIcs.STACS.2015.171.

[37] Sébastien Bubeck and Ronen Eldan. The entropic barrier: a simple and optimal universal self-concordant barrier. 2014. arXiv:1412.1587.

[38] Clément Canonne. A short note on poisson tail-bounds. Github repository link: https://github.com/ccanonne/probabilitydistributiontoolbox/blob/master/poissonconcentration.pdf, 2019.

[39] Chandra Chekuri and Kent Quanrud. Approximating the Held-Karp bound for metric TSP in nearly-linear time. In *58th Annual Symposium on Foundations of*

Computer Science, pages 789–800. IEEE, 2017. doi:10.1109/focs.2017.78.

[40] Chandra Chekuri and Kent Quanrud. Near-linear time approximation schemes for some implicit fractional packing problems. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 801–820. SIAM, 2017. doi:10.1137/1.9781611974782.51.

[41] Elliot Ward Cheney and Allen A. Goldstein. Newton's method for convex programming and Tchebycheff approximation. *Numerische Mathematik*, 1(1):253–268, 1959. doi:10.1007/bf01386389.

[42] Sergei Chubanov. A strongly polynomial algorithm for linear systems having a binary solution. *Mathematical Programming*, 134(2):533–570, 2011. doi:10.1007/s10107-011-0445-3.

[43] Sergei Chubanov. A polynomial algorithm for linear optimization which is strongly polynomial under certain conditions on optimal solutions. http://www.optimization-online.org/DB_HTML/2014/12/4710.html, 2014.

[44] Sergei Chubanov. A polynomial algorithm for linear feasibility problems given by separation oracles. *Optimization Online, Jan*, 2017. http://www.optimization-online.org/DB_HTML/2017/01/5838.html.

[45] Michael B. Cohen, Yin Tat Lee, and Zhao Song. Solving linear programs in the current matrix multiplication time. In *Proceedings of the 51st Annual ACM Symposium on Theory of Computing*, pages 938–942, 2019. doi:10.1145/3313276.3316303.

[46] Color02 - computational symposium: Graph coloring and its generalizations. available at, 2002. http://mat.gsia.cmu.edu/COLOR02.

[47] Daniel Dadush and Nicolai Hähnle. On the shadow simplex method for curved polyhedra. *Discrete & Computational Geometry*, 56(4):882–909, 2016. Preliminary version in SOCG '15. doi:10.1007/s00454-016-9793-3.

[48] Daniel Dadush, Christopher Hojny, Sophie Huiberts, and Stefan Weltge. A simple method for convex optimization in the oracle model. 2021. arXiv:2011.08557.

[49] Daniel Dadush, Christopher Hojny, Sophie Huiberts, and Stefan Weltge. Supplement: A simple method for convex optimization in the oracle model, 2021. URL: https://github.com/christopherhojny/supplement_simple-iterative-methods-linopt-convex-sets.

[50] Daniel Dadush and Sophie Huiberts. A friendly smoothed analysis of the simplex method. *SIAM Journal on Computing*, 49(5):STOC18–449, 2019. Preliminary version in STOC '18. doi:10.1137/18m1197205.

[51] Daniel Dadush and Sophie Huiberts. Smoothed analysis of the simplex method. In Tim Roughgarden, editor, *Beyond the Worst-Case Analysis of Algorithms*, pages 309–333. Cambridge University Press, 2021. doi:10.1017/9781108637435.019.

[52] Daniel Dadush, Sophie Huiberts, Bento Natura, and László A Végh. A scaling-invariant algorithm for linear programming whose running time depends only on the constraint matrix. In *Proceedings of the 52nd Annual ACM Symposium on Theory of Computing*, pages 761–774, 2020. doi:10.1145/3357713.3384326.

[53] Daniel Dadush, Bento Natura, and László A. Végh. Revisiting Tardos's framework for linear programming: Faster exact solutions using approximate solvers. In *61st Annual Symposium on Foundations of Computer Science*, pages 931–942. IEEE, 2020. doi:10.1109/focs46700.2020.00091.

[54] Daniel Dadush, László A. Végh, and Giacomo Zambelli. Rescaling algorithms for linear conic feasibility. *Mathematics of Operations Research*, 45(2):732–754, 2020. doi:10.1287/moor.2019.1011.

[55] Samuel I. Daitch and Daniel A. Spielman. Faster approximate lossy generalized flow via interior point algorithms. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 451–460, 2008. doi:10.1145/1374376.1374441.

[56] Daryl J. Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes*. Springer, 2008. doi:10.1007/978-0-387-49835-5.

[57] Valentina Damerow and Christian Sohler. Extreme points under random noise. In *European Symposium on Algorithms*, pages 264–274. Springer, 2004. doi:10.1007/978-3-540-30140-0_25.

[58] George B. Dantzig. Programming in a linear structure. Technical report, U.S. Air Force Comptroller, USAF, Washington, D.C., 1948.

[59] George B. Dantzig. Maximization of a linear function of variables subject to linear inequalities. In *Activity Analysis of Production and Allocation*, Cowles Commission Monograph No. 13, pages 339–347. John Wiley & Sons, Inc., New York, N. Y.; Chapman & Hall, Ltd., London, 1951. URL: https://cowles.yale.edu/sites/default/files/files/pub/mon/m13-all.pdf.

[60] George B. Dantzig. *Linear programming and extensions*. Princeton University Press, 1959. doi:10.7249/r366.

[61] George B. Dantzig. Converting a converging algorithm into a polynomially bounded algorithm. Technical report, Stanford University, 1991. URL: https://apps.dtic.mil/sti/pdfs/ADA234961.pdf.

[62] Jesús A. De Loera, Raymond Hemmecke, and Jon Lee. On augmentation algorithms for linear and integer-linear programming: From Edmonds–Karp to Bland and beyond. *SIAM Journal on Optimization*, 25(4):2494–2511, 2015. doi:10.1137/151002915.

[63] Jesús A. De Loera, Sean Kafer, and Laura Sanità. Pivot rules for circuit-augmentation algorithms in linear optimization. 2019. arXiv:1909.12863.

[64] Alberto Del Pia and Carla Michini. On the diameter of lattice polytopes. *Discrete & Computational Geometry*, 55(3):681–687, 2016. doi:10.1007/s00454-016-9762-x.

[65] Ayhan Demiriz, Kristin P Bennett, and John Shawe-Taylor. Linear programming boosting via column generation. *Machine Learning*, 46(1):225–254, 2002. doi:10.1023/a:1012470815092.

[66] Amit Deshpande and Daniel A. Spielman. Improved smoothed analysis of the shadow vertex simplex method. In *46th Annual Symposium on Foundations of Computer Science*, pages 349–356. IEEE, 2005. doi:10.1109/SFCS.2005.44.

[67] Olivier Devillers, Marc Glisse, Xavier Goaoc, and Rémy Thomasse. On the smoothed complexity of convex hulls. In *31st International Symposium on Computational Geometry*, pages 224–238, 2015. doi:10.4230/LIPIcs.SOCG.2015.224.

[68] Olivier Devillers, Marc Glisse, Xavier Goaoc, and Rémy Thomasse. Smoothed complexity of convex hulls by witnesses and collectors. *Journal of Computational Geometry*, 7(2):101–144, 2016. doi:10.20382/jocg.v7i2a6.

[69] Antoine Deza and Lionel Pournin. Improved bounds on the diameter of lattice polytopes. *Acta Mathematica Hungarica*, 154(2):457–469, 2018. doi:10.1007/s10474-017-0777-4.

[70] Robert Dial, Fred Glover, David Karney, and Darwin Klingman. A computational analysis of alternative algorithms and labeling techniques for finding shortest path trees. *Networks*, 9(3):215–248, 1979. doi:10.1002/net.3230090304.

[71] Ilya I. Dikin. Iterative solution of problems of linear and quadratic programming. *Doklady Akademii Nauk*, 174(4):747–748, 1967.

[72] Yann Disser, Oliver Friedmann, and Alexander V. Hopp. An exponential lower bound for Zadeh's pivot rule. 2020. arXiv:1911.01074.

[73] John Dunagan and Santosh Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. *Mathematical Programming*, 114(1):101–114, 2007. doi:10.1007/s10107-007-0095-7.

[74] Iain Dunning, Joey Huchette, and Miles Lubin. JuMP: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017. doi:10.1137/15M1020575.

[75] Martin Dyer and Alan Frieze. Random walks, totally unimodular matrices, and a randomised dual simplex algorithm. *Mathematical Programming*, 64(1):1–16, 1994. doi:10.1007/bf01582563.

[76] Jack Edmonds. Maximum matching and a polyhedron with 0,1-vertices. *Journal of Research of the National Bureau of Standards*, 69B(1–2):125–130, 1964. doi:10.6028/jres.069b.013.

[77] Friedrich Eisenbrand, Nicolai Hähnle, Alexander Razborov, and Thomas Rothvoss. Diameter of polyhedra: Limits of abstraction. *Mathematics of Operations Research*, 35(4):786–794, 2010. doi:10.1287/moor.1100.0470.

[78] Friedrich Eisenbrand and Santosh Vempala. Geometric random edge. *Mathematical Programming*, 164(1-2):325–339, 2017. doi:10.1007/s10107-016-1089-0.

[79] Matthias Englert, Heiko Röglin, and Berthold Vöcking. Worst case and probabilistic analysis of the 2-opt algorithm for the TSP. *Algorithmica*, 68(1):190–264, 2014. Preliminary version in SODA '07. doi:10.1007/s00453-013-9801-4.

[80] Kai-Tai Fang, Samuel Kotz, and Kai-Wang Ng. *Symmetric Multivariate and Related Distributions*. Monographs on Statistics and Applied Probability. Springer, 1990. doi:10.1201/9781351077040.

[81] John Fearnley and Rahul Savani. The complexity of the simplex method. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pages 201–208, 2015. doi:10.1145/2746539.2746558.

[82] John J. Forrest and Donald Goldfarb. Steepest-edge simplex algorithms for linear programming. *Mathematical Programming*, 57(3):341–374, 1992. doi:10.1007/BF01581089.

[83] András Frank. *Connections in Combinatorial Optimization*. Number 38 in Oxford Lecture Series in Mathematics and its Applications. Oxford University Press, 2011. URL: https://web.cs.elte.hu/~frank/cikkek/FrankB1.pdf.

[84] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956. doi:10.1002/nav.3800030109.

[85] Oliver Friedmann. A subexponential lower bound for Zadeh's pivoting rule for solving linear programs and games. In Oktay Günlük and Gerhard J. Woeginger, editors, *Proceedings of 15th International Conference on Integer Programming and Combinatorial Optimization*, pages 192–206. Springer, 2011. doi:10.1007/978-3-642-20807-2_16.

[86] Oliver Friedmann, Thomas Dueholm Hansen, and Uri Zwick. Subexponential lower bounds for randomized pivoting rules for the simplex algorithm. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, pages 283–292, 2011. doi:10.1145/1993636.1993675.

[87] Naveen Garg and Jochen Könemann. Faster and simpler algorithms for multicommodity flow and other fractional packing problems. *SIAM Journal on Computing*, 37(2):630–652, 2007. doi:10.1137/s0097539704446232.

[88] Saul Gass and Thomas Saaty. The computational algorithm for the parametric objective function. *Naval Research Logistics Quarterly*, 2:39–45, 1955. doi:10.1002/nav.3800020106.

[89] Marc Glisse, Sylvain Lazard, Julien Michel, and Marc Pouget. Silhouette of a random polytope. *Journal of Computational Geometry*, 7(1):14, 2016. doi:10.20382/jocg.v7i1a5.

[90] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, 1995. doi:10.1145/227683.227684.

[91] Jean-Louis Goffin. The relaxation method for solving systems of linear inequalities. *Mathematics of Operations Research*, 5(3):388–414, 1980. doi:10.1287/moor.5.3.388.

[92] Andrew V. Goldberg, Michael D. Grigoriadis, and Robert E. Tarjan. Use of dynamic trees in a network simplex algorithm for the maximum flow problem. *Mathematical Programming*, 50(3):277–290, 1991. doi:10.1007/BF01594940.

[93] Donald Goldfarb. Using the steepest-edge simplex algorithm to solve sparse linear programs. *Sparse matrix computations*, pages 227–240, 1976. doi:10.1016/B978-0-12-141050-6.50018-0.

[94] Donald Goldfarb. Worst case complexity of the shadow vertex simplex algorithm. Technical report, Columbia University, New York, 1983.

[95] Donald Goldfarb and Jianxiu Hao. A primal simplex algorithm that solves the maximum flow problem in at most nm pivots and $O(n^2m)$ time. *Mathematical Programming*, 47(1):353–365, 1990. doi:10.1007/bf01580869.

[96] Donald Goldfarb and Jianxiu Hao. Polynomial-time primal simplex algorithms for the minimum cost network flow problem. *Algorithmica*, 8(2):145–160, 1992. doi:10.1007/BF01758840.

[97] Donald Goldfarb, Jianxiu Hao, and Sheng-Roan Kai. Efficient shortest path simplex algorithms. *Operations Research*, 38(4):624–628, 1990. doi:10.1287/opre.38.4.624.

[98] Donald Goldfarb and William Y. Sit. Worst case behavior of the steepest edge simplex method. *Discrete Applied Mathematics*, 1(4):277–285, 1979. doi:10.1016/0166-218X(79)90004-0.

[99] Ralph E. Gomory. Outline of an algorithm for integer solutions to linear programs. *Bull. Amer. Math. Soc.*, 64:275–278, 1958. doi:10.1090/s0002-9904-1958-10224-4.

[100] Clovis C. Gonzaga. Path-following methods for linear programming. *SIAM review*, 34(2):167–224, 1992. doi:10.1137/1034048.

[101] Clovis C. Gonzaga and Hugo J. Lara. A note on properties of condition numbers. *Linear Algebra and its Applications*, 261(1):269 – 273, 1997. doi:10.1016/s0024-3795(96)00409-0.

[102] Richard C. Grinold. The Hirsch conjecture in Leontief substitution systems. *SIAM Journal on Applied Mathematics*, 21(3):483–485, 1971. doi:10.1137/0121052.

[103] Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer, 1988. doi:10.1007/978-3-642-78240-4.

[104] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2022. URL: https://www.gurobi.com.

[105] M. Haimovich. The simplex method is very good! On the expected number of pivot steps and related properties of random linear programs. Technical report, Columbia University, 1983.

[106] Thomas Dueholm Hansen and Uri Zwick. An improved version of the random-facet pivoting rule for the simplex algorithm. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pages 209–218, 2015. doi:10.1145/2746539.2746557.

[107] Jackie C.K. Ho and Levent Tunçel. Reconciliation of various complexity and condition measures for linear programming problems and a generalization of Tardos' theorem. In *Foundations of Computational Mathematics*, pages 93–147. World Scientific, 2002. doi:10.1142/9789812778031_0006.

[108] Gabriele Höfner. *Lineare Optimierung mit dem Schatteneckenalgorithmus: Untersuchungen zum mittleren Rechenaufwand und Entartungsverhalten*. PhD thesis, Universität Augsburg, 1995.

[109] Sophie Huiberts. How Large is the Shadow? Smoothed Analysis of the Simplex Method. Master's thesis, Utrecht University, the Netherlands, 2018. URL: https://dspace.library.uu.nl/handle/1874/359199.

[110] Ming S. Hung. A polynomial simplex method for the assignment problem. *Operations Research*, 31(3):595–600, 1983. doi:10.1287/opre.31.3.595.

[111] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. volume 28 of *Proceedings of Machine Learning Research*, pages 427–435. PMLR, 2013. URL: http://proceedings.mlr.press/v28/jaggi13.html.

[112] Svante Janson. Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms*, 24(3):234–248, 2004. doi:10.1002/rsa.20008.

[113] Robert G. Jeroslow. The simplex algorithm with the pivot rule of maximizing criterion improvement. *Discrete Mathematics*, 4:367–377, 1973. doi:10.1016/0012-365X(73)90171-4.

[114] Haotian Jiang, Yin Tat Lee, Zhao Song, and Sam Chiu-wai Wong. An improved cutting plane method for convex optimization, convex-concave games, and its applications. In *Proceedings of the 52nd Annual ACM Symposium on Theory of Computing*, page 944953, 2020. doi:10.1145/3357713.3384284.

[115] Satoshi Kakihara, Atsumi Ohara, and Takashi Tsuchiya. Information geometry and interior-point algorithms in semidefinite programs and symmetric cone programs. *Journal of Optimization Theory and Applications*, 157:749–780, 2013. doi:10.1007/s10957-012-0180-9.

[116] Satoshi Kakihara, Atsumi Ohara, and Takashi Tsuchiya. Curvature integrals and iteration complexities in SDP and symmetric cone programs. *Computational Optimization and Applications*, 57:623–665, 2014. doi:10.1007/s10589-013-9608-x.

[117] Gil Kalai. A subexponential randomized simplex algorithm. In *Proceedings of the 24th Annual ACM Symposium on Theory of Computing*, pages 475–482, 1992. doi:10.1145/129712.129759.

[118] Gil Kalai and Daniel J. Kleitman. A quasi-polynomial bound for the diameter of graphs of polyhedra. *Bull. Amer. Math. Soc.*, 26(2):315–317, 1992. doi:10.1090/s0273-0979-1992-00285-9.

[119] Narendra Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4(4):373–395, 1984. doi:10.1007/BF02579150.

[120] James E. Kelley, Jr. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 8(4):703–712, 1960. doi:10.1137/0108053.

[121] Jonathan A. Kelner and Daniel A. Spielman. A randomized polynomial-time simplex algorithm for linear programming. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 51–60, 2006. doi:10.1145/1132516.1132524.

[122] Leonid G. Khachiyan. A polynomial algorithm in linear programming (in Russian). *Doklady Akademiia Nauk SSSR 224*, 224:1093–1096, 1979. English Translation: Soviet Mathematics Doklady 20, 191-194.

[123] Tomonari Kitahara and Shinji Mizuno. A bound for the number of different basic solutions generated by the simplex method. *Mathematical Programming*, 137(1-2):579–586, 2013. doi:10.1007/s10107-011-0482-y.

[124] Tomonari Kitahara and Takashi Tsuchiya. A simple variant of the Mizuno–Todd–Ye predictor-corrector algorithm and its objective-function-free complexity. *SIAM Journal on Optimization*, 23(3):1890–1903, 2013. doi:10.1137/110835475.

[125] Victor Klee and George J. Minty. How good is the simplex algorithm. In *Inequalies : III : Proceedings of the 3rd Symposium on inequalities*, pages 159–175, 1972.

[126] Victor Klee and David W. Walkup. The d-step conjecture for polyhedra of dimension $d < 6$. *Acta Mathematica*, 117:53–78, 1967. doi:10.1007/bf02395040.

[127] Peter Kleinschmidt and Shmuel Onn. On the diameter of convex polytopes. *Discrete mathematics*, 102(1):75–77, 1992. doi:10.1016/0012-365x(92)90349-k.

[128] Jean-Philippe Labbé, Thibault Manneville, and Francisco Santos. Hirsch polytopes with exponentially long combinatorial segments. *Mathematical Programming*, 165(2):663–688, 2017. doi:10.1007/s10107-016-1099-y.

[129] Guanghui Lan, Renato D.C. Monteiro, and Takashi Tsuchiya. A polynomial predictor-corrector trust-region algorithm for linear programming. *SIAM Journal on Optimization*, 19(4):1918–1946, 2009. doi:10.1137/070693461.

[130] David G. Larman. Paths on polytopes. *Proc. London Math. Soc. (3)*, s3-20(1):161–178, 1970. doi:10.1112/plms/s3-20.1.161.

[131] Yin Tat Lee and Aaron Sidford. Efficient inverse maintenance and faster algorithms for linear programming. In *56th Symposium on Foundations of Computer Science*, pages 230–249. IEEE, 2015. doi:10.1109/focs.2015.23.

[132] Yin Tat Lee and Aaron Sidford. Solving linear programs with $\tilde{O}(\sqrt{\text{rank}})$ linear system solves. 2019. Preliminary version in FOCS '14. arXiv:1910.08033.

[133] Yin Tat Lee, Aaron Sidford, and Sam Chiu wai Wong. A faster cutting plane method and its implications for combinatorial and convex optimization. In *56th Annual Symposium on Foundations of Computer Science*, pages 1049–1065. IEEE, 2015. doi:10.1109/FOCS.2015.68.

[134] Carlton E. Lemke. Bimatrix equilibrium points and mathematical programming. *Management science*, 11(7):681–689, 1965. doi:10.1287/mnsc.11.7.681.

[135] Aleksander Madry. Navigating central path with electrical flows: From flows to matchings, and back. In *54th Annual Symposium on Foundations of Computer Science*, pages 253–262. IEEE, 2013. doi:10.1109/focs.2013.35.

[136] Jiří Matoušek. *Lectures on Discrete Geometry*. Springer, 2002. doi:10.1007/978-1-4613-0039-7.

[137] Jiří Matoušek and Bernd Gärtner. *Understanding and using linear programming*. Springer, 2007. doi:10.1007/978-3-540-30717-4.

[138] Jiri Matoušek, Micha Sharir, and Emo Welzl. A subexponential bound for linear programming. *Algorithmica*, 16(4-5):498–516, 1996. doi:10.1007/s004539900062.

[139] Nimrod Megiddo. Towards a genuinely polynomial algorithm for linear programming. *SIAM Journal on Computing*, 12(2):347–353, 1983. doi:10.1137/0212022.

[140] Nimrod Megiddo. A note on the generality of the self-dual algorithm with various starting points. *Methods of operations research*, 49:271–275, 1985. URL: http://theory.stanford.edu/~megiddo/pdf/selfdual.pdf.

[141] Nimrod Megiddo. Improved asymptotic analysis of the average number of steps performed by the self-dual simplex algorithm. *Mathematical Programming*, 35(2):140–172, 1986. doi:10.1007/BF01580645.

[142] Nimrod Megiddo, Shinji Mizuno, and Takashi Tsuchiya. A modified layered-step interior-point algorithm for linear programming. *Mathematical Programming*, 82(3):339–355, 1998. doi:10.1007/bf01580074.

[143] Sanjay Mehrotra. On the implementation of a primal-dual interior point method. *SIAM Journal on Optimization*, 2(4):575–601, 1992. doi:10.1137/0802028.

[144] Carla Michini and Antonio Sassano. The Hirsch Conjecture for the fractional stable set polytope. *Mathematical Programming*, 147(1-2):309–330, 2014. doi:10.1007/s10107-013-0723-3.

[145] Shinji Mizuno, Michael Todd, and Yinyu Ye. On adaptive-step primal-dual interior-point algorithms for linear programming. *Mathematics of Operations Research*, 18:964–981, 1993. doi:10.1287/moor.18.4.964.

[146] Renato D. C. Monteiro and Takashi Tsuchiya. A variant of the Vavasis-Ye layered-step interior-point algorithm for linear programming. *SIAM Journal on Optimization*, 13(4):1054–1079, 2003. doi:10.1137/s1052623401388926.

[147] Renato D.C. Monteiro and Takashi Tsuchiya. A new iteration-complexity bound for the MTY predictor-corrector algorithm. *SIAM Journal on Optimization*, 15(2):319–347, 2005. doi:10.1137/s1052623402416803.

[148] Renato D.C. Monteiro and Takashi Tsuchiya. A strong bound on the integral of the central path curvature and its relationship with the iteration-complexity of primal-dual path-following LP algorithms. *Mathematical Programming*, 115(1):105–149, 2008. doi:10.1007/s10107-007-0141-5.

[149] Katta G. Murty. Computational complexity of parametric linear programming. *Mathematical Programming*, 19(1):213–219, 1980. doi:10.1007/BF01581642.

[150] Denis Naddef. The Hirsch conjecture is true for (0, 1)-polytopes. *Mathematical Programming*, 45(1-3):109–110, 1989. doi:10.1007/bf01589099.

[151] Hariharan Narayanan, Rikhav Shah, and Nikhil Srivastava. A spectral approach to polytope diameter. 2021. arXiv:2101.12198.

[152] Arkadi Nemirovsky and David Yudin. Informational complexity and efficient methods for solution of convex extremal problems. *Ékonomika i Mathematicheskie Metody*, 12, 1983.

[153] Yurii Nesterov. Complexity estimates of some cutting plane methods based on the analytic barrier. *Mathematical Programming*, 69(1):149–176, 1995. doi:10.1007/bf01585556.

[154] Dianne P. O'Leary. On bounds for scaled projections and pseudoinverses. *Linear Algebra and its Applications*, 132:115–117, 1990. doi:10.1016/0024-3795(90)90056-i.

[155] Neil Olver and László A. Végh. A simpler and faster strongly polynomial algorithm for generalized flow maximization. In *Proceedings of the 49th Annual ACM Symposium on Theory of Computing*, pages 100–111, 2017. doi:10.1145/3055399.3055439.

[156] James B. Orlin. Genuinely polynomial simplex and non-simplex algorithms for the minimum cost flow problem. Technical report, Massachusetts Institute of Technology, 1984. URL: https://dspace.mit.edu/bitstream/handle/1721.1/48015/genuinelypolynom00orli.pdf.

[157] James B. Orlin, Serge A. Plotkin, and Éva Tardos. Polynomial dual network simplex algorithms. *Mathematical Programming*, 60(3):255–276, 1993. doi:10.1007/BF01580615.

[158] Serge A. Plotkin, David B. Shmoys, and Éva Tardos. Fast approximation algorithms for fractional packing and covering problems. *Mathematics of Operations Research*, 20(2):257–301, 1995. doi:10.1287/moor.20.2.257.

[159] Ian Post and Yinyu Ye. The simplex method is strongly polynomial for deterministic markov decision processes. *Mathematics of Operations Research*, 40(4):859–868, 2015. Preliminary version in SODA '14. doi:10.1287/moor.2014.0699.

[160] James Renegar. A polynomial-time algorithm, based on Newton's method, for linear programming. *Mathematical Programming*, 40(1):59–93, 1988. doi:10.1007/BF01580724.

[161] James Renegar. Is it possible to know a problem instance is ill-posed?: some foundations for a general theory of condition numbers. *Journal of Complexity*, 10(1):1–56, 1994. doi:10.1006/jcom.1994.1001.

[162] James Renegar. Incorporating condition measures into the complexity theory of linear programming. *SIAM Journal on Optimization*, 5(3):506–524, 1995. doi:10.1137/0805026.

[163] Aleksandr Reznikov and Edward B. Saff. The covering radius of randomly distributed points on a manifold. *International Mathematics Research Notices*, 2016(19):6065–6094, 2015. doi:10.1093/imrn/rnv342.

[164] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. doi:10.1037/h0042519.

[165] Laura Sanità. The diameter of the fractional matching polytope and its hardness implications. In *59th Annual Symposium on Foundations of Computer Science*, pages 910–921. IEEE, 2018. doi:10.1109/focs.2018.00090.

[166] Arvind Sankar, Daniel A. Spielman, and Shang-Hua Teng. Smoothed analysis of the condition numbers and growth factors of matrices. *SIAM Journal on Matrix Analysis and Applications*, 28(2):446–476, 2006. doi:10.1137/S0895479803436202.

[167] Francisco Santos. A counterexample to the Hirsch Conjecture. *Annals of Mathematics*, 176(1):383–412, 2012. doi:10.4007/annals.2012.176.1.7.

[168] Emanuel Schnalzger. *Lineare Optimierung mit dem Schatteneckenalgorithmus im Kontext probabilistischer Analysen*. PhD thesis, Universität Augsburg, 2014. English translation by K.H. Borgwardt.

[169] Rolf Schneider and Wolfgang Weil. *Stochastic and integral geometry*. Probability and its Applications. Springer, 2008. doi:10.1007/978-3-540-78859-1.

[170] Alexander Schrijver. *Combinatorial Optimization – Polyhedra and Efficiency*. Springer, 2003. doi:10.1002/9781118033142.

[171] Farhad Shahrokhi and David W. Matula. The maximum concurrent flow problem. *Journal of the ACM*, 37(2):318–334, 1990. doi:10.1145/77600.77620.

[172] Ron Shamir. The efficiency of the simplex method: a survey. *Management Science*, 33(3):301–334, 1987. doi:10.1287/mnsc.33.3.301.

[173] Steve Smale. On the average number of steps of the simplex method of linear programming. *Mathematical programming*, 27(3):241–262, 1983. doi:10.1007/bf02591902.

[174] Steve Smale. Mathematical problems for the next century. *The Mathematical Intelligencer*, 20(2):7–15, 1998. doi:10.1007/bf03025291.

[175] György Sonnevend. *New Algorithms in Convex Programming Based on a Notion of "Centre" (for Systems of Analytic Inequalities) and on Rational Extrapolation*, pages 311–326. Birkhäuser Basel, 1988. doi:10.1007/978-3-0348-9297-1_20.

[176] György Sonnevend, Josef Stoer, and Gongyun Zhao. On the complexity of following the central path of linear programs by linear extrapolation II. *Mathematical Programming*, 52(1-3):527–553, 1991. doi:10.1007/bf01582904.

[177] Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of termination of linear programming algorithms. *Mathematical Programming*, 97(1-2):375–404, 2003. doi:10.1007/s10107-003-0448-9.

[178] Daniel A. Spielman and Shang-Hua Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 81–90, 2004. doi:10.1145/1007352.1007372.

[179] Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: why the simplex algorithm usually takes polynomial time. *Journal of the ACM*, 51(3):385–463, 2004. doi:10.1145/990308.990310.

[180] Gilbert W. Stewart. On scaled projections and pseudoinverses. *Linear Algebra and its Applications*, 112:189–193, 1989. doi:10.1016/0024-3795(89)90594-6.

[181] Noriyoshi Sukegawa. An asymptotically improved upper bound on the diameter of polyhedra. *Discrete & Computational Geometry*, pages 1–10, 2018. doi:10.1007/s00454-018-0016-y.

[182] Éva Tardos. A strongly polynomial minimum cost circulation algorithm. *Combinatorica*, 5(3):247–255, 1985. doi:10.1007/bf02579369.

[183] Éva Tardos. A strongly polynomial algorithm to solve combinatorial linear programs. *Operations Research*, 34(2):250–256, 1986. doi:10.1287/opre.34.2.250.

[184] Michael J. Todd. Polynomial expected behavior of a pivoting algorithm for linear complementarity and linear programming problems. *Mathematical Programming*, 35(2):173–192, 1986. doi:10.1007/bf01580646.

[185] Michael J. Todd. A Dantzig–Wolfe-like variant of Karmarkar's interior-point linear programming algorithm. *Operations Research*, 38(6):1006–1018, 1990. doi:10.1287/opre.38.6.1006.

[186] Michael J. Todd. An improved Kalai–Kleitman bound for the diameter of a polyhedron. *SIAM Journal on Discrete Mathematics*, 28(4):1944–1947, 2014. doi:10.1137/140962310.

[187] Michael J. Todd, Levent Tunçel, and Yinyu Ye. Characterizations, bounds, and probabilistic analysis of two complexity measures for linear programming problems. *Mathematical Programming*, 90(1):59–69, 2001. doi:10.1007/pl00011420.

[188] Levent Tunçel. Approximating the complexity measure of Vavasis-Ye algorithm is NP-hard. *Mathematical Programming*, 86(1):219–223, 1999. doi:10.1007/s101070050087.

[189] UC Irvine Machine Learning Repository. https://archive-beta.ics.uci.edu/ml/datasets. accessed September 3, 2021.

[190] Pravin M. Vaidya. Speeding-up linear programming using fast matrix multiplication. In *30th Annual Symposium on Foundations of Computer Science*, pages 332–337. IEEE, 1989. doi:10.1109/sfcs.1989.63499.

[191] Pravin M. Vaidya. A new algorithm for minimizing convex functions over convex sets. *Mathematical Programming*, 73(3):291–341, 1996. doi:10.1007/bf02592216.

[192] Jan van den Brand. A deterministic linear program solver in current matrix multiplication time. In *Proceedings of the 31st Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 259–278. SIAM, 2020. doi:10.1137/1.9781611975994.16.

[193] Jan van den Brand, Yin Tat Lee, Yang P. Liu, Thatchaphol Saranurak, Aaron Sidford, Zhao Song, and Di Wang. Minimum cost flows, MDPs, and ℓ_1-regression in nearly linear time for dense instances. In *Proceedings of the 53rd Annual ACM Symposium on Theory of Computing*, page 859869, 2021. doi:10.1145/3406325.3451108.

[194] Jan van den Brand, Yin Tat Lee, Danupon Nanongkai, Richard Peng, Thatchaphol Saranurak, Aaron Sidford, Zhao Song, and Di Wang. Bipartite matching in nearly-linear time on moderately dense graphs. In *61st Annual Symposium on Foundations of Computer Science*, pages 919–930. IEEE, 2020. doi:10.1109/focs46700.2020.00090.

[195] Jan van den Brand, Yin Tat Lee, Aaron Sidford, and Zhao Song. Solving tall dense linear programs in nearly linear time. In *Proceedings of the 52nd Annual ACM Symposium on Theory of Computing*, pages 775–788, 2020. doi:10.1145/3357713.3384309.

[196] Virginia Vassilevska Williams. Multiplying matrices faster than Coppersmith-Winograd. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing*, pages 887–898, 2012. doi:10.1145/2213977.2214056.

[197] Stephen A. Vavasis. Stable numerical algorithms for equilibrium systems. *SIAM Journal on Matrix Analysis and Applications*, 15(4):1108–1131, 1994. doi:10.1137/s0895479892230948.

[198] Stephen A. Vavasis and Yinyu Ye. A primal-dual interior point method whose running time depends only on the constraint matrix. *Mathematical Programming*, 74(1):79–120, 1996. doi:10.1007/bf02592148.

[199] László A. Végh. A strongly polynomial algorithm for generalized flow maximization. *Mathematics of Operations Research*, 42(1):179–211, 2017. doi:10.1287/moor.2016.0800.

[200] Roman Vershynin. Beyond Hirsch conjecture: walks on random polytopes and smoothed complexity of the simplex method. *SIAM Journal on Computing*, 39(2):646–678, 2009. Preliminary version in FOCS '06. doi:10.1137/070683386.

[201] Stephen J. Wright. *Primal-Dual Interior-Point Methods*. SIAM, 1997. doi:10.1137/1.9781611971453.

[202] Yinyu Ye. *Interior-Point Algorithms: Theory and Analysis*. John Wiley and Sons, 1997. doi:10.1002/9781118032701.

[203] Yinyu Ye. A new complexity result on solving the markov decision problem. *Mathematics of Operations Research*, 30(3):733–749, 2005. doi:10.1287/moor.1050.0149.

[204] Yinyu Ye. Improved complexity results on solving real-number linear feasibility problems. *Mathematical Programming*, 106(2):339–363, 2006. doi:10.1007/s10107-005-0610-7.

[205] Yinyu Ye. The simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate. *Mathematics of Operations Research*, 36(4):593–603, 2011. doi:10.1287/moor.1110.0516.

Samenvatting

Lineaire programmering (LP) is een wiskundige manier om verschillende praktische optimalisatieproblemen te formuleren. Lineaire programmeringsproblemen komen op veel plekken in de praktijk voor, zoals bij productieplanning. Er bestaan daarom verschillende softwarepakketten die deze lineaire programmeringsproblemen kunnen inlezen en vervolgens een goede oplossing uitrekenen. In dit proefschrift bestuderen we enkele veelgebruikte algoritmes in deze software, en aan de hand van meetkundige principes analyseren we hoe snel deze algoritmes zijn.

Een LP-probleem wordt omschreven aan de hand van randvoorwaarden, gegeven door een matrix $A \in \mathbb{R}^{m \times n}$ en vector $b \in \mathbb{R}^m$, en een doelfunctie $c \in \mathbb{R}^n$. De taak is om een vector $x \in \mathbb{R}^n$ te vinden met zo groot mogelijk inproduct $c^\mathsf{T} x$, onder de voorwaarde dat x voldoet aan het stelsel van lineaire ongelijkheden $Ax \leq b$. Dit schrijven we ook al op als

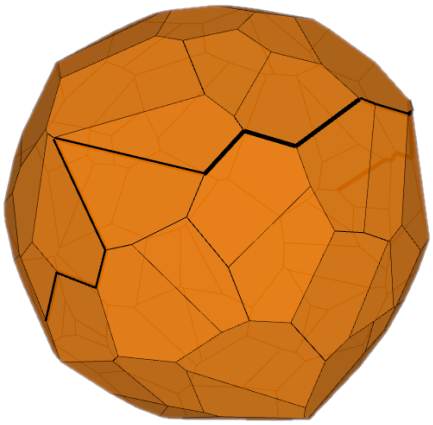
$$\begin{aligned} &\text{maximaliseer } c^\mathsf{T} x \\ &\text{met voorwaarde } Ax \leq b. \end{aligned}$$

De verzameling van alle vectoren die aan de randvoorwaarden voldoen noemen we het *toegestane gebied*. Meetkundig gezien vormt dit gebied een veelvlak. Gegeven dit veelvlak, vraagt een lineair programmeringsprobleem om een punt daarin te vinden dat zo ver mogelijk in een aangegeven richting ligt.

6.1 De simplexmethode

Er worden verschillende algoritmes gebruikt om LP-problemen op te lossen. De oudste hiervan is de *simplexmethode*, en vandaag de dag is dit nog steeds een van de snelste algoritmes.

De simplexmethode kiest in het begin een deelverzameling $B \subseteq \{1, \ldots, m\}$ van de rijen van de matrix A. Voor de deelmatrix A_B en vector b_B vindt het algoritme het punt $x_B \in \mathbb{R}^n$ dat voldoet aan het stelsel van vergelijkingen $A_B x_B = b_B$. De keuze van B wordt zo gedaan, dat het punt x_B uniek gedefinieerd is en voldoet aan alle randvoorwaarden $Ax_B \leq b$. Dit kunnen we meetkundig interpreteren als dat het algoritme begint op een hoekpunt van de toegestane verzameling.



Figuur 6.1: Een oranje veelvlak, met dikgedrukt een pad van hoekpunten verbonden met ribben.

In elke stap van de simplexmethode wordt een enkel element uit de verzameling B vervangen met een nieuwe index uit $\{1, \ldots, m\}$, op zo een manier dat het nieuwe punt x_B ook weer voldoet aan alle randvoorwaarden. Dit vervangen gebeurt zodanig dat het inproduct $c^\mathsf{T} x_B$ groter wordt. Dit proces herhaalt zich totdat het een optimale oplossing vind.

Meetkundig kunnen we ons voorstellen dat de simplexmethode op "reis" gaat langs de rand van het toegestane gebied. Deze reis gaat van hoekpunt naar hoekpunt over de ribben van dit veelvlak, en komt elke stap dichter bij de bestemming. Een voorbeeld van een toegestane gebied en een mogelijke reis is te vinden in Figure 6.1.

In Hoofdstuk 2 bestuderen we hoeveel stappen de simplexmethode nodig heeft om een LP-probleem op te lossen. Hier zullen we aannemen dat de data een kleine hoeveelheid toevallige "ruis" bevat. We omschrijven een nieuwe variant op de simplexmethode die weinig stappen nodig heeft. In Hoofdstuk 3 bestuderen we toevallige veelvlakken. We bewijzen boven- en ondergrenzen op de lengte van de kortste reis tussen hoekpunten.

6.2 Inwendige-punt methoden

De andere veelgebruikte algoritmen voor lineaire programmeringsproblemen zijn de *inwendige-punt methoden*. Deze algoritmen werken met punten in het inwendige van het toegestane gebied, in plaats van punten op de rand zoals de simplexmethode.

Veel inwendige-punt methoden volgen gedurende hun looptijd een *centraal pad* door het toegestane gebied, met een optimale geldige oplossing aan het einde van het pad. Deze methoden zijn in praktijk erg snel, en in theorie kunnen we de looptijd van deze methoden van boven begrensen met behulp van het "formaat" van de data A, b, en c.

Er bestaan inwendige-punt methoden die *schaal-invariant* zijn. Dit houdt in dat het voor het algoritme niet uit maakt in welke schaal de grootheden in A, b, en c zijn geformuleerd, zoals bijvoorbeeld in meters en kilogrammen of in yards en pounds. In theorie is dit een gewenste eigenschap.

Eerder onderzoek heeft uitgewezen dat er ook inwendige-punt methoden bestaan waarvan de looptijd enkel afhangt van het formaat van de matrix A, en dus niet van de vectoren b en c. In 2003 stelde Monteiro en Tsuchiya de vraag of er een methode bestaat die allebei deze eigenschappen heeft: zowel schaal-invariant en met looptijd begrensd met behulp van het formaat van A. In Hoofdstuk 4 laten we zien dat dit inderdaad het geval is. We omschrijven een inwendige-punt methode die op schaal-invariante wijze het *primaal-duale centrale pad* van een LP-probleem kan volgen om een oplossing te vinden. De looptijd van onze methode hangt enkel af van de matrix A.

We introduceren een nieuwe maat om te meten hoe "gebalanceerd" de matrix A is. We geven een methode om deze gebalanceerdheid bij benadering uit te rekenen, en begrenzen de looptijd van onze inwendige-punt methode met behulp van deze gebalanceerdheid.

6.3 Snijvlakmethode

Voor LP-problemen mogen we ervan uitgaan dat alle randvoorwaarden op voorhand bekend zijn. In Hoofdstuk 5 laten we deze aanname los. In plaats van een lijst met alle randvoorwaarden, mogen we nu vragen stellen aan een "orakel". We leggen dit orakel telkens een punt $x \in \mathbb{R}^d$ voor, en het orakel laat weten of dit punt in het toegestane gebied ligt. Als het punt niet toegestaan is, dan geeft het orakel daarvoor een "reden": we krijgen een lineaire ongelijkheid die geldig is voor het hele toegestane gebied maar ongeldig is voor ons voorgestelde punt x.

Algoritmes voor optimalisatie in dit orakel-model worden snijvlakmethoden genoemd, omdat elke nieuwe lineaire ongelijkheid punten "afsnijd" van het toegestane gebied. In Hoofdstuk 5 omschrijven we een nieuwe snijvlakmethode. Onze methode geeft zowel boven- als ondergrenzen op de waarde van het optimalisatieprobleem, en de twee grenzen bewegen bewijsbaar snel naar elkaar toe.

Ook doen we ook enkele experimenten met verschillende optimalisatieproblemen. Vergeleken met drie bestaande snijvlakmethoden, heeft ons algoritme maar weinig vragen aan het orakel nodig om een toegestane oplossing met goede waarde te vinden.

Curriculum vitae

Sophie Huiberts was born on 24 August 1994 in Heerenveen, The Netherlands. In 2012 she started her studies at Utrecht University, receiving her bachelor's degree in mathematics in 2016, her bachelor's degree in computer science in 2017 and her master's degree in mathematics in 2018. She wrote her master's thesis at Centrum Wiskunde & Informatica on smoothed analysis of the simplex method and continued working on related topics as a PhD candidate.

At CWI she has been part of the works council during 2018–2022 and was chair of this body in 2021–2022. She wrote the 'LaTeX in Slack' browser extension, a remote collaboration tool for mathematicians with over 3500 users. Sophie was a Rising Star speaker at the TCS Women Spotlight Workshop of STOC 2021 and was awarded a Junior Fellowship by the Simons Society of Fellows in 2022.

Acknowledgements

I have many people to be grateful for. I can only hope that these words are not the first time they learn of my gratitude.
First, my thanks are for Daniel. You have introduced me to questions and techniques, to communities and practices, to people and opportunities. Thank you for mentoring and supporting me in innumerable ways throughout this journey.
I am grateful to Gunther for his insights and strategy, as well as for his practical support in making all this possible.
The research I worked on over the past years could not have happened without my stellar co-authors. Bento, Christopher, Daniel, Galyna, Gilles, Laci, Samarth, Sander, Stefan and Uri, thank you for our collaboration. I would be delighted to work with you on other projects in the future. Special thanks to Stefan for hosting me at the Technical University of Munich in early 2020. I had a wonderful time working with you and interacting with the research group while there.
I would like to thank Robert Luce and Sebastian Pokutta for their valuable feedback on the work that became Chapter 5, and I thank all members of the assessment committee for their time and effort.
CWI has been the perfect place for me to work, and I thank all of my colleagues for making it such a pleasant environment. That includes, but is not limited to, the Networks & Optimization group, the works council, and everyone I met through the PhD Activitee Committee's events.
Cactus, Paulette, Stefan, Tim, you each have my special thanks for proofreading parts of this thesis. Marieke, you have been a great help in navigating the final steps of the PhD.
And of course, thanks to all my other friends and family. I appreciate each and every one of you so much.