ALGEBRAIC METHODS IN CONTINUOUS OPTIMIZATION

A Dissertation
Presented to
The Academic Faculty

By

Kevin Shu

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Algorithms, Combinatorics, and Optimization
Home Department: School of Mathematics

Georgia Institute of Technology

Mar 2024

ALGEBRAIC METHODS IN CONTINUOUS OPTIMIZATION

Thesis committee:

Dr. Greg Blekherman
School of Mathematics
Georgia Institute of Technology

Dr. Amir Ali Ahmadi
Department of Operations Research and Financial Engineering
Princeton University

Dr. Santanu Dey
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Shixuan Zhang
Department of Industrial & Systems Engineering
Texas A&M University

Dr. Mohit Singh
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date approved: TODO

Ad astra per aspera.

-Variant of a Latin saying

ACKNOWLEDGMENTS

I have been immensely well supported throughout my time as a graduate student. I would like to thank the following people specifically.

To my advisor, Greg Blekherman, who gave me a wealth of opportunities to learn and to grow intellectually. Beyond just teaching me new approaches to asking and solving questions, you have been patient and understanding when I was struggling. I could not have done this without your guidance and your support through these years.

To my many fantastic coauthors: Santanu Dey, Shengding Sun, Mario Kummer, Raman Sanyal, Spencer Gordon, Alex Wang, Akshay Ramachandran, Julia Lindberg, Alejandro Toriello, Diego Cifuentes, and Ben Grimmer. Thank you for your help and for teaching me so many new things.

To Bernd Sturmfels, for allowing me to visit the Max Plack Institute for a wonderful summer.

To my parents, who supported me even as I have moved back and forth throughout the country.

To my friends at Georgia Tech: Adam Brown, Abhishek Dhawan, Chris Dupre, Christina Giannitsi, Dan Minahan, Mirabel Reid, and more that I cannot enumerate. Thanks for keeping me sane. To my friends from Caltech that have continued to support me, and in particular Mark Gillespie.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis broadly concerns the usage of techniques from algebra, the study of higher order structures in mathematics, toward understanding difficult optimization problems. Of particular interest will be optimization problems related to systems of polynomial equations, algebraic invariants of topological spaces, and algebraic structures in convex optimization.

We will discuss various concrete examples of these kinds of problems. Firstly, we will describe new constructions for a class of polynomials known as hyperbolic polynomials which have connections to convex optimization. Secondly, we will describe how we can use ideas from algebraic geometry, notably the study of Stanley-Reisner varieties to study sparse structures in semidefinite programming. This will lead to quantitative bounds on some approximations for sparse problems and concrete connections to sparse linear regression and sparse PCA. Thirdly, we will use methods from algebraic topology to show that certain optimization problems on nonconvex topological spaces can be turned into convex problems due to a phenomenon known as 'hidden convexity'. Specifically, we give a sufficient condition for the image of a topological space under a continuous map to be convex, and give a number of examples of this phenomena with practical importance. This unifies and generalizes a number of existing results. Finally, we will describe how to use techniques inspired by the sum of squares method to find new variants of gradient descent which converge faster than typical gradient descent on smooth convex problems.

CHAPTER 1
INTRODUCTION

1.1 High level overview

Modern large scale algorithmic decision making problems have led to a recent explosion of mathematical problems related to optimization. These ask the question of how to best utilize some limited collection of resources or else how to make good decisions under complicated constraints. This thesis will describe how these questions connect naturally to ideas from algebra and in particular, the study of polynomial equations and algebraic invariants in topology.

The appearance of algebra in optimization has a similar underlying cause to that of convexity in optimization: both are ways of enforcing a global structure on an optimization problem. This global structure is important in cases in which some property of a solution is a hard requirement but difficult to enforce. In real world examples such as the management of a power plant or an aircraft, mistakes may not be acceptable. In these cases, mathematical guarantees that they will not fail in normal operation are needed, and such guarantees cannot be provided by heuristics or local optimization methods. These guarantees are also needed when these applications are then applied to other areas of formal mathematics, which require rigorous proof of correctness for these solutions. Algebraic methods offer these kinds of global guarantees.

From a mathematical perspective, interesting questions come about from considering computational problems with complicated constraints which arise from geometric or physical considerations. Much of our discussion will concern the interplay between convex optimization and algebraic geometry. While these subjects may seem vastly

different, in a sense, they can both be understood as different generalizations of linear algebra, which we may think of as the study of linear equations. Convex geometry can be understood loosely as the study of linear inequalities, while algebraic geometry can be understood loosely as the study of higher degree polynomial equations. For this reason, many higher order concepts from linear algebra such as matrix groups or the spectral theorem can be understood through an algebraic geometry lense or a convex geometry lense. Many such connections will play prominent roles in various parts of this thesis.

To give an explicit example of the type of work described here, we will give a high level overview of the results of Chapter 4. This chapter concerns a method for minimizing a class of functions using *gradient descent*. For a function f starting at some point x_0, the method iterates the update rule

$$x_t = x_{t-1} + h_{t-1}\nabla f(x_{t-1}).$$

Here, the choice of step size sequence h_t does not depend on the function f and is fixed in advance of the algorithm. While different methods for performing this minimization have been discovered, within this class of optimization methods, it was not known how to asymptotically improve on the sequence of step sizes where h_t is constant, in which case the gap in the value of f to the minimum shrinks at a rate of $O(\frac{1}{t})$. The basic issue is that if h_t are too large, then this method may begin to diverge, while steps that are too small will lead to slow convergence.

The 'meta-optimization' problem is therefore to find the choice of h_t so that this iteration converges to the minimum as fast as possible, and it is necessary to alternate between larger steps that move toward the minimum quickly and smaller steps that correct errors that the larger steps may introduce. Therefore, in order to improve the asymptotics of this method, global considerations are necessary due to the fact that

the objective value may not be decreasing at every iteration. For this, we consider the set of L-smooth convex functions, a natural class of functions which satisfy the inequalities that for any $x, y \in \mathbb{R}^n$,

$$f(x) - f(y) \geq \langle \nabla f(y), x - y \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2.$$

These are polynomial inequalities in terms of the values of f and its gradients, and while they are individually not very hard to prove, *by taking combinations of these inequalities*, we are able to find a new sequence of step sizes achieving a rate of $O(\frac{1}{t^{1+\epsilon}})$ for $\epsilon > 0$. Once an appropriate combination is found, elementary algebraic manipulations can be used to derive the appropriate inequalities showing the desired convergence rate.

1.1.1 Summaries of chapters

We will summarize the contents of the individual chapters here, though some terms have yet to be defined fully.

Chapter 2 concerns *hyperbolic polynomials*, a class of polynomials which have arisen in a number of different contexts ranging from differential equations to combinatorics. These polynomials can be viewed as satisfying a generalized version of the spectral theorem from linear algebra, and they have associated to them convex cones known as hyperbolicity cones which generalize the convex cone of positive semidefinite matrices. This chapter will primarily be focused on some new constructions of such hyperbolic polynomials which may have utility in future applications, though this chapter is primarily focused on their mathematical structure.

Chapter 3 concerns *sparsity in semidefinite programming*. There are a variety of different notions of sparsity. One notion concerns situations in which the input to some optimization problem uses far fewer parameters than a general input instance, and

how to exploit that structure to improve efficiency. Another notion concerns situations in which the output to be found is desired to be sparse in the sense of having many zero entries. This chapter describes a framework for understanding sparse structures in the context of a particularly prominent kind of optimization known as semidefinite programming. This offers both structural and quantitative analysis of these sparse problems and gives applications to sparse versions of linear regression and PCA, as well as some interesting optimization problems regarding eigenvalues of certain classes of matrices.

Chapter 4 concerns *hidden convexity*. These types of results reduce complicated optimization problems over nonconvex domains to convex optimization problems which are often tractable. In particular, the perspective taken in this chapter shows how to prove that the images of certain topological spaces under continuous maps are convex using algebraic topology. This proves both new results and unifies and simplifies a number of existing theorems concerning hidden convexity in the context of Lie groups. We also give an application to the problem of finding a rotation matrix that maps one set of points to another subject to a linear constraint on that matrix.

?? concerns *accelerating gradient descent using long steps*. Here, we consider the problem of choosing step sizes in a common iterative method for minimizing a convex function. We show that by occasionally taking exceptionally large steps, it is possible to achieve faster convergence rates to the optimum than is possible using steps of constant size. These results are shown using the algebraic technique of combining a number of simple inequalities regarding our function class in somewhat complicated ways.

Next, we will give some basic definitions that are common to a number of the chapters in this thesis.

1.2 Preliminary notions

1.2.1 Algebraic geometry

Algebraic geometry for our purposes will concern the study of polynomial equations and their solution sets. We will use $k[x_1, \ldots, x_n]$ to denote the vector space of polynomials in n variables with coefficients in a field k. Given a collection of polynomials $p_1, \ldots, p_m \in k[x_1, \ldots, x_n]$, we let

$$\mathcal{V}(p_1, \ldots, p_m) = \{x \in k^n : p_1(x) = 0 \text{ and } p_2(x) = 0 \ldots \text{ and } p_m(x) = 0\}.$$

A set of this form is said to be *Zariski closed*, and it is noteworthy that any Zariski closed set has the property that its complement is either empty or dense.

We will say a polynomial p is nonnegative if $p(x) \geq 0$ for all $x \in \mathbb{R}^n$, and we will denote this by writing $p \geq 0$.

We will not require much detailed theory from algebraic geometry here, though we will recount some varieties of particular interest to us. We will define $\mathbb{R}^{n \times n}$ to be the vector space of $n \times n$ real matrices. We will also define $\mathbb{R}^{n \times n}_{sym}$ to be the vector space of symmetric matrices. We define

$$\mathrm{SO}(n) = \{X \in \mathbb{R}^{n \times n} : X^\intercal X = I, \text{ and } \det(X) = 1\}.$$

Here, I denotes the identity matrix.

1.2.2 Convex geometry

A convex set C is a subset of $\mathbb{R}^n$ with the property that if $x, y \in C$ then the line segment joining x and y is also contained in C. Equivalently, a convex set is the set of points satisfying a (possibly infinite) collection of linear inequalities of the form $\langle v, x \rangle \geq c$ for some $v \in \mathbb{R}^n$ and $c \in \mathbb{R}$. The convex hull of a set of points $S \subseteq \mathbb{R}^n$ is

the intersection of all convex sets containing S, and we will denote this by $\text{conv}(S)$.

A convex cone is a convex set which has the property that if $x \in C$, then $\lambda x \in C$ for $\lambda > 0$.

An extreme point of a convex set C is a point x which is not of the form $x = (1-\lambda)z_1 + \lambda z_2$, where $z_1, z_2 \neq x$ and $\lambda \in [0,1]$. Similarly, we say that if C is a convex cone, and $x \in C$, then x spans an extreme ray of C if x is not of the form $x = (1-\lambda)z_1 + \lambda z_2$ for linearly independent z_1 and z_2.

Particularly important convex cones include the nonnegative orthant $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x_i \geq 0 \text{ for } i = 1 \dots n\}$ and the positive semidefinite (PSD) cone

$$\mathcal{S}^{n\times n}_+ = \{X \in \mathbb{R}^{n\times n}_{sym} : \text{for all } v \in \mathbb{R}^n,\ v^\intercal X v \geq 0\}.$$

If $X \in \mathcal{S}^{n\times n}_+$, we will write $X \succeq 0$.

We will often refer to conical optimization problems, which for a given convex cone C are optimization problems of the form

$$\begin{aligned} \max \quad & \langle c, x\rangle \\ \text{s.t.} \quad & Ax = b \\ & x \in C. \end{aligned} \tag{1.2.1}$$

Here, $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$.

A conical optimization problem over $\mathbb{R}^n_+$ is a *linear program*, and one over $\mathcal{S}^{n\times n}_+$ is a semidefinite program.

A convex function is a function satisfying the property that the epigraph $\{(x,y) : y \geq f(x)\}$ is convex, and a function is concave if its negative is convex.

1.2.3 Convex linear algebra

Here, we will recount some convexity properties of the eigenvalues of symmetric matrices, some of which we will need for our analysis. In some places, these historical results will serve as a backdrop for our further exploration of the connections between convex geometry and algebra. We will begin by recalling the spectral theorem.

Theorem 1.2.1. *If $X \in \mathbb{R}^{n\times n}_{sym}$, then there exists some $U \in SO(n)$ and a diagonal matrix $D \in \mathbb{R}^{n\times n}$ so that*

$$X = UDU^{\intercal}.$$

Equivalently, X has n real eigenvalues (counting multiplicity) and the eigenspaces associated to distinct eigenvalues are orthogonal.

While this theorem does not appear to have much to do with convexity, it can be seen to underlie the fact that the set of symmetric matrices with nonnegative eigenvalues is a convex set using the theory of hyperbolic polynomials described in Chapter 2. We will typically denote the eigenvalues of a symmetric matrix X by

$$\lambda_1(X) \leq \lambda_2(X) \leq \cdots \leq \lambda_n(X).$$

When referring to the largest or smallest eigenvalues of a matrix X, we will sometimes use the notation $\lambda_{min}(X)$ and $\lambda_{max}(X)$ to avoid confusion.

The spectral theorem is also closely connected to many other convexity results in linear algebra. For example, using the general theory of hyperbolic polynomials, the following can be shown (though other proofs are available):

Theorem 1.2.2. *The polynomial* $\log(\det(X))$ *is concave on* $\mathcal{S}^{n\times n}_{+}$.

Using this, we can show the *Hadamard inequality* (this proof was communicated to us by James Saunderson in private communication). For $X \in \mathbb{R}^{n\times n}$, we will use

the notation $\mathrm{diag}(X) \in \mathbb{R}^n$ to denote the diagonal of X, and for $x \in \mathbb{R}^n$, we use the notation $\mathrm{Diag}(x)$ to denote the diagonal matrix whose diagonal entries are the entries of x.

Theorem 1.2.3. *If $X \in \mathcal{S}_+^{n\times n}$, then $\det(X) \leq \prod_{i=1}^n X_{ii}$.*

Proof. Let $x = \mathrm{diag}(X)$. We have the following formula, which can be proven by considering the sum entry by entry:

$$\mathrm{Diag}(x) = \frac{1}{2^n} \sum_{s\in\{-1,1\}^n} \mathrm{Diag}(s) X \mathrm{Diag}(s).$$

By concavity, we have that

$$\begin{aligned}
\log\left(\prod_{i=1}^n X_{ii}\right) &= \log(\det(\mathrm{Diag}(x))) \\
&= \log\left(\det\left(\frac{1}{2^n} \sum_{s\in\{-1,1\}^n} \mathrm{Diag}(s) X \mathrm{Diag}(s)\right)\right) \\
&\geq \frac{1}{2^n} \sum_{s\in\{-1,1\}^n} \log\left(\det\left(\mathrm{Diag}(s) X \mathrm{Diag}(s)\right)\right) \\
&= \log\left(\det\left(X\right)\right).
\end{aligned}$$

Here, we have used the fact that $\det\left(\mathrm{Diag}(s) X \mathrm{Diag}(s)\right) = \det\left(X\right)$ when $s \in \{-1,1\}^n$. This shows the desired result. □

We will now recall the Schur-Horn theorem. For a permutation $\pi \in \mathfrak{S}_n$ and a vector $v \in \mathbb{R}^n$, we will write $\pi(v)$ to denote the vector where $\pi(v)_i = v_{\pi^{-1}(i)}$. We will also need to define the variety of symmetric matrices with fixed eigenvalues. That is, for $\mu \in \mathbb{R}^n$, we let

$$M_\mu^{\mathbb{R}} = \{X \in \mathbb{R}_{sym}^{n\times n} : \text{The eigenvalues of } X \text{ are } \mu\}.$$

Theorem 1.2.4. *The image of $M_\mu^{\mathbb{R}}$ under the linear map* diag *is precisely* $\mathrm{conv}\{\pi(\mu) :$

$\pi \in \mathfrak{S}_n\}$.

We will give related 'hidden convexity' theorems in Chapter 4.

We also recall the *Cauchy interlacing theorem*. If $S \subseteq [n]$ and X is an $n \times n$ matrix, then the principal submatrix of X indexed by S is the $|S| \times |S|$ matrix given by

$$X|_S = (X_{ij})_{i,j \in S}.$$

Theorem 1.2.5. *Let X be an $n \times n$ symmetric matrix. For each $i = 1, \ldots, k$,*

$$\lambda_i(X) \leq \lambda_i(Y) \leq \lambda_{i+n-k}(X).$$

It is known that in fact, the Cauchy interlacing inequalities completely specify how the eigenvalues of a principal submatrix of a general matrix relate to the eigenvalues of the whole matrix. In Chapter 3, we in give in a sense a converse inequality where if the eigenvalues of *all of the $k \times k$ principal submatrices* of the matrix X are constrained, then in fact the eigenvalues of X are constrained in a way that is not given by the interlacing inequalities.

We will end this section with two related facts that are corollaries of the Cauchy interlacing theorem, one of which appears to be new. Firstly, we recall that the *Hadamard product* of two $n \times n$ matrices is the matrix $X \cdot Y$ where for each $i, j \in [n]$, $(X \cdot Y)_{ij} = X_{ij} Y_{ij}$.

Theorem 1.2.6. *If X and Y are symmetric matrices, then*

$$\lambda_{min}(X \cdot Y) \geq \min_{i,j \in [n]} \lambda_i(X) \lambda_j(Y),$$

and in particular, if X and Y are PSD, then $X \cdot Y$ is as well.

Proof. Notice that while $\min_{i,j \in [n]} \lambda_i(X) \lambda_j(Y)$ is invariant to change of basis in both

X and Y, $X \cdot Y$ is highly basis dependent. For this reason, it is natural to relate $X \cdot Y$ to another, larger matrix which has better basis independent properties.

Consider the *Kronecker product* $X \otimes Y$, an $n^2 \times n^2$ which represents the tensor product of the linear maps associated to X and Y. In coordinates, we may index the entries of $X \otimes Y$ by ordered pairs, so that

$$(X \otimes Y)_{(ij)(k\ell)} = X_{ik} Y_{j\ell}.$$

It is possible to explicitly construct a basis of eigenvectors of $X \otimes Y$ to show that the eigenvalues of $X \otimes Y$ are precisely those real numbers which can be represented as $\lambda_i(X)\lambda_j(Y)$ for $i, j \in [n]$.

Now, note that $X \cdot Y$ is the principal submatrix of $X \otimes Y$ obtained by restricting to rows indexed by pairs (ii). The eigenvalue bounds then follow from the Cauchy interlacing formula. □

Secondly, we will recall that the *Schur complement* of a matrix X with block decomposition

$$X = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

with respect to the $k \times k$ principal submatrix A is $X/A = D - CA^{-1}B$. More generally, if $S \subseteq [n]$, we denote by $X \setminus S = X|_{S^c} - X_{S^c,S} X|_S X_{S,S^c}$. Here, $X_{S,T}$ denotes the (possibly nonprincipal) submatrix of X indexed by the sets S and T. We will first note a formula for the entries of $X \setminus A$, which is cited in [1] and which we found in [2].

Lemma 1.2.7. *Suppose $X \in \mathbb{R}^{n \times n}$ and $S \subseteq [n]$ satisfies that $X|_S$ is nonsingular. For any $i, j \in S^c$,*

$$(X \setminus S)_{ij} = \frac{1}{\det(X|_S)} \det(X_{S \cup i, S \cup j}).$$

Next, we will show an analogue of Theorem 1.2.6, which to our knowledge is novel.

Theorem 1.2.8. *If X is a symmetric matrix, and $S \subseteq [n]$ is of size k, then*

$$\lambda_{min}(X \setminus S) \geq \frac{1}{\det(X|_S)} \min_{\substack{T \subseteq [n] \\ |T|=k+1}} \prod_{i \in T} \lambda_i(X).$$

In particular, if X is PSD, then

$$\lambda_{min}(X \setminus S) \geq \frac{\prod_{i=1}^{k} \lambda_i(X)}{\det(X|_S)} \lambda_{k+1}(X).$$

Proof. Here, we make use of the notion of a wedge power of a matrix, which is correctly defined in terms of exterior algebras. For us, if X is an $n \times n$ matrix, we define the matrix $\wedge^k X$ to be a $\binom{n}{k} \times \binom{n}{k}$ matrix, whose entries are indexed by subsets of $[n]$ of size k, and for $S, T \subseteq [n]$ with $|S| = |T| = k$,

$$(\wedge^k X)_{ST} = \det(X_{ST}).$$

While it is not clear from this coordinate focused definition, in fact, this wedge power satisfies $\wedge^k(XY) = (\wedge^k X)(\wedge^k Y)$, and it can be shown by explicitly constructing eigenvectors that the eigenvalues of $\wedge^k X$ are precisely the possible values of $\prod_{i \in S} \lambda_i(X)$ as S ranges over sets of size k.

Next, we note that $X \setminus S$ is precisely the principal submatrix of $\frac{1}{\det(X|_S)} \wedge^{k+1} X$ corresponding to those sets of the form $S \cup i$ for $i \in S^c$ by the previous lemma. Therefore, the theorem follows from Cauchy interlacing. □

CHAPTER 2
HYPERBOLIC POLYNOMIALS

2.1 Preliminary notions

2.1.1 Definitions

The main subject of this chapter are polynomials which have real rootedness properties with interesting connections to convex optimization.

Definition 2.1.1. *A homogeneous polynomial* $p \in \mathbb{R}[x_1, \ldots, x_n]$ *is said to be* hyperbolic *with respect to* $v \in \mathbb{R}^n$ *if* $p(v) \neq 0$ *and for every* $x \in \mathbb{R}^n$*, the univariate polynomial*

$$p_x(t) = p(x + tv)$$

has only real roots.

Example 2.1.1. *If we let* X *be a symmetric matrix of indeterminants, then the polynomial* $\det(X) \in \mathbb{R}[X]$ *is hyperbolic with respect to the identity matrix* $I \in \mathbb{R}^{n \times n}_{sym}$*. This follows because the univariate polynomial* $\det(X + tI)$ *has roots equal to the* $-\lambda_1, \ldots, -\lambda_n$*, where the* λ_i *are the eigenvalues of* X*, and these are real by the spectral theorem.*

A related class of polynomials are *stable polynomials*, which are polynomials $p \in \mathbb{C}[z_1, \ldots, z_n]$ with the property that $p(z_1, \ldots, z_n) \neq 0$ when z_i are all complex numbers with positive imaginary part. A homogeneous polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]$ is stable if and only if it is hyperbolic with respect to every vector in the positive orthant.[3] Hyperbolic polynomials have been used in proofs of deep results in a variety of fields ranging from differential equations to combinatorics [4, 5, 6, 7, 3].

Associated to a given hyperbolic polynomial is a closed convex cone with nonempty interior known as its *hyperbolicity cone.*

Definition 2.1.2. *Let $p \in \mathbb{R}[x_1, \ldots, x_n]$ be hyperbolic with respect to $v \in \mathbb{R}^n$. The hyperbolicity cone $\Lambda_v(p)$ of p is defined in any of the following equivalent ways:*

1. $\Lambda_v(p) = \{x \in \mathbb{R}^n : p(x + tv) > 0 \text{ when } t > 0\}$.
2. *$\Lambda_v(p)$ is the set of x where all roots of the polynomial $p(x - tv)$ are nonnegative.*
3. *$\Lambda_v(p)$ is the closure of the connected component of $\mathbb{R}^n \setminus \mathcal{V}(p)$ containing v, where $\mathcal{V}(p)$ is the set of x for which $p(x) = 0$.*
4. *$\Lambda_v(p)$ is the set of points where the coefficients of the univariate polynomial $p_x(t)$ are nonnegative.*

In addition, if $u \in \Lambda_v(p)$, then p is hyperbolic with respect to u and $v \in \Lambda_u(p)$. It is not hard to see using these definitions that the hyperbolicity cone of the determinant polynomial is precisely the cone of positive semidefinite matrices. The success of semidefinite programming has lead to analogous interest for *hyperbolicity cone programming.*

For a given hyperbolic polynomial p, a hyperbolicity cone program is an optimization problem of the form

$$\begin{aligned} \max \quad & \langle c, x \rangle \\ \text{s.t.} \quad & Ax = b \\ & x \in \Lambda_v(p) \end{aligned} \tag{2.1.1}$$

Here, $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Such problems have been studied extensively [8, 9]. Of particular note is the fact that the function $-\log(p)$ serves as a self-concordant barrier function for the hyperbolicity cone of p [10], which enables interior point methods to be applied to this problem.

2.1.2 Hyperbolic optimization and polynomial nonnegativity

One application of such hyperbolicity cone programs is towards certifying the nonnegativity of polynomials. In [11], it was shown that $\Lambda_v(p)$ is a slice of the cone of nonnegative polynomials. Explicitly, if we let $D_u p$ denote the directional derivative of p with respect to the vector u, then

$$\Lambda_v(p) = \{u \in \mathbb{R}^n : D_u p D_v p - p D_u D_v p \geq 0\}.$$

Here, the polynomial $\Delta_{u,v} p = D_u p D_v p - p D_u D_v p$ is known as the *mixed derivative* of p.

This was further extended in [12], which shows that a matrix known as the *Bézoutian* of p, denoted $B_v(p)$ is positive semidefinite for all x. They use this to define the notion of a hyperbolic certificate of nonnegativity in such a way that a polynomial is a sums of squares if and only if it has a hyperbolic certificate of nonnegativity where the underlying polynomial is the determinant.

Remark 1. *It is not hard to show that every sum of squares polynomial can be written as $\Delta_{u,v} p(f_1, \ldots, f_k)$ for some quadratic polynomial p and polynomials $f_1, \ldots, f_k$, i.e. that every sum of squares polynomial has a hyperbolic certificate of nonnegativity in a certain sense. This is because it suffices to show that for each $k \geq 1$, the polynomial $\sum_{i=1}^k x_i^2$ can be written as $\Delta_{u,v} p$ for some hyperbolic polynomial. In this case, we may take*

$$p = 2zw + (z + w) \sum_{i=1}^n x_i + \sum_{i \neq j} x_i x_j.$$

This polynomial is stable, which can easily be checked because $p = \begin{pmatrix} z \\ w \\ x_1 \\ \dots \\ x_n \end{pmatrix} A \begin{pmatrix} z & w & x_1 & \dots & x_n \end{pmatrix}$, *where* A *is a matrix with nonnegative entries and exactly one positive eigenvector, and it is known that all such polynomials are stable [3]. Moreover, we have that* $\frac{\partial}{\partial z} p \frac{\partial}{\partial w} p - p \frac{\partial}{\partial w} \frac{\partial}{\partial z} p = \sum_{i=1}^{k} x_i^2$. *This construction is unsatisfying, because it does not give any insight as to how to find a sum of squares decomposition of* f.

While it is true that for any polynomial p hyperbolic with respect to v, the mixed derivative $\Delta_{v,v} p$ is nonnegative, the converse is not always true. However, it is noteworthy that there is a partial converse, which we will prove here for later reference.

Theorem 2.1.3. *Suppose that* $D_v p$ *is hyperbolic with respect to* v, *and also that* $\Delta_{v,v} p$ *is globally nonnegative. Also assume that there is some* x *so that the polynomials* $p_x(t)$ *and* $\frac{d}{dt} p_x(t)$ *are square-free (i.e. nonzero with no root of multiplicity greater than 1). Then* p *is hyperbolic with respect to* v.

Proof. We wish to show that for all x, $p_x(t)$ is real rooted. Firstly, because the property of being real rooted is closed in the set of univariate polynomials, it suffices to show that $p_x(t)$ is real rooted for a dense set of x. Because there is some x satisfying the condition that $p_x(t)$ and $\frac{d}{dt} p_x(t)$ are square-free, and the set of x failing this condition is Zariski closed, a dense set of x satisfies this condition.

We now note that for any fixed x,

$$\Delta_{v,v} p(x + tv) = ((D_v p)^2 - p D_v^2 p)(x + tv) = \left(\frac{d}{dt} p_x(t) \right)^2 - p_x(t) \frac{d^2}{dt^2} p_x(t).$$

Therefore, for each fixed x, the mixed derivative of $p_x(t)$ is nonnegative.

It remains to show that if $g(t)$ is a square-free univariate polynomial with a square

free derivative; $\frac{d}{dt}g(t)$ is real rooted, and the mixed derivative of g is nonnegative, then g is real rooted. For this, we note that

$$\left(\frac{d}{dt}g(t)\right)^2 - g(t)\frac{d^2}{dt^2}g(t) = -g(t)^2\frac{d}{dt}\left(\frac{\frac{d}{dt}g(t)}{g(t)}\right).$$

Consider the rational function $\frac{\frac{d}{dt}g(t)}{g(t)}$, and note that it vanishes at the $d-1$ roots of $\frac{d}{dt}g(t)$, since $g(t)$ and $\frac{d}{dt}g(t)$ have no common zeros. Also note that there are 2 additional 'zeros' of this function at ∞ and $-\infty$, in the sense that $\lim_{t\to\pm\infty}\frac{\frac{d}{dt}g(t)}{g(t)} = 0$ by considering the degree of this rational function. Let r and s be two consecutive zeros of $\frac{d}{dt}g(t)$, i.e. zeros of this polynomial so that there are no zeros in the interval (r,s). If $\frac{\frac{d}{dt}g(t)}{g(t)}$ were differentiable on the interval (r,s), then because $-g(t)^2\left(\frac{d}{dt}\frac{\frac{d}{dt}g(t)}{g(t)}\right) \geq 0$, $\frac{\frac{d}{dt}g(t)}{g(t)}$ would monotonic. However, a monotonic function on $[r,s]$ that vanishes at r and s would be identically zero, which is a contradiction. We can conclude that $\frac{\frac{d}{dt}g(t)}{g(t)}$ must have a pole in the interval $[r,s]$, implying that $g(t)$ must vanish in this interval.

There are d intervals of this form, so we conclude that $g(t)$ must have at d real roots, as desired. □

Remark 2. *Note that the example $g(t) = t^4 - 1$ has a real rooted derivative, and its mixed derivative is $4t^2(t^4+3) \geq 0$, but g has nonreal roots. This shows that the additional requirement that $\frac{d}{dt}p_x(t)$ be square-free for some x is required.*

2.1.3 Hyperbolicity preservers

We conclude our historical remarks on hyperbolic polynomials with some facts about hyperbolicity preservers. It follows from Rolle's theorem that if p is a real rooted univariate polynomial, then $\frac{d}{dt}p$ is real rooted. Similarly, if p is hyperbolic with respect to v, and $u \in \Lambda_v(p)$, then $D_u p$ is also hyperbolic with respect to v.

Motivated by this example, we consider the following definition of a hyperbolicity preserver:

Definition 2.1.4. *Let $U \subseteq \mathbb{R}[x_1, \ldots, x_n]$ and $V \subseteq \mathbb{R}[y_1, \ldots, y_m]$ be linear subspaces, and let $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$. We say that a linear map $T : U \to V$ is a hyperbolicity preserver if for every $p \in U$ that is hyperbolic with respect to u, $T(p)$ is hyperbolic with respect to v.*

We will also refer to univariate hyperbolicity preservers. It is not hard to see that if $g(t)$ is a univariate polynomial, then g is real rooted if and only if the homogenization of g, denoted here by $g^h(s,t)$, is hyperbolic with one of the two coordinate vectors. Similarly, if $\mathbb{R}[t]_n$ denotes the vector space of univariate polynomials of degree at most n, then a linear map $T : \mathbb{R}[t]_n \to \mathbb{R}[t]_d$ sends real rooted polynomials to real rooted polynomials if and only if $T^h : \mathbb{R}[s,t]_n \to \mathbb{R}[s,t]_d$ preserves hyperbolicity in a coordinate direction in the sense of the previous definition. For this reason, we will also say that the map $T : \mathbb{R}[t]_n \to \mathbb{R}[t]_m$ is a hyperbolicity preserver.

We will say that a linear map $T : \mathbb{R}[t]_n \to \mathbb{R}[t]_d$ is diagonal if there are constants γ_i so that $T(t^{n-i}) = \gamma_i t^{d-i}$ for each $i \leq \min\{d, n\}$. An important result of Schur and Pólya in [13] characterizes hyperbolicity preservers $T : \mathbb{R}[t]_n \to \mathbb{R}[t]_d$ which are diagonal.

Theorem 2.1.5. *Let $T : \mathbb{R}[t]_n \to \mathbb{R}[t]_d$ be a diagonal linear map. Then T is a hyperbolicity preserver if and only if $T((t-1)^n)$ has real roots, all with the same sign.*

We will also recount the theory of stability preservers developed by Borcea and Brändén in [6] for completeness. Fix some $\kappa \in \mathbb{N}^n$ and let $\mathbb{R}[x_1, \ldots, x_n]_\kappa$ denote the vector space of polynomials where the degree of x_i is at most κ_i for each i. If $\gamma \in \mathbb{N}^n$, we then say that a linear map $T : \mathbb{R}[x_1, \ldots, x_n]_\kappa \to \mathbb{R}[x_1, \ldots, x_n]_\gamma$ is a stability preserver if the image of every stable polynomial is also stable. We may extend T to a linear map $T : \mathbb{R}[x_1, \ldots, x_n, w_1, \ldots, w_n]_\kappa \to \mathbb{R}[x_1, \ldots, x_n]_\gamma$ by treating the w_i as constants, i.e. we let

$$T(w^\alpha x^\beta) = w^\alpha T(x^\beta).$$

Here, if $\alpha \in \mathbb{N}^n$, we denote by $w^\alpha = w_1^{\alpha_1} \cdot \cdots \cdot w_n^{\alpha_n}$ and $x^\alpha = x_1^{\alpha_1} \cdot \cdots \cdot x_n^{\alpha_n}$.

Theorem 2.1.6. *$T : \mathbb{R}[x_1, \ldots, x_n]_\kappa \to \mathbb{R}[x_1, \ldots, x_n]_\gamma$ is a stability preserver if and only if either the image of T is 1-dimensional and spanned by a stable polynomial, or*

$$T((z+w)^\kappa)$$

is stable.

2.2 Symmetric hyperbolic polynomials

A polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]$ is *symmetric* if it is invariant under permutations of its variables. A well known class of symmetric polynomials are the *elementary symmetric polynomials*, defined as

$$e_{n,k}(x) = \sum_{\substack{S \subseteq [n] \\ |S|=k}} \prod_{i \in S} x_i.$$

We will often suppress the dependence of $e_{n,k}$ on n when it is convenient. These polynomials generate the ring of symmetric polynomials in n variables, i.e. every symmetric polynomial in n variables can be written as $q(e_{n,1}(x), \ldots, e_{n,n}(x))$, where q is some polynomial in n variables.

The elementary symmetric polynomials are hyperbolic with respect to the vector of all one's, denoted by $\vec{1}$. We will generally be interested in *symmetric hyperbolic polynomials*, which are symmetric polynomials that are hyperbolic with respect to $\vec{1}$.

Definition 2.2.1. *A polynomial p is symmetric hyperbolic if it is symmetric and it is hyperbolic with respect to $\vec{1}$.*

Symmetry in this context has been exploited in the past to understand hyperbolicity cones, notably in [14], as well as in [15, 16]. The main goals of this section will be to characterize the symmetric hyperbolic polynomials of degree 3, and give connections

between symmetric hyperbolic polynomials and hyperbolicity preservers for more general degrees. The content of this section was originally proven in [17].

2.2.1 Hook-shaped polynomials and 0-sum hyperbolicity preservers

Definition 2.2.2. *We will let $\mathbb{R}[t]_{n,0}$ denote the subspace of $\mathbb{R}[t]_n$ consisting of polynomials where the coefficient of t^{n-1} is 0. If $T : \mathbb{R}[t]_{n,0} \to \mathbb{R}[t]_{d,0}$ is a linear map, then we say T is a 0-sum hyperbolicity preserver if T sends real rooted polynomials to real rooted polynomials. We say that T is diagonal if $T(t^{n-i}) = \gamma_i t^{d-i}$ whenever $d = 0, 2, \ldots, \min\{n, d\}$.*

We will say that a symmetric polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]$ is *hook-shaped* if it is of the form

$$p = \sum_{i=1}^{d} a_i e_1(x)^{d-i} e_i(x),$$

for some coefficients a_i.

Definition 2.2.3. *If p is a hook-shaped polynomial, then its* associated operator *is a map $T : \mathbb{R}[t]_{n,0} \to \mathbb{R}[t]_{d,0}$ defined as follows: if $g(t) \in \mathbb{R}_n[t]$ is a polynomial which factors as $a(t - r_1) \ldots (t - r_n)$ for $r_1, \ldots, r_n \in \mathbb{C}$, then we let $T(g)$ be a polynomial so that*

$$T(g) = ap(\vec{r} - t\vec{1}),$$

where $\vec{r}$ denotes a vector whose entries are $r_1, \ldots, r_n$. We then extend this definition to all elements of $\mathbb{R}[t]_{n,0}$ (including those of degree less than n) by continuity.

The fact that the T above in fact sends polynomials with real coefficients to polynomials with real coefficients follows from the fact that if a symmetric polynomial is evaluated at the roots of a polynomial with real coefficients, then the result is real. Perhaps more surprisingly, if p is hook-shaped, then T is in fact a diagonal linear map. We show this in the next theorem.

Theorem 2.2.4. *If p is a hook-shaped polynomial, then its associated operator is a linear map, and moreover, the map sending p to its associated operator is linear and invertible. p is symmetric hyperbolic if and only if $p(\vec{1}) \neq 0$ and its associated operator is a 0-sum hyperbolicity preserver.*

Proof. We can perform a direct computation to show that the associated operator is linear and diagonal. Fix a hook-shaped p of degree d in n variables. It will be convenient to write p in terms of the *elementary symmetric means*, defined as $\tilde{e}_i(x) = \frac{e_i(x)}{\binom{n}{i}}$. It is clear that because p is hook-shaped, then there exist coefficients $a_1, \ldots, a_d$ so that

$$p = \sum_{i=1}^{d} a_i \tilde{e}_1^{d-i} \tilde{e}_i(x).$$

Let $g(t) \in \mathbb{R}[t]_{n,0}$ be monic with roots $r_1, \ldots, r_n \in \mathbb{C}$. We may write $g(t) = \prod_{i=1}^{n}(t - r_i) = \sum_{i=0}^{n} \binom{n}{k} c_i t^{n-i}$, where $c_1 = 0$. In this case, we have that $c_i = (-1)^i \tilde{e}_i(r_1, \ldots, r_n)$. Now, the definition of associated operator gives that

$$T(g) = p(\vec{r} - \vec{1}t) = \sum_{i=1}^{d} a_i \tilde{e}_1(\vec{r} - \vec{1}t)^{d-i} \tilde{e}_i(\vec{r} - \vec{1}t).$$

It follows from a Taylor expanding $\tilde{e}_i(\vec{r} - \vec{1}t)$ in t that $\tilde{e}_i(\vec{r} - \vec{1}t) = \sum_{j=0}^{i} (-1)^{i-j} \binom{i}{j} \tilde{e}_j(\vec{r}) t^{i-j}$. In particular, because $\tilde{e}_1(\vec{r}) = 0$, we have that $\tilde{e}_1(\vec{r} - t\vec{1}) = -t$.

We may compute

$$\begin{aligned} T(g) &= \sum_{i=1}^{d} a_i \tilde{e}_1(\vec{r} - \vec{1}t)^{d-i} \tilde{e}_i(\vec{r} - \vec{1}t) \\ &= \sum_{i=1}^{d} a_i \sum_{j=0}^{i} (-1)^{d-j} \binom{i}{j} \tilde{e}_j(\vec{r}) t^{d-j} \\ &= \sum_{j=0}^{d} (-1)^{d-j} \left(\sum_{i=j}^{d} \binom{i}{j} a_i \right) \tilde{e}_j(\vec{r}) t^{d-j} \\ &= \sum_{j=0}^{d} \left(\sum_{i=j}^{d} (-1)^{d} \binom{i}{j} a_i \right) c_j t^{d-j} \end{aligned}$$

That is, $T(g) = \sum_{i=0}^{d} \gamma_j c_j t^{d-j}$, where $\gamma_j = \frac{1}{\binom{n}{j}} (-1)^d \sum_{i=j}^{d} (-1)^{d-j} \binom{i}{j} a_i$. This clearly implies that $T(g)$ is the linear map sending t^{d-j} to $\gamma_j t^{d-j}$, so that T is a diagonal linear map as desired. Moreover, the γ_j are the image of the a_j under an upper triangular linear map with nonzero diagonal entries, and so the linear map sending p to its associated operator is invertible.

We now want to show the equivalence of p being symmetric hyperbolic and $T(g)$ being a 0-sum hyperbolicity preserver. For this, note that if p is symmetric hyperbolic, then for any monic $g \in \mathbb{R}[t]_{n,0}$ with real roots $r_1, \ldots, r_n$, we have that the polynomial

$$T(g) = p(\vec{r} - \vec{1}t)$$

is real rooted, implying that T is a 0-sum hyperbolicity preserver.

On the other hand, if T is a 0-sum hyperbolicity preserver, then for any $x \in \mathbb{R}^n$, we have that

$$T((t - x_1)(t - x_2) \ldots (t - x_n)) = p(x - \vec{1}t)$$

is real rooted, showing that p is hyperbolic. □

2.2.2 Extendable linear maps

This theorem gives a bijective correspondence between symmetric hyperbolic polynomials which are hook-shaped and a certain class of linear hyperbolicity preservers. One can then apply Theorem 2.1.5 to obtain a recipe for constructing a hook-shaped symmetric hyperbolic polynomials of degree d from real rooted polynomials of degree d as follows: We may begin with any real rooted polynomial $g \in \mathbb{R}[t]_d$ with roots of the same sign, and then find the unique diagonal linear map $\hat{T} : \mathbb{R}[t]_n \to \mathbb{R}[t]_d$ so that $T((t+1)^n) = g$. Theorem 2.1.5 then implies that $\hat{T}$ is a hyperbolicity preserver, which then restricts to a 0-sum hyperbolicity preserver, $T : \mathbb{R}[t]_{n,0} \to \mathbb{R}[t]_{d,0}$. We may then find the unique hook-shaped polynomial whose associated operator is T, and this will be symmetric hyperbolic by Theorem 2.2.4.

Example 2.2.1. *We may take as an example the univariate polynomial with only nonnegative real roots*

$$g(t) = (t-1)(t-2)(t-3)(t-4) = t^4 - 10t^3 + 35t^2 - 50t + 24.$$

By Theorem 2.1.5, the unique diagonal map sending $(t-1)^4 = t^n - 4t^3 + 6t^2 - 4t + 1$ *to* $g(t)$ *is a hyperbolicity preserver.*

If we let T *be this diagonal map, then*

$$T((t-1)^3(t+3)) = t^4 - 35t^2 + 100t - 72.$$

We would therefore want to find a hook-shaped polynomial p so that

$$p\left(\begin{pmatrix}1\\1\\1\\-3\end{pmatrix} - t\begin{pmatrix}1\\1\\1\\1\end{pmatrix}\right) = t^4 - 35t^2 + 100t - 72.$$

Letting

$$p = a_1\tilde{e}_1^4 + a_2\tilde{e}_1^2\tilde{e}_2 + a_3\tilde{e}_1\tilde{e}_3 + a_4\tilde{e}_4,$$

we may solve for a_1, a_2, a_3, a_4 and find that

$$p = -6\tilde{e}_1^4 + 29\tilde{e}_1^2\tilde{e}_2 - 46\tilde{e}_1\tilde{e}_3 + 24\tilde{e}_4,$$

Our theorem then implies that p is symmetric hyperbolic.

A natural question is whether or not this recipe in fact yields all hook-shaped polynomials which are symmetric hyperbolic. The only step in this construction which is not obviously bijective is the step of restricting a hyperbolicity preserver to $\mathbb{R}[t]_{n,0}$. We will make the definition that a map $T : \mathbb{R}[t]_{n,0} \to \mathbb{R}[t]_{d,0}$ is *extendable* if there exists a hyperbolicity preserver $\hat{T} : \mathbb{R}[t]_n \to \mathbb{R}[t]_d$ that equals T on the subspace $\mathbb{R}[t]_{n,0}$.

We can give an equivalent condition for T to be extendable in terms of the value of $T((t+n-1)(t-1)^{n-1})$. For this, we need to consider a map $\delta_n : \mathbb{R}[t]_n \to \mathbb{R}[t]_n$, which is defined so that it is equal to the unique diagonal map sending $\delta_n((t-1)^n) = (t-n+1)(t-1)^{n-1}$.

Definition 2.2.5. *Let $\delta_n : \mathbb{R}[t]_n \to \mathbb{R}[t]_{n,0}$ be the diagonal linear map defined by*

$$\delta_n(t^{n-k}) = -(k-1)t^{n-k}$$

for all $k \in [n]$.

For any diagonal map $T : \mathbb{R}[t]_n \to \mathbb{R}[t]_d$, $T(\delta_n(g)) = \delta_d(T(g))$, since the coefficient of $\delta_n(t^{n-k})$ does not depend on n. For this reason, we will abbreviate δ_n as δ where this no ambiguity as to its domain.

Lemma 2.2.6. *A linear map $T : \mathbb{R}[t]_{n,0} \to \mathbb{R}[t]_{d,0}$ is extendable if and only if there exists some $g \in \mathbb{R}[t]_n$ with real roots of the same sign, so that $\delta(g) = T((t-n+1)(t-1)^{n-1})$.*

Proof. Suppose that $T((t+n-1)(t-1)^{n-1}) = \delta(g)$ for some g with real roots of the same sign. We claim that the unique diagonal map $\hat{T} = \mathbb{R}[t]_n \to \mathbb{R}[t]_d$ satisfying $\hat{T}((t-1)^n) = g$ extends T and is a hyperbolicity preserver. The fact that $\hat{T}$ is a hyperbolicity preserver follows from Theorem 2.1.5. The fact that $\hat{T}$ extends T follows because

$$\begin{aligned}\hat{T}((t+n-1)(t-1)^{n-1}) &= \hat{T}(\delta((t-1)^n)) \\ &= \delta(\hat{T}((t-1)^n)) \\ &= \delta(g) \\ &= T((t+n-1)(t-1)^{n-1}).\end{aligned}$$

Here, we make the observation that a diagonal map $T : \mathbb{R}[t]_{n,0} \to \mathbb{R}[t]_{d,0}$ is uniquely determined by the value of $T((t+n-1)(t-1)^{n-1})$.

On the other hand, if T is extendable by a map $\hat{T}$, then it follows from the same sequence of equalities that $\delta(\hat{T}((t-1)^n)) = T((t+n-1)(t-1)^{n-1})$, so the result follows from setting $g = \hat{T}((t-1)^n)$. □

Next, we will show that all 0-sum hyperbolicity preservers $T : \mathbb{R}[t]_{n,0} \to \mathbb{R}[t]_{3,0}$ are extendable.

Theorem 2.2.7. *For any $n \in \mathbb{N}$, every diagonal 0-sum hyperbolicity preserver $T : \mathbb{R}[t]_{n,0} \to \mathbb{R}[t]_{3,0}$ is extendable. Moreover, a diagonal linear map $T : \mathbb{R}[t]_{n,0} \to \mathbb{R}[t]_{3,0}$*

is extendable if and only if

$$T((t+n-1)(t-1)^{n-1})$$

is real rooted.

Proof. Note $(t+n-1)(t-1)^{n-1} \in \mathbb{R}[t]_{n,0}$, so if T is a 0-sum hyperbolicity preserver, then $T((t+n-1)(t-1)^{n-1})$ is real rooted.

Now, suppose that $q = T((t+n-1)(t-1)^{n-1})$ is real rooted. By Lemma 2.2.6, to show that T is extendable, it suffices to show that $q = \delta(g)$, for some g with real roots of the same sign.

We have therefore reduced the problem to showing that for any real rooted polynomial in $\mathbb{R}[t]_{3,0}$ is the image under δ of some polynomial with roots of the same sign.

To show this, we will use the facts that $\delta((t-\frac{1}{2})^3) = (t-\frac{1}{2})^2(t+1)$, $\delta(t^3) = t^3$, and $\delta((t-1)^2 t) = (t-1)(t+1)t$. We will define $\mathcal{H}_{3,+} \subseteq \mathbb{R}[t]$ to be the set of hyperbolic polynomials of degree at most 3 with nonnegative real roots of the same sign. Also note that the set of univariate polynomials with nonnegative real roots is path connected (as it is the image of $\mathbb{R}^3_+$ under a polynomial map), and in particular, there is a path $\tau : [0,1] \to \mathcal{H}_{3,+}$ so that $\tau(0) = (t-\frac{1}{2})^3$ and $\tau(1) = (t-1)(t+1)t$. Also note that without loss of generality, we may assume that τ does not intersect $\{at^3 - bt^2 : a, b \in \mathbb{R}\} \subseteq \mathcal{H}_{3,+}$, since this is a codimension 2 subset.

Next, let $\mathcal{H}_{3,0}$ denote the set of real rooted, degree 3 polynomials in $\mathbb{R}[t]_{3,0}$. We will need to define a continuous function $E : \mathcal{H}_{3,0} \setminus \{at^3 : a \in \mathbb{R}\} \to [0,1]$ as follows. If $g \in \mathcal{H}_{3,0} \setminus \{at^3 : a \in \mathbb{R}\}$, with roots $r_1 \geq r_2 \geq r_3$, then we let

$$E(g) = \min\{\left|\frac{r_1}{r_3}\right|, \left|\frac{r_3}{r_1}\right|\}.$$

Note that $E(g)$ is well defined for any $g \in \mathcal{H}_{3,0} \setminus \{at^3 : a \in \mathbb{R}\}$ since any such

polynomial must have a nonzero root, and since the sum $r_1 + r_2 + r_3 = 0$, this implies that it must have a positive and negative root. We also clearly have that $E(g) \in [0, 1]$, and it is continuous because the roots of a polynomial are continuous functions of the polynomial.

Note that if $g, q \in \mathcal{H}_{3,0}$, then $E(g) = E(q)$ if and only if there exist $a, b \in \mathbb{R}$ so that $ag(bt) = q(t)$ for all $t \in \mathbb{R}$.

Suppose now that there is some $g \in \mathcal{H}_{3,0}$ so that $g(t)$ is not in the image of δ. We see that because $g(t)$ is not in the image of δ, neither is $ag(bt)$ for any $a, b \neq 0$, the image of δ is invariant under these operations. We further see that $g(t)$ cannot be t^3 since that is in the image of δ.

Finally, we have reached a contradiction. As $E(\tau(0)) = 1$ and $E(\tau(1)) = 0$, so by the intermediate value theorem, there is some t so that $E(\tau(t)) = E(g)$, contradicting the fact that g is not in the image of δ. □

Using this, we can obtain a characterization of cubic symmetric hyperbolic polynomials.

Corollary 2.2.8. *Let b be a coordinate vector, and let p be a cubic symmetric polynomial. Then, p is symmetric hyperbolic if and only if $p(\vec{1}) \neq 0$ and $p(b - t\vec{1})$ has real roots.*

Proof. For this, note that p is symmetric hyperbolic if and only if its associated operator is a hyperbolicity preserver. In turn, if T is its associated operator, then T preserves hyperbolicity if and only if $T((t + n - 1)(t - 1)^{n-1})$ is real rooted. In

conclusion, p is symmetric hyperbolic if and only if

$$T((t+n-1)(t-1)^{n-1}) = p\left(\begin{pmatrix} -n+1 \\ 1 \\ \cdots \\ 1 \end{pmatrix} - t\vec{1}\right)$$
$$= -n^3 p\left(\begin{pmatrix} 1 \\ 0 \\ \cdots \\ 0 \end{pmatrix} - \frac{(t-1)}{n}\vec{1}\right)$$

has real roots, which is clearly equivalent to $p(b - \vec{1}t)$ having real roots. □

The paper [17] generalizes some parts of this argument and shows using a more complicated topological argument that in fact, all diagonal 0-sum hyperbolicity preservers $T : \mathbb{R}[t]_{n,0} \to \mathbb{R}[t]_{4,0}$ are extendable. This also leads to a characterization of the degree 4 hook-shaped symmetric hyperbolic polynomials, which we will not prove here.

Theorem 2.2.9. *For any $n \in \mathbb{N}$, every diagonal 0-sum hyperbolicity preserver $T : \mathbb{R}[t]_{n,0} \to \mathbb{R}[t]_{4,0}$ is extendable. Moreover, a diagonal linear map $T : \mathbb{R}[t]_{n,0} \to \mathbb{R}[t]_{4,0}$ is extendable if and only if*

$$T((t+n-1)(t-1)^{n-1})$$

has real roots, of which 3 are of the same sign.

Also, if p is a coordinate vector,

However, there is an example of a diagonal 0-sum hyperbolicity preserver $T : \mathbb{R}[t]_{n,0} \to \mathbb{R}[t]_{5,0}$ which is not extendable, which we discuss next.

Theorem 2.2.10. *Let*

$$p = 4500e_5 - 220e_1e_4 + 7e_1^2e_3 \in \mathbb{R}[x_1, \ldots, x_5]_5$$

Then, p is symmetric hyperbolic, but p's associated operator is not extendable.

Proof. We will note that the associated operator of p satisfies

$$T((t+4)(t-1)^2) = -750(t+6)(t-1)^2(t-2)^2.$$

To show that p is symmetric hyperbolic, we will use Theorem 2.1.3.

To see that $D_{\bar{1}}p$ is symmetric hyperbolic, we note that the associated operator of $D_{\bar{1}}p$ sends a univariate polynomial g to $\frac{d}{dt}T(g)$, so it is easy to check that $D_{\bar{1}}p$ satisfies the conditions of Theorem 2.2.9.

We next check that $\Delta_{\bar{1}\bar{1}}p$ is nonnegative. In fact, using computational methods, it is possible to show that this polynomial is in fact a sum of squares, immediately implying nonnegativity.

Finally, we note that $p_x(t)$ and $\frac{d}{dt}p_x(t)$ are square free when $x = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$, so that we may apply Theorem 2.1.3.

On the other hand, to see that T is not extendable, we use the following fact from [17]:

Theorem 2.2.11. *Let $g \in \mathbb{R}[t]$ be univariate. If $\delta(g)$ has a root of multiplicity $k > 1$ at $r > 0$, then g must have a root of multiplicity $k+1$ at r.*

Therefore, if $-750(t+6)(t-1)^2(t-2)^2 = \delta(g)$ for some g with real roots, then g must vanish at 1 and 2 with multiplicity 3, but this is a contradiction as g is a degree

5 polynomial. □

Remark 3. *This polynomial p has some rather remarkable properties, for example, it is not SOS-hyperbolic, as can be seen by computing $\Delta_{\vec{1},u}p$, where $u = (6, 1, 1, 1, 1)$. Moreover, it can be shown that if p is written in the $\tilde{e}_i$ basis, then in fact, the hyperbolicity of p is independent of the number of variables.*

2.3 Linear principal minor polynomials

A family of polynomials which are closely related to the elementary symmetric polynomials are what we will refer to as the *characteristic coefficients*. These are sometimes also called the k-determinants. For a symmetric matrix X, these are defined by

$$c_{n,k}(X) = \sum_{\substack{S\subseteq[n]\\|S|=k}} \det(X|_S), \tag{2.3.1}$$

where the sum runs over all principal submatrices of X.

Such characteristic coefficients are directly related to the elementary symmetric polynomials, in the sense that

$$c_{n,k}(X) = e_{n,k}(\lambda_1(X), \ldots, \lambda_n(X)),$$

where the $\lambda_i(X)$ are the eigenvalues of X.

As a consequence of the fact that $e_{n,k}$ is hyperbolic with respect to $\vec{1}$, $c_{n,k}$ is hyperbolic with respect to I, the identity matrix. In fact, $c_{n,k}$ is hyperbolic with respect to all matrices in the PSD cone. Motivated by this, we make the following definition:

Definition 2.3.1. *A polynomial $p \in \mathbb{R}[X_{ij} : i \leq j \leq n]$ is PSD stable if it is hyperbolic with respect to I and $\Lambda_I(p)$ contains the PSD cone.*

Amongst the set of stable polynomials are those which are *multiaffine*, meaning that the polynomial is of degree at most 1 in each variable. Such multiaffine stable polynomials arise as generating functions in combinatorics. A homogeneous multiaffine stable polynomial of degree d can be written in the form

$$p(x) = \sum_{\substack{S \subseteq [n] \\ |S|=d}} a_S \prod_{i \in S} x_i.$$

Comparing the representation of the elementary symmetric polynomial as a sum of monomials to the representation of $c_{n,k}$ as a sum of principal minors, we are motivated to consider polynomials which are generally of the form

$$P(x) = \sum_{\substack{S \subseteq [n] \\ |S|=d}} a_S \det(X|_S).$$

We will refer to these as linear principal minor polynomials (LPM for short).

Formally, there is a linear map Φ from the linear subspace of $\mathbb{R}[x_1, \ldots, x_n]$ spanned by square free monomials to the linear subspace of $\mathbb{R}[X_{i,j} : i \leq j \leq n]$ spanned by principal minors of X sending $\prod_{i \in S} x_i$ to $\det(X|_S)$. Throughout, we will use lower case letters to denote multiaffine polynomials and upper case letters to denote the corresponding LPM polynomial, so that for example, if p is a multiaffine polynomial, then P is $\Phi(p)$. Note that Φ has an inverse map, since $p = P(\mathrm{Diag}(x_1, \ldots, x_n))$.

Given the example of the characteristic coefficients, it is interesting to ask when LPM polynomials are PSD stable. We may characterize these polynomials naturally.

Theorem 2.3.2. *An LPM polynomial P is PSD stable if and only if its corresponding multiaffine polynomial p is stable.*

One direction of this theorem is clear; if P is PSD stable, then $p = P(\mathrm{Diag}(x_1, \ldots, x_n))$ is stable because $p(x + tv) = P(\mathrm{Diag}(x) + t\mathrm{Diag}(v))$, and the right side is real rooted

for all x and v in the positive orthant. Thus, we are mostly interested in showing that if p is stable, then P is PSD stable.

Before we go on, we need the following fact about hyperbolicity cones that can be found in [3].

Lemma 2.3.3. *Let $p \in \mathbb{R}[x_1, \ldots, x_n]$ be a homogeneous polynomial and $K \subset \mathbb{R}^n$ a cone. The following are equivalent:*

1. *p is hyperbolic with respect to all $a \in K$, and*
2. *$p(v + ia) \neq 0$ for all $v \in \mathbb{R}^n$ and $a \in K$.*

We now make an observation

Lemma 2.3.4. *Let $P \in \mathbb{R}[X_{ij} : i \leq j \leq n]$ be a homogeneous polynomial. Then P is PSD-stable if and only if the following two conditions hold:*

1. *$P(A) \neq 0$ for all positive definite matrices A;*
2. *$P(Diag(x_1, \ldots, x_n) + M) \in \mathbb{R}[x_1, \ldots, x_n]$ is stable for every real symmetric matrix M.*

Proof. First assume that P is PSD-stable and let A be a positive definite matrix. Because A is in the interior of $\Lambda_I(P)$, we have that $P(A) \neq 0$. Also, if for some M, $P(\text{Diag}(x_1, \ldots, x_n) + M)$ is not stable, then there is also some $\zeta \in \mathbb{C}^n$ so that for each i, the imaginary part of ζ_i is positive, and

$$P(\text{Diag}(\zeta_1, \ldots, \zeta_n) + M) = 0.$$

However, consider the univariate polynomial

$$P((\text{Re}(\text{Diag}(\zeta_1, \ldots, \zeta_n)) + M) + t\text{Im } (\text{Diag}(\zeta_1, \ldots, \zeta_n))).$$

This clearly has a root when $t = i$, so that P is not hyperbolic with respect to Im $(\mathrm{Diag}(\zeta_1, \ldots, \zeta_n)))$. This is a contradiction, as Im $(\mathrm{Diag}(\zeta_1, \ldots, \zeta_n)))$ is positive definite, and therefore in the hyperbolicity cone of P.

For the other direction we first observe that condition (2) implies that P is hyperbolic with respect to the identity matrix. We then see that because $P(A) \neq 0$ for $A \succ 0$, the connected component of $\mathbb{R}^{n\times n}_{sym} \setminus \mathcal{V}(P)$ containing I must contain all positive definite matrices, so that $\Lambda_I(P)$ contains the positive definite cone. □

Proof of Theorem 2.3.2. Let $p \in \mathbb{R}[x]$ be multiaffine, homogeneous and stable. Then by [18, Thm. 6.1] all nonzero coefficients of p have the same sign. Without loss of generality assume that all are positive. Then $P = \Phi(p)$ is clearly positive on positive definite matrices since the minors of a positive definite matrix are positive. Thus by Lemma 2.3.4, it remains to show that

$$P(\mathrm{Diag}(x_1, \ldots, x_n) + M) = \left(p^* \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_n}\right)\right) \det(\mathrm{Diag}(x_1, \ldots, x_n) + M)$$

is stable for every real symmetric matrix M. The polynomial $\det(\mathrm{Diag}(x_1, \ldots, x_n)+M)$ is stable as well as p^* by [18, Prop. 4.2]. Thus the polynomial $P(\mathrm{Diag}(x_1, \ldots, x_n)+M)$ is also stable by [19, Thm. 1.3]. □

The consequences of this result are explored in greater detail in [20]. In particular, various analogues of the Hadamard-Fischer theorem are given relating the values of P and p within their hyperbolicity cones.

We will highlight one consequence of this result here: a particular cubic polynomial p in 6 variables which is hyperbolic with respect to some vector v, but with the property that $\Delta_{uv}p$ is not a sum of squares for some $u, v \in \Lambda_v(p)$.

Remark 4. *In [12] Saunderson constructs a hyperbolic cubic in* 43 *variables whose Bézout matrix is not a matrix sum of squares, and noted that at the time, it was not known whether or not there existed such cubic polynomials with fewer variables. The*

polynomial we construct below in particular has this property, as $\Delta_{uv}p$ is one of the diagonal entries of the Bézout matrix of p. For all higher degrees, it is known precisely for which n all hyperbolic polynomials of that degree with n variables have a Bézout matrix that is not a matrix sum of squares.

Consider the complete graph K_4 on 4 vertices. We define the spanning tree polynomial of K_4 as the element of $\mathbb{R}[x_e : e \in E(K_4)]$ given by

$$t_{K_4}(x) \;=\; \sum_{\tau} \prod_{e \in \tau} x_e \,,$$

where $\tau \subset E(K_4)$ ranges over all edge sets of spanning trees of K_4. The polynomial t_{K_4} is multiaffine, homogeneous and stable [18, Thm. 1.1]. Let T be its corresponding LPM polynomial. Finally, let p be the polynomial obtained from T by evaluating T at the matrix of indeterminants

$$A = \begin{array}{c} \begin{array}{cccccc} 12 & 13 & 14 & 23 & 24 & 34 \end{array} \\ \begin{pmatrix} x_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & x_2 & a & b & c & 0 \\ 0 & a & x_2 & c & b & 0 \\ 0 & b & c & x_2 & a & 0 \\ 0 & c & b & a & x_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & x_3 \end{pmatrix} \end{array}.$$

Thus p is hyperbolic with respect to every positive definite matrix that can be obtained by specializing entries of A to some real numbers. In particular, the polynomial

$$W = \frac{\partial p}{\partial x_1} \cdot \frac{\partial p}{\partial x_3} - p \cdot \frac{\partial^2 p}{\partial x_1 \partial x_3}$$

is nonnegative. We will show that it is not a sum of squares.

Theorem 2.3.5. *The polynomial W is not a sum of squares.*

Proof. Explicitly,

$$\frac{1}{4}W = a^2b^2 + a^2c^2 + b^2c^2 + c^4 - 8abcx_2 + 2a^2x_2^2 + 2b^2x_2^2$$

We first note that if W were a sum of squares, then it is the sum of squares of quadratic forms. Indeed, by examining the Newton polytope of W, we see that if W were a sum of squares, then it would necessarily be a sum of squares of polynomials in the linear subspace

$$\text{span}\{ab, ac, ax_2, bc, bx_2, c^2\}.$$

The idea of considering the Newton polytope in finding such sum-of-squares decompositions was first discussed in [21].

W can be written as a sum of squares from elements in this subspace if and only if there is a PSD matrix A so that

$$W = v^\intercal A v, \tag{2.3.2}$$

where

$$v = \begin{pmatrix} ab \\ ac \\ ax_2 \\ bc \\ bx_2 \\ c^2 \end{pmatrix}.$$

Suppose that such an A existed. Expanding out Equation (2.3.2) in terms of the

entries of A, we obtain that A must be of the following form:

$$\begin{pmatrix} 1 & A_{ab,ac} & A_{ab,ax_2} & A_{ab,bc} & A_{ab,bx_2} & A_{ab,c^2} \\ A_{ab,ac} & 1 & A_{ac,ax_2} & A_{ac,bc} & A_{ac,bx_2} & A_{ac,c^2} \\ A_{ab,ax_2} & A_{ac,ax_2} & 2 & A_{ax_2,bc} & A_{ax_2,bx_2} & A_{ax_2,c^2} \\ A_{ab,bc} & A_{ac,bc} & A_{ax_2,bc} & 1 & A_{bc,bx_2} & A_{bc,c^2} \\ A_{ab,bx_2} & A_{ac,bx_2} & A_{ax_2,bx_2} & A_{bc,bx_2} & 2 & A_{bx_2,c^2} \\ A_{ab,c^2} & A_{ac,c^2} & A_{ax_2,c^2} & A_{bc,c^2} & A_{bx_2,c^2} & 1 \end{pmatrix},$$

and also satisfy the property that $A_{ax_2,bc} + A_{ac,bx_2} = -4$.

Here, we index the entries of A by the pair of monomials corresponding to that entry of A.

Consider now the matrix

$$B = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 12 & 0 & 0 & 9 & 0 \\ 0 & 0 & 8 & 9 & 0 & 0 \\ 0 & 0 & 9 & 12 & 0 & 0 \\ 0 & 9 & 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

This matrix is positive definite, and also satisfies the property that for any A of the above form, satisfying $A_{ax_2,bc} + A_{ac,bx_2} = -4$,

$$\mathrm{Tr}(AB) = -10.$$

This is negative, contradicting the fact that A was positive semidefinite. This implies that W is not a sum-of-squares. □

Remark 5. *The matrix B that certified that W was not a sum-of-squares can be found-*

ing using general semidefinite programming techniques. We used the SumOfSquares.jl Julia package [22, 23] for this problem.

Remark 6. *In the terminology of [12] this shows in particular that h is neither SOS-hyperbolic nor weakly SOS-hyperbolic.*

CHAPTER 3
SPARSITY IN SEMIDEFINITE PROGRAMMING

Many problems in optimization have a natural 'sparse' structure, meaning that either many of the entries of the input or the expected output are zero. In combinatorics for example, many inputs are graphs which have much smaller than the maximum number of edges possible for their number of vertices. Optimization problems associated to these graphs can inherit this sparse structure, as in the case of the Goemans-Williamson relaxation of the MAX-CUT problem which we discuss in more detail below. We may want to solve such sparse problems more quickly or with less memory than dense instances of the same problems.

A different situation where such sparse structure appears naturally is in sparse linear regression, known in the literature as the subset selection problem. This problem has as input a matrix $A \in \mathbb{R}^{m \times n}$, a vector $b \in \mathbb{R}^m$, and a sparsity parameter $k \in \mathbb{N}$, and asks to find

$$\min\{\|Ax - b\|_2^2 : \|x\|_0 \leq k\}. \tag{3.0.1}$$

Here, $\|x\|_0$ is the number of nonzero entries in x. This problem is of fundamental interest in data science and optimization, but is very difficult to solve with rigorous guarantees. This problem is known to be NP-hard even to approximate [24, 25] in general, so typically guarantees are more often sought after in special settings. Other problems involve finding solutions to optimization problems which have few nonzero entries.

To distinguish between the two types of sparse structure above, we may refer to these two situations as having 'input sparsity' and 'output sparsity' respectively. The goal of this section is to discuss an interesting structural framework that captures

aspects of these questions in the context of semidefinite programming. Section 3.1 defines various convex cones which are related to these questions of sparse semidefinite programming, and motivates how an understanding of these cones can lead to practical implications for problems possessing both input and output sparsity. Section 3.2 derives quantitative bounds on approximations to certain semidefinite programs with input sparsity. Finally, Section 3.3 connects these sparse semidefinite programming questions to hyperbolicity cones and problems with output sparsity.

3.1 Preliminary notions

Fix some collection of subsets $\Delta \subseteq [n]$, and consider the set of Δ-sparse vectors in $\mathbb{R}^n$,

$$\mathcal{X}(\Delta) = \{x \in \mathbb{R}^n : \text{supp}(x) \in \Delta\}.$$

Here $\text{supp}(x) = \{i \in [n] : x_i \neq 0\}$. It is natural to assume that Δ has the property that if $S \in \Delta$, then for any $T \subseteq S$, $T \in \Delta$, as this implies that $\mathcal{X}(\Delta)$ is closed. Indeed, for a simplicial complex Δ, the set $\mathcal{X}(\Delta)$ is a well known type of algebraic variety, known as a Stanley-Reisner variety. The connections between the algebraic structure of such Stanley-Reisner varieties and the topological properties of Δ are well known and are discussed, for example in [26, Chapter 1].

Given a simplicial complex Δ, we may define the convex cone

$$\mathcal{M}(\Delta) = \text{conv}\{xx^\intercal \in \mathbb{R}^{n\times n}_{sym} : x \in \mathcal{X}(\Delta)\}.$$

Let $G(\Delta) = \{S \in \Delta : |S| \leq 2\}$, which we may regard as a graph where the one element sets are vertices and the two element sets are edges. If we let $G = G(\Delta)$, then $\mathcal{M}(\Delta)$ is a full dimensional cone inside of

$$\mathbb{R}^G = \{X \in \mathbb{R}^{n\times n}_{sym} : X_{i,j} = 0 \text{ if } \{i,j\} \notin G\}.$$

The case in which $\Delta = \binom{[n]}{k}$, the set of all k-element subsets of $[n]$, is especially well studied. In this case, $\mathcal{M}(\Delta)$ is known as the *factor-width* k cone, which we denote by $\mathcal{F}^{n,k}$.

We can express sparse quadratically constrained quadratic programs in terms of the cone $\mathcal{M}(\Delta)$, as in the next theorem.

Theorem 3.1.1. *The following two optimization problems have the same optimal values, assuming A_1 is PSD.*

$$\begin{array}{lll} \min & x^\intercal A_0 x & \\ s.t. & x^\intercal A_1 x = 1 & \\ & \operatorname{supp}(x) \in \Delta & \end{array} \qquad (3.1.1) \qquad \begin{array}{lll} \min & \operatorname{Tr}(A_0 X) & \\ s.t. & \operatorname{Tr}(A_1 X) = 1 & \\ & X \in \mathcal{M}(\Delta) & \end{array} \qquad (3.1.2)$$

One optimization problem captured in this framework is the sparse PCA problem, defined for a given symmetric matrix A as

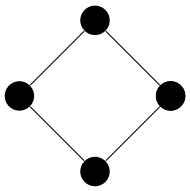
$$\begin{array}{lll} \min & x^\intercal A x & \\ \text{s.t.} & x^\intercal x = 1 & \\ & \|x\|_0 \leq k. & \end{array} \qquad (3.1.3)$$

The sparse linear regression problem defined above can also be captured in this framework for a given A and b as

$$\begin{array}{lll} \min & \operatorname{Tr}(A^\intercal b b^\intercal A X) & \\ \text{s.t.} & \operatorname{Tr}(A^\intercal A X) = 1 & \\ & X \in \mathcal{M}(\Delta). & \end{array} \qquad (3.1.4)$$

3.1.1 Nonnegative quadratic forms and sparse semidefinite programming

There is a dual vector space to $\mathbb{R}^G$, R^{G*} which we may think of as being as being 'partial matrices', i.e. matrices where the entries which correspond to nonedges of $G(\Delta)$ have been forgotten. To be more specific, a given $X \in \mathbb{R}^{G*}$, assigns a value X_{ij} for every $\{i,j\} \in \Delta$, but does not assign a value to other entries. Equivalently, we can think of an element of $\mathbb{R}^{G(\Delta)*}$ as being a quadratic form on $\mathcal{X}(\Delta)$, in the sense that if $x \in \mathcal{X}(\Delta)$, then we may define $x^\intercal X x = \sum_{i,j \in G(\Delta)} X_{ij} x_i x_j$. We may think of an element $S \in \Delta$ as indexing a submatrix of X where all of the entries are specified, which we denote $X|_S$.



$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & ? & 0 \\ 0 & 1 & 0 & ? \\ ? & 0 & 1 & 0 \\ 0 & ? & 0 & 1 \end{pmatrix}$$

Figure 3.1: An example of the projection of a matrix onto the edges of a cycle graph. This image was originally shown in [27].

Example 3.1.1.

There is a natural dual cone to $\mathcal{M}(\Delta)$, which is also a full dimensional cone inside $\mathbb{R}^{G(\Delta)*}$. We will denote by $\mathcal{P}(\Delta)$ this dual cone, and we may think of elements of $\mathcal{P}(\Delta)$ as being nonnegative quadratic forms on $\mathcal{X}(\Delta)$. Equivalently, a partial matrix X is in $\mathcal{P}(\Delta)$ if and only if for each $S \in \Delta$, $X|_S \succeq 0$.

Inside of $\mathcal{P}(\Delta)$ is another natural cone, the cone of *PSD-completable* matrices, which we denote by $\Sigma(G(\Delta))$. That is,

$$\Sigma(G) = \{X \in \mathbb{R}^G : \exists \hat{X} \succeq 0 \text{ where } \forall i,j \in G,\ X_{i,j} = \hat{X}_{i,j}\}.$$

These correspond to quadratic forms on $X(\Delta)$ which are sums of squares of linear forms on $X(\Delta)$.

It is natural to ask when $\Sigma(G(\Delta)) = \mathcal{P}(\Delta)$. This question was answered in different terms in [28], when the complexes are restricted to being *clique complexes*. The clique complex of a graph G, $\chi(G)$, is the set of all subsets of the vertices of G which induce a clique in G. [1] For clique complexes, the results in [28] showed that if G is a graph, then $\Sigma(G) = \mathcal{P}(\chi(G))$ if and only if G is *chordal* in the sense that G contains no induced cycles of length greater than 3. In [29], it was shown that this is a consequence of a more general fact about the differences between sums of squares and nonnegative quadratic forms on algebraic varieties using a result of [30]. As a consequence of this more general fact, it can easily be shown that in fact for any simplicial complex, $\Sigma(G(\Delta)) = \mathcal{P}(\Delta)$ if and only if Δ is the clique complex of a graph.

The fact that $\Sigma(G(\Delta)) = \mathcal{P}(\Delta)$ when Δ is the clique complex of a chordal graph is in fact extensively used in the study of semidefinite programs with sparse inputs. Given a semidefinite program

$$\begin{aligned} \min \quad & \mathrm{Tr}(A_0^\intercal X) \\ \text{s.t.} \quad & \mathrm{Tr}(A_i^\intercal X) = b_i \text{ for } i = 1, \ldots, m \\ & X \succeq 0, \end{aligned} \tag{3.1.5}$$

we may define the *joint sparsity* of this semidefinite program to be the smallest graph on n vertices so that $A_i \in \mathbb{R}^G$ for each $i = 0, \ldots, m$. In this case, rather than optimizing over the set of all positive semidefinite matrices, we can instead optimize over $\Sigma(G)$. When G is chordal, the fact that $\Sigma(G) = \mathcal{P}(\chi(G))$ implies that indeed we can equivalently optimize over $\mathcal{P}(\chi(G))$. To be concrete about the advantages of this replacement, we note that a partial matrix X is in $\mathcal{P}(\Delta)$ if and only if for

[1] In category theoretic terms, this is a right adjoint to the functor sending a simplicial complex to its underlying graph.

every maximal element $S \in \Delta$, the fully specified submatrix $X|_S$ is PSD. When G is a chordal graph, the number of maximal cliques of G is at most linear in the number of vertices, and so it is possible to check if a partial matrix is in $\mathcal{P}(\chi(G))$ in time $O(n\omega(G)^3)$, where $\omega(G)$ denotes the size of the largest clique of G. In cases in which $\omega(G)$ is much smaller than n, this can make algorithms for semidefinite programming much more efficient.

An example of a type of semidefinite program with joint sparsity are the *Goemans-Williamson relaxations* for the MAX-CUT problem, first given in [31]. The MAX-CUT problem is an optimization problem which takes in as input a graph G, and asks for a set of vertices of G, S, so that the number of edges with one vertex in S and the other in S^c is as large as possible. This problem is NP-hard, but by making S a uniformly random subset, it is easy to obtain a $\frac{1}{2}$-factor approximation. Goemans and Williamson in their breakthrough work show that the following semidefinite program achieves an approximation factor of at least 0.878.

$$\begin{aligned} \max \quad & \sum_{\{i,j\} \in G} \frac{1 - X_{ij}}{2} \\ \text{s.t.} \quad & X_{ii} = 1 \text{ for } i = 1, \ldots, n \\ & X \succeq 0. \end{aligned} \tag{3.1.6}$$

It is clear that in this case, the joint sparsity of this problem is G. We might therefore be interested in getting faster algorithms in cases where G is sparse than when G is dense. Indeed, various authors [32, 33] show that it is practically useful to take G to be a somewhat larger chordal graph and to instead optimize over $\mathcal{P}(\chi(G))$.

We will describe how we can extend this algorithmic value beyond the chordal case using quantitative analysis of the difference between $\mathcal{P}(\Delta)$ and $\Sigma(G(\Delta))$.

3.2 Approximate positive semidefinite completions

The results in this section were first described in [34].

It is desirable to extend the theory that allows us to work with chordal graphs to more general settings. Some noteworthy work in this direction are [35, 36], which show that if Δ is the clique complex of a series parallel graph G, a partial matrix $X \in \mathcal{P}(\Delta)$ is in $\Sigma(G)$ if and only if for every cycle C contained in G, the partial submatrix $X|_C$ is in $\Sigma(C)$.

In the cases in which $\Sigma(G(\Delta)) \neq \mathcal{P}(\Delta)$, it is natural to ask *how different are these two cones?* From an optimization perspective, there may be cases of inequality where it still may be computationally preferable to optimize over $\mathcal{P}(\Delta)$ instead of $\Sigma(G(\Delta))$, for example the number of elements of Δ and the sizes of the elements of Δ are small relative to n. Therefore, we might ask how much we stand to lose by replacing $\Sigma(G(\Delta))$ by $\mathcal{P}(\Delta)$ in an optimization problem.

To measure this gap, we first say that for a given element $X \in \mathbb{R}^G$, the minimum eigenvalue of X is $\lambda_{min}(X) = -\min\{\varepsilon \in \mathbb{R} : X + \varepsilon I \in \Sigma(G)\}$, where I is the image of the identity matrix inside of $\mathbb{R}^G$. This definition coincides with the minimum eigenvalue of a matrix X in the cases where G is the complete graph, and clearly, $\lambda_{min}(X) \geq 0$ if and only if $X \in \Sigma(G)$. We then define the *gap* of a simplicial complex as follows: let $\overline{\mathcal{P}(\Delta)} = \{X \in \mathcal{P}(\Delta) : \mathrm{Tr}(X) = 1\}$.

$$\varepsilon(\Delta) = -\min\{\lambda_{min}(X) : \forall X \in \overline{\mathcal{P}(\Delta)}\}.$$

We have that $\varepsilon(\Delta) \geq 0$ with equality if and only if Δ is the clique complex of a chordal graph.

We can also define this quantity in terms of $X \in \mathcal{M}(\Delta)$:

Theorem 3.2.1. *For any simplicial complex Δ,*

$$\varepsilon(\Delta) = \max_{\substack{X \in \mathbb{R}^{G(\Delta)} \\ X \succeq 0 \\ \mathrm{Tr}(X)=1}} \max\{\epsilon : X + \epsilon I \in \mathcal{M}(\Delta)\}.$$

Proof. Suppose there is some $X \in \mathbb{R}^{G(\Delta)}$ such that $X \succeq 0$ and $\mathrm{Tr}(X) = 1$, and so that $X + \epsilon(\Delta)I \notin \mathcal{M}(\Delta)$. Then by convex duality, there must be some Y so that $Y \in \mathcal{P}(\Delta)$ so that $\langle Y, X + \epsilon(\Delta)I\rangle < 0$, and we may take Y to have trace 1. We then have that

$$\langle Y, X + \epsilon(\Delta)I\rangle = \langle Y, X\rangle + \epsilon(\Delta)\mathrm{Tr}(Y) = \langle Y + \epsilon(\Delta)I, X\rangle < 0.$$

However, this is a contradiction, as this implies that $Y + \epsilon(\Delta)I$ cannot have a PSD completion, which is a contradiction of the definition of $\epsilon(\Delta)$.

Applying this sequence of equations in reverse gives the other direction. □

If $\varepsilon(\Delta)$ is small, then we can optimize over $\mathcal{P}(\Delta)$ as a substitute for optimizing over $\Sigma(\Delta)$ without too much difficulty. To formalize this, we will say that a semidefinite program with a maximization objective is of Goemans-Williamson type if it has joint sparsity G, the identity matrix is feasible with nonnegative objective value, and if every feasible point has trace n, so that for example, the Goemans-Williamson relaxation of MAX-CUT is of Goemans-Williamson type.

Theorem 3.2.2. *Let α denote the value of a Goemans-Williamson type SDP, and let α' be the value of the SDP obtained by replacing $\Sigma(G)$ by $\mathcal{P}(\Delta)$, for some Δ where $G(\Delta)$ contains G. Then*

$$\frac{1}{1 + \varepsilon(\Delta)n}\alpha' \leq \alpha \leq \alpha'.$$

Proof. To be explicit, the Goemans-Williamson type SDP is of the form

$$\begin{aligned} & \max && \mathrm{Tr}(A_0 X) \\ \alpha = & \text{ s.t.} && \mathrm{Tr}(A_i X) = b_i \text{ for } i = 1, \ldots, m \\ & && X \succeq 0. \end{aligned} \tag{3.2.1}$$

where each $A_i \in \mathbb{R}^G$. We can equivalently write this as

$$\begin{aligned} & \max && \mathrm{Tr}(A_0 X) \\ \alpha = & \text{ s.t.} && \mathrm{Tr}(A_i X) = b_i \text{ for } i = 1, \ldots, m \\ & && X \in \Sigma(G). \end{aligned} \tag{3.2.2}$$

If $G(\Delta)$ contains G, then we may relax this to

$$\begin{aligned} & \max && \mathrm{Tr}(A_0 X) \\ \alpha' = & \text{ s.t.} && \mathrm{Tr}(A_i X) = b_i \text{ for } i = 1, \ldots, m \\ & && X \in \mathcal{P}(\Delta). \end{aligned} \tag{3.2.3}$$

Because $\mathcal{P}(\Delta) \supseteq \Sigma(G)$, we have that $\alpha' \geq \alpha$. On the other hand, if we let X^* maximize Equation (3.2.3), then by definition of $\varepsilon(\Delta)$, we have that

$$X^* + n\varepsilon(\Delta) I \in \Sigma(G).$$

Because I and X^* satisfy the linear constraints, so does $\frac{1}{1+n\varepsilon(\Delta)}(X^* + n\varepsilon(\Delta)I)$, so that $\frac{1}{1+n\varepsilon(\Delta)}(X^* + n\varepsilon(\Delta)I)$ is feasible for the original problem. Therefore, because we have found a feasible point, we must have that

$$\alpha \geq \mathrm{Tr}\left(\frac{A_0}{1+n\varepsilon(\Delta)}(X^* + n\varepsilon(\Delta)I)\right) \geq \frac{1}{1+n\varepsilon(\Delta)}\mathrm{Tr}(A_0 X^*) = \frac{1}{1+n\varepsilon(\Delta)}\alpha'.$$

Here, we have used the fact that the objective value of I is nonnegative. $\square$

Here, we will show some results about this quantity for specific complexes, specifically cycles and the clique complexes of series parallel graphs. More results about this quantity can be found in [27, 34].

3.2.1 Gaps of cycles

We begin by describing how to compute $\varepsilon(C_n)$, where C_n is a cycle with n vertices, i.e. as a simplicial complex,

$$C_n = \{\{i\} : i \in [n]\} \cup \{\{i, (i+1 \mod n)\} : i \in [n]\}.$$

Theorem 3.2.3. *For $n \geq 3$, $\varepsilon(C_n) = \frac{1}{n}\left(\frac{1}{\cos(\frac{\pi}{n})} - 1\right)$.*

We will require some lemmas before proving this theorem.

Firstly, we say that a given $X \in \mathcal{P}(\Delta)$ is *locally rank 1* if for every $S \in \Delta$, $X|_S$ is rank at most 1. We also say that a complex Δ is locally rank 1 if every extreme ray of $\mathcal{P}(\Delta)$ is locally rank 1.

Lemma 3.2.4. *Every graph (in the sense of being a simplicial complex where every element has size at most 2) is locally rank 1.*

Proof. Let G be such a graph, and let X be an extreme ray of $\mathcal{P}(G)$. Suppose that for some $S = \{i, j\} \in G$, $X|_S$ has rank greater than 1. Because every element of C_n has size at most 2, we have that $X|_S$ would then have rank 2 and be positive definite.

We claim that if $D \in \mathbb{R}^G$ has the property that $D_{ij} = D_{ji} = \delta > 0$ sufficiently small, and $D_{k\ell} = 0$ for all other $\{k, \ell\} \in G$, then

$$X \pm D \in \mathcal{P}(G).$$

To see this, note that if $S \in G$ but $S \neq \{i, j\}$, then $(X \pm D)|_S = X|_S$. So we can conclude that for every $S \in \Delta$, $(X \pm D)|_S \succeq 0$ as long as δ is small enough that $(X \pm D)|_{\{i,j\}} \succeq 0$, and there exist positive such δ because $X|_{\{i,j\}}$ is positive definite.

This implies that X is not an extreme ray, as desired, since $X = \frac{1}{2}(X + D) + \frac{1}{2}(X - D)$, and X cannot be a multiple of D, since $D, -D \notin \mathcal{P}(G)$. □

It is not hard to see that clique complexes of chordal graphs are also locally rank 1. In [37], many other complexes besides these examples are shown to have this property, which leads naturally to other calculations of $\varepsilon(\Delta)$ for various other complexes Δ.

Lemma 3.2.5. *The minimum of λ_{min} on $\overline{\mathcal{P}(C_n)}$ is obtained at some X which is locally rank 1 and with exactly one negative entry.*

Proof. Note that the function λ_{min} is concave, and therefore maximized at an extreme point of $\overline{\mathcal{P}(C_n)}$, which by our previous lemma is locally rank 1. Now, let X' be such a locally rank 1 maximizer. We note that if D is a diagonal matrix where all of the diagonal entries are either 1 or -1, then D is unitary, and hence, conjugating a matrix by D preserves its eigenvalues. Given A, an extreme ray in $\mathcal{P}(C_n)$, we can conjugate it by an appropriate diagonal matrix D so that A has a minimal number of negative entries. There are two cases: either A can be conjugated so that all of its entries are nonnegative, or so that exactly one pair of entries is negative. If A can be conjugated so that all of its entries are nonnegative, then in fact, it is the projection of a rank 1 PSD matrix, and so, it is PSD completable. The only important case then is the case when A has exactly one pair of negative entries, and we will call this the *normal form* of an extreme ray A. □

Lemma 3.2.6. *If $X \in \mathcal{P}(C_n)$, then X has a rank 2 PSD completion if and only if*

there are some $a_i \in \{1, -1\}$ *so that*

$$\sum_{i=1}^{n} a_i \arccos\left(\frac{X_{ii+1}}{\sqrt{X_{ii}}\sqrt{X_{i+1i+1}}}\right) = 2k\pi, \qquad (*)$$

for some integer k*.*

Proof. We make a sequence of reductions to prove the result.

In terms of vector arrangements, X is completable to a PSD rank 2 matrix if and only if there are vectors $v_1, \ldots, v_n \in \mathbb{R}^2$ so that for each i

$$\|v_i\|^2 = X_{ii},$$
$$\langle v_i, v_{i+1}\rangle = X_{ii+1}.$$

For the sake of notation, let $\bar{X}_i = \frac{X_{ii+1}}{\sqrt{X_{ii}}\sqrt{X_{i+1i+1}}}$. If we renormalize these equations to make the v_i lie on the unit circle, we can equivalently ask for $v_i \in \mathbb{R}^2$ so that

$$\|v_i\|^2 = 1,$$
$$\langle v_i, v_{i+1}\rangle = \bar{X}_i$$

In this case, we will think of $a_i \arccos(\bar{X}_i)$ as the angle between v_i and v_{i+1}; the equation is equivalent to the condition that the sum of the angles between the vectors on the circle is a multiple of 2π.

Formally, for each $v_i \in \mathbb{R}^2$ so that $\langle v_i, v_i\rangle = 1$, we can express v_i in polar coordinates. This implies that there are some θ_i so that

$$v_i = (\cos(\theta_i), \sin(\theta_i)).$$

In this case, these equations reduce to the equations

$$\cos(\theta_{i+1} - \theta_i) = \bar{X}_i.$$

To see the necessity of the equation $(*)$, note that if there exist θ_i satisfying the previous equation, then using some basic facts about the cos function, there exist some $a_i \in \{-1, 1\}$ and $\ell_i \in \mathbb{Z}$ so that

$$\theta_{i+1} - \theta_i = a_i \arccos(\bar{X}_i) + 2\pi\ell_i.$$

Letting $k = -\sum_{i=1}^{n} \ell_i$, this implies that

$$\begin{aligned}
\sum_{i=1}^{n}(\theta_{i+1} - \theta_i) &= \sum_{i=1}^{n}(a_i \arccos(\bar{X}_i) + 2\pi\ell_i) \\
&= \sum_{i=1}^{n} a_i \arccos\left(\frac{X_{ii+1}}{\sqrt{X_{ii}}\sqrt{X_{i+1i+1}}}\right) - 2k\pi \\
&= 0.
\end{aligned}$$

which clearly implies equation $(*)$.

To see the sufficiency of equation $(*)$, suppose that there exist a_i and some k so that equation $(*)$ holds. Then, set $\theta_1 = 0$, and for each $1 \leq i < n$, set

$$\theta_{i+1} = \theta_i + a_i \arccos(\bar{X}_i).$$

Then, clearly, for $i < n$, we have the desired result that

$$\cos(\theta_{i+1} - \theta_i) = \cos(a_i \arccos(\bar{X}_i)) = \bar{X}_i.$$

and for $i = n$, we see that

$$\begin{aligned}\theta_n &= \sum_{i=1}^{n-1} a_i \arccos(\bar{X}_i) \\ &= \sum_{i=1}^{n-1} a_i \arccos\left(\frac{X_{ii+1}}{\sqrt{X_{ii}}\sqrt{X_{i+1i+1}}}\right) \\ &= 2k\pi - a_n \arccos\left(\frac{X_{1n}}{\sqrt{X_{11}}\sqrt{X_{nn}}}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\cos(\theta_1 - \theta_n) &= \cos\left(2k\pi - a_n \arccos\left(\frac{X_{1n}}{\sqrt{X_{11}}\sqrt{X_{nn}}}\right)\right) \\ &= \bar{X}_n,\end{aligned}$$

as we desired.

□

The previous lemma has particular application to a partial matrix in normal form.

Lemma 3.2.7. *If $X \in \mathcal{P}(C_n)$, and X is locally rank 1 with exactly one negative entry (say $X_{n\,1} < 0$), then $X + \varepsilon I_n$ has a rank 2 PSD completion if*

$$\sum_{i=1}^{n} \arccos\left(\frac{\sqrt{X_{ii}X_{i+1i+1}}}{\sqrt{X_{ii}+\varepsilon}\sqrt{X_{i+1i+1}+\varepsilon}}\right) = \pi.$$

Proof. This follows from lemma 3.2.6: after considering the sum

$$\begin{aligned}
&\sum_{i=1}^{n-1} \arccos\left(\frac{X_{ii+1}}{\sqrt{X_{ii}+\varepsilon}\sqrt{X_{i+1i+1}+\varepsilon}}\right) - \arccos\left(\frac{X_{1n}}{\sqrt{X_{11}+\varepsilon}\sqrt{X_{nn}+\varepsilon}}\right) \\
&= \sum_{i=1}^{n-1} \arccos\left(\frac{\sqrt{X_{ii}X_{i+1i+1}}}{\sqrt{X_{ii}+\varepsilon}\sqrt{X_{i+1i+1}+\varepsilon}}\right) - \arccos\left(-\frac{\sqrt{X_{11}X_{nn}}}{\sqrt{X_{11}+\varepsilon}\sqrt{X_{nn}+\varepsilon}}\right) \\
&= \sum_{i=1}^{n} \arccos\left(\frac{\sqrt{X_{ii}X_{i+1i+1}}}{\sqrt{X_{ii}+\varepsilon}\sqrt{X_{i+1i+1}+\varepsilon}}\right) - \pi \\
&= 0
\end{aligned}$$

where we have used the equation $\arccos(-x) = \pi - \arccos(x)$, and the hypothesis of the lemma.

□

The last, somewhat mysterious fact we will need is the following:

Lemma 3.2.8. *For any $\varepsilon \geq 0$, the function $f_\varepsilon : \mathbb{R}^2_+ \to \mathbb{R}$ given by*

$$f_\varepsilon(x,y) = \arccos\left(\frac{\sqrt{xy}}{\sqrt{x+\varepsilon}\sqrt{y+\varepsilon}}\right)$$

is convex.

Proof. We prove this by computing the Hessian matrix of f_ε.

$$H(f_\varepsilon) = \begin{pmatrix} \frac{\partial^2}{\partial x^2} f_\varepsilon & \frac{\partial^2}{\partial x \partial y} f_\varepsilon \\ \frac{\partial^2}{\partial x \partial y} f_\varepsilon & \frac{\partial^2}{\partial y^2} f_\varepsilon \end{pmatrix}.$$

This evaluates to

$$\begin{pmatrix} \frac{\varepsilon^2 y^2\left(\varepsilon^2+\varepsilon(5x+y)+x(4x+3y)\right)}{4(\varepsilon+x)^{7/2}(\varepsilon+y)^{3/2}(xy)^{3/2}\left(\frac{\varepsilon(\varepsilon+x+y)}{(\varepsilon+x)(\varepsilon+y)}\right)^{3/2}} & -\frac{\varepsilon^2}{4(\varepsilon+x)^{3/2}(\varepsilon+y)^{3/2}\sqrt{xy}\left(\frac{\varepsilon(\varepsilon+x+y)}{(\varepsilon+x)(\varepsilon+y)}\right)^{3/2}} \\ -\frac{\varepsilon^2}{4(\varepsilon+x)^{3/2}(\varepsilon+y)^{3/2}\sqrt{xy}\left(\frac{\varepsilon(\varepsilon+x+y)}{(\varepsilon+x)(\varepsilon+y)}\right)^{3/2}} & \frac{\varepsilon^2 x^2\left(\varepsilon^2+\varepsilon(x+5y)+y(3x+4y)\right)}{4(\varepsilon+x)^{3/2}(\varepsilon+y)^{7/2}(xy)^{3/2}\left(\frac{\varepsilon(\varepsilon+x+y)}{(\varepsilon+x)(\varepsilon+y)}\right)^{3/2}} \end{pmatrix}.$$

Note that if $\varepsilon, x, y \geq 0$, then the diagonal entries of this matrix are nonnegative.

Consider the Hessian determinant of this function if $x, y > 0$:

$$\det(H(f_\varepsilon)) = \frac{\varepsilon(\varepsilon(x+y)+3xy)}{4xy(\varepsilon+x)^2(\varepsilon+y)^2(\varepsilon+x+y)}.$$

This is also nonnegative on this domain.

These two facts about $H(f_\varepsilon)$ are enough to determine that it is positive semidefinite, and so f is convex. □

Proof of Theorem 3.2.3. Fix some X in normal form. We wish to show that there is some $\varepsilon \leq \frac{1}{n}(\frac{1}{\cos(\frac{\pi}{n})} - 1)$ so that $X + \varepsilon I_n$ is completable to a PSD matrix with rank 2. We want to apply the condition in lemma 3.2.7. Consider the function

$$g(\varepsilon) = \sum_{i=1}^{n} \arccos\left(\frac{\sqrt{X_{ii}X_{i+1i+1}}}{\sqrt{X_{ii}+\varepsilon}\sqrt{X_{i+1i+1}+\varepsilon}}\right).$$

Lemma 3.2.7 implies if ε is such that $g(\varepsilon) = \pi$, then in fact there is a rank 2 PSD completion of $X + \varepsilon I_n$. We will show that $g(0) = 0$ and $g(\frac{1}{n}(\frac{1}{\cos(\frac{\pi}{n})} - 1)) \geq \pi$, so that by the intermediate value theorem, there must be $\varepsilon \in [0, \frac{1}{n}(\frac{1}{\cos(\frac{\pi}{n})} - 1)]$ so that $g(\varepsilon) = \pi$, yielding the result.

First note that if $\varepsilon = 0$, then

$$\sum_{i=1}^{n} \arccos\left(\frac{\sqrt{X_{ii}X_{i+1i+1}}}{\sqrt{X_{ii}+\varepsilon}\sqrt{X_{i+1i+1}+\varepsilon}}\right) = \sum_{i=1}^{n} \arccos(1) = 0 < \pi,$$

Now, we want to show that if $\varepsilon = \frac{1}{n}(\frac{1}{\cos(\frac{\pi}{n})} - 1)$, then $g(\varepsilon) \geq \pi$. We use lemma

3.2.8 and the fact that $\mathrm{Tr}(X) = 1$ to see that

$$\begin{aligned} g(\varepsilon) &= n \sum_{i=1}^{n} \frac{1}{n} f_\varepsilon \left(X_{ii}, X_{i+1i+1} \right) \\ &\geq n f_\varepsilon \left(\frac{1}{n} \sum_{i=1}^{n} X_{ii}, \frac{1}{n} \sum_{i=1}^{n} X_{ii} \right) \\ &= n f \left(\frac{1}{n}, \frac{1}{n} \right) \\ &= n \arccos(\frac{1}{1+n\varepsilon}) \\ &= n \arccos(\cos(\frac{\pi}{n})) \\ &= \pi, \end{aligned}$$

as desired.

Thus, there is some $\varepsilon \leq \frac{1}{n}(\frac{1}{\cos(\frac{\pi}{n})} - 1)$ satisfying the condition of lemma 3.2.7.

To see that this value of ε is in fact attained for some X, we can use the matrix X where $X_{ii} = \frac{1}{n}$ for each i, $X_{12} = -\frac{1}{n}$, and each other specified off-diagonal entry is $\frac{1}{n}$. This matrix can be verified to have $\lambda_{min}(X) = \frac{1}{n}(1 - \frac{1}{\cos(\frac{\pi}{n})})$ using the cycle conditions found in [36]. □

We note that asymptotically $\frac{1}{n}\left(\frac{1}{\cos(\frac{\pi}{n})} - 1\right)$ is $O(\frac{1}{n^3})$.

3.2.2 Extensions of Theorem 3.2.3

We conclude this section by noting some extensions of Theorem 3.2.3 to other complexes. We will say that a graph G is *cycle dominated* if $\varepsilon(\chi(G)) = \max \varepsilon(C_n)$, where the maximum is over all induced cycles with at least 4 vertices contained in G.

It is a clear consequence of the the cycle completability results that if G is series parallel or chordal, then G is cycle dominated. In fact, [27] shows that a much larger range of graphs is cycle dominated beyond just these classes. This suggests that there may be some sense in which the requirement of being chordal is too stringent for use

in these semidefinite programming contexts, at least as far as approximation results are concerned.

3.3 Connections to hyperbolic polynomials

We will conclude this section on approximate PSD completions with some results concerning specifically the set $\Delta = \binom{[n]}{k}$, all sets of size at most k in $[n]$. We will make the notation $\mathcal{F}^{n,k} = \mathcal{M}(\binom{[n]}{k})$ and $\mathcal{S}^{n,k} = \mathcal{P}(\binom{[n]}{k})$. Explicitly, $\mathcal{S}^{n,k}$ is the set of matrices in $\mathbb{R}^{n\times n}_{sym}$ with the property that every $k \times k$ principal submatrix is positive semidefinite.

We will study this cone, and in particular, how it relates to the hyperbolicity cone $\Lambda_I(c_{n,k})$ defined in Equation (2.3.1).

Theorem 3.3.1. *For any $n \geq k \geq 1$, we have that*

$$\mathcal{S}^{n,k} \subseteq \Lambda_I(c_{n,k}).$$

Proof. We first show that for any X in the interior of $\mathcal{S}^{n,k}$,

$$c_{n,k}(X) > 0.$$

For this, recall that

$$c_{n,k}(X) = \sum_{S\subseteq[n]:|S|=k} \det(X|_S),$$

If X is in fact in the interior of $\mathcal{S}^{n,k}$, then for every $S \subseteq [n]$ with $|S| = k$, $X|_S \succ 0$, and so $\det(X|_S) > 0$. Because each of these summands is positive, we have that $c_{n,k}(X) > 0$.

Therefore, in particular, we have that $c_{n,k}(X)$ does not vanish on the interior of $\mathcal{S}^{n,k}$ Now, we note that I is in the interior of $\mathcal{S}^{n,k}$, and because $\mathcal{S}^{n,k}$ is convex and connected, it follows that the interior of $\mathcal{S}^{n,k}$ is contained in the connected component

of $\mathbb{R}^{n\times n}_{sym} \setminus \mathcal{V}(c_{n,k})$ containing I.

By Item 3, the interior of $\mathcal{S}^{n,k}$ is contained in $\Lambda_I(c_{n,k})$, and since $\mathcal{S}^{n,k}$ is full dimensional, and $\Lambda_I(c_{n,k})$ is closed, we must have that $\mathcal{S}^{n,k} \subseteq \Lambda_I(c_{n,k})$. □

Remark 7. *It is not hard to modify this proof to show that in fact, $\mathcal{S}^{n,k}$ is a subset of $\Lambda_I(P)$ for any PSD-stable LPM polynomial P of degree k.*

We can use this fact to characterize the gapof this complex explicitly:

Theorem 3.3.2. *We have that*

$$\varepsilon\left(\binom{[n]}{k}\right) = \frac{n-k}{n(k-1)}.$$

In words, for any $X \in \mathcal{S}^{n,k}$ with trace 1, $\lambda_{min}(X) \geq \frac{k-n}{n(k-1)}$ and there is some $X \in \mathcal{S}^{n,k}$ with trace 1 meeting this inequality with equality.

Proof. To see that $\varepsilon(\binom{[n]}{k}) \geq \frac{n-k}{n(k-1)}$, we first exhibit a matrix $X \in \mathcal{S}^{n,k}$ so that $\lambda_{min}(X) = \frac{k-n}{n(k-1)}$. We then let

$$G(n,k) = \frac{k}{n(k-1)} I_n - \frac{1}{n(k-1)} \vec{1}_n \vec{1}_n^\intercal.$$

Here, we use the subscripts to emphasize that these matrices are of size $n \times n$. Note that $\mathrm{Tr}(G(n,k)) = 1$, and that

$$\lambda_{min}(G(n,k)) = \frac{k}{n(k-1)} - \frac{1}{n(k-1)} \lambda_{max}(\vec{1}_n \vec{1}_n^\intercal) = \frac{k-n}{n(k-1)}.$$

To see that $G(n,k) \in \mathcal{S}^{n,k}$, we note that any $k \times k$ submatrix of $G(n,k)$ has the form

$$\frac{k}{n(k-1)} I_k - \frac{1}{n(k-1)} \vec{1}_k \vec{1}_k^\intercal,$$

which has minimum eigenvalue 0.

Now, we need to show that for any $X \in \mathcal{S}^{n,k}$ with trace 1, $\lambda_{min}(X) \geq \frac{k-n}{n(k-1)}$. We can in fact show that for any $X \in \Lambda_I(c_{n,k})$ with trace 1, $\lambda_{min}(X) \geq \frac{k-n}{n(k-1)}$. Suppose for a contradiction that $X \in \Lambda_I(c_{n,k})$ with trace 1 and $\lambda_{min}(X) < \frac{k-n}{n(k-1)}$. Let $\vec{\lambda}$ denote the vector of eigenvalues of X in ascending order, so that $\vec{\lambda}_1 = \lambda_{min}(X)$. We now claim that $\vec{\lambda} \in \Lambda_{\vec{1}}(e_{n,k})$. For this, we use the fact that

$$e_{n,k}(\vec{\lambda} + t\vec{1}) = c_{n,k}(X + tI).$$

Therefore, if X satisfies Item 1 in the definition of membership in $\Lambda_I(c_{n,k})$, then $\vec{\lambda}$ satisfies Item 1 for membership in $\Lambda_{\vec{1}}(e_{n,k})$.

Now, we consider the following minimization problem:

$$\begin{aligned} \min \quad & \vec{\lambda}_1 \\ \text{s.t.} \quad & \sum_{i=1}^{n} \vec{\lambda}_i = 1 \quad . \\ & \vec{\lambda} \in \Lambda_{\vec{1}}(e_{n,k}) \end{aligned} \tag{3.3.1}$$

This is a convex optimization problem, and because $\vec{\lambda} \in \Lambda_I(e_{n,k})$ and $\sum_{i=1}^{n} \vec{\lambda}_i = \mathrm{Tr}(X) = 1$, we have a feasible solution where $\vec{\lambda}_1 < \frac{k-n}{n(k-1)}$ Now, we note that this problem is symmetric with respect to the last $n-1$ variables, in the sense that if $\vec{\lambda}$ is feasible and π is a permutation of $[n]$ so that $\pi(1) = 1$, then $\pi(\vec{\lambda})$ is also feasible. Therefore, we may consider the symmetric solution $\sum_{\pi \in \mathfrak{S}_n : \pi(1)=1} \pi(\vec{\lambda})$ is feasible with the same objective value. We conclude that there is an optimal solution to this program satisfying $\vec{\lambda}_i = \vec{\lambda}_j$ when $n \geq i, j > 1$. Making the replacement $a = \vec{\lambda}_1$ and $b = \frac{1-a}{n-1}$ be the common value of $\vec{\lambda}_i$ for $i = 2, \ldots, n$, we have therefore reduced this to

a 1 dimensional optimization problem:

$$\begin{aligned} \min \quad & a \\ \text{s.t.} \quad & a + (n-1)b = 1 \\ & (a, b, b, \ldots, b) \in \Lambda_{\vec{1}}(e_{n,k}) \end{aligned} \quad . \tag{3.3.2}$$

We see that this must have a solution on the boundary of $\Lambda_{\vec{1}}(e_{n,k})$, i.e. we must have that at an optimal a, b,

$$e_{n,k}(a, b, b, \ldots, b) = a\binom{n-1}{k-1}b^{k-1} + \binom{n-1}{k}b^k = 0.$$

That is, $\binom{n-1}{k-1}a + \binom{n-1}{k}b = 0$. Solving the resulting linear system of equations for a and b yields

$$a = \frac{k-n}{n(k-1)},\ b = \frac{(n-1)k}{n(k-1)}$$

This contradicts the fact that there is a feasible solution with an objective value smaller than $\frac{k-n}{n(k-1)}$. □

Remark 8. *Various extensions of this result are discussed in [38]. For example, the same proof can be made to work if the constraint that* $\mathrm{Tr}(X) = 1$ *is replaced by other convex constraints that are invariant under orthogonal change of basis. Moreover, it can be shown that in fact,* $G(n, k)$ *is the unique minimizer of* $\lambda_{min}(X)$ *for any such constraint.*

This follows from the more general fact, shown in [38] that if $X \in \mathcal{S}^{n,k}$ *is nonsingular, but has the property that all* $k \times k$ *submatrices are singular, then* X *must be of the form* $DG(n, k)D$ *for some diagonal matrix* D*. Note that if* X *is on the boundary of* $\Lambda_I(c_{n,k})$*, then* $c_{n,k}(X) = 0$*, and for* $X \in \mathcal{S}^{n,k}$*, this can only occur when all* $k \times k$ *submatrices are singular.*

3.3.1 Connections to sparse quadratic programming

We conclude this section by describing how the containment $\mathcal{S}^{n,k} \subseteq \Lambda_I(c_{n,k})$ can be used to produce heuristics for sparse quadratic programming (which is defined in the introduction to this chapter). This work was originally conducted in [39].

If A_1 is positive semidefinite, the conical optimization problem

$$\begin{aligned} \min \quad & \mathrm{Tr}(A_0 X) \\ \text{s.t.} \quad & \mathrm{Tr}(A_1 X) = 1 \\ & X \in \mathcal{F}^{n,k} \end{aligned} \tag{3.3.3}$$

has a dual given by

$$\begin{aligned} \max \quad & t \\ \text{s.t.} \quad & A_0 - A_1 t \in \mathcal{S}^{n,k}. \end{aligned} \tag{3.3.4}$$

We denote the common optimal value of these programs by α.

We may consider replacing $\mathcal{S}^{n,k}$ by $\Lambda_I(c_{n,k})$ in the definition, to obtain a value α' so that $\alpha' \geq \alpha$. We may interpret this value as ensuring that there is some $X \in \mathcal{F}^{n,k}$ which obtains an optimal value of at most α'.

We will note though that the value of this program can in fact be easily computed in terms of the roots of univariate polynomials.

Theorem 3.3.3. *If $A_1 \succ 0$, then the optimal value of the program*

$$\begin{aligned} \max \quad & t \\ \textit{s.t.} \quad & A_0 - A_1 t \in \Lambda_I(c_{n,k}). \end{aligned} \tag{3.3.5}$$

is given by

$$\max\{t : c_{n,k}(A_1 t - A_0) = 0\}.$$

Proof. We note that because A_1 is positive definite, A_1 is in the interior of $\Lambda_I(c_{n,k})$

and $-A_1$ is not in $\Lambda_I(c_{n,k})$. We conclude that for t large enough, $A_1t - A_0 \in \Lambda_I(c_{n,k})$, and for t a sufficiently small negative number, $A_1t - A_0 \notin \Lambda_I(c_{n,k})$.

Therefore, the value of this problem is bounded, and its value is obtained when $A_1t - A_0$ is on the boundary of the hyperbolicity cone of $\Lambda_I(c_{n,k})$, which implies that $c_{n,k}(A_1t - A_0) = 0$ when t is at its optimum. To see that it is the largest value of t so that this is the case, note that if $X \in \Lambda_I(c_{n,k})$, then by the definition of the hyperbolicity cone, for all $c_{n,k}(X + tA_1) \neq 0$ for $t > 0$, and if we let $X = A_1t - A_0$, then this implies that there are no larger roots of this polynomial. □

We may apply this relaxation to some of the standard problems we listed above.

For example, if we consider the sparse PCA problem, we obtain that the largest k-sparse eigenvector of A is

$$\max\{v^\intercal Av : \|v\|_2 = 1,\ \|v\|_0 \leq k\} \geq \max\{t : c_{n,k}(It - A) = 0\}.$$

Compare this to the formula for the maximum eigenvalue of a matrix A, which is

$$\max\{t : \det(It - A) = 0\}.$$

We can give even nicer expressions for the sparse linear regression problem:

Theorem 3.3.4. *Let A have rank at least k, then for the sparse regression problem in Equation* (3.0.1)*, the lower bound arising from Theorem 3.3.3 is*

$$\frac{p(A^\intercal(I + bb^\intercal)A)}{p(A^\intercal A)} - 1.$$

Proof of Theorem 3.3.4. We consider the univariate polynomial

$$q(y) = c_{n,k}(A^\intercal Ay - A^\intercal bb^\intercal A) = 0.$$

Notice that when $y = 0$, we obtain $q(y) = c_{n,k}(A^\intercal bb^\intercal A)$. Now, notice that $X = A^\intercal bb^\intercal A$ is rank 1, and therefore, $\det(X|_S)$ vanishes to order $k-1$ at this point. Because $c_{n,k}$ is a linear combination of determinants, $c_{n,k}$ must then have a root of multiplicity at least $k-1$ at 0.

Because $c_{n,k}(A^\intercal A) \neq 0$, and $A^\intercal A$ is positive semidefinite, we have that any root of this polynomial must be nonnegative, so we have that

$$c_{n,k}(A^\intercal Ay - A^\intercal bb^\intercal A) = y^{k-1}(ay - b) = ay^k - by^{k-1}$$

for some $a, b \geq 0$. Hence, the maximal root of $c_{n,k}$ must be $\frac{b}{a}$.

We can compute a and b explicitly. Notice that

$$\lim_{y\to\infty} \frac{c_{n,k}(A^\intercal Ay - A^\intercal bb^\intercal A)}{y^k} = c_{n,k}(A^\intercal A) = a,$$

and that

$$c_{n,k}(-A^\intercal A - A^\intercal bb^\intercal A) = (-1)^k c_{n,k}(A^\intercal Ay + A^\intercal bb^\intercal A) = (-1)^k(a+b)$$

From this, we obtain that

$$\frac{b}{a} = \frac{c_{n,k}(A^\intercal Ay + A^\intercal bb^\intercal A)}{c_{n,k}(A^\intercal A)} - 1,$$

as desired. □

Remark 9. *The quantity described in Theorem 3.3.4 can also be interpreted as the expected value a random variable representing the loss resulting from first randomly choosing a set of k columns of A* from a determinantal point process *and then performing regression using those columns.*

We will conclude by asking some questions: is it possible to algorithmically find a

set k-sparse vector x meeting the bound described by Theorem 3.3.3? Perhaps more importantly, is it the case that this bound is in fact useful?

In fact, it is possible to find this value in polynomial time, as described in [39], but this bound by itself is not useful. Fortunately, the methods described here typically returns a sparse vector which has much better loss than what was originally stated.

For this, we will need to define a new class of polynomials, which we refer to as *conditional polynomials*. For a fixed, n, and $S \subseteq [n]$ with $|S| \leq k$, we let

$$c_{n,k,S}(X) = \sum_{\substack{S \subseteq T \subseteq [n] \\ |T|=k}} \det(X|_T).$$

This polynomial can easily be seen to be PSD stable using Theorem 2.3.2, and $\mathcal{S}^{n,k} \subseteq \Lambda_I(c_{n,k,S})$ for any S with $|S| \leq k$. The proof of Theorem 3.3.3 also shows that an upper bound for the optimal value of Equation (3.3.4) is given by

$$\max\{t : c_{n,k,S}(A_1 t - A_0) = 0\}.$$

If we let $\eta_S = \max\{t : c_{n,k,S}(A_1 t - A_0) = 0\}$, then we can show the following:

Theorem 3.3.5. *Fix any S with $|S| < k$, and for any $i \in [n]$ let $S_i = S \cup \{i\}$. Then, there exists $i \in S^c$ so that*

$$\eta_{S_i} \leq \eta_S.$$

Proof. We show this by noting the algebraic identity that if $S_i = S \cup \{i\}$, then

$$c_{n,k,S}(X) = \frac{1}{k - |S|} \sum_{i \in S^c} c_{n,k,S_i}(X).$$

This follows by considering each summand of $c_{n,k,S}(X)$ and noting that it appears in precisely $k - |S|$ polynomials in the sum on the right.

Therefore,

$$c_{n,k,S}(A_1\eta_S - A_0) = \frac{1}{k-|S|}\sum_{i\in S^c} c_{n,k,S_i}(A_1\eta_S - A_0) = 0.$$

There must therefore be some i so that $c_{n,k,S_i}(A_1\eta_S - A_0) \leq 0$. Fixing this i, since $\lim_{t\to\infty} c_{n,k,S_i}(A_1\eta_S - A_0) > 0$, we have by the intermediate value theorem that there is a root of $c_{n,k,S_i}(A_1 t - A_0)$ which is at least η_S, as desired. □

We can also compute this polynomial efficiently using the notion of *Schur complements*, which was defined in the introduction of this thesis.

Lemma 3.3.6. *We have that*

$$c_{n,k,S}(X) = \det(X|_S)c_{n-|S|,k-|S|}(X \setminus S).$$

Proof. We recall the Schur complement lemma, that for any $S \subseteq [n]$, $\det(X) = \det(X|_S)\det(X \setminus S)$.

It is not hard to see from the definition that for any $T \subseteq S^c$, $(X \setminus S)|_T = X|_{T\cup S} \setminus S$.

Therefore,

$$\begin{aligned}
c_{n,k,S}(X) &= \sum_{\substack{S\subseteq T\subseteq[n]\\|T|=k}} \det(X|_T) \\
&= \sum_{\substack{T\subseteq S^c\\|T|=k-|S|}} \det(X|_{S\cup T}) \\
&= \sum_{\substack{T\subseteq S^c\\|T|=k-|S|}} \det(X|_S)\det(X|_{S\cup T} \setminus S) \\
&= \det(X|_S) \sum_{\substack{T\subseteq S^c\\|T|=k-|S|}} \det((X \setminus S)|_T) \\
&= \det(X|_S)c_{n-|S|,k-|S|}(X \setminus S).
\end{aligned}$$

□

Putting these ideas together, we reach the greedy algorithm, Algorithm 1, for finding a good solution to the sparse quadratic programming. By carefully performing

Algorithm 1 The Greedy Conditioning Heuristic

$T \leftarrow \varnothing$
for $t = 1 \dots k$ **do**
  $j \leftarrow \operatorname{argmax} \eta_{T+j}$
  $T \leftarrow T + j$
end forreturn T

the linear algebraic manipulations needed to compute the η_T, it is possible to obtain a practically efficient algorithm for finding these values, as was done in [39].

CHAPTER 4
HIDDEN CONVEXITY AND ALGEBRAIC TOPOLOGY

We use the term 'hidden convexity' to refer to situations in which the image of a *nonconvex* set under a possibly *nonlinear* map is unexpectedly convex. While this phenomenon may seem rather esoteric, there are a wide range of examples occuring in the literature, and such results have been impactful in the study of nonconvex optimization.

There is a vast literature on the convexity of sets known as the *generalized numerical ranges* of a collection of matrices, which we will attempt to outline in Section 4.1 Our goal will be to give a proof of these results in a unified framework that involve facts about the algebraic topology of homogeneous spaces of Lie groups.

4.1 History and preliminary notions

There is a vast literature on the convexity of sets known as the *generalized numerical ranges* of a collection of matrices. These results generalize the well known Toeplitz-Hausdorff Theorem [40, 41, 42]. The Toeplitz-Hausdorff Theorem states that if A is an $n \times n$ hermitian matrix with $n \geq 3$, then the numerical range of A is convex, where the numerical range of A is defined as $\{x^*Ax : x^*x = 1\}$. Some generalizations of this result consider the image of the sphere under a larger number of quadratic maps [43, 44, 45]. Others regard more complicated domains such as the set of symmetric or hermitian matrices with fixed eigenvalues [46, 47], or even more generally, the orbit of an element of a Lie algebra under the adjoint action of an associated Lie group [48]. Such theorems have consequences related to the well known S-lemma in optimization [49].

4.2 Summary of results

We summarize our new approach as Theorem 4.3.1, after stating some definitions. We will also give a few examples of explicit applications of Theorem 4.3.1 to generalized numerical ranges. The main advantage of this approach is that once the framework is set up, specializing to the above examples requires only recalling some well-known theorems. In this way, our proofs help show how these hidden convexity results are really consequences of basic geometric and topological properties of the relevant groups.

As an application of Theorem 4.3.1, we give a common generalization of two known theorems, Theorem 4.2.2 and Theorem 4.2.1, in Theorem 4.5.1. While we will defer the statement of Theorem 4.3.1 until we have built up some more generalizations, we will state the two theorems which we aim to generalize here.

Theorem 4.2.1 concerns higher dimensional analogues of the Toeplitz-Hausdorff Theorem. To state it, we will need to recall that an eigenvalue of a matrix X is nondegenerate if its corresponding eigenspace is one dimensional. Following [50], we say that a subspace of $\mathbb{R}^{n\times n}_{sym}$ is noncrossing if every nonzero matrix in that subspace has only nondegenerate eigenvalues. we say that a subspace of $\mathbb{R}^{n\times n}_{sym}$ is k-weakly noncrossing if every nonzero matrix has the property that its k largest eigenvalues are all nondegenerate.

Theorem 4.2.1 (Theorem 5.1 of [45])**.** *Let $A_1, \ldots, A_k$ be $n \times n$ symmetric or hermitian matrices with $n \geq 3$, with the property that every nonzero matrix in their linear span has a nondegenerate maximum eigenvalue. Then, $\{(x^\intercal A_1 x, \ldots, x^\intercal A_k x) : \|x\| = 1\}$ is convex.*

Note that [45] shows more; rather than requiring that the maximum eigenvalue be nondegenerate, they only require that the corresponding eigenspace have constant dimension in the span of $A_1, \ldots, A_k$.

Theorem 4.2.2 is a generalization of Toeplitz-Hausdorff Theorem involving the set of symmetric matrices with fixed eigenvalues. For a field $\mathbb{F}$ which is either $\mathbb{C}$ or $\mathbb{R}$, we let $\mathbb{F}^{n\times n}_{sym}$ denote the vector space of symmetric and hermitian $n \times n$ matrices respectively. For $\lambda \in \mathbb{R}^n$, let $M^{\mathbb{F}}_\lambda$ denote the set of matrices in $\mathbb{F}^{n\times n}_{sym}$ whose eigenvalues are the entries of λ (counting multiplicity), in some order.

Theorem 4.2.2 (Theorems from [46, 47])**.** *Fix some $\lambda \in \mathbb{R}^n$. Let $T : \mathbb{R}^{n\times n}_{sym} \to \mathbb{R}^2$ be linear with $n \geq 3$, then $T(M^{\mathbb{R}}_\lambda)$ is convex. Similarly, let $T : \mathbb{C}^{n\times n}_{sym} \to \mathbb{R}^3$ be linear (as a map of real vector spaces) with $n \geq 3$, then $T(M^{\mathbb{C}}_\lambda)$ is convex.*

The gist of Theorem 4.5.1 is that as long as $A_1, \ldots, A_k \in \mathbb{F}^{n\times n}$ span a noncrossing subspace, then $T(M^{\mathbb{F}}_\lambda)$ is convex, as long as $O_{\mathbb{F}}(n)$ satisfies some homotopy theoretic conditions. These conditions are easily shown to be satisfied when specialized to these two cases.

We will also give another proof of a result from [48]. Fix some $R \in \mathbb{R}^{n\times m}$, and let S_R denote the orbit of R under the action of $\mathrm{SO}(n) \times \mathrm{SO}(m)$, i.e. $S_R = \{U^\intercal RV : U \in \mathrm{SO}(n), V \in \mathrm{SO}(m)\}$.

Theorem 4.2.3. *Let $n, m \geq 3$, then for any linear map $T : \mathbb{R}^{n\times m} \to \mathbb{R}^2$, $T(S_R)$ is convex.*

This result also generalizes a result in [51] stating that the image of $\mathrm{SO}(n)$ under a linear map into $\mathbb{R}^2$ is always convex.

The layout of this chapter is as follows: in Section 4.3, we define the notion of *continously maximized functions* and prove our main theorem. In Section 4.4, we give some examples of such continuously maximized functions. In Section 4.5, we show the aforementioned hidden convexity theorems. Finally, in Section 4.6, we give some applications to optimization over the set of rotation matrices which were originally discussed in [51].

4.3 Continuously maximized functions

Let X be a compact path-connected topological space and let S^k denote the unit sphere in $\mathbb{R}^{k+1}$. We say a continuous function $f : X \to \mathbb{R}^{k+1}$ is *continuously maximized* by a continuous function $\phi : S^k \to X$ if for every $v \in S^k$,

$$\max_{x \in X} \langle v, f(x) \rangle = \langle v, f(\phi(v)) \rangle.$$

We say f is continuously maximized (omitting the dependence on ϕ) if it is continuously maximized by some continuous function ϕ. The function

$$h(v) = \max_{x \in X} \langle v, f(x) \rangle$$

is the *support function* of $f(X)$, and is closely related to the convex hull of $f(X)$.

We will let $\pi_m(X, x_0)$ denote the m^{th} homotopy group of the topological space X with basepoint x_0. In cases in which X is path-connected (as it will be in all of our examples), we will omit the dependence on the basepoint. We will use 0 to denote the trivial group with one element, and say that a group homomorphism is nonzero if its image is not 0.

Theorem 4.3.1. *Let X be a compact path-connected topological space. Suppose that there is some m so that $\pi_m(S^k) \neq 0$, but there are no nonzero group homomorphisms from $\pi_m(X)$ to $\pi_m(S^k)$. Let $f : X \to \mathbb{R}^{k+1}$ be continuously maximized. Then the image $f(X)$ is convex.*

A special case of Theorem 4.3.1 (which follows immediately) is when $m = k$, in which case $\pi_k(S^k) = \mathbb{Z}$, and it suffices for the abelianization $\pi_k(X)'$ to be finite for the homotopy theoretic conditions to be satisfied.

Corollary 4.3.2. *Let X be a compact path-connected topological space, so that $\pi_k(X)'$ is finite. If $f : X \to \mathbb{R}^{k+1}$ is continuously maximized, then the image $f(X)$ is convex.*

The approach used in the proof of this theorem is similar to that of [45], which also considered the support function, though here we use the homotopy groups of X more explicitly. We next show some useful lemmas about continuously maximized functions, which we will use in our proof of Theorem 4.3.1 in Section 4.3.2.

Also, while we use homotopy groups in the statements of these theorems, the proof of this result can easily be modified to use homology or cohomology groups, which may be easier to calculate in some circumstances. We mostly appeal to homotopy groups in our settings, as we will only be considering spaces whose homotopy groups are already well understood.

4.3.1 Preliminaries On Continuously Maximized Functions

Recall that the convex hull of a set $S \subseteq \mathbb{R}^k$, denoted $\operatorname{conv} S$, is the intersection of all convex sets containing S. It is well known that if S is compact, then so is $\operatorname{conv} S$, and for any $w \in \mathbb{R}^k$, $\max_{z \in S} \langle w, z \rangle = \max_{z \in \operatorname{conv} S} \langle w, z \rangle$.

We use the notation Y^o to denote the topological interior of Y.

Lemma 4.3.3. *Let $f : X \to \mathbb{R}^{k+1}$ be continuously maximized by ϕ. Let $Y = \operatorname{conv} f(X)$. Then Y is either a single point, or it has nonempty interior.*

Proof. Suppose for contradiction that Y^o is empty and that Y is not a single point. Because Y is convex, Y^o is empty if and only if there is a $v \in S^k$ so that the value of $\langle v, y \rangle$ is constant for $y \in Y$. Because Y is not a single point, there is some $w \in S^k$ so that

$$\max_{y \in Y} \langle w, y \rangle \neq \min_{y \in Y} \langle w, y \rangle = -\max_{y \in Y} \langle -w, y \rangle.$$

Consider the curve $\gamma(\epsilon) = \frac{v + \epsilon w}{\|v + \epsilon w\|}$ which is well defined on the interval $[-\delta, \delta]$ for δ small enough. Note that if $\epsilon > 0$, then x maximizes $\langle \gamma(\epsilon), f(x) \rangle$ if and only if x maximizes $\langle w, f(x) \rangle$. Similarly, if x maximizes $\langle \gamma(\epsilon), f(x) \rangle$ for $\epsilon < 0$, then x maximizes $\langle -w, f(x) \rangle$.

By continuity of ϕ and γ, we have that

$$\max_{x\in X}\langle w, f(x)\rangle = \lim_{\epsilon\to 0^+}\langle w, f(\phi(\gamma(\epsilon)))\rangle = \lim_{\epsilon\to 0^-}\langle w, f(\phi(\gamma(\epsilon)))\rangle = -\max_{x\in X}\langle -w, f(x)\rangle,$$

which is a contradiction, as this implies that $\max_{y\in Y}\langle w, y\rangle = -\max_{y\in Y}\langle -w, y\rangle$

□

Our next lemma concerns properties of the function $f \circ \phi$, when f is continuously maximized by ϕ.

Lemma 4.3.4. *Let $f : X \to \mathbb{R}^{k+1}$ be continuously maximized by ϕ, and let $Y = \operatorname{conv} f(X)$. Let $\psi = f \circ \phi$. Then for any $y \in Y^o$, and any $w \in S^k$, $\langle w, \psi(v)\rangle > \langle w, y\rangle$.*

Proof. The definition of continuous maximization implies that

$$\max_{z\in f(X)}\langle w, z\rangle = \max_{x\in X}\langle w, f(x)\rangle = \langle w, f(\phi(w))\rangle = \langle w, \psi(w)\rangle.$$

On the other hand, for $y \in Y^o$, $\max_{z\in Y}\langle v, z\rangle > \langle v, y\rangle$. This shows the result □

Our next lemma is the main technical component of our homotopy argument. Intuitively, if we consider $\psi = f \circ \phi$, we would like to argue that this defines a 'nice' parameterization of the boundary of $\operatorname{conv} f(X)$. This is in the sense that for any y in the interior of $\operatorname{conv} f(X)$, the map ψ in fact defines a homotopy equivalence between the sphere and $\mathbb{R}^{k+1} \setminus \{y\}$. If this is the case, then it is visually intuitive that if the map ψ turns out to be contractible for homotopy theoretic reasons, then y will be in the image of f.

Lemma 4.3.5. *Let $f : X \to \mathbb{R}^{k+1}$ be continuously maximized by ϕ. Let $\psi : S^k \to \mathbb{R}^{k+1}$ be the composition of ϕ and f. Let $Y = \operatorname{conv} f(X)$. For any $y \in Y^o$, ψ is a homotopy equivalence between S^k and $\mathbb{R}^{k+1} \setminus \{y\}$.*

Proof. Fix a $y \in Y^o$ for the remainder of this argument. For $x \in \mathbb{R}^{k+1}$, we will denote $[x] = \frac{x}{\|x\|} \in S^k$, which is clearly continuous on $\mathbb{R}^{k+1} \setminus 0$.

We define the map $\tau : \mathbb{R}^{k+1} \setminus y \to S^k$ by letting $\tau(x) = [x - y] \in S^k$. This map is clearly continuous on its domain.

To show that ψ is a homotopy equivalence, it suffices to argue that $\psi \circ \tau$ is homotopic to the identity on S^k and $\tau \circ \psi$ is homotopic to the identity map on $\mathbb{R}^{k+1} \setminus y$.

First, we show $\tau \circ \psi$ is homotopic to the identity on S^k. Consider the function $h : S^k \times [0,1] \to \mathbb{R}^{k+1}$ defined by

$$h(v,t) = tv + (1-t)(\psi(v) - y).$$

Then define the normalized map $g : S^k \times [0,1] \to S^k$ by $g(v,t) = [h(v,t)]$. To see that g is well defined and continuous, we only need to show that $h(v,t) \neq 0$ for any $v, t \in S^k \times [0,1]$. For any $v \in S^k$,

$$\langle v, h(v,t)\rangle = t\langle v, v\rangle + (1-t)(\langle v, \psi(v)\rangle - \langle v, y\rangle).$$

Now, $\langle v, v\rangle = 1$ for all $v \in S^k$, and $\langle v, \psi(v)\rangle - \langle v, y\rangle > 0$ for all $v \in S^k$ by Lemma 4.3.4. In particular, $\langle v, h(v,t)\rangle > 0$ for all $v \in S^k$ and $t \in [0,1]$, so $h(v,t) \neq 0$ for any $v, t \in S^k \times [0,1]$.

Since $g(v,0) = \tau(\psi(v))$, and $g(v,1) = v$, g is a homotopy from $\tau \circ \psi$ to the identity map on S^k.

Now, we want to show that $\psi \circ \tau$ is homotopic to the identity on $\mathbb{R}^{k+1} \setminus y$. To see this, let

$$g(x,t) = tx + (1-t)(\psi(\tau(x))).$$

Again, we need to show that this is well defined in the sense that $g(x,t) \neq y$ for any

$x, t \in (\mathbb{R}^{k+1} \setminus y) \times [0,1]$. Consider

$$\langle x - y, g(x,t) - y \rangle = t\langle x - y, x - y \rangle + (1-t)\langle x - y, \psi(\tau(x)) - y \rangle.$$

Now, note that $\langle x - y, x - y \rangle > 0$ for $x \in \mathbb{R}^{k+1} \setminus y$, and

$$\langle x - y, \psi \circ \tau(x) - y \rangle = \|x - y\|(\langle \tau(x), \psi \circ \tau(x) \rangle - \langle \tau(x), y \rangle) > 0,$$

where we apply Lemma 4.3.4 to the vector $\tau(x)$.

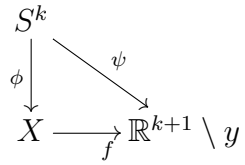
We conclude that $\langle x - y, g(x,t) - y \rangle > 0$ for all $x, t \in (\mathbb{R}^{k+1} \setminus y) \times [0,1]$, and so $g(x,t) \neq y$, as desired. $g(x,t)$ is also clearly continuous on its domain. Since $g(x,0) = \psi \circ \tau(x)$ and $g(x,1) = x$, we conclude that g is a homotopy from $\psi \circ \tau$ to the identity. □

4.3.2 Proof of Theorem 4.3.1

Let $Y = \operatorname{conv} f(X)$ be the convex hull of $f(X)$. We will show that $f(X) = Y$.

Clearly, if Y is a single point, then the result holds. We can thus apply Lemma 4.3.3 to say that the interior Y^o is nonempty. Now, suppose that there is some $y \in Y^o$ so that $y \notin f(X)$, so that we may think of f as a map from X to $\mathbb{R}^{k+1} \setminus \{y\}$.

If $\psi = f \circ \phi$, then the following diagram of continuous maps commutes:



This translates to a commutative diagram of homotopy groups

$$\begin{array}{ccc} \pi_m(S^k) & & \\ {\scriptstyle \phi^*}\downarrow & \searrow^{\psi^*} & \\ \pi_m(X) & \xrightarrow[f^*]{} & \pi_m(\mathbb{R}^{k+1} \setminus y) \end{array},$$

where h^* denotes the map on homotopy groups induced by the continuous map h. Now, we make a few notes on this diagram: ψ is a homotopy equivalence by Lemma 4.3.5, so ψ^* is an isomorphism; in particular it is a nonzero map since $\pi_m(S^k)$ is nonzero. However, this is a contradiction, as $\psi^* = f^* \circ \phi^*$, and $f^* : \pi_m(X) \to \pi_m(\mathbb{R}^{k+1} \setminus y) \cong \pi_m(S^k)$ is the zero map by our assumption that $\pi_m(X)$ has no nonzero homomorphisms to $\pi_m(S^k)$.

We conclude that f does not define a map to $\mathbb{R}^{k+1} \setminus y$, i.e. that y must be in the image of f. This implies that $Y^o \subseteq f(X)$, and by compactness, this implies that the closure of Y^o is also contained in $f(X)$. Since Y is compact, convex, and has nonempty interior, Y is the closure of its interior, so $f(X) = Y$, as desired.

4.4 Examples of continuously maximized functions from noncrossing subspaces

Here, we will give some examples of continuously maximized functions arising from *noncrossing subspaces*, as defined in [50].

Firstly, let us give notation for some familiar objects from linear algebra. Recall that $\mathbb{F}^{n\times n}_{sym}$ denotes the space of $n \times n$ symmetric real matrices when $\mathbb{F} = \mathbb{R}$ and the space of $n \times n$ hermitian matrices when $\mathbb{F} = \mathbb{C}$. Let $\mathrm{O}_{\mathbb{F}}(n)$ denote the orthogonal group when $\mathbb{F} = \mathbb{R}$ and the unitary group when $\mathbb{F} = \mathbb{C}$. For matrices A and B of the same dimensions, we let $\langle A, B\rangle = \mathrm{Tr}(A^\dagger B)$.

For $x \in \mathbb{F}^{n\times n}_{sym}$, let $\lambda(x) \in \mathbb{R}^n$ be the vector of eigenvalues of x (counting multiplicity)

in descending order. Fix $\mu \in \mathbb{R}^n$ with entries in descending order, and let

$$M_\mu^{\mathbb{F}} = \{x \in \mathbb{F}_{sym}^{n\times n} : \lambda(x) = \mu\} = \{U\mathrm{Diag}(\mu)U^\dagger : U \in \mathrm{O}_{\mathbb{F}}(n)\}.,$$

where $\mathrm{Diag}(\mu)$ denotes the diagonal matrix whose diagonal entries correspond to those of μ.

A linear subspace $L \subseteq \mathbb{F}_{sym}^{n\times n}$ is *noncrossing* if every nonzero matrix in L has only nondegenerate eigenvalues. More generally, we will say a linear subspace $L \subseteq \mathbb{F}_{sym}^{n\times n}$ is *k-weakly noncrossing* if every nonzero matrix in L has the property that its k largest eigenvalues are all nondegenerate. The *Von Neumann-Wigner noncrossing theorem* states that a generic 2 dimensional subspace of $\mathbb{R}_{sym}^{n\times n}$ is noncrossing, and that a generic 3 dimensional subspace of $\mathbb{C}_{sym}^{n\times n}$ is noncrossing. Higher dimensional noncrossing subspaces are not common, and noncrossing subspaces of dimension greater than 2 only exist for certain values of n due to Adam's theory of linearly independent vector fields on spheres [50].

Lemma 4.4.1. *Let $\mu \in \mathbb{R}^n$ be a vector with entries in descending order so that $\mu_i = 0$ for $i > k$. Let $A_1, \ldots, A_{d+1} \in \mathbb{F}_{sym}^{n\times n}$ be linearly independent elements of a k-weakly noncrossing subspace of $\mathbb{F}^{n\times n}$, then the function $f : M_\mu^{\mathbb{F}} \to \mathbb{R}^{d+1}$ defined by*

$$f(x) = (\langle A_1, x\rangle, \ldots, \langle A_{d+1}, x\rangle)$$

is continuously maximized.

Proof. For $v \in S^k$, let $A(v) = \sum_{i=1}^{d+1} v_i A_i$.

Because the A_i span a k-weakly noncrossing subspace, the largest k eigenvalues $\lambda_1(A(v)) > \cdots > \lambda_k(A(v))$ are nondegenerate on S^k. This implies that if we let $\nu_i(v)$ be any unit norm eigenvector of $A(v)$ with eigenvalue $\lambda_i(A(v))$, the rank 1 matrices $V_i(v) = \nu_i(v)\nu_i(v)^\dagger$ are continuous functions on S^k. [1]

[1]While this fact is standard, it can be seen explicitly by noting that $\lambda_i(\cdot)$ is a continuous function

We then define

$$\phi(v) = \sum_{i=1}^{k} \mu_i V_i(v).$$

Note that $\phi(v) \in M_\mu^{\mathbb{F}}$. ϕ is also clearly continuous because each V_i is.

We want to argue that f is continuously maximized by ϕ, i.e. that

$$\langle v, f(\phi(v))\rangle = \max_{x \in M_\mu^{\mathbb{F}}} \langle v, f(x)\rangle.$$

To see this, first note that

$$\langle v, f(x)\rangle = \sum_{i=1}^{k+1} v_i \langle A_i, x\rangle = \langle A(v), x\rangle.$$

Because $M_\mu^{\mathbb{F}}$ is invariant under $\mathrm{O}_{\mathbb{F}}(n)$,

$$\max_{x \in M_\mu^{\mathbb{F}}} \langle A(v), x\rangle = \max_{x \in M_\mu^{\mathbb{F}}} \langle \Lambda(v), x\rangle,$$

where $\Lambda(v) = \mathrm{Diag}(\lambda_1(A(v)), \ldots, \lambda_n(A(v)))$. It follows from the Schur component of the Schur-Horn that this is maximized when x is a diagonal matrix with ascending diagonal entries, i.e.

$$\max_{x \in M_\mu^{\mathbb{F}}} \langle A(v), x\rangle = \langle \Lambda(v), \mathrm{Diag}(\mu)\rangle = \langle v, f(\phi(v))\rangle.$$

□

We will also want to consider analogous concepts for singular values, rather than eigenvalues. For this, we say a subspace $L \subseteq \mathbb{F}^{n\times m}$ is *k-weakly singularly noncrossing* if the k largest singular values of every nonzero matrix in L are nondegenerate.

Let $\mathrm{SO}(n)$ be special orthogonal group, consisting of the elements of $\mathrm{O}_{\mathbb{R}}(n)$ which

on $\mathbb{F}_{sym}^{n\times n}$ and that $V(v) = \frac{1}{\mathrm{Tr}(X(v))} X(v)$, where X is the adjugate matrix of $A(v) - \lambda_i(A(v))I$.

have determinant 1. Let $R \in \mathbb{R}^{n\times m}$, and let

$$S_R = \{U_1 R U_2^\intercal : U_1 \in \mathrm{SO}(n), U_2 \in \mathrm{SO}(m)\}.$$

That is, S_R is the orbit of R under the action of $\mathrm{SO}(n) \times \mathrm{SO}(m)$ by multiplication on the left and right. The uniqueness properties of the singular value decomposition of a matrix implies that there is a unique element $D \in S_R$ which is diagonal and so that $D_{11} \geq D_{22} \geq \ldots D_{n-1n-1} \geq |D_{nn}|$, and we call these numbers the *special singular values* of the matrix R.

We will need a lemma regarding optimization of linear functions on S_R, which is an easy corollary of [52, Corollary 3] or [53].

Lemma 4.4.2. *Let $A, R \in \mathbb{R}^{n\times m}$, with $n < m$. Suppose that the special singular values of A are $a_1 \geq \cdots \geq |a_n|$, and that the special singular values of R are $r_1 \geq \cdots \geq |r_n|$. Then,*

$$\max_{X\in S_R} \mathrm{Tr}(A^\intercal X) = \sum_{i=1}^{n} a_i r_i.$$

Lemma 4.4.3. *Let $R \in \mathbb{R}^{n\times m}$ be diagonal so that $R_{i,i} = 0$ for $i > k$. Let $A_1, \ldots, A_{d+1}$ be linearly independent elements of a k-weakly singularly noncrossing subspace of $\mathbb{R}^{n\times m}$. Then the map*

$$f : S_R \to \mathbb{R}^{d+1}$$

defined by $f(x) = (\langle A_1, x\rangle, \ldots, \langle A_{d+1}, x\rangle)$ is continuously maximized.

Proof. As in the proof of Lemma 4.4.1, for $v \in S^k$, we let $A(v) = \sum_{i=1}^{d+1} v_i A_i$.

Because the A_i span a k-weakly singularly noncrossing subspace, the largest k singular values $\sigma_1(A(v)) > \cdots > \sigma_k(A(v))$ are nondegenerate on S^k. Let $u_i(v)$ and $w_i(v)$ be, respectively, the left and right unit singular vectors of $A(v)$ associated to the singular value $\sigma_i(A(v))$. The nondegeneracy of these singular values imply that

the functions

$$V_i(v) = u_i(v)w_i(v)^\mathsf{T}$$

are continuous functions on S^k.

We then define

$$\phi(v) = \sum_{i=1}^{k} R_{ii}V_i(v).$$

Note that $\phi(v) \in S_R$ because its special singular values are those of R. ϕ is also clearly continuous because each V_i is.

We want to argue that f is continuously maximized by ϕ, i.e. that

$$\langle v, f(\phi(v))\rangle = \max_{x \in S_R} \langle v, f(x)\rangle.$$

This follows because

$$\langle v, f(x)\rangle = \sum_{i=1}^{k+1} v_i \langle A_i, x\rangle = \langle A(v), x\rangle.$$

We may then apply Lemma 4.4.2 to see that

$$\max_{x \in S_R} \langle A(v), x\rangle = \sum_{i=1}^{n} \sigma_i(A(v))R_{ii} = \langle v, f(\phi(v))\rangle.$$

□

4.5 Some Hidden Convexity Theorems

Now that we have given some examples of continuously maximized functions, we can now see how to apply Theorem 4.3.1 in a few examples.

We start by stating the implication of Theorem 4.3.1 in light of Lemma 4.4.1. The proof of the following theorem is immediate given these two results:

Theorem 4.5.1. *Let $\mu \in \mathbb{R}^n$ be a vector with entries in descending order so that*

$\mu_i = 0$ for $i > k$. Let $A_1, \ldots, A_{d+1} \in \mathbb{F}^{n\times n}_{sym}$ be linearly independent elements of a k-weakly noncrossing subspace of $\mathbb{F}^{n\times n}$. If, for some m, $\pi_m(S^d) \neq 0$, and there are no nonzero group homomorphisms from $\pi_m(M^{\mathbb{F}}_\mu)$ to $\pi_m(S^d)$, then

$$\{(\langle A_1, x\rangle, \ldots, \langle A_1, x\rangle) : x \in M^{\mathbb{F}}_\mu\}$$

is convex.

First, we reprove Theorem 4.2.2 using our language.

Proof of Theorem 4.2.2. Fix $\mu \in \mathbb{R}^n$. Whether $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, it suffices to show the result for a dense set of choices of the linear maps T and vectors μ, since the remaining cases follow easily from continuity.

We start with the case in which $\mathbb{F} = \mathbb{R}$, so that we may write $T : \mathbb{R}^{n\times n}_{sym} \to \mathbb{R}^2$ in the form

$$T(X) = (\langle A_1, X\rangle, \langle A_2, X\rangle),$$

so that by Von Neumann-Wigner, a dense set of pairs $A_1, A_2 \in \mathbb{R}^{n\times n}_{sym}$ span a noncrossing subspace, so that by Lemma 4.4.1, we have that $T : M^{\mathbb{R}}_\mu \to \mathbb{R}^2$ is continuously maximized for a dense set of T.

Assuming that T is continuously maximized, we are in position to apply Theorem 4.3.1. It remains to show that $\pi_1(M^{\mathbb{R}}_\mu)$ is finite. This can seen quickly, as $M^{\mathbb{R}}_\mu$ is acted on transitively by $\mathrm{SO}(n)$, and when μ has distinct entries, the stabilizer of a point is finite, which implies that $M^{\mathbb{R}}_\mu$ is a finite quotient of a topological space with finite fundamental group. This implies that the fundamental group of $M^{\mathbb{R}}_\mu$ is finite.

We consider the case in which $\mathbb{F} = \mathbb{C}$, so that we may write $T : \mathbb{C}^{n\times n}_{sym} \to \mathbb{R}^3$ in the form

$$T(X) = (\langle A_1, X\rangle, \langle A_2, X\rangle, , \langle A_3, X\rangle),$$

so that by Von Neumann-Wigner, a dense set of triples $A_1, A_2, A_3 \in \mathbb{C}^{n\times n}_{sym}$ span

a noncrossing subspace, so that by Lemma 4.4.1, we have that $T : M_\mu^{\mathbb{C}} \to \mathbb{R}^3$ is continuously maximized for a dense set of T.

Assuming that T is continuously maximized, we are in position to apply Theorem 4.3.1. For this, we will show that $\pi_4(M_\mu^{\mathbb{C}}) = 0$ for $n \geq 3$, while $\pi_4(S^2) = \mathbb{Z}_2$ ([54]), so that Theorem 4.3.1 applies.

Note that $M_\mu^{\mathbb{C}}$ is acted on transitively by $\mathrm{O}_{\mathbb{C}}(n)$, and therefore is homeomorphic to $\mathrm{O}_{\mathbb{C}}(n)/\mathrm{O}_{\mathbb{C}}(n)_\mu$, where $\mathrm{O}_{\mathbb{C}}(n)_\mu$ denotes the stabilizer subgroup of μ. If the entries of μ are distinct, this stabilizer is $\mathrm{O}_{\mathbb{C}}(1)^n$, the group of diagonal matrices in $\mathrm{O}_{\mathbb{C}}(n)$. Therefore, $M_\mu^{\mathbb{C}}$ is homeomorphic to the *flag variety*, $\mathrm{O}_{\mathbb{C}}(n)/\mathrm{O}_{\mathbb{C}}(1)^n$. We then have that $M_\mu^{\mathbb{C}}$ can be thought of as a fiber bundle

$$\mathrm{O}_{\mathbb{C}}(1)^n \hookrightarrow \mathrm{O}_{\mathbb{C}}(n) \to M_\mu^{\mathbb{C}}$$

There is a long exact sequence of homotopy groups one of the entries of which is

$$0 \cong \pi_4(\mathrm{O}_{\mathbb{C}}(n)) \to \pi_4(M_\mu^{\mathbb{C}}) \to \pi_3(\mathrm{O}_{\mathbb{C}}(1)^n) \cong 0.$$

Here, the isomorphism $\pi_4(\mathrm{O}_{\mathbb{C}}(n)) \cong 0$ for $n \geq 3$ can be found in [55][Table 6.VII, Appendix A.] (for these dimensions, this lies in the 'stable range', which implies it is the same for all n larger than 3), and $\pi_3(\mathrm{O}_{\mathbb{C}}((1)^n) \cong 0$ is due to the fact that $\pi_3(\mathrm{O}_{\mathbb{C}}(1)) \cong 0$, and homotopy groups commute with products. □

We also reprove Theorem 4.2.1 here:

Proof of Theorem 4.2.1. In our language, we may take $A_1, \ldots, A_k$ to be linearly independent elements of a 1-weakly noncrossing subspace of $\mathbb{F}_{sym}^{n\times n}$. By [56], we have that $k < n$.

Let $\mu = \begin{pmatrix} 1 \\ 0 \\ \cdots \\ 0 \end{pmatrix} \in \mathbb{R}^n$. Let $\tau : S^d \to M_\mu^{\mathbb{F}}$ be such that $\tau(x) = xx^\intercal$. Note that

$$\{(x^\intercal A_1 x, x^\intercal A_2 x, \ldots, x^\intercal A_k x) : \|x\| = 1\} = \{(\langle A_1, x\rangle, \langle A_2, x\rangle, \ldots, \langle A_k, x\rangle) : x \in M_\mu^{\mathbb{F}}\}.$$

Now by Theorem 4.5.1, since $\pi_{k-1}(S^{k-1}) = \mathbb{Z}$, it suffices to show that there are no nonzero group homomorphisms from $\pi_k(M_\mu^{\mathbb{R}})$ to $\mathbb{Z}$. Note that τ is a double covering, so that $\pi_{k-1}(M_\mu^{\mathbb{F}})$ is finite, since $\pi_{k-1}(S^n)$ is finite. Therefore, this homotopy theoretic condition is satisfied, and we may conclude the theorem. □

Next, we prove Theorem 4.2.3.

Proof of Theorem 4.2.3. Without loss of generality, we may take $n \leq m$. By combining Lemma 4.4.3 and the Von Neumann-Wigner theorem for singular values, we may take T to be continuously maximized, as the set of such maps are dense. We may also assume that R is generic, so that R has distinct singular values. It remains to show that $\pi_1(S_R)$ is finite.

For this, let $\mathrm{St}_{m,n}$ denote the Stiefel manifold, consisting of $m \times n$ matrices whose columns are orthogonal (and if $m = n$, we let $\mathrm{St}_{n,n} = \mathrm{SO}(n)$). We may think of S_R as being the quotient of $\mathrm{SO}(n) \times \mathrm{St}_{m,n}$ by a finite group action. Specifically, if we let D denote the diagonal matrix of special singular values of R, then we have a covering map $c : \mathrm{SO}(n) \times \mathrm{St}_{m,n} \to S_R$ given by

$$c(U, V) = UDV^\intercal.$$

Because of uniqueness properties of the singular value decomposition, $c(U, V) = c(U', V')$ if and only if there exists a signed permutation matrix P in $\mathrm{SO}(n)$ so

that $U' = PU$ and $V' = PV$. From this, it follows that S_R is a finite quotient of $\mathrm{SO}(n) \times \mathrm{St}_{m,n}$, so it has finite fundamental group if and only if $\mathrm{SO}(n) \times \mathrm{St}_{m,n}$ does.

We then note that $\pi_1(\mathrm{SO}(n) \times \mathrm{St}_{m,n}) = \pi_1(\mathrm{SO}(n)) \times \pi_1(\mathrm{St}_{m,n})$, and that $\pi_1(\mathrm{SO}(n)) = \mathbb{Z}_2$ for $n \geq 3$, and that when $m, n \geq 3$, $\pi_1(\mathrm{St}_{m,n})$ is either $\mathbb{Z}_2$ or trivial.

We conclude that S_R has finite fundamental group, so that the result follows from Theorem 4.3.1. □

Remark 10. *Note that there are likely other hidden convexity theorems that can be obtained about linear images of S_R than what was stated in Theorem 4.2.3. However, it is unclear what applications there are of such theorems, as singularly noncrossing subspaces are less well studied than noncrossing subspaces.*

4.6 Application to orientation finding

The contents of this section were originally proven in [51]. As an application, we will discuss *constrained* variants of Wahba's problem, which was first discussed in [51]. To set up Wahba's problem, imagine that a satellite in space wants to determine its relative rotation (with respect to a reference rotation) given the observed direction of some number of far-away stars (or other objects).

Formally, we are given a set of (unit) vectors $v_1, \ldots, v_k \in \mathbb{R}^3$, corresponding to the known directions of the k stars in the reference rotation, and (unit) vectors $u_1, \ldots, u_k \in \mathbb{R}^3$, corresponding to the observed directions of the k stars in the satellite's frame. Our goal is to find a rotation minimizing the observation error

$$\begin{aligned} \min \quad & \sum_{i=1}^{k} \|Xu_i - v_i\|_2^2 \\ \text{s.t.} \quad & X \in \mathrm{SO}(3) \end{aligned} \tag{4.6.1}$$

In [57], it was observed that this is equivalent to a linear optimization problem over

SO(3)

$$\begin{aligned} \max \quad & \langle \sum_{i=1}^{k} u_i v_i^\intercal,\, X \rangle \\ \text{s.t.} \quad & X \in \mathrm{SO}(3). \end{aligned} \tag{4.6.2}$$

This problem can be solved using a singular value decomposition (SVD) computation, in the following sense: The *special singular value decomposition* of A is a decomposition $A = U\Sigma V^\intercal$, where Σ is a diagonal matrix with at most one negative entry, and $U, V \in \mathrm{SO}(n)$. It is easy to compute the special singular value decomposition of A given its usual SVD. Then if $A \in \mathbb{R}^{n\times n}$ has special SVD given by $A = U\Sigma V^\intercal$, then the maximizer of $\langle A, X\rangle$ for $X \in \mathrm{SO}(n)$ is given by $VU^\intercal$. For a given matrix A, we will denote by $\mathrm{str}(A)$ the maximum value of $\langle A, X\rangle$ for $X \in \mathrm{SO}(n)$.

Now, suppose we are given additional information about the true rotation X^*. We will incorporate this additional information as hard constraints into Wahba' problem to get a constrained optimization problem over SO(3).

For example, we may know that the true rotation X^* is within some angle, δ, of another rotation $X_0 \in \mathrm{SO}(3)$. In this case, we would need to solve the problem

$$\begin{aligned} \max \quad & \langle \sum_{i=1}^{k} u_i v_i^\intercal,\, X \rangle \\ \text{s.t.} \quad & \langle X_0, X\rangle \geq 1 + 2\cos(\delta) \\ & X \in \mathrm{SO}(3). \end{aligned} \tag{4.6.3}$$

In general, we will consider the constrained version of Wahba's problem, which for matrices $A, B \in \mathbb{R}^{n\times n}$ is given by

$$\begin{aligned} \max \quad & \langle A,\, X\rangle \\ \text{s.t.} \quad & \langle B, X\rangle \in [a, b] \\ & X \in \mathrm{SO}(n). \end{aligned} \tag{4.6.4}$$

The observation that we will make that makes this tractable is that if we let $f : \mathbb{R}^{n\times n} \to \mathbb{R}^2$ be given by $f(X) = (\langle A, X\rangle, \langle B, X\rangle)$, then Theorem 4.2.3 implies that $f(\mathrm{SO}(n))$ is in fact convex. Then, Equation (4.6.4) is equivalent to

$$\begin{aligned} \max \quad & x \\ \text{s.t.} \quad & y \in [a, b] \\ & (x, y) \in f(\mathrm{SO}(n)). \end{aligned} \tag{4.6.5}$$

We can then solve this convex problem using the *ellipsoid algorithm*. If $C \subseteq \mathbb{R}^n$ is a compact convex set and $x \notin C$, then there is a hyperplane that separates x and C. This *separating hyperplane* is given by a nonzero vector $y \in \mathbb{R}^n$ so that $\langle y, x\rangle \geq \max\{\langle y, c\rangle : c \in C\}$. A *$\epsilon$-weak separation oracle* for C is an oracle that on an input $x \in \mathbb{R}^n$, either correctly declares $x \in C + \mathbb{B}_\infty(0, \epsilon)$, or outputs $y \in \mathbb{R}^n$ so that y is a separating hyperplane between x and C. Here, $\mathbb{B}_\infty(a, r)$ is the ball of radius r in the L_∞ norm centered at a. The algorithmic equivalence between weak separation oracles and approximate optimization over convex sets is outlined in [58].

The ellipsoid algorithm as described in [59] provides the following guarantee for optimization in $\mathbb{R}^2$.

Theorem 4.6.1. *Suppose we have access to an ϵ-weak separation oracle for closed compact $C \subseteq \mathbb{R}^2$, we are given a $R \in \mathbb{R}$ so that $C \subseteq \mathbb{B}_2(0, R)$ and C includes a ball of radius at least ϵ. There is an algorithm that optimizes a linear function with unit L_2 norm over C within an additive error of ϵ using at most $O(\log(\frac{R}{\epsilon}))$ calls to the weak separation oracle.*

We now describe how the ellipsoid algorithm applies in this setting. Let $\|A\|_{\mathrm{Tr}}$ denote the sum of the singular values of A.

Theorem 4.6.2. *Let $n \geq 3$, A, $B \in \mathbb{R}^{n\times n}$ with $\|A\|_{\mathrm{Tr}} = \|B\|_{\mathrm{Tr}} = 1$. Here $\|\cdot\|_{\mathrm{Tr}}$ is the trace norm, defined as the sum of the singular values. Let X^* be the optimal solution*

to (4.6.4). *We can compute* $\langle A, X^* \rangle$ *and* $\langle B, X^* \rangle$ *within an additive error of* ϵ *in time*

$$O\left(n^3 \log\left(\frac{1}{\epsilon}\right)^2\right).$$

Here, n^3 *is the time complexity of computing the SVD of an* $n \times n$ *matrix.*

Moreover, we will return $\alpha, \beta \in \mathbb{R}$ *so that* $|\alpha| + |\beta| = 1$ *and*

$$\langle \alpha A + \beta B, X^* \rangle + \epsilon \geq \max\{\langle \alpha A + \beta B, X \rangle : X \in SO(n)\}.$$

Remark 11. *Let* α, β *denote the quantities returned in Theorem 4.6.2. While Theorem 4.6.2 does not directly return a minimizer of* (4.6.4)*, we believe that any element of*

$$\operatorname*{argmax}_{X \in SO(n)} \langle \alpha A + \beta B, X \rangle$$

should be a good approximation of a true minimizer under mild conditions. Such an element can be computed from $\alpha A + \beta B$ *in the time of a single SVD decomposition. Analyzing this procedure is outside the scope of the current paper and we leave this question for future work.*

Theorem 4.6.1 implies that to show Theorem 4.6.2, it suffices to provide a weak separation oracle for the set $C = \pi(\mathrm{SO}(n)) \cap \{x_2 \in [a, b]\}$. As the second constraint is trivial to separate over, we focus on separating over the convex set $\pi(\mathrm{SO}(n))$.

We denote by $f^* : \mathbb{R}^2 \to \mathbb{R}^{n \times n}$ the dual map to the map f defined above. It will be useful to have a subroutine for minimizing $h(y)$ over the unit ℓ_1-ball in $\mathbb{R}_2$, where

$$h(y) \coloneqq \mathrm{str}(f^*(y)) - \langle y, x \rangle.$$

Lemma 4.6.3. *Given* $x \in \mathbb{R}^2$ *with* $\|x\|_\infty \leq 1 + \epsilon$*, we can construct* $\widehat{y}$ *with* $\|\widehat{y}\|_1 = 1$

so that

$$h(\widehat{y}) - \epsilon \leq \min_{y} \{h(y) : \|y\|_1 = 1\}$$

using at most $O\left(\log\left(\frac{1}{\epsilon}\right)\right)$ *evaluations of* h *and additional computations.*

Proof. We note that $\{y : \|y\|_1 = 1\}$ is a union of 4 line segments, so minimizing h on this set can be done by minimizing the following 4 univariate functions on $[0, 1]$:

$$g_{\sigma_1\sigma_2}(\alpha) = h(\sigma_1\alpha, \sigma_2(1-\alpha)) = \mathrm{str}(\sigma_1\alpha A + \sigma_2(1-\alpha)B) - \sigma_1\alpha x_1 - \sigma_2(1-\alpha)x_2,$$

indexed by $\sigma_1, \sigma_2 \in \{\pm 1\}$. Each of the four functions $g_{\sigma_1\sigma_2}$ is a one-dimensional convex function with Lipschitz constant bounded by

$$\|A\|_{\mathrm{Tr}} + \|B\|_{\mathrm{Tr}} + \|x\|_1 \leq 4 + 2\epsilon.$$

For each $\sigma_1\sigma_2 \in \{\pm 1\}$, we may use golden section search [60] to find a $\widehat{\alpha}_{\sigma_1\sigma_2} \in [0, 1]$ such that

$$g_{\sigma_1\sigma_2}(\widehat{\alpha}_{\sigma_1\sigma_2}) \leq \min_{\alpha\in[0,1]} g_{\sigma_1\sigma_2}(\alpha) + \epsilon.$$

Each application of golden section search requires $O\left(\log\left(\frac{1}{\epsilon}\right)\right)$ evaluations of $g_{\sigma_1\sigma_2}$, or equivalently, evaluations of h. □

Lemma 4.6.4. *Let* $n \geq 3$ *and* $A, B \in \mathbb{R}^{n\times n}$ *with* $\|A\|_{\mathrm{Tr}} = \|B\|_{\mathrm{Tr}} = 1$. *There is a weak separation oracle for the set* $f(SO(n))$ *that runs in time* $O(n^3 \log(\frac{1}{\epsilon}))$.

Proof. Suppose we are given $A, B \in \mathbb{R}^{n\times n}$ and $x \in \mathbb{R}^2$. If $\|x\|_\infty > 1 + \epsilon$, then in fact, $x \notin f(\mathrm{SO}(n)) + \mathbb{B}_\infty(0, \epsilon)$ as, by Holder's inequality,

$$X \in \mathrm{SO}(n) \implies \max\{|\langle A, X\rangle|, |\langle B, X\rangle|\} \leq \|X\|_{op} \max\{\|A\|_{\mathrm{Tr}}, \|B\|_{\mathrm{Tr}}\} \leq 1,$$

where the last step was by our assumption $\|A\|_{\mathrm{Tr}} = \|B\|_{\mathrm{Tr}} = 1$. Therefore, in this case, we may immediately terminate with one of $(\pm 1, 0)$ or $(0, \pm 1)$ as a separating hyperplane. For the remainder, we assume that $\|x\|_\infty \leq 1 + \epsilon$.

A nonzero vector $y \in \mathbb{R}^2$ defines a separating hyperplane between x and $f(\mathrm{SO}(n))$ if and only if

$$\langle y, x \rangle \geq \max_{X \in \mathrm{SO}(n)} \langle y, f(X) \rangle.$$

Recalling the definition of $\mathrm{str}(\cdot)$, the expression on the right can be written as

$$\max_{X \in \mathrm{SO}(n)} \langle y, f(X) \rangle = \max_{X \in \mathrm{SO}(n)} \langle f^*(y), X \rangle = \mathrm{str}(f^*(y)).$$

Thus, a nonzero $y \in \mathbb{R}^2$ defines a separating hyperplane if and only if $h(y) \leq 0$. Note that we can compute $h(y)$ for a given y using a single SVD. As h is 1-homogeneous, such a y exists if and only if one exists with $\|y\|_1 = 1$.

Now, we apply Lemma 4.6.3 to compute $\widehat{y}$ approximately minimizing $h(y)$ on the unit ℓ_1 ball.

If $h(\widehat{y}) \leq 0$, then we may output $\widehat{y}$ as a separating hyperplane. For the remainder of the proof, suppose $h(\widehat{y}) > 0$. By Lemma 4.6.3 and 1-homogeneity, $h(y) > -\epsilon$ for all $y \in \mathbb{B}_1(0, 1)$. We claim that $x \in f(\mathrm{SO}(n)) + \mathbb{B}_\infty(0, \epsilon)$. If, to the contrary, $x \notin f(\mathrm{SO}(n)) + \mathbb{B}_\infty(0, \epsilon)$, then by the separating hyperplane theorem, there would be some y so that

$$\begin{aligned} \langle y, x \rangle &\geq \max\{\langle y, c + \delta \rangle : c \in f(\mathrm{SO}(n)), \delta \in \mathbb{B}_\infty(0, \epsilon)\} \\ &= \max\{\langle y, c \rangle : c \in f(\mathrm{SO}(n))\} + \epsilon \|y\|_1. \end{aligned}$$

In particular, there would be some y with $\|y\|_1 = 1$ such that

$$h(y) = \mathrm{str}(f^*(y)) - \langle y, x \rangle \leq -\epsilon,$$

which is a contradiction. □

CHAPTER 5

LONG STEP GRADIENT DESCENT

5.1 Introduction to long step gradient descent

This work was originally put forward in [61], and has since been modified with the coauthors Ben Grimmer and Alex Wang.

We will consider minimizing a L-smooth convex function $f\colon \mathbb{R}^n \to \mathbb{R}$ via gradient descent. A function is L-smooth if it is continuously differentiable, and its gradient is L-Lipschitz. We will specifically consider iterations of the following form:

$$x_{i+1} = x_i - \frac{h_i}{L}\nabla f(x_i) \tag{5.1.1}$$

with (normalized) stepsizes $h_i > 0$ starting from some $x_0 \in \mathbb{R}^n$. We assume a minimizer $x_\star$ of f exists.

Our main goal is to show that there exists a sequence $h_i \in \mathbb{R}^T$ achieving the following rate (which we will state loosely here, and more precisely in Theorem 5.2.1):

$$f(x_T) - f(x_\star) = O\left(\frac{1}{T^{1.27}}\right), \tag{5.1.2}$$

where the O hides constants related to the starting point x_0.

The main tool that we will be using for this is a collection of quadratic inequalities that characterizes first order information associated to a smooth convex function at a finite set of points. Formally, we will note the following basic inequalities for smooth convex functions: if f is L-smooth, and $\|\cdot\|$ denotes the ℓ_2 norm, then for

any $x, y \in \mathbb{R}^n$,

$$f(x) - f(y) \geq \langle \nabla f(y), x - y \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2$$

The idea of systematically using these inequalities to analyze the convergence of gradient based methods was is proposed in [62]. In fact, a strong converse to this set of inequalities was proven in [63]:

Theorem 5.1.1. *Let $x_1, \ldots, x_m \in \mathbb{R}^n$, let $f_1, \ldots, f_m \in \mathbb{R}$ and let $g_1, \ldots, g_m \in \mathbb{R}^n$. Then, there is an L-smooth convex function f so that $f(x_i) = f_i$ and $\nabla f(x_i) = g_i$ for all $i \in [m]$ if and only if for each $i, j \in [m]$,*

$$f_i - f_j \geq \langle g_j, x_i - x_j \rangle + \frac{1}{2L} \|g_i - g_j\|^2.$$

The main idea we will be using to prove our convergence rates is *taking nonnegative combinations of these inequalities*. In the rest of this section, we will note prior work on the subject and then give our step size sequence.

5.1.1 Prior work

When utilizing *constant stepsizes*, until recently, the best known guarantee was the textbook result [64] that fixing $h_i = 1$ ensures $f(x_T) - f(x_\star) \leq LD^2/2T$. This was improved by the tight convergence theory of Teboulle and Vaisbourd [65], showing a rate of

$$f(x_T) - f(x_\star) \leq \frac{LD^2}{4T}$$

when the stepsizes $h_i = 1$. Utilizing nonconstant stepsizes monotonically converging up to 2, they further showed a rate approaching $LD^2/8T$. These coefficient improvements were first conjectured by [63].

By utilizing *nonconstant periodically long stepsizes*, Grimmer [66] showed improved

convergence rates are possible outside the classic range of stepsizes $(0, 2)$. We refer to steps with $h_i > 2$ as long steps since they go beyond the classic regime $h_i \in (0, 2)$ where descent on the objective value is guaranteed. Their strongest result, resulting from a computer-aided semidefinite programming proof technique, showed repeating a cycle of 127 stepsizes $h_0, \ldots, h_{126}$ ranging from 1.4 to 370.0 gives a rate of

$$\min_{i \leq T} f(x_i) - f(x_\star) \leq \frac{LD^2}{5.83463T} + O(1/T^2).$$

Note, bounding $\min_{i \leq T} f(x_i) - f(x_\star)$ (or a similar quantity) is natural for such long step methods as monotone decrease of the objective is no longer ensured. By considering longer and more complex patterns, increasing gains in the coefficient appear to follow. However, the reliance on numerically solving semidefinite programs with size depending on the pattern length limited this prior work's ability to explore and prove continued improvements in convergence rates. Grimmer conjectured at least a $O(1/T \log(T))$ rate would follow if one could design and analyze (algebraically) cyclic patterns of generic length.

Das Gupta et. al. [67] produced numerically globally optimal stepsize selections via a branch-and-bound procedure for gradient descent with a fixed number of steps $T \in [0, 25]$. By fitting to asymptotics of their numerical guarantees [67, Figure 2], they conjectured a $O(1/T^{1.178})$ rate may be possible and may be best possible. Our work leaves open the gap between our $O(1/T^{1.0564})$ rate and their conjecture, as well as the gap between their conjecture and the known lower bound for general gradient methods of $O(1/T^2)$ [68].

At the same time that this work was released, Altschuler and Parrilo [69, 70] also showed an accelerated rate through the inclusion of long steps. In their second work in this series, they showed an improved convergence rate of roughly $\frac{1}{2T^{1.2716}}$.

Stronger guarantees for gradient descent with variable stepsizes are known in

specialized settings, like μ-strongly convex minimization. Classically, gradient descent with constant stepsizes $h_i = 1/L$ produces an ϵ-minimizer in $O(\kappa \log(1/\epsilon))$ iterations, where $\kappa = L/\mu$.

In the further specialized case of minimizing strongly convex quadratics, the optimal stepsizes were given by [71], which attain the optimal $O(\kappa^{1/2} \log(1/\epsilon))$ rate. For nonconvex optimization, exact worst-case guarantees for gradient descent with short steps $h_k \in (0, 1]$ were given by Abbaszadehpeivasti et al. [72].

5.1.2 The Proposed Stepsizes

In this subsection, we will define the step sizes which we will prove achieve accelerated convergence for smooth convex functions. Before we do this, we will need some preliminary definitions.

A quantity which appears prominently in our work (as well as that in [70]) is the *silver ratio*, defined as $\rho = 1+\sqrt{2}$. We will also define the numbers $\beta_i = 1+(1+\sqrt{2})^{i-1}$.

For $i \in \mathbb{N}$, we let $\nu(i)$ be the largest k so that 2^k divides i with the convention that $\nu(0) = \infty$. This is sometimes also known as the 2-adic valuation. For $\ell \geq 0$, let $\pi^{(\ell)} \in \mathbb{R}^{2^\ell - 1}$ be the vector where $\pi_i^{(\ell)} = \beta_{\nu(i)}$, with the convention that $\pi^{(0)} = []$, the empty sequence. We list the first few vectors $\pi^{(\ell)}$ for concreteness:

$$\begin{aligned}
\pi^{(-1)} &= \emptyset \text{ is the empty vector,} \\
\pi^{(1)} &= [\beta_0], \\
\pi^{(2)} &= [\beta_0, \beta_1, \beta_0], \text{ and} \\
\pi^{(3)} &= [\beta_0, \beta_1, \beta_0, \beta_2, \beta_0, \beta_1, \beta_0].
\end{aligned}$$

This was shown to be a good sequence of step sizes for gradient descent in [70]. The step sizes we will consider are somewhat more complicated.

We will also need to define two sequences, α_i and μ_i in a mutually recursive fashion.

For $i \geq 0$, define

$$\mu_i \coloneqq 2\sum_{\ell=0}^{i-1} \alpha_\ell + \sum_{\ell=0}^{i-2} 2(2^{i-\ell-1}-1)\beta_\ell + 1.$$

Here, we say that the empty sum is 0, so that $\mu_0 = 1$, and $\mu_1 = 2\alpha_0 + 1$. In general, μ_i only depends on α_ℓ for $\ell \leq i-1$.

We may then inductively define

$$\begin{aligned} \alpha_i \coloneqq &\text{ the unique root larger than 1 of} \\ &q_i(x) \coloneqq 2(x-1)^2 + \mu_i(x-1) - \rho^i \mu_i. \end{aligned} \tag{5.1.3}$$

Note that $q_i(1) = -\rho^i \mu_i < 0$, so that q_i has a unique root larger than 1 and α_i is well-defined.

We will finally define our sequence $\mathfrak{h}^{(k)} \in \mathbb{R}^{2^k-1}$ inductively as follows: $h^{(1)} = [\frac{3}{2}]$, and for $k > 0$,

$$\mathfrak{h}^{(k)} = [\pi^{(k-1)}, \alpha_{k-1}, h^{(k-1)}].$$

Here, if $v_1, \ldots, v_k$ are vectors or numbers, then the notation $[v_1, \ldots, v_k]$ denotes the concatenation of these vectors. We also use the convention that $\mathfrak{h}^{(i)}$ is 0-indexed. As a historical note, the original version of this work in [61] considered a somewhat different sequence, which was discovered after extensive computer search. This original sequence satisfied a condition called 'straightforwardness', originally proposed by [66], which streamlined this computer search. Since then, this original sequence has been truncated and a neater proof of its convergence properties have been discovered, which we will give here.

Following this construction, the first four sequences are, for example, given by

$$\begin{aligned}
\mathfrak{h}^{(1)} &= \left[\frac{3}{2}\right], \\
\mathfrak{h}^{(2)} &= \left[\sqrt{2},\ 1+\sqrt{2}, \frac{3}{2}\right], \\
\mathfrak{h}^{(3)} &= \left[\sqrt{2}, 2, \sqrt{2}, -\frac{1}{2} - \sqrt{2} + \frac{3\sqrt{5}}{2} + \sqrt{10}, \sqrt{2}, 1+\sqrt{2}, \frac{3}{2}\right].
\end{aligned}$$

In Section 5.3, we will give algebraic equations relating these various constants and sequences. We will also provide bounds on how the quantities α_k and μ_k grow asymptotically.

Lemma 5.1.2. *For all $k \geq 1$, it holds that*

$$\begin{aligned}
\beta_k \leq \alpha_k \leq \beta_{k+1}, \\
2(1+\sqrt{2})^k \leq \mu_k
\end{aligned}$$

From this, we see these building block patterns not only have exponentially large steps occur periodically, but also have exponentially large average stepsizes.

5.2 Proof of convergence rate

To state our convergence theorem, it will be helpful to have defined yet another sequence. We will let $c^{(k)} \in \mathbb{R}^{2^k}$ be defined as follows: $c_1 = \begin{pmatrix} 1 & 1 \end{pmatrix}$ and for $k > 1$,

$$c^{(k)} = [\frac{1}{\sqrt{\mu_k}}\pi^{(k-1)}, \frac{1}{\sqrt{\mu_k}}\beta_k, c_{k-1}].$$

We will show the following:

Theorem 5.2.1. *Fix $f : \mathbb{R}^n \to \mathbb{R}$ to be a 1-smooth convex function with a global minimizer x_*, and let $x_0 \in \mathbb{R}^n$. Fix some $k \geq 1$, and let $x_{i+1} = x_i - \mathfrak{h}_i^{(k)} \nabla f(x_i)$ for $i = 0, \ldots, 2^k - 2$. Also define a sequence y_i as follows: $y_0 = x_0$ and $y_{i+1} =$*

$y_i - \sqrt{\mu_k} c_i^{(k)} \nabla f(x_i)$. *With these definitions, we have that*

$$f(x_{2^k-1}) - f(x_*) \leq \frac{\|x_0 - x_*\|^2 - \|y_{2^k-1} - x_*\|^2}{2\mu_k}.$$

Remark 12. *The sequence y_i is not quite the result of a gradient descent procedure with different step sizes, as the updates are made using the gradients of f at x_i and not y_i. Nevertheless, it is interesting that the gap between $f(x_{2^k-1}) - f(x_*)$ improves if $\|y_{2^k-1} - x_*\|^2$ is large. In particular, it is guaranteed that either $f(x_{2^k-1})$ is especially close to $f(x_*)$, or y_{2^k-1} is close to x_* in Euclidean distance.*

Because our sequences have length 2^n, if we let T be the number of steps in this sequence, then $\mu_k = \Theta(T^{\log_2(1+\sqrt{2})})$, where $\log_2(1+\sqrt{2}) = 1.27$.

As we have stated, we will prove this by combining the basic inequalities on the values of 1-smooth convex functions.

Fix some T. For each $i \in \{0, \ldots, T\}$, we will let $x_i, g_i \in \mathbb{R}^n$ and $f_i \in \mathbb{R}$ be indeterminants. We will define quadratic polynomials for $i \neq j \in \{0, \ldots, T\}$,

$$Q_{ij} = f_i - f_j - \langle g_j, x_i - x_j \rangle - \frac{1}{2}\|g_i - g_j\|^2 \in \mathbb{R}[f_i, f_j, g_i, g_j, x_i, x_j].$$

We will also define quadratic polynomials

$$Q_{i*} = f_i - f_* - \frac{1}{2}\|g_i\|^2, \text{ and}$$

$$Q_{*i} = f_* - f_i - \langle g_i, x_* - x_i \rangle - \frac{1}{2}\|g_i\|^2 \in \mathbb{R}[f_*, f_i, g_i, x_*, x_i].$$

Now, if f is a 1-smooth convex function, $x_0, \ldots, x_T \in \mathbb{R}^n$, and $g_i = \nabla f(x_i)$ for each $i \in \{0, \ldots, T\}$, then inputting these values into some Q_{ij} will result in nonnegative numbers. If, moreover, we set $x_i = x_0 - \sum_{i=0}^{i-1} h_i g_i$ for some constant $h \in \mathbb{R}^{T-1}$, as is the case when the x_i are the iterates arising from gradient descent, then Q_{ij} is a polynomial in just f and g variables for $i \neq j \in \{0, \ldots, n\}$, while Q_{*i} and Q_{i*} are

polynomials in the f variables, g variables and $x_* - x_0$.

If we set $x_i = x_0 - \sum_{i=0}^{i-1} h_i g_i$ for some fixed $h \in \mathbb{R}^{T-1}$, then we may then define the *conical hull* of these polynomials to be the set of all nonnegative linear combinations of the Q_{ij} for $i, j \in \{*, 0, \ldots, T\}$ (thought of as being polynomials in $\mathbb{R}[f_1, \ldots, f_n, g_1, \ldots, g_n, x_* - x_0]$). We will denote this conical hull by $\mathcal{Q}(h)$.

In this context, we will define $y_i = x_0 - \sum_{j=0}^{i-1} \sqrt{\mu_k} c_i^{(k)} g_i$. Our main theorem can be shown to be equivalent to the claim that

$$\frac{\|x_* - x_0\|^2 - \|y_{2^k-1} - x_*\|^2}{2\mu_k} + f_* - f_{2^k} \in \mathcal{Q}(\mathfrak{h}^{(k)}).$$

To prove this, we will need to first construct an auxiliary polynomial in $\mathcal{Q}(\mathfrak{h}^{(k)})$.

$$\begin{aligned} d_k = &\frac{1}{\sqrt{\mu_k}} \sum_{i=0}^{2^k-2} c_i^{(k)} (f_i - f_{2^k-1}) \\ &+ \frac{1}{\sqrt{\mu_k}} \sum_{i=0}^{2^k-1} \langle c_i^{(k)} g_i, y_i - x_i \rangle \\ &+ \sum_{i=0}^{2^k-1} \left(\frac{c_i^{(k)}}{2\sqrt{\mu_k}} - \frac{(c_i^{(k)})^2}{2} \right) \|g_i\|^2. \end{aligned}$$

Lemma 5.2.2. *For any $k \geq 1$, $d_k \in \mathcal{Q}(\mathfrak{h}^{(k)})$.*

Before we prove Lemma 5.2.2, we will show how we can use it to prove Theorem 5.2.1.

Proof of Theorem 5.2.1. Consider

$$\begin{aligned} a_k &= \sum_{i=0}^{2^k-1} c_i^{(k)} Q_{*i} \\ &= \sum_{i=0}^{2^k-1} c_i^{(k)} (f_* - f_i) \\ &\quad - \sum_{i=0}^{2^k-1} c_i^{(k)} \langle g_i, x_i - x_* \rangle \\ &\quad - \frac{1}{2} \sum_{i=0}^{2^k-1} c_i^{(k)} \|g_i\|^2 \end{aligned}$$

From this representation, it is clear that $a_k \in \mathcal{Q}(\mathfrak{h}^{(k)})$, since the entries of $c^{(k)}$ are all nonnegative. We also have from Lemma 5.3.4 that $\left(\sum_{i=0}^{2^k-1} c_i^{(k)}\right) = \sqrt{\mu_k}$.

Consider $\sqrt{\mu_k} a_k + \mu_k d_k$, which is

$$\mu_k (f_* - f_{2^k-1}) - \sum_{i=0}^{2^k-1} \langle \sqrt{\mu_k} c_i^{(k)} g_i, y_i - x_* \rangle - \sum_{i=0}^{2^k-1} \frac{(\mu_k)(c_i^{(k)})^2}{2} \|g_i\|^2 \geq 0.$$

Now, note that $\sqrt{\mu_k} c_i^{(k)} g_i = y_{i+1} - y_i$, so that

$$\begin{aligned} \sum_{i=0}^{2^k-1} \langle \sqrt{\mu_k} c_i^{(k)} g_i, y_i - x_* \rangle + \frac{1}{2} \sum_{i=0}^{2^k-1} (\sqrt{\mu_k} c_i^{(k)})^2 \|g_i\|^2 = \\ = \frac{1}{2} \sum_{i=0}^{2^k-1} (2\langle y_{i+1} - y_i, y_i - x_* \rangle + \|y_{i+1} - y_i\|^2) \\ = \frac{1}{2} \sum_{i=0}^{2^k-1} (\langle y_{i+1} - y_i, 2y_i - 2x_* \rangle + \langle y_{i+1} - y_i, y_{i+1} - y_i \rangle) \\ = \frac{1}{2} \sum_{i=0}^{2^k-1} \langle y_{i+1} - y_i, y_{i+1} + y_i - 2x_* \rangle \\ = \frac{1}{2} \sum_{i=0}^{2^k-1} (\|y_{i+1} - x_*\|^2 - \|y_i - x_*\|^2) \\ = \frac{\|y_0 - x_*\|^2 - \|y_{2^k-1} - x_*\|^2}{2} \\ = \frac{\|x_0 - x_*\|^2 - \|y_{2^k-1} - x_*\|^2}{2} \end{aligned}$$

We conclude that for any 1-smooth convex function f, with minimizer x_*,

$$f(x_{2^k-1}) - f(x_*) \leq \frac{\|x_0 - x_*\|^2 - \|y_{2^k-1} - x_*\|^2}{2\mu_k}.$$

Clearly, this implies Theorem 5.2.1. □

The remainder of this section is devoted to the proof of Lemma 5.2.2

5.2.1 Proof of Lemma 5.2.2

In order to prove Lemma 5.2.2, we will need to define an auxiliary polynomial.

$$b_k = \frac{1}{\rho}\sum_{i=0}^{2^k-2} \pi_i^{(k)}(f_i - f_{2^k-1}) - \frac{1}{2\rho}\sum_{i=0}^{2^k-2} \pi_i^{(k)}(\pi_i^{(k)} - 1)\|g_i\|^2 - \frac{\rho^{k-1}}{2}(\rho^k - 1)\|g_{2^k-1}\|^2.$$

Lemma 5.2.3. *For any $k \geq 1$, $b_k \in \mathcal{Q}(\pi^{(k)})$.*

Proof. We will prove this using induction. For notational convenience, we will let $T = 2^k - 1$, the length of $\pi^{(k)}$ and note that $2T + 1 = 2^{k+1} - 1$ is the length of $\pi^{(k+1)}$.

Firstly, as a base case, we have that

$$\begin{aligned}
b_1 &= \frac{1}{\rho}\pi_0^{(1)}(f_0 - f_1) \\
&\quad - \frac{1}{2\rho}\pi_0^{(1)}(\pi_0^{(1)} - 1)\|g_0\|^2 - \frac{1}{2}(\rho - 1)\|g_1\|^2 \\
&= \frac{\sqrt{2}}{\rho}f_0 - \frac{\sqrt{2}}{\rho}f_1 - \frac{\sqrt{2}}{2\rho^2}\|g_0\|^2 - \frac{1}{\sqrt{2}}\|g_1\|^2 \\
&= (f_0 - f_1 - \langle g_1, x_0 - x_1\rangle - \frac{1}{2}\|g_1 - g_0\|^2) \\
&\quad + \frac{1}{\rho}(f_1 - f_0 - \langle g_0, x_1 - x_0\rangle - \frac{1}{2}\|g_0 - g_1\|^2) \\
&= Q_{0,1} + \frac{1}{\rho}Q_{1,0}.
\end{aligned}$$

It is important to note that because the first and last T steps in $\pi^{(k+1)}$ are the same as $\pi^{(k)}$, if $p \in \mathcal{Q}(\pi^{(k)})$, then this implies that $p \in \mathcal{Q}(\pi^{(k+1)})$ and that the *shift of*

p, defined by adding 2^k to the indices of all variables in p, is in $\mathcal{Q}(\pi^{(k+1)})$. We will denote this shift by $S_k p$.

Next, for $k > 1$, we claim that

$$b_{k+1} = b_k + \rho^2 S_k b_k + \Delta,$$

where

$$\Delta = \sum_{i=T+1}^{2T} \pi_i^{(k+1)}(Q_{T\,i} + Q_{2T+1\,i}) + Q_{T\,2T+1} + \rho^{k-1} Q_{2T+1\,T}.$$

Note that $b_k + \rho^2 S_k b_k \in \mathcal{Q}(\pi^{(k+1)})$ by induction and our remarks about shifting. Therefore, if this equality holds, then $b_{k+1} \in \mathcal{Q}(\pi^{(k+1)})$, and the inductive step holds, and so does the conclusion of the theorem.

We expand as follows, using the observation that $\pi_i^{(k)} = \pi_i^{(k+1)} = \pi_{i+2^k}^{(k+1)}$ for $i = \{0, \ldots, T-1\}$:

$$\begin{aligned} b_k + \rho^2 S_k b_k = \frac{1}{\rho} \sum_{i=0}^{T-1} \pi_i^{(k+1)}(f_i - f_T) - \frac{1}{2\rho} \sum_{i=0}^{T-1} \pi_i^{(k+1)}(\pi_i^{(k+1)} - 1)\|g_i\|^2 - \frac{\rho^{k-1}}{2}(\rho^k - 1)\|g_T\|^2 \\ + \rho \sum_{i=T+1}^{2T} \pi_i^{(k+1)}(f_i - f_{2T+1}) - \frac{\rho}{2} \sum_{i=T+1}^{2T} \pi_i^{(k+1)}(\pi_i^{(k+1)} - 1)\|g_i\|^2 \\ - \frac{\rho^{k+1}}{2}(\rho^k - 1)\|g_{2T+1}\|^2, \end{aligned}$$

and

$$\begin{aligned} \Delta = \sum_{i=T+1}^{2T} \pi_i^{(k+1)}(f_T + f_{2T+1} - 2f_i) \\ - \sum_{i=T+1}^{2T} \pi_i^{(k+1)}(\langle g_i, x_T + x_{2T+1} - 2x_i \rangle + \frac{1}{2}\|g_T - g_i\|^2 + \frac{1}{2}\|g_{2T+1} - g_i\|^2) \\ + (f_T - f_{2T+1} - \langle g_{2T+1}, x_T - x_{2T+1} \rangle - \frac{1}{2}\|g_T - g_{2T+1}\|^2) \\ + \rho^{k-1}(f_{2T+1} - f_T - \langle g_T, x_{2T+1} - x_T \rangle - \frac{1}{2}\|g_{2T+1} - g_T\|^2) \end{aligned}$$

We note the following about this expression: $x_T + x_{2T+1} - 2x_i = \sum_{j=T}^{i-1} \pi_j^{(k+1)} g_i - \sum_{j=i}^{2T} \pi_j^{(k+1)} g_i$, so

$$\begin{aligned}\sum_{i=T+1}^{2T} \pi_i^{(k+1)} \langle g_i, x_T + x_{2T+1} - 2x_i \rangle &= \sum_{i=T+1}^{2T} \sum_{j=T}^{i-1} \pi_i^{(k+1)} \pi_j^{(k+1)} \langle g_i, g_j \rangle \\ &\quad - \sum_{i=T+1}^{2T} \sum_{j=i}^{2T} \pi_i^{(k+1)} \pi_j^{(k+1)} \langle g_i, g_j \rangle \\ &= \langle \sum_{i=T+1}^{2T} \pi_i^{(k+1)} g_i, \pi_T^{(k+1)} g_T \rangle - \sum_{i=T+1}^{2T} (\pi_i^{(k+1)})^2 \|g_i\|^2.\end{aligned}$$

We also note that

$$\begin{aligned}\sum_{i=T+1}^{2T} \pi_i^{(k+1)} (\|g_i - g_T\|^2 + \|g_{2T+1} - g_i\|^2) &= \sum_{i=T+1}^{2T} \pi_i^{(k+1)} (2\|g_i\|^2 + \|g_T\|^2 + \|g_{2T+1}\|^2) \\ &\quad - 2\langle \sum_{i=T+1}^{2T} \pi_i^{(k+1)} g_i, g_T + g_{2T+1} \rangle.\end{aligned}$$

Note that $\sum_{i=T}^{2T+1} \pi_i^{(k)} g_i = x_T - x_{2T+1}$. Combining and noting that the expressions of

the form $\langle g_i, x_{2^k-1} - x_{2^{k-1}-1}\rangle$ cancel, yields that

$$\begin{aligned}
\Delta = &\sum_{i=T+1}^{2T} \pi_i^{(k+1)}(f_T + f_{2T+1} - 2f_i) \\
&- \langle x_T - x_{2T+1} - \pi_T^{(k+1)} g_T, \pi_T^{(k+1)} g_T\rangle + \sum_{i=T+1}^{2T} (\pi_i^{(k+1)})^2 \|g_i\|^2 \\
&- \frac{1}{2}\sum_{i=T+1}^{2T} \pi_i^{(k+1)}(2\|g_i\|^2 + \|g_T\|^2 + \|g_{2T+1}\|^2) \\
&+ \langle x_T - x_{2T+1} - \pi_T^{(k+1)} g_T, g_T + g_{2T+1}\rangle \\
&+ (f_T - f_{2T+1} - \langle g_{2T+1}, x_T - x_{2T+1}\rangle - \frac{1}{2}\|g_T - g_{2T+1}\|^2) \\
&+ \rho^{k-1}(f_{2T+1} - f_T - \langle g_T, x_{2T+1} - x_T\rangle - \frac{1}{2}\|g_{2T+1} - g_T\|^2) \\
= &\left(\sum_{i=T+1}^{2T} \pi_i^{(k+1)} + 1 - \rho^{k-1}\right) f_T + \left(\sum_{i=T+1}^{2T} \pi_i^{(k+1)} - 1 + \rho^{k-1}\right) f_{2T+1} \\
&- 2\sum_{i=T+1}^{2T} \pi_i^{(k+1)} f_i - \frac{1}{2}\left(1 + \rho^{k-1} + \sum_{i=T+1}^{2T} \pi_i^{(k+1)} + 2\pi_T^{(k+1)}(\pi_T^{(k+1)} - 1)\right)\|g_T\|^2 \\
&- \frac{1}{2}\left(1 + \rho^{k-1} + \sum_{i=T+1}^{2T} \pi_i^{(k+1)}\right)\|g_{2T+1}\|^2 + \sum_{i=T+1}^{2T}\left(\pi_i^{(k+1)}(\pi_i^{(k+1)} - 1)\right)\|g_i\|^2.
\end{aligned}$$

From Lemma 5.3.4, and the fact that $\pi_{i+T+1}^{(k+1)} = \pi_i^{(k)}$ for $i \in \{0, \ldots, T\}$ makes this expression equal to

$$\begin{aligned}
\Delta = &\left(\rho^k - \rho^{k-1}\right)(f_T - f_{2T+1}) - 2\sum_{i=T+1}^{2T} \pi_i^{(k+1)}(f_i - f_{2T+1}) \\
&- \frac{1}{2}\left(\rho^{k-1} + \rho^k + 2\pi_T^{(k+1)}(\pi_T^{(k+1)} - 1)\right)\|g_T\|^2 \\
&+ \sum_{i=T+1}^{2T}\left(\pi_i^{(k+1)}(\pi_i^{(k+1)} - 1)\right)\|g_i\|^2 - \frac{1}{2}\left(\rho^k + \rho^{k-1}\right)\|g_{2T+1}\|^2.
\end{aligned}$$

Finally, we combine, apply the fact that $\rho - 2 = \frac{1}{\rho}$, and see that

$$\begin{aligned}
b_{k+1} = b_k + \rho^2 S_k b_k + \Delta &= \frac{1}{\rho}\sum_{i=0}^{T-1} \pi_i^{(k+1)}(f_i - f_T) + \left(\rho^k - \rho^{k-1}\right)(f_T - f_{2T+1}) \\
&\quad - \frac{1}{2\rho}\sum_{i=0}^{T-1} \pi_i^{(k+1)}(\pi_i^{(k+1)} - 1)\|g_i\|^2 \\
&\quad - \left(\frac{\rho^{k-1}}{2}(\rho^k - 1) + \frac{1}{2}(\rho^{k-1} + \rho^k) + \pi_T^{(k+1)}(\pi_T^{(k+1)} - 1)\right)\|g_T\|^2 \\
&\quad + (\rho - 2)\sum_{i=T+1}^{2T} \pi_i^{(k+1)}(f_i - f_{2T+1}) \\
&\quad - \sum_{i=T+1}^{2T}\left(\frac{\rho}{2}\pi_i^{(k+1)}(\pi_i^{(k+1)} - 1) - \left(\pi_i^{(k+1)}(\pi_i^{(k+1)} - 1)\right)\right)\|g_i\|^2 \\
&\quad - \left(\frac{\rho^{k+1}}{2}(\rho^k - 1) + \frac{1}{2}\left(\rho^k + \rho^{k-1}\right)\right)\|g_{2T+1}\|^2 \\
&= \frac{1}{\rho}\sum_{i=0}^{2T} \pi_i^{(k+1)}(f_i - f_T) - \frac{1}{2\rho}\sum_{i=0}^{2T} \pi_i^{(k)}(\pi_i^{(k)} - 1)\|g_i\|^2 \\
&\quad - \frac{\rho^k}{2}(\rho^{k+1} - 1)\|g_{2T+1}\|^2 \\
&= b_{k+1}.
\end{aligned}$$

Here, we have noted that $\left(\rho^k - \rho^{k-1}\right)(f_T - f_{2T+1}) = \frac{1}{\rho}\sum_{i=0}^{T} \pi_i^{(k+1)}(f_T - f_{2T+1})$, and applied Lemma 5.3.3. □

Next, we will prove our main result of this subsection.

Proof of Lemma 5.2.2. We will show this by induction. First note that $\mu_1 = 5$, $c_0^{(1)} = c_1^{(1)} = 1$, so that $y_1 = x_0 - 2g_0$.

$$d_1 = \frac{1}{2}(f_0 - f_1) - \frac{1}{4}\langle g_1, g_0\rangle - \frac{1}{4}\|g_0\|^2 - \frac{1}{4}\|g_1\|^2.$$

We recognize this as $\frac{1}{2}Q_{0,1} \in \mathcal{Q}(\mathfrak{h}^{(1)})$.

From here out, we let $T = 2^k - 1$, the length of $\mathfrak{h}^{(k)}$, so that $2T + 1 = 2^{k+1} - 1$ is the length of $\mathfrak{h}^{(k+1)}$.

One important note is that because the last T step sizes in the $\mathfrak{h}^{(k+1)}$ sequence are the same as $\mathfrak{h}^{(k)}$, we once again have that if $p \in \mathcal{Q}(\mathfrak{h}^{(k)})$, then $S_k p \in \mathcal{Q}(\mathfrak{h}^{(k+1)})$, where S_k is the shift operator that adds 2^k to all indices of variables. Also, because the first $2^k - 1$ steps in the $\mathfrak{h}^{(k+1)}$ are the same as $\pi^{(k)}$, we have that if $p \in \mathcal{Q}(\pi^{(k)})$, then $p \in \mathcal{Q}(\mathfrak{h}^{(k+1)})$.

We then claim that

$$d_{k+1} = S_k d_k + \frac{\rho}{\mu_{k+1}} b_k + \left(\frac{1}{\sqrt{\mu_k}} - \frac{1}{\sqrt{\mu_{k+1}}} \right) \Delta_k,$$

where

$$\Delta_k = \sum_{i=T+1}^{2T+1} c_i^{(k)} Q_{T\,i}.$$

By the inductive hypothesis, Lemma 5.2.3, and the fact that Δ is clearly in $\mathcal{Q}(\mathfrak{h}^{(k)})$, the conclusion will follow. This finishes the inductive step, and thus implies the conclusion of the theorem.

To apply the shift operator, we will need to expand the definition of y_i in d_{k-1}:

$$\begin{aligned} d_k = \frac{1}{\sqrt{\mu_k}} & \sum_{i=0}^{T-1} c_i^{(k)} (f_i - f_T) \\ & + \sum_{i=0}^{T} \langle c_i^{(k)} g_i, \sum_{j=0}^{i-1} \left(c_j^{(k)} - \frac{\mathfrak{h}_j^{(k)}}{\sqrt{\mu_k}} \right) g_j \rangle \\ & + \sum_{i=0}^{T} \left(\frac{c_i^{(k)}}{2\sqrt{\mu_k}} - \frac{(c_i^{(k)})^2}{2} \right) \|g_i\|^2. \end{aligned}$$

Shifting, we obtain

$$\begin{aligned} S_k d_k = &\frac{1}{\sqrt{\mu_k}} \sum_{i=T+1}^{2T} c_i^{(k+1)}(f_i - f_T) \\ &+ \sum_{i=T+1}^{2T+1} \langle c_i^{(k+1)} g_i, \sum_{j=T+1}^{i-1} \left(c_j^{(k+1)} - \frac{\mathfrak{h}_j^{(k+1)}}{\sqrt{\mu_k}} \right) g_j \rangle \\ &+ \sum_{i=T+1}^{2T+1} \left(\frac{c_i^{(k+1)}}{2\sqrt{\mu_k}} - \frac{(c_i^{(k+1)})^2}{2} \right) \|g_i\|^2. \end{aligned}$$

Here, it is important to note that the last 2^k entries of $c^{(k+1)}$ are the same as $c^{(k)}$, so we may substitute $c_i^{(k)}$ for $c_{i-2^{k-1}}^{(k-1)}$ for $i \in [2^{k-1}, 2^k - 1]$. A similar situation holds for similarly for $\mathfrak{h}^{k+1}$.

We will also recall

$$\begin{aligned} \frac{\rho}{\mu_{k+1}} b_k &= \frac{1}{\mu_{k+1}} \sum_{i=0}^{T-1} \pi_i^{(k)}(f_i - f_T) - \frac{1}{2} \sum_{i=0}^{T-1} \frac{\pi_i^{(k)}(\pi_i^{(k)} - 1)}{\mu_{k+1}} \|g_i\|^2 \\ &\quad - \frac{\rho^k(\rho^k - 1)}{2\mu_{k+1}} \|g_T\|^2 \\ &= \frac{1}{\sqrt{\mu_{k+1}}} \sum_{i=0}^{T-1} c_i^{(k+1)}(f_i - f_T) - \frac{1}{2} \sum_{i=0}^{T-1} \left((c_i^{(k+1)})^2 - \frac{c_i^{(k+1)}}{\sqrt{\mu_{k+1}}} \right) \|g_i\|^2 \\ &\quad - \frac{\rho^k(\rho^k - 1)}{2\mu_{k+1}} \|g_T\|^2 \end{aligned}$$

Here, we make use of the facts that $c_i^{(k+1)} = \frac{\pi_i^{(k)}}{\sqrt{\mu_k}}$ if $i \in \{0, \ldots, 2^{k-1} - 2\}$.

We next expand Δ_k:

$$\begin{aligned} \Delta_k = &\sum_{i=T+1}^{2T+1} c_i^{(k+1)}(f_T - f_i - \langle g_i, x_i - x_T \rangle) \\ &- \frac{1}{2} \sum_{i=T+1}^{2T+1} c_i^{(k+1)} \|g_T - g_i\|^2 \end{aligned}$$

We note that $x_T - x_i = \sum_{j=T}^{i-1} \mathfrak{h}_j^{(k+1)} g_j$, so that

$$\begin{aligned}\sum_{i=T+1}^{2T+1} c_i^{(k+1)} \langle g_i, x_i - x_T \rangle &= -\sum_{i=T+1}^{2T+1} c_i^{(k+1)} \langle g_i, \sum_{j=T}^{i-1} \mathfrak{h}_j^{(k+1)} g_j \rangle \\ &= -\sum_{i=T+1}^{2T+1} \sum_{j=T}^{i-1} \langle c_i^{(k+1)} g_i, \mathfrak{h}_j^{(k+1)} g_j \rangle.\end{aligned}$$

We also note that, since $\sum_{i=T+1}^{2T+1} c_i^{(k+1)} = \sum_{i=0}^{T} c_i^{(k)} = \sqrt{\mu_k}$ by Lemma 5.3.4,

$$\sum_{i=T+1}^{2T+1} c_i^{(k+1)} (\|g_T\|^2 - 2\langle g_T, g_i \rangle + \|g_i\|^2) =$$
$$\sqrt{\mu_k} \|g_T\|^2 - 2\langle g_T, \sum_{i=T+1}^{2T+1} c_i^{(k+1)} g_i \rangle + \sum_{i=T+1}^{2T+1} c_i^{(k)} \|g_i\|^2.$$

Finally, we will note that $c_j^{(k+1)} = \frac{\pi_j^{(k)}}{\sqrt{\mu_{k+1}}} = \frac{\mathfrak{h}_j^{(k)}}{\sqrt{\mu_{k+1}}}$ when $j < T$, and so in particular, we can write

$$0 = \sum_{i=0}^{2T+1} \langle c_i^{(k)} g_i, \sum_{j=0}^{\min i-1, T-1} \left(c_j^{(k)} - \frac{\mathfrak{h}_j^{(k)}}{\sqrt{\mu_k}} \right) \rangle.$$

Combining, we have that $S_k d_k + \frac{\rho}{\mu_{k+1}} b_k + \left(\frac{1}{\sqrt{\mu_k}} - \frac{1}{\sqrt{\mu_{k+1}}}\right) \Delta_k$ is

$$
\begin{aligned}
&\frac{1}{\sqrt{\mu_k}} \sum_{i=T+1}^{2T} c_i^{(k+1)} (f_i - f_T) \\
&\quad + \sum_{i=T+1}^{2T+1} \langle c_i^{(k+1)} g_i, \sum_{j=T+1}^{i-1} \left(c_j^{(k+1)} - \frac{\mathfrak{h}_j^{(k+1)}}{\sqrt{\mu_k}} \right) g_j \rangle \\
&\quad + \sum_{i=T+1}^{2T+1} \left(\frac{c_i^{(k+1)}}{2\sqrt{\mu_k}} - \frac{(c_i^{(k+1)})^2}{2} \right) \|g_i\|^2 \\
&\quad + \frac{1}{\sqrt{\mu_{k+1}}} \sum_{i=0}^{T-1} c_i^{(k+1)} (f_i - f_T) - \frac{1}{2} \sum_{i=0}^{T-1} \left((c_i^{(k+1)})^2 - \frac{c_i^{(k+1)}}{\sqrt{\mu_{k+1}}} \right) \|g_i\|^2 \\
&\quad - \frac{\rho^k (\rho^k - 1)}{2(\mu_{k+1})} \|g_T\|^2 \\
&\quad + \left(\frac{1}{\sqrt{\mu_k}} - \frac{1}{\sqrt{\mu_{k+1}}} \right) \sum_{i=T+1}^{2T+1} c_i^{(k+1)} (f_T - f_i) \\
&\quad + \left(\frac{1}{\sqrt{\mu_k}} - \frac{1}{\sqrt{\mu_{k+1}}} \right) \sum_{i=T+1}^{2T+1} \sum_{j=T}^{i-1} \langle c_i^{(k+1)} g_i, \mathfrak{h}_j^{(k+1)} g_j \rangle \\
&\quad - \frac{1}{2} \left(\frac{1}{\sqrt{\mu_k}} - \frac{1}{\sqrt{\mu_{k+1}}} \right) \left(\sqrt{\mu_k} \|g_T\|^2 - 2 \langle g_T, \sum_{i=T+1}^{2T+1} c_i^{(k)} g_i \rangle \right) \\
&\quad - \frac{1}{2} \left(\frac{1}{\sqrt{\mu_k}} - \frac{1}{\sqrt{\mu_{k+1}}} \right) \sum_{i=T+1}^{2T+1} c_i^{(k+1)} \|g_i\|^2.
\end{aligned}
$$

Carefully combining terms shows that this expression is d_{k+1}, as desired. □

5.3 Equations and bounds related to constants

Lemma 5.3.1. *For any k,*

$$\sum_{i=0}^{2^k-1} \pi_i^{(k)} = \rho^k - 1.$$

Proof. In $\pi^{(k)}$, there are 2^{k-1} entries which are equal to β_0, 2^{k-2} entries which are equal to β_1 and so on, so that

$$\sum_{i=0}^{2^k-1} \pi_i^{(k)} = \sum_{i=0}^{k-1} 2^{k-i-1} \beta_i = \sum_{i=0}^{k} 2^{k-i} (\rho^{i-1} + 1).$$

This sum is a geometric series, which can be easily computed. □

Lemma 5.3.2. *For any $k \geq 1$,*

$$\mu_k = \mu_{k-1} + 2(\alpha_{k-1} + \rho^k - 1)$$

Proof. It follows quickly from the definition that

$$\mu_k - \mu_{k-1} = 2\alpha_{k-1} + 2\rho^k - 2.$$

□

Lemma 5.3.3. *For any $k \geq 1$,*

$$2\rho^{k-1} + \sqrt{\mu_{k-1}\mu_k} = \mu_k.$$

Proof. This is equivalent to

$$\mu_{k-1}\mu_k = (\mu_k - 2\rho^{k-1})^2.$$

We note that it is clear from the definition of μ_k that $\mu_{k-1} + 2(\alpha_k + \rho^k - 1) = \mu_k$. Applying this identity results in

$$\mu_{k-1}(\mu_{k-1} + 2(\alpha_k + \rho^k - 1)) = (\mu_{k-1} + 2(\alpha_k - 1))^2.$$

Rearranging this identity yields

$$0 = 2((\alpha_k - 1)^2 + \mu_{k-1}(a - 1) - \rho^k\mu_{k-1}).$$

This is the defining equation for α_k. □

Lemma 5.3.4. *For any $k \geq 1$,*

$$\sum_{i=0}^{2^k-1} c_i^{(k)} = \sqrt{\mu_k}$$

Proof. We will show this by induction. As a base case, note that $\sum_{i=0}^{2^k-1} c_i^{(k)} = \sqrt{\mu_0} = 1$.

For the inductive step, note that

$$\sum_{i=0}^{2^{k+1}-1} c_i^{(k+1)} = \frac{\sum_{i=0}^{2^k-1} \pi_i^{(k)} + \beta_{k+1}}{\sqrt{\mu_{k+1}}} + \sum_{i=2^k}^{2^{k+1}-1} c_i^{(k)}.$$

We have that $\sum_{i=0}^{2^k-1} \pi_i^{(k)} = \rho^k - 1$, $\beta_{k+1} = \rho^k + 1$, and $\sum_{i=2^k}^{2^{k+1}-1} c_i^{(k)} = \sqrt{\mu_k}$, so that

$$\sum_{i=0}^{2^{k+1}-1} c_i^{(k+1)} = \frac{2\rho^k}{\sqrt{\mu_{k+1}}} + \sqrt{\mu_k}$$

This is equivalent to Lemma 5.3.3 □

Finally, we prove the asymptotics stated in Lemma 5.1.2.

Proof of Lemma 5.1.2. First we verify $\alpha_k \leq \beta_{k+1}$. The defining equation of α_k is that α_k is the unique root larger than 1 of q_k. It is clear that $\beta_{k+1} \geq 1$, thus to show $\alpha_k \leq \beta_{k+1}$ suffices to show that $q_k(\beta_{k+1}) > 0$. We compute

$$q_k(\beta_{k+1}) = 2(\beta_{k+1} - 1)^2 + \mu_k(\beta_{k+1} - 1) - (\beta_{k+1} - 1)\mu_k = 2(\beta_{k+1} - 1)^2 > 0.$$

We next note that

$$\mu_k \geq \sum_{\ell=0}^{k-2} 2(2^{k-\ell-1} - 1)\beta_\ell = \sqrt{2}(\rho^k - 1) - 2k$$

Next, we note that $\beta_k \leq \alpha_k$, which follows because $\sqrt{2} = \beta_1 \leq \alpha_1 = \frac{3}{2}$, and for

$k \geq 2$,

$$\begin{aligned} q_k(\beta_k) &= (\beta_k - 1)^2 + \mu_k(\beta_k - 1) - \rho^k \mu_k \\ &= \rho^{2(k-1)} - \sqrt{2}\rho^{k-1}\mu_k \\ &\leq \rho^{(k-1)}((1-2\rho)\rho^{k-1} + 2 + 2\sqrt{2}k) \end{aligned}$$

which is negative.

This leads to the improved lower bound that

$$\mu_k \geq 2 \sum_{i=0}^{2^k-1} \pi^{(k)} = 2(\rho^k - 1).$$

□

REFERENCES

[1] D. Carlson, "What are schur complements, anyway?" *Linear Algebra and its Applications*, vol. 74, pp. 257–275, 1986.

[2] F. Zhang, *The Schur complement and its applications*. Springer Science & Business Media, 2006, vol. 4.

[3] D. G. Wagner, "Multivariate stable polynomials: Theory and applications," *Bull. Amer. Math. Soc. (N.S.)*, vol. 48, no. 1, pp. 53–84, 2011.

[4] N. Anari, S. O. Gharan, A. Saberi, and M. Singh, "Nash social welfare, matrix permanent, and stable polynomials," in *8th Innovations in Theoretical Computer Science Conference*, ser. LIPIcs. Leibniz Int. Proc. Inform. Vol. 67, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2017, Art. No. 36, 12.

[5] J. Borcea and P. Brändén, "Applications of stable polynomials to mixed determinants: Johnson's conjectures, unimodality, and symmetrized fischer products," *Duke Mathematical Journal*, vol. 143, no. 2, pp. 205–223, 2008.

[6] J. Borcea and P. Brändén, "The Lee-Yang and Pólya-Schur programs. I. Linear operators preserving stability," *Invent. Math.*, vol. 177, no. 3, pp. 541–569, 2009.

[7] R. Pemantle, "Hyperbolicity and stable polynomials in combinatorics and probability," in *Current developments in mathematics, 2011*, Int. Press, Somerville, MA, 2012, pp. 57–123.

[8] J. Renegar, "Hyperbolic programs, and their derivative relaxations," Cornell University Operations Research and Industrial Engineering, Tech. Rep., 2004.

[9] Y. Nesterov and L. Tunçel, "Local superlinear convergence of polynomial-time interior-point methods for hyperbolicity cone optimization problems," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 139–170, 2016.

[10] O. Güler, "Hyperbolic polynomials and interior point methods for convex programming," *Mathematics of Operations Research*, vol. 22, no. 2, pp. 350–377, 1997.

[11] M. Kummer, D. Plaumann, and C. Vinzant, "Hyperbolic polynomials, interlacers, and sums of squares," *Math. Program.*, vol. 153, no. 1, Ser. B, pp. 223–245, 2015.

[12] J. Saunderson, "Certifying polynomial nonnegativity via hyperbolic optimization," *SIAM J. Appl. Algebra Geom.*, vol. 3, no. 4, pp. 661–690, 2019.

[13] I. Schur and G. Polya, "Über zwei arten von faktorenfolgen in der theorie der algebraischen gleichungen.," 1914.

[14] H. H. Bauschke, O. Güler, A. S. Lewis, and H. S. Sendov, "Hyperbolic polynomials and convex analysis," *Canad. J. Math.*, vol. 53, no. 3, pp. 470–488, 2001.

[15] R. Sanyal and J. Saunderson, "Spectral polyhedra," *arXiv preprint arXiv:2001.04361*, 2020.

[16] M. Kummer, "Spectral linear matrix inequalities," *Adv. Math.*, vol. 384, Paper No. 107749, 36, 2021.

[17] G. Blekherman, J. Lindberg, and K. Shu, "Symmetric hyperbolic polynomials," *arXiv preprint arXiv:2308.09653*, 2023.

[18] Y.-B. Choe, J. G. Oxley, A. D. Sokal, and D. G. Wagner, "Homogeneous multivariate polynomials with the half-plane property," in 1-2, vol. 32, 2004, pp. 88–187.

[19] J. Borcea and P. Brändén, "Multivariate Pólya-Schur classification problems in the Weyl algebra," *Proc. Lond. Math. Soc. (3)*, vol. 101, no. 1, pp. 73–104, 2010.

[20] G. Blekherman, M. Kummer, R. Sanyal, K. Shu, and S. Sun, "Linear principal minor polynomials: Hyperbolic determinantal inequalities and spectral containment," *International Mathematics Research Notices*, vol. 2023, no. 24, pp. 21 346–21 380, 2023.

[21] B. Reznick, "Extremal psd forms with few terms," *Duke mathematical journal*, vol. 45, no. 2, pp. 363–374, 1978.

[22] T. Weisser, B. Legat, C. Coey, L. Kapelevich, and J. P. Vielma, "Polynomial and moment optimization in julia and jump," in *JuliaCon*, 2019.

[23] B. Legat, C. Coey, R. Deits, J. Huchette, and A. Perry, "Sum-of-squares optimization in Julia," in *The First Annual JuMP-dev Workshop*, 2017.

[24] W. J. Welch, "Algorithmic complexity: Three np-hard problems in computational statistics," *Journal of Statistical Computation and Simulation*, vol. 15, no. 1, pp. 17–25, 1982.

[25] M. Magdon-Ismail, "Np-hardness and inapproximability of sparse pca," *Information Processing Letters*, vol. 126, pp. 35–38, 2017.

[26] E. Miller and B. Sturmfels, *Combinatorial commutative algebra*. Springer Science & Business Media, 2004, vol. 227.

[27] K. Shu, "Approximate psd-completion for generalized chordal graphs," *arXiv preprint arXiv:2107.11436*, 2021.

[28] R. Grone, C. R. Johnson, E. M. Sá, and H. Wolkowicz, "Positive definite completions of partial hermitian matrices," *Linear Algebra and its Applications*, vol. 58, pp. 109–124, 1984.

[29] G. Blekherman, R. Sinn, and M. Velasco, "Do sums of squares dream of free resolutions?" *SIAM J. Appl. Algebra Geom.*, vol. 1, no. 1, pp. 175–199, 2017.

[30] R. Fröberg, "On Stanley-Reisner rings," in *Topics in algebra, Part 2 (Warsaw, 1988)*, ser. Banach Center Publ. Vol. 26, PWN, Warsaw, 1990, pp. 57–70.

[31] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," *Journal of the ACM (JACM)*, vol. 42, no. 6, pp. 1115–1145, 1995.

[32] M. Fukuda, M. Kojima, K. Murota, and K. Nakata, "Exploiting sparsity in semidefinite programming via matrix completion i: General framework," *SIAM Journal on Optimization*, vol. 11, Feb. 1970.

[33] L. Vandenberghe and M. S. Andersen, "Chordal graphs and semidefinite optimization," *Foundations and Trends in Optimization*, vol. 1, no. 4, pp. 241–433, 2015.

[34] G. Blekherman and K. Shu, "Sums of squares and sparse semidefinite programming," *SIAM Journal on Applied Algebra and Geometry*, vol. 5, no. 4, pp. 651–674, 2021.

[35] M. Laurent, "The real positive semidefinite completion problem for series-parallel graphs," *Linear Algebra and its Applications*, vol. 252, no. 1-3, pp. 347–366, 1997.

[36] W. Barrett, C. R. Johnson, and P. Tarazaga, "The real positive definite completion problem for a simple cycle," *Linear Algebra and its Applications*, vol. 192, pp. 3–31, 1993.

[37] K. Shu, "Extreme nonnegative quadratics over stanley reisner varieties," *arXiv preprint arXiv:2106.13894*, 2021.

[38] G. Blekherman, S. S. Dey, K. Shu, and S. Sun, *Hyperbolic relaxation of k-locally positive semidefinite matrices*, 2021. arXiv: 2012.04031 [math.OC].

[39] K. Shu, "Quadratic programming with sparsity constraints via polynomial roots," *arXiv preprint arXiv:2208.11143*, 2022.

[40] O. Toeplitz, "Das algebraische analogon zu einem satze von fejér," *Mathematische Zeitschrift*, vol. 2, no. 1-2, pp. 187–197, 1918.

[41] F. Hausdorff, "Der wertvorrat einer bilinearform," *Mathematische Zeitschrift*, vol. 3, no. 1, pp. 314–316, 1919.

[42] C. Davis, "The toeplitz-hausdorff theorem explained," *Canadian Mathematical Bulletin*, vol. 14, no. 2, pp. 245–246, 1971.

[43] L. Brickman, "On the field of values of a matrix," *Proceedings of the American Mathematical Society*, vol. 12, no. 1, pp. 61–66, 1961.

[44] P. Binding, "Hermitian forms and the fibration of spheres," *Proceedings of the American Mathematical Society*, vol. 94, no. 4, pp. 581–584, 1985.

[45] E. Gutkin, E. A. Jonckheere, and M. Karow, "Convexity of the joint numerical range: Topological and differential geometric viewpoints," *Linear Algebra and its Applications*, vol. 376, pp. 143–171, 2004.

[46] Y. H. Au-Yeung and N.-K. Tsing, "An extension of the hausdorff-toeplitz theorem on the numerical range," *Proceedings of the American Mathematical Society*, vol. 89, no. 2, pp. 215–218, 1983.

[47] Y.-H. Au-yeung and N.-K. Tsing, "Some theorems on the generalized numerical ranges," *Linear and Multilinear Algebra*, vol. 15, no. 1, pp. 3–11, 1984.

[48] C.-K. Li and T.-Y. Tam, "Numerical ranges arising from simple lie algebras," *Canadian Journal of Mathematics*, vol. 52, no. 1, pp. 141–171, 2000.

[49] I. Pólik and T. Terlaky, "A survey of the s-lemma," *SIAM review*, vol. 49, no. 3, pp. 371–418, 2007.

[50] S. Friedland, J. W. Robbin, and J. H. Sylvester, *On the crossing rule*. University of Wisconsin-Madison. Mathematics Research Center, 1982.

[51] A. Ramachandran, K. Shu, and A. L. Wang, "Hidden convexity, optimization, and algorithms on rotation matrices," *arXiv preprint arXiv:2304.08596*, 2023.

[52] T.-Y. Tam, "Kostant's convexity theorem and the compact classical groups," *Linear and Multilinear Algebra*, vol. 43, no. 1-3, pp. 87–113, 1997.

[53] T.-Y. Tam, "A lie theoretic approach to thompson's theorems on singular values–diagonal elements and some related results," *Journal of the London Mathematical Society*, vol. 60, no. 2, pp. 431–448, 1999.

[54] A. Hatcher, *Algebraic topology*. Cambridge: Cambridge University Press, 2002, pp. xii+544, ISBN: 0-521-79160-X; 0-521-79540-0.

[55] K. Itō, *Encyclopedic dictionary of mathematics*. MIT press, 1993, vol. 1.

[56] S. Friedland and R. Loewy, "Subspaces of symmetric matrices containing matrices with a multiple first eigenvalue," *Pacific Journal of Mathematics*, vol. 62, no. 2, pp. 389–399, 1976.

[57] G. Wahba, "A least squares estimate of satellite attitude," *SIAM Rev.*, vol. 7, no. 3, pp. 409–409, 1965.

[58] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric algorithms and combinatorial optimization*. Springer Science & Business Media, 2012, vol. 2.

[59] M. Grötschel, L. Lovász, and A. Schrijver, "The ellipsoid method and its consequences in combinatorial optimization," *Combinatorica*, vol. 1, pp. 169–197, 1981.

[60] J. Kiefer, "Sequential minimax search for a maximum," vol. 4, no. 3, pp. 502–506, 1953.

[61] B. Grimmer, K. Shu, and A. L. Wang, "Accelerated gradient descent via long steps," *arXiv preprint arXiv:2309.09961*, 2023.

[62] Y. Drori and M. Teboulle, "Performance of first-order methods for smooth convex minimization: A novel approach," *Mathematical Programming*, vol. 145, pp. 451–482, 2012.

[63] A. Taylor, J. Hendrickx, and F. Glineur, "Smooth strongly convex interpolation and exact worst-case performance of first-order methods," *Mathematical Programming*, vol. 161, pp. 307–345, 2017.

[64] D. P. Bertsekas, "Convex optimization algorithms," 2015.

[65] M. Teboulle and Y. Vaisbourd, "An elementary approach to tight worst case complexity analysis of gradient based methods," *Math. Program.*, vol. 201, no. 1–2, pp. 63–96, 2022.

[66] B. Grimmer, "Provably Faster Gradient Descent via Long Steps," *arxiv:2307.06324*, 2023.

[67] S. D. Gupta, B. P. V. Parys, and E. Ryu, "Branch-and-bound performance estimation programming: A unified methodology for constructing optimal optimization methods," *Mathematical Programming*, 2023.

[68] Y. E. Nesterov, *Introductory Lectures on Convex Optimization - A Basic Course* (Applied Optimization). Springer, 2004, vol. 87, ISBN: 978-1-4613-4691-3.

[69] J. M. Altschuler and P. A. Parrilo, *Acceleration by stepsize hedging i: Multi-step descent and the silver stepsize schedule*, 2023. arXiv: 2309.07879 [math.OC].

[70] J. M. Altschuler and P. A. Parrilo, "Acceleration by stepsize hedging ii: Silver stepsize schedule for smooth convex optimization," *arXiv preprint arXiv:2309.16530*, 2023.

[71] D. Young, "On richardson's method for solving linear systems with positive definite matrices," *Journal of Mathematics and Physics*, vol. 32, no. 1-4, pp. 243–255, 1953. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sapm1953321243.

[72] H. Abbaszadehpeivasti, E. de Klerk, and M. Zamani, "The exact worst-case convergence rate of the gradient method with fixed step lengths for l-smooth functions," *Optimization Letters*, vol. 16, pp. 1649–1661, 2021.