Semidefinite representations with applications in estimation and inference

by

James Francis Saunderson
BE Electrical Engineering and BSc Mathematics, The University of Melbourne, 2008
S.M. Electrical Engineering and Computer Science, MIT, 2011

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical Engineering and Computer Science
at the Massachusetts Institute of Technology

June 2015

Signature of Author: ______________________________
Department of Electrical Engineering and Computer Science
April 27, 2015

Certified by: ______________________________
Pablo A. Parrilo
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by: ______________________________
Alan S. Willsky
Edwin Sibley Webster Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by: ______________________________
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Semidefinite representations with applications in estimation and inference

by James Francis Saunderson
BE Electrical Engineering and BSc Mathematics, The University of Melbourne, 2008
S.M. Electrical Engineering and Computer Science, MIT, 2011

Submitted to the Department of Electrical Engineering
and Computer Science on April 27, 2015
in Partial Fulfillment of the Requirements for the Degree
of Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

Semidefinite optimization problems are an expressive family of convex optimization problems that can be solved efficiently. We develop semidefinite optimization-based formulations and approximations for a number of families of optimization problems, including problems arising in spacecraft attitude estimation and in learning tree-structured statistical models.

We construct explicit exact reformulations of two families of optimization problems in terms of semidefinite optimization. The first family are linear optimization problems over the derivative relaxations of spectrahedral cones. The second family are linear optimization problems over rotation matrices, i.e. orthogonal matrices with unit determinant. We use our semidefinite description of linear optimization problems over rotation matrices to express a joint spin-rate and attitude estimation problem for a spinning spacecraft exactly as a semidefinite optimization problem.

For families of optimization problems that are, in general, intractable, one cannot hope for efficient semidefinite optimization-based formulations. Nevertheless, there are natural ways to develop approximations for these problems called semidefinite relaxations. We analyze one such relaxation of a broad family of optimization problems with multiple variables interacting pairwise, including, for instance, certain multivariate optimization problems over rotation matrices. We characterize the worst-case gap between the optimal value of the original problem and a particular semidefinite relaxation, and develop systematic methods to round solutions of the semidefinite relaxation to feasible points of the original problem. Our results establish a correspondence between the analysis of rounding schemes for these problems and a natural geometric optimization problem that we call the normalized maximum width problem.

We also develop semidefinite optimization-based methods for a statistical modeling problem. The problem involves realizing a given multivariate Gaussian distribution as the marginal distribution among a subset of variables in a Gaussian tree model. This is desirable because Gaussian tree models enjoy certain conditional independence relations that allow for very efficient inference. We reparameterize this realization problem

as a structured matrix decomposition problem and show how it can be approached using a semidefinite optimization formulation. We establish sufficient conditions on the parameters and structure of an underlying Gaussian tree model so that our methods can recover it from the marginal distribution on its leaf-indexed variables.

Thesis Supervisor: Pablo A. Parrilo
Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Alan S. Willsky
Edwin Sibley Webster Professor of Electrical Engineering and Computer Science

Acknowledgments

I have been very fortunate to have Pablo Parrilo and Alan Willsky as thesis advisors. They have given me great intellectual freedom and encouragment and have infected me with their intellectual enthusiasm and exposed me to a very broad range of ideas (most of which are still stewing in a giant confused soup in my head). They have provided opportunities for me to travel, introduced me to many people, and have always ensured I have been well supported financially[1]. I would like to thank Sanjoy Mitter for serving on my thesis committee, for encouraging me to focus on the big picture, and always being willing and available to talk.

I am grateful for Venkat Chandrasekaran's mentorship since my first years in graduate school. Matt Johnson single-handedly made 32-D572 a place I wanted to be (to work, learn, and laugh), and I look forward to my next project with Hamza Fawzi, whatever and whenever it may be.

I have met many other wonderful people during my time at LIDS. I shared an office (and interesting discussions) with Igor Kadota, Sidhant Misra, Rajat Talak, Theja Tulabandhula, and Kush Varshney. Rachel Cohen, Lynne Dell, Lisa Gaumond, Brian Jones, and Debbie Wright have always been there to provide help of all kinds. I would particularly like to thank Jennifer Donovan for encouraging me to be involved in LIDS, not just to be in LIDS. Elie Adam made organizing a conference enjoyable, and I was always happy to run into Christina Lee and George Chen. Being part of the Stochastic Systems Group gave me the chance to interact with, and learn from, Jason Chang, Jin Choi, Ying Liu, and Vincent Tan. I have also had the great benefit of time spent (in locations near and far) with Amir Ali Ahmadi, Ozan Candogan, Diego Cifuentes, Frank Permenter, Parikshit Shah, and Takashi Tanaka.

Beyond LIDS, I have enjoyed many good times with generous and kind people, most notably Lily and Matt Johnson, Been Kim, Alejandro Morales, and Nick Sheridan.

Finally, I would like to thank Jenny Huang for providing compelling reasons (and constant encouragement) to complete my degree, and most of all my family for always supporting me, in the most unobtrusive way, no matter what I chose to do.

[1] The research described in this thesis was funded in part by the Air Force Office of Scientific Research under grants FA9550-12-1-0287 and FA9550-11-1-0305.

Contents

Chapter 1

Introduction

Developing useful mathematical models of the behavior of natural and engineered systems is of fundamental importance in science and engineering. Probabilistic models, i.e. models that describe quantities of interest in terms of random variables, are particularly useful since they explicitly account for and describe uncertainty. When modeling, there is a natural tension between finding a model that fits the observations, and a model that is simple in the sense that it can be described succinctly.

Given a model, we typically use it to try to answer questions about a real system, by translating them into mathematical problems that we can try to solve algorithmically. For instance, we may want to determine the best possible prediction of the state of a system in the future, given the noisy partial observations we currently have. It is important that the answers to such inference queries can be computed or approximated efficiently, otherwise all we have done is translate a hard problem about the real world, into a computationally hard mathematical problem.

Many such inference queries can be phrased as optimization problems, or even families of related optimization problems where the particular instance may depend, e.g., on observed data. Optimization problems involve maximizing (or minimizing) some objective function subject to constraints that the variables in the problem must satisfy.

Computationally, the most desirable optimization problems are those for which the complexity of describing the problem and the computational complexity of globally solving the problem, are similar. In these cases, 'simple' models lead directly to efficient inference methods. These families of problems generally enjoy many good properties, such as convexity (to help us certify global optimality of solutions) and algebraic structure (underlying the development of efficient algorithms for their solution with guaranteed running times). More importantly, their good properties are transparent from their description. One family of optimization problems with these good properties are *semidefinite optimization problems*. These play a central role in this thesis.

Another situation that can occur is when the 'natural' formulation of a family of optimization problem obscures its good properties. In these cases it may be possible to *reformulate* the problems, describing them in a better way that immediately leads to good algorithms for their global solution. Chapter 4 is about reformulating certain families of ostensibly non-linear and non-convex optimization problems as instances of the much nicer class of semidefinite optimization problems.

A third situation occurs when a family of optimization problems can be stated succinctly but is, in general, difficult to solve (in a sense that could be formalized using ideas from computational complexity theory). In such cases there is a discrepancy between model complexity and the computational complexity of solving certain inference problems. To deal with this, it is natural to try to *approximate* the family of problems by problems that can be solved globally and efficiently, and to establish guarantees on the quality of the approximation. It is typical to seek approximations with the additional property that their optimal values always underestimate the objective value for minimization problems (or overestimate the objective value for maximization problems). These approximations that keep track of additional *global* bounds are called *relaxations*, and provide useful information that is not available to local optimization methods. Chapter 5 is related to understanding the approximation quality of a semidefinite optimization-based relaxation for a family of hard optimization problems.

Returning to the probabilistic modeling context, similar considerations apply directly to probabilistic models themselves, not just to optimization problems that arise in performing inference in such models. Indeed finding 'good' alternative descriptions (or approximations) of probabilistic models with good computational properties is of central importance. These alternative descriptions often arise by describing the given model as the marginal distribution among a subset of variables in a more structured *latent* model.

For example, one family of probabilistic models in which inference queries can be carried out very efficiently are Gaussian tree models. These are the subject of Chapter 6. One way to recognize these models is that the inverse of the covariance matrix is sparse, with the sparsity pattern being that of a tree. If we are given only the covariance among a subset of the variables of a Gaussian tree model (the other variables being unobserved), the inverse of the given covariance typically has no interesting sparsity structure. In other words, the form in which we get to see the model obscures the fact that it actually comes from a model with computationally beneficial structure. In this case, it may be possible to recover this better description of the given covariance, allowing us to take advantage of its good computational properties. Furthermore, if we are given an arbitrary covariance matrix, it is natural to try to *approximate* it with a covariance that has structure allowing us to perform inference queries efficiently. For instance, we may aim to approximate it as the marginal covariance among a subset of variables in a Gaussian tree model. Such an approximation problem is exactly what arises when learning a latent variable model from data.

1.1 Outline and Contributions

We now provide an outline of the thesis, summarizing the main content and contributions of each chapter. Details of related previous work are discussed separately in each chapter.

Chapter 2: Background

Chapter 2 provides a summary of some of the technical background and notation that appears in multiple places throughout the thesis. It includes basic information and facts about convex sets and functions, semidefinite optimization, hyperbolic optimization, and the interaction between convexity ideas and symmetry.

Section 2.2 of Chapter 2 has more of a tutorial nature than the other sections. It gives a high-level summary, via concrete examples, of much of the technical context for Chapters 3, 4, and 5. In particular it explains the way in which a family of non-linear and non-convex optimization problems can be transformed into the problem of maximizing a linear functional over a convex set. This basic (and well-known) transformation allows us to focus on good descriptions of convex sets, whenever we are more broadly interested in good descriptions of families of optimization problems. This provides justification for our focus on various families of convex sets in Chapters 3 and 4. Section 2.2 also discusses the idea of semidefinite relaxations and associated rounding schemes. This basic idea is central to the discussion of rounding schemes in Chapter 5.

Chapter 3: Semidefinite descriptions of derivative relaxations of spectrahedral cones

Spectrahedral cones are convex cones related to the feasible regions of semidefinite optimization problems. There is a very appealing way to construct a family of outer approximations to any spectrahedral cone, called *derivative relaxations* [102]. These outer approximations have many interesting algebraic and geometric properties. Among their remarkable properties is the way their faces relate to the faces of the spectrahedral cone from which they are constructed. These relaxations preserve low-dimensional faces, and successively relax high-dimensional faces of the original spectrahedral cone. This property may play an important role in future applications of these cones, since in many settings (e.g. structured linear inverse problems [28], convex approaches to combinatorial optimization problems [54]), we are most interested the solutions lying on low-dimensional faces of the feasible region (which often correspond to sparse vectors or low-rank matrices).

We can solve optimization problems involving spectrahedral cones using semidefinite optimization. However, it is not obvious, from the construction of derivative relaxations, that we should also be able to solve optimization problems over derivative relaxations of spectrahedral cones using semidefinite optimization. The main contribution of Chapter 3 is the construction of explicit semidefinite representations of the derivative relaxations of spectrahedral cone. The descriptions we give of these cones are all of polynomial size in the 'size' parameter of the spectrahedral cone and the 'relaxation' parameter of the derivative relaxation and are the first known representations (of any size) of these cones. Our constructions show that we can, indeed, solve optimization problems involving the derivative relaxations of spectrahedral cones using semidefinite optimization. Chapter 3 is based on the work in [115].

Chapter 4: Semidefinite descriptions of the convex hull of rotation matrices

The set of rotation matrices, i.e. $n \times n$ orthogonal matrices with determinant one, describes linear isometries of Euclidean space that *preserve orientation*. These matrices form a group under matrix multiplication, called the *special orthogonal group*, denoted $SO(n)$. Optimization problems over rotation matrices arise whenever we want to optimize over the configuration spaces of rigid bodies. Examples of rigid bodies of interest may be satellites [97], molecules [123], mobile robots [25], or cameras [128].

The main contribution of Chapter 4 is the construction of an explicit semidefinite description of the convex hull of $SO(n)$ of size 2^{n-1}. The construction is the first known semidefinite representation (of any size) of this convex body. These descriptions give a natural way to come up with semidefinite relaxations for problems involving multiple rotation-matrix-valued variables. Moreover, using these semidefinite descriptions, we show how to reformulate a family of optimization problems involving $n \times n$ rotation matrices and trigonometric polynomials as semidefinite optimization problems. In the case $n = 3$, optimization problems of this form occur in a joint attitude and spin-rate estimation problem for spacecraft [97]. This estimation problem is discussed briefly in Chapter 4 and in more detail in [116]. Chapter 4 is based on the work in [117].

Chapter 5: Rounding semidefinite relaxations for pairwise optimization problems

While Chapters 3 and 4 show how semidefinite optimization can *exactly* capture certain specific families of optimization problems, Chapter 5 focuses on how semidefinite optimization can be used to *approximate* a certain class of optimization problems. In particular it focuses on semidefinite relaxations of optimization problems in which

- there are multiple variables, each taking values in some subset $\mathcal{X}$ of $m \times d$ contractions, and
- the variables $X_i, X_j \in \mathcal{X}$ interact *pairwise* via terms in the objective function of the form $\mathrm{tr}(C_{ij} X_i^T X_j)$.

Examples of constraint sets $\mathcal{X}$ of interest include the following.

- Rotation matrices, i.e. $\mathcal{X} = SO(3)$. The problem class we study arises in the context of discrete-time optimal filtering problems on $SO(3)$ [118], cryo-electron miscroscopy [123] where the rotation-valued variables correspond to the orientations of many different molecules, and in pose estimation problems in robotics [25].

- Permutation matrices, i.e. $\mathcal{X}$ consists of $d \times d$ matrices with all entries either zero or one, and exactly one non-zero entry in each row and column. In this case the problem class we study appears naturally in the joint matching problems studied, for instance, in [66] and [30] in the context of computer graphics and computer vision.

We study a particular simple semidefinite relaxation of this general family of pairwise optimization problems. This is a semidefinite optimization problem that always produces upper bounds on the value of the maximization problems of interest. We are interested in understanding the (worst-case) ratio between the optimal value of the semidefinite optimization problem, and the optimal value of the original pairwise optimization problem over $\mathcal{X}^n$. We are also interested in *rounding*, i.e. constructing, from the solution of a semidefinite relaxation, feasible points for the original problem that have near-optimal objective values.

The main contributions of Chapter 5 are the following.

1. We characterize the worst possible gap between the optimal value of the semidefinite relaxation and the optimal value of the original pairwise optimization problem over $\mathcal{X}^n$ when the objective function is positive semidefinite and $\mathcal{X}$ has certain symmetry properties. The gap is exactly the optimal value of a geometric problem related to $\mathcal{X}$ and is *independent of n*. We call this geometric optimization problem the *normalized maximum width problem*.

2. We show how to construct a randomized rounding scheme from any feasible point of the normalized maximum width problem. If we can find a maximizer of the normalized maximum width problem, the corresponding rounding scheme is optimal (in an appropriate sense). Finally, the rounding scheme can be *computed efficiently* whenever we can maximize a linear functional over $\mathcal{X}$ efficiently.

One way to think about these results is as follows. If we can solve a *linear* optimization problem over $\mathcal{X}$, then we can approximately maximize a large class of convex *quadratic forms* over $\mathcal{X}^n$. Furthermore, the approximation factor depends only on a geometric quantity related to $\mathcal{X}$, but is *independent of n*. The results of Chapter 5 unify many special cases described in the literature (these are explicitly summarized in Section 5.4

of Chapter 5), as well as providing a systematic approach to the design and analysis of new rounding schemes (with certain optimality properties) for many problems.

Chapter 6: A convex approach to learning Gaussian latent tree models

In Chapter 6 we shift from reformulating (or approximating) optimization problems in terms of semidefinite optimization, and focus on the probabilistic modeling setting. In particular we develop methods to express multivariate Gaussian random variables in terms of the particularly tractable subclass of Gaussian tree models. Gaussian tree models are collections of jointly Gaussian random variables indexed by the vertices of a tree in which the edges of the tree describe certain additional relations, called conditional independence relations, among the random variables [72]. While such relations reduce the expressiveness of the model, they significantly improve the tractability of performing inference in the model. We do not restrict ourselves to tree models in which the variables are scalar, so this is a flexible class of models in which inference is very efficient only if all the variables have small dimensions.

Much more expressive, and still computationally tractable, are *Gaussian latent tree models* (see, e.g. [93, 31]). These are the jointly Gaussian random variables obtained as the marginal distribution on a *subset* of the variables in a Gaussian tree model. If we are only given such a marginal distribution, it generally has no interesting conditional independence relations. For us to expose the underlying latent tree structure for computation, we need to be able to recognize that the marginal distribution has such an alternative description and reconstruct that description.

Specifically, in Chapter 6 we consider the problem of (approximately) realizing a given covariance matrix as the marginal covariance among the leaf-indexed variables of a Gaussian tree model. This is the problem of modeling a given covariance as a Gaussian latent tree model. Our approach is based on reparameterizing the marginal covariance matrices among the leaves of Gaussian tree models in terms of certain structured matrix decompositions.

We devise two (closely related) semidefinite optimization-based methods to (approximately) construct a latent tree model for a given covariance matrix. The two methods differ in the amount of information about the structure of the tree that they fix. The first (and simpler) method fixes the entire structure of the tree. The second fixes only certain information about how the observed variables are associated with the (possibly non-scalar) leaf-indexed variables of the tree. Our main contribution is to provide conditions on an underlying Gaussian tree model under which these methods, when given the covariance among the leaf-indexed variables of that model (and the corresponding information about the tree structure), can exactly recover the full model.

Chapter 2

Background

2.1 Introduction

This chapter has two main aims. The first is to collect notation, terminology, and basic facts that are used repeatedly throughout the thesis. As such we discuss various aspects of convex geometry and convex optimization, as well as the interplay between group symmetry and these topics. The second aim is to explain a systematic approach to optimization via convexification. This approach involves transforming optimization problems into equivalent convex optimization problems, and then seeking tractable descriptions (or approximations) of these convex optimization problems. This approach highlights the importance of good descriptions of convex optimization problems. We pay particular attention to descriptions of optimization problems as semidefinite optimization problems.

The rest of the chapter is organized as follows. In Section 2.2 we describe the basic idea of convex reformulations and relaxations of optimization problems. This section provides technical context for many of the problems studied in this thesis. In Section 2.3 we summarize basic notation and terminology related to convex sets and functions. Section 2.4 focuses on semidefinite optimization, explaining basic facts and terminology related to this family of convex optimization problems. We emphasize properties of the convex sets that arise as feasible regions of semidefinite optimization problems. In Section 2.5 we describe hyperbolic polynomials, hyperbolicity cones, and hyperbolic optimization problems. These are a family of optimization problems that generalize semidefinite optimization problems, and have nice algebraic properties. In Section 2.6 we discuss the interaction between symmetry and convex geometry and optimization. We establish some basic results that allow us to exploit symmetry properties, whenever possible, throughout the thesis.

2.2 Convex reformulations and relaxations

In this section we describe a well-known way to reformulate a family of (finite dimensional) optimization problems in terms of maximizing linear functionals over a convex set. The reformulation is completely formal and so is not obviously useful. However, it does shift our viewpoint away from the issue of whether a problem is convex, to the issue of whether it has a *tractable* convex description. We then briefly discuss the situation in which we only have a tractable approximation (i.e. a convex *relaxation*) to the convex problem we would like to solve.

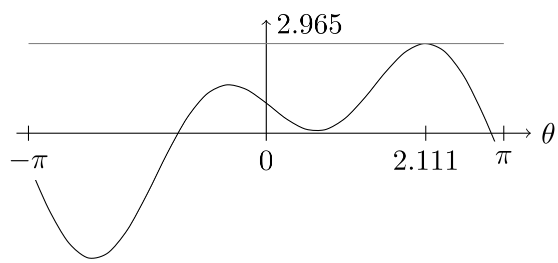
Throughout this section we use terminology related to convex sets and functions that we do not introduce until later in the chapter. In these cases we provide forward references where appropriate. Nevertheless in this section we encourage the reader to ignore unfamiliar terminology as much as possible.

Throughout this section we discuss a simple example of a non-linear and non-convex optimization problem. This is not a difficult problem to solve. We use it because it is simple enough that we can explicitly illustrate the basic approach.

Example 2.2.1. Consider the following optimization problem over the unit circle (which we parameterize by $(\cos(\theta), \sin(\theta))$):

$$\max_{\theta\in[-\pi,\pi]} \cos(\theta) + 2\sin(\theta) - 2\sin(2\theta). \tag{2.2.1}$$

A plot of this function on the interval $[-\pi, \pi]$ is shown below. The unique global maximum occurs at $\theta \approx 2.111$ and the maximum value is approximately 2.965.



This is an instance of a whole family of problems of the form

$$\max_{\theta\in[-\pi,\pi]} a_1\cos(\theta) + a_2\sin(\theta) + a_3\sin(2\theta) \tag{2.2.2}$$

where a_1, a_2, and a_3 are real parameters.

Generally, suppose we have a compact subset $S \subseteq \mathbb{R}^n$ of a finite-dimensional real vector space and a collection of continuous functions $f_i : S \to \mathbb{R}$ for $i = 1, 2, \ldots, m$[1].

[1] The compactness assumption on S and the continuity assumption on the f_i are made to avoid certain technicalities.

Let $F : S \to \mathbb{R}^m$ be defined by

$$F(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{bmatrix}.$$

Suppose we are interested in the family of optimization problems

$$\max_{x \in S} \sum_{i=1}^{m} a_i f_i(x) = \max_{x \in S} \langle a, F(x) \rangle \tag{2.2.3}$$

parameterized by a real vector $a \in \mathbb{R}^m$. Each problem in this family involves maximizing some function $f \in \operatorname{span}\{f_1, f_2, \ldots, f_m\}$, a particular subspace of functions on S. In this section, when we refer to the *instance* of (2.2.3) defined by a particular vector a we mean the optimization problem (2.2.3) with that particular a as its parameter.

This subspace is typically determined by the basic structure of a problem. It may be reasonable to assume it is known in advance. This is the case, for instance, in the attitude estimation problem discussed in Chapter 4. The parameters a_i of such a problem family are often specified by data and only known at 'run-time'. As such it can be worthwhile to invest considerable effort in reformulating and understanding the whole family of problems with the aim of developing solution methods that are valid for any problem instance.

■ 2.2.1 Reparameterization

Example 2.2.2. In our running example, the set is the interval $S = [-\pi, \pi]$ (which is convex in this case, but need not be). A collection of functions f_i is given by $f_1(\theta) = \cos(\theta)$, $f_2(\theta) = \sin(\theta)$, and $f_3(\theta) = \sin(2\theta)$. Note that any collection of functions that span the same three-dimensional space would do equally well.

Instead of thinking about the decision variable as $\theta \in [-\pi, \pi]$ we reparameterize the problem, and think of the decision variable as being a three-dimensional vector z taking values in

$$F([-\pi, \pi]) = \left\{ \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \\ \sin(2\theta) \end{bmatrix} : \theta \in [-\pi, \pi] \right\} \subset \mathbb{R}^3. \tag{2.2.4}$$

This subset of $\mathbb{R}^3$ is the curve shown on the left in Figure 2.1. Then we can rewrite the family of optimization problems (2.2.2) as

$$\max_{z \in F([-\pi, \pi])} a_1 z_1 + a_2 z_2 + a_3 z_3, \tag{2.2.5}$$

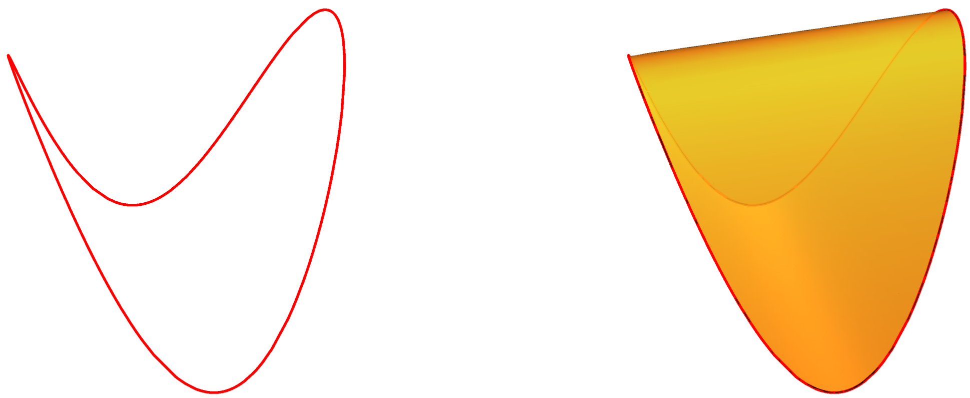


Figure 2.1: On the left is the subset $F([-\pi,\pi])$ of $\mathbb{R}^3$ described in (2.2.4). On the right is its convex hull.

the maximization of the real-valued linear function specified by a_1, a_2, and a_3 over $F([-\pi,\pi])$.

In the general setting we change from regarding $x \in S \subset \mathbb{R}^n$ as the decision variable of the optimization problem to regarding $z \in \mathbb{R}^m$ as the decision variable and constraining it to take values in

$$F(S) = \left\{ \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{bmatrix} : x \in S \right\} \subset \mathbb{R}^m,$$

which is compact (by the continuity of F and compactness of S). The family of optimization problems (2.2.3) can then be expressed as

$$\max_{z \in F(S)} \langle a, z \rangle. \tag{2.2.6}$$

Note that in this parameterization the objective function is linear in the decision variable z. All the complexity of the problem is in the (non-convex) constraint set $F(S)$. In this section we often refer to the function $z \mapsto \langle a, z \rangle$ as the *linear objective function defined by a*.

2.2.2 Convexification

Our problem now involves maximizing a linear objective function over a set. Momentarily, let us assume that all we care about is the optimal value of the problem. (We return to the issue of recovering an optimal point in Section 2.2.3 to follow.) This value is the same whether we optimize over $F(S)$ or over its (necessarily compact[2]) convex hull, denoted $\operatorname{conv}(F(S))$, i.e.

$$\max_{z \in F(S)} \langle a, z\rangle = \max_{z \in \operatorname{conv}(F(S))} \langle a, z\rangle \tag{2.2.7}$$

(see, e.g., [105, Theorem 32.2].) The latter of these problems involves maximizing a linear objective function over a closed convex set and so is a convex optimization problem. The family of problems we are interested in is parameterized by all vectors a that define the linear objective function $z \mapsto \langle a, z\rangle$. As such, the entire family of convex optimization problems is captured by the convex set $\operatorname{conv}(F(S))$.

Example 2.2.3. In our running example, the convex hull of $F([-\pi,\pi])$ is shown on the right in Figure 2.1. On the left of Figure 2.2 is the set $F(S)$ together with the level set $\{z : z_1 + 2z_2 - 2z_3 = 2.965\ldots\}$[3] of the objective function, highlighting the intersection. On the right of Figure 2.2 is the set $\operatorname{conv}(F(S))$ together with the same level set of the objective function, again highlighting the intersection. Observe that the intersection of $\operatorname{conv}(F(S))$ with this level set is also a point of $F(S)$.

Indeed the number $2.965\ldots$ is the optimal value of both the optimization problem

$$\max_{z \in F([-\pi,\pi])} z_1 + 2z_2 - 2z_3$$

and the convex optimization problem

$$\max_{z \in \operatorname{conv}(F([-\pi,\pi]))} z_1 + 2z_2 - 2z_3.$$

We have not yet explained how we solved these optimization problems. We do this in Example 2.2.4 to follow.

Returning to the general setting, the reformulation in (2.2.7) has not yet used any properties of the set S or the functions f_i. Hence our transformations, alone, have not achieved anything beyond changing our viewpoint. Nevertheless, this viewpoint puts many families of optimization problems on a common footing, reducing them to

[2]The convex hull of a compact set is again compact [105, Theorem 17.2]

[3]We write $2.965\ldots$ as a decimal followed by an ellipsis to emphasize that this is not a rational number but an algebraic number. In fact it is an algebraic number of degree four, i.e. a root of a degree four polynomial with rational coefficients (see, e.g., [14, p. 220]).

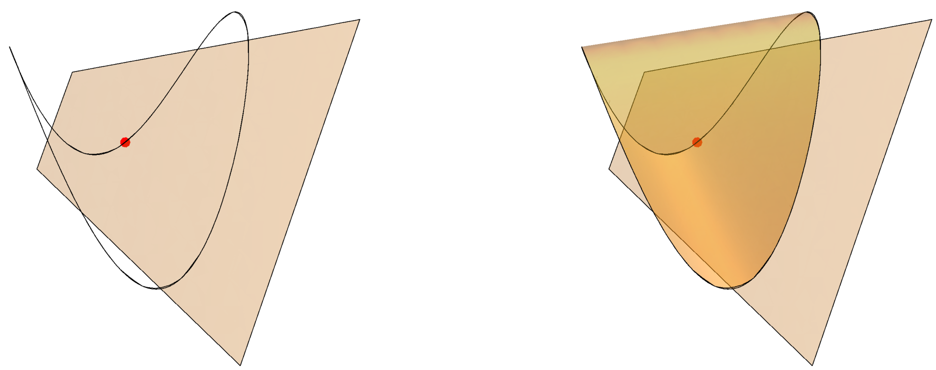


Figure 2.2: Shown on the left is the set $F([-\pi,\pi])$, the hyperplane $\{z : z_1+2z_2-2z_3 = 2.965\ldots\}$, and their intersection (in red). Shown on the right is the set $\mathrm{conv}(F([-\pi,\pi]))$, the same hyperplane, and their intersection (in red). In both cases the intersection is the same.

understanding a corresponding geometric object, namely $\mathrm{conv}(F(S))$. This allows us to focus on the problem of maximizing a linear objective function over a convex set as a prototypical optimization problem.

Tractable descriptions

Given a family of optimization problems, the major challenge is to find efficient algorithms to solve its convex reformulation

$$\max_{x\in\mathrm{conv}(F(S))} \langle a, x\rangle. \tag{2.2.8}$$

Rather than directly developing algorithms for (2.2.8) on a case-by-case basis, a conceptually simpler approach is to try to give an alternative mathematical description of $\mathrm{conv}(F(S))$ in a standard form for which algorithms have already been developed. A typical notion of a standard form is to fix a convex cone K (see Section 2.3.2 to follow) that is well-understood, and try to express the convex set in *conic form* as

$$\mathrm{conv}(F(S)) = \pi(K \cap L)$$

where π is a linear map and L is an affine subspace (see Section 2.3.1 to follow). Such a description packages all that is difficult (and interesting) about the problem into the cone K.

Semidefinite representations

In this thesis we mostly focus on *semidefinite representations* of convex sets (see Definition 2.4.5 to follow). These are descriptions of convex sets in conic form in which the cone K is a cone of positive semidefinite matrices. A semidefinite representation has *size* m if $K = \mathcal{S}_+^m$ is the cone of $m \times m$ positive semidefinite matrices. If we have a semidefinite representation of $\operatorname{conv}(F(S))$ of size m we can solve (2.2.8) in time polynomial in m using algorithms for *semidefinite optimization* (see Section 2.4 to follow).

Example 2.2.4. In our running example, we want a tractable description of

$$\operatorname{conv}(F([-\pi,\pi])) = \operatorname{conv}\left\{\begin{bmatrix}\cos(\theta) & \sin(\theta) & \sin(2\theta)\end{bmatrix}^T : \theta \in [-\pi,\pi]\right\}.$$

By specializing Proposition 4.8.5 of Chapter 4 we obtain a semidefinite representation of size 3 for this convex set as

$$\operatorname{conv}(F([-\pi,\pi])) = \left\{\begin{bmatrix}u_1 & v_1 & v_2\end{bmatrix}^T \in \mathbb{R}^3 : \exists u_2 \in \mathbb{R}, \begin{bmatrix}1+v_2 & u_1+v_1 & u_2\\ u_1+v_1 & 1 & u_1-v_1\\ u_2 & u_1-v_1 & 1-v_2\end{bmatrix} \in \mathcal{S}_+^3\right\}.$$

This is the projection of a convex set in $\mathbb{R}^4$ in the variables (u_1, v_1, u_2, v_2) onto the variables (u_1, v_1, v_2). We can then solve any instance of our family of optimization problems by solving the semidefinite optimization problem

$$\max_{u_1,u_2,v_1,v_2} a_1u_1 + a_2v_1 + a_3v_2 \quad \text{subject to} \quad \begin{bmatrix}1+v_2 & u_1+v_1 & u_2\\ u_1+v_1 & 1 & u_1-v_1\\ u_2 & u_1-v_1 & 1-v_2\end{bmatrix} \in \mathcal{S}_+^3.$$

In general we do not expect to be able to find tractable descriptions, as semidefinite representations with fairly small size, of arbitrary convex sets.

Example 2.2.5. For a concrete example, consider the family of binary quadratic optimization problems, i.e. problems of the form

$$\max_{x\in\{-1,1\}^n} \sum_{1\leq i<j\leq n} A_{ij}x_ix_j.$$

If we take take $S = \{-1,1\}^n$ and $F(x) = (x_ix_j)_{1\leq i<j\leq n}$ then we can reformulate this

family as

$$\max_{Z \in \mathrm{conv}(F(S))} \sum_{1 \le i < j \le n} A_{ij} Z_{ij}.$$

The convex set $\mathrm{conv}(F(S))$ is called the *cut polytope* [38]. If we could efficiently maximize arbitrary linear objective functions over the cut polytope (by some method), we could efficiently solve arbitrary binary quadratic optimization problems, including well-known NP-hard problems such as MAX-CUT [96, 71]. Concerning semidefinite representations in particular, it has recently been shown that any semidefinite representation of the cut polytope must have size exponential in n [76].

Even when solving the optimization problem (2.2.8) is hard, the convexified viewpoint is still very useful. This is because it suggests an approach to approximately solving the problem via convex relaxations (see Section 2.2.4 to follow).

■ 2.2.3 Extracting an optimal point

For now, let us assume we can solve the optimization problem (2.2.8) for some fixed parameter vector a of interest. We now briefly discuss how we can obtain an optimal point for the original problem from the solution to the convex reformulation.

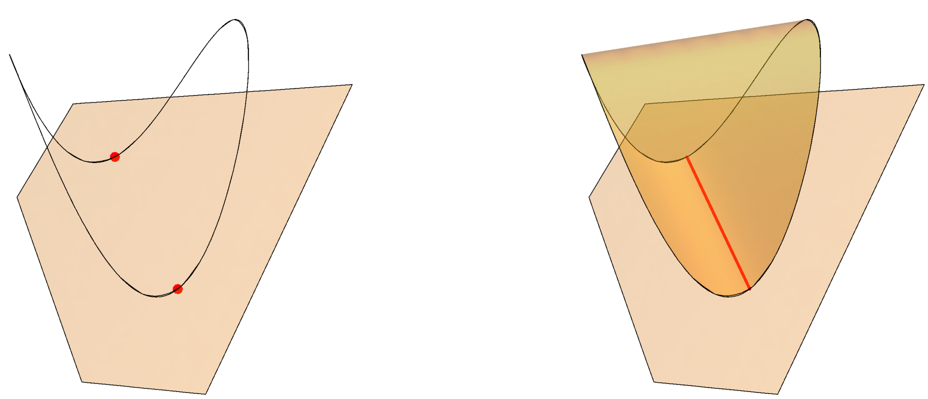


Figure 2.3: Shown on the left is the set $F([-\pi,\pi])$, the hyperplane $\{z : z_1 + z_2 - (5/2)z_3 = 13/5\}$, and their intersection (in red). Shown on the right is the set $\mathrm{conv}(F([-\pi,\pi]))$, the same hyperplane, and their intersection (in red). Observe that in this case the intersection with $F([-\pi,\pi])$ consists of two points, whereas the intersection with $\mathrm{conv}(F([-\pi,\pi]))$ is the convex hull of these two points.

Example 2.2.6. In our running example, we have seen in Figure 2.2 that there is a unique optimal solution to the convex optimization problem

$$\max_{z \in \operatorname{conv}(F([-\pi,\pi]))} (z_1 + 2z_2 - 2z_3).$$

This point is $z_\star = (-0.514\ldots, 0.858\ldots, -0.882\ldots)$. This is also the unique optimal solution to

$$\max_{z \in F([-\pi,\pi])} (z_1 + 2z_2 - 2z_3).$$

To obtain a solution $\theta \in [-\pi, \pi]$ we need to solve the following nonlinear equation for θ:

$$(\cos(\theta), \sin(\theta), \sin(2\theta)) = F(\theta) = z_\star = (-0.514\ldots, 0.858\ldots, -0.882\ldots).$$

Doing so we obtain $\theta \approx 2.111$ which matches what we observed in the figure immediately after (2.2.1). If, instead, we choose the parameters $a_1 = 1, a_2 = 1$, and $a_3 = -5/2$ we see from the right of Figure 2.3 that the convex optimization problem

$$\max_{z \in \operatorname{conv}(F([-\pi,\pi]))} z_1 + z_2 - (5/2)z_3$$

has multiple optimal solutions. Indeed the optimal face (see Section 2.3.2) is given by all convex combinations of $z_\star^{(1)} = (-3/5, 4/5, -24/25)$ and $z_\star^{(2)} = (4/5, -3/5, -24/25)$ and is the intersection of $\operatorname{conv}(F([-\pi, \pi]))$ with the hyperplane $\{z : z_1 + z_2 - (5/2)z_3 = 13/5\}$. The extreme points (see Section 2.3.2) of this optimal face are $z_\star^{(1)}$ and $z_\star^{(2)}$, which are elements of $F([-\pi, \pi])$. From the left of Figure 2.3 we see that these are precisely the optimal solutions of

$$\max_{z \in F([-\pi,\pi])} z_1 + z_2 - (5/2)z_3.$$

We obtain the corresponding solutions for θ by solving the equations

$$F(\theta) = z_\star^{(1)} \quad \text{and} \quad F(\theta) = z_\star^{(2)}.$$

The plot of $\cos(\theta) + \sin(\theta) - (5/2)\sin(2\theta)$ shown in Figure 2.4 confirms that we expect multiple global optima on the interval $[-\pi, \pi]$.

We now consider the general setting. Let $C = \operatorname{conv}(F(S))$ for brevity. Suppose we solve (2.2.8) for some fixed instance determined by a. We discuss the basic process of extracting optimal points by considering separately the cases in which the convex reformulation has a unique optimal point or multiple optimal points, respectively.

First, suppose that the point $z_\star$ is the unique optimal point for the convex reformulation (2.2.8). Then using the fact that every extreme point of $\operatorname{conv}(F(S))$ is actually in $F(S)$ [105, Corollary 18.3.1], we can deduce that $z_\star \in F(S)$. Any solution for x for

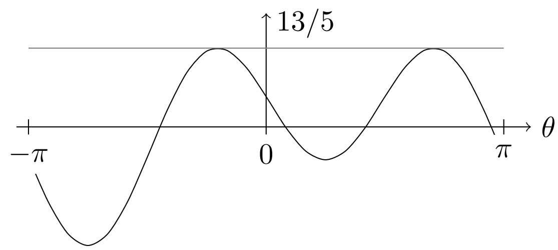


Figure 2.4: A plot of $\cos(\theta) + \sin(\theta) - (5/2)\sin(2\theta)$ on $[-\pi, \pi]$ showing that it has multiple global maxima.

the nonlinear equation

$$F(x) = z_\star \tag{2.2.9}$$

is optimal for the original optimization problem $\max_{x \in S}\langle a, F(x)\rangle$.

Suppose, now, that the convex reformulation (2.2.8) has more than one optimal solution. The set of optimal solutions is convex (indeed it is a face of C) and the extreme points of the set of optima are all elements of $F(S)$ [105, Theorem 18.3]. We can obtain an optimal point for the original problem by taking any extreme point $z_\star$ of this optimal face, and solving the non-linear equation (2.2.9) for x.

There are constructive versions of Carathéodory's theorem [105, Section 17] that allow us, in principle, to find extreme points of the optimal face. In practice the best way to do this algorithmically depends significantly on the problem structure.

■ 2.2.4 Convex relaxations

It is often the case that no 'good' description of the set $\mathrm{conv}(F(S))$ is known, and so we are not able to solve the convex reformulation (2.2.8) efficiently. In this case, a natural approach is to find a convex set C such that

$$C \supseteq \mathrm{conv}(F(S))$$

and such that we can optimize linear objective functions over C efficiently. It is common to use the term *convex relaxation* to describe the problem of optimizing a linear objective function over C, since we have enlarged or *relaxed* the feasible region from $\mathrm{conv}(F(S))$ to C.

Example 2.2.7. An important example of this is the standard convex relaxation for binary quadratic optimization (see Example 2.2.5). Here we have $S = \{-1, 1\}^n$ and $F(x) = (x_i x_j)_{1 \le i < j \le n}$. The set $\mathrm{conv}(F(S))$ is the cut polytope. It is shown in Figure 2.5 for $n = 3$. A well-known convex relaxation of the cut polytope is the *elliptope* defined by

$$C = \left\{ (Z_{ij})_{1 \le i < j \le n} \in \mathbb{R}^{\binom{n}{2}} : \begin{bmatrix} 1 & Z_{12} & Z_{13} & \cdots & Z_{1n} \\ Z_{12} & 1 & Z_{23} & \cdots & Z_{2n} \\ Z_{13} & Z_{23} & 1 & \cdots & Z_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_{1n} & Z_{2n} & Z_{3n} & \cdots & 1 \end{bmatrix} \in \mathcal{S}^n_+ \right\}.$$

The set C has a semidefinite representation of size n, so we can maximize a linear objective function over it in time polynomial in n. To see that $C \supseteq \operatorname{conv}(F(\{-1,1\}^n))$, let $x \in \{-1,1\}^n$ be arbitrary. Then

$$\begin{bmatrix} 1 & x_1x_2 & x_1x_3 & \cdots & x_1x_n \\ x_1x_2 & 1 & x_2x_3 & \cdots & x_2x_n \\ x_1x_3 & x_2x_3 & 1 & \cdots & x_3x_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1x_n & x_2x_n & x_3x_n & \cdots & 1 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \end{bmatrix} \in \mathcal{S}^n_+$$

which shows that $F(x) \in C$. Hence $F(\{-1,1\}^n) \subseteq C$. Convex relaxations generalizing this one play an important role in Chapter 5.

Bounds on the optimal value

The inequality

$$\max_{z \in C} \langle a, z \rangle \ge \max_{z \in \operatorname{conv}(F(S))} \langle a, z \rangle = \max_{x \in S} \langle a, F(x) \rangle \tag{2.2.10}$$

always holds because $C \supseteq \operatorname{conv}(F(S))$. This inequality tells us that convex relaxations give computationally tractable methods to obtain upper bounds on the global maximum. Such global upper bounds complement optimization methods that search within the feasible region of the original problem. This is because whenever $y \in S$ we have that

$$\max_{z \in C} \langle a, z \rangle \ge \max_{x \in S} \langle a, F(x) \rangle \ge \langle a, F(y) \rangle$$

where the first inequality is repeated from (2.2.10). As such, the optimal value of a convex relaxation can be use to assess how close the function value at a local optimum (obtained by local optimization methods) is to the global optimal value.

Exact instances for convex relaxations

Often a convex outer approximation C to $\operatorname{conv}(F(S))$ preserves some of the faces of $\operatorname{conv}(F(S))$. Consequently, for certain linear objective functions defined by vectors a, maximizing $\langle a, x \rangle$ over $x \in \operatorname{conv}(F(S))$ or over $x \in C$ gives the same optimal value and

Figure 2.5: On the left is the cut polytope for $n = 3$. In this case it is the tetrahedron conv $\{(1,1,1),(1,-1,-1),(-1,1,-1),(-1,-1,1)\}$. On the right is corresponding elliptope, a convex outer approximation of the cut polytope for $n = 3$. All of the zero- and one-dimensional faces of the cut polytope for $n = 3$ are also faces of the elliptope.

the same optimal face.

Example 2.2.8. In Figure 2.5 we can see that all of the 0- and 1-dimensional faces of the cut polytope for $n = 3$ are also faces of the corresponding elliptope. Suppose, for instance, we maximize the linear objective function defined by $A_{12} = 1$, $A_{13} = -2$, $A_{23} = 0$ over the elliptope, i.e. solve the semidefinite optimization problem

$$\max_{Z_{12},Z_{13},Z_{23}} Z_{12} - 2Z_{13} + 0Z_{23} \quad \text{subject to} \quad \begin{bmatrix} 1 & Z_{12} & Z_{13} \\ Z_{12} & 1 & Z_{23} \\ Z_{13} & Z_{23} & 1 \end{bmatrix} \in \mathcal{S}^3_+.$$

The unique optimal point is $(Z_{12}, Z_{13}, Z_{23}) = (1, -1, -1)$ (see Figure 2.6). This is a 0-dimensional face of the elliptope that is also a face of the cut polytope. We can deduce this by observing that $(1, -1, -1) = F((1, 1, -1))$. On the other hand, if we maximize the linear objective function defined, for instance, by $A_{12} = -3/2$, $A_{13} = 1$, $A_{13} = 1$ over the elliptope, the maximum value is $11/6$ and the unique optimal point is $(-7/9, 1/3, 1/3)$ which is not in $F(\{-1, 1\}^3)$ (see Figure 2.6).

In the general setting, when we maximize the linear objective function defined by a over C, one of two things could happen.

1. It could happen that the optimal face of C is a face of $\mathrm{conv}(F(S))$. In this case we have equality in (2.2.10). Furthermore, the extreme points of the optimal face

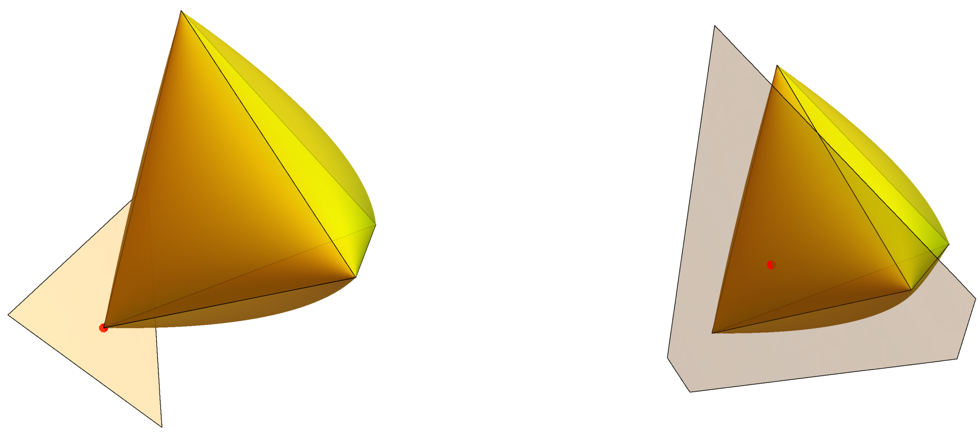


Figure 2.6: On the left is the elliptope, the level set $\{(Z_{12}, Z_{13}, Z_{23}) : Z_{12} - 2Z_{23} = 3\}$ of the linear objective function $Z_{12} - 2Z_{23}$, and their intersection $(1, -1, -1)$. The maximum of this function over the elliptope is achieved at $(1, -1, -1)$, which is also an element of the cut polytope. As such the convex relaxation is *exact* for this instance. On the right is the elliptope, the level set $\{(Z_{12}, Z_{13}, Z_{23}) : -(3/2)Z_{12} + Z_{13} + Z_{23} = 11/6\}$ of the linear objective function $-(3/2)Z_{12} + Z_{13} + Z_{23}$, and their intersection at $(-7/9, 1/3, 1/3)$. The maximum of this linear function over the elliptope is achieved at $(-7/9, 1/3, 1/3)$ which is *not* an element of the cut polytope.

are in $F(S)$, and these correspond to global optima of the instance of the original optimization problem defined by a. In this case, we say that the relaxation is *exact* for the problem instance defined by a.

2. Otherwise the optimal face of C is not a face of $\text{conv}(F(S))$. In this case, we have a strict inequality in (2.2.10) and the extreme points of the optimal face are not in $F(S)$.

Suppose we have a way to check whether a given point is in $F(S)$. Then we can use the solution (in general the optimal face) of our convex relaxation to detect whether the relaxation was exact for the instance we just solved. To do this, we find the extreme points of the optimal face and check whether they are in the set $F(S)$. If they are, all of these extreme points must correspond to optimal solutions for the original optimization problem (by (2.2.10)).

As we have seen, we can typically detect whether a convex relaxation is exact for a given instance *after* we have solved it. Nevertheless, it is of considerable interest to give simple sufficient conditions on problem instances (i.e. the vectors a) that ensure a convex relaxation is exact for those instances. We can then guarantee, beforehand, that these instances of the original problem can be efficiently solved.

Rounding

Suppose $C \supseteq \text{conv}(F(S))$ defines a convex relaxation of our problem of interest. We have seen that for some problem instances defined by vectors a the maximum of $z \mapsto \langle a, z \rangle$ over C occurs on a face of C that is not a face of $\text{conv}(F(S))$. In other words the instance defined by a is not exact. Nevertheless, we would expect that the optimal point of the convex relaxation contains useful global information about the original problem.

Rounding schemes are procedures that map the extreme points of C to $\text{conv}(F(S))$ in a way that aims to preserve the value of the objective function as well as possible (in a sense that depends on the specific aims of the problem). Such a procedure allows us to take an optimal point for the convex relaxation defined by C and produce a feasible point for the convex reformulation defined by $\text{conv}(F(S))$. Usually these procedures are described as randomized algorithms that map extreme points of C to samples from a probability distribution on $F(S)$. The map to $\text{conv}(F(S))$ is obtained by taking the expectation of such a distribution on $F(S)$.

These randomized algorithms allow us to produce feasible points for the original optimization problem over $F(S)$. In some cases it can be shown that these 'rounded' feasible points have objective value that is close to the true global optimum of the original optimization problem. It is in this way that tractable convex relaxations, followed by a well-designed rounding scheme, can lead to approximation algorithms.

This approach to approximation algorithms for combinatorial optimization problems has been extensively studied by researchers in algorithms and complexity theory (see, e.g., [136, Chapter 6]). The approach is less developed (particularly the problem of designing rounding schemes) for continuous optimization problems. Chapter 5 considers this for a particular family of optimization problems, and a particular family of convex relaxations.

2.3 Convex sets and functions

In this section we summarize basic facts about convex sets and convex functions. All the material in this section is standard, and is presented rather briefly.

2.3.1 Preliminaries

Throughout the thesis we always work in a finite-dimensional real vector space V. For convenience we usually endow V with a real-valued inner product $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ and identify V with its dual space[4] via this choice of inner product. If V and W are two real vector spaces (with associated inner products) and $A : V \to W$ is a linear map, the *adjoint* of A is the map $A^* : W \to V$ defined by

$$\langle Av, w \rangle = \langle v, A^* w \rangle \quad \text{for all } v \in V \text{ and all } w \in W.$$

The main concrete examples of interest are $\mathbb{R}^n$ (real n-vectors), $\mathbb{R}^{n\times m}$ (real $n \times m$ matrices), and $\mathcal{S}^m$ (symmetric $m \times m$ matrices). We also work with $\mathbb{C}^n$ (complex n-vectors), $\mathbb{C}^{n\times m}$ (complex $n \times m$ matrices) and $\mathcal{H}^m$ (Hermitian $m \times m$ matrices). Throughout, these last three cases are almost always regarded as *real vector spaces* of real dimension $2n$, $2nm$, and m^2 respectively.

We denote the transpose of $A \in \mathbb{R}^{n\times m}$ by A^T, the entry-wise complex conjugate of $A \in \mathbb{C}^{n\times m}$ by $\bar{A}$, and the conjugate transpose of $A \in \mathbb{C}^{n\times m}$ by A^*. If $A \in \mathbb{C}^{n\times m}$ then $\mathrm{Re}\,[A] = (A + \bar{A})/2$ and $\mathrm{Im}\,[Z] = (A - \bar{A})/(2i)$ denote the real and imaginary parts respectively. Note that $A \mapsto \bar{A}$, $A \mapsto A^*$, $A \mapsto \mathrm{Re}\,[A]$ and $A \mapsto \mathrm{Im}\,[A]$ are linear maps when we regard $\mathbb{C}^{n\times m}$ as a $2nm$ dimensional real vector space.

We equip the real vector spaces of interest with *real-valued* inner products as follows. The inner product on

- $\mathbb{R}^n$ is the standard inner product $\langle x, y \rangle := \sum_{i=1}^n x_i y_i$;
- $\mathbb{R}^{n\times m}$ is the trace inner product $\langle X, Y \rangle := \mathrm{tr}(X^T Y) = \sum_{i=1}^n \sum_{j=1}^m X_{ij} Y_{ij}$;

[4] If V is a finite dimensional vector space its dual space is $V^* := \{\ell : V \to \mathbb{R} : \ell \text{ is linear}\}$, the real vector space of $\mathbb{R}$-valued linear functionals on V.

- $\mathcal{S}^m$ is the restriction of the trace inner product to symmetric matrices, i.e. $\langle X, Y\rangle = \mathrm{tr}(XY)$;
- $\mathbb{C}^n$ is the real inner product $\langle x, y\rangle := \mathrm{Re}\,[x^*y] = \langle \mathrm{Re}\,[x], \mathrm{Re}\,[y]\rangle + \langle \mathrm{Im}\,[x], \mathrm{Im}\,[y]\rangle$;
- $\mathbb{C}^{n\times m}$ is the real trace inner product $\langle X, Y\rangle := \mathrm{Re}\,[\mathrm{tr}(X^*Y)] = \langle \mathrm{Re}\,[X], \mathrm{Re}\,[Y]\rangle + \langle \mathrm{Im}\,[X], \mathrm{Im}\,[Y]\rangle$;
- $\mathcal{H}^m$ is the restriction of the real trace inner product to Hermitian matrices, i.e. $\langle X, Y\rangle = \mathrm{Re}\,[\mathrm{tr}(XY)]$.

Associated with each of these inner products is a norm $\|\cdot\| = \langle\,\cdot\,,\,\cdot\,\rangle^{1/2}$. In the case of $\mathbb{R}^n$ and $\mathbb{C}^n$ this is the usual Euclidean norm and is simply denoted $\|\cdot\|$. In the cases of $\mathbb{R}^{n\times m}$, $\mathcal{S}^m$, $\mathbb{C}^{n\times m}$ and $\mathcal{H}^m$ the corresponding norm is the *Frobenius norm* and is denoted $\|\cdot\|_F$.

Linear and affine subspaces

A subset U of a finite dimensional real vector space V is a *(linear) subspace* if whenever $u_1, u_2 \in U$ and $\lambda_1, \lambda_2 \in \mathbb{R}$ then $\lambda_1 u_1 + \lambda_2 u_2 \in U$, i.e. U is closed under taking *linear combinations*. If $S \subseteq V$ then the *span* of S, denoted $\mathrm{span}(S)$, is the set of all linear combinations of finitely many elements of S. Equivalently, $\mathrm{span}(S)$ is the intersection of all linear subspaces of V containing S.

A subset L of a finite dimensional real vector space V is an *affine subspace* if whenever $u_1, u_2 \in L$ and $\lambda_1, \lambda_2 \in \mathbb{R}$ satisfy $\lambda_1 + \lambda_2 = 1$ then $\lambda_1 u_1 + \lambda_2 u_2 \in L$, i.e. L is closed under taking *affine combinations*. If $S \subseteq V$ then the *affine span* of S, denoted $\mathrm{aff}(S)$, is the set of all affine combinations of finitely many elements of S. Equivalently, $\mathrm{aff}(S)$ is the intersection of all affine subspaces of V containing S.

Topological notions

If V is a finite dimensional real inner product space we equip V with the usual topology induced by the metric $d(x, y) = \|x - y\|$ on the vector space V. The interior of a set is denoted $\mathrm{int}(S)$, the closure is denoted $\mathrm{cl}(S)$. Any relative topological notions related to a set S are taken with respect to the affine span of S. For instance the *relative interior* of S, denoted $\mathrm{relint}(S)$, is the interior of S when thought of as a subset of $\mathrm{aff}(S)$ with the relative topology induced from V.

Subspaces associated with linear maps

If $A : V \to W$ is a linear map between two finite dimensional real vector spaces V and W then the *column space*[5] of A, denoted $\mathrm{col}(A)$, is the subspace $A(V) = \{Av : v \in V\}$

[5] Since this is an abstract linear map it would be more usual to call this the *image* or the *range space* of A. Throughout much of the thesis we work concretely with matrices, so we use the more

of W. The *nullspace* of A, denoted $\mathrm{null}(A)$, is the subspace $\{v \in V : Av = 0\}$ of V.

Dimension and explicit parameterization of affine subspaces

An affine subspace L can always be expressed as $L = \{x_0\}+U = \{x_0+u : u \in U\}$ where $x_0 \in L$ is any point in L and U is a linear subspace uniquely determined by L. The *dimension* of an affine subspace $L = \{x_0\} + U$ is the dimension of the corresponding linear subspace U.

If $U = \mathrm{col}(A)$ for some $n \times m$ matrix A, then the affine subspace $\{x_0\} + U$ of $\mathbb{R}^n$ can be expressed concretely as

$$\{x_0\} + U = \{x_0 + Ax : x \in \mathbb{R}^m\}. \tag{2.3.1}$$

Similarly if $V = \mathrm{null}(B)$ for some $m \times n$ matrix B, then the affine subspace $\{x_0\} + V$ of $\mathbb{R}^n$ can be expressed concretely as

$$\{x_0\} + V = \{x \in \mathbb{R}^n : x - x_0 \in \mathrm{null}(B)\} = \{x \in \mathbb{R}^n : Bx = Bx_0\}. \tag{2.3.2}$$

Orthogonal projections

If U is a subspace of a finite dimensional real inner product space V then the orthogonal projection onto U is $P_U : V \to V$ defined by being self-adjoint and satisfying $\mathrm{col}(P_U) = U$ and $P_U^2 = P_U$. We also define $\Pi_U : V \to U$ and its adjoint $\Pi_U^* : U \to V$ by $P_U = \Pi_U^*\Pi_U$ and $I = \Pi_U\Pi_U^*$.

■ 2.3.2 Convex sets

A subset K of a finite dimensional real vector space V is a *convex cone* if whenever $u_1, u_2 \in S$ and $\lambda_1, \lambda_2 \in \mathbb{R}$ such that $\lambda_1, \lambda_2 \geq 0$ then $\lambda_1 u_1 + \lambda_2 u_2 \in K$, i.e. K is closed under taking *conic combinations*. If $S \subseteq V$ then the *conic hull* of S, denoted $\mathrm{cone}(S)$, is the set of all conic combinations of finitely many elements of S. This is also the intersection of all convex cones in V containing S.

A subset C of a finite dimensional real vector space V is *convex* if whenever $u_1, u_2 \in S$ and $\lambda_1, \lambda_2 \in \mathbb{R}$ such that $\lambda_1 + \lambda_2 = 1$ and $\lambda_1, \lambda_2 \geq 0$ then $\lambda_1 u_1 + \lambda_2 u_2 \in C$, i.e. C is closed under taking *convex combinations*. If $S \subseteq V$ then the *convex hull* of S, denoted $\mathrm{conv}(S)$, is the set of all convex combinations of finitely many elements of S. This is also the intersection of all convex subsets of V containing S.

A convex cone K is *pointed* if $K \cap (-K) = \{0\}$, *solid* if the span of K is all of V, and *proper* if it is pointed, solid, and closed. Any pointed cone $K \subseteq V$ defines a partial order on V which we denote by $x \preceq_K y$ if and only if $y - x \in K$.

---

matrix-oriented term *column space*.

Isomorphism of convex sets

We now describe the appropriate notion of equiavalence for convex cones. Suppose V_1 and V_2 are finite dimensional real vector spaces. A pair of convex cones $K_1 \subseteq V_1$ and $K_2 \subseteq V_2$ are *linearly isomorphic* if there is a bijective linear map $A : \mathrm{span}(K_1) \to \mathrm{span}(K_2)$ such that $A(K_1) = K_2$. Note that the ambient spaces of V_1 and V_2 need not be the same, and K_1 and K_2 need not be full-dimensional.

In the case of general convex sets the appropriate maps to consider are *affine maps*. These are maps $T : U \to W$ between finite dimensional real vector spaces that have the form $T(u) = Au + b$ where $A : U \to W$ is linear and $b \in W$. A pair of convex sets $C_1 \subseteq V_1$ and $C_2 \subseteq V_2$ are *affinely isomorphic* if there is a bijective affine map $T : \mathrm{aff}(C_1) \to \mathrm{aff}(C_2)$ such that $T(C_1) = C_2$. Again the ambient spaces V_1 and V_2 need not be the same, and C_1 and C_2 need not be full-dimensional.

Positive semidefinite matrices

A symmetric matrix $A \in \mathcal{S}^m$ is *positive semidefinite* if it satisfies $\langle Ax, x\rangle \geq 0$ for all $x \in \mathbb{R}^m$. An Hermitian matrix $A \in \mathcal{H}^m$ is *positive semidefinite* if it satisfies $\langle Ax, x\rangle \geq 0$ for all $x \in \mathbb{C}^n$. The set of all $m \times m$ symmetric positive semidefinite matrices forms a proper convex cone (in $\mathcal{S}^m$) denoted $\mathcal{S}^m_+$. We denote the corresponding partial order by $A \succeq B$ if $A - B \in \mathcal{S}^m_+$. The set of all $m \times m$ Hermitian positive semidefinite matrices forms a proper convex cone (in $\mathcal{H}^m$) denoted $\mathcal{H}^m_+$.

The interior of $\mathcal{S}^m_+$ is the open convex cone of strictly positive definite $m \times m$ symmetric matrices, i.e. those symmetric matrices A satisfying $\langle Ax, x\rangle > 0$ whenever $x \in \mathbb{R}^m \setminus \{0\}$. If A is strictly positive definite we write $A \succ 0$. Similarly the interior of $\mathcal{H}^m_+$ is the open convex cone of strictly positive definite $m \times m$ Hermitian matrices, i.e. those Hermitian matrices A satisfying $\langle Ax, x\rangle > 0$ whenever $x \in \mathbb{C}^m \setminus \{0\}$.

Faces

Suppose C is a convex set. An element $x \in C$ is an *extreme point* of C if whenever $y, z \in C$ and $\lambda \in (0, 1)$ are such that $x = \lambda y + (1-\lambda)z$ then $x = y = z$. A convex subset F of C is a *face* of C if whenever $y, z \in C$ and $\lambda \in (0, 1)$ are such that $\lambda y + (1-\lambda)z \in F$ then $y, z \in F$. The *dimension* of a face F is the dimension of its affine span $\mathrm{aff}(F)$. By comparing the definitions, we can see that the extreme points of C are precisely the zero-dimensional faces of C. A face F of a convex set C is *exposed* if there is an affine hyperplane (i.e. codimension one affine subspace) L such that $L \cap C = F$. An exposed point is a zero-dimensional exposed face. If K is a convex cone, an *extreme ray* of K is a one-dimensional face that is of the form $\mathrm{cone}(x)$ for some $x \in K$.

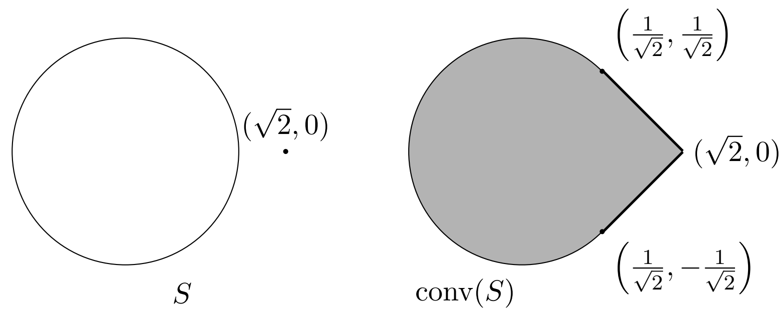


Figure 2.7: On the left is the set $S = \{(x,y) \in \mathbb{R}^2 : x^2+y^2 = 1\} \cup \{(\sqrt{2},0)\}$, the union of a circle and a point. On the right is the convex hull of S. The convex set $\mathrm{conv}(S)$ has two faces that are not exposed, these are $\left(\frac{1}{\sqrt{2}},\pm\frac{1}{\sqrt{2}}\right)$. The two one-dimensional faces of $\mathrm{conv}(S)$ are indicated with a thicker line. See Example 2.3.1 for a discussion of the faces of this convex set.

Example 2.3.1. On the right in Figure 2.7 is the convex set

$$\mathrm{conv}(S) = \mathrm{conv}(\{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\} \cup \{(\sqrt{2},0)\}),$$

the convex hull of the unit circle and a point outside the circle. We now describe the faces of this convex set. The one-dimensional faces are the convex sets $F_1 = \mathrm{conv}\{(\sqrt{2},0),(1/\sqrt{2},1/\sqrt{2})\}$ and $F_2 = \mathrm{conv}\{(\sqrt{2},0),(1/\sqrt{2},-1/\sqrt{2})\}$. These are exposed faces since, for instance, $F_1 = \mathrm{conv}(S) \cap L$ where $L = \{(x,y) \in \mathbb{R}^2 : x + y = \sqrt{2}\}$. The zero-dimensional faces (i.e. extreme points) are $(\sqrt{2},0)$ and each point of the form $(\cos(\theta),\sin(\theta))$ for $\theta \in [\pi/4, 7\pi/4]$. All of these are exposed points *except* the points $(\cos(\pi/4),\sin(\pi/4)) = (1/\sqrt{2},1/\sqrt{2})$ and $(\cos(7\pi/4),\sin(7\pi/4)) = (1/\sqrt{2},-1/\sqrt{2})$. Any affine line passing through either of these points necessarily intersects with other points of $\mathrm{conv}(S)$.

Example 2.3.2 (Faces of the positive semidefinite cone)**.** The faces of $\mathcal{S}^m_+$, the cone of $m \times m$ positive semidefinite matrices, are in bijection with subspaces U of $\mathbb{R}^m$. If U is a subspace of $\mathbb{R}^m$ then $F_U := \{A \in \mathcal{S}^m_+ : \mathrm{col}(A) \subseteq U\}$ is a face of $\mathcal{S}^m_+$. Moreover all of the faces of $\mathcal{S}^m_+$ are of this form (see, e.g., [101]). These are all exposed faces since each F_U can alternatively be expressed as $F_U = \mathcal{S}^m_+ \cap \{A \in \mathcal{S}^m : \langle P_{U^\perp}, A\rangle = 0\}$ where $P_{U^\perp} = I - P_U$ is the orthogonal projector onto the orthogonal complement of U.

■ 2.3.3 Convex functions

Let V be a finite-dimensional real vector space. When discussing convex functions on V it is useful to allow functions that take values in the extended real line $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$

with arithmetic on $-\infty$ and ∞ defined appropriately (see, e.g., [105, Section 4]). With any function $f : V \to \overline{\mathbb{R}}$ we associate its *epigraph*, the set

$$\operatorname{epi}(f) = \{(x,t) \in V \times \mathbb{R} : f(x) \leq t\}.$$

A function $f : V \to \overline{\mathbb{R}}$ is *convex* if its epigraph is a convex set. A function $f : V \to \overline{\mathbb{R}}$ is concave if $-f$ is convex. If f does not take the value $-\infty$ then a more familiar definition of convexity makes sense. Indeed $f : V \to \mathbb{R} \cup \{\infty\}$ is convex if and only if $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$ for all $x, y \in V$ and all $\alpha \in [0,1]$.

It can also be useful to regard sets as (extended real-valued) functions. If $S \subseteq V$ is a set then its *indicator function* is $\iota_S : V \to \overline{\mathbb{R}}$ defined by

$$\iota_S(x) = \begin{cases} 0 & \text{if } x \in S \\ \infty & \text{otherwise.} \end{cases}$$

If S is a convex set then ι_S is a convex function.

■ 2.3.4 Notions of duality

Notions of duality are central to convex geometry, analysis, and optimization. These allow us to view convex sets not just in terms of the points in the set but also in terms of the values that linear functionals take on the set.

Definition 2.3.3. If $S \subseteq V$ is a subset of a real inner product space V then the *dual cone* of S is

$$S^* = \{y \in V : \langle y, x\rangle \geq 0, \;\; \text{for all } x \in S\}.$$

The dual cone S^* is always a closed convex cone since it is given by the intersection of half-spaces passing through the origin in V, each of which is a closed convex cone. The dual cone of a set S is shown in Figure 2.8.

Definition 2.3.4. If $S \subseteq V$ is a subset of a real inner product space V then the *polar* of S is

$$S^\circ = \{y \in V : \langle y, x\rangle \leq 1, \;\; \text{for all } x \in S\}.$$

The polar S° is a closed convex set (again it is the intersection of half-spaces) containing the origin in V.

Example 2.3.5. If $a \neq 0$ is a point in $\mathbb{R}^n$ then its polar is

$$\{a\}^\circ = \{y \in \mathbb{R}^n : \langle a, y\rangle \leq 1\},$$

a closed *half-space* containing the origin.

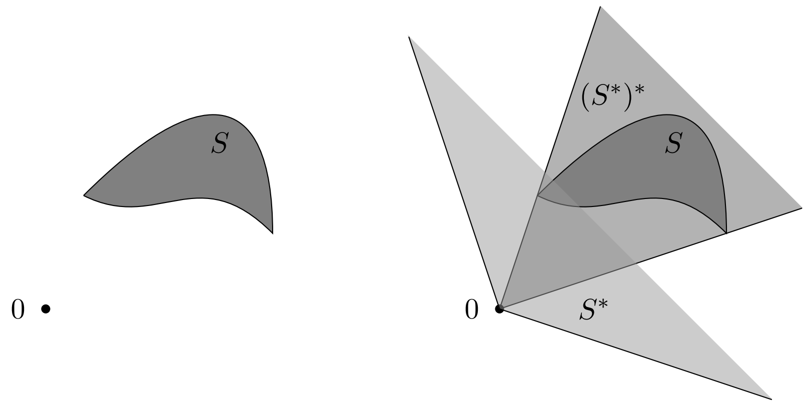


Figure 2.8: Shown on the left is a set S together with the origin. Shown on the right is S together with its dual cone S^*, and the dual of the dual cone $(S^*)^*$. Both S^* and $(S^*)^*$ are closed convex cones. Furthermore, $(S^*)^*$ is the closure of the conic hull of S (by Proposition 2.3.7).

Example 2.3.6. If K is a convex cone then $K^\circ = -K^*$. This holds because K is closed under non-negative scaling so that

$$\begin{aligned} K^\circ &= \{y : \langle x, y\rangle \le 1, \text{ for all } x \in K\} \\ &= \{y : \langle tx, y\rangle \le 1, \text{ for all } t \ge 0 \text{ and all } x \in K\} \quad (\text{since } tK = K \text{ for all } t \ge 0) \\ &= \{y : \langle x, y\rangle \le 0, \text{ for all } x \in K\} = -K^*. \end{aligned}$$

The following biduality results give concrete ways to describe the (closure of the) conic and convex hulls of S.

Proposition 2.3.7 ([105, Section 14])**.** *Let $S \subseteq V$ be a subset of a real inner product space V. Then*

$$\mathrm{cl}(\mathrm{cone}(S)) = (S^*)^* \quad \textit{and} \quad \mathrm{cl}(\mathrm{conv}(S \cup \{0\})) = (S^\circ)^\circ.$$

Note that this biduality theorem gives an alternative characterization of the closure of the convex and conic hulls of a set. Rather than describing these in terms of combinations of points in S, biduality describes these sets in terms of half-spaces that contain S. A set S together with the dual of its dual cone $(S^*)^*$ is shown in Figure 2.8

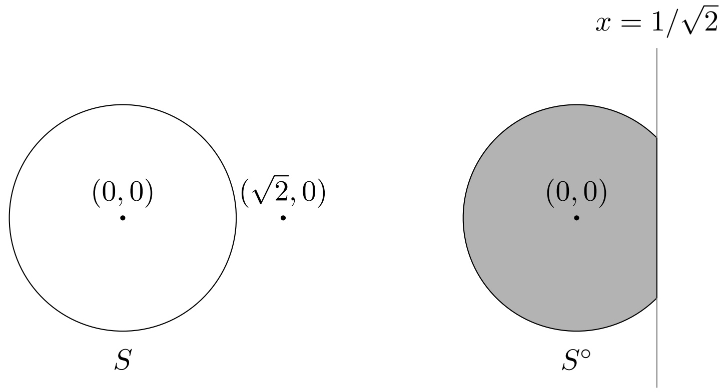


Figure 2.9: On the left is the set $S = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\} \cup \{(\sqrt{2},0)\}$ and on the right is its polar $S^\circ = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1,\ x \leq 1/\sqrt{2}\}$.

Useful duality results

The following lemma is useful for understanding the polars of the unions of sets. It follows in a straightforward way from the definition of the polar of a set.

Lemma 2.3.8 ([105, Corollary 16.5.2])**.** *If S_1, S_2 are subsets of a finite-dimensional real inner product space V then*

$$(S_1 \cup S_2)^\circ = S_1^\circ \cap S_2^\circ.$$

Example 2.3.9. Consider the set $S = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\} \cup \{(\sqrt{2},0)\}$ shown on the left in Figure 2.9. Its polar is the intersection of the polar of a circle and the polar of a point, i.e.

$$S^\circ = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}^\circ \cap \{(\sqrt{2},0)\}^\circ.$$

The polar of the unit circle is the unit disk, and the polar of a point is a half-space. Hence

$$S^\circ = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1,\ x \leq 1/\sqrt{2}\}$$

as shown on the right in Figure 2.9.

The following result tells us about the polar of the image of a set under a linear map. It is also straightforward consequence of the definitions.

Lemma 2.3.10 ([105, Corollary 16.3.2])**.** *Let V, W be finite dimensional real inner product spaces and $B : W \to V$ a linear map. If $C \subseteq V$ is a subset of V then*

$$(B(C))^\circ = \{w \in W : B^*(w) \in C^\circ\}.$$

Duality relations are less straightforward when they involve the intersection of a convex set and a subspace (taking a 'slice' of the convex set). Either additional closure operations are required (as in (2.3.3) to follow) or additional assumptions (such as the existence of x_0 in Lemmas 2.3.11 and 2.3.12 to follow) are required that ensure the intersection is sufficiently 'general'. Note also that while Lemma 2.3.10 holds for an arbitrary set C, the following result requires a convex set C.

Lemma 2.3.11 ([105, Corollary 16.3.2])**.** *Let V and W be finite-dimensional real inner product spaces and $A : W \to V$ a linear map. If $C \subseteq V$ is a convex subset of V then*

$$\{x \in W : A(x) \in \mathrm{cl}(C)\}^\circ = \mathrm{cl}(A^*(C^\circ)). \tag{2.3.3}$$

If there exists $x_0 \in W$ such that $A(x_0)$ is in the relative interior of C then

$$\{x \in W : A(x) \in C\}^\circ = A^*(C^\circ).$$

Combining the previous two results allows us to express the dual cone of a projection of a slice of a closed convex cone K in terms of a projection of a slice of the dual cone K^*.

Lemma 2.3.12. *Let W, V_1, V_2 be finite-dimensional real inner product spaces. Suppose $K_1 \subseteq V_1$ is a closed convex cone and $A : W \to V_1$ and $B : W \to V_2$ are linear maps. Let*

$$K_2 = \{B(x) : A(x) \in K_1\} \subseteq V_2.$$

Furthermore, assume there is some $x_0 \in V_1$ such that $A(x_0)$ is in the relative interior of K_1. Then

$$K_2^* = \{w \in V_2 : \exists y \in K_1^* \text{ s.t. } B^*(w) = A^*(y)\}.$$

Proof. If K_2 is a convex cone then $K_2^* = -K_2^\circ$ (by Example 2.3.6), so we can apply the results of Lemmas 2.3.10 and 2.3.11 to find K_2° and then appropriately change the sign. Let $C = \{x \in W : A(x) \in K_1\}$. Then since K_1 is closed and convex and there is $x_0 \in W$ such that $A(x_0)$ is in the relative interior of K_1, it follows from Lemma 2.3.11 that

$$C^\circ = A^*(K_1^\circ) = \{A^*(y) : y \in K_1^\circ\}.$$

Combining this with Lemma 2.3.10 we see that

$$K_2^\circ = (B(C))^\circ = \{w \in W : B^*(y) \in C^\circ\} = \{w \in W : \exists y \in K_1^\circ \text{ s.t. } B^*(w) = A^*(y)\}.$$

To complete the argument we use the fact that $K_2^\circ = -K_2^*$ and that $K_1^\circ = -K_1^*$. $\square$

Convex conjugates and support functions

If $f : V \to \overline{\mathbb{R}}$ is a function then its *Fenchel conjugate* is a function $f^* : V \to \mathbb{R}$ defined by

$$f^*(y) = \sup_x \langle y, x\rangle - f(x). \tag{2.3.4}$$

The conjugate function f^* is always convex since its epigraph is the intersection of half-spaces, one for each value of x. It follows that $(f^*)^* : V \to \mathbb{R}$ is a convex function, called the *convex envelope* of f. Observe that

$$(f^*)^*(x) = \sup_y \langle x, y\rangle - f^*(y)$$

is the pointwise maximum of a collection of affine functions $x \mapsto \langle x, y\rangle - f^*(y)$. Each of these affine functions is a lower bound on $x \mapsto f(x)$ since the inequality

$$f(x) \geq \langle x, y\rangle - f^*(y) \quad \text{for all } x \text{ and all } y$$

is a restatement of (2.3.4). As such the convex envelope is the pointwise maximum of all the affine functions that are global lower bounds on f.

An important case of this construction is the conjugate function of an indicator function. If $S \subseteq V$ is a set then its *support function* is $h_S : V \to \overline{\mathbb{R}}$ defined by

$$h_S(y) = \iota_S^*(y) = \sup_{x\in S} \langle y, x\rangle. \tag{2.3.5}$$

Note that h_S is a convex function for any set S and is the conjugate function of the indicator function of S. Furthermore, $h_S(y) = h_{\mathrm{cl}(\mathrm{conv}\, S)}(y)$ for all y (see, e.g., [105, Theorem 32.2] specialized to the case of linear functions).

2.4 Semidefinite representations and semidefinite optimization

In this section we summarize basic facts about the family of semidefinite optimization problems before focusing our attention on the convex sets that arise as the feasible regions of semidefinite optimization problems.

2.4.1 Semidefinite optimization

A *semidefinite optimization problem* is an optimization problem of the form

$$\min_X \langle C, X\rangle \quad \text{subject to} \quad X \in L \cap \mathcal{S}_+^m \tag{2.4.1}$$

where L is an affine subspace of $\mathcal{S}^m$ and $\mathcal{S}_+^m$ is the cone of $m \times m$ symmetric positive semidefinite matrices. It is typical to write semidefinite optimization problems via an explicit parameterization of the affine subspace L appearing in (2.4.1). Suppose we parameterize L as

$$L = \{X \in \mathcal{S}^m : \mathcal{A}(X) = b\} = X_0 + \mathrm{null}(\mathcal{A})$$

where $b \in \mathbb{R}^n$, $\mathcal{A} : \mathcal{S}^m \to \mathbb{R}^n$ is a linear map, $X_0 \in \mathcal{S}^m$ satisfies $\mathcal{A}(X_0) = b$ and $\mathrm{null}(\mathcal{A})$ is the nullspace of $\mathcal{A}$. Using this description for L, (2.4.1) becomes

$$\min_X \langle C, X \rangle \quad \text{subject to} \quad X \succeq 0,\ \ \mathcal{A}(X) = b. \tag{2.4.2}$$

An optimization problem that is dual to (2.4.2) (in a sense that is discussed under the heading 'duality results' to follow) is

$$\max_y \langle b, y \rangle \quad \text{subject to} \quad C - \mathcal{A}^*(y) \succeq 0. \tag{2.4.3}$$

This is again a semidefinite optimization problem. To see this note that if $\mathrm{col}(\mathcal{A}^*)$ denotes the column space of $\mathcal{A}^*$ then the feasible region of (2.4.3) is affinely isomorphic to the intersection of $\mathcal{S}_+^m$ with the affine subspace $C + \mathrm{col}(\mathcal{A}^*)$. Moreover the objective function is linear.

Semidefinite optimization problems are of interest because they can be solved efficiently to numerical accuracy (both in theory and in practice) using interior point methods [89] and can also model a wide range of optimization problems (see, e.g., [87]). The main property of semidefinite optimization problems that allows the application of interior point methods is that their feasible regions have certain well-behaved barrier functions called *self-concordant barrier functions* [89]. In addition to being self-concordant, to obtain efficient algorithms it is necessary that the gradient and the Hessian of the barrier function can be evaluated efficiently, and that a quantity called the *barrier parameter* [89] is not too large. In particular, the function $X \mapsto -\log\det(X)$ is a self-concordant barrier function for $\mathcal{S}_+^m$ with barrier parameter m. Restricting the function $-\log\det(\cdot)$ to the affine subspace L gives a self-concordant barrier function for $L \cap \mathcal{S}_+^m$ with barrier parameter m.

Duality results

We now summarize basic duality results for the pair (2.4.2) and (2.4.3). The first result is known as *weak duality*, and holds just by virtue of the structure of the primal and dual problems.

Lemma 2.4.1. *If X is feasible for (2.4.2) and y is feasible for (2.4.3) then $\langle C, X \rangle \geq \langle y, b \rangle$.*

Proof. Since X is feasible for (2.4.2) we have that $\mathcal{A}(X) = b$. Hence

$$\langle C, X\rangle - \langle y, b\rangle = \langle C, X\rangle - \langle y, \mathcal{A}(X)\rangle = \langle C, X\rangle - \langle \mathcal{A}^*(y), X\rangle = \langle C - \mathcal{A}^*(y), X\rangle \geq 0$$

where the last equality holds because $C - \mathcal{A}^*(y) \succeq 0$ and $X \succeq 0$. □

Let $p_\star$ denote the optimal value of (2.4.2) and $d_\star$ the optimal value of (2.4.3), adopting the usual convention that $p_\star = -\infty$ if the problem is unbounded and $p_\star = \infty$ if the problem is infeasible, and that $d_\star = \infty$ if the problem is unbounded and $d_\star = -\infty$ if the problem is infeasible (these are reversed because one is a maximization problem, the other a minimization problem). We could restate Lemma 2.4.1 in this language as $d_\star \leq p_\star$.

In general strong duality does not hold, i.e. we do not always have $p_\star = d_\star$. The optimal value of the primal problem and the optimal value of the dual problem can both be finite and yet not coincide (for instance see the example after Theorem 3.1 of [131]). It is also possible that the optimal primal or dual values are not achieved by any feasible point, i.e. there is a sequence of feasible points with objective value approaching the optimal value, but no feasible point with objective value equal to the optimal value (see, for instance, the same example after Theorem 3.1 of [131]). Nevertheless, under fairly mild additional hypotheses, these situations can be ruled out. The following paraphrases [131, Theorem 3.1].

Theorem 2.4.2. *If either*

1. *there is $X_0 \succ 0$ such that $\mathcal{A}(X_0) = b$ (i.e. the problem (2.4.2) is strictly feasible) or*

2. *there is y such that $C - \mathcal{A}^*(y) \succ 0$ (i.e. the problem (2.4.3) is strictly feasible)*

then $p_\star = d_\star$. Furthermore if both conditions 1 and 2 hold, then the primal and dual optimal values are both achieved.

Under similar hypotheses to those that ensure strong duality holds, there is a simple characterization of the optimality conditions for the semidefinite optimization problem (2.4.2). We choose this rather asymmetric statement of the optimality conditions because it is most natural in Chapter 6.

Theorem 2.4.3. *If $X_\star$ is feasible for (2.4.2) and there exists $y_\star$ such that*

$$C - \mathcal{A}^*(y_\star) \succeq 0 \quad \textit{and} \quad X_\star(C - \mathcal{A}^*(y_\star)) = 0 \tag{2.4.4}$$

then $X_\star$ is optimal for (2.4.2) (and $y_\star$ is optimal for (2.4.3)). Conversely, if the primal (2.4.2) and the dual (2.4.3) are strictly feasible and $X_\star$ is optimal for (2.4.2) then there exists $y_\star$ such that (2.4.4) holds.

Proof. If some $y_\star$ satisfying (2.4.4) exists then

$$0 = \mathrm{tr}(X_\star(C - \mathcal{A}^*(y_\star))) = \langle C, X_\star\rangle - \langle \mathcal{A}(X_\star), y_\star\rangle = \langle C, X_\star\rangle - \langle b, y_\star\rangle. \tag{2.4.5}$$

Let X be any primal feasible point. Then by weak duality (and the fact that $y_\star$ is dual feasible), $\langle C, X\rangle \geq \langle b, y_\star\rangle$. Then by (2.4.5) we have $\langle C, X\rangle \geq \langle b, y_\star\rangle = \langle C, X_\star\rangle$. Hence $X_\star$ is optimal.

On the other hand let $X_\star$ be optimal for (2.4.2). If both the primal and dual problems are strictly feasible then by Theorem 2.4.2 the primal and dual optimal values are the same and are achieved. Hence there exists $y_\star$ that satisfies $C - \mathcal{A}^*(y_\star) \succeq 0$ and $\langle y_\star, b\rangle = \langle C, X_\star\rangle$. Equivalently $\langle C - \mathcal{A}^*(y_\star), X_\star\rangle = 0$. Since both $X_\star \succeq 0$ and $C - \mathcal{A}^*(y_\star) \succeq 0$ it follows that $(C - \mathcal{A}^*(y_\star))X_\star = 0$[6]. Hence $y_\star$ satisfies (2.4.4). □

■ 2.4.2 Spectrahedral and semidefinite representations

In this section we discuss two families of convex sets related to semidefinite optimization. The first family are the *spectrahedra* (see Definition 2.4.4 to follow). These are the feasible regions (up to the notion of affine isomorphism discussed in Section 2.3.2) of semidefinite optimization problems. In their foundational paper [101], Ramana and Goldman coined the term 'spectrahedra' for these sets and established many of the basic properties of these convex sets.

Definition 2.4.4. A convex set C is a *spectrahedron* if there exists a positive integer m and an affine subspace L of $\mathcal{S}^m$ such that C is affinely isomorphic to $L \cap \mathcal{S}^m_+$.

The second family of convex sets we discuss in this section are the *semidefinitely representable* sets (see Definition 2.4.5 to follow). These are the images of spectrahedra under linear maps that are possibly not injective. Consequently, semidefinitely representable convex sets are also known as *projected spectrahedra*.

Definition 2.4.5. A convex set $C \subseteq \mathbb{R}^n$ is *semidefinitely representable* if there exists a positive integer m, an affine subspace L of $\mathcal{S}^m$ and a linear map $\pi : \mathcal{S}^m \to \mathbb{R}^n$ such that $C = \pi(L \cap \mathcal{S}^m_+)$.

At first glance these definitions look very similar. The key difference is the map $\pi : \mathcal{S}^m \to \mathbb{R}^n$ appearing in the definition of a semidefinitely representable set. Any spectrahedron is also semidefinitely representable. However, it is not at all clear from the definition whether the family of semidefinitely representable sets is strictly larger than the family of spectrahedra, i.e. whether we can obtain more sets by allowing maps π that

[6] We have used the fact that if $A, B \in \mathcal{S}^m_+$ then $\langle A, B\rangle = 0$ if and only if $AB = 0$ [14, Corollary A.24].

are not injective. It turns out that this is the case; there are semidefinitely representable sets that are not spectrahedra. We describe one such set in Example 2.4.10.

By comparing the definition of a spectrahedron (Definition 2.4.4) with the description of a semidefinite optimization problem in (2.4.2) we can see that spectrahedra are the feasible regions of semidefinite optimization problems. We are also interested in semidefinitely representable sets in the context of semidefinite optimization, because we can optimize a linear functional over a semidefinitely representable set using semidefinite optimization. We now explain why this is the case. Suppose $C = \pi(L \cap \mathcal{S}^m_+)$ is a semidefinitely representable set where L is an affine subspace of $\mathcal{S}^m$ and $\pi : \mathcal{S}^m \to \mathbb{R}^n$ is a linear map. Then we can rewrite the optimization problem

$$\min_x \langle c, x\rangle \quad \text{s.t.} \quad x \in C = \pi(L \cap \mathcal{S}^m_+)$$

as a semidefinite optimization problem. Indeed an equivalent problem is

$$\min_X \langle c, \pi(X)\rangle \quad \text{s.t.} \quad X \in L \cap \mathcal{S}^m_+.$$

By rewriting the objective function we obtain a problem in the form

$$\min_X \langle \pi^*(c), X\rangle \quad \text{s.t.} \quad X \in L \cap \mathcal{S}^m_+,$$

i.e. the standard form of a semidefinite optimization problem (2.4.2). In other words, instead of optimizing the linear functional defined by c over the set C, we can optimize the linear functional defined by $\pi^*(c)$ over the set $L \cap \mathcal{S}^m_+$, which is a semidefinite optimization problem.

The cost of solving a semidefinite optimization problem depends on the size of the positive semidefinite matrices involved in the formulation (i.e. on the parameter m in (2.4.2)). As such, it is useful to have refinements of the notions of spectrahedra and semidefinitely representable sets that keep track of this size parameter.

Definition 2.4.6. A convex set C has a *spectrahedral representation of size* m if there exists an affine subspace L of $\mathcal{S}^m$ such that C is affinely isomorphic to $L \cap \mathcal{S}^m_+$.

This differs from the definition of a spectrahedron only in that it includes the size parameter m in the definition.

Definition 2.4.7. A convex set $C \subseteq \mathbb{R}^n$ has a *semidefinite representation of size* m if there exists an affine subspace L of $\mathcal{S}^m$ and a linear map $\pi : \mathcal{S}^m \to \mathbb{R}^n$ such that $C = \pi(L \cap \mathcal{S}^m_+)$.

Again, this differs from the definition of a semidefinitely representable set only in that it includes the size parameter m in the definition.

We can explicitly parameterize the affine subspaces L that appear in Definitions 2.4.6 and 2.4.7. Any affine subspace L of $\mathcal{S}^m$ with dimension at most n can be parameterized as

$$L = \left\{ A_0 + \sum_{i=1}^{n} A_i x_i \;:\; x \in \mathbb{R}^n \right\}$$

where $A_0, A_1, \ldots, A_n \in \mathcal{S}^m$ (see (2.3.1)). As such, C has a spectrahedral representation of size m if and only if C is affinely isomorphic to a set of the form

$$\left\{ x \in \mathbb{R}^n : A_0 + \sum_{i=1}^{n} A_i x_i \in \mathcal{S}_+^m \right\}$$

where $A_0, A_1, \ldots, A_n \in \mathcal{S}^m$. Similarly a convex set $C \subseteq \mathbb{R}^n$ has a semidefinite representation of size m if and only if it is of the form

$$C = \left\{ x \in \mathbb{R}^n : \exists y \;\text{ s.t. }\; x = \pi(y) \;\text{ and }\; A_0 + \sum_{i=1}^{n} A_i y_i \in \mathcal{S}_+^m \right\}$$

where $\pi : \mathcal{S}^m \to \mathbb{R}^n$ and $A_0, A_1, \ldots, A_n \in \mathcal{S}^m$.

Polars of spectrahedra

In Lemma 2.4.8 to follow we give explicit descriptions for the polar of a spectrahedron. These descriptions show that the polar of a spectrahedron is semidefinitely representable. Since the second description in Lemma 2.4.8 is not so standard, we include a proof.

Lemma 2.4.8. *Let $\mathcal{A} : \mathcal{S}^m \to \mathbb{R}^n$ be a linear map with adjoint $\mathcal{A}^* : \mathbb{R}^n \to \mathcal{S}^m$. Suppose $C = \{x \in \mathbb{R}^n : A_0 - \mathcal{A}^*(x) \in \mathcal{S}_+^m\}$ is a spectrahedron with non-empty interior and suppose there is $e \in \mathbb{R}^n$ such that $A_0 - \mathcal{A}^*(e) \succ 0$. Then*

$$C^\circ = \{y \in \mathbb{R}^n : \exists Z \;\; s.t. \;\; y = \mathcal{A}(Z), \;\; \langle A_0, Z\rangle \le 1, \;\; Z \in \mathcal{S}_+^m\}. \tag{2.4.6}$$

Suppose, in addition, there is some $Z_0 \in \mathcal{S}_+^m$ such that $\langle A_0, Z_0\rangle = 1$ and $\mathcal{A}(Z_0) = 0$. Then

$$C^\circ = \{y \in \mathbb{R}^n : \exists Z \;\; s.t. \;\; y = \mathcal{A}(Z), \;\; \langle A_0, Z\rangle = 1, \;\; Z \in \mathcal{S}_+^m\}. \tag{2.4.7}$$

Proof. The characterization in (2.4.6) is given, for instance, in [55]. (It could also be deduced directly from Lemma 2.3.11 in Section 2.3 applied to the set $\{A_0\} - \mathcal{S}_+^m$ and observing that its polar is $\{Z \in \mathcal{S}^m : \langle A_0, Z\rangle \le 1, \;\; Z \in \mathcal{S}_+^m\}$.)

We now establish the second description of C° in (2.4.7). Suppose there is $Z_0 \in \mathcal{S}_+^m$ such that $\langle A_0, Z_0\rangle = 1$ and $\mathcal{A}(Z_0) = 0$. Under this additional assumption we show that

$$\{y \in \mathbb{R}^n : \exists Z \text{ s.t. } y = \mathcal{A}(Z),\ \ \langle A_0, Z\rangle \leq 1,\ \ Z \in \mathcal{S}_+^m\} \subseteq$$
$$\{y \in \mathbb{R}^n : \exists Z \text{ s.t. } y = \mathcal{A}(Z),\ \ \langle A_0, Z\rangle = 1,\ \ Z \in \mathcal{S}_+^m\} \qquad (2.4.8)$$

which is enough to establish (2.4.7) (since the other inclusion is obvious). To establish (2.4.8), define

$$Z' = Z + (1 - \langle A_0, Z\rangle)Z_0.$$

Then since $Z \in \mathcal{S}_+^m$ and $Z_0 \in \mathcal{S}_+^m$ it follows that $Z' \in \mathcal{S}_+^m$. Furthermore,

$$\langle A_0, Z'\rangle = \langle A_0, Z\rangle + \langle A_0, Z_0\rangle - \langle A_0, Z\rangle\langle A_0, Z_0\rangle = \langle A_0, Z\rangle + 1 - \langle A_0, Z\rangle = 1.$$

Finally $\mathcal{A}(Z') = \mathcal{A}(Z + (1 - \langle A_0, Z\rangle)Z_0) = \mathcal{A}(Z) = y$ since $\mathcal{A}(Z_0) = 0$. Hence Z' is an element of the right hand of (2.4.8) as we require. □

Faces of spectrahedra

If $L \cap \mathcal{S}_+^m$ is a spectrahedron, then all of its faces are of the form $L \cap F_U$ where F_U is the face of the positive semidefinite cone corresponding to the subspace $U \subseteq \mathbb{R}^m$ (see Example 2.3.2). An important characteristic that spectrahedra inherit from the positive semidefinite cone is that all of their faces are exposed.

Lemma 2.4.9 (Ramana and Goldman [101])**.** *All boundary faces of a spectrahedron are exposed.*

One may wonder whether the faces of semidefinitely representable sets are always exposed. This is not the case as the following example shows.

Example 2.4.10. In this example we describe a semidefinitely representable set with non-exposed faces.

Let $S = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\} \cup \{(\sqrt{2}, 0)\}$ be the set shown on the left in Figure 2.7. Consider its polar, the convex set $C = S^\circ = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1,\ x \leq 1/\sqrt{2}\}$ shown on the right of Figure 2.9. First note that C has a spectrahedral representation as

$$C = \left\{(x, y) \in \mathbb{R}^2 : \begin{bmatrix} 1/\sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + x \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + y \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \succeq 0 \right\}.$$

We now explain why this representation is valid. Since the matrices in the representation of C are block diagonal, and after a little rearrangement, we obtain the equivalent description

$$C = \left\{ (x,y) \in \mathbb{R}^2 : x \leq 1/\sqrt{2},\ \begin{bmatrix} 1-x & y \\ y & 1+x \end{bmatrix} \succeq 0 \right\}.$$

Since a 2×2 symmetric matrix A is positive semidefinite if and only if $\mathrm{tr}(A) \geq 0$ and $\det(A) \geq 0$ we see that the conditions

$$\begin{bmatrix} 1-x & y \\ y & 1+x \end{bmatrix} \succeq 0 \quad \text{and} \quad \det\left(\begin{bmatrix} 1-x & y \\ y & 1+x \end{bmatrix} \right) \geq 0 \quad \text{and} \quad x^2 + y^2 \leq 1$$

are all equivalent.

The polar of C is $C^\circ = (S^\circ)^\circ = \mathrm{conv}(S)$ (where we have used the biduality result Proposition 2.3.7 and the fact that $\mathrm{conv}(S)$ is closed and $0 \in \mathrm{conv}(S)$). The set C° is the convex set shown on the right in Figure 2.7. On the one hand this set is semidefinitely representable because it is the polar of a spectrahedron, and the polar of a spectrahedron is semidefinitely representable by Lemma 2.4.8. On the other hand, we have already seen (in Example 2.3.1) that C° has non-exposed faces. Hence C° is a semidefinitely representable set with non-exposed faces.

Since the faces of spectrahedra are always exposed, it follows that the family of spectrahedra is a strict subset of the family of semidefinitely representable sets. Moreover, this example also shows that the polar of a spectrahedron need not be a spectrahedron.

Size of spectrahedral and semidefinite representations

Suppose C has a *spectrahedral* representation of size m. This means we can optimize a linear functional over C by solving a semidefinite optimization problem involving $m \times m$ positive semidefinite matrices. It may be the case that C has a (different) *semidefinite* representation of much smaller size. In other words we may be able to express C as the projection of a spectrahedron in a higher dimensional space that has a much *simpler* description.

An example of this arises in Chapter 4. Let C be the nuclear norm ball in $\mathbb{R}^{n \times n}$, i.e. the set of $n \times n$ real matrices such that the sum of the singular values is at most one. This is a spectrahedron (see Theorem 4.1.2 from Chapter 4) and the size of the smallest possible spectrahedral representation of this set is 2^n (see Theorem 4.1.4 from Chapter 4). On the other hand it is known that C has a semidefinite representation of size $2n$ (see Example 4.1.5 from Chapter 4).

Expressive power and limitations of semidefinite optimization

Studying which convex sets have semidefinite representations tells us which optimization problems we can solve via semidefinte optimization. By also paying attention to the size of such representations, we can also keep track of the complexity of the resulting semidefinite optimization problems.

All semidefinitely representable sets are convex and semialgebraic[7]. Helton and Nie [63] have investigated the general question of which convex semialgebraic sets have semidefinite representations, leading to the following conjecture.

Conjecture 2.4.11 (Helton-Nie [63])**.** *Every convex semialgebraic subset of $\mathbb{R}^n$ is semidefinitely representatable.*

This conjecture has been resolved in the case of subsets of $\mathbb{R}^2$ by Scheiderer.

Theorem 2.4.12 (Scheiderer [119, Theorem 6.7])**.** *Every convex semialgebraic subset of $\mathbb{R}^2$ is semidefinitely representatable.*

In practice, we are interested not just in whether such representations exist, but also the size of the smallest such representation. Very few techniques are available for establishing lower bounds on the size of semidefinite representations of convex sets. Most of the work carried out in this direction has focused on lower bounds on the size of semidefinite representations of polytopes, i.e. convex sets obtained by taking the convex hull of finitely many points. Perhaps the only strong lower bound on the size of semidefinite representations of an explicit polytope is in the recent work of Lee et al. [76]. In that work it is established (among other results) that any semidefinite representation of the cut polytope, i.e. $\operatorname{conv}\{(x_ix_j)_{1\leq i<j\leq n} : x_i \in \{-1,1\} \quad \text{for } i = 1,2,\ldots,n\}$ must have exponential size. For a survey of what is known about the limitations on the expressive power of semidefinite representations and related topics, and a long list of interesting open problems, see [42].

■ 2.5 Hyperbolic polynomials and hyperbolicity cones

In this section we describe a family of convex cones that are constructed from multivariate polynomials with certain restrictions on their zeros. These *hyperbolicity cones* (see Definition 2.5.2 to follow) include many familiar convex cones, most notably the cone of real symmetric $m \times m$ positive semidefinite matrices. Hyperbolic optimization problems are problems involving maximizing (or minimizing) a linear functional over the intersection of a hyperbolicity cone and an affine subspace. In Section 2.5.4 we briefly discuss what is known, and what is conjectured to be true, about the relationship between semidefinite optimization and hyperbolic optimization.

[7]A subset of $\mathbb{R}^n$ is *basic semialgebraic* if it has the form $\{x \in \mathbb{R}^n : p_0(x) \neq 0, p_1(x) \geq 0, \ldots, p_m(x) \geq 0\}$ for some positive integer m and polynomials $p_0, p_1, \ldots, p_m$. A subset of $\mathbb{R}^n$ is *semialgebraic* if it is the union of finitely many basic semialgebraic sets (see, e.g., [14, Appendix A.4.4])

2.5.1 Hyperbolic polynomials

The following definition dates back (at least) to the 1959 work of Gårding [49] on partial differential equations.

Definition 2.5.1. A homogeneous polynomial p of degree m in n variables is *hyperbolic with respect to* $e \in \mathbb{R}^n$ if $p(e) \neq 0$ and whenever $x \in \mathbb{R}^n$ the univariate polynomial $t \mapsto p(x - te)$ has only real roots in t.

Similar definitions, that place related restrictions on the location of the zeros of multivariate polynomials, appear in combinatorics, probability theory, control theory, and statistical mechanics (see, e.g., the recent surveys [94, 134]).

Examples

- Let $p(x) = \prod_{i=1}^n x_i$ and let $e \in \mathbb{R}^n$ be the vector with all entries equal to one. Then p is hyperbolic with respect to e because if $x \in \mathbb{R}^n$, the univariate polynomial $t \mapsto p(x - te) = \prod_{i=1}^n (x_i - t)$ has roots $x_1, x_2, \ldots, x_n \in \mathbb{R}$.

- Let $p(x) = (x_1^2 + x_2^2 + \cdots + x_{n-1}^2) - x_n^2$ and let $e = (0, 0, \ldots, 0, 1)$. Then p is hyperbolic with respect to e because for any $x \in \mathbb{R}^n$ the univariate polynomial
$$t \mapsto p(x - te) = (x_1^2 + x_2^2 + \cdots + x_{n-1}^2) - (x_n - t)^2$$
has only the real roots $t = x_n \pm (x_1^2 + \cdots + x_{n-1}^2)^{1/2}$.

- Let X be a symmetric $m \times m$ matrix of indeterminates. Then the polynomial $p(X) = \det(X)$ is hyperbolic with respect to $I \in \mathcal{S}^m$. This is because if $X \in \mathcal{S}^m$ then the univariate polynomial $t \mapsto p(X - tI)$ is the characteristic polynomial of a symmetric matrix and so its roots are the (real) eigenvalues $\lambda_1(X) \geq \cdots \geq \lambda_n(X)$ of X.

- Let $A_1, A_2, \ldots, A_n \in \mathcal{S}^m$ and $e \in \mathbb{R}^n$ be such that $A_0 := \sum_{i=1}^n A_i e_i \succ 0$. Then the polynomial
$$p(x) = \det\left(\textstyle\sum_{i=1}^n A_i x_i\right)$$
is hyperbolic with respect to e. This is because if $A_0^{1/2}$ is the unique positive definite square root of A_0 we have
$$t \mapsto p(x - te) = \det(A_0) \det\left(\textstyle\sum_{i=1}^n A_0^{-1/2} A_i A_0^{-1/2} x_i - tI\right)$$
is (up to positive scaling) the characteristic polynomial of the symmetric matrix $\sum_{i=1}^n A_0^{-1/2} A_i A^{-1/2} x_i$. Hence the roots of $t \mapsto p(x - te)$ are the (real) eigenvalues of $\sum_{i=1}^{n-1} A_0^{-1/2} A_i A_0^{-1/2} x_i$.

It is less obvious that the next two examples are, in fact, hyperbolic polynomials. The first example is discussed in Section 3.1.2 of Chapter 3. For the second example, see [138, Theorem 10.2].

- Let $[n] = \{1, 2, \ldots, n\}$ and let
$$p(x) = \sum_{\substack{I\subseteq[n]\\|I|=k}} \prod_{i\in I} x_i$$
be the *elementary symmetric polynomial of degree k in n variables*. This polynomial is hyperbolic with respect to $e = (1, 1, \ldots, 1)$.

- Let M be an entry-wise non-negative $k \times n$ matrix (with $k \leq n$). Then define a polynomial in n variables of degree k by
$$p(x) = \sum_{\substack{I\subseteq[n]\\|I|=k}} \mathrm{Per}(M_I) \prod_{i\in I} x_i$$
where M_I is the square $k \times k$ submatrix of M with columns indexed by the set I and $\mathrm{Per}(A)$ is the *permanent* of a $k \times k$ matrix A, i.e.
$$\mathrm{Per}(A) = \sum_{\sigma} \prod_{i\in[k]} A_{i\sigma(i)}$$
where the sum is over all permutations σ of k symbols. The polynomial p is hyperbolic with respect to $e = (1, 1, \ldots, 1)$ [138, Theorem 10.2].

Note that we recover (up to scaling) the first example from the second by taking M to be the $k \times n$ matrix of all ones. This last example is particularly interesting because it shows that hyperbolic polynomials are not necessarily easy to evaluate. Indeed suppose a subset $I \subseteq [n]$ has $|I| = k$ and define $\epsilon_I \in \mathbb{R}^n$ to have ones in the positions indexed by I and zeros elsewhere. Then to compute $p(\epsilon_I)$ we must compute $\mathrm{Per}(M_I)$ which for a general non-negative $k \times k$ integer matrix M_I is #P-complete [130].

■ 2.5.2 Hyperbolicity cones

Definition 2.5.2. If p is hyperbolic with respect to e then the *hyperbolicity cone* corresponding to (p, e) is the connected component of $\{x \in \mathbb{R}^n : p(x) \neq 0\}$ containing e. We denote this cone by $\Lambda_{++}(p, e)$ and its closure by $\Lambda_+(p, e)$.

It is clear from the definition that the closure $\Lambda_+(p, e)$ of $\Lambda_{++}(p, e)$ is closed under non-negative scaling (i.e. it is a closed cone). The remarkable fact is that it is a *convex* cone, i.e. it is also closed under addition. This was first established by Gårding in the context of partial differential equations.

Theorem 2.5.3 (Gårding [49])**.** *If p is hyperbolic with respect to e then $\Lambda_+(p,e)$ is a closed convex cone.*

There are a number of alternative characterizations of the hyperbolicity cone, perhaps the simplest being the following.

Theorem 2.5.4. *If p is hyperbolic with respect to $e \in \mathbb{R}^n$ then*

$$\Lambda_+(p,e) = \{x \in \mathbb{R}^n : \text{all roots of } t \mapsto p(x-te) \text{ are non-negative}\}.$$

Throughout Chapter 3 we make extensive use of yet another characterization of $\Lambda_+(p,e)$, in terms of polynomial inequalities (see (3.1.1) of Chapter 3).

Examples

We now describe the hyperbolicity cones corresponding to the first group of examples from Section 2.5.1. These are all familiar convex cones.

- If $e = (1,1,\ldots,1)$ and $p(x) = \prod_{i\in[n]} x_i$ then the roots of $p(x-te)$ are simply $x_1, x_2, \ldots, x_n$. Hence the corresponding closed hyperbolicity cone is

  $$\Lambda_+(p,e) = \{x \in \mathbb{R}^n : x_i \geq 0 \ \text{ for all } i \in [n]\}.$$

- If $e = (0,0,\ldots,0,1)$ and $p(x) = \sum_{i=1}^{n-1} x_i^2 - x_n^2$ then the roots of $p(x-te)$ are $x_n \pm \left(\sum_{i=1}^{n-1} x_i^2\right)^{1/2}$. Hence the corresponding closed hyperbolicity cones is

  $$\Lambda_+(p,e) = \left\{x \in \mathbb{R}^n : \left(\sum_{i=1}^{n-1} x_i^2\right)^{1/2} \leq x_n\right\}.$$

- If $e = I$ and $p(X) = \det(X)$ (where X is symmetric) then the roots of $p(x-te)$ are the eigenvalues of X. Hence the corresponding closed hyperbolicity cone is

  $$\Lambda_+(p,e) = \{X \in \mathcal{S}^m : \lambda_i(X) \geq 0, \ \text{ for all } i = 1,2,\ldots,m\} = \{X \in \mathcal{S}^m : X \succeq 0\}.$$

- Suppose $A_1, A_2, \ldots, A_n \in \mathcal{S}^m$ and $e \in \mathbb{R}^n$ is such that $A_0 = \sum_{i=1}^n A_i e_i \succ 0$. Then the roots of $p(x-te)$ are the eigenvalues of $\sum_{i=1}^n A_0^{-1/2} A_i A_0^{-1/2} x_i$. Hence the corresponding closed hyperbolicity cones is

  $$\Lambda_+(p,e) = \{x \in \mathbb{R}^n : \textstyle\sum_{i=1}^n A_i x_i \succeq 0\},$$

i.e. it is a spectrahedral cone.

When $p(x)$ is the elementary symmetric polynomial of degree k and e is the vector with all entries equal to one, the corresponding hyperbolicity cone is called the *kth derivative relaxation* of the orthant (see Equation (3.1.4) of Chapter 3). These are central objects of interest in Chapter 3, so we defer further discussion of these cones to that chapter.

All of the boundary faces of hyperbolicity cones are exposed, a result due to Renegar [102, Theorem 23]. This allows us to deduce that the dual cones of hyperbolicity cones are *not* typically hyperbolicity cones. We show this by modifying Example 2.4.10.

Example 2.5.5. Consider polynomial, homogeneous of degree 3, defined by

$$p(x, y, z) = (x^2 + y^2 - z^2)(z/\sqrt{2} - x)$$

and let $e = (0, 0, 1)$. The polynomial p is hyperbolic with respect to e since, for any (x, y, z) the roots of $t \mapsto p(x, y, z - t)$ are $t = z - \sqrt{2}x, t = z + (x^2 + y^2)^{1/2}, t = z - (x^2 + y^2)^{1/2}$. The corresponding hyperbolicity cone is

$$\Lambda_+(p, e) = \{(x, y, z) \in \mathbb{R}^3 : (x^2 + y^2)^{1/2} \leq z,\ x \leq z/\sqrt{2}\},$$

i.e. the cone over the set C in Example 2.4.10. Its dual cone is

$$\{(x, y, z) \in \mathbb{R}^3 : (x^2 + y^2)^{1/2} \leq z\} + \text{cone}\{(-\sqrt{2}, 0, 1)\}$$

the sum of a quadratic cone and a single ray. This cone has non-exposed extreme rays. They are the extreme rays generated by $(-1/\sqrt{2}, 1/\sqrt{2}, 1)$ and $(-1/\sqrt{2}, -1/\sqrt{2}, 1)$ (this follows by a very similar argument to that from Example 2.3.1). Since all boundary faces of hyperbolicity cones are exposed, the dual cone $(\Lambda_+(p, e))^*$ of the hyperbolicity cone $\Lambda_+(p, e)$ is not a hyperbolicity cone.

■ 2.5.3 Hyperbolic optimization

Interest in hyperbolicity cones in the context of optimization was stimulated by Güler [58]. Among other things, he established that there is a natural self-concordant barrier function associated with any hyperbolicity cone.

Theorem 2.5.6 (Güler). *If p has degree m and is hyperbolic with respect to $e \in \mathbb{R}^n$ then $F(x) = -\log p(x)$ is a self-concordant barrier function with barrier parameter m.*

As such, as long as F and its gradient and Hessian can be evaluated efficiently, polynomial time (in n and m) interior point methods [89, 58] can be devised for the corresponding *hyperbolic optimization problem*:

$$\min_x \langle c, x\rangle \quad \text{subject to} \quad \mathcal{A}(x) = b,\ \ x \in \Lambda_+(p, e) \tag{2.5.1}$$

where $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^k$ is a linear map and $b \in \mathbb{R}^k$. The corresponding conic dual problem is

$$\max_y \langle y, b\rangle \quad \text{subject to} \quad c - \mathcal{A}^*(y) \in (\Lambda_+(p,e))^* \tag{2.5.2}$$

where $(\Lambda_+(p,e))^*$ is the dual cone of $\Lambda_+(p,e)$. Foundational work on this family of optimization problems has been carried out by Güler [58] and more recently by Renegar [102].

One of the challenges with hyperbolic optimization in general is that we do not have a good understanding of the dual cones $(\Lambda_+(p,e))^*$. This makes the development of primal-dual algorithms for this class of optimization problems challenging. Recent progress in this direction has been made by Renegar and Sondjaja [103] and Myklebust and Tunçel [83].

■ 2.5.4 Hyperbolic vs semidefinite optimization

Chapter 3 is devoted to describing explicit polynomial-size semidefinite representations for a particular family of hyperbolicity cones. It can be viewed in the context of a broader family of questions about the relationship between hyperbolicity cones, spectrahedral cones, and semidefinitely representable cones.

Since all spectrahedral cones are also hyperbolicity cones, and there are no known properties of spectrahedral cones that are not also properties of hyperbolicity cones, it is natural to wonder if these two classes of convex bodies are the same. This is known to be true for three-dimensional hyperbolicity cones. In fact the following stronger result was conjectured by Lax [74], and established (in different language) by Helton and Vinnikov [64] (see also [77]).

Theorem 2.5.7 (Helton-Vinnikov [64])**.** *Suppose that the degree m polynomial $p : \mathbb{R}^3 \to \mathbb{R}$ is hyperbolic with respect to $e \in \mathbb{R}^3$. Then there exist matrices $A_1, A_2, A_3 \in \mathcal{S}^m$ such that $A_1e_1 + A_2e_2 + A_3e_3 \succ 0$ and*

$$p(x_1, x_2, x_3) = \det\left(A_1x_1 + A_2x_2 + A_3x_3\right). \tag{2.5.3}$$

In dimensions greater than three, a parameter-counting argument [18] rules out the possibility that every hyperbolic polynomial has a determinantal representation generalizing (2.5.3). Nevertheless, it is possible for different hyperbolic polynomials (of different degrees) to have the same hyperbolicity cone (see, e.g., [18]). The following conjecture, which is a consequence of the Helton-Vinnikov theorem in dimension 3, is often called the generalized Lax conjecture.

Conjecture 2.5.8. *Every hyperbolicity cone is spectrahedral.*

Although it is very appealing, at present there is little evidence, positive or negative, for this conjecture.

Recall that semidefinitely representable sets (see Definition 2.4.5) are a strictly larger family of convex sets than spectrahedra. If the following were to hold then every hyperbolic optimization problem could, in principle, be solved using semidefinite optimization[8].

Conjecture 2.5.9. *Every hyperbolicity cone is semidefinitely representable.*

This conjecture is a simultaneous weakening of the generalized Lax conjecture (Conjecture 2.5.8) and the Helton-Nie conjecture (Conjecture 2.4.11). Netzer and Sanyal established that all smooth hyperbolicity cones are semidefinitely representable [90] by carefully applying results of Helton and Nie [63] to the setting of hyperbolicity cones. The resulting semidefinite representations are very large and are not explicitly defined.

2.6 Symmetry, representations, and convexity

Many of the convex bodies we study in this thesis have a large symmetry group. This is a more common situation than one might imagine. Many problems of interest involve optimization over the configurations of a system, which may have a great deal of symmetry. For instance in Chapter 4 we study optimization problems over rotation matrices, i.e. orientation-preserving rigid transformations. In this section we collect some basic results about the way group symmetry interacts with convex optimization and convex geometry. A number of these simple results are used repeatedly throughout the thesis.

In Section 2.6.1 we gather some basic definitions relating to groups and their action on vector spaces, i.e. linear representations of groups. Section 2.6.2 contains a number of useful results about the interaction between symmetry and convexity. In Section 2.6.3 we briefly discuss equivariant semidefinite representations of symmetric convex sets. These are semidefinite representations that respect, in a sense we make precise, the symmetries of the underlying set.

2.6.1 Basic definitions

Definition 2.6.1. A *group* is a set G together with a binary operation $(a, b) \mapsto ab$ such that

1. if $a, b \in G$ then $ab \in G$;
2. if $a, b, c \in G$ then $(ab)c = a(bc)$;

[8] Although this would tell us nothing about the complexity of directly solving the hyperbolic optimization problem vs solving the semidefinite optimization problem. For more refined questions along these lines that also take into account complexity issues, see [129, Chapter 12].

3. there exists $e \in G$ such that $ae = ea = a$ for all $a \in G$; and

4. for all $a \in G$ there exists $b \in G$ such that $ba = ab = e$

The maps that preserve group structure are called *group homomorphisms*.

Definition 2.6.2. If G and H are groups then $\phi : G \to H$ is a *homomorphism* if $\phi(g_1 g_2) = \phi(g_1)\phi(g_2)$ for all $g_1, g_2 \in G$.

Groups naturally arise when considering the symmetries of sets. If V is a set then the *automorphism group* $\mathrm{Aut}(V)$ of V is the group consisting of bijections $f : V \to V$ with the group operation being composition. An *action* of a group G on a set V is a homomorphism $\rho : G \to \mathrm{Aut}(V)$. In other words, this associates with each group element $g \in G$ a bijection $\rho(g) : V \to V$ so that $\rho(gh) = \rho(g) \circ \rho(h)$ for all $g, h \in G$.

If the set V has additional structure, then it is fruitful to restrict to group actions that preserve that structure. Our primary interest is in convex subsets of finite dimensional real vector spaces V. The structure-preserving automorphisms of V are the invertible linear maps. We denote these by $GL(V)$. When $V = \mathbb{R}^n$ we instead use the notation $GL(n)$. Concretely $GL(n)$ can be thought of as the group of $n \times n$ invertible matrices with the group operation being matrix multiplication.

The space V together with a linear group action is a (linear) representation of that group.

Definition 2.6.3. If G is a group and (V, ρ) is a pair consisting of a finite dimensional real vector space V and homomorphism $\rho : G \to GL(V)$, we call (V, ρ) a *representation* of G over $\mathbb{R}$.

Since we often fix an inner product on V (and use it to identify V and its dual space V^*), it is natural to further restrict to automorphisms of V that are linear and preserve the inner product. The linear maps $Q : V \to V$ such that $\langle Qu, Qv \rangle = \langle u, v \rangle$ for all $u, v \in V$ are called *orthogonal*. The group of orthogonal transformations of V is denoted $O(V)$. Again when $V = \mathbb{R}^n$ we denote this group by $O(n)$. Concretely we can think of $O(n)$ as consisting of $n \times n$ matrices that satisfy $Q^T Q = I$. We now define the representations that additionally preserve inner products.

Definition 2.6.4. A representation (V, ρ) of G over $\mathbb{R}$ is *orthogonal* if $\langle \rho(g)x, \rho(g)y \rangle = \langle x, y \rangle$ for all $x, y \in V$ and all $g \in G$.

There are obvious parallels of Definitions 2.6.3 and 2.6.4 if V is a complex vector space equipped with an Hermitian inner product. In this case we would refer to a *(unitary) representation of G over* $\mathbb{C}$ rather than an orthogonal representation of G over $\mathbb{R}$.

2.6.2 Convexity and the fixed point subspace

If a group acts on a vector space by linear transformations, the set of points that are fixed by the action is a subspace. More precisely, if (V, ρ) is a representation of G then the *fixed-point subspace* is

$$V^\rho := \{v \in V : \rho(g)v = v \text{ for all } g \in G\}.$$

Example 2.6.5. Let G be the group (under matrix multiplication) of diagonal matrices with all entries either 1 or -1. We call these *diagonal sign matrices*. Let $V = \mathcal{S}^n$ be the space of $n \times n$ symmetric matrices. The group G acts on V by conjugation, i.e. $\rho(g)(X) = gXg^T$ for all $g \in G$. Note that this is an orthogonal representation since $\langle gXg^T, gYg^T \rangle = \langle X, Y \rangle$ for all symmetric matrices X and Y. The fixed-point subspace of this action is precisely the subspace of diagonal matrices.

Suppose G is a finite group and (V, ρ) is an orthogonal representation of G. One reason the fixed point subspace is very useful in the context of convex optimization and convex geometry is that the orthogonal projector onto the fixed point subspace is a convex combination of the $\rho(g)$.

Lemma 2.6.6. *Let G be a finite group and (V, ρ) a finite-dimensional orthogonal representation of G. Then the orthogonal projector P_{V^ρ} onto the fixed-point subspace V^ρ is given by*

$$P_{V^\rho} = \frac{1}{|G|}\sum_{g\in G}\rho(g).$$

Proof. Let $P = \frac{1}{|G|}\sum_{g\in G}\rho(g)$. We will show that P is symmetric, satisfies $P^2 = P$, and has column space V^ρ. To see that P is symmetric note that

$$P^T = \frac{1}{|G|}\sum_{g\in G}\rho(g)^T = \frac{1}{|G|}\sum_{g\in G}\rho(g^{-1}) = \frac{1}{|G|}\sum_{h\in G}\rho(h)$$

where for the last equality we have made the change of variables $h = g^{-1}$ in the sum. If $g \in G$ then

$$\rho(g)P = \rho(g)\frac{1}{|G|}\sum_{h\in G}\rho(h) = \frac{1}{|G|}\sum_{h\in G}\rho(gh) = \frac{1}{|G|}\sum_{h'\in G}\rho(h') = P$$

where for the second last equality we have made the change of variables $h' = gh$ in the sum. Hence $P^2 = \frac{1}{|G|}\sum_{g\in G}\rho(g)P = P$. If $v \in V^\rho$ then

$$Pv = \frac{1}{|G|}\sum_{g\in G}\rho(g)v = \frac{1}{|G|}\sum_{g\in G} v = v$$

so the column space of P contains V^ρ. Conversely if $w \in V$ is arbitrary (so Pw is an arbitrary element of the column space of P) then $\pi(g)(Pw) = (\pi(g)P)w = Pw$ for all $g \in G$. It follows that $Pw \in V^\rho$ and so that the column space of P is contained in V^ρ. Hence the column space of P is V^ρ, completing the proof. $\square$

This result also holds for unitary representations, and for a much larger class of groups than finite groups; for instance it holds for any unimodular group [22, Chapter 1]. For these groups there is a unique probability measure on the group that is invariant under left translation (the left Haar measure [22, Chapter 1]), and a unique probability measure on the group that is invariant under right translation (the right Haar measure [22, Chapter 1]), and these measures coincide. All the steps in the proof above hold if we replace the sum with the integral with respect to such a measure.

Example 2.6.7. In the case where the group G consists of $n \times n$ diagonal sign matrices acting on $\mathcal{S}^n$ by conjugation, we have seen that the fixed-point subspace is the subspace of diagonal matrices. As such, the orthogonal projection onto diagonal matrices is given by

$$\frac{1}{2^n}\sum_{g \in G} gXg^T.$$

We note that this mapping sends positive semidefinite matrices to positive semidefinite matrices since if $X \succeq 0$ then $gXg^T \succeq 0$ for all g. Thus we recover the simple fact that the diagonal elements of positive semidefinite matrices are non-negative.

Consequences for convex optimization

The following result tells us that if a convex optimization problem has an objective function and constraint set that are invariant under the action of a group, then it has a solution that is fixed by the group.

Lemma 2.6.8. *Suppose G is a finite group and (V, ρ) is an orthogonal representation of G. Let $f : V \to \mathbb{R}$ be a convex function and $C \subseteq V$ be a convex set such that*

$$f(\rho(g)v) = f(v) \quad \text{for all } v \in V \text{ and all } g \in G \quad \text{and} \quad \rho(g)C = C \quad \text{for all } g \in G.$$

Then if the convex optimization problem $\min_{x \in C} f(x)$ has an optimal solution, it has an optimal solution in V^ρ.

Proof. Let x be any optimal solution of $\min_{x\in C} f(x)$. We claim that $P_{V^\rho}x \in V^\rho$ is also optimal. To see this note that since f is convex and invariant,

$$f(P_{V^\rho}x) = f\left(\tfrac{1}{|G|}\textstyle\sum_{g\in G}\rho(g)x\right) \leq \frac{1}{|G|}\sum_{g \in G} f(\rho(g)x) = f(x).$$

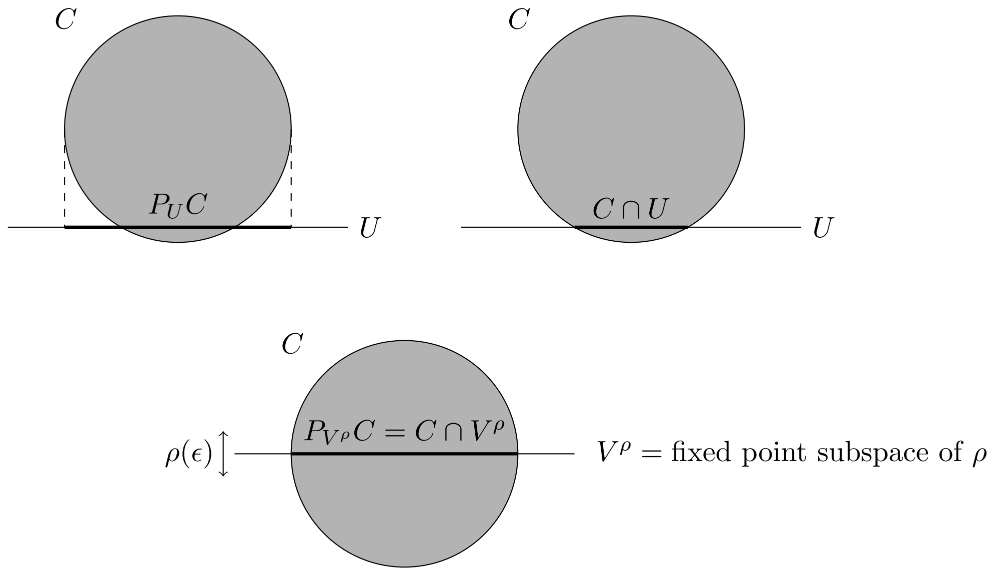


Figure 2.10: On the top left is shown a convex set C (shaded), a subspace U (the thin solid line) and the *projection* $P_U C$ of the convex set onto the subspace (the thick solid line). The dashed lines are only present to help visualize the projection. On the top right is shown the same convex set C, the same subspace U, and the *intersection* $C \cap U$ of the convex set and the subspace (the thick solid line). Comparing the top left and top right diagram, we see that $P_U C \supseteq C \cap U$. On the bottom is shown a convex set C (shaded) and a different subspace V^ρ (the thin line). In the bottom diagram the group $\{1, \epsilon\}$ consisting of two elements acts on $\mathbb{R}^2$ in such a way that $\rho(1)$ is the identity map and $\rho(\epsilon)$ is the reflection in the horizontal line shown. Clearly V^ρ is the fixed point subspace of this action and C is invariant under the action. In this case $P_{V^\rho} C = C \cap V^\rho$, i.e. the projection of C onto V^ρ and the intersection of C with V^ρ are the same (see part 1 of Lemma 2.6.9 to follow).

Furthermore since $\rho(g)x \in C$ for all $g \in G$ we have that

$$P_{V^\rho} x = \frac{1}{|G|} \sum_{g \in G} \rho(g) x$$

is a convex combination of elements of C and so is in C. Hence $P_{V^\rho} x \in C$ and has cost no larger than the cost of x, from which it follows that $P_{V^\rho} x$ is also optimal. $\square$

Geometric consequences

Suppose $C \subseteq V$ is a convex set and $U \subseteq V$ is a subspace. In general the orthogonal projection of C onto the subspace U contains the intersection of C with the subspace U. Furthermore, the containment is usually strict (see the top of Figure 2.10). In the special case where C is invariant under the action of a group and V^ρ is the fixed point subspace of the action, the projection onto V^ρ and the intersection with V^ρ are actually the same (see the bottom of Figure 2.10). We prove this in part 1 of Lemma 2.6.9 to follow. This allows us to deduce a useful description of the polar of the intersection of C and the fixed point subspace (part 2 of Lemma 2.6.9).

Lemma 2.6.9. *Suppose G is a finite group and (V, ρ) is an orthogonal representation of G. Let $C \subseteq V$ be a convex set such that $\rho(g)C = C$ for all $g \in G$. Then*

1. *$V^\rho \cap C = P_{V^\rho} C$ and*
2. *$[\Pi_{V^\rho}(V^\rho \cap C)]^\circ = \Pi_{V^\rho}(V^\rho \cap C^\circ)$.*

Proof. To see that $V^\rho \cap C \subseteq P_{V^\rho} C$ simply note that if $x \in V^\rho \cap C$ then $P_{V^\rho} x = x$ so $x \in P_{V^\rho} C$. It is the reverse inclusion that uses our assumptions on C and properties of V^ρ. Indeed let $x \in P_{V^\rho} C$. Hence there is $y \in C$ such that

$$x = P_{V^\rho} y = \frac{1}{|G|} \sum_{g \in G} \rho(g) y.$$

Since $\rho(g)C = C$ for all $g \in G$ we have that $\rho(g)y \in C$ for all $g \in G$. Since C is convex it follows that $x \in C$. Furthermore, by construction $x = P_{V^\rho} y \in V^\rho$. Hence $x \in V^\rho \cap C$ establishing the reverse inclusion and the first part of the statement of the Lemma.

We now prove the second part of the statement. We have already seen from the first part that

$$[\Pi_{V^\rho}(V^\rho \cap C)]^\circ = [\Pi_{V^\rho} P_{V^\rho} C]^\circ = [\Pi_{V^\rho} C]^\circ.$$

Applying Lemma 2.3.10 that characterizes the polar of the image of a convex set under a linear map, we see that

$$[\Pi_{V^\rho} C]^\circ = \{y : \Pi_{V^\rho}^*(y) \in C^\circ\} = \Pi_{V^\rho}(V^\rho \cap C^\circ),$$

completing the proof. □

■ 2.6.3 Equivariant semidefinite representations

Suppose (V, ρ) is a representation of a group G, and $C \subseteq V$ is a convex set that is invariant under the group action. Recall that a semidefinite representation of C is a description of the form $C = \pi(\mathcal{S}_+^m \cap L)$ where $\pi : \mathcal{S}^m \to V$ is a linear map, and $L \subseteq \mathcal{S}^m$ is an affine subspace. If C has symmetry, it is natural to study semidefinite representations of C that respect this symmetry. The idea is that for every linear transformation that preserves C, there should be a corresponding linear transformation of the space of symmetric matrices that preserves $\mathcal{S}_+^m$ and L. This is made precise in the following definition [43].

Definition 2.6.10. Let (V, λ) be a representation of a group G, and let $C \subseteq V$ be a convex set such that $\lambda(g)C = C$ for all $g \in G$. A semidefinite representation $C = \pi(\mathcal{S}_+^d \cap L)$ of C is λ-equivariant if there is a representation $(\rho, \mathbb{R}^m)$ such that

1. $\rho(g)L\rho(g)^T = L$ for all $g \in G$
2. $\pi(\rho(g)X\rho(g)^T) = \lambda(g)\pi(X)$ for all $X \in \mathcal{S}^m$ and all $g \in G$.

It is common just to assume that G acts on V without explicitly naming the representation λ. In this case we use the term G-equivariant rather then λ-equivariant.

One major motivation for studying equivariant semidefinite representations comes from attempts to prove lower bounds on the size of semidefinite representations of convex sets with symmetry. As we discussed at the end of Section 2.4, very few techniques are currently available to establish such lower bounds, even for polytopes. In the absence of good tools for establishing such lower bounds, one way to make partial progress is to seek lower bounds that holds under additional restrictions on the semidefinite representations allowed. For convex sets with symmetry, restricting attention to equivariant semidefinite lifts is a natural choice. Doing so means that basic tools from representation theory can be applied to the problem. See, for instance, [43] which studies the structure of equivariant semidefinite representations of a class of symmetric convex bodies known as orbitopes (see Chapter 4).

Chapter 3

Polynomial-sized Semidefinite Representations of Derivative Relaxations of Spectrahedral Cones

3.1 Introduction

In Section 2.5 of Chapter 2 we described a construction due to Gårding that associates convex cones with certain multivariate polynomials, called hyperbolic polynomials. The associated cones, called hyperbolicity cones, include the non-negative orthant, the second-order cone, and the positive semidefinite cone as special cases. A hyperbolic optimization problem is an optimization problem that involves maximizing (or minimizing) a linear functional over an affine slice of a hyperbolicity cone (see Section 2.5).

Since the family of hyperbolicity cones contains the positive semidefinite cone, every semidefinite optimization problem is an instance of a hyperbolic optimization problem. Understanding the extent to which hyperbolic optimization is more expressive than semidefinite optimization is an important part of the general program of understanding the power and limitations of different classes of convex optimization problems. One natural approach to these questions is purely geometric. The aim is to understand which hyperbolicity cones have semidefinite representations (see Definition 2.4.5 of Chapter 2), and if so, how large those representations are.

Finding semidefinite representations of hyperbolicity cones has other benefits beyond understanding the relationships between hyperbolic and semidefinite optimization problems. Semidefinite optimization problems enjoy a simple and well-understood duality theory. As such, giving a semidefinite representation of a hyperbolicity cone provides insight into its dual cone. This is useful because, in general, the dual cones of hyperbolicity cones are not hyperbolicity cones [58], and their properties remain poorly-understood. In addition there are many well-developed numerical routines to solve semidefinite optimization problems. As such a semidefinite representation of a hyperbolicity cone also allows us to use these existing numerical routines to solve the corresponding hyperbolic optimization problem. At least in theory, such a transformation is not likely to be a good idea, since it typically increases the theoretical complexity of solving the optimization problem. Nevertheless, in practice semidefinite optimization solvers are fairly well developed, are widely available, and are easy to use. This is not yet the case for numerical solvers for hyperbolic optimization problems, although this may change in the light of recent algorithmic developments [103, 83].

There are a number of algebraic operations that allow us to construct new hyperbolic polynomials from existing hyperbolic polynomials. One of the most important is the operation of taking directional derivatives (in directions that are in the corresponding hyperbolicity cone). The fact that directional derivatives preserve hyperbolicity is central to a number of recent applications of hyperbolic polynomials, most notably the affirmative resolution of the Kadison-Singer problem by Marcus, Spielman, and Srivastava [81]. Geometrically, the hyperbolicity cone of such a directional derivative always contains the hyperbolicity cone corresponding to the original polynomial (see Equation (3.1.3) in Section 3.1.2). As such, the hyperbolicity cones obtained by taking directional derivatives are often called 'derivative relaxations' or 'Renegar derivatives' (after J. Renegar, who studied this construction, for instance, in [102]). One particularly noteworthy feature of this construction is that derivative relaxations preserve all the (sufficiently) low-dimensional faces of the original hyperbolicity cone (see, e.g., [102]).

In this chapter we study these derivative relaxations of the non-negative orthant and the positive semidefinite cone from the point of view of semidefinite representability. (We give precise definitions of these families of convex cones in Section 3.1.2 to follow.) Our main contribution is to construct explicit, polynomial-sized semidefinite representations of these convex cones.

The rest of this chapter is organized as follows. In Sections 3.1.1 and 3.1.2 we summarize basic facts about hyperbolic polynomials, hyperbolicity cones, and derivative relaxations. Section 3.1.3 summarizes prior work directly related to the problem considered in this chapter. In Section 3.2 we state our constructions, explaining their recursive structure and establishing their size. Section 3.3 gives the proofs of the key steps in our constructions. Section 3.4 proves the correctness of our semidefinite descriptions of the corresponding dual cones. We conclude by describing, in Section 3.5, some related problems for future work.

■ 3.1.1 Hyperbolic polynomials and hyperbolicity cones

We begin by recalling (from Section 2.5) the definition of hyperbolic polynomials and their hyperbolicity cones. A homogeneous polynomial p of degree m in n variables is *hyperbolic* with respect to $e \in \mathbb{R}^n$ if $p(e) \neq 0$ and if for all $x \in \mathbb{R}^n$ the univariate polynomial $t \mapsto p(x-te)$ has only real roots. Gårding's foundational work on hyperbolic polynomials [49] establishes that if p is hyperbolic with respect to e then the connected component of $\{x \in \mathbb{R}^n : p(x) \neq 0\}$ containing e is an open convex cone. This cone is called the *hyperbolicity cone* corresponding to (p, e). We denote it by $\Lambda_{++}(p, e)$, and its closure by $\Lambda_+(p, e)$. Note that p is hyperbolic with respect to e if and only if $-p$ is hyperbolic with respect to e. As such we assume throughout that $p(e) > 0$.

In this chapter we use an equivalent description of the hyperbolicity cone $\Lambda_+(p, e)$, which we now state. Since p is a polynomial we can expand $p(x + te)$ as

$$p(x + te) = p(e)\left[t^m + a_1(x)t^{m-1} + a_2(x)t^{m-2} + \cdots + a_{m-1}(x)t + a_m(x)\right]$$

where the $a_i(x)$ are polynomials that are homogeneous of degree i. The hyperbolicity cone $\Lambda_+(p, e)$ can be expressed in terms of inequalities on the polynomials $a_i(x)$

$$\Lambda_+(p, e) = \{x \in \mathbb{R}^n : a_1(x) \geq 0, \;\; a_2(x) \geq 0, \;\; \ldots, \;\; a_m(x) \geq 0\}, \tag{3.1.1}$$

a description due to Renegar [102, Theorem 20]. We use this description of $\Lambda_+(p, e)$ throughout the chapter.

Basic examples: We revisit the basic examples of the non-negative orthant and the positive semidefinite cone. We saw that these are hyperbolicity cones in Section 2.5.2. We now use (3.1.1) to describe these cones in terms of elementary symmetric polynomials.

- The polynomial $p(x_1, x_2, \ldots, x_n) = x_1 x_2 \cdots x_n$ is hyperbolic with respect to $e = \mathbf{1}_n := (1, 1, \ldots, 1)$. The associated closed hyperbolicity cone is the non-negative orthant, $\mathbb{R}^n_+$. Since

  $$p(x + t\mathbf{1}_n) = t^n + e_1(x)t^{n-1} + \cdots + e_{n-1}(x)t + e_n(x)$$

  where
  $$e_k(x) = \sum_{1 \leq i_1 < \cdots < i_k \leq n} x_{i_1} \cdots x_{i_k} \tag{3.1.2}$$

  is the elementary symmetric polynomial of degree k in the variables $x_1, x_2, \ldots, x_n$, we have that

  $$\Lambda_+(p, e) = \mathbb{R}^n_+ = \{x \in \mathbb{R}^n : e_1(x) \geq 0, \;\; e_2(x) \geq 0, \;\; \ldots, \;\; e_n(x) \geq 0\}.$$

- Let X be an $n \times n$ symmetric matrix of indeterminates. The polynomial $p(X) = \det(X)$ is hyperbolic with respect to $e = I_n$, the $n \times n$ identity matrix. The associated closed hyperbolicity cone is the positive semidefinite cone, $\mathcal{S}^n_+$. Since

  $$p(X + tI_n) = t^n + E_1(X)t^{n-1} + \cdots + E_{n-1}(X)t + E_n(X)$$

where the $E_k(X)$ are the coefficients of the characteristic polynomial of X, we have that

$$\Lambda_+(p,e) = \mathcal{S}_+^n = \{X : E_1(X) \geq 0,\;\; E_2(X) \geq 0,\;\; \ldots,\;\; E_n(X) \geq 0\}.$$

Observe that $E_k(X) := e_k(\lambda(X))$ is the elementary symmetric polynomial of degree k in the eigenvalues of X so the positive semidefinite cone can also be described in terms of polynomial inequalities on the eigenvalues of X as

$$\mathcal{S}_+^n = \{X : e_1(\lambda(X)) \geq 0,\;\; e_2(\lambda(X)) \geq 0,\;\; \ldots,\;\; e_n(\lambda(X)) \geq 0\}.$$

■ 3.1.2 Derivative relaxations

If p is hyperbolic with respect to e then (essentially by Rolle's theorem [106]) the directional derivative of p in the direction e, *viz.*

$$p_e^{(1)}(x) := \left.\frac{d}{dt}p(x+te)\right|_{t=0}$$

is also hyperbolic with respect to e, a construction that goes back to Gårding [49]. If p has degree m, by repeatedly differentiating in the direction e we construct a sequence of polynomials $p, p_e^{(1)}, p_e^{(2)}, \ldots, p_e^{(m-1)}$ each hyperbolic with respect to e.

The corresponding hyperbolicity cones can be expressed nicely in terms of polynomial inequalities. Indeed if $p(x+te) = p(e)\left[t^m + \sum_{i=1}^m a_i(x)t^{m-i}\right]$ then differentiating k times with respect to t we see that

$$p_e^{(k)}(x+te) = p(e)\left[c_0 a_{m-k}(x) + c_1 a_{m-k-1}(x)t + \cdots + c_{m-k}t^{m-k}\right]$$

where $c_i = (k+i)!/i! > 0$. By (3.1.1) the corresponding hyperbolicity cone is

$$\Lambda_+^{(k)}(p,e) := \Lambda_+(p_e^{(k)}, e) = \{x \in \mathbb{R}^n : a_1(x) \geq 0,\;\; a_2(x) \geq 0,\;\; \ldots,\;\; a_{m-k}(x) \geq 0\}$$

and can be obtained from (3.1.1) by removing k of the inequality constraints. As a result, the hyperbolicity cones $\Lambda_+^{(k)}(p,e)$ provide a sequence of outer approximations to the original hyperbolicity cone that satisfy

$$\Lambda_+(p,e) \subset \Lambda_+^{(1)}(p,e) \subset \cdots \subset \Lambda_+^{(m-1)}(p,e). \tag{3.1.3}$$

The last of these, $\Lambda_+^{(m-1)}(p,e)$, is simply the closed half-space defined by e. The work of Renegar [102] highlights the many nice properties of this sequence of approximations.

Note that we abuse terminology by referring to the cones $\Lambda_+^{(k)}(p,e)$ as *derivative relaxations* of the hyperbolicity cone $\Lambda_+(p,e)$. The abuse is that $\Lambda_+^{(k)}(p,e)$ does not depend only on the *geometric* object $\Lambda_+(p,e)$ but on its particular *algebraic* description via p and e.

Examples:

- In the case of $p(x) = x_1x_2\cdots x_n = e_n(x)$ and $e = \mathbf{1}_n$, we have that $p_e^{(k)}(x) = k!e_{n-k}(x)$. Consequently the kth derivative relaxation of the orthant, which we denote by $\mathbb{R}_+^{n,(k)}$, is the hyperbolicity cone $\Lambda_+(e_{n-k},\mathbf{1}_n)$. It can be expressed as

$$\mathbb{R}_+^{n,(k)} = \{x\in\mathbb{R}^n : e_1(x)\geq 0,\;\; e_2(x)\geq 0,\;\; \ldots,\;\; e_{n-k}(x)\geq 0\}. \tag{3.1.4}$$

  Consistent with these descriptions we define $\mathbb{R}_+^{n,(n)} := \mathbb{R}^n$. Note, also, that $\mathbb{R}_+^{n,(0)} = \mathbb{R}_+^n$.

- In the case of $p(X) = \det(X) = E_n(X)$ and $e = I_n$, we have that $p_e^{(k)}(x) = k!E_{n-k}(X)$. The kth derivative relaxation of the positive semidefinite cone, which we denote by $\mathcal{S}_+^{n,(k)}$, can be described as

$$\begin{aligned}\mathcal{S}_+^{n,(k)} &= \{X\in\mathcal{S}^n : E_1(x)\geq 0,\;\; E_2(x)\geq 0,\;\; \ldots,\;\; E_{n-k}(x)\geq 0\} &(3.1.5)\\ &= \{X\in\mathcal{S}^n : e_1(\lambda(X))\geq 0,\;\; e_2(\lambda(X))\geq 0,\;\; \ldots,\;\; e_{n-k}(\lambda(X))\geq 0\}\,. &(3.1.6)\end{aligned}$$

  Again we define $\mathcal{S}_+^{n,(n)} := \mathcal{S}^n$, the set of $n\times n$ symmetric matrices and note that $\mathcal{S}_+^{n,(0)} = \mathcal{S}_+^n$. Since $E_i(\mathrm{diag}(x)) = e_i(x)$ for all i, the diagonal slice of $\mathcal{S}_+^{n,(k)}$ is exactly $\mathbb{R}_+^{n,(k)}$.

Symmetry: Suppose G is a group acting by linear transformations on $\mathbb{R}^n$ by $x\mapsto g\cdot x$ for all $g\in G$. Suppose *both* p *and* e are invariant under the group action, i.e., $g\cdot e = e$ and $(g\cdot p)(x) := p(g^{-1}\cdot x) = p(x)$ for all $g\in G$. Then for all $t\in\mathbb{R}$, $x\in\mathbb{R}^n$ and $g\in G$

$$p(x+te) = (g\cdot p)(x+te) = p(g^{-1}\cdot(x+te)) = p((g^{-1}\cdot x)+te).$$

Hence the hyperbolicity cone $\Lambda_+(p,e)$ and all of its derivative cones $\Lambda_+^{(k)}(p,e)$ are invariant under this same group action.

For our purposes an important example of this is the symmetry of the cones $\mathcal{S}_+^{n,(k)}$. The action of $O(n)$ by conjugation on symmetric matrices leaves the polynomial $p(X) = \det(X)$ invariant *and* preserves the direction $e = I_n$. Hence all of the derivative relaxations of the positive semidefinite cone are invariant under conjugation by orthogonal matrices. As such, the cones $\mathcal{S}_+^{n,(k)}$ are *spectral sets*, in the sense that whether a symmetric matrix X belongs to $\mathcal{S}_+^{n,(k)}$ depends only on the eigenvalues of X. This is evident from the description of $\mathcal{S}_+^{n,(k)}$ in (3.1.6).

3.1.3 Related work

Previous work has focused on semidefinite and spectrahedral representations (see Section 2.4) of the derivative relaxations of the orthant. Zinchenko [140] used a decomposition approach to give semidefinite representations of $\mathbb{R}_+^{n,(1)}$ and its dual cone. Sanyal [111] subsequently gave spectrahedral representations of $\mathbb{R}_+^{n,(1)}$ and $\mathbb{R}_+^{n,(n-2)}$ and conjectured that all of the derivative relaxations of the orthant admit spectrahedral representations.

Recently Brändén [17] settled this conjecture in the affirmative giving spectrahedral representations of $\mathbb{R}_+^{n,(n-k)}$ for $k = 1, 2, \ldots, n-1$ of size $O(n^{k-1})$. For each $1 \leq k < n$ Brändén constructs a graph $G_{n,k} = (V, E)$ together with edge weights $(w_e(x))_{e \in E}$ that are linear forms in x so that

$$\mathbb{R}_+^{n,(n-k)} = \left\{x \in \mathbb{R}^n : L_{G_{n,k}}(x) \succeq 0\right\} \tag{3.1.7}$$

where $L_{G_{n,k}}(x)$ is the $|V| \times |V|$ edge-weighted Laplacian of $G_{n,k}$. Since $L_{G_{n,k}}(x)$ is linear in the edge weights, and the edge weights are linear forms in x, (3.1.7) is a spectrahedral representation of size $|V|$. With the exception of two distinguished vertices, the vertices of $G_{n,k}$ are indexed by all ℓ-tuples (for $1 \leq \ell \leq k-1$) consisting of distinct elements of $\{1, 2, \ldots, n\}$. Hence $|V| = 2 + \sum_{\ell=1}^{k-1} \ell!\binom{n}{\ell}$ showing that Brändén's spectrahedral representation of $\mathbb{R}_+^{n,(n-k)}$ has size $O(n^{k-1})$. While Brändén's construction is of considerable theoretical interest, these representations (unlike ours) are not practical for optimization due to their prohibitive size.

A spectrahedral representation of $\mathbb{R}_+^{n,(1)}$ is implicit in the work of Choe et al. [138] that studies the relationships between matroids[1] and hyperbolic polynomials. Choe et al. observe that if $\mathcal{M}$ is a regular matroid[2] represented by the rows of a totally unimodular[3] matrix V then $\det(V^T \mathrm{diag}(x) V)$ is the basis generating polynomial of $\mathcal{M}$. In particular, the uniform matroid U_n^{n-1} is regular and has $e_{n-1}(x)$ as its basis

[1] A *matroid* is a finite set S together with a collection $\mathcal{I}$ of subsets of S satisfying 1. if $I \in \mathcal{I}$ and $J \subseteq I$ then $J \in \mathcal{I}$ and 2. if $I, J \in \mathcal{I}$ and $|I| < |J|$ then there is $z \in J \setminus I$ such that $I \cup \{z\} \in \mathcal{I}$ [121, Section 39.1]. A matroid $(S, \mathcal{I})$ is *representable over a field* $\mathbb{K}$ if there is a matrix V with entries in $\mathbb{K}$ and columns indexed by S such that $I \in \mathcal{I}$ if and only if the columns of V indexed by I are linearly independent over $\mathbb{K}$.

[2] A *regular matroid* is a matroid that is representable over every field, and can always be represented by the columns of a totally unimodular matrix [121, Section 39.4].

[3] A matrix is *totally unimodular* if all of its square submatrices have determinant $-1, 0$, or 1 [121, Section 5.16].

generating polynomial, yielding a symmetric determinantal representation of $e_{n-1}(x)$ and hence a spectrahedral representation of $\mathbb{R}_+^{n,(n-1)}$.

From a computational perspective, Güler [58] showed that if p has degree m and is hyperbolic with respect to e then $-\log p$ is a self-concordant barrier function (with barrier parameter m) for the hyperbolicity cone $\Lambda_+(p,e)$. As such, as long as p and its gradient and Hessian can be computed efficiently, one can use interior point methods to minimize a linear functional over an affine slice of $\Lambda_+(p,e)$ efficiently. Renegar [102, Section 9] gave an efficient interpolation-based method for computing $p_e^{(k)}$ (and its gradient and Hessian) whenever p (and its gradient and Hessian) can be evaluated efficiently. Güler and Renegar's observations together yield efficient computational methods to optimize a linear functional over an affine slice of a derivative relaxation of a spectrahedral cone. Our results complement these, giving a method to solve optimization problems of this type using existing numerical procedures for semidefinite programming.

■ 3.2 Results

Our main contribution is to construct two different explicit polynomial-sized semidefinite representations of the derivative relaxations of the positive semidefinite cone. We call our two representations the *derivative-based* and *polar derivative-based* representations respectively. In this section we describe these representations, and outline the proof of our main theoretical result.

Theorem 3.2.1. *For each positive integer n and each $k = 1, 2, \ldots, n-1$, the cone $\mathcal{S}_+^{n,(k)}$ has a semidefinite representation of size $O(\min\{k, n-k\}n^2)$.*

We defer detailed proofs of the correctness of our representations to Sections 3.3 and 3.4. At this stage, we just highlight that there is essentially one basic algebraic fact that underlies all of our results. Whenever V_n is an $n \times (n-1)$ matrix with orthonormal columns that are each orthogonal to $\mathbf{1}_n$, i.e. $V_n^T V_n = I_{n-1}$ and $V_n^T \mathbf{1}_n = 0$, then

$$e_{n-1}(x) = n \det(V_n^T \operatorname{diag}(x) V_n).$$

We give a proof of this identity in Section 3.3. Note that this identity is independent of the particular choice of V_n satisfying $V_n^T V_n = I_{n-1}$ and $V_n^T \mathbf{1}_n = 0$. In fact, all of the results expressed in terms of V_n (notably Propositions 3.2.3, 3.2.4, 3.2.3D, and 3.2.4D) are similarly independent of the particular choice of V_n.

Both of the representations are recursive in nature. The derivative-based representation is based on recursively applying two basic propositions (Propositions 3.2.2 and 3.2.3, to follow) to construct a chain of semidefinite representations of the form

$$\boxed{\mathcal{S}_+^{n,(k)}} \xleftarrow[\text{Prop. 3.2.2}]{O(n^2)} \mathbb{R}_+^{n,(k)} \xleftarrow[\text{Prop. 3.2.3}]{0} \boxed{\mathcal{S}_+^{n-1,(k-1)}} \xleftarrow[\text{Prop. 3.2.2}]{O((n-1)^2)} \mathbb{R}_+^{n-1,(k-1)} \leftarrow \cdots \qquad (3.2.1)$$
$$\cdots \leftarrow \mathbb{R}_+^{n-k+1,(1)} \xleftarrow[\text{Prop. 3.2.3}]{0} \boxed{\mathcal{S}_+^{n-k,(0)}.}$$

The annotated arrow $C \xleftarrow[\text{Prop. } a]{m} K$ indicates that given a semidefinite representation of K of size m' we can construct a semidefinite representation of C of size $m' + m$, and that an explicit description of the construction is given in Proposition a.

The base case of the recursion is just the positive semidefinite cone $\mathcal{S}_+^{n-k,(0)} = \mathcal{S}_+^{n-k}$, which has a trivial semidefinite representation. Hence starting from $\mathcal{S}_+^{n-k,(0)}$ (which has a semidefinite representation of size $n - k$), we can apply Proposition 3.2.3 to obtain a semidefinite representation of $\mathbb{R}_+^{n-k+1,(1)}$ of size $n - k$, then apply Proposition 3.2.2 to obtain a semidefinite representation of $\mathcal{S}_+^{n-k+1,(1)}$ of size $(n - k) + O((n - k + 1)^2)$, and so on.

The polar derivative-based representation is based on recursively applying Proposition 3.2.2 together with a third basic proposition (Proposition 3.2.4, to follow) to construct a slightly different chain of semidefinite representations of the form

$$\boxed{\mathcal{S}_+^{n,(k)}} \xleftarrow[\text{Prop. 3.2.2}]{O(n^2)} \mathbb{R}_+^{n,(k)} \xleftarrow[\text{Prop. 3.2.4}]{n} \boxed{\mathcal{S}_+^{n-1,(k)}} \xleftarrow[\text{Prop. 3.2.2}]{O(n^2)} \mathbb{R}_+^{n-1,(k)} \leftarrow \cdots \qquad (3.2.2)$$
$$\cdots \leftarrow \mathbb{R}_+^{k+2,(k)} \xleftarrow[\text{Prop. 3.2.4}]{n} \boxed{\mathcal{S}_+^{k+1,(k)}.}$$

Note that the base case of the recursion is just $\mathcal{S}_+^{k+1,(k)} = \{X \in \mathcal{S}^{k+1} : \operatorname{tr}(X) \geq 0\}$, a half-space.

■ 3.2.1 Building blocks of the two recursions

We now describe the constructions related to each of the types of arrows in the recursions sketched above. The arrows labeled by Proposition 3.2.2 assert that we can construct a semidefinite representation of $\mathcal{S}_+^{n,(k)}$ from a semidefinite representation of $\mathbb{R}_+^{n,(k)}$. This can be done in the following way.

Proposition 3.2.2. *If $\mathbb{R}_+^{n,(k)}$ has a semidefinite representation of size m, then $\mathcal{S}_+^{n,(k)}$ has a semidefinite representation of size $m + O(n^2)$. Indeed*

$$\mathcal{S}_+^{n,(k)} = \left\{X \in \mathcal{S}^n : \exists z \in \mathbb{R}^n \;\; s.t. \;\; z \in \mathbb{R}_+^{n,(k)}, \;\; (X, z) \in \mathrm{SH}_n\right\}, \qquad (3.2.3)$$

where SH_n *is the* Schur-Horn cone *defined as*

$$\mathrm{SH}_n = \left\{(X,z): \quad z_1 \geq z_2 \geq \cdots \geq z_n, \quad X \in \mathrm{conv}_{Q\in O(n)}\{Q^T\mathrm{diag}(z)Q\}\right\}$$

i.e. the set of pairs (X,z) *such that* X *is in the convex hull of all symmetric matrices with ordered spectrum* z. *The Schur-Horn cone has the semidefinite characterization*

$$\begin{aligned} (X,z)\in\mathrm{SH}_n \quad &\textit{if and only if} \quad && z_1 \geq z_2 \geq \cdots \geq z_n \;\; \textit{and}\\ &\textit{there exist} && t_2,\ldots,t_{n-1}\in\mathbb{R},\;\; Z_2,\ldots,Z_{n-1}\succeq 0\\ &\textit{such that} && \mathrm{tr}(X) = \textstyle\sum_{j=1}^n z_j,\;\; X \preceq z_1 I,\;\; \textit{and}\\ \textit{for } \ell = 2,\ldots,n-1, & && X \preceq t_\ell I + Z_\ell \;\; \textit{and} \;\; \ell\cdot t_\ell + \mathrm{tr}(Z_\ell) \leq \textstyle\sum_{j=1}^{\ell} z_j. \end{aligned}$$

Proposition 3.2.2 holds because of the *symmetry* of $\mathcal{S}_+^{n,(k)}$. In particular it is a spectral set—invariant under conjugation by orthogonal matrices. The other reason this representation works is that the diagonal slice of $\mathcal{S}_+^{n,(k)}$ is $\mathbb{R}_+^{n,(k)}$. We discuss this result in more detail in Section 3.4.

The arrows in (3.2.1) labeled by Proposition 3.2.3 appear only in the derivative-based recursion. They assert that we can obtain a semidefinite representation of $\mathbb{R}_+^{n,(k)}$ from a semidefinite representation of $\mathcal{S}_+^{n-1,(k-1)}$. Indeed we establish in Section 3.3.1 that $\mathbb{R}_+^{n,(k)}$ is actually a slice of $\mathcal{S}_+^{n-1,(k-1)}$.

Proposition 3.2.3. *If* $1 \leq k \leq n-1$ *then*

$$\mathbb{R}_+^{n,(k)} = \left\{x\in\mathbb{R}^n : V_n^T\mathrm{diag}(x)V_n \in \mathcal{S}_+^{n-1,(k-1)}\right\}.$$

The arrows in (3.2.2) labeled by Proposition 3.2.4 appear only in the polar derivative-based recursion. They assert that we can obtain a semidefinite representation of $\mathbb{R}_+^{n,(k)}$ from a semidefinite representation of $\mathcal{S}_+^{n-1,(k)}$. We establish the following in Section 3.3.2.

Proposition 3.2.4. *If* $1 \leq k \leq n-2$ *then*

$$\mathbb{R}_+^{n,(k)} = \left\{x\in\mathbb{R}^n : \exists Z \in \mathcal{S}_+^{n-1,(k)} \;\; \textit{s.t.}\;\; \mathrm{diag}(x) \succeq V_n Z V_n^T\right\}.$$

■ 3.2.2 Size of the representations

Recall that each arrow $C \xleftarrow{m} K$ in (3.2.1) and (3.2.2) is labeled with the *additional size* m required to implement the representation of C given a semidefinite representation of K. Since the derivative-based recursion has $2k$ arrows, it is immediate from (3.2.1) that the derivative-based semidefinite representation of $\mathcal{S}_+^{n,(k)}$ has size $O(kn^2)$ and so is of polynomial size.

On the other hand, this approach gives a disappointingly large semidefinite representation of the half-space $\mathcal{S}_+^{n,(n-1)} = \{X \in \mathcal{S}^n : \mathrm{tr}(X) \geq 0\}$ of size $O(n^3)$. The derivative-based approach cannot exploit the fact that this is a very simple cone. This is why we also consider the polar derivative-based representation, as it is designed around the fact that $\mathcal{S}_+^{n,(n-1)}$ has a simple semidefinite representation.

It is immediate from (3.2.2) that the polar derivative-based semidefinite representation of $\mathcal{S}_+^{n,(k)}$ has size $O((n-k)n^2)$ and so is also of polynomial size. Furthermore, it gives small representations of size $O(n^2)$ exactly when the derivative-based representations are large, of size $O(n^3)$. For any given pair (n,k) we should always use the derivative-based representation of $\mathcal{S}_+^{n,(k)}$ if $k < n/2$ and the polar derivative-based representation when $k > n/2$. Theorem 3.2.1 combines our two size estimates, stating that $\mathcal{S}_+^{n,(k)}$ has a semidefinite representation of size $O(\min\{k, n-k\}n^2)$.

■ 3.2.3 Pseudocode for our derivative-based representation

We do not write out any of our semidefinite representations in full because the recursive descriptions given here are actually more naturally suited to implementation. To illustrate this, we give pseudocode for the MATLAB-based high-level modeling language YALMIP [78] that ‘implements’ the derivative-based representations of $\mathcal{S}_+^{n,(k)}$ and $\mathbb{R}_+^{n,(k)}$. Decision variables are declared by expressions like `x = sdpvar(n,1);` which creates a decision variable `x` taking values in $\mathbb{R}^n$. A linear matrix inequality (LMI) object is a list of equality constraints and positive semidefinite matrix inequality constraints that are linear in any declared decision variables.

Suppose we have a function `SH(X,z)` that takes a pair of decision variables and returns an LMI object corresponding to the constraint that $(X,z) \in \mathrm{SH}_n$. This is easy to construct from the explicit semidefinite representation in Proposition 3.2.2. Then the function `psdcone` takes an $n \times n$ symmetric matrix-valued decision variable `X` and returns an LMI object for the constraint $X \in \mathcal{S}_+^{n,(k)}$.

```
function K = psdcone(X,k)
    if k==0
        K = [X >= 0];
    else
        z = sdpvar(size(X,1),1);
        K = [orthant(z,k), SH(X,z)];
    end
```

It calls a function `orthant` that takes a decision variable `x` in $\mathbb{R}^n$ and returns an LMI object for the constraint $x \in \mathbb{R}_+^{n,(k)}$.

```
1:  function K = orthant(x,k)
2:      if k==0
3:          K = [x >= 0];
4:      else
5:          V = null(ones(size(x))');
6:          K = [psdcone(V'*diag(x)*V,k-1)];
7:      end
```

It is straightforward to adapt these two functions for the polar derivative-based representation, one needs only to change the base cases (lines 2–4 of each) and to adapt line 6 of `orthant` to reflect Proposition 3.2.4.

3.2.4 Dual cones

If a cone is semidefinitely representable, so is its dual cone. In fact there are explicit procedures to take a semidefinite representation for a cone and produce a semidefinite representation for its dual cone [85, Section 4.1.1]. Here we describe two explicit semidefinite representations of the dual cones $(\mathcal{S}_+^{n,(k)})^*$ that enjoy the same recursive structure as the corresponding semidefinite representations of $\mathcal{S}_+^{n,(k)}$.

To construct them, we essentially dualize all the relationships given by the arrows in (3.2.1) and (3.2.2). By straightforward applications of a conic duality argument, in Section 3.3.3 we establish the following dual analogues of Propositions 3.2.3 and 3.2.4.

Proposition 3.2.3D. *If* $1 \leq k \leq n-1$ *then*

$$(\mathbb{R}_+^{n,(k)})^* = \left\{\operatorname{diag}(V_n Y V_n^T) : \; Y \in (\mathcal{S}_+^{n-1,(k-1)})^*\right\}.$$

Proposition 3.2.4D. *If* $1 \leq k \leq n-2$ *then*

$$(\mathbb{R}_+^{n,(k)})^* = \left\{\operatorname{diag}(Y) : \; Y \succeq 0, \; V_n^T Y V_n \in (\mathcal{S}_+^{n-1,(k)})^*\right\}.$$

We could also obtain a dual version of Proposition 3.2.2 by directly applying conic duality to the semidefinite representation in Proposition 3.2.2. This would involve dualizing the semidefinite representation of SH_n. Instead we give another, perhaps simpler, representation of $(\mathcal{S}_+^{n,(k)})^*$ in terms of $(\mathbb{R}_+^{n,(k)})^*$ that is not obtained by directly applying conic duality to Proposition 3.2.2.

Proposition 3.2.2D. *If $(\mathbb{R}_+^{n,(k)})^*$ has a semidefinite representation of size m, then $(\mathcal{S}_+^{n,(k)})^*$ has a semidefinite representation of size $m + O(n^2)$ given by*

$$(\mathcal{S}_+^{n,(k)})^* = \left\{ W \in \mathcal{S}^n : \exists y \in \mathbb{R}^n \;\; s.t. \;\; y \in (\mathbb{R}_+^{n,(k)})^*, \;\; (W, y) \in \mathrm{SH}_n \right\}. \tag{3.2.4}$$

Recall that Proposition 3.2.2 holds because $\mathcal{S}_+^{n,(k)}$ is invariant under orthogonal conjugation and $\mathbb{R}_+^{n,(k)}$ is the diagonal slice of $\mathcal{S}_+^{n,(k)}$. While it is immediate that $(\mathcal{S}_+^{n,(k)})^*$ is also orthogonally invariant, it is a less obvious result that the diagonal slice of $(\mathcal{S}_+^{n,(k)})^*$ is $(\mathbb{R}_+^{n,(k)})^*$. We prove this in Section 3.4.

The recursions underlying the derivative-based and polar derivative-based representations of $(\mathcal{S}_+^{n,(k)})^*$ then take the form

$$(\mathcal{S}_+^{n,(k)})^* \leftarrow (\mathbb{R}_+^{n,(k)})^* \leftarrow (\mathcal{S}_+^{n-1,(k-1)})^* \leftarrow \cdots \leftarrow (\mathbb{R}_+^{n-k+1,(1)})^* \leftarrow (\mathcal{S}_+^{n-k,(0)})^* \tag{3.2.5}$$

and, respectively,

$$(\mathcal{S}_+^{n,(k)})^* \leftarrow (\mathbb{R}_+^{n,(k)})^* \leftarrow (\mathcal{S}_+^{n-1,(k)})^* \leftarrow \cdots \leftarrow (\mathbb{R}_+^{k+2,(k)})^* \leftarrow (\mathcal{S}_+^{k+1,(k)})^*. \tag{3.2.6}$$

Note that for the dual derivative-based representation, the base case is $(\mathcal{S}_+^{n-k,(0)})^* = \mathcal{S}_+^{n-k}$ (since the positive semidefinite cone is self dual). For the dual polar derivative-based representation the base case is $(\mathcal{S}_+^{k+1,(k)})^* = \{tI_{k+1} : t \geq 0\}$, the ray generated by the identity matrix in $\mathcal{S}^{k+1}$.

■ 3.2.5 Derivative relaxations of spectrahedral cones

So far we have focused on the derivative relaxations of the positive semidefinite cone. It turns out that the derivative relaxations of spectrahedral cones are just slices of the associated derivative relaxations of the positive semidefinite cone.

Proposition 3.2.5. *Suppose $p(x) = \det(\sum_{i=1}^n A_i x_i)$ where the A_i are $m \times m$ symmetric matrices and $e \in \mathbb{R}^n$ is such that $\sum_{i=1}^n A_i e_i = B$ is positive definite. Then for $k = 0, 1, \ldots, m-1$,*

$$\Lambda_+^{(k)}(p, e) = \left\{ x \in \mathbb{R}^n : \sum_{i=1}^n B^{-1/2} A_i B^{-1/2} x_i \in \mathcal{S}_+^{m,(k)} \right\}.$$

Proof. Let $\mathcal{A}(x) = \sum_{i=1}^n B^{-1/2} A_i B^{-1/2} x_i$. Then $\mathcal{A}(e) = I$ and for all $x \in \mathbb{R}^n$ and all $t \in \mathbb{R}$

$$p(x + te) = \det(B)\det(\mathcal{A}(x + te)) = \det(B)\det(\mathcal{A}(x) + tI).$$

This implies that all the derivatives of p in the direction e are exactly the same as the corresponding derivatives of $\det(B)\det(X)$ in the direction I evaluated at $X = \mathcal{A}(x)$. Since $\det(B) > 0$, it follows that for $k = 0, 1, \ldots, m-1$, $x \in \Lambda_+^{(k)}(p,e)$ if and only if $\mathcal{A}(x) \in \mathcal{S}_+^{m,(k)}$. □

We conclude this section with an example of these constructions.

Example 3.2.6 (Derivative relaxations of a 3-ellipse)**.** Given foci $(0,0)$, $(0,4)$ and $(3,0)$ in the plane, the 3-ellipse consisting of points such that the sum of distances to the foci equals 8 is shown in Figure 3.1. This is one connected component of the real algebraic curve of degree 8 given by $\{(x,y) \in \mathbb{R}^2 : \det \mathcal{E}(x,y,1) = 0\}$ where $\mathcal{E}$ is defined in (3.2.7) (see Nie et al. [91]). The region enclosed by this 3-ellipse is the $z = 1$ slice of the spectrahedral cone defined by $\mathcal{E}(x,y,z) \succeq 0$ where

$$\mathcal{E}(x,y,z) = \begin{bmatrix} 5z+3x & y & y-4z & 0 & y & 0 & 0 & 0 \\ y & 5z+x & 0 & y-4z & 0 & y & 0 & 0 \\ y-4z & 0 & 5z+x & y & 0 & 0 & y & 0 \\ 0 & y-4z & y & 5z-x & 0 & 0 & 0 & y \\ y & 0 & 0 & 0 & 11z+x & y & y-4z & 0 \\ 0 & y & 0 & 0 & y & 11z-x & 0 & y-4z \\ 0 & 0 & y & 0 & y-4z & 0 & 11z-x & y \\ 0 & 0 & 0 & y & 0 & y-4z & y & 11z-3x \end{bmatrix}. \quad (3.2.7)$$

Note that $\mathcal{E}(0,0,1) \succ 0$ and so $e = (0,0,1)$ is a direction of hyperbolicity for $p(x,y,z) = \det \mathcal{E}(x,y,z)$. The left of Figure 3.1 shows the $z = 1$ slice of the cone $\Lambda_+(p,e)$ and its first three derivative relaxations $\Lambda_+^{(1)}(p,e), \Lambda_+^{(2)}(p,e)$, and $\Lambda_+^{(3)}(p,e)$. The right of Figure 3.1 shows the $z = 1$ slice of the cones $(\Lambda_+(p,e))^*, (\Lambda_+^{(1)}(p,e))^*, (\Lambda_+^{(2)}(p,e))^*$, and $(\Lambda_+^{(3)}(p,e))^*$. All of these convex bodies were plotted by computing 200 points on their respective boundaries by optimizing 200 different linear functionals over them. We performed the optimization by modeling our semidefinite representations of these cones in YALMIP [78] which numerically solved the corresponding semidefinite program using SDPT3 [127].

■ 3.3 The derivative-based and polar derivative-based recursive constructions

In this section we prove Proposition 3.2.3 which relates $\mathbb{R}_+^{n,(k)}$ and $\mathcal{S}_+^{n-1,(k-1)}$ as well as Proposition 3.2.4 which relates $\mathbb{R}_+^{n,(k)}$ and $\mathcal{S}_+^{n-1,(k)}$. These relationships are the geometric consequences of polynomial identities between elementary symmetric polynomials and determinants.

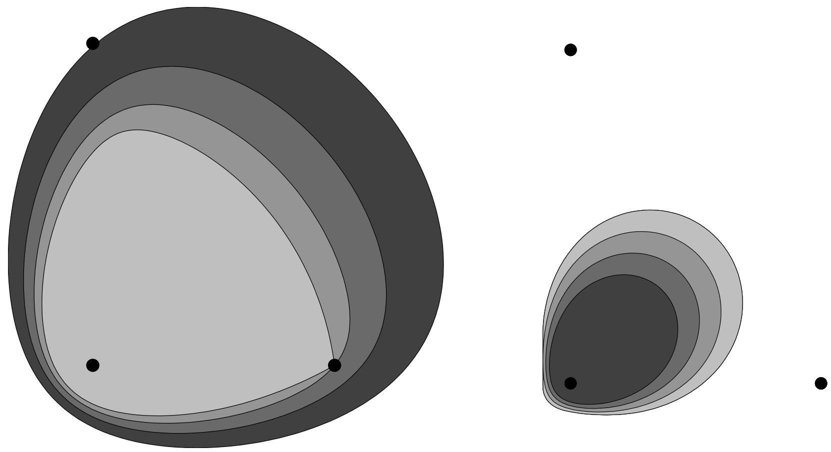


Figure 3.1: On the left, the inner region is the 3-ellipse consisting of points with sum-of-distances to $(0,0), (0,4)$, and $(3,0)$ equal to 8, i.e. the $z = 1$ slice of the spectrahedral cone defined by (3.2.7). The outer three regions are the $z = 1$ slices of the first three derivative relaxations of this spectrahedral cone in the direction $(0,0,1)$. On the right are the $z = 1$ slices of the dual cones of the cones shown on the left, with dual pairs having the same shading.

Specifically the proof of Proposition 3.2.3 makes use of a determinantal representation (Equation (3.3.3) in Section 3.3.1) of the derivative

$$\frac{\partial}{\partial t} e_n(sx + t\mathbf{1}_n)\big|_{s=1} = \left[1 \cdot e_{n-1}(x) + \cdots + (n-1) \cdot e_1(x)t^{n-2} + n \cdot t^{n-1}\right]. \qquad (3.3.1)$$

(Note that s plays no role in (3.3.1), we include it to highlight the relationship with (3.3.2).) Similarly the proof of Proposition 3.2.4 relies on a determinantal expression (Equation (3.3.6) in Section 3.3.2) for the polar derivative

$$\frac{\partial}{\partial s} e_n(sx + t\mathbf{1}_n)\big|_{s=1} = \left[n \cdot e_n(x) + (n-1) \cdot e_{n-1}(x)t + \cdots + 1 \cdot e_1(x)t^{n-1}\right]. \qquad (3.3.2)$$

This explains why we call one the *derivative-based representation*, and the other the *polar derivative-based representation.*

■ 3.3.1 The derivative-based recursion: relating $\mathbb{R}_+^{n,(k)}$ and $\mathcal{S}_+^{n-1,(k-1)}$

Let V_n denote an (arbitrary) $n \times (n-1)$ matrix satisfying $V_n^T V_n = I_{n-1}$ and $V_n^T \mathbf{1}_n = 0$. Our results in this section and the next stem from the following identity.

Lemma 3.3.1. *For all $x \in \mathbb{R}^n$ and all $t \in \mathbb{R}$,*

$$\frac{\partial}{\partial t} e_n(sx + t\mathbf{1}_n)\big|_{s=1} = e_{n-1}(x + t\mathbf{1}_n) = n \det(V_n^T \operatorname{diag}(x) V_n + tI_{n-1}). \qquad (3.3.3)$$

This is a special case of an identity established by Choe et al. [138, Corollary 8.2] and is closely related to Sanyal's result [111, Theorem 1.1]. The proof of Choe et al. uses the Cauchy-Binet identity. Here we provide an alternative proof.

Proof. The polynomial $e_{n-1}(x_1, x_2, \ldots, x_n)$ is characterized by satisfying $e_{n-1}(\mathbf{1}_n) = n$, and by being symmetric, homogeneous of degree $n-1$ and of degree one in each of the x_i. We show, below, that $n\det(V_n^T\operatorname{diag}(x)V_n)$ also has these properties and so that $e_{n-1}(x) = n\det(V_n^T\operatorname{diag}(x)V_n)$. The stated result then follows because $V_n^TV_n = I_{n-1}$ implies

$$e_{n-1}(x + t\mathbf{1}_n) = n\det(V_n^T\operatorname{diag}(x + t\mathbf{1}_n)V_n) = n\det(V_n^T\operatorname{diag}(x)V_n + tI_{n-1}).$$

Now, it is clear that $\det(V_n^T\operatorname{diag}(x)V_n)$ is homogeneous of degree $n-1$ and that

$$n\det(V_n^T\operatorname{diag}(\mathbf{1}_n)V_n) = n\det(I_{n-1}) = n.$$

It remains to establish that $\det(V_n^T\operatorname{diag}(x)V_n)$ is symmetric and of degree one in each of the x_i. To do so we repeatedly use the fact that if V_n and U_n both have orthonormal columns that span the orthogonal complement of $\mathbf{1}_n$ then $\det(V_n^T\operatorname{diag}(x)V_n) = \det(U_n^T\operatorname{diag}(x)U_n)$.

The polynomial $\det(V_n^T\operatorname{diag}(x)V_n)$ is symmetric because for any $n \times n$ permutation matrix P the columns of V_n and PV_n respectively are both orthonormal and each spans the orthogonal complement of $\mathbf{1}_n$ (because $P\mathbf{1}_n = \mathbf{1}_n$). Hence

$$\det(V_n^T\operatorname{diag}(Px)V_n) = \det((PV_n)^T\operatorname{diag}(x)(PV_n)) = \det(V_n^T\operatorname{diag}(x)V_n).$$

We finally show that $\det(V_n^T\operatorname{diag}(x)V_n)$ is of degree one in each x_i by a convenient choice of V_n. For any i, we can always choose V_n to be of the form

$$V_n^T = \begin{bmatrix} v_1 & \cdots v_{i-1} & \sqrt{\tfrac{n-1}{n}}\epsilon_i & v_{i+1} & \cdots & v_n \end{bmatrix}$$

where ϵ_i is the ith standard basis vector in $\mathbb{R}^{n-1}$. Then

$$\det(V_n^T\operatorname{diag}(x)V_n) = \det\left(x_i\left(\tfrac{n-1}{n}\right)\epsilon_i\epsilon_i^T + \textstyle\sum_{j\neq i}x_jv_jv_j^T\right)$$

which is of degree one in x_i by the linearity of the determinant in its ith column. □

As observed by Sanyal, such a determinantal identity for $e_{n-1}(x)$ establishes that $\mathbb{R}_+^{n,(1)}$ is a slice of $\mathcal{S}_+^{n-1} = \mathcal{S}_+^{n-1,(1-1)}$. We now have two expressions for the derivative $\frac{\partial}{\partial t}e_n(sx + t\mathbf{1}_n)\big|_{s=1}$, one from the definition (3.3.1) and one from (3.3.3). Comparing them allows us to deduce Proposition 3.2.3, that $\mathbb{R}_+^{n,(k)}$ is a slice of $\mathcal{S}_+^{n-1,(k-1)}$ for all

$1 \leq k \leq n-1$.

Proof of Proposition 3.2.3. From (3.3.1) and (3.3.3) we see that

$$\begin{aligned}\tfrac{\partial}{\partial t}e_n(sx+t\mathbf{1}_n)\big|_{s=1} &= \left[1\cdot e_{n-1}(x)+\cdots+(n-1)\cdot e_1(x)t^{n-2}+n\cdot t^{n-1}\right]\\ &= n\left[E_{n-1}(V_n^T\operatorname{diag}(x)V_n)+\cdots+E_1(V_n^T\operatorname{diag}(x)V_n)t^{n-2}+t^{n-1}\right].\end{aligned}$$

Comparing coefficients of powers of t we see that for $i = 0, 1, \ldots, n-1$

$$nE_{(n-1)-(i-1)}(V_n^T\operatorname{diag}(x)V_n) = (n-i)e_{n-i}(x).$$

Hence for $k = 1, 2, \ldots, n-1$, $x \in \mathbb{R}_+^{n,(k)}$ if and only if $V_n^T\operatorname{diag}(x)V_n \in \mathcal{S}_+^{n-1,(k-1)}$. □

■ 3.3.2 The polar derivative-based recursion: relating $\mathbb{R}_+^{n,(k)}$ and $\mathcal{S}_+^{n-1,(k)}$

In this section we relate $\mathbb{R}_+^{n,(k)}$ with $\mathcal{S}_+^{n-1,(k)}$, eventually proving Proposition 3.2.4. Our argument follows a pattern similar to the previous section. First we give a determinantal expression for the polar derivative $\frac{\partial}{\partial s}e_n(sx+t\mathbf{1}_n)\big|_{s=1}$, and then interpret it geometrically.

While our approach here is closely related to the approach of the previous section, things are a little more complicated. This is not surprising because our construction aims to express $\mathbb{R}_+^{n,(k)}$, which has an algebraic boundary of degree $n-k$, in terms of $\mathcal{S}_+^{n-1,(k)}$, which has an algebraic boundary of *smaller* degree, $n-k-1$. Hence it is not possible for $\mathbb{R}_+^{n,(k)}$ simply to be a slice of $\mathcal{S}_+^{n-1,(k)}$.

Block matrix notation: Let $\hat{\mathbf{1}}_n = \mathbf{1}_n/\sqrt{n}$ and define $Q_n = \begin{bmatrix}V_n & \hat{\mathbf{1}}_n\end{bmatrix}$ noting that Q_n is orthogonal. It is convenient to introduce the block matrix

$$\begin{aligned}M(x) := Q_n^T\operatorname{diag}(x)Q_n &= \begin{bmatrix}V_n^T\operatorname{diag}(x)V_n & V_n^T\operatorname{diag}(x)\hat{\mathbf{1}}_n\\ \hat{\mathbf{1}}_n^T\operatorname{diag}(x)V_n & \hat{\mathbf{1}}_n^T\operatorname{diag}(x)\hat{\mathbf{1}}_n\end{bmatrix}\\ &=: \begin{bmatrix}M_{11}(x) & M_{12}(x)\\ M_{12}(x)^T & M_{22}(x)\end{bmatrix}\end{aligned} \tag{3.3.4}$$

which reflects the fact that it is natural to work in coordinates that are adapted to the symmetry of the problem. (Indeed $\hat{\mathbf{1}}_n$ and the columns of V_n each span invariant subspaces for the permutation action on the coordinates of $\mathbb{R}^n$.)

Schur complements: In this section our results are expressed naturally in term of the *Schur complement* $(M/M_{22})(x) := M_{11}(x) - M_{12}(x)M_{22}(x)^{-1}M_{12}(x)^T$ which is well defined whenever $e_1(x) = nM_{22}(x) \neq 0$. The following lemma summarizes the main properties of the Schur complement that we use.

Lemma 3.3.2. *If* $M = \left[\begin{smallmatrix} M_{11} & M_{12} \\ M_{12}^T & M_{22}\end{smallmatrix}\right]$ *is a partitioned symmetric matrix with non-zero scalar* M_{22} *and* $M/M_{22} := M_{11} - M_{12}M_{22}^{-1}M_{12}^T$ *then*

$$\begin{bmatrix} M_{11} & M_{12} \\ M_{12}^T & M_{22}\end{bmatrix} = \begin{bmatrix} I_{n-1} & M_{12}M_{22}^{-1} \\ 0 & I_1 \end{bmatrix}\begin{bmatrix} M/M_{22} & 0 \\ 0 & M_{22}\end{bmatrix}\begin{bmatrix} I_{n-1} & 0 \\ M_{22}^{-1}M_{12}^T & I_1\end{bmatrix}. \tag{3.3.5}$$

This factorization immediately implies the following properties.

- *If* M *is invertible then the* $(1,1)$ *block of* M^{-1} *is given by* $[M^{-1}]_{11} = (M/M_{22})^{-1}$.
- *If* $M_{22} > 0$ *then*
$$M \succeq 0 \Longleftrightarrow M/M_{22} \succeq 0.$$

We now establish our determinantal expression for the polar derivative.

Lemma 3.3.3. *If* $e_1(x) = nM_{22}(x) \neq 0$ *then*

$$\tfrac{\partial}{\partial s} e_n(sx + t\mathbf{1}_n)\big|_{s=1} = e_1(x)\det((M/M_{22})(x) + tI_{n-1}). \tag{3.3.6}$$

Proof. First assume $x_i \neq 0$ for $i = 1, 2, \ldots, n$. If $x \in \mathbb{R}^n$ let x^{-1} denote its entry-wise inverse. Exploiting our determinantal (3.3.3) expression for the derivative we see that

$$\begin{aligned}
\tfrac{\partial}{\partial s} e_n(sx + t\mathbf{1}_n) &= e_n(x)\tfrac{\partial}{\partial s} e_n(s\mathbf{1}_n + tx^{-1}) \\
&\overset{*}{=} e_n(x)\, n\, \det(V_n^T \mathrm{diag}(tx^{-1} + s\mathbf{1}_n)V_n) \\
&= e_n(x)\, n\, \det(V_n^T\mathrm{diag}(x^{-1})V_n)\det(tI_{n-1} + s(V_n^T\mathrm{diag}(x^{-1})V_n)^{-1}) \\
&\overset{*}{=} e_n(x)e_{n-1}(x^{-1})\det(tI_{n-1} + s(V_n^T\mathrm{diag}(x^{-1})V_n)^{-1}) \\
&= e_1(x)\det(tI_{n-1} + s(V_n^T\mathrm{diag}(x^{-1})V_n)^{-1})
\end{aligned} \tag{3.3.7}$$

where the equalities marked with an asterisk are due to (3.3.3). Since Q_n is orthogonal $M(x)^{-1} = (Q_n^T\mathrm{diag}(x)Q_n)^{-1} = Q_n^T\mathrm{diag}(x^{-1})Q_n = M(x^{-1})$. Hence using a property of the Schur complement from Lemma 3.3.2 we see that

$$(V_n^T\mathrm{diag}(x^{-1})V_n)^{-1} = [M(x^{-1})]_{11}^{-1} = [M(x)^{-1}]_{11}^{-1} = (M/M_{22})(x).$$

Substituting this into (3.3.7) establishes the stated identity, which, by continuity, is valid for all x such that $e_1(x) = nM_{22}(x) \neq 0$. □

We now have two expressions for the polar derivative, namely (3.3.2) and (3.3.6). One comes from the definition of polar derivative, the other from the determinantal representation of Lemma 3.3.3. Expanding each and equating coefficients gives the following identities.

Lemma 3.3.4. *Let $x \in \mathbb{R}^n$ be such that $e_1(x) = nM_{22}(x) \neq 0$. Then for $k = 0, 1, 2, \ldots, n-1$*

$$e_1(x)E_{n-1-k}((M/M_{22})(x)) = (n-k)e_{n-k}(x).$$

Proof. Expanding the polar derivative two ways (from Lemma 3.3.3 and (3.3.2)) we obtain

$$\begin{aligned}\tfrac{\partial}{\partial s}e_n(sx+t\mathbf{1}_n)\big|_{s=1} &= \left[n\cdot e_n(x) + (n-1)\cdot e_{n-1}(x)t + \cdots + 1\cdot e_1(x)t^{n-1}\right]\\ &= e_1(x)\left[E_{n-1}((M/M_{22})(x)) + E_{n-2}((M/M_{22})(x))t + \cdots + t^{n-1}\right].\end{aligned}$$

The result follows by equating coefficients of t^k. $\square$

We are now in a position to prove the main result of this section.

Proof of Proposition 3.2.4. From the definition of $M(x)$ in (3.3.4), observe that because Q_n is orthogonal, the constraint $\operatorname{diag}(x) \succeq V_n Z V_n^T$ holds if and only if

$$M(x) = Q_n^T \operatorname{diag}(x) Q_n \succeq Q_n^T(V_n Z V_n^T)Q_n = \left[\begin{smallmatrix} Z & 0\\ 0 & 0\end{smallmatrix}\right].$$

Hence we aim to establish the following statement that is equivalent to Proposition 3.2.4

$$\mathbb{R}_+^{n,(k)} = \left\{x \in \mathbb{R}^n : \exists Z \in \mathcal{S}_+^{n-1,(k)} \text{ s.t. } M(x) \succeq \begin{bmatrix} Z & 0\\ 0 & 0\end{bmatrix}\right\} \quad \text{for } k = 1, 2, \ldots, n-2.$$

The arguments that follow repeatedly use the fact (from Lemma 3.3.2) that if $e_1(x) = nM_{22}(x) > 0$ then

$$M(x) \succeq \begin{bmatrix} Z & 0\\ 0 & 0\end{bmatrix} \quad \Longleftrightarrow \quad (M/M_{22})(x) \succeq Z. \tag{3.3.8}$$

With these preliminaries established, we turn to the proof of Proposition 3.2.4. First suppose there is $Z \in \mathcal{S}_+^{n-1,(k)}$ such that $M(x) - \left[\begin{smallmatrix} Z & 0\\ 0 & 0\end{smallmatrix}\right] \succeq 0$. There are two cases to consider, depending on whether $M_{22}(x)$ is positive or zero.

Suppose we are in the case where $e_1(x) = nM_{22}(x) > 0$. Then $(M/M_{22})(x) \succeq Z$, so there is some $Z' \in \mathcal{S}_+^{n-1}$ such that

$$(M/M_{22})(x) = Z + Z' \in \mathcal{S}_+^{n-1,(k)} + \mathcal{S}_+^{n-1} = \mathcal{S}_+^{n-1,(k)}$$

where the last equality holds because $\mathcal{S}_+^{n-1,(k)} \supset \mathcal{S}_+^{n-1}$. It follows that $x \in \mathbb{R}_+^{n,(k)}$ because $e_1(x) > 0$ (by assumption) and by Lemma 3.3.4,

$$ie_i(x) = e_1(x)E_{i-1}((M/M_{22})(x)) \geq 0 \quad \text{for } i = 2, 3, \ldots, n-k.$$

Now consider the case where $e_1(x) = nM_{22}(x) = 0$. Since

$$\begin{bmatrix} M_{11}(x) - Z & M_{12}(x) \\ M_{12}(x)^T & M_{22}(x) \end{bmatrix} = \begin{bmatrix} M_{11}(x) - Z & V_n^T x/\sqrt{n} \\ x^T V_n/\sqrt{n} & 0 \end{bmatrix} \succeq 0$$

it follows that $V_n^T x = 0$. Since, $\hat{\mathbf{1}}_n^T x = 0$ we see that $Q_n^T x = 0$ so $x = 0 \in \mathbb{R}_+^{n,(k)}$.

Consider the reverse inclusion and suppose $x \in \mathbb{R}_+^{n,(k)}$. Again there are two cases depending on whether $e_1(x)$ is positive or zero. If $e_1(x) > 0$ take $Z = (M/M_{22})(x)$. Then, by (3.3.8), $M(x) \succeq \left[\begin{smallmatrix} Z & 0 \\ 0 & 0 \end{smallmatrix}\right]$. To see that $Z \in \mathcal{S}_+^{n-1,(k)}$ note that by Lemma 3.3.4,

$$E_i((M/M_{22})(x)) = (i+1)\frac{e_{i+1}(x)}{e_1(x)} \geq 0 \quad \text{for } i = 1, 2, \ldots, n-1-k.$$

If $x \in \mathbb{R}_+^{n,(k)}$ and $e_1(x) = 0$ then we use the assumption that $k \leq n-2$. Under this assumption $x \in \mathbb{R}_+^{n,(k)} \cap \{x : e_1(x) = 0\} = \{0\}$. In this case we can simply take $Z = 0 \in \mathcal{S}_+^{n-1,(k)}$ since $M(x) = 0 \succeq 0 = \left[\begin{smallmatrix} Z & 0 \\ 0 & 0 \end{smallmatrix}\right]$. □

■ 3.3.3 Dual relationships

We conclude this section by establishing Propositions 3.2.3D and 3.2.4D, the dual versions of Propositions 3.2.3 and 3.2.4. Both follow from Lemma 2.3.12 which we restate here for convenience.

Lemma 2.3.12. *Let W, V_1, V_2 be finite-dimensional real inner product spaces. Suppose $K_1 \subseteq V_1$ is a closed convex cone and $\mathcal{A} : W \to V_1$ and $\mathcal{B} : W \to V_2$ are linear maps. Let*

$$K_2 = \{\mathcal{B}(x) : \mathcal{A}(x) \in K_1\} \subseteq V_2.$$

Furthermore, assume there is some $x_0 \in V_1$ such that $\mathcal{A}(x_0)$ is in the relative interior of K_1. Then

$$K_2^* = \{w \in V_2 : \exists y \in K_1^* \text{ s.t. } \mathcal{B}^*(w) = \mathcal{A}^*(y)\}.$$

Proof of Proposition 3.2.3D. Define $\mathcal{A} : \mathbb{R}^n \to \mathcal{S}^{n-1}$ by $\mathcal{A}(x) = V_n^T \operatorname{diag}(x) V_n$ and define $\mathcal{B}$ to be the identity on $\mathbb{R}^n$. Then by Proposition 3.2.3

$$\mathbb{R}_+^{n,(k)} = \{\mathcal{B}(x) : \mathcal{A}(x) \in \mathcal{S}_+^{n-1,(k-1)}\}.$$

Clearly $\mathcal{B}^*$ is the identity on $\mathbb{R}^n$ and $\mathcal{A}^* : \mathcal{S}^{n-1} \to \mathbb{R}^n$ is given by $\mathcal{A}^*(Y) = \operatorname{diag}(V_n Y V_n^T)$. Since $\mathcal{A}(\mathbf{1}_n) = I_{n-1}$ is in the interior of $\mathcal{S}_+^{n-1,(k-1)}$, applying Lemma 2.3.12 we obtain

$$(\mathbb{R}_+^{n,(k)})^* = \{w \in \mathbb{R}^n : \exists Y \in (\mathcal{S}_+^{n-1,(k-1)})^* \text{ s.t. } w = \operatorname{diag}(V_n Y V_n^T).\}$$

Eliminating w gives the statement in Proposition 3.2.3D. □

Proof of Proposition 3.2.4D. Define $\mathcal{A} : \mathbb{R}^n \times \mathcal{S}^{n-1} \to \mathcal{S}^n \times \mathcal{S}^{n-1}$ by

$$\mathcal{A}(x, Z) = (\operatorname{diag}(x) - V_n Z V_n^T, Z)$$

and $\mathcal{B} : \mathbb{R}^n \times \mathcal{S}^{n-1} \to \mathbb{R}^n$ by $\mathcal{B}(x, Z) = x$. Then by Proposition 3.2.4

$$\mathbb{R}_+^{n,(k)} = \{\mathcal{B}(x, Z) : \mathcal{A}(x, Z) \in \mathcal{S}_+^n \times \mathcal{S}_+^{n-1,(k)}\}.$$

A straightforward computation shows that $\mathcal{B}^* : \mathbb{R}^n \to \mathbb{R}^n \times \mathcal{S}^{n-1}$ is given by $\mathcal{B}^*(w) = (w, 0)$. Furthermore $\mathcal{A}^* : \mathcal{S}^n \times \mathcal{S}^{n-1}$ is given by $\mathcal{A}^*(Y, W) = (\operatorname{diag}(Y), W - V_n^T Y V_n)$. Since $\mathcal{A}(2\mathbf{1}_n, I_{n-1})$ is in the interior of $\mathcal{S}_+^n \times \mathcal{S}_+^{n-1,(k)}$, applying Lemma 2.3.12 we obtain

$$(\mathbb{R}_+^{n,(k)})^* = \{w \in \mathbb{R}^n : \exists (Y, W) \in \mathcal{S}_+^n \times (\mathcal{S}_+^{n-1,(k)})^* \text{ s.t. } w = \operatorname{diag}(W),\ \ V_n^T Y V_n = W\}.$$

Eliminating W and w gives the statement in Proposition 3.2.4D. □

■ 3.4 Exploiting symmetry: relating $\mathcal{S}_+^{n,(k)}$ and $\mathbb{R}_+^{n,(k)}$ and their dual cones

In the introduction we observed that $\mathcal{S}_+^{n,(k)}$ is invariant under the action of orthogonal matrices by conjugation on $\mathcal{S}^n$ and that its diagonal slice is $\mathbb{R}_+^{n,(k)}$. In this section we explain how to use these properties to construct the semidefinite representation of $\mathcal{S}_+^{n,(k)}$ in terms of $\mathbb{R}_+^{n,(k)}$ stated in Proposition 3.2.2. We then discuss how the duals of these two cones relate. The material in this section is well known so in some places we give appropriate references to the literature rather than providing proofs.

■ 3.4.1 Relating $\mathcal{S}_+^{n,(k)}$ and $\mathbb{R}_+^{n,(k)}$: proof of Proposition 3.2.2

Let $O(n)$ denote the group of $n \times n$ orthogonal matrices. The *Schur-Horn cone* is

$$\mathrm{SH}_n = \left\{(X, z) :\ \ z_1 \geq z_2 \geq \cdots \geq z_n,\ \ X \in \operatorname{conv}_{Q \in O(n)}\{Q^T \operatorname{diag}(z) Q\}\right\}, \tag{3.4.1}$$

the set of pairs (X, z) such that z is in weakly decreasing order and X is in the convex hull of symmetric matrices with ordered spectrum z. We call this the Schur-Horn cone because all *symmetric Schur-Horn orbitopes* [110] appear as slices of SH_n of the form $\{X : (X, z_0) \in \mathrm{SH}_n\}$ where z_0 is fixed and in weakly decreasing order.

Whenever a convex subset $C \subset \mathcal{S}^n$ is invariant under orthogonal conjugation, i.e. C is a spectral set, we can express C in terms of the Schur-Horn cone and the (hopefully simpler) diagonal slice of C as follows.

Lemma 3.4.1. *If $C \subset \mathcal{S}^n$ is convex and invariant under orthogonal conjugation then*

$$C = \{X \in \mathcal{S}^n : \exists z \in \mathbb{R}^n \ \ s.t. \ \ (X, z) \in \mathrm{SH}_n, \ \ \mathrm{diag}(z) \in C\}.$$

Proof. Suppose $X \in C$. Take $z = \lambda(X)$, the ordered vector of eigenvalues of X. Then there is some $Q \in O(n)$ such that $X = Q^T \mathrm{diag}(\lambda(X))Q$ so $(X, \lambda(X)) \in \mathrm{SH}_n$. By the orthogonal invariance of C, $X \in C$ implies that $QXQ^T = \mathrm{diag}(\lambda(X)) \in C$.

For the reverse inclusion, suppose there is $z \in \mathbb{R}^n$ such that $(X, z) \in \mathrm{SH}_n$ and $\mathrm{diag}(z) \in C$. Then by the orthogonal invariance of C, $Q^T\mathrm{diag}(z)Q \in C$ for all $Q \in O(n)$. Since C is convex, $\mathrm{conv}_{Q\in O(n)}\{Q^T \mathrm{diag}(z)Q\} \subseteq C$. Hence $(X, z) \in \mathrm{SH}_n$ implies that

$$X \in \mathrm{conv}_{Q\in O(n)}\{Q^T \mathrm{diag}(z)Q\} \subseteq C.$$

□

The first statement in Proposition 3.2.2 follows from Lemma 3.4.1 by recalling that $\mathcal{S}_+^{n,(k)}$ is orthogonally invariant and $\mathbb{R}_+^{n,(k)} = \{z \in \mathbb{R}^n : \mathrm{diag}(z) \in \mathcal{S}_+^{n,(k)}\}$.

Proving the remainder of Proposition 3.2.2 then reduces to establishing the correctness of the stated semidefinite representation of SH_n. This can be deduced from the following two well-known results.

Lemma 3.4.2. *If $\lambda(X)$ is ordered so that $\lambda_1(X) \geq \cdots \geq \lambda_n(X)$ then $(X, z) \in \mathrm{SH}_n$ if and only if $z_1 \geq z_2 \geq \cdots \geq z_n$,*

$$\mathrm{tr}(X) = \sum_{i=1}^{n} \lambda_i(X) = \sum_{i=1}^{n} z_i, \quad \textit{and} \quad \sum_{i=1}^{\ell} \lambda_i(X) \leq \sum_{i=1}^{\ell} z_i \quad \textit{for } \ell = 1, 2, \ldots, n-1.$$

In other words $(X, z) \in \mathrm{SH}_n$ if and only if z is weakly decreasing and $\lambda(X)$ is *majorized* by z. This is discussed, for example, in [110, Corollary 3.2]. To turn this characterization into a semidefinite representation, it suffices to have semidefinite representations of the epigraphs (see Section 2.3) of the convex functions $s_\ell(X) := \sum_{i=1}^{\ell} \lambda_i(X)$. These are given by Nesterov and Nemirovski in [89, Section 6.4.3, Example 7].

Lemma 3.4.3. *If $2 \leq \ell \leq n-1$, the epigraph of the convex function $s_\ell(X) = \sum_{i=1}^{\ell} \lambda_i(X)$ has a semidefinite representation of size $O(n)$ given by*

$$\begin{aligned}&\{(X, t) : s_\ell(X) \leq t\} = \\ &\qquad\qquad \{(X, t) : \exists s \in \mathbb{R}, \ Z \in \mathcal{S}^n \quad s.t. \quad Z \succeq 0, \ \ X \preceq Z + sI, \ \ \mathrm{tr}(Z) + s\,\ell \leq t\}.\end{aligned}$$

The epigraph of $s_1(X)$ has a simpler semidefinite representation as

$$\{(X, t) : s_1(X) \leq t\} = \{(X, t) : X \preceq tI\}.$$

3.4.2 Relating the corresponding dual cones: proof of Proposition 3.2.2D

We now turn to the relationship between $(\mathcal{S}_+^{n,(k)})^*$ and $(\mathbb{R}_+^{n,(k)})^*$. Note that $(\mathcal{S}_+^{n,(k)})^*$ is invariant under orthogonal conjugation. So the claim (Proposition 3.2.2D) that

$$(\mathcal{S}_+^{n,(k)})^* = \{Y \in \mathcal{S}^n : \exists w \in \mathbb{R}^n \text{ s.t. } w \in (\mathbb{R}_+^{n,(k)})^*, \ \ (Y, w) \in \mathrm{SH}_n\}$$

would follow from Lemma 3.4.1 once we know that the diagonal slice of $(\mathcal{S}_+^{n,(k)})^*$ is $(\mathbb{R}_+^{n,(k)})^*$. The next lemma establishes this.

Lemma 3.4.4. *The intersection of $(\mathcal{S}_+^{n,(k)})^*$ with the subspace of diagonal matrices is $(\mathbb{R}_+^{n,(k)})^*$, i.e.*

$$\{y \in \mathbb{R}^n : \mathrm{diag}(y) \in (\mathcal{S}_+^{n,(k)})^*\} = \{z \in \mathbb{R}^n : \mathrm{diag}(z) \in \mathcal{S}_+^{n,(k)}\}^* = (\mathbb{R}_+^{n,(k)})^*. \tag{3.4.2}$$

Proof. For every subset $I \subseteq \{1, 2, \ldots, n\}$ let Δ_I denote the diagonal matrix with $[\Delta_I]_{ii} = 1$ if $i \in I$ and $[\Delta_I]_{ii} = -1$ otherwise. The Δ_I are all orthogonal, form a group under matrix multiplication, and act on symmetric matrices by $X \mapsto \Delta_I X \Delta_I^T$. A symmetric matrix is fixed by the action of all the Δ_I if and only if it is diagonal.

Since $\mathcal{S}_+^{n,(k)}$ is invariant under conjugation by orthogonal matrices, it is invariant under conjugation by all the Δ_I. The result then follows by applying Lemma 2.6.9 from Chapter 2. □

3.5 Concluding remarks

We conclude with some comments about ways to simplify our representations a little. We also discuss some open questions.

3.5.1 Simplifications

If we can simplify a representation of $\mathbb{R}_+^{n,(k)}$ or $\mathcal{S}_+^{n,(k)}$ for some $k = i$, that allows us to simplify the derivative-based representations for $k \geq i$ and the polar derivative-based representations for $k \leq i$. For example $\mathbb{R}_+^{n,(n-2)}$ can be succinctly expressed in terms of the second-order cone $Q_+^{n+1} = \{x \in \mathbb{R}^{n+1} : (\sum_{i=1}^n x_i^2)^{1/2} \leq x_{n+1}\}$ as

$$\mathbb{R}_+^{n,(n-2)} = \{x \in \mathbb{R}^n : (x, e_1(x)) \in Q_+^{n+1}\}.$$

Then we can represent $\mathcal{S}_+^{n,(n-2)}$ in terms of the second-order cone as

$$\mathcal{S}_+^{n,(n-2)} = \{Z \in \mathcal{S}^n : (Z, \mathrm{tr}(Z)) \in Q_+^{n^2+1}\}$$

because $\mathrm{tr}(Z) = \sum_{i=1}^{n} \lambda_i(Z)$ and $\sum_{i,j=1}^{n} Z_{ij}^2 = \sum_{i=1}^{n} \lambda_i(Z)^2$. This should be used as a base case instead of $\mathcal{S}_+^{n,(n-1)}$ in the polar derivative-based representations.

As an example of this, Proposition 3.2.4 can be used to give a concise representation of $\mathbb{R}_+^{n,(n-3)}$ in terms of the second-order cone as

$$\begin{aligned} x \in \mathbb{R}_+^{n,(n-3)} \quad \iff \quad & \exists Z \in \mathcal{S}^{n-1} \ \text{ such that} \\ & \mathrm{diag}(x) \succeq V_n Z V_n^T \ \text{ and } \ (Z, \mathrm{tr}(Z)) \in Q_+^{(n-1)^2+1}. \end{aligned}$$

■ 3.5.2 Lower bounds on the size of representations

The explicit constructions given in this chapter establish upper bounds on the minimum size of semidefinite representations of $\mathcal{S}_+^{n,(k)}$ and $\mathbb{R}_+^{n,(k)}$. To assess how good our representations are, it is interesting to establish corresponding *lower* bounds on the size of semidefinite representations of $\mathbb{R}_+^{n,(k)}$ and $\mathcal{S}_+^{n,(k)}$. Since $\mathbb{R}_+^{n,(k)}$ is a slice of $\mathcal{S}_+^{n,(k)}$, any lower bound on the size of a semidefinite representation of $\mathbb{R}_+^{n,(k)}$ also provides a lower bound on the size of a semidefinite representation of $\mathcal{S}_+^{n,(k)}$. Hence we focus our discussion on $\mathbb{R}_+^{n,(k)}$.

In the case of $\mathbb{R}_+^{n,(n-1)}$, a halfspace, the obvious semidefinite representation of size one is clearly of minimum size. Less trivial is the case of $\mathbb{R}_+^{n,(0)}$, the non-negative orthant. It has been shown by Gouveia et al. [56, Section 5] that $\mathbb{R}_+^n$ does not admit a semidefinite representation of size smaller than n. Hence the obvious representation of $\mathbb{R}_+^n$ as the restriction of $\mathcal{S}_+^n$ to the diagonal is of minimum size.

For each k, the slice of $\mathbb{R}_+^{n,(k)}$ obtained by setting the last k variables to zero is $\mathbb{R}_+^{n-k}$. Hence any semidefinite representation of $\mathbb{R}_+^{n,(k)}$ has size at least $n-k$, the minimum size of a semidefinite representation of $\mathbb{R}_+^{n-k}$. This argument establishes that Sanyal's spectrahedral representation of $\mathbb{R}_+^{n,(1)}$ of size $n-1$ is actually a minimum size semidefinite representation of $\mathbb{R}_+^{n,(1)}$.

Problem 3.5.1. *Find lower bounds on the size of semidefinite representations of the cones $\mathbb{R}_+^{n,(k)}$ for $2 \leq k \leq n-2$.*

The cones $\mathbb{R}_+^{n,(k)}$ are invariant under permutation of coordinates. Furthermore, the semidefinite representations of $\mathbb{R}_+^{n,(k)}$ given in this chapter are *equivariant* with respect to this action of the symmetric group. (See Definition 2.6.10 for a precise definition of this notion.) As such it would also be interesting to establish lower bounds on the size of equivariant semidefinite representations of the derivative relaxations of the non-negative orthant.

Problem 3.5.2. *Find lower bounds on the size of equivariant semidefinite representations of the cones $\mathbb{R}_+^{n,(k)}$ for $2 \leq k \leq n-2$.*

This may be significantly easier than Problem 3.5.1. This is because the additional equivariance assumptions of Problem 3.5.2 allow us to use tools from representation theory to find obstructions to the existence of semidefinite representations. Indeed such representation-theoretic tools have already proved to be quite effective in giving such lower bounds for orbitopes [43].

3.5.3 Spectrahedral representations of the cones $\mathcal{S}_+^{n,(k)}$

We have shown that the cones $\mathbb{R}_+^{n,(k)}$ and $\mathcal{S}_+^{n,(k)}$ have semidefinite representations of polynomial size. Brändén [17] has shown that the cones $\mathbb{R}_+^{n,(k)}$ also have spectrahedral representations (see Section 3.1.3). It is obvious that $\mathcal{S}_+^{n,(k)}$ has a spectrahedral representation when $k = 0, n-2, n-1$, since in these cases $\mathcal{S}_+^{n,(k)}$ is the positive semidefinite cone, a quadratic cone, and a half-space, respectively.

Question 3.5.3. *Does $\mathcal{S}_+^{n,(k)}$ have a spectrahedral representation for all n and all $1 \leq k \leq n-3$?*

Resolving this question is a natural step in understanding the generalized Lax conjecture (Conjecture 2.5.8) which asks whether all hyperbolicity cones are spectrahedral.

Chapter 4

Semidefinite descriptions of the convex hull of rotation matrices

4.1 Introduction

Optimization problems in which the decision variables are constrained to be in the set of orthogonal matrices

$$O(n) := \{X \in \mathbb{R}^{n\times n} : \; X^T X = I\} \tag{4.1.1}$$

arise in many contexts (see, e.g., [84, 86] and references therein), particularly when searching over Euclidean isometries or orthonormal frames. In some situations, especially those arising from physical problems, we require the additional constraint that the decision variables be in the set of rotation matrices

$$SO(n) := \{X \in \mathbb{R}^{n\times n} : \; X^T X = I, \;\; \det(X) = 1\} \tag{4.1.2}$$

representing Euclidean isometries that also *preserve orientation*. For example, these additional constraints arise in problems involving attitude estimation for spacecraft [97], pose estimation in computer vision applications [65], or in understanding protein folding [80]. The unit determinant constraint is important in these situations because we typically cannot reflect physical objects such as spacecraft or molecules.

The set of $n \times n$ rotation matrices is non-convex, so optimization problems over rotation matrices are ostensibly non-convex optimization problems. An important approach to global non-convex optimization is to approximate the original non-convex problem with a tractable convex optimization problem. In some circumstances it may even be possible to *exactly reformulate* the original non-convex problem as a tractable convex problem. This approach to global optimization via convexification has been very influential in combinatorial optimization [121], and more generally in polynomial optimization via the machinery of moments and sums of squares [14].

As an example of a problem amenable to this approach, in Section 4.2 we describe the problem of jointly estimating the attitude and spin-rate of a spinning satellite and show how to reformulate this ostensibly non-convex problem as a convex optimization problem that, using the constructions in this chapter, can be expressed as a semidefinite program.

When we attempt to convexify optimization problems involving rotation matrices two natural geometric objects arise. The first of these is the *convex hull* of $SO(n)$ which we denote, throughout, by $\operatorname{conv} SO(n)$. The second convex body of interest in this chapter is the *polar* of $SO(n)$, the set of linear functionals that take value at most one on $SO(n)$, i.e.,

$$SO(n)^\circ = \{Y \in \mathbb{R}^{n\times n} : \langle Y, X\rangle \leq 1 \text{ for all } X \in SO(n)\}$$

where we have identified $\mathbb{R}^{n\times n}$ with its dual space via the trace inner product $\langle Y, X\rangle = \operatorname{tr}(Y^T X)$. These two convex bodies are closely related. Since $\operatorname{conv} SO(n)$ is closed and contains the origin it follows from a basic result of convex analysis (Proposition 2.3.7 in Chapter 2) that $\operatorname{conv} SO(n) = (SO(n)^\circ)^\circ$.

We also study the convex hull and the polar of orthogonal matrices in this chapter. It is well-known that these correspond to commonly used matrix norms (see, e.g., [110]). The convex hull of $O(n)$ is the *operator norm ball*, the set of $n \times n$ matrices with largest singular value at most one, and the polar of $O(n)$ is the *nuclear norm ball*, the set of $n \times n$ matrices such that the sum of the singular values is at most one, i.e.

$$\operatorname{conv} O(n) = \left\{X \in \mathbb{R}^{n\times n} : \sigma_1(X) \leq 1\right\} \text{ and } O(n)^\circ = \left\{X \in \mathbb{R}^{n\times n} : \sum_{i=1}^{n} \sigma_i(X) \leq 1\right\}.$$

Note that $O(n)$ is the (disjoint) union of $SO(n)$ and the set $SO^-(n) := \{X \in \mathbb{R}^{n\times n} : X^T X = I, \det(X) = -1\}$. As such, it follows from basic properties of the polar (see Lemma 2.3.8 in Chapter 2) that

$$O(n)^\circ = SO(n)^\circ \cap SO^-(n)^\circ, \tag{4.1.3}$$

allowing us to deduce properties of $O(n)^\circ$ from those of $SO(n)^\circ$. On the other hand we show in Proposition 4.4.6 that for $n \geq 3$,

$$\operatorname{conv} SO(n) = (\operatorname{conv} O(n)) \cap (n-2) SO^-(n)^\circ, \tag{4.1.4}$$

allowing us to deduce properties of $\operatorname{conv} SO(n)$ from properties of $\operatorname{conv} O(n)$ and $SO^-(n)^\circ$. Figure 4.1 illustrates the differences between $\operatorname{conv} SO(n)$ and $\operatorname{conv} O(n)$ and the relationship described in (4.1.3).

The convex bodies $\operatorname{conv} SO(n)$ and $\operatorname{conv} O(n)$ are examples of *orbitopes*, a family of

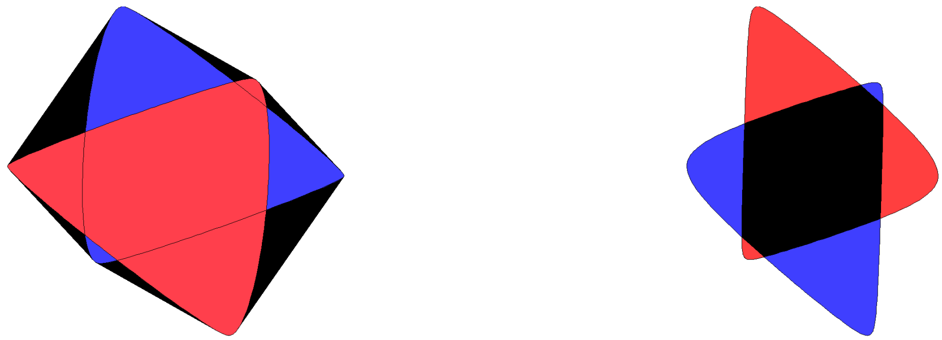


Figure 4.1: On the left are shown 2-dimensional projections of $\operatorname{conv} SO(3)$ (red), $\operatorname{conv} SO^-(3)$ (blue), and $\operatorname{conv} O(3) = \operatorname{conv}[SO(3) \cup SO^-(3)]$ (black). On the right are shown the corresponding 2-dimensional sections of $SO(3)^\circ$ (red), $SO^-(3)^\circ$ (blue), and $O(3)^\circ = SO(3)^\circ \cap SO^-(3)^\circ$ (black). These were created by optimizing 100 linear functionals over each of these sets to obtain 100 boundary points. The optimization was performed by implementing our spectrahedral representations in the parser YALMIP [78], and solving the semidefinite programs numerically using SDPT3 [127].

highly symmetric convex bodies that arise from representations of groups [8, 10, 110]. Suppose a compact group G acts on $\mathbb{R}^n$ by linear transformations and $x_0 \in \mathbb{R}^n$. Then the *orbit* of x_0 under G is

$$G \cdot x_0 = \{g \cdot x_0 : g \in G\} \subseteq \mathbb{R}^n$$

and the corresponding *orbitope* is $\operatorname{conv}(G \cdot x_0)$, the convex hull of the orbit. The sets $O(n)$ and $SO(n)$ defined above can be thought of as the orbit of the identity matrix $I \in \mathbb{R}^{n\times n}$ under the linear action of the groups $O(n)$ and $SO(n)$, respectively, by right multiplication on $n \times n$ matrices. The corresponding orbitopes are known as the *tautological $O(n)$ orbitope* and the *tautological $SO(n)$ orbitope* respectively [110]. The set $SO^-(n)$ can be viewed as the orbit of $R := \operatorname{diag}^*(1, 1, \ldots, 1, -1)$, the diagonal matrix with diagonal entries $(1, 1, \ldots, 1, -1)$, under the same $SO(n)$ action on $n \times n$ matrices. Note that $SO^-(n)$ is then the image of $SO(n)$ under the invertible linear map $X \mapsto R \cdot X$.

Spectrahedra and semidefinite representations As we discussed in Chapter 2, for convex reformulations or relaxations involving the convex hull of $SO(n)$ to be useful from a computational point of view we need an effective description of the convex body $\operatorname{conv} SO(n)$. One effective way to describe a convex body is to express it as a *spectrahedron*, the intersection of the cone of symmetric positive semidefinite matrices with an affine subspace. We discuss the properties of spectrahedra in Section 2.4 of Chapter 2.

We briefly recall their algebraic description here, since we use such descriptions throughout the chapter. If $C \subset \mathbb{R}^n$ contains the origin in its interior[1], it is a spectrahedron if it can be expressed as

$$C = \left\{ x \in \mathbb{R}^n : I_m + \sum_{i=1}^{n} A^{(i)} x_i \succeq 0 \right\} \tag{4.1.5}$$

where I_m is the $m \times m$ identity matrix, $A^{(1)}, A^{(2)}, \ldots, A^{(n)}$ are $m \times m$ real symmetric matrices. If the matrices $A^{(i)}$ are $m \times m$, we call the description (4.1.5) a *spectrahedral representation of size m*.

We are also interested in semidefinite representations of convex sets, i.e. descriptions as the image of a spectrahedron under a linear map (see Section 2.4 of Chapter 2). Throughout much of this chapter we consider only spectrahedral representations, confining our discussion of semidefinite representations to Section 4.5.2.

Doubly spectrahedral convex sets In this chapter we are interested in both $SO(n)^\circ$ and $\operatorname{conv} SO(n)$, and so study both from the point of view of semidefinite programming. For finite sets S, both S° and $\operatorname{conv} S$ are polyhedra. On the other hand, for infinite sets S, usually neither S° nor $\operatorname{conv} S$ are spectrahedra. Even if a convex set is a spectrahedron, typically its polar is not a spectrahedron (see Section 4.6). We use the term *doubly spectrahedral convex sets* to refer to those very special convex sets C with the property that both C and C° are spectrahedra.

Main contribution The main contribution of this chapter is to establish that $\operatorname{conv} SO(n)$ is doubly spectrahedral and to give explicit spectrahedral representations of both $SO(n)^\circ$ and $\operatorname{conv} SO(n)$.

Main proof technique The main idea behind our representations is that we start with a *parameterization* of $SO(n)$, rather than working with the defining equations in (4.1.2). The parameterization is a direct (and classical) generalization of the widely used unit quaternion parameterization of $SO(3)$. In higher dimensions the unit quaternions are replaced with $\mathrm{Spin}(n)$, a multiplicative subgroup of the invertible elements of a Clifford algebra. In the cases $n = 2$ and $n = 3$ it is relatively straightforward to produce our semidefinite representations directly from this parameterization. For $n \geq 4$ the parameterization does not immediately yield our semidefinite representations. The additional arguments required to establish the correctness of our representations for $n \geq 4$ form the main technical contribution of the chapter.

[1] We can assume this without loss of generality by translating C and restricting to its affine hull

4.1.1 Statement of results

In this section we explicitly state the spectrahedral representations that we prove are correct in subsequent sections of the chapter. In particular we state spectrahedral representations for $SO(n)^\circ$ and $\operatorname{conv} SO(n)$, as well as a spectrahedral representation of $O(n)^\circ$, the nuclear norm ball. All the spectrahedral representations stated in this section are of minimum size (see Theorem 4.1.4). The reader primarily interested in implementing our semidefinite representations should find all the information necessary to do so in this section.

Matrices of the spectrahedral representations Our main results are stated in terms of a collection of symmetric $2^{n-1} \times 2^{n-1}$ matrices denoted $(A^{(ij)})_{1\leq i,j\leq n}$. We give concrete descriptions of them here in terms of the Kronecker product of 2×2 matrices, deferring more invariant descriptions to Section 4.7. The matrices $A^{(ij)}$ can be expressed as

$$A^{(ij)} = -P_{\text{even}}^T \lambda_i \rho_j P_{\text{even}} \tag{4.1.6}$$

where $(\lambda_i)_{i=1}^n$ and $(\rho_i)_{i=1}^n$ are the $2^n \times 2^n$ skew-symmetric matrices defined concretely by

$$\lambda_i = \overbrace{\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix} \otimes \cdots \otimes \begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}}^{i-1} \otimes \begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix} \otimes \overbrace{\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \otimes \cdots \otimes \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}}^{n-i}$$

and

$$\rho_j = \underbrace{\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \otimes \cdots \otimes \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}}_{j-1} \otimes \begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix} \otimes \underbrace{\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix} \otimes \cdots \otimes \begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}}_{n-j}$$

and where P_{even} is the $2^n \times 2^{n-1}$ matrix with orthonormal columns

$$P_{\text{even}} = \frac{1}{2}\begin{bmatrix}1\\ 1\end{bmatrix} \otimes \overbrace{\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \otimes \cdots \otimes \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}}^{n-1} + \frac{1}{2}\begin{bmatrix}1\\ -1\end{bmatrix} \otimes \overbrace{\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix} \otimes \cdots \otimes \begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}}^{n-1}.$$

Note that $P_{\text{even}}^T M P_{\text{even}}$ just selects a particular $2^{n-1} \times 2^{n-1}$ principal submatrix of M. For any $1 \leq i \leq n$, λ_i and ρ_i are both skew-symmetric since they are formed by taking the Kronecker product of $n-1$ symmetric matrices and one skew-symmetric matrix. Furthermore, for any pair $1 \leq i,j \leq n$ the product $\lambda_i \rho_j$ is symmetric. This is because if $i \geq j$, $\lambda_i\rho_j$ is the Kronecker product of n symmetric matrices, and if $i < j$, $\lambda_i\rho_j$ is the Kronecker product of $n-2$ symmetric matrices and two skew-symmetric matrices. It follows that each $A^{(ij)}$ is symmetric. Furthermore since λ_i and ρ_j are signed permutation matrices, so is $-\lambda_i\rho_j$. From this we can see that all of the entries of the $A^{(ij)}$ are 0, 1, or -1.

Spectrahedral representations The following, which we prove in Section 4.4, is the main technical result of this chapter.

Theorem 4.1.1. *The polar of $SO(n)$ is a spectrahedron. Explicitly*

$$SO(n)^\circ = \left\{ Y \in \mathbb{R}^{n\times n} : \sum_{i,j=1}^{n} A^{(ij)} Y_{ij} \preceq I_{2^{n-1}} \right\} \tag{4.1.7}$$

where the $2^{n-1} \times 2^{n-1}$ matrices $A^{(ij)}$ are defined in (4.1.6).

Since $O(n) = SO(n) \cup SO^-(n)$ as a corollary of Theorem 4.1.1 we obtain a spectrahedral representation of $O(n)^\circ = SO(n)^\circ \cap SO^-(n)^\circ$.

Theorem 4.1.2. *The polar of $O(n)$ is a spectrahedron. Explicitly*

$$O(n)^\circ = \left\{ Y \in \mathbb{R}^{n\times n} : \sum_{i,j=1}^{n} A^{(ij)} Y_{ij} \preceq I_{2^{n-1}},\ \sum_{i,j=1}^{n} A^{(ij)} [RY]_{ij} \preceq I_{2^{n-1}} \right\}.$$

where $R = \operatorname{diag}^*(1, 1, \ldots, 1, -1)$.

Just because a convex set C is a spectrahedron does not, in general, mean that its polar is also spectrahedron (see Section 4.6 for a simple example). Even if we are in the special case where C is doubly spectrahedral, it is not straightforward to obtain a spectrahedral representation of C° from a spectrahedral representation of C. For example, if C is a polyhedron (and so certainly doubly spectrahedral) this is the problem of computing a facet description of C° (i.e. the vertices of C) from a facet description of C.

Nevertheless, we obtain a *spectrahedral* representation of $\operatorname{conv} SO(n)$ by showing that, for $n \geq 3$, $\operatorname{conv} SO(n) = (\operatorname{conv} O(n)) \cap (n-2)SO^-(n)^\circ$ (Proposition 4.4.6), expressing $\operatorname{conv} SO(n)$ as the intersection of two spectrahedra. We explain how this works in detail in Section 4.4.3.

Theorem 4.1.3. *The convex hull of $SO(n)$ is a spectrahedron. Explicitly*

$$\operatorname{conv} SO(n) = \left\{ X \in \mathbb{R}^{n\times n} : \begin{bmatrix} 0 & X \\ X^T & 0 \end{bmatrix} \preceq I_{2n},\ \sum_{i,j=1}^{n} A^{(ij)} [RX]_{ij} \preceq (n-2) I_{2^{n-1}} \right\}. \tag{4.1.8}$$

In the special cases $n = 2$ and $n = 3$ we have

$$\operatorname{conv} SO(2) = \left\{ \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \in \mathbb{R}^{2\times 2} : \begin{bmatrix} 1+c & s \\ s & 1-c \end{bmatrix} \succeq 0 \right\} \quad \textit{and} \tag{4.1.9}$$

$$\operatorname{conv} SO(3) = \left\{ X \in \mathbb{R}^{3\times 3} : \sum_{i,j=1}^{3} A^{(ij)}[RX]_{ij} \preceq I_4 \right\} \tag{4.1.10}$$

$$= \left\{ X \in \mathbb{R}^{3\times 3} : \right. \tag{4.1.11}$$

$$\left. \begin{bmatrix} 1-X_{11}-X_{22}+X_{33} & X_{13}+X_{31} & X_{12}-X_{21} & X_{23}+X_{32} \\ X_{13}+X_{31} & 1+X_{11}-X_{22}-X_{33} & X_{23}-X_{32} & X_{12}+X_{21} \\ X_{12}-X_{21} & X_{23}-X_{32} & 1+X_{11}+X_{22}+X_{33} & X_{31}-X_{13} \\ X_{23}+X_{32} & X_{12}+X_{21} & X_{31}-X_{13} & 1-X_{11}+X_{22}-X_{33} \end{bmatrix} \succeq 0 \right\}.$$

We note that the representation of $\operatorname{conv} SO(3)$ described in Sanyal et al. [110, Proposition 4.1] can be obtained from the spectrahedral representation for $\operatorname{conv} SO(3)$ given here by conjugating by a signed permutation matrix, establishing that the two representations are equivalent.

In Section 4.5 we prove that our spectrahedral representations in Theorems 4.1.1, 4.1.2, 4.1.3 are of minimum size. We do so by establishing lower bounds on the minimum size of spectrahedral representations of $SO(n)^\circ$, $\operatorname{conv} SO(n)$ and $O(n)^\circ$ that match the upper bounds given by our constructions.

Theorem 4.1.4. *If $n \geq 1$ the minimum size of a spectrahedral representation of $O(n)^\circ$ is 2^n. If $n \geq 2$ the minimum size of a spectrahedral representation of $SO(n)^\circ$ is 2^{n-1}. If $n \geq 4$ the minimum size of a spectrahedral representation of $\operatorname{conv} SO(n)$ is $2^{n-1}+2n$. The minimum size of a spectrahedral representation of $\operatorname{conv} SO(3)$ is 4.*

Semidefinite representations Given a spectrahedral representation of size m of a convex set C (with the origin in its interior), by applying a straightforward conic duality argument (Lemma 2.4.8 of Chapter 2) we can obtain a semidefinite representation of C°. This representation, however, is usually *not* a spectrahedral representation.

Example 4.1.5. Theorems 4.1.2 and 4.1.4 tell us that the smallest spectrahedral representation of $O(n)^\circ$, the nuclear norm ball, has size 2^n. Yet by dualizing the size $2n$ spectrahedral representation of $\operatorname{conv} O(n)$ (given in Proposition 4.4.7 to follow) we obtain a semidefinite representation of $O(n)^\circ$ of size $2n$

$$O(n)^\circ = \left\{ Z \in \mathbb{R}^{n\times n} : \exists X, Y \ \text{ s.t. } \begin{bmatrix} X & Z \\ Z^T & Y \end{bmatrix} \succeq 0,\ \operatorname{tr}(X) + \operatorname{tr}(Y) = 2 \right\}.$$

This is equivalent to the representation given by Fazel [45] for the nuclear norm ball.

By dualizing, in a similar fashion, the spectrahedral representation of $SO(n)^\circ$ we obtain a representation of conv $SO(n)$ as the projection of a spectrahedron, i.e. a semidefinite representation of conv $SO(n)$. In some situations it may be preferable to use this representation of conv $SO(n)$ rather than the spectrahedral representation in Theorem 4.1.3.

Corollary 4.1.6. *The convex hull of $SO(n)$ can be expressed as a projection of the $2^{n-1} \times 2^{n-1}$ positive semidefinite matrices with trace one as*

$$\operatorname{conv} SO(n) = \left\{ \begin{bmatrix} \langle A^{(11)}, Z\rangle & \langle A^{(12)}, Z\rangle & \cdots & \langle A^{(1n)}, Z\rangle \\ \langle A^{(21)}, Z\rangle & \langle A^{(22)}, Z\rangle & \cdots & \langle A^{(2n)}, Z\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle A^{(n1)}, Z\rangle & \langle A^{(n2)}, Z\rangle & \cdots & \langle A^{(nn)}, Z\rangle \end{bmatrix} : Z \succeq 0,\ \operatorname{tr}(Z) = 1 \right\}.$$

Note that we obtained this by applying Lemma 2.4.8. In doing so we use the fact that there is a point Z_0 satisfying $\operatorname{tr}(Z_0) = 1$, $Z_0 \succeq 0$ and $\langle A^{(ij)}, Z_0\rangle = 0$ for all $1 \leq i, j \leq n$. Indeed one can take $Z_0 = I/2^{n-1}$, since $\operatorname{tr}(A^{(ij)}) = 0$ for all $1 \leq i, j \leq n$, a fact we establish in Lemma 4.7.4 to follow using properties of the linear maps represented by the matrices $A^{(ij)}$.

■ 4.1.2 Related work

That the convex hull of $O(n)$ is a spectrahedron is a classical result. (We give a self-contained proof of this fact in Proposition 4.4.7 to follow.) It was not until recently that Sanyal, Sottile, and Sturmfels [110] established that $O(n)^\circ$ is a spectrahedron by explicitly giving a (non-optimal) size $\binom{2n}{n}$ spectrahedral representation. In the same paper, Sanyal, Sottile, and Sturmfels study numerous $SO(n)$- and $O(n)$-orbitopes considering both convex geometric aspects such as their facial structure and Carathéodory number [59, Definition 2.4], and algebraic aspects such as their algebraic boundary and whether they are spectrahedra. They describe (previously known) spectrahedral representations of conv $SO(2)$ and conv $SO(3)$. The representation for conv $SO(3)$ given in [110, Eq. 4.1] is equivalent to our representation in Theorem 4.1.3, and the representation given in [110, Eq. 4.2] is equivalent to

$$\operatorname{conv} SO(3) = \left\{ \begin{bmatrix} Z_{11}-Z_{22}-Z_{33}+Z_{44} & -2Z_{13}-2Z_{24} & -2Z_{12}+2Z_{34} \\ 2Z_{13}-2Z_{24} & Z_{11}+Z_{22}-Z_{33}-Z_{44} & -2Z_{14}-2Z_{23} \\ 2Z_{12}+2Z_{34} & 2Z_{14}-2Z_{23} & Z_{11}-Z_{22}+Z_{33}-Z_{44} \end{bmatrix} : Z \succeq 0,\ \operatorname{tr}(Z) = 1 \right\}$$

which can be obtained by specializing Corollary 4.1.6. Sanyal, Sottile, and Sturmfels raise the general question of whether conv $SO(n)$ is a spectrahedron for all n (which we answer in the affirmative), and more broadly ask for a classification of the $SO(n)$-orbitopes that are spectrahedra.

Earlier work on orbitopes in the context of convex geometry includes the work of Barvinok and Vershik [10] who consider orbitopes of finite groups in the context of combinatorial optimization, Barvinok and Blekherman [8], who used asymptotic volume computations to show that there are many more non-negative polynomials than sums of squares (among other things), and Longinetti et al. [80] who studied $SO(3)$-orbitopes with a view to applications in protein structure determination. More recently Sinn [99] has studied in detail the algebraic boundary of four-dimensional $SO(2)$-orbitopes as well as the Barvinok-Novik orbitopes [9, 133].

4.1.3 Notation

In this brief section we recall some notation from Chapter 2 and define some notation that is not explicitly defined elsewhere. If $\mathcal{U} \subseteq \mathbb{R}^n$ is a subspace then $\Pi_{\mathcal{U}} : \mathbb{R}^n \to \mathcal{U}$ is the orthogonal projector onto $\mathcal{U}$ and $\Pi_{\mathcal{U}}^* : \mathcal{U} \to \mathbb{R}^n$ is its adjoint. If the subspace in question is the subspace of diagonal matrices $\mathcal{D} \subseteq \mathbb{R}^{n\times n}$ we occasionally also use $\operatorname{diag} := \Pi_{\mathcal{D}}$ and $\operatorname{diag}^* := \Pi_{\mathcal{D}}^*$. We frequently use the matrix $R = \operatorname{diag}^*(1, 1, \ldots, 1, -1) \in \mathbb{R}^{n\times n}$. It could be replaced, throughout, by any orthogonal self-adjoint matrix with determinant -1. We use the shorthand $[n]$ for the set $\{1, 2, \ldots, n\}$ and $\mathcal{I}_{\text{even}}$ for the set of subsets of $[n]$ with even cardinality.

4.1.4 Outline

The remainder of the chapter is organized as follows. In Section 4.2 we describe a problem in satellite attitude estimation that can be reformulated as a semidefinite program using the ideas in this chapter. Section 4.3 focuses on the symmetry properties of $\operatorname{conv} SO(n)$ and $\operatorname{conv} O(n)$, as well as certain convex polytopes that naturally arise when studying these convex bodies. With these preliminaries established, Section 4.4 outlines the main arguments required to establish the correctness of the spectrahedral representations of $SO(n)^\circ$, $O(n)^\circ$, $\operatorname{conv} SO(n)$ and $\operatorname{conv} O(n)$. Details of some of the constructions required for these arguments are deferred to Section 4.7. Section 4.5 establishes lower bounds on the size of spectrahedral representations of $SO(n)^\circ$, $O(n)^\circ$, $\operatorname{conv} SO(n)$ and $\operatorname{conv} O(n)$ as well as a lower bound on the size of equivariant semidefinite representations of $\operatorname{conv} SO(n)$.

Many of the properties of the convex bodies of interest in this chapter are summarized in Table 4.1 which may serve as a useful navigational aid for the reader.

4.2 An illustrative application—joint satellite attitude and spin-rate estimation

In this section we discuss a problem in satellite attitude estimation that can be reformulated as a semidefinite optimization problem using the representation of $SO(n)^\circ$

S	$SO(n)$		$O(n)$	
Definition	$\{X \in \mathbb{R}^{n\times n} : X^T X = I,\, \det(X) = 1\}$		$\{X \in \mathbb{R}^{n\times n} : X^T X = I\}$	
S°	$SO(n)^\circ$		$O(n)^\circ =$ Nuclear norm ball	
Diagonal slice	Polar of parity polytope	(Prop. 4.3.3)	Cross-polytope	(Prop. 4.3.3)
Spectrahedral representation	Size: 2^{n-1} Optimal? Yes	(Thm 4.1.1) (Thm 4.1.4)	Size: 2^n Optimal? Yes	(Thm 4.1.2) (Thm 4.1.4)
Semidefinite representation	Size: 2^{n-1} Optimal? Unknown(Cor. 4.5.5, Q. 4.6.2)		Size: $2n$	(Eg. 4.1.5)
$(S^\circ)^\circ = \operatorname{conv} S$	$\operatorname{conv} SO(n)$		$\operatorname{conv} O(n) =$ Operator norm ball	
Diagonal slice	Parity polytope	(Prop. 4.3.3)	Hypercube	(Prop. 4.3.3)
Spectrahedral representation	Size: $\begin{cases} 2^{n-1} + 2n & n \geq 4 \\ 4 & n = 3 \end{cases}$ Optimal? Yes	(Thm 4.1.3) (Thm 4.1.4)	Size: $2n$ Optimal? Yes	(Prop. 4.4.7) (Thm 4.1.4)
Semidefinite representation	Size: 2^{n-1} Optimal? Unknown(Cor. 4.5.5, Q. 4.6.2)	(Cor. 4.1.6)	Size: $2n$	

Table 4.1: Summary of results related to the convex bodies considered in this chapter.

described in Section 4.1.1. Our aim here is to give a concrete example of situations where the semidefinite representations we describe in this paper arise naturally. The problem of interest is one of estimating the attitude (i.e. orientation) and spin-rate of a spinning satellite, and is a slight generalization of a problem posed recently by Psiaki [97]. We first focus on describing the basic attitude estimation problem in Section 4.2.1, before describing the joint attitude and spin-rate estimation problem in Section 4.2.2. We show how to reformulate the joint attitude and spin-rate estimation problem as a semidefinte optimization problem in Section 4.2.3 (with some proofs deferred to Section 4.8).

■ 4.2.1 Attitude estimation

The *attitude* of a satellite is the element of $SO(3)$ that transforms a reference coordinate system (the *inertial system*) in which, say, the sun is fixed, into a local coordinate system fixed with respect to the satellite's body (the *body system*). We are given unit vectors $x_1, x_2, \ldots, x_T$ (e.g., the alignment of the Earth's magnetic field, directions of landmarks such as the sun or other stars, etc.) in the inertial coordinate system, and noisy measurements $y_1, y_2, \ldots, y_T$ of these directions in the body coordinate system. Let $Q \in SO(3)$ denote the unknown attitude of the satellite. The aim is to estimate (in the maximum likelihood sense) Q given the y_k, the x_k and a description of the measurement noise.

The simplest noise model assumes that each y_k is independent and has a von Mises-Fisher distribution [82] (a natural family of probability distributions on the sphere) with mean Qx_k and concentration parameter κ, i.e. its probability density function is, up to a proportionality constant that does not depend on Q, $p(y_k; Q) \propto \exp(\kappa\langle y_k, Qx_k\rangle)$. Then the maximum likelihood estimate of Q is found by solving

$$\max_{Q\in SO(3)} \sum_{k=1}^{T} \kappa\langle y_k, Qx_k\rangle = \max_{Q\in SO(3)} \langle Q, \kappa \sum_{k=1}^{T} y_k x_k^T\rangle = \max_{Q\in \operatorname{conv} SO(3)} \langle Q, \kappa \sum_{k=1}^{T} y_k x_k^T\rangle. \quad (4.2.1)$$

This is a probabilistic interpretation of a problem known as *Wahba's problem* in the astronautical literature, posed by Grace Wahba in the July 1965 *SIAM Review* problems and solutions section [135, Problem 65-1].

Our spectrahedral representation of $\operatorname{conv} SO(n)$ allows us to express the optimization problem in (4.2.1) as a semidefinite optimization problem. In the astronautical literature it is common to solve this problem via the q-method [69] which involves parameterizing $SO(3)$ in terms of unit quaternions and solving a symmetric eigenvalue problem. Our semidefinite optimization-based formulation could be thought of as a much more flexible generalization of this eigenvalue problem-based approach that works for any n, not just the case $n = 3$.

■ 4.2.2 Joint attitude and spin-rate estimation

A significant benefit of having a semidefinite optimization-based description of a problem (such as Wahba's problem), is that it can lead to semidefinite optimization-based solutions to more complicated related problems by composing semidefinite representations in different ways. An example of this is given by the following generalization of Wahba's problem posed by Psiaki [97].[2]

Consider a satellite rotating at a constant unknown angular velocity ω rad/sample around a known axis (e.g. its major axis). Assume the body coordinate system is chosen so that the rotation is around the axis defined by the first coordinate direction. Then the attitude matrix at the kth sample instant is of the form

$$Q(k) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(k\omega) & -\sin(k\omega) \\ 0 & \sin(k\omega) & \cos(k\omega) \end{bmatrix} Q$$

[2] Psiaki's formulation only considers the $\kappa_2 = 0$ case, where measurements of the spin rate are not considered.

where $Q \in SO(3)$ is the initial attitude. Suppose, now, the satellite *sequentially* obtains measurements $y_0, y_1, \ldots, y_T$ in the body coordinate system of known landmarks in the directions $x_0, x_1, \ldots, x_T$ in the inertial coordinate system. As before assume that the y_k are independent and have von Mises-Fisher distribution with mean $Q(k)x_k$ and concentration parameter κ_1. Furthermore, the satellite obtains a sequence $\omega_1, \omega_2, \ldots, \omega_T$ of noisy measurements of the unknown constant spin rate ω. Suppose the ω_k are independent and each ω_k has a von Mises distribution [82] (a natural distribution for angular-valued quantities) with mean ω and concentration parameter κ_2, i.e., its probability density function (up to a constant independent of ω) is given by $p(\omega_k; \omega) \propto \exp(\kappa_2 \cos(\omega_k - \omega))$. If the ω_k and the y_k are independent then the maximum likelihood estimate of Q and ω can be found by solving

$$\max_{\substack{Q \in SO(3) \\ \omega \in [0, 2\pi)}} \sum_{k=0}^{T} \left\langle y_k, \kappa_1 \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(k\omega) & -\sin(k\omega) \\ 0 & \sin(k\omega) & \cos(k\omega) \end{bmatrix} Q x_k \right\rangle + \kappa_2 \sum_{k=0}^{T} \cos(\omega_k - \omega). \tag{4.2.2}$$

For appropriate collections of matrices $(A_k)_{k=0}^T$, $(B_k)_{k=1}^T$ and scalars a_1 and b_1 that depend on the problem data (i.e. the reference directions $(x_k)_{k=0}^T$, the measured directions $(y_k)_{k=0}^T$, the measured spin-rates $(\omega_k)_{k=0}^T$, and the weights κ_1 and κ_2), the optimization problem (4.2.2) can be rewritten as

$$\max_{\substack{Q \in SO(3) \\ \omega \in [0, 2\pi)}} a_1 \cos(\omega) + b_1 \sin(\omega) + \langle A_0, Q \rangle + \sum_{k=1}^{T} \langle A_k, \cos(k\omega) Q \rangle + \langle B_k, \sin(k\omega) Q \rangle. \tag{4.2.3}$$

This can be thought of as the maximization of a linear functional over

$$\begin{aligned} \mathcal{M}_{3,T} = & \\ & \left\{ (\cos(\omega), \sin(\omega), Q, \cos(\omega)Q, \sin(\omega)Q, \ldots, \cos(T\omega)Q, \sin(T\omega)Q) \in \mathbb{R}^2 \times (\mathbb{R}^{3 \times 3})^{2T} : \right. \\ & \hspace{20em} \left. Q \in SO(3),\ \omega \in [0, 2\pi) \right\}. \end{aligned}$$

In this formulation, the problem data all appear in the linear functional defined by the scalars a_1 and b_1 and the matrices $(A_k)_{k=0}^T$ and $(B_k)_{k=1}^T$.

Following the approach described in Section 2.2 of Chapter 2, we can reformulate this family of problems in terms of semidefinite optimization if we have a semidefinite representation of $\text{conv}(\mathcal{M}_{3,T})$. This is because the optimization problem (4.2.3) is equivalent to the maximization of the same linear functional over $\text{conv}(\mathcal{M}_{3,T})$.

■ 4.2.3 A semidefinite representation of $\operatorname{conv}\mathcal{M}_{3,T}$

To solve optimization problems of the form (4.2.2) using semidefinite optimization, it is sufficient to have a semidefinite representation of $\operatorname{conv}\mathcal{M}_{3,T}$. In this section we present a semidefinite representation of $\operatorname{conv}\mathcal{M}_{3,T}$ of size $4(T+1)$. A representation of this form exists because $SO(3)^\circ$ has a spectrahedral representation of size 4.

To describe this representation concisely, we introduce some additional notation that is used only in this section and Section 4.8. If $W_0, W_1, \ldots, W_T$ are $d\times d$ matrices, define the corresponding $d(T+1)\times d(T+1)$ block Toeplitz matrix by

$$\operatorname{Toep}(W_{-T},\ldots,W_{-1},W_0,W_1,\ldots,W_T) = \begin{bmatrix} W_0 & W_1 & W_2 & \cdots & W_T \\ W_{-1} & W_0 & W_1 & \ddots & \vdots \\ W_{-2} & W_{-1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & W_1 \\ W_{-T} & \cdots & \cdots & W_{-1} & W_0 \end{bmatrix}. \tag{4.2.4}$$

Note that if $\operatorname{Toep}(W_{-T},\ldots,W_{-1},W_0,W_1,\ldots,W_T)$ is symmetric then $W_{-k} = W_k^T$ for $k = 0,1,2,\ldots,T$. If $S_1, S_2, \ldots, S_{2T+1}$ are $d\times d$ matrices define the corresponding $d(T+1)\times d(T+1)$ block Hankel matrix by

$$\operatorname{Hank}(S_1,S_2,\ldots,S_{2T+1}) = \begin{bmatrix} S_1 & S_2 & \cdots & S_T & S_{T+1} \\ S_2 & & & S_{T+1} & S_{T+2} \\ \vdots & & ⋰ & & \vdots \\ S_T & S_{T+1} & & & S_{2T} \\ S_{T+1} & S_{T+2} & \cdots & S_{2T} & S_{2T+1} \end{bmatrix}. \tag{4.2.5}$$

Note that if $\operatorname{Hank}(S_1,S_2,\ldots,S_{2T+1})$ is symmetric then $S_k = S_k^T$ for $k = 1,2,\ldots,2T+1$.

Our semidefinite representation of $\operatorname{conv}\mathcal{M}_{3,T}$ can then be described as follows.

Proposition 4.2.1. *The convex hull of $\mathcal{M}_{3,T}$ has a semidefinite representation of size $4(T+1)$ as*

$$\begin{aligned}\operatorname{conv}\mathcal{M}_{3,T} = \{(\operatorname{tr}(X_1),\operatorname{tr}(Y_1),\mathcal{A}(X_0),\mathcal{A}(X_1),\mathcal{A}(Y_1),\ldots,\mathcal{A}(X_T),\mathcal{A}(Y_T)) :&\\ \operatorname{Toep}(X_T,\ldots,X_1,X_0,X_1,\ldots,X_T)+&\\ \operatorname{Hank}(Y_T,Y_{T-1},\ldots,Y_1,0,-Y_1,\ldots,-Y_{T-1},-Y_T) \succeq 0\}&\end{aligned}$$

where $\mathcal{A} : \mathcal{S}^4 \to \mathbb{R}^{3\times 3}$ *is defined by*

$$\mathcal{A}(Z) = \begin{bmatrix} Z_{11} - Z_{22} - Z_{33} + Z_{44} & -2Z_{13} - 2Z_{24} & -2Z_{12} + 2Z_{34} \\ 2Z_{13} - 2Z_{24} & Z_{11} + Z_{22} - Z_{33} - Z_{44} & -2Z_{14} - 2Z_{23} \\ 2Z_{12} + 2Z_{34} & 2Z_{14} - 2Z_{23} & Z_{11} - Z_{22} + Z_{33} - Z_{44} \end{bmatrix}.$$

We prove this result in Section 4.8. The only fact we use about $SO(3)$ in the proof is that $\operatorname{conv} SO(3)$ has a semidefinite representation of the form $\{\mathcal{A}(Z) : Z \succeq 0,\ \operatorname{tr}(Z) = 1\}$ (see Corollary 4.1.6).

4.3 Basic properties of $\operatorname{conv} SO(n)$ and $\operatorname{conv} O(n)$

In this section we consider the convex bodies $\operatorname{conv} SO(n)$ and $\operatorname{conv} O(n)$ purely from the point of view of convex geometry leaving the discussion of aspects related to their semidefinite representations for Section 4.4. In this section we describe their symmetries, and how the full space $\mathbb{R}^{n\times n}$ of $n \times n$ matrices decomposes with respect to these symmetries, via the (special) singular value decomposition. To a large extent one can characterize $\operatorname{conv} SO(n)$ and $\operatorname{conv} O(n)$ in terms of their intersections with the subspace of diagonal matrices. These diagonal sections are well-known polytopes—the parity polytope and the hypercube respectively. The properties of these diagonal sections are crucial to establishing our spectrahedral representation of $\operatorname{conv} SO(n)$ in Section 4.4.3 and the lower bounds on the size of spectrahedral representations given in Section 4.5.

All of the results in this section are (sometimes implicitly) in the literature in various forms. Here we aim for a brief yet unified presentation to make the chapter as self-contained as possible.

4.3.1 Symmetry and the special singular value decomposition

In this section we describe the symmetries of $\operatorname{conv} O(n)$ and $\operatorname{conv} SO(n)$.

The group $O(n) \times O(n)$ acts on $\mathbb{R}^{n\times n}$ by $(U, V) \cdot X = UXV^T$. This action leaves the set $O(n)$ invariant, and hence leaves the convex bodies $\operatorname{conv} O(n)$ and $O(n)^\circ$ invariant. It is also useful to understand how the ambient space of $n\times n$ matrices decomposes under this group action. Indeed by the well-known singular value decomposition every element $X \in \mathbb{R}^{n\times n}$ can be expressed as $X = U\Sigma V^T = (U, V) \cdot \Sigma$ where $(U, V) \in O(n) \times O(n)$, and Σ is diagonal with $\Sigma_{11} \geq \cdots \geq \Sigma_{nn} \geq 0$. These diagonal elements are the *singular values*. We denote them by $\sigma_i(X) = \Sigma_{ii}$. Note that for most of what follows, we only use the fact that Σ is diagonal, not that its elements can be taken to be non-negative and sorted.

Similarly the group

$$S(O(n) \times O(n)) = \{(U, V) : U, V \in O(n),\ \det(U)\det(V) = 1\}$$

acts on $\mathbb{R}^{n\times n}$ by $(U,V)\cdot X = UXV^T$. This action leaves the sets $SO(n)$ and $SO^-(n)$ invariant, and hence leaves the convex bodies $\operatorname{conv} SO(n)$, $\operatorname{conv} SO^-(n)$, $SO(n)^\circ$, $SO^-(n)^\circ$, $\operatorname{conv} O(n)$ and $O(n)^\circ$ invariant. A variant on the singular value decomposition, known as the *special singular value decomposition* [110] describes how the space of $n\times n$ matrices decomposes under this group action. Indeed every $X\in\mathbb{R}^{n\times n}$ can be expressed as $X = U\tilde{\Sigma}V^T = (U,V)\cdot\tilde{\Sigma}$ where $(U,V)\in S(O(n)\times O(n))$ and $\tilde{\Sigma}$ is diagonal with $\tilde{\Sigma}_{11}\geq\cdots\geq\tilde{\Sigma}_{n-1,n-1}\geq|\tilde{\Sigma}_{nn}|$. These diagonal elements are the *special singular values*. We denote them by $\tilde{\sigma}_i(X) = \tilde{\Sigma}_{ii}$. Again in what follows we typically only use the fact that $\tilde{\Sigma}$ is diagonal for our arguments.

The special singular value decomposition can be obtained from the singular value decomposition. Suppose $X = U\Sigma V^T$ is a singular value decomposition of X so that $(U,V)\in O(n)\times O(n)$. If $\det(U)\det(V) = 1$ this is also a valid special singular value decomposition. Otherwise, if $\det(U)\det(V) = -1$ then $X = UR(R\Sigma)V^T$ gives a decomposition where $(UR,V)\in S(O(n)\times O(n))$ and $R\Sigma$ is again diagonal, but with the last diagonal entry being negative. As such the singular values and special singular values of an $n\times n$ matrix are related by $\sigma_i(X) = \tilde{\sigma}_i(X)$ for $i = 1,2,\ldots,n-1$ and $\tilde{\sigma}_n(X) = \operatorname{sign}(\det(X))\sigma_n(X)$.

The importance of these decompositions of $\mathbb{R}^{n\times n}$ under the action of $O(n)\times O(n)$ and $S(O(n)\times O(n))$ is that they allow us to reduce many arguments, by invariance properties, to arguments about diagonal matrices.

4.3.2 Polytopes associated with $\operatorname{conv} O(n)$ and $\operatorname{conv} SO(n)$

The convex hull of $O(n)$ is closely related to the *hypercube*

$$\mathrm{C}_n = \operatorname{conv}\{x\in\mathbb{R}^n : x_i^2 = 1,\;\; \text{for } i\in[n]\}; \tag{4.3.1}$$

the convex hull of $SO(n)$ is closely related to the *parity polytope*

$$\mathrm{PP}_n = \operatorname{conv}\{x\in\mathbb{R}^n : \textstyle\prod_{i=1}^n x_i = 1,\;\; x_i^2 = 1,\;\; \text{for } i\in[n]\}; \tag{4.3.2}$$

the convex hull of $SO^-(n)$ is closely related to the *odd parity polytope*

$$\mathrm{PP}_n^- = \operatorname{conv}\{x\in\mathbb{R}^n : \textstyle\prod_{i=1}^n x_i = -1,\;\; x_i^2 = 1,\;\; \text{for } i\in[n]\}. \tag{4.3.3}$$

In this section we briefly discuss properties of these polytopes and show that they are the diagonal sections of $\operatorname{conv} O(n)$, $\operatorname{conv} SO(n)$ and $\operatorname{conv} SO^-(n)$ respectively.

Facet descriptions The hypercube has $2n$ facets corresponding to the linear inequality description

$$\mathrm{C}_n = \{x\in\mathbb{R}^n : -1\leq x_i\leq 1 \text{ for } i\in[n]\}. \tag{4.3.4}$$

The parity polytope PP_n has the linear inequality description

$$\mathrm{PP}_n = \left\{ x \in \mathbb{R}^n : -1 \leq x_i \leq 1 \text{ for } i \in [n], \sum_{i \notin I} x_i - \sum_{i \in I} x_i \leq n-2 \text{ for } I \subseteq [n],\ |I| \text{ odd} \right\}. \tag{4.3.5}$$

This description is due to Jeroslow [98] (see, e.g., [32, Theorem 5.3] for a self-contained proof). If $n \geq 4$, all $2n+2^{n-1}$ linear inequalities in (4.3.5) define facets. By symmetry it suffices to check one inequality of each type. Indeed if we remove the inequality $x_1 \leq 1$ then $(n-2, 0, \ldots, 0)$ satisfies all the other inequalities but is not in PP_n (for $n \geq 4$). Similarly if we remove the inequality $-x_1 + x_2 + \cdots + x_n \leq n-2$ then $(-1, 1, \ldots, 1)$ satisfies all the other inequalities but is not in PP_n. In the cases $n = 2$ and $n = 3$, (4.3.5) simplifies to

$$\mathrm{PP}_2 = \left\{ \left[\begin{smallmatrix} x \\ x \end{smallmatrix}\right] \in \mathbb{R}^2 : -1 \leq x \leq 1 \right\} \quad \text{and} \tag{4.3.6}$$

$$\begin{aligned} \mathrm{PP}_3 = \{ x \in \mathbb{R}^3 : & \ x_1 - x_2 + x_3 \leq 1,\ -x_1 + x_2 + x_3 \leq 1, \\ & \ x_1 + x_2 - x_3 \leq 1,\ -x_1 - x_2 - x_3 \leq 1 \}, \end{aligned} \tag{4.3.7}$$

showing that PP_3 has only four facets.

The polar of the hypercube is the *cross-polytope*. We denote it by C_n°. It is clear from (4.3.1) that C_n° has 2^n facets and corresponding linear inequality description

$$\mathrm{C}_n^\circ = \left\{ x \in \mathbb{R}^n : \sum_{i \notin I} x_i - \sum_{i \in I} x_i \leq 1 \quad \text{for } I \subseteq [n] \right\}. \tag{4.3.8}$$

The polar of the parity polytope is denoted by PP_n°. It is clear from (4.3.2) that PP_n° has 2^{n-1} facets and corresponding linear inequality description

$$\mathrm{PP}_n^\circ = \left\{ x \in \mathbb{R}^n : \sum_{i \notin I} x_i - \sum_{i \in I} x_i \leq 1 \quad \text{for } I \subseteq [n],\ |I| \text{ even} \right\}. \tag{4.3.9}$$

Similarly

$$\mathrm{PP}_n^{-\circ} = \left\{ x \in \mathbb{R}^n : \sum_{i \notin I} x_i - \sum_{i \in I} x_i \leq 1 \quad \text{for } I \subseteq [n],\ |I| \text{ odd} \right\}. \tag{4.3.10}$$

To get a sense of the importance of these polytopes for understanding $\operatorname{conv} SO(n)$ it may be instructive to compare (4.3.5) with (4.1.8), (4.3.6) with (4.1.9), (4.3.7) with (4.1.10), and (4.3.9) with (4.1.7).

We conclude the discussion of these polytopes with another description of PP_n.

Lemma 4.3.1. *If $n \geq 3$, the parity polytope can be expressed as*

$$\mathrm{PP}_n = \mathrm{C}_n \cap (n-2) \cdot \mathrm{PP}_n^{-\circ}.$$

If $n = 3$ this simplifies to $\mathrm{PP}_3 = \mathrm{PP}_3^{-\circ}$. If $n = 2$, $\mathrm{PP}_2 = \mathrm{C}_2 \cap \mathrm{span}(1,1)$.

Proof. For the general case, we need only examine the facet descriptions in (4.3.4), (4.3.5), and (4.3.10). If $n = 3$ the result follows by comparing (4.3.7) with (4.3.10). The case $n = 2$ is a restatement of (4.3.6). □

Diagonal projections and sections We now establish the link between the hypercube and the convex hull of $O(n)$, and the parity polytope and the convex hull of $SO(n)$. The key fact that relates the parity polytope and the convex hull of $SO(n)$ is the following celebrated theorem of Horn [1].

Theorem 4.3.2 (Horn)**.** *The projection onto the diagonal of $SO(n)$ is the parity polytope, i.e. $\Pi_{\mathcal{D}}(SO(n)) = \mathrm{PP}_n$.*

Note that we do not need the full strength of Horn's theorem. We only use the corollaries that

$$\Pi_{\mathcal{D}}(\mathrm{conv}\, SO(n)) = \mathrm{conv}\, \Pi_{\mathcal{D}}(SO(n)) = \mathrm{conv}\, \mathrm{PP}_n = \mathrm{PP}_n \quad \text{and} \tag{4.3.11}$$

$$\begin{aligned}\Pi_{\mathcal{D}}(\mathrm{conv}\, SO^-(n)) &= \Pi_{\mathcal{D}}(R \cdot \mathrm{conv}\, SO(n)) \\ &= R \cdot \Pi_{\mathcal{D}}(\mathrm{conv}\, SO(n)) = R \cdot \mathrm{PP}_n = \mathrm{PP}_n^-. \end{aligned}\tag{4.3.12}$$

We are now in a position to establish the main result of this section (Proposition 4.3.3 to follow). It follows directly from Horn's theorem and Lemma 2.6.9 from Chapter 2 that describes the interaction between a convex body, invariant under an orthogonal group action, and the fixed-point subspace of that group action.

Proposition 4.3.3. *Let $\mathcal{D} \subseteq \mathbb{R}^{n\times n}$ denote the subspace of diagonal matrices. Then*

$$\begin{aligned}
\Pi_{\mathcal{D}}(\mathcal{D} \cap \mathrm{conv}\, O(n)) &= \mathrm{C}_n, & \Pi_{\mathcal{D}}(\mathcal{D} \cap O(n)^\circ) &= \mathrm{C}_n^\circ,\\
\Pi_{\mathcal{D}}(\mathcal{D} \cap \mathrm{conv}\, SO(n)) &= \mathrm{PP}_n, & \Pi_{\mathcal{D}}(\mathcal{D} \cap SO(n)^\circ) &= \mathrm{PP}_n^\circ,\\
\Pi_{\mathcal{D}}(\mathcal{D} \cap \mathrm{conv}\, SO^-(n)) &= \mathrm{PP}_n^-, & \Pi_{\mathcal{D}}(\mathcal{D} \cap SO^-(n)^\circ) &= \mathrm{PP}_n^{-\circ}.
\end{aligned}$$

Proof. First note that by (4.3.11) and (4.3.12) we know that $\Pi_{\mathcal{D}}(\mathrm{conv}\, SO(n)) = \mathrm{PP}_n$ and that $\Pi_{\mathcal{D}}(\mathrm{conv}\, SO^-(n)) = \mathrm{PP}_n^-$. Consequently

$$\Pi_{\mathcal{D}}(\mathrm{conv}\, O(n)) = \mathrm{conv}\, \Pi_{\mathcal{D}}(SO(n) \cup SO^-(n)) = \mathrm{conv}\,(\mathrm{PP}_n \cup \mathrm{PP}_n^-) = \mathrm{C}_n. \tag{4.3.13}$$

Let G denote the group of diagonal matrices with non-zero entries in $\{-1,1\}$ under matrix multiplication. Observe that $g \in G$ acts on $n \times n$ matrices by $X \mapsto gXg^T$. This action is orthogonal with respect to the trace inner product on $n \times n$ matries. Furthermore, the fixed-point subspace of the action is precisely $\mathcal{D}$. Since each of $\operatorname{conv} O(n)$, $\operatorname{conv} SO(n)$, $\operatorname{conv} SO^-(n)$ is invariant under conjugation by diagonal sign matrices we can apply Lemma 2.6.9 from Chapter 2. This result tells us that if C is any of $\operatorname{conv} O(n)$, $\operatorname{conv} SO(n)$ or $\operatorname{conv} SO^-(n)$ then

$$\Pi_{\mathcal{D}}(C \cap \mathcal{D}) = \Pi_{\mathcal{D}}(C) \quad \text{and} \quad \Pi_{\mathcal{D}}(C^\circ \cap \mathcal{D}) = \Pi_{\mathcal{D}}(C \cap \mathcal{D})^\circ = \Pi_{\mathcal{D}}(C)^\circ.$$

Combining this with the characterization of $\Pi_{\mathcal{D}}(C)$ for each of these convex bodies from (4.3.11), (4.3.12), and (4.3.13), completes the proof. □

4.4 Spectrahedral representations of $SO(n)^\circ$ and $\operatorname{conv} SO(n)$

This section is devoted to outlining the proofs of Theorems 4.1.1, 4.1.2 and 4.1.3, giving spectrahedral representations of $SO(n)^\circ$, $O(n)^\circ$ and $\operatorname{conv} SO(n)$. For the sake of exposition, we initially focus on $SO(2)^\circ$ as in this case all the ideas are familiar. Low-dimensional coincidences do mean that some issues are simpler in the 2×2 case than in general. After discussing the 2×2 case, in Section 4.4.2 we generalize the argument, deferring some details to Section 4.7. Finally in Section 4.4.3 we construct our spectrahedral representation of $\operatorname{conv} SO(n)$.

4.4.1 The 2×2 case

We begin by giving a spectrahedral representation of $SO(2)^\circ$. We make crucial use of the trigonometric identities $\cos(\theta) = \cos^2(\theta/2) - \sin^2(\theta/2)$ and $\sin(\theta) = 2\cos(\theta/2)\sin(\theta/2)$. Recall that elements of $SO(2)$ have the form

$$\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} = \begin{bmatrix} \cos^2(\frac{\theta}{2}) - \sin^2(\frac{\theta}{2}) & -2\cos(\frac{\theta}{2})\sin(\frac{\theta}{2}) \\ 2\cos(\frac{\theta}{2})\sin(\frac{\theta}{2}) & \cos^2(\frac{\theta}{2}) - \sin^2(\frac{\theta}{2}) \end{bmatrix}$$

and that $(\cos(\theta/2), \sin(\theta/2))$ parameterizes the unit circle in $\mathbb{R}^2$. Hence $SO(2)$ is the image of the unit circle $\{(x_1, x_2) : x_1^2 + x_2^2 = 1\}$ under the quadratic map

$$Q(x_1, x_2) = \begin{bmatrix} x_1^2 - x_2^2 & -2x_1x_2 \\ 2x_1x_2 & x_1^2 - x_2^2 \end{bmatrix}.$$

As such, $Y \in SO(2)^\circ$ if and only if, for all (x_1, x_2) in the unit circle,

$$\begin{aligned}\langle Y, Q(x_1,x_2)\rangle &= \left\langle \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22}\end{bmatrix}, \begin{bmatrix} x_1^2 - x_2^2 & -2x_1x_2 \\ 2x_1x_2 & x_1^2 - x_2^2\end{bmatrix}\right\rangle \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} Y_{11}+Y_{22} & Y_{21}-Y_{12} \\ Y_{21}-Y_{12} & -Y_{11}-Y_{22}\end{bmatrix} \begin{bmatrix} x_1 \\ x_2\end{bmatrix} \leq 1.\end{aligned}$$

This is equivalent to the spectrahedral representation

$$SO(2)^\circ = \left\{ Y : \begin{bmatrix} Y_{11}+Y_{22} & Y_{21}-Y_{12} \\ Y_{21}-Y_{12} & -Y_{11}-Y_{22}\end{bmatrix} \preceq I \right\}$$

which coincides with the $n = 2$ case of Theorem 4.1.1.

To summarize, the main idea of the argument is that we use a parameterization of $SO(2)$ as the image of the unit circle under a quadratic map. This parameterization allows us to rewrite the maximum of a linear functional on $SO(2)$ as the maximum of a quadratic form on the unit circle which can be expressed as a spectrahedral condition.

We note that a very similar argument works in the case $n = 3$ to directly produce the representations of $SO(3)^\circ$ and $\operatorname{conv} SO(3)$ in Theorem 4.1.1 and Corollary 4.1.6 respectively. Indeed the unit quaternion parameterization of rotations gives a parameterization of $SO(3)$ as the image of the unit sphere in $\mathbb{R}^4$ under a quadratic mapping. This allows us to rewrite the maximum of a linear functional on $SO(3)$ as the maximum of a quadratic form on the unit sphere or, equivalently, as a spectrahedral condition.

■ 4.4.2 Outline of the general argument

In this section we outline the argument for general n. We do not define all of the mathematical objects (in particular $\mathrm{Spin}(n)$, $\mathrm{Cl}^0(n)$ and Φ) that play a role in the argument in this section. Instead we just describe their relevant properties, and give definitions in Section 4.7.

For the general case, we first need a quadratic parameterization of $SO(n)$. There is a classical construction of a quadratic map $Q : \mathbb{R}^{2^{n-1}} \to \mathbb{R}^{n\times n}$ and a subset $\mathrm{Spin}(n)$ of the unit sphere in $\mathbb{R}^{2^{n-1}}$ such that $SO(n) = Q(\mathrm{Spin}(n))$. (We summarize this construction in Section 4.7, only discussing those aspects relevant for our argument here.)

Unfortunately, for $n \geq 4$, $\mathrm{Spin}(n)$ is a *strict* subset of the unit sphere in $\mathbb{R}^{2^{n-1}}$, so we cannot simply follow the argument for the $n = 2$ case verbatim. The key difficulty is that we need a spectrahedral characterization of the maximum *over* $\mathrm{Spin}(n)$ of the quadratic form $x \mapsto \langle Y, Q(x)\rangle$ (for arbitrary Y). It is not obvious how to do this when $\mathrm{Spin}(n)$ is a strict subset of the sphere.

We achieve this by showing that, for any Y, the maximum of the quadratic form

$x \mapsto \langle Y, Q(x)\rangle$ over the entire sphere coincides with its maximum over the strict subset $\mathrm{Spin}(n)$ of the sphere (see Proposition 4.4.4, to follow). To establish this we exploit additional structure in $\mathrm{Spin}(n)$ and certain equivariance properties of Q. The specific properties we use are stated in Propositions 4.4.1, 4.4.2, and 4.4.3. We prove these in Section 4.7.

Proposition 4.4.1. *There is a 2^{n-1}-dimensional inner product space*[3]*, $\mathrm{Cl}^0(n)$, a subset $\mathrm{Spin}(n)$ of the unit sphere in $\mathrm{Cl}^0(n)$ and a quadratic map $Q : \mathrm{Cl}^0(n) \to \mathbb{R}^{n\times n}$ such that $Q(\mathrm{Spin}(n)) = SO(n)$.*

From now on fix $\mathrm{Cl}^0(n)$, $\mathrm{Spin}(n)$, and Q that satisfy the previous proposition and are explicitly constructed in Section 4.7. The quadratic mapping Q interacts well with left and right multiplication by elements of $SO(n)$.

Proposition 4.4.2. *If $U, V \in SO(n)$ then there is a corresponding invertible linear map $\Phi_{(U,V)} : \mathrm{Cl}^0(n) \to \mathrm{Cl}^0(n)$ such that for any $x \in \mathrm{Cl}^0(n)$, $UQ(x)V^T = Q(\Phi_{(U,V)}x)$ and $\Phi_{(U,V)}(\mathrm{Spin}(n)) = \mathrm{Spin}(n)$.*

Let $\mathcal{I}_{\text{even}}$ denote the collection of subsets of $[n]$ of even cardinality.

Proposition 4.4.3. *Given any orthonormal basis $u_1, \ldots, u_n$ for $\mathbb{R}^n$, there is a corresponding orthonormal basis $(u_I)_{I\in\mathcal{I}_{even}}$ for $\mathrm{Cl}^0(n)$ such that*

- *$u_I \in \mathrm{Spin}(n)$ for all $I \in \mathcal{I}_{even}$ and*
- *for all $i \in [n]$, if $x = \sum_{I\in\mathcal{I}_{even}} x_I u_I \in \mathrm{Cl}^0(n)$ then*

$$\langle u_i, Q\left(x\right) u_i\rangle = \sum_{I\in\mathcal{I}_{even}} x_I^2 \langle u_i, Q(u_I)u_i\rangle.$$

The following proposition, the crux of our argument, implies that for any $n \times n$ matrix Y, the maximum of the quadratic form $x \mapsto \langle Y, Q(x)\rangle$ over the whole sphere and over the (strict) subset $\mathrm{Spin}(n)$, coincide.

Proposition 4.4.4. *Given any $Y \in \mathbb{R}^{n\times n}$ the quadratic form $x \mapsto \langle Y, Q(x)\rangle$ has a basis of eigenvectors*[4] *that are elements of $\mathrm{Spin}(n)$.*

Proof. Suppose $Y \in \mathbb{R}^{n\times n}$ is arbitrary. Then by the special singular value decomposition Y can be expressed as $Y = U^T DV$ where U and V are in $SO(n)$ and D is diagonal. Then by Proposition 4.4.2

$$\langle Y, Q(x)\rangle = \langle U^T DV, Q(x)\rangle = \langle D, UQ(x)V^T\rangle = \langle D, Q(\Phi_{(U,V)}x)\rangle.$$

[3]By choosing a basis we can identify this with $\mathbb{R}^{2^{n-1}}$ equipped with the standard inner product.

[4]Suppose $x \mapsto M(x) = \langle Mx, x\rangle$ is a quadratic form represented by a self-adjoint linear map M. By the *eigenvectors* of the quadratic form we mean the eigenvectors of the associated linear map M.

Consider the quadratic form $z \mapsto \langle D, Q(z)\rangle$ and let $e_1,\ldots,e_n$ denote the standard basis for $\mathbb{R}^n$. By Proposition 4.4.3 there is a basis $(e_I)_{I\in\mathcal{I}_{\text{even}}}$ such that if $z = \sum_{I\in\mathcal{I}_{\text{even}}} z_I e_I$ then

$$\langle D, Q(z)\rangle = \sum_{i=1}^n D_{ii}\langle e_i, Q(z)e_i\rangle = \sum_{I\in\mathcal{I}_{\text{even}}} z_I^2\left(\sum_{i=1}^n D_{ii}\langle e_i, Q(e_I)e_i\rangle\right).$$

Hence $z \mapsto \langle D, Q(z)\rangle$ has $(e_I)_{I\in\mathcal{I}_{\text{even}}}$ as a basis of eigenvectors. Hence the quadratic form $x \mapsto \langle Y, Q(x)\rangle$ has $\Phi_{(U,V)}^{-1}e_I$ for $I \in \mathcal{I}_{\text{even}}$ as a basis of eigenvectors. Since the e_I are in Spin(n) (by Proposition 4.4.3), $\Phi_{(U,V)}$ is invertible, and $\Phi_{(U,V)}^{-1}$ preserves Spin(n) (by Proposition 4.4.2), we can conclude that the quadratic form $x \mapsto \langle Y, Q(x)\rangle$ has a basis of eigenvectors all of which are elements of Spin(n). □

Assuming Propositions 4.4.1 and 4.4.4 we can prove Theorem 4.1.1 using an embellishment of the same argument we used in the 2×2 case.

Theorem 4.1.1. *The polar of $SO(n)$ is a spectrahedron. Explicitly*

$$SO(n)^\circ = \left\{Y \in \mathbb{R}^{n\times n} : \sum_{i,j=1}^n A^{(ij)}Y_{ij} \preceq I_{2^{n-1}}\right\}$$

where the $2^{n-1}\times 2^{n-1}$ matrices $A^{(ij)}$ are defined in (4.1.6).

Proof. Since the image of Spin(n) under Q is $SO(n)$, an $n\times n$ matrix Y is in $SO(n)^\circ$ if and only if

$$\max_{X\in SO(n)}\langle Y, X\rangle = \max_{x\in \text{Spin}(n)}\langle Y, Q(x)\rangle \le 1.$$

Since Spin(n) is a subset of the unit sphere in $\text{Cl}^0(n)$, we have that

$$\max_{x\in\text{Spin}(n)}\langle Y, Q(x)\rangle \le \max_{\substack{x\in \text{Cl}^0(n)\\ \langle x,x\rangle=1}}\langle Y, Q(x)\rangle.$$

The maximum of the quadratic form $x \mapsto \langle Y, Q(x)\rangle$ over the unit sphere in $\text{Cl}^0(n)$ occurs at any eigenvector corresponding to the largest eigenvalue of the quadratic form. By Proposition 4.4.4 we can always find such an eigenvector in Spin(n), establishing that

$$\max_{x\in\text{Spin}(n)}\langle Y, Q(x)\rangle = \max_{\substack{x\in \text{Cl}^0(n)\\ \langle x,x\rangle=1}}\langle Y, Q(x)\rangle.$$

Hence $Y \in SO(n)^\circ$ if and only if for all $x \in \text{Cl}^0(n)$ such that $\langle x, x\rangle = 1$,

$$\langle Y, Q(x)\rangle = \sum_{i,j=1}^n Y_{ij}\langle e_i, Q(x)e_j\rangle \le 1. \tag{4.4.1}$$

In Section 4.7.4 we explicitly describe a choice of coordinates for $\mathrm{Cl}^0(n)$ such that the matrix representing the quadratic form $x \mapsto \langle e_i, Q(x)e_j\rangle$ in those coordinates is precisely the matrix $A^{(ij)}$ defined in (4.1.6). Hence (4.4.1) is equivalent to the spectrahedral representation given in Theorem 4.1.1. □

Remark 4.4.5. We briefly describe a more geometric dual interpretation of the arguments that establish Theorem 4.1.1. Throughout this remark let $S = \{x \in \mathrm{Cl}^0(n) : \langle x, x\rangle = 1\}$ be the unit sphere in $\mathrm{Cl}^0(n)$. We have seen that there is a quadratic map Q such that $SO(n) = Q(\mathrm{Spin}(n)) \subseteq Q(S)$ with the inclusion being strict for $n \geq 4$. The remainder of the proof of Theorem 4.1.1 shows, from this viewpoint, that $\mathrm{conv}\, SO(n) = \mathrm{conv}\, Q(\mathrm{Spin}(n)) = \mathrm{conv}\, Q(S)$, i.e. all the points in S that are not in $\mathrm{Spin}(n)$ are mapped by Q inside the convex hull of $Q(\mathrm{Spin}(n))$. One may wonder whether $Q(S) = \mathrm{conv}\, SO(n)$, i.e. whether the image of the sphere under Q is actually convex. This is not the case—already for $n = 2$ we can see that $Q(S) = SO(2) \neq \mathrm{conv}\, SO(2)$.

It is now straightforward to prove Theorem 4.1.2, giving a spectrahedral representation of $O(n)^\circ$ of size 2^n. We restate the result here for convenience.

Theorem 4.1.2. *The polar of $O(n)$ is a spectrahedron. Explicitly*

$$O(n)^\circ = \left\{ Y \in \mathbb{R}^{n\times n} : \sum_{i,j=1}^{n} A^{(ij)} Y_{ij} \preceq I_{2^{n-1}},\ \sum_{i,j=1}^{n} A^{(ij)} [RY]_{ij} \preceq I_{2^{n-1}} \right\}.$$

where $R = \mathrm{diag}^*(1, 1, \ldots, 1, -1)$.

Proof. Since $O(n)^\circ = SO(n)^\circ \cap SO^-(n)^\circ$ (see (4.1.3)) and we have already constructed a spectrahedral representation of $SO(n)^\circ$, it remains to give a spectrahedral representation of $SO^-(n)^\circ$. Since $SO^-(n) = R \cdot SO(n)$, it follows that $Y \in SO^-(n)^\circ$ if and only if $\langle Y, RX\rangle = \langle RY, X\rangle \leq 1$ for all $X \in SO(n)$. Hence $Y \in SO^-(n)^\circ$ if and only if $RY \in SO(n)^\circ$.

The stated spectrahedral representation of $O(n)^\circ$ of size 2^n follows from these observations and Theorem 4.1.1. □

■ 4.4.3 A spectrahedral representation of $\mathrm{conv}\, SO(n)$

In this section we give a spectrahedral representation of $\mathrm{conv}\, SO(n)$ using a description of $\mathrm{conv}\, SO(n)$ which is inherited from the corresponding description of the parity polytope.

Proposition 4.4.6. *If $n \geq 3$, the convex hull of $SO(n)$ can be expressed as*

$$\mathrm{conv}\, SO(n) = (\mathrm{conv}\, O(n)) \cap (n-2)SO^-(n)^\circ.$$

If $n = 3$ this simplifies to $\operatorname{conv} SO(3) = SO^-(3)^\circ$*. In the case $n = 2$,*

$$\operatorname{conv} SO(2) = (\operatorname{conv} O(2)) \cap \operatorname{span}\left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \right\}.$$

Proof. Suppose $X \in \mathbb{R}^{n\times n}$ is arbitrary and $n \geq 3$. By the special singular value decomposition $X = U\tilde{\Sigma}V^T$ where $(U, V) \in S(O(n) \times O(n))$ and $\tilde{\Sigma} = \operatorname{diag}^*(\tilde{\sigma})$ is diagonal. Then since $SO(n)$ is invariant under the action of $S(O(n) \times O(n))$, it follows that $X \in \operatorname{conv} SO(n)$ if and only if $\tilde{\Sigma} \in (\operatorname{conv} SO(n)) \cap \mathcal{D}$. Similarly since $\operatorname{conv} O(n)$ and $SO^-(n)^\circ$ are invariant under the action of $S(O(n) \times O(n))$, it follows that $X \in (\operatorname{conv} O(n)) \cap (n-2)SO^-(n)^\circ$ if and only if $\tilde{\Sigma} \in (\operatorname{conv} O(n)) \cap \mathcal{D}$ and $\tilde{\Sigma} \in (n-2)SO^-(n)^\circ \cap \mathcal{D}$.

Since the diagonal section of $\operatorname{conv} SO(n)$ is the parity polytope, $X \in \operatorname{conv} SO(n)$ if and only if $\tilde{\sigma} \in \mathrm{PP}_n$. Since the diagonal section of $\operatorname{conv} O(n)$ is the hypercube, $\tilde{\sigma} \in \mathrm{C}_n$ if and only if $\tilde{\Sigma} \in (\operatorname{conv} O(n)) \cap \mathcal{D}$. Since the diagonal section of $SO^-(n)^\circ$ is $\mathrm{PP}_n^{-\circ}$, $\tilde{\sigma} \in (n-2)\mathrm{PP}_n^{-\circ}$ if and only if $\tilde{\Sigma} \in (n-2)SO^-(n)^\circ \cap \mathcal{D}$.

Finally we use the fact that $\mathrm{PP}_n = \mathrm{C}_n \cap (n-2)\mathrm{PP}_n^{-\circ}$ (see Lemma 4.3.1). Then $X \in \operatorname{conv} SO(n)$ if and only if $\tilde{\sigma} \in \mathrm{PP}_n$ which occurs if and only if $\tilde{\sigma} \in \mathrm{C}_n$ and $\tilde{\sigma} \in (n-2)\mathrm{PP}_n^{-\circ}$, which occurs if and only if $X \in (\operatorname{conv} O(n)) \cap (n-2)SO^-(n)^\circ$.

In the case $n = 3$ the description $\mathrm{PP}_n = \mathrm{C}_n \cap (n-2)\mathrm{PP}_n^{-\circ}$ simplifies to $\mathrm{PP}_3 = \mathrm{PP}_3^{-\circ}$. The corresponding simplification propagates through the above argument to give $\operatorname{conv} SO(3) = SO^-(3)^\circ$. The result in the case $n = 2$ follows from the same argument but using the description $\mathrm{PP}_2 = \mathrm{C}_2 \cap \operatorname{span}(1, 1)$ and the fact that $\tilde{\sigma} \in \operatorname{span}(1, 1)$ if and only if $X \in \operatorname{span}\left\{\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]\right\}$. □

Since the description of $\operatorname{conv} SO(n)$ in Proposition 4.4.6 involves $\operatorname{conv} O(n)$, we first give the well-known spectrahedral representation of $\operatorname{conv} O(n)$.

Proposition 4.4.7. *The convex hull of $O(n)$ is a spectrahedron. An explicit spectrahedral representation of size $2n$ is given by*

$$\operatorname{conv} O(n) = \left\{ X \in \mathbb{R}^{n\times n} : \begin{bmatrix} 0 & X \\ X^T & 0 \end{bmatrix} \preceq I_{2n} \right\}. \tag{4.4.2}$$

Proof. Let $Q \in O(n)$ be arbitrary. Then since $Q^TQ = I_n$ it follows that

$$\begin{bmatrix} I_n & -Q \\ -Q^T & I_n \end{bmatrix} = \begin{bmatrix} I_n \\ -Q^T \end{bmatrix} \begin{bmatrix} I_n & -Q \end{bmatrix} \succeq 0$$

and so Q is an element of the right hand side of (4.4.2). Since the right hand side of (4.4.2) is convex, it follows that $\operatorname{conv} O(n) \subseteq \left\{X \in \mathbb{R}^{n\times n} : \left[\begin{smallmatrix} 0 & X \\ X^T & 0 \end{smallmatrix}\right] \preceq I_{2n}\right\}$.

For the reverse inclusion, suppose X is an element of the right hand side of (4.4.2). By the singular value decomposition there is a diagonal matrix Σ such that $X = U\Sigma V^T$ where $U, V \in O(n)$. Conjugating by the orthogonal matrix $\left[\begin{smallmatrix} U^T & 0 \\ 0 & V^T \end{smallmatrix}\right]$ we see that

$$\begin{bmatrix} 0 & X \\ X^T & 0 \end{bmatrix} \preceq I_{2n} \iff \begin{bmatrix} 0 & \Sigma \\ \Sigma & 0 \end{bmatrix} \preceq I_{2n}$$

which is equivalent to $-1 \leq \Sigma_{ii} \leq 1$ for $i \in [n]$. Since $\Pi_{\mathcal{D}}(\mathcal{D} \cap \operatorname{conv} O(n))$ is the hypercube it follows that $\Sigma \in \mathcal{D} \cap \operatorname{conv} O(n)$ and so that $U\Sigma V^T \in \operatorname{conv} O(n)$. $\square$

We now restate (omitting the explicit description of $\operatorname{conv} SO(3)$) and prove Theorem 4.1.3.

Theorem 4.1.3. *The convex hull of $SO(n)$ is a spectrahedron. Explicitly*

$$\operatorname{conv} SO(n) = \left\{ X \in \mathbb{R}^{n\times n} : \begin{bmatrix} 0 & X \\ X^T & 0 \end{bmatrix} \preceq I_{2n},\ \sum_{i,j=1}^{n} A^{(ij)}[RX]_{ij} \preceq (n-2)I_{2^{n-1}} \right\}.$$

In the special cases $n = 2$ and $n = 3$ we have

$$\operatorname{conv} SO(2) = \left\{ \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \in \mathbb{R}^{2\times 2} : \begin{bmatrix} 1+c & s \\ s & 1-c \end{bmatrix} \succeq 0 \right\} \quad \textit{and}$$

$$\operatorname{conv} SO(3) = \left\{ X \in \mathbb{R}^{3\times 3} : \sum_{i,j=1}^{3} A^{(ij)}[RX]_{ij} \preceq I_4 \right\}.$$

Proof. Since we now have a spectrahedral representation of $\operatorname{conv} O(n)$ (from (4.4.2)) and of $SO^-(n)^\circ$ (from the proof of Theorem 4.1.2), by Proposition 4.4.6 their intersection gives the spectrahedral representation of $\operatorname{conv} SO(n)$ valid for $n \geq 3$. In the case $n = 3$ Proposition 4.4.6 tells us that $\operatorname{conv} SO(3) = SO^-(3)^\circ$ giving the stated simplification (which can be expressed explicitly as in (4.1.11) by using the definition of the $A^{(ij)}$ in (4.1.6)). In the case $n = 2$, from Proposition 4.4.6 we have that

$$\operatorname{conv} SO(2) = \left\{ \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \in \mathbb{R}^{2\times 2} : \begin{bmatrix} 1 & 0 & -c & s \\ 0 & 1 & -s & -c \\ -c & -s & 1 & 0 \\ s & -c & 0 & 1 \end{bmatrix} \succeq 0 \right\}.$$

This is still a spectrahedral representation of size 4, but the constraint has symmetry—it is invariant under simultaneously reversing the order of the rows and columns—

suggesting that it can be block diagonalized [50]. Under the change of coordinates

$$\frac{1}{2}\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & -1 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 & -c & s \\ 0 & 1 & -s & -c \\ -c & -s & 1 & 0 \\ s & -c & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & -1 & 0 \end{bmatrix}^T = \begin{bmatrix} 1+c & s & 0 & 0 \\ s & 1-c & 0 & 0 \\ 0 & 0 & 1+c & s \\ 0 & 0 & s & 1-c \end{bmatrix} \tag{4.4.3}$$

we see that the size 4 spectrahedral representation in (4.4.3) is actually two copies of the same size 2 representation, giving the stated result. □

4.5 Lower bounds on the size of representations

4.5.1 Spectrahedral representations

Whenever a convex set has a polyhedral section, we can immediately obtain a simple lower bound on the possible size of a spectrahedral representation of that convex set in terms of the number of facets of that polyhedron. The bound is based on the following result of Ramana [100, Corollary 2.5].

Lemma 4.5.1. *If $P \subseteq \mathbb{R}^p$ is a polyhedron with f facets and P has a spectrahedral representation of size m, then $m \geq f$.*

The following combines Ramana's result with the simple fact that restricting a spectrahedral representation of C to an affine subspace U gives a spectrahedral representation of $C \cap U$ of the same size.

Lemma 4.5.2. *Suppose $C \subseteq \mathbb{R}^n$ has a spectrahedral representation of size m. If $U \subseteq \mathbb{R}^n$ is an affine subspace and $C \cap U$ is a polyhedron with f facets, then $m \geq f$.*

Proof. Parameterize the subspace U as $U = \{Ax + b : x \in \mathbb{R}^p\}$ where $A \in \mathbb{R}^{n\times p}$ and $b \in \mathbb{R}^n$. Let C have a spectrahedral representation $C = \{x : \sum_{i=1}^n A^{(i)}x_i + A^{(0)} \succeq 0\}$ of size m, so the symmetric matrices $A^{(i)}$ are $m \times m$. Let $B^{(j)} = \sum_{i=1}^n A^{(i)}A_{ij}$ for $j = 1, 2, \ldots, p$ and let $B^{(0)} = A^{(0)} + \sum_{i=1}^n A^{(i)}b_i$. Then $C \cap U$ has a spectrahedral representation of size m given by $C \cap U = \{x \in \mathbb{R}^p : \sum_{j=1}^p B^{(j)}x_j + B^{(0)} \succeq 0\}$. Since $C \cap U$ has f facets, it follows from Ramana's result (Lemma 4.5.1) that $m \geq f$. □

Remarkably this simple technique allows us to establish that our spectrahedral representations are of minimum size.

Theorem 4.1.4. *If $n \geq 1$, the minimum size of a spectrahedral representation of $O(n)^\circ$ is 2^n. If $n \geq 2$, the minimum size of a spectrahedral representation of $SO(n)^\circ$ is 2^{n-1}. If $n \geq 4$, the minimum size of a spectrahedral representation of* $\operatorname{conv} SO(n)$ *is $2^{n-1}+2n$. The minimum size of a spectrahedral representation of* $\operatorname{conv} SO(3)$ *is 4.*

Proof. The diagonal slice of $O(n)^\circ$ is the cross-polytope, which (for $n \geq 1$) has 2^n facets. Hence, for $n \geq 1$, any spectrahedral representation of $O(n)^\circ$ has size at least 2^n. The diagonal slice of $SO(n)^\circ$ is the polar of the parity polytope, which (for $n \geq 2$) has 2^{n-1} facets. Hence, for $n \geq 2$, any spectrahedral representation of $SO(n)^\circ$ has size at least 2^{n-1}. The diagonal slice of $\operatorname{conv} SO(n)$ is the parity polytope, which for $n \geq 4$ has $2^{n-1} + 2n$ facets, and for $n = 3$ has 4 facets. It follows that any spectrahedral representation of $\operatorname{conv} SO(n)$ has size at least $2^{n-1} + 2n$ for $n \geq 4$ and size at least 4 for $n = 3$. □

The spectrahedral representations we construct in Section 4.4 achieve these lower bounds and so are of minimum size.

■ 4.5.2 Equivariant semidefinite representations

As is established in Theorem 4.1.4, our spectrahedral representations are necessarily of exponential size. While they are useful in practice for very small n (such as the physically relevant $n = 3$ case), this is not the case for larger n. Recall from Section 2.4 of Chapter 2 that if C is a spectrahedron, it may be possible to give a much smaller semidefinite representation of C, i.e. a description of C as the image of a spectrahedron under a linear map.

It is straightforward to show that if C has a semidefinite representation of size m, then C° also has a semidefinite representation of size m [56]. This simple observation already yields examples of convex bodies for which there is an exponential gap between the size of the smallest spectrahedral representation and the size of the smallest semidefinite representation. For instance, as demonstrated in Example 4.1.5, the smallest possible spectrahedral representation of $O(n)^\circ$ has size 2^n and yet it has a semidefinite representation of size $2n$. Since the spectrahedral representations of both $\operatorname{conv} SO(n)$ and $SO(n)^\circ$ are large, a similar argument cannot yield polynomial-sized semidefinite representations for $\operatorname{conv} SO(n)$.

Equivariant semidefinite representations Since the convex sets of interest in this chapter are highly symmetric, we can try to understand whether there are small semidefinite representations of these convex sets that respect their symmetries. Recall that we call such symmetry-respecting semidefinite representations, equivariant semidefinite representations (see Definition 2.6.10 from Chapter 2).

In the remainder of this section we show that any semidefinite representation of $\operatorname{conv} SO(n)$ that is equivariant with respect to the action of $S(O(n) \times O(n))$, must have size exponential in n. The argument works by showing that from any semidefinite representations of $\operatorname{conv} SO(n)$ that is equivariant with respect to the action of $S(O(n) \times O(n))$ we can construct a semidefinite representation of the parity polytope that is equivariant with respect to a certain group action on $\mathbb{R}^n$. We then apply a recent result that gives an exponential lower bound on the size of appropriately equivariant semidefinite representation of the parity polytope.

For convenience we slightly restate, in more concrete language, the definition of an equivariant semideinite representation (Definition 2.6.10 from Section 2.6).

Definition 4.5.2. Let $C \subseteq \mathbb{R}^n$ be a convex body invariant under the action of a group G by linear transformations. Assume $C = \pi(L \cap \mathcal{S}_+^m)$ is a semidefinite representation of C of size m. The lift is called *G-equivariant* if there is a group homomorphism $\rho : G \to GL(m)$ such that

$$\begin{aligned} \rho(g)X\rho(g)^T &\in L && \text{for all } g \in G \text{ and all } X \in L \quad \text{and} \\ \pi(\rho(g)X\rho(g)^T) &= g \cdot \pi(X) && \text{for all } g \in G \text{ and all } X \in L \cap \mathcal{S}_+^m. \end{aligned} \tag{4.5.1}$$

Let Γ_{parity} denote the symmetry group of the parity polytope. Here Γ_{parity} can be thought of concretely as the group of evenly signed permutation matrices—signed permutation matrices where there are an even number of entries that take the value -1. These act on $\mathbb{R}^n$ by matrix multiplication.

In the present setting we are interested in two particular cases of equivariant semidefinite representations: $S(O(n) \times O(n))$-equivariant representations of $\operatorname{conv} SO(n)$, and Γ_{parity}-equivariant representations of the parity polytope. We show in Proposition 4.5.3 to follow that if we could construct a semidefinite representation of $\operatorname{conv} SO(n)$ that is $S(O(n) \times O(n))$-equivariant, then we could construct a semidefinite representation of PP_n that is Γ_{parity}-equivariant. This is a useful connection to make because Γ_{parity}-equivariant semidefinite representations have been studied in [43]. In particular they are necessarily of exponential size in n (see Theorem 4.5.4 to follow).

Proposition 4.5.3. *If* $\operatorname{conv} SO(n)$ *has an equivariant semidefinite representation size* m *then* PP_n *has an equivariant semidefinite representation size* m.

Proof. Suppose $\operatorname{conv} SO(n) = \pi(L \cap \mathcal{S}_+^m)$ is a $S(O(n) \times O(n))$-equivariant semidefinite representation $\operatorname{conv} SO(n)$ of size m and let $\rho : S(O(n) \times O(n)) \to GL(m)$ be the associated homomorphism. Since the projection of $\operatorname{conv} SO(n)$ onto the subspace of diagonal matrices is PP_n (Theorem 4.3.2) it follows that

$$\mathrm{PP}_n = (\Pi_{\mathcal{D}} \circ \pi)(L \cap \mathcal{S}_+^m)$$

is a semidefinite representation PP_n of size m. It remains to show that this semidefinite representation of PP_n is Γ_{parity}-equivariant. In other words we need to construct a homomorphism $\tilde{\rho} : \Gamma_{\text{parity}} \to GL(m)$ satisfying the requirements of Definition 4.5.2.

First observe that any element of Γ_{parity} can be uniquely expressed as as DP where D is a diagonal sign matrix with determinant one, and P is a permutation matrix. Furthermore, note that if D_1P_1 and D_2P_2 are elements of Γ_{parity}, then

$$(D_1P_1)(D_2P_2) = (D_1P_1D_2P_1^T)(P_1P_2)$$

gives a factorization into a diagonal sign matrix $D_1P_1D_2P_1^T$ and a permutation matrix P_1P_2. Hence define $\phi : \Gamma_{\text{parity}} \to S(O(n) \times O(n))$ by $\phi(DP) = (DP, P)$. Observe that this is a homomorphism because

$$\begin{aligned}\phi((D_1P_1)(D_2P_2)) &= \phi((D_1P_1D_2P_1^T)(P_1P_2)) \\ &= ((D_1P_1)(D_2P_2), P_1P_2) = \phi(D_1P_1) \cdot \phi(D_2P_2).\end{aligned}$$

Define a homomorphism $\tilde{\rho} : \Gamma_{\text{parity}} \to GL(m)$ by $\tilde{\rho} = \rho \circ \phi$. For any symmetric matrix X it is the case that $DP \cdot \Pi_{\mathcal{D}}(X) = \Pi_{\mathcal{D}}(DPXP^T)$. Hence the following establishes that the lift is Γ_{parity}-equivariant:

$$\begin{aligned}DP \cdot \Pi_{\mathcal{D}}(\pi(X)) &= \Pi_{\mathcal{D}}(DP\pi(X)P^T) \\ &= \Pi_{\mathcal{D}}(\phi(DP) \cdot \pi(X)) \\ &\overset{*}{=} \Pi_{\mathcal{D}}(\pi(\rho(\phi(DP))X\rho(\phi(DP))^T)) \\ &= \Pi_{\mathcal{D}}(\pi(\tilde{\rho}(DP)X\tilde{\rho}(DP)^T)) \quad \text{by the definition of } \tilde{\rho}\end{aligned}$$

where the equality marked with an asterisk holds because the lift of $\operatorname{conv} SO(n)$ is equivariant. $\square$

The following lower bound on the size of Γ_{parity}-equivariant semidefinite representations of the parity polytope is one of the main results of [43].

Theorem 4.5.4. *If $n \geq 8$, any Γ_{parity}-equivariant semidefinite representation of PP_n must have size at least $\binom{n}{\lceil \frac{n}{4} \rceil}$.*

Combining Proposition 4.5.3 with Proposition 4.5.4 we obtain the following exponential lower bound on the size of any equivariant semidefinite representation of $\operatorname{conv} SO(n)$.

Corollary 4.5.5. *If $n \geq 8$, any $S(O(n) \times O(n))$-equivariant semidefinite representation of $\operatorname{conv} SO(n)$ must have size at least $\binom{n}{\lceil \frac{n}{4} \rceil}$.*

■ 4.6 Summary and open questions

In this chapter we have constructed minimum size spectrahedral representations for the convex hull of $SO(n)$ and its polar. We have also constructed a minimum-size spectrahedral representation of $O(n)^\circ$ (the nuclear norm ball). We conclude the chapter by discussing some natural questions raised by our results.

■ 4.6.1 Doubly spectrahedral convex sets

We have seen that both the convex hull of $SO(n)$ and its polar are spectrahedra. The same is true of the convex hull of $O(n)$ (the operator norm ball) and its polar (the nuclear norm ball), as established by Sanyal et al. [110, Corollary 4.9]. This is a very special phenomenon—the polar of a spectrahedron is not, in general, a spectrahedron. For example, the intersection of the second-order cone $\{(x, y, z) : z \geq \sqrt{x^2 + y^2}\}$ and the non-negative orthant is a spectrahedron, but its polar has non-exposed faces and so is not a spectrahedron [101] (see Section 2.4).

If a convex set C and its polar are both spectrahedra, we say that C is a *doubly spectrahedral* convex set. Apart from $\operatorname{conv} O(n)$ and $\operatorname{conv} SO(n)$, two distinct families of doubly spectrahedral convex sets are the following:

Polyhedra Every polyhedron is a spectrahedron, and the polar of a polyhedron is again a polyhedron. Hence polyhedra are doubly spectrahedral.

Homogeneous cones A convex cone K is *homogeneous* if the automorphism group (see Section 2.6) of K acts transitively on the interior of K. Using Vinberg's classification of homogeneous cones in terms of T-algebras [40], Chua gave spectrahedral representations for all homogeneous cones [23]. Furthermore, K is homogeneous if and only its dual cone $K^* = -K^\circ$ is homogeneous [40, Proposition 9]. From these two observations it follows that any homogeneous cone is doubly spectrahedral.

We have seen that the doubly spectrahedral convex sets are a strict subset of all spectrahedra that includes all polyhedra, all homogeneous convex cones, and $\operatorname{conv} O(n)$ and $\operatorname{conv} SO(n)$.

Problem 4.6.1. *Characterize doubly spectrahedral convex sets.*

■ 4.6.2 Non-equivariant semidefinite representations

In Section 4.5 we showed that our spectrahedral representations of $\operatorname{conv} SO(n)$ and $SO(n)^\circ$ are necessarily of exponential size and that any $S(O(n) \times O(n))$-equivariant semidefinite representation of $\operatorname{conv} SO(n)$ must also have exponential size. Our lower bound on the size of $S(O(n)\times O(n))$-equivariant semidefinite representation of $\operatorname{conv} SO(n)$ used the fact that any Γ_{parity}-equivariant semidefinite representation of the parity polytope has exponential size. Nevertheless, the parity polytope is known to have a semidefinite representation of size $4(n-1)$ (in fact it is a description as a projection of a polytope with $4(n-1)$ facets) [26, Section 2.6.3] that is *not* Γ_{parity}-equivariant. It is quite possible that by appropriately breaking symmetry we can find a small semidefinite representation of $\operatorname{conv} SO(n)$.

Question 4.6.2. *Does* $\operatorname{conv} SO(n)$ *have a semidefinite representation with size polynomial in* n*?*

4.7 Clifford algebras and $\operatorname{Spin}(n)$

In this section we describe and establish the key properties of the quadratic mapping Q from Proposition 4.4.1 that underlies our spectrahedral representation of $SO(n)^\circ$ given in Theorem 4.1.1. The mapping Q is most naturally described in terms of an algebraic structure known as a Clifford algebra, which generalizes some properties of complex numbers and quaternions. The first part of this section is devoted to describing the basic properties of Clifford algebras we require. In Section 4.7.2 we define the set $\operatorname{Spin}(n)$ and establish some of its properties. In Section 4.7.3 we describe the mapping Q, and establish Propositions 4.4.1, 4.4.2, and 4.4.3. Section 4.7.4 gives explicit constructions of the matrices $A^{(ij)}$ appearing in our spectrahedral representations.

Many of the constructions and properties we describe here are standard and can be found, for example, in [79, 41]. We highlight those aspects of the development that are novel as they arise.

4.7.1 Clifford algebras

The Clifford algebra $\mathrm{Cl}(n)$ is the associative algebra[5] (over the reals) with generators $e_1, e_2, \ldots, e_n$ and relations

$$e_i^2 = -\mathbf{1} \quad \text{and} \quad e_ie_j = -e_je_i \quad \text{for } i \neq j. \tag{4.7.1}$$

Here $\mathbf{1}$ denotes the multiplicative identity in the algebra.

Standard basis As a real vector space $\mathrm{Cl}(n)$ has dimension 2^n. A basis for $\mathrm{Cl}(n)$ is given by all elements of the form

$$e_I := e_{i_1}e_{i_2}\cdots e_{i_k}$$

where $I = \{i_1, i_2, \ldots, i_k\}$ is a subset of $[n]$ and $i_1 < i_2 < \cdots < i_k$. By convention $e_\emptyset := \mathbf{1}$. Let us call $(e_I)_{I\subseteq[n]}$ the *standard basis* for $\mathrm{Cl}(n)$. With respect to this basis

[5] An associative algebra is a vector space equipped with an associative bilinear product. That the generators and relations in (4.7.1) define an associative algebra that is unique up to isomorphism follows because it can be realized as a quotient of the tensor algebra (see, e.g., [41, Definition 9.4]).

we can think of an arbitrary element $x \in \mathrm{Cl}(n)$ as

$$x = \sum_{I \subseteq [n]} x_I e_I$$

where the $x_I \in \mathbb{R}$. We equip $\mathrm{Cl}(n)$ with the inner product $\langle x, y \rangle = \sum_{I \subseteq [n]} x_I y_I$. Clearly the standard basis is orthonormal with respect to this inner product.

Left and right multiplication Any element $x \in \mathrm{Cl}(n)$ acts linearly on $\mathrm{Cl}(n)$ by left multiplication and by right multiplication. In other words, given $x \in \mathrm{Cl}(n)$ there are linear maps $\lambda_x, \rho_x : \mathrm{Cl}(n) \to \mathrm{Cl}(n)$ defined by $\lambda_x(y) = xy$ and $\rho_x(y) = yx$ for all $y \in \mathrm{Cl}(n)$.

Conjugation A straightforward computation shows that with respect to the inner product on $\mathrm{Cl}(n)$, the adjoint of left multiplication by e_i is left multiplication by $-e_i$, i.e., $\lambda^*_{e_i} = \lambda_{-e_i}$. Similarly the adjoint of right multiplication by e_i is right multiplication by $-e_i$. In fact, it is the case that for any $x \in \mathrm{Cl}(n)$ there is $\overline{x} \in \mathrm{Cl}(n)$ such that $\lambda^*_x = \lambda_{\overline{x}}$ and $\rho^*_x = \rho_{\overline{x}}$. To see this define a *conjugation* map $x \mapsto \overline{x}$ on the standard basis by

$$\overline{e_I} = (-1)^{|I|} e_{i_k} \cdots e_{i_2} e_{i_1} \quad \text{where } I = \{i_1, i_2, \dots, i_k\}$$

and extend by linearity. We use this conjugation map repeatedly in the sequel, usually via the relations

$$\langle xy, z \rangle = \langle \lambda_x y, z \rangle = \langle y, \lambda^*_x z \rangle = \langle y, \lambda_{\overline{x}} z \rangle = \langle y, \overline{x} z \rangle \tag{4.7.2}$$

and

$$\langle yx, z \rangle = \langle \rho_x y, z \rangle = \langle y, \rho^*_x z \rangle = \langle y, \rho_{\overline{x}} z \rangle = \langle y, z\overline{x} \rangle. \tag{4.7.3}$$

Copy of $\mathbb{R}^n$ in $\mathrm{Cl}(n)$ Throughout this appendix, we use the notation $\mathbb{R}^n$ to denote the n-dimensional subspace of $\mathrm{Cl}(n)$ spanned by the generators $e_1, e_2, \dots, e_n$. Elements of $\mathbb{R}^n \subseteq \mathrm{Cl}(n)$ satisfy the following coordinate-free version of the defining relations of $\mathrm{Cl}(n)$ given in (4.7.1).

Lemma 4.7.1. *If $u, v \in \mathbb{R}^n$ then $uv + vu = -2\langle u, v \rangle \mathbf{1}$.*

Proof. First note that $uv + uv = -2\langle u, v \rangle \mathbf{1}$ is bilinear in u and v so it suffices to verify the identity for $u = e_i$ and $v = e_j$ (for all $1 \le i, j \le n$). That the statement holds for $u = e_i$ and $v = e_j$ (for all $1 \le i, j \le n$) is equivalent to the relations (4.7.1) (since $\langle e_i, e_j \rangle = \delta_{ij}$). □

The sphere in $\mathbb{R}^n$ We use the notation $S^{n-1} \subseteq \mathbb{R}^n \subseteq \mathrm{Cl}(n)$ to denote the set of elements $x \in \mathbb{R}^n$ satisfying $\langle x, x \rangle = 1$. We next state and prove some basic properties of the elements of $S^{n-1} \subseteq \mathrm{Cl}(n)$.

Lemma 4.7.2. *If $u \in S^{n-1} \subseteq \mathrm{Cl}(n)$ then $u\overline{u} = \mathbf{1} = \overline{u}u$. Consequently $\langle uy, uz\rangle = \langle y, z\rangle = \langle yu, zu\rangle$ for all $y, z \in \mathrm{Cl}(n)$.*

Proof. The second statement follows from the first together with (4.7.2) and (4.7.3). That $u\overline{u} = \mathbf{1}$ whenever $u \in S^{n-1}$ follows from a direct application of Lemma 4.7.1. □

The following can be established by repeatedly applying Lemma 4.7.2.

Corollary 4.7.3. *If $u_1, u_2, \ldots, u_k \in S^{n-1}$ then $\langle u_1u_2\cdots u_k, u_1u_2\cdots u_k\rangle = 1$.*

Even subalgebra Consider the subspaces $\mathrm{Cl}^0(n)$ and $\mathrm{Cl}^1(n)$ of $\mathrm{Cl}(n)$ defined by

$$\mathrm{Cl}^0(n) = \mathrm{span}\{e_I : I \subseteq [n],\ |I| \text{ even}\} \quad \text{and} \quad \mathrm{Cl}^1(n) = \mathrm{span}\{e_I : I \subseteq [n],\ |I| \text{ odd}\}.$$

It is straightforward to show that if $x, y \in \mathrm{Cl}^0(n)$ then $xy \in \mathrm{Cl}^0(n)$, and if $x, y \in \mathrm{Cl}^1(n)$ then $xy \in \mathrm{Cl}^0(n)$. The first of these properties states that $\mathrm{Cl}^0(n)$ is a subalgebra of $\mathrm{Cl}(n)$, which we call the *even subalgebra*. With these properties we have that the product of an even number of elements of S^{n-1} is in the even subalgebra.

Lemma 4.7.4. *If $u_1, u_2, \ldots, u_{2k} \in S^{n-1}$ then $x = u_1u_2\cdots u_{2k} \in \mathrm{Cl}^0(n)$.*

Proof. Since $S^{n-1} \subseteq \mathbb{R}^n \subseteq \mathrm{Cl}^1(n)$, each $u_i \in \mathrm{Cl}^1(n)$. Hence $u_{2i-1}u_{2i} \in \mathrm{Cl}^0(n)$ for $i = 1, 2, \ldots, k$. So $u_1u_2\cdots u_{2k} = (u_1u_2)(u_3u_4)\cdots(u_{2k-1}u_{2k})$ is the product of elements in $\mathrm{Cl}^0(n)$ so is itself an element of $\mathrm{Cl}^0(n)$. □

4.7.2 Spin(n)

We now define Spin(n) and establish some of its basic properties.

Definition 4.7.5. Spin(n) is the set of all even length products of elements of S^{n-1}, i.e.

$$\mathrm{Spin}(n) = \{x \in \mathrm{Cl}(n) : x = u_1u_2\cdots u_{2k} \text{ where } k \text{ is a positive integer and } u_1, \ldots, u_{2k} \in S^{n-1}\}.$$

Although we do not require this fact, it can be shown that in the above definition it is enough to take $k = \lfloor n/2 \rfloor$. We note that a common alternative definition [79] is to take Spin(n) to be the elements of $\mathrm{Cl}^0(n)$ satisfying $x\overline{x} = \mathbf{1}$ and $xv\overline{x} \in \mathbb{R}^n$ for every $v \in \mathbb{R}^n$ (which defines a real algebraic variety specified by the vanishing of a collection of quadratic equations). It is fairly straightforward to establish that these two definitions are equivalent.

The following observation follows directly from Lemma 4.7.4 and Corollary 4.7.3.

Lemma 4.7.6. *The set* Spin(n) *is a subset of the unit sphere in* $\mathrm{Cl}^0(n)$*, i.e.,* $\mathrm{Spin}(n) \subseteq \{x \in \mathrm{Cl}^0(n) : \langle x, x\rangle = 1\}$.

The next result establishes that $\mathrm{Spin}(n)$ is a group under multiplication.

Lemma 4.7.7. *If $x \in \mathrm{Spin}(n)$ then $\overline{x}x = x\overline{x} = \mathbf{1}$. If $x, y \in \mathrm{Spin}(n)$ then $xy \in \mathrm{Spin}(n)$.*

Proof. That $\mathrm{Spin}(n)$ is closed under multiplication is clear from the definition. That conjugation and inversion coincide on $\mathrm{Spin}(n)$ follows from Lemma 4.7.2. □

4.7.3 The quadratic mapping

We now define and establish the relevant properties of the quadratic mapping $Q : \mathrm{Cl}^0(n) \to \mathbb{R}^{n\times n}$ that plays a prominent role in Section 4.4.2. First define $\widetilde{Q} : \mathrm{Cl}(n) \to \mathbb{R}^{n\times n}$ by

$$\widetilde{Q}(x)(u) = \Pi_{\mathbb{R}^n}\lambda_x\rho_{\overline{x}}(u) = \Pi_{\mathbb{R}^n}(xu\overline{x}).$$

Note that $\widetilde{Q}(x)$ is quadratic in x. Then define $Q : \mathrm{Cl}^0(n) \to \mathbb{R}^{n\times n}$ as the restriction of $\widetilde{Q}$ to the subalgebra $\mathrm{Cl}^0(n)$.

When we express the linear map $\widetilde{Q}(x)$ as a matrix (with respect to the standard basis) we see that $[\widetilde{Q}(x)]_{ij} = \langle e_i, xe_j\overline{x}\rangle$. Furthermore $\widetilde{Q}$ (and hence Q) interacts nicely with the conjugation map.

Lemma 4.7.8. *If $x \in \mathrm{Cl}(n)$ then $\widetilde{Q}(\overline{x}) = \widetilde{Q}(x)^T$.*

Proof. Simply observe that $[\widetilde{Q}(x)]_{ij} = \langle e_i, xe_j\overline{x}\rangle = \langle \overline{x}e_ix, e_j\rangle = [\widetilde{Q}(\overline{x})]_{ji}$. □

The definition of $\widetilde{Q}$ is motivated by the fact that if $u \in S^{n-1}$ then $-\widetilde{Q}(u)$ is the reflection in the hyperplane orthogonal to u.

Lemma 4.7.9. *Let $u \in S^{n-1}$. Then whenever $v \in \mathbb{R}^n$, $-uv\overline{u} \in \mathbb{R}^n$ is the reflection of v in the hyperplane normal to u. In particular $-uv\overline{u} \in \mathbb{R}^n$.*

Proof. Let $u \in S^{n-1}$. Then by Lemma 4.7.1, if $v \in \mathbb{R}^n$ then $-uv = 2\langle u, v\rangle\mathbf{1} + vu$. Since $u\overline{u} = \mathbf{1}$ and $\overline{u} = -u$, it follows that

$$-uv\overline{u} = 2\langle u, v\rangle\overline{u} + vu\overline{u} = v - 2\langle u, v\rangle u,$$

which is the reflection in the hyperplane orthogonal to u and is certainly in $\mathbb{R}^n$. □

Note that our definition of $\widetilde{Q}$ is one possible extension to all of $\mathrm{Cl}(n)$ of the map that sends $u \in S^{n-1}$ to the reflection in the hyperplane orthogonal to u. It is specifically chosen so as to be quadratic on all of $\mathrm{Cl}(n)$. Our choice is different from the typical extension used in the literature—the *twisted adjoint representation* [79]— which is *not* quadratic in x on all of $\mathrm{Cl}(n)$ and is not suitable for our purposes.

Lemma 4.7.10. *Let $x \in \mathrm{Cl}(n)$ and $u \in S^{n-1}$. Then*

$$\widetilde{Q}(xu) = \widetilde{Q}(x)\widetilde{Q}(u) \quad \text{and} \quad \widetilde{Q}(ux) = \widetilde{Q}(u)\widetilde{Q}(x)$$

where the product on the right hand side in each case is composition of linear maps.

Proof. If $u \in S^{n-1}$, we know from the previous lemma that $v \mapsto uv\overline{u}$ leaves the subspace $\mathbb{R}^n$ (and hence its orthogonal complement) invariant. So by the definition of $\widetilde{Q}$ we see that

$$\widetilde{Q}(xu)(v) = \Pi_{\mathbb{R}^n}(xuv\overline{u}\,\overline{x}) = \Pi_{\mathbb{R}^n}(xP_{\mathbb{R}^n}(uv\overline{u})\overline{x}) = \widetilde{Q}(x)(\widetilde{Q}(u)(v)).$$

Similarly since $P_{\mathbb{R}^n} + P_{\mathbb{R}^{n\perp}} = I$,

$$\begin{aligned}\widetilde{Q}(ux)(v) &= \Pi_{\mathbb{R}^n}(uxv\overline{x}\,\overline{u})\\ &= \Pi_{\mathbb{R}^n}(uP_{\mathbb{R}^n}(xv\overline{x})\overline{u}) + \Pi_{\mathbb{R}^n}(uP_{\mathbb{R}^{n\perp}}(xv\overline{x})\overline{u}) = Q(u)(Q(x)(v)) + 0\end{aligned}$$

where we have used the fact that $uy\overline{u} \in \mathbb{R}^{n\perp}$ whenever $y \in \mathbb{R}^{n\perp}$. □

We are now in a position to prove Propositions 4.4.1, 4.4.2, and 4.4.3. We restate them here for convenience.

Proposition 4.4.1. *There is a 2^{n-1}-dimensional inner product space, $\mathrm{Cl}^0(n)$, a subset $\mathrm{Spin}(n)$ of the unit sphere in $\mathrm{Cl}^0(n)$ and a quadratic map $Q : \mathrm{Cl}^0(n) \to \mathbb{R}^{n\times n}$ such that $Q(\mathrm{Spin}(n)) = SO(n)$.*

Proof. The construction of $\mathrm{Cl}^0(n)$ is given in Section 4.7.1. The set $\mathrm{Spin}(n)$ is defined in 4.7.5. That $\mathrm{Spin}(n)$ is a subset of the sphere in $\mathrm{Cl}^0(n)$ is the content of Lemma 4.7.6. The quadratic mapping Q is defined in Section 4.7.3. It remains to show that $Q(\mathrm{Spin}(n)) = SO(n)$.

Let $X \in SO(n)$. By the Cartan-Dieudonné theorem [48] any such X can be expressed as the composition of an even number (at most n) of reflections in hyperplanes with normal vectors, say, $u_1, u_2, \ldots, u_{2k} \in S^{n-1}$. Let $x = u_1u_2\cdots u_{2k-1}u_{2k} \in \mathrm{Spin}(n)$. Then by Lemma 4.7.9 and Lemma 4.7.10 and the fact that Q is the restriction of $\widetilde{Q}$ to $\mathrm{Cl}^0(n)$,

$$X = \widetilde{Q}(u_1)\widetilde{Q}(u_2)\cdots\widetilde{Q}(u_{2k-1})\widetilde{Q}(u_{2k}) = \widetilde{Q}(x) = Q(x) \in Q(\mathrm{Spin}(n)).$$

Hence $SO(n) \subseteq Q(\mathrm{Spin}(n))$. On the other hand, if $x = u_1u_2\cdots u_{2k-1}u_{2k} \in \mathrm{Spin}(n)$ then $Q(x)$ is the product of an even number of reflections in hyperplanes and so is an element of $SO(n)$, establishing the reverse inclusion. □

Proposition 4.4.2. *If $U, V \in SO(n)$ then there is a corresponding invertible linear* *map* $\Phi_{(U,V)} : \mathrm{Cl}^0(n) \to \mathrm{Cl}^0(n)$ *such that for any* $x \in \mathrm{Cl}^0(n)$, $UQ(x)V^T = Q(\Phi_{(U,V)}x)$ *and* $\Phi_{(U,V)}(\mathrm{Spin}(n)) = \mathrm{Spin}(n)$.

Proof. By Proposition 4.4.1 there are $u, v \in \mathrm{Spin}(n)$ such that $Q(u) = U$ and $Q(v) = V$. Define $\Phi_{(U,V)} : \mathrm{Cl}^0(n) \to \mathrm{Cl}^0(n)$ by $\Phi_{(U,V)}(x) = ux\overline{v}$. Then $\Phi_{(U,V)}$ is invertible with inverse $\Phi^{-1}_{(U,V)}(x) = \overline{u}xv$. By Lemmas 4.7.8 and 4.7.10, for any $x \in \mathrm{Cl}^0(n)$,

$$UQ(x)V^T = Q(u)Q(x)Q(v)^T = Q(u)Q(x)Q(\overline{v}) = Q(ux\overline{v}).$$

Finally, if $x \in \mathrm{Spin}(n)$ then $\Phi_{(U,V)}(x) = ux\overline{v} \in \mathrm{Spin}(n)$ by Lemma 4.7.7. Hence $\Phi_{(U,V)}(\mathrm{Spin}(n)) \subseteq \mathrm{Spin}(n)$. For the reverse inclusion, if $x \in \mathrm{Spin}(n)$ then $\Phi^{-1}_{(U,V)}(x) = \overline{u}xv \in \mathrm{Spin}(n)$ by Lemma 4.7.7. Hence $\Phi_{(U,V)}(\mathrm{Spin}(n)) \supseteq \mathrm{Spin}(n)$, establishing that $\Phi_{(U,V)}(\mathrm{Spin}(n)) = \mathrm{Spin}(n)$. □

Proposition 4.4.3. *Given any orthonormal basis* $u_1, \ldots, u_n$ *for* $\mathbb{R}^n$, *there is a corresponding orthonormal basis* $(u_I)_{I \in \mathcal{I}_{even}}$ *for* $\mathrm{Cl}^0(n)$ *such that*

- $u_I \in \mathrm{Spin}(n)$ *for all* $I \in \mathcal{I}_{even}$ *and*
- *for all* $i \in [n]$, *if* $x = \sum_{I \in \mathcal{I}_{even}} x_I u_I$ *then*

$$\langle u_i, Q\,(x)\,u_i\rangle = \sum_{I \in \mathcal{I}_{even}} x_I^2 \langle u_i, Q(u_I)u_i\rangle.$$

Proof. Let $u_1, u_2, \ldots, u_n \in \mathbb{R}^n$ be orthonormal with respect to the usual inner product on $\mathbb{R}^n$. When thought of as elements of $\mathbb{R}^n \subset \mathrm{Cl}(n)$ these satisfy $u_i^2 = -\mathbf{1}$ for all i and $u_iu_j = -u_ju_i$ when $i \neq j$ (by Lemma 4.7.1). As such we can construct from $u_1, u_2, \ldots, u_n$ a basis for $\mathrm{Cl}^0(n)$ just as we did for the standard basis. Indeed let $I = \{i_1, \ldots, i_{2k}\} \subseteq [n]$ where $i_1 < i_2 < \cdots < i_{2k}$, and define $u_I = u_{i_1}u_{i_2}\cdots u_{i_{2k}}$. This realizes u_I as the product of an even number of elements of S^{n-1}, showing that $u_I \in \mathrm{Spin}(n)$. For the second statement, note that if $x = \sum_{I \in \mathcal{I}_{\mathrm{even}}} x_I u_I$ and $i \in [n]$,

$$\begin{aligned}
\langle u_i, Q(x)u_i\rangle &= \langle u_i, xu_i\overline{x}\rangle \\
&= \sum_{I,J \in \mathcal{I}_{\mathrm{even}}} x_I x_J \langle u_i, u_I u_i \overline{u_J}\rangle \\
&\overset{*}{=} \sum_{I,J \in \mathcal{I}_{\mathrm{even}}} x_I x_J \delta_{IJ} \langle u_i, u_I u_i \overline{u_I}\rangle \\
&= \sum_{I \in \mathcal{I}_{\mathrm{even}}} x_I^2 \langle u_i, Q(u_I)u_i\rangle
\end{aligned}$$

where $\delta_{IJ} = 1$ if $I = J$ and zero otherwise, and the equality marked with an asterisk follows directly from the coordinate-free version of the defining relations of the Clifford algebra (Lemma 4.7.1). $\square$

4.7.4 Matrices of the quadratic mapping

For $1 \leq i, j \leq n$, let $A^{(ij)} : \mathrm{Cl}^0(n) \to \mathrm{Cl}^0(n)$ be the self-adjoint linear map such that, for all $x \in \mathrm{Cl}^0(n)$,

$$[Q(x)]_{ij} = \langle e_i, x e_j \overline{x} \rangle = \langle x, A^{(ij)} x \rangle.$$

First we note that the $A^{(ij)}$ have trace zero.

Lemma 4.7.4. *For $1 \leq i, j \leq n$, $\mathrm{tr}(A^{(ij)}) = 0$.*

Proof. For $i \in [n]$ and $I \in \mathcal{I}_{\text{even}}$ define $\delta_{[i \in I]} = 1$ if $i \in I$ and 0 otherwise. Observe that from the definition of $A^{(ij)}$ and the defining relations of the Clifford algebra,

$$\mathrm{tr}(A^{(ij)}) = \sum_{I \in \mathcal{I}_{\text{even}}} \langle e_I, A^{(ij)} e_I \rangle = \sum_{I \in \mathcal{I}_{\text{even}}} \langle e_i, e_I e_j \overline{e_I} \rangle = \sum_{I \in \mathcal{I}_{\text{even}}} (-1)^{\delta_{[j \in I]}} \langle e_i, e_j \rangle.$$

If $i \neq j$ every term in the sum vanishes. If $i = j$ observe that there are 2^{n-2} elements of $\mathcal{I}_{\text{even}}$ containing j and 2^{n-2} elements of $\mathcal{I}_{\text{even}}$ not containing j, hence $\sum_{I \in \mathcal{I}_{\text{even}}} (-1)^{\delta_{[j \in I]}} \langle e_j, e_j \rangle = 0$. $\square$

For the remainder of the section we show that with respect to the standard basis $(e_I)_{I \in \mathcal{I}_{\text{even}}}$ for $\mathrm{Cl}^0(n)$, the $A^{(ij)}$ are represented by the $2^{n-1} \times 2^{n-1}$ symmetric matrices described in (4.1.6).

Let $\widetilde{A}^{(ij)} : \mathrm{Cl}(n) \to \mathrm{Cl}(n)$ be the self-adjoint linear map such that, for all $x \in \mathrm{Cl}(n)$, $[\widetilde{Q}(x)]_{ij} = \langle e_i, x e_j \overline{x} \rangle = \langle x, \widetilde{A}^{(ij)} x \rangle$. Since

$$\langle e_i, x e_j \overline{x} \rangle = \langle e_i x, x e_j \rangle = \langle x, \lambda_{\overline{e_i}} \rho_{e_j} x \rangle = -\langle x, \lambda_{e_i} \rho_{e_j} x \rangle$$

it follows that $\widetilde{A}^{(ij)} = -\lambda_{e_i} \rho_{e_j}$. Since $A^{(ij)}$ is the restriction of $\widetilde{A}^{(ij)}$ to the subspace $\mathrm{Cl}^0(n)$ we have that

$$A^{(ij)} = \Pi_{\mathrm{Cl}^0(n)} (-\lambda_{e_i} \rho_{e_j}) \Pi^*_{\mathrm{Cl}^0(n)}. \tag{4.7.4}$$

It remains to derive the matrices that represent the λ_{e_i} and ρ_{e_i} for $i = 1, 2, \ldots, n$, as well as the matrix representing $\Pi_{\mathrm{Cl}^0(n)}$, in terms of the standard basis $(e_I)_{I \subseteq [n]}$ for $\mathrm{Cl}(n)$ (ordered in a particular way). In what follows, if $v \in \mathrm{Cl}(n)$ we write $[v]$ for its coordinate representation as a vector in $\mathbb{R}^{2^n}$ with respect to the particular ordered basis we use. We use brackets in a similar way to express an abstractly defined linear map in these coordinates as a matrix.

To describe the ordered basis, define $\delta_{[i \in I]} = 1$ if $i \in I$ and zero otherwise, and define $\delta_{[i \notin I]} = 1$ if $i \notin I$ and zero otherwise. We order the basis elements in such a way that, in coordinates,

$$[e_I] = \begin{bmatrix} \delta_{[1\notin I]} \\ \delta_{[1\in I]} \end{bmatrix} \otimes \begin{bmatrix} \delta_{[2\notin I]} \\ \delta_{[2\in I]} \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} \delta_{[n\notin I]} \\ \delta_{[n\in I]} \end{bmatrix}.$$

It is straightforward to verify (by checking that the relations of (4.7.1) are satisfied) that in these coordinates,

$$\lambda_i := [\lambda_{e_i}] = \overbrace{\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}}^{i-1} \otimes \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \otimes \overbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}^{n-i}$$

and

$$\rho_i := [\rho_{e_i}] = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{i-1} \otimes \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \otimes \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}}_{n-i}.$$

Now $P_{\mathrm{Cl}^0(n)} : \mathrm{Cl}(n) \to \mathrm{Cl}(n)$ is represented in these coordinates by

$$[P_{\mathrm{Cl}^0(n)}] = \frac{1}{2} \overbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}^{n} + \frac{1}{2} \overbrace{\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}}^{n}.$$

This can be verified by noting that if $|I|$ is odd, $[P_{\mathrm{Cl}^0(n)}][e_I] = 0$, and if $|I|$ is even, $[P_{\mathrm{Cl}^0(n)}][e_I] = [e_I]$. Defining the $2^n \times 2^{n-1}$ matrix

$$P_{\text{even}} = \frac{1}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \otimes \overbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}^{n-1} + \frac{1}{2} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \otimes \overbrace{\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}}^{n-1}$$

and checking that it satisfies $P_{\text{even}} P_{\text{even}}^T = [P_{\mathrm{Cl}^0(n)}]$ and that the columns of P_{even} are orthonormal, establishes that $P_{\text{even}} = [\Pi^*_{\mathrm{Cl}^0(n)}]$. It then follows that in these coordinates

$$A^{(ij)} = -P_{\text{even}}^T \lambda_i \rho_j P_{\text{even}}$$

as stated in (4.1.6).

4.8 Semidefinite representations of a generalized trigonometric moment curve

The trigonometric moment curve (first studied by Carathéodory [24]) is the following curve:

$$\{(1, \cos(\omega), \sin(\omega), \ldots, \cos(T\omega), \sin(T\omega)) : \omega \in [0, 2\pi)\} \subset \mathbb{R}^{2T+1}. \tag{4.8.1}$$

In Section 2.2 of Chapter 2 we considered the projection $\{(\cos(\omega), \sin(\omega), \sin(2\omega)) : \omega \in [0, 2\pi)\}$ of this curve to illustrate the idea of reformulating a family of optimization problems as linear optimization over a convex set.

In this section we study semidefinite representations of the convex hull of the following generalization of the trigonometric moment curve to tuples of symmetric matrices

$$\mathcal{TM}_{m,T}^{\mathbb{R}} = \{(Z, \cos(\omega)Z, \sin(\omega)Z, \ldots, \cos(T\omega)Z, \sin(T\omega)Z) : \\ Z \in \mathcal{S}_+^m,\ \ \operatorname{tr}(Z) = 1\Big\} \subset (\mathcal{S}^m)^{2T+1}. \tag{4.8.2}$$

If $m = 1$ it is clear that $\mathcal{TM}_{m,T}^{\mathbb{R}}$ reduces to the curve given in (4.8.1). We are interested in semidefinite representations of the convex hull of $\mathcal{TM}_{m,T}^{\mathbb{R}}$ in the context of the joint attitude and spin-rate estimation problem described in Section 4.2. This is because the convex hull of $\mathcal{M}_{3,T}$, described in Proposition 4.2.1, is a projection of the convex hull of $\mathcal{TM}_{4,T}^{\mathbb{R}}$ (see Section 4.8.3 to follow).

The following is a natural complex Hermitian counterpart of $\mathcal{TM}_{m,T}^{\mathbb{R}}$,

$$\mathcal{TM}_{m,T}^{\mathbb{C}} = \Big\{(e^{-i\omega T}Z, \ldots, e^{-i\omega T}Z, Z, e^{i\omega}Z, e^{i2\omega}Z, \ldots, e^{iT\omega}Z) : \\ Z \in \mathcal{H}_+^m,\ \ \operatorname{tr}(Z) = 1\Big\} \subset (\mathbb{C}^{m\times m})^{2T+1}. \tag{4.8.3}$$

(Throughout this section we use the notation $\mathcal{H}_+^m$ for the cone of $m \times m$ Hermitian positive semidefinite matrices.) The main result we use is the following Hermitian semidefinite representation of the conic hull (see Section 2.3 from Chapter 2) of $\mathcal{TM}_{m,T}^{\mathbb{C}}$. This appears in [51, Section 6.2] for instance.

Theorem 4.8.1. *The conic hull of $\mathcal{TM}_{m,T}^{\mathbb{C}}$ is*

$$\operatorname{cone}(\mathcal{TM}_{m,T}^{\mathbb{C}}) = \{(W_{-T}, W_{-T+1}, \ldots, W_{T-1}, W_T) : \\ \operatorname{Toep}(W_{-T}, W_{-T+1}, \ldots, W_{T-1}, W_T) \in \mathcal{H}_+^{(T+1)m}\Big\}.$$

We refer to [62, Theorem 2.1.6] for a detailed proof of this well-known result. The basic idea is that the dual cone $(\mathcal{TM}_{m,T}^{\mathbb{C}})^*$ of $\mathcal{TM}_{m,T}^{\mathbb{C}}$ consists of the coefficients of

$\mathcal{H}_+^m$-valued trigonometric polynomials. These have a semidefinite representation via the matrix version of the Fejér-Riesz theorem [108, 39] (also known as the spectral factorization theorem). Taking the dual of this semidefinite representation gives the statement of Theorem 4.8.1.

In Lemma 4.8.2 of Section 4.8.1 we show that $\text{conv}(\mathcal{TM}_{m,T}^{\mathbb{R}})$ is a section of the cone $\text{cone}(\mathcal{TM}_{m,T}^{\mathbb{C}})$. Combining this with Theorem 4.8.1 immediately gives an Hermitian semidefinite representation of $\text{conv}(\mathcal{TM}_{m,T}^{\mathbb{R}})$ of size $(T+1)m$. In Section 4.8.2 we show that this particular slice has a special property that allows us to convert our Hermitian semidefinite representation of size $(T+1)m$ to a real symmetric semidefinite representation of the same size, $(T+1)m$. We also establish sufficient conditions under which this phenomenon occurs in general (see Lemma 4.8.4). These conditions seem to be new, and may be useful in other situations. We give the final form of our semidefinite representation of $\text{conv}(\mathcal{TM}_{m,T}^{\mathbb{R}})$ in Proposition 4.8.5. In Section 4.8.3 we use it to prove Proposition 4.2.1 that gives an explicit semidefinite representation of $\mathcal{M}_{3,T}$.

4.8.1 Relating $\text{conv}(\mathcal{TM}_{m,T}^{\mathbb{R}})$ and $\text{cone}(\mathcal{TM}_{m,T}^{\mathbb{C}})$

We now show how to recover $\text{conv}(\mathcal{TM}_{m,T}^{\mathbb{R}})$ as a section of $\text{cone}(\mathcal{TM}_{m,T}^{\mathbb{C}})$.

Lemma 4.8.2.

$$\begin{aligned}\text{conv}(\mathcal{TM}_{m,T}^{\mathbb{R}}) = \{&(X_0, X_1, Y_1, \ldots, X_T, Y_T) \in (\mathcal{S}^m)^{2T+1} : \\ &(X_0, X_1 + iY_1, \ldots, X_T + iY_T) \in \text{cone}(\mathcal{TM}_{m,T}^{\mathbb{C}}),\ \text{tr}(X_0) = 1\}.\end{aligned} \tag{4.8.4}$$

Proof. Suppose $Z \in \mathcal{S}_+^m$ and $\omega \in [0, 2\pi)$ so that

$$(Z, Z\cos(\omega), Z\sin(\omega), \ldots, Z\cos(T\omega), Z\sin(T\omega) \in \mathcal{TM}_{m,T}^{\mathbb{R}}.$$

Then since $\text{tr}(Z) = 1$ and $Z\cos(k\omega) + iZ\sin(k\omega) = Ze^{ik\omega}$ for $k = 1, 2, \ldots, T$ it follows that

$$(Z, Z\cos(\omega) + iZ\sin(\omega), \ldots, Z\cos(T\omega) + iZ\sin(T\omega)) \in \mathcal{TM}_{m,T}^{\mathbb{C}} \subseteq \text{cone}(\mathcal{TM}_{m,T}^{\mathbb{C}}).$$

Hence $\text{conv}(\mathcal{TM}_{m,T}^{\mathbb{R}})$ is a subset of the right hand side of (4.8.4).

For the reverse inclusion, suppose $(X_0, X_1 + iY_1, \ldots, X_T + iY_T) \in \text{cone}(\mathcal{TM}_{m,T}^{\mathbb{C}})$ and $\text{tr}(X_0) = 1$. Then there is a positive integer r, and (for $j = 1, 2, \ldots, r$) there are non-negative scalars λ_j, there are $\omega_j \in [0, 2\pi)$, there are W_j^{Re} (real symmetric) and W_j^{Im} (real skew symmetric) such that $W_j^{\text{Re}} + iW_j^{\text{Im}} \in \mathcal{H}_+^m$, $\text{tr}(W_j^{\text{Re}} + iW_j^{\text{Im}}) = \text{tr}(W_j^{\text{Re}}) = 1$,

$$X_0 = \sum_{j=1}^{r} \lambda_j (W_j^{\text{Re}} + iW_j^{\text{Im}}) \tag{4.8.5}$$

and for $k = 1, 2, \ldots, T$,

$$X_k + iY_k = \sum_{j=1}^{r} \lambda_j (W_j^{\text{Re}} + iW_j^{\text{Im}})(\cos(k\omega_j) + i\sin(k\omega_j)). \tag{4.8.6}$$

Since $X_0 = X_0^T$ it follows that we also have

$$X_0 = \sum_{j=1}^{r} \lambda_j (W_j^{\text{Re}} - iW_j^{\text{Im}}) \tag{4.8.7}$$

and since for $k = 1, 2, \ldots, T$, $X_k = X_k^T$ and $Y_k = Y_k^T$, we have that

$$X_k + iY_k = \sum_{j=1}^{r} \lambda_j (W_j^{\text{Re}} - iW_j^{\text{Im}})(\cos(k\omega_j) + i\sin(k\omega_j)). \tag{4.8.8}$$

By taking the average of these two decompositions of X_0 (from (4.8.5) and (4.8.7)) and similarly taking the average of the two decompositions we have for each $X_k + iY_k$ for $k = 1, 2, \ldots, T$ (from (4.8.6) and (4.8.8)) we see that

$$X_0 = \sum_{j=1}^{r} \lambda_j W_j^{Re} \quad \text{and} \quad X_k + iY_k = \sum_{j=1}^{r} \lambda_j \cos(k\omega_j) W_j^{\text{Re}} + i \sum_{j=1}^{r} \lambda_j \sin(k\omega_j) W_j^{\text{Re}}. \tag{4.8.9}$$

Note that $W_j^{\text{Re}} \in \mathcal{S}_+^m$ since $W_j^{\text{Re}} + iW_j^{\text{Im}} \in \mathcal{H}_+^m$ and $W_j^{Re} - iW_j^{\text{Im}} \in \mathcal{H}_+^m$ so their average is real symmetric and positive semidefinite.

Observe that

$$1 = \operatorname{tr}(X_0) = \sum_{j=1}^{r} \lambda_j \operatorname{tr}(W_j^{\text{Re}}) = \sum_{j=1}^{r} \lambda_j.$$

Hence (4.8.9) gives a realization of $(X_0, X_1, Y_1, \ldots, X_T, Y_T)$ as a convex combination of the tuples $(W_j^{\text{Re}}, \cos(\omega_j)W_j^{\text{Re}}, \sin(\omega_j)W_j^{\text{Re}}, \ldots, \cos(T\omega_j)W_j^{\text{Re}}, \sin(T\omega_j)W_j^{\text{Re}}) \in \mathcal{TM}_{m,T}^{\mathbb{R}}$ as required. $\square$

By directly combining the results of Theorem 4.8.1 and Lemma 4.8.2 we obtain the following Hermitian semidefinite representation of $\operatorname{conv}(\mathcal{TM}_{m,T}^{\mathbb{R}})$ of size $(T+1)m$.

Proposition 4.8.3.

$$\begin{aligned}\operatorname{conv}(\mathcal{TM}_{m,T}^{\mathbb{R}}) = \{\, & (X_0, X_1, Y_1, \ldots, X_T, Y_T) \in (\mathcal{S}^m)^{2T+1} : \\ & \operatorname{Toep}(X_T - iY_T, \ldots, X_1 - iY_1, X_0, X_1 + iY_1, \ldots, X_T + iY_T) \in \mathcal{H}_+^{(T+1)m} \,\}.\end{aligned}$$

■ 4.8.2 Real symmetric semidefinite representations

We have an Hermitian semidefinite representation of $\operatorname{conv}(\mathcal{TM}^{\mathbb{R}}_{m,T})$ of size $(T+1)m$. It is possible to express any Hermitian semidefinite representation of size d as a real symmetric semidefinite representation of size $2d$. This is because

$$L \in \mathcal{H}^d_+ \quad \text{if and only if} \quad \begin{bmatrix} \mathrm{Re}\,[L] & -\mathrm{Im}\,[L] \\ \mathrm{Im}\,[L] & \mathrm{Re}\,[L] \end{bmatrix} \in \mathcal{S}^{2d}_+$$

(a well-known fact that we reestablish in the course of the proof of Lemma 4.8.4 to follow). The Hermitian semidefinite representation of size $(T+1)m$ in Proposition 4.8.3 has additional special structure. This structure actually allows us to rewrite it as a real symmetric semidefinite representation of *the same size* rather than twice the size. Lemma 4.8.4, which we state and prove next, describes this structure in general.

Lemma 4.8.4. *Let $\mathcal{L}$ be an affine subspace (over the reals) of $\mathcal{H}^d$. Suppose there is some orthogonal $J \in O(d)$ such that $J^2 = I$ and*

$$JLJ^T = \bar{L} \quad \textit{for all } L \in \mathcal{L},$$

i.e. congruence by J restricted to $\mathcal{L}$ is entry-wise complex conjugation. Then

$$\{L \in \mathcal{L} : L \in \mathcal{H}^d_+\} = \{L \in \mathcal{L} : \mathrm{Re}\,[L] - J\mathrm{Im}\,[L] \in \mathcal{S}^d_+\}.$$

Proof. First note that $L \in \mathcal{H}^d_+$ if and only if $\bar{L} \in \mathcal{H}^d_+$ which holds if and only if the block diagonal matrix $\left[\begin{smallmatrix} L & 0 \\ 0 & \bar{L}\end{smallmatrix}\right] \in \mathcal{H}^{2d}_+$. Conjugating by a unitary matrix we obtain

$$\begin{bmatrix} \frac{1}{\sqrt{2}}I & \frac{1}{\sqrt{2}}I \\ \frac{i}{\sqrt{2}}I & -\frac{i}{\sqrt{2}}I \end{bmatrix} \begin{bmatrix} L & 0 \\ 0 & \bar{L} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}}I & \frac{1}{\sqrt{2}}I \\ \frac{i}{\sqrt{2}}I & -\frac{i}{\sqrt{2}}I \end{bmatrix}^* = \begin{bmatrix} \mathrm{Re}\,[L] & \mathrm{Im}\,[L] \\ -\mathrm{Im}\,[L] & \mathrm{Re}\,[L] \end{bmatrix}. \tag{4.8.10}$$

We have simply recovered the familiar realization of $\mathcal{H}^d_+$ as a section of $\mathcal{S}^{2d}_+$, and have not yet used any special properties of $\mathcal{L}$. To complete the proof it remains to carefully choose a $2d \times 2d$ orthogonal matrix Q (depending on J) such that

$$Q \begin{bmatrix} \mathrm{Re}\,[L] & \mathrm{Im}\,[L] \\ -\mathrm{Im}\,[L] & \mathrm{Re}\,[L] \end{bmatrix} Q^T = \begin{bmatrix} \mathrm{Re}\,[L] - J\mathrm{Im}\,[L] & 0 \\ 0 & \mathrm{Re}\,[L] - J\mathrm{Im}\,[L] \end{bmatrix} \quad \text{for all } L \in \mathcal{L}.$$

Observe that $J^2 = I$ and $J^TJ = I$ imply that $J = J^T$. Since $JLJ^T = \bar{L}$ we have that for all $L \in \mathcal{L}$,

$$\mathrm{Re}\,[L] = \frac{L + JLJ}{2} \quad \text{and} \quad \mathrm{Im}\,[L] = \frac{L - JLJ}{2i}. \tag{4.8.11}$$

It follows that for all $L \in \mathcal{L}$, $\mathrm{Re}\,[L]$ and $\mathrm{Im}\,[L]$ commute and anti-commute respectively with J, i.e.,

$$J\mathrm{Re}\,[L] = \mathrm{Re}\,[L]J \quad \text{and} \quad J\mathrm{Im}\,[L] = -\mathrm{Im}\,[L]J. \tag{4.8.12}$$

Choosing Q to be the orthogonal matrix $Q = \frac{1}{\sqrt{2}}\left[\begin{smallmatrix} I & J \\ -J & I \end{smallmatrix}\right]$ we obtain

$$\begin{bmatrix} \frac{1}{\sqrt{2}}I & \frac{1}{\sqrt{2}}J \\ -\frac{1}{\sqrt{2}}J & \frac{1}{\sqrt{2}}I \end{bmatrix}\begin{bmatrix} \mathrm{Re}\,[L] & \mathrm{Im}\,[L] \\ -\mathrm{Im}\,[L] & \mathrm{Re}\,[L] \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{2}}I & \frac{1}{\sqrt{2}}J \\ -\frac{1}{\sqrt{2}}J & \frac{1}{\sqrt{2}}I \end{bmatrix}^T = \begin{bmatrix} \mathrm{Re}\,[L] - J\mathrm{Im}\,[L] & 0 \\ 0 & \mathrm{Re}\,[L] - J\mathrm{Im}\,[L] \end{bmatrix}.$$

Clearly this last matrix is positive semidefinite if and only if the real symmetric matrix $\mathrm{Re}\,[L] - J\mathrm{Im}\,[L]$ is positive semidefinite, completing the proof. □

We now apply Lemma 4.8.4 to reduce the Hermitian semidefinite representation of $\mathrm{conv}(\mathcal{TM}_{m,T}^{\mathbb{R}})$ from 4.8.3 to a real symmetric semidefinite representation of $\mathrm{conv}(\mathcal{TM}_{m,T}^{\mathbb{R}})$ of the same size.

Proposition 4.8.5. *The convex hull of $\mathcal{TM}_{m,T}^{\mathbb{R}}$ has a real symmetric semidefinite representation of size $(T+1)m$ given by*

$$\begin{aligned}\mathrm{conv}(\mathcal{TM}_{m,T}^{\mathbb{R}}) = \Big\{(X_0, X_1, Y_1, \ldots, X_T, Y_T) \in (\mathcal{S}^m)^{T+1} :\ & \mathrm{tr}(X_0) = 1, \\ & \mathrm{Toep}(X_T, \ldots, X_1, X_0, X_1, \ldots, X_T) + \\ & \mathrm{Hank}(Y_T, Y_{T-1}, \ldots, Y_1, 0, -Y_1, \ldots, -Y_{T-1}, -Y_T) \in \mathcal{S}_+^{(T+1)m}\Big\}.\end{aligned}$$

Proof. Suppose J is the $m(T+1) \times m(T+1)$ matrix that is block anti-diagonal with $m \times m$ identity matrices on the block anti-diagonal. More precisely, if $0 \leq k, \ell \leq T$ then the $m \times m$ block of J indexed by (k, ℓ) is

$$[J]_{k\ell} = \begin{cases} I & \text{if } k + \ell = T \\ 0 & \text{otherwise.} \end{cases}$$

Note that $L \mapsto JLJ^T$ has the effect of reversing the block rows and columns of the matrix L. Specifically

$$\begin{aligned} & J\mathrm{Toep}((X_T + iY_T)^*, \ldots, (X_1 + iY_1)^*, X_0, X_1 + iY_1, \ldots, X_T + iY_T)J^T = \\ & \qquad \mathrm{Toep}(X_T + iY_T, \ldots, X_1 + iY_1, X_0, (X_1 + iY_1)^*, \ldots, (X_T + iY_T)^*). \end{aligned} \tag{4.8.13}$$

While this has the effect of taking the conjugate transpose of the block entries, in general this is not enough to apply Lemma 4.8.4. Since, in our situation, $X_k = X_k^T$ for all $k = 0, 1, \ldots, T$ and $Y_k = Y_k^T$ for all $1, 2, \ldots, T$ we have that $(X_k + iY_k)^* = X_k - iY_k$ for all $k = 1, 2, \ldots, T$. These additional observations show that the right side of (4.8.13) is the *entry-wise* complex conjugate of $\mathrm{Toep}((X_T + iY_T)^*, \ldots, (X_1 + iY_1)^*, X_0, X_1 + iY_1, \ldots, X_T + iY_T)$.

We are in a position to apply Lemma 4.8.4 to the Hermitian semidefinite representation of $\mathrm{conv}(\mathcal{TM}_{m,T}^{\mathbb{R}})$ from Proposition 4.8.3. To do so we first compute

$$\mathrm{Re}\,[\mathrm{Toep}((X_T + iY_T)^*, \ldots, (X_1 + iY_1)^*, X_0, X_1 + iY_1, \ldots, X_T + iY_T)] = \mathrm{Toep}(X_T, \ldots, X_1, X_0, X_1, \ldots, X_T)$$

and

$$-J\mathrm{Im}\,[\mathrm{Toep}((X_T + iY_T)^*, \ldots, (X_1 + iY_1)^*, X_0, X_1 + iY_1, \ldots, X_T + iY_T)] = \mathrm{Hank}(Y_T, \ldots, Y_1, 0, -Y_1, \ldots, -Y_T).$$

The stated semidefinite representation of $\mathrm{conv}(\mathcal{TM}_{m,T}^{\mathbb{R}})$ then follows directly from Lemma 4.8.4. $\square$

■ 4.8.3 Proof of Proposition 4.2.1

We now prove the correctness of the description of $\mathrm{conv}(\mathcal{M}_{3,T})$ given in Section 4.2.

Proof of Proposition 4.2.1. We begin by recalling some useful facts. First, recall that in Corollary 4.1.6 we showed that $\mathrm{conv}\, SO(3) = \{\mathcal{A}(Z) : Z \in \mathcal{S}_+^4,\ \mathrm{tr}(Z) = 1\}$ where $\mathcal{A} : \mathcal{S}^4 \to \mathbb{R}^{3\times 3}$ is the linear map defined in the statement of Proposition 4.2.1. Define the linear map $\tilde{\mathcal{A}} : (\mathcal{S}^4)^{2T+1} \to \mathbb{R}^2 \times (\mathbb{R}^{3\times 3})^{2T+1}$ by

$$\tilde{\mathcal{A}}(X_0, X_1, Y_1, \ldots, X_T, Y_T) = (\mathrm{tr}(X_1), \mathrm{tr}(Y_1), \mathcal{A}(X_0), \mathcal{A}(X_1), \mathcal{A}(Y_1), \ldots, \mathcal{A}(X_T), \mathcal{A}(Y_T)).$$

Recall, also, that for any linear map $\mathcal{B}$ and any set S, $\mathrm{conv}(\mathcal{B}(S)) = \mathcal{B}(\mathrm{conv}(S))$ (see Section 2.3 of Chapter 2).

In light of Proposition 4.8.5 (our semidefinite represetation of $\mathrm{conv}(\mathcal{TM}_{m,T}^{\mathbb{R}})$), an alternative way to write the statement we are trying to prove is that

$$\mathrm{conv}(\mathcal{M}_{3,T}) = \tilde{\mathcal{A}}(\mathrm{conv}(\mathcal{TM}_{4,T}^{\mathbb{R}})).$$

We use this version of the statement in what follows.

Suppose that $Q \in SO(3)$ and $\omega \in [0, 2\pi)$. Then since $Q \in \mathrm{conv}\, SO(3)$ there is some $Z \succeq 0$ with $\mathrm{tr}(Z) = 1$ such that $\mathcal{A}(Z) = Q$ and so that

$$(\cos(\omega), \sin(\omega), Q, Q\cos(\omega), Q\sin(\omega), \ldots, Q\cos(T\omega), Q\sin(T\omega)) = \tilde{\mathcal{A}}(Z, Z\cos(\omega), Z\sin(\omega), \ldots, Z\cos(T\omega), Z\sin(T\omega)).$$

Since the left hand side is an arbitrary element of $\mathcal{M}_{3,T}$ and the right hand side is an element of $\tilde{\mathcal{A}}(\mathcal{TM}^{\mathbb{R}}_{4,T})$ we have that $\mathrm{conv}(\mathcal{M}_{3,T}) \subseteq \mathrm{conv}(\tilde{\mathcal{A}}(\mathcal{TM}^{\mathbb{R}}_{4,T})) = \tilde{\mathcal{A}}(\mathrm{conv}(\mathcal{TM}^{\mathbb{R}}_{4,T}))$.

To establish the reverse inclusion, suppose that $Z \in \mathcal{S}^4_+$ (with $\mathrm{tr}(Z) = 1$) and $\omega \in [0, 2\pi)$ and let $Q = \mathcal{A}(Z) \in \mathrm{conv}\, SO(3)$. Then

$$\begin{aligned}\tilde{\mathcal{A}}((Z, Z\cos(\omega), Z\sin(\omega), \ldots, Z\cos(T\omega), Z\sin(T\omega))) = \\ (\cos(\omega), \sin(\omega), Q, Q\cos(\omega), Q\sin(\omega), \ldots, Q\cos(T\omega), Q\sin(T\omega)).\end{aligned} \tag{4.8.14}$$

Since $Q \in \mathrm{conv}\, SO(3)$, it follows that the right hand side of (4.8.14) is in $\mathrm{conv}(\mathcal{M}_{3,T})$. Since the left hand side of (4.8.14) is an arbitrary element of $\tilde{\mathcal{A}}(\mathcal{TM}^{\mathbb{R}}_{4,T})$ it follows that $\tilde{\mathcal{A}}(\mathcal{TM}^{\mathbb{R}}_{4,T}) \subseteq \mathrm{conv}(\mathcal{M}_{3,T})$. It then follows that $\tilde{\mathcal{A}}(\mathrm{conv}(\mathcal{TM}^{\mathbb{R}}_{4,T})) = \mathrm{conv}(\tilde{\mathcal{A}}(\mathcal{TM}^{\mathbb{R}}_{4,T})) \subseteq \mathrm{conv}(\mathcal{M}_{3,T})$, completing the proof of the reverse inclusion. □

Chapter 5

Rounding semidefinite relaxations for pairwise optimization problems

5.1 Introduction

Tractable semidefinite relaxations are used to give globally valid approximations to difficult optimization problems. For many instances they are *exact*, in the sense that the semidefinite relaxation has an optimal solution from which we can efficiently construct an optimal solution to the original problem. When a semidefinite relaxation fails to be exact, we would like to use its solution to help us find feasible solutions to the original problem that are close to optimal. This is the idea of *rounding* a semidefinite relaxation.

Since the work of Goemans and Williamson [54] giving the best known approximation algorithm for the MAX-CUT problem, the idea of devising approximation algorithms for combinatorial optimization problems via constructing semidefinite relaxations and associated rounding schemes has led to many new approximation algorithms. More recently, this approach has been applied to continuous optimization problems [12, 124, 7]. Many of the semidefinite relaxations proposed are now well understood, being typically the first level of a hierarchy of semidefinite relaxations based on sums-of-squares [92, 73]. On the other hand, we do not yet have systematic methods to construct and analyze rounding schemes.

In this chapter we focus on a family of optimization problems that involve many variables interacting pairwise, each of which takes values in some set $\mathcal{X}$ of $m \times d$ matrices that are contractions (Problem 5.2.1 to follow). We consider the problem of *designing* rounding schemes for a particular simple semidefinite relaxation of these problems. The design aim is to maximize the achievable approximation ratio over problem instances with an objective function defined by a positive semidefinite matrix. Our main result describes, explicitly, the structure of the optimal rounding schemes when the set $\mathcal{X}$ obeys a certain symmetry property (see Definition 5.2.11 to follow). It generalizes and unifies many special cases appearing in the literature (see Section 5.4.1 to follow), as well as providing a systematic way to design rounding schemes for new problems.

Perhaps more importantly, it reduces the problem of designing a rounding scheme for these problems to a finite dimensional optimization problem related to the *geometry* of $\mathcal{X}$. We call this problem the *normalized maximum width problem* (Problem 5.2.10 to follow). The aim is to find a linear transformation Y with Frobenius norm one so that the Gaussian width (a natural measure of the size of a set) of $Y\mathcal{X}$ is maximized. We show how to construct a rounding scheme from any feasible point of the normalized maximum width problem. Under this correspondence, the objective value in the normalized maximum width problem can be directly related with the approximation ratio achieved by the rounding scheme. Moreover, any optimal point for the normalized maximum width problem gives a rounding scheme that is optimal in a sense that we make precise in Section 5.2.1.

■ 5.1.1 Notation

We briefly summarize notation used throughout the chapter that is not explicitly defined elsewhere. Throughout this chapter we write many statements that are valid over the real numbers $\mathbb{R}$ and over the complex numbers $\mathbb{C}$. To do this concisely we use $\mathbb{K}$ to denote a field that is either $\mathbb{R}$ or $\mathbb{C}$. If $X \in \mathbb{K}^{m\times d}$ then $X^* \in \mathbb{K}^{d\times m}$ is the transpose if $\mathbb{K} = \mathbb{R}$ and the transpose of the complex conjugate if $\mathbb{K} = \mathbb{C}$. We denote real symmetric $n \times n$ matrices by $\mathcal{H}^n_{\mathbb{R}}$ and complex Hermitian $n \times n$ matrices by $\mathcal{H}^n_{\mathbb{C}}$. We collectively refer to these as *self-adjoint matrices* when the field is left unspecified. We denote $n\times n$ symmetric positive semidefinite matrices by $\mathcal{H}^n_{\mathbb{R},+}$ and $n\times n$ complex Hermitian positive semidefinite matrices by $\mathcal{H}^n_{\mathbb{C},+}$. We use the notation $[n]$ to denote the set $\{1, 2, \ldots, n\}$.

Recall from Section 2.3 that for $X, Y \in \mathbb{K}^{m\times d}$ we define a *real-valued* inner product $\langle\cdot,\cdot\rangle : \mathbb{K}^{m\times d} \times \mathbb{K}^{m\times d} \to \mathbb{R}$ by

$$\langle X, Y\rangle = \mathrm{Re}\,[\mathrm{tr}(X^*Y)].$$

We note that if $\mathbb{K} = \mathbb{C}$ this is the usual Euclidean inner product when we regard $\mathbb{C}^{m\times d}$ as a real $2md$-dimensional vector space. Similarly we define the Frobenius norm to be $\|X\|_F = \langle X, X\rangle^{1/2} = \mathrm{Re}\,[\mathrm{tr}(X^*X)]^{1/2}$ for any $X \in \mathbb{K}^{m\times d}$.

■ 5.1.2 Chapter Outline

The remainder of the chapter is organized as follows. In Section 5.2 we describe, precisely, the main problems of interest in the chapter, and state our main result (Theorem 5.2.12 to follow). To illustrate the result, we apply it to the example of binary quadratic optimization, showing how it recovers the rounding scheme of Goemans and Williamson [54] and the associated approximation result of Nesterov [88]. Our main result is stated in terms of a geometric optimization problem called the normalized maximum width problem. Section 5.3 is devoted to discussing this problem. The normalized maximum with problem may be of interest in its own right, and Section 5.3 is self-contained and could be read independently of the rest of the chapter. Section 5.4 describes related work some examples applying our main result.

Sections 5.5 and 5.6 are devoted to proving our main result (Theorem 5.2.12 to follow). The main result involves characterizing a quantity we call the positive semidefinite integrality gap (see Definition 5.2.4 to follow). Part of the main result involves establishing an upper bound on this quantity. This is the subject of Section 5.5. The other part of the main result involves establishing a lower bound on this quantity, via constructing a rounding scheme. This is the subject of Section 5.6. The proof of the main result appears explicitly in Section 5.6, drawing on results from earlier sections.

Finally Section 5.7 describes a variation (Theorem 5.7.1 to follow) on our main result. This variation considers how the approximation guarantees change when we are only able to approximately compute the optimal rounding scheme.

■ 5.2 Problem statements and main result

In this section we explain the basic problems we address in this chapter, and state our main result.

■ 5.2.1 Pairwise quadratic optimization problems and a semidefinite relaxation

In this chapter we focus on semidefinite relaxations and associated rounding methods for a class of quadratic optimization problems over a collection of variables that interact pairwise.

Problem 5.2.1. *Let $\mathcal{X} \subset \mathbb{K}^{m\times d}$ (where $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$) be a compact set of contractions (i.e. satisfying $X^*X \preceq I$ for all $X \in \mathcal{X}$). Let $\mathcal{C} \subset \mathcal{H}^{nd}_{\mathbb{K}}$ be a collection of $nd \times nd$ self-adjoint matrices (thought of as $n \times n$ matrices consisting of $d \times d$ blocks). For each $C \in \mathcal{C}$ define the optimization problem*

$$\textsc{opt}_{\mathcal{X}}(C) := \max_{X:[n]\to\mathcal{X}} \sum_{i,j=1}^{n} \langle C_{ij}, X_i^* X_j\rangle \tag{5.2.1}$$

where C_{ij} denotes the $d \times d$ block of C indexed by i and j.

We note that the family of instances $\mathcal{C}$ plays no explicit role in the optimization problem (5.2.1). We include it in the statement to emphasize that for a given set $\mathcal{X}$ we are interested not just in single instances of the corresponding pairwise optimization problem, but in families of instances described by $\mathcal{C}$. In this chapter we focus on the case where $\mathcal{C} = \mathcal{H}^{nd}_{\mathbb{K},+}$ is the cone of positive semidefinite matrices.

Also observe that the notation $X : [n] \to \mathcal{X}$ in (5.2.1) just means a collection $X_1, X_2, \ldots, X_n$ of n elements of $\mathcal{X}$. We use this functional notation because it generalizes nicely to the situation where $[n]$ is replaced with a measure space (as occurs in Section 5.5 to follow).

With a view towards convex relaxations, we can also think of (5.2.1) as the problem of solving

$$\textsc{opt}_{\mathcal{X}}(C) = \max_{Z \in \mathcal{G}_{\mathcal{X}}} \langle C, Z \rangle \tag{5.2.2}$$

where we define

$$\mathcal{G}^n_{\mathcal{X}} := \operatorname{conv} \left\{ \begin{bmatrix} X_1^* X_1 & X_1^* X_2 & \cdots & X_1^* X_n \\ X_2^* X_1 & X_2^* X_2 & \cdots & X_2^* X_n \\ \vdots & \vdots & \ddots & \vdots \\ X_n^* X_1 & X_n^* X_2 & \cdots & X_n^* X_n \end{bmatrix} \; : \; X : [n] \to \mathcal{X} \right\}$$

to be the convex hull of the set of (generalized) Gram matrices of elements of $\mathcal{X}$.

Semidefinite relaxation

We focus on a single, simple, semidefinite relaxation for (5.2.1) that is valid for all sets $\mathcal{X} \subset \mathbb{K}^{m \times d}$ (for fixed d and $\mathbb{K}$) of contractions. To relaxation is obtained by finding a simple outer approximation for the set $\mathcal{G}^n_{\mathcal{X}}$. Indeed observe that

$$\mathcal{G}^n_{\mathcal{X}} \subseteq \{ Z \in \mathcal{H}^{nd}_{\mathbb{K},+} : Z_{ii} \preceq I \} =: \mathcal{G}^{n,d,\mathbb{K}}_{\text{sdp}}.$$

Note that $\mathcal{G}^{n,d,\mathbb{K}}_{\text{sdp}}$ is the feasible region of a semidefinite optimization problem. Hence $\textsc{opt}_{\mathcal{X}}(C)$ is always bounded above by the optimal value $\textsc{opt}^{d,\mathbb{K}}_{\text{sdp}}(C)$ of the following semidefinite optimization problem:

$$\textsc{opt}^{d,\mathbb{K}}_{\text{sdp}}(C) := \max_{Z \in \mathcal{H}^{nd}_{\mathbb{K}}} \sum_{i,j=1}^{n} \langle C_{ij}, Z_{ij} \rangle \quad \text{s.t.} \quad Z \succeq 0 \;\text{ and }\; Z_{ii} \preceq I \;\text{ for } i \in [n]. \tag{5.2.3}$$

Again we can express this more compactly as

$$\textsc{opt}^{d,\mathbb{K}}_{\text{sdp}}(C) = \max_{Z \in \mathcal{G}^{n,d,\mathbb{K}}_{\text{sdp}}} \langle C, Z \rangle. \tag{5.2.4}$$

We frequenly omit the parameters d and $\mathbb{K}$ from the notation $\mathcal{G}^{n,d,\mathbb{K}}_{\text{sdp}}$ and the notation $\textsc{opt}^{d,\mathbb{K}}_{\text{sdp}}$ when they are clear from the context.

5.2.2 Key terminology and questions

We now outline the questions we address in this chapter. These questions focus on the relationships between the original pairwise optimization problem (5.2.1) and its semidefinite relaxation (5.2.3) both in terms of the optimal objective value and optimal or near-optimal feasible points for each problem. To describe the problems of interest concisely and precisely, we define the key notions of *integrality gap*, *randomized rounding*, and *approximation ratio* in the present setting. We do so for general families $\mathcal{C}$ of instances, later specializing to the case of positive semidefinite objective functions. We note that these terms are borrowed from the literature on (discrete) approximation algorithms where variables are often constrained to be integer-valued, hence the use of the words 'integrality' and 'rounding'.

Integrality gap

The integrality gap is the worst case (over a family of problem instances) ratio between the optimal objective value for the original problem (5.2.1) and its semidefinite relaxation (5.2.3). Such a ratio only makes sense for families of objective functions for which the functions $\textsc{opt}_{\mathcal{X}}(\cdot)$ and $\textsc{opt}_{\mathrm{sdp}}(\cdot)$ have certain non-negativity properties.

Definition 5.2.2. Given a set $\mathcal{X} \subset \mathbb{K}^{m\times d}$ of contractions and a positive integer n, let $\mathcal{C}$ be a subset of $\mathcal{H}^{nd}_{\mathbb{K}}$. The family $\mathcal{C}$ is *approximable with respect to* $\mathcal{X}$ if $\textsc{opt}_{\mathcal{X}}(C) \geq 0$ for all $C \in \mathcal{C}$ and $\textsc{opt}_{\mathrm{sdp}}(C) > 0$ for all $C \in \mathcal{C} \setminus \{0\}$.

We now define the integrality gap for approximable families of pairwise optimization problems.

Definition 5.2.3 (Integrality gap)**.** Given a set $\mathcal{X} \subset \mathbb{K}^{m\times d}$ of contractions and a family $\mathcal{C}$ of objective functions that is approximable with respect to $\mathcal{X}$, the *integrality gap* over $\mathcal{C}$ between the original optimization problem (5.2.1) and its semidefinite relaxation (5.2.3) is

$$\alpha_{\mathcal{X}}(\mathcal{C}) := \inf_{C \in \mathcal{C}\setminus\{0\}} \frac{\textsc{opt}_{\mathcal{X}}(C)}{\textsc{opt}_{\mathrm{sdp}}(C)}. \tag{5.2.5}$$

The operational meaning of the integrality gap is the following. Suppose we solve the semidefinite relaxation (5.2.3) as a proxy for the hard optimization problem (5.2.1). The semidefinite relaxation always gives an upper bound on the optimal value of the original problem. The integrality gap describes how good (or bad) this upper bound is for the family of problems $\mathcal{C}$ since it satisfies an inequality of the form

$$\alpha_{\mathcal{X}}(\mathcal{C})\textsc{opt}_{\mathrm{sdp}}(C) \leq \textsc{opt}_{\mathcal{X}}(C) \leq \textsc{opt}_{\mathrm{sdp}}(C) \quad \text{for all } C \in \mathcal{C}. \tag{5.2.6}$$

Observe that for any set $\mathcal{X}$ of contractions and any $\mathcal{C}$ that is approximable with respect to $\mathcal{X}$ we have that $0 \leq \alpha_{\mathcal{X}}(\mathcal{C}) \leq 1$. The case where $\mathcal{C} = \mathcal{H}^{nd}_{\mathbb{K},+}$ is the set of $nd \times nd$

positive semidefinite matrices is the focus of this chapter, especially in the limit as $n \to \infty$ (for fixed d). As such, we introduce special terminology and notation for this case.

Definition 5.2.4. Given a set $\mathcal{X} \subset \mathbb{K}^{m \times d}$ of contractions, the *positive semidefinite integrality gap* is

$$\alpha_{\mathcal{X}} := \lim_{n \to \infty} \alpha_{\mathcal{X}}(\mathcal{H}^{nd}_{\mathbb{K},+}). \tag{5.2.7}$$

This limit exists because the sequence $\alpha_{\mathcal{X}}(\mathcal{H}^{nd}_{\mathbb{K},+})$ is non-increasing (since the family of objective functions becomes larger as n increases) and bounded below by zero (since for any positive integer n, $\mathcal{H}^{nd}_{\mathbb{K},+}$ is approximable with respect to any set $\mathcal{X}$ of contractions). Given this definition we state the first central problem of this chapter.

Problem 5.2.5 (Integrality gap)**.** *Given any set $\mathcal{X} \subset \mathbb{K}^{m \times d}$ of contractions find the corresponding positive semidefinite integrality gap $\alpha_{\mathcal{X}}$.*

Rounding

The integrality gap $\alpha_{\mathcal{X}}(\mathcal{C})$ carries two pieces of information. First, it tells us that there is a sequence of difficult instances $C_p \in \mathcal{C}$, such that the optimal value of the original problem is at least a factor of $\alpha_{\mathcal{X}}(\mathcal{C})$ smaller than the optimal value of the semidefinite relaxation. On the other hand, it also tells us that given any problem instance $C \in \mathcal{C}$ and a corresponding solution Z to the semidefinite relaxation (5.2.3), it must be possible to construct a feasible point $\hat{X} : [n] \to \mathcal{X}$ for the original optimization problem (5.2.1) so that

$$\kappa \sum_{i,j=1}^{n} \langle C_{ij}, Z_{ij} \rangle \leq \sum_{i,j=1}^{n} \langle C_{ij}, \hat{X}_i^* \hat{X}_j \rangle \tag{5.2.8}$$

where $\kappa = \alpha_{\mathcal{X}}(\mathcal{C})$. Indeed if this were not the case, the infimum in (5.2.5) would be smaller. A map from $\mathcal{G}^n_{\text{sdp}}$ to feasible points $\hat{X} : [n] \to \mathcal{X}$ for the original problem is called a *rounding scheme*. Note that a rounding scheme allows us to explicitly map solutions of the semidefinite relaxation to feasible points of the original optimization problem that have near-optimal (assuming κ is close to one) value.

Randomized rounding

It is often fruitful to allow randomized rounding schemes. These are maps R that assign to each element $Z \in \mathcal{G}^n_{\text{sdp}}$ a random variable $R(Z) : [n] \to \mathcal{X}$. We assess their quality by considering inequalities of the form in (5.2.8) in expectation.

Definition 5.2.6. A randomized rounding scheme R *achieves an approximation ratio* *of* $\kappa(R,\mathcal{C})$ *with respect to* $\mathcal{C} \subset \mathcal{H}_{\mathbb{K}}^{nd}$ if

$$\kappa(R,\mathcal{C}) \sum_{i,j=1}^{n} \langle C_{ij}, Z_{ij}\rangle \leq \sum_{i,j=1}^{n} \langle C_{ij}, \mathbb{E}[R(Z)_i^* R(Z)_j]\rangle$$

for all $C \in \mathcal{C}$ and all $Z \in \mathcal{G}_{\mathrm{sdp}}^n$.

An alternative view of a randomized rounding scheme R is that it defines a map $F_R : \mathcal{G}_{\mathrm{sdp}}^n \to \mathcal{G}_{\mathcal{X}}^n$ by

$$F_R(Z) = \mathbb{E}\left[\begin{bmatrix} R(Z)_1^* R(Z)_1 & R(Z)_1^* R(Z)_2 & \cdots & R(Z)_1^* R(Z)_n \\ R(Z)_2^* R(Z)_1 & R(Z)_2^* R(Z)_2 & \cdots & R(Z)_2^* R(Z)_n \\ \vdots & \vdots & \ddots & \vdots \\ R(Z)_n^* R(Z)_1 & R(Z)_n^* R(Z)_2 & \cdots & R(Z)_n^* R(Z)_n \end{bmatrix}\right] \in \mathcal{G}_{\mathcal{X}}^n. \tag{5.2.9}$$

Hence if R achieves approximation ratio $\kappa(R,\mathcal{C})$ then

$$\kappa(R,\mathcal{C})\langle C, Z\rangle \leq \langle C, F_R(Z)\rangle \leq \mathrm{OPT}_{\mathcal{X}}(C) \tag{5.2.10}$$

for all $C \in \mathcal{C}$ and all $Z \in \mathcal{G}_{\mathrm{sdp}}$. Choosing Z to be an optimal point of the semidefinite relaxation with objective function defined by C, we see that $\kappa(R,\mathcal{C})\mathrm{OPT}_{\mathrm{sdp}}(C) \leq \langle C, F_R(Z)\rangle \leq \mathrm{OPT}_{\mathcal{X}}(C)$. Hence for any randomized rounding scheme, any achievable approximation ratio is a lower bound on the corresponding integrality gap.

Since our focus is on the case where $\mathcal{C} = \mathcal{H}_{\mathbb{K},+}^{nd}$ and n is allowed to grow, we need to understand the approximation ratios of a sequence of rounding schemes.

Definition 5.2.7. A sequence of randomized rounding schemes (R_n) *achieves a positive semidefinite approximation ratio* of κ_∞ if each rounding scheme R_n in the sequence achieves an approximation ratio of $\kappa(R_n, \mathcal{H}_{\mathbb{K},+}^{nd})$ (see Definition 5.2.6) and $\kappa(R_n, \mathcal{H}_{\mathbb{K},+}^{nd}) \geq \kappa_\infty$ for all n.

The key constructive question we address in this chapter is the following.

Problem 5.2.8. *Given any set* $\mathcal{X} \subset \mathbb{K}^{m\times d}$ *of contractions* design *a sequence of randomized rounding schemes with positive semidefinite approximation ratio equal to the positive semidefinite integrality gap* $\alpha_{\mathcal{X}}$*.*

Throughout the chapter we call a rounding scheme *optimal* if it satisfies the criterion in Problem 5.2.8. Our main result establishes the structure of optimal rounding schemes whenever $\mathcal{X}$ has an additional symmetry property (Definition 5.2.11 to follow).

Local randomized rounding schemes

In this chapter we search for such a sequence (R_n) of rounding schemes among a particular class that we call *local randomized rounding schemes*. Local randomized rounding schemes are parameterized by a function $\hat{X} : \mathbb{K}^{p\times d} \to \mathcal{X}$ (defined up to sets of Gaussian measure zero). Given such a function, and an optimal solution $Z \in \mathcal{H}^{nd}_{\mathbb{K},+}$ of the semidefinite relaxation satisfying $Z_{ii} = I$[1] for $i \in [n]$, we round by carrying out the procedure described in Algorithm 5.1 to follow.

Algorithm 5.1 The local randomized rounding scheme defined by $\hat{X} : \mathbb{K}^{p\times d} \to \mathcal{X}$ applied to a solution Z of the semidefinite relaxation.

Input: A positive semidefinite matrix $Z \in \mathcal{H}^{nd}_{\mathbb{K},+}$ satisfying $Z_{ii} = I$ for $i \in [n]$.

1. Sample a zero-mean Gaussian matrix $W = \begin{bmatrix} W_1 & W_2 & \cdots & W_n \end{bmatrix} \in \mathbb{K}^{p\times nd}$ such that the rows of W are i.i.d. with covariance Z.
2. Define the random variable $R_n(Z) : [n] \to \mathcal{X}$ by

$$[R_n(Z)]_i = \hat{X}(W_i).$$

Output: The random variable $R_n(Z)$

We refer to such schemes as 'local' because we apply the same deterministic map $\hat{X}$ to each of the W_i, rather than allowing a map that arbitrarily combines the W_i.

■ 5.2.3 The normalized maximum width problem

Our main results are described in terms of an auxiliary optimization problem that plays a central role in this chapter. As such, we briefly state and discuss this problem before stating our main results. Section 5.3 is devoted to a more detailed study of the problem. The optimization problem takes as input a set $\mathcal{X} \subset \mathbb{K}^{m\times d}$ and searches over linear transformations of this set with Frobenius norm one that maximize a natural notion of the width of the set.

We begin by defining the appropriate notion of width. Here, and throughout, we denote by $\gamma^{\mathbb{K}}_{p\times d}$ the standard Gaussian probability measure on $\mathbb{K}^{p\times d}$. Explicitly this is the measure with density

$$\begin{cases} (2\pi)^{-pd/2}\exp(-\frac{1}{2}\mathrm{tr}(U^*U)) & \text{if } \mathbb{K} = \mathbb{R} \\ \pi^{-pd}\exp(-\mathrm{tr}(U^*U)) & \text{if } \mathbb{K} = \mathbb{C}. \end{cases}$$

with respect to Lesbegue measure. When the dimensions and the field are clear from

[1] Such a solution always exists when the objective function is defined by a positive semidefinite matrix.

the context, we denote this measure by γ.

Definition 5.2.9. Given a set $\mathcal{X} \subset \mathbb{K}^{p\times d}$ the *Gaussian width* is

$$w(\mathcal{X}) := \mathbb{E}_{U\sim\gamma}[\max_{X\in\mathcal{X}}\langle X, U\rangle].$$

In other words, the Gaussian width is the expected maximum of a random (standard Gaussian) linear functional over the set $\mathcal{X}$. We can also rephrase this in terms of the *support function* of $\mathcal{X}$, the convex function defined as

$$h_{\mathcal{X}}(U) := \max_{X\in\mathcal{X}}\langle X, U\rangle. \tag{5.2.11}$$

The Gaussian width can then be expressed concisely as $\mathbb{E}_\gamma[h_{\mathcal{X}}]$.

We now describe the *normalized maximum width problem*, the problem with respect to which all the main results of this chapter are expressed.

Problem 5.2.10 (Normalized maximum width)**.** *If* $\mathcal{X} \subset \mathbb{K}^{m\times d}$ *then the* normalized maximum width problem *for* $\mathcal{X}$ *is*

$$w_\star(\mathcal{X}) := \sup_{\substack{p\in\mathbb{N}\\ p\geq m}} \max_{\substack{Y\in\mathbb{K}^{m\times p}\\ \|Y\|_F=1}} w(Y^*\mathcal{X}). \tag{5.2.12}$$

We discuss this problem in Section 5.3, including explicitly solving some examples, and giving a number of reformulations. For now we note only two properties of the problem. First, one can show (see Lemma 5.3.1) that there is always an optimal Y for (5.2.12) with $p = m$ and $Y \succeq 0$. We take the apparently more general formulation (5.2.12) as the definition because the extra flexibility is useful. Second, note that the function $Y \mapsto w(Y^*\mathcal{X})$ is convex. Hence the normalized maximum width problem (for fixed p) involves maximizing a convex function over the unit sphere. Typically such problems are not easy to solve globally. A main aim of Section 5.3 is to identify families of sets $\mathcal{X}$ for which the normalized maximum width problem can be reformulated as a convex optimization problem.

■ 5.2.4 Symmetry assumption on $\mathcal{X}$

Our main result (Theorem 5.2.12 to follow) applies to sets $\mathcal{X} \subset \mathbb{K}^{m\times d}$ of contractions that satisfy an additional symmetry assumption. The assumption is *always* satisfied when $d = 1$ (see the comments after Definition 5.2.11 to follow). It is only a restriction for $d > 1$.

To state the assumption we need some basic terminology about representations of groups (see Section 2.6 of Chapter 2). Recall from Section 2.6 that a unitary representation of a group G on $\mathbb{C}^d$ is a homomorphism $\rho : G \to U(d)$. A subspace (over

$\mathbb{C}$[2]) V of $\mathbb{C}^d$ is an *invariant subspace* if $\rho(g)V = V$ for all $g \in G$. A representation $\rho : G \to U(d)$ is *irreducible over* $\mathbb{C}$ if its only invariant subspaces (over $\mathbb{C}$) are $\{0\}$ and $\mathbb{C}^d$. We have introduced these terms for unitary representations on $\mathbb{C}^d$. The same terminology makes sense for orthogonal representation on $\mathbb{R}^d$, by thinking of them as unitary representations that act on $\mathbb{C}^d$ (by acting separately on the real and imaginary parts).

We now formally state our symmetry assumption on $\mathcal{X}$.

Definition 5.2.11 (Right symmetry)**.** We call a set $\mathcal{X} \subset \mathbb{K}^{m\times d}$ *right symmetric* if there is a group G and an irreducible (over $\mathbb{C}$) orthogonal/unitary representation ρ such that $\mathcal{X}\rho(g) = \mathcal{X}$ for all $g \in G$.

If $\mathcal{X} \subset \mathbb{K}^{m\times 1}$ (i.e. if $d = 1$) then $\mathcal{X}$ is *automatically right symmetric*. Indeed if we take $G = \{e\}$ to be the group with one element, and $\rho : G \to \mathbb{K}$ to be the trivial representation $\rho(e) = 1$, this is an irreducible representation that leaves $\mathcal{X}$ invariant. In Section 5.4 we see numerous non-trivial examples of sets $\mathcal{X}$ that are right symmetric.

5.2.5 Main result

We are now in a position to state our main result for this chapter. This result characterizes the positive semidefinite integrality gap and describes a corresponding optimal rounding scheme for pairwise quadratic optimization problems over a compact set $\mathcal{X} \subset \mathbb{K}^{m\times d}$ of contractions that are right symmetric (see Definition 5.2.11). In particular it shows that we can construct rounding schemes from feasible points Y of the normalized maximum width problem in such a way that the approximation ratio they achieve is related to the objective value $w(Y^*\mathcal{X})$ for the normalized maximum width problem. Furthermore an optimal rounding scheme is obtained by performing this construction for a maximizer of the normalized maximum width problem.

Theorem 5.2.12. *If $\mathcal{X} \subset \mathbb{K}^{m\times d}$ (where $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$) is a compact right symmetric set of contractions then the following hold.*

1. *The positive semidefinite integrality gap is $\alpha_{\mathcal{X}} = \left[\frac{w_\star(\mathcal{X})}{d}\right]^2$.*
2. *If Y is any $m \times p$ matrix with $\|Y\|_F = 1$ then the local randomized rounding scheme specified by $\hat{X} : \mathbb{K}^{p\times d} \to \mathcal{X}$ defined (almost everywhere) by*

$$\hat{X}(U) = \arg\max_{X\in\mathcal{X}} \langle X, YU\rangle, \tag{5.2.13}$$

achieves an approximation ratio of $\left[\frac{w(Y^*\mathcal{X})}{d}\right]^2$.

[2]Here V is closed under addition and multiplication by *complex* scalars.

3. *If Y is any argument of maximum in the normalized maximum width problem (5.2.12) the rounding scheme specified in part 2. achieves the optimal approximation ratio of* $\left[\frac{w_\star(\mathcal{X})}{d}\right]^2$.

We restate and prove Theorem 5.2.12 in Section 5.6. The proof is based on the results developed in Sections 5.5 and 5.6.

We now illustrate Theorem 5.2.12 by applying it to the case $\mathcal{X} = \{-1, 1\}$. Section 5.4 gives more examples of applying Theorem 5.2.12 to obtain explicit approximation ratios for various problems, including all the special cases of this problem (of which we are aware) that have been previously studied in the literature.

Example 5.2.13. Suppose $\mathcal{X} = \{-1, 1\} \subset \mathbb{R}^{1\times 1}$. Here $m = 1$ and $d = 1$. This is a compact set. It is right symmetric since $d = 1$ (see the comments after Definition 5.2.11). The maximum width problem for this set is

$$w_\star(\{-1,1\}) = \sup_{\substack{p\in\mathbb{N}\\ p\geq 1}} \max_{\substack{Y\in\mathbb{R}^{1\times p}\\ \|Y\|_F=1}} w(Y^*\{-1,1\}) \tag{5.2.14}$$

Here we have used the definition of the normalized maximum width problem in (5.2.12). Rather than working directly with this definition, we make use of the general comment made after (5.2.12) (and proved in Lemma 5.3.1 to follow) that we can always find an optimum with $p = m$ and $Y \succeq 0$. In this case $p = m = 1$ so we deduce that there is an optimal Y that is 1×1, non-negative, and has Frobenius norm 1. In other words taking $p = 1$ and $Y = 1$ gives an optimal point for (5.2.14). Then by the definition of Gaussian width (Definition 5.2.9)

$$w_\star(\{-1,1\}) = w(1\cdot\{-1,1\}) = \mathbb{E}_{U\sim\gamma^{\mathbb{R}}_{1\times 1}}[\max_{X\in\{-1,1\}}\langle X, U\rangle] = \mathbb{E}_{U\sim\gamma^{\mathbb{R}}_{1\times 1}}[|U|].$$

The expected value of the absolute value of a standard Gaussian random variable is $\sqrt{\frac{2}{\pi}}$, showing that $w_\star(\{-1,1\}) = \sqrt{\frac{2}{\pi}}$.

Substituting this into the first part of Theorem 5.2.12 we see that the positive semidefinite integrality gap is $\left[\frac{w_\star(\{-1,1\})}{1}\right]^2 = \frac{2}{\pi}$ (as established in [88]). We now describe the corresponding optimal rounding scheme by applying parts 2 and 3 of Theorem 5.2.12 with $p = 1$ and $Y = 1$ (which are optimal for (5.2.14)). In this case the optimal rounding scheme is the local randomized rounding scheme (Algorithm 5.1) specified by the function $\hat{X} : \mathbb{R}^{1\times 1} \to \{-1, 1\}$ defined (almost everywhere) by

$$\hat{X}(U) = \arg\max_{X\in\{-1,1\}} \langle X, 1\cdot U\rangle = \operatorname{sgn}(U)$$

the mapping that takes value 1 when $U > 0$ and -1 when $U < 0$. This is the classical hyperplane rounding scheme of Goemans and Williamson [54].

To conclude the section we make a number of general remarks about Theorem 5.2.12.

1. Theorem 5.2.12 characterizes the integrality gap in terms of the normalized maximum width of $\mathcal{X}$, and specifies the approximation ratio of each of a family of rounding schemes (including optimal ones) parameterized by matrices Y with Frobenius norm one. The integrality gap, and the approximation ratios achieved by these rounding schemes are constants that are *independent of* n.

2. Note that there is a unique argument of maximum in Equation (5.2.13) except on a set of measure zero (for U). Hence the map $\hat{X}$ is well-defined almost everywhere, which is all we require.

3. To implement the rounding scheme corresponding to a choice of Y, we need only be able to sample Gaussian matrices and solve the problem of maximizing a linear functional over the set $\mathcal{X}$ (to evaluate the function $\hat{X}$). As such, one way to view our results is that whenever we can maximize a linear functional over $\mathcal{X}$, we obtain a constant factor approximation algorithm for any pairwise quadratic optimization problem (with objective function defined by a positive semidefinite matrix) over *any number* of copies of $\mathcal{X}$.

4. If we cannot efficiently maximize a linear functional over $\mathcal{X}$ we can still obtain an overall approximation algorithm if we have access to an approximation algorithm for (5.2.13) with a multiplicative approximation ratio. We discuss this situation in Section 5.7. This observation means that whenever we have an approximation algorithm for maximizing a linear functional over $\mathcal{X}$ we obtain a constant factor approximation algorithm for any pairwise quadratic optimization problem (with objective function defined by a positive semidefinite matrix) over *any number* of copies of $\mathcal{X}$.

■ 5.3 The normalized maximum width problem

In this section we focus on the *normalized maximum width problem* (Problem 5.2.10) that plays a central role in our main results. Indeed finding explicit solutions to the normalized maximum width problem allows us to design explicit rounding schemes for the problem class of interest in this chapter with best possible approximation guarantees. Furthermore, Theorem 5.2.12 tells us that given any feasible point for the normalized maximum width problem we can immediately construct a (possibly non-optimal) rounding scheme and deduce its associated approximation ratio. As such, even if the maximum width problem cannot always be solved globally, any feasible point with a positive objective value still allows us to construct a rounding schemes with a non-trivial approximation guarantee.

This section is self-contained, and may be of interest independent of its relationship to the rounding problems studied in the rest of this chapter. Indeed from this point on in the section no mention is made of these rounding problems.

The normalized maximum width problem (Problem 5.2.10) involves finding a linear transformation (with Frobenius norm one) that puts a set into a position where its Gaussian width is maximized. Extremal positions of convex bodies with respect to various functionals such as width, volume, etc. often arise in convex geometry (see, e.g., [52, 53]). Nevertheless we are not aware of any systematic prior study of this particular problem. We begin by giving a number of simple reformulations of the normalized maximum width problem. The last formulation in Lemma 5.3.1 (to follow) is particularly appealing because in that case the constraint set is a compact convex set. We used this formulation in Example 5.2.13 to simplify the discussion.

Lemma 5.3.1. *Given a set $\mathcal{X} \subset \mathbb{K}^{m\times d}$ we have that*

$$w_\star(\mathcal{X}) := \sup_{\substack{p\in\mathbb{N}\\ p\geq m}} \max_{\substack{Y\in\mathbb{K}^{m\times p}\\ \|Y\|_F=1}} w(Y^*\mathcal{X}) \tag{5.3.1}$$

$$= \max_{\substack{Y\in\mathbb{K}^{m\times p'}\\ \|Y\|_F=1}} w(Y^*\mathcal{X}) \text{ for all } p' \geq m \tag{5.3.2}$$

$$= \max_{\substack{P\in\mathcal{H}_+^m\\ \operatorname{tr}(P)=1}} w(P^{1/2}\mathcal{X}). \tag{5.3.3}$$

Proof. The main observation we need is that if $Q \in \mathbb{K}^{p\times m}$ has orthonormal columns then $w(Q\mathcal{X}) = w(\mathcal{X})$. This holds because if U is a standard $p \times d$ Gaussian matrix then Q^*U is a standard $m \times d$ Gaussian matrix. Hence using the notation $h_{\mathcal{X}}$ for the support function of $\mathcal{X}$ (see (5.2.11)) we see that

$$w(Q\mathcal{X}) = \mathbb{E}_{U\sim\gamma_{p\times d}^{\mathbb{K}}}[h_{Q\mathcal{X}}(U)] = \mathbb{E}_{U\sim\gamma_{p\times d}^{\mathbb{K}}}[h_{\mathcal{X}}(Q^*U)] = \mathbb{E}_{W\sim\gamma_{m\times d}^{\mathbb{K}}}[h_{\mathcal{X}}(W)] = w(\mathcal{X}).$$

With that established the proof is straightforward but a little tedious. Clearly (5.3.1) is greater than or equal to (5.3.2). For the reverse inequality we show that for any $p' \geq m$,

$$\max_{\substack{Y\in\mathbb{K}^{m\times p'}\\ \|Y\|_F=1}} w(Y^*\mathcal{X}) = \max_{\substack{Y\in\mathcal{H}_+^m\\ \|Y\|_F=1}} w(Y^*\mathcal{X}). \tag{5.3.4}$$

To do so, let $Y \in \mathbb{K}^{m\times p'}$ be optimal for the left hand side problem. Let $Y = RV^*$ be its (reduced) polar decomposition, where $R \in \mathcal{H}_+^m$ is positive semidefinite and $V \in \mathbb{K}^{m\times p'}$

has orthonormal columns. Then $w(Y^*\mathcal{X}) = w(VR\mathcal{X}) = w(R\mathcal{X})$ and so R is feasible for the right hand optimization problem in (5.3.4) and has the same objective value. On the other hand, let $Y \in \mathcal{H}_+^m$ be optimal for the right-hand side of (5.3.4). Then $Y\begin{bmatrix} I_{m\times m} & 0_{m\times p'}\end{bmatrix} \in \mathbb{K}^{m\times p'}$ is feasible for the left-hand side of (5.3.4) and has the same objective value.

We now show that (5.3.1) and (5.3.3) are equal. First observe that if P is optimal for (5.3.3) then $Y = P^{1/2}$ has $\|Y\|_F^2 = \mathrm{tr}(Y^*Y) = \mathrm{tr}(P) = 1$ is feasible for (5.3.1) (with $p = m$), establishing that (5.3.3) is at most (5.3.1). For the reverse inequality, from (5.3.4) we can see that (5.3.1) always has an optimal solution Y with $p = m$ and $Y \in \mathcal{H}_+^m$. Taking $P = Y^2$ gives a positive semidefinite matrix with trace one (so feasible for (5.3.3)) such that $w(P^{1/2}\mathcal{X}) = w(Y\mathcal{X})$. □

We state and prove some other simple properties of the normalized maximum width in Section 5.3.2 to follow.

The normalized maximum width problem is qualitatively different based on certain properties of the convex hull of $\mathcal{X}$. Whether or not the convex hull of $\mathcal{X} \subset \mathbb{K}^m$ is a type of convex body called a *zonoid* (see Definition 5.3.2 to follow) affects the properties of the normalized maximum width problem. Similary, whether or not the convex hull of $\mathcal{X} \subset \mathbb{K}^{m\times d}$ is a generalization of a zonoid that we call a *contraction zonoid* (see Definition 5.3.6) similarly affects the properties of the normalized maximum width problem when $d > 1$.

■ 5.3.1 (Contraction) zonoids

Zonoids are a family of convex bodies that are Minkowski sums (or limits of Minkowski sums) of line segments. We focus on centrally symmetric zonoids. There are many equivalent characterizations of zonoids (see, e.g. [15]). For our purposes the most useful is the following characterization of the support functions of zonoids which we take as the definition (see [120, Theorem 1.2]).

Definition 5.3.2. A convex set $\mathcal{Z} \subset \mathbb{R}^m$ is a *centrally symmetric zonoid* if there exists a positive measure μ on $\mathbb{R}^m$ such that the support function $h_{\mathcal{Z}}$ of $\mathcal{Z}$ has a representation as

$$h_{\mathcal{Z}}(u) = \int_{v\in\mathbb{R}^m} |\langle u, v\rangle|\, d\mu(v). \tag{5.3.5}$$

We remark that in general, there are many positive measures μ associated with a given centrally symmetric zonoid $\mathcal{Z}$. Nevertheless, given a centrally symmetric zonoid $\mathcal{Z}$, there is a unique positive measure μ supported on the sphere S^{m-1} that is even, i.e. that satisfies $\mu(-A) = \mu(A)$ [120, Theorem 1.4].

If $\mathcal{Z}$ is a zonoid and we want to keep track of some positive measure μ on $\mathbb{R}^m$ such that (5.3.5) holds, we write $\mathcal{Z}_\mu$ instead of just $\mathcal{Z}$. For any set $\mathcal{Z} \subset \mathbb{R}^m$ we have that

$h_{\mathcal{Z}} = h_{\mathrm{cl}(\mathrm{conv}\,\mathcal{Z})}$. Hence for a non-convex set $\mathcal{Z}$, the closure of the convex hull of $\mathcal{Z}$ is a zonoid if and only if there is a positive measure μ on $\mathbb{R}^m$ such that (5.3.5) holds.

Example 5.3.3 (Zonotopes)**.** A zonoid that is also a polytope is called a *zonotope*. Every zonotope in $\mathbb{R}^m$ is of the form $A[-1,1]^k$ where $A : \mathbb{R}^k \to \mathbb{R}^m$ is a linear map and $[-1,1]^k$ is the hypercube in $\mathbb{R}^k$ [15]. This satisfies the definition in (5.3.5) by taking the measure μ on $\mathbb{R}^m$ to consist of point masses at each of the columns of A.

Example 5.3.4 (Two-dimensional zonotopes)**.** Every two dimensional centrally symmetric polytope is a zonotope, and hence a zonoid [120]. Note that this fails to hold in higher dimensions. In fact a polytope is a zonotope if and only if all of its two-dimensional faces are centrally symmetric [120].

Example 5.3.5 (ℓ_p norm balls for $p \geq 2$)**.** The sets

$$B_p^m := \left\{x \in \mathbb{R}^m : \left(\textstyle\sum_{i=1}^m |x_i|^p\right)^{1/p} \leq 1\right\} \qquad \text{for } 2 \leq p \leq \infty$$

are zonoids. In the case $p = 2$ this is easy to see from the definition we have given since for any $u \in \mathbb{R}^m$

$$h_{B_2^m}(u) = \|u\|_2 = \sqrt{\frac{\pi}{2}} \int_{v \in \mathbb{R}^m} |\langle u, v\rangle| \, d\gamma_{m\times 1}^{\mathbb{R}}(v).$$

To see this it is enough to observe that the right-hand side is positively homogeneous of degree one, invariant under $u \mapsto Qu$ for any orthogonal matrix Q, and that for any unit vector $\hat{u}$ the right-hand side evaluates to one.

It is not so easy to see from our definition that B_p^m is a zonoid when $2 < p < \infty$. This can be deduced from an alternative characterization of the support functions of zonoids in terms of positive definite functions (see, e.g., [15]).

We now introduce an analogue of zonoids for subsets of $\mathbb{K}^{m\times d}$ that generalizes the properties of zonoids we need. We call these *contraction zonoids* and define these convex sets via their support functions. Clearly putting $d = 1$ in the following definition recovers the definition of a centrally symmetric zonoid.

Definition 5.3.6. A convex set $\mathcal{Z} \subset \mathbb{K}^{m\times d}$ is a *contraction zonoid* if there is a positive integer p and a positive measure μ on $\mathbb{K}^{m\times p}$ such that

$$h_{\mathcal{Z}}(U) = \int_{V \in \mathbb{K}^m} \|U^* V\|_* \, d\mu(V) \tag{5.3.6}$$

where $\|X\|_*$ denotes the nuclear norm (i.e. sum of singular values) of a $d \times p$ matrix, or equivalently the support function of $\mathrm{St}_{\mathbb{K}}(d, p)$.

As for zonoids, even if $\mathcal{Z} \subset \mathbb{K}^{m\times d}$ is not convex, if the closure of the convex hull of $\mathcal{Z}$ is a contraction zonoid then $h_{\mathcal{Z}}$ has a representation as in (5.3.6). Again we write

$\mathcal{Z}_\mu$ if the convex hull of $\mathcal{Z}$ is a contraction zonoid and we want to keep track of some measure that is valid for (5.3.6).

Example 5.3.7 (Stiefel manifold)**.** Let $\mathrm{St}_{\mathbb{K}}(d,m) = \{X \in \mathbb{K}^{m\times d} : X^*X = I\}$ be the Stiefel manifold consisting of $m \times d$ matrices with $\mathbb{K}$-orthonormal columns. Each of these sets has support function $h_{\mathrm{St}_{\mathbb{K}}(d,m)}(U) = \|U\|_*$ where $\|U\|_*$ is the nuclear norm or sum of the singular values of the $d \times m$ matrix U. Clearly $\mathrm{St}_{\mathbb{K}}(d,m)$ is a contraction zonoid because we can take μ to be the atomic measure on $\mathbb{K}^{m\times m}$ with an atom at the identity matrix.

The normalized maximum width problem for contraction zonoids

The maximum width problem is qualitatively different for contraction zonoids and sets that are not contraction zonoids because when the convex hull of $\mathcal{X}$ is a contraction zonoid, the objective function in (5.3.3) is concave in P, and so the normalized maximum width problem is then a finite dimensional convex optimization problem.

Theorem 5.3.8. *If $\mathcal{Z}_\mu \subset \mathbb{K}^{m\times d}$ is a contraction zonoid with associated measure μ on $\mathbb{K}^{m\times p}$ then the map $f : \mathcal{H}_+^m \to \mathbb{R}$ defined by $f(P) = w(P^{1/2}\mathcal{Z}_\mu)$ is concave.*

Proof. Interchanging the order of integration in the definition of f we see that

$$\begin{aligned} f(P) &= \mathbb{E}_{U\sim\gamma}\left[\int_{V\in\mathbb{K}^{m\times p}} \|V^*P^{1/2}U\|_*\, d\mu(V)\right] \\ &= \int_{V\in\mathbb{K}^{m\times p}} \mathbb{E}_{U\sim\gamma}[\|V^*P^{1/2}U\|_*]\, d\mu(V). \end{aligned}$$

For fixed $V \in \mathbb{K}^{m\times p}$ the quantity $V^*P^{1/2}U$ is a $p \times d$ Gaussian random matrix with zero mean and i.i.d. columns each having covariance V^*PV. Hence

$$\begin{aligned} f(P) &= \int_{V\in\mathbb{K}^{m\times p}} \mathbb{E}_{W\sim\gamma}[\|(V^*PV)^{1/2}W\|_*]\, d\mu(V) \\ &= \int_{V\in\mathbb{K}^{m\times p}} \mathbb{E}_{W\sim\gamma}\left[\mathrm{tr}\left[(W^*(V^*PV)W)^{1/2}\right]\right]\, d\mu(V). \end{aligned}$$

Now, the function $T \mapsto \mathrm{tr}[T^{1/2}]$ is concave for positive semidefinite T. (One way to see this is to use Davis' characterization of concave spectral functions [35] together with the fact that the permutation invariant function $\sum_{i=1}^m t_i^{1/2}$ on the non-negative orthant is concave.) It then follows that $f(P)$ is a non-negative combination of concave functions and so is concave. □

The fact that f is a concave function of P, and so the normalized maximum width problem is a convex optimization problem, means that for contraction zonoids we can hope to solve this problem globally using numerical routines. Perhaps more importantly, if the contraction zonoid has enough symmetry, we can often deduce the solution of the normalized maximum width problem without doing any computation at all.

Contraction zonoids with symmetry

For zonoids with enough symmetry, we can simplify the normalized maximum width problem, and in some cases even solve it explicitly. This uses the fact that convex optimization problems with symmetry have symmetric solutions.

Lemma 5.3.9. *Suppose $\mathcal{Z}$ is a contraction zonoid in $\mathbb{K}^{m\times d}$, G is a subgroup of the group of $m \times m$ orthogonal/unitary matrices, and $g\mathcal{Z} = \mathcal{Z}$ for all $g \in G$. Then the normalized maximum width problem* (5.3.3) *has an optimal solution P that satisfies $g^*Pg = P$ for all $g \in G$.*

Proof. First we observe that the objective function of (5.3.3) is invariant under the action of G. Indeed for any fixed $g \in G$,

$$w((g^*Pg)^{1/2}\mathcal{Z}) = w(g^*P^{1/2}g\mathcal{Z}) = w(P^{1/2}\mathcal{Z})$$

where the first equality holds since g is orthogonal/unitary, and the second holds because $g\mathcal{Z} = \mathcal{Z}$ and because $w(g\mathcal{X}) = w(\mathcal{X})$ for any orthogonal/unitary g and any set $\mathcal{X}$. Similarly the constraint set $P \succeq 0$ and $\mathrm{tr}(P) = 1$ is invariant under $P \mapsto g^*Pg$. Hence there is an optimal solution that satisfies $g^*Pg = P$ for all $g \in G$. □

Example 5.3.10 (ℓ_p norm balls for $p \geq 2$)**.** Recall from Example 5.3.5 that the unit ℓ_p norm balls in $\mathbb{R}^m$ are zonoids for $p \geq 2$. These are invariant under the group of signed permutations. As such, there is a maximizer for P in the corresponding maximum width problem that satisfies $gPg^T = P$ for all signed permutation matrices g. It follows that P is a multiple of the identity. Since $\mathrm{tr}(P) = 1$, we can conclude that $P = I/m$ is a solution of (5.3.3) when $\mathcal{X}$ is an ℓ_p-norm ball in $\mathbb{R}^m$ with $p \geq 2$.

Example 5.3.11 (The Stiefel manifolds $\mathrm{St}_{\mathbb{K}}(d,m)$)**.** Recall from Example 5.3.7 that the Stiefel manifolds $\mathrm{St}_{\mathbb{K}}(d,m)$ consisting of elements of $\mathbb{K}^{m\times d}$ with orthonormal columns are contraction zonoids. These are invariant under left multiplication by orthogonal/unitary matrices hence satisfy 5.3.9 with g being the full orthogonal/unitary group. Hence there is an optimal solution P to (5.3.3) such that $g^*Pg = P$ for all orthogonal/unitary g. As such $P = I/m$ is a solution of (5.3.3) and so $w_\star(\mathrm{St}_{\mathbb{K}}(d,m)) = \frac{1}{\sqrt{m}}w(\mathrm{St}_{\mathbb{K}}(d,m))$.

We conclude this section by establishing some properties of $w_\star(\mathrm{St}_{\mathbb{K}}(d,m))$ that play an important role in Section 5.5. The following was established (in different language) by Bandeira, Kennedy, and Singer [7].

Lemma 5.3.12. *Fix a positive integer d. Then*

$$w_\star(\mathrm{St}_{\mathbb{K}}(d,m)) \leq d \text{ for all } m \geq d \quad \text{and} \quad \lim_{m\to\infty} w_\star(\mathrm{St}_{\mathbb{K}}(d,m)) = d.$$

Proof. Recall from Example 5.3.11 that $w_\star(\mathrm{St}_{\mathbb{K}}(d,m)) = \frac{1}{\sqrt{m}}w(\mathrm{St}_{\mathbb{K}}(d,m))$. Since the support function of the Stiefel manifold is the nuclear norm we have that

$$\begin{aligned} w_\star(\mathrm{St}_{\mathbb{K}}(d,m)) &= \mathbb{E}_{U\sim\gamma}\left[\tfrac{1}{\sqrt{m}}\|U\|_*\right] \\ &= \mathbb{E}_{U\sim\gamma}\left[\mathrm{tr}\left[\left(\tfrac{1}{m}U^*U\right)^{1/2}\right]\right] \\ &\leq \mathrm{tr}\left[\left(\mathbb{E}_{U\sim\gamma}\left[\tfrac{1}{m}U^*U\right]\right)^{1/2}\right] \\ &= \mathrm{tr}[I_d] = d \end{aligned}$$

where the inequality follows from applying Jensen's inequality to the concave function $T \mapsto \mathrm{tr}[T^{1/2}]$ (restricted to the positive semidefinite cone).

To see that the limit is actually d we use the fact that $\frac{1}{\sqrt{m}}\|U\|_* \geq d\sigma_{\min}\left(\frac{1}{\sqrt{m}}U\right)$ allowing us to use standard results about this the expected smallest singular value of a Gaussian matrix. In particular Gordan's theorem (see, e.g., [132, Theorem 5.32]) tells us that

$$\mathbb{E}_{U\sim\gamma}\left[\sigma_{\min}\left(\frac{1}{\sqrt{m}}U\right)\right] \geq 1 - \sqrt{\frac{d}{m}}.$$

Hence $\mathbb{E}_{U\sim\gamma}[\frac{1}{\sqrt{m}}\|U\|_*] \geq d\mathbb{E}_{U\sim\gamma}[\sigma_{\min}\left(\frac{1}{\sqrt{m}}U\right)] \geq d - d\sqrt{\frac{d}{m}} \to d$ as $m \to \infty$. □

■ 5.3.2 General sets

For sets that are not contraction zonoids, the maximum normalized width problem seems to be more complicated. In this section we use a simple method to obtain lower bounds on the normalized maximum width of a set based on the normalized maximum width of its subsets. This allows us to lower bound the normalized maximum width of some sets that are not contraction zonoids such as the unit norm balls for the ℓ_1 norm, the convex body that is perhaps as far as possible from being a zonoid (being the polar of the hypercube which is the canonical zonotope). We also give characterization of when $P = I/m$ satisfies the first-order optimality conditions of the normalized maximum width problem (5.3.3) and show that this holds whenever $\mathcal{X}$ has certain symmetry properties.

A simple lower bound on $w_\star(\mathcal{X})$

The following observation gives a surprisingly useful lower bound on the normalized maximum width.

Lemma 5.3.13. *Let $\mathcal{X} \subset \mathbb{K}^{m\times d}$ and let $\mathcal{X}' \subset \operatorname{conv}\mathcal{X}$. Then $w_\star(\mathcal{X}) \geq w_\star(\mathcal{X}')$.*

Proof. Let Y be optimal for the normalized maximum width problem for $\mathcal{X}'$, i.e. $w(Y^*\mathcal{X}') = w_\star(\mathcal{X}')$. Then $Y^*\mathcal{X}' \subseteq Y^*\operatorname{conv}\mathcal{X}$ and so

$$\begin{aligned} w_\star(\mathcal{X}') &= w(Y^*\mathcal{X}') \\ &\leq w(Y^*\operatorname{conv}\mathcal{X}) \\ &= w(\operatorname{conv} Y^*\mathcal{X}) \\ &= w(Y^*\mathcal{X}) \\ &\leq w_\star(\mathcal{X}) \end{aligned}$$

where we have used the fact that $w(\operatorname{conv}\mathcal{X}) = w(\mathcal{X})$ for any set $\mathcal{X}$ and the fact that $A\operatorname{conv}(\mathcal{X}) = \operatorname{conv}(A\mathcal{X})$ for any set $\mathcal{X}$ and any linear map A. □

A simple situation in which this result is useful is when a subset of $\mathbb{R}^m$ contains a line segment, i.e. a set of the form $[-v, v] := \operatorname{conv}\{-v, v\}$ where $v \in \mathbb{R}^m$, in its convex hull.

Example 5.3.14 (Centrally symmetric subsets of $\mathbb{R}^{m\times 1}$ with radius one)**.** As an illustration of the previous result, we consider sets $\mathcal{X} \subset \mathbb{R}^{m\times 1}$ that are centrally symmetric (i.e. satisfy $\mathcal{X} = -\mathcal{X}$) and have radius one. Then there some unit vector $v \in \mathcal{X}$ and so the line segment $[-v, v]$ of length two is a subset of $\mathcal{X}$. Hence $w_\star(\mathcal{X}) \geq w_\star([-v, v]) = \sqrt{\frac{2}{\pi}}$.

The result of the previous example applies to the ℓ_1-norm ball, the convex body that is, in some sense, as far as possible from being a zonoid.

Example 5.3.15 (ℓ_1-norm ball)**.** In the case of ℓ_1^m, the unit ball for the ℓ_1-norm in $\mathbb{R}^m$, $h_{\ell_1^m}(u) = \max_{i\in[m]} |u_i|$. Hence the maximum width problem can be expressed as the expected maximum absolute value of a Gaussian process on a set of cardinality m that has covariance of unit trace, i.e.

$$w_\star(\ell_1^m) = \max_{P\succeq 0} \mathbb{E}[\max_{i\in[m]} |U_i|] \quad \text{s.t.} \quad U \sim \mathcal{N}(0, P), \;\; \operatorname{tr}(P) = 1.$$

This quantity is clearly bounded below by $\sqrt{\frac{2}{\pi}}$ by Example 5.3.14. In fact, we conjecture that $w_\star(\ell_1^m) = \sqrt{\frac{2}{\pi}}$ for all m. This holds for $m = 1$ and $m = 2$ because in these cases ℓ_1^m is a zonotope.

We note that taking $Y = \frac{1}{\sqrt{m}}I$ in the normalized maximum width problem is far from optimal in the case of the ℓ_1 norm ball. This is because

$$w\left(\frac{1}{\sqrt{m}}\ell_1^m\right) = \mathbb{E}[\max_{i\in[m]}|U_i|] \quad \text{s.t.} \quad U \sim \mathcal{N}(0, I),$$

and the expected maximum absolute value of m standard Gaussian random variables is at most $\sqrt{2\log(m)}$. Hence $\frac{1}{\sqrt{m}}w(\ell_1^m) \leq \sqrt{\frac{2\log(m)}{m}}$ which is much smaller than the lower bound of $\sqrt{\frac{2}{\pi}}$ we obtained in the previous paragraph. This calculation suggests that the solution to the maximum width problem is to project ℓ_1^m onto a coordinate axis (the extreme opposite to taking $Y = I/\sqrt{m}$) although we do not have a proof of this.

First-order optimality conditions

For general sets that are not contraction zonoids, it is natural to at least understand what form the stationary points of the normalized maximum width problem might take. We establish only the simplest result in this direction, giving a characterization of when $P = \frac{1}{m}I$ is a stationary point for the reformulation of the normalized maximum width problem given in (5.3.3) of Lemma 5.3.1.

Proposition 5.3.16. *Suppose $\mathcal{X} \subset \mathbb{K}^{m\times d}$. Then $P = \frac{1}{m}I$ is a stationary point for*

$$\max_P w(P^{1/2}\mathcal{X}) \quad \textit{s.t.} \quad P \in \mathcal{H}_+^m, \;\; \mathrm{tr}(P) = 1 \tag{5.3.7}$$

if and only if there is a non-negative scalar κ such that

$$\mathbb{E}_{U\sim\gamma_{m\times d}^{\mathbb{K}}}[UU^* h_{\mathcal{X}}(U)] = \kappa I.$$

We prove this result in Section 5.9 to follow. Note that this result does not tell us the nature of the stationary point. It would be interesting to study the second-order optimality conditions to obtain conditions under which $P = \frac{1}{m}I$ is a local maximum of (5.3.3).

■ 5.4 Related work and examples

■ 5.4.1 Related work

Many authors, in the contexts of functional analysis [57], systems and control [12], optimization (see, e.g., [88, 124]), and theoretical computer science (see, e.g., [54]) have considered problems closely related to the dicussion here. In this section we briefly summarize those contributions directly related to understanding the positive semidefinite integrality gap, and associated rounding methods.

In the case where the objective functions are defined by positive semidefinite matrices and $\mathcal{X} = \{-1, 1\}$ the problem of understanding the positive semidefinite integrality gap is often called the 'positive semidefinite Grothendieck problem' or the 'little Grothendieck problem' [95]. A number of generalization of the binary problem have appeared in the literature, all of which are special cases of the results of this chapter. In the case where $\mathcal{X} = \{-1, 1\}$, the problem was first studied by Grothendieck [57]. The problem instance that establishes an upper bound of $\frac{2}{\pi}$ on the integrality gap in this case already appears in [57]. An analysis showing that $\frac{2}{\pi}$ is also a lower bound on the integrality gap in the positive semidefinite case is due to Rietz [104]. An algorithmic rounding scheme with approximation ratio that achieves the integrality gap is due to Nesterov [88]. In the case where $\mathcal{X} = U(1)$ is the set of unit complex numbers, Haagerup [60] (see also the independent work of Ben-Tal, Nemirovski, and Roos [12] and Zhang and Huang [139]) established a lower bound of $\frac{\pi}{4}$ on the integrality gap. So, Zhang, and Ye [124] established the corresponding upper bound. Furthermore So, Zhang, and Ye [124] extended the analysis to the case where $\mathcal{X}$ is the set of kth roots of unity, giving a rounding scheme that achieves an approximation ratio of $\frac{(k\sin(\pi/k))^2}{4\pi}$. The case where $\mathcal{X} = S^{m-1}$ is the unit sphere is considered by Briët, Oliveira, and Vallentin [21]. They give a rounding scheme that achieves an approximation ratio of $\frac{2}{m}\left[\frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})}\right]^2$. Briët, Buhrman, and Toner [20] establish a matching upper bound on the integrality gap. All of these examples are special cases of Theorem 5.2.12 where $Y = I/\sqrt{m}$.

The problem when $\mathcal{X} = \mathrm{St}(m, d)$ is the real (or complex) Stiefel manifold is studied by Bandeira, Kennedy, and Singer [7], where the authors show that the integrality gap is precisely $\left(\frac{1}{d}\mathbb{E}[\|G\|_*]\right)^2$ where G is a $d \times m$ random matrix with i.i.d. real (or complex) $\mathcal{N}(0, 1/m)$ entries and $\|\cdot\|_*$ is the nuclear norm (i.e. the sum of the singular values or, equivalently, the support function of $\mathrm{St}(m, d)$). These results are a special case of Theorem 5.2.12 again with $Y = I/\sqrt{m}$. We discuss this example in more detail in Section 5.4.2, to follow. We note that much of the work in this chapter was inspired by the techniques and problems considered by Bandeira, Kennedy, and Singer [7].

■ 5.4.2 Special cases previously studied

We now show how to apply our main result Theorem 5.2.12 to two of the examples summarized in the related work section (Section 5.4.1). The first is the case of complex roots of unity from [124]. The second is the case of the Stiefel manifolds from [7] which contains all of the others as special cases.

Example 5.4.1 (Roots of unity)**.** Suppose $\mathcal{X} = \{1, e^{i2\pi/k}, \ldots, e^{i2\pi(k-1)/k}\} \subset \mathbb{C}^{1\times 1}$ are the kth roots of unity. Here we have $\mathbb{K} = \mathbb{C}$ and $m = d = 1$. Since $d = 1$ this set is automatically right symmetric. Furthermore, since these are unit complex numbers they are certainly contractions. Hence the assumptions of Theorem 5.2.12 are satisfied.

We first find an optimal point of the normalized maximum width problem. We know that the normalized maximum width problem always has an optimal point $Y \in \mathcal{H}_+^m$ with Frobenius norm one. Since $m = 1$, we have that there is an optimal Y that is real and non-negative and has Frobenius norm one. Hence $Y = 1$ is optimal for the normalized maximum width problem. In other words $w_\star(\mathcal{X}) = w(\mathcal{X})$. The actual computation of the Gaussian width of $\mathcal{X}$ is carried out in [124]. The value of the Gaussian width of $\mathcal{X}$ (and hence $w_\star(\mathcal{X})$) is $w(\mathcal{X}) = \frac{k\sin(\pi/k)}{2\sqrt{\pi}}$.

Applying Theorem 5.2.12 we see that the positive semidefinite integrality gap is $\left[\frac{w_\star(\mathcal{X})}{1}\right]^2 = w(\mathcal{X})^2 = \frac{(k\sin(\pi/k))^2}{4\pi}$. The associated rounding scheme is the local randomized rounding scheme (Algorithm 5.1) defined by the map $\hat{X} : \mathbb{C} \to \mathcal{X}$ given by

$$\hat{X}(U) = \arg\max_{X\in\mathcal{X}} \langle X, U\rangle.$$

Concretely, this maps a complex number U to the nearest kth root of unity, i.e.

$$\hat{X}(U) = e^{i2\pi\ell/k} \quad \text{if and only if} \quad \frac{\pi(2\ell-1)}{k} < \arg(U) < \frac{\pi(2\ell+1)}{k}$$

(where the inequalities are to be interpreted modulo 2π). The local randomized rounding scheme defined by $\hat{X}$ is precisely the rounding scheme in [124, Equation 6].

Example 5.4.2 (Stiefel manifolds)**.** Suppose $\mathcal{X} = \mathrm{St}_{\mathbb{K}}(d,m) \subset \mathbb{K}^{m\times d}$ is the set of $m \times d$ matrices with entries in $\mathbb{K}$ satisfying $X^*X = I$. Since the representation of $O(d)$ (or $U(d)$) by orthogonal (or unitary) matrices on $\mathbb{K}^d$ is irreducible (over $\mathbb{C}$) it follows that $\mathrm{St}_{\mathbb{K}}(d,m)$ is right symmetric. Since $X^*X = I$ for all $X \in \mathrm{St}_{\mathbb{K}}(d,m)$ it certainly follows that $X^*X \preceq I$ for all $X \in \mathrm{St}_{\mathbb{K}}(d,m)$. Hence the assumptions of Theorem 5.2.12 are satisfied.

We have seen, from Example 5.3.11, that $Y = \frac{1}{\sqrt{m}} I$ is optimal for the normalized maximum width problem when $\mathcal{X} = \mathrm{St}_{\mathbb{K}}(d,m)$. Hence $w_\star(\mathrm{St}_{\mathbb{K}}(d,m)) = \frac{1}{\sqrt{m}} w(\mathrm{St}_{\mathbb{K}}(d,m))$. Now the support function of the Stiefel manifold is the nuclear norm, or sum of the singular values of a matrix. Hence the value of the normalized maximum width is

$$w_\star(\mathrm{St}_{\mathbb{K}}(d,m)) = \frac{1}{\sqrt{m}} \mathbb{E}_{U\sim\gamma^{\mathbb{K}}_{m\times d}}[\|U\|_*],$$

i.e. up to scaling the expected nuclear norm of a standard Gaussian matrix.

Applying Theorem 5.2.12 we see that the positive semidefinite integrality gap is

$$\left[\frac{w_\star(\mathrm{St}_{\mathbb{K}}(d,m))}{d}\right]^2 = \frac{1}{d^2 m} \mathbb{E}_{U\sim\gamma}[\|U\|_*]^2.$$

This is the same as the result of Bandeira, Kennedy, and Singer [7] (note that we use a Gaussian matrix U with standard Gaussian entries accounting for the difference between this expression and that given in [7]). The associated rounding scheme is the local randomized rounding scheme (Algorithm 5.1) defined by the map $\hat{X} : \mathbb{K}^{m\times d} \to \mathrm{St}_{\mathbb{K}}(d,m)$ given by

$$\hat{X}(U) = \arg \max_{X\in \mathrm{St}_{\mathbb{K}}(d,m)} \langle X, U\rangle.$$

This map can be described concretely as follows. Let U be an $m \times d$ matrix with singular value decomposition $U = V_1\Sigma V_2^*$ (here $V_1 \in \mathrm{St}_{\mathbb{K}}(d,m)$, Σ is $d\times d$ and diagonal with non-negative entries, and $V_2 \in \mathrm{St}_{\mathbb{K}}(d,d)$). Then $\arg\max_{X\in\mathrm{St}_{\mathbb{K}}(d,m)}\langle X, U\rangle = V_1V_2^*$, which recovers the rounding scheme proposed in [7]. We note that this example includes the cases $\mathcal{X} = O(d)$ and $\mathcal{X} = U(d)$ by setting $d = m$. It also generalizes the binary case, the case of unit complex numbers and the case of the sphere.

■ 5.4.3 Pairwise optimization problems on irreducible tautological orbits

We now describe a new family of sets $\mathcal{X}$ for which we can construct rounding schemes (and analyze their approximation properties) using our main result (Theorem 5.2.12). The family of examples significantly generalizes the examples of orthogonal matrices and unitary matrices studied by Bandeira, Kennedy, and Singer [7]. As particular special cases, it includes the case where $\mathcal{X} = SO(d)$, the set of $d \times d$ rotation matrices (see Chapter 4), as well as the case where $\mathcal{X}$ is a set that is affinely isomorphic to the set of $d \times d$ permutation matrices.

Let G be a group and ρ be an orthogonal/unitary representation on $\mathbb{K}^d$ that is irreducible over $\mathbb{C}$. Let $\mathcal{X} = \rho(G) \subset \mathbb{K}^{d\times d}$ be the image of G under the representation. We call this a *irreducible tautological orbit* of G (following the terminology 'tautological orbitope' from [110]). Then $\rho(G)$ is a subset of orthogonal/unitary matrices (and hence contractions), that is clearly right symmetric with respect to the action of ρ. Hence any such $\rho(G)$ satisfies the assumptions of Theorem 5.2.12.

To obtain optimal approximation algorithms for these problems, we need to solve the normalized maximum width problem and be able to implement the resulting rounding scheme. In general, irreducible tautological orbits are not contraction zonoids (see Definition 5.3.6), so we cannot automatically deduce that the normalized maximum width problem reduces to a convex optimization problem for these sets. In fact we do not know the solution of the normalized maximum width problem for general irreducible tautological orbits. Nevertheless, the following lemma shows that taking $Y = \frac{1}{\sqrt{d}}I$ is at least a candidate to be optimal for the normalized maximum width problem.

Lemma 5.4.3. *Let $\rho(G)$ be an irreducible tautological orbit of a group G. Then $P = \frac{1}{d}I$ is a stationary point for the reformulation of the normalized maximum width problem* *from* (5.3.3), *i.e.*

$$\max_{\substack{P\in\mathcal{H}_+^d\\ \mathrm{tr}(P)=1}} w(P^{1/2}\rho(G)).$$

We prove this result in Section 5.9. It uses the fact that irreducible tautological orbits are not just right symmetric (as noted above) but also *left* symmetric (the precise definition is given in Section 5.9). If $P = \frac{1}{d}I$ were in fact optimal for the reformulation (5.3.3) of the normalized maximum width problem, then the corresponding choice of Y in Theorem 5.2.12 would be $Y = \frac{1}{\sqrt{d}}I$.

A benefit of our analysis is that we do not need optimal solutions to the normalized maximum width problem to construct interesting rounding schemes and analyze their approximation properties. Any feasible solution (such as taking $Y = \frac{1}{\sqrt{d}}I$ in this case) gives a rounding scheme, the achievable approximation ratio of which we can, in principle, analyze. To implement the resulting rounding scheme for $\mathcal{X} = \rho(G)$ (an irreducible tautological orbit) with $Y = \frac{1}{\sqrt{d}}I$ we need to be able to evaluate the map

$$\hat{X}(U) = \arg\max_{X\in\rho(G)} \frac{1}{\sqrt{d}}\langle X, U\rangle. \tag{5.4.1}$$

We discuss the form of the optimization problem we need to solve for two examples below. To evaluate the approximation guarantee this rounding scheme achieves, we need to compute

$$\left[\frac{1}{d}w\left(\frac{1}{\sqrt{d}}\rho(G)\right)\right]^2 = \frac{1}{d^3}\mathbb{E}_{U\sim\gamma_{d\times d}^{\mathbb{K}}}\left[\max_{X\in\rho(G)}\langle X, U\rangle\right]^2. \tag{5.4.2}$$

Rotation matrices

The pairwise optimization problems that arise in the case where $\mathcal{X} = SO(3)$ are an example of pairwise optimization problems on irreducible tautological orbits. These problems arise, for instance, in molecular imaging applications [123] and discrete-time optimal filtering problems for rotation matrix-valued variables [118]. In this case the optimization problem (5.4.1) that defines the rounding map is known as Wahba's problem [135]. We discuss it in detail in Chapter 4.

Permutation matrices

Suppose S_d is the symmetric group of d symbols and ρ is the irreducible (over $\mathbb{C}$) representation of S_d on the subspace $\mathbf{1}^\perp$ of $\mathbb{R}^d$ consisting of vectors $x \in \mathbb{R}^d$ with $\sum_{i=1}^d x_i = 0$. Concretely, the irreducible tautological orbit $\rho(S_d)$ consists of all matrices of the form V^TPV where P is a $d\times d$ permutation matrix and V is a $d\times(d-1)$ matrix with columns an orthonormal basis for $\mathbf{1}^\perp$. This set is affinely isomorphic to the set of permutation matrices. Pairwise optimization problems over $\rho(S_d)$ are (up to an additive constant) the same as pairwise optimization problems over permutation matrices. These arise, for instance, in the multi-reference alignment problem [6].

The optimization problem (5.4.1) involves maximizing the linear functional VUV^T over permutation matrices, an instance of the classical transportation or bipartite matching problem [121]. To compute the associated approximation ratio, one needs to find the expected value of the maximum bipartite matching with weights VUV^T where U is an i.i.d. $(d-1)\times(d-1)$ standard Gaussian matrix. This is a very natural problem that may be interesting to study in its own right.

■ 5.5 Upper bounds on the positive semidefinite integrality gap

In this section we establish the following upper bound on the positive semidefinite integrality gap for any set $\mathcal{X} \subset \mathbb{K}^{m\times d}$ of right symmetric contractions. This is one key component in the proof of our main result (Theorem 5.2.12), which is given in Section 5.6 to follow.

Proposition 5.5.1. *If $\mathcal{X} \subset \mathbb{K}^{m\times d}$ is a right symmetric set of contractions then the positive semidefinite integrality gap is at most $\left[\frac{w_\star(\mathcal{X})}{d}\right]^2$.*

Since the integrality gap is defined as an infimum (see Equation 5.2.5) we can obtain upper bounds on the positive semidefinite integrality gap by evaluating the ratio $\text{OPT}_{\mathcal{X}}(C)/\text{OPT}_{\text{sdp}}(C)$ for some objective function defined by $C \succeq 0$. Rather than a single $nd \times nd$ positive semidefinite matrix, we construct a sequence of matrices C_p^n of increasing dimension nd and rank p such that $\text{OPT}_{\mathcal{X}}(C_p^n)/\text{OPT}_{\text{sdp}}(C_p^n) \to \left[\frac{w_\star(\mathcal{X})}{d}\right]^2$ as $n, p \to \infty$.

In fact, rather than finite sized matrices C_p^n (and hence problem instances indexed by the finite set $[n]$), we consider problem instances indexed by a probability space that are specified by positive semidefinite kernels on that probability space. The idea behind the construction goes back to the work of Grothendieck [57], and has been repeatedly reused and slightly generalized by many authors (see, e.g., [124, 20, 7]). This section can be thought of as a further generalization and simplification of the arguments by Bandeira, Kennedy, and Singer [7] for the case where $\mathcal{X} = \text{St}_{\mathbb{K}}(d, m)$.

Indeed we define a sequence of positive semidefinite kernels on the measure space $(\mathbb{K}^{p\times d}, \gamma_{p\times d}^{\mathbb{K}})$ consisting of $\mathbb{K}^{p\times d}$ equipped with the standard Gaussian measure. On this space define the matrix-valued mapping $C_p : \mathbb{K}^{p\times d} \times \mathbb{K}^{p\times d} \to \mathbb{K}^{d\times d}$ by $C_p(U, V) = \frac{1}{p}U^*V$. Note that C_p is normalized so that $\mathbb{E}_{U\sim\gamma}[\text{tr}(C_p(U, U))] = d$. The corresponding instance of a pairwise quadratic optimization problem on $\mathcal{X}$ is to solve

$$\text{OPT}_{\mathcal{X}}(C_p) := \sup_{X:\mathbb{R}^{p\times d}\to\mathcal{X}} \mathbb{E}_{U\sim\gamma}[\mathbb{E}_{V\sim\gamma}[\langle C_p(U, V), X(U)^*X(V)\rangle]]. \tag{5.5.1}$$

It may be instructive to compare this with the expression for a finite problem instance (5.2.1). One can see (5.5.1) as a limit of a finite instance of a pairwise optimization problem on $\mathcal{X}$ obtained by replacing the expectations (i.e. integrals) with their definitions as the limit of finite sums. Similarly we can define $\mathrm{OPT}_{\mathrm{sdp}}(C_p)$ as the appropriate limit of a sequence of semidefinite optimization problems.

Towards establishing Proposition 5.5.1, our upper bound on the integrality gap, we characterize the optimal value in (5.5.1) in terms of the normalized maximum width of $\mathcal{X}$.

Proposition 5.5.2. *For any integer $p \geq m$,*

$$p\,\mathrm{OPT}_{\mathcal{X}}(C_p) = \sup_{X:\mathbb{R}^p\to\mathcal{X}} \|\mathbb{E}_{U\sim\gamma}[X(U)U^*]\|_F^2 = w_\star(\mathcal{X})^2. \tag{5.5.2}$$

Proof. We begin by establishing the right-hand equality. To show this we use the well-known variational characterization of the Frobenius norm as $\|W\|_F = \max_{\|Y\|_F=1}\langle Y, W\rangle$ (which can be deduced from the Cauchy-Schwarz inequality, for example) to write

$$\begin{aligned}\sup_{X:\mathbb{R}^{p\times d}\to\mathcal{X}} \|\mathbb{E}_{U\sim\gamma}[X(U)U^*]\|_F &= \sup_{X:\mathbb{R}^{p\times d}\to\mathcal{X}}\max_{\|Y\|_F=1}\mathbb{E}_{U\sim\gamma}[\langle Y, X(U)U^*\rangle]\\ &= \max_{\|Y\|_F=1}\sup_{X:\mathbb{R}^{p\times d}\to\mathcal{X}}\mathbb{E}_{U\sim\gamma}[\langle X(U), YU\rangle].\end{aligned}$$

For fixed Y, the supremum over $X:\mathbb{R}^{p\times d}\to\mathcal{X}$ is achieved by maximizing the function inside the expectation pointwise, that is by taking $X(U)\in\arg\max_{X\in\mathcal{X}}\langle X, YU\rangle$ for each $U\in\mathbb{R}^{p\times d}$. Substituting this choice of X back into the objective function we obtain

$$\begin{aligned}\max_{\|Y\|_F=1}\mathbb{E}_{U\sim\gamma}[\langle\arg\max_{X\in\mathcal{X}}\langle X, YU\rangle, YU\rangle] &= \max_{\|Y\|_F=1}\mathbb{E}_{U\sim\gamma}[\max_{X\in\mathcal{X}}\langle X, YU\rangle]\\ &= \max_{\|Y\|_F=1}\mathbb{E}_{U\sim\gamma}[\max_{X\in\mathcal{X}}\langle Y^*X, U\rangle]\\ &= w_\star(\mathcal{X})\end{aligned}$$

(where for the last equality we have used Lemma 5.3.1). This establishes the right-hand equality in (5.5.2). The left-hand equality in (5.5.2) holds because

$$\begin{aligned}p\,\mathrm{OPT}_{\mathcal{X}}(C_p) &= \sup_{X:\mathbb{R}^{p\times d}\to\mathcal{X}}\mathbb{E}_{U\sim\gamma}[\mathbb{E}_{V\sim\gamma}[\langle U^*V, X(U)^*X(V)\rangle]]\\ &= \sup_{X\in\mathbb{R}^{p\times d}\to\mathcal{X}}\mathbb{E}_{U\sim\gamma}[\mathbb{E}_{V\sim\gamma}[\langle X(U)U^*, X(V)V^*\rangle]]\\ &= \sup_{X\in\mathbb{R}^{p\times d}\to\mathcal{X}}\|\mathbb{E}_{U\sim\gamma}[X(U)U^*]\|_F^2.\end{aligned}$$

□

We now consider the optimal value of the semidefinite relaxation for the sequence (C_p) of instances. This tells us about the denominator of the ratio that defines the integrality gap.

Lemma 5.5.3. *For any positive integer $p \geq d$,*

$$p\,\mathrm{OPT}_{\mathrm{sdp}}^{d,\mathbb{K}}(C_p) \geq p\,\mathrm{OPT}_{\mathrm{St}_{\mathbb{K}}(d,p)}(C_p) = w_\star(\mathrm{St}_{\mathbb{K}}(d,p))^2.$$

Proof. The first equality holds because $\mathrm{OPT}_{\mathrm{sdp}}(C) \geq \mathrm{OPT}_{\mathcal{X}}(C)$ for every C and every set $\mathcal{X} \subset \mathbb{K}^{p\times d}$ of contractions (and so, in particular, for $\mathcal{X} = \mathrm{St}_{\mathbb{K}}(d,p)$). The second equality is established in Proposition 5.5.2. □

This observation is useful because we have a good understanding of the quantity $w_\star(\mathrm{St}_{\mathbb{K}}(d,p))$. We restate this result from Section 5.3.1.

Lemma 5.3.12. *For any positive integer d,*

$$w_\star(\mathrm{St}_{\mathbb{K}}(d,p)) \leq d \;\; \textit{for all } p \geq d \quad \textit{and} \quad \lim_{p\to\infty} w_\star(\mathrm{St}_{\mathbb{K}}(d,p)) = d.$$

We are now in a position to establish the upper bound on the positive semidefinite integrality gap stated as Proposition 5.5.1 at the start of this section.

Proof of Proposition 5.5.1. Since each C_p is positive semidefinite, we have that

$$\inf_{C\succeq 0, C\neq 0} \frac{\mathrm{OPT}_{\mathcal{X}}(C)}{\mathrm{OPT}_{\mathrm{sdp}}(C)} \leq \inf_{p\geq m} \frac{\mathrm{OPT}_{\mathcal{X}}(C_p)}{\mathrm{OPT}_{\mathrm{sdp}}(C_p)}.$$

Then by applying Lemma 5.5.3 and Lemma 5.3.12,

$$\inf_{p\geq m} \frac{\mathrm{OPT}_{\mathcal{X}}(C_p)}{\mathrm{OPT}_{\mathrm{sdp}}(C_p)} \leq \inf_{p\geq m} \frac{w_\star(\mathcal{X})^2}{w_\star(\mathrm{St}_{\mathbb{K}}(d,p))^2} = \left[\frac{w_\star(\mathcal{X})}{d}\right]^2.$$

□

■ 5.6 Designing optimal rounding schemes

Given a set $\mathcal{X} \subset \mathbb{K}^{m\times d}$ of contractions we now consider the problem of constructing a sequence of rounding schemes that achieves the best possible approximation ratio. Since the approximation ratio is a lower bound on the integrality gap, we know from Proposition 5.5.1 that it cannot exceed $\left[\frac{w_\star(\mathcal{X})}{d}\right]^2$. In this section we design a sequence of randomized rounding schemes for any right symmetric set $\mathcal{X}$ of contractions that is optimal since it achieves an approximation ratio of $\left[\frac{w_\star(\mathcal{X})}{d}\right]^2$. This construction allows us to establish our main result, Theorem 5.2.12, which we restate and prove at the end of this section.

Our basic strategy is to search over local randomized rounding schemes (see Section 5.2.2) to find a rounding scheme with the largest approximation ratio. We aim to do this by finding a lower bound on the achievable approximation ratio for an arbitrary local randomized rounding scheme, and then maximizing this lower bound. This strategy works when $d = 1$, but for $d > 1$ we need an additional assumption on $\mathcal{X}$ (i.e. that $\mathcal{X}$ is right symmetric) to carry it out. We also restrict our search over rounding schemes to local randomized rounding schemes that are also equivariant with respect to the group action on $\mathcal{X}$ (see Definition 5.6.2 to follow). We then find a lower bound on the approximation ratio achieved by any equivariant local randomized rounding scheme. We optimize this lower bound to obtain the best possible equivariant local randomized rounding scheme. We find that it achieves an approximation ratio of $\left[\frac{w_\star(\mathcal{X})}{d}\right]^2$ and so is, in fact, optimal among all randomized rounding schemes.

We note that the discussion in this section could be presented in a different way. Indeed the argument would work perfectly well if we simply proposed the family of rounding schemes described in the statement of Theorem 5.2.12 and analyzed the approximation ratio they achieve directly (using the fact that they are equivariant). We have chosen to present the material in the way we have in an effort not just to show that Theorem 5.2.12 is true, but also to explain where our optimal rounding schemes come from.

Achievable approximation ratio

Consider an arbitrary sequence of local randomized rounding schemes (R_n) specified by a function $\hat{X} : \mathbb{K}^{p\times d} \to \mathcal{X}$ (see Section 5.2.2). Recall that with any such rounding scheme R_n we define a mapping $F_{R_n} : \mathcal{G}^n_{\mathrm{sdp}} \to \mathcal{G}^n_{\mathcal{X}}$ from the feasible region of the semidefinite relaxation to $\mathcal{G}^n_{\mathcal{X}}$ by (5.2.9). By using (5.2.10) we can obtain a lower bound on the positive semidefinite approximation ratio achieved by computing the largest constant κ such that

$$\kappa\langle C, Z\rangle \leq \langle C, F_{R_n}(Z)\rangle$$

for all $C \succeq 0$, all $Z \in \mathcal{G}_{\mathrm{sdp}}$, and all $n \in \mathbb{N}$. By using the fact that the positive semidefinite cone is self-dual (i.e. $\langle X, Y\rangle \geq 0$ for all $Y \succeq 0$ holds if and only if $X \succeq 0$) we obtain the equivalent problem:

$$\max_{\kappa}\ \kappa \quad \text{s.t.} \quad F_{R_n}(Z) - \kappa Z \succeq 0 \ \text{ for all } Z \in \mathcal{G}^n_{\mathrm{sdp}} \text{ and all } n \in \mathbb{N}. \tag{5.6.1}$$

From now on we use (5.6.1) to give a lower bound on the approximation ratio achievable by the sequence (R_n) of local randomized rounding schemes specified by $\hat{X}$.

All the results of this section follow from the next somewhat technical lemma. The idea is to obtain an expression something like $F_{R_n}(Z) - \kappa Z \succeq 0$ for all $Z \in \mathcal{G}_{\text{sdp}}$, but we are not quite able to do this in such generality. The approach taken here has its origins in the proof method of Rietz [104] and its generalization in [7].

Lemma 5.6.1. *Define matrices*

$$A_{\hat{X}} = \mathbb{E}_{U\sim\gamma}[\hat{X}(U)U^*]/d \in \mathbb{K}^{m\times d}, \qquad B_{\hat{X}} = \mathbb{E}_{U\sim\gamma}\mathbb{E}_{V\sim\gamma}[\hat{X}(U)^*\hat{X}(V)V^*U]/d \in \mathbb{K}^{d\times d}$$

and $\tilde{B}_{\hat{X}} = \operatorname{blkdiag}(B_{\hat{X}},\ldots,B_{\hat{X}}) \in \mathbb{K}^{nd\times nd}$*. Then for all* $Z \succeq 0$ *such that* $Z_{ii} = I$ *for* $i \in [n]$*,*

$$F_{R_n}(Z) - Z\tilde{B}_{\hat{X}}^* - \tilde{B}_{\hat{X}}Z + \|A_{\hat{X}}\|_F^2 Z \succeq 0. \tag{5.6.2}$$

Proof. We establish (5.6.2) by showing that the i,j block entry has a factorization as

$$[F_{R_n}(Z)]_{ij} - Z_{ij}B_{\hat{X}}^* - B_{\hat{X}}Z_{ij} + \|A_{\hat{X}}\|_F^2 Z_{ij} = \mathbb{E}[(\hat{X}(W_i) - A_{\hat{X}}W_i)^*(\hat{X}(W_j) - A_{\hat{X}}W_j)] \tag{5.6.3}$$

where $W_1, W_2, \ldots, W_n$ are the random variables in the definition of F_{R_n} (i.e. in Algorithm 5.1). In particular the Gaussian matrices $W_1, W_2, \ldots, W_n$ are defined by taking $W \in \mathbb{K}^{p\times nd}$ to have are i.i.d. Gaussian rows with mean zero and covariance Z, and then taking the W_i to satisfy

$$W = \begin{bmatrix} W_1 & W_2 & \cdots & W_n \end{bmatrix}.$$

Because the rows of W are i.i.d. it is straightforward to check that $\mathbb{E}[W_i^* w w^* W_j] = Z_{ij}$ for all unit vectors $w \in \mathbb{K}^p$.

With these preliminary facts established, we now show that (5.6.3) holds. To do this we just expand the expectation on the right hand side and simplify each term.

- We note that $\mathbb{E}[\hat{X}(W_i)^*\hat{X}(W_j)] = [F_{R_n}(Z)]_{ij}$ by the definition of the rounding scheme R_n and the map F_{R_n}.
- We claim that $\mathbb{E}[\hat{X}(W_i)^*A_{\hat{X}}W_j] = B_{\hat{X}}Z_{ij}$. To see this we decompose W_j into a component independent of W_i and a component in the direction of W_i. This decomposition is

  $$W_j = (W_iZ_{ij}) + (W_j - W_iZ_{ij}). \tag{5.6.4}$$

  The matrices W_iZ_{ij} and $(W_j - W_iZ_{ij})$ are independent because because for each unit vector $w \in \mathbb{K}^p$ we have

  $$\mathbb{E}[W_i^* w w^*(W_j - W_iZ_{ij})] = Z_{ij} - Z_{ii}Z_{ij} = 0 \quad (\text{since } Z_{ii} = I).$$

Hence using the expression for W_j in (5.6.4) we see that

$$\mathbb{E}[\hat{X}(W_i)^* A_{\hat{X}} W_j] = \mathbb{E}[\hat{X}(W_i)^* A_{\hat{X}} W_i] Z_{ij} - \mathbb{E}[x(W_i)^* A_{\hat{X}}(W_j - W_i Z_{ij})].$$

Since W_i is a standard Gaussian matrix, the first of these terms is, by the definitions of $A_{\hat{X}}$ and $B_{\hat{X}}$, equal to $B_{\hat{X}} Z_{ij}$. The second vanishes because W_i and $W_j - W_i Z_{ij}$ are independent and have zero mean.

- Similarly $\mathbb{E}[W_i^* A_{\hat{X}}^* \hat{X}(W_j)] = [B_{\hat{X}} Z_{ji}]^* = Z_{ji}^* B_{\hat{X}}^* = Z_{ij} B_{\hat{X}}^*$.
- Finally $\mathbb{E}[W_i^* A_{\hat{X}}^* A_{\hat{X}} W_j] = \|A_{\hat{X}}\|_F^2 Z_{ij}$. To see this decompose $A_{\hat{X}}^* A_{\hat{X}}$ as

$$A_{\hat{X}}^* A_{\hat{X}} = \sum_{k=1}^{p} \sigma_k^2 w_k w_k^*$$

where σ_k are the singular values of $A_{\hat{X}}$ and w_k are the associated (unit) right singular vectors. Then

$$\mathbb{E}[W_i^* A_{\hat{X}}^* A_{\hat{X}} W_j] = \sum_{k=1}^{p} \sigma_k^2 \mathbb{E}[W_i^* w_k w_k^* W_j] = \|A_{\hat{X}}\|_F^2 Z_{ij}.$$

□

We have shown that $F_{R_n}(Z) - Z\tilde{B}_{\hat{X}}^* - \tilde{B}_{\hat{X}} Z + \|A_{\hat{X}}\|_F^2 Z \succeq 0$ which is almost in the form $F_{R_n}(Z) - \kappa Z \succeq 0$ from which we could deduce a lower bound on the approximation ratio. Indeed if $d = 1$ then $B_{\hat{X}}$ is a scalar and so $\tilde{B}_{\hat{X}}$ is a multiple of the identity and we *can* deduce a lower bound on the approximation ratio. To proceed in a similar fashion for the general d case, we impose additional assumptions on the problem to ensure that $B_{\hat{X}}$ (and hence $\tilde{B}_{\hat{X}}$) is a scalar multiple of the identity.

■ 5.6.1 Equivariant local randomized rounding

The key additional assumption we impose is that $\mathcal{X} \subset \mathbb{K}^{m \times d}$ is right symmetric with respect to a group G and an orthogonal/unitary representation ρ (see Definition 5.2.11). We also further restrict our attention to local randomized rounding schemes that respect the symmetries of $\mathcal{X}$ in a sense made precise by Definition 5.6.2 to follow. While this is a restriction, we will show, in our proof (to follow) of Theorem 5.2.12 that when $\mathcal{X}$ is right-symmetric there is an optimal rounding scheme that has this form.

Definition 5.6.2. Suppose $\mathcal{X} \subset \mathbb{K}^{m \times d}$ is right symmetric with respect to the group G and orthogonal/unitary representation ρ. The local randomized rounding scheme defined by $\hat{X} : \mathbb{K}^{p \times d} \to \mathcal{X}$ is *equivariant* if $\mathcal{X}(U)\rho(g) = \mathcal{X}(U\rho(g))$ for all $g \in G$.

Under these assumptions, we can deduce that the matrix $B_{\hat{X}}$ is a multiple of the identity. To do so, we use a fundamental result of representation theory, known as Schur's lemma. For completeness we state and prove the version of Schur's lemma we require.

Lemma 5.6.3. *Suppose ρ is a unitary representation of a group G on $\mathbb{C}^d$ that is irreducible over $\mathbb{C}$. Let $A : \mathbb{C}^d \to \mathbb{C}^d$ satisfy*

$$A\rho(g) = \rho(g)A$$

for all $g \in G$. Then there is some $\alpha \in \mathbb{C}$ such that $A = \alpha I$.

Proof. Let $\alpha \in \mathbb{C}$ be any eigenvalue of A and let $U \subseteq \mathbb{C}^d$ denote the corresponding (non-zero) eigenspace. Then for any $u \in U$ we have that $(\alpha I - A)(\rho(g)u) = \rho(g)(\alpha I - A)u = 0$. Hence U is an invariant subspace of $\mathbb{C}^d$. Since U is non-zero by assumption and ρ is irreducible over $\mathbb{C}$ it follows that $U = \mathbb{C}^d$ and so $A = \alpha I$. □

With Schur's lemma established, to see that $B_{\hat{X}}$ should be a multiple of the identity, it is enough to show that it commutes with an irreducible representation.

Lemma 5.6.4. *Suppose $\mathcal{X}$ is right symmetric with respect to G and ρ, and $\hat{X}$ is equivariant with respect to G and ρ. Then*

$$\rho(g)B_{\hat{X}} = B_{\hat{X}}\rho(g) \quad \textit{for all } g \in G$$

and so $B_{\hat{X}} = \|A_{\hat{X}}\|_F^2 I$.

Proof. For the first assertion note that for all $g \in G$

$$\begin{aligned}
\rho(g)(d\,B_{\hat{X}}) &= \mathbb{E}_{U,V}[\rho(g)\hat{X}(U)^*\hat{X}(V)V^*U] \\
&= \mathbb{E}_{U,V}[(\hat{X}(U)\rho(g)^*)^*\hat{X}(V)\rho(g)^*\rho(g)V^*U] \\
&= \mathbb{E}_{U,V}[\hat{X}(U\rho(g)^*)^*\hat{X}(V\rho(g)^*)(V\rho(g)^*)^*U] \\
&= \mathbb{E}_{\tilde{U},\tilde{V}}[\hat{X}(\tilde{U})^*\hat{X}(\tilde{V})\tilde{V}^*\tilde{U}]\rho(g) \\
&= (d\,B_{\hat{X}})\rho(g).
\end{aligned}$$

where the first equality is the definition of $B_{\hat{X}}$, the second holds because $\rho(g)$ is orthogonal/unitary, the third holds because $\hat{X}$ is equivariant, and the last holds by making the substitution $\tilde{U} = U\rho(g)^*$ and $\tilde{V} = V\rho(g)^*$ and using the fact that the joint distribution of U, V and $\tilde{U}, \tilde{V}$ are the same since $\rho(g)$ is orthogonal/unitary.

Since ρ is irreducible over $\mathbb{C}$ (by assumption) it follows from Schur's lemma that $B_{\hat{X}} = \beta I$ for some scalar $\beta \in \mathbb{C}$. In fact by taking the trace of both sides of $B_{\hat{X}} = \beta I$

we have that

$$\beta = \frac{\mathrm{tr}(B_{\hat{X}})}{d} = \frac{1}{d^2}\mathbb{E}_{U,V}[\mathrm{tr}(\hat{X}(U)^*\hat{X}(V)V^*U)]$$
$$= \frac{1}{d^2}\mathrm{tr}(\mathbb{E}_V[\hat{X}(V)V^*]\mathbb{E}_U[\hat{X}(U)U^*]^*) = \|A_{\hat{X}}\|_F^2.$$

□

By combining Lemmas 5.6.1 and 5.6.4, we have a lower bound on the approximation ratio achieved by *any* equivariant rounding scheme.

Proposition 5.6.5. *If $\mathcal{X} \subset \mathbb{K}^{m\times d}$ is right symmetric and $\hat{X} : \mathbb{K}^{p\times d} \to \mathcal{X}$ is equivariant then the corresponding equivariant local randomized rounding achieves an approximation ratio of at least*

$$\|A_{\hat{X}}\|_F^2 = \left[\frac{\|\mathbb{E}_{U\sim\gamma}[\hat{X}(U)U^*]\|_F}{d}\right]^2. \tag{5.6.5}$$

We have already encountered the quantity appearing on the right hand side in Equation (5.6.5). (Up to scaling) this appears in Section 5.5 as the function to be optimized to compute $\mathrm{OPT}_{\mathcal{X}}(C_p)$. Using the same idea as in the proof of Proposition 5.5.2 we now maximize this bound over equivariant functions $\hat{X} : \mathbb{K}^{p\times d} \to \mathcal{X}$ to obtain an optimal equivariant local randomized rounding scheme. This is the content of the second part of our main result. As such, we now restate and prove that result.

Theorem 5.2.12. *If $\mathcal{X} \subset \mathbb{K}^{m\times d}$ (where $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$) is a right symmetric set of contractions then the following hold.*

1. *The positive semidefinite integrality gap is $\alpha_{\mathcal{X}} = \left[\frac{w_\star(\mathcal{X})}{d}\right]^2$.*
2. *If Y is any $m \times p$ matrix with $\|Y\|_F = 1$ then the local randomized rounding scheme specified by any function $\hat{X} : \mathbb{R}^{p\times d} \to \mathcal{X}$ satisfying*

$$\hat{X}(U) \in \arg\max_{X\in\mathcal{X}}\langle X, YU\rangle, \tag{5.6.6}$$

   achieves an approximation ratio of $\left[\frac{w(Y^*\mathcal{X})}{d}\right]^2$.
3. *If Y is any argument of maximum in the normalized maximum width problem (5.2.12) the rounding scheme specified in 2. achieves the optimal approximation ratio of $\left[\frac{w_\star(\mathcal{X})}{d}\right]^2$.*

Proof. The fact that $\alpha_{\mathcal{X}} \leq \left[\frac{w_\star(\mathcal{X})}{d}\right]^2$ was established in Proposition 5.5.1. We establish the reverse inequality at the end of the proof.

We have seen in Proposition 5.6.5 that for any equivariant map $\hat{X} : \mathbb{K}^{p\times d} \to \mathcal{X}$ a lower bound on the approximation ratio achieved by the corresponding equivariant local randomized rounding scheme is given by

$$\left[\frac{\|\mathbb{E}_{U\sim\gamma}[\hat{X}(U)U^*]\|_F}{d}\right]^2 = \frac{1}{d^2}\left[\max_{\substack{Y\in\mathbb{K}^{m\times p}\\ \|Y\|_F=1}} \mathbb{E}_{U\sim\gamma}[\langle \hat{X}(U), YU\rangle]\right]^2.$$

As such, to find an optimal equivariant rounding scheme it is enough to solve

$$\sup_{\hat{X}:\mathbb{K}^{p\times d}\to\mathcal{X}} \max_{\substack{Y\in\mathbb{K}^{m\times p}\\ \|Y\|_F=1}} \mathbb{E}_{U\sim\gamma}[\langle \hat{X}(U), YU\rangle] \quad \text{s.t.} \quad \hat{X}(U\rho(g)) = \hat{X}(U)\rho(g) \ \text{ for all } g\in G.$$

First we consider solving the problem without the equivariance constraint on $\hat{X}$, i.e. solving

$$\sup_{\hat{X}:\mathbb{K}^{p\times d}\to\mathcal{X}} \max_{\substack{Y\in\mathbb{K}^{m\times p}\\ \|Y\|_F=1}} \mathbb{E}_{U\sim\gamma}[\langle \hat{X}(U), YU\rangle] = \max_{\substack{Y\in\mathbb{K}^{m\times p}\\ \|Y\|_F=1}} \sup_{\hat{X}:\mathbb{K}^{p\times d}\to\mathcal{X}} \mathbb{E}_{U\sim\gamma}[\langle \hat{X}(U), YU\rangle].$$

We have seen in the proof of Proposition 5.5.2 that for fixed Y with $\|Y\|_F = 1$ any $\hat{X}$ satisfying

$$\hat{X}(U) = \arg\max_{X\in\mathcal{X}} \langle X, YU\rangle$$

is optimal for the inner optimization problem and achieves a value of $w(Y^*\mathcal{X})$. Note that this map is equivariant because

$$\hat{X}(U\rho(g)) = \arg\max_{X\in\mathcal{X}}\langle X, YU\rho(g)\rangle = \arg\max_{X\in\mathcal{X}}\langle X\rho(g)^*, YU\rangle = [\arg\max_{X\in\mathcal{X}}\langle X, YU\rangle]\rho(g)$$

where the last equality uses the right symmetry of $\mathcal{X}$. Hence, for any fixed Y with $\|Y\|_F = 1$ we have constructed an equivariant local randomized rounding scheme. It achieves an approximation ratio of

$$\frac{1}{d^2}\left[\max_{\substack{Y'\in\mathbb{K}^{m\times p}\\ \|Y'\|_F=1}} \mathbb{E}_{U\sim\gamma}[\langle \arg\max_{X\in\mathcal{X}}\langle X, YU\rangle, Y'U\rangle]\right]^2 \geq \left[\frac{w(Y^*\mathcal{X})}{d}\right]^2$$

where the last inequality holds by choosing $Y' = Y$ in the maximum over Y'. This completes the proof of the second part of the statement.

The third part of the statement simply follows by optimizing over the choice of Y in the second part of the statement. Finally since we have a randomized rounding scheme that achieves an approximation ratio of $\left[\frac{w_\star(\mathcal{X})}{d}\right]^2$ it follows that $\alpha_{\mathcal{X}} \geq \left[\frac{w_\star(\mathcal{X})}{d}\right]^2$. This is the reverse inequality on the positive semidefinite integrality gap required to complete the proof. □

5.7 Approximating the optimal rounding scheme

To implement the rounding schemes that arise from Theorem 5.2.12, we need to be able to solve optimization problems of the form

$$\max_{X \in \mathcal{X}} \langle X, U \rangle,$$

i.e. to maximize a linear functional over the set $\mathcal{X}$. This is computationally tractable for many cases of interest (such as when $\mathcal{X}$ consists of permutation matrices or rotation matrices, as discussed in Section 5.4). In other cases this optimization problem itself may be hard. In this section we assume we only have access to an algorithm that given $U \in \mathbb{K}^{m \times d}$ produces a feasible point $\tilde{X}(U) \in \mathcal{X}$ satisfying

$$\beta \left[\max_{X \in \mathcal{X}} \langle X, U \rangle\right] \leq \langle \tilde{X}(U), U \rangle \leq \max_{X \in \mathcal{X}} \langle X, U \rangle$$

for some positive constant $0 < \beta \leq 1$. In other words we have a β-approximation algorithm for the maximization of a linear functional over $\mathcal{X}$. The idea is to use this β-approximation algorithm as part of a local randomized rounding scheme and understand how β enters in the achievable approximation ratio. As before we assume that $\mathcal{X} \subset \mathbb{K}^{m \times d}$ is right symmetric with respect to ρ and G. We correspondingly assume that the available β-approximation algorithm defined by $U \mapsto \tilde{X}(U)$ is equivariant in the sense that $\tilde{X}(U\rho(g)) = \tilde{X}(U)\rho(g)$ for all $g \in G$.

Before, for a fixed $Y \in \mathbb{K}^{m \times d}$ with $\|Y\|_F = 1$ we considered the equivariant local randomized rounding scheme parameterized by the equivariant function $\hat{X} : \mathbb{K}^{p \times d} \to \mathcal{X}$ where $\hat{X}(U) = \arg\max_{X \in \mathcal{X}} \langle X, YU \rangle$. This achieved an approximation ratio of $\left[\frac{w(Y^*\mathcal{X})}{d}\right]^2$. Now we analyze the rounding scheme parameterized by the equivariant function $U \mapsto \tilde{X}(YU)$. Doing so we obtain a variation on Theorem 5.2.12 that applies in this setting.

Theorem 5.7.1. *Let $\mathcal{X} \subset \mathbb{K}^{m \times d}$ be a right-symmetric (w.r.t. G and ρ) subset of contractions. Let $U \mapsto \tilde{X}(U)$ satsify*

$$\beta \left[\max_{X \in \mathcal{X}} \langle X, U \rangle\right] \leq \langle \tilde{X}(U), U \rangle \leq \max_{X \in \mathcal{X}} \langle X, U \rangle$$

for all $U \in \mathbb{K}^{m \times d}$ and $\tilde{X}(U\rho(g)) = \tilde{X}(U)\rho(g)$ for all $g \in G$. The equivariant local ran-

domized rounding scheme parameterized by the equivariant map $U \mapsto \tilde{X}(YU)$ achieves an approximation ratio of $\left[\frac{\beta w(Y^*U)}{d}\right]^2$. By choosing Y to be the argument of maximum in the normalized maximum width problem for $\mathcal{X}$ we obtain a rounding scheme that achieves and approximation ratio of $\left[\frac{\beta w_\star(\mathcal{X})}{d}\right]^2$.

Proof. Since the local randomized rounding scheme defined by $\hat{X}(U) = \tilde{X}(YU)$ is equivariant, it follows from 5.6.5 that it achieves an approximation ratio of

$$\|A_{\hat{X}}\|_F^2 = \left[\frac{\|\mathbb{E}_{U\sim\gamma}\tilde{X}(YU)U^*\|_F}{d}\right]^2.$$

Now observe that if $\|Y\|_F = 1$ then

$$\begin{aligned}\left\|\mathbb{E}_{U\sim\gamma}[\tilde{X}(YU)U^*]\right\|_F &\geq \mathbb{E}[\langle YU, \tilde{X}(YU)\rangle] \\ &\geq \beta\mathbb{E}_{U\sim\gamma}\left[\max_{X\in\mathcal{X}}\langle X, YU\rangle\right] \\ &= \beta w(Y^*\mathcal{X}).\end{aligned}$$

Hence the local randomized rounding scheme defined by $\hat{X}(U) = \tilde{X}(YU)$ achieves an approximation ratio of at least $\left[\frac{\beta w(Y^*\mathcal{X})}{d}\right]^2$. Choosing Y to be an argument of maximum in the normalized maximum width problem for $\mathcal{X}$ we obtain a rounding scheme with an approximation ratio of at least $\left[\frac{\beta w_\star(\mathcal{X})}{d}\right]^2$. □

5.8 Summary and future directions

In this chapter we considered a class of optimization problems where each of the n variables is constrained to lie in a set $\mathcal{X} \subset \mathbb{K}^{m\times d}$, and where the objective function is linear in the quantities $X_i^*X_j$ for $1 \leq i, j \leq n$. For this class we analyzed the integrality gap (with respect to objective functions defined by positive semidefinite matrices) of a simple semidefinite relaxation that is valid whenever $\mathcal{X}$ consists only of contractions. Our main result (Theorem 5.2.12) exactly characterizes the integrality gap, and shows how to construct rounding schemes that achieve an approximation ratio equal to the integrality gap, in the case where $\mathcal{X}$ satisfies an additional symmetry assumption.

For this specific family of problems we now discuss some possible future research directions. It would be very interesting to understand the situation where $\mathcal{X}$ is not right symmetric. It is very likely that in this case, the gap instances (i.e. the difficult objective functions that produce the tightest upper bounds on the integrality gap) are more complicated because they should exploit the asymmetry of $\mathcal{X}$ (from the right). Nevertheless it may be possible to extend the techniques we use in this chapter to understand this case. Second, it is important to be able to explicitly compute integrality gaps for interesting sets $\mathcal{X}$, or develop good techniques for finding lower bounds on the integrality gaps. While exact values are of considerable interest, tools to numerically solve the normalized maximum width problem and to estimate the approximation ratio achieved by the associated rounding scheme would be useful.

In this chapter we only studied the case where the objective functions correspond to positive semidefinite matrices. The problem is significantly more difficult to understand in the case where the objective function is an arbitrary bilinear form in variables from $\mathcal{X}$. Even in the case where $\mathcal{X} = \{-1, 1\}$ and with the simplest possible semidefinite relaxation, the integrality gap and an optimal rounding scheme remain elusive (see, e.g., [19]). It is possible that thinking about more general sets $\mathcal{X}$ may reveal which structures in the problem are due to binary variables, and which are due to working with arbitrary bilinear forms. The problem has been studied for $\mathcal{X}$ being the sphere and $O(d)$ and $U(d)$. The only situation where an optimal rounding scheme is known is for $U(d)$ in the limit as $d \to \infty$ [61, 84].

The semidefinite relaxation that we analyze is very simple. Tighter relaxations are available for a variety of problems of this type (see, e.g., [118] in the case $\mathcal{X} = SO(d)$). Such relaxations impose more constraints that should be satisfied by elements of $\mathcal{G}_{\mathcal{X}}$. Designing rounding schemes that can exploit more general moment constraints than those appearing in our basic semidefinite relaxation could open the door to being able to automatically design rounding schemes for semidefinite relaxations for a much broader variety of problems.

■ 5.9 Proofs of stationarity results

In this section, we establish the following characterization of when $P = \frac{1}{m}I$ is a stationary point for the normalized maximum width problem. This is restated from Section 5.3.2. Recall that $h_{\mathcal{X}}$ is the support function of $\mathcal{X}$.

Proposition 5.3.16. *Suppose $\mathcal{X} \subset \mathbb{K}^{m\times d}$. Then $P = \frac{1}{m}I$ is a stationary point for*

$$\max_P w(P^{1/2}\mathcal{X}) \quad s.t. \quad P \in \mathcal{H}_+^m, \;\; \mathrm{tr}(P) = 1 \tag{5.9.1}$$

if and only if there is a non-negative scalar κ such that

$$\mathbb{E}_{U\sim\gamma_{m\times d}^{\mathbb{K}}}[UU^* h_{\mathcal{X}}(U)] = \kappa I.$$

Before establishing this result, we use it to prove the following generalization of Lemma 5.4.3 (which dealt with the case where $\mathcal{X}$ is an irreducible tautological orbit). The generalization (Lemma 5.9.1 to follow) uses the notion of a set $\mathcal{X} \subset \mathbb{K}^{m\times d}$ being left-symmetric, which means that there is a group G and an irreducible (over $\mathbb{C}$) orthogonal/unitary representation $\rho : G \to \mathbb{K}^m$ such that $\rho(g)\mathcal{X} = \mathcal{X}$ for all $g \in G$. This is the same as the definition for $\mathcal{X}$ being right-symmetric (Definition 5.2.11) except that now the representation ρ acts on the left.

Lemma 5.9.1. *If $\mathcal{X} \subset \mathbb{K}^{m\times d}$ is left-symmetric then $P = \frac{1}{m}I$ is a stationary point of*

$$\max_P w(P^{1/2}\mathcal{X}) \quad s.t. \quad P \in \mathcal{H}_+^m,\ \operatorname{tr}(P) = 1.$$

Proof. It is enough, by Proposition 5.3.16 to show that if $\mathcal{X}$ is left-symmetric then there is some non-negative κ such that

$$M := \mathbb{E}_{U\sim\gamma^{\mathbb{K}}_{m\times d}}[UU^* h_{\mathcal{X}}(U)] = \kappa I. \tag{5.9.2}$$

Since $\mathcal{X}$ is left symmetric there is a group G and an irreducible (over $\mathbb{C}$) orthogonal/unitary representation ρ such that $\rho(g)\mathcal{X} = \mathcal{X}$ for all $g \in G$. We now show that the matrix M on the left-hand side of (5.9.2) commutes with $\rho(g)$ for all $g \in G$. To do so we first note that since $\rho(g)\mathcal{X} = \mathcal{X}$ for all $g \in G$ we have that

$$h_{\mathcal{X}}(\rho(g)^*W) = h_{\rho(g)\mathcal{X}}(W) = h_{\mathcal{X}}(W)$$

for all $W \in \mathbb{K}^{m\times d}$ and all $g \in G$. Now, for any $g \in G$ we have that

$$\begin{aligned}
\rho(g)\mathbb{E}_{U\sim\gamma^{\mathbb{K}}_{m\times d}}[UU^* h_{\mathcal{X}}(U)] &= \mathbb{E}_{U\sim\gamma^{\mathbb{K}}_{m\times d}}[(\rho(g)U)(\rho(g)U)^* h_{\mathcal{X}}(U)]\rho(g)\\
&= \mathbb{E}_{W\sim\gamma^{\mathbb{K}}_{m\times d}}[WW^* h_{\mathcal{X}}(\rho(g)^*W)]\rho(g)\\
&= \mathbb{E}_{W\sim\gamma^{\mathbb{K}}_{m\times d}}[WW^* h_{\mathcal{X}}(W)]\rho(g)
\end{aligned}$$

where the first equality uses the fact that $\rho(g)^*\rho(g) = I$, the second uses the fact that the Gaussian measure is invariant under left multiplication by orthogonal/unitary matrices, and the third uses the left-symmetry of $\mathcal{X}$. Since M commutes with $\rho(g)$ for all $g \in G$ and ρ is irreducible over $\mathbb{C}$, it follows from Schur's lemma (Lemma 5.6.3) that $M = \kappa I$ for some $\kappa \in \mathbb{C}$. Since M is a non-negative combination of the positive semidefinite matrices UU^*, it follows that M is also positive semidefinite, and so that κ is non-negative. □

Proof of Lemma 5.4.3. Lemma 5.4.3 is the special case of Lemma 5.9.1 when $\mathcal{X} = \rho(G)$ is an irreducible tautological orbit. It holds because any irreducible tautological orbit is left-symmetric. □

We now turn our attention to establishing Proposition 5.3.16.

Proof of Proposition 5.3.16. Let $f(P) = \log w(P^{1/2}\mathcal{X})$ and let $\beta_{\mathbb{K}} = 1$ if $\mathbb{K} = \mathbb{R}$ and $\beta_{\mathbb{K}} = 2$ if $\mathbb{K} = \mathbb{C}$. Note that we can rewrite f in terms of a Gaussian integral as

$$\begin{aligned} f(P) &= \log\left(\int_{U\in\mathbb{K}^{m\times d}} h_{P^{1/2}\mathcal{X}}(U)\exp\left(-\frac{\beta_{\mathbb{K}}}{2}\operatorname{tr}(U^*U)\right)\, dU\right) - \frac{md\beta_{\mathbb{K}}}{2}\log\left(\frac{2\pi}{\beta_{\mathbb{K}}}\right) \\ &= \log\left(\int_{V\in\mathbb{K}^{m\times d}} h_{\mathcal{X}}(V)\exp\left(-\frac{\beta_{\mathbb{K}}}{2}\operatorname{tr}(V^*P^{-1}V)\right)\, dV\right) - \\ &\qquad \frac{md\beta_{\mathbb{K}}}{2}\log\left(\frac{2\pi}{\beta_{\mathbb{K}}}\right) - \frac{\beta_{\mathbb{K}}}{2}\log\det(P). \end{aligned}$$

Here the first equality holds from the definition of the standard Gaussian measure, and the second holds by making the change of variables $V = P^{1/2}U$ in the integral. If P is strictly positive definite then the gradient of $\operatorname{tr}(V^*P^{-1}V)$ is precisely $-P^{-1}VV^*P^{-1}$ and the gradient of $\log\det(P)$ is P^{-1}. So taking the gradient of f with respect to P we obtain

$$\nabla f(P) = -\frac{\beta_{\mathbb{K}}}{2}P^{-1} + \frac{\beta_{\mathbb{K}}}{2}\frac{\int_{V\in\mathbb{K}^{m\times d}} P^{-1}VV^*P^{-1}h_{\mathcal{X}}(V)\exp\left(-\frac{\beta_{\mathbb{K}}}{2}\operatorname{tr}(V^*P^{-1}V)\right)\, dV}{Z(P)} \tag{5.9.3}$$

where

$$Z(P) = \int_{V\in\mathbb{K}^{m\times d}} h_{\mathcal{X}}(V)\exp\left(-\frac{\beta_{\mathbb{K}}}{2}\operatorname{tr}(V^*P^{-1}V)\right)\, dV$$

is a positive scalar-valued function since $h_{\mathcal{X}}(V)$ is positive for almost all V. Making the change of variables $U = P^{-1/2}V$ in the integral on the right-hand side of (5.9.3) we obtain

$$\begin{aligned} \nabla f(P) &= -\frac{\beta_{\mathbb{K}}}{2}P^{-1} + \frac{1}{Z(P)}\frac{\beta_{\mathbb{K}}}{2}\int_{U\in\mathbb{K}^{m\times d}} P^{-\frac{1}{2}}UU^*P^{-\frac{1}{2}}h_{P^{\frac{1}{2}}\mathcal{X}}(U)\exp\left(-\frac{\beta_{\mathbb{K}}}{2}\operatorname{tr}(U^*U)\right)\, dU \\ &= -\frac{\beta_{\mathbb{K}}}{2}P^{-1} + \frac{1}{Z(P)}\frac{\beta_{\mathbb{K}}}{2}P^{-\frac{1}{2}}\mathbb{E}_{U\sim\gamma^{\mathbb{K}}_{m\times d}}[UU^*h_{P^{\frac{1}{2}}\mathcal{X}}(U)]P^{-\frac{1}{2}}. \end{aligned}$$

Now on the interior of the positive semidefinite cone, the only active constraint for the optimization problem (5.9.1) is the constraint $\operatorname{tr}(P) = 1$. Thus introducing the appropriate Lagrange multiplier we see that a positive definite point P is stationary for (5.9.1) if and only if there exists some scalar λ such that

$$\nabla f(P) = -\frac{\beta_{\mathbb{K}}}{2}P^{-1} + \frac{1}{Z(P)}\frac{\beta_{\mathbb{K}}}{2}P^{-1/2}\mathbb{E}_{U\sim\gamma^{\mathbb{K}}_{m\times d}}[UU^*h_{P^{1/2}\mathcal{X}}(U)]P^{-1/2} = \lambda I.$$

Now, if $P = \frac{1}{m}I$ is a stationary point then there exists $\lambda \in \mathbb{R}$ such that

$$-\frac{\beta_{\mathbb{K}} m}{2} I + \frac{\beta_{\mathbb{K}} m}{2Z(I/m)} \mathbb{E}_{U \sim \gamma^{\mathbb{K}}_{m \times d}}[UU^* h_{\mathcal{X}/\sqrt{m}}(U)] = \lambda I.$$

This implies that

$$\mathbb{E}_{U \sim \gamma^{\mathbb{K}}_{m \times d}}[UU^* h_{\mathcal{X}}(U)] = \sqrt{m}\mathbb{E}_{U \sim \gamma^{\mathbb{K}}_{m \times d}}[UU^* h_{\mathcal{X}/\sqrt{m}}(U)] = \kappa I \tag{5.9.4}$$

for some real κ (where we have used the fact that the support function of $\mathcal{X}$ is positively homogeneous, i.e. $h_{t\mathcal{X}} = t h_{\mathcal{X}}$ for all non-negative t). Since UU^* is positive semidefinite for all U, it follows that the left hand side of (5.9.4) is a positive semidefinite matrix and so that $\kappa \geq 0$.

Conversely, suppose that

$$\mathbb{E}_{U \sim \gamma^{\mathbb{K}}_{m \times d}}[UU^* h_{\mathcal{X}}(U)] = \kappa I = \sqrt{m}\mathbb{E}_{U \sim \gamma^{\mathbb{K}}_{m \times d}}[UU^* h_{\mathcal{X}/\sqrt{m}}(U)]$$

for some non-negative κ. Then taking $P = \frac{1}{m}I$ we see that

$$\nabla f(I/m) = -\frac{\beta_{\mathbb{K}} m}{2} I + \frac{\beta_{\mathbb{K}} \kappa}{2Z(I/m)} I$$

and so we can take the Lagrange multiplier $\lambda = -\frac{\beta_{\mathbb{K}} m}{2} + \frac{\beta_{\mathbb{K}} \sqrt{m}\kappa}{2Z(I/m)}$ to establish that $P = I/m$ is a stationary point. $\square$

Chapter 6

A convex approach to learning Gaussian latent tree models

6.1 Introduction

The central problem of statistical modeling is to find a simple probabilistic model that approximates (in an appropriate sense) observations of quantities of interest. Insisting on simple models avoids overfitting and (depending on the model class) may mean that subsequent inference tasks can be performed more efficiently. From the viewpoint of probabilistic graphical models, simpler models are typically those for which many non-trivial conditional independence relations hold among the variables. These relations can be encoded in a graph, and if this graph has low treewidth many inference queries can be performed efficiently[1].

Even if a collection of random variables has many non-trivial conditional independence relations, the marginal distribution on a subset of variables often has none at all. Put another way, even if there does not seem to be any interesting conditional independence relations among a collection of observed random variables, we could just be observing a *subset* of variables in a model with more variables and a much simpler structure. This observation suggests modeling given data as observations of a subset of the variables in a larger, simple model with additional *latent variables*. In some situations these latent variables may have natural interpretations, as true unobserved effects that explain correlation among the data. In general they may not have such clear interpretations, but still allow us to work with more expressive models that are of low complexity and may allow for efficient inference.

Particularly interesting are latent variable models where the latent variables have a hierarchical dependence structure. Models with this flavor range from 'deep Boltzmann machines' (see, e.g., [109]) to the multiresolution models used in applications such as geophysics and oceanography [137]. These models are able to capture correlations oc-

[1]This is the case, at least, when all the variables are jointly Gaussian or all the variables take finitely many values

curring at multiple scales in the data, while (in some cases) maintaining computational tractability.

The simplest hierarchical latent variable models are latent tree models. Here the overall model consists of a collection of (not necessarily scalar) random variables that are Markov with respect to a tree (see Section 6.2). The observed variables are often (but not necessarily) the variables corresponding to the leaves of the tree. These models have the distinct advantage that inference tasks can typically be carried out via naturally asynchronous and distributed message-passing algorithms on the tree, with complexity that is linear in the number of vertices in the tree (but depends also on the complexity of 'local' inference queries at each node).

In this chapter we focus on latent tree models where all the variables are jointly Gaussian and where we observe the leaf-indexed variables of the tree model. Our overall aim is to develop tractable methods to learn the parameters, the dimensions of the latent variables, and the tree structure of a Gaussian latent tree model given only second-order statistics among the leaf-indexed variables.

For the purposes of analysis, we are interested in a planted version of the problem, where we assume there is a true underlying tree model with low-dimensional latent variables and we aim to (approximately) recover it. On the other hand, even if we have no reason to believe that our data come from a tree model we may still want to approximate it as such for computational reasons. With this in mind, we aim to develop global methods that always produce valid Gaussian latent tree models, and do not rely on the underlying data having any particular structure. In this setting, our aim is to balance the quality of the approximation with complexity of the latent tree model, which is dominated by the largest dimension of a latent variable.

■ 6.1.1 Basic approach and main contributions

We now explain the basic approach taken in the chapter and our main contributions.

The key observation we use is that the covariance matrices among the leaf-indexed variables in a Gaussian tree model are characterized by admitting a structured matrix decomposition of the form $\Sigma = \sum_{v\in\mathcal{V}} X_v$ with one term for every vertex of the tree. The terms X_v in the decomposition are all positive semidefinite, the *supports* of the terms encode the combinatorial structure of the tree, and the *ranks* of certain linear combinations of elements of the X_v encode the dimensions of the latent variables. We call such a decomposition of a positive semidefinite matrix a *latent tree covariance decomposition* (LTCD), a notion defined precisely in Section 6.3 to follow.

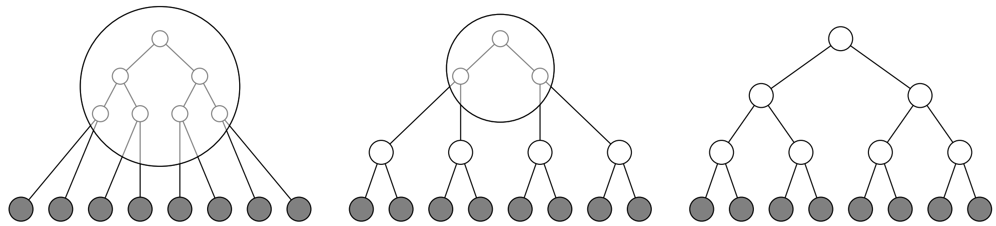


Figure 6.1: There are many valid latent tree models corresponding to the covariance among the leaves of a tree model. Our method for finding the tree structure involves first finding a latent tree model with one 'large' (higher-dimensional) latent variable, as shown on the left. It then searches for additional hierarchical structure in this latent variable (as shown in the center diagram) until no more structure can be found (as shown on the right).

Known tree structure

If the tree structure is fixed and assumed known, then the supports of all the terms in an LTCD are fixed and known. As such, the set of all possible LTCDs of a given positive semidefinite matrix X with respect to a fixed tree is a convex set defined by linear equality constraints and positive semidefiniteness conditions (see Section 6.3). This means that instead of searching over explicit parameterizations of Gaussian latent tree models (Markov with respect to our fixed tree), we can search for LTCDs. Instead of explicitly fixing the dimensions of the latent variables in a parameterization, we search for LTCDs for which the appropriate linear combinations of the terms have low rank, as these correspond to Gausian latent tree models with low-dimensional latent variables. By minimizing different convex functions as surrogates for the rank function, we obtain different convex optimization-based methods to attempt to construct LTCDs.

Our first contribution is to propose and analyze a convex optimization-based method that aims to perform LTCDs with respect to a fixed, given, tree structure. We establish sufficient conditions on an underlying Gaussian latent tree model that ensure our method can exactly recover the model parameters (including dimensions of the latent variables) from the covariance among the leaf-indexed variables. The conditions are geometric in nature, and are expressed in terms of the principal angles between certain subspaces of the observation space associated with each variable in the latent tree model.

Unknown tree structure

Our second contribution is to propose and analyze a method to recover the tree structure (in addition to the dimensions of the latent variables and the model parameters) in the case in which we are only given the covariance among the leaf-indexed variables. Our method uncovers the unknown tree structure in stages from the leaf to the root. At each stage we solve a convex optimization problem, extract additional partial information about the tree structure from its solution, and then use this information to define a more refined convex optimization problem, allowing us to progress to the next stage.

The main idea behind the method is that the covariance among the leaf-indexed variables of a Gaussian tree model has many LTCDs with respect to many trees (see Figure 6.1). Indeed if a matrix has an LTCD with respect to one rooted tree, it also has an LTCD with respect to any tree obtained by contracting into a single variable, an entire rooted subtree of the original tree. The extreme case of this is to group together all of the non-leaf vertices, in which case the observed covariance has an LTCD with respect to a star-shaped tree. Our method proceeds by first performing an LTCD of the given leaf-covariance with respect to this star-shaped tree. We do this using our convex optimization-based method for constructing LTCDs. We can recover the next layer of the tree from the column space of the root-indexed term in the decomposition as it tells us how the leaves are connected to their parents. We then repeat the process, each time finding an LTCD with respect to a more refined tree based on the additional structure we have uncovered, until there is no remaining structure in the root-indexed latent variable.

Throughout our discussion of recovering the tree structure, we restrict attention to trees where all the leaves are at the same distance from the root. When the underlying Gaussian model is non-singular, we show that our procedure correctly recovers all of these structures under very similar conditions to those that ensure our convex optimization-based method works when the tree structure is given.

■ 6.1.2 Related work

The simplest Gaussian latent tree model is the factor analysis model, introduced by Spearman in 1904 [125]. In the system identification literature this model is referred to as the 'Frisch scheme' [70] after the Norwegian economist Ragnar Frisch. The problem of fitting the model parameters of a factor analysis model remains a challenge, and is an actively studied problem (see, e.g., [13] for a recent survey of new algorithmic approaches). As with most latent variable models, the dominant approach in practice is to use the expectation maximization algorithm (EM) [37]. Being based on local optimization, this approach offers few guarantees, (although see, e.g., [5] for recent developments in understanding EM). Furthermore, approaches based on EM do not naturally extend to more complicated latent tree models where the tree structure itself is unknown.

In the multiscale signal processing literature, Gaussian latent tree models are referred to as Multiscale Auto-Regressive (MAR) processes [11]. The stochastic realization problem for MAR processes involves studying which models can arise as the marginal distribution among the observed variables (typically leaves) of MAR processes. This problem is considered, for instance, in [33, 68, 47]. In these works the tree structure and dimensions of the latent variables are fixed, a priori. The work of Daoudi et al. [33] focuses on establishing connections between the realization problem and wavelets. In [68], a method based on the idea of canonical correlations is proposed to fit MAR processes to data. A related, yet more computationally efficient, approach appears in [47], but is restricted to so-called internal models, a strict subset of all MAR processes [67]. These methods are inherently local in nature, aiming to fit parameters in such a way that certain subsets of variables are decorrelated according to the constraints imposed by the tree structure and the fixed state dimensions.

Convex optimization-based matrix decomposition methods for learning Gaussian latent variable models date at least to the 1940 work of Ledermann [75]. Ledermann gave a characterization of when the minimum trace factor analysis heuristic (a special case of the method we propose in Section 6.4) recovers factor analysis models with a single, one-dimensional, latent variable. Much more recently, Chandrasekaran, Parrilo, and Willsky [27] propose and carefully analyze a convex optimization-based matrix decomposition approach to a more complex latent variable modeling problem. In that work the aim is to decompose the inverse covariance of the observations as a sparse positive semidefinite matrix minus a low-rank positive semidefinite matrix, corresponding to a sparse graphical model structure among the observed variables, and a single low-dimensional latent variable. In contrast, latent tree models focus on uncovering structure among the latent variables, while assuming the simplest possible structure among the observed variables.

Recently, methods have been developed to learn the tree structure of latent tree models from data. Choi et al. [31] developed an approach based on the fact that the information distances among the observed variables in a latent tree model must form a tree metric. From this observation, the authors develop a method based on quartet tests to learn the latent tree structure when the latent variables are scalar Gaussians or discrete random variables with a common finite alphabet. They extend this to a more global method that first learns a tree model among the observed variables, and then adds latent structure to the tree via quartet tests. Notably, the methods of Choi et al. can deal with the case where any subset of the variables is observed, not just the leaves. These ideas are extended to the case where the latent variables are non-scalar by Anandkumar et al. [4] who developed a spectral generalization of the quartet tests of Choi et al. [31].

6.1.3 Outline

The remainder of the chapter is organized as follows. Section 6.2 summarizes terminology and notational conventions, establishes some useful linear algebraic facts that are used repeatedly in the chapter, and describes basic facts about Gaussian tree models, including the structure of their covariance matrices. In Section 6.3 we examine Gaussian latent tree models, and relate them to LTCDs, which are certain structured matrix decompositions (see Definition 6.3.4). The main result of the section is Theorem 6.3.9. It characterizes the covariance among a subset of the variables of a Gaussian tree model in terms of such a matrix decomposition. In Section 6.4 we propose a convex optimization-based method to perform LTCDs when the tree structure is given but the dimensions of the latent variables are not given. Section 6.4.2 is focused on establishing conditions on an underlying Gaussian latent tree model under which our convex optimization method can exactly decompose the covariance among the leaves, allowing us to recover the full tree model. We conclude the section with a discussion of alternative objective functions in our convex optimization problem that are complicated to analyze but may lead to better results in practice. In Section 6.5 we no longer assume that the tree structure is known. We show that the convex optimization methods from Section 6.4 actually reveal information about the tree structure in the optimal dual variables. We propose a method that, by solving a sequence of convex optimization problems, uncovers the tree structure and performs an LTCD to recover the model parameters.

6.2 Preliminaries

In this section we first introduce notation and terminology related to trees and matrices used throughout the chapter. The second part of the section introduces Gaussian tree models with a focus on the structure of their covariance matrices.

6.2.1 Trees

An undirected *tree* $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ is a connected acyclic graph with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. The *leaves* of a tree $\mathcal{T}$ are the vertices of degree one. There is a unique path joining any pair of vertices $u, v \in \mathcal{V}$. A *rooted tree* is an undirected tree $\mathcal{T}$ together with a distinguished vertex called the *root*. A choice of root induces a natural partial order, the *tree order*, on the vertices of a tree, with $u \preceq v$ if and only if the path from the root to u passes through v. The length of the path joining vertices u and v is the *distance* $d(u, v)$ from u to v. A vertex u is the *parent* of v, denoted $u = \mathcal{P}(v)$, if $u \succ v$ and $d(u, v) = 1$. The *children* of v, denoted $\mathcal{C}(v)$, are those vertices with parent v. The set of *ancestors* of v is $\{u \in \mathcal{V} : u \succeq v\}$ and the set of *descendants* of v is $\mathcal{D}(v) := \{u \in \mathcal{V} : u \preceq v\}$. For any pair of vertices $u, v \in \mathcal{V}$ there is a least upper bound

$u \vee v \in \mathcal{V}$ that is furthest from the root among all common ancestors of u and v. See Figure 6.2 for a diagram showing this notation.

Trees and collections of subsets

Often, in what follows, we are interested in a tree with additional structure. In addition to a tree we usually have n scalar variables that are partitioned among the k leaves of the tree (where $k \leq n$). As such, the structure of interest consists of a tree with k leaves together with a partition of $[n]$ into k disjoint sets. For example, the tree on the right in Figure 6.2 has $k = 8$ leaves and $n = 10$. The $k = 8$ shaded sets form a partition of $[10]$. Suppose we have established this correspondence between leaves $\mathcal{L}$ of a tree and a partition of $[n]$. Let the sets in the partition be $(S_\ell)_{\ell \in \mathcal{L}}$. Then we can think of the entire tree structure just in terms of subsets of $[n]$. We associate with a vertex w in the tree, the subset S_w obtained by taking the (disjoint) union of the sets S_ℓ over all leaves ℓ that are descendants of w. For example, the vertex labeled w on the left in Figure 6.2 corresponds to the subset $S_w = \{1, 2, 3, 4\}$. This is because the leaves that are descendants of w correspond to the sets $\{1, 2\}, \{3\}, \{4\}$ and the union of these sets is $\{1, 2, 3, 4\}$. Similarly, since all of the leaves are descendants of the root r, the set corresponding to the root is $\cup_{\ell \in \mathcal{L}} S_\ell = [10]$.

This view of trees via subsets is notationally very useful throughout the chapter. The following terminology captures which collections of subsets correspond to trees.

Definition 6.2.1. A collection $(S_v)_{v \in \mathcal{V}}$ of subsets of $[n]$ *forms a tree* if there exists a rooted tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ so that for all vertices u that are not leaves,

$$S_u = \dot{\bigcup}_{v \in \mathcal{C}(u)} S_v \tag{6.2.1}$$

where $\dot{\cup}$ denotes disjoint union.

With any partially ordered set there is a standard way to construct a directed acyclic graph, called the *Hasse diagram* [126, Chapter 3], of the partial order. In the present context, the vertices of the diagram are labeled with the sets S_v. A set S_u is a parent of S_v if $S_u \supsetneq S_v$ and there does not exist another set S_w between S_u and S_v in the sense that $S_u \supsetneq S_w \supsetneq S_v$. The Hasse diagram of the following collection of fourteen subsets of $[10]$ is shown on the right in Figure 6.2[2]

$$\{1, 2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8, 9\}, \{10\}, \{1, 2, 3\},$$
$$\{1, 2, 3, 4\}, \{5, 6\}, \{8, 9, 10\}, \{7, 8, 9, 10\}, \{1, 2, 3, \ldots, 10\}.$$

[2]It is conventional not to explicitly show the orientation of the edges in a Hasse diagram.

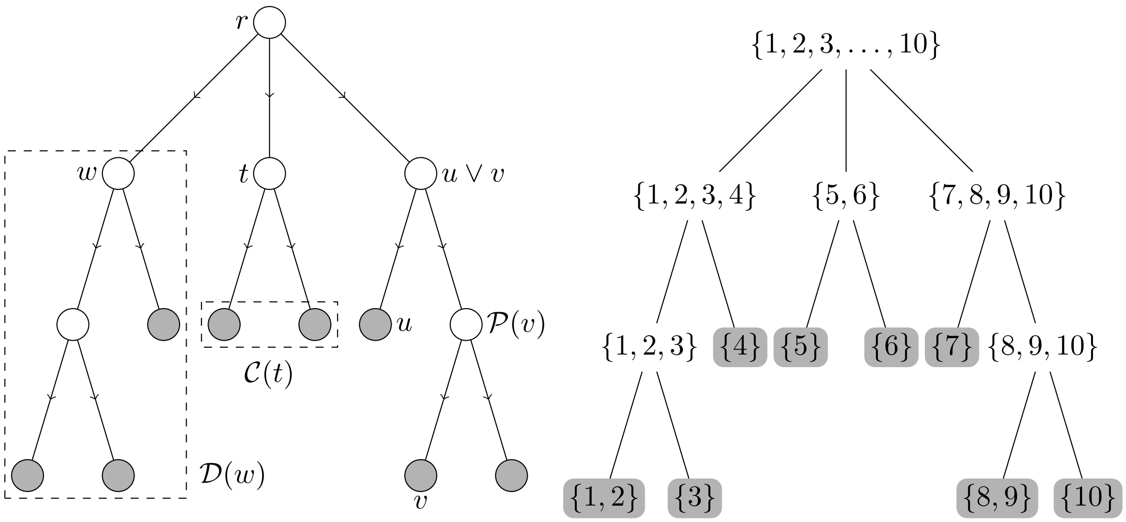


Figure 6.2: On the left is a rooted tree illustrating various notational conventions used throughout the paper. The eight leaves are shaded. On the right is the Hasse diagram (see the discussion following Definition 6.2.1) of a collection of fourteen subsets of $\{1, 2, 3, \ldots, 10\}$. This collection of subsets forms the rooted tree (in the sense of Definition 6.2.1) appearing on the left. The leaves are shaded to aid comparison of the two diagrams. Correspondences between the tree and the collection of subsets are explained under the heading 'Trees and collections of subsets' in Section 6.2.1.

Concretely, a collection of subsets of $[n]$ forms a tree if its Hasse diagram is a tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ and the condition (6.2.1) is satisfied with respect to that tree.

■ 6.2.2 Matrices and linear algebra

General notation

Let $\mathbb{R}^n$ denote real vectors with n components, $\mathbb{R}^{n \times m}$ denote $n \times m$ matrices with real entries, and $\mathcal{S}^n$ the space of symmetric $n \times n$ matrices. We equip $\mathbb{R}^n$ with the standard inner product $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$. We equip $\mathcal{S}^n$ with the trace inner product $\langle X, Y \rangle = \operatorname{tr}(XY)$.

If $x \in \mathbb{R}^n$ and $I \subseteq [n]$ we say that x is *supported on* I if $x_i \neq 0$ implies $i \in I$. If $I \subseteq [n]$ we use the notation $\mathbb{R}^I$ to denote the coordinate subspace of $\mathbb{R}^n$ consisting of vectors supported on I. We use the notation $\mathcal{S}^I$ to denote the subspace of $\mathcal{S}^n$ consisting of symmetric matrices X such that if $X_{ij} \neq 0$ then $i, j \in I$.

We denote the column space of an $n \times m$ matrix A by $\operatorname{col}(A) \subseteq \mathbb{R}^n$. Given a subspace

$U \subseteq \mathbb{R}^n$ we define its *support* to be

$$\mathrm{supp}(U) := \bigcap_{\substack{I \subseteq [n] \\ \mathbb{R}^I \supseteq U}} I$$

i.e. I is the inclusion-wise minimal set on which every element of U is supported. Observe that if $X \in \mathcal{S}^n$ and $I = \mathrm{supp}(\mathrm{col}(X))$ then I is the inclusion-wise minimal subset of $[n]$ such that $X_{ij} \neq 0$ whenever $i, j \in I$.

If U is a subspace of $\mathbb{R}^n$ let $P_U : \mathbb{R}^n \to \mathbb{R}^n$ be the orthogonal projector onto U, i.e. the positive semidefinite linear map such that $\mathrm{col}(P_U) = U$, $P_U^2 = P_U$ and $\mathrm{tr}(P_U) = \dim(U)$. Similarly let $\pi_U : \mathbb{R}^n \to U$ be the linear map such that $\pi_U \pi_U^*$ is the identity map on U and $P_U = \pi_U^* \pi_U$. If $I \subseteq [n]$ and $\mathbb{R}^I$ is the associated coordinate subspace, we use the shorthand P_I and π_I instead of $P_{\mathbb{R}^I}$ and $\pi_{\mathbb{R}^I}$ to simplify the notation. Furthermore, if X is an $n \times n$ matrix and $I, J \subseteq [n]$ we use the shorthand $X_{IJ} := \pi_I X \pi_J^*$ if it is convenient to do so.

Positive semidefinite matrices

Let $\mathcal{S}_+^n$ denote $n \times n$ positive semidefinite matrices and if $I \subseteq [n]$ denote by $\mathcal{S}_+^I$ the $|I| \times |I|$ positive semidefinite matrices with rows and columns indexed by I. We write $B \succeq A$ if $B - A \in \mathcal{S}_+^n$ for some n.[3] The following simple observation about the column spaces of positive semidefinite matrices is used at various points throughout the chapter.

Lemma 6.2.2. *If A and B be symmetric matrices satisfying $B \succeq A \succeq 0$ then $\mathrm{col}(B) \supseteq \mathrm{col}(A)$.*

Proof. If M is any positive semidefinite matrix then $\mathrm{col}(M) = \{x : \langle Mx, x\rangle > 0\} \cup \{0\}$. Since $B \succeq A \succeq 0$, if $x \in \mathrm{col}(A) \setminus \{0\}$ then $\langle Bx, x\rangle \geq \langle Ax, x\rangle > 0$ and so $x \in \mathrm{col}(B) \setminus \{0\}$. □

■ 6.2.3 Gaussian tree models

Let $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ be a tree and $(x_v)_{v \in \mathcal{V}}$ a collection of jointly Gaussian random variables with mean zero[4]. Suppose x_v takes values in $\mathbb{R}^{n_v}$ for each $v \in \mathcal{V}$. We call $\mathbb{R}^{n_v}$ the *state space* of x_v and call n_v the *dimension* of x_v. The collection $(x_v)_{v \in \mathcal{V}}$ is *Markov* with respect to $\mathcal{T}$ if whenever $\{u, v\} \notin \mathcal{E}$ and $w \notin \{u, v\}$ is on the (unique) path between u and v, the variables x_u and x_v are conditionally independent given x_w. We say such a collection of Gaussian random variables follows a *Gaussian tree model*.

[3] Recall that we also use the notation $u \succeq v$ for the partial order on the vertices of a rooted tree. The intended interpretation will be clear from the context.

[4] We make this assumption throughout to simplify the exposition.

If $(x_v)_{v\in\mathcal{V}}$ follows a Gaussian tree model with associated tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ (rooted at r) and dimensions $(n_v)_{v\in\mathcal{V}}$ then it has a parameterization as a generative process as follows. Let $(Q_v)_{v\in\mathcal{V}}$ be $n_v \times n_v$ positive semidefinite matrices and let $(A_v)_{v\in\mathcal{V}\setminus\{r\}}$ be $n_v \times n_{\mathcal{P}(v)}$ real matrices. If $(x_v)_{v\in\mathcal{V}}$ is defined via the root-to-leaves generative process

$$x_r \sim \mathcal{N}(0, Q_r) \quad \text{and} \quad x_v | x_{\mathcal{P}(v)} \sim \mathcal{N}(A_v x_{\mathcal{P}(v)}, Q_v) \tag{6.2.2}$$

then $(x_v)_{v\in\mathcal{V}}$ follows a Gaussian tree model with tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ (rooted at r) and dimensions $(n_v)_{v\in\mathcal{V}}$. Furthermore, every Gaussian tree model can be obtained this way. We can express (6.2.2) more explicitly in state space form as

$$x_r = w_r \quad \text{and} \quad x_v = A_v x_{\mathcal{P}(v)} + w_v \quad \text{for all } v \neq r \tag{6.2.3}$$

where $w_v \sim \mathcal{N}(0, Q_v)$ for all $v \in \mathcal{V}$.[5] We do not require that the covariance of the entire tree model $(x_v)_{v\in\mathcal{V}}$ be strictly positive definite in this parameterization. This additional flexibility is important in this chapter.

Covariance of Gaussian tree models

We now describe Σ, the covariance of the collection of random variables $(x_v)_{v\in\mathcal{V}}$. We first do so explicitly in terms of $(A_v)_{v\in\mathcal{V}\setminus\{r\}}$ and $(Q_v)_{v\in\mathcal{V}}$. We then rewrite this description in a more compact way so that the notation is more concise in subsequent sections.

If $v \in \mathcal{V}$ then from (6.2.2) it is clear that

$$\Sigma_{vv} := \mathbb{E}[x_v x_v^T] = Q_v + A_v \Sigma_{\mathcal{P}(v)} A_v^T.$$

Recursively applying this formula we see that if $v = v_1, v_2, \ldots, v_k, v_{k+1} = r$ is the path from v to the root r then

$$\Sigma_{vv} = Q_{v_1} + A_{v_1} Q_{v_2} A_{v_1}^T + (A_{v_1} A_{v_2}) Q_{v_3} (A_{v_1} A_{v_2})^T + \cdots + \\ (A_{v_1} A_{v_2} \cdots A_{v_k}) Q_{v_{k+1}} (A_{v_1} A_{v_2} \cdots A_{v_k})^T.$$

From this we can see that the covariance of a single variable x_v depends only on its ancestors. If $v, w \in \mathcal{V}$, let $v \vee w$ denote their least common ancestor (see Figure 6.2). Let the path from v to w be $v = v_1, v_2, \ldots, v_k, v_{k+1} = v \vee w = w_{\ell+1}, w_\ell, \ldots, w_2, w_1$. Then the covariance between v and w is given by

$$\Sigma_{vw} = (A_{v_1} A_{v_2} \cdots A_{v_k}) \Sigma_{v\vee w\, v\vee w} (A_{w_1} A_{w_2} \cdots A_{w_\ell})^T$$

[5] We emphasize that in (6.2.3) the index of the updated state x_v and the noise w_v are the same. This is different from a typical state space description of the form $x_{k+1} = Ax_k + w_k$, where the index of the updated state x_{k+1} and the noise w_k are different.

and so only depends on the ancestors of $v \vee w$, i.e. the common ancestors of v and w.

We now introduce notation to describe Σ, the covariance of the $(x_v)_{v\in\mathcal{V}}$, more concisely. Given matrices $(A_v)_{v\in\setminus\{r\}}$ with $A_v \in \mathbb{R}^{n_v \times n_{\mathcal{P}(v)}}$ define a block matrix N by

$$N_{vu} = \begin{cases} A_v & \text{if } u = \mathcal{P}(v) \\ 0 & \text{otherwise.} \end{cases} \tag{6.2.4}$$

If we order the vertices of the tree in a perfect elimination ordering[6] then N is block strictly upper triangular. Define a block matrix Φ by $\Phi := (I-N)^{-1}$. It can be verified (by computing $(I-N)\Phi$, for instance) that the block entries of Φ are given by the products of the A_v along directed paths in $\mathcal{T}$, i.e. by

$$\Phi_{vu} = \begin{cases} 0 & \text{if } u \text{ and } v \text{ are incomparable} \\ I & \text{if } u = v \\ A_v\Phi_{\mathcal{P}(v)u} & \text{if } v \prec u. \end{cases} \tag{6.2.5}$$

Observe that $\Phi_{vu} \in \mathbb{R}^{n_v \times n_u}$ describes a transition map from the state space for x_u to the state space for x_v. In this notation it is straightforward to check that

$$\Sigma_{vw} := \mathbb{E}[x_v x_w^T] = \sum_{u \succeq v \vee w} \Phi_{vu} Q_u \Phi_{wu}^T \tag{6.2.6}$$

because x_v and x_w can only be correlated via their common ancestors. Again if we order $\mathcal{V}$ in a perfect elimination ordering then Φ is upper triangular with identity matrices on the block diagonal. Let Q be the block diagonal matrix with the Q_v on the block diagonal, ordered in the same way as Φ. Then the covariance of the corresponding Gaussian tree model is

$$\Sigma = (I-N)^{-1}Q(I-N)^{-T} = \Phi Q \Phi^T = \sum_{v\in\mathcal{V}} \Phi_{\mathcal{V}v} Q_v \Phi_{\mathcal{V}v}^T \tag{6.2.7}$$

where $\Phi_{\mathcal{V}v}$ is the block column of Φ indexed by v. This is consistent with (6.2.6) because the column of Φ indexed by v is only supported on the blocks corresponding to the descendants of v.

To clarify this notation we now describe an example. We return to this example in Section 6.3.

Example 6.2.3. Consider the rooted tree shown on the left in Figure 6.3. Suppose the collection of matrices $A_{v_1}, A_{v_2}, A_{w_1}, A_{w_2}, \ldots, A_{w_5}$ and $Q_r, Q_{v_1}, Q_{v_2}, Q_{w_1}, \ldots, Q_{w_5}$

[6] For a tree, a perfect elimination ordering is an ordering of the vertices such that for any vertex u, the only neighbor of u occurring after u in the ordering is its parent (see, e.g., [107]).

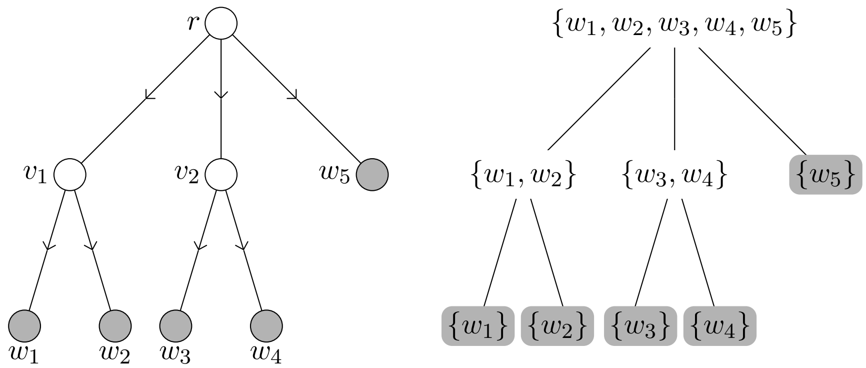


Figure 6.3: On the left is the tree structure for Examples 6.2.3 and 6.3.2. On the right is the Hasse diagram of the collection $(\mathcal{D}(u) \cap \mathcal{L})_{u \in \mathcal{V}}$ of subsets of leaves of the tree on the left. Note that, for example, the vertex v_1 on the left corresponds to the set $\mathcal{D}(v_1) \cap \mathcal{L}$ on the right.

parameterize a Gaussian tree model with respect to this tree. Then the covariance Σ of the model has a factorization as $\Phi Q \Phi^T$. Writing this explicitly we obtain

$$\begin{bmatrix} I \\ A_{v_1} & I \\ A_{v_2} & 0 & I \\ A_{w_1}A_{v_1} & A_{w_1} & 0 & I \\ A_{w_2}A_{v_1} & A_{w_2} & 0 & 0 & I \\ A_{w_3}A_{v_2} & 0 & A_{w_3} & 0 & 0 & I \\ A_{w_4}A_{v_2} & 0 & A_{w_4} & 0 & 0 & 0 & I \\ A_{w_5} & 0 & 0 & 0 & 0 & 0 & 0 & I \end{bmatrix} \begin{bmatrix} Q_r \\ & Q_{v_1} \\ & & Q_{v_2} \\ & & & Q_{w_1} \\ & & & & Q_{w_2} \\ & & & & & Q_{w_3} \\ & & & & & & Q_{w_4} \\ & & & & & & & Q_{w_5} \end{bmatrix} \times$$
$$\begin{bmatrix} I \\ A_{v_1} & I \\ A_{v_2} & 0 & I \\ A_{w_1}A_{v_1} & A_{w_1} & 0 & I \\ A_{w_2}A_{v_1} & A_{w_2} & 0 & 0 & I \\ A_{w_3}A_{v_2} & 0 & A_{w_3} & 0 & 0 & I \\ A_{w_4}A_{v_2} & 0 & A_{w_4} & 0 & 0 & 0 & I \\ A_{w_5} & 0 & 0 & 0 & 0 & 0 & 0 & I \end{bmatrix}^T$$

(where any missing entries are zeros). Each column of the matrix on the left corresponds to $\Phi_{\mathcal{V}u}$ for some vertex $u \in \mathcal{V}$. For example,

$$\Phi_{\mathcal{V}r} = \begin{bmatrix} I \\ A_{v_1} \\ A_{v_2} \\ A_{w_1}A_{v_1} \\ A_{w_2}A_{v_1} \\ A_{w_3}A_{v_2} \\ A_{w_4}A_{v_2} \\ A_{w_5} \end{bmatrix} \quad \text{and} \quad \Phi_{\mathcal{V}v_2} = \begin{bmatrix} 0 \\ 0 \\ I \\ 0 \\ 0 \\ A_{w_3} \\ A_{w_4} \\ 0 \end{bmatrix}.$$

6.3 Gaussian latent tree models and matrix decompositions

A collection of jointly Gaussian random variables follows a *Gaussian latent tree model* if they can be realized as the marginal distribution among a subset of variables in a Gaussian tree model. In this section we first discuss notions of minimality and singularity for Gaussian latent tree models, and see that these are quite distinct. We then characterize the marginal covariance among the leaf-indexed variables of a Gaussian tree model in terms of structured matrix decompositions. We are particularly interested in Gaussian latent tree models with low-dimensional latent variables. The dimensions of the latent variables appear as the ranks of linear combinations of the matrices appearing in the corresponding matrix decompositions.

6.3.1 Minimality and singularity

In this section we discuss two aspects of latent tree models. The first is minimality, whether the dimensions of the latent variables can be reduced while maintaining the same marginal distribution among the observed variables and the same Markov structure among all the variables. The second is singularity, whether the joint distribution of all the variables in the model is singular. We present a simple example (Example 6.3.1 to follow) that shows that these notions are different.

Minimality

Suppose $(x_v)_{v\in\mathcal{V}}$ is Markov with respect to $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ and is parameterized by $(A_v)_{v\in\mathcal{V}\setminus\{r\}}$ and $(Q_v)_{v\in\mathcal{V}}$. The marginal distribution at the leaf-indexed variables follows (by definition) a latent tree model with respect to $\mathcal{T}$. The joint model can fail to be a minimal realization with respect ot $\mathcal{T}$ in one of two ways.

1. The marginal covariance Σ_{vv} at some variable $v \in \mathcal{V}$ could be singular. In this case there are parts of the state space at v on which x_v is deterministically zero. We can reparameterize the latent variable to create a new realization in which x_v

has a lower dimensional state space without affecting the marginal distribution at the leaves. We do this by taking $U = \text{col}(\Sigma_{vv})$, replacing A_v with $\pi_U A_v$, replacing A_w with $A_w \pi_U^*$ for $w \in \mathcal{C}(v)$, and replacing Q_v with $\pi_U Q_v \pi_U^*$. The new state space has dimension $\dim(U)$ and the marginal covariance is now $\pi_U^* \Sigma_{vv} \pi_U$ which is non-singular.

2. The map $\Phi_{\mathcal{L}v}$ from the state space at v to the observation space could fail to be injective (i.e. have a non-trivial nullspace). In this case there are some parts of the state space at v that have no effect on the observed variables, and so need not be represented. We do this explicitly by taking U to be the nullspace of $\Phi_{\mathcal{L}v}$, replacing A_v with $\pi_{U^\perp} A_v$, replacing A_w with $A_w \pi_{U^\perp}^*$ for $w \in \mathcal{C}(v)$, and replacing Q_v with $\pi_{U^\perp}^* Q_v \pi_{U^\perp}$.

Given a tree model, applying these operations (where possible) reduces it to a minimal one. Our objective, overall, is to construct latent tree models. The methods we describe always produce minimal models (via the construction in Theorem 6.3.9 to follow).

If either Σ_{vv} is singular or $\Phi_{\mathcal{L}v}$ is not injective then the column space of

$$\Phi_{\mathcal{L}v} \Sigma_{vv} \Phi_{\mathcal{L}v}^T$$

has dimension that is smaller than the dimension of the state space (the inner dimension in the factorization of this matrix). As such, the dimensions of the column spaces of the matrices $\Phi_{\mathcal{L}v} \Sigma_{vv} \Phi_{\mathcal{L}v}^T$ (for all latent variables v) tell us the dimensions of the state spaces in a corresponding minimal model. The column spaces of these matrices play an important role in this chapter. We revisit them in Definition 6.3.3 to follow.

We have seen that failures of minimiality occur because of rank deficiencies in quantities related to a single latent variable. These can be fixed by making local coordinate changes. Since these coordinate changes are local to each latent variable they do not affect the overall Markov structure of the model.

Singularity

A quite distinct phenomenon occurs when the *overall* latent tree model has singular covariance. We call these *singular models*. It is often not possible to fix this without breaking the Markov structure among the variables. Indeed it is quite possible for the overall model to be singular even when

1. the model is minimal and

2. the marginal covariance among all of the leaf variables is non-singular.

The following example gives a simple illustration of this.

Example 6.3.1. Consider a tree with three vertices, a root r and two leaves u and v. Define a Gaussian tree model via

$$x_r = w_r, \quad x_u = \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_r + w_u, \quad \text{and} \quad x_v = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} x_r + w_v$$

where w_r, w_u, and w_v each have mean zero and covariances, respectively,

$$Q_r = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad Q_u = Q_v = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

We claim that the model is minimal and that the marginal covariance among the leaves is non-singular, and yet the joint model is singular.

The joint covariance of the entire model is (by (6.2.7))

$$\Sigma = \begin{bmatrix} 1 & 0 & & & & \\ 0 & 1 & & & & \\ 1 & -1 & 1 & 0 & & \\ 0 & 0 & 0 & 1 & & \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & & & & \\ 0 & 2 & & & & \\ & & 0 & 0 & & \\ & & 0 & 1 & & \\ & & & & 0 & 0 \\ & & & & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & & & & \\ 0 & 1 & & & & \\ 1 & -1 & 1 & 0 & & \\ 0 & 0 & 0 & 1 & & \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}^T$$

which is singular since the left and right matrices are invertible and the matrix in the middle is singular. On the other hand, the model is minimal. The marginal covariances at the leaves u and v are

$$\Sigma_{uu} = \Sigma_{vv} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$$

which are clearly non-singular. The marginal covariance at the root is

$$\Sigma_r = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

which is also non-singular. Furthermore, the map

$$\Phi_{\mathcal{L}r} = \begin{bmatrix} 1 & -1 \\ 0 & 0 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}$$

is injective. Hence the model is minimal. Finally, the joint covariance of the leaf-indexed variables is, by a straightforward calculation

$$\begin{bmatrix} \Sigma_{uu} & \Sigma_{uv} \\ \Sigma_{vu} & \Sigma_{vv} \end{bmatrix} = \begin{bmatrix} 3 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

which is non-singular.

Such models, that are minimal and yet singular, arise naturally when connecting wavelets and ideas in multi-scale signal processing (see, e.g., [46, 33]). As such, we do not want to rule them out. We revisit issues related to singular and non-singular models at the end of Section 6.3.2 to follow in connection with the matrix decomposition parameterization of latent tree models developed in that section. Furthermore, differences between singular and non-singular models are important in Section 6.5 in which we analyze a method to learn the tree structure associated with a Gaussian latent tree model.

■ 6.3.2 Latent tree covariance decompositions

Recall from Section 6.2 that the covariance of a Gaussian tree model has a decomposition as

$$\Sigma = \sum_{v\in\mathcal{V}} \Phi_{\mathcal{V}v} Q_v \Phi_{\mathcal{V}v}^T.$$

The terms in this decomposition are all positive semidefinite. Furthermore, the column space of $\Phi_{\mathcal{V}v}$ has support contained in the coordinate subspace corresponding to $\mathcal{D}(v)$, the descendants of v. As such, the supports of the terms encode the tree structure.

Suppose, now, we observe only the variables indexed by $\mathcal{L} \subset \mathcal{V}$, the leaves of the tree. The covariance among the observed variables is simply the principal submatrix $\Sigma_{\mathcal{L}\mathcal{L}}$ of Σ indexed by $\mathcal{L}$. Then the marginal covariance has a similar decomposition to the full covariance as

$$\Sigma_{\mathcal{L}\mathcal{L}} = \sum_{v\in\mathcal{V}} \Phi_{\mathcal{L}v} Q_v \Phi_{\mathcal{L}v}^T. \tag{6.3.1}$$

Before discussing the key properties of this decomposition, we consider an example.

Example 6.3.2. Continuing from Example 6.2.3, we now consider the submatrix of Σ that is indexed by the leaves of the tree on the left in Figure 6.3, i.e. the vertices

$w_1, w_2, \ldots, w_5$. This submatrix $\Sigma_{\mathcal{LL}}$ has the following factorization

$$\begin{bmatrix} A_{w_1}A_{v_1} & A_{w_1} & 0 & I & & & & \\ A_{w_2}A_{v_1} & A_{w_2} & 0 & 0 & I & & & \\ A_{w_3}A_{v_2} & 0 & A_{w_3} & 0 & 0 & I & & \\ A_{w_4}A_{v_2} & 0 & A_{w_4} & 0 & 0 & 0 & I & \\ A_{w_5} & 0 & 0 & 0 & 0 & 0 & 0 & I \end{bmatrix} \begin{bmatrix} Q_r & & & & & & & \\ & Q_{v_1} & & & & & & \\ & & Q_{v_2} & & & & & \\ & & & Q_{w_1} & & & & \\ & & & & Q_{w_2} & & & \\ & & & & & Q_{w_3} & & \\ & & & & & & Q_{w_4} & \\ & & & & & & & Q_{w_5} \end{bmatrix} \times$$

$$\begin{bmatrix} A_{w_1}A_{v_1} & A_{w_1} & 0 & I & & & & \\ A_{w_2}A_{v_1} & A_{w_2} & 0 & 0 & I & & & \\ A_{w_3}A_{v_2} & 0 & A_{w_3} & 0 & 0 & I & & \\ A_{w_4}A_{v_2} & 0 & A_{w_4} & 0 & 0 & 0 & I & \\ A_{w_5} & 0 & 0 & 0 & 0 & 0 & 0 & I \end{bmatrix}^T$$

(where again the missing entries are zeros). We obtain the corresponding matrix decomposition (6.3.1) by expanding this as a sum of block outer products. Such a decomposition takes the explicit form

$$\Sigma_{\mathcal{LL}} = \begin{bmatrix} A_{w_1}A_{v_1} \\ A_{w_2}A_{v_1} \\ A_{w_3}A_{v_2} \\ A_{w_4}A_{v_2} \\ A_{w_5} \end{bmatrix} Q_r \begin{bmatrix} A_{w_1}A_{v_1} \\ A_{w_2}A_{v_1} \\ A_{w_3}A_{v_2} \\ A_{w_4}A_{v_2} \\ A_{w_5} \end{bmatrix}^T + \begin{bmatrix} A_{w_1} \\ A_{w_2} \\ 0 \\ 0 \\ 0 \end{bmatrix} Q_{v_1} \begin{bmatrix} A_{w_1} \\ A_{w_2} \\ 0 \\ 0 \\ 0 \end{bmatrix}^T + \begin{bmatrix} 0 \\ 0 \\ A_{w_3} \\ A_{w_4} \\ 0 \end{bmatrix} Q_{v_2} \begin{bmatrix} 0 \\ 0 \\ A_{w_3} \\ A_{w_4} \\ 0 \end{bmatrix}^T +$$

$$\begin{bmatrix} I \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} Q_{w_1} \begin{bmatrix} I \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}^T + \begin{bmatrix} 0 \\ I \\ 0 \\ 0 \\ 0 \end{bmatrix} Q_{w_2} \begin{bmatrix} 0 \\ I \\ 0 \\ 0 \\ 0 \end{bmatrix}^T + \begin{bmatrix} 0 \\ 0 \\ I \\ 0 \\ 0 \end{bmatrix} Q_{w_3} \begin{bmatrix} 0 \\ 0 \\ I \\ 0 \\ 0 \end{bmatrix}^T + \begin{bmatrix} 0 \\ 0 \\ 0 \\ I \\ 0 \end{bmatrix} Q_{w_4} \begin{bmatrix} 0 \\ 0 \\ 0 \\ I \\ 0 \end{bmatrix}^T + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ I \end{bmatrix} Q_{w_5} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ I \end{bmatrix}^T.$$

We now describe the important features of the decomposition (6.3.1).

Positivity

Each term in the decomposition (6.3.1) is of the form $\Phi_{\mathcal{L}v}Q_v\Phi_{\mathcal{L}v}^T$ and so is positive semidefinite.

Support

The column space of a term $\Phi_{\mathcal{L}v}Q_v\Phi_{\mathcal{L}v}^T$ in the decomposition (6.3.1) has support contained in the coordinate subspace corresponding to $\mathcal{D}(v) \cap \mathcal{L}$. This is because if $\ell \in \mathcal{L}$ then $\Phi_{\ell v} \neq 0$ if and only if ℓ is also a descendant of v. In Example 6.3.2 we see that the third term in the decomposition, i.e. $\Phi_{\mathcal{L}v_2}Q_{v_2}\Phi_{\mathcal{L}v_2}^T$, has column space contained in the coordinate subspace corresponding to the leaves w_3 and w_4. From Figure 6.3 these are the leaves that are also descendants of v_2, i.e. $\mathcal{D}(v_2) \cap \mathcal{L} = \{w_3, w_4\}$.

The collection of sets $\mathcal{D}(v) \cap \mathcal{L}$ for $v \in \mathcal{V}$ forms the rooted tree (in the sense of Definition 6.2.1) with respect to which the Gaussian model is parameterized. To see this, suppose we construct the Hasse diagram (see Section 6.2.1) of this collection of sets ordered by inclusion. Doing so we obtain the original tree. For instance, suppose we take the tree shown on the left in Figure 6.3 and form the collection $(\mathcal{D}(v) \cap \mathcal{L})_{v\in\mathcal{V}}$ of subsets of the leaves. The corresponding Hasse diagram of the sets is shown on the right in Figure 6.3 and is clearly is the same as the original tree structure on the left.

Subspaces corresponding to the latent variables

We can change basis in the state space of a latent variable (and appropriately change parameterization) without affecting the covariance at the leaves. Indeed any information we can hope to extract about a latent variable x_v from observations of the leaves must be invariant under such coordinate changes in the latent state spaces. One such quantity is the matrix $\Phi_{\mathcal{L}v}\Sigma_{vv}\Phi_{\mathcal{L}v}^T$. We have seen in Section 6.3.1 that the dimension of the column space of this matrix reflects the dimension of the corresponding latent variable in a minimal realization. This is intuitive because the column space of $\Phi_{\mathcal{L}v}\Sigma_{vv}\Phi_{\mathcal{L}v}^T$ is the subspace of the observation space that can be affected by the random variable x_v.

Since the column spaces of the $\Phi_{\mathcal{L}v}\Sigma_{vv}\Phi_{\mathcal{L}v}^T$ for $v \in \mathcal{V}$ capture important structural information about the latent variables, and play an important role in the rest of the chapter, we introduce specific notation and terminology for them.

Definition 6.3.3. Suppose $(x_v)_{v\in\mathcal{V}}$ is Gaussian tree model with covariance Σ, and suppose we observe the leaf-indexed variables. Define the *subspace corresponding to* $v \in \mathcal{V}$ to be $U_v := \mathrm{col}(\Phi_{\mathcal{L}v}\Sigma_{vv}\Phi_{\mathcal{L}v}^T)$.

We next explain how these subspaces arise in the decomposition (6.3.1). In particular we relate the subspaces U_v with the column spaces of certain matrices in the decomposition (6.3.1).

First, suppose we are in the situation where all of the Q_v are positive definite. Then the column space of $\Phi_{\mathcal{L}v}\Sigma_{vv}\Phi_{\mathcal{L}v}$ is the same as the column space of $\Phi_{\mathcal{L}v}Q_v\Phi_{\mathcal{L}v}^T$. Hence the subspaces U_v appear directly as the column spaces of the terms in the decomposition (6.3.1).

Things are more subtle in the general case where some of the Q_v may be singular. In this case the column space of the term $\Phi_{\mathcal{L}v}Q_v\Phi_{\mathcal{L}v}^T$ of the decomposition (6.3.1) is a strict subspace of U_v. As such, we cannot recover U_v from this single term of the decomposition. Nevertheless we can write the matrix $\Phi_{\mathcal{L}v}\Sigma_{vv}\Phi_{\mathcal{L}v}^T$ as a linear combination of submatrices of the terms in the decomposition (6.3.1). In the course of the proof of Theorem 6.3.9 (in Section 6.7.1) we show that

$$\Phi_{\mathcal{L}v}\Sigma_{vv}\Phi_{\mathcal{L}v}^T = \sum_{u\succeq v} P_{\mathcal{L}\cap\mathcal{D}(v)}(\Phi_{\mathcal{L}u}Q_u\Phi_{\mathcal{L}u}^T)P_{\mathcal{L}\cap\mathcal{D}(v)}.$$

This is a sum of (appropriately zero-padded) submatrices of terms in the decomposition (6.3.1). In particular its entries are *linear* in the terms of the decomposition. This shows that, in general, it should be possible to extract the subspaces U_v given only the terms of the decomposition (6.3.1).

Definition of a latent tree covariance decomposition

Thus far in this section we have seen that the marginal covariance among the leaves of a Gaussian tree model admits a particular decomposition (6.3.1). We have seen that the terms of the decomposition are positive semidefinite and the supports of their column spaces form a tree. Definition 6.3.4, to follow, describes those matrices that admit a decomposition into positive semidefinite terms with column spaces the supports of which form a tree.

Furthermore, we have seen that the decomposition (6.3.1) contains information about the structure of the tree and the subspaces U_v which encode important structural information about the latent variables such as their minimal dimensions. Definition 6.3.4, to follow, also describes how these structures appear in the more abstract matrix decomposition setting.

The following definition plays a central role in the chapter, as it gives an alternative description (see Theorem 6.3.9 to follow) of the possible matrices that can appear as the marginal covariance among the leaf-indexed variables of a Gaussian latent tree model.

Definition 6.3.4. Let $\mathcal{V}$ be a finite set and $X \in \mathcal{S}_+^n$ be a positive semidefinite matrix. A decomposition $X = \sum_{v\in\mathcal{V}} X_v$ is a *latent tree covariance decomposition (LTCD)* of X if

- $X_v \succeq 0$ for all $v \in \mathcal{V}$

- there is a collection $(S_v)_{v\in\mathcal{V}}$ of distinct subsets of $[n]$ such that
  - the $(S_v)_{v\in\mathcal{V}}$ form a tree (see Definition 6.2.1)
  - $S_v \supseteq \mathrm{supp}(\mathrm{col}(X_v))$ for all $v \in \mathcal{V}$.

Associated with an LTCD we define

- the *subspaces* of the decomposition by

$$U_v := \mathrm{col}\left(\textstyle\sum_{u\succeq v} P_{S_v} X_u P_{S_v}\right) \subseteq \mathbb{R}^{S_v} \quad \text{for all } v \in \mathcal{V}; \tag{6.3.2}$$

- the *dimensions* of the decomposition by $n_v := \dim(U_v)$ for all $v \in \mathcal{V}$; and
- the *structure* of the decomposition to be the collection $(S_v)_{v\in\mathcal{V}}$ of subsets of $[n]$.

The most important part of the definition is the notion of the dimension of an LTCD. Indeed any positive semidefinite matrix has an LTCD (see Examples 6.3.5). On the other hand, if a matrix has an LTCD where all of the dimensions are small, then this is a very special structure that can be exploited algorithmically.

Examples of LTCDs

Since Definition 6.3.4 is rather complicated, we illustrate it with some examples.

Example 6.3.5 (Trivial LTCDs)**.** Every $n \times n$ positive semidefinite matrix X has an LTCD with one term $X = X_r$. In this case the structure of the decomposition is $S_r = [n]$ (which forms a rooted tree consisting of a single vertex), the subspace of the decomposition is $U_r = \mathrm{col}(X)$, and the dimension is $n_r = \mathrm{rank}(X)$.

In Examples 6.3.6, 6.3.7, and 6.3.8 to follow we work with explicit LTCDs of 5×5 matrices. It should not be obvious how to systematically come up with decompositions such as these. The problem of constructing LTCDs given just the sum of the terms and the structure of the tree is the subject of Section 6.4 to follow.

Example 6.3.6 (An LTCD of a 5×5 matrix)**.** Consider the following decomposition

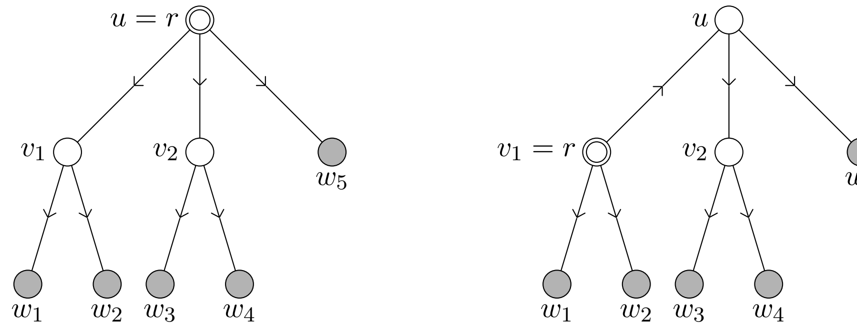


Figure 6.4: The rooted trees on the left and on the right are the same *except* for the choice of root (which is indicated with a double circle in each case). On the left, the root is the vertex labeled $u = r$. On the right, the root is the vertex labeled $v_1 = r$. Note that this different choice of root reverses the orientation of the edge between u and v_1. Example 6.3.7 shows how these different choices of root give rise to different LTCDs.

of a 5×5 positive definite matrix (where omitted entries are zero).

$$\begin{bmatrix} 6 & -5 & 2 & 2 & -6 \\ -5 & 7 & -2 & -2 & 6 \\ 2 & -2 & 6 & 3 & -3 \\ 2 & -2 & 3 & 5 & -3 \\ -6 & 6 & -3 & -3 & 10 \end{bmatrix} = \begin{bmatrix} 1 & & & & \\ & 2 & & & \\ & & 3 & & \\ & & & 2 & \\ & & & & 1 \end{bmatrix} + \begin{bmatrix} 1 & -1 & & & \\ -1 & 1 & & & \\ & & 2 & 2 & \\ & & 2 & 2 & \\ & & & & \end{bmatrix} + \begin{bmatrix} 4 & -4 & 2 & 2 & -6 \\ -4 & 4 & -2 & -2 & 6 \\ 2 & -2 & 1 & 1 & -3 \\ 2 & -2 & 1 & 1 & -3 \\ -6 & 6 & -3 & -3 & 9 \end{bmatrix}.$$

If we treat each of the diagonal blocks in the three matrices on the right hand side as separate terms, then we can think of this as a decomposition into eight terms (five from the diagonal matrix, two from the block diagonal matrix in the middle, and one from the full matrix). Each is positive semidefinite. Furthermore, if we list the supports of column spaces of these terms we obtain the following list of subsets of $\{1, 2, 3, 4, 5\}$:

$$\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{1, 2\}, \{3, 4\}, \{1, 2, 3, 4, 5\}.$$

It is straightforward to check that these sets form the rooted tree shown on the left in Figure 6.4. Hence we have identified that this is a valid LTCD. Suppose we index the terms in the decomposition in the same way that we labeled the vertices of the tree on the left in Figure 6.4. Hence the *structure* of the LTCD is $S_{w_1} = \{1\}$, $S_{w_2} = \{2\}$, $S_{w_3} = \{3\}$, $S_{w_4} = \{4\}$, $S_{w_5} = \{5\}$, $S_{v_1} = \{1,2\}$, $S_{v_2} = \{3,4\}$, and $S_u = \{1,2,3,4,5\}$.

We now describe the subspaces associated with two of the terms in the decomposition using (6.3.2). Indeed we have that the column space U_u of the term in the decomposition corresponding to the root is

$$U_u = \mathrm{col}(X_u) = \mathrm{col}\left(\begin{bmatrix} 4 & -4 & 2 & 2 & -6 \\ -4 & 4 & -2 & -2 & 6 \\ 2 & -2 & 1 & 1 & -3 \\ 2 & -2 & 1 & 1 & -3 \\ -6 & 6 & -3 & -3 & 9 \end{bmatrix}\right) = \mathrm{span}\left\{\begin{bmatrix} 2 \\ -2 \\ 1 \\ 1 \\ -3 \end{bmatrix}\right\}.$$

By (6.3.2), the subspace U_{v_1} is the column space of the matrix

$$\begin{aligned} X_{v_1} + P_{S_{v_1}} X_u P_{S_{v_1}} &= \begin{bmatrix} 1 & -1 & & & \\ -1 & 1 & & & \\ & & & & \\ & & & & \\ & & & & \end{bmatrix} + P_{\{1,2\}} \begin{bmatrix} 4 & -4 & 2 & 2 & -6 \\ -4 & 4 & -2 & -2 & 6 \\ 2 & -2 & 1 & 1 & -3 \\ 2 & -2 & 1 & 1 & -3 \\ -6 & 6 & -3 & -3 & 9 \end{bmatrix} P_{\{1,2\}} \\ &= \begin{bmatrix} 5 & -5 & & & \\ -5 & 5 & & & \\ & & & & \\ & & & & \\ & & & & \end{bmatrix} \end{aligned}$$

and so $U_{v_1} = \mathrm{span}\left\{\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \end{bmatrix}^T\right\}$. Both U_u and U_{v_1} are one-dimensional subspaces of $\mathbb{R}^5$, hence the corresponding dimensions are $n_u = \dim(U_u) = 1$ and $n_{v_1} = \dim(U_{v_1}) = 1$.

The next two examples deal with different issues related to non-uniqueness of LTCDs. The first illustrates that the same matrix typically has different LTCDs corresponding to different choices of root.

Example 6.3.7 (Multiple LTCDs based on choices of root)**.** Consider the two rooted trees shown in Figure 6.4. The LTCD in Example 6.3.6 has structure that forms the rooted tree on the left in Figure 6.4. It also follows from Theorem 6.3.9 to follow, that the same matrix has a different LTCD with structure that forms the rooted tree on the right in Figure 6.4. A corresponding alternative decomposition is

$$\begin{bmatrix} 6 & -5 & 2 & 2 & -6 \\ -5 & 7 & -2 & -2 & 6 \\ 2 & -2 & 6 & 3 & -3 \\ 2 & -2 & 3 & 5 & -3 \\ -6 & 6 & -3 & -3 & 10 \end{bmatrix} = \begin{bmatrix} 1 & & & & \\ & 2 & & & \\ & & 3 & & \\ & & & 2 & \\ & & & & 1 \end{bmatrix} + \frac{1}{5}\begin{bmatrix} & & & & \\ & & & & \\ & & 1 & 1 & -3 \\ & & 1 & 1 & -3 \\ & & -3 & -3 & 9 \end{bmatrix} +$$
$$\begin{bmatrix} & & & & \\ & & & & \\ & & 2 & 2 & \\ & & 2 & 2 & \\ & & & & \end{bmatrix} + \frac{1}{5}\begin{bmatrix} 25 & -25 & 10 & 10 & -30 \\ -25 & 25 & -10 & -10 & 30 \\ 10 & -10 & 4 & 4 & -12 \\ 10 & -10 & 4 & 4 & -12 \\ -30 & 30 & -12 & -12 & 36 \end{bmatrix}.$$

Example 6.3.8 (Multiple LTCDs based on grouping terms)**.** Another way a matrix can have multiple LTCDs is by grouping terms in the sum in such a way that the resulting decomposition is still a valid LTCD. For example, by grouping all the terms corresponding to non-leaf vertices in Example 6.3.6 we obtain another valid LTCD

$$\begin{bmatrix} 6 & -5 & 2 & 2 & -6 \\ -5 & 7 & -2 & -2 & 6 \\ 2 & -2 & 6 & 3 & -3 \\ 2 & -2 & 3 & 5 & -3 \\ -6 & 6 & -3 & -3 & 10 \end{bmatrix} = \begin{bmatrix} 1 & & & & \\ & 2 & & & \\ & & 3 & & \\ & & & 2 & \\ & & & & 1 \end{bmatrix} + \begin{bmatrix} 5 & -5 & 2 & 2 & -6 \\ -5 & 5 & -2 & -2 & 6 \\ 2 & -2 & 1 & 1 & -3 \\ 2 & -2 & 1 & 1 & -3 \\ -6 & 6 & -3 & -3 & 9 \end{bmatrix}.$$

This LTCD has structure $S_{w_1} = \{1\}, S_{w_2} = \{2\}, S_{w_3} = \{3\}, S_{w_4} = \{4\}, S_{w_5} = \{5\}, S_r = \{1, 2, 3, 4, 5\}$ which forms a star-shaped tree with five leaves and the center being the root r. In this case

$$U_r = \operatorname{col}\left(\begin{bmatrix} 5 & -5 & 2 & 2 & -6 \\ -5 & 5 & -2 & -2 & 6 \\ 2 & -2 & 1 & 1 & -3 \\ 2 & -2 & 1 & 1 & -3 \\ -6 & 6 & -3 & -3 & 9 \end{bmatrix}\right) = \operatorname{span}\left\{\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \\ 1 \\ 1 \\ -3 \end{bmatrix}\right\}$$

is now two-dimensional. This idea of constructing LTCDs by grouping terms is important in Section 6.5 as it gives a way to incorporate partial knowledge of the tree structure into a decomposition.

6.3.3 LTCDs and Gaussian latent tree models

Our main interest in LTCDs is that they give a useful alternative characterization of the possible covariances among the leaf-indexed variables of a Gaussian latent tree model. This characterization shows both how to construct an LTCD from a Gaussian tree model but also, and more importantly, how to take an LTCD with given dimensions and structure and construct from it a Gaussian latent tree model with tree and state dimensions related to the structure and dimensions of the LTCD. We provide a proof of this result in Section 6.7.1.

Theorem 6.3.9. *If X is positive definite and has an LTCD with subspaces $(U_v)_{v\in\mathcal{V}}$, dimensions $(n_v)_{v\in\mathcal{V}}$, and structure $(S_v)_{v\in\mathcal{V}}$ (that forms a tree $\mathcal{T}$) then X can be realized as the covariance among the leaf-indexed variables of a Gaussian tree model (with state dimensions $(n_v)_{v\in\mathcal{V}}$ and tree $\mathcal{T}$) parameterized by:*

$$A_v = \pi_{U_v}\pi^*_{U_{\mathcal{P}(v)}} \qquad \text{for } v \in \mathcal{V}\setminus r \tag{6.3.3}$$

$$Q_v = \pi_{U_v}X_v\pi^*_{U_v} \qquad \text{for } v \in \mathcal{V}. \tag{6.3.4}$$

Conversely, if X can be realized as the covariance among the leaf-indexed variables of a Gaussian tree model with state dimensions $(n_v)_{v\in\mathcal{V}}$ and tree $\mathcal{T}$ then X has an LTCD with dimensions $(m_v)_{v\in\mathcal{V}}$ (which satisfy $m_v \leq n_v$ for all $v \in \mathcal{V}$) and structure $(S_v)_{v\in\mathcal{V}}$ (which forms the tree $\mathcal{T}$).

The inequality $m_v \leq n_v$ for all $v \in \mathcal{V}$ in the statement of Theorem 6.3.9 is related to the discussion of minimal models in Section 6.3.1. Indeed suppose we take a non-minimal parameterization of a latent tree model with given leaf-covariance, compute the corresponding LTCD, and then use (6.3.3) and (6.3.4) to construct a new explicit parameterization that realizes the same leaf covariance. The parameterization we have constructed is in fact minimal, and hence may have smaller state dimension m_v than the state dimensions n_v of the parameterization with which we started.

LTCDs for singular and non-singular models

Given an LTCD of a positive definite matrix X, the corresponding Gaussian tree model constructed by Theorem 6.3.9 may be a singular Gaussian model. This is useful because there are situations where we want to recover singular models (see Section 6.3.1). In this section we briefly discuss the difference between singular and non-singular models from the point of view of LTCDs.

Definition 6.3.10. An LTCD of a positive definite matrix X is *singular* if the joint covariance of the Gaussian tree model defined by (6.3.3) and (6.3.4) is singular. If an LTCD is not singular we call it *non-singular*.

Suppose that the matrices $(A_v)_{v\in\mathcal{V}\setminus\{r\}}$ and $(Q_v)_{v\in\mathcal{V}}$ parameterize a Gaussian tree model. This model is non-singular if and only if all of the Q_v are positive definite. In Lemma 6.3.11 to follow, we reformulate this simple characterization into one about the relationship between the subspaces U_v and the columns spaces of the terms in an LTCD. To provide some intuition for Lemma 6.3.11 we first describe its main point working with an explicit parameterization.

The subspaces $(U_v)_{v\in\mathcal{V}}$ are the column spaces of $\Phi_{\mathcal{L}v}\Sigma_{vv}\Phi_{\mathcal{L}v}^T$ (see Definition 6.3.3), whereas the terms in the LTCD can be written in our explicit parameterization as the matrices $\Phi_{\mathcal{L}v}Q_v\Phi_{\mathcal{L}}^T$. Since $\Sigma_{vv} = Q_v + A_v\Sigma_{\mathcal{P}(v)\,\mathcal{P}(v)}A_v^T$, if all the Q_v are positive definite, then so are all of the Σ_{vv}. It follows that the column spaces of the matrices

$$\Phi_{\mathcal{L}v}\Sigma_{vv}\Phi_{\mathcal{L}v}^T \quad \text{and} \quad \Phi_{\mathcal{L}v}Q_v\Phi_{\mathcal{L}v}^T$$

are the same for all $v\in\mathcal{V}$. The column space of the term on the left is the subspace U_v. In the language of LTCDs, the column space of the term on the right is the column space of the term X_v in the matrix decomposition corresponding to the latent variable v.

To summarize, we have seen that if the model is non-singular then

$$U_v = \operatorname{col}(\Phi_{\mathcal{L}v}Q_v\Phi_{\mathcal{L}v}^T) \quad \text{for all } v\in\mathcal{V}.$$

In the general (possibly non-singular) case, it is only necessary that $U_v \supseteq \operatorname{col}(\Phi_{\mathcal{L}v}Q_v\Phi_{\mathcal{L}v}^T)$.

We now formally state the characterization of those LTCDs that correspond to non-singular models. We provide a proof in Section 6.7.1. We remark that we make no use of the second item in Lemma 6.3.11 in this chapter. It is included only to aid comparison with the work in [112] and [114].

Lemma 6.3.11. *Suppose $X \succ 0$ has an LTCD $X = \sum_{v\in\mathcal{V}} X_v$ with subspaces $(U_v)_{v\in\mathcal{V}}$ and structure $(S_v)_{v\in\mathcal{V}}$. Then the following are equivalent:*

1. *the LTCD $X = \sum_{v\in\mathcal{V}} X_v$ is non-singular;*
2. $\operatorname{col}(X_v) \supseteq \operatorname{col}(P_{S_v}X_{\mathcal{P}(v)}P_{S_v})$ *for all $v\in\mathcal{V}\setminus r$;*
3. $\operatorname{col}(X_v) = U_v := \operatorname{col}\left(\sum_{u\succeq v} P_{S_v}X_uP_{S_v}\right)$ *for all $v\in\mathcal{V}$.*

We conclude this section with an example of a singular LTCD. By explicit computations we show that the third item in Lemma 6.3.11 fails for this example. We also apply the procedure in Theorem 6.3.9 for constructing a parameterization of a Gaussian latent tree model from an LTCD and see that it produces a singular model.

Example 6.3.12 (A singular LTCD)**.** This example arises by slightly modifying Example 6.3.6 (which was a non-singular LTCD) so that the column spaces no longer satisfy the appropriate constraints for a non-singular model. Our focus is on explicitly seeing why the following LTCD corresponds to a singular Gaussian latent tree model:

$$\begin{bmatrix} 3 & -3 & 1 & 1 & -3 \\ -3 & 7 & -2 & -2 & 6 \\ 1 & -2 & 6 & 3 & -3 \\ 1 & -2 & 3 & 5 & -3 \\ -3 & 6 & -3 & -3 & 10 \end{bmatrix} = \begin{bmatrix} 1 & & & & \\ & 2 & & & \\ & & 3 & & \\ & & & 2 & \\ & & & & 1 \end{bmatrix} + \begin{bmatrix} 1 & -1 & & & \\ -1 & 1 & & & \\ & & 2 & 2 & \\ & & 2 & 2 & \\ & & & & \end{bmatrix} + \begin{bmatrix} 1 & -2 & 1 & 1 & -3 \\ -2 & 4 & -2 & -2 & 6 \\ 1 & -2 & 1 & 1 & -3 \\ 1 & -2 & 1 & 1 & -3 \\ -3 & 6 & -3 & -3 & 9 \end{bmatrix}.$$

Here the supports of the terms in the decomposition are the same as in Example 6.3.6. As such, they form the rooted tree shown on the left in Figure 6.4, and we label the terms according the vertices of that tree. In this case we have that

$$X_r = X_u = \begin{bmatrix} 1 & -2 & 1 & 1 & -3 \\ -2 & 4 & -2 & -2 & 6 \\ 1 & -2 & 1 & 1 & -3 \\ 1 & -2 & 1 & 1 & -3 \\ -3 & 6 & -3 & -3 & 9 \end{bmatrix} \quad \text{and that} \quad X_{v_1} = \begin{bmatrix} 1 & -1 & & & \\ -1 & 1 & & & \\ & & & & \\ & & & & \\ & & & & \end{bmatrix}.$$

From this we can see that $\operatorname{col}(X_{v_1}) = \operatorname{span}\{\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \end{bmatrix}^T\}$. We now compute

$$U_{v_1} = \operatorname{col}\left(X_{v_1} + P_{S_{v_1}} X_r P_{S_{v_1}}\right) = \operatorname{col}\left(\begin{bmatrix} 1 & -1 & & & \\ -1 & 1 & & & \\ & & & & \\ & & & & \\ & & & & \end{bmatrix} + \begin{bmatrix} 1 & -2 & & & \\ -2 & 4 & & & \\ & & & & \\ & & & & \\ & & & & \end{bmatrix}\right)$$

$$= \operatorname{span}\left\{\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}\right\}.$$

We can see that $\operatorname{col}(X_{v_1}) \neq U_{v_1}$ so Lemma 6.3.11 tells us that it this must correspond to a singular model. Let us see this explicitly by computing part of a parameterization of this model via (6.3.4). Indeed we can compute the orthogonal projector onto the two-dimensional subspace U_{v_1} to obtain

$$\pi_{U_{v_1}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

Then applying (6.3.4) we see that

$$Q_{v_1} = \pi_{U_{v_1}} X_{v_1} \pi^*_{U_{v_1}} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

which is clearly singular.

■ 6.4 Finding LTCDs given the tree structure

Suppose we are given a matrix $\hat{X}$ and a tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ and our aim is to approximately realize $\hat{X}$ as the leaf covariance of a Gaussian model that is Markov with respect to $\mathcal{T}$ and has small state dimensions. From Theorem 6.3.9 we know that we can, instead, search for a matrix X that is close to $\hat{X}$ and that has an LTCD with structure $(S_v)_{v\in\mathcal{V}}$ that forms the tree $\mathcal{T}$ and has small dimensions. With this shift of perspective from Gaussian tree models to more abstract matrix decompositions, from now on we change notation to use X for the matrix we are decomposing where we would have used $\Sigma_{\mathcal{L}\mathcal{L}}$ in Sections 6.2 and 6.3.

A natural convex optimization-based approach to performing approximate LTCDs is to design a convex regularizer $X \mapsto \mathcal{R}(X; (S_v)_{v\in\mathcal{V}})$ that induces X to have such a decomposition. To approximate $\hat{X}$, we could then solve

$$\min_X \gamma\, \mathcal{R}(X; (S_v)_{v\in\mathcal{V}}) + \mathcal{L}(X, \hat{X}) \tag{6.4.1}$$

for some convex loss function $\mathcal{L}$ that penalizes error between X and $\hat{X}$ and some choice of regularization parameter $\gamma > 0$. (We use $\mathcal{L}(X, \hat{X}) = \frac{1}{2}\|X - \hat{X}\|_F^2$ throughout for concreteness.) Of course, we want $\mathcal{R}(X; (S_v)_{v\in\mathcal{V}})$ to do much more than simply regularize. We would also like it to *produce* an LTCD of X with structure $(S_v)_{v\in\mathcal{V}}$ having small dimensions. In this section we describe and analyze a convex optimization problem that aims to produce such LTCDs of X with a *given* structure $(S_v)_{v\in\mathcal{V}}$ and small dimensions. The optimal value function then has an interpretation as a convex regularizer $\mathcal{R}(X; (S_v)_{v\in\mathcal{V}})$.

To find an LTCD of X with dimensions $(n_v)_{v\in\mathcal{V}}$ that are all small it is natural to try to minimize a non-negative combination of the dimensions

$$n_v = \operatorname{rank}\left(\textstyle\sum_{u\succeq v} P_{S_v} X_u P_{S_v}\right)$$

where the expression on the right is from (6.3.2). As such, one would expect the following rank minimization problem to find an LTCD of X with dimensions that are small, if such a decomposition exists:

$$\min_{(Z_v)_{v\in\mathcal{V}}} \sum_{v\in\mathcal{V}} \kappa_v \operatorname{rank}\left(\sum_{u\succeq v}[\pi_{S_u}^* Z_u \pi_{S_u}]_{S_v S_v}\right) \quad \text{s.t.} \quad \begin{cases} Z_v \in \mathcal{S}_+^{S_v} \quad \text{for all } v\in\mathcal{V} \\ \displaystyle\sum_{v\in\mathcal{V}} \pi_{S_v}^* Z_v \pi_{S_v} = X. \end{cases} \tag{6.4.2}$$

Note that the constraint set in (6.4.2) is a convex set. Furthermore, since we are fixing the support of the variables in the LTCD we parameterize them as $\pi_{S_v}^* Z_v \pi_{S_v}$ where $Z_v \in \mathcal{S}_+^{S_v}$ rather than using $n\times n$ variables X_v that are constrained to have many zeros.

Since the only non-convex part of (6.4.2) is the objective, it makes sense to replace the rather complicated function $(A_1, A_2, \ldots, A_k) \mapsto \operatorname{rank}(A_1 + A_2 + \cdots + A_k)$ with a convex function $f(A_1, A_2, \ldots, A_k)$ designed to penalize collections $(A_1, A_2, \ldots, A_k)$ of matrices where $A_1 + A_2 + \cdots + A_k$ has high rank. Fixing a choice of convex surrogate f for rank, we obtain a convex optimization problem

$$\min_{(Z_v)_{v\in\mathcal{V}}} \sum_{v\in\mathcal{V}} \kappa_v\, f\left(\left([\pi_{S_u}^* Z_u \pi_{S_u}]_{S_v S_v}\right)_{u\succeq v}\right) \quad \text{s.t.} \quad \begin{cases} Z_v \in \mathcal{S}_+^{S_v} \quad \text{for all } v\in\mathcal{V} \\ \displaystyle\sum_{v\in\mathcal{V}} \pi_{S_v}^* Z_v \pi_{S_v} = X \end{cases} \tag{6.4.3}$$

the solutions of which are always valid LTCDs with structure $(S_v)_{v\in\mathcal{V}}$.

The remainder of this section is structured as follows. In Section 6.4.1 we use the trace as a surrogate for $\operatorname{rank}(\cdot)$ in (6.4.3) and discuss basic properties of the resulting semidefinite optimization problem. In Section 6.4.2 we provide conditions on the subspaces of an LTCD of X under which the solution of the semidefinite optimization problem always recovers the underlying LTCD.

■ 6.4.1 Minimum trace covariance decomposition

It is well-known that the convex envelope of $\operatorname{rank}(A)$ restricted to $\{X : 0 \preceq X \preceq I\}$ is $\operatorname{tr}(A)$ [45]. Based on this justification, in this section we use

$$f(A_1, A_2, \ldots, A_k) = \operatorname{tr}(A_1 + A_2 + \cdots + A_k)$$

in (6.4.3) as a surrogate for $\text{rank}(A_1 + A_2 + \cdots + A_k)$. Doing so yields a semidefinite optimization problem that we call *minimum trace covariance decomposition (MTCD)*:

$$\min_{(Z_v)_{v\in\mathcal{V}}} \sum_{v\in\mathcal{V}} \kappa_v \,\text{tr}\left(\sum_{u\succeq v}[\pi_{S_u}^* Z_u \pi_{S_u}]_{S_v S_v}\right) \quad \text{s.t.} \quad \begin{cases} Z_v \in \mathcal{S}_+^{S_v} \quad \text{for all } v\in\mathcal{V} \\ \sum_{v\in\mathcal{V}} \pi_{S_v}^* Z_v \pi_{S_v} = X. \end{cases} \tag{6.4.4}$$

By restricting our choice of the weights κ_v and using the linearity of the trace, we can simplify the form of the objective in (6.4.4). Our aim is to rewrite it as $\sum_{v\in\mathcal{V}} \lambda_v \text{tr}(Z_v)$, i.e. as a linear combination of the quantities $\text{tr}(Z_u)$ for $u \in \mathcal{V}$. This is only possible if we put certain restrictions on the κ_v and corresponding restrictions on the λ_v.

Definition 6.4.1. A collection $(\lambda_v)_{v\in\mathcal{V}}$ of scalars indexed by the vertices of a rooted tree $\mathcal{T}$ is *order-preserving* if $u \succ v$ implies $\lambda_u > \lambda_v$. A collection $(\lambda_v)_{v\in\mathcal{V}}$ is *constant on children* if $\lambda_v = \lambda_w$ whenever v and w have a common parent u. In this case we write $\lambda_{\mathcal{C}(u)}$ for this common value and define $\lambda_{\mathcal{C}(w)} = 0$ whenever $w \in \mathcal{L}$.

The restriction that $\lambda_u > \lambda_v$ whenever $u \succ v$ can be thought of as penalizing the rank of terms in the decomposition more as we go towards the root. One could justify this from a model-complexity perspective since terms indexed by vertices nearer the root have larger support, so increasing the rank of those terms increases the number of parameters more than increasing the rank of terms with small support.

We now describe the way in which the objective of (6.4.4) simplifies. We provide a proof in Section 6.7.2.

Lemma 6.4.2. *Suppose $(\lambda_v)_{v\in\mathcal{V}}$ is a collection of positive scalars that is order-preserving and constant on children. If $\kappa_v := \lambda_v - \lambda_{\mathcal{C}(v)}$ for all $v \in \mathcal{V}$ then the κ_v are all positive and*

$$\sum_{v\in\mathcal{V}} \kappa_v \text{tr}\left(\sum_{u\succeq v}[\pi_{S_u}^* Z_u \pi_{S_u}]_{S_v S_v}\right) = \sum_{v\in\mathcal{V}} \lambda_v \text{tr}(Z_v).$$

MTCD reformulated As such, we assume that the $(\lambda_v)_{v\in\mathcal{V}}$ are positive, order-preserving, and constant on children and focus on the following reformulation of MTCD:

$$\mathcal{R}(X;(S_v)_{v\in\mathcal{V}}) := \min_{(Z_v)_{v\in\mathcal{V}}} \sum_{v\in\mathcal{V}} \lambda_v \text{tr}(Z_v) \quad \text{s.t.} \quad \begin{cases} Z_v \in \mathcal{S}_+^{S_v} \quad \text{for all } v\in\mathcal{V} \\ \sum_{v\in\mathcal{V}} \pi_{S_v}^* Z_v \pi_{S_v} = X. \end{cases} \tag{6.4.5}$$

Dual of MTCD The dual of MTCD (6.4.5) can be obtained, e.g., by applying conic duality (see (2.4.2) and (2.4.3) of Chapter 2). It is

$$\max_Y \langle X, Y\rangle \quad \text{s.t.} \quad Y_{S_v S_v} \preceq \lambda_v I \;\text{ for all } v \in \mathcal{V}. \tag{6.4.6}$$

Moreover since the dual is strictly feasible (take, e.g., $Y = -I$ since all the λ_v are positive) and its objective function is bounded above by $(\max_{v\in\mathcal{V}} \lambda_v)\,\mathrm{tr}(X)$ it follows from Theorem 2.4.2 in Chapter 2 that strong duality holds.

This means that the optimal value function $\mathcal{R}(X;(S_v)_{v\in\mathcal{V}})$ for (6.4.5) and the optimal value function of (6.4.6) are the same. The optimal value function of (6.4.6), as a function of X, is the support function of the constraint set, and so is convex (see (2.3.5) of Chapter 2). Hence $\mathcal{R}(X;(S_v)_{v\in\mathcal{V}})$ is a convex function of X. We can think of $X \mapsto \mathcal{R}(X;(S_v)_{v\in\mathcal{V}})$ as a convex regularizer that aims to induce its argument to have an LTCD with low dimensions and structure $(S_v)_{v\in\mathcal{V}}$.

6.4.2 Exact recovery

In this section we provide sufficient conditions on an underlying LTCD of a given matrix, with given structure, that ensure that minimum trace covariance decomposition (MTCD) recovers that underlying LTCD.

Definition 6.4.3. Fix an LTCD of $X = \sum_{v\in\mathcal{V}} X_v$ with a given structure $(S_v)_{v\in\mathcal{V}}$. We say that MTCD *recovers* the LTCD of X if $(Z_v)_{v\in\mathcal{V}}$ is the unique optimal solution to MTCD and

$$\pi_{S_v}^* Z_v \pi_{S_v} = X_v \quad \text{for all } v \in \mathcal{V}.$$

The optimality conditions of MTCD give us a way to certify that a given LTCD is indeed optimal by explicitly constructing a feasible point for the dual problem that satisfies additional complementarity conditions.

Proposition 6.4.4. *Suppose $X \in \mathcal{S}^n$ is positive definite and has an LTCD $X = \sum_{v\in\mathcal{V}} X_v$ with structure $(S_v)_{v\in\mathcal{V}}$. Suppose, further, that all the λ_v are positive and order-preserving. Then MTCD recovers the LTCD of X if and only if there exists a symmetric matrix Y such that for all $v \in \mathcal{V}$,*

$$\langle Yx, x\rangle \leq \lambda_v \|x\|^2 \quad \textit{for all } x \in \mathbb{R}^{S_v} \;\; \textit{and} \tag{6.4.7}$$

$$\langle Yx, x\rangle = \lambda_v \|x\|^2 \quad \textit{for all } x \in \mathrm{col}(X_v) \subseteq \mathbb{R}^{S_v}. \tag{6.4.8}$$

We give a proof in Section 6.7.2. Most of the argument involves establishing that MTCD has a unique solution. The required conditions on the matrix Y are just obtained from the standard optimality conditions for semidefinite optimization. We have written them in terms of the quadratic form $\langle Yx, x\rangle$ rather than in terms of the symmetric matrix Y only because this makes the notation simpler in subsequent parts of this section.

The star-shaped case

Suppose that the supports $(S_v)_{v\in\mathcal{V}}$ form a star-shaped tree, so that $(S_v)_{v\in\mathcal{L}}$ partition $[n]$ and $S_r = [n]$. We consider this case separately for two reasons. First, our results for the general case depend crucially on our understanding of the star-shaped case. Second, this case has been previously studied, for instance, in [113] and so explaining it separately makes it easier to establish connections to the terminology and results used in previous work.

If X has an LTCD with star-shaped structure $(S_v)_{v\in\mathcal{V}}$ and dimensions $(n_v)_{v\in\mathcal{V}}$ then it has a decomposition as

$$X = \sum_{v\in\mathcal{L}} X_v + X_r$$

where X_r has rank n_r and $\sum_{v\in\mathcal{L}} X_v$ is block diagonal with respect to the partition $(S_v)_{v\in\mathcal{L}}$ of $[n]$. As discussed in Section 6.1.2 this special case corresponds to the classical factor analysis model. The corresponding specialization of MTCD is known as (constrained) minimum trace factor analysis [122]. The discussion in this section relies on the results of [113] that studies such block diagonal and low-rank matrix decompositions and recovery properties of MTCD specialized to this case. The main results of that work are stated in terms of a property that relates the subspace $\mathrm{col}(X_r)$ and the partition $(S_v)_{v\in\mathcal{L}}$ of $[n]$.

Definition 6.4.5. A subspace $U \subseteq \mathbb{R}^n$ is *realizable* with respect to the partition $(S_v)_{v\in\mathcal{L}}$ of $[n]$ if and only if there exists Y such that

1. $\langle Yx, x\rangle = 0 \qquad$ for all $x \in \mathbb{R}^{S_v}$ and all $v \in \mathcal{L}$;
2. $\langle Yx, x\rangle \leq \|x\|^2$ for all $x \in \mathbb{R}^n$ and
3. $\langle Yx, x\rangle = \|x\|^2$ for all $x \in U$.

We remark that this differs from [113, Definition 5.2] by the transformation $Y \mapsto I - Y$. The parameterization in Definition 6.4.5 is more convenient for our purposes.

We now give some intuition behind this definition. Suppose U is realizable with respect to a partition $(S_v)_{v\in\mathcal{L}}$ of $[n]$. Then Definition 6.4.5 tells us that U comes equipped with a particular linear functional Y that separates symmetric matrices with column space U and those with column space given by any of the coordinate subspaces $\mathbb{R}^{S_v}$. Indeed if $\mathrm{col}(X_r) = U$ and $\mathrm{col}(X_v) = \mathbb{R}^{S_v}$ then the matrix Y in Definition 6.4.5 satisfies $\langle Y, X_r + X_v\rangle = \mathrm{tr}(X_r)$ (by parts 1 and 3 of the definition). This is a necessary condition if we hope to uniquely decompose sums of matrices with these column spaces. Indeed one can construct a linear functional with this property whenever $U \cap \mathbb{R}^{S_v} = \{0\}$ for all $v \in \mathcal{L}$. It is the addition of part 2 of the definition that makes Y useful.

Intuitively, it tells us that with respect to Y, matrices with column space U are not only distinguished from those with column space $\mathbb{R}^{S_v}$ (for any v), but are also distinguished from matrices with any other column space (in a way that makes them easy to find via convex optimization).

In the star-shaped case, if $X = X_r + \sum_{v\in\mathcal{L}} X_v$ is an LTCD of X then realizability of $\mathrm{col}(X_r)$ is sufficient for MTCD to recovers the LTCD of X.

Lemma 6.4.6. *Suppose $(S_v)_{v\in\mathcal{V}}$ form a star-shaped tree with root r. Suppose X has an LTCD with structure $(S_v)_{v\in\mathcal{V}}$ and suppose the subspace corresponding to the root is $U_r = \mathrm{col}(X_r)$. If U_r is realizable with respect to $(S_w)_{w\in\mathcal{L}}$ then MTCD recovers the LTCD of X.*

Proof. Simply note that if U_r is realizable with respect to $(S_w)_{w\in\mathcal{L}}$ then the matrix Y in Definition 6.4.5 satisfies the optimality conditions of Proposition 6.4.4. □

This means that sufficient conditions for realizability (with respect to a partition of $[n]$) translate into sufficient conditions for exact recovery of the decomposition in the star-shaped case. Two such sufficient conditions are given in [113], one in terms of the angles between certain subspaces, the other based on the notion of 'balanced' subspaces [36]. In this chapter we only make use of the simpler angle-based condition, as it is easier to work with in the general setting.

Definition 6.4.7. Let U, V be subspaces of $\mathbb{R}^n$. The *principal angle* between U and V is the angle $0 \leq \theta(U,V) \leq \pi/2$ such that

$$\cos(\theta(U,V)) = \max_{x\in U}\max_{y\in V}\frac{\langle x,y\rangle}{\|x\|\|y\|}.$$

It is straightforward to see that an alternative description of the cosine of the principal angle between subspaces U and V is

$$\cos(\theta(U,V)) = \|\pi_U\pi_V^*\| \tag{6.4.9}$$

where the norm used here is the spectral norm, i.e. the largest singular value. The main result of [113] is the following sufficient condition for realizability of a subspace.

Proposition 6.4.8 ([113, Corollary 5.11])**.** *Let U be a subspace of $\mathbb{R}^n$ and $(S_v)_{v\in\mathcal{L}}$ a partition of $[n]$. If $\cos(\theta(U,\mathbb{R}^{S_v}))^2 < 1/2$ for all $v\in\mathcal{L}$ then U is realizable with respect to $(S_v)_{v\in\mathcal{L}}$.*

To get a sense for what Proposition 6.4.8 means, we mention two examples from [113], both dealing with the case where $|S_v| = 1$ for all $v\in\mathcal{L}$. First, suppose the number of leaves is fixed at n and we consider how large the dimension of the subspace U can be so that the principal angle between U and each one-dimensional coordinate direction is greater than $\pi/4$. For this to hold we must have $\dim(U) \leq n/2$. Moreover, with high probability a random subspace of dimension $n(1/2 - \epsilon)$ (for small positive ϵ) makes an angle of at least $\pi/4$ with any coordinate direction. In the other extreme direction, if U has dimension one, we can ask how small n, the number of leaves, can be so that U still makes an angle of greater than $\pi/4$ with each coordinate axis. If $n = 2$ this is impossible, but it is possible for $n \geq 3$. That is, for an LTCD with respect to a star shaped tree to satisfy the condition of Proposition 6.4.8 the tree must have at least three leaves.

The general case

We now turn to the case where X has an LTCD with arbitrary structure. Our aim is to give sufficient conditions on the underlying decomposition of X that ensure the semidefinite optimization problem MTCD (6.4.5) recovers that decomposition. By the optimality conditions of Proposition 6.4.4 the problem reduces to finding a dual certificate Y satisfying (6.4.7) and (6.4.8). We can simplify the task of constructing a global dual certificate Y by constructing such a Y as a sum of local dual certificates Y_v that depend only on the relationship between a non-leaf vertex and its children. This is the sense in which understanding the star-shaped case is enough to understand the general tree case.

The next result shows how to construct a global dual certificate from local ones, and is the main technical lemma of this section. We note that we again require the assumption that the $(\lambda_v)_{v\in\mathcal{V}}$ are positive, order-preserving and constant on children.

Lemma 6.4.9. *Let $X \succ 0$ have an LTCD with subspaces $(U_v)_{v\in\mathcal{V}}$ and structure $(S_v)_{v\in\mathcal{V}}$. Suppose the $(\lambda_u)_{u\in\mathcal{V}}$ are positive, order-preserving and constant on children. Suppose that, for every $v \in \mathcal{V}$, $\pi_{S_v} U_v \subseteq \mathbb{R}^{S_v}$ is realizable with respect to the partition $(S_u)_{u\in\mathcal{C}(v)}$ of $\mathbb{R}^{S_v}$ and let $Y_v \in \mathcal{S}^{S_v}$ be the associated matrix in the definition of realizability (Definition 6.4.5). Then*

$$Y = \sum_{v\in\mathcal{V}} (\lambda_v - \lambda_{\mathcal{C}(v)}) \pi_{S_v}^* Y_v \pi_{S_v}.$$

satisfies the optimality conditions (6.4.7) and (6.4.8) and so MTCD recovers the LTCD of X.

Proof. We provide a proof in Section 6.7.2. □

By combining the sufficient condition for a subspace to be realizable (Proposition 6.4.8) with the fact that we can obtain global dual certificates from local ones

(Lemma 6.4.9) we arrive at the main result of this section. It gives a geometric condition on the subspaces $(U_v)_{v\in\mathcal{V}}$ of an LTCD of X that ensure it can be recovered by MTCD.

Theorem 6.4.10. *Suppose $X \succ 0$ and $X = \sum_{v\in\mathcal{V}} X_v$ is an LTCD of X with structure $(S_v)_{v\in\mathcal{V}}$ and subspaces $(U_v)_{v\in\mathcal{V}}$. Suppose the $(\lambda_v)_{v\in\mathcal{V}}$ are chosen to be positive, order-preserving, and constant on children. If the subspaces $(U_v)_{v\in\mathcal{V}}$ satisfy*

$$\cos(\theta(U_{\mathcal{P}(w)}, U_w))^2 < 1/2 \quad \textit{for all } w \in \mathcal{V}\setminus\{r\}$$

then MTCD recovers the LTCD of X.

Proof. Whenever $u \in \mathcal{V}\setminus\mathcal{L}$ and v is a child of u then by a simple technical result that exploits certain relationships between the supports and subspaces of an LTCD (Lemma 6.7.3, in Section 6.7.2) we have that

$$1/2 > \cos(\theta(U_u, U_v))^2 = \cos(\theta(\pi_{S_u}U_u, \pi_{S_u}\mathbb{R}^{S_v}))^2.$$

Hence by Proposition 6.4.8 we have that whenever $u \notin \mathcal{L}$, $\pi_{S_u}U_u$ is realizable with respect to the partition $(S_v)_{v\in\mathcal{C}(u)}$ of S_u. Then it follows from Lemma 6.4.9 that MTCD recovers the LTCD of X. □

Theorem 6.4.10 says that MTCD can recover any LTCD where the state space of any latent variable (when viewed from the leaves) makes an angle of more than $\pi/4$ with the state space of any of its children (again when viewed from the leaves). A condition of this form is very natural, as we expect it to be difficult to distinguish a parent and a child if their state spaces look similar from the point of view of the observation space. In the case where all the latent variables are one-dimensional, we note that for U_v to make and angle of more than $\pi/4$ with $U_{\mathcal{P}(v)}$ for every u it is necessary that every (non-leaf) vertex have at least three children. Indeed it becomes easier to satisfy the angle condition of Theorem 6.4.10 as the number of children each vertex has in a tree grows (with the dimensions of the variables being fixed).

■ 6.5 Uncovering the tree structure

In Section 6.4 we assumed a tree structure was given to us. Our aim was to take a matrix X that admits an LTCD with that structure and small dimensions, and recover it by solving a convex optimization problem. In this section we no longer assume the structure is given. Our aim is to take an $n \times n$ matrix X that admits an LTCD with some (unspecified) structure and small dimensions, and recover the decomposition and the structure.

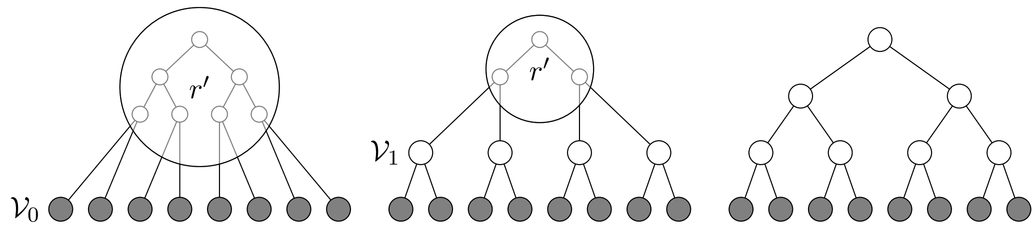


Figure 6.5: Trees corresponding to partial LTCDs at different scales of a constant depth tree.

There is one minor structural assumption we make about the tree that we now clarify. If we observe n scalar variables, we assume we are given a partition of $[n]$ into k disjoint sets. This partition tells us that the tree has k leaves and how to associate the n observed variables with the k leaves. In most previous work on this topic, it is silently assumed that the observed variables are scalar, and so that this initial partition is just $\{\{1\}, \{2\}, \ldots, \{n\}\}$. Our more general setting forces us to make this assumption explicit. We briefly discuss approaches to removing the reliance on an initial partition in Section 6.6.2.

With this assumption clarified, we now sketch our approach to finding the tree structure. Our approach is based on uncovering the tree in stages from the leaves to the root. If we know part of the tree structure, but do not know a connected subtree that contains the root, we can think of the remaining unknown part of the tree as a consisting of a single root node r' (see Figure 6.5). This grouping of variables still yields an LTCD (much as in Example 6.3.8), which we can try to recover using MTCD. We describe this in further detail in Section 6.5.2. The additional observation we exploit is that the column space of the term $X_{r'}$ in this new decomposition has additional structure. Indeed if the underlying model is non-singular then the column space of $X_{r'}$ decomposes into pieces that are supported on coordinate subspaces. This block structure contains information about the unknown structure in the tree.

Any optimal variable Y for the dual of MTCD obeys a complementarity condition with respect to $X_{r'}$ and so contains information about the column space of $X_{r'}$. In fact, any block structure in the column space of $X_{r'}$ is more readily available in these dual variables than in $X_{r'}$ itself. Indeed we show that if MTCD recovers the LTCD of X then the dual always has an optimal solution that is block diagonal with this block structure. To find the block structure in a robust way, we propose a method based on regularizing the dual of MTCD to induce block diagonal solutions.

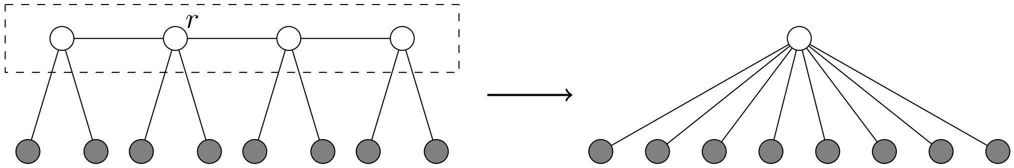


Figure 6.6: The tree on the left does not have constant depth. This is because no matter which variable is chosen as the root, the leaves will not all be at the same distance from the root. For instance if we choose the vertex labeled r as the root, the left-most leaf is at distance two from r and the right-most leaf is at distance three from r. If all of the non-leaf vertices are grouped together as a single vertex, the result (shown on the right) is a constant depth tree.

■ 6.5.1 Constant depth trees

In this section, we restrict our attention to rooted trees where all the leaves are at the same distance from the root. We call these *constant depth trees*. Any tree can be thought of as a constant depth tree by possibly grouping together many of the non-leaf nodes. Clearly this may significantly change the tree structure (see, e.g., Figure 6.6).

One convenient aspect of working only with constant depth trees is that we can organize the vertices into scales based on their distance from the leaves (or equivalently from the root). If $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ is a constant depth tree we define $\mathcal{V}_0 := \mathcal{L}$ and $\mathcal{V}_k = \{v \in \mathcal{V} : \mathcal{C}(v) \in \mathcal{V}_{k-1}\}$. and call this set of vertices *scale* k. We also define $\mathcal{V}_{\leq k} := \bigcup_{0 \leq m \leq k} \mathcal{V}_k$ to be the vertices with scale at most k. We define $\mathcal{V}_{\geq k}, \mathcal{V}_{<k}$, and $\mathcal{V}_{>k}$ in a similar way.

Suppose $(S_v)_{v \in \mathcal{V}}$ is a collection of subsets of $[n]$ that forms a constant depth tree with $r = \mathcal{V}_d$. Then for each $0 \leq k \leq d$ the sets $(S_v)_{v \in \mathcal{V}_k}$ form a partition of $[n]$. Furthermore, if $k_1 > k_2$ then the partition $(S_v)_{v \in \mathcal{V}_{k_2}}$ refines the partition $(S_v)_{\mathcal{V}_{k_1}}$ in the sense that if $v \in \mathcal{V}_{k_1}$ then S_v is a disjoint union of sets S_w with each $w \in \mathcal{V}_{k_2}$.

■ 6.5.2 Partial LTCDs

Suppose $X \succ 0$ has an LTCD $X = \sum_{v \in \mathcal{V}} X_v$ with structure $(S_v)_{v \in \mathcal{V}}$ (that forms a rooted, constant depth, tree $\mathcal{T}$) and subspaces $(U_v)_{v \in \mathcal{V}}$. Suppose, however, that we only have access to part of the structure of the LTCD. Specifically suppose there is some scale $0 < k \leq d$ such that we know $(S_v)_{v \in \mathcal{V}_{<k}}$, the structure of the tree from the leaves to scale $k-1$, but we do not know the structure of the tree at or above scale k. As such, we define a new tree by grouping all nodes in $\mathcal{V}_{\geq k}$ into a single root node denoted r' and connecting r' to all vertices at scale $k-1$. Correspondingly we define $S_{r'} = [n]$ and $\mathcal{V}' = \mathcal{V}_{<k} \cup \{r'\}$. With this construction established, we see that X has

an LTCD with structure $(S_v)_{v\in\mathcal{V}'}$. Explicitly this decomposition is given by

$$X = X_{r'} + \sum_{v\in\mathcal{V}_{<k}} X_v \quad \text{where} \quad X_{r'} := \sum_{u\in\mathcal{V}_{\geq k}} X_u.$$

We call this the *partial LTCD at scale* k.

The next result (which we prove in Section 6.7.3) investigates the column space of $X_{r'}$ in the partial LTCD at scale k.

Lemma 6.5.1. *Suppose that $X \succ 0$ has an LTCD with subspaces $(U_v)_{v\in\mathcal{V}}$ and structure that forms a rooted tree $\mathcal{T}$ of constant depth d. Then the corresponding partial LTCD at depth $k < d$ has*

$$\operatorname{col}(X_{r'}) \subseteq \bigoplus_{v\in\mathcal{V}_k} U_v.$$

Suppose, in addition, that the LTCD of X is non-singular. Then the partial LTCD at scale k is non-singular and

$$\operatorname{col}(X_{r'}) = \bigoplus_{v\in\mathcal{V}_k} \operatorname{col}(X_v) = \bigoplus_{v\in\mathcal{V}_k} U_v.$$

In the statement of Lemma 6.5.1 (and subsequently) if $U_1, U_2, \ldots, U_k \subset \mathbb{R}^n$ are subspaces such that $U_i \cap U_j = \{0\}$ for all $i \neq j$ we use the notation $\bigoplus_{i=1}^k U_i$ to mean the sum of the subspaces $U_1 + \cdots + U_k$. The notation applies here because the subspaces U_v for $v \in \mathcal{V}_k$ are actually supported on disjoint coordinate subspaces, and so certainly satisfy $U_v \cap U_{v'} = \{0\}$ for all $v, v' \in \mathcal{V}_k$.

Lemma 6.5.1 tells us that in the non-singular case, the column space of $X_{r'}$ has additional structure that tells us about the (unknown) kth scale of the tree. In particular the column space of $X_{r'}$ decomposes as a sum of subspaces that are supported on *disjoint coordinate subspaces*. If we can recover these coordinate subspaces from $X_{r'}$ then we can uncover the structure of the next scale of the tree. The following result (which we prove in Section 6.7.3) describes two ways to recover these coordinate subspaces. It forms the basis of our methods to find the entire tree structure.

If $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ is a constant depth tree we say that a collection $(\lambda_v)_{v\in\mathcal{V}}$ of scalars is *constant on scales* if $\lambda_u = \lambda_v$ whenever $u, v \in \mathcal{V}_k$ for some k. We use the notation λ_k to denote this common value at scale k.

Theorem 6.5.2. *Suppose $X \succ 0$ is $n \times n$ and has a non-singular LTCD with structure $(S_v)_{v\in\mathcal{V}}$ that forms a constant depth tree. Suppose, also, that the LTCD can be recovered by MTCD (with scalars $(\lambda_v)_{v\in\mathcal{V}}$ that are positive, order preserving, and constant on scales). If λ_k denotes the common value of λ_v for $v \in \mathcal{V}_k$, then for each scale $k > 0$ the following hold.*

1. *The partial LTCD at scale k can be recovered by partial MTCD at scale k, i.e. the unique optimum of*

$$\min_{(Z_v)_{v\in\mathcal{V}'}} \sum_{v\in\mathcal{V}_{<k}} \lambda_v \mathrm{tr}(Z_v) + \lambda_k \mathrm{tr}(Z_{r'}) \quad s.t. \quad \begin{cases} X = \sum_{v\in\mathcal{V}'} \pi_{S_v}^* Z_v \pi_{S_v} \\ X_v \succeq 0 \ \ \text{for all } v \in \mathcal{V}' \end{cases} \tag{6.5.1}$$

satisfies $\pi_{S_v}^* Z_v \pi_{S_v} = X_v$ *for all* $v \in \mathcal{V}_{<k}$ *and* $Z_{r'} = X_{r'}$.

2. *If $Z_{r'}$ is optimal for (6.5.1) then the orthogonal projector $P_{\mathrm{col}(Z_{r'})}$ is block diagonal with respect to the partition $(S_v)_{v\in\mathcal{V}_k}$ of $[n]$.*

3. *The dual of (6.5.1),*

$$\max_{Y} \langle X, Y\rangle \quad s.t. \quad \begin{cases} Y_{S_vS_v} \preceq \lambda_v I \ \ \text{for all } v \in \mathcal{V}_{<k} \\ Y \preceq \lambda_k I, \end{cases} \tag{6.5.2}$$

has an optimal solution that is block diagonal with respect to the partition $(S_v)_{v\in\mathcal{V}_k}$ of $[n]$.

Theorem 6.5.2 tells us that (under certain conditions), if we know the tree structure up to scale $k-1$ then we can *recognize* the tree structure at scale k in two different ways: by inspecting the block structure of $P_{\mathrm{col}(Z_{r'})}$, or by finding a block diagonal optimal solution (among the many optimal solutions) to the dual (6.5.2). If all we want to do is recognize when this structure appears, computing $P_{\mathrm{col}(Z_{r'})}$ is much simpler than finding a block diagonal solution to the dual. The advantage of considering dual solutions is that we are working directly with the decision variables of the optimization problem. Hence we can try to *induce* block diagonal dual solutions by appropriately regularizing the dual problem. Doing so is particularly useful in the approximate decomposition setting because it allows us to bias our search for approximate LTCDs towards those where the term $X_{r'}$ has a column space with certain block decomposition properties. This forms the basis of our method (described in Section 6.5.4) to perform approximate LTCDs and uncover the tree structure.

■ 6.5.3 A first procedure to recover constant depth trees

Theorem 6.5.2 tells us how to extract the tree structure at scale k from the tree structure at scale $k-1$. By repeatedly applying the result we see that for non-singular models, under the same conditions we need to recover an LTCD when the tree is known, we can also recover the tree structure. An explicit procedure to do this is given in Algorithm 6.1. It takes as input an $n \times n$ positive definite matrix X and a partition $(S_v)_{v\in\mathcal{V}_0}$ of $[n]$ corresponding to the support of the leaf-indexed variables. (Recall, from the introductory discussion of Section 6.5, that we always assume such a partition is given.) It produces a collection $(S_v)_{v\in\mathcal{V}}$ of subsets of $[n]$ that form a constant depth tree as well as an LTCD of X with structure $(S_v)_{v\in\mathcal{V}}$.

Algorithm 6.1 Recovering tree structure using block-diagonal orthogonal projectors

Input: $X \succ 0$, $(S_v)_{v\in\mathcal{V}_0}$
$S_{r'} \leftarrow [n]$
$k \leftarrow 0$
repeat
  $k \leftarrow k+1$ ▷ increment scale
  $Z_{r'} \leftarrow \textsc{mtcd}(X; (S_v)_{v\in(\mathcal{V}_k\cup r')})$ ▷ compute partial LTCD
  $(S_v)_{v\in\mathcal{V}_k} \leftarrow \textsc{blocks}(P_{\mathrm{col}(Z_{r'})}, (S_v)_{v\in\mathcal{V}_{k-1}})$ ▷ get next scale from blocks of $P_{\mathrm{col}(Z'_r)}$
until $|\mathcal{V}_k| = 1$ ▷ stop when $P_{\mathrm{col}(Z_{r'})}$ has no interesting block structure

Algorithm 6.1 makes use of two sub-procedures. The first, $Z_{r'} \leftarrow \textsc{mtcd}(X; (S_v)_{v\in\mathcal{V}_k\cup r'})$ involves solving (6.5.1) with input X and structure $(S_v)_{v\in\mathcal{V}_k\cup r'}$ and returning the optimal solution for $Z_{r'}$. The second, $\textsc{blocks}$, involves computing the block diagonal structure of $P_{\mathrm{col}(Z_{r'})}$ with respect to the partition $(S_v)_{v\in\mathcal{V}_{k-1}}$ of $[n]$. Indeed $(S_v)_{v\in\mathcal{V}_k} \leftarrow \textsc{blocks}(A, (S_v)_{v\in\mathcal{V}_{k-1}})$ can be implemented as follows.

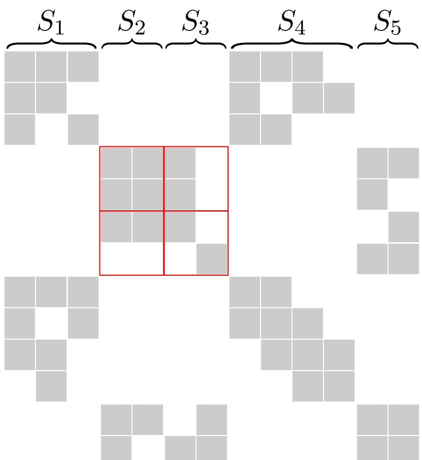
1. Form a graph with vertex set $\mathcal{V}_{k-1}$ and an edge between $u, v \in \mathcal{V}_{k-1}$ if and only if the corresponding block A_{S_u,S_v} is a non-zero matrix.

2. Find the connected components of this graph and let $\mathcal{V}_k$ index these connected components.

3. For each $v \in \mathcal{V}_k$, if $\{u_1, \ldots, u_p\} \subseteq \mathcal{V}_{k-1}$ is the corresponding connected component then define $S_v := \bigcup_{i=1}^{p} S_{u_i}$.

Observe that $\textsc{blocks}(A, (S_v)_{v\in\mathcal{V}_{k-1}})$ always produces a partition of $[n]$ that is refined by $(S_v)_{v\in\mathcal{V}_{k-1}}$. The procedure $\textsc{blocks}$ is illustrated on a specific example in Figure 6.7.

It is clear that, under the assumptions of Theorem 6.5.2, Algorithm 6.1 recovers the tree and the LTCD of X. We now discuss the differences between the conditions needed for recovery when the tree is known, compared with the conditions required for the procedure in Algorithm 6.1 to recover the LTCD and the tree structure. The differences are that Theorem 6.5.2 only holds when X has a non-singular LTCD and when the tree has constant depth. The assumption that the LTCD is non-singular seems essential to our approach, since the column space of $X_{r'}$ simply does not have the same block structure in the singular case.



$$\begin{bmatrix} A_{S_1S_1} & & & A_{S_1S_4} & \\ & A_{S_2S_2} & A_{S_2S_3} & & A_{S_2S_5} \\ & A_{S_3S_2} & A_{S_3S_3} & & A_{S_3S_5} \\ A_{S_4S_1} & & & A_{S_4S_4} & \\ & A_{S_5S_2} & A_{S_5S_3} & & A_{S_5S_5} \end{bmatrix}$$

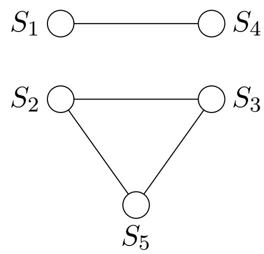


Figure 6.7: An illustration of aspects of the BLOCKS procedure applied to a 13×13 matrix with the corresponding input partition being $\mathcal{V}_{k-1} = \{S_1, S_2, S_3, S_4, S_5\}$ where $S_1 = \{1,2,3\}, S_2 = \{4,5\}, S_3 = \{6,7\}, S_4 = \{8,9,10,11\}, S_5 = \{12,13\}$. The sparsity pattern of the matrix A is shown on the left (we explain the additional red higlights in the following sentences). The input partition specifies that we should think of A as a block matrix, with the block entries being the submatrices $A_{S_iS_j}$. The non-zero entries in the block matrix are shown in the matrix on the top right. Observe that $A_{S_2S_2}, A_{S_2S_3}, A_{S_3S_2}$ and $A_{S_3S_3}$ are all non-zero, since the corresponding blocks of A (highlighted by red squares) are all non-zero matrices. On the bottom right is the graph obtained by putting an edge between S_i and S_j if and only if the matrix $A_{S_iS_j}$ is non-zero. Constructing this graph is the first step of BLOCKS. Observe that the vertices of this graph are labeled by the elements of $\mathcal{V}_{k-1}$. This graph has two connected components, $\{S_1, S_4\}$ and $\{S_2, S_3, S_5\}$. Hence $\mathcal{V}_k$ consists of two elements S_6 and S_7 (step two of BLOCKS). There are given by $S_6 = S_1 \cup S_4 = \{1,2,3,8,9,10,11\}$ and $S_7 = S_2 \cup S_3 \cup S_5 = \{4,5,6,7,12,13\}$ (step three of BLOCKS).

■ 6.5.4 Finding both a tree structure and an approximate LTCD

While Algorithm 6.1 can recover the tree structure (under appropriate assumptions) we do not expect natural variants of it to work in the approximate decomposition setting. Directly inspecting $P_{Z_{r'}}$ in Algorithm 6.1 only allows us to *recognize* when it is block diagonal and extract the associated block structure. There is no mechanism in Algorithm 6.1 that *encourages* partial LTCDs where the column space of $Z_{r'}$ has the appropriate block structure. Such a mechanism is unnecessary in the exact decomposition setting, but crucial in the approximate decomposition setting.

Part 3 of Theorem 6.5.2 suggests that finding block diagonal solutions to the dual of MTCD gives an alternative way to recover the tree structure in the exact decomposition setting. The following is a partial converse to part 3 of Theorem 6.5.2. It tells us that if X has an LTCD that can be recovered by MTCD and the dual has a block diagonal solution that obeys a strict complementarity condition, then the term X_r in the decomposition can be further decomposed in a non-trivial way. We prove this result in Section 6.7.3.

Proposition 6.5.3. *Suppose $X \succ 0$ is $n \times n$ and has a LTCD $\sum_{v\in\mathcal{V}} X_v$ with structure $(S_v)_{v\in\mathcal{V}}$ that can be recovered by MTCD. Suppose that the dual of MTCD has an optimal solution Y that is block diagonal with respect to the partition $(S_i)_{i=1}^k$ of $[n]$. If $\lambda_r I - Y$ and X_r satisfy the strict complementarity condition*

$$\operatorname{rank}(\lambda_r I - Y) + \operatorname{rank}(X_r) = n \tag{6.5.3}$$

then

$$X_r = X_0 + \sum_{i=1}^{k} X_i$$

where

1. $\operatorname{col}(X_r) \subseteq \bigoplus_{i=1}^k \operatorname{col}(X_i)$;
2. $\operatorname{rank}(X_i) = |S_i| - \operatorname{rank}(Y_{S_iS_i})$ *and* $\operatorname{supp}(\operatorname{col}(X_i)) \subseteq S_i$, *for all* $i = 1, 2, \ldots, k$;
3. $\operatorname{rank}(X_0) < \operatorname{rank}(X_r)$.

We remark that this result is non-trivial precisely because $\operatorname{rank}(X_0) < \operatorname{rank}(X_r)$, otherwise we could have just taken all $X_i = 0$ and $X_0 = X_r$.

Proposition 6.5.3 is intended to justify an approach to recovering the tree structure via regularizing the dual of MTCD to induce block diagonal solutions. Indeed if we bias our search for LTCDs towards those where the dual has block diagonal solutions, then we are, implicitly, biasing our search towards LTCDs where the term corresponding to the root can be further decomposed. This is the basic idea behind our procedure to construct approximate LTCDs without being initially given a tree structure.

Dual regularized approximate MTCD

Let $h : \mathcal{S}^n \to \mathbb{R}$ be a convex function that is designed to induce block-diagonal structure. More specifically, if $(S_1, S_2, \ldots, S_k)$ is a partition of $[n]$ we would like $h(Y; (S_i)_{i=1}^k)$ to encourage Y, thought of as a matrix with block entries $Y_{S_iS_j}$ to be block diagonal. Since such a matrix has only a small number of the block-entries $Y_{S_iS_j}$ being non-zero, one possible choice for h is

$$h(Y; (S_i)_{i=1}^k) = \sum_{i \neq j} \|Y_{S_iS_j}\|$$

where the norm on the blocks is the spectral norm. Let $h^*(\,\cdot\,; (S_i)_{i=1}^k)$ be the convex conjugate of h. This is given, explicitly, by

$$h^*(X; (S_i)_{i=1}^k) = \begin{cases} 0 & \text{if } X_{S_iS_i} = 0 \text{ for all } i \in [k] \text{ and } \|X_{S_iS_j}\|_* \leq 1 \text{ for all } i \neq j \\ \infty & \text{otherwise} \end{cases}$$

where $\|M\|_*$ is the nuclear norm of a matrix M, i.e. the sum of its singular values.

Let $\hat{X}$ be a given $n \times n$ matrix matrix for which we would like to find an approximate LTCD where the dual optimal solution is block diagonal. Then it is natural to solve a variant of approximate MTCD that has an additional regularization penalty in the dual. The regularization is based on $h(Y; (S_v)_{v\in\mathcal{C}(r)})$. Note that if $(S_v)_{v\in\mathcal{V}}$ form a rooted tree then $(S_v)_{v\in\mathcal{C}(r)}$ is a partition of $[n]$. It is the natural choice here because our aim is to find additional structure in the term corresponding to the root in the LTCD.

Using this regularization, the resulting pair of convex optimization problems is

$$\min_{\substack{\hat{X}=W+X+X_e \\ (Z_v)_{v\in\mathcal{V}}}} \gamma\left(\sum_{v\in\mathcal{V}} \lambda_v \operatorname{tr}(Z_v)\right) + \delta\, h^*(W/\delta; (S_v)_{v\in\mathcal{C}(r)}) + \frac{1}{2}\|X_e\|_F^2$$
$$\text{subject to} \quad \begin{cases} X = \sum_{v\in\mathcal{V}} \pi_{S_v}^* Z_v \pi_{S_v} \\ Z_v \in \mathcal{S}_+^{S_v} \text{ for all } v \in \mathcal{V} \end{cases} \tag{6.5.4}$$

and

$$\max_{Y} \langle \hat{X}, Y\rangle - \left[\delta\, h(Y; (S_v)_{v\in\mathcal{C}(r)}) + \frac{1}{2}\|Y\|_F^2\right]$$
$$\text{subject to} \quad Y_{S_vS_v} \preceq \gamma\lambda_v I \text{ for all } v \in \mathcal{V}. \tag{6.5.5}$$

where δ and γ are positive regularization parameters.

We have stated this pair of optimization problems for a particular structure $(S_v)_{v\in\mathcal{V}}$ that forms a tree. In Algorithm 6.2 we repeatedly use these optimization problems as a subroutine. Each time they are called, the input structure $(S_v)_{v\in\mathcal{V}}$ is *different*. In particular, the first time they are called the structure is the initial partition that tells us how to associate observed variables with leaves of the tree.

The primal problem (6.5.4) aims to decompose the given matrix as a sum of a matrix X that has an LTCD with structure $(S_v)_{v\in\mathcal{V}}$ and small dimensions, a matrix W that is small (in a sense determined by h^*), and a residual that has small Frobenius norm. As such, the effect of regularizing the dual (6.5.5) with h is to allow for another error term, measured differently, in the primal decomposition.

Even in this regularized approximate decomposition setting, if the dual has a block diagonal solution it is an indication that we can refine the decomposition we have found in the primal. After all, the optimal X in (6.5.4) necessarily has an LTCD (defined by the $(Z_v)_{v\in\mathcal{V}}$) that can be recovered by MTCD. Furthermore, any optimal dual solution Y satisfies the optimality conditions of MTCD with respect to the optimal LTCD of X. As such, if the optimal Y is block diagonal (and the strict complementarity condition in (6.5.3) holds) then by Proposition 6.5.3 it follows that Z_r can be decomposed further according to the block structure in Y.

This suggests the procedure in Algorithm 6.2 to construct both an LTCD of a matrix $\hat{X}$ and a corresponding tree structure. The subroutine BLOCKS is defined in the same

Algorithm 6.2 Recovering a tree for approximate LTCDs by dual regularization

Input: $\hat{X} \succ 0$, $(S_v)_{v\in\mathcal{V}_0}$
  $S_r \leftarrow [n]$
  $k \leftarrow 0$
  repeat
    $k \leftarrow k+1$ ▷ increment scale
    $Y \leftarrow$ A-D-MTCD$_{\delta,\gamma}(\hat{X};(S_v)_{v\in(\mathcal{V}_{<k}\cup r)})$ ▷ solve (6.5.5) for Y
    $(S_v)_{v\in\mathcal{V}_k} \leftarrow$ BLOCKS$(Y,(S_v)_{v\in\mathcal{V}_{k-1}})$ ▷ get next scale from blocks of Y
  until $|\mathcal{V}_k| = 1$ ▷ no interesting block structure in Y

way as in Section 6.5.3. As such, it always produces a partition that is refined by its input partition, and so Algorithm 6.2 necessarily returns $(S_v)_{v\in\mathcal{V}}$ that forms a constant depth tree. The subroutine A-D-MTCD$_{\delta,\gamma}(\hat{X};(S_v)_{v\in\mathcal{V}_{<k}\cup r'})$ returns an optimal solution to (6.5.5), the approximate dual version of LTCD.

We remark that in the limit as δ and γ go to zero, Algorithm 6.2 essentially reduces to Algorithm 6.1 but using part 3 rather part 2 of Theorem 6.5.2 to uncover the block structure of interest. To see this, note that taking the limits $\gamma \to 0$ and $\delta \to 0$ in (6.5.4) is the same as solving MTCD with input $\hat{X}$. Taking the same limits in the dual, (6.5.5), it is the case that the limit points of Y/γ are the solutions to the dual of LTCD that minimize $h(Y)$ over the dual optimal face. For the concrete choice of h given above, under the assumptions of Theorem 6.5.2, the minimum of h over the dual optimal face has the appropriate block diagonal structure (we show this in Lemma 6.7.5 of Section 6.7.3).

6.6 Summary and future work

6.6.1 Summary of contributions

In this chapter we considered the problem of approximating a given positive semidefinite matrix as the covariance among the leaf-indexed variables of a Gaussian tree model. We established a connection between the marginal covariance at the leaves of a Gaussian tree model and a certain family of matrix decomposition problems that we call latent tree covariance decompositions (LTCDs).

If the tree structure is fixed, the set of valid matrix decompositions is a convex set, while the complexity of the corresponding tree model is captured by the ranks of certain linear combinations of terms in the decomposition. A natural approach to searching for such decompositions is minimize a convex function that encourages low-rank solutions over valid decompositions. Using trace as a surrogate for rank, we obtain a formulation that we call minimum trace covariance decomposition (MTCD). We analyze this method, giving geometric conditions on an underlying Gaussian tree model that ensure it can be recovered from its leaf covariance using MTCD.

If the tree structure is unknown we proposed and analyzed a method based on solving a sequence of LTCD problems. The idea is to start with the simplest possible tree structure (a star), find an LTCD with respect to this structure, and then try to identifiy additional properties of the root-indexed latent variable that suggest it can be further decomposed. Our method to do this is based on regularizing the dual of MTCD. This has the benefit of allowing us to bias our search for decompositions towards those where the term corresponding to the root has the additional structure of interest.

6.6.2 Problems for future study

There are many possible directions for future research based on the approach to learning Gaussian latent trees presented in this chapter. We conclude the chapter by summarizing a selection of these.

Minimality

For any problem of latent variable modeling, it is desirable to learn minimal models. In the case where the tree structure is fixed, a minimal model is one where the dimensions of each of the latent state spaces cannot be made any smaller. In the case where all the latent variables have dimension one, notions of minimality of the tree structure (in terms of contraction and deletion of edges) are studied in [31] (see also [93]). When both the structure and the dimensions of the latent variables can vary, it is not completely obvious what the right *definition* of minimality is. In any case, our view of Gaussian latent tree models in terms of matrix decompositions may help in understanding and recovering minimal models. Indeed by refining the definition of an LTCD to require that the subsets that make up the structure satisfy additional properties, it may be possible to ensure that non-minimal trees cannot be obtained from these more refined LTCDs.

Removing parameterization dependence

Throughout this chapter we work with covariance parameterizations of Gaussian tree models and Gaussian latent tree models. This is particularly convenient when working with singular models since the inverse covariance parameterization, where the tree structure can be interpreted in terms of sparsity, is not well defined in that case. If the model is non-singular it is fairly easy to show that the inverse of the covariance among the leaves admits a structured matrix decomposition that is very similar to an LTCD. In the singular case, one needs to be careful since only some generalized inverses of the covariance of a Gaussian tree model preserve Markov structure.

Furthermore the LTCD approach depends on the choice of root for the tree. Since this choice is arbitrary, it should not play a role in any good method to learn Gaussian tree models. It would be desirable to develop an approach that still gives a convex parameterization of the set of models of interest, but does not require this arbitrary choice.

Approximate recovery, consistency and sample complexity

We have focused on establishing conditions under which our methods exactly recover the tree structure and the parameters of a Gaussian latent tree model from the exact covariance among the leaf variables. Nevertheless, our approach is specifically geared towards providing algorithms for learning Gaussian latent trees that are robust. In particular if the given matrix is close to the leaf-covariance of a Gaussian latent tree model (that can be recovered by MTCD), we expect that our methods produce an approximate LTCD that is close to the underlying exact LTCD. It would be interesting to investigate this recovery error in terms of natural distance measures on the terms in an LTCD and also in terms of structural properties such as the dimensions of the subspace U_v and the recovered tree structure.

When the given matrix is a sample-based approximation to the exact leaf-covariance of a Gaussian tree model, understanding the approximate recovery properties of our methods translates into understanding their consistency properties. Since we are interested in multiple structures, it may even be the case that the methods we present can consistently recover the underlying tree structure and the parameters of the model, but not, say, the dimensions of the subspaces U_v. If this were the case then we would like to modify the methods to ensure they are consistent in all senses of interest.

With consistency established the next natural question is to understand how many samples are required so that our methods applied to the given sample covariance can obtain the correct structural features of the model, and also the rate at which the model parameters converge to the true parameters. This is closely related to providing guidelines to choose the regularization parameters as a function of the number of samples available. Since the methods proposed in this chapter are much more global in nature than competing methods for learning Gaussian latent tree models, we expect that when correctly tuned they may, in practice, require fewer samples than other methods (even though the theoretical guarantees may be the same). It would also be interesting to investigate differences in the sample complexity of determining the combinatorial structure of the tree and the ranks of the latent variables. It is possible that one or the other of these quantities is significantly easier to learn in this sense.

Improved methods to learn the tree

The method we propose to learn the tree structure only works for non-singular models where all of the leaves are at the same distance from the root. It is likely that a modification of the method can remove this second assumption. Dealing with singular models seems difficult. This is because in the singular case the subspaces we seek to recover do not have the block structure our method exploits. Another place for improvement in this method is the choice of regularizer h in the dual of MTCD. The problem of regularizing to induce block diagonal structure is closely related to convex optimization-based methods for problems in clustering (see, e.g., [3, 2]) and to convex relaxations for sparse PCA (see, e.g., [34]). As such, we expect there are opportunities both to improve the choice of block-diagonal regularization we use here, and more broadly develop improved techniques for numerous problems where this structure naturally arises.

Furthermore, it would be good to remove the reliance of our methods on an initial partition that assigns observed variables to leaf-indexed vertices of a tree. In terms of matrix decompositions, this choice of initial partition tells us that we should start by solving a block-diagonal and low-rank decomposition problem, with the block diagonal structure specified by the initial partition. One obvious alternative is to start by seeking a decomposition of the given matrix as a sum of a sparse matrix plus a low-rank matrix, a problem that has been studied by Chandrasekaran, Parrilo, and Willsky [27]. A more sophisticated approach would aim to decompose the given matrix as the sum of a block diagonal matrix and a low-rank matrix, but with the block diagonal structure being unknown, using the same sort of block-diagonal regularization methods discussed in the previous paragraph.

6.7 Proofs for Chapter 6

6.7.1 Proofs for Section 6.3

The following result gathers useful facts about the relationship between the subspaces and the supports in an LTCD.

Lemma 6.7.1. *If $X = \sum_{v\in\mathcal{V}} X_v$ be an LTCD with supports $(S_v)_{v\in\mathcal{V}}$ and subspaces $(U_v)_{v\in\mathcal{V}}$ then*

1. $P_{S_w}U_u \subseteq U_w$ *whenever* $w \preceq u$;
2. $\pi_{U_v}(\pi_{U_w}^*\pi_{U_w})\pi_{U_w}^* = \pi_{U_v}\pi_{U_u}^*$ *whenever* $v \prec w \prec u$.

Proof. First note that since $w \preceq u$ we have that $S_w \subseteq S_u$ and so $P_{S_w}P_{S_u} = P_{S_w}$. Then

$$\left(\sum\nolimits_{t\succeq w} P_{S_w}X_tP_{S_w}\right) = P_{S_w}\left(\sum\nolimits_{t\succeq w} P_{S_u}X_tP_{S_u}\right)P_{S_w} \succeq P_{S_w}\left(\sum\nolimits_{t\succeq u} P_{S_u}X_tP_{S_u}\right)P_{S_w} \succeq 0 \tag{6.7.1}$$

where the second inequality holds because each X_u is positive semidefinite, and the first because, in addition, $\{t : t \succeq w\} \supseteq \{t : t \succeq u\}$. The column space of the left hand side of (6.7.1) is U_w and the column space of the right hand side of (6.7.1) is $P_{S_w}U_u$. Applying Lemma 6.2.2, which tells us that $B \succeq A \succeq 0$ implies $\mathrm{col}(B) \supseteq \mathrm{col}(A)$, we see that $P_{S_w}U_u \subseteq U_w$, establishing the first part of the Lemma.

For the second part, since $U_w \subseteq \mathbb{R}^{S_w}$ we have that $P_{U_w} = P_{U_w}P_{S_w}$ and so

$$\pi_{U_v}(\pi_{U_w}^*\pi_{U_w})\pi_{U_u} = \pi_{U_v}P_{U_w}P_{S_w}\pi_{U_u}.$$

From the first part of the lemma we know that $P_{S_w}U_u \subseteq U_w$ which means that $P_{U_w}(P_{S_w}\pi_{U_u}^*) = P_{S_w}\pi_{U_u}^*$. Combining these two observations with the fact that $U_v \subseteq \mathbb{R}^{S_v} \subseteq \mathbb{R}^{S_w}$ we obtain

$$\pi_{U_v}(\pi_{U_w}^*\pi_{U_w})\pi_{U_u} = \pi_{U_v}P_{U_w}(P_{S_w}\pi_{U_u}) = \pi_{U_v}P_{S_w}\pi_{U_u}^* = \pi_{U_v}\pi_{U_u}^*.$$

$\square$

Proof of Theorem 6.3.9

The facts established in the previous lemma are important in the proof of Theorem 6.3.9, that the covariances among the leaves of a Gaussian latent tree model are characterized by having an LTCD.

Proof of Theorem 6.3.9. We begin by establishing the converse. Most of the argument is sketched in the discussion prior to the definition of an LTCD; we make it precise here. Suppose X is the $n \times n$ covariance among the leaf-indexed variables of a Gaussian tree model with state dimensions $(n_v)_{v\in\mathcal{V}}$ and tree $\mathcal{T}$ with root r. Then there are $(A_v)_{v\in\mathcal{V}\setminus r}$ and $(Q_v)_{v\in\mathcal{V}}$ such that $Q_v \in \mathcal{S}_+^{n_v}$ for all $v \in \mathcal{V}$ and $A_v \in \mathbb{R}^{n_{\mathcal{P}(v)}\times n_v}$ for all $v \in \mathcal{V}\setminus r$ such that

$$X = \sum_{v\in\mathcal{V}} \Phi_{\mathcal{L}v} Q_v \Phi_{\mathcal{L}v}^T$$

(where Φ is defined according to (6.2.5)). Let $X_v = \Phi_{\mathcal{L}v} Q_v \Phi_{\mathcal{L}v}^T$ and note that each of these is positive semidefinite. For each $w \in \mathcal{L}$ define $S_w \subset [n]$ to index the state space of that leaf (so that $|S_w| = n_w$). For each $v \in \mathcal{V}\setminus\mathcal{L}$ define $S_v = \bigcup_{w\in\mathcal{D}(v)\cap\mathcal{L}} S_w$, the (disjoint) union of the state spaces of the leaves descended from v. Note that $S_v \subset S_u$ if and only if $v \preceq u$ in the tree order, so the collection $(S_v)_{v\in\mathcal{V}}$ forms the rooted tree $\mathcal{T}$. Furthermore, since $\Phi_{\mathcal{L}v}$ is supported only on the block rows indexed by $\mathcal{L}\cap\mathcal{D}(v)$ it follows that

$$S_v \supseteq \operatorname{supp}\left(\operatorname{col}\left(\Phi_{\mathcal{L}v} Q_v \Phi_{\mathcal{L}v}^T\right)\right) = \operatorname{supp}\left(\operatorname{col}(X_v)\right).$$

This establishes that $X = \sum_{v\in\mathcal{V}} X_v$ is a valid LTCD with structure $(S_v)_{v\in\mathcal{V}}$. The subspaces of this decomposition are $(U_v)_{v\in\mathcal{V}}$ where (by definition) U_v is the column space of

$$\begin{aligned}
\sum_{u\succeq v} P_{S_v} X_u P_{S_v} &= \sum_{u\succeq v} P_{S_v}\Phi_{\mathcal{L}u} Q_u \Phi_{\mathcal{L}u}^T P_{S_v} && \text{(by definition of } X_u\text{)}\\
&\overset{*}{=} \sum_{u\succeq v} \Phi_{\mathcal{L}v}\Phi_{vu} Q_u \Phi_{vu}^T \Phi_{\mathcal{L}v}^T && \text{(see comments below)}\\
&= \Phi_{\mathcal{L}v}\Sigma_{vv}\Phi_{\mathcal{L}v}^T && \text{(by the expression for } \Sigma_{vv} \text{ in (6.2.6)).}
\end{aligned}$$

The equality marked with an asterisk requires additional comment. It holds because if $v \preceq u$ and $w \in \mathcal{L}$ then

$$\Phi_{wv}\Phi_{vu} = \begin{cases} \Phi_{wu} & \text{if } w \preceq u\\ 0 & \text{otherwise.}\end{cases}$$

An equivalent way to write this is that $\Phi_{\mathcal{L}v}\Phi_{vu} = P_{S_v}\Phi_{\mathcal{L},u}$ whenever $v \preceq u$.

Since U_v is the column space of $\Phi_{\mathcal{L}v}\Sigma_{vv}\Phi_{\mathcal{L}v}^T$ we have that the corresponding dimension $m_v = \dim(U_v)$ of the decomposition satisfies $m_v \leq \mathrm{rank}(\Sigma_{vv}) \leq n_v$. This completes the proof of the converse.

Now suppose $X = \sum_{v\in\mathcal{V}} X_v$ is an LTCD with subspaces $(U_v)_{v\in\mathcal{V}}$, dimensions $(n_v)_{v\in\mathcal{V}}$ and supports $(S_v)_{v\in\mathcal{V}}$ that are subsets of $[n]$ and form the rooted tree $\mathcal{T}$. Then the explicit parameterization given in terms of the $(A_v)_{v\in\mathcal{V}\setminus r}$ and $(Q_v)_{v\in\mathcal{V}}$ certainly defines a Gaussian tree model Markov with respect to $\mathcal{T}$ and with state dimensions $(n_v)_{v\in\mathcal{V}}$.

It remains to check that the covariance among the leaves in this model is, indeed, X. Much of the work is done by Lemma 6.7.1 which describes relationships between the subspaces U_v for $v \in \mathcal{V}$. In particular part 2 of Lemma 6.7.1 tells us that whenever $v \prec w \prec u$ we have $\pi_{U_v}\pi_{U_w}^*\pi_{U_w}\pi_{U_u}^* = \pi_{U_v}\pi_{U_u}^*$. Repeatedly applying part 2 of Lemma 6.7.1 we see that if $v \prec u$ and w is the child of u satisfying $v \preceq w \prec u$ then

$$\Phi_{vu} = A_v A_{\mathcal{P}(v)}\cdots A_w = \pi_{U_v}\pi_{U_{\mathcal{P}(v)}}^*\pi_{U_{\mathcal{P}(v)}}\cdots\pi_{U_w}\pi_{U_w}\pi_{U_u}^* = \pi_{U_v}\pi_{U_u}^*. \tag{6.7.2}$$

Whenever $w \in \mathcal{L}$ we have that $U_w = \mathrm{col}(X_{S_wS_w})$. Because $X \succ 0$ (by assumption) we have that $U_w = \mathrm{col}(X_{S_wS_w}) = \mathbb{R}^{S_w}$ whenever $w \in \mathcal{L}$. Applying (6.7.2) we see that if $w \in \mathcal{L}$, $\Phi_{wu} = \pi_{S_w}\pi_{U_u}^*$ for any $u \in \mathcal{V}$. Hence $\Phi_{\mathcal{L}u} = \pi_{U_u}^*$ for any $u \in \mathcal{V}$. Hence the covariance among the leaves of the Gaussian tree model defined by the $(Q_v)_{v\in\mathcal{V}}$ and the $(A_v)_{v\in\mathcal{V}\setminus\{r\}}$ is

$$\begin{aligned}\sum_{v\in\mathcal{V}}\Phi_{\mathcal{L}v}Q_v\Phi_{\mathcal{L}v}^T &= \sum_{v\in\mathcal{V}}\pi_{U_v}^*Q_v\pi_{U_v} && \text{(by (6.7.2))}\\ &= \sum_{v\in\mathcal{V}}P_{U_v}X_vP_{U_v} && \text{(by definition of } Q_v)\\ &= \sum_{v\in\mathcal{V}}X_v && \text{(since } \mathrm{col}(X_u) \subseteq U_u \text{ by Lemma 6.2.2)}\end{aligned}$$

which is X by the definition of an LTCD. □

Proof of Lemma 6.3.11

We now establish the two characterizations of non-singular LTCDs given in Lemma 6.3.11.

Proof of Lemma 6.3.11. We begin by observing that

$$0 \preceq X_v \preceq X_v + P_{S_v}X_{\mathcal{P}(v)}P_{S_v} \preceq \sum_{u\succeq v}P_{S_v}X_uP_{S_v}.$$

Hence by Lemma 6.2.2 the following inclusions always hold

$$\mathrm{col}(X_v) \subseteq \mathrm{col}(X_v + P_{S_v}X_{\mathcal{P}(v)}P_{S_v}) \subseteq \mathrm{col}\left(\textstyle\sum_{u\succeq v}P_{S_v}X_uP_{S_v}\right) = U_v. \tag{6.7.3}$$

If part 3 holds, i.e. $U_v = \mathrm{col}(X_v)$ for all $v \in \mathcal{V}$ then all the subspaces in (6.7.3) are equal. Hence $\mathrm{col}(X_v) = \mathrm{col}(X_v + P_{S_v} X_{\mathcal{P}(v)} P_{S_v})$. Equivalently $\mathrm{col}(P_{S_v} X_{\mathcal{P}(v)} P_{S_v}) \subseteq \mathrm{col}(X_v)$ so part 2 holds.

Conversely suppose part 2 holds. We establish part 3 by induction on the distance between v and the root. If $d(v,r) = 0$ then part 3 holds vacuously. The induction hypothesis is that whenever $d(v,r) = k$ then $\mathrm{col}(X_v) = \mathrm{col}\left(\sum_{u \succeq v} P_{S_v} X_u P_{S_v}\right)$. Assume $d(v,r) = k+1$ and so that $d(\mathcal{P}(v), r) = k$. Then

$$\begin{aligned}
\mathrm{col}(X_v) &\supseteq \mathrm{col}\left(P_{S_v} X_{\mathcal{P}(v)} P_{S_v}\right) && \text{since part 2 holds}\\
&= P_{S_v} \mathrm{col}(X_{\mathcal{P}(v)})\\
&= P_{S_v} \mathrm{col}\left(\textstyle\sum_{u \succeq \mathcal{P}(v)} P_{S_{\mathcal{P}(v)}} X_u P_{S_{\mathcal{P}(v)}}\right) && \text{by the induction hypothesis}\\
&= \mathrm{col}\left(\textstyle\sum_{u \succeq \mathcal{P}(v)} P_{S_v} X_u P_{S_v}\right) && \text{since } S_v \subseteq S_{\mathcal{P}(v)}.
\end{aligned}$$

Hence $\mathrm{col}(X_v) = \mathrm{col}\left(X_v + \sum_{u \succeq \mathcal{P}(v)} P_{S_v} X_u P_{S_v}\right) = U_v$ and so part 3 holds.

We have shown that part 2 and part 3 are equivalent. We now show that part 1 and part 3 are equivalent. The LTCD $X = \sum_{v \in \mathcal{V}} X_v$ is non-singular if and only if $Q_v = \pi_{U_v} X_v \pi_{U_v}^* \succ 0$ (this is clear from the factorization $\Sigma = \Phi Q \Phi^T$ from (6.2.6), where Q has the Q_v on the block diagonal and Φ is non-singular). If $\pi_{U_v} X_v \pi_{U_v}^* \succ 0$ it follows that $\mathrm{col}(X_v) \supseteq U_v$. Since $U_v \subseteq \mathrm{col}(X_v)$ by definition, we have that $\mathrm{col}(X_v) = U_v$ and so part 3 holds. Conversely if part 3 holds then $Q_v = \pi_{U_v} X_v \pi_{U_v}^T \succ 0$ so part 1 holds. □

■ 6.7.2 Proofs for Section 6.4

We begin with a lemma that establishes the basic reason why we require scalars $(\lambda_v)_{v \in \mathcal{V}}$ to be constant on children in a number of places in the discussion.

Lemma 6.7.2. *If $(\lambda_v)_{v \in \mathcal{V}}$ are constant on children (see Definition 6.4.1) then for all $u \in \mathcal{V}$*

$$\lambda_u P_{S_u} = \sum_{v \preceq u} (\lambda_v - \lambda_{\mathcal{C}(v)}) P_{S_v}.$$

Proof. We argue by induction on the maximum distance from u to any descendant of u, i.e. $\max_{t \preceq u} d(u,t)$. For the base case, suppose $u \in \mathcal{L}$ so that $\max_{t \preceq u} d(u,t) = 0$. Then since $\lambda_{\mathcal{C}(u)} = 0$ and $\{v : v \preceq u\} = \{u\}$,

$$\sum_{v \preceq u} (\lambda_v - \lambda_{\mathcal{C}(v)}) P_{S_v} = (\lambda_u - \lambda_{\mathcal{C}(u)}) P_{S_u} = \lambda_u P_{S_u}.$$

Assume that the result holds for all vertices u such that $\max_{t \preceq u} d(u,t) = k$. Suppose $u \in \mathcal{V}$ is such that $\max_{t \preceq u} d(u,t) = k+1$. Then $\max_{t \preceq w} d(w,t) = k$ for all $w \in \mathcal{C}(u)$.

Now we can break up the sum over descendants of u into a term for u and a sum over the descendants of each of the children of u as

$$\sum_{v \preceq u}(\lambda_v - \lambda_{\mathcal{C}(v)})P_{S_v} = (\lambda_u - \lambda_{\mathcal{C}(u)})P_u + \sum_{w \in \mathcal{C}(u)} \left[\sum_{v \preceq w}(\lambda_v - \lambda_{\mathcal{C}(v)})P_{S_v}\right].$$

Applying the induction hypothesis to the right hand side we obtain

$$\sum_{v \preceq u}(\lambda_v - \lambda_{\mathcal{C}(v)})P_{S_v} = (\lambda_u - \lambda_{\mathcal{C}(u)})P_u + \sum_{w \in \mathcal{C}(u)} \lambda_w P_{S_w}.$$

Since $\lambda_w = \lambda_{\mathcal{C}(u)}$ for all $w \in \mathcal{C}(u)$ and $\sum_{w \in \mathcal{C}(u)} P_{S_w} = P_{S_u}$ it follows that

$$\sum_{v \preceq u}(\lambda_v - \lambda_{\mathcal{C}(v)})P_{S_v} = (\lambda_u - \lambda_{\mathcal{C}(u)} + \lambda_{\mathcal{C}(u)})P_{S_u} = \lambda_u P_{S_u}$$

completing the proof. □

Proof of Lemma 6.4.2

This enables us to establish the correctness of our simplification of the objective function of MTCD.

Proof of Lemma 6.4.2. First, since the $(\lambda_v)_{v\in\mathcal{V}}$ are non-negative, order preserving, and constant on children it follows that $\kappa_v := \lambda_v - \lambda_{\mathcal{C}(v)} > 0$ for all $v \in \mathcal{V}$. Then since trace is a linear function, we can rewrite the objective function of (6.4.4) in a simpler form as

$$\begin{aligned}
\operatorname{tr}\left[\sum_{v\in V}\sum_{u \succeq v} \kappa_v [\pi_{S_u}^* Z_u \pi_{S_u}]_{S_v S_v}\right] &= \sum_{u\in\mathcal{V}} \operatorname{tr}\left[\sum_{v \preceq u} \kappa_v [\pi_{S_u}^* Z_u \pi_{S_u}]_{S_v S_v}\right] \\
&= \sum_{u\in\mathcal{V}} \operatorname{tr}\left[\pi_{S_u}^* Z_u \pi_{S_u}\left(\sum_{v \preceq u}\kappa_v P_{S_v}\right)\right].
\end{aligned}$$

Here the first equality holds by changing the order of summation and using linearity of the trace. The second equality holds by linearity of trace and the fact that $\operatorname{tr}([A]_{SS}) = \operatorname{tr}(\pi_S A \pi_S^*) = \operatorname{tr}(AP_S)$. Since the $(\lambda_v)_{v\in\mathcal{V}}$ are constant on children we can apply Lemma 6.7.2 to obtain

$$\sum_{u\in\mathcal{V}} \operatorname{tr}\left[\pi_{S_u}^* Z_u \pi_{S_u}\left(\sum_{v \preceq u}\kappa_v P_{S_v}\right)\right] = \sum_{u\in\mathcal{V}} \operatorname{tr}\left[\pi_{S_u}^* Z_u \pi_{S_u}\left(\lambda_u P_{S_u}\right)\right] = \sum_{u\in\mathcal{V}} \lambda_u \operatorname{tr}(Z_u)$$

completing the proof. ☐

Proof of Proposition 6.4.4

Proof of Proposition 6.4.4. The condition that there exists Y such that for all $v \in \mathcal{V}$,

$$\begin{aligned} \langle Yx, x\rangle &\leq \lambda_v \|x\|^2 \quad \text{for all } x \in \mathbb{R}^{S_v} \text{ and} \\ \langle Yx, x\rangle &= \lambda_v \|x\|^2 \quad \text{for all } x \in \mathrm{col}(X_v) \subseteq \mathbb{R}^{S_v}. \end{aligned}$$

is equivalent to the condition that there exists Y such that for all $v \in \mathcal{V}$,

$$Y_{S_vS_v} \preceq \lambda_v I \tag{6.7.4}$$

$$Y_{S_vS_v} Z_v = \lambda_v Z_v. \tag{6.7.5}$$

If X is strictly positive definite then the primal problem is strictly feasible. We have already seen that the dual problem is strictly feasible. The conditions (6.7.4) and (6.7.5) are precisely a statement of the optimality conditions for semidefinite optimization under strict feasibility of the primal and dual problems (Theorem 2.4.3 in Chapter 2).

It remains to show that if MTCD is feasible then it always has a unique optimal point. Suppose $(Z_v^{(1)})_{v\in\mathcal{V}}$ and $(Z_v^{(2)})_{v\in\mathcal{V}}$ are both optimal for MTCD. Then by convexity, the averages $((Z_v^{(1)} + Z_v^{(2)})/2)_{v\in\mathcal{V}}$ are also optimal for MTCD. Let Y satisfy, for all $v \in \mathcal{V}$,

$$Y_{S_vS_v} \preceq \lambda_v I \quad \text{and} \quad Y_{S_vS_v}(Z_v^{(1)} + Z_v^{(2)}) = \lambda_v(Z_v^{(1)} + Z_v^{(2)})$$

which exists by the argument in the previous paragraph. Let V_{λ_v} denote the λ_v-eigenspace of Y and note that $\mathrm{col}(Z_v^{(1)}+Z_v^{(2)}) \subseteq V_{\lambda_v}$. In addition, since $Z_v^{(1)} \preceq Z_v^{(1)}+Z_v^{(2)}$ and $Z_v^{(2)} \preceq Z_v^{(1)} + Z_v^{(2)}$ we can conclude (from Lemma 6.2.2) that $\mathrm{col}(Z_v^{(1)}) \subseteq V_{\lambda_v}$ and $\mathrm{col}(Z_v^{(2)}) \subseteq V_{\lambda_v}$ and so that

$$Y_{S_vS_v} Z_v^{(1)} = \lambda_v Z_v^{(1)} \quad \text{and} \quad Y_{S_vS_v} Z_v^{(2)} \quad \text{for all } v \in \mathcal{V}. \tag{6.7.6}$$

Assume, seeking a contradiction, that there is some v such that $Z_v^{(1)} \neq Z_v^{(2)}$, and choose such a v that is maximal with this property, i.e. so that if $t \succ v$ then $Z_t^{(1)} = Z_t^{(2)}$. Observe that

$$\Sigma_{\mathcal{D}(v)\cap\mathcal{L},\mathcal{D}(v)\cap\mathcal{L}} = \sum_{w \prec v} [\pi_{S_w}^* Z_w^{(i)} \pi_{S_w}]_{S_vS_v} + Z_v^{(i)} + \sum_{t \succ v} [\pi_{S_t}^* Z_t^{(i)} \pi_{S_t}]_{S_vS_v}$$

for $i = 1, 2$. Subtracting these two equations we see that

$$\begin{aligned} 0 &= \sum_{w \prec v} [\pi_{S_w}^* (Z_w^{(1)} - Z_w^{(2)}) \pi_{S_w}]_{S_v S_v} + (Z_v^{(1)} - Z_v^{(2)}) + \sum_{t \succ v} [\pi_{S_t}^* (Z_t^{(1)} - Z_t^{(2)}) \pi_{S_t}]_{S_v S_v} \\ &= \sum_{w \prec v} [\pi_{S_w}^* (Z_w^{(1)} - Z_w^{(2)}) \pi_{S_w}]_{S_v S_v} + (Z_v^{(1)} - Z_v^{(2)}) \end{aligned} \tag{6.7.7}$$

where the second equality holds by our choice of v. Since $Z_v^{(1)} \neq Z_v^{(2)}$ (by assumption) for (6.7.7) to hold there must be some $w \prec v$ such that $Z_w^{(1)} \neq Z_w^{(2)}$. Let $w_1, w_2, \ldots, w_k$ be the set of maximal elements of $\{w \in \mathcal{V} : w \prec v \text{ and } Z_w^{(1)} \neq Z_w^{(2)}\}$. Then observe that we can rewrite (6.7.7) as

$$Z_v^{(2)} - Z_v^{(1)} = \sum_{i=1}^{k} \sum_{u_i \preceq w_i} [\pi_{S_{u_i}}^* (Z_{u_i}^{(1)} - Z_{u_i}^{(2)}) \pi_{S_{u_i}}]_{S_v S_v}.$$

It follows that $Z_v^{(2)} - Z_v^{(1)}$ is block diagonal with support on the blocks indexed by S_{w_i} for $i = 1, 2, \ldots, k$. These blocks are disjoint because the w_i are maximal and so are incomparable in the tree order. Because of the support of $Z_v^{(2)} - Z_v^{(1)}$, for each $i = 1, 2, \ldots, k$ we have that

$$Y_{S_{w_i} S_{w_i}} [Z_v^{(2)} - Z_v^{(1)}]_{S_{w_i} S_{w_i}} = [Y(Z_v^{(2)} - Z_v^{(1)})]_{S_{w_i} S_{w_i}} = \lambda_v [Z_v^{(2)} - Z_v^{(1)}]_{S_{w_i} S_{w_i}} \tag{6.7.8}$$

where the last equality is from (6.7.6). Now, for all $i = 1, 2, \ldots, k$ we have that $\lambda_v > \lambda_{w_i}$ because the $(\lambda_v)_{v \in \mathcal{V}}$ are order preserving. Hence for all $i = 1, 2, \ldots, k$ we have that

$$\lambda_v I - Y_{S_{w_i} S_{w_i}} \succ \lambda_{w_i} I - Y_{S_{w_i} S_{w_i}} \succeq 0$$

where the right hand side is positive semidefinite because Y satisfies (6.7.4). Hence $\lambda_v I - Y_{S_{w_i} S_{w_i}}$ is invertible for $i = 1, 2, \ldots, k$. Then we can see from (6.7.8) that

$$(\lambda_v I - Y_{S_{w_i} S_{w_i}}) [Z_v^{(1)} - Z_v^{(2)}]_{S_{w_i} S_{w_i}} = 0$$

for all $i = 1, 2, \ldots, k$. It then follows that $Z_v^{(1)} = Z_v^{(2)}$, a contradiction. □

Proof of Lemma 6.4.9

The main technical lemma of Section 6.4 establishes that we can construct a global dual certificate for MTCD out of local dual certificates. We now establish this result.

Proof of Lemma 6.4.9. Our aim is to understand

$$\langle Yx, x\rangle = \sum_{v\in\mathcal{V}} (\lambda_v - \lambda_{\mathcal{C}(v)})\langle \pi_{S_v}^* Y_v \pi_{S_v} x, x\rangle$$

when $x \in \mathbb{R}^{S_u}$ (and also when $x \in \mathrm{col}(X_u) \subseteq \mathbb{R}^{S_u}$). To this end we begin by studying $\langle \pi_{S_v}^* Y_v \pi_{S_v} x, x\rangle$ when $x \in \mathbb{R}^{S_u}$ and $u, v \in \mathcal{V}$ are arbitrary. We claim that

$$\langle \pi_{S_v}^* Y_v \pi_{S_v} x, x\rangle = 0 \quad \text{for all } x \in \mathbb{R}^{S_u} \text{ whenever } u, v \in \mathcal{V} \text{ and } u \not\succeq v. \tag{6.7.9}$$

First assume u and v are incomparable. Then $\mathbb{R}^{S_u}$ and $\mathbb{R}^{S_v}$ are orthogonal subspaces so $\pi_{S_v} x = 0$ for all $x \in \mathbb{R}^{S_u}$. Now if u and v are comparable and $u \not\succeq v$ then $u \prec v$. Then there is some w such that w is a child of v and $u \preceq w \prec v$. Then $\mathbb{R}^{S_u} \subseteq \mathbb{R}^{S_w}$ so that whenever $x \in \mathbb{R}^{S_u}$ we have that $\langle \pi_{S_v}^* Y_v \pi_{S_v} x, x\rangle = 0$ (by the first assumption on Y_v), establishing (6.7.9).

It follows from this claim that whenever $x \in \mathbb{R}^{S_u}$

$$\langle Yx, x\rangle = \sum_{v\in\mathcal{V}} (\lambda_v - \lambda_{\mathcal{C}(v)})\langle \pi_{S_v}^* Y_v \pi_{S_v} x, x\rangle = \sum_{v\preceq u} (\lambda_v - \lambda_{\mathcal{C}(v)})\langle \pi_{S_v}^* Y_v \pi_{S_v} x, x\rangle \tag{6.7.10}$$

since the remaining terms in the sum are zero.

Now we establish that whenever $(\lambda_v)_{v\in\mathcal{V}}$ is constant on children and $\lambda_v > \lambda_{\mathcal{C}(v)}$ for all v then Y satisfies the first of the optimality conditions (6.4.7). To do so observe that if $u \in \mathcal{V}$ and $x \in \mathbb{R}^{S_u}$

$$\begin{aligned}
\langle Yx, x\rangle &= \sum_{v\preceq u} (\lambda_v - \lambda_{\mathcal{C}(v)})\langle \pi_{S_v}^* Y_v \pi_{S_v} x, x\rangle \\
&= \sum_{v\preceq u} (\lambda_v - \lambda_{\mathcal{C}(v)})\langle \pi_{S_v}^* Y_v \pi_{S_v} P_{S_v} x, P_{S_v} x\rangle & (6.7.11)\\
&\leq \sum_{v\preceq u} (\lambda_v - \lambda_{\mathcal{C}(v)})\|P_{S_v} x\|^2 & (6.7.12)\\
&= \lambda_u \|P_{S_u} x\|^2 & (6.7.13)\\
&= \lambda_u \|x\|^2 & (6.7.14)
\end{aligned}$$

where the inequality holds by the second assumption on Y_v (and the fact that $\lambda_v > \lambda_{\mathcal{C}(v)}$ for all v), the second-last equality follows from Lemma 6.7.2 and the last equality holds since $x \in \mathbb{R}^{S_u}$. We have now established that Y satisfies (6.4.7).

To see that (6.4.8) holds, we use a similar argument. The key additional observation we need is that whenever $u \succeq v$ then $P_v\mathrm{col}(X_u) \subseteq U_v$. This holds because

$$U_v = \mathrm{col}\left(\textstyle\sum_{w \succeq v} P_{S_v} X_w P_{S_v}\right) \supseteq \mathrm{col}(P_{S_v} X_u P_{S_v}) = P_{S_v}\mathrm{col}(X_u)$$

where the inclusion follows from Lemma 6.2.2. With this observation established, if $u \in \mathcal{V}$ and $x \in \mathrm{col}(X_u) \subseteq \mathbb{R}^{S_u}$ then (reusing the steps from (6.7.11), (6.7.13), and (6.7.14)) we have that

$$\begin{aligned}\langle Yx, x\rangle &= \sum_{v \preceq u}(\lambda_v - \lambda_{\mathcal{C}(v)})\langle \pi_{S_v}^* Y_v \pi_{S_v} P_{S_v} x, P_{S_v} x\rangle \\ &= \sum_{v \preceq u}(\lambda_v - \lambda_{\mathcal{C}(v)})\|P_{S_v} x\|^2 \qquad (6.7.15)\\ &= \lambda_u \|x\|^2\end{aligned}$$

where (6.7.15) follows from the fact that $P_{S_v}x \in P_{S_v}\mathrm{col}(X_u) \subseteq U_v$ and from the third assumption on Y_v. This completes the proof. □

The following technical result allows us to state our sufficient conditions for MTCD to recover an LTCD (Theorem 6.4.10) in terms of the angles between the subspaces U_v and $U_{\mathcal{P}(v)}$ from an LTCD. Without this lemma the same result would hold, but it would be stated in a less aesthetically appealing (but completely equivalent) way in terms of the angle between the subspaces $\pi_{S_{\mathcal{P}(v)}} U_{\mathcal{P}(v)}$ and $\pi_{S_{\mathcal{P}(v)}}\mathbb{R}^{S_v}$.

Lemma 6.7.3. *Suppose $X = \sum_{v\in\mathcal{V}} X_v$ is an LTCD with supports $(S_v)_{v\in\mathcal{V}}$ and subspaces $(U_v)_{v\in\mathcal{V}}$. Then*

$$\cos(\theta(U_u, U_w))^2 = \cos(\,\theta(\pi_{S_u} U_u, \pi_{S_u}\mathbb{R}^{S_w})\,)^2 \quad \textit{whenever } w \preceq u.$$

Proof. We start by observing that since $U_u \subseteq \mathbb{R}^{S_u}$ and $\mathbb{R}^{S_w} \subseteq \mathbb{R}^{S_u}$ it follows that $\cos(\,\theta(\pi_{S_u} U_u, \pi_{S_u}\mathbb{R}^{S_w})\,)^2 = \cos(\,\theta(U_u, \mathbb{R}^{S_w})\,)^2$. To conclude the proof we use the results of Lemma 6.7.1 and the definition of the principal angle between subspaces to see that:

$$\begin{aligned}\cos(\theta(U_u, \mathbb{R}^{S_w}))^2 &= \max_{x\in U_u} \frac{\|P_{S_w} x\|^2}{\|x\|^2} \\ &= \max_{x\in U_u} \frac{\|P_{U_w} P_{S_w} x\|^2}{\|x\|^2} \quad \text{(since } P_{S_w} U_u \subseteq U_w \text{ from part 1 of Lemma 6.7.1)} \\ &= \max_{x\in U_u} \frac{\|P_{U_w} x\|^2}{\|x\|^2} \quad \text{(since } U_w \subseteq \mathbb{R}^{S_w}\text{)} \\ &= \cos(\theta(U_u, U_w))^2.\end{aligned}$$

□

6.7.3 Proofs for Section 6.5

We begin with a useful result that shows how to write the projection onto the block diagonal part of a symmetric matrix as a convex combination of symmetric matrices.

If $(S_i)_{i=1}^k$ is a partition of $[n]$ then define $G((S_i)_{i=1}^k)$ to be the group consisting of matrices of the form

$$g = \sum_{i=1}^{k} \epsilon_i P_{S_i} \quad \text{where } \epsilon_i \in \{-1, 1\} \text{ for all } i \in [k].$$

This is a group because $P_{S_i} P_{S_j} = \delta_{ij} P_{S_i}$. It consists of diagonal sign matrices that are constant on the part of the diagonal indexed by S_i, for each $i = 1, 2, \ldots, k$.

Lemma 6.7.4. *If $(S_i)_{i=1}^k$ is a partition of $[n]$ and $G = G((S_i)_{i=1}^k)$ then the orthogonal projection onto symmetric matrices that are block diagonal with respect to $(S_i)_{i=1}^k$ is*

$$X \mapsto \sum_{i=1}^{k} P_{S_i} X P_{S_i} = \frac{1}{|G|} \sum_{g \in G} g X g^T.$$

Proof. The action $g \cdot X = gXg^T$ is a valid action of the group G on symmetric matrices. Moreover the action preserves the trace inner product, i.e. $\langle X, Y \rangle = \langle gXg^T, gYg^T \rangle$ for all $X, Y \in \mathcal{S}^n$. The subspace of symmetric matrices satisfying $X = gXg^T$ for all $g \in G$ is precisely the set of matrices block diagonal with respect to $(S_i)_{i=1}^k$. The result then follows from Lemma 2.6.6 of Chapter 2, which states that the orthogonal projector onto the fixed-point subspace of a group action that preserves inner products is given by averaging over the group action. □

Proof of Lemma 6.5.1

Proof of Lemma 6.5.1. Let $G = G(\,(S_v)_{v \in \mathcal{V}_k}\,)$. Then by Lemma 6.7.4

$$\sum_{v \in \mathcal{V}_k} P_{S_v} \left(\sum_{u \succeq v} X_u \right) P_{S_v} = \sum_{v \in \mathcal{V}_k} P_{S_v} X_{r'} P_{S_v} = \frac{1}{|G|} \sum_{g \in G} g X_{r'} g^T \succeq \frac{1}{|G|} X_{r'} \succeq 0.$$

It then follows from Lemma 6.2.2 that

$$\bigoplus_{v \in \mathcal{V}_k} U_v = \mathrm{col}\left(\sum_{v \in \mathcal{V}_k} P_{S_v} \left(\sum_{u \succeq v} X_u \right) P_{S_v} \right) \supseteq \mathrm{col}(X_{r'}).$$

If the LTCD is non-singular then by Lemma 6.3.11 $\mathrm{col}(X_v) = U_v$ for all $v \in \mathcal{V}$. Hence

$$\bigoplus_{v\in\mathcal{V}_k} U_v = \bigoplus_{v\in\mathcal{V}_k} \mathrm{col}(X_v) = \mathrm{col}\left(\sum_{v\in\mathcal{V}_k} X_v\right) \subseteq \mathrm{col}\left(\sum_{v\in\mathcal{V}_{\geq k}} X_v\right) = \mathrm{col}(X_{r'})$$

where the inclusion again holds by Lemma 6.2.2. This establishes the second part of the statement of the lemma. □

Proof of Theorem 6.5.2

Proof of Theorem 6.5.2. Since the LTCD can be recovered by MTCD, it follows from Prop 6.4.4 that there is a dual certificate Y satisfying

$$Y_{S_vS_v} \preceq \lambda_v I \quad \text{and} \quad Y_{S_vS_v}[X_v]_{S_vS_v} = [X_v]_{S_vS_v} \quad \text{for all } v \in \mathcal{V}. \tag{6.7.16}$$

Let Y' denote the matrix that is block diagonal with respect to $(S_v)_{v\in\mathcal{V}_k}$ with diagonal blocks $Y'_{S_vS_v} = Y_{S_vS_v}$ for all $v \in \mathcal{V}_k$. Then Y' satisfies (6.7.16) for $v \in \mathcal{V}_{<k}$ simply because $Y_{S_vS_v} = Y'_{S_vS_v}$ whenever $v \in \mathcal{V}_{<k}$. It remains to show that Y' satisfies (6.7.16) for $v = r'$, i.e. that

$$Y' \preceq \lambda_k I \quad \text{and} \quad Y'X_{r'} = \lambda_k X_{r'}.$$

The first of these conditions holds because Y' is block diagonal, the blocks satisfy $Y'_{S_vS_v} = Y_{S_vS_v} \preceq \lambda_v I$ for all $v \in \mathcal{V}_k$, and λ_k is (by definition) the common value of λ_v for all $v \in \mathcal{V}_k$.

For the second, by Lemma 6.5.1 and the fact that the LTCD of X is non-singular, we have that $\mathrm{col}(X_{r'}) = \bigoplus_{v\in\mathcal{V}_k} \mathrm{col}(X_v)$. This means that if $x \in \mathrm{col}(X_{r'})$ then $x = \sum_{v\in\mathcal{V}_k} x_v$ where each $x_v \in \mathrm{col}(X_v) \subseteq \mathbb{R}^{S_v}$ for $v \in \mathcal{V}_k$. Then since Y' is block diagonal with respect to $(S_v)_{v\in\mathcal{V}_k}$,

$$Y'x = \sum_{v\in\mathcal{V}_k} Y'x_v = \sum_{v\in\mathcal{V}_k} \pi^*_{S_v}(Y_{S_vS_v})\pi_{S_v}x_v = \sum_{v\in\mathcal{V}_k} \lambda_v x_v = \lambda_k x.$$

Since this holds for any $x \in \mathrm{col}(X_{r'})$ it follows that $Y'X_{r'} = \lambda_k X_{r'}$.

This argument establishes part 1 and part 3 of the statement of Theorem 6.5.2. To establish part 2, we note that since $\mathrm{col}(X_{r'}) = \bigoplus_{v\in\mathcal{V}_k} \mathrm{col}(X_v)$ and $\mathrm{col}(X_v) \subseteq \mathbb{R}^{S_v}$ for $v \in \mathcal{V}_k$ it follows that $P_{\mathrm{col}(X_{r'})}$ is block diagonal with respect to $(S_v)_{v\in\mathcal{V}_k}$ with the diagonal block indexed by S_v being $P_{\mathrm{col}(X_v)}$. Finally by part 1 we have that $P_{\mathrm{col}(Z_{r'})} = P_{\mathrm{col}(X_{r'})}$ so it also has this block diagonal structure. □

Proof of Proposition 6.5.3

Proof of Proposition 6.5.3. By the optimality conditions for MTCD we know that $YX_r = \lambda_r X_r$. Since $\text{rank}(\lambda_r I - Y) + \text{rank}(X_r) = n$ we know that the column space of X_r equals the nullspace of $\lambda_r I - Y$. Since $\lambda_r I - Y$ is block diagonal it follows X_r has a factorization as

$$X_r = P \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_k \end{bmatrix} M \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_k \end{bmatrix}^T P^T$$

where P is a permutation matrix, M is positive definite, and each A_i has columns that are a basis for the nullspace of $\lambda_r I - Y_{S_iS_i}$. Decompose M as the sum of a positive semidefinite block diagonal matrix D (with blocks corresponding to the partition induced by the columns of the A_i) and a rank-deficient positive semidefinite matrix L. One could do this using MTCD, for instance. Then the decomposition

$$\begin{aligned} X_r &= P \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_k \end{bmatrix} (D+L) \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_k \end{bmatrix}^T P^T \\ &= \left(\sum_{i,j=1}^{k} \pi_{S_i}^* A_i L_{ij} A_j^T \pi_{S_j} \right) + \sum_{i=1}^{k} (\pi_{S_i}^* A_i D_{ii} A_i^T \pi_{S_i}) \end{aligned}$$

satisfies the conclusion of the Proposition. □

At the end of Section 6.5.4 we briefly describe how, in the limit as the regularization parameters δ and γ go to zero, the optima of dual regularized approximate LTCD (6.5.5) approach optimal solutions of the dual of MTCD for which h is minimized. The following result gives conditions on an underlying LTCD and on a regularizer h, that ensures that the minimum of h over the dual optimal face is block diagonal.

Lemma 6.7.5. *Suppose $X \succ 0$ is $n \times n$ and has a non-singular LTCD $X = \sum_{v\in\mathcal{V}} X_v$ with structure $(S_v)_{v\in\mathcal{V}}$ that forms a constant depth tree. Suppose, also, that the LTCD can be recovered by MTCD (with scalars $(\lambda_v)_{v\in\mathcal{V}}$ that are positive, order preserving, and constant on scales). Let $h : \mathcal{S}^n \to \mathbb{R}$ be a convex function that satisfies, for each $k > 0$, $h(gYg^T) = h(Y)$ for all $g \in G_{k-1} := G((S_v)_{v\in\mathcal{V}_{k-1}})$. Then for each scale $k > 0$*

$$\min_Y h(Y; (S_v)_{v\in\mathcal{V}_{k-1}}) \quad s.t. \quad Y \text{ is dual optimal}$$

is block diagonal with respect to $(S_v)_{v\in\mathcal{V}_k}$.

Proof. Under these hypotheses, in the proof of Theorem 6.5.2 we showed that if Y is any element of the dual optimal set then the matrix Y' obtained by setting $Y'_{S_vS_w} = \delta_{vw}Y_{S_vS_v}$ for $v, w \in \mathcal{V}_k$ is also in the dual optimal set. Let $G_k = G((S_v)_{v\in\mathcal{V}_k})$ and note that because $(S_v)_{v\in\mathcal{V}_{k-1}}$ is a refinement of $(S_v)_{v\in\mathcal{V}_k}$, G_k is a subgroup of G_{k-1}. By Lemma 6.7.4 we have that

$$Y' = \frac{1}{|G|}\sum_{g\in G_k} gYg^T.$$

Then by the convexity of h and by the fact that $h(gYg^T) = h(Y)$ for all $g \in G_{k-1}$ (and hence all $g \in G_k$), we can conclude that

$$h(Y') \leq \frac{1}{|G|}\sum_{g\in G_k} h(gYg^T) = h(Y).$$

Hence the minimizer of h must be block diagonal with respect to $(S_v)_{v\in\mathcal{V}_k}$. $\square$

Chapter 7

Conclusion

This thesis is focused on giving new descriptions of both optimization-based and probabilistic models. In particular the emphasis is on descriptions that allow problems to be solved (or approximately solved) *globally* and with both computational efficiency (and approximation) guarantees. Semidefinite optimization problems play a central role, either by providing

- exact reformulations (Chapters 3 and 4) for families of optimization problems
- approximations for families of optimization problems with provable approximation guarantees (Chapter 5), or
- methods to search for computationally efficient approximations for certain probabilistic models (Chapter 6).

In Section 7.1, we summarize the main contributions of the thesis. In Section 7.2 we suggest avenues for future work. The discussion in Section 7.2 is intended to complement the concrete problems and suggestions for future work at the end of Chapters 3, 4, 5, and 6, by discussing some broader possible research themes related to the ideas in this thesis.

■ 7.1 Summary of contributions

In this section we summarize the main contributions of the thesis.

Chapter 3: Semidefinite representations of derivative relaxations of spectrahedral cones

The work in Chapter 3 is broadly related to understanding how different families of convex optimization problems are related, both in terms of their expressive power and in terms of the complexity of describing and solving the associated optimization problems. In particular we consider the problem of giving semidefinite representations of a family of hyperbolicity cones called the derivative relaxations of spectrahedral cones. These derivative relaxations form a sequence of outer approximations to a spectrahedral cone. This sequence of approximations starts with a spectrahedral cone, and then successively relaxes the high-dimensional faces of the cone, preserving lower-dimensional faces, until only a half-space containing the original cone is left.

Our main contribution is to construct explicit semidefinite representations of these derivative relaxations. Our representations have size that is polynomial in both the size of the original spectrahedral cone and the relaxation parameter. These are the first known semidefinite representations for this family of convex cones.

Chapter 4: Semidefinite descriptions of the convex hull of rotation matrices

In Chapter 4 we study $n \times n$ rotation matrices (i.e. orthogonal matrices with determinant one) from the point of view of semidefinite optimization. Our main contribution is to give the first semidefinite representations of the convex hull of $n \times n$ rotation matrices. Optimization problems with variables that are constrained to be rotations arise naturally in many application areas. Our semidefinite representations open up the possibility of constructing semidefinite representations and relaxations for a range of problems involving rotations that are typically tackled by local optimization methods. As an example, in Chapter 4 we use our representations to exactly reformulate a joint attitude and spin-rate estimation problem for a spinning spacecraft as a semidefinite optimization problem.

Chapter 5: Rounding semidefinite relaxations for pairwise optimization problems

In Chapter 5 we consider the problem of rounding semidefinite relaxations for a class of pairwise optimization problems that includes natural multivariate extensions of the problems we solve exactly in Chapter 4. We generalize well-known constant-factor approximation guarantees and associated rounding schemes for quadratic optimization over $\{-1, 1\}^n$ to a much larger class of problems than has been previously considered. The problem class we work with replaces binary variables with subsets $\mathcal{X}$ of $m \times d$ matrices such that

- $X^*X \preceq I$ for all $X \in \mathcal{X}$;
- $\mathcal{X}$ has a certain symmetry property;
- we can (approximately) maximize linear functionals over $\mathcal{X}$.

An example of such a set $\mathcal{X}$ is the set of $d \times d$ rotation matrices.

Our main result shows that to design an optimal (in an appropriate sense) rounding scheme for a particular semidefinite relaxation, requires the solution of a geometric optimization problem, the *normalized maximum width problem*, related to the constraint set $\mathcal{X}$. Feasible points for this geometric problem can be translated into rounding schemes. The corresponding objective value is related to the approximation guarantee the rounding scheme achieves.

Chapter 6: A convex approach to learning Gaussian latent tree models

In Chapter 6 we shift from developing tractable descriptions of optimization problems to developing tractable descriptions of multivariate Gaussian random variables. In particular we propose methods to approximate a given covariance matrix as the marginal covariance among the leaf-indexed variables of a Gaussian tree model. Gaussian tree models are a tractable family of models that allow very efficient decentralized inference algorithms. Typical methods for constructing such latent variable models involve local optimization approaches such as the expectation-maximization (EM) algorithm.

The method we propose to construct such tree models is based on semidefinite optimization and is global in nature. The main contribution of Chapter 6 is to analyze our method in the case where we are given the covariance among the leaf-indexed variables of a latent tree model, and we aim to recover the full underlying latent tree model. We give sufficient conditions on the underlying model that ensure our method recovers the model parameters, the dimensions of the latent variables, and the combinatorial structure of the tree.

■ 7.2 Future directions

In this final section we describe some natural avenues for future research related to the work in this thesis.

Breaking symmetry in semidefinite representations

Many of the sets for which we would like to find semidefinite representations have a large symmetry group. The semidefinite representations developed in Chapters 3 and 4 of this thesis, as well as many others in the literature [110], are equivariant in the sense that the semidefinite representations respect the symmetries of the underlying convex set being represented. Restricting to equivariant representations allows us to use systematic approaches (see, e.g., [43, 44]) to construct semidefinite representations. On the other hand, as was discussed briefly in Chapter 4, in some cases it is known that *breaking* symmetry allows for dramatic reductions in the size of semidefinite representations. Developing systematic approaches to reducing the size of semidefinite representations by breaking symmetry is a natural topic for future study.

Many of the representations known that do break symmetry proceed by decomposing the (extreme points of the) convex set to be represented into simpler pieces (in a way that breaks some symmetry), representing the constituent pieces, and then reassembling them to form a representation of the overall object. Understanding how to systematically search for decompositions of a given set that are beneficial in this context seems like one natural approach to systematically breaking symmetry in semidefinite representations.

Semidefinite relaxations for multivariate optimization problems over rotation matrices

Optimization problems with multiple 3×3 rotation-matrix-valued variables are an important family of non-linear non-convex optimization problems. These arise naturally in many estimation and control problems for, say, rigid bodies. The multiple variables could arise, for example, from discretizing time in a dynamic model to obtain a time-series taking values in rotation matrices, or from having multiple static objects of interest, or having multiple such time-series. Our semidefinite representations for the convex hull of 3×3 rotation matrices from Chapter 4 give rise to exact semidefinite representations for a very restricted class of multivariate problems, and can also be used to obtain tighter semidefinite relaxations for a broader class of problems [118].

Nevertheless there has not yet been a systematic approach taken to developing and analyzing methods based on semidefinite optimization for multivariate optimization problems on 3×3 rotation matrices. This is even the case for the simple class of time-series models discussed in [118]. On the theoretical end of the spectrum, it would be useful to find explicit and exact semidefinite reformulations of these multivariate problems (even though, if such representations exist, they are likely of exponential size in the number of variables). Such descriptions could be used to construct more tractable relaxations by systematically weakening some of their constraints. On the other end of the spectrum, developing efficient specialized numerical algorithms to solve semidefinite relaxations for these classes of problems could have significant practical impact. For recent initial work in this direction see, for instance, [16].

Simultaneous design of semidefinite relaxations and rounding schemes

Numerous tools have been developed over the last 10–15 years to construct hierarchies of semidefinite relaxations for polynomial optimization problems. These are essentially based on the basic idea of certifying non-negativity of functions by writing them as sums of squares [92] (or, dually, by constructing relaxations for the moments of probability measures supported on a given set [73]). Since these procedures, generally, only produce relaxations of the convex sets of interest, it is desirable to have associated rounding schemes with good approximation guarantees. Thus far there has been little progress, in general, in systematically developing rounding schemes for relaxations of polynomial optimization problems coming from these hierarchies (beyond the first level of the hierarchies).

One natural (and very broad) problem area for future research would be to develop families of semidefinite relaxations that are *designed with rounding in mind*. Any rounding scheme must map the extreme points of the relaxation into the exact convex hull of the feasible points for the original problem. As such it would be important that for a family of relaxations developed with rounding in mind we have a good characterization of the extreme points of the relaxation. This is typically not the case for hierarchies based on sums-of-squares, for instance.

Convex approaches for hierarchical matrix decomposition problems

One interpretation of the latent tree covariance decomposition problem from Chapter 6 is that it is a hierarchical version of a (block) diagonal and low-rank matrix decomposition problem, where the hierarchical nature comes from additional structure within the low-rank term. In the past five years, convex optimization-based methods for related matrix decomposition problems, such as sparse and low-rank decompositions [29, 27], have been developed, analyzed, and applied to numerous problem areas. It would be interesting to develop hierarchical analogs of matrix decompositions involving low-rank matrices, where the low-rank term has additional hierarchical structure, as well as associated convex optimization-based methods to find such decompositions. The family of matrices that are well approximated by matrices with, say, a hierarchical sparse and low-rank decomposition, is likely significantly larger than those that are well approximated by a single-scale sparse and low-rank decomposition. On the other hand, due to the multi-scale structure one would imagine that the subsequent cost of carrying out computations with a hierarchical approximation would be similar to that of the single-scale approximation. It is likely that many of the issues that arise in the hierarchical diagonal and low-rank decomposition setting of Chapter 6 will be similar to the issues that appear in hierarchical versions of more complex matrix decomposition problems.

Bibliography

[1] A. Horn. "Doubly stochastic matrices and the diagonal of a rotation matrix". In: *American J. Math.* 76.3 (1954), pp. 620–630.

[2] B. P. W. Ames and S. A. Vavasis. "Convex optimization for the planted k-disjoint-clique problem". In: *Mathematical Programming* 143.1-2 (2014), pp. 299–337.

[3] B. P. W. Ames and S. A. Vavasis. "Nuclear norm minimization for the planted clique and biclique problems". In: *Mathematical Programming* 129.1 (2011), pp. 69–89.

[4] A. Anandkumar, K. Chaudhuri, D. J. Hsu, S. M. Kakade, L. Song, and T. Zhang. "Spectral methods for learning multivariate latent tree structure". In: *Advances in Neural Information Processing Systems*. 2011, pp. 2025–2033.

[5] S. Balakrishnan, M. J. Wainwright, and B. Yu. *Statistical guarantees for the EM algorithm: From population to sample-based analysis*. 2014. eprint: arXiv:1408.2156.

[6] A. S. Bandeira, M. Charikar, A. Singer, and A. Zhu. "Multireference alignment using semidefinite programming". In: *Proceedings of the 5th conference on Innovations in theoretical computer science*. ACM. 2014, pp. 459–470.

[7] A. S. Bandeira, C. Kennedy, and A. Singer. *Approximating the little Grothendieck problem over the orthogonal group*. 2013. eprint: arXiv:1308.5207.

[8] A. Barvinok and G. Blekherman. "Convex geometry of orbits". In: *Combinatorial and Computational Geometry*. Vol. 52. Cambridge University Press Cambridge, 2005, pp. 51–77.

[9] A. Barvinok and I. Novik. "A centrally symmetric version of the cyclic polytope". In: *Discrete & Computational Geometry* 39.1-3 (2008), pp. 76–99.

[10] A. Barvinok and A. M. Vershik. "Convex hulls of orbits of representations of finite groups and combinatorial optimization". In: *Funct. Anal. Appl.* 22.3 (1988), pp. 224–225.

[11] M. Basseville, A. Benveniste, and A. S. Willsky. "Multiscale autoregressive processes. I. Schur-Levinson parametrizations". In: *IEEE Transactions on Signal Processing* 40.8 (1992), pp. 1915–1934.

[12] A. Ben-Tal, A. Nemirovski, and C. Roos. "Extended matrix cube theorems with applications to μ-theory in control". In: *Mathematics of Operations Research* 28.3 (2003), pp. 497–523.

[13] D. Bertsimas and R. Mazumder. *Factor analysis via a modern optimization lens.* 2014.

[14] G. Blekherman, P. A. Parrilo, and R. R. Thomas, eds. *Semidefinite Optimization and Convex Algebraic Geometry.* MOS-SIAM Series on Optimization. SIAM, Philadelphia, 2013.

[15] E. D. Bolker. "A class of convex bodies". In: *Transactions of the American Mathematical Society* 145 (1969), pp. 323–345.

[16] N. Boumal. "Optimization and estimation on manifolds". PhD thesis. Université catholique de Louvain, 2014.

[17] P. Brändén. "Hyperbolicity cones of elementary symmetric polynomials are spectrahedral". In: *Optim. Lett.* 8.5 (2014), pp. 1773–1782.

[18] P. Brändén. "Obstructions to determinantal representability". In: *Advances in Mathematics* 226.2 (2011), pp. 1202–1212.

[19] M. Braverman, K. Makarychev, Yu. Makarychev, and A. Naor. "The Grothendieck constant is strictly smaller than Krivine's bound". In: *Forum of Mathematics, Pi.* Vol. 1. Cambridge Univ Press. 2013, pp. 1–42.

[20] J. Briët, H. Buhrman, and B. Toner. "A generalized Grothendieck inequality and nonlocal correlations that require high entanglement". In: *Communications in Mathematical Physics* 305.3 (2011), pp. 827–843.

[21] J. Briët, F. M. Oliveira, and F. Vallentin. "The positive semidefinite Grothendieck problem with rank constraint". In: *Automata, Languages and Programming.* Springer, 2010, pp. 31–42.

[22] D. Bump. *Lie groups.* Springer, 2013.

[23] C. B. Chua. "Relating homogeneous cones and positive definite cones via T-algebras". In: *SIAM J. Optim.* 14.2 (2003), pp. 500–506.

[24] C. Carathéodory. "Über den Variabilitätsbereich der Koeffizienten von Potenzreihen, die gegebene Werte nicht annehmen". In: *Mathematische Annalen* 64.1 (1907), pp. 95–115.

[25] L. Carlone, R. Tron, K. Daniilidis, and F. Dellart. “Initialization techniques for 3D SLAM: a survey on rotation estimation and its use in pose graph optimization”. In: *Proc. IEEE International Conference on Robotics and Automation (ICRA), 2015.* to appear.

[26] R. D. Carr and G. Konjevod. “Polyhedral combinatorics”. In: *Tutorials on emerging methodologies and applications in Operations Research.* Springer, 2005, pp. 2-1–2-48.

[27] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. “Latent Variable Graphical Model Selection via Convex Optimization”. In: *Annals of Statistics* 40.4 (2012), pp. 1935–1967.

[28] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. “The convex geometry of linear inverse problems”. In: *Foundations of Computational Mathematics* 12.6 (2012), pp. 805–849.

[29] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. “Rank-sparsity incoherence for matrix decomposition”. In: *SIAM J. Optim.* 21.2 (2011), pp. 572–596.

[30] Y. Chen, L. J. Guibas, and Q.-X. Huang. *Near-optimal joint object matching via convex relaxation.* 2014. eprint: arXiv:1402.1473.

[31] M.-J. Choi, V. Y. F. Tan, A. Anandkumar, and A. S. Willsky. “Learning latent tree graphical models”. In: *J. Machine Learning Research* 12 (2011), pp. 1771–1812.

[32] M. Conforti, G. Cornuéjols, and G. Zambelli. “Extended formulations in combinatorial optimization”. In: *Ann. Oper. Res.* 204.1 (2013), pp. 97–143.

[33] K. Daoudi, A. B. Frakt, and A. S. Willsky. “Multiscale autoregressive models and wavelets”. In: *IEEE Trans. Info. Theory* 45.3 (1999), pp. 828–845.

[34] A. d’Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. “A direct formulation for sparse PCA using semidefinite programming”. In: *SIAM Rev.* 49.3 (2007), pp. 434–448.

[35] C. Davis. “All convex invariant functions of Hermitian matrices”. In: *Archiv der Mathematik* 8.4 (1957), pp. 276–278.

[36] C. Delorme and S. Poljak. “Combinatorial properties and the complexity of a max-cut approximation”. In: *European Journal of Combinatorics* 14.4 (1993), pp. 313–333.

[37] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *J. Royal Stat. Soc. Ser. B (methodological)* (1977), pp. 1–38.

[38] M. Deza and M. Laurent. *Geometry of cuts and metrics*. Vol. 15. Algorithms and Combinatorics. Springer, 1997.

[39] M. A. Dritschel and J. Rovnyak. “The operator Fejér-Riesz theorem”. In: *A Glimpse at Hilbert Space Operators*. Springer, 2010, pp. 223–254.

[40] È. B. Vinberg. “The theory of homogeneous convex cones”. In: *Trans. Moscow Math. Soc.* 12 (1965), pp. 340–403.

[41] F. R. Harvey. *Spinors and Calibrations*. Vol. 9. Perspectives in Mathematics. Academic Press, Inc., Boston, MA, 1990.

[42] H. Fawzi, J. Gouveia, P. A. Parrilo, R. Z. Robinson, and R. R. Thomas. *Positive semidefinite rank*. 2014. eprint: arXiv:1407.4095.

[43] H. Fawzi, J. Saunderson, and P. A. Parrilo. *Equivariant semidefinite lifts and sum-of-squares hierarchies*. 2013. eprint: arXiv:1312.6662.

[44] H. Fawzi, J. Saunderson, and P. A. Parrilo. *Sparse sum-of-squares certificates on finite abelian groups*. 2015. eprint: arXiv:1503.01207.

[45] M. Fazel. “Matrix rank minimization with applications”. PhD thesis. Stanford University, 2002.

[46] P. W. Fieguth and A. S. Willsky. “Fractal estimation using models on multiscale trees”. In: *IEEE Transactions on Signal Processing* 44.5 (1996), pp. 1297–1300.

[47] A. B. Frakt and A. S. Willsky. “Computationally efficient stochastic realization for internal multiscale autoregressive models”. In: *Multidimensional Systems and Signal Processing* 12.2 (2001), pp. 109–142.

[48] J. Gallier. “The Cartan-Dieudonné Theorem”. In: *Geometric Methods and Applications*. Springer, 2001, pp. 197–247.

[49] L. Gårding. “An inequality for hyperbolic polynomials”. In: *J. Math. Mech* 8.6 (1959), pp. 957–965.

[50] K. Gatermann and P. A. Parrilo. “Symmetry groups, semidefinite programs, and sums of squares”. In: *J. Pure Appl. Algebra* 192.1 (2004), pp. 95–128.

[51] Y. Genin, Y. Hachez, Yu. Nesterov, and P. Van Dooren. “Optimization problems over positive pseudopolynomial matrices”. In: *SIAM J. Matrix Anal. Appl.* 25.1 (2003), pp. 57–79.

[52] A. A. Giannopoulos and V. D. Milman. “Extremal problems and isotropic positions of convex bodies”. In: *Israel Journal of Mathematics* 117.1 (2000), pp. 29–60.

[53] A. A. Giannopoulos, V. D. Milman, and M. Rudelson. "Convex bodies with minimal mean width". In: *Geometric Aspects of Functional Analysis*. Springer, 2000, pp. 81–93.

[54] M. X. Goemans and D. P. Williamson. "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming". In: *Journal of the ACM* 42.6 (1995), pp. 1115–1145.

[55] J. Gouveia and T. Netzer. "Positive polynomials and projections of spectrahedra". In: *SIAM J. Optim.* 21.3 (2011), p. 960.

[56] J. Gouveia, P. A. Parrilo, and R. R. Thomas. "Lifts of convex sets and cone factorizations". In: *Mathematics of Operations Research* 38.2 (2013), pp. 248–264.

[57] A. Grothendieck. "Résumé de la théorie métrique des produits tensoriels topologiques". In: *Boll. Soc. Mat. São Paulo* 8 (1953), pp. 1–79.

[58] O. Güler. "Hyperbolic polynomials and interior point methods for convex programming". In: *Math. Oper. Res.* 22.2 (1997), pp. 350–377.

[59] O. Güler and L. Tunçel. "Characterization of the barrier parameter of homogeneous convex cones". In: *Mathematical Programming* 81.1 (1998), pp. 55–76.

[60] U. Haagerup. "A new upper bound for the complex Grothendieck constant". In: *Israel Journal of Mathematics* 60.2 (1987), pp. 199–224.

[61] U. Haagerup. "The Grothendieck inequality for bilinear forms on C^*-algebras". In: *Advances in Mathematics* 56.2 (1985), pp. 93–116.

[62] Y. Hachez. "Convex optimization over non-negative polynomials: structured algorithms and applications". PhD thesis. Université catholique de Louvain, 2003.

[63] J. W. Helton and J. Nie. "Semidefinite representation of convex sets". In: *Mathematical Programming* 122.1 (2010), pp. 21–64.

[64] J. W. Helton and V. Vinnikov. "Linear matrix inequality representation of sets". In: *Comm. Pure Appl. Math* 60.5 (2007), pp. 654–674.

[65] M. B. Horowitz, N. Matni, and J. W. Burdick. "Convex relaxations of SE(2) and SE(3) for visual pose estimation". In: *Proc. IEEE International Conference on Robotics and Automation (ICRA), 2014*. IEEE. 2014, pp. 1148–1154.

[66] Q.-X. Huang and L. Guibas. "Consistent shape maps via semidefinite programming". In: *Computer Graphics Forum*. Vol. 32. 5. Wiley Online Library. 2013, pp. 177–186.

[67] W. W. Irving. "Multiscale stochastic realization and model identification with applications to large-scale estimation problems". PhD thesis. Massachusetts Institute of Technology, 1995.

[68] W. W. Irving and A. S. Willsky. "A canonical correlations approach to multiscale stochastic realization". In: *IEEE Trans. Automatic Control* 46.10 (2001), pp. 1514–1528.

[69] J. E. Keat. *Analysis of Least-Squares Attitude Determination Routine DOAOP*. Tech. rep. CSC/TM-77/6034. Computer Sciences Corporation, Feb. 1977.

[70] R. E. Kalman. "Identification of noisy systems". In: *Russian Mathematical Surveys* 40.4 (1985), pp. 25–42.

[71] R. M. Karp. *Reducibility among combinatorial problems*. Springer, 1972.

[72] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[73] J. B. Lasserre. "Global optimization with polynomials and the problem of moments". In: *SIAM J. Optim.* 11.3 (2001), pp. 796–817.

[74] P. D. Lax. "Differential equations, difference equations and matrix theory". In: *Communications on Pure and Applied Mathematics* 11.2 (1958), pp. 175–194.

[75] W. Ledermann. "On a problem concerning matrices with variable diagonal elements." In: *Proceedings of the Royal Society of Edinburgh* 60.01 (1940), pp. 1–17.

[76] J. R. Lee, P. Raghavendra, and D. Steurer. *Lower bounds on the size of semidefinite programming relaxations*. 2014. eprint: arXiv:1411.6317.

[77] A. S. Lewis, P. A. Parrilo, and M. V. Ramana. "The Lax conjecture is true". In: *Proc. Amer. Math. Soc.* 133.9 (2005), pp. 2495–2500.

[78] J. Löfberg. "YALMIP: A Toolbox for Modeling and Optimization in MATLAB". In: *Proceedings of the CACSD Conference*. Taipei, Taiwan, 2004. URL: http://users.isy.liu.se/johanl/yalmip.

[79] M. F. Atiyah, R. Bott, and A. Shapiro. "Clifford modules". In: *Topology* 3 (1964), pp. 3–38.

[80] M. Longinetti, L. Sgheri, and F. Sottile. "Convex hulls of orbits and orientations of a moving protein domain". In: *Discrete Comput. Geom.* 43.1 (2010), pp. 54–77.

[81] A. Marcus, D. A. Spielman, and N. Srivastava. "Interlacing families II: Mixed characteristic polynomials and the Kadison-Singer problem". In: *Annals of Mathematics* 182.1 (2015), pp. 327–350.

[82] K. V. Mardia and P. E. Jupp. *Directional statistics*. Vol. 494. John Wiley & Sons, 2009.

[83] T. Myklebust and L. Tunçel. *Interior-point algorithms for convex optimization based on primal-dual metrics*. 2014. eprint: arXiv:1411.2129.

[84] A. Naor, O. Regev, and T. Vidick. "Efficient rounding for the noncommutative Grothendieck inequality". In: *Proceedings of the 45th annual ACM symposium on theory of computing*. ACM. 2013, pp. 71–80. eprint: arXiv:1210.7656.

[85] A. Nemirovski. "Advances in convex optimization: conic programming". In: *Proceedings of the International Congress of Mathematicians: Madrid, August 22–30, 2006: invited lectures*. 2006, pp. 413–444.

[86] A. Nemirovski. "Sums of random symmetric matrices and quadratic optimization under orthogonality constraints". In: *Mathemathcal Programming* 109.2-3 (2007), pp. 283–317.

[87] A. Nemirovski and A. Ben-Tal. *Lectures on Modern Convex Optimization: Analysis, Algorithms and Engineering Applications*. MOS-SIAM Series on Optimization. SIAM, 2001.

[88] Yu. Nesterov. "Semidefinite relaxation and nonconvex quadratic optimization". In: *Optimization methods and software* 9.1-3 (1998), pp. 141–160.

[89] Yu. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*. Vol. 13. SIAM studies in applied and numerical mathematics. SIAM, Philadelphia, 1994.

[90] T. Netzer and R. Sanyal. "Smooth hyperbolicity cones are spectrahedral shadows". In: *Mathematical Programming* (to appear). eprint: arXiv:1208.0441.

[91] J. Nie, P. A. Parrilo, and B. Sturmfels. "Semidefinite Representation of the k-Ellipse". In: *Algorithms in Algebraic Geometry*. Vol. 146. The IMA Volumes in Mathematics and its Applications. Springer New York, 2008, pp. 117–132.

[92] P. A. Parrilo. "Semidefinite programming relaxations for semialgebraic problems". In: *Mathematical Programming* 96.2 (2003), pp. 293–320.

[93] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1997.

[94] R. Pemantle. "Hyperbolicity and stable polynomials in combinatorics and probability". In: *Current Developments in Mathematics, 2011*. Int. Press, Somerville, MA, 2012, pp. 57–123.

[95] G. Pisier. "Grothendieck's theorem, past and present". In: *Bulletin of the American Mathematical Society* 49.2 (2012), pp. 237–323.

[96] S. Poljak and Z. Tuza. “Maximum cuts and large bipartite subgraphs”. In: *Combinatorial Optimization*. Ed. by W. Cook, L. Lovász, and P. D. Seymour. Vol. 20. DIMACS series in Discrete Mathematics and Theoretical Computer Science. American Mathematical Society, 1995.

[97] M. L. Psiaki. “Generalized Wahba Problems for Spinning Spacecraft Attitude and Rate Determination”. In: *J. Astronautical Sciences* 57.1-2 (2009), pp. 73–92.

[98] R. G. Jeroslow. “On defining sets of vertices of the hypercube by linear inequalities”. In: *Discrete Math.* 11.2 (1975), pp. 119–124.

[99] R. Sinn. “Algebraic Boundaries of SO(2)-Orbitopes”. In: *Discrete Comput. Geom.* 50.1 (2013), pp. 219–235.

[100] M. V. Ramana. “Polyhedra, spectrahedra, and semidefinite programming”. In: *Topics in semidefinite and interior-point methods*. Ed. by P. M. Pardalos and H. Wolkowicz. Vol. 18. Fields Institute Communications. American Mathematical Society, 1998, pp. 27–38.

[101] M. Ramana and A. J. Goldman. “Some geometric results in semidefinite programming”. In: *J. Global Optim.* 7.1 (1995), pp. 33–50.

[102] J. Renegar. “Hyperbolic Programs, and Their Derivative Relaxations”. In: *Foundations of Computational Mathematics* 6 (1 2006), pp. 59–79.

[103] J. Renegar and M. Sondjaja. *A polynomial-time affine-scaling method for semidefinite and hyperbolic programming*. 2014. eprint: arXiv:1410.6734.

[104] R. E. Rietz. “A proof of the Grothendieck inequality”. In: *Israel Journal of Mathematics* 19.3 (1974), pp. 271–276.

[105] R. T. Rockafellar. *Convex analysis*. Vol. 28. Princeton University Press, 1997.

[106] M. Rolle. *Démonstration d’une Méthode pour resoudre les Egalitez de tous les degrez*. 1691.

[107] D. J. Rose, R. E. Tarjan, and G. S. Lueker. “Algorithmic aspects of vertex elimination on graphs”. In: *SIAM J. Computing* 5.2 (1976), pp. 266–283.

[108] M. Rosenblatt. “A multi-dimensional prediction problem”. In: *Arkiv för Matematik* 3.5 (1958), pp. 407–424.

[109] R. Salakhutdinov and G. E. Hinton. “Deep Boltzmann Machines”. In: *International Conference on Artificial Intelligence and Statistics*. 2009, pp. 448–455.

[110] R. Sanyal, F. Sottile, and B. Sturmfels. “Orbitopes”. In: *Mathematika* 57.02 (2011), pp. 275–314.

[111] Raman Sanyal. “On the derivative cones of polyhedral cones”. In: *Adv. Geom.* 13.2 (2013), pp. 315–321.

[112] J. Saunderson. "Subspace identification via convex optimization". S.M. thesis. Massachusetts Institute of Technology, 2011.

[113] J. Saunderson, V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. "Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting". In: *SIAM J. Matrix Anal. Appl.* 33.4 (2012), pp. 1395–1416.

[114] J. Saunderson, V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. "Tree-structured statistical modeling via convex optimization". In: *Proc. 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*. IEEE. 2011, pp. 2883–2888.

[115] J. Saunderson and P. A. Parrilo. "Polynomial-sized semidefinite representations of derivative relaxations of spectrahedral cones". In: *Mathematical Programming* (to appear). eprint: arXiv:1208.1443.

[116] J. Saunderson, P. A. Parrilo, and A. S. Willsky. "A convex solution to a joint attitude and spin-rate estimation problem". In: *Journal of Guidance, Control, and Dynamics* (to appear). eprint: arXiv:1410.2841.

[117] J. Saunderson, P. A. Parrilo, and A. S. Willsky. "Semidefinite descriptions of the convex hull of rotation matrices". In: *SIAM J. Optimization* (to appear). eprint: arXiv:1403.4914.

[118] J. Saunderson, P. A. Parrilo, and A. S. Willsky. "Semidefinite relaxations for optimization problems over rotation matrices". In: *Proc. 53rd IEEE Conference on Decision and Control (CDC)*. IEEE. 2014, pp. 160–166.

[119] C. Scheiderer. *Semidefinite representation for convex hulls of real algebraic curves*. 2012. eprint: arXiv:1208.3865.

[120] R. Schneider and W. Weil. "Zonoids and related topics". In: *Convexity and its Applications*. Springer, 1983, pp. 296–317.

[121] A. Schrijver. *Combinatorial optimization: polyhedra and efficiency*. Springer, 2003.

[122] A. Shapiro. "Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis". In: *Psychometrika* 47.2 (1982), pp. 187–199.

[123] A. Singer and Y. Shkolnisky. "Three-dimensional structure determination from common lines in cryo-EM by eigenvectors and semidefinite programming". In: *SIAM Journal on Imaging Sciences* 4.2 (2011), pp. 543–572.

[124] A. M.-C. So, J. Zhang, and Y. Ye. "On approximating complex quadratic optimization problems via semidefinite programming relaxations". In: *Mathematical Programming* 110.1 (2007), pp. 93–110.

[125] C. Spearman. "“General Intelligence,” objectively determined and measured". In: *American J. Psychology* 15.2 (1904), pp. 201–292.

[126] R. P. Stanley. *Enumerative Combinatorics*. Vol. 1. Cambridge University Press, 2011.

[127] K. C. Toh, M. J. Todd, and R. H. Tütüncü. "SDPT3—a MATLAB software package for semidefinite programming, version 1.3". In: *Optim. Methods Softw.* 11.1-4 (1999), pp. 545–581.

[128] R. Tron and R. Vidal. "Distributed image-based 3-d localization of camera sensor networks". In: *Proc. 48th IEEE Conference on Decision and Control, 2009.* IEEE. 2009, pp. 901–908.

[129] L. Tunçel. *Polyhedral and semidefinite programming methods in combinatorial optimization*. Vol. 27. American Mathematical Society, 2010.

[130] L. G. Valiant. "The complexity of computing the permanent". In: *Theoretical computer science* 8.2 (1979), pp. 189–201.

[131] L. Vandenberghe and S. Boyd. "Semidefinite programming". In: *SIAM Rev.* 38.1 (1996), pp. 49–95.

[132] R. Vershynin. *Introduction to the non-asymptotic analysis of random matrices.* 2010. eprint: arXiv:1011.3027.

[133] C. Vinzant. "Edges of the Barvinok–Novik orbitope". In: *Discrete & Computational Geometry* 46.3 (2011), pp. 479–487.

[134] D. Wagner. "Multivariate stable polynomials: theory and applications". In: *Bulletin of the American Mathematical Society* 48.1 (2011), pp. 53–84.

[135] G. Wahba. "A least squares estimate of satellite attitude". In: *SIAM Rev.* 7.3 (1965), pp. 409–409.

[136] D. P. Williamson and D. B. Shmoys. *The design of approximation algorithms.* Cambridge University Press, 2011.

[137] A. S. Willsky. "Multiresolution Markov models for signal and image processing". In: *Proc. IEEE* 90.8 (2002), pp. 1396–1458.

[138] Y.-B. Choe, J. G. Oxley, A. D. Sokal, and D. G. Wagner. "Homogeneous multivariate polynomials with the half-plane property". In: *Adv. Appl. Math.* 32.1–2 (2004), pp. 88–187.

[139] S. Zhang and Y. Huang. "Complex quadratic optimization and semidefinite programming". In: *SIAM J. Optim.* 16.3 (2006), pp. 871–890.

[140] Y. Zinchenko. "On hyperbolicity cones associated with elementary symmetric polynomials". In: *Optim. Lett.* 2.3 (2008), pp. 389–402.