UNIVERSITY OF CALIFORNIA
Santa Barbara

Quadratic Convexity and Sums of Squares

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Mathematics

by

Martin Ames Harrison

Committee in Charge:

Professor Mihai Putinar, Chair

Professor Carlos Javier Garcia-Cervera

Professor Kenneth Ralph Goodearl

December 2013

The Dissertation of
Martin Ames Harrison is approved:

Professor Carlos Javier Garcia-Cervera

Professor Kenneth Ralph Goodearl

Professor Mihai Putinar, Committee Chairperson

October 2013

Quadratic Convexity and Sums of Squares

by

Martin Ames Harrison

Acknowledgements

I thank Professor Mihai Putinar for the guidance and encouragement he has provided me over the last five years. I thank my committee members Professor Carlos Garcia-Cervera and Professor Ken Goodearl for several helpful discussions. I thank Medina Price for her inexhaustible patience and empathy. Finally, I thank Akshat Kumar for doing mathematics with me.

Curriculum Vitæ

Martin Ames Harrison

Education

BS, Mathematics, California Polytechnic State University-San Luis Obispo, 2007

MA, Mathematics, University of California-Santa Barbara, 2009

PhD, Mathematics, University of California-Santa Barbara, 2013 (expected)

Positions

Teaching Assistant, University of California-Santa Barbara, 2007-2013

Teaching Associate, University of California-Santa Barbara, Summer 2009

Publication

"Pfister's Theorem Fails in the Free Case", *Mathematical Methods in Systems, Optimization and Control*, pp. 189-194, Birkhäuser Basel, 2012.

Fields of Study

Functional Analysis and Convex Geometry

Advisor: Mihai Putinar

Abstract

Quadratic Convexity and Sums of Squares

Martin Ames Harrison

The *length* of a sum of squares σ in a ring R is the smallest natural k such that σ can be realized as a sum of k squares in R. For a set $S \subseteq R$, the *pythagoras number* of S, denoted by $\mathcal{P}(S)$, is the maximum value of length over all $\sigma \in S$. This dissertation is motivated by the following simple question: if $R = \mathbb{R}[x_1, \dots, x_n]$ and $S = \mathbb{R}[x_1, \dots, x_n]_{2d}$ (the span of forms of degree $2d$), then what is $\mathcal{P}(S)$?

By parametrizing the set of sums of k squares, we obtain a new formulation of the problem: when is the image $A(\mathbb{R}^N)$ of a quadratic map $A : \mathbb{R}^N \to \mathbb{R}^M$ convex? We prove several results on the structure of quadratic images and of the set of quadratic maps in general. In particular, we give conditions under which convexity of $A(\mathbb{R}^N)$ is equivalent to convexity of its compact intersection with an affine hyperplane. We prove, given assumptions on $A(\mathbb{R}^N)$, a relationship between convexity of $A(\mathbb{R}^N)$ and rank of the derivative of A. We then show how an arbitrary quadratic map A can be modified so that convexity of $A(\mathbb{R}^N)$ is preserved and the result on the derivative may be exploited. A necessary condition for *quadratic convexity* is thus derived.

With that necessary condition we develop a method to determine the smallest natural k for which there is an open subset of $\Sigma_{n,d}$ with pythagoras number k. This result yields a lower bound on $\mathcal{P}(\mathbb{R}[x_1, \ldots, x_n]_{2d})$ that is exact in all cases for which the value is known. Finally, other approaches based on measure and convex optimization are described.

Contents

List of Figures

Chapter 1

Introduction

Over the past few decades, advances in real algebraic geometry and convex optimization have led to efficient algorithms for computing (approximate) global extrema of polynomials. The key observation is that a computationally accessible criterion for positivity provides the means to find absolute minima of polynomial functions restricted to certain sets. Concisely, *Positivstellensätze* yield semidefinite programming problems via the equation

$$\inf\{p(x) \mid x \in C\} = \sup\{y \mid (p - y)|_C > 0\},$$

which holds even when p is not bounded below on C (recall that $\sup \emptyset = -\infty$).

The cone of sums of squares of polynomials (*SOS* cone) is a proper subset of the cone of nonnegative polynomials. In cases of low dimension and degree, however, the two sets differ by little or even coincide. In general, the results alluded to above enable one to express nonnegative polynomials in terms of sums of squares. The SOS cone is therefore of great practical importance, but it is also an interesting object of study in its own right.

Questions of quadratic convexity arise in the relaxation of systems of quadratic equations and in quantum control. The joint numerical range of a pair of matrices is famously convex, but for larger tuples this is not the case. Some progress has been made toward a general characterization of convexity of joint numerical range, but current results give conditions which are necessary or sufficient, but not both.

The remainder of this chapter outlines the history of these subjects, introduces the terminology and conventions to be used, and provides an overview of the content of this dissertation.

1.1 Background

Hilbert's 17^{th} problem is to determine whether every positive polynomial can be expressed as a sum of squares of *rational* functions; Hilbert already knew that not every positive polynomial is a sum of squares, although it was not until Motzkin gave an explicit form in 1957 that a concrete example was known. Artin developed his theory of real closed fields to prove that all positive polynomials are in fact sums of squares of rational functions, thus settling Hilbert's question (in [1]).

A more practical but equally difficult challenge is to test positivity of a given polynomial. It took Tarski's substantial contributions in logic to show that this problem is decidable (see [40]). Positivstellensätze (theorems describing positive polynomials) were proved by Krivine ([17]), and later in independent work of Stengle ([39]). More recently, results of Schmüdgen and Putinar (see [37] and [30], respectively) and others have given characterizations of positivity amenable to semidefinite programming. Nesterov and Nemiroski, building on earlier work in linear programming, initiated in [24] the adaptation of interior point methods to semidefinite programming, which has myriad applications ranging from graph theory to control engineering. Work by Parrilo and others demonstrate this, and there are several refined software packages (including [29], [22] and [15]) implementing these algorithms.

Further study of the PSD and SOS cones themselves, as geometric objects, has been carried out by Blekherman, who proved that there are, in terms of the Lebesgue measures of compact "slices", significantly more positive polynomials than sums of squares. This gap, however, is either absent or negligible in many of the instances arising from practical applications.

The question of computing Pythagoras numbers in the abstract has been studied by Scheiderer (see [36] for this and a general discussion of positivity and sums of squares). Reznick, Choi and Lam, in [7], set upper and lower bounds by careful analysis of certain convex sets associated to the monomials appearing in forms. A remark on length and pythagoras number appears in [5].

On the topic of quadratic convexity, considerably less progress has been made. A result of Hausdorff and Toeplitz states that the joint numerical range of a pair of matrices is convex (see [2] for a proof of the *Toeplitz-Hausdorff Theorem*). Other authors have described sufficient conditions which are not necessary (see [9],[10],[28], and [20]). Ramana gave a somewhat superficial characterization in [31] and proved that the problem is NP-Hard. He was approaching from the angle of relaxing quadratic systems of equations, which is closely related to the motivation for this dissertation. Barvinok, in his book [2], gives some interesting examples and provides a bound on rank in an arbitrary spectrahedron, which is essentially equivalent to the upper bound established in [7]. His more recent work employs ideas from information theory in bounding the distance from the image of a quadratic map to an arbitrary point in its convex hull. Finally, problems in quantum control have led to investigation of the convexity of joint numerical range of tuples of Hermitian matrices (see [9], [21]).

1.2 Notation and Preliminaries

In this section we establish the notational conventions to be followed, define some of the terms to be used, and record some standard results which will be invoked directly in proofs or referenced in later discussion.

As mentioned above, $\mathbb{R}[x_1, \ldots, x_n]_{2d}$ will denote the real vector space of homogeneous polynomials of degree $2d$ (including the zero polynomial), which is spanned by monomials of degree $2d$ in the indeterminates $x_1, \ldots, x_n$. For any multi-index $\alpha = (\alpha_1, \ldots, \alpha_n) \in (\mathbb{N} \cup \{0\})^n$, we denote by x^α the monomial $x_1^{\alpha_1} \cdots x_n^{\alpha_n}$. The degree of x^α is $|\alpha| \equiv \sum_i \alpha_i$. Note that this is unambiguous because the monomials are in one-to-one correspondence with multi-indices. Throughout this document, we use graded lexicographical order on the set of multi-indices; other orderings would suffice, but we need to fix one explicitly. We denote by $\Sigma_{n,d}$ the subset of $\mathbb{R}[x_1, \ldots, x_n]_{2d}$ consisting of sums of squares. Additionally, for $k \in \mathbb{N}$ we write

$$\Sigma_{n,d}(k) = \{\sigma \mid \exists g_1, \ldots, g_k \in \mathbb{R}[x_1, \ldots, x_n]_d \text{ such that } \sigma = g_1^2 + \ldots + g_k^2\}.$$

The next two terms deserve emphasis.

Definition 1.2.1. The *length* of $\sigma \in \Sigma_{n,d}$ is the smallest natural number k for which $\sigma \in \Sigma_{n,d}(k)$. We write $l(\sigma)$ for the length of σ.

Definition 1.2.2. The *pythagoras number* of a set $S \subseteq \mathbb{R}[x_1, \dots, x_n]_{2d}$, denoted by $\mathcal{P}(S)$, is the maximum of the set $\{l(\sigma) \mid \sigma \in S\}$.

Because we have restricted degree, the pythagoras numbers of all sets to be considered are finite (see discussion following 1.2.7). But if degree is not restricted, then forms of arbitrarily large length can be constructed.

The term *SOS* will be abused slightly, serving as an adjective ("σ is SOS"), and as a noun denoting a set ("...faces of SOS"). We will see in the next chapter how SOS can be realized as the linear image of a special set of matrices.

We now turn to matrices and convexity. The algebra of $N \times N$ matrices over $\mathbb{R}$ is denoted by $M_N(\mathbb{R})$. The set $\mathcal{S}_N \subseteq M_N(\mathbb{R})$ consists of the symmetric matrices. If $A \in \mathcal{S}_N$ has only nonnegative eigenvalues, then we write $A \in \mathcal{S}_N^+$ and $A \succeq 0$, and we call A *positive semidefinite*. If $A \in \mathcal{S}_N^+$ has only positive eigenvalues, then we call A *positive definite* and write $A \succ 0$. Equivalently, A is positive definite if and only if $x^T A x > 0$ for all $x \neq 0$. The set of positive definite matrices is the interior of the set of positive semidefinite matrices (with respect to the standard topology on $\mathbb{R}^{\binom{N+1}{2}}$), and is denoted by $\mathcal{S}_N^{++}$. It will be convenient to have a name for the intersection of $\mathcal{S}_N^+$ with the affine subspace $\{x \in M_N(R) \mid \text{trace}(x) = 1\}$; we call it K_N^+, and remark that it is a first nontrivial example of the *spectrahedra* of semidefinite programming to be defined later. We record below a standard result describing positive semidefinite (PSD) matrices.

Theorem 1.2.3. *(Characterization of PSD Matrices) For $A \in \mathcal{S}_N$, the following are equivalent:*

i) A is positive semidefinite.

ii) $A = B^2$ for some $B \in \mathcal{S}_N$.

iii) $x^T Ax \geq 0$ for all $x \in \mathbb{R}^N$.

iv) The principal minors of A are nonnegative.

v) $A = LL^T$ for some $L \in M_N(\mathbb{R})$.

There is a bit more to say about condition *iv*), pertaining to rank, which will appear in later chapters. Most of the sets introduced so far share a special property called *convexity.*

Definition 1.2.4. A subset S of a real vector space is called *convex* if for all $t \in [0, 1]$ and all $x, y \in S$, the linear combination $tx + (1 - t)y$ belongs to S.

In general, a linear combination $t_1x_1 + \ldots + t_kx_k$, where the x_i are vectors and the t_i are nonnegative reals summing to 1, is called a *convex combination.* If C is a convex set, then a convex set $F \subseteq C$ is a *face* of C if for all $t \in (0, 1)$ and for all $x, y \in C$, $tx + (1 - t)y \in F$ implies $x, y \in F$. An *extreme point* is a face consisting of a singe point. A face $F \subseteq C$ is *exposed* if it is "cut out" by a hyperplane, i.e. if there is a linear functional Λ bounded above by $\ell \in \mathbb{R}$ on C such that $F = C \cap \{x \mid \Lambda(x) = \ell\}$. In polyhedra, of course, all faces are exposed faces. While this is not true in general (see Figure 1.1), some of the familiar properties of polyhedra extend.

Theorem 1.2.5. *(Partitioning a Convex Set) Let C be a nonempty convex set, and let U be the collection of all relative interiors of nonempty faces of C. Then U is a partition of C, i.e. the sets in U are disjoint and their union is C.*

The above statement is an excerpt of the one appearing in [34], where a proof is given. The next result generalizes a property trivially possessed by polytopes (see [2] for a proof).

Theorem 1.2.6. *(Krein-Milman for Euclidean Space) If $K \subset \mathbb{R}^d$ is a compact convex set, then K is the convex hull of its extreme points.*

The *convex hull* $\mathrm{conv}(S)$ of a set S is the smallest convex set containing S. It is equal to the set of all convex combinations of elements of S. The *conic hull* $\mathrm{co}(S)$ is defined to be the set of all *conic combinations* of elements of S: $\mathrm{co}(S) = \{t_1 x_1 + \ldots + t_k x_k \mid x_i \in S, t_i \geq 0\}$. In general, a *cone* is any subset of a real vector space closed under nonnegative scalar multiplication. For example, the sets $\mathcal{S}_N^+$ and $\Sigma_{n,d}$ are convex cones. The problem of finding pythagoras numbers is therefore just a special case of a general problem in convex geometry described below.

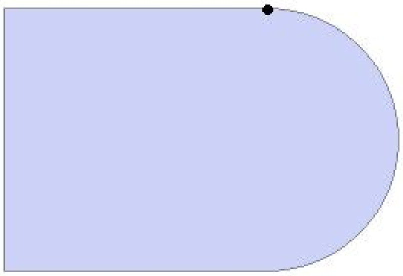

Figure 1.1: The emphasized point is extreme but not exposed.

As already mentioned, the convex hull of a set $S \subset \mathbb{R}^n$ consists of convex combinations $t_1 x_1 + \ldots + t_k x_k$, but what is k? More precisely: given a set S, what is the smallest k for which everything in $conv(S)$ is a convex combination of no more than k points in S? It is not immediately obvious that there is such a k, but a theorem of Carathéodory gives the best upper bound possible given its hypotheses (see Figure 1.2).

Theorem 1.2.7. *(Carathéodory Convex Hull Theorem) If $S \subset \mathbb{R}^d$, then every element of $conv(S)$ may be expressed as a convex combination of $d+1$ or fewer elements of S*

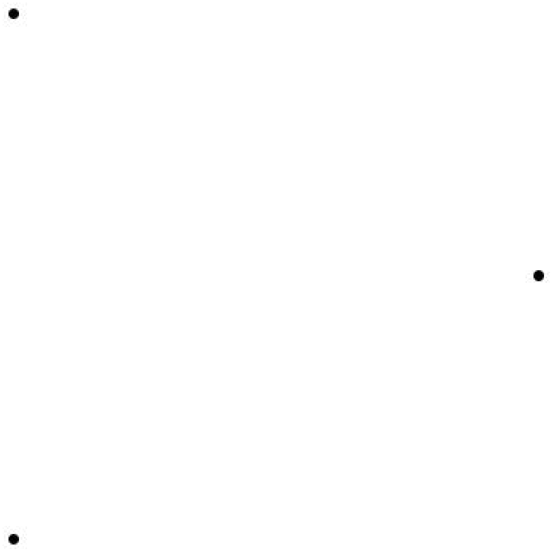

Figure 1.2: A set requiring $d+1$ summands

While the bound given by Carathéodory's theorem is in this sense sharp, it is too crude for our purposes. We can say much more about convex hulls of specific sets for which we have explicit descriptions. For now we have that the pythagoras number of $\Sigma_{n,d}$ itself is no more than $\binom{n+2d-1}{2d}+1$, since $\dim \mathbb{R}[x_1, \ldots, x_n]_{2d}$ can be shown by a "stars and bars" counting argument to be $\binom{n+2d-1}{2d}$.

We conclude with some intuitive results which will be useful in the remaining chapters, all of which can be found in [2].

Theorem 1.2.8. *Let $S \subset \mathbb{R}^m$ be a convex set. If $\mathring{S} = \emptyset$, then S is contained in a proper affine subspace of* $\mathbb{R}^m$

The next will be used to show that $\Sigma_{n,d}$ is closed. See discussion preceding 2.2.3 for a definition of *base.*

Theorem 1.2.9. *If $C \subset \mathbb{R}^m$ is a cone with a compact base, then C is closed.*

The last has several variants used extensively in functional analysis. We will use it in Chapter 3.

Theorem 1.2.10. *(Separating Hyperplane Theorem) If $C \subset \mathbb{R}^m$ is a closed convex set and $x \notin C$, then there is an affine hyperplane strictly separating x and C, that is, there exists a linear functional Λ on $\mathbb{R}^m$ and a real number r such that $\Lambda(y) < r$ for all $y \in C$ and $\Lambda(x) > r$.*

1.3 Overview

In the next chapter we discuss some well-known properties of the SOS cone and explain in detail its connection with the PSD cone. We survey relevant work and summarize what is currently known about pythagoras numbers. In addition, we explore noncommutative sums of ‘squares’, and show how a theorem of Pfister fails to extend into this setting. Chapter 3 contains the main contribution of this dissertation and deals with the more general problem of quadratic convexity.

Related results from the literature are included. After reducing the cases to be considered, we prove the main result which is in turn used to derive lower bounds on pythagoras numbers. In addition, we employ the *Approximate Carathéodory Theorem* in bounding the distance from the sums of k squares to an arbitrary sum of squares. We conclude with Chapter 4, where other approaches are discussed which might be used to characterize quadratic convexity.

Chapter 2

Sums of Squares and PSD Matrices

We now explain the connection between PSD matrices and polynomials. We will see how the problem of computing length may be recast as a rank minimization problem. The first two sections are intended as a survey of standard techniques and well-known results. The chapter concludes with some work in the noncommutative setting.

2.1 PSD, SDP and Rank

Let us fix $p = \sum_\alpha p_\alpha x^\alpha \in \Sigma_{n,d}$. There are forms $g_1, \ldots, g_k \in \mathbb{R}[x_1, \ldots, x_n]_d$ such that $p = g_1^2 + \ldots + g_k^2$. In fact, any g_i satisfying this equation must be homogeneous of degree exactly d because the largest and smallest monomials occurring with nonzero coefficients in the g_i will contribute nonzero terms to the sum $g_1^2 + \ldots + g_k^2$. For each i, we may express g_i as $\sum_\alpha g_{i,\alpha} x^\alpha$. Defining $\mathbf{g_i} = (g_{i,\alpha})^T_{|\alpha|=d} \in \mathbb{R}^N$ and $\mathbf{m} = (x^\alpha)^T_{|\alpha|=d} \in \mathbb{R}[x_1, \ldots, x_n]_d^N$, where $N = \dim \mathbb{R}[x_1, \ldots, x_n]_d$, we get $g_i^2 = \mathbf{g_i}^T \mathbf{m}\mathbf{m}^T \mathbf{g_i} = \text{trace}(\mathbf{g_i}\mathbf{g_i}^T \mathbf{m}\mathbf{m}^T)$, and therefore $p = \text{trace}(\mathbf{m}\mathbf{m}^T (\sum_i \mathbf{g_i}\mathbf{g_i}^T))$. This equation is really a system of $\binom{n+2d-1}{2d}$ equations of the form $p_\alpha = \text{trace}(A_\alpha \sum_i \mathbf{g_i}\mathbf{g_i})$, where A_α is the $N \times N$ matrix, indexed by multi-indices, whose entries are given by

$$(A_\alpha)_{\beta,\gamma} = \begin{cases} 1, & \text{if } \beta + \gamma = \alpha \\ \\ 0, & \text{otherwise} \end{cases}$$

With these matrices we may now define the *spectrahedron associated to* p:

$$S(p) \equiv \{X \in S_N^+ \mid \text{trace}(A_\alpha X) = p_\alpha \text{ for all } \alpha \text{ with } |\alpha| = 2d\}.$$

In general, a *spectrahedron* is a set of the form $S_m^+ \cap V$, where V is an affine subspace of $\mathbb{R}^m$. In the case of $S(p)$, the set V is the codimension $\binom{n+2d-1}{2d}$ (since the A_α are disjointly supported and therefore comprise a linearly independent set) affine subspace

$$\{X \in M_N(\mathbb{R}) \mid \text{trace}(A_\alpha X) = p_\alpha \text{ for all } \alpha \text{ with } |\alpha| = 2d\}$$

An element of $S(p)$ is called a *Gram matrix* for p. It was demonstrated in the discussion above that each decomposition $g_1^2 + \ldots + g_k^2$ of p into a sum of squares yields an element of the spectrahedron $S(p)$, namely the matrix $\sum_i \mathbf{g_i g_i}$. It is also true that every element of $S(p)$ yields a decomposition of p into a sum of squares. To see how, fix $X \in S(p)$ and apply 1.2.3 to obtain

$$X = L^T L,$$

where $L \in M(\mathbb{R})_N$. For each $1 \leq i \leq N$, we now define f_i to be the polynomial in the i^{th} entry of $L\mathbf{m}$. Since $X = L^T L \in S(p)$, we get $p = \sum_i f_i^2$. This yields a dramatic improvement of the bound obtained from 1.2.7, but the improved bound does not generally give the length of a sum of squares. In most cases, the Gram matrix is not unique. In fact, the set $S(p)$ is generally not finite. Below is an example illustrating this non-uniqueness, followed by a characterization of the cases in which $S(p)$ is a singleton.

Example 2.1.1. (An explicit spectrahedron) Take $n = 2$ and $d = 2$. We then have $\binom{2+4-1}{4} = 5$ matrices $A_\alpha \in \mathcal{S}_3$ which describe $p \in \Sigma_{n,d}$: $A_{(4,0)} = \left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right)$, $A_{(3,1)} = \left(\begin{smallmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right)$, $A_{(2,2)} = \left(\begin{smallmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{smallmatrix}\right)$, $A_{(1,3)} = \left(\begin{smallmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{smallmatrix}\right)$, and $A_{(0,4)} = \left(\begin{smallmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right)$. If we take $p(x, y) = x^4 + x^2y^2 + y^4$, then $X = \left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right)$ belongs to $S(p)$, but $S(p)$ is in fact a line segment in $\mathcal{S}_3^+$ which we'll now describe.

Note that $X \longmapsto (\text{trace}(A_\alpha X))_{|\alpha|=4}$ defines a surjective linear map from $\mathcal{S}_3$ to $\mathbb{R}^5$. The kernel of this map is therefore spanned by a single matrix in $\mathcal{S}_3$. One such matrix is $X_0 \equiv \left(\begin{smallmatrix} 0 & 0 & 1 \\ 0 & -2 & 0 \\ 1 & 0 & 0 \end{smallmatrix}\right)$. We conclude that the set $S(p)$ consists of exactly the linear combinations $X + tX_0$ which are PSD. Invoking 1.2.3 again we find that $X + tX_0 \succeq 0$ exactly when $-1 \leq t \leq \frac{1}{2}$. It is interesting to note that each of the two boundary values yields a Gram matrix of rank 2. It is easily verified that 2 is the minimum value of rank on the set $S(p)$. More can be said; we will see in 2.3 that $\mathcal{P}(\Sigma_{2,d}) = 2$ for any d. ■

There is much in the preceding example to be fleshed out and generalized. Recalling the linear transformation defined above, we extend to arbitrary n and d in order to understand when it happens that $S(p)$ consists of the unique Gram matrix for p.

Proposition 2.1.2. *For any $n, d \in \mathbb{N}$, the linear transformation $T : \mathcal{S}_N \to \mathbb{R}^M$ defined by $T(X) = (\operatorname{trace}(A_\alpha X))_{|\alpha|=2d}$, where $N = \binom{n+d-1}{d}$ and $M = \binom{n+2d-1}{2d}$, is surjective. The transformation T is a bijection if and only if $d = 1$ or $n = 1$.*

Proof. There are exactly M of the A_α. Furthermore, we have $\operatorname{trace}(A_\alpha A_\beta) = 0$ for $\alpha \neq \beta$ from the definition. It follows that the map T is surjective.

Now we assume that $d = 1$ and let $n \in \mathbb{N}$ be arbitrary. In this case, we have $\dim \mathcal{S}_N = \binom{N+1}{2} = \binom{n+1}{2} = \binom{n+2-1}{2}$. By the rank-nullity theorem, T must be a bijection. If we take $n = 1$, then both spaces are $1-$dimensional and T is a bijection.

Conversely, suppose that $d \geq 2$ and $n \geq 2$. Define an equivalence relation $\sim$ on the set of pairs of monomials in $\mathbb{R}[x_1, \ldots, x_n]_d$ by

$$(u, v) \sim (w, z) \iff uv = wz.$$

This partitions the set of pairs into M disjoint equivalence classes (this will be justified rigorously in the next section), but some of these classes are not singletons. Consider for instance the monomial $x_1^d x_2^d = (x_1^{d-1} x_2)(x_1 x_2^{d-1})$, and note that because $d - 1 > 0$, these are really two distinct decompositions. □

It follows from this proposition and the convexity of $S(p)$ that it is finite only in the simple cases $d = 1$ and $n = 1$. And we obtain another description of this set in terms of the kernel of the map T above.

Proposition 2.1.3. *If $p \in \Sigma_{n,d}$ and T is as above, then for any X_p satisfying $T(X_p) = (p_\alpha)_{|\alpha|=2d}$ we have*

$$S(p) = \{X \in \mathcal{S}_N^+ \mid X - X_p \in \ker T\}$$

We also saw in 2.1.1 that the length of our particular p was obtained on the boundary of $S(p)$, at Gram matrices of smallest possible rank. This holds in general.

Proposition 2.1.4. *If $p \in \Sigma_{n,d}$, then $\ell(p) = \min\{\mathrm{rank}(X) \mid X \in S(p)\}$.*

Proof. Let $X \in S(p)$. Since X is symmetric, the spectral theorem yields a decomposition $U^T DU = X$, where U is orthogonal and D diagonal, with the eigenvalues λ_i of X along the diagonal. If we denote by U_i the i^{th} row of U, then we have

$$p = \text{trace}(\mathbf{m}\mathbf{m}^T X) = \text{trace}(\mathbf{m}^T U^T DU\mathbf{m}) = \sum_i \lambda_i (U_i \mathbf{m})^2$$

where $\mathbf{m}$ is the monomial vector introduced in the discussion preceding 2.1.1. The rank of X is exactly the number of nonzero eigenvalues λ_i, which in turn is the number of squares in the sum. □

This observation does not make our task easier. In fact, it is shown in [41] that rank minimization is NP-Hard. Heuristics exist for rank minimization which succeed in many cases, as demonstrated in [32] and [6]. They succeed sometimes. The work of Recht, Fazel and others ([32]) does not help us here because we are not concerned with restricted isometries. But 2.1.1 does in fact give an example where trace minimization produces solutions of minimum rank.

We just saw a rudimentary instance of semidefinite programming which could be done by hand. Formally, a *semidefinite program* (SDP) is an optimization problem of the form

$$\begin{aligned} \text{minimize} \quad & \operatorname{trace}(CX) \\ \text{subject to} \quad & \operatorname{trace}(A_iX) = b_i \\ & X \succeq 0, \end{aligned}$$

where C and A_i are square matrices and b_i are real numbers. The formulation above is called the *primal* SDP. The *dual* to this problem is:

$$\begin{aligned} \text{maximize} \quad & a^T y \\ \text{subject to} \quad & \sum_i y_i A_i + S = C \\ & S \succeq 0. \end{aligned}$$

In both of these optimization problems, the feasible region is a spectrahedron (or at least isometrically isomorphic to one). Because we can lift any functional to a functional on a higher dimensional space, the feasible regions of SDPs are exactly linear images of spectrahedra. In the next section, we show that the set $\Sigma_{n,d}$ is exactly such an image. In fact, certain other sets of polynomials related to $\Sigma_{n,d}$ can be realized as such, and therefore optimization of linear functionals over those sets can be done via SDP.

The *quadratic module* generated by $g_1, \ldots, g_m \in \mathbb{R}[x_1, \ldots, x_n]$, denoted $Q(g_1, \ldots, g_m)$, is the set of all sums of the form $\sum_{i=0}^{m} g_i \sigma_i$, where $g_0 \equiv 1$ and each σ_i is a sum of squares. The k^{th} *truncated quadratic module* is the set

$$Q_k(g_1, \ldots, g_m) = \Big\{ \sum_{i=0}^{m} g_i \sigma_i \mid \sigma_i \text{ is SOS and } \deg(g_i\sigma_i) \leq 2k \text{ for every } i \Big\}.$$

A truncated qudratic module is also a projected spectrahedron, since it is an image under a linear transformation of another projected spectrahedron, namely $\Sigma_{n,d}$ (recall 2.2.2). Under additional assumptions on the polynomials $g_1, \ldots, g_m$, it has been shown (in [18], for instance) that the optima

$$\lambda_k = \max \lambda \text{ s.t } f - \lambda \in Q_k(g_1, \ldots, g_m)$$

converge to the minimum of the polynomial f on the set

$$\{x \mid g_i(x) \geq 0 \text{ for all } i\}.$$

2.2 Properties of the SOS Cone

We present some important properties of the SOS cone and justify the claims made in the proof of 2.1.2. In some of the results below, polynomials are identified with their coefficient vectors in $\mathbb{R}^M$; the monomials are identified with the standard basis vectors for $\mathbb{R}^M$, and we use the euclidean topology.

Proposition 2.2.1. *The set $\Sigma_{n,d}$ is a convex cone of full dimension in $\mathbb{R}[x_1, \ldots, x_n]_{2d}$.*

The idea of the proof is to show that $\Sigma_{n,d}$ is not contained in a proper subspace of $\mathbb{R}[x_1, \ldots, x_n]_{2d}$. Since it is convex, this implies that it has nonempty interior.

Proof. The convexity of $\Sigma_{n,d}$ is trivial: if $g_1^2 + \ldots + g_k^2$, $f_1^2 + \ldots + f_\ell^2 \in \Sigma_{n,d}$, and $t \in [0, 1]$, then $t(g_1^2 + \ldots + g_k^2) + (1 - t)(f_1^2 + \ldots + f_\ell^2) =$

$$(\sqrt{t}g_1)^2 + \ldots + (\sqrt{t}g_k)^2 + (\sqrt{1-t}f_1)^2 + \ldots + (\sqrt{1-t}f_\ell)^2 \in \Sigma_{n,d}.$$

It remains to show that the set of squares of polynomials in $\mathbb{R}[x_1, \ldots, x_n]_d$ is contained in no proper affine subspace of $\mathbb{R}^M$. Together with 1.2.8, this will prove the claim. We prove that the squares span $\mathbb{R}^M$ (rather, their coefficient vectors do).

Let $x^\alpha \in \mathbb{R}[x_1, \ldots, x_n]_{2d}$. Now construct β and γ as follows: if α_k is even, then we set $\beta_k = \gamma_k = \frac{1}{2}\alpha_k$. If α_k is odd, then we set one of β_k and γ_k equal to $\alpha_k - 1$ and the other to $\alpha_k + 1$. Since $2d$ is even, *the number of odd α_ks must be even.* We can therefore assign half of the decremented values $(\alpha_k - 1)$ and half of the incremented values $(\alpha_k + 1)$ to each of β and γ to obtain x^β and x^γ in $\mathbb{R}[x_1, \ldots, x_n]_d$ satisfying $x^\alpha = x^\beta x^\gamma$. Finally, we compute

$$x^\alpha = \frac{1}{4}(x^\beta + x^\gamma)^2 - \frac{1}{4}(x^\beta - x^\gamma)^2,$$

which completes the proof. □

The next proposition will be used in Chapter 4 to give an alternative formulation of the problem.

Proposition 2.2.2. *The set $\Sigma_{n,d}$ is a projected spectrahedron.*

Proof. Recall the map from 2.1.2. In the remarks preceding 2.1.1, we proved that $\Sigma_{n,d} = T(\mathcal{S}_N^+)$. $\square$

Before stating 2.2.3, we need a new term. A convex subset K of a convex cone C is called a *base* if for every nonzero $x \in C$ there is exactly one positive $t \in \mathbb{R}$ for which $tx \in K$.

Proposition 2.2.3. *The set $\Sigma_{n,d}$ has a compact base.*

Proof. Define $K = \{p \in \Sigma_{n,d} \mid \int_{[0,1]^n} p = 1\}$, where $\int_{[0,1]^n}$ is the integral with respect to Lebesgue measure over the unit cube in $\mathbb{R}^n$. We show that K is a compact base for $\Sigma_{n,d}$. If $p \in \Sigma_{n,d}$ is nonzero, then $\int_{[0,1]^n} p > 0$. This follows from the fact that a nonzero polynomial is a continuous function nonzero almost everywhere; since p is positive at some point, p is positive on an open subset of $[0,1]^n$. Defining t by $t \int_{[0,1]^n} p = 1$, we get $tp \in K$. Uniqueness of t is obvious.

It remains to show that K is compact and convex. Convexity of K follows easily from linearity of $\int_{[0,1]^n}$ as a linear functional. To prove that K is compact, we realize K as the image of a compact set under a continuous map. Define

$$K' = \left\{p \in \mathbb{R}[x_1, \ldots, x_n]_d \mid \|p\|_2 = 1\right\}^N \times \Delta^{N-1},$$

and $f : (\mathbb{R}[x_1, \ldots, x_n]_d)^N \times \Delta^{N-1} \to \mathbb{R}[x_1, \ldots, x_n]_{2d}$ by $f(p_1, \ldots, p_N, t_1, \ldots, t_N) = \sum_j t_j p_j^2$, where Δ^{N-1} is the standard simplex in $\mathbb{R}^N$, and $\| \cdot \|_2$ is the L^2 norm. The set K is equal to $f(K')$ by construction, and is therefore compact. □

A cone with a compact base is called *proper*. The terms is used in other senses, but this will do for all cases handled here. In Chapter 3 we will realize 2.2.3 as a special case of a general theorem about quadratic maps; we could find a compact base by constructing a positive definite linear combination of the A_α.

Here is one last property of the cone $\Sigma_{n,d}$ itself which we will meet again, but in more generality, in the next chapter.

Corollary 2.2.4. *The set $\Sigma_{n,d}$ is closed.*

Proof. This follows from 2.2.3 and 1.2.9. □

Example 2.2.5. Here is a concrete example of a compact base for $\Sigma_{n,d}$. Taking $n = d = 2$, with notation as in 2.1.1, we note that the linear combination

$$B \equiv A_{(4,0)} + \tfrac{1}{2}A_{(3,1)} + \tfrac{1}{3}A_{(2,2)} + \tfrac{1}{4}A_{(1,3)} + \tfrac{1}{5}A_{(0,4)}$$

is positive definite (since B is a truncated Hankel matrix of Lebesgue measure on the unit interval). Define

$$C = \{X \in \mathcal{S}_3^+ \mid \operatorname{trace}(XB) = 1\},$$

and $K = T(C)$. Since C is a compact (by positive definiteness of B) convex set and T is linear, K must be a compact convex set. Suppose that $p \in \Sigma_{n,d}$ is not zero. Then there exists $X_p \in \mathcal{S}_3^+$ such that $T(X_p) = p$. Since $X_p \succeq 0$ and $B \succ 0$, we have trace$(X_p B) > 0)$. Thus, $\frac{1}{\text{trace}(X_p B)} p \in K$. To see that this is the unique positive multiple of p belonging to K, note that whenever $X \in S(p)$ we must have trace$(A_\alpha(X - X_p)) = 0$ for all α, so that trace$(B(X - X_p)) = 0$. That is, the functional trace$(B\cdot)$ is constant on $S(p)$. ■

Finally, we mention the *Motzkin form*, a nonnegative polynomial which is not SOS.

Example 2.2.6. Define $M(x, y, z) = x^2y^4 + x^4y^2 + z^6 - 3x^2y^2z^2$. That M is nonnegative follows from the arithmetic-geometric mean inequality. Using 2.3.2, it can be shown that M is not SOS. To see how, let us examine the *Newton polytope* C(M) of M.

$$C(M) = \text{conv}\left\{ \left(\begin{smallmatrix} 2 \\ 4 \\ 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} 4 \\ 2 \\ 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} 0 \\ 0 \\ 6 \end{smallmatrix}\right), \left(\begin{smallmatrix} 2 \\ 2 \\ 2 \end{smallmatrix}\right) \right\},$$

so that if $M = g_1^2 + \ldots + g_k^2$ we must have for each j

$$C(g_j) \subset \text{conv}\left\{ \left(\begin{smallmatrix} 1 \\ 2 \\ 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} 2 \\ 1 \\ 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} 0 \\ 0 \\ 3 \end{smallmatrix}\right) \right\}.$$

Since $\left(\begin{smallmatrix} 1 \\ 2 \\ 0 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 2 \\ 1 \\ 0 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 1 \\ 1 \\ 1 \end{smallmatrix}\right)$, and $\left(\begin{smallmatrix} 0 \\ 0 \\ 3 \end{smallmatrix}\right)$ are the only integer points in $\frac{1}{2}C(M)$ (see Figure 2.1), the assumption that $M = g_1^2 + \ldots + g_k^2$ implies that $x^2y^2z^2$ must appear with a positive coefficient in M. Since this is not the case, we conclude that M is not a sum of squares. ■

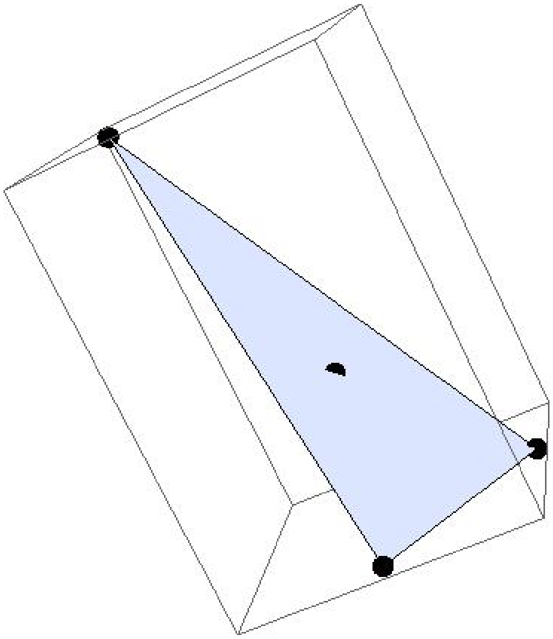

Figure 2.1: $\frac{1}{2}C(M)$. Emphasized points are integer lattice points.

Combining the example above with 2.2.4, we conclude that for some values of n and d the cone of nonnegative polynomials in $\mathbb{R}[x_1, \ldots, x_n]_{2d}$ contains an open ball disjoint from $\Sigma_{n,d}$. It is natural to ask how much "smaller" $\Sigma_{n,d}$ is. One way to compare the two cones is to compare corresponding compact bases - nonnegative polynomials integrating to 1 on the unit cube also comprise a compact base for the nonnegative cone. In [4], Blekherman showed that if degree is fixed and the number of indeterminates allowed to grow, then the base of SOS shrinks to an arbitrarily small proportion of the base of the nonnegative forms.

2.3 Bounds on $\mathcal{P}(\mathbb{R}[x_1, \ldots, x_n]_{2d})$

It is known that $\mathcal{P}(\Sigma_{2,d}) = 2$, that $\mathcal{P}(\Sigma_{n,1}) = n$ and that $\mathcal{P}(\Sigma_{3,2}) = 3$. For arbitrary n and d there are explicit upper and lower bounds on $\mathcal{P}(\Sigma_{n,d})$ which generally do not coincide. As we saw in Chapter 2, it holds for all n and d that $\mathcal{P}(\mathbb{R}[x_1, \ldots, x_n]_{2d}) \leq \dim \mathbb{R}[x_1, \ldots, x_n]_d$. Better bounds are given in [7], where the authors use the method of *cages* - a generalization of the Newton polytope.

Definition 2.3.1. For a form $p = \sum_\alpha p_\alpha x^\alpha \in \mathbb{R}[x_1, \ldots, x_n]_{2d}$, we define the *Newton polytope* of p, denoted $C(p)$, to be the convex hull of the set $\{\alpha \mid p_\alpha \neq 0\}$ (recall that the α belong to $\mathbb{R}^n$).

A property of the Newton polytope used in [7] (and proved in [33]) to obtain bounds on the the pythagoras numbers of certain subsets of $\mathbb{R}[x_1, \ldots, x_n]_{2d}$ also provides an easy way to show that the Motzkin form is not SOS. While eminently plausible, this result is not trivial. In particular, it is much stronger than the observation that $\alpha = \frac{1}{2}(2\alpha)$.

Theorem 2.3.2. *If $f = g_1^2 + \ldots + g_k^2$, then $C(g_i) \subset \frac{1}{2}C(f)$ for each $i = 1, \ldots, k$.*

Let a denote the dimension of $\mathbb{R}[x_1, \ldots, x_n]_{2d}$, and e that of $\mathbb{R}[x_1, \ldots, x_n]_d$. Here are the bounds given in [7]:

Theorem 2.3.3.

$$\frac{2e+1-\sqrt{(2e+1)^2-8a}}{2} \leq \mathcal{P}(\mathbb{R}[x_1, \ldots, x_n]_{2d}) \leq \frac{\sqrt{1+8a}-1}{2}$$

The lower bound is sharp in many cases, but the upper bound agrees in only a few cases. The upper bound can be obtained from a bound on rank in spectrahedra presented in [2]. We now turn to sums of squares of *noncommutative* polynomials to see how their lengths behave differently.

2.4 The NC Setting and Pfister's Theorem

In this section we present related work on sums of squares of noncommutative polynomials originally published in [11], reused here with kind permission of Springer Science+Business Media. First, we must define new objects and mention some noncommutative results. We work in the real free $*$-algebra $\mathbb{R}\langle X, X^* \rangle$ generated by the n noncommuting (NC) variables $X_1, \ldots, X_n$ and their *adjoints* X_j^*. We can think of these variables as linear operators, and the $*$ function on $\mathbb{R}\langle X, X^* \rangle$ as the adjoint operation. In particular, $*$ respects addition and multiplication by scalars and is defined on monomials by $(X_{j_1} \cdots X_{j_k})^* = X_{j_k}^* \cdots X_{j_1}^*$ and $(X_j^*)^* = X_j$. We use multi-indices α, tuples of non-negative integers from

0 to $2n$, to index monomials: $X^\alpha \equiv X_{\alpha_1} X_{\alpha_2} \cdots X_{\alpha_k}$. The word $X^\emptyset$ is simply the empty word, denoted by 1. For $0 < j \leq n$, we define $X_{j+n} \equiv X_j^*$. Conjugation and concatenation of multi-indices α and β is defined by the equations $X^{\alpha^*} = (X^\alpha)^*$ and $X^{\alpha \circ \beta} = X^\alpha X^\beta$.

Evaluation of $p \in \mathbb{R}\langle X, X^* \rangle$ at a tuple $(M_1, \ldots, M_n)$ of square matrices of common size is defined by the substitution of M_j for X_j and M_j^T for X_j^*.

We say that $p \in \mathbb{R}\langle X, X^* \rangle$ is symmetric when $p^* = p$. Such a polynomial p is said to be *matrix positive* if the matrix $p(M)$ is positive semidefinite (or *PSD*) for every tuple M of square matrices. It was shown by Helton in [13] that every matrix positive polynomial is a sum of squares. The minimal number of squares required to express a matrix positive polynomial as a sum of squares is not known in general, although upper bounds are easy to obtain.

Optimization in certain quantum physics problems is done over feasible regions of operators on Hilbert spaces, and so NC variables are useful there. Several examples and a general framework for such problems are presented in [26], where the semidefinite programming relaxations of Lasserre are extended to the NC setting. Motivation for the study of NC polynomials from control theory is discussed in [12].

To any symmetric polynomial $p \in \mathbb{R}\langle X, X^* \rangle$ we can associate a real, symmetric matrix M with the property

$$V^*MV = p$$

where $V^* = (X^{\alpha^*})_{|\alpha|\leq d}$, and V is the column vector $(X^\alpha)^T_{|\alpha|\leq d}$ (with the monomials in graded lexicographical order). The matrix M is not unique, in fact the set of all such matrices (for a fixed p) forms an affine space which we will denote $\mathcal{M}_p$.

By the rank of p, we mean the minimum of rank(M) over all $M \in \mathcal{M}_p$. For a positive polynomial, this minimum is to be taken over only the PSD matrices. The following lemma helps us obtain a lower bound on rank.

Lemma 2.4.1. *If A is a symmetric matrix satisfying $V^*AV = 0$, then the $(2n)^d \times (2n)^d$ lower right submatrix of A is the zero matrix.*

Proof. Let B denote the block in question, and $\hat{V}$ the tautological vector of just the monomials of degree d. Then $V^*AV = 0$ implies that $\hat{V}^*B\hat{V} = 0$ as well since the product $\hat{V}^*B\hat{V}$ yields exactly the degree $2d$ terms of the polynomial V^*AV. But the entries of B are exactly the coefficients of the distinct monomials in $\hat{V}^*B\hat{V}$, hence B is the zero matrix. □

The lemma above shows that there is no freedom in choosing the block corresponding to the degree $2d$ terms of the polynomial. Since the rank of this block gives a lower bound on the rank of the whole matrix, taking the block to be the the $(2n)^d \times (2n)^d$ identity yields a polynomial with rank at least $(2n)^d$.

As already mentioned, the cone of positive polynomials properly contains the SOS cone (recall the Motzkin form). In contrast, the NC setting offers the nice result, proved by Helton in [13], that any positive polynomial is a sum of squares. Here, a square takes the form f^*f, so that a sum of squares is positive in the sense defined above. Just as in the commutative case, we have a simple relationship between SOS and the PSD cones.

Lemma 2.4.2. *A polynomial p is matrix positive exactly when it can be expressed $p = V^*MV$, with M a PSD matrix.*

The proof is straightforward. It follows that the *length* of p is

$$\begin{aligned} \min \quad & \text{rank } X \\ \text{s.t. } & V^*XV = p, \\ & X \succeq 0, \end{aligned}$$

just as in the commutative setting.

As a simple example of this problem consider the polynomial

$$P = 1 + X^*X + XX^*,$$

clearly a sum of squares. The polynomial P is a sum of 3 squares, but can be expressed as a sum of 2 squares (and no fewer). To see why we parametrize the affine space $\mathcal{M}_P$ by the single parameter $t \in \mathbb{R}$. As usual $V = (1, X, X^*)^T$, and so $P = V^*V = X^*X + XX^* + 1$. Defining

$$M = \left(\begin{smallmatrix} 0 & 1 & -1 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \end{smallmatrix}\right),$$

we get $\mathcal{M}_P = \{I + tM \mid t \in \mathbb{R}\}$, and find the minimal SOS representation

$$P = \left(X + \frac{\sqrt{2}}{2}\right)^* \left(X + \frac{\sqrt{2}}{2}\right) + \left(X^* - \frac{\sqrt{2}}{2}\right)^* \left(X^* - \frac{\sqrt{2}}{2}\right)$$

on the boundary of the region where $I + tM \succ 0$. Note that in this example trace is constant on $\mathcal{M}_P \cap PSD$, and that the given solution is obtained by maximizing t over $\{t \mid I + tM \succeq 0\}$.

Pfister's Theorem (proved in [25]) gives a bound on the number of rational functions in the SOS representation of a PSD polynomial. The bound is remarkable because it does not depend on the degree of the polynomial in question. D'Angelo and Lebl proved in [8] that this result fails for Hermitian polynomials. We'll show that it fails for noncommutative polynomials. The first theorem below is needed for the second. It is easy to check that the polynomial S below has length $(2n)^d$, but more is true.

Theorem 2.4.3. *Suppose that $q \in \mathbb{R}\langle X, X^* \rangle$ and define $S = \sum_{|\alpha|=d} X^{\alpha^*} X^{\alpha}$. Then $p = q^*Sq$ has length at least $(2n)^d$. Here, $(2n)^d$ is the dimension of span$\{X^\alpha\}_{|\alpha|=d}$.*

Proof. Since p is matrix positive, it is a sum of squares, and so we may write $p = V^*MV$, appending V with the necessary monomials. Let q be such that $q^*Sq = p$, and write $q = \sum_\alpha q_\alpha X^\alpha$. Let $\hat{\alpha}$ be maximal, with respect to lexicographical ordering, among all α such that $q_\alpha \neq 0$.

We have $V^*MV = p = q^*Sq = q^*(\sum_{|\alpha|=d} X^{\alpha^*}X^\alpha)q = \sum_{|\alpha|=d}(X^\alpha q)^*(X^\alpha q)$. For each α, write $X^\alpha q = Q_\alpha V$, where Q_α is the row vector of the coefficients of $X^\alpha q$. Forming the matrix Q whose rows are the Q_α we get $V^*MV = p = V^*Q^*QV$, hence $V^*(M - Q^*Q)V = 0$.

The polynomials $X^\alpha q$ form a linearly independent set, and in fact have the distinct leading terms $q_{\hat{\alpha}} X^{\alpha \circ \hat{\alpha}}$. It follows that the last $(2n)^{d+deg(q)}$ columns of Q form a block of rank at least $(2n)^d$. Writing Q in block form $Q = \begin{bmatrix} A & B \end{bmatrix}$ where B is a $(2n)^d \times (2n)^{d+deg(q)}$ matrix, we compute

$$p = V^*Q^TQV = V^* \begin{bmatrix} A^T \\ B^T \end{bmatrix} \begin{bmatrix} A & B \end{bmatrix} V = V^* \begin{bmatrix} A^TA & A^TB \\ B^TA & B^TB \end{bmatrix} V.$$

The V above includes all monomials up to degree $(2n)^{d+deg(q)}$. Since $V^*(M - Q^*Q)V = 0$, we know from the lemma that M cannot differ from Q^*Q in its $(2n)^{d+deg(q)} \times (2n)^{d+deg(q)}$ lower right block; this block equals B^TB. Therefore M, an arbitrary matrix representation for p, has rank at least $(2n)^d$. □

Alternatively, one might ask whether a Pfister's Theorem holds for products of the usual form. Consider what it would take for q^*qS to be a SOS. Because q^*q is symmetric, we note that since SOS are symmetric we must have $q^*qS = (q^*qS)^* = Sq^*q$, so that q^*q and S commute. Since we evaluate these polynomials on tuples of matrices, it is tempting to treat them as symmetric matrices. In particular, one might guess that if two of them commute, then they are both polynomials in a third polynomial. This happens to be true, and it follows from the following more general theorem from combinatorics.

Theorem 2.4.4. *(Bergman's Centralizer Theorem) Let K be a field, and $K\langle X\rangle$ the ring of polynomials over K in noncommuting variables $X_1, \ldots, X_n$. Then the centralizer of a nonscalar element in $K\langle X\rangle$ is isomorphic to $K[t]$ for a single variable t.*

The proof is a bit lengthy and can be found in [23]. It uses the fact that such a centralizer is integrally closed in its field of fractions together with an easier result in the formal series setting.

Theorem 2.4.5. *(Cohn's Centralizer Theorem) Let K be a field and $K\langle\langle X\rangle\rangle$ the ring for formal power series over K in noncommuting variables $X_1, \ldots, X_n$. Then the centralizer of a nonscalar element in $K\langle\langle X\rangle\rangle$ is isomorphic to $K[t]$ for a single variable t.*

These theorems apply despite the superficial difference that we are working with indeterminates $X_1, \ldots, X_n, X_1^*, \ldots, X_n^*$ for which $(X_i^*)^* = X_i$; there are no *polynomial* relations among them, and so we can regard them as $2n$ non-commuting variables $Y_1, \ldots, Y_{2n}$. Armed with 2.4.4, we are ready to give the counterexample.

Theorem 2.4.6. *If $p \in \mathbb{R}\langle X, X^* \rangle$, a matrix positive polynomial, is of the form q^*qS with $S = \sum_{|\alpha|=d} X^{\alpha^*} X^{\alpha}$, then $l(p) \geq (2n)^d$.*

Proof. We will use the previous 2.4.3 together with Bergman's Centralizer Theorem. The main difficulty lies in showing that under the hypotheses, q^*q is actually a polynomial in S.

Invoking the centralizer theorem we write $q^*q = f(h(X, X^*))$ and $S = g(h(X, X^*))$ for $h(X, X^*) \in \mathbb{R}\langle X, X^* \rangle$ and $f(t), g(t) \in \mathbb{R}[t]$. It follows from the equation $S = g(h(X, X^*))$ that g must have degree 1. To see why, write

$$h(X, X^*) = c_1 X^{\alpha_1} + \ldots + c_l X^{\alpha_l} + (\textit{lower degree terms}), \qquad g(t) = a_k t^k + \ldots + a_0$$

with $c_j, a_i \in \mathbb{R}$. We note that each term $X^{\alpha_{j_1}} \cdots X^{\alpha_{j_k}}$ is symmetric since it must be one of the monomials $X^{\alpha^*} X^{\alpha}$ in S. Supposing $k > 1$, we have always that $\alpha_{j_1} = \alpha_{j_k}^*$. This implies that there is just one α_j, which is certainly not the case. Therefore $\deg(g) = 1$ and we write $g(t) = at + b$ so that $S = g(h(X, X^*)) = ah(X, X^*) + b$ or $h(X, X^*) = 1/a(S - b)$.

Now we have $q^*q = f(1/a(S - b)) = r(S)$ for some polynomial $r(t) \in \mathbb{R}[t]$. Since $r(S)$ has length equal to 1(it can be expressed as a single noncommutative square), it follows that $r(t)$ is of even degree. If not, write $r(t) = r_{2k+1}t^{2k+1} + \ldots + r_0$ with $r_{2k+1} \neq 0$. Then $r(S) = r_{2k+1}S^{2k+1} + (\textit{lower degree terms})$ and we have by 2.4.3 that $S^{2k+1} = S^k S S^k$ and therefore $r(S)$ itself has length at least $(2n)^d > 1$, a contradiction. Finally, $tr(t)$ has odd degree and therefore another application of 2.4.3 lets us conclude that $p = Sr(S)$ has length at least $(2n)^d$. $\square$

Chapter 3

Quadratic Convexity

3.1 The Basic Problem

Given a tuple $A = (A_1, \ldots, A_M)$ of matrices $A_j \in \mathcal{S}_N$, we are interested in determining whether the sets $A(\mathbb{R}^N) \equiv \{(x^T A_1 x, \ldots, x^T A_M x) \mid x \in \mathbb{R}^N\}$ and $A(S^{N-1}) \equiv \{(x^T A_1 x, \ldots, x^T A_M x) \mid x \in S^{N-1}\}$ are convex. We will show how the problems are related to each other, and we will derive necessary conditions for convexity of $A(\mathbb{R}^N)$ and $A(S^{N-1})$. First, we mention some related work.

In Chapter 1 we mentioned *joint numerical range*, defined for tuple $A = (A_1, \ldots, A_M)$ of $N \times N$ Hermitian matrices to be the set

$$w(A) = \{(z^* A_1 z, \ldots, z^* A_M z)^T \mid z \in \mathbb{C}^N \text{ and } \|z\| = 1\},$$

where $z^* \equiv (\bar{z}_1, \ldots, \bar{z}_N)$ and $\|z\|^2 \equiv \sum_j \bar{z}_j z_j$ for any vector $z \in \mathbb{C}^N$. It was shown in [9] that $w(A)$ is convex under certain conditions on the lead eigenvalue of combinations $\sum_j \eta_j A$ with $\|\eta\| = 1$, though the converse is false (as is demonstrated in [9]). A simpler sufficient condition is the simultaneous diagonalizability of the A_j or, equivalently, the condition that all pairs commute. This appears in [2] as an exercise, and in [3] alongside results relating boundary points of $w(A)$ to the joint spectrum of A.

But the joint numerical range is just the image of a sphere under a special kind of quadratic map. Indeed, the map $z \mapsto (z^* A_1 z, \ldots, z^* A_M z)$ can be regarded as a quadratic map from $\mathbb{R}^{2n}$ to $\mathbb{R}^{2M}$ by identifying $\mathbb{C}$ with $\mathbb{R}^2$. Under this identification, the set $\{z : \|z\| = 1\}$ is of course identified with $S^{2N-1} \subset \mathbb{R}^{2N}$.

3.2 Quadratic Maps and a Necessary Condition

We begin with a theorem which enables us to modify a quadratic map A in such a way that convexity (or lack thereof) is preserved, but the hypotheses of our main theorem are satisfied. A *quadratic map* from $\mathbb{R}^N$ to $\mathbb{R}^M$ is a function given by $x \mapsto (x^T A_1 x, \ldots, x^T A_M x)$, where the A_j, for all $j = 1, \ldots, M$, are symmetric matrices. If $A = (A_1, \ldots, A_M)$ is a tuple of symmetric matrices, we will write $A(x)$ for the vector $(x^T A_1 x, \ldots, x^T A_M x)^T \in \mathbb{R}^M$, and abuse notation by calling A itself a quadratic map.

Theorem 3.2.1. *Suppose $A : \mathbb{R}^N \to \mathbb{R}^M$ is a quadratic map. Then there is a linear operator $T : \mathbb{R}^M \to \mathbb{R}^k$ satisfying*

i) $(T \circ A)(\mathbb{R}^N)$ *is contained in no proper subspace of* $\mathbb{R}^k$*, and*

ii) $(T \circ A)(\mathbb{R}^N)$ *is convex if and only if* $A(\mathbb{R}^N)$ *is convex.*

Proof. Take U to be the orthogonal projection onto the span of $A(\mathbb{R}^N)$, and V to be a linear isometry from that span to $\mathbb{R}^k$, where $k = \dim \operatorname{span} A(\mathbb{R}^N)$. Set $T = VU$. Then property *i)* is immediate. For property *ii)*, note that $A(\mathbb{R}^N) = U(A(\mathbb{R}^N)) \simeq T(A(\mathbb{R}^N))$ □

While the above observation is fairly straightforward, it is slightly less obvious that the analogous thing can be done in the case of $A(S^{N-1})$.

Theorem 3.2.2. *Suppose $A : \mathbb{R}^N \to \mathbb{R}^M$ is a quadratic map. Then there is a linear operator $T : \mathbb{R}^M \to \mathbb{R}^k$ satisfying*

i) $(T \circ A)(S^{N-1})$ *is contained in no proper affine subspace of* $\mathbb{R}^k$*, and*

ii) $(T \circ A)(S^{N-1})$ *is convex if and only if* $A(S^{N-1})$ *is convex.*

Proof. Define the linear transformation $L : \mathcal{S}_N \to \mathbb{R}^M$ by $L(X)_i = \operatorname{trace}(A_i X)$ for all $i = 1, \ldots, M$, and let W be the subspace $L(\ker(\operatorname{trace}))$ of $\mathbb{R}^M$. Now take U to be the orthogonal projection onto W and suppose that F is an affine subspace of

W containing $U(A(S^{N-1}))$. Then F also contains the affine hull of $U(A(S^{N-1}))$, which is exactly the set $UL(\{X \in \mathcal{S}_N \mid \operatorname{trace}(X) = 1\}) = W \simeq \mathbb{R}^{\dim W}$. Once again, take T to be the composition of U with a linear isometry from W to $\mathbb{R}^{\dim W}$.

To establish property *ii)*, we simply show that the two sets $U(A(S^{N-1}))$ and $A(S^{N-1})$ are translates of each other. Fix $p \in S^{N-1}$ and note that

$$-A(p) + A(S^{N-1}) = L(-pp^t) + L(\{xx^t : x \in S^{n-1}\}) = L(\{xx^t - pp^t : x \in S^{n-1}\})$$
$$= UL(\{xx^t - pp^t : x \in S^{n-1}\} = -U(A(p)) + U(A(S^{n-1})).$$

□

We reiterate that the above theorems are relevant because they enable to modify an arbitrary quadratic map so that the hypotheses of our two main results are satisfied. While the the convexity of $A(S^{N-1})$ and $A(\mathbb{R}^N)$ are not equivalent, there is an implication in one direction.

Proposition 3.2.3. *Suppose that $A : \mathbb{R}^N \to \mathbb{R}^M$ is a quadratic map. Let B denote the unit ball $\{x \in \mathbb{R}^N \mid \|x\| \leq 1\}$ in $\mathbb{R}^N$. Then $A(\mathbb{R}^N)$ is convex whenever $A(B)$ is convex, and $A(B)$ is convex whenever $A(S^{N-1})$ is convex.*

Proof. Let A be as above and suppose that $A(B)$ is convex. Fix $x_1, x_2 \in \mathbb{R}^N$, and $t \in [0, 1]$ and assume $0 \neq \|x_1\| \geq \|x_2\|$. Then $(1/\|x_1\|)x_j \in B$ for $j = 1, 2$, and by convexity of $A(B)$ there exists $x \in B$ such that

$$A(x) = tA\Big(\frac{1}{\|x_1\|}x_1\Big) + (1-t)A\Big(\frac{1}{\|x_1\|}x_2\Big).$$

Multiplying this equation by $\|x_1\|^2$ completes the proof of the first implication (since $A(cv) = c^2A(v)$ for all $c \in \mathbb{R}$ and all $v \in \mathbb{R}^N$).

Now assume that $A(S^{N-1})$ is convex, and fix $x_1, x_2 \in B$, and $t \in [0,1]$. Assume that both x_j are nonzero (the alternative being a trivial case). There exists $x \in S^{N-1}$ such that

$$A(x) = \frac{t\|x_1\|^2}{t\|x_1\|^2 + (1-t)\|x_2\|^2}A\Big(\frac{1}{\|x_1\|}x_1\Big) + \frac{(1-t)\|x_2\|^2}{t\|x_1\|^2 + (1-t)\|x_2\|^2}A\Big(\frac{1}{\|x_2\|}x_2\Big).$$

Multplying by $t\|x_1\|^2 + (1-t)\|x_2\|^2$ we find

$$A\Big(\sqrt{t\|x_1\|^2 + (1-t)\|x_2\|^2}x\Big) = tA(x_1) + (1-t)A(x_2).$$

Since $x \in S^{N-1}$ and $\sqrt{t\|x_1\|^2 + (1-t)\|x_2\|^2} \leq 1$, we have shown that $A(B)$ is convex. □

Now we give two theorems regarding quadratic maps whose hypotheses generalize the properties of quadratic maps arising from parametrization of sums in $\Sigma_{n,d}$. The first reduces a broad class of quadratic maps to positive definite quadratic maps, and yields a corollary which can be seen as a partial converse to 3.2.3. This can be combined with the *Approximate Carathéodory Theorem* to bound the distance of elements in the convex hull to the original image. The second gives a simple necessary condition for quadratic convexity which yields, through recent results in convex algebraic geometry, an algebraic necessary condition for quadratic convexity.

Theorem 3.2.4. *Suppose $A : \mathbb{R}^N \to \mathbb{R}^M$ is a quadratic map whose image spans $\mathbb{R}^M$. Then the following are equivalent*

i) $0 \notin A(K_N^+) \equiv \operatorname{conv} A(S^{N-1})$

ii) There is a linear functional ℓ on $\mathbb{R}^M$ such that $\ell \cdot A \equiv \sum_j \ell_j A_j$ is positive definite.

iii) There is an invertible linear operator T on $\mathbb{R}^M$ such that $T \circ A$ is positive definite in each coordinate.

Proof. Supposing that $0 \notin \operatorname{conv} A(S^{N-1})$, we will show that $A(\mathbb{R}^N)$ is the conic hull of a compact convex set not containing the origin. If $L : \mathcal{S}_N \to \mathbb{R}^M$ is the linear transformation defined as in 3.2.2, then $\operatorname{conv} A(\mathbb{R}^N) = L(\mathcal{S}_N^+)$, and $L(K_N^+) = \operatorname{conv} A(S^{N-1})$, the last equality following from the spectral theorem. Furthermore, $L(\mathcal{S}_N^+)$ is the conic hull of the compact set $L(K_N^+)$, which is disjoint from $\{0\}$ by assumption. By 1.2.10, there is a linear functional ℓ separating 0 from $L(K_N^+)$, and we may assume that $\ell(X) > 1$ for all $X \in K_N^+$. The functional ℓ corresponds naturally to a vector with entries $\ell_j, 1 \leq j \leq M$. Let $\ell \cdot A$ denote the linear combination $\sum_j \ell_j A_j$. To prove that this combination is positive definite we note that for any nonzero vector $x \in \mathbb{R}^N$

$$x^T(\ell \cdot A)x = \|x\|^2 \operatorname{trace}(\tfrac{1}{\|x\|^2} x x^T (\ell \cdot A)) > \|x\|^2 > 0,$$

which says that $\ell \cdot A$ is positive definite.

We now prove that the second condition implies the third. Suppose that $\ell = (\ell_1, \ldots, \ell_M)^T$ as above is given. Since $\ell \cdot A$ belongs to the interior of $\mathcal{S}_N^+$, and since the linear map $y \mapsto y \cdot A$ is continuous, there is an open ball $B \subset \mathbb{R}^M$ centered at ℓ such that $y \cdot A \succ 0$ for all $y \in B$. But any open subset of $\mathbb{R}^M$ spans $\mathbb{R}^M$, and so we may choose a basis of vectors from B. If T is defined as left multiplication on $\mathbb{R}^M$ by the matrix whose rows are given by the coordinate vectors of the elements of this basis, then T is a linear operator satisfying the third condition.

Finally, we prove that the third implies the first. Suppose that T is as above. Then $T(\operatorname{conv} A(S^{N-1}))$ is contained in the strictly positive orthant, and therefore does not contain 0. By linearity of T, $\operatorname{conv} A(S^{N-1})$ itself does not contain 0. □

The conditions of 3.2.4 are a strengthening of a necessary condition for convexity:

Theorem 3.2.5. *If $A : \mathbb{R}^M \to \mathbb{R}^N$ is a quadratic map and $A(\mathbb{R}^N) \neq \mathbb{R}^M$ is convex, then there is a linear functional ℓ on $\mathbb{R}^M$ such that $\ell \cdot A \succeq 0$.*

Proof. This follows directly from a variant of 1.2.10 and the observation that 0 belongs to the boundary of $A(\mathbb{R}^N) \neq \mathbb{R}^M$. □

In the language of semidefinite programming, these conditions become statements about spectrahedra. In 3.2.4 we have described spectrahedra with nonempty interior; in 3.2.5 we give a sufficient condition for a spectrahedron to be nonempty. The latter gives a necessary condition for convexity which can be expressed, using work of Klep and Schweighofer, in an algebraic form (see [16]).

Theorem 3.2.6. *(Nonlinear Farkas' Lemma) There exists $\ell \in \mathbb{R}^M$ such that $\ell \cdot A \succeq 0$ if and only if $-1 \in M_A$, where*

$$M_A \equiv \{\sigma + \sum_i V_i^T A_i V_i \mid \sigma \text{ is SOS, and } V_i \in \mathbb{R}[x_1, \ldots, x_n]^M\}$$

is the quadratic module associated to the tuple A.

In the language of this result, 3.2.5 states that whenever A is such that $A(\mathbb{R}^N) \neq \mathbb{R}^M$ is convex the quadratic module M_A cannot contain -1.

The following corollary shows that certain instances of the quadratic convexity problem can be recast as questions about joint numerical range, a subject on which more has been written. And it gives an interesting partial result in our motivating context: *there is a compact subset of $\Sigma_{n,d}(k)$ whose convexity is equivalent to that of the entire set.* Of course, we have already described such a subset (in 2.2.3), but 3.2.7 describes them all.

Corollary 3.2.7. *(to 3.2.4)If $A : \mathbb{R}^N \to \mathbb{R}^M$ is a quadratic map and* $\operatorname{conv} A(S^{N-1})$ *does not contain* 0*, then there is an invertible linear operator L on $\mathbb{R}^N$ such that $A(\mathbb{R}^N)$ is convex exactly when $A \circ L(S^{N-1})$ is convex.*

Proof. Let $A : \mathbb{R}^N \to \mathbb{R}^M$ be a quadratic map such that $\operatorname{conv} A(S^{N-1})$ does not contain 0. Fix ℓ such that $\ell \cdot A \succ 0$. By 3.2.4 there is an invertible linear operator B such that $\ell \cdot A = B^T B$. The convex hull of

$$A(\{x \in \mathbb{R}^N \mid x^T(\ell \cdot A)x = 1\}) = A(\{x \in \mathbb{R}^N \mid (Bx)^T Bx = 1\})$$

is a compact base for $\operatorname{conv} A(\mathbb{R}^N)$, and therefore $A(\mathbb{R}^N)$ is convex exactly when $A(\{x \in \mathbb{R}^N \mid (Bx)^T Bx = 1\})$ is convex. Since

$$A(\{x \in \mathbb{R}^N \mid (Bx)^T Bx = 1)\} = A \circ B^{-1}(S^{N-1}),$$

the map B^{-1} is the desired L as in the statement. □

Finally, we show that 3.2.7 can be strengthened to a statement about topological and geometric properties of the set $A(\mathbb{R}^N)$ itself. The conditions of 3.2.4 imply that $A(\mathbb{R}^N)$ is closed and contains no lines. While the converse is not true, we still have the following:

Theorem 3.2.8. *Suppose that $A : \mathbb{R}^N \to \mathbb{R}^M$ is a quadratic map whose image $A(\mathbb{R}^N)$ is closed and contains no lines. Then there is a linear transformation T from $\mathbb{R}^k$ to $\mathbb{R}^N$, for some $k \leq N$, such that $A(\mathbb{R}^N)$ is convex exactly when $A \circ T(S^{k-1})$ is convex.*

Proof. Assuming the above hypotheses, we first show that $A(\mathbb{R}^N)$ has a compact base. Since $A(\mathbb{R}^N)$ is closed, so is the set $U = \operatorname{conv} A(\mathbb{R}^N) \cap S^{M-1}$. In fact, U is a compact convex set not containing the origin. Just as in the proof of 3.2.4, we conclude that there is a linear functional ℓ on $\mathbb{R}^M$ strictly separating U from the origin, and bounded below by 1 on U. Then $K \equiv \ell^{-1}(1) \cap A(\mathbb{R}^N)$ is a compact base for $A(\mathbb{R}^N)$.

Now we define an invertible linear operator on $\mathbb{R}^M$ which takes K into the positive orthant. Since K is compact, sufficiently small perturbations of ℓ will also be strictly positive on K; as in the proof of 3.2.4, we take a basis for $\mathbb{R}^M$ from an open ball around ℓ (regarded as an element of $\mathbb{R}^M$) all of whose elements are positive on K. Say this basis consists of the vectors $v_1, \ldots, v_M$, and define

$$V = \begin{pmatrix} v_1^T \\ \vdots \\ v_M^T \end{pmatrix}.$$

By contrivance, $V(K)$ (the image of K under multiplication by V) is contained in the strictly positive orthant, and it is a compact base for the cone $V(A(\mathbb{R}^N))$. The origin is therefore the only point of $V(A(\mathbb{R}^N))$ whose entries are not all strictly positive. This gives $\ker v_j \cdot A = \ker v_i \cdot A$ for all i and j. Take T to be a linear transformation from $\mathbb{R}^k$ onto the range of $v_1 \cdot A$ (or that of any of the other $v_1 \cdot A$), where $k = \operatorname{rank} v_1 \cdot A$. Now apply 3.2.7 to the quadratic map $A \circ T : \mathbb{R}^k \to \mathbb{R}^M$. □

We are almost ready for the main theorem. In the next section we show how it enables one to compute lower bounds on pythagoras numbers, and demonstrate that those bounds agree in all cases for which pythagoras numbers are now known. The proof relies on Sard's theorem (see [35]). We state a simplified version below for the sake of clarity and completeness.

Theorem 3.2.9. *(Sard's Theorem) If $f : N \to \mathbb{R}^m$ is a smooth map on a smooth manifold N, then the set of critical values of f has measure zero. That is, the set of values*

$$\{f(x) : Df(x) \textit{ has rank less than } m\} \subset \mathbb{R}^m$$

has Lebesgue measure zero.

An obvious corollary is that the set of critical values of such a map must have empty interior. Therefore, if every value of f is a critical value, then the image of f must have an empty interior. In other words, if the image of f has nonempty interior, then the critical values of f constitute a proper subset of the image of f, and it follows that there must be some x for which $Df(x)$ has full rank. Conversely, if $Df(x)$ has full rank for some $x \in N$, then an application of the inverse function theorem guarantees nonemptiness of the interior of the image of f.

Note that $Df(x)$ has full rank exactly when $Df(x)Df(x)^T$ (here identifying $Df(x)$ with the standard matrix representation of $Df(x)$) is invertible (and therefore positive definite), which in turn is equivalent to the inequality $\det(Df(x)Df(x)^T) > 0$. Combined with 1.2.8, these observations give us a necessary condition for quadratic convexity in terms of a polynomial in $\mathbb{R}[x_1, \ldots, x_N]$.

Theorem 3.2.10. *If $A : \mathbb{R}^N \to \mathbb{R}^M$ is a quadratic map and $A(\mathbb{R}^N)$ spans $\mathbb{R}^M$, then $A(\mathbb{R}^N)$ is convex only if the polynomial $\det(DA(x)DA(x)^T) \in \mathbb{R}[x_1, \ldots, x_N]$ is different from the zero polynomial. The derivative $DA(x)$ therefore has full rank for almost every x provided it has full rank at some x.*

Proof. Since A is a polynomial function, its jacobian consists of polynomial entries. The expression $\det(DA(x)DA(x)^T)$ is therefore a polynomial. We have from the above remarks that $\det(DA(x)DA(x)^T) \neq 0 \in \mathbb{R}[x_1, \ldots, x_N]$ when $A(\mathbb{R}^N)$ is convex and spans $\mathbb{R}^M$. Since the zero set of a nontrivial polynomial has Lebesgue measure 0, we conclude that $DA(x)$ has full rank at almost every x exactly when it has full rank for some x. □

The above argument can be modified to prove an analog of 3.2.10 for the question of convexity of $A(S^{N-1})$.

Theorem 3.2.11. *If $A(S^{N-1}) \subset \mathbb{R}^M$ is convex and not contained in a proper affine subspace of $\mathbb{R}^M$, then the polynomial*

$$\det\Big([DA(x)(\|x\|^2 I - xx^T)][DA(x)(\|x\|^2 I - xx^T)]^T\Big)$$

is not the zero polynomial.

Proof. Fix $v \in S^{N-1}$, with A as in the statement. The tangent space to S^{N-1} at v is $\{y : v^T y = 0\}$ (as a subspace of $\mathbb{R}^N$), and the orthogonal projection onto this tangent space is simply $I - vv^T$. The requirement that $Df(v)$ have full rank implies that $DA(v)(I - vv^T)$ must have full rank, and therefore

$$\det\Big([DA(v)(\|x\|^2 I - xx^T)][DA(x)(\|x\|^2 I - xx^T)]^T\Big) \neq 0,$$

and so the claim is proved. $\square$

3.3 Application to SOS

As in 2.1.2, we fix $n, d \in \mathbb{N}$ and define the linear transformation $T : \mathcal{S}_N \to \mathbb{R}^M$ by $T(X) = (\text{trace}(A_\alpha X))_{|\alpha|=2d}$, where $N = \binom{n+d-1}{d}$, $M = \binom{n+2d-1}{2d}$ and A_α are the $N \times N$ matrices indexed by multi-indices, the entries of which are given by

$$(A_\alpha)_{\beta,\gamma} = \begin{cases} 1, & \text{if } \beta + \gamma = \alpha \\ 0, & \text{otherwise.} \end{cases}$$

In the first chapter we showed that $\Sigma_{n,d}$ can be realized as the image of $\mathcal{S}_N^+$ under a linear map, and therefore coincides with the convex hull of the quadratic map $x \longmapsto (x^T A_\alpha x)_{|\alpha|=2d}$. In fact, for any k the set $\Sigma_{n,d}(k)$ is the image of a quadratic map, and coincides with $\Sigma_{n,d}$ itself if and only if it is convex. This key observation is recorded below.

Proposition 3.3.1. *The set $\Sigma_{n,d}(k)$ is the image of the rank-k PSD matrices under T. Equivalently, $\Sigma_{n,d}(k)$ is the image of the quadratic map associated to the tuple $I_k \otimes A \equiv (I_k \otimes A_\alpha)_{|\alpha|=2d}$, where I_k is the $k \times k$ identity matrix and $\otimes$ is the Kronecker product.*

Furthermore, $I_k \otimes A(\mathbb{R}^{kN})$ is convex if and only if $\mathcal{P}(\mathbb{R}[x_1, \ldots, x_n]_{2d}) \leq k$.

The main significance of 3.2.4 is that it lets us replace an unbounded set with a compact one (3.2.7). Another application is to estimate the distance from an arbitrary $\sigma \in \Sigma_{n,d}$ to the subset $\Sigma_{n,d}(k)$. First, we need another result (see [27]).

Theorem 3.3.2. *(Approximate Carathéodory Theorem) If $S \subset \Delta^{N-1}$, then any $x \in \operatorname{conv}(S)$ can be approximated within error $\frac{1}{\sqrt{k}}$ (in the ℓ^2 norm) by a convex combination of k elements of S.*

From 3.2.4 and 3.2.7 we conclude that there is a linear operator T on $\mathbb{R}^M$ (identying $\mathbb{R}^M$ with $\mathbb{R}[x_1, \ldots, x_n]_{2d}$) such that $T(\Sigma_{n,d})$ lies in the positive orthant. Let us define

$$K = T(\Sigma_{n,d}) \cap \Delta^{M-1}.$$

By 3.3.2, any element of $T(x) \in K$ can be approximated within $\frac{1}{\sqrt{k}}$ by $T(\sigma)$ for some $\sigma \in \Sigma_{n,d}(k)$. Applying T^{-1}, we infer that any x in the compact base

$$T^{-1}(K) = \Sigma_{n,d} \cap \{x \mid \ell \circ T(x) = 1\}$$

can be approximated within $\frac{\|T^{-1}\|}{\sqrt{k}}$ by σ in

$$\Sigma_{n,d}(k) \cap \{x \mid \ell \circ T(x) = 1\} \subset T^{-1}(K).$$

It would be interesting to find bounds on $\|T^{-1}\|$ in terms of n and d. For a given map, we can obtain an estimate no less precise than that of 3.3.2. One approach is to start with the compact base for $\Sigma_{n,d}$ presented in 2.2.3. Let λ denote the linear functional defined there in terms of integration over a box. We can first choose an orthogonal transformation U which takes λ to

$$\tfrac{\|\lambda\|}{\sqrt{M}}(1, \ldots, 1)^T \in \mathbb{R}^M.$$

Then, it is a matter of stretching the image along the direction of $(1, \ldots, 1)^T$ until it fits in the positive orthant. Equivalently one could shrink the orthogonal complement of $(1, \ldots, 1)^T$; both can be achieved by semidefinite programming. This procedure makes uniform the distortion of distance in moving (linearly) the compact base in 2.2.3 into the standard simplex Δ^{M-1}.

The results of the preceding section say slightly more than is expressed in 3.3.1. In fact, we have shown how to find *the smallest k such that there is an open subset of $\Sigma_{n,d}$ with pythagoras number k.* The procedure to 'compute' this value is straightforward: for each k, calculate the determinant in 3.2.10 for the tuple $I_k \otimes A$ and stop when the result is not the zero polynomial. This is spelled out explicitly in the pseudocode below.

Algorithm for finding $\min\{\mathcal{P}(U) \mid U(\neq \emptyset) \subset \Sigma_{n,d} \text{ is open}\}$

```
Input: $A$, the $N-$tuple associated to $\Sigma_{n,d}$

Output: The natural number $\min\{\mathcal{P}(U) \mid U(\neq \emptyset) \subset \Sigma_{n,d} \text{ is open}\}$

  $k := 1$

  while $\det[D(I_k \otimes A)(x) D(I_k \otimes A)(x)^T] = 0 \in \mathbb{R}[x_1, \ldots, x_{kN}]$ do

    $k := k + 1$

  end while

  return k
```

Surprisingly, it happens that pythagoras number is equal to this lower bound in all cases for which it is known.

Example 3.3.3. Recall that $\mathcal{P}(\Sigma_{2,d}) = 2$, that $\mathcal{P}(\Sigma_{n,1}) = n$ and that $\mathcal{P}(\Sigma_{3,2}) = 3$ for all n and d. The purpose of this example is to show that in all of these cases the derivative $D(I_k \otimes A)$ does not achieve full rank for k less than the pythagoras number. Taking A to be the quadratic map parametrizing single squares, we note that A maps from a space of dimension $\binom{2+d-1}{d} = d + 1$ to a space of dimension $\binom{2+2d-1}{2d} = 2d + 1$. By 2.2.1, DA cannot have full rank anywhere. Since $I_2 \otimes A$ gives us the entire cone, $D(I_2 \otimes A)$ must have full rank somewhere. The case $n = 3, d = 4$ is handled in the same way, since $\binom{3+4-1}{4} = 15 > 12 = 2\binom{3+2-1}{2}$.

When looking at quadratic forms, however, we cannot get by with a simple dimension count. Indeed, the map $I_{n-1} \otimes A$ in this case has domain $\mathbb{R}^{n(n-1)}$ and codomain $\mathbb{R}^{n(n+1)/2}$. Instead, we recall 2.1.2, which says that $T : \mathcal{S}_N \to \mathbb{R}^M$ defined by $T(X) = (\mathrm{trace}(A_\alpha X))_{|\alpha|=2}$, where $N = n$ and $M = n(n+1)/2$, is a bijection. Since. We also have $I_{n-1} \otimes A = T \circ F$, where $F : \mathbb{R}^{N\times(N-1)} \to \mathcal{S}_N$ is defined by $F(X) = XX^T$ (here we identify $\mathbb{R}^{N\times(N-1)}$ with $N \times (N-1)$ matrices). But the image of F consists of exactly the singular elements of $\mathcal{S}_N^+$, which has dimension strictly less than M. By the chain rule (see [38]), $D(I_{n-1} \otimes A) = D(T \circ F) = (DT)(DF) = T(DF)$, so that if $D(I_{n-1} \otimes A)$ has rank M somewhere, then so must DF, contradicting the previous sentence. ■

We conclude this chapter with a conjecture about quadratic maps which includes as a special case the conjecture that our derivative condition detects the pythagoras number.

Conjecture 3.3.4. *If A is positive definite in each entry and $A(\mathbb{R}^N) \subset \mathbb{R}^M$, then*

$$\begin{aligned}&\min\{k \mid I_k \otimes A(\mathbb{R}^{kN}) \textit{ is convex.}\}\\ &= \min\{k \mid D(I_k \otimes A) \textit{ has full rank somewhere}\}.\end{aligned}$$

Chapter 4

Other Approaches

Below we briefly discuss other approaches to the question of quadratic convexity.

4.1 Moments and Separation

In 3.2.2 we showed how to modify a quadratic map A so that the convexity of $A(S^{N-1})$ is preserved while the resulting image lies in no proper affine subspace of $\mathbb{R}^M$ (the codomain of the modified map). Assuming A to be such a map, then it follows from 1.2.8 that the closed set $A(S^{N-1})$ is convex if and only if $m(A(S^{N-1})) = m(\operatorname{conv} A(S^{N-1}))$, where m is Lebesgue measure on $\mathbb{R}^M$. Equivalently, $A(S^{N-1})$ fails to be convex if and only if $\operatorname{conv} A(S^{N-1}) \setminus A(S^{N-1})$ contains an open ball. From this it follows that $A(S^{N-1})$ and $\operatorname{conv} A(S^{N-1})$ coincide (i.e.

$A(S^{N-1})$ is convex) if and only if all quadratic polynomials nonnegative on one are also nonnegative on the other.

Proposition 4.1.1. *If for all quadratic $p \in \mathbb{R}[x_1, \ldots, x_M]$ it holds that $p \geq 0$ on $A(S^{N-1})$ only if $p \geq 0$ on $\operatorname{conv} A(S^{N-1})$, then $A(S^{N-1})$ is convex.*

Proof. If $A(S^{N-1})$ is not convex, then there exist $x_0 \in \operatorname{conv} A(S^{N-1}) \setminus A(S^{N-1})$ and $\varepsilon > 0$ such that the polynomial $\|x - x_0\|^2 - \varepsilon$ is positive on $A(S^{N-1})$ (and not on $\operatorname{conv} A(S^{N-1})$). □

Another interpretation of the above statement is that the quadratic moments of the two sets coincide if and only if $A(S^{N-1})$ is convex. That is, the set of truncated sequences of moments of monomials of degree at most 2 (henceforth *quadratic moment sequences*), with respect to probability measures on $A(S^{N-1})$, coincide only if $A(S^{N-1})$ is convex. In this view, the coefficients of a quadratic polynomial positive on $A(S^{N-1})$ but not on $\operatorname{conv} A(S^{N-1})$ define a linear functional separating a single quadratic moment sequence on $\operatorname{conv} A(S^{N-1})$ (namely, a point evaluation) from the compact convex set of quadratic moment sequences of probability measures on $A(S^{N-1})$.

In [14], the optimization techniques from [19] are employed in computing volumes and moments of compact basic semialgebraic sets. While $A(S^{N-1})$ is not such a set, it is the image of such a set under a linear transformation, and for this reason the same methods can, at least in theory, be applied. In practice, the convergence of the approximations of [14] is thwarted by round-off error.

4.2 Faces and Extreme Points

Let $K \subset \mathbb{R}^N$ be a compact convex set with extreme points $E(K)$. If

$$T : \mathbb{R}^N \to \mathbb{R}^M$$

is a linear transformation, under what additional assumptions (regarding K, $E(K)$ and T) may we conclude that $T(K) = T(E(K))$? This basic question is posed in [42], and some sufficient conditions are proved there. Roughly, if K has exactly one nonsingleton proper face (or if all proper faces are singletons) and $\ker T$ contains a difference $x - y \neq 0$ with $x \in K$ and $y \in E(K)$, then $T(K) = T(E(K))$. While this result holds in very general settings (in any real vector space), the conditions on the structure of K are rather strong. In particular, these results do not help us to decide quadratic convexity. Rather, the question of quadratic convexity is a question of when "extreme points suffice".

Recall from Chapter 3 the $M-$tuple $I_k \otimes A$ defined in terms of the $M-$tuple $A = (A_1, \ldots, A_M)$. Assuming that each A_j is $N \times N$, we define T_k on $\mathcal{S}_{kN}$ by $T_k(X) = (\text{trace}((I_k \otimes A_j)X))_{j=1}^M$. We therefore have that the set of convex combinations of k elements of $A(S^{N-1})$ is exactly

$$I_k \otimes A(S^{kN-1}) = T_k(E(K_{kN}^+)),$$

since the rank-one projections are the extreme points of K_{kN}^+. Therefore, combinations of k terms suffice for conv $A(S^{N-1})$ if and only if $T_k(E(K_{kN}^+)) = T_k(K_{kN}^+)$.

Moving from the compact to a convex cone, we remark on a related idea; that the dimension of the largest proper face of a cone provides a bound on the number of terms required in an arbitrary convex combination. More precisely, we have the following observation.

Proposition 4.2.1. *If $C \subset \mathbb{R}^M$ is a (nonempty) convex cone with compact base K, then each element of C may be expressed as a conic combination of $m+2$ extreme points of K, where m is the maximum dimension among all proper faces of C.*

Proof. A variant of 1.2.6 (see [2]) adapted for cones gives the first part of the claim, namely that every element of C can be realized as a conic combination of extreme points of K. We combine this with 1.2.7 as follows. Since C is a cone, it suffices to prove the result for all elements of K. Let $x \in K \setminus E(K)$ (if there is no such x, then the result follows easily). Choose $y \in E(K)$ (there is such a y by 1.2.6). Since y is extreme, C contains combinations $tx + (1-t)y$ only for $t \geq 0$. Since K is compact, K contains combinations $tx + (1-t)y$ only for t up to some finite maximum, call it t_m. Since K is a compact base for C, the same is true of C. Now $t_m x + (1 - t_m)y$ must lie in some proper face F of C. If $\dim F = d$, then 1.2.7 implies that $t_m x + (1 - t_m)y$ may be expressed as a convex combination of $d + 1$ extreme points of $F \cap K$, all of which must also be extreme points of K itself.

Writing $t_m x + (1 - t_m)y = \lambda_1 x_1 + \ldots + \lambda_{d+1} x_{d+1}$, we finally have

$$x = \frac{1}{t_m}(y + t_m(x - y) + (t_m - 1)y) = \frac{1}{t_m}(\lambda_1 x_1 + \ldots + \lambda_{d+1} x_{d+1} + (t_m - 1)y).$$

□

With a more complete picture of the facial structure of the SOS cone, it might be possible to improve known bounds on $\mathcal{P}(\Sigma_{n,d})$ using this approach.

Bibliography

[1] Emil Artin. Über die Zerlegung definiter Funktionen in Quadrate. *Abh. Math. Sem. Univ. Hamburg*, 5(1):100–115, 1927.

[2] Alexander Barvinok. *A course in convexity*, volume 54 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002.

[3] Paul Binding and Chi-Kwong Li. Joint ranges of Hermitian matrices and simultaneous diagonalization. *Linear Algebra Appl.*, 151:157–167, 1991.

[4] Grigoriy Blekherman. There are significantly more nonnegative polynomials than sums of squares. *Israel J. Math.*, 153:355–380, 2006.

[5] Jacek Bochnak, Michel Coste, and Marie-Françoise Roy. *Real algebraic geometry*, volume 36 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1998. Translated from the 1987 French original, Revised by the authors.

[6] Emmanuel J. Candès and Terence Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2010.

[7] M. D. Choi, T. Y. Lam, and B. Reznick. Sums of squares of real polynomials. In *K-theory and algebraic geometry: connections with quadratic forms and division algebras (Santa Barbara, CA, 1992)*, volume 58 of *Proc. Sympos. Pure Math.*, pages 103–126. Amer. Math. Soc., Providence, RI, 1995.

[8] John P. D'Angelo and Jiří Lebl. Pfister's theorem fails in the Hermitian case. *Proc. Amer. Math. Soc.*, 140(4):1151–1157, 2012.

[9] Eugene Gutkin, Edmond A. Jonckheere, and Michael Karow. Convexity of the joint numerical range: topological and differential geometric viewpoints. *Linear Algebra Appl.*, 376:143–171, 2004.

[10] Eugene Gutkin and Karol Życzkowski. Joint numerical ranges, quantum maps, and joint numerical shadows. *Linear Algebra Appl.*, 438(5):2394–2404, 2013.

[11] M. Harrison. Pfister's theorem fails in the free case. In M. Putinar, M. De Oliveira, and H. Dym, editors, *Mathematical Methods in Systems, Optimization and Control*, pages 189–194. Birkhäuser Basel, 2012.

[12] J. W. Helton, F. Dell Kronewitter, W. M. McEneaney, and Mark Stankus. Singularly perturbed control systems using non-commutative computer algebra. *Internat. J. Robust Nonlinear Control*, 10(11-12):983–1003, 2000. George Zames commemorative issue.

[13] J. William Helton. "Positive" noncommutative polynomials are sums of squares. *Ann. of Math. (2)*, 156(2):675–694, 2002.

[14] D. Henrion, J. B. Lasserre, and C. Savorgnan. Approximate volume and integration for basic semialgebraic sets. *SIAM Rev.*, 51(4):722–743, 2009.

[15] Didier Henrion, Jean-Bernard Lasserre, and Johan Löfberg. GloptiPoly 3: moments, optimization and semidefinite programming. *Optim. Methods Softw.*, 24(4-5):761–779, 2009.

[16] I. Klep and M. Schweighofer. An exact duality theory for semidefinite programming based on sums of squares. *ArXiv e-prints*, July 2012.

[17] J.-L. Krivine. Anneaux préordonnés. *J. Analyse Math.*, 12:307–326, 1964.

[18] Jean B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM J. Optim.*, 11(3):796–817 (electronic), 2000/01.

[19] Jean B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM J. Optim.*, 11(3):796–817, 2000/01.

[20] Chi-Kwong Li and Yiu-Tung Poon. Convexity of the joint numerical range. *SIAM J. Matrix Anal. Appl.*, 21(2):668–678, 1999.

[21] Chi-Kwong Li and Yiu-Tung Poon. Generalized numerical ranges and quantum error correction. *J. Operator Theory*, 66(2):335–351, 2011.

[22] 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.

[23] M. Lothaire. *Combinatorics on words*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1997. With a foreword by Roger Lyndon and a preface by Dominique Perrin, Corrected reprint of the 1983 original, with a new preface by Perrin.

[24] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13 of *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994.

[25] Albrecht Pfister. Zur Darstellung definiter Funktionen als Summe von Quadraten. *Invent. Math.*, 4:229–237, 1967.

[26] S. Pironio, M. Navascués, and A. Acín. Convergent relaxations of polynomial optimization problems with noncommuting variables. *SIAM J. Optim.*, 20(5):2157–2180, 2010.

[27] G. Pisier. Remarques sur un résultat non publié de B. Maurey. In *Seminar on Functional Analysis, 1980–1981*, pages Exp. No. V, 13. École Polytech., Palaiseau, 1981.

[28] Yiu Tung Poon. Generalized numerical ranges, joint positive definiteness and multiple eigenvalues. *Proc. Amer. Math. Soc.*, 125(6):1625–1634, 1997.

[29] S. Prajna, A. Papachristodoulou, P. Seiler, and P. A. Parrilo. *SOSTOOLS: Sum of squares optimization toolbox for MATLAB*, 2004.

[30] Mihai Putinar. Positive polynomials on compact semi-algebraic sets. *Indiana Univ. Math. J.*, 42(3):969–984, 1993.

[31] Motakuri Ramana and A.J. Goldman. Quadratic maps with convex images. Technical report, RUTCOR, 1995.

[32] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010.

[33] Bruce Reznick. Extremal PSD forms with few terms. *Duke Math. J.*, 45(2):363–374, 1978.

[34] R. Tyrrell Rockafellar. *Convex analysis.* Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[35] Arthur Sard. Images of critical sets. *Ann. of Math. (2)*, 68:247–259, 1958.

[36] Claus Scheiderer. Positivity and sums of squares: a guide to recent results. In *Emerging applications of algebraic geometry*, volume 149 of *IMA Vol. Math. Appl.*, pages 271–324. Springer, New York, 2009.

[37] Konrad Schmüdgen. The K-moment problem for compact semi-algebraic sets. *Math. Ann.*, 289(2):203–206, 1991.

[38] Michael Spivak. *Calculus on manifolds. A modern approach to classical theorems of advanced calculus.* W. A. Benjamin, Inc., New York-Amsterdam, 1965.

[39] Gilbert Stengle. A nullstellensatz and a positivstellensatz in semialgebraic geometry. *Math. Ann.*, 207:87–97, 1974.

[40] Alfred Tarski. A decision method for elementary algebra and geometry. In *Quantifier elimination and cylindrical algebraic decomposition (Linz, 1993)*, Texts Monogr. Symbol. Comput., pages 24–84. Springer, Vienna, 1998.

[41] Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM Rev.*, 38(1):49–95, 1996.

[42] Michael D. Wills and Douglas Baker. When are extreme points enough? *J. Convex Anal.*, 17(2):565–582, 2010.