Semidefinite Optimization

Mastermath
Spring 2012

Monique Laurent

Centrum Wiskunde & Informatica
Science Park 123
1098 XG Amsterdam
The Netherlands

monique@cwi.nl

Frank Vallentin

Delft Institute of Applied Mathematics
Technical University of Delft
P.O. Box 5031
2600 GA Delft
The Netherlands

f.vallentin@tudelft.nl

May 22, 2012

CONTENTS

Part I

Theory and algorithms for semidefinite optimization

CHAPTER 1

BACKGROUND: CONVEX SETS AND POSITIVE SEMIDEFINITE MATRICES

A set C is called convex if, given any two points x and y in C, the straight line segment connecting x and y lies completely inside of C. For instance, cubes, balls or ellipsoids are convex sets whereas a torus is not. Intuitively, convex sets do not have holes or dips.

Usually, arguments involving convex sets are easy to visualize by two-dimensional drawings. One reason being that the definition of convexity only involves three points which always lie in some two-dimensional plane. On the other hand, convexity is a very powerful concept which appears (sometimes unexpected) in many branches of mathematics and its applications. Here are a few areas where convexity is an important concept: mathematical optimization, high-dimensional geometry, analysis, probability theory, system and control, harmonic analysis, calculus of variations, game theory, computer science, functional analysis, economics, and there are many more.

Our aim is to work with convex sets algorithmically. So we have to discuss ways to represent them in the computer, in particular which data do we want to give to the computer. Roughly speaking, there are two convenient possibilities to represent convex sets: By an implicit description as an intersection of halfspaces or by an explicit description as the convex combination of extreme points. The goal of this chapter is to discuss these two representations. In the context of functional analysis they are connected to two famous theorems, the Hahn-Banach theorem and the Krein-Milman theorem. Since we are only working in finite-dimensional Euclidean spaces (and not in the more general setting of infinite-dimensional topological vector spaces) we can derive the statements

using simple geometric arguments.

Later we develop the theory of convex optimization in the framework of conic programs. For this we need a special class of convex sets, namely convex cones. The for optimization most relevant convex cones are at the moment two involving vectors in $\mathbb{R}^n$ and two involving symmetric matrices in $\mathbb{R}^{n\times n}$, namely the non-negative orthant, the second order cone, the cone of positive semidefinite matrices, and the cone of copositive matrices. Clearly, the cone of positive semidefinite matrices plays the main role here. As background information we collect a number of basic properties of positive semidefinite matrices.

1.1 Some fundamental notions

Before we turn to convex sets we recall some fundamental geometric notions. The following is a brief review, without proofs, of some basic definitions and notations appearing frequently in the sequel.

1.1.1 Euclidean space

Let E be an n-dimensional *Euclidean space* which is an n-dimensional real vector space having an inner product. We usually use the notation $x \cdot y$ for the inner product between the vectors x and y. This inner product defines a norm on E by $\|x\| = \sqrt{x \cdot x}$ and a metric by $d(x, y) = \|x - y\|$.

For sake of concreteness we will work with coordinates most of the time: One can always identify E with $\mathbb{R}^n$ where the inner product of the column vectors $x = (x_1, \ldots, x_n)^\mathsf{T}$ and $y = (y_1, \ldots, y_n)^\mathsf{T}$ is the usual one: $x \cdot y = x^\mathsf{T} y = \sum_{i=1}^n x_i y_i$. This identification involves a linear transformation $T : E \to \mathbb{R}^n$ which is an isometry, i.e. $x \cdot y = Tx \cdot Ty$ holds for all $x, y \in E$. Then the norm is the Euclidean norm (or ℓ_2-norm): $\|x\|_2 = \sqrt{\sum_i x_i^2}$ and $d(x, y) = \|x - y\|_2$ is the Euclidean distance between two points $x, y \in \mathbb{R}^n$.

1.1.2 Topology in finite-dimensional metric spaces

The *ball* with center $x \in \mathbb{R}^n$ and radius r is

$$B(x, r) = \{y \in \mathbb{R}^n : d(x, y) \leqslant r\}.$$

Let A be a subset of n-dimensional Euclidean space. A point $x \in A$ is an *interior point* of A if there is a positive radius $\varepsilon > 0$ so that $B(x, \varepsilon) \subseteq A$. The set of all interior points of A is denoted by $\operatorname{int} A$. We say that a set A is *open* if all points of A are interior points, i.e. if $A = \operatorname{int} A$. The set A is *closed* if its complement $\mathbb{R}^n \backslash A$ is open. The *(topological) closure* $\overline{A}$ of A is the smallest (inclusion-wise) closed set containing A. One can show that a set A in $\mathbb{R}^n$ is closed if and only if every converging sequence of points in A has a limit which also lies in A. A point $x \in A$ belongs to the *boundary* ∂A of A if for every $\varepsilon > 0$ the ball $B(x, \varepsilon)$ contains points in A and in $\mathbb{R}^n \backslash A$. The boundary ∂A is a closed

set and we have $\overline{A} = A \cup \partial A$, and $\partial A = \overline{A} \setminus \operatorname{int} A$. The set A is *compact* if every sequence in A contains a convergent subsequence. The set A is compact if and only if it is closed and bounded (i.e. it is contained in a ball of sufficiently large, but finite, radius).

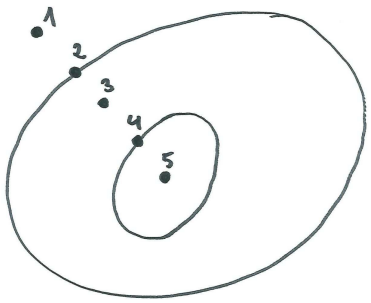


Figure 1.1: A compact, non-convex set A. Which points lie in $\operatorname{int} A$, $\overline{A}$, ∂A?

For instance, the boundary of the ball with radius 1 and center 0 is the *unit sphere*

$$\partial B(0,1) = \{y \in \mathbb{R}^n : d(0,y) = 1\} = \{x \in \mathbb{R}^n : x^\mathsf{T} x = 1\}.$$

Traditionally, it is called the $(n-1)$-dimensional unit sphere, denoted as $\mathbb{S}^{n-1}$, where the superscript $n-1$ indicates the dimension of the manifold.

1.1.3 Affine geometry

A subset $A \subseteq \mathbb{R}^n$ is called an *affine subspace* of $\mathbb{R}^n$ if it is a translated linear subspace: One can write A in the form

$$A = x + L = \{x + y : y \in L\}$$

where $x \in \mathbb{R}^n$ and where L is a linear subspace of $\mathbb{R}^n$. The *dimension* of A is defined as $\dim A = \dim L$. Affine subspaces are closed under *affine linear combinations*:

$$\forall N \in \mathbb{N}\ \forall x_1, \ldots, x_N \in A\ \forall \alpha_1, \ldots, \alpha_N \in \mathbb{R} : \sum_{i=1}^N \alpha_i = 1 \Longrightarrow \sum_{i=1}^N \alpha_i x_i \in A.$$

The smallest affine subspace containing a set of given points is its *affine hull*. The affine hull of $A \subseteq \mathbb{R}^n$ is the set of all possible affine linear combinations

$$\operatorname{aff} A = \left\{ \sum_{i=1}^N \alpha_i x_i : N \in \mathbb{N}, x_1, \ldots, x_N \in A, \alpha_1, \ldots, \alpha_N \in \mathbb{R}, \sum_{i=1}^N \alpha_i = 1, \right\}.$$

A fact which requires a little proof (exercise). The dimension of an arbitrary set A is $\dim A = \dim(\operatorname{aff} A)$. One-dimensional affine subspaces are *lines* and $(n-1)$-dimensional affine subspaces are *hyperplanes*. A hyperplane can be specified as

$$H = \{x \in \mathbb{R}^n : c^{\mathsf{T}} x = \beta\},$$

where $c \in \mathbb{R}^n \backslash \{0\}$ is the normal of H (which lies orthogonal to H) and where $\beta \in \mathbb{R}$. Sometimes we write $H_{c,\beta}$ for it.

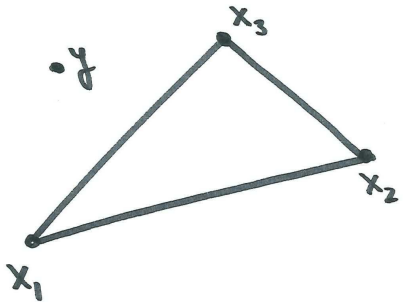


Figure 1.2: Determine (as accurate as possible) the coefficients $\alpha_1, \alpha_2, \alpha_3$ of the affine combination $y = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3$ with $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

If the dimension of $A \subseteq \mathbb{R}^n$ is strictly smaller than n, then A does not have an interior, $\operatorname{int} A = \varnothing$. In this situation one is frequently interested in the interior points of A relative to the affine subspace $\operatorname{aff} A$. We say that a point $x \in A$ belongs to the *relative interior* of A when there is a ball $B(x, \varepsilon)$ with strictly positive radius $\varepsilon > 0$ so that $\operatorname{aff} A \cap B(x, \varepsilon) \subseteq A$. We denote the set of all relative interior points of A by $\operatorname{relint} A$. Of course, if $\dim A = n$, then the interior coincides with the relative interior: $\operatorname{int} A = \operatorname{relint} A$.

1.2 Convex sets

A subset $C \subseteq \mathbb{R}^n$ is called a *convex set* if for every pair of points $x, y \in C$ also the entire line segment between x and y is contained in C. The *line segment* between the points x and y is defined as

$$[x, y] = \{(1 - \alpha)x + \alpha y : 0 \leqslant \alpha \leqslant 1\}.$$

Convex sets are closed under *convex combinations*:

$$\forall N \in \mathbb{N}\ \forall x_1, \ldots, x_N \in C\ \forall \alpha_1, \ldots, \alpha_N \in \mathbb{R}_{\geqslant 0} : \sum_{i=1}^{N} \alpha_i = 1 \Longrightarrow \sum_{i=1}^{N} \alpha_i x_i \in C.$$

The convex hull of $A \subseteq \mathbb{R}^n$ is the smallest convex set containing A. It is

$$\operatorname{conv} A = \left\{ \sum_{i=1}^{N} \alpha_i x_i : N \in \mathbb{N}, x_1, \ldots, x_N \in A, \alpha_1, \ldots, \alpha_N \in \mathbb{R}_{\geqslant 0}, \sum_{i=1}^{N} \alpha_i = 1 \right\},$$

which requires an argument. We can give a mechanical interpretation of the convex hull of finitely many point $\mathrm{conv}\{x_1, \ldots, x_N\}$: The convex hull consists of all centres of gravity of point masses $\alpha_1, \ldots, \alpha_N$ at the positions $x_1, \ldots, x_N$.

The convex hull of finitely many points is called a *polytope*. Two-dimensional, planar, polytopes are polygons. Other important examples of convex sets are balls, halfspaces, and line segments. Furthermore, arbitrary intersections of convex sets are convex again. The *Minkowski sum* of convex sets C, D given by

$$C + D = \{x + y : x \in C, y \in D\}$$

is a convex set.

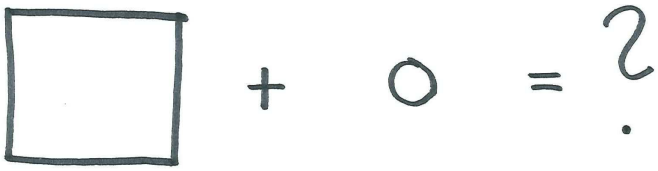


Figure 1.3: Exercise: What is the Minkowski sum of a square and a disk?

Here are two useful properties of convex sets. The first result gives an alternative description of the relative interior of a convex set and the second one permits to embed a convex set with an empty interior into a lower dimensional affine space.

Lemma 1.2.1. *Let $C \subseteq \mathbb{R}^n$ be a convex set. A point $x \in C$ lies in the relative interior of C if and only if*

$$\forall y \in C\ \exists z \in C, \alpha \in (0,1) : x = \alpha y + (1-\alpha)z,$$

where $(0,1)$ denotes the open interval $0 < \alpha < 1$.

Theorem 1.2.2. *Let $C \subseteq \mathbb{R}^n$ be a convex set. If $\operatorname{int} C = \varnothing$ then the dimension of its affine closure is at most $n-1$.*

1.3 Implicit description of convex sets

In this section we show how one can describe a closed convex set implicitly as the intersection of halfspaces (Theorem 1.3.7). For this we show the intuitive fact that through every of its boundary points there is a hyperplane which has the convex set on only one of its sides (Lemma 1.3.5). We also prove an important fact which we will need later: Any two convex sets whose relative interiors do not intersect can be properly separated by a hyperplane (Theorem 1.3.8). After giving the definitions of separating and supporting hyperplanes we look

at the metric projection which is a useful tool to construct these separating hyperplanes.

The *hyperplane* at a point $x \in \mathbb{R}^n$ with normal vector $c \in \mathbb{R}^n \backslash \{0\}$ is

$$H = \{y \in \mathbb{R}^n : c^\mathsf{T} y = c^\mathsf{T} x\}.$$

It is an affine subspace of dimension $n-1$. The hyperplane H divides $\mathbb{R}^n$ into two closed *halfspaces*

$$H^+ = \{y \in \mathbb{R}^n : c^\mathsf{T} y \geqslant c^\mathsf{T} x\}, \quad H^- = \{y \in \mathbb{R}^n : c^\mathsf{T} y \leqslant c^\mathsf{T} x\}.$$

A hyperplane H is said to *separate* two sets $A \subseteq \mathbb{R}^n$ and $B \subseteq \mathbb{R}^n$ if they lie on different sides of the hyperplane, i.e., if $A \subseteq H^+$ and $B \subseteq H^-$ or conversely. In other words, A and B are separated by a hyperplane if there exists a non-zero vector $c \in \mathbb{R}^n$ and a scalar $\beta \in \mathbb{R}$ such that

$$\forall x \in A, y \in B : c^\mathsf{T} x \leqslant \beta \leqslant c^\mathsf{T} y.$$

The separation is said to be *strict* if both inequalities are strict, i.e.,

$$\forall x \in A, y \in B : c^\mathsf{T} x < \beta < c^\mathsf{T} y.$$

The separation is said to be *proper* when H separates A and B but does not contain both A and B.

A hyperplane H is said to *support* A at $x \in A$ if $x \in H$ and if A is contained in one of the two halfspaces H^+ or H^-, say H^-. Then H is a *supporting hyperplane* of A at x and H^- is a supporting halfspace.

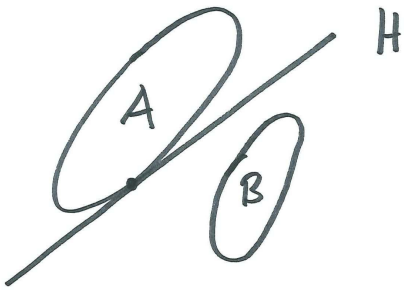


Figure 1.4: The hyperplane H supports A and separates A and B.

1.3.1 Metric projection

Let $C \in \mathbb{R}^n$ be a non-empty closed convex set. One can project every point $x \in \mathbb{R}^n$ onto C by simply taking the point in C which is closest to it. This fact is very intuitive and in the case when C is a linear subspace we are talking simply about the orthogonal projection onto C.

Lemma 1.3.1. *Let C be a non-empty closed convex set in $\mathbb{R}^n$. Let $x \in \mathbb{R}^n \backslash C$ be a point outside of C. Then there exists a unique point $\pi_C(x)$ in C which is closest to x. Moreover, $\pi_C(x) \in \partial C$.*

Proof. The argument for *existence* is a compactness argument: As C is not empty, pick $z_0 \in C$ and consider the intersection C' of C with the ball $B(z_0, r)$ centered at z_0 and with radius $r = \|z_0 - x\|$. Then C' is closed, convex and bounded. Moreover the minimum of the distance $\|y-x\|$ for $y \in C$ is equal to the minimum taken over C'. As we minimize a continuous function over a compact set, the minimum is attained. Hence there is at least one closest point to x in C.

The argument for *uniqueness* requires convexity: Let y and z be two distinct points in C, both having minimum distance to x. In this case, the midpoint of y and z, which lies in C, would even be closer to x, because the distance $d(x, \frac{1}{2}(y+z))$ is the height of the isosceles triangle with vertices x, y, z.

Hence there is a unique point in C which is at minimum distance to x, which we denote by $\pi_C(x)$. Clearly, $\pi_C(x) \in \partial C$, otherwise one would find another point in C closer to x lying in some small ball $B(\pi_C(x), \varepsilon) \subseteq C$. □

Thus, the map $\pi_C : \mathbb{R}^n \to C$ defined by the property

$$\forall y \in C : d(y, x) \geqslant d(\pi_C(x), x)$$

is well-defined. This map is called *metric projection* and sometimes we refer to the vector $\pi_C(x)$ as the *best approximation* of x in the set C.

The metric projection π_C is a contraction:

Lemma 1.3.2. *Let C be a non-empty closed and convex set in $\mathbb{R}^n$. Then,*

$$\forall x, y \in \mathbb{R}^n : d(\pi_C(x), \pi_C(y)) \leqslant d(x, y).$$

In particular, the metric projection π_C is a Lipschitz continuous map.

Proof. We can assume that $d(\pi_C(x), \pi_C(y)) \neq 0$. Consider the line segment $[\pi_C(x), \pi_C(y)]$ and the two parallel hyperplanes H_x and H_y at $\pi_C(x)$ and at $\pi_C(y)$ both having normal vector $\pi_C(x) - \pi_C(y)$. The points x and $\pi_C(y)$ are separated by H_x because otherwise there would be a point in $[\pi_C(x), \pi_C(y)] \subseteq C$ which is closer to x than to $\pi_C(x)$, which is impossible. In the same way, y and $\pi_C(x)$ are separated by H_y. Hence, x and y are on different sides of the "slab" bounded by the parallel hyperplanes H_x and by H_y. So their distance $d(x, y)$ is at least the width of the slab, which is $d(\pi_C(x), \pi_C(y))$. □

The metric projection can reach every point on the boundary of C:

Lemma 1.3.3. *Let C be a non-empty closed and convex set in $\mathbb{R}^n$. Then, for every boundary point $y \in \partial C$ there is a point x lying outside of C so that $y = \pi_C(x)$.*

Proof. First note that one can assume that C is bounded (since otherwise replace C by its intersection with a ball around y). Since C is bounded it is contained in a ball B of sufficiently large radius. We will construct the desired

point x which lies on the boundary ∂B by a limit argument. For this choose a sequence of points $y_i \in \mathbb{R}^n \backslash C$ such that $d(y, y_i) < 1/i$, and hence $\lim_{i\to\infty} y_i = y$. Because the metric projection is a contraction (Lemma 1.3.2) we have

$$d(y, \pi_C(y_i)) = d(\pi_C(y), \pi_C(y_i)) \leqslant d(y, y_i) < 1/i.$$

By intersecting the line $\mathrm{aff}\{y_i, \pi_C(y_i)\}$ with the boundary ∂B one can determine a point $x_i \in \partial B$ so that $\pi_C(x_i) = \pi_C(y_i)$. Since the boundary ∂B is compact there is a convergent subsequence (x_{i_j}) having a limit $x \in \partial B$. Then, because of the previous considerations and because π_C is continuous

$$\begin{aligned} y = \pi_C(y) = \pi_C\left(\lim_{j\to\infty} y_{i_j}\right) = \lim_{j\to\infty} \pi_C(y_{i_j}) \\ = \lim_{j\to\infty} \pi_C(x_{i_j}) = \pi_C\left(\lim_{j\to\infty} x_{i_j}\right) = \pi_C(x), \end{aligned}$$

which proves the lemma. □

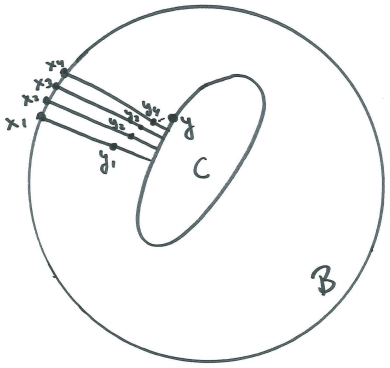


Figure 1.5: The construction which proves Lemma 1.3.3.

1.3.2 Separating and supporting hyperplanes

One can use the metric projection to construct separating and supporting hyperplanes:

Lemma 1.3.4. *Let C be a non-empty closed convex set in $\mathbb{R}^n$. Let $x \in \mathbb{R}^n \backslash C$ be a point outside C and let $\pi_C(x)$ its closest point in C. Then the following holds.*

(i) The hyperplane through x with normal $x - \pi_C(x)$ supports C at $\pi_C(x)$ and thus it separates $\{x\}$ and C.

(ii) The hyperplane through $(x + \pi_C(x))/2$ with normal $x - \pi_C(x)$ strictly separates $\{x\}$ and C.

Proof. It suffices to prove (i) and then (ii) follows directly. Consider the hyperplane H through x with normal vector $c = x - \pi_C(x)$, defined by

$$H = \{y \in \mathbb{R}^n : c^\mathsf{T} y = c^\mathsf{T} \pi_C(x)\}.$$

As $c^\mathsf{T} x > c^\mathsf{T} \pi_C(x)$, x lies in the open halfspace $\{y : c^\mathsf{T} y > c^\mathsf{T} \pi_C(x)\}$. We show that C lies in the closed halfspace $\{y : c^\mathsf{T} y \leqslant c^\mathsf{T} \pi_C(x)\}$. Suppose for a contradiction that there exists $y \in C$ such that $c^\mathsf{T}(y - \pi_C(x)) > 0$. Then select a scalar $\lambda \in (0, 1)$ such that $0 < \lambda < \frac{2c^\mathsf{T}(y - \pi_C(x))}{\|y - \pi_C(x)\|^2} < 1$ and set $w = \lambda y + (1-\lambda)\pi_C(x)$ which is a point C. Now verify that $\|w - x\| < \|\pi_C(x) - x\| = \|c\|$, which follows from

$$\|w - x\|^2 = \|\lambda(y - \pi_C(x)) - c\|^2 = \|c\|^2 + \lambda^2 \|y - \pi_C(x)\|^2 - 2\lambda c^\mathsf{T}(y - \pi_C(x))$$

and which contradicts the fact that $\pi_C(x)$ is the closest point in C to x. □

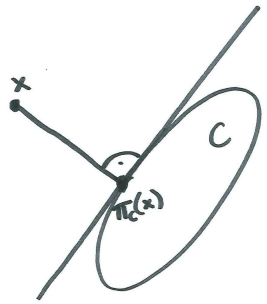


Figure 1.6: A separating hyperplane constructed using π_C.

Combining Lemma 1.3.3 and Lemma 1.3.4 we deduce that one can construct a supporting hyperplane at every boundary point.

Lemma 1.3.5. *Let $C \subseteq \mathbb{R}^n$ be a closed convex set and let $x \in \partial C$ be a point lying on the boundary of C. Then there is a hyperplane which supports C at x.*

One can generalize Lemma 1.3.4 (i) and remove the assumption that C is closed.

Lemma 1.3.6. *Let $C \subseteq \mathbb{R}^n$ be a non-empty convex set and let $x \in \mathbb{R}^n \backslash C$ be a point lying outside C. Then, $\{x\}$ and C can be separated by a hyperplane.*

Proof. In view of Lemma 1.3.1 we only have to show the result for non-closed convex sets C. We are left with two cases: If $x \notin \overline{C}$, then a hyperplane separating $\{x\}$ and the closed and convex set $\overline{C}$ also separates $\{x\}$ and C. If $x \in \overline{C}$, then $x \in \partial\overline{C}$. By Lemma 1.3.5 there is a hyperplane supporting $\overline{C}$ at x. In particular, it separates $\{x\}$ and C. □

As a direct application of the strict separation result in Lemma 1.3.4 (ii), we can formulate the following fundamental structural result for closed convex sets.

Theorem 1.3.7. *A non-empty closed convex set is the intersection of its supporting halfspaces.*

This is an implicit description as it gives a method to verify whether a point belongs to the closed convex set in question: One has to check whether the point lies in all these supporting halfspaces. If the closed convex set is given as an intersection of finitely many halfspaces, then it is called a *polyhedron* and the test we just described is a simple algorithmic membership test.

We conclude with the following result which characterizes when two convex sets can be separated properly. When both sets are closed and one of them is bounded, one can show a strict separation. These separation results will be the basis in our discussion of the duality theory of conic programs.

Theorem 1.3.8. *Let $C, D \subseteq \mathbb{R}^n$ be non-empty convex sets.*

(i) C and D can be properly separated if and only if their relative interiors do not have a point in common: $\operatorname{relint} C \cap \operatorname{relint} D = \emptyset$.

(ii) Assume that C and D are closed and that at least one of them is bounded. If $C \cap D = \emptyset$, then there is a hyperplane strictly separating C and D.

Proof. (i) *The "only if" part ($\Longrightarrow$):* Let $H_{c,\beta}$ be a hyperplane properly separating C and D with $C \subseteq H^-$ and $D \subseteq H^+$, i.e.,

$$\forall x \in C, y \in D : c^\mathsf{T} x \leqslant \beta \leqslant c^\mathsf{T} y.$$

Suppose there is a point $x_0 \in \operatorname{relint} C \cap \operatorname{relint} D$. Then $c^\mathsf{T} x_0 = \beta$, i.e., $x_0 \in H$. Pick any $x \in C$. By Lemma 1.2.1 there exists $x' \in C$ and $\alpha \in (0,1)$ such that $x_0 = \alpha x + (1-\alpha)x'$. Now

$$\beta = c^\mathsf{T} x_0 = \alpha c^\mathsf{T} x + (1-\alpha) c^\mathsf{T} x' \leqslant \alpha\beta + (1-\alpha)\beta = \beta,$$

hence all inequalities have to be tight and so $c^\mathsf{T} x = \beta$. Thus C is contained in the hyperplane H. Similarly, $D \subseteq H$. This contradicts the assumption that the separation is proper.

The "if part" ($\Longleftarrow$): Consider the set

$$E = \operatorname{relint} C - \operatorname{relint} D = \{x - y : x \in \operatorname{relint} C, y \in \operatorname{relint} D\},$$

which is convex. By assumption, the origin 0 does not lie in E. By Lemma 1.3.6 there is a hyperplane H separating $\{0\}$ and E which goes through the origin. Say $H = H_{c,0}$ and

$$\forall x \in \operatorname{relint} C, y \in \operatorname{relint} D : c^\mathsf{T}(x - y) \geqslant 0.$$

Define

$$\beta = \inf\{c^{\mathsf{T}} x : x \in \operatorname{relint} C\}.$$

Then,

$$C \subseteq \{x \in \mathbb{R}^n : c^{\mathsf{T}} x \geqslant \beta\},$$

and we want to show that

$$D \subseteq \{y : c^{\mathsf{T}} y \leqslant \beta\}.$$

For suppose not. Then there is a point $y \in \operatorname{relint} D$ so that $c^{\mathsf{T}} y > \beta$. Moreover, by definition of the infimum there is a point $x \in \operatorname{relint} C$ so that $\beta \leqslant c^{\mathsf{T}} x < c^{\mathsf{T}} y$. But then we find $c^{\mathsf{T}}(x - y) < 0$, a contradiction. Thus, C and D are separated by the hyperplane $H_{c,\beta}$.

If $C \cup D$ lies in some lower dimensional affine subspace, then the argument above gives a hyperplane in the affine subspace $\operatorname{aff}(C \cup D)$ which can be extended to a hyperplane in $\mathbb{R}^n$ which properly separates C and D.

(ii) Assume that C is bounded and $C \cap D = \varnothing$. Consider now the set

$$E = C - D$$

which is closed (check it) and convex. As the origin 0 does not lie in E, by Lemma 1.3.4 (ii), there is a hyperplane strictly separating $\{x\}$ and E: There is a non-zero vector c and a positive scalar β such that

$$\forall x \in C, y \in D : c^{\mathsf{T}}(x - y) > \beta > 0.$$

This implies

$$\inf_{x \in C} c^{\mathsf{T}} x \geqslant \beta + \sup_{y \in D} c^{\mathsf{T}} y > \frac{\beta}{2} + \sup_{y \in D} c^{\mathsf{T}} y > \sup_{y \in D} c^{\mathsf{T}} y.$$

Hence the hyperplane $H_{c,\alpha}$ with $\alpha = \frac{\beta}{2} + \sup_{y \in D} c^{\mathsf{T}} y$ strictly separates C and D. □

1.4 Explicit description of convex sets

Now we turn to an explicit description of convex sets. An explicit description gives an easy way to generate points lying in the convex set.

We say that a point $x \in C$ is *extreme* if it is not a relative interior point of any line segment in C. In other words, if x cannot be written in the form $x = (1 - \alpha)y + \alpha z$ with $y, z \in C$ and $0 < \alpha < 1$. The set of all extreme points of C we denote by $\operatorname{ext} C$.

Theorem 1.4.1. *Let $C \subseteq \mathbb{R}^n$ be a compact and convex set. Then,*

$$C = \operatorname{conv}(\operatorname{ext} C).$$

Proof. We prove the theorem by induction on the dimension n. If $n = 0$, then C is a point and the result follows.

Let the dimension n be at least one. If the interior of C is empty, then C lies in an affine subspace of dimension at most $n - 1$ and the theorem follows from the induction hypothesis. Suppose that $\operatorname{int} C \neq \varnothing$. We have to show that every $x \in C$ can be written as the convex hull of extreme points of C. We distinguish between two cases:

First case: If x lies on the boundary of C, then by Lemma 1.3.5 there is a supporting hyperplane H of C through x. Consider the set $F = H \cap C$. This is a compact and convex set which lies in an affine subspace of dimension at most $n - 1$ and hence we have by the induction hypotheses $x \in \operatorname{conv}(\operatorname{ext} F)$. Since $\operatorname{ext} F \subseteq \operatorname{ext} C$, we are done.

Second case: If x does not lie on the boundary of C, then the intersection of a line through x with C is a line segment $[y, z]$ with $y, z \in \partial C$. By the previous argument we have $y, z \in \operatorname{conv}(\operatorname{ext} C)$. Since x is a convex combination of y and z, the theorem follows. □

1.5 Convex cones

We will develop the theory of convex optimization using the concept of conic programs. Before we can say what a "conic program" is, we have to define convex cones.

Definition 1.5.1. *A non-empty subset K of $\mathbb{R}^n$ is called a* convex cone *if it is closed under non-negative linear combinations:*

$$\forall \alpha, \beta \in \mathbb{R}_{\geqslant 0}\ \forall x, y \in K : \alpha x + \beta y \in K.$$

Moreover, K is pointed *if*

$$x, -x \in K \implies x = 0.$$

One can easily check that convex cones are indeed convex sets. Furthermore, the *direct product*

$$K \times K' = \{(x, x') \in \mathbb{R}^{n+n'} : x \in K, x' \in K'\}$$

of two convex cones $K \subseteq \mathbb{R}^n$ and $K' \subseteq \mathbb{R}^{n'}$ is a convex cone again.

A pointed convex cone in $\mathbb{R}^n$ defines a *partial order* on $\mathbb{R}^n$ by

$$x \succeq y \iff x - y \in K$$

for $x, y \in \mathbb{R}^n$. This partial order satisfies the following conditions:

reflexivity:

$$\forall x \in \mathbb{R}^n : x \succeq x$$

antisymmetry:

$$\forall x, y \in \mathbb{R}^n : x \succeq y, y \succeq x \Longrightarrow x = y$$

transitivity:

$$\forall x, y, z \in \mathbb{R}^n : x \succeq y, y \succeq z \Longrightarrow x \succeq z$$

homogenity:

$$\forall x, y \in \mathbb{R}^n \ \forall \alpha \in \mathbb{R}_{\geqslant 0} : x \succeq y \Longrightarrow \alpha x \succeq \alpha y$$

additivity:

$$\forall x, y, x', y' \in \mathbb{R}^n : x \succeq y, x' \succeq y' \Longrightarrow x + x' \succeq y + y'.$$

In order that a convex cone is useful for practical algorithmic optimization methods we will need two additional properties to eliminate undesired degenerate conditions: A convex cone should be closed and full-dimensional, that is, it has a non-empty interior. Then, we define strict inequalities by:

$$x \succ y \Longleftrightarrow x - y \in \operatorname{int} K.$$

Let $(x_i)_{i \in \mathbb{N}}$ and $(y_i)_{i \in \mathbb{N}}$ be sequences of elements in $\mathbb{R}^n$ which have limits x and y, then we can pass to limits in the inequalities:

$$(\exists N \in \mathbb{N} \ \forall i \geqslant N : x_i \succeq y_i) \Longleftrightarrow x \succeq y.$$

The separation result from Lemma 1.3.4 specializes to convex cones in the following way.

Lemma 1.5.2. *Let $C \subseteq \mathbb{R}^n$ be a closed convex cone and let $x \in \mathbb{R}^n \backslash C$ be a point outside of C. Then there is a linear hyperplane separating $\{x\}$ and C. Even stronger, there is a non-zero vector $c \in \mathbb{R}^n$ such that*

$$\forall y \in C : c^{\mathsf{T}} y \geqslant 0 > c^{\mathsf{T}} x,$$

thus with the strict inequality $c^{\mathsf{T}} x < 0$.

1.6 Examples

The convex cone generated by a set of vectors $A \subseteq \mathbb{R}^n$ is the smallest convex cone containing A. It is

$$\operatorname{cone} A = \left\{ \sum_{i=1}^{N} \alpha_i x_i : N \in \mathbb{N}, x_1, \ldots, x_N \in A, \alpha_1, \ldots, \alpha_N \in \mathbb{R}_{\geqslant 0} \right\}.$$

Furthermore, every linear subspace of E is a convex cone, however a somewhat boring one. More interesting are the following examples. We will use them, especially cone of positive semidefinite matrices, very often.

1.6.1 The non-negative orthant and linear programming

The convex cone which is connected to linear programming is the non-negative orthant. It lies in the Euclidean space $\mathbb{R}^n$ with the standard inner product. The *non-negative orthant* is defined as

$$\mathbb{R}^n_{\geqslant 0} = \{x = (x_1, \ldots, x_n)^\mathsf{T} \in \mathbb{R}^n : x_1, \ldots, x_n \geqslant 0\}.$$

It is a pointed, closed and full-dimensional cone. A *linear program* is an optimization problem of the following form

$$\begin{aligned} \text{maximize} \quad & c_1 x_1 + \cdots + c_n x_n \\ \text{subject to} \quad & a_{11} x_1 + \cdots + a_{1n} x_n \geqslant b_1 \\ & a_{21} x_1 + \cdots + a_{2n} x_n \geqslant b_2 \\ & \qquad \vdots \\ & a_{m1} x_1 + \cdots + a_{mn} x_n \geqslant b_m. \end{aligned}$$

One can express the above linear program more conveniently using the partial order defined by the non-negative orthant $\mathbb{R}^n_{\geqslant 0}$:

$$\begin{aligned} &\text{maximize } c^\mathsf{T} x \\ &\text{subject to } Ax \succeq b, \end{aligned}$$

where $c = (c_1, \ldots, c_n)^\mathsf{T} \in \mathbb{R}^n$ is the *objective vector*, $A = (a_{ij}) \in \mathbb{R}^{m \times n}$ is the matrix of linear *constraints*, $x = (x_1, \ldots, x_n)^\mathsf{T} \in \mathbb{R}^n$ is the *optimization variable*, and $b = (b_1, \ldots, b_m)^\mathsf{T} \in \mathbb{R}^m$ is the *right hand side*. Here, the partial order $x \succeq y$ means inequality coordinate-wise: $x_i \geqslant y_i$ for all $i \in [n]$.

1.6.2 The second-order cone

While the non-negative orthant is a polyhedron, the following cone is not. The *second-order cone* is defined in the Euclidean space $\mathbb{R}^{n+1} = \mathbb{R}^n \times \mathbb{R}$ with the standard inner product. It is

$$\mathcal{L}^{n+1} = \left\{ (x, t) \in \mathbb{R}^n \times \mathbb{R} : \|x\|_2 = \sqrt{x_1^2 + \cdots + x_n^2} \leqslant t \right\}.$$

Sometimes it is also called *ice cream cone* (make a drawing to convince yourself) or *Lorentz cone*. The second-order cone will turn out to be connected to conic quadratic programming.

1.6.3 The cone of semidefinite matrices

The convex cone which will turn out to be connected to semidefinite programming is the cone of positive semidefinite matrices. It lies in the $n(n+1)/2$-dimensional Euclidean space of $n \times n$-symmetric matrices $\mathcal{S}^n$ with the trace

inner product. Namely, for two matrices $X, Y \in \mathbb{R}^{n\times n}$,

$$\langle X, Y\rangle = \text{Tr}(X^{\mathsf{T}}Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}Y_{ij}, \quad \text{where Tr}X = \sum_{i=1}^{n} X_{ii}.$$

Here we identify the Euclidean space $\mathcal{S}^n$ with $\mathbb{R}^{n(n+1)/2}$ by the isometry $T : \mathcal{S}^n \to \mathbb{R}^{n(n+1)/2}$ defined by

$$T(X) = (X_{11}, \sqrt{2}X_{12}, \sqrt{2}X_{13}, \ldots, \sqrt{2}X_{1n}, X_{22}, \sqrt{2}X_{23}, \ldots, \sqrt{2}X_{2n}, \ldots, X_{nn})$$

where we only consider the upper triangular part (in good old FORTRAN 77 tradition) of the matrix X.

The *cone of semidefinite matrices* is

$$\mathcal{S}^n_{\geq 0} = \{X \in \mathcal{S}^n : X \text{ is positive semidefinite}\},$$

where a matrix X is *positive semidefinite* if

$$\forall x \in \mathbb{R}^n : x^{\mathsf{T}}Xx \geqslant 0.$$

More characterizations are given in Section 1.7 below.

1.6.4 The copositive cone

The copositive cone is a cone in $\mathcal{S}^n$ which contains the semidefinite cone. It is the basis of copositive programming and it is defined as the set of all copositive matrices:

$$\mathcal{C}^n = \{X \in \mathcal{S}^n : \ x^{\mathsf{T}}Xx \geqslant 0 \ \ \forall x \in \mathbb{R}^n_{\geqslant 0}\}.$$

Unlike for the semidefinite cone no easy characterization (for example in terms of eigenvalues) of copositive matrices is known. Even stronger: Unless the complexity classes $\mathcal{P}$ and $\mathcal{NP}$ coincide no easy characterization (meaning one which is polynomial-time computable) exists.

1.7 Positive semidefinite matrices

1.7.1 Basic facts

A matrix P is *orthogonal* if $PP^{\mathsf{T}} = I_n$ or, equivalently, $P^{\mathsf{T}}P = I_n$, i.e. the rows (resp., the columns) of P form an orthonormal basis of $\mathbb{R}^n$. By $\mathcal{O}(n)$ we denote the set of $n \times n$ orthogonal matrices which forms a group under matrix multiplication.

The spectral decomposition theorem is probably the most important theorem about real symmetric matrices.

Theorem 1.7.1. (Spectral decomposition theorem) *Any real symmetric matrix $X \in \mathcal{S}^n$ can be decomposed as*

$$X = \sum_{i=1}^{n} \lambda_i u_i u_i^{\mathsf{T}},$$

where $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ are the eigenvalues of X and where $u_1, \ldots, u_n \in \mathbb{R}^n$ are the corresponding eigenvectors which form an orthonormal basis of $\mathbb{R}^n$. In matrix terms, $X = PDP^{\mathsf{T}}$, where D is the diagonal matrix with the λ_i's on the diagonal and P is the orthogonal matrix with the u_i's as its columns.

Theorem 1.7.2. (Positive semidefinite matrices) *Let $X \in \mathcal{S}^n$ be a symmetric matrix. The following assertions are equivalent.*

(1) X is positive semidefinite, written as $X \succeq 0$, which is defined by the property: $x^T X x \geqslant 0$ for all $x \in \mathbb{R}^n$.

(2) The smallest eigenvalue of X is non-negative, i.e., the spectral decomposition of X is of the form $X = \sum_{i=1}^{n} \lambda_i u_i u_i^T$ with all $\lambda_i \geqslant 0$.

(3) $X = LL^{\mathsf{T}}$ for some matrix $L \in \mathbb{R}^{n \times k}$ (for some $k \geqslant 1$), called a Cholesky decomposition *of X.*

(4) There exist vectors $v_1, \ldots, v_n \in \mathbb{R}^k$ (for some $k \geqslant 1$) such that $X_{ij} = v_i^{\mathsf{T}} v_j$ for all $i, j \in [n]$; the vectors v_i's are called a Gram representation *of X.*

(5) All principal minors of X are non-negative.

The set $\mathcal{S}^n_{\succeq 0}$ of all positive semidefinite matrices is a pointed, closed, convex, full-dimensional cone in $\mathcal{S}^n$. Moreover, it is generated by rank one matrices, i.e.

$$\mathcal{S}^n_{\succeq 0} = \mathrm{cone}\{xx^{\mathsf{T}} : x \in \mathbb{R}^n\}.$$

Matrices lying in the interior of the cone $\mathcal{S}^n_{\succeq 0}$ are called *positive definite*. The above result extends to positive definite matrices. A matrix X is positive definite (denoted as $X \succ 0$) if it satisfies any of the following equivalent properties: (1) $x^T X x > 0$ for all $x \in \mathbb{R}^n \backslash \{0\}$, (2) all eigenvalues are strictly positive, (3) in a Cholesky decomposition the matrix L is non-singular, (4) any Gram representation has full rank n, and (5) all the principal minors are positive (in fact already positivity of all the leading principal minors implies positive definiteness; Sylvester's criterion).

1.7.2 The trace inner product

The *trace* of an $n \times n$ matrix A is defined as $\mathrm{Tr}(A) = \sum_{i=1}^{n} A_{ii}$. The trace is a linear form on $\mathbb{R}^{n \times n}$ and satisfies the following properties: $\mathrm{Tr}(A) = \mathrm{Tr}(A^{\mathsf{T}})$, $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$. Moreover, if A is symmetric then the trace of A is equal to the sum of the eigenvalues of A.

One can define an inner product on $\mathbb{R}^{n\times n}$ by setting

$$\langle A, B\rangle = \text{Tr}(A^\mathsf{T} B) = \sum_{i,j=1}^{n} A_{ij}B_{ij}.$$

This defines the *Frobenius norm* on $\mathbb{R}^{n\times n}$ by setting $\|A\| = \sqrt{\langle A, A\rangle}$. For a vector $x \in \mathbb{R}^n$ we have $x^\mathsf{T} Ax = \langle A, xx^\mathsf{T}\rangle$. For positive semidefinite matrices we have the following result.

Lemma 1.7.3. *For a symmetric matrix $A \in \mathcal{S}^n$,*

$$A \succeq 0 \iff \forall B \in \mathcal{S}^n_{\geq 0} : \langle A, B\rangle \geqslant 0.$$

In other words, the cone $\mathcal{S}^n_{\geq 0}$ is self-dual: $(\mathcal{S}^n_{\geq 0})^* = \mathcal{S}^n_{\geq 0}$.

Proof. Direct verification using the conditions (1) and (2) in Theorem 1.7.2. □

1.7.3 Hoffman-Wielandt inequality

Here is a nice inequality to know about eigenvalues.

Theorem 1.7.4. *(**Hoffman, Wielandt (1953)**) Let $A, B \in \mathcal{S}^n$ be symmetric matrices with respective eigenvalues $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$ ordered as follows: $\alpha_1 \leqslant \ldots \leqslant \alpha_n$ and $\beta_1 \geqslant \ldots \geqslant \beta_n$. Then,*

$$\sum_{i=1}^{n} \alpha_i\beta_i = \min\{\text{Tr}(AXBX^T) : X \in \mathcal{O}(n)\}. \tag{1.1}$$

In particular,

$$\text{Tr}(AB) \geqslant \sum_{i=1}^{n} \alpha_i\beta_i.$$

Proof. Write $A = PDP^\mathsf{T}$ and $B = QEQ^\mathsf{T}$ where $P, Q \in \mathcal{O}(n)$ and D (resp., E) is the diagonal matrix with diagonal entries α_i (resp. β_i). As $\text{Tr}(AXBX^\mathsf{T}) = \text{Tr}(DYEY^\mathsf{T})$ where $Y = P^\mathsf{T}XQ \in \mathcal{O}(n)$, the optimization problem (1.1) is equivalent to

$$\min\{\text{Tr}(DXEX^\mathsf{T}) : X \in \mathcal{O}(n)\}.$$

We want to prove that the minimum is $\text{Tr}(DE)$, which is attained $X = I$. For this consider the linear program[1]

$$\max_{x,y\in\mathbb{R}^n} \left\{ \sum_{i=1}^{n} x_i + \sum_{j=1}^{n} y_j : \alpha_i\beta_j - x_i - y_j \geqslant 0 \ \forall i, j \in [n] \right\} \tag{1.2}$$

[1]We now use only the dual linear program (1.3), but we will use also the primal linear program (1.2) in Chapter 2 for reformulating program (1.1) as a a semidefinite program.

and its dual linear program

$$\min_{Z\in\mathbb{R}^{n\times n}} \left\{ \sum_{i,j=1}^{n} \alpha_i\beta_j Z_{ij} : \sum_{i=1}^{n} Z_{ij} = 1 \ \forall j \in [n],\ \sum_{j=1}^{n} Z_{ij} = 1 \ \forall i \in [n],\ Z \geqslant 0 \right\}. \tag{1.3}$$

Note that the feasible region of the linear program (1.3) is the set of all *doubly-stochastic matrices,* i.e., the matrices with non-negative entries where all rows and all columns sum up to one. By Birkhoff's theorem the set of doubly-stochastic matrices is equal to the convex hull of all permutation matrices. In other words, the minimum of (1.3) is equal to the minimum value of $\sum_{i=1}^{n} \alpha_i\beta_{\sigma(i)}$ taken over all permutations σ of $[n]$. It is an easy exercise to verify that this minimum is attained for the identity permutation. This shows that the optimum value of (1.3) (and thus of (1.2)) is equal to $\sum_i \alpha_i\beta_i = \mathrm{Tr}(DE)$.

Now pick any $X \in \mathcal{O}(n)$. Observe that the matrix $Z = ((X_{ij})^2)_{i,j=1}^n$ is doubly-stochastic (by the definition that X is orthogonal) and that

$$\mathrm{Tr}(DXEX^\mathsf{T}) = \sum_{i,j=1}^{n} \alpha_i\beta_j (X_{ij})^2,$$

which implies that $\mathrm{Tr}(DXEX^\mathsf{T})$ is at least the minimum $\mathrm{Tr}(DE)$ of program (1.3). This shows that the minimum of (1.1) is at least $\mathrm{Tr}(DE)$, which finishes the proof of the theorem. □

1.7.4 Schur complements

The following notion of *Schur complement* can be very useful for showing positive semidefiniteness.

Definition 1.7.5. (Schur complement) *Consider a symmetric matrix X in block form*

$$X = \begin{pmatrix} A & B \\ B^\mathsf{T} & C \end{pmatrix}, \tag{1.4}$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times l}$ and $C \in \mathbb{R}^{l\times l}$. Assume that A is non-singular. Then, the matrix $C - B^\mathsf{T}A^{-1}B$ is called the Schur complement *of A in X.*

Lemma 1.7.6. *Let $X \in \mathcal{S}^n$ be in block form (1.4) where A is non-singular. Then,*

$$X \succeq 0 \iff A \succeq 0 \text{ and } C - B^\mathsf{T}A^{-1}B \succeq 0.$$

Proof. The following identity holds:

$$X = P^\mathsf{T} \begin{pmatrix} A & 0 \\ 0 & C - B^\mathsf{T}A^{-1}B \end{pmatrix} P, \quad \text{where } P = \begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix}.$$

As P is non-singular, we deduce that $X \succeq 0$ if and only if $(P^{-1})^\mathsf{T}XP^{-1} \succeq 0$ which is thus equivalent to $A \succeq 0$ and $C - B^\mathsf{T}A^{-1}B \succeq 0$.

1.7.5 Block-diagonal matrices

Given matrices $X_1 \in \mathcal{S}^{n_1}, \ldots, X_r \in \mathcal{S}^{n_r}$, $X_1 \oplus \cdots \oplus X_r$ denotes the following block-diagonal matrix $X \in \mathcal{S}^n$, where $n = n_1 + \cdots + n_r$,

$$X = X_1 \oplus \cdots X_r = \begin{pmatrix} X_1 & 0 & \ldots & 0 \\ 0 & X_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & X_r \end{pmatrix}. \tag{1.5}$$

Then, X is positive semidefinite if and only if all the blocks $X_1, \ldots, X_r$ are positive semidefinite.

Given two sets of matrices $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{A} \oplus \mathcal{B}$ denotes the set of all matrices $X \oplus Y$, where $X \in \mathcal{A}$ and $Y \in \mathcal{B}$. Moreover, for an integer $m \geqslant 1$, $m\mathcal{A}$ denotes $\mathcal{A} \oplus \cdots \oplus \mathcal{A}$, the m-fold sum.

From an algorithmic point of view it is much more economical to deal with positive semidefinite matrices in block-form like (1.5).

For instance, if we have a set $\mathcal{A}$ of matrices that pairwise commute, then it is well known that they admit a common set of eigenvectors. In other words, there exists an orthogonal matrix $P \in \mathcal{O}(n)$ such that the matrices $P^{\mathsf{T}} X P$ are diagonal for all $X \in \mathcal{A}$.

In general one may use the following powerful result about C^*-algebras which permits to show that certain sets of matrices can be block-diagonalized.

Consider a non-empty set $\mathcal{A} \subseteq \mathbb{C}^{n \times n}$ of matrices. $\mathcal{A}$ is said to be a *C^*-algebra* if it satisfies the following conditions:

1. $\mathcal{A}$ is closed under matrix addition and multiplication, and under scalar multiplication.

2. For any matrix $A \in \mathcal{A}$, its conjugate transpose A^* also belongs to $\mathcal{A}$.

For instance, the full matrix algebra $\mathbb{C}^{n \times n}$ is a simple instance of C^*-algebra, and the algebra $\bigoplus_{i=1}^{r} m_i \mathbb{C}^{n_i \times n_i}$ as well, where n_i, m_i are integers. The following fundamental result shows that up to an orthogonal transformation this is the general form of a C^*-algebra.

Theorem 1.7.7. (Wedderburn-Artin theorem) *Assume $\mathcal{A}$ is a C^*-algebra of matrices in $\mathbb{C}^{n \times n}$ containing the identity matrix. Then there exists a unitary matrix P (i.e., such that $PP^* = I_n$) and integers $r, n_1, m_1, \ldots, n_r, m_r \geqslant 1$ such that the set $P^*\mathcal{A}P = \{P^*XP : X \in \mathcal{A}\}$ is equal to*

$$m_1 \mathbb{C}^{n_1 \times n_1} \oplus \ldots \oplus m_r \mathbb{C}^{n_r \times n_r}.$$

See e.g. the thesis of Gijswijt [4] for a detailed exposition and its use for bounding the size of error correcting codes in finite fields.

1.7.6 Kronecker and Hadamard products

Given two matrices $A = (A_{ij}) \in \mathbb{R}^{n\times m}$ and $B = (B_{hk}) \in \mathbb{R}^{p\times q}$, their *Kronecker product* is the matrix $A \otimes B \in \mathbb{R}^{np\times mq}$ with entries

$$A_{ih,jk} = A_{ij}B_{hk} \ \forall i \in [n], j \in [m], h \in [p],\ k \in [q].$$

It can also be seen as the $n \times m$ block matrix whose ij-th block is the $p \times q$ matrix $A_{ij}B$ for all $i \in [n], j \in [m]$.

This includes in particular defining the Kronecker product $u \otimes v \in \mathbb{R}^{np}$ of two vectors $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^p$, with entries $(u \otimes v)_{ih} = u_i v_h$ for $i \in [n], h \in [p]$.

Given two matrices $A, B \in \mathbb{R}^{n\times m}$, their *Hadamard product* is the matrix $A \circ B \in \mathbb{R}^{n\times m}$ with entries

$$(A \circ B)_{ij} = A_{ij}B_{ij} \ \forall i \in [n], j \in [m].$$

Note that $A \circ B$ coincides with the principle submatrix of $A \otimes B$ indexed by the subset of all 'diagonal' pairs of indices of the form (ii, jj) for $i \in [n], j \in [m]$.

Here are some (easy to verify) facts about these products, where the matrices and vectors have the appropriate sizes.

1. $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$.
2. In particular, $(A \otimes B)(u \otimes v) = (Au) \otimes (Bv)$.
3. Assume $A \in \mathcal{S}^n$ and $B \in \mathcal{S}^p$ have, respectively, eigenvalues $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_p$. Then $A \otimes B \in \mathcal{S}^{np}$ has eigenvalues $\alpha_i\beta_h$ for $i \in [n], h \in [p]$. In particular,
$$A, B \succeq 0 \Longrightarrow A \otimes B \succeq 0 \ \text{ and } \ A \circ B \succeq 0,$$
$$A \succeq 0 \Longrightarrow A^{\circ k} = ((A_{ij})^k) \succeq 0 \ \forall k \in \mathbb{N}.$$

□

1.8 Historical remarks

The history of convexity is astonishing: On the one hand, the notion of convexity is very natural and it can be found even in prehistoric arts. For instance, the Platonic solids are convex polyhedra and carved stone models of some of them were crafted by the late neolithic people of Scotland more than 4,000 years ago. For more information on the history, which unearthed some good hoax, see also John Baez' discussion of "Who discovered the icosahedron?" `http://math.ucr.edu/home/baez/icosahedron/`.

On the other hand, the first mathematician who realized how important convexity is as a geometric concept was the brilliant Hermann Minkowski (1864–1909) who in a series of very influential papers "Allgemeine Lehrsätze über die konvexen Polyeder" (1897), "Theorie der konvexen Körper, insbesondere

Begründung ihres Oberflächenbegriffs" (published posthumously) initiated the mathematical study of convex sets and their properties. All the results in this chapter on the implicit and the explicit representation of convex sets can be found there (although with different proofs).

Not much can be added to David Hilbert's (1862–1943) praise in his obituary of his close friend Minkowski:

> Dieser Beweis eines tiefliegenden zahlentheoretischen Satzes[2] ohne rechnerische Hilfsmittel wesentlich auf Grund einer geometrisch anschaulichen Betrachtung ist eine Perle Minkowskischer Erfindungskunst. Bei der Verallgemeinerung auf Formen mit n Variablen führte der Minkowskische Beweis auf eine natürlichere und weit kleinere obere Schranke für jenes Minimum M, als sie bis dahin Hermite gefunden hatte. Noch wichtiger aber als dies war es, daß der wesentliche Gedanke des Minkowskischen Schlußverfahrens nur die Eigenschaft des Ellipsoids, daß dasselbe eine konvexe Figur ist und einen Mittelpunkt besitzt, benutzte und daher auf beliebige konvexe Figuren mit Mittelpunkt übertragen werden konnte. Dieser Umstand führte Minkowski zum ersten Male zu der Erkenntnis, daß überhaupt der *Begriff des konvexen Körpers* ein fundamentaler Begriff in unserer Wissenschaft ist und zu deren fruchtbarsten Forschungsmitteln gehört.
>
> Ein konvexer (nirgends konkaver) Körper ist nach Minkowski als ein solcher Körper definiert, der die Eigenschaft hat, daß, wenn man zwei seiner Punkte in Auge faßt, auch die ganze geradlinige Strecke zwischen denselben zu dem Körper gehört.[3]

Until the end of the 1940s convex geometry was a small discipline in pure mathematics. This changed dramatically when in 1947 the breakthrough of general linear programming came. Then Dantzig formulated the linear programming problem and designed the simplex algorithm for solving it. Nowadays, convex geometry is an important toolbox for researchers, algorithm designers and practitioners in mathematical optimization.

---

[2] Hilbert is refering to Minkowski's lattice point theorem. It states that for any invertible matrix $A \in \mathbb{R}^{n\times n}$ defining a lattice $A\mathbb{Z}^n$ and any convex set in $\mathbb{R}^n$ which is symmetric with respect to the origin and with volume greater than $2^n \det(A)^2$ contains a non-zero lattice point.

[3] It is not easy to translate Hilbert's praise into English without losing its poetic tone, but here is an attempt. This proof of a deep theorem in number theory contains little calculation. Using chiefly geometry, it is a gem of Minkowski's mathematical craft. With a generalization to forms having n variables Minkowski's proof lead to an upper bound M which is more natural and also much smaller than the bound due to Hermite. More important than the result itself was his insight, namely that the only salient features of ellipsoids used in the proof were that ellipsoids are convex and have a center, thereby showing that the proof could be immediately generalized to arbitrary convex bodies having a center. This circumstances led Minkowski for the first time to the insight that the notion of a convex body is a fundamental and very fruitful notion in our scientific investigations ever since.

Minkowski defines a convex (nowhere concave) body as one having the property that, when one looks at two of its points, the straight line segment joining them entirely belongs to the body.

1.9 Further reading

Two very good books which emphasize the relation between convex geometry and optimization are by Barvinok [1] and by Gruber [5] (available online). Less optimization but more convex geometry is discussed in the little book of Bonnesen, Fenchel [3] and the encyclopedic book by Schneider [7]. The first one is now mainly interesting for historical reasons. Somewhat exceptional, and fun to read, is Chapter VII in the book of Berger [2] (available online) where he gives a panoramic view on the concept of convexity and its many relations to modern higher geometry.

Let us briefly mention connections to functional analysis. Rudin in his classical book "Functional analysis" discusses Theorem 1.3.8 and Theorem 1.4.1 in an infinite-dimensional setting. Although we will not need these more general theorems, they are nice to know.

The Hahn-Banach *separation theorem* is Theorem 3.4 in Rudin.

Theorem 1.9.1. *Suppose A and B are disjoint, nonempty, convex sets in a topological vector space X.*

(a) If A is open there exist $\Lambda \in X^*$ and $\gamma \in \mathbb{R}$ such that

$$\Re\Lambda x < \gamma \leqslant \Re\Lambda y$$

for every $x \in A$ and for every $y \in B$. (Here, $\Re z$ is the real part of the complex number z.)

(b) If A is compact, B is closed, and X is locally convex, there exist $\Lambda \in X^*$, $\gamma_1 \in \mathbb{R}$, $\gamma_2 \in \mathbb{R}$, such that

$$\Re\Lambda x < \gamma_1 < \gamma_2 < \Re\Lambda y$$

for every $x \in A$ and for every $y \in B$.

The Krein-Milman theorem is Theorem 3.23 in Rudin.

Theorem 1.9.2. *Suppose X is a topological vector space on which X^* separates points. If K is a nonempty compact convex set in X, then K is the closed convex hull of the set of its extreme points.*

In symbols, $K = \overline{\mathrm{conv}(\mathrm{ext}(K))}$.

In his blog "What's new?" Terry Tao [8] gives an insightful discussion of the finite-dimensional Hahn-Banach theorem.

The book "Matrix analyis" by Horn and Johnson [6] contains a wealth of very useful information, more than 70 pages, about positive definite matrices.

1.10 Exercises

1.1. Give a proof for the following statement:

Let $C \subseteq \mathbb{R}^n$ be a convex set. If $C \neq \varnothing$, then $\mathrm{relint}\, C \neq \varnothing$

1.2. Give a proof for the following statement:

Let $C \subseteq \mathbb{R}^n$ be a closed convex set and let $x \in \mathbb{R}^n \backslash C$ a point lying outside of C. A separating hyperplane H is defined in Lemma 1.3.4. Consider a point y on the line $\mathrm{aff}\{x, \pi_C(x)\}$ which lies on the same side of the separating hyperplane H as x. Then, $\pi_C(x) = \pi_C(y)$.

1.3. (a) Prove or disprove: Let $A \subseteq \mathbb{R}^n$ be a subset. Then,

$$\overline{\mathrm{conv}\, A} = \mathrm{conv}\, \overline{A}.$$

(b) Construct two convex sets $C, D \subseteq \mathbb{R}^2$ so that they can be separated by a hyperplane but which cannot be properly separated.

1.4. Show that the l_p^n unit ball

$$\left\{ (x_1, \ldots, x_n)^\mathsf{T} \in \mathbb{R}^n : \|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \leqslant 1 \right\}$$

is convex for $p = 1$, $p = 2$ and $p = \infty$ ($\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$). Determine the extreme points and determine a supporting hyperplane for every boundary point.

(*) What happens for the other p?

BIBLIOGRAPHY

[1] A. Barvinok, *A Course in Convexity*, American Mathematical Society, 2002.

[2] M. Berger, *Geometry revealed, a Jacob's ladder to modern higher geometry*, Springer, 2010.

http://www.springerlink.com/content/978-3-540-71132-2

[3] T. Bonnesen, W. Fenchel, *Theorie der konvexen Körper*, Springer, 1934.

[4] D.C. Gijswijt, *Matrix algebras and semidefinite programming techniques for codes*, Ph.D. thesis, University of Amsterdam, 2005.

http://arxiv.org/abs/1007.0906

[5] P.M. Gruber, *Convex and Discrete Geometry*, Springer, 2007.

http://www.springerlink.com/content/978-3-540-71132-2

[6] R.A. Horn, C.R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

[7] R. Schneider, *Convex bodies: the Brunn-Minkowski theory*, Cambridge University Press, 1993.

[8] T. Tao, *What's new? The Hahn-Banach theorem, Mengers theorem, and Hellys theorem*, 2007.

http://terrytao.wordpress.com/2007/11/30/
the-hahn-banach-theorem-mengers-theorem-and-hellys-theorem/

CHAPTER 2

SEMIDEFINITE PROGRAMS: BASIC FACTS AND EXAMPLES

In this chapter we introduce semidefinite programs, give some basic properties, and we present several problems that can be modeled as instances of semidefinite programs, arising from optimization, geometry and algebra.

For convenience we briefly recall some notation that we will use in this chapter. Most of it has already been introduced in Section 1.7. $\mathcal{S}^n$ denotes the set of symmetric $n \times n$ matrices. For a matrix $X \in \mathcal{S}^n$, $X \succeq 0$ means that X is positive semidefinite and $\mathcal{S}^n_{\succeq 0}$ is the cone of positive semidefinite matrices. Analogously, $X \succ 0$ means that X is positive definite and $\mathcal{S}^n_{\succ 0}$ is the open cone of positive definite matrices.

Throughout I_n (or simply I when the dimension is clear from the context) denotes the $n \times n$ identity matrix, e denotes the all-ones vector, i.e., $e = (1, \ldots, 1)^{\mathsf{T}} \in \mathbb{R}^n$, and $J_n = ee^{\mathsf{T}}$ (or simply J) denotes the all-ones matrix. The vectors $e_1, \ldots, e_n$ are the standard unit vectors in $\mathbb{R}^n$, and the matrices $E_{ij} = (e_i e_j^{\mathsf{T}} + e_j e_i^{\mathsf{T}})/2$ form the standard basis of $\mathcal{S}^n$. $\mathcal{O}(n)$ denotes the set of orthogonal matrices, where A is orthogonal if $AA^{\mathsf{T}} = I_n$ or, equivalently, $A^{\mathsf{T}}A = I_n$.

We consider the *trace inner product*: $\langle A, B \rangle = \mathrm{Tr}(A^{\mathsf{T}}B) = \sum_{i,j=1}^n A_{ij}B_{ij}$ for two matrices $A, B \in \mathbb{R}^{n \times n}$. Here $\mathrm{Tr}(A) = \langle I_n, A \rangle = \sum_{i=1}^n A_{ii}$ denotes the trace of A. Recall that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$; in particular, $\langle QAQ^{\mathsf{T}}, QBQ^{\mathsf{T}} \rangle = \langle A, B \rangle$ if Q is an orthogonal matrix. A well known property of the positive semidefinite cone $\mathcal{S}^n_{\succeq 0}$ is that it is self-dual: for a matrix $X \in \mathcal{S}^n$, $X \succeq 0$ if and only if $\langle X, Y \rangle \geqslant 0$ for all $Y \in \mathcal{S}^n_{\succeq 0}$.

For a matrix $A \in \mathcal{S}^n$, $\mathrm{diag}(A)$ denotes the vector in $\mathbb{R}^n$ with entries are the diagonal entries of A and, for a vector $a \in \mathbb{R}^n$, $\mathrm{Diag}(a) \in \mathcal{S}^n$ is the diagonal matrix with diagonal entries the entries of a.

2.1 Primal and dual semidefinite programs

2.1.1 Primal form

The typical form of a semidefinite program (often abbreviated as SDP) is a maximization problem of the form

$$p^* = \sup_X \{\langle C, X\rangle : \langle A_j, X\rangle = b_j \ (j \in [m]),\ X \succeq 0\}. \tag{2.1}$$

Here $A_1, \ldots, A_m \in \mathcal{S}^n$ are given $n \times n$ symmetric matrices and $b \in \mathbb{R}^m$ is a given vector, they are the *data* of the semidefinite program (2.1). The matrix X is the *variable*, which is constrained to be positive semidefinite and to lie in the affine subspace

$$\mathcal{W} = \{X \in \mathcal{S}^n \mid \langle A_j, X\rangle = b_j \ (j \in [m])\}$$

of $\mathcal{S}^n$. The goal is to maximize the linear objective function $\langle C, X\rangle$ over the *feasible region*

$$\mathcal{F} = \mathcal{S}^n_{\succeq 0} \cap \mathcal{W},$$

obtained by intersecting the positive semidefinite cone $\mathcal{S}^n_{\succeq 0}$ with the affine subspace $\mathcal{W}$.

A feasible solution $X \in \mathcal{F}$ is said to be *strictly feasible* if X is positive definite. The program (2.1) is said to be *strictly feasible* if it admits at least one strictly feasible solution.

One can also handle minimization problems, of the form

$$\inf_X \{\langle C, X\rangle : \langle A_j, X\rangle = b_j \ (j \in [m]),\ X \succeq 0\}$$

since they can be brought into the above standard maximization form using the fact that $\inf\langle C, X\rangle = -\sup\langle -C, X\rangle$.

Note that we write a *supremum* in (2.1) rather than a *maximum*. This is because the optimum value p^* might not be attained in (2.1). In general, $p^* \in \mathbb{R} \cup \{\pm\infty\}$, with $p^* = -\infty$ if the problem (2.1) is infeasible (i.e., $\mathcal{F} = \varnothing$) and $p^* = +\infty$ might occur in which case we say that the problem is unbounded.

We give a small example as an illustration. Consider the problem of minimizing/maximizing X_{11} over the feasible region

$$\mathcal{F}_a = \left\{X \in \mathcal{S}^2 : X = \begin{pmatrix} X_{11} & a \\ a & 0 \end{pmatrix} \succeq 0\right\} \text{ where } a \in \mathbb{R} \text{ is a given parameter.}$$

Note that $\det(X) = -a^2$ for any $X \in \mathcal{F}_a$. Hence, if $a \neq 0$ then $\mathcal{F}_a = \varnothing$ (the problem is infeasible). Moreover, if $a = 0$ then the problem is feasible but not strictly feasible. The minimum value of X_{11} over $\mathcal{F}_0$ is equal to 0, attained at $X = 0$, while the maximum value of X_{11} over $\mathcal{F}_0$ is equal to ∞ (the problem is unbounded).

As another example, consider the problem

$$p^* = \inf_{X \in \mathcal{S}^2} \left\{X_{11} : \begin{pmatrix} X_{11} & 1 \\ 1 & X_{22} \end{pmatrix} \succeq 0\right\}.$$

Then the infimum is $p^* = 0$ which is reached at the limit when $X_{11} = 1/X_{22}$ and letting X_{22} tend to ∞. So the infimum is not attained.

In the special case when the matrices A_j, C are diagonal matrices, with diagonals $a_j, c \in \mathbb{R}^n$, then the program (2.1) reduces to the linear program (LP):

$$\max \left\{ c^\mathsf{T} x : a_j^\mathsf{T} x = b_j \ (j \in [m]),\ x \geqslant 0 \right\}.$$

Indeed, let x denote the vector consisting of the diagonal entries of the matrix X, so that $x \geqslant 0$ if $X \succeq 0$, and $\langle C, X \rangle = c^\mathsf{T} x$, $\langle A_j, X \rangle = a_j^\mathsf{T} x$. Hence semidefinite programming contains linear programming as a special instance.

2.1.2 Dual form

The program (2.1) is often referred to as the *primal SDP* in standard form. One can define its *dual SDP*, which takes the form:

$$d^* = \inf_y \left\{ \sum_{j=1}^m b_j y_j = b^\mathsf{T} y : \sum_{j=1}^m y_j A_j - C \succeq 0 \right\}. \tag{2.2}$$

Thus the dual program has variables y_j, one for each linear constraint of the primal program. The positive semidefinite constraint arising in (2.2) is also named a *linear matrix inequality (LMI)*. The following facts relate the primal and dual SDP's. They are simple, but very important.

Lemma 2.1.1. *Let (X, y) be a primal/dual pair of feasible solutions, i.e., X is a feasible solution of (2.1) and y is a feasible solution of (2.2).*

1. **(weak duality)** *We have that $\langle C, X \rangle \leqslant b^\mathsf{T} y$ and thus $p^* \leqslant d^*$.*
2. **(complementary slackness)** *Assume that the primal program attains its supremum at X, that the dual program attains its infimum at y, and that $p^* = d^*$. Then the equalities $\langle C, X \rangle = b^\mathsf{T} y$ and $\langle X, \sum_{j=1}^m y_j A_j - C \rangle = 0$ hold.*
3. **(optimality criterion)** *If equality $\langle C, X \rangle = b^\mathsf{T} y$ holds, then the supremum of (2.1) is attained at X, the infimum of (2.2) is attained at y and $p^* = d^*$.*

Proof. If (X, y) is a primal/dual pair of feasible solutions, then

$$0 \leqslant \langle X, \sum_j y_j A_j - C \rangle = \sum_j \langle X, A_j \rangle y_j - \langle X, C \rangle = \sum_j b_j y_j - \langle X, C \rangle = b^\mathsf{T} y - \langle C, X \rangle.$$

The left most inequality follows from the fact that both X and $\sum_j y_j A_j - C$ are positive semidefinite and we use the fact that $\langle A_j, X \rangle = b_j$ to get the second equality. This implies that

$$\langle C, X \rangle \leqslant p^* \leqslant d^* \leqslant b^\mathsf{T} y.$$

The rest of the lemma follows by direct verification. □

The quantity $d^* - p^*$ is called the *duality gap*. In general there might be a positive duality gap between the primal and dual SDP's. When there is no duality gap, i.e., $p^* = d^*$, one says that *strong duality* holds, a very desirable sitiuation. This topic and criteria for strong duality will be discussed in detail in the next chapter. For now we only quote the following result on strong duality which will be proved in the next chapter (in the general setting of conic programming).

Theorem 2.1.2. (Strong duality: no duality gap) *Consider the pair of primal and dual programs (2.1) and (2.2).*

1. *Assume that the dual program (2.2) is bounded from below ($d^* > -\infty$) and that it is strictly feasible. Then the primal program (2.1) attains its supremum (i.e., $p^* = \langle C, X \rangle$ for some $X \in \mathcal{F}$) and there is no duality gap: $p^* = d^*$.*

2. *Assume that the primal program (2.1) is bounded from above ($p^* < \infty$) and that it is strictly feasible. Then the dual program (2.2) attains its infimum (i.e., $d^* = b^{\mathsf{T}} y$ for some dual feasible y) and there is no duality gap: $p^* = d^*$.*

In the rest of this chapter we discuss several examples of semidefinite programs.

2.2 Eigenvalue optimization

Given a matrix $C \in \mathcal{S}^n$, let $\lambda_{\min}(C)$ (resp., $\lambda_{\max}(C)$) denote its smallest (resp., largest) eigenvalue. One can express them (please check it) as follows:

$$\lambda_{\max}(C) = \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{x^{\mathsf{T}} C x}{\|x\|^2} = \max_{x \in \mathbb{S}^{n-1}} x^{\mathsf{T}} C x, \tag{2.3}$$

where $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n \mid \|x\| x = 1\}$ denotes the unit sphere in $\mathbb{R}^n$, and

$$\lambda_{\min}(C) = \min_{x \in \mathbb{R}^n \setminus \{0\}} \frac{x^{\mathsf{T}} C x}{\|x\|^2} = \min_{x \in \mathbb{S}^{n-1}} x^{\mathsf{T}} C x. \tag{2.4}$$

(This is known as the Rayleigh principle.) As we now see the largest and smallest eigenvalues can be computed via a semidefinite program. Namely, consider the semidefinite program

$$p^* = \sup \{\langle C, X \rangle : \mathrm{Tr}(X) = \langle I, X \rangle = 1, X \succeq 0\} \tag{2.5}$$

and its dual program

$$d^* = \inf_{y \in \mathbb{R}} \{y : yI - C \succeq 0\}. \tag{2.6}$$

In view of (2.3), we have that $d^* = \lambda_{\max}(C)$. The feasible region of (2.5) is bounded (all entries of any feasible X lie in $[0, 1]$) and contains a positive definite matrix (e.g., the matrix I_n/n), hence the infimum is attained in (2.6).

Analogously, the program (2.6) is bounded from below (as $y \geqslant \lambda_{\max}(C)$ for any feasible y) and strictly feasible (pick y large enough), hence the infimum is attained in (2.6). Moreover there is no duality gap: $p^* = d^*$. Here we have applied Theorem 2.1.2. Thus we have shown:

Lemma 2.2.1. *The largest and smallest eigenvalues of a symmetric matrix $C \in \mathcal{S}^n$ can be expressed with the following semidefinite programs:*

$$\begin{array}{rlll} \lambda_{\max}(C) = & \max & \langle C, X\rangle & \\ & \textit{s.t.} & \mathrm{Tr}(X) = 1, X \succeq 0 & \end{array} \quad = \quad \begin{array}{ll} \min & y \\ \textit{s.t.} & yI_n - C \succeq 0 \end{array}$$

$$\begin{array}{rlll} \lambda_{\min}(C) = & \min & \langle C, X\rangle & \\ & \textit{s.t.} & \mathrm{Tr}(X) = 1, X \succeq 0 & \end{array} \quad = \quad \begin{array}{ll} \max & y \\ \textit{s.t.} & C - yI_n \succeq 0 \end{array}$$

More generally, also the sum of the k largest eigenvalues of a symmetric matrix can be computed via a semidefinite program.

Theorem 2.2.2. (Fan's theorem) *Let $C \in \mathcal{S}^n$ be a symmetric matrix with eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_n$. Then the sum of its k largest eigenvalues is given by any of the following two programs:*

$$\lambda_1 + \cdots + \lambda_k = \max_{X \in \mathcal{S}^n} \left\{ \langle C, X\rangle : \mathrm{Tr}(X) = k,\ I_n \succeq X \succeq 0 \right\}, \tag{2.7}$$

$$\lambda_1 + \cdots + \lambda_k = \max_{Y \in \mathbb{R}^{n\times k}} \left\{ \langle C, YY^{\mathsf{T}}\rangle : Y^{\mathsf{T}}Y = I_k \right\}. \tag{2.8}$$

There is a simple, elegant proof for this result which relies on a geometric insight about the feasible regions of the two programs (2.7) and (2.8):

$$\mathcal{K}_1 = \{X \in \mathcal{S}^n \mid I \succeq X \succeq 0,\ \mathrm{Tr}(X) = k\}, \tag{2.9}$$

$$\mathcal{K}_2 = \{YY^{\mathsf{T}} \mid Y \in \mathbb{R}^{n\times k},\ Y^{\mathsf{T}}Y = I_k\}. \tag{2.10}$$

The (non-convex) set $\mathcal{K}_2$ consists of all projection matrices of rank k and is clearly contained in the (convex) set $\mathcal{K}_1$. As the next lemma shows, $\mathcal{K}_1$ coincides with the convex hull of $\mathcal{K}_2$.

Lemma 2.2.3. *$\mathcal{K}_2$ is the set of extreme points of the convex set $\mathcal{K}_1$. Therefore equality $\mathcal{K}_1 = \mathrm{conv}(\mathcal{K}_2)$ holds.*

Proof. The proof uses the following simple observation: For any orthogonal matrix $P \in \mathcal{O}(n)$, $X \in \mathcal{K}_1$ if and only if $PXP^{\mathsf{T}} \in \mathcal{K}_1$ and, moreover, X is an extreme point of $\mathcal{K}_1$ if and only if PXP^{T} is an extreme point of $\mathcal{K}_1$. This observation allows us to deal with diagonal matrices and to reduce the lemma to a claim about the extreme points of the following polytope:

$$\mathcal{P} = \{x \in [0,1]^n : e^{\mathsf{T}}x = k\}. \tag{2.11}$$

Indeed, consider $X \in \mathcal{S}^n$, written as $X = PDP^{\mathsf{T}}$, where $P \in \mathcal{O}(n)$, D is the diagonal matrix with the eigenvalues of X as diagonal entries, and define the

vector $d = \text{diag}(D) \in \mathbb{R}^n$. Then, X belongs to (resp., is an extreme point of) $\mathcal{K}_1$ if and only if D belongs to (resp., is an extreme point of) $\mathcal{K}_1$ or, equivalently, d belongs to (resp., is an extreme point of) $\mathcal{P}$.

Now it suffices to observe that the extreme points of the polytope $\mathcal{P}$ are the vectors $d \in \{0,1\}^n$ with $e^\mathsf{T} d = k$. This implies that X is an extreme point of $\mathcal{K}_1$ if and only if it has k non-zero eigenvalues, all equal to 1, which precisely means that $X \in \mathcal{K}_2$. □

We can now conclude the proof of Theorem 2.2.2.

Proof. (of Theorem 2.2.2). In program (2.8), we maximize the linear objective function $\langle C, X\rangle$ over the set $\mathcal{K}_2$, while in (2.7) we maximize it over the set K_1. In (2.7) we may assume that the maximum is attained at an extreme point of $\mathcal{K}_1$ which, by Lemma 2.2.3, belongs to $\mathcal{K}_2$. Therefore, both programs have the same optimum value, denoted as p^*. We now show that $p^* = \lambda_1 + \ldots + \lambda_k$.

Let $u_1, \ldots, u_n$ be an orthonormal set of eigenvectors corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_n$ of C and let Y be the $n \times k$ matrix with columns $u_1, \cdots, u_k$. Then $YY^\mathsf{T} \in \mathcal{K}_2$ and $\langle C, YY^\mathsf{T}\rangle = \sum_{i=1}^k \lambda_i$, thus showing $\lambda_1 + \ldots + \lambda_k \leqslant p^*$.

Denote by Q the orthogonal matrix with columns $u_1, \ldots, u_n$ and by D the diagonal matrix with the λ_i's on the diagonal, so that $C = QDQ^\mathsf{T}$. Then $\langle Q, YY^\mathsf{T}\rangle = \langle D, ZZ^\mathsf{T}\rangle$ after setting $Z = Y^\mathsf{T} Q$; note that $Z^\mathsf{T} Z = I_k$ and thus $ZZ^\mathsf{T} \in \mathcal{K}_2$. Hence p^* is the maximum value of $\langle D, M\rangle = \sum_{i=1}^n \lambda_i M_{ii}$ taken over $M \in \mathcal{K}_2$. The constraints for $M \in \mathcal{K}_2$ imply that the vector $\text{diag}(M)$ belongs to the polytope $\mathcal{P}$ from (2.11). Therefore the maximum of $\sum_i \lambda_i M_{ii}$ is at most the maximum of $\sum_i \lambda_i x_i$ taken over $x \in \mathcal{P}$. Now the latter maximum is attained at an extreme point of $\mathcal{P}$, from which one derives that it is equal to $\lambda_1 + \ldots + \lambda_k$. This gives the reverse inequality: $p^* \leqslant \lambda_1 + \ldots + \lambda_k$. □

We mention another result of the same flavor: Given two symmetric matrices A, B, one can reformulate as a semidefinite program the following optimization problem over orthogonal matrices:

$$\min \left\{ \text{Tr}(AXBX^\mathsf{T}) : X \in \mathcal{O}(n) \right\}. \tag{2.12}$$

We already considered this problem in Section 1.7.2 in relation with the Hoffman-Wielandt inequality for eigenvalues (recall Theorem 1.7.4). The semidefinite reformulation uses Kronecker products of matrices (introduced in Section 1.7.6).

Theorem 2.2.4. *Let $A, B \in \mathcal{S}^n$. Then the program (2.12) is equivalent to the semidefinite program*

$$\max \left\{ \text{Tr}(S) + \text{Tr}(T) : A \otimes B - I_n \otimes T - S \otimes I_n \succeq 0 \right\}. \tag{2.13}$$

Moreover its optimum value is

$$\sum_{i=1}^n \alpha_i \beta_i,$$

where the α_i's are the eigenvalues of A ordered in ascending order: $\alpha_1 \leqslant \ldots \leqslant \alpha_n$, and the β_i's are the eigenvalues of B ordered in descending order: $\beta_1 \geqslant \ldots \geqslant \beta_n$.

Proof. Let D be the diagonal matrix whose diagonal entries are the α_i's and let E be the diagonal matrix whose diagonal entries are the β_i's. As in the proof of Theorem 1.7.4, the program (2.12) is equivalent to

$$\min \left\{ \text{Tr}(DXEX^{\mathsf{T}} : X \in \mathcal{O}(n) \right\} \tag{2.14}$$

which in turn is equivalent to the linear program (1.2), repeated here for convenience:

$$\max_{s,t \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} s_i + \sum_{j=1}^{n} t_j : \alpha_i \beta_j - s_i - t_j \geqslant 0 \ \forall i, j \in [n] \right\}. \tag{2.15}$$

We now show that the linear program (2.15) is equivalent to the following semidefinite program

$$\max_{S,T \in \mathcal{S}^n} \left\{ \text{Tr}(S) + \text{Tr}(T) : E \otimes F - I_n \otimes T - S \otimes I_n \succeq 0 \right\}. \tag{2.16}$$

To see it, let S, T be feasible for (2.16) and define the vectors $s = \text{diag}(S)$, $t = \text{diag}(T)$. Then, as $E \otimes F$ is a diagonal matrix, the diagonal matrices $\text{Diag}(s)$ and $\text{Diag}(t)$ are feasible for (2.16) with the same objective value: $\text{Tr}(S) + \text{Tr}(T) = \text{Tr}(\text{Diag}(s)) + \text{Tr}(\text{Diag}(t))$. Now, program (2.16) with the additional condition that S, T are diagonal matrices can be reformulated as (2.15). Finally, write $A = PDP^{\mathsf{T}}$ and $B = QEQ^{\mathsf{T}}$ where $P, Q \in \mathcal{O}(n)$ and observe that

$$(P \otimes Q)(E \otimes F - I_n \otimes T - S \otimes I_n)(P \otimes Q)^{\mathsf{T}} = A \otimes B - I_n \otimes (QTQ^{\mathsf{T}}) - (PSP^{\mathsf{T}}) \otimes I_n.$$

Hence S, T is feasible for (2.16) if and only if $S' = PSP^{\mathsf{T}}$, $T' = QTQ^{\mathsf{T}}$ is feasible for (2.13), and $\text{Tr}(S) + \text{Tr}(T) = \text{Tr}(S') + \text{Tr}(T')$. From this follows the desired equivalence of (2.12) and (2.13). The fact that the optimum value is $\sum_i \alpha_i \beta_i$ was computed in Theorem 1.7.4. □

2.3 Convex quadratic constraints

Consider a quadratic constraint for a vector $x \in \mathbb{R}^n$ of the form

$$x^{\mathsf{T}} A x \leqslant b^{\mathsf{T}} x + c, \tag{2.17}$$

where $A \in \mathcal{S}^n$, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. In the special case when $A \succeq 0$, then the feasible region defined by this constraint is convex and it turns out that it can be equivalently defined by a semidefinite constraint.

Lemma 2.3.1. *Assume $A \succeq 0$. Say, $A = LL^{\mathsf{T}}$, where $L \in \mathbb{R}^{n \times k}$. Then, for any $x \in \mathbb{R}^n$,*

$$x^{\mathsf{T}} A x \leqslant b^{\mathsf{T}} x + c \iff \begin{pmatrix} I_k & L^{\mathsf{T}} x \\ x^{\mathsf{T}} L & b^{\mathsf{T}} x + c \end{pmatrix} \succeq 0.$$

Proof. The equivalence follows as a direct application of Lemma 1.7.6: Choose here $A = I_k$, $B = L^\mathsf{T} x \in \mathbb{R}^{k\times 1}$ and $C = b^\mathsf{T} x + c \in \mathbb{R}^{1\times 1}$. And take the Schur complement of the submatrix I_k in the block-matrix on the right hand side. □

As a direct application, the Euclidean unit ball can be represented by an LMI:

$$\{x \in \mathbb{R}^n : \|x\| \leqslant 1\} = \left\{x \in \mathbb{R}^n : \begin{pmatrix} 1 & x^\mathsf{T} \\ x & I_n \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & I_n \end{pmatrix} + \sum_{i=1}^n x_i \begin{pmatrix} 0 & e_i^\mathsf{T} \\ e_i & 0 \end{pmatrix} \succeq 0\right\}$$

as well as its homogenization:

$$\mathcal{L}^{n+1} = \{(x,t) \in \mathbb{R}^{n+1} : \|x\| \leqslant t\} = \left\{x \in \mathbb{R}^n : \begin{pmatrix} t & x^\mathsf{T} \\ x & tI_n \end{pmatrix} \succeq 0\right\}.$$

So at $t = t_0$, we have in the x-space the ball of radius t_0. The set $\mathcal{L}^{n+1}$ is a cone, known as the *second-order cone* (or *Lorentz cone*), to which we will come back in the next chapter.

The fact that one can reformulate linear optimization over the Euclidean ball as a maximization or minimization semidefinite program can be very useful as we will see in the next section.

Corollary 2.3.2. *Given $c \in \mathbb{R}^n$, the following holds:*

$$\begin{aligned} \min_{\|x\|\leqslant 1} c^\mathsf{T} x &= \min_{x\in\mathbb{R}^n} c^\mathsf{T} x \ \textit{ s.t. } \begin{pmatrix} 1 & x^\mathsf{T} \\ x & I_n \end{pmatrix} \succeq 0 \\ &= \max_{X\in\mathcal{S}^{n+1}} -\mathrm{Tr}(X) \ \textit{ s.t. } X_{0i} = c_i \ (i \in [n]),\ X \succeq 0. \end{aligned} \tag{2.18}$$

Proof. Apply Lemma 2.3.1 combined with the duality theorem (Theorem 2.1.2). □

2.4 Robust optimization

We indicate here how semidefinite programming comes up when dealing with some robust optimization problems.

Consider the following linear programming problem:

$$\max\{c^\mathsf{T} x : a^\mathsf{T} x \geqslant b\},$$

where $c, a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ are given data, with just one constraint for simplicity of exposition. In practical applications the data a, b might be given through experimental results and might not be known exactly with 100% certainty, which is in fact the case in most of the real world applications of linear programming. One may write $a = a(z)$ and $b = b(z)$ as functions of an uncertainty parameter z assumed to lie in a given uncertainty region $\mathcal{Z} \subseteq \mathbb{R}^k$. Then one wants to find an optimum solution x that is *robust* against this uncertainty, i.e., that satisfies

the constraints $a(z)^\mathsf{T} x \geqslant b(z)$ for *all* values of the uncertainty parameter $z \in \mathcal{Z}$. That is, solve

$$\max\{c^\mathsf{T} x : a(z)^\mathsf{T} x \geqslant b(z)\ \forall z \in \mathcal{Z}\}. \tag{2.19}$$

Depending on the set $\mathcal{Z}$ this problem might have infinitely many constraints. However, for certain choices of the functions $a(z), b(z)$ and of the uncertainty region $\mathcal{Z}$, one can reformulate the problem as a semidefinite programming problem, thus tractable.

Suppose that the uncertainty region $\mathcal{Z}$ is the unit ball and that $a(z), b(z)$ are linear functions in the uncertainty parameter $z = (\zeta_1, \cdots, \zeta_k) \in \mathbb{R}^k$, of the form

$$a(z) = a_0 + \sum_{j=1}^{k} \zeta_j a_j,\ b(z) = b_0 + \sum_{j=1}^{k} \zeta_j b_j \tag{2.20}$$

where $a_j, b_j \in \mathbb{R}^n$ are known. Then the robust optimization problem (2.19) can be reformulated as a semidefinite programming problem involving the variable $x \in \mathbb{R}^n$ and a new matrix variable $Z \in \mathcal{S}^k$. The proof relies on the result from Corollary 2.3.2, where we made use in a crucial manner of the duality theory for semidefinite programming, for showing the equivalence of both problems in (2.18).

Theorem 2.4.1. *Suppose that the functions $a(z)$ and $b(z)$ are given by (2.20) and that $\mathcal{Z} = \{z \in \mathbb{R}^m : \|z\| \leqslant 1\}$. Then problem (2.19) is equivalent to the problem:*

$$\max_{x \in \mathbb{R}^n, Z \in \mathcal{S}^{k+1}} c^\mathsf{T} x \ \text{ such that } \quad \begin{array}{l} a_j^\mathsf{T} x - Z_{0j} = b_j\ (j \in [k]) \\ a_0^\mathsf{T} x - \mathrm{Tr}(Z) \geqslant b_0,\ Z \succeq 0. \end{array} \tag{2.21}$$

Proof. Fix $x \in \mathbb{R}^n$, set $\alpha_j = a_j^\mathsf{T} x - b_j$ for $j = 0, 1, \ldots, k$, and define the vector $\alpha = (\alpha_j)_{j=1}^k \in \mathbb{R}^k$ (which depends on x). Then the constraints: $a(z)^\mathsf{T} x \geqslant b(z)$ $\forall z \in Z$ can be rewritten as

$$\alpha^\mathsf{T} z \geqslant -\alpha_0\ \forall z \in \mathcal{Z}.$$

Therefore, we find the problem of deciding whether $p^* \geqslant -\alpha_0$, where

$$p^* = \min_{\|z\| \leqslant 1} \alpha^\mathsf{T} z.$$

Now the above problem fits precisely within the setting considered in Corollary 2.3.2. Hence, we can rewrite it using the second formulation in (2.18) – the one in maximization form – as

$$p^* = \max_{Z \in \mathcal{S}^{m+1}} \{-\mathrm{Tr}(Z) : Z_{0j} = \alpha_j\ (j \in [k]), Z \succeq 0\}.$$

So, in problem (2.19), we can substitute the condition: $a(z)^\mathsf{T} x \geqslant b(z)\ \forall z \in \mathcal{Z}$ by the condition:

$$\exists Z \in \mathcal{S}^{m+1}_{\succeq 0} \ \text{ s.t. } \ -\mathrm{Tr}(Z) \geqslant -\alpha_0,\ Z_{0j} = \alpha_j\ (j \in [k]).$$

The crucial fact here is that the quantifier "$\forall z$" has been replaced by the existential quantifier "$\exists Z$". As problem (2.19) is a maximization problem in x, it is equivalent to the following maximization problem in the variables x and Z:

$$\max_{x\in\mathbb{R}^n, Z\in\mathcal{S}^{m+1}} \left\{c^{\mathsf{T}}x : a_0^{\mathsf{T}}x - \mathrm{Tr}(Z) \geqslant b_0,\ a_j^{\mathsf{T}}x - Z_{0j} = b_j\ (j\in[k])\right\}$$

(after substituting back in α_j their expression in terms of x). □

2.5 Examples in combinatorial optimization

Semidefinite programs provide a powerful tool for constructing useful convex relaxations for combinatorial optimization problems. We will treat this in detail in a later chapter. For now we illustrate the main idea on the following two examples: finding a maximum independent set and a maximum cut in a graph.

2.5.1 The maximum independent set problem

Consider a graph $G = (V, E)$ with vertex set $V = [n]$, the edges are unordered pairs of distinct vertices. A set of nodes (or vertices) $S \subseteq V$ is said to be *independent* (or *stable*) if it does not contain an edge and the maximum cardinality of an independent set is denoted as $\alpha(G)$, known as the *stability number* of G. The *maximum independent set problem* asks to compute $\alpha(G)$. This problem is $\mathcal{N}P$-hard.

Here is a simple recipe for constructing a semidefinite programming upper bound for $\alpha(G)$. It is based on the following observation: Let S be an independent set in G and let $x \in \{0,1\}^n$ be its incidence vector, with $x_i = 1$ if $i \in S$ and $x_i = 0$ otherwise. Define the matrix $X = xx^{\mathsf{T}}/|S|$. Then the matrix X satisfies the following conditions: $X \succeq 0$, $X_{ij} = 0$ for all edges $\{i,j\} \in E$, $\mathrm{Tr}(X) = 1$, and $\langle J, X\rangle = |S|$. It is therefore natural to consider the following semidefinite program

$$\vartheta(G) = \max_{X\in\mathcal{S}^n}\{\langle J, X\rangle : \mathrm{Tr}(X) = 1,\ X_{ij} = 0\ (\{i,j\}\in E),\ X \succeq 0\}, \tag{2.22}$$

whose optimum value $\vartheta(G)$ is known as the *theta number* of G. It follows from the above discussion that $\vartheta(G)$ is an upper bound for the stability number. That is,

$$\alpha(G) \leqslant \vartheta(G).$$

The dual semidefinite program reads

$$\min_{y\in\mathbb{R}^E, t\in\mathbb{R}} \left\{ t : tI + \sum_{\{i,j\}\in E} y_{ij}E_{ij} - J \succeq 0 \right\}, \tag{2.23}$$

and its optimum value is equal to $\vartheta(G)$ (because (6.13) is strictly feasible and bounded – check it). Here we have used the elementary matrices E_{ij} introduced in the abstract of the chapter.

We will come back to the theta number in a later chapter. As we will see there, there is an interesting class of graphs for which $\alpha(G) = \vartheta(G)$, the so-called *perfect graphs*. For these graphs, the maximum independent set problem can be solved in polynomial time. This result is one of the first breakthrough applications of semidefinite programming obtained in the early eighties.

2.5.2 The maximum cut problem

Consider again a graph $G = (V, E)$ where $V = [n]$. Given a subset $S \subseteq V$, the *cut* $\delta_G(S)$ consists of all the edges $\{i, j\}$ of G that are cut by the partition $(S, V\backslash S)$, i.e., exactly one of the two nodes i, j belongs to S. The *maximum cut problem* (or *max-cut*) asks to find a cut of maximum cardinality. This is an $\mathcal{NP}$-hard problem.

One can encode the max-cut problem using variables $x \in \{\pm 1\}^n$. Assign $x_i = 1$ to the nodes $i \in S$ and -1 to the nodes $i \in V\backslash S$. Then the cardinality of the cut $\delta_G(S)$ is equal to $\sum_{\{i,j\}\in E}(1 - x_i x_j)/2$. Therefore max-cut can be formulated as

$$\text{max-cut} = \max_{x\in\mathbb{R}^n} \left\{ \sum_{\{i,j\}\in E} (1 - x_i x_j)/2 : x \in \{\pm 1\}^n \right\}. \tag{2.24}$$

Again there is a simple recipe for constructing a semidefinite relaxation for max-cut: Pick a vector $x \in \{\pm 1\}^n$ (arising in the above formulation of max-cut) and consider the matrix $X = xx^{\mathsf{T}}$. This matrix X satisfies the following conditions: $X \succeq 0$ and $X_{ii} = 1$ for all $i \in [n]$. Therefore, it is natural to consider the following semidefinite relaxation for max-cut:

$$\text{sdp} = \max_{X\in\mathcal{S}^n} \left\{ \sum_{\{i,j\}\in E} (1 - X_{ij})/2 : X \succeq 0,\ X_{ii} = 1\ (i \in [n]) \right\}. \tag{2.25}$$

As we will see later this semidefinite program provides a very good approximation for the max-cut problem: $\text{sdp} \leqslant 1.13 \cdot \text{max-cut}$. This is a second breakthrough application of semidefinite programming, obtained in the early nineties.

Let $L_G \in \mathcal{S}^n$ denote the Laplacian matrix of G: its (i, i)th diagonal entry is the degree of node i in G, and the (i, j)th off-diagonal entry is -1 if $\{i, j\}$ is an edge and 0 otherwise. Note that

$$x^{\mathsf{T}} L_G x = \sum_{\{i,j\}\in E} (x_i - x_j)^2\ \forall x \in \mathbb{R}^n, \quad \frac{1}{4} x^{\mathsf{T}} L_G x = \frac{1}{2} \sum_{\{i,j\}\in E} (1 - x_i x_j)\ \forall x \in \{\pm 1\}^n.$$

The first item shows that $L_G \succeq 0$, and the second item shows that one can reformulate max-cut using the Laplacian matrix. Analogously one can reformulate

the semidefinite program (2.25) as

$$\text{sdp } = \max\left\{\frac{1}{4}\langle L_G, X\rangle : X \succeq 0,\ X_{ii} = 1\ (i \in [n])\right\}. \tag{2.26}$$

Given a positive semidefinite matrix A, consider the following quadratic problem

$$\text{opt } = \max\{x^T Ax : \|x\|_\infty \leqslant 1\}. \tag{2.27}$$

where $\|x\|_\infty = \max_i |x_i|$ is the ℓ_∞-norm. As we maximize a convex function over the convex set $[-1,1]^n$, the maximum is attained at a vertex, i.e., at a point of $\{\pm 1\}^n$. This shows that (2.27) is equivalent to

$$\text{opt} = \max\{x^T Ax : x \in \{\pm 1\}^n\}. \tag{2.28}$$

This problem is $\mathcal{NP}$-hard – indeed it contains the max-cut problem, obtained when choosing $A = L_G/4$.

Note that if we would replace in (2.27) the cube $[-1,1]^n$ by the Euclidean unit ball, then we find the problem of computing the largest eigenvalue of A which, as we saw earlier, can be modeled as a semidefinite program.

Just as for max-cut one can formulate the following semidefinite relaxation of (2.28) (and thus of (2.27)):

$$\text{sdp } = \max\{\langle A, X\rangle : X \succeq 0,\ X_{ii} = 1\ \forall i \in [n]\}. \tag{2.29}$$

We will see later that this semidefinite program too gives a good approximation of the quadratic problem (2.27): sdp $\leqslant \frac{\pi}{2}$ opt .

2.6 Examples in geometry

Given vectors $u_1, \ldots, u_n \in \mathbb{R}^k$, let $d = (d_{ij})$ denote the vector consisting of their pairwise squared Euclidean distances, i.e., $d_{ij} = \|u_i - u_j\|^2$ for all $i, j \in [n]$. Thus $d_{ii} = 0$ for all i. Now, think of the vectors u_i as representing the locations of some objects (atoms of a molecule, or sensors in a sensor network). One might be able to determine the pairwise distances d_{ij} by making some measurements. However, in general, one can determine these distances d_{ij} only for a subset of pairs, corresponding to the edges of a graph G. Then the problem arises whether one can reconstruct the locations of the objects (the vectors u_i) from these partial measurements (the distances d_{ij} for the edges $\{i,j\}$ of G).

In mathematical terms, given a graph $G = (V = [n], E)$ and $d \in \mathbb{R}^E_{\geqslant 0}$, decide whether there exist vectors $u_1, \ldots, u_n \in \mathbb{R}^k$ such that

$$\|u_i - u_j\|^2 = d_{ij} \text{ for all } \{i,j\} \in E.$$

Of course, this problem comes in several flavors. One may search for such vectors u_i lying in a space of prescribed dimension k; then typically $k = 2$, 3 or

4 would be of interest. This is in fact a hard problem. However, if we relax the bound on the dimension and simply ask for the existence of the u_i's in $\mathbb{R}^k$ for *some* $k \geqslant 1$, then the problem can be cast as the problem of deciding feasibility of a semidefinite program.

Lemma 2.6.1. *Given $d \in \mathbb{R}^E_{\geqslant 0}$, there exist vectors $u_1, \ldots, u_n \in \mathbb{R}^k$ (for some $k \geqslant 1$) if and only if the following semidefinite program is feasible:*

$$X \succeq 0,\ X_{ii} + X_{jj} - 2X_{ij} = d_{ij} \text{ for } \{i,j\} \in E.$$

Moreover, such vectors exist in the space $\mathbb{R}^k$ if and only if the above semidefinite program has a feasible solution of rank at most k.

Proof. Directly, using the fact that $X \succeq 0$ if and only if X admits a Gram representation $u_1, \ldots, u_n \in \mathbb{R}^k$ (for some $k \geqslant 1$), i.e., $X_{ij} = u_i^{\mathsf{T}} u_j$ for all $i, j \in [n]$. Moreover, the rank of X is equal to the rank of the system $\{u_1, \ldots, u_n\}$. □

Thus arises naturally the problem of finding *low rank* solutions to a semidefinite program. We will come back to this topic in a later chapter.

2.7 Examples in algebra

Another, maybe a bit unexpected at first sight, application of semidefinite programming is for testing whether a multivariate polynomial can be written as a sum of squares of polynomials.

First recall a bit of notation. $\mathbb{R}[x_1, \ldots, x_n]$ (or simply $\mathbb{R}[x]$ for simplicity) denotes the ring of polynomials in n variables. A polynomial $p \in \mathbb{R}[x]$ can be written as $p = \sum_\alpha p_\alpha x^\alpha$, where $p_\alpha \in \mathbb{R}$ and x^α stands for the monomial $x_1^{\alpha_1} \cdots x_n^{\alpha_n}$. The sum is finite and the maximum value of $|\alpha| = \sum_{i=1}^n \alpha_i$ for which $p_\alpha \neq 0$ is the degree of p. For an integer d, $[x]_d$ denotes the vector consisting of all monomials of degree at most d, which has $\binom{n+d}{d}$ entries. Denoting by $\boldsymbol{p} = (p_\alpha)$ the vector of coefficients of p, we can write

$$p = \sum_\alpha p_\alpha x^\alpha = \boldsymbol{p}^{\mathsf{T}} [x]_d. \tag{2.30}$$

Definition 2.7.1. *A polynomial p is said to be a sum of squares (SOS) if p can be written as a sum of squares of polynomials, i.e., $p = \sum_{j=1}^m (q_j)^2$ for some polynomials q_j.*

It turns out that checking whether p is SOS can be reformulated via a semidefinite program. Clearly, we may assume that p has even degree $2d$ (else p is not SOS) and the polynomials q_j arising in a SOS decomposition will have degree at most d.

Let us now make the following simple manipulation, based on (2.30):

$$\sum_j q_j^2 = \sum_j [x]_d^{\mathsf{T}} \boldsymbol{q_j} \boldsymbol{q_j}^{\mathsf{T}} [x]_d = [x]_d^{\mathsf{T}} \Big(\sum_j \boldsymbol{q_j} \boldsymbol{q_j}^{\mathsf{T}} \Big) [x]_d = [x]_d^{\mathsf{T}} Q [x]_d,$$

after setting $Q = \sum_j \boldsymbol{q_j}\boldsymbol{q_j}^\mathsf{T}$. Having such a decomposition for the matrix Q amounts to requiring that Q is positive semidefinite. Therefore, we have just shown that the polynomial p is SOS if and only if

$$p = [x]_d^\mathsf{T} Q [x]_d \text{ for some matrix } Q \succeq 0.$$

Linear conditions on Q arise by equating the coefficients of the polynomials on both sides in the above identity.

Summarizing, one can test whether p can be written as a sum of squares by checking the feasibility of a semidefinite program. If p has degree $2d$, this SDP involves a variable matrix Q of size $\binom{n+d}{d}$ (the number of monomials of degree at most d) and $\binom{n+2d}{2d}$ (the number of monomials of degree at most $2d$) linear constraints.

One can sometimes restrict to smaller matrices Q. For instance, if the polynomial p is homeogeneous (i.e, all its terms have degree $2d$), then we may assume without loss of generality that the polynomials q_j appearing in a SOS decomposition are homogeneous of degree d. Hence Q will be indexed by the $\binom{n+d-1}{d}$ monomials of degree *equal to* d.

Why bother about sums of squares of polynomials? A good reason is that they can be useful to recognize and certify positive polynomials and to approximate optimization problems dealing with polynomials. Let us just give a glimpse on this.

Suppose that one wants to compute the infimum $p^{\min}$ of a polynomial p over the full space $\mathbb{R}^n$. In other words, one wants to find the largest scalar λ for which $p(x) - \lambda \geqslant 0$ for all $x \in \mathbb{R}^n$. This is in general a hard problem. However, if we relax the positivity condition on $p - \lambda$ and instead require that $p - \lambda$ is a sum of squares, then it follows from the above considerations that we can compute the maximum λ for which $p - \lambda$ is SOS using semidefinite programming. This gives a tractable bound p^* satisfying: $p^* \leqslant p^{\min}$.

In general p^* might be distinct from $p^{\min}$. However in the univariate case ($n = 1$), equality holds: $p^{\min} = p^*$. (This will follow from the result in Problem 2.2.) Equality holds also in the quadratic case: $d = 2$, and in one exceptional case: $n = 2$ and $d = 4$. This was shown by Hilbert in 1888.

We will return to this topic in a later chapter.

2.8 Further reading

A detailed treatment about Fan's theorem (Theorem 2.2.2) can be found in Overton and Womersley [8] and a detailed discussion about Hoffman-Wielandt inequality, Theorem 2.2.4 and applications (e.g. to quadratic assignment) can be found in Anstreicher and Wolkowicz [2].

The recent monograph of Ben-Tal, El Ghaoui and Nemirovski [3] offers a detailed treatment of robust optimization. The result presented in Theorem 2.4.1 is just one of the many instances of problems which admit a robust counterpart which is a tractable optimization problem. Although we formulated it in terms

of semidefinite programming (to fit our discussion), it can in fact be formulated in terms of second-order conic optimization, which admits faster algorithms.

The theta number $\vartheta(G)$ was introduced in the seminal work of Lovász [3]. A main motivation of Lovász was to give good bounds for the Shannon capacity of a graph, an information theoretic measure of the graph. Lovász succeeded to determine the exact value of the Shannon capacity of C_5, the circuit on five nodes, by computing $\vartheta(C_5) = \sqrt{5}$. This work of Lovász can be considered as the first breakthrough application of semidefinite programming, although the term *semidefinite programming* was coined only later. Chapter 33 of [1] gives a beautiful treatment of this result. The monograph by Grötschel, Lovász and Schrijver [1] treats in detail algorithmic questions related to semidefinite programming and, in particular, to the theta number. Polynomial time solvability based on the ellipsoid method is treated in detail there.

Using semidefinite programming to approximate max-cut was pioneered by the work of Goemans and Williamson [5]. This novel approach and their result had a great impact on the area of combinatorial optimization. It indeed spurred a lot of research activity for getting tight approximations for various problems. This line of research is now also very active in theoretical computer science, where the *unique games conjecture* has been formulated that is directly relevant to the basic semidefinite relaxation (2.25) for max-cut – cf. e.g. the survey by Trevisan [10].

Sums of squares of polynomials are a classical topic in mathematics and they have many applications e.g. to control theory and engineering. In the late 1800s David Hilbert classified the parameters degree/number of variables for which any positive polynomial can be written as a sum of squares of polynomials. He posed the question whether any positive polynomial can be written as a sum of squares of rational functions, known as Hilbert's 17th problem. This was solved by Artin in 1927, a result which started the field of real algebraic geometry. The survey by Reznick [6] gives a nice overview and historical perspective and the monograph by Delzell and Prestell [4] gives an in-depth treatment of positivity.

2.9 Exercises

2.1. (a) Formulate the dual SDP of the program (2.7).

(b) Give a semidefinite programming formulation for the following problem:

$$\min\{\lambda_1(X) + \ldots + \lambda_k(X) : \langle A_j, X\rangle = b_j \ (j \in [m])\},$$

which asks for a matrix $X \in \mathcal{S}^n$ satisfying a system of linear constraints and for which the sum of the k largest eigenvalues of X is minimum.

2.2. Let p be a univariate polynomial.

(a) Show that p can be written as a sum of squares if and only if p is non-negative over $\mathbb{R}$, i.e., $p(x) \geqslant 0 \ \forall x \in \mathbb{R}$.

(b) Show that if p is non-negative over $\mathbb{R}$ then it can be written as sum of two squares.

2.3**. (a) Build the dual of the semidefinite programming (2.26) and show that it is equivalent to

$$\frac{n}{4} \min_{u \in \mathbb{R}^n} \{\lambda_{\max}(\text{Diag}(u) + L_G) : e^{\mathsf{T}} u = 0\},$$

where Diag(u) is the diagonal matrix with diagonal entries $u_1, \ldots, u_n$.

(b) Show that the maximum cardinality of a cut is at most

$$\frac{n}{4} \lambda_{\max}(L_G),$$

where $\lambda_{\max}(L_G)$ is the maximum eigenvalue of the Laplacian matrix of G.

(c) Show that the maximum cardinality of a cut in G is at most

$$\frac{1}{2}|E| - \frac{n}{4} \lambda_{\min}(A_G)$$

where A_G is the adjacency matrix of G.

(d) Show that both bounds in (b) and (c) coincide when G is a regular graph (i.e., all nodes have the same degree).

2.4. Consider the polynomial in two variables x and y

$$p = x^4 + 2x^3y + 3x^2y^2 + 2xy^3 + 2y^4.$$

(a) Build a semidefinite program permitting to recognize whether p can be written as sum of squares.

(b) Describe all possible sums of squares decompositions for p.

(c) What can you say about the number of squares needed?

BIBLIOGRAPHY

[1] M. Aigner and G. Ziegler. *Proofs from THE BOOK*. Springer, 2003.

[2] K. Anstreicher and H. Wolkowicz. On Lagrangian relaxation of quadratic matrix constraints. *SIAM Journal on Matrix Analysis and its Applications* **22(1)**:41–55, 2000.

[3] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*, Princeton University Press, 2009.

[4] A. Prestel and C.N. Delzell. *Positive Polynomials - From Hilberts 17th Problem to Real Algebra*. Springer, 2001.

[5] M.X. Goemans and D. Williamson. Improved approximation algorithms for maximum cuts and satisfiability problems using semidefinite programming. *Journal of he ACM* **42**:1115–1145, 1995.

[6] M. Grötschel, L. Lovász and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer, 1988.

[7] L. Lovász. On the Shannon capacity of a graph. *IEEE Transactions on Information Theory* **IT-25**:1–7, 1979.

[8] M. Overton and R.S. Womersley. On the sum of the k largest eigenvalues of a symmetric matrix. *SIAM Journal on Matrix Analysis and its Applications* **13(1)**:41–45, 1992.

[9] B. Reznick. Some concrete aspects of Hilbert's 17th problem. In *Real Algebraic Geometry and Ordered Structures*. C.N. Delzell and J.J. Madden (eds.), *Contemporary Mathematics* **253**:251–272, 2000.

[10] L. Trevisan. On Khot's unique games conjecture. *Bull. Amer. Math. Soc.* **49**:91-111, 2012.

CHAPTER 3

DUALITY IN CONIC PROGRAMMING

Traditionally, convex optimization problems are of the form

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_1(x) \leqslant 0, \ldots, f_N(x) \leqslant 0, \\ & a_1^{\mathsf{T}} x = b_1, \ldots, a_M^{\mathsf{T}} x = b_M, \end{aligned}$$

where the *objective function* $f_0 : D \to \mathbb{R}$ and the *inequality constraint functions* $f_i : D \to \mathbb{R}$ which are defined on a convex domain $D \subseteq \mathbb{R}^n$ are *convex*, i.e. their *epigraphs*

$$\operatorname{epi} f_i = \{(x, \alpha) : D \times \mathbb{R} : f_i(x) \leqslant \alpha\}, \quad i = 0, \ldots, N,$$

are convex sets in $D \times \mathbb{R} \subseteq \mathbb{R}^{n+1}$. Equivalently, the function f_i is convex if and only if

$$\forall x, y \in D \ \forall \alpha \in [0,1] : f_i((1-\alpha)x + \alpha x) \leqslant (1-\alpha) f_i(x) + \alpha f_i(y).$$

The *equality constraints* are given by vectors $a_j \in \mathbb{R}^n \backslash \{0\}$ and right hand sides $b_j \in \mathbb{R}$. The convex set of *feasible solutions* is the intersection of N convex sets with M hyperplanes

$$\bigcap_{i=1}^{N} \{x \in D : f_i(x) \leqslant 0\} \cap \bigcap_{j=1}^{M} \{x \in \mathbb{R}^n : a_j^{\mathsf{T}} x = b_j\}.$$

The set-up for conic programming is slightly different. We start by considering a fixed convex cone K lying in the n-dimensional Euclidean space $\mathbb{R}^n$. The

task of conic programming is the following: One wants to maximize (or minimize) a linear function over the feasible region which is given as the intersection of the convex cone K with an affine subspace:

$$\begin{aligned} &\text{maximize} && c^\mathsf{T} x \\ &\text{subject to} && x \in K, \\ & && a_1^\mathsf{T} x = b_1, \ldots, a_m^\mathsf{T} x = b_m. \end{aligned}$$

This differs only slightly from a traditional convex optimization problem: The objective function is linear and feasibility with respect to the inequality constraint functions is replaced by membership in the fixed convex cone K. In principle, one can transform every convex optimization problem into a conic program. However, the important point in conic programming is that it seems that a vast majority of convex optimization problems which come up in practice can be formulated as conic programs using the three standard cones:

1. the non-negative orthant $\mathbb{R}^n_{\geqslant 0}$ – giving linear programming (LP),
2. the second-order cone $\mathcal{L}^{n+1}$ – giving second-order cone programming (CQP),
3. or the cone of positive semidefinite matrices $\mathcal{S}^n_{\succeq 0}$ – giving semidefinite programming (SDP).

As we will see in the next lecture, these three cones have particular nice analytic properties: They have a self-concordant barrier function which is easy to evaluate. This implies that there are theoretically (polynomial-time) and practically efficient algorithms to solve these standard problems.

In addition to this, the three examples are ordered by their "difficulty", which can be pictured as

$$\text{LP} \subseteq \text{CQP} \subseteq \text{SDP}.$$

This means that one can formulate every linear program as a conic quadratic program and one can formulate every conic quadratic program as a semidefinite program.

Why do we care about conic programming in general and do not focus on these three most important special cases?

The answer is that conic programming gives a unifying framework to design algorithms, to understand the basic principles of its geometry and duality, and to model optimization problems. Moreover this offers the flexibility of dealing with new cones obtained e.g. by taking direct products of the three standard types of cones.

3.1 Fundamental properties

3.1.1 Local minimizers are global minimizers

A first fundamental property of convex optimization problems is that every local minimizer is at the same time a global minimizer. A *local minimizer* of the convex optimization problem is a feasible solution $x \in D$ having the property that there is a positive ϵ so that

$$f_0(x) = \inf\{f_0(y) : y \text{ is feasible and } d(x,y) \leqslant \epsilon\}.$$

Here and throughout we use the notation $d(x,y)$ to denote the Euclidean distance $\|x-y\|_2$ between $x, y \in \mathbb{R}^n$. To see that local optimality implies global optimality assume that x is a local but *not* a global minimizer, then there is a feasible solution y so that $f_0(y) < f_0(x)$. Clearly, $d(x,y) > \epsilon$. Define $z \in [x,y]$ by setting

$$z = (1-\alpha)x + \alpha y, \quad \alpha = \frac{\epsilon}{2d(x,y)},$$

which is a feasible solution because of convexity. Then, $d(x,z) = \epsilon/2$ and again by convexity

$$f_0(z) \leqslant (1-\alpha)f_0(x) + \alpha f_0(y) < f_0(x),$$

which contradicts the fact that x is a local minimizer.

3.1.2 Karush-Kuhn-Tucker condition

A second fundamental property of convex optimization problems is that one has necessary and sufficient conditions for x being a local (and hence a global) minimizer. Stating and analyzing these kind of conditions is central to the theory of non-linear programming and convex analysis. We just state one fundamental result here without proving it. A proof can be found for instance in the book [2, Chapter 5] by Boyd and Vandenberghe.

We assume that the convex optimization problem satisfies the following condition, known as *Slater's condition*:

There exists a point $x \in \operatorname{relint} D$ such that $f_i(x) < 0$ for all $i = 1, \ldots, N$ and such that $a_j^{\mathsf{T}} x = b_j$ for all $j = 1, \ldots, M$.

This point is called a *strictly feasible solution* since the inequality constraints hold with strict inequality. Furthermore, we assume that the objective function and that the inequality constraint functions are differentiable. Under these conditions a feasible solution is a global minimizer if and only if the Karush-Kuhn-Tucker (KKT) condition holds: There are $\lambda_1, \ldots, \lambda_N \in \mathbb{R}_{\geqslant 0}$ and $\mu_1, \ldots, \mu_M \in \mathbb{R}$ so that the following equations are satisfied:

$$\lambda_1 f_1(x) = 0, \ldots, \lambda_N f_N(x) = 0,$$
$$\nabla f_0(x) + \sum_{i=1}^{N} \lambda_i \nabla f_i(x) + \sum_{j=1}^{M} \mu_j a_j = 0.$$

The KKT-condition is an extension of the method of Lagrange multipliers where one also can consider inequalities instead of only equalities.

3.2 Primal and dual conic programs

When defining conic programming we need a "nice" cone K, satisfying the following properties: K is closed, convex, pointed, and has a non-empty interior or, equivalently, it is full-dimensional.

3.2.1 Primal conic programs

Let $K \subseteq \mathbb{R}^n$ be a pointed, closed, convex cone with non-empty interior.

Definition 3.2.1. *Given $c \in \mathbb{R}^n$, $a_1, \ldots, a_m \in \mathbb{R}^n$, and $b_1, \ldots, b_m \in \mathbb{R}$, a* primal conic program (in standard form) *is the following maximization problem:*

$$\sup\{c^\mathsf{T} x : x \in K,\ \ a_1^\mathsf{T} x = b_1, \ldots, a_m^\mathsf{T} x = b_m\},$$

which can also be written in a more compact form as

$$\sup\{c^\mathsf{T} x : x \in K,\ Ax = b\},$$

where A is the $m \times n$ matrix with rows $a_1^\mathsf{T}, \ldots, a_m^\mathsf{T}$ and $b = (b_1, \ldots, b_m)^\mathsf{T} \in \mathbb{R}^m$.

We say that $x \in \mathbb{R}^n$ is a *feasible solution (of the primal)* if it lies in the cone K and if it satisfies the equality constraints. It is a *strictly feasible solution* if it additionally lies in the interior of K.

Note that we used a supremum here instead of a maximum. The reason is simply that sometimes the supremum is not attained. We shall see examples in Section 3.5.

3.2.2 Dual conic programs

The principal problem of duality is to find upper bounds for the primal conic program (a maximization problem), in a systematic, or even mechanical way. This is helpful e.g. in formulating optimality criteria and in the design of efficient algorithms. Duality is a powerful technique, and sometimes translating primal problems into dual problems gives unexpected benefits and insights. To define the dual conic program we need the dual cone K^*.

Definition 3.2.2. *Let $K \subseteq \mathbb{R}^n$ be a cone. The* dual cone *K^* of K is*

$$K^* = \{y \in \mathbb{R}^n : y^\mathsf{T} x \geqslant 0 \text{ for all } x \in K\}.$$

Lemma 3.2.3. *If K is a pointed, closed, convex cone with non-empty interior, then the same holds for its dual cone K^*.*

You will prove this in Exercise 3.1. The following property of cones will be useful — you will prove it in Exercise 3.2.

Lemma 3.2.4. *Let K be a closed convex full-dimensional cone. Then we have the equivalence*

$$x \in \text{int } K \iff \forall y \in K^* \backslash \{0\} : y^\mathsf{T} x > 0.$$

Definition 3.2.5. *Let*

$$\sup\{c^\mathsf{T} x : x \in K,\ a_1^\mathsf{T} x = b_1, \ldots, a_m^\mathsf{T} x = b_m\} = \sup\{c^\mathsf{T} x : x \in K,\ Ax = b\}$$

be a primal conic program. Its dual conic program *is the following minimization problem*

$$\inf \left\{ \sum_{j=1}^m y_j b_j : y_1, \ldots, y_m \in \mathbb{R}, \sum_{j=1}^m y_j a_j - c \in K^* \right\},$$

or more compactly,

$$\inf\{b^\mathsf{T} y : y \in \mathbb{R}^m,\ A^\mathsf{T} y - c \in K^*\}.$$

We say that $y \in \mathbb{R}^m$ is a *feasible solution (of the dual)* if $\sum_{j=1}^m y_j a_j - c \in K^*$. It is a *strictly feasible solution* if $\sum_{j=1}^m y_j a_j - c \in \text{int } K^*$.

3.2.3 Geometric interpretation of the primal-dual pair

At first sight, the dual conic program does not look like a conic program, i.e. optimizing a linear function over the intersection of a convex cone by an affine subspace. Although the expression $z = \sum_{i=1}^m y_i a_i - c$ ranges over the intersection of the convex cone K^* with an affine subspace, it might be less clear a priori why the objective function $\sum_{i=1}^m y_i b_i$ has the right form (a linear function in $z = \sum_{i=1}^m y_i a_i - c$).

The following explanation shows how to view the primal and the dual conic program geometrically. This also will bring the dual program into the right form. For this consider the linear subspace

$$L = \{x \in \mathbb{R}^n : a_1^\mathsf{T} x = 0, \ldots, a_m^\mathsf{T} x = 0\},$$

and its orthogonal complement

$$L^\perp = \left\{ \sum_{j=1}^m y_j a_j \in \mathbb{R}^n : y_1, \ldots, y_m \in \mathbb{R} \right\}.$$

We may assume that there exists a point $x_0 \in \mathbb{R}^n$ satisfying $Ax_0 = b$ for, if not, the primal conic program would not have a feasible solution. Note then that

$$b^\mathsf{T} y = x_0^\mathsf{T} A^\mathsf{T} y = x_0^\mathsf{T} \left(\sum_{j=1}^m a_j y_j \right) = x_0^\mathsf{T} \left(\sum_{j=1}^m a_j y_j - c \right) + x_0^\mathsf{T} c.$$

Therefore, the primal conic program can be written as

$$\sup\{c^\mathsf{T} x : x \in K \cap (x_0 + L)\}$$

and the dual conic program as

$$c^\mathsf{T} x_0 + \inf\{x_0^\mathsf{T} z : z \in K^* \cap (-c + L^\perp)\}.$$

Now both the primal and the dual conic programs have the right form and the symmetry between the primal and the dual conic program becomes more clear.

What happens when one builds the dual of the dual? Then one gets a conic program which is equivalent to the primal. This is due to the following lemma.

Lemma 3.2.6. *Let $K \subseteq \mathbb{R}^n$ be a closed convex cone. Then, $(K^*)^* = K$.*

Proof. The inclusion $K \subseteq (K^*)^*$ is easy to verify using the definition only. For the reverse inclusion, one needs the separation theorem (Lemma 1.5.2). Let $x \in \mathbb{R}^n \backslash K$. Then $\{x\}$ and K can be separated by a hyperplane of the form $H = \{z \in \mathbb{R}^n : c^\mathsf{T} z = 0\}$ for some $c \in \mathbb{R}^n \backslash \{0\}$. Say, $K \subseteq H^+ = \{z : c^\mathsf{T} z \geqslant 0\}$ and $c^\mathsf{T} x < 0$. The inclusion $K \subseteq H^+$ shows that $c \in K^*$ and then the inequality $c^\mathsf{T} x < 0$ shows that $x \notin (K^*)^*$ □

3.3 Examples

Now we specialize the cone K to the first three examples of Section 1.5. These three examples are useful for a huge spectrum of applications.

3.3.1 Linear programming (LP)

A conic program where K is the non-negative orthant $\mathbb{R}^n_{\geqslant 0}$ is a *linear program*. We write a primal linear program (in standard form) as

$$\sup\{c^\mathsf{T} x : x \geqslant 0, a_1^\mathsf{T} x = b_1, \ldots, a_m^\mathsf{T} x = b_m\} = \sup\{c^\mathsf{T} x : x \geqslant 0,\ Ax = b\}.$$

The non-negative orthant is self-dual: $(\mathbb{R}^n_{\geqslant 0})^* = \mathbb{R}^n_{\geqslant 0}$. The dual linear program is

$$\inf\left\{\sum_{j=1}^m b_j y_j : y_1, \ldots, y_m \in \mathbb{R}, \sum_{j=1}^m y_j a_j - c \geqslant 0\right\} = \inf\{b^\mathsf{T} y : A^\mathsf{T} y - c \geqslant 0\}.$$

In the case when the problems are not unbounded we could replace the supremum/infimum by maximum/minimum. This is because we are optimizing a linear function over a polyhedron, which is equivalent to optimizing over its set of extreme points, and any polyhedron has finitely many extreme points.

3.3.2 Conic quadratic programming (CQP)

A conic program where K is the second-order cone $\mathcal{L}^{n+1}$ is a *conic quadratic program*. We write a primal conic quadratic program (in standard form) as

$$\sup\{(c,\gamma)^{\mathsf{T}}(x,t) : (x,t) \in \mathcal{L}^{n+1}, (a_1,\alpha_1)^{\mathsf{T}}(x,t) = b_1, \dots (a_m,\alpha_m)^{\mathsf{T}}(x,t) = b_m\}.$$

Here (x,t) stands for the (column) vector in $\mathbb{R}^{n+1}$ obtained by appending a new entry $t \in \mathbb{R}$ to $x \in \mathbb{R}^n$, we use this notation to emphasize the different nature of the vector's components. Recall the definition of the second-order cone $\mathcal{L}^{n+1}$:

$$(x,t) \in \mathcal{L}^{n+1} \text{ if and only if } \|x\|_2 \leqslant t.$$

The second-order cone is self-dual, too — you will show this in Exercise 3.3

$$(\mathcal{L}^{n+1})^* = \mathcal{L}^{n+1}.$$

The dual conic quadratic program is

$$\inf\left\{\sum_{j=1}^m y_j b_j : y_1,\dots,y_m \in \mathbb{R}, \sum_{j=1}^m y_j(a_j,\alpha_j) - (c,\gamma) \in \mathcal{L}^{n+1}\right\}.$$

This can be written in a nicer and more intuitive form using the Euclidean norm. Define the matrix $B \in \mathbb{R}^{n\times m}$ which has a_i as its i-th column, and the vectors $b = (b_j)_{j=1}^m$, $\alpha = (\alpha_j)_{j=1}^m$ and $y = (y_j)_{j=1}^m$ in $\mathbb{R}^m$. Then the dual conic quadratic program can be reformulated as

$$\inf\left\{b^{\mathsf{T}}y : y \in \mathbb{R}^m, \|By - c\|_2 \leqslant \alpha^{\mathsf{T}}y - \gamma\right\}.$$

3.3.3 Semidefinite programming (SDP)

A conic program where K is the cone of semidefinite matrices $\mathcal{S}^n_{\succcurlyeq 0}$ is a *semidefinite program*. We write a primal semidefinite program (in standard form) as

$$\sup\{\langle C, X\rangle : X \succeq 0, \langle A_1, X\rangle = b_1, \dots, \langle A_m, X\rangle = b_m\}.$$

We have already seen earlier that the cone of semidefinite matrices is self-dual:

$$(\mathcal{S}^n_{\succeq 0})^* = \mathcal{S}^n_{\succeq 0}.$$

The dual semidefinite program is

$$\inf\left\{\sum_{j=1}^m y_j b_j : y_1,\dots,y_m \in \mathbb{R}, \sum_{j=1}^m y_j A_j - C \succeq 0\right\}.$$

Engineers and applied mathematicians like to call an inequality of the form $\sum_{i=1}^m y_i A_i - C \succeq 0$ a *linear matrix inequality (LMI)* between the parameters $y_1,\dots,y_m$. It is a convenient way to express a convex constraint posed on the vector $y = (y_1,\dots,y_m)^{\mathsf{T}}$.

3.4 Duality theory

Duality is concerned with understanding the relation between the primal conic program and the dual conic program. We denote the supremum of the primal conic program by p^* and the infimum of the dual conic program by d^*. What is the relation between p^* and d^*? As we see in the next theorem it turns out that in many cases one has equality $p^* = d^*$ and that the supremum as well as the infimum are attained. In these cases duality theory can be very useful because sometimes it is easier to work with the dual problem instead of the primal problem.

Theorem 3.4.1. *Suppose we are given a pair of primal and dual conic programs. Let p^* be the supremum of the primal and let d^* be the infimum of the dual.*

1. ***(weak duality)** Suppose x is a feasible solution of the primal conic program, and y is a feasible solution of the dual conic program. Then,*

   $$c^T x \leqslant b^T y.$$

   In particular $p^* \leqslant d^*$.

2. ***(complementary slackness)** Suppose that the primal conic program attains its supremum at x, and that the dual conic program attains its infimum at y, and that $p^* = d^*$. Then*

   $$\left(\sum_{i=1}^m y_i a_i - c \right)^T x = 0.$$

3. ***(optimality criterion)** Suppose that x is a feasible solution of the primal conic program, and y is a feasible solution of the dual conic program, and equality*

   $$\left(\sum_{i=1}^m y_i a_i - c \right)^T x = 0$$

   holds. Then the supremum of the primal conic program is attained at x and the infimum of the dual conic program is attained at y.

4. ***(strong duality; no duality gap)** If the dual conic program is bounded from below and if it is strictly feasible, then the primal conic program attains its supremum and there is no duality gap: $p^* = d^*$.*

   If the primal conic program is bounded from above and if it is strictly feasible, then the dual conic programs attains its infimum and there is no duality gap.

Before the proof one more comment about the usefulness of weak duality: Suppose you want to solve a primal conic program. If the oracle of Delft, gives you y, then it might be wise to check whether $\sum_{i=1}^m y_i a_i - c$ lies in K^*. If so, then this gives immediately an upper bound for p^*.

The difference $d^* - p^*$ is also called the *duality gap* between the primal conic program and dual conic program.

One last remark: If the dual conic program is not bounded from below: $d^* = -\infty$, then weak duality implies that $p^* = -\infty$, i.e., the primal conic program is infeasible.

Proof. The proof of **weak duality** is important and simple. It reveals the origin of the definition of the dual conic program: We have

$$\sum_{j=1}^{m} y_j b_j = \sum_{j=1}^{m} y_j (a_j^{\mathsf{T}} x) = \left(\sum_{j=1}^{m} y_j a_j \right)^{\mathsf{T}} x \geqslant c^{\mathsf{T}} x,$$

where the last inequality is implied by $\sum_{i=1}^{m} y_i a_i - c \in K^*$ and $x \in K$.

Now **complementary slackness** and the **optimality criterion** immediately follow from this.

Strong duality needs considerably more work. It suffices to prove the first statement (since the second one follows using the symmetry between the primal and dual problems). So we assume that $d^* > -\infty$ and that the dual program has a strict feasible solution. Using these assumptions we will construct a primal feasible solution x^* with $c^{\mathsf{T}} x^* \geqslant d^*$. Then, weak duality implies $p^* = d^*$ and hence x^* is a maximizer of the primal conic program.

Consider the set

$$M = \left\{ \sum_{j=1}^{m} y_j a_j - c : y \in \mathbb{R}^m, b^{\mathsf{T}} y \leqslant d^* \right\}.$$

If $b = 0$ then $d^* = 0$ and setting $x^* = 0$ proves the result immediately. Hence we may assume that there is an index i so that b_i is not zero, and then M is not empty. We first claim that

$$M \cap \operatorname{int} K^* = \varnothing.$$

For suppose not. Then there exists $y \in \mathbb{R}^m$ such that $\sum_{j=1}^{m} y_j a_j - c \in \operatorname{int} K^*$ and $y^{\mathsf{T}} b \leqslant d^*$. Assume without loss of generality that $b_1 < 0$. Then for a small enough $\epsilon > 0$ one would have $(y_1 + \epsilon) a_1 + \sum_{j=2}^{m} y_j a_j - c \in K^*$ with $(y_1 + \epsilon) b_1 + \sum_{j=2}^{m} y_j b_j < y^{\mathsf{T}} b \leqslant d^*$. This contradicts the fact that d^* is the infimum of the dual conic program.

Since M and K^* are both convex sets whose relative interiors do not intersect, we can separate them by an affine hyperplane, according to Theorem 1.3.8. Hence, there is a non-zero vector $x \in \mathbb{R}^n$ so that

$$\sup\{x^{\mathsf{T}} z : z \in M\} \leqslant \inf\{x^{\mathsf{T}} z : z \in K^*\}. \tag{3.1}$$

We shall use this point x to construct a maximizer of the primal conic program which we do in three steps.

First step: $x \in K$.

To see it, it suffices to show that

$$\inf_{z \in K^*} x^\mathsf{T} z \geqslant 0, \tag{3.2}$$

as this implies that $x \in (K^*)^* = K$. We show the inequality by contradiction. Suppose there is a vector $z \in K^*$ with $x^\mathsf{T} z < 0$. Then, for any positive λ, the vector λz lies in the convex cone K^*. Making λ extremely large drives $x^\mathsf{T} \lambda z$ towards $-\infty$. But we reach a contradiction since, by (3.1), the infimum of $x^\mathsf{T} z$ over $z \in K^*$ is lower bounded since $M \neq \emptyset$.

Second step: There exists $\mu > 0$ so that $a_j^\mathsf{T} x = \mu b_j$ ($j \in [m]$) and $x^\mathsf{T} c \geqslant \mu d^*$.

Since $0 \in K^*$ we also have that the infimum of (3.2) is at most 0. So we have shown that the infimum of (3.2) is equal to 0. Therefore, by (3.1), $\sup_{z \in M} x^\mathsf{T} z \leqslant 0$. In other words, by the definition of M, for any $y \in \mathbb{R}^m$,

$$y^\mathsf{T} b \leqslant d^* \implies x^\mathsf{T} \Big(\sum_{j=1}^m y_j a_j - c \Big) \leqslant 0$$

or, equivalently,

$$y^\mathsf{T} b \leqslant d^* \implies \sum_{j=1}^m y_j (x^\mathsf{T} a_j) \leqslant x^\mathsf{T} c.$$

This means that the halfspace $\{y : y^\mathsf{T} b \leqslant d^*\}$ is contained into the halfspace $\{y : y^\mathsf{T} (x^\mathsf{T} a_j)_j \leqslant x^\mathsf{T} c\}$. Hence their normal vectors b and $(x^\mathsf{T} a_j)_j$ point in the same direction. In other words there exists a scalar $\mu \geqslant 0$ such that

$$x^\mathsf{T} a_j = \mu b_j \ (j = 1, \ldots, m), \ \mu d^* \leqslant x^\mathsf{T} c.$$

It suffices now to verify that μ is positive. Indeed suppose that $\mu = 0$. Then, on the one hand, we have that $x^\mathsf{T} c \geqslant 0$. On the other hand, using the assumption that the conic dual program is **strictly feasible**, there exists $\bar{y} \in \mathbb{R}^m$ such that $\sum_j \bar{y}_j a_j - c \in \operatorname{int} K$. This implies

$$0 < \Big(\sum_{j=1}^m \bar{y}_j a_j - c \Big)^\mathsf{T} x = -c^\mathsf{T} x,$$

where strict inequality follows from $\sum_j \bar{y}_j a_j - c \in \operatorname{int} K$ and $x \in K \backslash \{0\}$ (use here Lemma 3.2.4). This gives $c^\mathsf{T} x < 0$, a contradiction.

Third step: $x^* = x/\mu$ is a maximizer of the primal conic program.

This follows directly from the fact that x^* is a primal feasible solution (since we saw above that $x^* \in K$ and $a_j^\mathsf{T} x^* = b_j$ for $j \in [m]$) with $c^\mathsf{T} x^* \geqslant d^*$. □

3.5 Some pathological examples

If you know linear programming and its duality theory you might wonder why do we always write $\sup$ and $\inf$ instead of $\max$ and $\min$ and why do we care about strictly feasibility in Theorem 3.4.1. Why doesn't strong duality always hold? Here are some examples of semidefinite programs showing that we indeed have to be more careful.

3.5.1 Dual infimum not attained

Consider the semidefinite program

$$p^* = \sup\left\{\left\langle \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}, X \right\rangle : X \succeq 0,\ \left\langle \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, X \right\rangle = 1, \left\langle \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, X \right\rangle = 0\right\}$$

and its dual

$$d^* = \inf\left\{y_1 : y_1 \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + y_2 \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} y_1 & 1 \\ 1 & y_2 \end{pmatrix} \succeq 0\right\}.$$

In this example, $p^* = d^* = 0$ and the supremum is attained in the primal, but the infimum is not attained in the dual. Note indeed that the primal is not strictly feasible (since $X_{22} = 0$ for any feasible solution).

3.5.2 Positive duality gap

There can be a duality gap between the primal and the dual conic programs. Consider the primal semidefinite program with data matrices

$$C = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, A_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, A_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

and $b_1 = 0$, $b_2 = 1$. It reads

$$p^* = \sup\{-X_{11} - X_{22} : X_{11} = 0,\ 2X_{13} + X_{22} = 1,\ X \succeq 0\}$$

and its dual reads

$$d^* = \inf\left\{y_2 : y_1 A_1 + y_2 A_2 - C = \begin{pmatrix} y_1 + 1 & 0 & y_2 \\ 0 & y_2 + 1 & 0 \\ y_2 & 0 & 0 \end{pmatrix} \succeq 0\right\}.$$

Then any primal feasible solution satisfies $X_{13} = 0$, $X_{22} = 1$, so that the primal optimum value is equal to $p^* = -1$, attained at the matrix $X = E_{22}$. Any dual feasible solution satisfies $y_2 = 0$, so that the dual optimum value is equal to $d^* = 0$, attained at $y = 0$. Hence there is a positive duality gap: $d^* - p^* = 1$.

3.6 Strong and weak infeasibility

Consider the following two conic programming systems

$$Ax = b,\ x \in K, \tag{3.3}$$

$$\sum_{j=1}^{m} y_j a_j = A^{\mathsf{T}} y \in K^*,\ b^{\mathsf{T}} y < 0. \tag{3.4}$$

Clearly, if (3.3) has a solution then (3.4) has no solution: If x is feasible for (3.3) and y is feasible for (3.4) then

$$0 \leqslant (A^{\mathsf{T}} y)^{\mathsf{T}} x = y^{\mathsf{T}} A x = y^{\mathsf{T}} b < 0,$$

giving a contradiction. When K is the non-negative orthant then the converse also holds: If (3.3) has no solution then (3.4) has a solution. This fact follows by applying the separation theorem (Lemma 1.5.2). Indeed, assume that (3.3) has no solution. Then b does not belong to the cone generated by the columns of A. By Lemma 1.5.2, there exists a hyperplane, having normal $y \in \mathbb{R}^m$, separating $\{b\}$ and this cone spanned by column vectors. So we have the inequalities $A^{\mathsf{T}} y \geqslant 0$ and $y^{\mathsf{T}} b < 0$. This shows that y is feasible for (3.4). We just proved Farkas' lemma for linear programming.

Theorem 3.6.1. (Farkas' lemma for linear programming)
Given $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, exactly one of the following two alternatives holds:

(1) Either the linear system $Ax = b$, $x \geqslant 0$ has a solution,

(2) Or the linear system $A^{\mathsf{T}} y \geqslant 0$, $b^{\mathsf{T}} y < 0$ has a solution.

For general conic programming, it is not true that infeasibility of (3.3) implies feasibility of (3.4). As an illustration, consider the following semidefinite systems:

$$\langle E_{11}, X\rangle = 0, \langle E_{12}, X\rangle = 1,\ X \succeq 0, \tag{3.5}$$

$$y_1 E_{11} + y_2 E_{12} \succeq 0,\ y_2 < 0, \tag{3.6}$$

which are both infeasible.

However, one can formulate the following analogous, although weaker, theorem of alternatives, which needs some strict feasibility condition.

Theorem 3.6.2. *Let $K \subseteq \mathbb{R}^n$ be a full dimensional, pointed, closed and convex cone, let $A \in \mathbb{R}^{m\times n}$ with rows $a_1^{\mathsf{T}}, \ldots, a_m^{\mathsf{T}}$ and let $b \in \mathbb{R}^m$. Assume that the system $Ax = b$ has a solution x_0. Then exactly one of the following two alternatives holds:*

(1) Either there exists $x \in \operatorname{int} K$ such that $Ax = b$.

(2) Or there exists $y \in \mathbb{R}^m$ such that $\sum_{j=1}^m y_j a_j = A^{\mathsf{T}} y \in K^* \backslash \{0\}, b^{\mathsf{T}} y \leqslant 0$.

Proof. Again one direction is clear: If $x \in \operatorname{int} K$ satisfies $Ax = b$ and y satisfies $A^{\mathsf{T}}y \in K^*\backslash\{0\}$ and $b^{\mathsf{T}}y \leqslant 0$, then we get $0 \leqslant (A^{\mathsf{T}}y)^{\mathsf{T}}x = y^{\mathsf{T}}Ax = y^{\mathsf{T}}b \leqslant 0$, implying $(A^{\mathsf{T}}y)^{\mathsf{T}}x = 0$. This gives a contradiction since $x \in \operatorname{int} K$ and $A^{\mathsf{T}}y \in K^*\backslash\{0\}$ (recall Lemma 3.2.4).

Assume now that the system in (1) has no solution. By assumption, the affine space $L = \{x : Ax = b\}$ is not empty, as $x_0 \in L$. Define the linear space

$$\mathcal{L} = \{x : Ax = 0\} = \{x : a_1^{\mathsf{T}}x = 0, \ldots, a_m^{\mathsf{T}}x = 0\}$$

so that $L = \mathcal{L} + x_0$. By assumption, $L \cap \operatorname{int} K = \varnothing$. By the separation theorem (Theorem 1.3.8), there exists a hyperplane separating L and $\operatorname{int} K$: There exists a non-zero vector $c \in \mathbb{R}^n$ and a scalar β such that

$$\forall x \in K : c^{\mathsf{T}}x \geqslant \beta \;\text{ and }\; \forall x \in L : c^{\mathsf{T}}x \leqslant \beta.$$

Then $\beta \leqslant 0$ (as $0 \in K$) and $c \in K^*$ (as $c^{\mathsf{T}}tx \geqslant \beta$ for all $x \in K$ and $t > 0$, which implies that $c^{\mathsf{T}}x \geqslant 0$). Moreover, for any $x \in \mathcal{L}$ and any scalar $t \in \mathbb{R}$, we have that $c^{\mathsf{T}}(tx + x_0) \leqslant \beta$ which implies $c^{\mathsf{T}}x = 0$. Therefore $c \in \mathcal{L}^{\perp}$ and thus c is a linear combination of the a_j's, say $c = \sum_{j=1}^m y_j a_j = A^{\mathsf{T}}y$ for some $y = (y_j) \in \mathbb{R}^m$. So we already have that $A^{\mathsf{T}}y \in K^*\backslash\{0\}$. Finally, $y^{\mathsf{T}}b = y^{\mathsf{T}}Ax_0 = c^{\mathsf{T}}x_0 \leqslant \beta \leqslant 0$ (as $x_0 \in L$). □

Consider again the above example: the system (3.5) is not strictly feasible, and indeed there is a feasible solution to (3.6) after replacing the condition $y_2 < 0$ by $y_2 \leqslant 0$ and adding the condition $y_1E_{11} + y_2E_{12} \neq 0$.

We now further investigate the situation when the primal system (3.3) is infeasible. According to the above discussion, there are two possibilities:

1. Either (3.4) is feasible: There exists $y \in \mathbb{R}^m$ such that $\sum_{j=1}^m y_j a_j \in K^*$ and $b^{\mathsf{T}}y < 0$. Then we say that the system (3.3) is *strongly infeasible*.

2. Or (3.4) is not feasible.

As we will show below, this second alternative corresponds to the case when the system (3.3) is "weakly infeasible", which roughly means that it is infeasible but any small perturbation of it becomes feasible. Here is the exact definition.

Definition 3.6.3. *The system $Ax = b,\ x \in K$ is* weakly infeasible *if it is infeasible and, for any $\epsilon > 0$, there exists $x \in K$ such that $\|Ax - b\| \leqslant \epsilon$.*

For instance, the system (3.5) is weakly infeasible: For any $\epsilon > 0$ the perturbed system $\langle E_{11}, X\rangle = \epsilon$, $\langle E_{12}, X\rangle = 1$, $X \succeq 0$ is feasible.

Theorem 3.6.4. *Consider the two systems (3.3) and (3.4). Assume that the system (3.3) is infeasible. Then exactly one of the following two alternatives holds.*

(1) Either (3.3) is strongly infeasible: There exists $y \in \mathbb{R}^m$ such that $b^{\mathsf{T}}y < 0$ and $\sum_{j=1}^m y_j a_j \in K^*$.

(2) Or (3.3) is weakly infeasible: For every $\epsilon > 0$ there exists $x \in K$ satisfying $\|Ax - b\| \leqslant \epsilon$.

Proof. Assume that (3.3) is not strongly infeasible. Then the two convex sets $\{y : A^\mathsf{T} y \in K^*\}$ and $\{y : b^\mathsf{T} y < 0\}$ are disjoint. By the separation theorem (Theorem 1.3.8) there exists a non-zero vector $c \in \mathbb{R}^m$ such that

$$\inf\{c^\mathsf{T} y : A^\mathsf{T} y \in K^*\} \geqslant \sup\{c^\mathsf{T} y : b^\mathsf{T} y < 0\}.$$

As $0 \in K^*$ the infimum is at most 0. Hence, $b^\mathsf{T} y < 0$ implies $c^\mathsf{T} y \leqslant 0$. This implies that $c = \lambda b$ for some positive λ and, up to rescaling, we can assume that $c = b$. Therefore,

$$\sum_{j=1}^m a_j y_j \in K^* \implies b^\mathsf{T} y \geqslant 0. \tag{3.7}$$

We show that (3.3) is weakly infeasible. For this consider the following program, where we have two new variables $z, z' \in \mathbb{R}^m$:

$$p^* = \inf_{x \in \mathbb{R}^n, z, z' \in \mathbb{R}^m} \{e^\mathsf{T} z + e^\mathsf{T} z' : Ax + z - z' = b,\ x \in K, z, z' \in \mathbb{R}^m_{\geqslant 0}\}, \tag{3.8}$$

where $e = (1, \ldots, 1)^\mathsf{T}$ is the all-ones vector. It suffices now to show that the infimum of (3.8) is equal to 0, since this implies directly that (3.3) is weakly infeasible. For this consider the dual program of (3.8), which can be written as (check it)

$$d^* = \sup_{y \in \mathbb{R}^m} \{b^\mathsf{T} y : -A^\mathsf{T} y \in K^*,\ -e \leqslant y \leqslant e\}. \tag{3.9}$$

Clearly the primal (3.8) is strictly feasible and $d^* \geqslant 0$ (since $y = 0$ is feasible). Moreover, $d^* \leqslant 0$ by (3.7). Hence $d^* = 0$ and thus $p^* = d^* = 0$ since there is no duality gap (applying Theorem 3.4.1). □

Of course the analogous result holds for the dual conic program (which follows using symmetry between primal/dual programs).

Theorem 3.6.5. *Assume that the system*

$$\sum_{j=1}^m y_j a_j - c \in K^* \tag{3.10}$$

is infeasible. Then exactly one of the following two alternatives holds.

(1) Either (3.10) is strongly infeasible: There exists $x \in K$ such that $Ax = 0$ and $c^\mathsf{T} x > 0$.

(2) Or (3.10) is weakly infeasible: For every $\epsilon > 0$ there exist $y \in \mathbb{R}^m$ and $z \in K^*$ such that $\|(\sum_{j=1}^m y_j a_j - c) - z\| \leqslant \epsilon$.

3.7 More on the difference between linear and conic programming

We have already seen above several differences between linear programming and semidefinite programming: there might be a duality gap between the primal and dual programs and the supremum/infimum might not be attained even though they are finite. We point out some more differences regarding rationality and bit size of optimal solutions.

In the classical bit (Turing machine) model of computation an integer number p is encoded in binary notation, so that its bit size is $\log p + 1$ (logarithm in base 2). Rational numbers are encoded as two integer numbers and the bit size of a vector or a matrix is the sum of the bit sizes of its entries.

Consider a linear program

$$\max\{c^{\mathsf{T}} x : Ax = b, x \geqslant 0\} \tag{3.11}$$

where the data A, b, c is *rational*-valued. From the point of view of computability this is a natural assumption and it would be desirable to have an optimal solution which is also rational-valued. A fundamental result in linear programming asserts that this is indeed the case: If program (3.11) has an optimal solution, then it has a *rational* optimal solution $x \in \mathbb{Q}^n$, whose bit size is polynomially bounded in terms of the bit sizes of A, b, c.

On the other hand it is easy to construct instances of semidefinite programming where the data are rational valued, yet there is no rational optimal solution. For instance, the following program

$$\max\left\{x : \begin{pmatrix} 1 & x \\ x & 2 \end{pmatrix} \succeq 0\right\}$$

attains its maximum at $x = \pm\sqrt{2}$.

Consider now the semidefinite program, with variables $x_1, \ldots, x_n$,

$$\inf\left\{x_n : \begin{pmatrix} 1 & 2 \\ 2 & x_1 \end{pmatrix} \succeq 0, \begin{pmatrix} 1 & x_{i-1} \\ x_{i-1} & x_i \end{pmatrix} \succeq 0 \text{ for } i = 2, \ldots, n\right\}.$$

Then any feasible solution satisfies $x_n \geqslant 2^{2^n}$. Hence the bit-size of an optimal solution is exponential in n, thus exponential in terms of the bit-size of the data.

3.8 Further reading

Conic programs, especially linear programs, conic quadratic programs, and semidefinite programs are the central topic in the text book of Ben-Tal and Nemirovski [3]. There also many interesting engineering applications (synthesis of filters and antennas, truss topology design, robust optimization, optimal control, stability analysis and synthesis, design of chips) are covered. This book

largely overlaps with Nemirovski's lecture notes [5] which are available online. A nutshell version of these lecture notes is Nemirovski's plenary talk "Advances in convex optimization: conic programming" at the International Congress of Mathematicians in Madrid 2006 for which a paper and a video is available online: [6]. It is astonishing how much material Nemirovski covers in only 60 minutes.

A second excellent text book on convex optimization is the book by Boyd and Vandenberghe [2] (available online). Here the treated applications are: approximation and fitting, statistical estimation, and geometric problems. Videos of Boyd's course held at Stanford can also be found there.

The duality theory for linear programming which does not involve duality gaps is explained in every book on linear programming. For example, Schrijver [7, Chapter 7] is a good source.

3.9 Historical remarks

The history of conic programming is difficult to trace. Only recently researchers recognized that they give a unifying framework for convex optimization.

In 1956, Duffin in a short paper "Infinite programs" [3] introduced conic programs. His approach even works in infinite dimensions and he focused on these cases. However, the real beginning of conic programming seems to be 1993 when the book "Interior-Point Polynomial Algorithms in Convex Optimization" by Yurii Nesterov and Arkadi Nemirovski was published. There they described for the first time a unified theory of polynomial-time interior point methods for convex optimization problems based on their conic formulations. Concerning the history of conic programs they write:

> Duality for convex program involving "non-negativity constraints" defined by a general-type convex cone in a Banach space is a relatively old (and, possibly, slightly forgotten by the mathematical programming community) part of convex analysis (see, e.g. [ET76]). The corresponding general results, as applied to the case of conic problems (i.e., finite-dimensional problems with general-type non-negativity constraints and *affine* functional constraints), form the contents of $\S 3.2$. To our knowledge, in convex analysis, there was no special interest to conic problems, and consequently to the remarkable symmetric form of the aforementioned duality in this particular case. The only previous result in spirit of this duality known to us it the dual characterization of the Lovasz capacity number $\theta(\Gamma)$ of a graph (see [Lo79]).

3.10 Exercises

3.1 Let $K \subseteq \mathbb{R}^n$ be a cone and let K^* be its dual cone.

(a) Show that K^* is a closed convex cone.

(b) If K is pointed, closed, convex and full-dimensional, show that the same holds for K^*.

3.2 Let K be a closed convex full dimensional cone. Show that

$$x \in \operatorname{int} K \iff y^\mathsf{T} x > 0 \ \forall y \in K^* \backslash \{0\}.$$

3.3 (a) For the Lorentz cone, show that $(\mathcal{L}^{n+1})^* = \mathcal{L}^{n+1}$.

(b) Determine the dual cone of the cone of copositive matrices.

3.4 Consider the following location problem: We are given N locations in the plane $x_1, \ldots, x_N \in \mathbb{R}^2$. Find a point $y \in \mathbb{R}^2$ which minimizes the sum of the distances to the N locations:

$$\min_{y \in \mathbb{R}^2} \sum_{i=1}^{N} d(x_i, y).$$

(a) Formulate this problem as a conic program using the cone

$$\mathcal{L}^{2+1} \times \mathcal{L}^{2+1} \times \cdots \times \mathcal{L}^{2+1}.$$

(b) Determine its dual.

(c) Is there a duality gap?

BIBLIOGRAPHY

[1] A. Ben-Tal, A. Nemirovski, *Lectures on Modern Convex Optimization*, SIAM 2001.

[2] S. Boyd, L. Vandenberghe, *Convex optimization*, Cambridge, 2004.
`http://www.stanford.edu/~boyd/cvxbook/`

[3] R.J. Duffin, *Infinite programs*, pages 157–170 in *Linear Equalities and Related Systems* (A.W. Tucker (ed.)), Princeton University Press, 1956.

[4] M. Grötschel, L. Lovász, A. Schrijver, *Geometric Algorithms in Combinatorial Optimization*, Springer, 1988.

[5] A. Nemirovski, *Lectures on Modern Convex Optimization*,
`http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf`

[6] A. Nemirovski, *Advances in convex optimization: Conic programming*, pages 413–444 in: *Proceedings of International Congress of Mathematicians, Madrid, August 22-30, 2006, Volume 1* (M. Sanz-Sol, J. Soria, J.L. Varona, J. Verdera, Eds.), European Mathematical Society Publishing House, 2007.
Paper: `http://www.icm2006.org/proceedings/Vol_I/21.pdf`
Video: `http://www.icm2006.org/video/` (Eighth Session)

[7] A. Schrijver, *Theory of linear and integer programming*, John Wiley & Sons, 1986.

CHAPTER 4

INTERIOR POINT METHODS

In this lecture we consider the problem of solving a conic program numerically. First, we recall the situation. Let $K \subseteq \mathbb{R}^n$ be a pointed, closed, convex cone with non-empty interior. Given are $c \in \mathbb{R}^n$, $a_1, \ldots, a_m \in \mathbb{R}^n$, and $b_1, \ldots, b_m \in \mathbb{R}$. The primal conic program in standard form is the following maximization problem:

$$\sup\{c^{\mathsf{T}} x : x \in K, a_1^{\mathsf{T}} x = b_1, \ldots, a_m^{\mathsf{T}} x = b_m\}.$$

Since the beginning of the 1990's the theory of efficient interior point methods was developed which basically says that if the cone K is "nice" (this can be made mathematically precise: a sufficient condition is the existence of a self-concordant barrier function which is computable in polynomial time; for the details we refer to the literature), then there exists a polynomial time algorithm which solves the conic program. Solving in polynomial time means that one can in polynomially many steps approximate an optimal solution within any desired precision where the precision is part of the input.

Here we only sketch the rough idea of interior point methods. The idea is to provide only some some background knowledge without giving many details. [This is just enough to implement a program which solves a small conic program with a few variables for instance.] We will ignore many, many technical details: How to guarantee a polynomial time running time? How to implement a method which is numerically stable? Going through all the details (very fascinating applied mathematics!) fills a complete advanced course, namely the LNMB course "Interior point methods".

For the details we refer to the comprehensive books of Nesterov and Nemirovski [5] and of Boyd and Vandenberghe [2] and to the literature given in Section 4.5.

First we present the classical barrier method. The principal ideas developed there form the backbone of the modern polynomial time interior point methods.

Then, we look at the most important properties of the central path of the primal-dual pair of a semidefinite program.

4.1 Classical barrier methods

To explain the basic idea of interior point method we need two ingredients: Newton's method for equality constrained minimization and barrier functions.

4.1.1 Newton's method

We start by recalling Newton's method for unconstrained minimization. Newton's method is an iterative method for finding roots of equations in one or more dimensions. It is one of the most important algorithms in numerical analysis and scientific computing. In convex optimization it can be used to find minimizers of convex differentiable functions. The Newton method is also the fundamental algorithm for the design of fast interior point algorithms.

Unconstrained minimization

Newton's method is quite general. It is natural to define it in the setting of Banach spaces. Chapter XVIII of the book "Functional analysis in normed spaces" by L.V. Kantorovich and G.P. Akilov is a classical resource for this which also includes the first thorough analysis of the convergence behavior of Newton's method. Nowadays every comprehensive book on numerical analysis contains a chapter stating explicit conditions for the convergence speed of Newton's method.

To keep it as simple and concrete as possible we define it here only for $\mathbb{R}^n$. Let Ω be an open set of $\mathbb{R}^n$ and let $f : \Omega \to \mathbb{R}$ be a strictly convex, differentiable function. The Taylor approximation of the function f around the point a is

$$f(a+x) = \left(f(a) + \nabla f(a)^{\mathsf{T}} x + \frac{1}{2} x^{\mathsf{T}} \nabla^2 f(a) x \right) + \text{h.o.t.},$$

where $\nabla f(a) \in \mathbb{R}^n$ is the *gradient* of f at a with entries

$$\nabla f(a) = \left(\frac{\partial}{\partial x_1} f(a), \ldots, \frac{\partial}{\partial x_n} f(a) \right)^{\mathsf{T}},$$

and where $\nabla^2 f(a) \in \mathbb{R}^{n \times n}$ is the *Hessian matrix* of f at a with entries

$$[\nabla^2 f(a)]_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f(a),$$

and where h.o.t. stands for "higher order terms". Since the function is strictly convex, the Hessian matrix is positive definite, $\nabla^2 f(a) \in \mathcal{S}^n_{\succ 0}$. By $q : \mathbb{R}^n \to \mathbb{R}$

we denote the quadratic function which we get by truncating the above Taylor approximation

$$q(x) = f(a) + \nabla f(a)^{\mathsf{T}} x + \frac{1}{2} x^{\mathsf{T}} \nabla^2 f(a) x.$$

This is a strictly convex quadratic function and so it has a unique minimizer $x^* \in \mathbb{R}^n$ which can be determined by setting the gradient of q to zero:

$$\begin{aligned} 0 &= \nabla q(x^*) \\ &= \left(\frac{\partial}{\partial x_1} q(x^*), \ldots, \frac{\partial}{\partial x_n} q(x^*) \right)^{\mathsf{T}} \\ &= \nabla f(a) + \nabla^2 f(a) x^*. \end{aligned}$$

Hence, we find the unique minimizer x^* of q by solving a system of linear equations

$$x^* = -\left(\nabla^2 f(a)\right)^{-1} \nabla f(a).$$

Now Newton's method is based on approximating the function f locally at a starting point a by the quadratic function q, finding the minimizer (the *Newton direction*) x^* of the quadratic function, updating the starting point to $a + x^*$ and repeating this until the desired accuracy is reached:

repeat

$x^* \leftarrow -\left(\nabla^2 f(a)\right)^{-1} \nabla f(a)$

$a \leftarrow a + x^*$

until a stopping criterion is fulfilled.

The following fact about Newton's method are important.

First the good news: If the starting point is close to the minimizer, then the Newton method converges quadratically (for instance the series $n \mapsto \frac{1}{10^{2^n}}$ converges quadratically to its limit 0), i.e. in every step the number of accurate digits is multiplied by a constant number.

However, if the starting point is not close to the minimizer or if the function is close to being not strictly convex, then Newton's method does not converge well. Consider for example the convex but not strictly convex univariate function $f(z) = 1/4z^4 - z$. Then $f'(z) = z^3 - 1$ and $f''(z) = 3z^2$. So if one starts the Newton iteration at $a = 0$, one immediately is in trouble: division by zero. If one starts at $a = -\sqrt[3]{1/2}$, then one can perform a Newton step and one is in trouble again, etc. Figure 4.1.1 shows the fractal structure which is behind Newton's method for solving the equation $f'(z) = z^3 - 1 = 0$ in the complex number plane. One has similar figures for other functions.

This pure Newton method is an idealization and sometimes it cannot be performed at all because it can very well happen, that $a + x^* \notin \Omega$. One can circumvent these problems by replacing the Newton step $a \leftarrow a + x^*$ by a *damped Newton step* $a \leftarrow a + \theta x^*$ with some step size $\theta > 0$ which is chosen

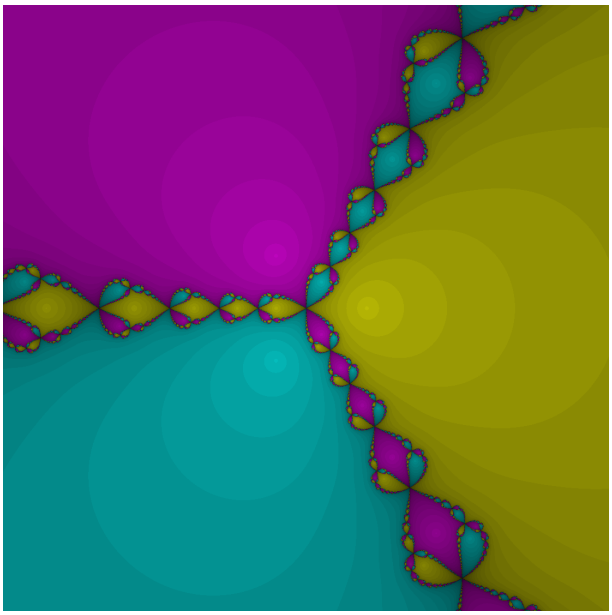


Figure 4.1: Newton fractal of $z^3 - 1 = 0$. The three colors indicate the region of attraction for the three roots. The shade of the color indicates the number of steps needed to come close to the corresponding root. (Source: wikipedia).

to ensure e.g. that $a + \theta x^* \in \Omega$. Choosing the right θ using a line search can be done in many ways. A popular choice is backtracking line search using the Armijo-Goldstein condition.

Let us discuss stopping criteria a bit: One possible stopping criterion is for example if the the norm of the gradient is small, i.e. for some predefined positive ϵ we do the iteration until

$$\|\nabla f(a)\|^2 \leqslant \epsilon. \tag{4.1}$$

We now derive a stopping criterion in the case when the function f is not only strictly convex but also *strongly convex*. This means that there is a positive constant m so that the smallest eigenvalue of all Hessian matrices of f is at least m:

$$\forall a \in \Omega : \lambda_{\min}(\nabla^2 f(a)) \geqslant m.$$

By the Lagrange form of the Taylor expansion we have

$$\forall a, a + x \in \Omega \; \exists \xi \in [a, a + x] : f(a + x) = f(a) + \nabla f(a)^\mathsf{T} x + \frac{1}{2} x^\mathsf{T} \nabla^2 f(\xi) x$$

and the strong convexity of f together with the variational characterization of

the the smallest eigenvalue, which says that

$$\lambda_{\min}(\nabla^2 f(\xi)) = \min_{x \in \mathbb{R}^n \setminus \{0\}} \frac{x^\mathsf{T} \nabla^2 f(\xi) x}{\|x\|^2},$$

gives

$$f(a+x) \geqslant f(a) + \nabla f(a)^\mathsf{T} x + \frac{1}{2} m \|x\|^2.$$

Consider the function of the right hand side

$$x \mapsto f(a) + \nabla f(a)^\mathsf{T} x + \frac{1}{2} m \|x\|^2.$$

It is a convex quadratic function with gradient

$$x \mapsto \nabla f(a) + mx,$$

hence its minimum is attained at

$$x^* = -\frac{1}{m} \nabla f(a).$$

So we have for the minimum μ^* of f

$$\begin{aligned} \mu^* &\geqslant f(a) + \nabla f(a)^\mathsf{T} (-\frac{1}{m} \nabla f(a)) + \frac{1}{2} m \left\| \frac{1}{m} \nabla f(a) \right\|^2 \\ &= f(a) - \frac{1}{2m} \|\nabla f(a)\|^2, \end{aligned}$$

which says that whenever the stopping criterion (4.1) is fulfilled we know that $f(a)$ and μ^* are at most $\epsilon/(2m)$ apart. Of course, the drawback of this consideration is that one has to know or estimate the constant m in advance which is often not easy. Nevertheless the consideration at least shows that the stopping criterion is sensible.

Equality-constrained minimization

In the next step we show how to modify Newton's method if we want to find the minimum of a strictly convex, differentiable function $f : \Omega \to \mathbb{R}$ in an affine subspace given by the equations

$$a_1^\mathsf{T} x = b_1, a_2^\mathsf{T} x = b_2, \ldots, a_m^\mathsf{T} x = b_m,$$

where $a_1, \ldots, a_m \in \mathbb{R}^n$ and $b_1, \ldots, b_m \in \mathbb{R}$.

We define the Lagrange function

$$L(x, \lambda_1, \ldots, \lambda_m) = f(x) + \sum_{i=1}^m \lambda_i a_i^\mathsf{T} x,$$

and the method of *Lagrange multipliers* says that if a point y^* lies in the affine space

$$a_1^\mathsf{T} y^* = b_1, \ldots, a_m^\mathsf{T} y^* = b_m,$$

then it is the unique minimizer of f if and only if

$$\nabla L(y^*) = 0.$$

To find this point y^* we approximate the function f using the Taylor approximation around the point a by

$$q(x) = f(a) + \nabla f(a)^\mathsf{T} x + \frac{1}{2} x^\mathsf{T} \nabla^2 f(a) x$$

and solve the linear system (in the variables x^* and $\lambda_1, \ldots, \lambda_m$)

$$\begin{aligned} a_1^\mathsf{T}(a + x^*) = b_1, \ldots, a_m^\mathsf{T}(a + x^*) = b_m \\ \nabla f(a) + \nabla^2 f(a) x^* + \sum_{i=1}^{m} \lambda_i a_i = 0 \end{aligned}$$

to find the Newton direction x^*. Then we can do the same Newton iterations using damped Newton steps as in the case of unconstrained optimization.

4.1.2 Barrier method

In this section it will be more convenient to consider the following minimization problem instead of the original maximization problem (which is completely equivalent to the maximization problem by switching the sign of the vector c)

$$\inf\{c^\mathsf{T} x : x \in K, a_1^\mathsf{T} x = b_1, \ldots, a_m^\mathsf{T} x = b_m\}.$$

Using Newton's method for equality constrained minimization we know how to deal with the minimization problem

$$\inf\{c^\mathsf{T} x : a_1^\mathsf{T} x = b_1, \ldots, a_m^\mathsf{T} x = b_m\}.$$

Now we have to answer the question: How do we deal with the constraint $x \in K$? The idea will be to start with a point x lying in the interior of K and lying in the affine subspace defined by the m equations and then apply Newton's method for equality constrained minimization always assuring that the next point will lie in the interior of K. For this we add to the objective function $c^\mathsf{T} x$ a barrier function $\phi(x)$ so that we want to minimize

$$c^\mathsf{T} x + \phi(x) \tag{4.2}$$

instead of $c^\mathsf{T} x$. The ideal barrier function would be

$$\phi(x) = \begin{cases} 0, & \text{if } x \in K, \\ \infty, & \text{otherwise.} \end{cases}$$

When we minimize (4.2) we will never consider points not lying in K. This "almost" works, but Newton's method is not applicable. The solution is to replace the ideal barrier function by a barrier function ϕ which is a function that is strictly convex and has the property that

$$x \to \partial K \Longrightarrow \phi(x) \to \infty.$$

Then, Newton's method becomes applicable.

Example 4.1.1. *Examples for barrier functions:*

- $K = \mathbb{R}^n_{\geqslant 0}$: $\phi(x) = -\ln(x_1 \cdots x_n)$.
- $K = \mathcal{L}^{n+1}$: $\phi(x,t) = -\ln(t^2 - x_1^2 - \cdots - x_n^2)$.
- $K = \mathcal{S}^n_{\succcurlyeq 0}$: $\phi(X) = -\ln\det X$.
  We have $\nabla\phi(X) = -X^{-1}$ *and* $(\nabla^2\phi(X))H = X^{-1}HX^{-1}$.

So for a positive parameter $t > 0$ we can solve

$$\inf\{t(c^\mathsf{T}x) + \phi(x) : a_1^\mathsf{T}x = b_1, \ldots, a_m^\mathsf{T}x = b_m\}.$$

using Newton's method. The optimal solution $x(t)$ of this minimization problem is called the *central path*. One can show that if t tends to infinity then $x(t)$ tends to an optimal solution of the original problem. In the next section we show this in the case of semidefinite programming. Solving $x(t)$ for large t is computationally expensive since Newton's method does not converge fast enough when we start from a point which is not close to the optimal solution. Now the idea of the barrier method is to find $x(t)$'s successively for increasing values of t. Then one can use the old $x(t)$ as a starting point for the next Newton method and making use of it's quadratic convergence.

In summary the barrier method has the following scheme:

input:
objective function c, constraints $a_1, \ldots, a_m$, $b_1, \ldots, b_m$,
interior point $x \in \operatorname{int} K$ with $a_1^\mathsf{T}x = b_1, \ldots, a_m^\mathsf{T}x = b_m$,
parameter t, parameter μ (for example $\mu = 10$)
repeat
compute $x(t)$ by Newton's method starting from x
$x \leftarrow x(t)$
$t \leftarrow \mu t$
until a stopping criterion is fulfilled.

4.1.3 Finding a starting point

We are still left with the problem of finding a first interior point x which we need as the input for the previous algorithm. In the case of semidefinite programs the

following approach works: We simply solve another semidefinite program (this is called Phase 1 of the algorithm) for which every symmetric matrix which lies in the affine subspace provides an interior point

$$\inf\{\lambda : X + \lambda I \in \mathcal{S}^n_{\succcurlyeq 0}, \langle A_1, X\rangle = b_1, \ldots, \langle A_m, X\rangle = b_m\},$$

where I denotes the identity matrix. For this problem it is easy to find a strictly feasible solution: One computes a matrix Y in the affine subspace $\langle Y, A_j\rangle = b_j$ and determines its smallest eigenvalue $\lambda_{\min}(Y)$. Hence, the matrix $X = Y + (\epsilon - \lambda_{\min}(Y))$ is a strictly feasible solution for every $\epsilon > 0$. Then we can start to minimize λ. If we find a matrix X is so that λ is negative, it can be used as a starting point for Phase 2, the original problem. If no such negative λ exists, the original problem is infeasible.

4.2 Central path of a semidefinite program

In this section we want to study the central path of a semidefinite program in more detail. These properties give the first ideas for developing a polynomial time interior point algorithm. They also show that the barrier method indeed converges to the right values when t tends to infinity and they even give the rate of convergence.

In the following we consider a primal-dual pair of a semidefinite program where both the primal and the dual are strictly feasible. Then by strong duality in Theorem 3.4.1 the primal attains the supremum, the dual attains the infimum and there is no duality gap. Let us recall the geometric formulation of the primal-dual pair of a semidefinite program

$$\begin{aligned}&\max\{\langle C, X\rangle : X \succcurlyeq 0, \langle A_1, X\rangle = b_1, \ldots, \langle A_m, X\rangle = b_m\}\\ =&\min\left\{\sum_{i=1}^m y_i b_i : y_1, \ldots, y_m \in \mathbb{R}, \sum_{i=1}^m y_i A_i - C \succcurlyeq 0\right\}.\end{aligned}$$

which is (see Section 3.2.3)

$$\begin{aligned}&\max\{\langle C, X\rangle : X \succcurlyeq 0, X \in X_0 + L\}\\ =&\langle X_0, C\rangle + \min\{\langle X_0, Y\rangle : Y \succcurlyeq 0, Y \in -C + L^\perp\},\end{aligned}$$

with linear subspace

$$L = \{X \in \mathcal{S}^n : \langle A_1, X\rangle = 0, \ldots, \langle A_m, X\rangle = 0\},$$

and matrix $X_0 \in \mathcal{S}^n$ with $\langle A_i, X_0\rangle = b_i$ for $i = 1, \ldots, m$.

Let $t > 0$ be a positive parameter. Consider the strictly convex functions

$$\begin{aligned}P_t(X) &= -t\langle C, X\rangle - \ln\det X,\\ D_t(Y) &= t\langle X_0, Y\rangle - \ln\det Y.\end{aligned}$$

Because the original primal-dual pair is strictly feasible one can show (with some effort) that P_t attains a unique minimizer $X(t)$ on the affine subspaces $X_0 + L$ which is strictly feasible for the primal, and that D_t attains a unique minimizer $Y(t)$ on $-C + L^\perp$ which is strictly feasible for the dual. Hence, this defines the *primal-dual central* path $(X(t), Y(t))$.

This primal-dual central path has many nice properties. Some of them are given in the following theorem.

Theorem 4.2.1. *For every $t > 0$ we have the augmented optimality condition*

$$X(t)Y(t) = \frac{1}{t}I. \tag{4.3}$$

Furthermore, the primal dual central path measures the duality gap between the solutions $X(t)$ and $Y(t)$:

$$\langle X_0, C\rangle + \langle X_0, Y(t)\rangle - \langle C, X(t)\rangle = \frac{n}{t}.$$

Proof. Using Lagrangian multipliers we see (write down the condition explicitly) that a matrix X^* is the unique minimizer of the strictly convex function P_t if and only if

$$X^* \succ 0, \;\; X^* \in X_0 + L, \;\; \text{and} \;\; -tC - (X^*)^{-1} \in L^\perp.$$

In the same way, Y^* is the unique minimizer of D_t if and only if

$$Y^* \succ 0, \;\; Y^* \in -C + L^\perp, \;\; \text{and } tX_0 - (Y^*)^{-1} \in L.$$

Hence, $\frac{1}{t}X(t)^{-1}$ is a strictly feasible solution of the dual, and $\frac{1}{t}Y(t)^{-1}$ is a strictly feasible solution of the primal. The gradient of D_t at $\frac{1}{t}X(t)^{-1}$ equals

$$\nabla D_t\left(\frac{1}{t}X(t)^{-1}\right) = tX_0 - \left(\frac{1}{t}X(t)^{-1}\right)^{-1} = tX_0 - tX(t) \in L,$$

Hence, by the characterization of the unique minimizer of P_t we have $Y(t) = \frac{1}{t}X(t)^{-1}$. In the same way one shows symmetrically that $X(t) = \frac{1}{t}Y(t)^{-1}$. This implies the first statement.

The second statement follows easily from the first: Let $y_1, \ldots, y_m$ be so that $\sum_{i=1}^m y_i b_i = \langle X_0, C\rangle + \langle X_0, Y(t)\rangle$, then

$$\begin{aligned}
& \langle X_0, C\rangle + \langle X_0, Y(t)\rangle - \langle C, X(t)\rangle \\
= \;& \sum_{i=1}^m y_i b_i - \langle C, X(t)\rangle \\
= \;& \sum_{i=1}^m y_i \langle A_i, X(t)\rangle - \langle C, X(t)\rangle \\
= \;& \langle Y(t), X(t)\rangle \\
= \;& \text{Tr}\left(\frac{1}{t}I\right) \\
= \;& \frac{n}{t}.
\end{aligned}$$

□

Compare (4.3) to the optimality condition in Theorem 3.4.1. In particular, it shows that if $t \to \infty$, then $X(t)$ converges to a primal optimal solution and $Y(t)$ converges to a dual optimal solution. Another important point for the analysis of interior point algorithms is that the theorem gives the rate of convergence which is proportional to $\frac{1}{t}$.

4.3 Software

One very good thing about conic programs such as linear programs, convex quadratic programs, and semidefinite programs is that they can be solved efficiently in **theory** and in **practice**. That they can be solved efficiently in theory means that they can be solved in polynomial time to any given precision. That they can be solved efficiently in practice means that there are software packages available which can be used to solve these problems up to some decent sizes.

ILOG CPLEX is known to be a high-performance mathematical programming solver for linear programming, mixed integer programming and quadratic programming. It can be used to solve very large, real-world optimization problems. ILOG CPLEX contains interior point methods for linear programming as well as for convex quadratic programming (but no semidefinite programming). It is free for academic use.

`http://www.ibm.com/software/integration/optimization/cplex-optimizer/`

Semidefinite program solvers are currently slightly less powerful but at least they can solve problems of moderate size involving matrices having size 1000×1000.

One semidefinite programming solver which is easy to use is CVXOPT by Joachim Dahl and Lieven Vandenberghe:

`http://abel.ee.ucla.edu/cvxopt/userguide/index.html`

It is also part of sage. Sage is a free open-source mathematics software system licensed under the GPL. It combines the power of many existing open-source packages into a common python-based interface. In particular it is not difficult to install.

`http://www.sagemath.org/`

Many more software packages for semidefinite programs can be found for example on the NEOS server for optimization:

`http://neos.mcs.anl.gov/neos/solvers/sdp:csdp/SPARSE_SDPA.html`

Here one can also submit the optimization problem online. This has the advantage that one does not have to install the software locally. The input format is explained here:

`http://plato.asu.edu/ftp/sdpa_format.txt`

Essentially one has to specify the matrix sizes and the nonzero entries of the matrices C, A_i and the values of b_i. One important hint! Note that the role of the primal and dual are "switched" in the documentation.

4.4 Historical remarks

Looking at the milestones of the history of mathematical programming shows that interior point methods for conic programs can be seen as the result of the development of efficient, polynomial-time algorithms.

1947 Dantzig invented the simplex algorithm for linear programming. The simplex algorithm works extremely good in practice, but until today nobody really understands why (although there are meanwhile good theoretical indications). It is fair to say that the simplex algorithm is one of the most important algorithms invented in the last century.

1972 Klee and Minty found a linear program for which the simplex algorithm is extremely slow (when one uses Dantzig's most-negative-entry pivoting rule): It uses exponentially many steps.

1979 Khachian invented the ellipsoid method for linear programming which runs in polynomial time. It is a great theoretical algorithm but until today it did not have any practical impact.

1984 Karmakar showed that one can use interior-point methods for designing a polynomial-time algorithm for linear programming. Nowadays, interior-point methods can compete with the simplex algorithm.

1994 Nesterov and Nemirovski generalized Karmarkar's result to conic programming with the use of self-concordant barrier functions.

since 1994 Every day conic programming becomes more useful (in theory and practice).

It is fair to say that during the last twenty years there has been a revolution in mathematical optimization based on the development of efficient interior point algorithms for convex optimization problems.

Margaret H. Wright begins her survey "The interior-point revolution in optimization: History, recent developments, and lasting consequences" [10] with:

> REVOLUTION:
>
> (i) a sudden, radical, or complete change;
>
> (ii) a fundamental change in political organization, especially the overthrow or renunciation of one government or ruler and the substitution of another.
>
> It can be asserted with a straight face that the field of continuous optimization has undergone a revolution since 1984 in the sense of the

> first definition and that the second definition applies in a philosophical sense: Because the interior-point presence in optimization today is ubiquitous, it is easy to lose sight of the magnitude and depth of the shifts that have occurred during the past twenty years. Building on the implicit political metaphor of our title, successful revolutions eventually become the status quo.
>
> The interior-point revolution, like many other revolutions, includes old ideas that are rediscovered or seen in a different light, along with genuinely new ideas. The stimulating interplay of old and new continues to lead to increased understanding as well as an ever-larger set of techniques for an ever-larger array of problems, familiar and heretofore unexplored. Because of the vast size of the interior-point literature, it would be impractical to cite even a moderate fraction of the relevant references, but more complete treatments are mentioned throughout. The author regrets the impossibility of citing all important work individually.

4.5 Further reading

There are quite some books on interior point methods. The classical barrier method is developed in the book [3] by Fiacco and McCormick. The standard reference is the book [5] by Nesterov and Nemirovski which is not completely easy to read. Boyd and Vandenberghe [2] and Ye [11] as well as [3, Chapter 6] are very helpful. Then, the books [6] by Renegar and [7] by Roos, Terlaky, Vial consider interior point methods for linear programs. There are some surveys available: Nemirovski, Todd [4], Vandenberghe, Boyd [9], Todd [8].

BIBLIOGRAPHY

[1] A. Ben-Tal, A. Nemirovski, *Lectures on Modern Convex Optimization*, SIAM 2001.

[2] S. Boyd, L. Vandenberghe, *Convex optimization*, Cambridge, 2004. `http://www.stanford.edu/~boyd/cvxbook/`

[3] A.V. Fiacco, G.P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley & Sons, 1968.

[4] A.S. Nemirovski, M.J. Todd, *Interior-point methods for optimization*, Acta Numerica (2008), 191-234.

`http://www2.isye.gatech.edu/~nemirovs/Published.pdf`

[5] Y. Nesterov, A. Nemirovski, *Interior-point polynomial methods in convex programming*, SIAM, 1994.

[6] J. Renegar *A mathematical view of interior-point methods in convex optimization*, SIAM, 2001.

[7] C. Roos, T. Terlaky, J.-Ph. Vial, *Theory and Algorithms for Linear Optimization: An Interior Point Approach*, John Wiley & Sons, 1997.

[8] M.J. Todd, *Semidefinite optimization*, Acta Numerica **10** (2001), 515–560.

`http://people.orie.cornell.edu/~miketodd/soa5.pdf`

[9] L. Vandenberghe, S. Boyd, *Semidefinite programming*, SIAM Review 38 (1996), 49-95.

`http://stanford.edu/~boyd/papers/sdp.html`

[10] Margaret H. Wright, *The interior-point revolution in optimization: History, recent developments, and lasting consequence*, Bull. Amer. Math. Soc. 42 (2005), 39–56.

`http://www.ams.org/journals/bull/2005-42-01/S0273-0979-04-01040-7/S0273-0979-04-01040-7.pdf`

[11] Y. Ye, *Interior Point Algorithms, Theory and Analysis*, John Wiley & Sons, 1997.

Part II

Applications in combinatorics

CHAPTER 5

$0/1$ OPTIMIZATION

Linear optimization problems in which the variables only can attain the values 0 or 1 are called $0/1$ linear optimization problems. A $0/1$ *linear optimization problem in standard form* is of the form

$$\begin{aligned} \max \quad & c^{\mathsf{T}} x \\ & x \in \{0,1\}^n, \\ & Ax \leqslant b, \end{aligned} \tag{5.1}$$

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

Many problems in combinatorial optimization can be written as $0/1$ linear optimization problems. One example is finding the independence number $\alpha(G)$ of a graph $G = (V, E)$:

$$\begin{aligned} \alpha(G) = \max \quad & \sum_{v \in V} x_v \\ & x \in \{0,1\}^V, \\ & x_u + x_v \leqslant 1 \quad \text{for all } \{u, v\} \in E. \end{aligned}$$

Another example is finding the domination number of a graph G. A *dominating set* of the graph G is a subset $U \subseteq V$ of its vertices so that every vertex in V is connected to at least one vertex in U. The cardinality of a smallest dominating set is the *domination number* $\gamma(G)$:

$$\begin{aligned} \gamma(G) = \min \quad & \sum_{v \in V} x_v \\ & x \in \{0,1\}^V, \\ & \sum_{u : \{u,v\} \in E} x_u \geqslant 1 \quad \text{for all } v \in V. \end{aligned}$$

These two problems, like many other problems in combinatorial optimization, are difficult to solve computationally. They are NP-hard. So we do not expect that there is an efficient algorithm solving them which runs in polynomial time.

One possibility to deal with NP-hard optimization problems is to *relax* them. The set of feasible solutions

$$\mathcal{F} = \{x \in \mathbb{R}^n : x \in \{0,1\}^n, Ax \leqslant b\}$$

of a $0/1$ linear optimization problem is a subset of the vertices of the cube $[0,1]^n$. We denote the convex hull of all feasible solutions, which is a polytope, by $P = \operatorname{conv} \mathcal{F}$. So solving (5.1) is equivalent to solving $\max\{c^\mathsf{T} x : x \in P\}$ because the maximum of the linear function $c^\mathsf{T} x$ is attained at an extreme point of the polytope P. By relaxing (5.1) we mean that we replace P by a larger set P', not necessarily a polytope, which contains P and for which we can solve $\max\{c^\mathsf{T} x : x \in P'\}$ efficiently. This maximum value provides an upper bound for the original maximization problem (5.1).

In Section 5.1 we explain a simple method how to construct such a set P' using semidefinite optimization. Here the theta number and the semidefinite relaxation of max cut of a graph will be important examples.

In Section 5.2 we will go much further. We will consider $0/1$ polynomial optimization problems which are much more general than $0/1$ linear optimization problems. We explain a hierarchy of stronger and stronger relaxations which even converges to the original problem. This method is one of the strongest general purpose techniques to attack difficult combinatorial optimization problems.

5.1 Relaxations using quadratic optimization

Another standard class of optimization problems are *quadratic optimization problems*. They are of the form

$$\begin{aligned} \max \quad & x^\mathsf{T} Q_0 x + b_0^\mathsf{T} x + \alpha_0 \\ & x \in \mathbb{R}^n, \\ & x^\mathsf{T} Q_j x + b_j^\mathsf{T} x + \alpha_j = 0 \quad \text{for all } j \in [m], \end{aligned} \tag{5.2}$$

where $Q_j \in \mathcal{S}^n$ are symmetric matrices, $b_j \in \mathbb{R}^n$ vectors, and $\alpha_j \in \mathbb{R}^n$ scalars.

It is easy to see that one can transform $0/1$ linear optimization problems into quadratic optimization problems. The constraint

$$x_i^2 - x_i = 0$$

forces feasible solutions to be $0/1$-valued. For inequality constraints $a_j^\mathsf{T} x \leqslant b_j$ we introduce a slack variable s_j and the quadratic equality constraint

$$s_j^2 + a_j^\mathsf{T} x - b_j = 0.$$

Example 5.1.1. *The independence number can be formulated as*

$$\begin{aligned}\alpha(G) = \max \quad & \sum_{v \in V} x_v \\ & x \in \mathbb{R}^V, s \in \mathbb{R}^E, \\ & x_v^2 - x_v = 0 \quad \textit{for all } v \in V, \\ & s_e^2 + x_u + x_v - 1 = 0 \quad \textit{for all } e = \{u, v\} \in E.\end{aligned}$$

Sometimes problems in combinatorial optimization also come naturally as quadratic optimization problems. One example which we already saw is the max-cut problem

$$\begin{aligned}\mathrm{MAXCUT}(G) = \max \quad & \sum_{\{u,v\} \in E} (1 - x_u x_v)/2 \\ & x \in \mathbb{R}^V, \\ & x_v^2 = 1 \quad \textit{for all } v \in V.\end{aligned}$$

The $-1/+1$-constraint $x_v = \pm 1$ is equivalent to the quadratic constraint $x_v^2 = 1$. Also the independence number has a natural quadratic formulation

$$\begin{aligned}\alpha(G) = \max \quad & \sum_{v \in V} x_v^2 \\ & x \in \mathbb{R}^V, \\ & x_v^2 - x_v = 0 \quad \textit{for all } v \in V, \\ & x_u x_v = 0 \quad \textit{for all } \{u, v\} \in E.\end{aligned}$$

Now we would like to relax the quadratic optimization problem by a semidefinite program. For this we rewrite quadratic expressions as trace inner products of symmetric matrices with $n + 1$ rows and columns

$$x^\mathsf{T} Q_j x + b_j^\mathsf{T} x + \alpha_j = \left\langle \begin{pmatrix} \alpha_j & \frac{1}{2} b_j^\mathsf{T} \\ \frac{1}{2} b_j & Q_j \end{pmatrix}, \begin{pmatrix} 1 & x^\mathsf{T} \\ x & x x^\mathsf{T} \end{pmatrix} \right\rangle.$$

Note that the optimization variable has the following special structure: The matrix

$$Y = \begin{pmatrix} 1 & x^\mathsf{T} \\ x & x x^\mathsf{T} \end{pmatrix}$$

is a semidefinite matrix of rank 1. In $0/1$ linear programming, the constraint $x_i^2 - x_i = 0$ translates into $Y_{i0} = Y_{ii}$. When we are dealing with $-1/+1$-valued variables, the constraint $x_i^2 = 1$ translates into $Y_{ii} = 1$.

One can get a semidefinite relaxation of the quadratic optimization problem

(5.2) if one simply ignores that the optimization variable Y has rank 1

$$\begin{aligned} \max \quad & \left\langle \begin{pmatrix} \alpha_0 & \frac{1}{2}b_0^\mathsf{T} \\ \frac{1}{2}b_0 & Q_0 \end{pmatrix}, Y \right\rangle \\ & Y \in \mathcal{S}^{n+1}_{\succcurlyeq 0}, \\ & Y_{00} = 1, \\ & \left\langle \begin{pmatrix} \alpha_j & \frac{1}{2}b_j^\mathsf{T} \\ \frac{1}{2}b_j & Q_j \end{pmatrix}, Y \right\rangle = 0 \quad \text{for all } j \in [m]. \end{aligned} \tag{5.3}$$

It is clear, as we now longer impose rank-1 constraint of the solution, that the optimal values of this maximization problem (5.3) is an upper bound of the original quadratic optimization problem (5.2). The set of feasible solutions of (5.2)

$$\{x \in \mathbb{R}^n : x^\mathsf{T} Q_j x + b_j^\mathsf{T} x + \alpha_j = 0, j \in [m]\}$$

is contained in

$$\left\{ (Y_{11}, \ldots, Y_{nn})^\mathsf{T} \in \mathbb{R}^n : Y \in \mathcal{S}^{n+1}_{\succcurlyeq 0}, Y_{00} = 1, \left\langle \begin{pmatrix} \alpha_j & \frac{1}{2}b_j^\mathsf{T} \\ \frac{1}{2}b_j & Q_j \end{pmatrix}, Y \right\rangle = 0, j \in [m] \right\}.$$

Geometrically this means that we first lift the n-dimensional situation of the quadratic optimization problem into a space of dimension $\binom{n+2}{2} - 1$ in which the relaxed semidefinite optimization problem lives. The matrix Y has $\binom{n+2}{2} - 1$ variable entries. We associate with variable x_i the variable Y_{i0} and with the product $x_i x_j$ the variable Y_{ij}. In the higher dimensional space the quadratic equalities translate into linear inequalities. Then we project the (convex) feasible region of semidefinite program back to n dimensions.

Dualizing (5.3) yields

$$\begin{aligned} \min \quad & y_0 \\ & y_0, \ldots, y_m \in \mathbb{R}, \\ & y_0 \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \sum_{j=1}^m y_j \begin{pmatrix} \alpha_j & \frac{1}{2}b_j^\mathsf{T} \\ \frac{1}{2}b_j & Q_j \end{pmatrix} - \begin{pmatrix} \alpha_0 & \frac{1}{2}b_0^\mathsf{T} \\ \frac{1}{2}b_0 & Q_0 \end{pmatrix} \succcurlyeq 0. \end{aligned}$$

By weak duality every feasible solution of the dual provides an upper bound of the original quadratic optimization problem. Thus this provides a rather simple way to *prove* upper bounds of the original problem. So we can certify the quality of solutions by this.

Example 5.1.2. *Let us apply this relaxation to the quadratic formulations of α and* MAXCUT. *For the independence number we get*

$$\begin{aligned} \alpha(G) \leqslant \max \quad & \sum_{v \in V} Y_{vv} \\ & Y \in \mathcal{S}^{V \cup \{0\}}_{\succcurlyeq 0}, \\ & Y_{00} = 1, \\ & Y_{0v} = Y_{vv} \quad \textit{for all } v \in V, \\ & Y_{uv} = 0 \quad \textit{for all } \{u, v\} \in E. \end{aligned}$$

The matrix variable Y has one row/column more than the number of vertices of the graph. This extra row/column is indexed by the index 0. In Exercise 5.1 you will show that this semidefinite optimization problem is equivalent to the theta number $\vartheta(G)$. For MAXCUT *one gets that* $\mathrm{MAXCUT}(G)$ *is upper bounded by the semidefinite optimization problem*

$$\begin{aligned} \max \quad & \left\langle \begin{pmatrix} |E| & 0^\mathsf{T} \\ 0 & -\frac{1}{2} \sum_{\{u,v\} \in E} E_{uv} \end{pmatrix}, Y \right\rangle \\ & Y \in \mathcal{S}^{V \cup \{0\}}_{\succcurlyeq 0}, \\ & Y_{00} = 1, \\ & Y_{vv} = 1 \quad \textit{for all } v \in V. \end{aligned}$$

(recall $E_{uv} = \frac{1}{2}(e_u e_v^\mathsf{T} + e_v e_u^\mathsf{T})$). Here it is obvious that this optimization problem is equivalent to the semidefinite relaxation sdp from Chapter 2.

5.2 A hierarchy of semidefinite programs

Now we consider $0/1$ polynomial optimization problems where the constraints are allowed to be polynomial inequalities rather than linear inequalities

$$\begin{aligned} \max \quad & c^\mathsf{T} x \\ & x \in \{0,1\}^n, \\ & p_j(x) \leqslant b_j \quad \text{for all } j \in [m]. \end{aligned}$$

with polynomials $p_1, \ldots, p_m \in \mathbb{R}[x_1, \ldots, x_n]$. An optimal solution of this $0/1$ polynomial optimization problem is attained at a vertex of the polytope

$$P = \operatorname{conv}\left(\{x \in \mathbb{R}^n : p_1(x) \leqslant b_1, \ldots, p_m(x) \leqslant b_m\} \cap \{0,1\}^n\right)$$

The following standard form of a $0/1$ polynomial optimization problems will become handy

$$\begin{aligned} \max \quad & c^\mathsf{T} x \\ & x \in \{0,1\}^n, \\ & p_j(x) \geqslant 0, \text{for all } j \in [m]. \end{aligned} \tag{5.4}$$

The goal of the lecture is to construct a hierarchy of relaxations of P consisting out of convex bodies which are projections of feasible regions of semidefinite programs (known as *spectrahedra*) and which find the exact polytope P after at most n steps:

$$K_1 \supseteq K_2 \supseteq K_3 \supseteq \ldots \supseteq K_n = P.$$

One attractive feature of this construction will be that one can optimize a linear function $c^\mathsf{T} x$ over each step of the hierarchy K_t in polynomial time once t is a fixed, which is independent of the input size.

We start by giving the construction of the last step in the hierarchy K_n. For this we need some facts from combinatorics.

5.2.1 Harmonic analysis on power sets

Let X be a finite set. For many applications this will be simply $X = [n] = \{1, \ldots, n\}$. By 2^X we denote the set of all subsets of X, the *power set* of X. The space

$$L^2(2^X) = \{f : 2^X \to \mathbb{R}\}$$

of real-valued functions is a Euclidean space with inner product

$$(f, g) = \sum_{A \in 2^X} f(A)g(A), \text{ with } f, g \in L^2(2^X).$$

If $X = [n]$ then the space $L^2(2^X)$ is simply the vector space $\mathbb{R}^{2^n}$ and (f, g) is the standard inner product between vectors, and $f(\{i_1, \ldots, i_j\})$ is the component of the vector f with index $\{i_1, \ldots, i_j\}$.

We like to write $L^2(2^X)$ nevertheless because then resemblance with concepts in harmonic (Fourier) analysis, like Fourier coefficients, convolutions, functions of positive type, will become more pronounced.

Biorthogonal bases

Two explicit bases of $L^2(2^X)$ will play an important role for us.

The first one is defined as follows: For $B \in 2^X$ we define $\chi_B \in L^2(2^X)$ by

$$\chi_B(A) = \begin{cases} 1 & \text{if } A \subseteq B, \\ 0 & \text{otherwise.} \end{cases}$$

The fact that the function χ_B, with $B \in 2^X$, forms a basis of $L^2(2^X)$ can be easily seen: If one considers the function χ_B as column vectors in $\mathbb{R}^{2^X}$ where we order the elements of 2^X by increasing cardinality, then the matrix $(\chi_B)_{B \in 2^X}$ is an upper triangular matrix in which all diagonal elements are not zero. Note that the value $\chi_B(A)$ coincides with $\xi(A, B)$ where ξ is the zeta-function of the Boolean lattice 2^X, see for instance Aigner [1, Chapter IV.1]. From the definition it is immediate, but extremely important, that this basis is *multiplicative*:

$$\chi_B(A \cup A') = \chi_B(A)\chi_B(A').$$

The second basis is the dual basis χ_B^* of χ_B which is defined by the biorthogonality relation

$$(\chi_B, \chi_{B'}^*) = \begin{cases} 1 & \text{if } B = B', \\ 0 & \text{otherwise.} \end{cases}$$

Although we will not need it here, one can write down the second basis χ_B^* explicitly using the Möbius function μ of the Boolean lattice 2^X, namely

$$\chi_B^*(A) = \mu(A, B) = \begin{cases} (-1)^{|B|-|A|} & \text{if } A \subseteq B, \\ 0 & \text{otherwise.} \end{cases}$$

Example 5.2.1. *For the set $X = \{1, 2, 3\}$ the matrix with column vectors χ_B is*

$$\begin{array}{c} \\ \varnothing \\ 1 \\ 2 \\ 3 \\ 12 \\ 13 \\ 23 \\ 123 \end{array} \begin{array}{c} \begin{array}{cccccccc} \varnothing & 1 & 2 & 3 & 12 & 13 & 23 & 123 \end{array} \\ \left(\begin{array}{cccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right) \end{array}$$

and its inverse contains as row vectors χ_B^*

$$\begin{array}{c} \\ \varnothing \\ 1 \\ 2 \\ 3 \\ 12 \\ 13 \\ 23 \\ 123 \end{array} \begin{array}{c} \begin{array}{cccccccc} \varnothing & 1 & 2 & 3 & 12 & 13 & 23 & 123 \end{array} \\ \left(\begin{array}{cccccccc} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 0 & 1 & 0 & 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 & 0 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right) \end{array}.$$

Every function $f \in L^2(2^X)$ comes in this way with two basis expansions:

$$f(A) = \sum_{B \in 2^X} \widehat{f}(B)\chi_B(A) = \sum_{B \in 2^X} \check{f}(B)\chi_B^*(A),$$

where

$$\widehat{f}(B) = (f, \chi_B^*), \text{ and } \check{f}(B) = (f, \chi_B).$$

Here again we use the notation $\widehat{f}$ to pronounce the similarity to harmonic analysis. One can think as $\widehat{f}$ as the "Fourier coefficient" of f. Classically Fourier coefficients of a function $f \in L^2([0, 2\pi])$ is given by the expansion

$$f(x) = \sum_{n=-\infty}^{\infty} \widehat{f}(n)e^{2\pi inx}$$

In our case the basis functions χ_B play the role of the exponential basis functions and the multiplicativity

$$\chi_B(A \cup A') = \chi_B(A)\chi_B(A')$$

corresponds to

$$e^{2\pi i(n+m)x} = e^{2\pi inx}e^{2\pi imx}.$$

Formally we are doing harmonic analysis on the semigroup $(2^X, \cup)$.

We have *Plancherel's identity*

$$(f,g) = \sum_{B\in 2^X} \widehat{f}(B)\breve{g}(B) = \sum_{B\in 2^X} \breve{f}(B)\widehat{g}(B).$$

For a subset $A \in 2^X$ we define the *shifted function* $f_A \in L^2(2^X)$ by

$$f_A(A') = f(A \cup A').$$

Then,

$$\widehat{f_A}(B) = \widehat{f}(B)\chi_B(A)$$

because

$$f_A(A') = f(A \cup A') = \sum_{B\in 2^X} \widehat{f}(B)\chi_B(A \cup A') = \sum_{B\in 2^X} \widehat{f}(B)\chi_B(A)\chi_B(A').$$

The *convolution* between two functions $f, g \in L^2(2^X)$ is

$$(f * g)(A) = (f, g_A) = \sum_{B\in 2^X} \breve{f}(B)\widehat{g}(B)\chi_B(A),$$

and so

$$\widehat{f * g}(B) = \breve{f}(B)\widehat{g}(B).$$

Note that the convolution is not commutative.

Functions of positive type

Definition 5.2.2. *We say that a function $f \in L^2(2^X)$ is of positive type if the symmetric matrix $M_f \in \mathcal{S}^{2^X}$ defined by*

$$M_f(A,B) = f(A \cup B)$$

is positive semidefinite.

Because of the special structure of the matrix M_f, the entry $M_f(A,B)$ only depends on the union $A \cup B$, it is sometimes called the moment matrix of f. The following theorem gives a characterization of functions of positive type.

> A side remark: The theorem is in the spirit of Bochner's theorem in harmonic analysis which says that functions of positive type are the Fourier transform of nonnegative measures.

Theorem 5.2.3. *A function $f \in L^2(2^X)$ is of positive type if and only if it is of the form*

$$f(A) = \sum_{B\in 2^X} \widehat{f}(B)\chi_B(A), \quad \widehat{f}(B) \geqslant 0.$$

In other words, the cone of all positive semidefinite moment matrices is a polyhedral cone, even simplicial cone, with extreme rays

$$M_B(A,A') = \chi_B(A \cup A'), \quad B \in 2^X.$$

Proof. One direction follows immediately from the multiplicativity of χ_B. For the other direction suppose that $f \in L^2(2^X)$ has the expansion

$$f(A) = \sum_{B \in 2^X} \widehat{f}(B)\chi_B(A),$$

where $\widehat{f}(B') < 0$ for some $B' \in 2^X$. Then,

$$\begin{aligned}
\sum_{A,A' \in 2^X} M_f(A \cup A')\chi^*_{B'}(A)\chi^*_{B'}(A') &= \sum_{A,A',B \in 2^X} \widehat{f}(B)\chi_B(A \cup A')\chi^*_{B'}(A)\chi^*_{B'}(A') \\
&= \sum_{B \in 2^X} \widehat{f}(B) \left(\sum_{A \in 2^X} \chi_B(A)\chi^*_{B'}(A) \right)^2 \\
&= \widehat{f}(B') \\
&< 0,
\end{aligned}$$

hence M_f is not positive semidefinite and f is not of positive type. □

5.2.2 Lasserre's hierarchy

Now we are ready to construct the hierarchy and prove that it converges.

Equivalent reformulation

Since we are dealing with $0/1$ problems we can assume that the polynomials $p_1, \ldots, p_m$ are square free, i.e. if one considers the polynomial p_i as the univariate polynomial in the variable x_j then its degree is at most one.

We identify square free monomials

$$m = x_{i_1}x_{i_2}\cdots x_{i_j} \in \mathbb{R}[x_1, \ldots, x_n]$$

with elements in $L^2(2^X)$ by

$$m = e_{\{i_1,i_2,\ldots,i_j\}} = \sum_{A : A \supseteq \{i_1,i_2,\ldots,i_j\}} \chi^*_A \in L^2(2^X),$$

where $e_{\{i_1,i_2,\ldots,i_j\}}$ denotes an element of the standard basis in $L^2(X)$ (so we even work with a third basis of $L^2(2^X)$ here) and we extend this by linearity to every square free polynomial.

Let $a \in \{0,1\}^X$ be a binary vector and let $A \in 2^X$ be the corresponding subset, i.e. we have $a_x = 1$ iff $x \in A$. Warning: We will switch between $0/1$ vectors and subsets without (another) warning.

Then we can evaluate a square free polynomial p by

$$p(a) = (p, \chi_A) = \breve{p}(A),$$

since

$$(m, \chi_A) = \begin{cases} 1 & \text{if } \{i_1, \ldots, i_j\} \subseteq A, \\ 0 & \text{otherwise.} \end{cases}$$

Example 5.2.4. *Given are the set $X = \{1, 2, 3\}$ and the polynomial $p(x_1, x_2, x_3) = 2x_1 + 3x_1x_2$. Then*

$$p(1, 1, 0) = (p, \chi_{12}) = (2(\chi_1^* + \chi_{12}^* + \chi_{13}^* + \chi_{123}^*) + 3(\chi_{12}^* + \chi_{123}^*), \chi_{12}) = 5.$$

Theorem 5.2.5. *Let*

$$\mathcal{F} = \{x \in \{0, 1\}^n : p_1(x) \geqslant 0, \ldots, p_m(x) \geqslant 0\}$$

be the set of feasible solutions of the $0/1$ polynomial optimization problem. Then a function $f \in L^2(2^{[n]})$ is of the form

$$f(A) = \sum_{b \in \mathcal{F}} \widehat{f}(B)\chi_B(A), \quad \textit{with } \widehat{f}(B) \geqslant 0$$

if and only if

a) f is of positive type,

b) $p_j * f$ is of positive type for all $j = 1, \ldots, m$.

> The geometric content of this theorem is actually pretty easy: Every feasible solution $b \in \mathcal{F}$ which is a $0/1$-vectors corresponds to a subset $B \subseteq \{1, \ldots, n\}$. Now we consider the cone $2^{[n]}$ which is spanned by all feasible solutions, meaning it is spanned by basis vectors χ_B where B corresponds to $b \in \mathcal{F}$. This is a polyhedral cone with $|\mathcal{F}|$ extreme rays $\mathbb{R}_{\geqslant 0}\chi_B$. Now the Theorem says that $f \in L^2(2^{[n]})$ is an element of this polyhedral cone iff conditions a) and b) are fulfilled. The advantage of these two conditions is that they have an equivalent formulation in terms of semidefinite matrices, namely that the moment matrices M_f and $M_{p_j * f}$ are positive semidefinite.

Proof. We have

$$(p_j * f)(A) = \sum_{B \in 2^X} \breve{p_j}(B)\widehat{f}(B)\chi_B(A).$$

Suppose conditions a) and b) are satisfied. By Theorem 5.2.3 the function f can be represented as

$$f(A) = \sum_{B \in 2^X} \widehat{f}(B)\chi_B(A)$$

with $\widehat{f}(B) \geqslant 0$. If strict inequality $\widehat{f}(B) > 0$ holds for some B then for every $j \in \{1, \ldots, m\}$ we have again by Theorem 5.2.3

$$\breve{p_j}(B)\widehat{f}(B) = p_j(b)\widehat{f}(B) \geqslant 0,$$

because $p_j * f$ is of positive type. Thus, $p_j(b) \geqslant 0$ for all j and hence $b \in \mathcal{F}$. The other implication follows with the same arguments. □

Corollary 5.2.6. *The* $0/1$ *polynomial optimization problem* (5.4) *is equivalent to the linear optimization problem*

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} c_i f(\{i\}) \\ & f \in L^2(2^X) \text{ is of positive type}, \\ & f(\varnothing) = 1, \\ & p_j * f \text{ is of positive type for all } j \in [m], \end{aligned}$$

which is equivalent to the semidefinite program

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} c_i M_f(\{i\}, \{i\}) \\ & M_f \in \mathcal{S}^{2^X}_{\geqslant 0}, \\ & M_f(\varnothing, \varnothing) = 1, \\ & M_{p_j * f} \in \mathcal{S}^{2^X}_{\geqslant 0} \quad \text{for all } j \in [m]. \end{aligned}$$

Proof. The fact that these two optimization problems are equivalent follows immediately from the definition of positive type functions.

Since we are maximizing a linear function we can assume that the optimum f_0 of the first problem is attained at the extreme ray of the cone of positive type functions. So it is of the form

$$f_0 = \alpha \chi_B$$

for some $\alpha \geqslant 0$ and by the previous theorem for some b with $p_j(b) \geqslant 0$ for all $j \in [m]$. Furthermore,

$$1 = f_0(\varnothing) = \alpha \chi_B(\varnothing) = \alpha$$

makes sure that one can recover an optimal $0/1$ solution from f_0. The objective value is

$$\sum_{i=1}^{n} c_i f(\{i\}) = \sum_{i=1}^{n} c_i \chi_B(\{i\}).$$

□

Probabilistic interpretation

The two conditions $f(\varnothing) = 1$ and f is of positive type have a simple probabilistic interpretation: The second condition says that f is the Fourier transform of a positive measure whereas the first condition says that the positive measure is in fact a probability measure

$$1 = f(\varnothing) = \sum_{B \in 2^X} \widehat{f}(B) \chi_B(\varnothing) = \sum_{B \in 2^X} \widehat{f}(B).$$

Hence, f determines a probability distribution on the power set 2^X. The set B is picked with probability $\widehat{f}(B)$.

Relaxation

Computationally, the equivalent reformulation in Corollary 5.2.6 is a rather useless statement: We exponentially increased the dimension of the problem and so in a sense we are enumerating all feasible solutions. However, the reformulation can be used to generate systematically valid non-linear inequalities, i.e. projected LMI inequalities, for the $0/1$ polytope $P = \operatorname{conv} \mathcal{F}$.

Instead of working with full moment matrices lying in $\mathcal{S}^{2^X}_{\succcurlyeq 0}$ we only consider in the t-th step/round of Lasserre's hierarchy truncated moment matrices where the rows/columns are indexed by all subsets with cardinality at most $t+1$; Notation: $\binom{X}{\leqslant t+1}$.

For $f \in \mathbb{R}^{\binom{X}{\leqslant 2t+2}}$ define the *truncated moment matrix* $M_f^{t+1} \in \mathcal{S}^{\binom{X}{\leqslant t+1}}$ by

$$M_f^{t+1}(A,B) = f(A \cup B)$$

with $A, B \in \binom{X}{\leqslant t+1}$. Let p be a squarefree polynomial of degree d

$$p = \sum_I p_I m_I$$

with coefficients p_I and monomials m_I where I runs through a subset J of 2^X where every $I \in J$ has cardinality at most d. Then,

$$M_{p*f}^{t+1-\lceil d/2 \rceil}(A,B) = (p*f)(A \cup B) = \sum_I p_I f(I \cup A \cup B)$$

with $A, B \in \binom{X}{\leqslant t+1-\lceil d/2 \rceil}$.

Definition 5.2.7. *Let $p_1, \ldots, p_m \in \mathbb{R}[x_1, \ldots, x_n]$ be polynomials with degrees $d_1, \ldots, d_m$. Furthermore, let*

$$v = \max\left\{\lceil d_j/2 \rceil : j \in [m]\right\}.$$

We define for $t \geqslant v-1$ the t-th step in Lasserre's hierarchy *by*

$$\begin{aligned} K_t = \{(f(\{1\}), \ldots, f(\{n\}))^{\mathsf{T}} \in \mathbb{R}^n \ : & f \in \mathbb{R}^{\binom{X}{\leqslant 2t+2}}, \\ & f(\varnothing) = 1, \\ & M_f^{t+1} \succcurlyeq 0, \\ & M_{p_j*f}^{t+1-\lceil d_j/2 \rceil} \succcurlyeq 0,\ j \in [m]\Big\}. \end{aligned}$$

Optimizing over K_t can be done in polynomial time once $t \geqslant v-1$ is a fixed constant, i.e. independent of the input size. By the previous considerations we have

$$K_{v-1} \supseteq K_v \supseteq K_{v+1} \supseteq \ldots \supseteq K_{n+v-1} = P.$$

5.2.3 Example: Independence number

Let us apply Lasserre's hierarchy to the problem of finding the independence number.

For this we consider the polynomials

$$p_{uv}(x) = 1 - x_u - x_v \in \mathbb{R}[x_v : v \in V] \quad \text{for all } \{u,v\} \in E.$$

The $0/1$ polytope

$$P = \mathrm{ST}(G) = \mathrm{conv}\{x \in \{0,1\}^V : p_{uv} \geqslant 0 \text{ for all } \{u,v\} \in E\}$$

is called the *stable set polytope*[1].

The following theorem says that one can simplify the conditions of K_t here. In fact the semidefinite conditions on the matrices $M^t_{p_{uv}*f}$ can be replaced by the conditions $f(\{u,v\}) = 0$ in the moment matrix M^{t+1}_f.

Theorem 5.2.8. *For $t \geqslant 1$ and $f \in \binom{V}{\leqslant 2t+2}$ the following three statements are equivalent:*

a) $M^{t+1}_f \succeq 0$ *and* $M^{t+1}_{p_{uv}*f} \succeq 0$ *for all* $\{u,v\} \in E$*,*

b) $M^{t+1}_f \succeq 0$ *and* $f(\{u,v\}) = 0$ *for every* $\{u,v\} \in E$*,*

c) $M^{t+1}_f \succeq 0$ *and* $f(U) = 0$ *for every subset* $U \in \binom{V}{\leqslant 2t+2}$ *of the vertex set which is not an independent set.*

Proof. From $M^{t+1}_f \succeq 0$ we can conclude that $f(U) \geqslant 0$ for all $U \in \binom{V}{\leqslant 2t+2}$ because $f(U)$ is a diagonal element of the positive semidefinite matrix M^{t+1}_f.

a) implies b): Consider the $(\{u\},\{u\})$-entry of $M^t_{p_{uv}*f}$. It equals (check this identity!)

$$p_{uv} * f(\{u\}) = -f(\{u,v\})$$

Since it is a diagonal entry, it has to be non-negative. So $f(\{u,v\}) = 0$.

b) implies c): Suppose $u,v \in U$ are connected by an edge.

First case: $|U| \leqslant t+1$, then the $(\{u,v\},\{u,v\})$-entry of M^{t+1}_f equals 0, and so the $(\{u,v\},U)$-entry, too (why?). Hence also $f(U) = M^{t+1}_f(U,U) = 0$.

Second case: $|U| > t+1$. Split U into $U = U_1 \cup U_2$ with $U_1, U_2 \in \binom{V}{\leqslant t+1}$ and assume $\{u,v\} \subseteq U_1$. Then,

$$M^{t+1}_f(U_1,U_1) = 0 \Longrightarrow M^{t+1}_f(U_1,U_2) = 0 \Longrightarrow f(U) = M^{t+1}_f(U,U) = 0.$$

c) implies a): We shall prove that the matrix $M^t_{p_{uv}*f}$ is positive semidefinite. Define $P_0 = \binom{V\setminus\{u,v\}}{\leqslant t}$ and $P_w = \{U \cup \{w\} : U \in P_0\}$ for $w \in \{u,v\}$. Then, the

[1] Sometimes independent sets are called stable sets, but who prefers stability over independence?

principal submatrix of M_f^{t+1} indexed by $P_0 \cup P_u \cup P_v$ is of the form

$$\begin{array}{c} \\ P_0 \\ P_u \\ P_v \end{array} \begin{array}{c} \begin{array}{ccc} P_0 & P_u & P_u \end{array} \\ \begin{pmatrix} C & A & B \\ A & A & 0 \\ B & 0 & B \end{pmatrix} \end{array}.$$

Since this matrix is positive semidefinite by assumption it follows that the matrix $C - A - B$ is positive semidefinite because

$$\begin{pmatrix} -x \\ x \\ x \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} C & A & B \\ A & A & 0 \\ B & 0 & B \end{pmatrix} \begin{pmatrix} -x \\ x \\ x \end{pmatrix} = x^{\mathsf{T}}(C - A - B)x.$$

Now consider the matrix $M^t_{p_{uv} * f}$ where we partition the rows/columns into P_0 and its complement $P_0' = \binom{V}{\leqslant t+1} \backslash P_0$. It has the form

$$\begin{array}{c} \\ P_0 \\ P_0' \end{array} \begin{array}{c} \begin{array}{cc} P_0 & P_0' \end{array} \\ \begin{pmatrix} C - A - B & 0 \\ 0 & 0 \end{pmatrix} \end{array}$$

and the result follows. □

One can show (Exercise 5.1.c)) that Lasserre's hierarchy already converges at step $\alpha(G) - 1$ to $\mathrm{ST}(G)$:

Theorem 5.2.9. *For a graph G with $\alpha(G) \geqslant 2$ we have*

$$\mathrm{ST}(G) = K_{\alpha(G)-1}.$$

5.3 Further reading

Shor [4] was the first who realized that one can use semidefinite programming to relax quadratic optimization problems. Meanwhile there are many different possibilities available to construct semidefinite programming hierarchies for $0/1$ polynomial optimization problems. The one we studied in Section 5.2 is due to Lasserre [2]. Lasserre's hierarchy and other hierarchies are presented and compared by Laurent in [3].

5.4 Exercises

5.1.** a) Show that the following two semidefinite programs give the same optimal value:

$$\begin{aligned} \max \quad & \sum_{v \in V} Y_{vv} \\ & Y \in \mathcal{S}_{\geqslant 0}^{V \cup \{0\}}, \\ & Y_{00} = 1, \\ & Y_{0v} = Y_{vv} \quad \text{for all } v \in V, \\ & Y_{uv} = 0 \quad \text{for all } \{u, v\} \in E, \end{aligned}$$

and

$$\begin{aligned} \vartheta(G) = \max \quad & \langle J, X \rangle \\ & X \in \mathcal{S}_{\geqslant 0}^{V}, \\ & \langle I, X \rangle = 1, \\ & X_{uv} = 0 \quad \text{for all } \{u, v\} \in E. \end{aligned}$$

b) Prove that the theta number of the cycle graph with five vertices C_5 is *exactly* $\sqrt{5}$.

c) Show that $K_{\alpha(G)-1} = \mathrm{ST}(G)$.

d) True or false? For all graphs $G = (V, E)$ the following equality holds

$$\vartheta(G) = \max \left\{ \sum_{v \in V} x_v : x \in K_1 \right\}.$$

5.2. (Computer exercise)

a) Do the following two ellipsoids intersect?

$$E_1 = \left\{ x \in \mathbb{R}^3 : x^{\mathsf{T}} \begin{pmatrix} 1/9 & 1/9 & 1/9 \\ 1/9 & 13/36 & 13/36 \\ 1/9 & 13/36 & 49/36 \end{pmatrix} x \leqslant 1 \right\},$$

$$E_2 = \left\{ x \in \mathbb{R}^3 : x^{\mathsf{T}} \begin{pmatrix} \sqrt{2} & \sqrt{2} & \sqrt{2} \\ \sqrt{2} & 2\sqrt{2} & 2\sqrt{2} \\ \sqrt{2} & 2\sqrt{2} & 2\sqrt{2}+1 \end{pmatrix} x + \begin{pmatrix} -6\sqrt{2} \\ -10\sqrt{2} \\ -10\sqrt{2} \end{pmatrix}^{\mathsf{T}} x \leqslant 1 - 13\sqrt{2} \right\}$$

Solve a semidefinite program to justify your answer.

b) Let $x_1, \ldots, x_N$ be N points in $\mathbb{R}^2$. Find a point $x \in \mathbb{R}^2$ which minimizes the maximum Euclidean distance to these points. Compute the point x for the cities Amsterdam, Athens, Berlin, Copenhagen, Lisbon, Moscow, Prague, Warsaw. For this assume that Europe is part of flatland...

BIBLIOGRAPHY

[1] M. Aigner, *Combinatorial theory*, Springer, 1979.

[2] J.B. Lasserre, *An explicit exact SDP relaxation for nonlinear 0-1 programs*, pp. 293–303 in *Integer Programming and Combinatorial Optimization, 8th International IPCO Conference* (K. Aardal, A.M.H. Gerards, ed.)., Springer, 2001.

[3] M. Laurent, *A comparison of the Sherali-Adams, Lovász-Schrijver and Lasserre relaxations for 0-1 programming*, Math. Oper. Res. **28** (2003), 470–496.

`http://www.cwi.nl/~monique/files/lasserrefinal.ps`

[4] N.Z. Shor, *Quadratic optimization problems*, Soviet J. Comput. Systems Sci. **25** (1987), 1–11.

CHAPTER 6

GRAPH COLORING AND INDEPENDENT SETS

In this chapter we revisit in detail the theta number $\vartheta(G)$, which has already been introduced in earlier chapters. In particular, we present several equivalent formulations for $\vartheta(G)$, we discuss its geometric properties, and present some applications: for bounding the Shannon capacity of a graph, and for computing in polynomial time maximum stable sets and minimum colorings in perfect graphs.

Here are some additional definitions used in this chapter. Let $G = (V, E)$ be a graph. Then, $\overline{E}$ denotes the set of pairs $\{i, j\}$ of distinct nodes that are not adjacent in G. The graph $\overline{G} = (V, \overline{E})$ is called the *complementary graph* of G. G is *self-complementary* if G and $\overline{G}$ are isomorphic graphs. Given a subset $S \subseteq V$, $G[S]$ denotes the *subgraph induced by* S: its node set is S and its edges are all pairs $\{i, j\} \in E$ with $i, j \in S$. The graph C_n is the circuit (or cycle) of length n, with node set $[n]$ and edges the pairs $\{i, i+1\}$ (for $i \in [n]$, indices taken modulo n). For a set $S \subseteq V$, its *characteristic vector* is the vector $\chi^S \in \{0, 1\}^V$, whose i-th entry is 1 if $i \in S$ and 0 otherwise. As before, e denotes the all-ones vector.

6.1 Preliminaries on graphs

6.1.1 Stability and chromatic numbers

A subset $S \subseteq V$ of nodes is said to be *stable* (or *independent*) if no two nodes of S are adjacent in G. Then the *stability number* of G is the parameter $\alpha(G)$ defined as the maximum cardinality of an independent set in G.

A subset $C \subseteq V$ of nodes is called a *clique* if every two distinct nodes in C are adjacent. The maximum cardinality of a clique in G is denoted $\omega(G)$, the *clique number* of G. Clearly,

$$\omega(G) = \alpha(\overline{G}).$$

Computing the stability number of a graph is a hard problem: Given a graph G and an integer k, deciding whether $\alpha(G) \geqslant k$ is an $\mathcal{NP}$-complete problem.

Given an integer $k \geqslant 1$, a *k-coloring* of G is an assignment of numbers (view them as *colors*) from $\{1, \cdots, k\}$ to the nodes in such a way that two adjacent nodes receive distinct colors. In other words, this corresponds to a partition of V into k stable sets: $V = S_1 \cup \cdots \cup S_k$, where S_i is the stable set consisting of all nodes that received the i-th color. The *coloring* (or *chromatic*) *number* is the smallest integer k for which G admits a k-coloring, it is denoted as $\chi(G)$.

Again it is an $\mathcal{NP}$-complete problem to decide whether a graph is k-colorable. In fact, it is $\mathcal{NP}$-complete to decide whether a planar graph is 3-colorable. On the other hand, it is known that every planar graph is 4-colorable – this is the celebrated 4-color theorem. Moreover, observe that one can decide in polynomial time whether a graph is 2-colorable, since one can check in polynomial time whether a graph is bipartite.

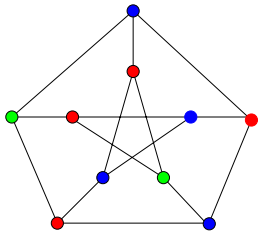


Figure 6.1: The Petersen graph has $\alpha(G) = 4$, $\omega(G) = 2$ and $\chi(G) = 3$

Clearly, any two nodes in a clique of G must receive distinct colors. Therefore, for any graph, the following inequality holds:

$$\omega(G) \leqslant \chi(G). \tag{6.1}$$

This inequality is strict, for example, when G is an odd circuit, i.e., a circuit of odd length at least 5, or its complement. Indeed, for an odd circuit C_{2n+1} ($n \geqslant 2$), $\omega(C_{2n+1}) = 2$ while $\chi(C_{2n+1}) = 3$. Moreover, for the complement $G = \overline{C_{2n+1}}$, $\omega(G) = n$ while $\chi(G) = n + 1$. For an illustration see the cycle of length 7 and its complement in Figure 6.2.

6.1.2 Perfect graphs

It is intriguing to understand for which graphs equality $\omega(G) = \chi(G)$ holds. Note that any graph G with $\omega(G) < \chi(G)$ can be embedded in a larger graph $\hat{G}$

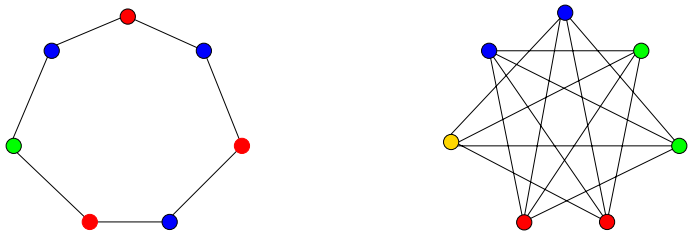


Figure 6.2: For C_7 and its complement $\overline{C_7}$: $\omega(C_7) = 2$, $\chi(C_7) = 3$, $\omega(\overline{C_7}) = \alpha(C_7) = 3$, $\chi(\overline{C_7}) = 4$

with $\omega(\hat{G}) = \chi(\hat{G})$, simply by adding to G a clique of size $\chi(G)$. This justifies the following definition, introduced by C. Berge in the early sixties, which makes the problem well posed.

Definition 6.1.1. *A graph G is said to be* perfect *if equality*

$$\omega(H) = \chi(H)$$

holds for all induced subgraphs H of G (including $H = G$).

For instance, bipartite graphs are perfect. It follows from the definition and the above observation about odd circuits that if G is a perfect graph then it does not contain an odd circuit of length at least 5 or its complement as an induced subgraph. Berge already conjectured that *all* perfect graphs arise in this way. Resolving this conjecture has haunted generations of graph theorists. It was finally settled in 2004 by Chudnovsky, Robertson, Seymour and Thomas who proved the following result, known as the *strong perfect graph theorem*:

Theorem 6.1.2. (The strong perfect graph theorem) *A graph G is perfect if and only if it does not contain an odd circuit of length at least 5 or its complement as an induced subgraph.*

This implies the following structural result about perfect graphs, known as the *perfect graph theorem*, already proved by Lovász in 1972.

Theorem 6.1.3. (The perfect graph theorem) *If G is a perfect graph, then its complement $\overline{G}$ too is a perfect graph.*

We will mention below some other, more geometric, characterizations of perfect graphs.

6.2 Linear programming bounds

Let $\mathrm{ST}(G)$ denote the polytope in $\mathbb{R}^V$ defined as the convex hull of the characteristic vectors of the stable sets of G:

$$\mathrm{ST}(G) = \mathrm{conv}\{\chi^S : S \subseteq V,\ S \text{ is a stable set in } G\},$$

called the *stable set polytope* of G. Hence, computing $\alpha(G)$ is linear optimization over the stable set polytope:

$$\alpha(G) = \max\{e^{\mathsf{T}}x : x \in \mathsf{ST}(G)\}.$$

We have now defined the stable set polytope by listing explicitly its extreme points. Alternatively, it can also be represented by its hyperplanes representation, i.e., in the form

$$\mathsf{ST}(G) = \{x \in \mathbb{R}^V : Ax \leqslant b\}$$

for some matrix A and some vector b. As computing the stability number is a hard problem one cannot hope to find the full linear inequality description of the stable set polytope (i.e., the explicit A and b). However some partial information is known: several classes of valid inequalities for the stable set polytope are known. For instance, if C is a clique of G, then the *clique inequality*

$$x(C) = \sum_{i \in C} x_i \leqslant 1 \tag{6.2}$$

is valid for $\mathsf{ST}(G)$: any stable set can contain at most one vertex from the clique C. The clique inequalities define the polytope

$$\mathsf{QST}(G) = \left\{x \in \mathbb{R}^V : x \geqslant 0,\ x(C) \leqslant 1\ \forall C \text{ clique of } G\right\} \tag{6.3}$$

and maximizing the linear function $e^{\mathsf{T}}x$ over it gives the parameter

$$\alpha^*(G) = \max\{e^{\mathsf{T}}x : x \in \mathsf{QST}(G)\}, \tag{6.4}$$

known as the *fractional stability number* of G. Clearly, $\mathsf{QST}(G)$ is a relaxation of the stable set polytope:

$$\mathsf{ST}(G) \subseteq \mathsf{QST}(G). \tag{6.5}$$

The parameter $\alpha^*(G)$ is nested between $\alpha(G)$ and $\chi(\overline{G})$, and it can also be interpreted in terms of *fractional colorings* of $\overline{G}$.

Lemma 6.2.1. *For any graph G, we have*

$$\alpha(G) \leqslant \alpha^*(G) \leqslant \chi(\overline{G}), \tag{6.6}$$

and $\alpha^*(G)$ is equal to the optimal value of the linear program

$$\min\left\{\sum_{C \text{ clique of } G} y_C : \sum_{C \text{ clique of } G} y_C \chi^C = e,\ y_C \geqslant 0\ \forall C \text{ clique of } G\right\}. \tag{6.7}$$

Proof. The left most inequality in (6.6) follows from the inclusion (6.5) and the right most one from the definitions: If $x \in \mathsf{QST}(G)$ and $V = C_1 \cup \cdots \cup C_k$ is a partition into k cliques, then

$$x^{\mathsf{T}}e = x^{\mathsf{T}}\left(\sum_{i=1}^{k} \chi^{C_i}\right) = \sum_{i=1}^{k} x(C_i) \leqslant \sum_{i=1}^{k} 1 = k.$$

We now show that the optimal value of (6.7) is equal to $\alpha^*(G)$. For this, first note that in the linear program (6.4) the condition $x \geqslant 0$ can be removed without changing the optimal value; that is,

$$\alpha^*(G) = \max\{e^T x : x(C) \leqslant 1 \ \forall C \text{ clique of } G\}.$$

Now, applying linear programming duality to this linear program, we obtain the linear program (6.7). □

For instance, for an odd circuit C_{2n+1} ($n \geqslant 2$), $\alpha^*(C_{2n+1}) = \frac{2n+1}{2}$ (check it) lies strictly between $\alpha(C_{2n+1}) = n$ and $\chi(\overline{C_{2n+1}}) = n+1$.

Assume that G is a perfect graph. Then equality holds throughout in relation (6.6). Hence, when $w \in \mathbb{R}^V$ is the all-ones vector, maximizing the objective function $w^T x$ over the stable set polytope $\mathrm{STAB}(G)$ or over its linear relaxation $\mathrm{QSTAB}(G)$ gives the same optimal values. The same holds if w is $0/1$ valued since this amounts to replacing G by its subgraph induced by the set of nodes with $w_v = 1$, which is again perfect. One can show that the same holds for *any integer* vector $w \in \mathbb{Z}^V_{\geqslant 0}$, which implies that the two polytopes $\mathrm{STAB}(G)$ and $\mathrm{QSTAB}(G)$ coincide. Moreover, this property characterizes perfect graphs.

Theorem 6.2.2. *A graph G is perfect if and only if* $\mathrm{STAB}(G) = \mathrm{QSTAB}(G)$.

Although an explicit linear inequality description is known for the stable set polytope of a perfect graph (given by the clique inequalities), it is not clear how to use this information in order to give an efficient algorithm for optimizing over the stable set polytope. As we see later in Section 6.5.2 there is yet another description of $\mathrm{STAB}(G)$ – in terms of semidefinite programming, using the theta body $\mathrm{TH}(G)$ – that allows to give an efficient algorithm.

6.3 Semidefinite programming bounds

6.3.1 The theta number

Definition 6.3.1. *Given a graph $G = (V, E)$, consider the following semidefinite program*

$$\max_{X \in \mathcal{S}^n} \{\langle J, X\rangle : \mathrm{Tr}(X) = 1,\ X_{ij} = 0\ \forall \{i,j\} \in E,\ X \succeq 0\}. \tag{6.8}$$

Its optimal value is denoted as $\vartheta(G)$, and called the theta number *of G.*

This parameter was introduced by Lovász [3]. He proved the following simple, but crucial result – called the Sandwich Theorem by Knuth [2] – which shows that $\vartheta(G)$ provides a bound for both the stability number of G and the chromatic number of the complementary graph $\overline{G}$.

Theorem 6.3.2. (Lovász' sandwich theorem) *For any graph G, we have that*

$$\alpha(G) \leqslant \vartheta(G) \leqslant \chi(\overline{G}).$$

Proof. Given a stable set S of cardinality $|S| = \alpha(G)$, define the matrix

$$X = \frac{1}{|S|}\chi^S(\chi^S)^\mathsf{T} \in \mathcal{S}^n.$$

Then X is feasible for (6.8) with objective value $\langle J, X\rangle = |S|$ (check it). This shows the inequality $\alpha(G) \leqslant \vartheta(G)$.

Now, consider a matrix X feasible for the program (6.8) and a partition of V into k cliques: $V = C_1 \cup \cdots \cup C_k$. Our goal is now to show that $\langle J, X\rangle \leqslant k$, which will imply $\vartheta(G) \leqslant \chi(\overline{G})$. For this, using the relation $e = \sum_{i=1}^k \chi^{C_i}$, observe that

$$Y := \sum_{i=1}^k \left(k\chi^{C_i} - e\right)\left(k\chi^{C_i} - e\right)^\mathsf{T} = k^2 \sum_{i=1}^k \chi^{C_i}(\chi^{C_i})^\mathsf{T} - kJ.$$

Moreover,

$$\left\langle X, \sum_{i=1}^k \chi^{C_i}(\chi^{C_i})^\mathsf{T} \right\rangle = \mathrm{Tr}(X).$$

Indeed the matrix $\sum_i \chi^{C_i}(\chi^{C_i})^\mathsf{T}$ has all its diagonal entries equal to 1 and it has zero off-diagonal entries outside the edge set of G, while X has zero off-diagonal entries on the edge set of G. As $X, Y \succeq 0$, we obtain

$$0 \leqslant \langle X, Y\rangle = k^2\mathrm{Tr}(X) - k\langle J, X\rangle$$

and thus $\langle J, X\rangle \leqslant k\ \mathrm{Tr}(X) = k$. □

An alternative argument for the inequality $\vartheta(G) \leqslant \chi(\overline{G})$, showing an even more transparent link to coverings by cliques, will be given in the paragraph after the proof of Lemma 6.4.2.

6.3.2 Computing maximum stable sets in perfect graphs

Assume that G is a graph satisfying $\alpha(G) = \chi(\overline{G})$. Then, as a direct application of Theorem 6.3.2, $\alpha(G) = \chi(\overline{G}) = \vartheta(G)$ can be computed by solving the semidefinite program (6.8), it suffices to solve this semidefinite program with precision $\epsilon < 1/2$ as one can then find $\alpha(G)$ by rounding the optimal value to the nearest integer. In particular, combining with the perfect graph theorem (Theorem 6.1.3):

Theorem 6.3.3. *If G is a perfect graph then $\alpha(G) = \chi(\overline{G}) = \vartheta(G)$ and $\omega(G) = \chi(G) = \vartheta(\overline{G})$.*

Hence one can compute the stability number and the chromatic number in polynomial time for perfect graphs. Moreover, one can also find a maximum stable set and a minimum coloring in polynomial time for perfect graphs. We now indicate how to construct a maximum stable set – we deal with minimum graph colorings in the next section.

Let $G = (V, E)$ be a perfect graph. Order the nodes of G as $v_1, \cdots, v_n$. Then we construct a sequence of induced subgraphs $G_0, G_1, \cdots, G_n$ of G. Hence each G_i is perfect, also after removing a node, so that we can compute in polynomial time the stability number of such graphs. The construction goes as follows: Set $G_0 = G$. For each $i = 1, \cdots, n$ do the following:

1. Compute $\alpha(G_{i-1}\backslash v_i)$.
2. If $\alpha(G_{i-1}\backslash v_i) = \alpha(G)$, then set $G_i = G_{i-1}\backslash v_i$.
3. Otherwise, set $G_i = G_{i-1}$.

By construction, $\alpha(G_i) = \alpha(G)$ for all i. In particular, $\alpha(G_n) = \alpha(G)$. Moreover, the node set of the final graph G_n is a stable set and, therefore, it is a maximum stable set of G. Indeed, if the node set of G_n is not stable then it contains a node v_i for which $\alpha(G_n\backslash v_i) = \alpha(G_n)$. But then, as G_n is an induced subgraph of G_{i-1}, one would have that $\alpha(G_n\backslash v_i) \leqslant \alpha(G_{i-1}\backslash v_i)$ and thus $\alpha(G_{i-1}\backslash v_i) = \alpha(G)$, so that node v_i would have been removed at Step 2.

Hence, the above algorithm permits to construct a maximum stable set in a perfect graph G in polynomial time – namely by solving $n + 1$ semidefinite programs for computing $\alpha(G)$ and $\alpha(G_{i-1}\backslash v_i)$ for $i = 1, \cdots, n$.

More generally, given integer weights $w \in \mathbb{Z}^V_{\geqslant 0}$ on the nodes, one can compute in polynomial time a stable set S of maximum weight $w(S)$. For this, one can apply the algorithm just described for computing a maximum cardinality stable set in the new graph G' defined in the following way: Replace each node $i \in V$ by a set W_i of w_i nodes pairwise non-adjacent, and make two nodes $x \in W_i$ and $y \in W_j$ adjacent if i and j are adjacent in G. One can verify that G' is perfect and that $\alpha(G')$ is the maximum weight $w(S)$ of a stable set S in G.

6.3.3 Minimum colorings of perfect graphs

We now describe an algorithm for computing a minimum coloring of a perfect graph G in polynomial time. This will be reduced to several computations of the theta number which we will use for computing the clique number of some induced subgraphs of G.

Let $G = (V, E)$ be a perfect graph. Call a clique of G *maximum* if it has maximum cardinality $\omega(G)$.

The crucial observation is that it suffices to find a stable set S in G which meets all maximum cliques. Indeed, if such S is found then one can recursively color $G\backslash S$ with $\omega(G) - 1$ colors (in polynomial time), and thus G with $\omega(G)$ colors. (Clearly, such a stable set S exists: any color class S in a $\omega(G)$-coloring must meet all maximum cliques as $\omega(G\backslash S) = \chi(G\backslash S) = \omega(G) - 1$.)

The algorithm goes as follows: For $t \geqslant 1$, grow a list $\mathcal{L}$ of t maximum cliques $C_1, \cdots, C_t$. Suppose $C_1, \cdots, C_t$ have been found. Then do the following:

1. We find a stable set S meeting each of the cliques $C_1, \cdots, C_t$ (see below).

2. Compute $\omega(G\backslash S)$.

3. If $\omega(G\backslash S) < \omega(G)$ then S meets all maximum cliques and we are done.

4. Otherwise, compute a maximum clique C_{t+1} in $G\backslash S$, which is thus a new maximum clique of G, and we add it to the list $\mathcal{L}$.

The first step can be done as follows: Set $w = \sum_{i=1}^{t} \chi^{C_i} \in \mathbb{Z}^V_{\geqslant 0}$ and compute a stable set S having maximum possible weight $w(S)$, then $w(S) = t$ and S meets $C_1, \cdots, C_t$.

The above algorithm has polynomial running time, since the number of iterations is bounded by $|V|$. To see this, define the affine space $L_t \subseteq \mathbb{R}^V$ defined by the equations $x(C_1) = 1, \cdots, x(C_t) = 1$ corresponding to the cliques in the current list $\mathcal{L}$. Then, L_t contains strictly L_{t+1}, since $\chi^S \in L_t \backslash L_{t+1}$ for the set S constructed in the first step, and thus the dimension decreases at least by 1 at each iteration.

6.4 Other formulations of the theta number

6.4.1 Dual formulation

We now give several equivalent formulations for the theta number obtained by applying semidefinite programming duality and some further elementary manipulations.

Lemma 6.4.1. *The theta number can be expressed by any of the following programs:*

$$\vartheta(G) = \min_{t\in\mathbb{R}, A\in\mathcal{S}^n} \{t : tI + A - J \succeq 0,\ A_{ij} = 0\ (i = j \text{ or } \{i,j\} \in \overline{E})\}, \tag{6.9}$$

$$\vartheta(G) = \min_{t\in\mathbb{R}, B\in\mathcal{S}^n} \left\{t : tI - B \succeq 0,\ B_{ij} = 1\ (i = j \text{ or } \{i,j\} \in \overline{E})\right\}, \tag{6.10}$$

$$\vartheta(G) = \min_{t\in\mathbb{R}, C\in\mathcal{S}^n} \{t : C - J \succeq 0,\ C_{ii} = t\ (i \in V),\ C_{ij} = 0\ (\{i,j\} \in \overline{E})\}, \tag{6.11}$$

$$\vartheta(G) = \min_{B\in\mathcal{S}^n} \left\{\lambda_{\max}(B) : B_{ij} = 1\ (i = j \text{ or } \{i,j\} \in \overline{E})\right\}. \tag{6.12}$$

Proof. First we build the dual of the semidefinite program (6.8), which reads:

$$\min_{t\in\mathbb{R}, y\in\mathbb{R}^E} \left\{ t : tI + \sum_{\{i,j\}\in E} y_{ij} E_{ij} - J \succeq 0 \right\}. \tag{6.13}$$

As both programs (6.8) and (6.13) are strictly feasible, there is no duality gap: the optimal value of (6.13) is equal to $\vartheta(G)$, and the optimal values are attained in both programs – here we have applied the duality theorem (Theorem 3.4.1).

Setting $A = \sum_{\{i,j\}\in E} y_{ij} E_{ij}$, $B = J - A$ and $C = tI + A$ in (6.13), it follows that the program (6.13) is equivalent to (6.9), (6.10) and (6.11). Finally the formulation (6.12) follows directly from (6.10) after recalling that $\lambda_{\max}(B)$ is the smallest scalar t for which $tI - B \succeq 0$. □

6.4.2 Two more (lifted) formulations

We give here two more formulations for the theta number. They rely on semidefinite programs involving symmetric matrices of order $1+n$ which we will index by the set $\{0\} \cup V$, where 0 is an additional index that does not belong to V.

Lemma 6.4.2. *The theta number $\vartheta(G)$ is equal to the optimal value of the following semidefinite program:*

$$\min_{Z \in \mathcal{S}^{n+1}} \{Z_{00} : Z \succeq 0,\ Z_{0i} = Z_{ii} = 1\ (i \in V),\ Z_{ij} = 0\ (\{i,j\} \in \overline{E})\}. \tag{6.14}$$

Proof. We show that the two semidefinite programs in (6.9) and (6.14) are equivalent. For this, observe that

$$tI + A - J \succeq 0 \Longleftrightarrow Z := \begin{pmatrix} t & e^{\mathsf{T}} \\ e & I + \frac{1}{t}A \end{pmatrix} \succeq 0,$$

which follows by taking the Schur complement of the upper left corner t in the block matrix Z. Hence, if (t, A) is feasible for (6.9), then Z is feasible for (6.14) with same objective value: $Z_{00} = t$. The construction can be reversed: if Z is feasible for (6.14), then one can construct (t, A) feasible for (6.9) with $t = Z_{00}$. Hence both programs are equivalent. □

From the formulation (6.14), the link to the chromatic number is even more transparent. Indeed, let $k = \chi(\overline{G})$ and consider a partition $V = C_1 \cup \cdots \cup C_k$ of the node set into k cliques. For each clique C_i define the extended characteristic vector $z_i = (1\ \chi^{C_i}) \in \mathbb{R}^{1+n}$ obtained by appending an entry 1 to the characteristic vector of C_i. Define the matrix $Z = \sum_{i=1}^{k} z_i z_i^{\mathsf{T}} \in \mathcal{S}^{1+n}$. Then one can easily check that the matrix Z is feasible for the program (6.14) with objective value $Z_{00} = k$. Hence this shows again the inequality $\vartheta(G) \leqslant \chi(\overline{G})$.

Applying duality to the semidefinite program (6.14), we obtain[1] the following formulation for $\vartheta(G)$.

Lemma 6.4.3. *The theta number $\vartheta(G)$ is equal to the optimal value of the following semidefinite program:*

$$\max_{Y \in \mathcal{S}^{n+1}} \left\{ \sum_{i \in V} Y_{ii} : Y \succeq 0,\ Y_{00} = 1,\ Y_{0i} = Y_{ii}\ (i \in V),\ Y_{ij} = 0\ (\{i,j\} \in E) \right\}. \tag{6.15}$$

Proof. One can verify that the dual program of (6.14) reads

$$\max \left\{ -\sum_{i \in V} Y_{ii} + 2Y_{0i} : Y \succeq 0,\ Y_{00} = 1,\ Y_{ij} = 0\ (\{i,j\} \in E) \right\} \tag{6.16}$$

[1] Of course there is more than one road leading to Rome: one can also show directly the equivalence of the two programs (6.8) and (6.15).

(check it). As (6.14) is strictly feasible (check it) there is no duality gap, the optimal value of (6.16) is attained and it is equal to $\vartheta(G)$. Note that the program (6.15) amounts to adding the constraints $Y_{ii} = Y_{0i}$ ($i \in V$) to the program (6.16). In order to show that both programs (6.15) and (6.16) are equivalent, it suffices to show that (6.16) admits an optimal solution satisfying these additional constraints.

For this pick an optimal solution Y to (6.16). In a first step, we show that $Y_{0i} + Y_{ii} = 0$ for all $i \in V$. Indeed, assume that $Y_{0i} + Y_{ii} \neq 0$ for some $i \in V$. Then, $Y_{ii} \neq 0$. Let us multiply the i-th column and the i-th row of the matrix Y by the scalar $-\frac{X_{0i}}{X_{ii}}$. In this way we obtain a new matrix Y' which is still feasible for (6.16), but with a better objective value: Indeed, $Y'_{ii} = Y_{ii}\left(-\frac{Y_{0i}}{Y_{ii}}\right)^2 = \frac{Y_{0i}^2}{Y_{ii}}$ and $Y'_{0i} = -\frac{Y_{0i}^2}{Y_{ii}}$, so that the i-th term in the new objective value is

$$-(Y'_{ii} + 2Y'_{0i}) = \frac{Y_{0i}^2}{Y_{ii}} > -(Y_{ii} + 2Y_{0i}).$$

Hence, $Y_{0i} = -Y_{ii}$ for all $i \in V$. Now, we can change the signs on the first row and column of Y (indexed by the index 0). In this way we obtain a new optimal solution of (6.16) which now satisfies the conditions: $Y_{ii} = Y_{0i}$ for $i \in V$. □

As explained in Chapter 5 one can define a hierarchy of semidefinite bounds for $\alpha(G)$, strengthening the theta number $\vartheta(G)$. While $\vartheta(G)$ is defined using matrices indexed by $\{0\} \cup V$, these stronger bounds are obtained by considering matrices indexed by larger index sets, thus lifting the problem in even larger dimensions – see Section 5.2.3 for details.

6.5 Geometric properties of the theta number

6.5.1 Orthonormal representations

Definition 6.5.1. *An* orthonormal representation *of G, abbreviated as* ONR*, consists of a set of unit vectors $\{u_1, \ldots, u_n\} \subseteq \mathbb{R}^d$ (for some $d \geqslant 1$) satisfying*

$$u_i^\mathsf{T} u_j = 0 \quad \forall \{i, j\} \in \overline{E}.$$

If S is a stable set in G and the u_i's form an ONR of G of dimension d, then the vectors u_i labeling the nodes of S are pairwise orthogonal, which implies that $d \geqslant \alpha(G)$. It turns out that the stronger lower bound: $d \geqslant \vartheta(G)$ holds.

Lemma 6.5.2. *The minimum dimension d of an orthonormal representation of a graph G satisfies: $\vartheta(G) \leqslant d$.*

Proof. Let $u_1, \cdots, u_n \in \mathbb{R}^d$ be an ONR of G. Define the matrices $U_0 = I_d$, $U_i = u_i u_i^\mathsf{T} \in \mathcal{S}^d$ for $i \in [n]$. Now we define a symmetric matrix $Z \in \mathcal{S}^{n+1}$ by setting $Z_{ij} = \langle U_i, U_j \rangle$ for $i, j \in \{0\} \cup [n]$. One can verify that Z is feasible for the program (6.14) defining $\vartheta(G)$ (check it) with $Z_{00} = d$. This gives $\vartheta(G) \leqslant d$. □

6.5.2 The theta body $\mathrm{TH}(G)$

It is convenient to introduce the following set of matrices $X \in \mathcal{S}^{n+1}$, where columns and rows are indexed by the set $\{0\} \cup V$:

$$\mathcal{M}_G = \{Y \in \mathcal{S}^{n+1} : Y_{00} = 1,\ Y_{0i} = Y_{ii}\ (i \in V),\ Y_{ij} = 0\ (\{i,j\} \in E),\ Y \succeq 0\}, \tag{6.17}$$

which is thus the feasible region of the semidefinite program (6.15). Now let $\mathrm{TH}(G)$ denote the convex set obtained by projecting the set $\mathcal{M}_G$ onto the subspace $\mathbb{R}^V$ of the diagonal entries:

$$\mathrm{TH}(G) = \{x \in \mathbb{R}^V : \exists Y \in \mathcal{M}_G \text{ such that } x_i = Y_{ii}\ \forall i \in V\}, \tag{6.18}$$

called the *theta body* of G. It turns out that $\mathrm{TH}(G)$ is nested between $\mathrm{ST}(G)$ and $\mathrm{QST}(G)$.

Lemma 6.5.3. *For any graph G, we have that $\mathrm{ST}(G) \subseteq \mathrm{TH}(G) \subseteq \mathrm{QST}(G)$.*

Proof. The inclusion $\mathrm{ST}(G) \subseteq \mathrm{TH}(G)$ follows from the fact that the characteristic vector of any stable set lies in $\mathrm{TH}(G)$ (check it). We now show the inclusion $\mathrm{TH}(G) \subseteq \mathrm{QST}(G)$. For this pick a vector $x \in \mathrm{TH}(G)$ and a clique C of G; we show that $x(C) \leqslant 1$. Say $x_i = Y_{ii}$ for all $i \in V$, where $Y \in \mathcal{M}_G$. Consider the principal submatrix Y_C of Y indexed by $\{0\} \cup C$, which is of the form

$$Y_C = \begin{pmatrix} 1 & x_C^{\mathsf{T}} \\ x_C & \mathrm{Diag}(x_C) \end{pmatrix},$$

where we set $x_C = (x_i)_{i \in C}$. Now, $Y_C \succeq 0$ implies that $\mathrm{Diag}(x_C) - x_C x_C^{\mathsf{T}} \succeq 0$ (taking a Schur complement). This in turn implies: $e^{\mathsf{T}}(\mathrm{Diag}(x_C) - x_C x_C^{\mathsf{T}})e \geqslant 0$, which can be rewritten as $x(C) - (x(C))^2 \geqslant 0$, giving $x(C) \leqslant 1$. □

In view of Lemma 6.4.3, maximizing the all-ones objective function over $\mathrm{TH}(G)$ gives the theta number:

$$\vartheta(G) = \max_{x \in \mathbb{R}^V} \{e^{\mathsf{T}} x : x \in \mathrm{TH}(G)\}.$$

As maximizing $e^{\mathsf{T}}x$ over $\mathrm{QST}(G)$ gives the LP bound $\alpha^*(G)$, Lemma 6.5.3 implies directly that the SDP bound $\vartheta(G)$ dominates the LP bound $\alpha^*(G)$:

Corollary 6.5.4. *For any graph G, we have that $\alpha(G) \leqslant \vartheta(G) \leqslant \alpha^*(G)$.*

Combining the inclusion from Lemma 6.5.3 with Theorem 6.2.2, we deduce that $\mathrm{TH}(G) = \mathrm{ST}(G) = \mathrm{QST}(G)$ for perfect graphs. It turns out that these equalities characterize perfect graphs.

Theorem 6.5.5. *For any graph G the following assertions are equivalent.*

1. *G is perfect.*
2. *$\mathrm{TH}(G) = \mathrm{ST}(G)$*
3. *$\mathrm{TH}(G) = \mathrm{QST}(G)$.*
4. *$\mathrm{TH}(G)$ is a polytope.*

6.5.3 More on the theta body

There is a beautiful relationship between the theta bodies of a graph G and of its complementary graph $\overline{G}$:

Theorem 6.5.6. *For any graph G,*

$$\mathrm{TH}(G) = \{x \in \mathbb{R}^V_{\geqslant 0} : x^T z \leqslant 1 \ \forall z \in \mathrm{TH}(\overline{G})\}.$$

In other words, we know an explicit linear inequality description of $\mathrm{TH}(G)$; moreover, the normal vectors to the supporting hyperplanes of $\mathrm{TH}(G)$ are precisely the elements of $\mathrm{TH}(\overline{G})$. One inclusion is easy:

Lemma 6.5.7. *If $x \in \mathrm{TH}(G)$ and $z \in \mathrm{TH}(\overline{G})$ then $x^{\mathsf{T}} z \leqslant 1$.*

Proof. Let $Y \in \mathcal{M}_G$ and $Z \in \mathcal{M}_{\overline{G}}$ such that $x = (Y_{ii})$ and $z = (Z_{ii})$. Let Z' be obtained from Z by changing signs in its first row and column (indexed by 0). Then $\langle Y, Z' \rangle \geqslant 0$ as $Y, Z' \succeq 0$. Moreover, $\langle Y, Z' \rangle = 1 - x^{\mathsf{T}} z$ (check it), thus giving $x^{\mathsf{T}} z \leqslant 1$. □

Next we observe how the elements of $\mathrm{TH}(G)$ can be expressed in terms of orthonormal representations of $\overline{G}$.

Lemma 6.5.8. *For $x \in \mathbb{R}^V_{\geqslant 0}$, $x \in \mathrm{TH}(G)$ if and only if there exist an orthonormal representation $v_1, \ldots, v_n$ of $\overline{G}$ and a unit vector d such that $x = ((d^{\mathsf{T}} v_i)^2)_{i \in V}$.*

Proof. Let d, v_i be unit vectors where the v_i's form an ONR of $\overline{G}$; we show that $x = ((d^{\mathsf{T}} v_i)^2) \in \mathrm{TH}(G)$. For this, let $Y \in \mathcal{S}^{n+1}$ denote the the Gram matrix of the vectors d and $(v_i^{\mathsf{T}} d) v_i$ for $i \in V$, so that $x = (Y_{ii})$. One can verify that $Y \in \mathcal{M}_G$, which implies $x \in \mathrm{TH}(G)$.

For the reverse inclusion, pick $Y \in \mathcal{M}_G$ and a Gram representation w_0, w_i ($i \in V$) of Y. Set $d = w_0$ and $v_i = w_i / \|w_i\|$ for $i \in V$. Then the conditions expressing membership of Y in $\mathcal{M}_G$ imply that the v_i's form an ONR of $\overline{G}$, $\|d\| = 1$, and $Y_{ii} = (d^{\mathsf{T}} v_i)^2$ for all $i \in V$. □

To conclude the proof of Theorem 6.5.6 we use the following result, which characterizes which partially specified matrices can be completed to a positive semidefinite matrix – you will prove it in Exercise 6.1.

Proposition 6.5.9. *Let $H = (W, F)$ be a graph and let a_{ij} ($i = j \in W$ or $\{i, j\} \in F$) be given scalars, corresponding to a vector $a \in \mathbb{R}^{W \cup F}$. Define the convex set*

$$K_a = \{Y \in \mathcal{S}^W : Y \succeq 0, \ Y_{ij} = a_{ij} \ \forall i = j \in W \text{ and } \{i, j\} \in F\} \tag{6.19}$$

(consisting of all possible positive semidefinite completions of a) and the cone

$$\mathcal{C}_H = \{Z \in \mathcal{S}^W : Z \succeq 0, \ Z_{ij} = 0 \ \forall \{i, j\} \in \overline{F}\} \tag{6.20}$$

(consisting of all positive semidefinite matrices supported by the graph H). Then, $K_a \neq \varnothing$ if and only if

$$\sum_{i \in W} a_{ii} Z_{ii} + 2 \sum_{\{i,j\} \in F} a_{ij} Z_{ij} \geqslant 0 \ \ \forall Z \in \mathcal{C}_H. \tag{6.21}$$

Proof. (*of Theorem 6.5.6*). Let $x \in \mathbb{R}^V_{\geqslant 0}$ such that $x^\mathsf{T} z \leqslant 1$ for all $z \in \mathrm{TH}(\overline{G})$; we show that $x \in \mathrm{TH}(G)$. For this we need to find a matrix $Y \in \mathcal{M}_G$ such that $x = (Y_{ii})_{i \in V}$. In other words, the entries of Y are specified already at the following positions: $Y_{00} = 1$, $Y_{0i} = Y_{ii} = x_i$ for $i \in V$, and $Y_{\{i,j\}} = 0$ for all $\{i,j\} \in E$, and we need to show that the remaining entries (at the positions of non-edges of G) can be chosen in such a way that $Y \succeq 0$.

To show this we apply Proposition 6.5.9, where the graph H is G with an additional node 0 adjacent to all $i \in V$. Hence it suffices now to show that $\langle Y, Z \rangle \geqslant 0$ for all $Z \in \mathcal{S}^{\{0\} \cup V}_{\succeq 0}$ with $Z_{ij} = 0$ if $\{i,j\} \in \overline{E}$. Pick such Z, with Gram representation $w_0, w_1, \cdots, w_n$. Then $w_i^\mathsf{T} w_j = 0$ if $\{i,j\} \in \overline{E}$. We can assume without loss of generality that all w_i are non-zero (use continuity if some w_i is zero) and up to scaling that w_0 is a unit vector. Then the vectors $w_i/\|w_i\|$ ($i \in V$) form an ONR of G. By Lemma 6.5.8 (applied to $\overline{G}$), the vector $z \in \mathbb{R}^V$ with $z_i = (w_0^\mathsf{T} w_i)^2/\|w_i\|^2$ belongs to $\mathrm{TH}(\overline{G})$ and thus $x^\mathsf{T} z \leqslant 1$ by assumption. Therefore, $\langle Y, Z \rangle$ is equal to

$$1 + 2\sum_{i \in V} x_i w_0^\mathsf{T} w_i + \sum_{i \in V} x_i \|w_i\|^2 \geqslant \sum_{i \in V} x_i \left(\frac{(w_0^\mathsf{T} w_i)^2}{\|w_i\|^2} + 2 w_0^\mathsf{T} w_i + \|w_i\|^2 \right)$$

$$= \sum_{i \in V} x_i \left(\frac{w_0^\mathsf{T} w_i}{\|w_i\|} + \|w_i\| \right)^2 \geqslant 0.$$

□

6.6 The theta number for vertex-transitive graphs

First we mention an inequality relating the theta numbers of a graph and its complement.

Proposition 6.6.1. *For any graph $G = (V, E)$, we have that $\vartheta(G)\vartheta(\overline{G}) \geqslant |V|$.*

Proof. Using the formulation of the theta number from (6.11), we obtain matrices $C, C' \in \mathcal{S}^n$ such that $C - J, C' - J \succeq 0$, $C_{ii} = \vartheta(G)$, $C'_{ii} = \vartheta(\overline{G})$ for $i \in V$, $C_{ij} = 0$ for $\{i,j\} \in \overline{E}$ and $C'_{ij} = 0$ for $\{i,j\} \in E$. Then, we have that $\langle C, J \rangle, \langle C', J \rangle \geqslant \langle J, J \rangle = n^2$, and $\langle C, C' \rangle = n\vartheta(G)\vartheta(\overline{G})$. Combining these facts yields the desired inequality. □

We now show that equality $\vartheta(G)\vartheta(\overline{G}) = |V|$ holds for certain symmetric graphs, namely for vertex-transitive graphs. In order to show this, one exploits in a crucial manner the symmetry of G, which permits to show that the semidefinite program defining the theta number has an optimal solution with a special (symmetric) structure. We need to introduce some definitions.

Let $G = (V, E)$ be a graph. A permutation σ of the node set V is called an *automorphism* of G if it preserves edges, i.e., $\{i,j\} \in E$ implies $\{\sigma(i), \sigma(j)\} \in E$. Then the set $\mathrm{Aut}(G)$ of automorphisms of G is a group. The graph G is said to

be *vertex-transitive* if for any two nodes $i, j \in V$ there exists an automorphism $\sigma \in \mathrm{Aut}(G)$ mapping i to j: $\sigma(i) = j$.

The group of permutations of V acts on symmetric matrices X indexed by V. Namely, if σ is a permutation of V and P_σ is the corresponding permutation matrix (with $P_\sigma(i,j) = P_{\sigma(i),\sigma(j)}$ for all $i, j \in V$), then one can build the new symmetric matrix

$$\sigma(X) := P_\sigma X P_\sigma^{\mathsf{T}} = (X_{\sigma(i),\sigma(j)})_{i,j\in V}.$$

If σ is an automorphism of G, then it preserves the feasible region of the semidefinite program (6.8) defining the theta number $\vartheta(G)$. This is an easy, but very useful fact.

Lemma 6.6.2. *If X is feasible for the program (6.8) and σ is an automorphism of G, then $\sigma(X)$ is again feasible for (6.8), moreover with the same objective value as X.*

Proof. Directly from the fact that $\langle J, \sigma(X)\rangle = \langle J, X\rangle$, $\mathrm{Tr}(\sigma(X)) = \mathrm{Tr}(X)$ and $\sigma(X)_{ij} = X_{\sigma(i)\sigma(j)} = 0$ if $\{i, j\} \in E$ (since σ is an automorphism of G). □

Lemma 6.6.3. *The program (6.8) has an optimal solution X^* which is invariant under action of the automorphism group of G, i.e., satisfies $\sigma(X^*) = X^*$ for all $\sigma \in \mathrm{Aut}(G)$.*

Proof. Let X be an optimal solution of (6.8). By Lemma 6.6.2, $\sigma(X)$ is again an optimal solution for each $\sigma \in \mathrm{Aut}(G)$. Define the matrix

$$X^* = \frac{1}{|\mathrm{Aut}(G)|} \sum_{\sigma\in \mathrm{Aut}(G)} \sigma(X),$$

obtained by averaging over all matrices $\sigma(X)$ for $\sigma \in \mathrm{Aut}(G)$. As the set of optimal solutions of (6.8) is convex, X^* is still an optimal solution of (6.8). Moreover, by construction, X^* is invariant under action of $\mathrm{Aut}(G)$. □

Corollary 6.6.4. *If G is a vertex-transitive graph then the program (6.8) has an optimal solution X^* satisfying $X^*_{ii} = 1/n$ for all $i \in V$ and $X^*e = \frac{\vartheta(G)}{n}e$.*

Proof. By Lemma 6.6.3, there is an optimal solution X^* which is invariant under action of $\mathrm{Aut}(G)$. As G is vertex-transitive, all diagonal entries of X^* are equal: Indeed, let $i, j \in V$ and $\sigma \in \mathrm{Aut}(G)$ such that $\sigma(i) = j$. Then, $X_{jj} = X_{\sigma(i)\sigma(i)} = X_{ii}$. As $\mathrm{Tr}(X^*) = 1$ we must have $X^*_{ii} = 1/n$ for all i. Analogously, the invariance of X^* implies that $\sum_{k\in V} X^*_{ik} = \sum_{k\in V} X^*_{jk}$ for all i, j, i.e., $X^*e = \lambda e$ for some scalar λ. Combining with the condition $\langle J, X^*\rangle = \vartheta(G)$ we obtain that $\lambda = \frac{\vartheta(G)}{n}$. □

Proposition 6.6.5. *If G is a vertex-transitive graph, then $\vartheta(G)\vartheta(\overline{G}) = |V|$.*

Proof. By Corollary 6.6.4, there is an optimal solution X^* of the program (6.8) defining $\vartheta(G)$ which satisfies $X^*_{ii} = 1/n$ for $i \in V$ and $X^*e = \frac{\vartheta(G)}{n}e$. Then $\frac{n^2}{\vartheta(G)}X^* - J \succeq 0$ (check it). Hence, $t = \frac{n}{\vartheta(G)}$ and $C = \frac{n^2}{\vartheta(G)}X^*$ define a feasible solution of the program (6.11) defining $\vartheta(\overline{G})$, which implies $\vartheta(\overline{G}) \leqslant n/\vartheta(G)$. Combining with Proposition 6.6.1 we get the equality $\vartheta(G)\vartheta(\overline{G}) = |V|$. □

For instance, the cycle C_n is vertex-transitive, so that

$$\vartheta(C_n)\vartheta(\overline{C_n}) = n. \tag{6.22}$$

In particular, as C_5 is isomorphic to $\overline{C_5}$, we deduce that

$$\vartheta(C_5) = \sqrt{5}. \tag{6.23}$$

For n even, C_n is bipartite (and thus perfect), so that $\vartheta(C_n) = \alpha(C_n) = \frac{n}{2}$ and $\vartheta(\overline{C_n}) = \omega(C_n) = 2$. For n odd, one can compute $\vartheta(C_n)$ using the above symmetry reduction:

Proposition 6.6.6. *For any odd $n \geqslant 3$,*

$$\vartheta(C_n) = \frac{n\cos(\pi/n)}{1+\cos(\pi/n)} \quad \textit{and} \quad \vartheta(\overline{C_n}) = \frac{1+\cos(\pi/n)}{\cos(\pi/n)}.$$

Proof. As $\vartheta(C_n)\vartheta(\overline{C_n}) = n$, it suffices to compute $\vartheta(C_n)$. We use the formulation (6.12). As C_n is vertex-transitive, there is an optimal solution B whose entries are all equal to 1, except $B_{ij} = 1 + x$ for some scalar x whenever $|i - j| = 1$ (modulo n). In other words, $B = J + xA_{C_n}$, where A_{C_n} is the adjacency matrix of the cycle C_n. Thus $\vartheta(C_n)$ is equal to the minimum value of $\lambda_{\max}(B)$ for all possible x. The eigenvalues of A_{C_n} are known: They are $\omega^k + \omega^{-k}$ (for $k = 0, 1, \cdots, n-1$), where $\omega = e^{\frac{2i\pi}{n}}$ is an n-th root of unity. Hence the eigenvalues of B are

$$n + 2x \text{ and } x(\omega^k + \omega^{-k}) \text{ for } k = 1, \cdots, n-1. \tag{6.24}$$

We minimize the maximum of the values in (6.24) when choosing x such that

$$n + 2x = -2x\cos(\pi/n)$$

(check it). This gives $\vartheta(C_n) = \lambda_{\max}(B) = -2x\cos(\pi/n) = \frac{n\cos(\pi/n)}{1+\cos(\pi/n)}$. □

6.7 Bounding the Shannon capacity

The theta number was introduced by Lovász [3] in connection with the problem of computing the Shannon capacity of a graph, a problem in coding theory considered by Shannon. We need some definitions.

Definition 6.7.1. (Strong product) *Let $G = (V, E)$ and $H = (W, F)$ be two graphs. Their* strong product *is the graph denoted as $G \cdot H$ with node set $V \times W$ and with edges the pairs of distinct nodes $\{(i, r), (j, s)\} \in V \times W$ with ($i = j$ or $\{i, j\} \in E$) and ($r = s$ or $\{r, s\} \in F$).*

If $S \subseteq V$ is stable in G and $T \subseteq W$ is stable in H then $S \times T$ is stable in $G \cdot H$. Hence, $\alpha(G \cdot H) \geqslant \alpha(G)\alpha(H)$. Let G^k denote the strong product of k copies of G, we have that

$$\alpha(G^k) \geqslant (\alpha(G))^k.$$

Based on this, one can verify that

$$\Theta(G) := \sup_{k \geqslant 1} \sqrt[k]{\alpha(G^k)} = \lim_{k \to \infty} \sqrt[k]{\alpha(G^k)}. \tag{6.25}$$

The parameter $\Theta(G)$ was introduced by Shannon in 1956, it is called the *Shannon capacity* of the graph G. The motivation is as follows. Suppose V is a finite alphabet, where some pairs of letters could be confused when they are transmitted over some transmission channel. These pairs of confusable letters can be seen as the edge set E of a graph $G = (V, E)$. Then the stability number of G is the largest number of one-letter messages that can be sent without danger of confusion. Words of length k correspond to k-tuples in V^k. Two words $(i_1, \cdots, i_k)$ and $(j_1, \cdots, j_k)$ can be confused if at every position $h \in [k]$ the two letters i_h and j_h are equal or can be confused, which corresponds to having an edge in the strong product G^k. Hence the largest number of words of length k that can be sent without danger of confusion is equal to the stability number of G^k and the Shannon capacity of G represents the rate of correct transmission of the graph.

For instance, for the 5-cycle C_5, $\alpha(C_5) = 2$, but $\alpha((C_5)^2) \geqslant 5$. Indeed, if $1, 2, 3, 4, 5$ are the nodes of C_5 (in this cyclic order), then the five 2-letter words $(1, 1)$, $(2, 3)$, $(3, 5)$, $(4, 2)$, $(5, 4)$ form a stable set in G^2. This implies that $\Theta(C_5) \geqslant \sqrt{5}$.

Determining the exact Shannon capacity of a graph is a very difficult problem in general, even for small graphs. For instance, the exact value of the Shannon capacity of C_5 was not known until Lovász [3] showed how to use the theta number in order to upper bound the Shannon capacity: Lovász showed that $\Theta(G) \leqslant \vartheta(G)$ and $\vartheta(C_5) = \sqrt{5}$, which implies that $\Theta(C_5) = \sqrt{5}$. For instance, although the exact value of the theta number of C_{2n+1} is known (cf. Proposition 6.6.6), the exact value of the Shannon capacity of C_{2n+1} is not known, already for C_7.

Theorem 6.7.2. *For any graph G, we have that $\Theta(G) \leqslant \vartheta(G)$.*

The proof is based on the multiplicative property of the theta number from Lemma 6.7.3 – which you will prove in Exercise 6.2 – combined with the fact that the theta number upper bounds the stability number: For any integer k, $\alpha(G^k) \leqslant \vartheta(G^k) = (\vartheta(G))^k$ implies $\sqrt[k]{\alpha(G^k)} \leqslant \vartheta(G)$ and thus $\Theta(G) \leqslant \vartheta(G)$.

Lemma 6.7.3. *The theta number of the strong product of two graphs G and H satisfies $\vartheta(G \cdot H) = \vartheta(G)\vartheta(H)$.*

As an application one can compute the Shannon capacity of any graph G which is vertex-transitive and self-complementary (e.g., like C_5).

Theorem 6.7.4. *If $G = (V, E)$ is a vertex-transitive graph, then $\Theta(G \cdot \overline{G}) = |V|$. If, moreover, G is self-complementary, then $\Theta(G) = \sqrt{|V|}$.*

Proof. We have $\Theta(G \cdot \overline{G}) \geqslant \alpha(G \cdot \overline{G}) \geqslant |V|$, since the set of diagonal pairs $\{(i, i) : i \in V\}$ is stable in $G \cdot \overline{G}$. The reverse inequality follows from Lemma 6.7.3 combined with Proposition 6.6.5: $\Theta(G \cdot \overline{G}) \leqslant \vartheta(G \cdot \overline{G}) = \vartheta(G)\vartheta(\overline{G}) = |V|$. If G is isomorphic to $\overline{G}$ then $\Theta(G \cdot \overline{G}) = \Theta(G^2) = (\Theta(G))^2$ (check the rightmost equality). This gives $\Theta(G) = \sqrt{|V|}$. □

6.8 Further reading

In his seminal paper [3], Lovász gives several equivalent formulations for the theta number, and relates it to the Shannon capacity and to some eigenvalue bounds. It is worth noting that Lovász' paper was published in 1979, thus before the discovery of polynomial time algorithms for semidefinite programming. In 1981, together with Grötschel and Schrijver, he derived the polynomial time algorithms for maximum stable sets and graph colorings in perfect graphs, based on the ellipsoid method for solving semidefinite programs. As of today, this is the only known polynomial time algorithm – in particular, no purely combinatorial algorithm is known.

Detailed information about the theta number can also be found in the survey of Knuth [2] and a detailed treatment about the material in this chapter can be found in Chapter 9 of Grötschel, Lovász and Schrijver [1]. In particular, proofs of the geometric characterizations of perfect graphs in Theorems 6.2.2 and 6.5.5 can be found there. Weighted versions of the theta number are considered there, replacing the all-ones objective function $e^{\mathsf{T}}x$ by $w^{\mathsf{T}}x$ where $w \in \mathbb{Z}^V_{\geqslant 0}$. One can give equivalent characterizations, analogue to those we have given for the all-ones weight function. We have restricted our exposition to the all-ones weight function for the sake of simplicity.

6.9 Exercises

6.1 Show the result of Proposition 6.5.9.

6.2** The goal is to show the result of Lemma 6.7.3 about the theta number of the strong product of two graphs $G = (V, E)$ and $H = (W, F)$:

$$\vartheta(G \cdot H) = \vartheta(G)\vartheta(H).$$

(a) Show that $\vartheta(G \cdot H) \geqslant \vartheta(G)\vartheta(H)$.

(b) Show that $\vartheta(G \cdot H) \leqslant \vartheta(G)\vartheta(H)$.

Hint: Use the primal formulation (6.8) for (a), and the dual formulation (6.9) for (b), and think of using Kronecker products of matrices in order to build feasible solutions.

6.3 Let $G = (V = [n], E)$ be a graph. Consider the graph parameter

$$\vartheta_1(G) = \min_{c, u_i} \max_{i \in V} \frac{1}{(c^{\mathsf{T}} u_i)^2},$$

where the minimum is taken over all unit vectors c and all orthonormal representations $u_1, \cdots, u_n$ of G.

Show that $\vartheta(G) = \vartheta_1(G)$.

Hint: For the inequality $\vartheta(G) \leqslant \vartheta_1(G)$ think of using the properties of the theta body from Section 6.5.2. For the inequality $\vartheta_1(G) \leqslant \vartheta(G)$, use an optimal solution B of the dual formulation (6.10) for $\vartheta(G)$ to build the vectors c, u_i.

BIBLIOGRAPHY

[1] M. Grötschel, L. Lovász, A. Schrijver. *Geometric Algorithms in Combinatorial Optimization*, Springer, 1988.

`http://www.zib.de/groetschel/pubnew/paper/groetschellovaszschrijver1988.pdf`

[2] D.E. Knuth. The Sandwich Theorem. *The Electronic Journal of Combinatorics* **1, A1**, 1994.

`http://www.combinatorics.org/ojs/index.php/eljc/article/view/v1i1a1`

[3] L. Lovász. On the Shannon capacity of a graph. *IEEE Transactions on Information Theory* **IT-25**:1–7, 1979.

CHAPTER 7

APPROXIMATING MAX CUT AND THE CUT NORM

The maximum cut problem (MAX CUT) is the following problem in combinatorial optimization. Let $G = (V, E)$ be an undirected graph with vertex set V and edge set $E \subseteq \binom{V}{2}$, where edges $e = \{u, v\} \in E$ are two-element subsets of the vertex set. With every edge $e = \{u, v\}$ we associate a nonnegative weight w_{uv}. Since the graph is undirected we assume that the weights are symmetric $w_{uv} = w_{vu}$. Furthermore, $w_{uv} = 0$ whenever $\{u, v\} \notin E$. We incorporate all the weights into a symmetric matrix $W = (w_{uv}) \in \mathcal{S}^V$. The MAX CUT problem seeks for a partition of the vertex set V into two parts V^-, V^+, cutting the graph into two pieces, so that the sum of edges connecting V^- and V^+, the *weight of the cut* $w(V^-, V^+)$, is maximized:

$$w(V^-, V^+) = \sum_{u \in V^-, v \in V^+} w_{uv}.$$

It is known that the maximum cut problem is an NP-complete problem. So unless the complexity classes P and NP coincide there is no efficient polynomial-time algorithm which solves MAX CUT exactly. In fact, MAX CUT was one of the first problems which were proved to be NP-complete: It is one of Karp's 21 NP-complete problems. Even stronger, Håstad in 2001 showed that it is NP-hard to approximate MAX CUT within a factor of $\frac{16}{17} = 0.941\ldots$ This is in sharp contrast to the MIN CUT problem, where we want to minimize the weight of a non-trivial cut. The MIN CUT problem (and its dual, the MAX FLOW problem) can be solved using linear programming.

On the positive side, one can compute an $0.878\ldots$-approximation of MAX CUT in polynomial time, using semidefinite programming. This algorithm, due

to Goemans and Williamson [2], is one of the most influential approximation algorithms which are based on semidefinite programming.

A problem which is related to (in fact a generalization of) MAX CUT is finding the cut norm of a matrix. Let $A = (A_{ij}) \in \mathbb{R}^{m \times n}$ be a real matrix. The *cut norm* of A is

$$\|A\|_{\square} = \max_{I \subseteq [m], J \subseteq [n]} \left| \sum_{i \in I} \sum_{j \in J} A_{ij} \right|.$$

Computing the cut norm of a matrix has many applications in combinatorics, especially in graph theory. Examples are finding Szemerédi partitions in graphs, or finding maximum acyclic subgraphs. As the cut norm is a generalization of the MAX CUT problem we only can hope for efficient approximations. Today in this lecture we link the cut norm with Grothendieck's inequality, a famous inequality in functional analysis from an even more famous mathematician. Thereby we will derive another approximation algorithm based on semidefinite programming which is due to Krivine from 1979 (although this connection was only found in 2006).

7.1 The algorithm of Goemans and Williamson

7.1.1 Semidefinite relaxation

We now want to describe the Goemans-Williamson algorithm. For this we first reformulate MAX CUT as a (non-convex) quadratic optimization problem having quadratic equality constraints. We start by recalling the construction of the semidefinite relaxation of the MAX CUT problem which we already saw in Chapter 2 and Chapter 5.

With every vertex of the graph $u \in V$ we associate a binary variable $x_u \in \{-1, +1\}$ which indicates whether u lies in V^- or V^+, i.e. $u \in V^-$ if $x_u = -1$ and $u \in V^+$ if $x_u = +1$. We model the binary constraint $x_u \in \{-1, +1\}$ as a quadratic equality constraint

$$x_u^2 = 1, \quad u \in V.$$

For two vertices $u, v \in V$ we have

$$1 - x_u x_v \in \{0, 2\}.$$

This value equals to 0 if u and v lie on the same side of the cut and the value equals to 2 if u and v lie on different sides of the cut. Hence, one can express the weight of a cut, which is defined by the variables $x_u \in \{-1, +1\}$, by

$$w(V^-, V^+) = \frac{1}{2} \left(\frac{1}{2} \sum_{u,v \in V} w_{uv}(1 - x_u x_v) \right).$$

Now, the MAX CUT problem can be equivalently formulated as

$$\mathrm{MAXCUT}(W) = \max\left\{\frac{1}{4}\sum_{u,v\in V} w_{uv}(1-x_u x_v) : x_u^2 = 1, u \in V\right\}.$$

If we replace in this optimization problem the scalar variables $x_u \in \{-1,+1\}$ by vector variables $y_u \in \mathbb{R}^{|V|}$ lying in $|V|$-dimensional Euclidean space, and the product $x_u x_v$ by the inner product $y_u \cdot y_v = y_u^{\mathsf{T}} y_v$, then the we get the following *vector optimization problem*:

$$\mathrm{SDP}(W) = \max\left\{\frac{1}{4}\sum_{u,v\in V} w_{uv}(1-y_u \cdot y_v) : y_u \cdot y_u = 1, u \in V\right\}.$$

Because $y_u \cdot y_u = 1$, we see that the vectors y_u lie on the unit sphere $S^{|V|-1}$. Note also, that every feasible solution x_u of the original problem can be transformed into a feasible solution of the vector optimization problem. We simply set $y_u = (x_u, 0, \ldots, 0)^{\mathsf{T}}$. This means that the maximum value of the vector optimization problem is at least the value of MAX CUT, thus $\mathrm{SDP}(W) \geqslant \mathrm{MAXCUT}(W)$.

We proceed by showing two things: First, it is not difficult to realize that the vector optimization problem can be reformulated as a semidefinite program. Second, we shall prove that the maximum of the vector optimization problem is not too far from the optimal value of the original MAX CUT problem. We show that the inequality

$$\mathrm{SDP}(W) \geqslant \mathrm{MAXCUT}(W) \geqslant 0.878\ldots \cdot \mathrm{SDP}(W)$$

holds for all symmetric weight matrices $W = (w_{uv})$ with nonnegative entries w_{uv}.

To show that the vector optimization problem is a semidefinite program we introduce the inner product matrix of the vectors y_u:

$$Y = (y_u \cdot y_v)_{u,v\in V}.$$

The matrix Y is a positive semidefinite matrix whose diagonal elements are all equal to one. Furthermore,

$$\frac{1}{4}\sum_{u,v\in V} w_{uv}(1-y_u \cdot y_v) = \frac{1}{4}\sum_{u,v\in V} w_{uv} - \frac{1}{4}\langle W, Y\rangle.$$

So it suffices to minimize the trace inner product $\langle W, Y\rangle$ in order to solve the vector optimization problem. Hence, solving the following semidefinite program gives the value of $\mathrm{SDP}(W)$:

$$\mathrm{SDP}(W) = \frac{1}{4}\sum_{u,v\in V} w_{uv} - \frac{1}{4}\min\left\{\langle W, Y\rangle : Y \succeq 0, \langle E_{uu}, Y\rangle = 1, u \in V\right\},$$

where E_{uu} denotes the matrix which has a one at position (u,u) and zeros in all other positions.

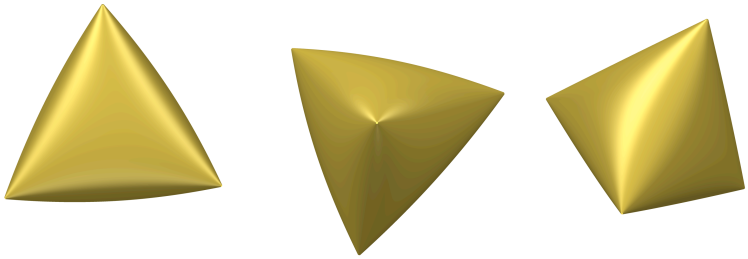


Figure 7.1: Views on spectrahedron behind the semidefinite relaxation.

The figure above illustrates the set of feasible solutions in the case of 3×3 matrices. It is an inflated tetrahedron. These figures were generated by the program `jSurfer` (`http://www.imaginary-exhibition.com/`).

$$Y = \begin{pmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{pmatrix} \succeq 0 \Longleftrightarrow 1 + 2xyz - x^2 - y^2 - z^2 \geqslant 0,\ x, y, z \in [-1, 1].$$

7.1.2 Analysis of the algorithm

Theorem 7.1.1. *For all matrices $W = (w_{uv})$ with nonnegative weights we have the inequality*

$$\mathrm{SDP}(W) \geqslant \mathrm{MAXCUT}(W) \geqslant 0.878\ldots \cdot \mathrm{SDP}(W).$$

Proof. The proof is algorithmic and it gives an approximation algorithm which approximates the MAX CUT problem within a ratio of $0.878\ldots$. The Goemans-Williamson algorithm has five steps:

1. Solve $\mathrm{SDP}(W)$ obtaining an optimal solution Y.
2. Perform a Cholesky decomposition of Y to find $y_u \in S^{|V|-1}$, with $u \in V$.
3. Choose a random vector $r \in S^{|V|-1}$ according to the rotationally invariant probability distribution on the unit sphere.
4. Define a cut by $x_u = \mathrm{sign}(y_u \cdot r)$, for $u \in V$.
5. Check whether $\frac{1}{4} \sum_{u,v \in V} w_{uv}(1 - x_u x_v) \geqslant 0.878\ldots \cdot \mathrm{SDP}(W)$. If not, go to step 3.

The following lemma, also known as *Grothendieck's identity*, is the key to the proof of the theorem.

Lemma 7.1.2. *Let x, y be unit vectors and let r be a random unit vector chosen according to the rotationally invariant probability distribution on the unit sphere. Then, the expectation of the random variable* $\operatorname{sign}(x \cdot r)\operatorname{sign}(y \cdot r) \in \{-1, +1\}$ *equals*

$$\mathbb{E}[\operatorname{sign}(x \cdot r)\operatorname{sign}(y \cdot r)] = \frac{2}{\pi}\arcsin x \cdot y.$$

Proof. By definition, the expectation can be computed as

$$\begin{aligned}\mathbb{E}[\operatorname{sign}(x \cdot r)\operatorname{sign}(y \cdot r)] =& (+1) \cdot \mathbb{P}[\operatorname{sign}(x \cdot r) = \operatorname{sign}(y \cdot r)] \\ &+ (-1) \cdot \mathbb{P}[\operatorname{sign}(x \cdot r) \neq \operatorname{sign}(y \cdot r)].\end{aligned}$$

Note that

$$\mathbb{P}[\operatorname{sign}(x \cdot r) \neq \operatorname{sign}(x \cdot r)] = 1 - \mathbb{P}[\operatorname{sign}(x \cdot r) = \operatorname{sign}(y \cdot r)],$$

so that we only have to compute the probability that the signs of $x \cdot r$ and $y \cdot r$ are the same. Since the probability distribution from which we sample the unit vector r is rotationally invariant we can assume that x, y and r lie in a common plane and hence on a unit circle and that r is chosen according to the uniform distribution on this circle. Then the probability that the signs of $x \cdot r$ and $y \cdot r$ are the same only depends on the angle between x and y. Using a figure (draw one!) it is easy to see that

$$\mathbb{P}[\operatorname{sign}(x \cdot r) = \operatorname{sign}(y \cdot r)] = 2 \cdot \frac{1}{2\pi}\arccos x \cdot y = \frac{1}{\pi}\arccos x \cdot y.$$

Now,

$$\begin{aligned}\mathbb{E}[\operatorname{sign}(x \cdot r)\operatorname{sign}(y \cdot r)] &= \frac{1}{\pi}\arccos x \cdot y - (1 - \frac{1}{\pi}\arccos x \cdot y) \\ &= \frac{2}{\pi}\arcsin x \cdot y,\end{aligned}$$

where we used the trigonometric identity

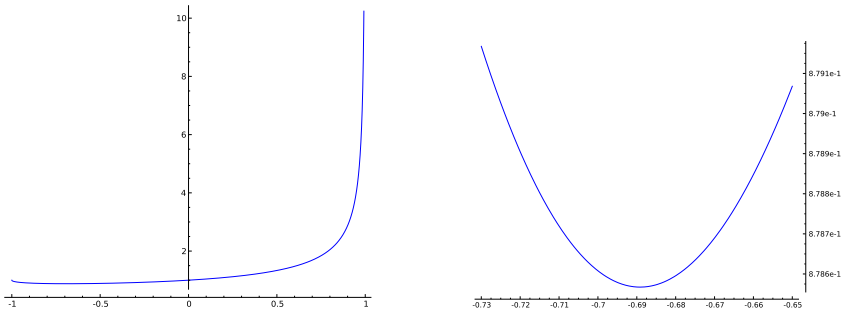
$$\arcsin t + \arccos t = \frac{\pi}{2},$$

to get the desired result. □

Let us apply Grothendieck's identity: Using elementary univariate calculus one can show that

$$\frac{1 - \mathbb{E}[\operatorname{sign}(x \cdot r)\operatorname{sign}(y \cdot r)]}{1 - x \cdot y} = \frac{1 - \frac{2}{\pi}\arcsin t}{1 - t} \geqslant 0.878\ldots, \tag{7.1}$$

holds, where $t = x \cdot y \in [-1, 1]$. To "see" this one can also plot the function using SAGE:



```
plot((1-2/pi*arcsin(x))/(1-x), (x,-1,1))
plot((1-2/pi*arcsin(x))/(1-x), (x,-0.73,-0.62))
```

This can be used to estimate the ratio between $\mathrm{MAXCUT}(W)$ and $\mathrm{SDP}(W)$. Clearly,

$$\mathrm{MAXCUT}(W) \geqslant \mathbb{E}\left[\frac{1}{4}\sum_{u,v\in V} w_{uv}(1 - x_u x_v)\right].$$

By linearity of expectation,

$$\mathbb{E}\left[\frac{1}{4}\sum_{u,v\in V} w_{uv}(1 - x_u x_v)\right] = \frac{1}{4}\sum_{u,v\in V} w_{uv}(1 - \mathbb{E}[x_u x_v]).$$

Since w_{uv} is nonnegative we can go on by estimating

$$1 - \mathbb{E}[x_u x_v] = 1 - \mathbb{E}[\mathrm{sign}(y_u \cdot r)\,\mathrm{sign}(y_v \cdot r)]$$

in every summand using (7.1) getting

$$\mathbb{E}[(1 - x_u x_v)] \geqslant 0.878\ldots(1 - y_u \cdot y_v).$$

Putting it together,

$$\begin{aligned}
\mathrm{MAXCUT}(W) &\geqslant \mathbb{E}\left[\frac{1}{4}\sum_{u,v\in V} w_{uv}(1 - x_u x_v)\right] \\
&\geqslant 0.878\ldots\frac{1}{4}\sum_{u,v\in V} w_{uv}(1 - y_u \cdot y_v) \\
&= 0.878\ldots\cdot\mathrm{SDP}(W),
\end{aligned}$$

which proves the theorem. □

We finish the explanation of the Goemans-Williamson algorithm by some further remarks. The steps 3. and 4. in the algorithm are called a *randomized rounding procedure* because a solution of a semidefinite program is "rounded" (or better: projected) to a solution of the original combinatorial problem with the help of randomness.

Note also that because the expectation of the constructed solution is at least $0.878\cdots\,\mathrm{SDP}(W)$ the algorithm eventually terminates; it will pass step 5 and without getting stuck in an endless loop. One can show that with high probability we do not have to wait long until the condition in step 5 is fulfilled.

7.1.3 Remarks on the algorithm

One can modify the Goemans-Williamson algorithm so that it becomes an algorithm which runs deterministically (without the use of randomness) in polynomial time and which gives the same approximation ratio. This was done by Mahajan and Ramesh in 1995.

It remains to give a procedure which samples a random vector from the unit sphere. This can be done if one can sample random numbers from the standard normal (Gaussian) distribution (with mean zero and variance one) which has probability density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Many software packages include a procedure which produces random numbers from the standard normal distribution. SAGE:

```
sage: T = RealDistribution('gaussian', 1)
sage: T.get_random_element()
0.818610064197
```

If we sample n real numbers $x_1, \ldots, x_n$ independently uniformly at random from the standard normal distribution, then, the vector

$$r = \frac{1}{\sqrt{x_1^2 + \cdots + x_n^2}} (x_1, \ldots, x_n)^{\mathsf{T}} \in S^{n-1}$$

is distributed according to the rotationally invariant probability measure on the unit sphere.

7.2 Cut norm and Grothendieck's inequality

7.2.1 Cut norm of a matrix

The cut norm of a matrix $A = (A_{ij}) \in \mathbb{R}^{m \times n}$ is

$$\|A\|_{\square} = \max_{I \subseteq [m], J \subseteq [n]} \left| \sum_{i \in I} \sum_{j \in J} A_{ij} \right|$$

Related to the cut norm is the following norm which is given as a quadratic optimization problem

$$\|A\|_{\infty \to 1} = \max \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} x_i y_j : x_i^2 = y_j^2 = 1, i \in [m], j \in [n] \right\}.$$

The notation $\|A\|_{\infty\to 1}$ indicates that this is an operator norm of an operator mapping the space ℓ_∞^n to the space ℓ_1^m. We will not use this fact here.

Lemma 7.2.1. *Let $A \in \mathbb{R}^{m\times n}$ be a matrix.*

a) We have the relation

$$4\|A\|_\square \geqslant \|A\|_{\infty\to 1} \geqslant \|A\|_\square.$$

b) If the row sum and the column sum of A are both 0, we have

$$\|A\|_\square = \frac{1}{4}\|A\|_{\infty\to 1}.$$

c) There exists a matrix $B \in \mathbb{R}^{(m+1)\times(n+1)}$ with row sum and column sum equal to zero such that

$$\|A\|_\square = \|B\|_\square.$$

Proof. a) For $x_i, y_j \in \{\pm 1\}$ we split

$$\begin{aligned}\sum_{i=1}^m \sum_{j=1}^n A_{ij}x_i y_j = &\sum_{(i,j):x_i=1,y_j=1} A_{ij} + \sum_{(i,j):x_i=-1,y_j=-1} A_{ij} \\ &- \sum_{(i,j):x_i=1,y_j=-1} A_{ij} - \sum_{(i,j):x_i=-1,y_j=1} A_{ij}\end{aligned}$$

and every term is bounded in absolute value from above by $\|A\|_\square$, hence the first inequality $4\|A\|_\square \geqslant \|A\|_{\infty\to 1}$ follows.

For the other inequality let $I \subseteq [m]$ and $J \subseteq [n]$ be given so that

$$\|A\|_\square = \left|\sum_{i\in I, j\in J} A_{ij}\right|.$$

Define $x_i = 1$ if $i \in I$ and $x_i = -1$ if $i \notin I$ and similarly $y_j = \pm 1$. Then,

$$\begin{aligned}\|A\|_\square &= \sum_{i,j} A_{ij}\frac{1+x_i}{2}\frac{1+y_j}{2} \\ &= \frac{1}{4}\left(\sum_{i,j} A_{ij} + \sum_{i,j} A_{ij}x_i + \sum_{i,j} A_{ij}y_j + \sum_{i,j} A_{ij}x_i y_j\right).\end{aligned} \tag{7.2}$$

This proves the second inequality $\|A\|_\square \leqslant \|A\|_{\infty\to 1}$ because the absolute value of every of the four summands is at most $\|A\|_{\infty\to 1}$.

b) The second statement of the lemma follows by looking at (7.2) and applying the additional assumption.

c) Simply construct B by adding a row and a column to A such that the row sum and the column sum are equal to 0. Checking that $\|A\|_\square$ equals $\|B\|_\square$ is an exercise. □

The following construction shows that computing the cut norm of a matrix is at least as difficult as computing the MAX CUT of a graph. Let $G = (V, E)$ be a graph with n vertices $v_1, \dots, v_n$ and m edges $e_1, \dots, e_m$, and let $W = (w_{jk}) \in \mathcal{S}^V$ be a weight matrix with nonnegative weights. Now we define a matrix $A \in \mathbb{R}^{2m \times n}$ whose cut norm coincides with $\mathrm{MAXCUT}(W)$. For this orient the edges of the graph in some arbitrary way. If e_i is arc from v_j to v_k then we set

$$A_{2i-1,j} = A_{2i,k} = w_{jk}, \quad A_{2i-1,k} = A_{2i,j} = -w_{jk}.$$

All other entries of A are zero.

Lemma 7.2.2. *Using the construction above we have*

$$\mathrm{MAXCUT}(W) = \frac{1}{4} \cdot \|A\|_{\infty \to 1} = \|A\|_{\square}.$$

Proof. Exercise 7.3 (a). □

7.2.2 Grothendieck's inequality

The semidefinite relaxation of $\|A\|_{\infty \to 1}$ is

$$\mathrm{SDP}_{\infty \to 1}(A) = \max \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} u_i v_j : \|u_i\| = \|v_j\| = 1, i \in [m], j \in [n] \right\},$$

where we optimize over $m + n$ unit vectors $u_i, v_j \in \mathbb{R}^{m+n}$. Note that this optimization problem is indeed a semidefinite program (why?).

Theorem 7.2.3 (Grothendieck's inequality)**.** *There is a constant K so that for all matrices $A \in \mathbb{R}^{m \times n}$ the inequality*

$$\|A\|_{\infty \to 1} \leqslant \mathrm{SDP}_{\infty \to 1}(A) \leqslant K \|A\|_{\infty \to 1}$$

holds.

The smallest constant K for which the second inequality holds, is called the *Grothendieck constant* K_G. It is known to lie between $1.676\ldots$ and $1.782\ldots$ but its exact value is currently not known. In the following we will prove that

$$K_G \leqslant \frac{\pi}{2 \ln(1 + \sqrt{2})} = 1.782\ldots$$

The argument will also rely on an approximation algorithm which uses randomized rounding (in a tricky way).

Thereby, and using Lemma 7.2.1, we find an algorithm which approximates the cut norm within a factor of $(1.782\ldots)^{-1} = 0.561\ldots$

7.2.3 Proof of Grothendieck's inequality

1. Solve $\mathrm{SDP}_{\infty\to 1}(A)$. This gives optimal unit vectors
$$u_1, \dots, u_m, v_1, \dots, v_n \in S^{m+n-1}.$$
2. Use these unit vectors to construct new unit vectors
$$u'_1, \dots, u'_m, v'_1, \dots, v'_n \in S^{m+n-1}$$
according to Krivine's trick presented in Lemma 7.2.4 below.
3. Choose a random vector $r \in S^{m+n-1}$ according to the rotationally invariant probability distribution on the unit sphere.
4. Randomized rounding: Set
$$x_i = \operatorname{sign}(u'_i \cdot r),\ y_j = \operatorname{sign}(v'_j \cdot r).$$

We analyze the expected quality of the solution x_i, y_j. By linearity of expectation we have

$$\mathrm{SDP}_{\infty\to 1}(A) \geqslant \mathbb{E}\left[\sum_{i=1}^{m}\sum_{j=1}^{m} A_{ij} x_i \cdot y_j\right] = \sum_{i=1}^{m}\sum_{j=1}^{m} A_{ij}\mathbb{E}\left[\operatorname{sign}(u'_i \cdot r)\operatorname{sign}(v'_j \cdot r)\right].$$

Now by Lemma 7.2.4 the last expectation will turn out to be $\beta u_i \cdot v_j$. Then the total sum will be equal $\beta\,\mathrm{SDP}_{\infty\to 1}(A)$ and hence $K_G \leqslant \beta^{-1}$.

Now the following lemma, Krivine's trick, finishes the proof of Theorem 7.2.3.

Lemma 7.2.4. *Let $u_1, \dots, u_m$ and $v_1, \dots, v_n$ be unit vectors in $\mathbb{R}^{m+n}$. Then there exist unit vectors $u'_1, \dots, u'_m$ and $v'_1, \dots, v'_n$ in $\mathbb{R}^{m+n}$ such that*

$$\mathbb{E}\left[\operatorname{sign}(u'_i \cdot r)\operatorname{sign}(v'_j \cdot r)\right] = \beta u_i \cdot v_j,$$

holds with

$$\beta = \frac{2}{\pi}\ln(1+\sqrt{2}) = 0.561\dots$$

Proof. Define the function $E : [-1,+1] \to [-1,+1]$ by $E(t) = \frac{2}{\pi}\arcsin t$. Then by Grothendieck's identity, Lemma 8.2.2,

$$\mathbb{E}\left[\operatorname{sign}(u'_i \cdot r)\operatorname{sign}(v'_j \cdot r)\right] = E(u'_i \cdot v'_j).$$

Now the idea is to invert the function E so that we have

$$u'_i \cdot v'_j = E^{-1}(\beta u_i \cdot v_j)$$

and use the series expansion

$$E^{-1}(t) = \sum_{r=0}^{\infty} g_{2r+1} t^{2r+1},$$

which is valid for all $t \in [-1,1]$ to define u_i' and v_j'.

For this define the infinite dimensional Hilbert space

$$H = \bigoplus_{r=0}^{\infty} (\mathbb{R}^{m+n})^{\otimes 2r+1},$$

and the vectors $u_i', v_j' \in H$ componentwise by

$$(u_i')_r = \mathrm{sign}(g_{2r+1})\sqrt{|g_{2r+1}|\beta^{2r+1}}\, u_i^{\otimes 2r+1}$$

and

$$(v_j')_r = \sqrt{|g_{2r+1}|\beta^{2r+1}}\, v_j^{\otimes 2r+1}.$$

Then

$$u_i' \cdot v_j' = \sum_{r=0}^{\infty} g_{2r+1}\beta^{2r+1}(u_i \cdot v_j)^{2r+1} = E^{-1}(\beta u_i \cdot v_j)$$

and

$$1 = u_i' \cdot u_i' = v_j' \cdot v_j' = \sum_{r=0}^{\infty} |g_{2r+1}|\beta^{2r+1},$$

which defines the value of β uniquely.

It's a fun exercise to work out β explicitly: We have

$$E(t) = \frac{2}{\pi}\arcsin t,$$

and so

$$E^{-1}(t) = \sin\left(\frac{\pi}{2}t\right) = \sum_{r=0}^{\infty} \frac{(-1)^{2r+1}}{(2r+1)!}\left(\frac{\pi}{2}t\right)^{2r+1}.$$

Hence,

$$1 = \sum_{r=0}^{\infty} \left|\frac{(-1)^{2r+1}}{(2r+1)!}\right| \left(\frac{\pi}{2}\beta\right)^{2r+1} = \sinh\left(\frac{\pi}{2}\beta\right),$$

which implies

$$\beta = \frac{2}{\pi}\mathrm{arsinh}\ 1 = \frac{2}{\pi}\ln(1+\sqrt{2})$$

because $\mathrm{arsinh}\ t = \ln(t + \sqrt{t^2+1})$. □

Last concern: How to find/approximate u_i', v_j' in polynomial time? Answer: we approximate the inner product matrix

$$\left(u_i' \cdot v_j'\right) = \sum_{r=0}^{\infty} g_{2r+1}\beta^{2r+1}\left((u_i \cdot v_j)^{2r+1}\right)$$

by its series expansion which converges fast enough and then we use its Cholesky decomposition.

7.3 Further reading

How good is the MAX CUT algorithm? Are there graphs where the value of the semidefinite relaxation and the value of the maximal cut are a factor of $0.878\ldots$ apart or is this value $0.878\ldots$, which maybe looks strange at first sight, only an artefact of our analysis? It turned out that the value is optimal. In Exercise 7.1 (b) you will see that already for the 5-cycle C_5 the gap is close to $0.878\ldots$. In 2002 Feige and Schechtmann gave an infinite family of graphs for which the ratio $\mathrm{MAXCUT}\,/\,\mathrm{SDP}$ converges to exactly $0.878\ldots$. This proof uses a lot of nice mathematics (continuous graphs, Voronoi regions, isoperimetirc inequality) and it is explained in detail in the Chapter 8 of the book *Approximation Algorithms and Semidefinite Programming* of Gärtner and Matoušek.

Rather recently in 2007, Khot, Kindler, Mossel, O'Donnell showed that the algorithm is optimal in the following sense: If the unique games conjecture is true, then there is no polynomial time approximation algorithm achieving a better approximation ratio than $0.878\ldots$ unless $\mathrm{P} = \mathrm{NP}$. Currently, the validity and the implications of the unique games conjecture are under heavy investigation. The topic of the unique games conjecture is too hot for this course, although it is very fascinating. The book of Gärtner and Matoušek also contains an introduction to the unique games conjecture.

How good is the upper bound K_G? Finding the value of K_G is an longstanding open problem. The best-known lower bound is $1.676\ldots$ by unpublished work of Davie and Reeds.

It was widely believed that Krivine's analysis gives the right value of K_G. So it came as a shock (at least to the author of these notes) when Braverman, Makarychev, Makarychev, and Naor proved in 2011 that one can improve it slightly, by a clever modification of Krivine's trick and much more complicated computations.

Alon and Naor [2] discovered the connection between the cut norm and Grothendieck's inequality in 2006. Since then Grothendieck's inequality became a unifying concept in combinatorial optimization; see the survey [3] of Khot and Naor.

7.4 Historical remarks and anecdotes

In 2000, Goemans and Williamson won the Fulkerson prize (sponsored jointly by the Mathematical Programming Society and the AMS) which recognizes outstanding papers in the area of discrete mathematics for this result.

About the finding of the approximation ratio $0.878\ldots$ Knuth writes in the article "Mathematical Vanity Plates":

Sometimes people obtain mathematically significant license plates purely by accident, without making a personal selection. A striking example of this phenomenon is the case of Michel Goemans, who received the following innocuous-looking plate from the Massachusetts Registry of Motor Vehicles when he and his wife purchased a Subaru at the beginning of September 1993:

Two weeks later, Michel got together with his former student David Williamson, and they suddenly realized how to solve a problem that they had been working on for some years: to get good approximations for maximum cut and satisfiability problems by exploiting semidefinite programming. Lo and behold, their new method—which led to a famous, award-winning paper [15]—yielded the approximation factor .878! There it was, right on the license, with `C`, `S`, and `W` standing respectively for cut, satisfiability, and Williamson.

Winfried Scharlau tries to answer the question: Who is Alexander Grothendieck?

`http://www.ams.org/notices/200808/tx080800930p.pdf`

7.5 Questions

7.1** (a) Find an approximation algorithm which approximates MAX CUT without using semidefinite programming: (Hint: What happens if one assigns $x_u \in \{-1, +1\}$ uniformly at random with probability $1/2$?)

(b) Let W_n be the adjacency matrix of the n-cycle C_n. Find a closed formula for $\mathrm{SDP}(W_n)$. How does this compare to $\mathrm{MAXCUT}(W_n)$?

(c) Let $A \in \mathcal{S}^n_{\geq 0}$ be a positive semidefinite matrix. Consider the quadratic optimization problem

$$\mathrm{BQP}(A) = \max \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j : x_i^2 = 1, i \in [n] \right\}$$

and its semidefinite relaxation

$$\mathrm{SDP}_{\mathrm{BQP}}(A) = \max \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} u_i \cdot u_j : \|u_i\|^2 = 1, i \in [n] \right\}$$

Show that

$$\mathrm{SDP}_{\mathrm{BQP}}(A) \geqslant \mathrm{BQP}(A) \geqslant \frac{\pi}{2} \mathrm{SDP}_{\mathrm{BQP}}(A)$$

Hint: Let $X = (X_{ij}) \in \mathcal{S}^n_{\geq 0}$ be a positive semidefinite matrix. Then the matrix

$$\left(\tfrac{2}{\pi} \arcsin X_{ij}\right)_{1 \leqslant i,j \leqslant n} - \frac{2}{\pi} X$$

is positive semidefinite as well (recall: Taylor series of $\arcsin$ and Hadamard product from Chapter 1).

7.2** Define the Erdős-Rényi random graph model $G(n,p)$ with $p \in [0,1]$ as follows: Choose a graph with n vertices uniformly at random by connecting two vertices with probability p. Define accordingly the random weight matrix $W = (w_{uv}) \in \mathbb{R}^{n \times n}$ by setting $w_{uv} = 1$ if $\{u,v\} \in E$ and $w_{uv} = 0$ if $\{u,v\} \notin E$.

Implement the Goemans-Williamson algorithm in SAGE. Use it to estimate numerically for $n = 100$ and $p = \frac{m}{10}$ with $m = 1, 2, \ldots, 9$ the average value of $\mathrm{SDP}(W)$ and the average value of the Goemans-Williamson algorithm.

7.3 (a) Prove Lemma 7.2.2.

(b) Prove

$$\|A\|_{\infty \to 1} = \max \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} x_i y_j : x_i, y_j \in [-1,1], i \in [m], j \in [n] \right\}.$$

BIBLIOGRAPHY

[1] N. Alon, A. Naor, *Approximating the cut-norm via Grothendieck's inequality*, SIAM J. Comp. **35** (2006), 787–803.

[2] M.X. Goemans, D.P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. ACM **42** (1995), 1115–1145.

[3] S. Khot, A. Naor, *Grothendieck-type inequalities in combinatorial optimization*, arXiv:1108.2464 [cs.DS]

`http://arxiv.org/abs/1108.2464`

CHAPTER 8

GENERALIZATIONS OF GROTHENDIECK'S INEQUALITY AND APPLICATIONS

In the second part of the last lecture we considered Grothendieck's inequality: There is a constant K so that for all matrices $A \in \mathbb{R}^{m \times n}$ the inequality

$$\|A\|_{\infty \to 1} \leqslant \mathrm{SDP}_{\infty \to 1}(A) \leqslant K \|A\|_{\infty \to 1}$$

holds, where:

$$\|A\|_{\infty \to 1} = \max \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} x_i y_j : x_i^2 = y_j^2 = 1, i \in [m], j \in [n] \right\}.$$

and where the semidefinite relaxation equals

$$\mathrm{SDP}_{\infty \to 1}(A) = \max \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} u_i \cdot v_j : \|u_i\| = \|v_j\| = 1, i \in [m], j \in [n] \right\}.$$

We saw that $\|A\|_{\infty \to 1}$ is closely related to the cut norm which is useful in many graph theoretic applications.

The number $\|A\|_{\infty \to 1}$ also has a meaning in theoretical physics. It can be used to find ground states in the Ising model. The Ising model (named after the physicist Ernst Ising), is a mathematical model of ferromagnetism in statistical mechanics. The model consists of discrete variables called spins that can be in

one of two states, namely $+1$ or -1, UP or DOWN. The spins are arranged in a graph, and each spin only interacts with its neighbors.

In many cases, the interaction graph is a finite subgraph of the integer lattice $\mathbb{Z}^n$ where the vertices are the lattice points and where two vertices are connected if their Euclidean distance is one. These graphs are bipartite since they can be partitioned into even and odd vertices, corresponding to the parity of the sum of the coordinates. Let $G = (V, E)$ be a bipartite interaction graph. The potential function is given by a symmetric matrix $A = (A_{uv}) \in \mathcal{S}^V$. $A_{uv} = 0$ if u and v are not adjacent, A_{uv} is positive if there is ferromagnetic interaction between u and v, and A_{uv} is negative if there is antiferromagnetic interaction. The particles possess a spin $x \in \{-1, +1\}^V$. In the absence of an external field, the total energy of the system is given by

$$-\sum_{\{u,v\}\in E} A_{uv} x_u x_v.$$

The ground state of this model is a configuration of spins $x \in \{-1, +1\}^V$ which minimizes the total energy. So computing the $x_u \in \{-1, +1\}$ which give $\|A\|_{\infty\to 1}$ is equivalent to finding the ground state and computing $\mathrm{SDP}_{\infty\to 1}$ amounts to approximate this ground state energy.

In this lecture we consider two generalizations of this bipartite Ising model.

We start by studying the Ising model for *arbitrary interaction graphs* and we find approximations of the ground state energy. The quality of this approximation will clearly depend on properties of the interaction graph. In particular, the theta number will appear here in an unexpected way.

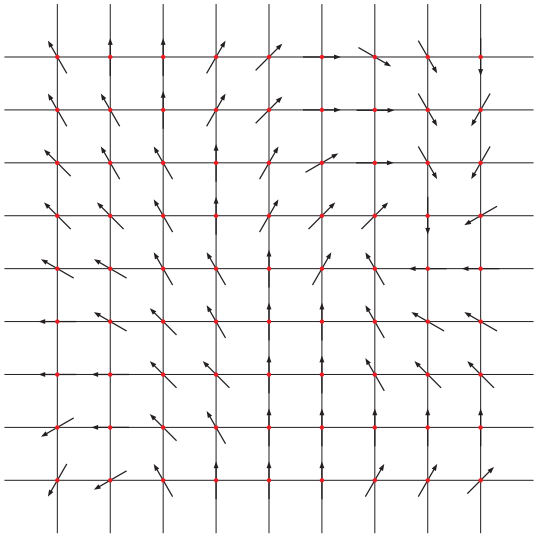


Figure 8.1: Spins in the XY model

Another generalization will be the consideration of *more complicated spins*. Instead of looking only at spins attaining the values -1 and $+1$ as in the Ising model, the r-vector model considers spins which are vectors in the unit sphere $S^{r-1} = \{x \in \mathbb{R}^r : x \cdot x = 1\}$. The case $r = 1$ corresponds to the Ising model, the case $r = 2$ to the XY model, the case $r = 3$ to the Heisenberg model, and the case $r = |V|$ to the Berlin-Kac spherical model. We will derive approximations of ground state energies

$$-\max \sum_{\{u,v\}\in E} A_{u,v} x_u \cdot x_v, \quad \text{for } x_u \in S^{r-1} \text{ and } u \in V$$

for fixed r and for bipartite graphs.

In principle a mixture of both generalizations is possible. We do not give it here as it would require adding even more technical details.

8.1 The Grothendieck constant of a graph

The *Grothendieck constant* of an undirected graph $G = (V, E)$ is the smallest constant[1] $K(G) = K$ so that for every symmetric matrix $A \in \mathcal{S}^V$ the inequality

$$\max \left\{ \sum_{\{u,v\}\in E} A_{uv} f_u \cdot f_v : f_u \in \mathbb{R}^V, u \in V, \|f_u\| = 1 \right\}$$
$$\leqslant K \max \left\{ \sum_{\{u,v\}\in E} A_{uv} x_u x_v : x_u = \pm 1, u \in V \right\}$$

holds true. The left hand side is the semidefinite relaxation of the right hand side. Furthermore, the original Grothendieck constant which we studied in the last lecture is equal to the supremum of $K(G)$ over all bipartite graphs G; see Exercise 8.1 (a).

The following theorem gives a surprising connection between the Grothendieck constant of a graph and the theta number.

Theorem 8.1.1. *There is a constant C so that for any graph G we have*

$$K(G) \leqslant C \ln \vartheta(\overline{G}),$$

where $\vartheta(\overline{G})$ is the theta number of the complementary graph of G.

The proof of this theorem will again be based on an approximation algorithm which performs randomized rounding of the solution of the semidefinite relaxation.

[1] Do not confuse the $K(G)$ with K_G of the last lecture.

8.1.1 Randomized rounding by truncating

In the algorithm we use the constant $M = 3\sqrt{1 + \ln \vartheta(\overline{G})}$. The meaning of it will become clear when we analyze the algorithm.

1. Solve the semidefinite relaxation

$$\Gamma_{\max} = \max \left\{ \sum_{\{u,v\} \in E} A_{uv} f_u \cdot f_v : f_u \in \mathbb{R}^V, u \in V, \|f_u\| = 1 \right\}$$

2. Choose a random vector $z = (z_u) \in \mathbb{R}^V$ so that every entry z_u is distributed independently according to the standard normal distribution with mean 0 and variance 1: $z_u \sim N(0,1)$.

3. Round to real numbers $y_u = z \cdot f_u$ for all $u \in V$.

4. Truncate y_u by setting

$$x_u = \begin{cases} y_u & \text{if } |y_u| \leqslant M, \\ 0 & \text{otherwise} \end{cases}$$

We denote by Δ the optimal value of the ± 1-constrained problem

$$\Delta = \max \left\{ \sum_{\{u,v\} \in E} A_{uv} x_u x_v : x_u = \pm 1, u \in V \right\}.$$

Important note. The solution x_u which the algorithm determines does not satisfy the ± 1-constraint. It only lies in the interval $[-M, M]$ by construction. However, it is easy to show (how? see Exercise 7.3 (b)) that

$$M^2 \Delta = \max \left\{ \sum_{\{u,v\} \in E} A_{uv} x_u x_v : x_u \in [-M, M], u \in V \right\}$$

holds. Similarly,

$$\Gamma_{\max} = \max \left\{ \sum_{\{u,v\} \in E} A_{uv} f_u \cdot f_v : f_u \in \mathbb{R}^V, u \in V, \|f_u\| \leqslant 1 \right\}$$

In the remainder of this section we shall prove the theorem by giving an explicit bound on the ratio $\Gamma_{\max}/\Delta$.

8.1.2 Quality of expected solution

The expected quality of the solution x_u is

$$\begin{aligned}
&\sum_{\{u,v\}\in E} A_{uv}\mathbb{E}[x_u x_v] \\
&= \sum_{\{u,v\}\in E} A_{uv}(\mathbb{E}[y_u y_v] - \mathbb{E}[y_u(y_v - x_v)] \\
&\qquad - \mathbb{E}[y_v(y_u - x_u)] + \mathbb{E}[(y_u - x_u)(y_v - x_v)]) \\
&= \Gamma_{\max} - \mathbb{E}\left[\sum_{\{u,v\}\in E} A_{uv}((y_u(y_v - x_v) + y_v(y_u - x_u))\right] \\
&\quad + \mathbb{E}\left[\sum_{\{u,v\}\in E} A_{uv}(y_u - x_u)(y_v - x_v)\right]
\end{aligned} \tag{8.1}$$

because $\mathbb{E}[y_u y_v] = f_u \cdot f_v$ (Exercise 8.1 (b)).

8.1.3 A useful lemma

To estimate the second and third summand in (8.1) we use the following lemma.

Lemma 8.1.2. *Let X_u, Y_u be random variables with $u \in V$. Assume*

$$\mathbb{E}[X_u^2] \leqslant A \quad \textit{and} \quad \mathbb{E}[Y_u^2] \leqslant B.$$

Then,

$$\mathbb{E}\left[\sum_{\{u,v\}\in E} A_{uv}(X_u Y_v + X_v Y_u)\right] \leqslant 2\sqrt{AB}(\Gamma_{\max} - \Gamma_{\min}),$$

where

$$\Gamma_{\min} = \min\left\{\sum_{\{u,v\}\in E} A_{uv} f_u \cdot f_v : f_u \in \mathbb{R}^V, u \in V, \|f_u\| \leqslant 1\right\}.$$

Proof. If $\mathbb{E}[X_u^2] \leqslant 1$, then

$$\mathbb{E}\left[\sum_{\{u,v\}\in E} A_{uv} X_u X_v\right] \in [\Gamma_{\min}, \Gamma_{\max}]. \tag{8.2}$$

This follows from the fact we can write

$$\big(\mathbb{E}[X_u X_v]\big)_{u,v\in V} = \big(f_u \cdot f_v\big)_{u,v\in V}$$

because the matrix on the left hand side is positive semidefinite (Exercise 8.1 (c)) and thus has a Cholesky factorization.

We introduce new variables U_u and V_u to be able to apply (8.2). The new variables are

$$U_u = \frac{1}{2}\left(X_u/\sqrt{A} + Y_u/\sqrt{B}\right), \quad V_v = \frac{1}{2}\left(X_u/\sqrt{A} - Y_u/\sqrt{B}\right).$$

Then $\mathbb{E}[U_u^2] \leqslant 1$ and $\mathbb{E}[V_u^2] \leqslant 1$ (verify it). So we can apply (8.2)

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\{u,v\}\in E} A_{uv}(X_uY_v + X_vY_u)\right] \\
&= 2\sqrt{AB}\left(\mathbb{E}\left[\sum_{\{u,v\}\in E} A_{uv}U_uU_v\right] - \mathbb{E}\left[\sum_{\{u,v\}\in E} A_{uv}V_uV_v\right]\right) \\
&\leqslant 2\sqrt{AB}(\Gamma_{\max} - \Gamma_{\min}).
\end{aligned}$$

□

8.1.4 Estimating A and B in the useful lemma

It is clear, that $\mathbb{E}[y_u^2] = 1$. We find an upper bound for $\mathbb{E}[(y_u - x_u)^2]$ in the following lemma.

Lemma 8.1.3.

$$\mathbb{E}[(y_u - x_u)^2] = 2\frac{1}{\sqrt{2\pi}}\int_M^\infty t^2 e^{-t^2/2}dt \leqslant Me^{-M^2/2}.$$

Proof. The equation follows from the definition of y_u and x_u and the normal distribution. The inequality is coming from the following simple but noteworthy trick to estimate the integrand by an expression which can be integrated:

$$t^2e^{-t^2/2} \leqslant (t^2 + t^{-2})e^{-t^2/2}.$$

Then,

$$\int (t^2 + t^{-2})e^{-t^2/2}dt = -\frac{t^2+1}{t}e^{-t^2/2} + \text{constant of integration},$$

and the lemma follows by

$$2\frac{1}{\sqrt{2\pi}}\int_M^\infty t^2e^{-t^2/2}dt \leqslant \sqrt{\frac{2}{\pi}}(M + 1/M)e^{-M^2/2} \leqslant Me^{-M^2/2}$$

because the fact $M \geqslant 2$ implies that

$$\sqrt{\frac{2}{\pi}}(M + 1/M) \leqslant \frac{4}{5}(M + 1/M) \leqslant M.$$

□

8.1.5 Applying the useful lemma

In (8.1) we estimate the second summand by applying the useful lemma with $X_u = y_u$, $Y_u = y_u - x_u$. We get

$$-\mathbb{E}\left[\sum_{\{u,v\}\in E} A_{uv}((y_u(y_v - x_v) + y_v(y_u - x_u))\right] \geqslant -2\sqrt{Me^{-M^2/2}}(\Gamma_{\max} - \Gamma_{\min}).$$

The third summand in (8.1) we estimate by applying the useful lemma with $X_u = y_u - x_u$, $Y_u = -(y_u - x_u)$. We get

$$\mathbb{E}\left[\sum_{\{u,v\}\in E} A_{uv}(y_u - x_u)(y_v - x_v)\right] \geqslant -Me^{-M^2/2}(\Gamma_{\max} - \Gamma_{\min}).$$

Altogether,

$$\sum_{\{u,v\}\in E} A_{uv}\mathbb{E}[x_u x_v] \geqslant \Gamma_{\max} - \left(2\sqrt{Me^{-M^2/2}} + Me^{-M^2/2}\right)(\Gamma_{\max} - \Gamma_{\min}).$$

8.1.6 Connection to the theta number

The connection to the theta number comes in the following lemma.

Lemma 8.1.4.

$$\frac{\Gamma_{\max} - \Gamma_{\min}}{\Gamma_{\max}} \leqslant \vartheta(\overline{G}).$$

Proof. Exercise 8.2. □

In particular, we have

$$M \geqslant \tilde{M} = 3\sqrt{1 + \ln((\Gamma_{\max} - \Gamma_{\min})/\Gamma_{\max})}.$$

Furthermore ($\ln x \leqslant x - 1$),

$$\tilde{M} \leqslant 3\sqrt{(\Gamma_{\max} - \Gamma_{\min})/\Gamma_{\max}}$$

From this it follows that

$$Me^{-M^2/2} \leqslant \tilde{M}e^{-\tilde{M}^2/2} \leqslant \frac{1}{10}\left(\frac{\Gamma_{\max}}{\Gamma_{\max} - \Gamma_{\min}}\right)^2.$$

So,

$$\sum_{\{u,v\}\in E} A_{uv}\mathbb{E}[x_u x_v] \geqslant \Gamma_{\max} - \frac{2}{\sqrt{10}}\Gamma_{\max} - \frac{1}{10}\frac{\Gamma_{\max}}{\Gamma_{\max} - \Gamma_{\min}}\Gamma_{\max},$$

since $\Gamma_{\max} - \Gamma_{\min} \geqslant \Gamma_{\max}$ this leads to

$$\sum_{\{u,v\}\in E} A_{uv}\mathbb{E}[x_u x_v] \geqslant \frac{1}{4}\Gamma_{\max}.$$

Finally we can put everything together: There is a positive constant C (which is not difficult to estimate) so that

$$\Delta \geqslant \frac{1}{M^2} \sum_{\{u,v\} \in E} A_{uv} \mathbb{E}[x_u x_v] \geqslant \frac{1}{C \ln \vartheta(\overline{G})} \Gamma_{\max},$$

which finishes the proof of Theorem 8.1.1.

8.2 Higher rank Grothendieck inequality

Now we model finding ground states in the r-vector model. Given positive integers m, n, r and a matrix $A = (A_{ij}) \in \mathbb{R}^{m \times n}$, the *Grothendieck problem with rank-r-constraint* is defined as

$$\mathrm{SDP}_r(A) = \max\left\{ \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}\ x_i \cdot y_j : x_1, \ldots, x_m \in S^{r-1}, y_1, \ldots, y_n \in S^{r-1} \right\},$$

where $S^{r-1} = \{x \in \mathbb{R}^r : x \cdot x = 1\}$ is the unit sphere; the inner product matrix of the vectors $x_1, \ldots, x_m, y_1, \ldots, y_n$ has rank at most r. When $r = 1$, then $\mathrm{SDP}_1(A) = \|A\|_{\infty \to 1}$ because $S^0 = \{-1, +1\}$.

When r is a constant that does not depend on the matrix size m, n there is no polynomial-time algorithm known which solves SDP_r. However, it is not known if the problem SDP_r is NP-hard when $r \geqslant 2$. On the other hand the *semidefinite relaxation* of $\mathrm{SDP}_r(A)$ defined by

$$\mathrm{SDP}_{m+n}(A) = \max\left\{ \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}\ u_i \cdot v_j : u_1, \ldots, u_m, v_1, \ldots, v_n \in S^{m+n-1} \right\}$$

can be computed in polynomial time using semidefinite programming.

Theorem 8.2.1. *For all matrices $A \in \mathbb{R}^{m \times n}$ we have*

$$\mathrm{SDP}_r(A) \leqslant \mathrm{SDP}_{m+n}(A) \leqslant \frac{1}{2\gamma(r) - 1} \mathrm{SDP}_r(A),$$

where

$$\gamma(r) = \frac{2}{r} \left(\frac{\Gamma((r+1)/2)}{\Gamma(r/2)} \right)^2,$$

and where Γ is the usual Gamma function, which is the *extension of the factorial function.*

The first three values of $\frac{1}{2\gamma(r)-1}$ are:

$$\begin{aligned}
\frac{1}{2\gamma(1) - 1} &= \frac{1}{4/\pi - 1} = 3.65979\ldots, \\
\frac{1}{2\gamma(2) - 1} &= \frac{1}{\pi/2 - 1} = 1.75193\ldots, \\
\frac{1}{2\gamma(3) - 1} &= \frac{1}{16/(3\pi) - 1} = 1.43337\ldots
\end{aligned}$$

For $r \to \infty$ the values $\frac{1}{2\gamma(r)-1}$ converge to 1. In particular, the proof of the theorem gives another proof of the original Grothendieck's inequality albeit with a worse constant $K_G \leqslant \frac{1}{4/\pi-1}$.

8.2.1 Randomized rounding by projecting

The approximation algorithm which we use to prove the theorem is the following three-step process.

1. By solving $\mathrm{SDP}_{m+n}(A)$ we obtain the vectors $u_1, \ldots, u_m, v_1, \ldots, v_n \in S^{m+n-1}$.
2. Choose $Z = (Z_{ij}) \in \mathbb{R}^{r\times(m+n)}$ so that every matrix entry Z_{ij} is distributed independently according to the standard normal distribution with mean 0 and variance 1: $Z_{ij} \sim N(0,1)$.
3. Project $x_i = Zu_i/\|Zu_i\| \in S^{r-1}$ with $i = 1, \ldots, m$, and $y_j = Zv_j/\|Zv_j\| \in S^{r-1}$ with $j = 1, \ldots, n$.

8.2.2 Extension of Grothendieck's identity

The quality of the feasible solution $x_1, \ldots, x_m, y_1, \ldots, y_n$ for SDP_r is measured by the expectation

$$\mathrm{SDP}_r(A) \geqslant \mathbb{E}\Big[\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}\, x_i \cdot y_j\Big].$$

Lemma 8.2.2. *Let u, v be unit vectors in $\mathbb{R}^{m+n}$ and let $Z \in \mathbb{R}^{r\times(m+n)}$ be a random matrix whose entries are distributed independently according to the standard normal distribution with mean 0 and variance 1. Then,*

$$\mathbb{E}\left[\frac{Zu}{\|Zu\|}\cdot\frac{Zv}{\|Zv\|}\right] = \frac{2}{r}\left(\frac{\Gamma((r+1)/2)}{\Gamma(r/2)}\right)^2 \sum_{k=0}^{\infty} \frac{(1\cdot 3\cdots(2k-1))^2}{(2\cdot 4\cdots 2k)((r+2)\cdot(r+4)\cdots(r+2k))}(u\cdot v)^{2k+1}.$$

The case $r = 1$ specializes to Grothendieck's identity from the previous chapter:

$$\begin{aligned}\mathbb{E}[\mathrm{sign}(Zu)\mathrm{sign}(Zv)] &= \frac{2}{\pi}\arcsin(u\cdot v)\\ &= \frac{2}{\pi}\left(u\cdot v + \left(\frac{1}{2}\right)\frac{(u\cdot v)^3}{3} + \left(\frac{1\cdot 3}{2\cdot 4}\right)\frac{(u\cdot v)^5}{5} + \cdots\right).\end{aligned}$$

The proof of Lemma 8.2.2 requires quite some integration. The computation starts of by

$$\mathbb{E}\left[\frac{Zu}{\|Zu\|}\cdot\frac{Zv}{\|Zv\|}\right] = (2\pi\sqrt{1-t^2})^{-r}\int_{\mathbb{R}^r}\int_{\mathbb{R}^r}\frac{x}{\|x\|}\cdot\frac{y}{\|y\|}\exp\left(-\frac{x\cdot x - 2tx\cdot y + y\cdot y}{2(1-t^2)}\right)dxdy,$$

where $t = u \cdot v$. We will omit the tedious calculation here. For those who cannot resist a definite integral (like G.H. Hardy): it can be found in [4].

The only three facts which will be important is that the power series expansion

$$\mathbb{E}\left[\frac{Zu}{\|Zu\|} \cdot \frac{Zv}{\|Zv\|}\right] = \sum_{k=0}^{\infty} f_{2k+1}(u \cdot v)^{2k+1}$$

has the following three properties:

1. the leading coefficient f_1 equals $\gamma(r)$
2. all coefficients f_{2k+1} are nonnegative
3. $\sum_{k=0}^{\infty} f_{2k+1} = 1$.

8.2.3 Proof of the theorem

Now we have

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}\, x_i \cdot y_j\right] &= \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}\mathbb{E}\left[\frac{Zu_i}{\|Zu_i\|} \cdot \frac{Zv_j}{\|Zv_j\|}\right] \\ &= f_1 \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij} u_i \cdot v_j + \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij} \sum_{k=1}^{\infty} f_{2k+1}(u_i \cdot v_j)^{2k+1}.\end{aligned}$$

The first summand equals $f_1 \mathrm{SDP}_{m+n}(A)$. The second summand is bounded *in absolute value* by $(1 - f_1)\mathrm{SDP}_{m+n}(A)$ as you will prove in Exercise 8.1 (d).

Thus for the second sum we have

$$\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij} \sum_{k=1}^{\infty} f_{2k+1}(u_i \cdot v_j)^{2k+1} \geqslant (f_1 - 1)\mathrm{SDP}_{m+n}(A),$$

which finishes the proof.

8.3 Further reading

Section 8.1: The result is from [1] and the presentation of the proof is closely following Chapter 10 of the book *Approximation Algorithms and Semidefinite Programming* of Gärtner and Matoušek, which mostly follows K. Makarychev's thesis.

Section 8.2: The proof is from [3] and it follows the idea of Alon and Naor [2, Section 4] which in turn relies on ideas of Rietz.

More on the definite integral: When working with power series expansions it is sometimes useful to use hypergeometric functions for this. For instance we

have

$$\sum_{k=0}^{\infty} \frac{(1 \cdot 3 \cdots (2k-1))^2}{(2 \cdot 4 \cdots 2k)((r+2) \cdot (r+4) \cdots (r+2k))} (u \cdot v)^{2k+1}$$
$$= (u \cdot v)\, {}_2F_1 \left(\begin{matrix} 1/2, 1/2 \\ r/2+1 \end{matrix} ; (u \cdot v)^2 \right),$$

where ${}_2F_1$ is a hypergeometric function. Hypergeometric functions are a classical subject in mathematics. In fact, many (all?) functions you know, are hypergeometric functions. However the topic of hypergeometric functions seems somehow to be too classical for many modern universities.
In case you want to know more about them: The book "A=B" by Petkovsek, Wilf and Zeilberger

`http://www.math.upenn.edu/~wilf/AeqB.html`

is a good start.

8.4 Exercises

8.1** (a) Why does Theorem 8.1.1 give a proof of the original Grothendieck inequality? Which explicit upper bound for K_G does it provide? (Determine a concrete number.)

(b) Show that $\mathbb{E}[y_u y_v] = f_u \cdot f_v$ holds.

(c) Prove that the matrix

$$\big(\mathbb{E}[X_u X_v]\big)_{u,v \in V}$$

is positive semidefinite.

(d) Show that

$$\left| \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} \sum_{k=1}^{\infty} f_{2k+1} (u_i \cdot v_j)^{2k+1} \right| \leqslant (1 - f_1) \mathrm{SDP}_{m+n}(A).$$

8.2 Let $G = (V, E)$ be a graph. A *vector k-coloring* of G is a collection of unit vectors $f_u \in \mathbb{R}^V$ so that

$$f_u \cdot f_v = -\frac{1}{k-1} \quad \text{if } \{u, v\} \in E.$$

(a) Show that if G is colorable with k colors, then it also has a vector k-coloring.

(b) Find a connection between vector k-colorings and the theta number.

(c) Prove Lemma 8.1.4.

BIBLIOGRAPHY

[1] N. Alon, K. Makarychev, Y. Makarychev, A. Naor, *Quadratic forms on graphs*, Invent. Math. **163** (2006), 499–522.

[2] N. Alon, A. Naor, *Approximating the cut-norm via Grothendieck's inequality*, SIAM J. Comp. **35** (2006), 787–803.

[3] J. Briët, F.M. de Oliveira Filho, F. Vallentin, *The Grothendieck problem with rank constraint*, pp. 111–113 in Proceedings of the 19th Symposium on Mathematical Theory of Networks and Systems, 2010

[4] J. Briët, F.M. de Oliveira Filho, F. Vallentin, *Grothendieck inequalities for semidefinite programs with rank constraint*, arXiv:1011.1754v1 [math.OC]

`http://arxiv.org/abs/1011.1754`

Part III

Applications in geometry

CHAPTER 9

OPTIMIZING WITH ELLIPSOIDS AND DETERMINANTS

What is an ellipsoid?

There are two convenient ways to represent an ellipsoid.

1. We can define ellipsoids explicitly as the image of the unit ball under an invertible affine transformation

$$\mathcal{E}_{A,c} = \{Ax + c : x \in B\}, \quad \text{where} \quad B = \{x \in \mathbb{R}^n : \|x\| \leqslant 1\},$$

is the unit ball, where $A \in \mathbb{R}^{n\times n}$ is an invertible matrix, and where $c \in \mathbb{R}^n$ is a translation vector.

From linear algebra it is known that every invertible matrix A has a factorization of the form $A = XP$ where $X \in \mathcal{S}^n_{\succ 0}$ is a positive definite matrix and $P \in \mathcal{O}(n)$ is an orthogonal matrix. So we may assume in the following that the matrix A which defines the ellipsoid $\mathcal{E}_{A,c}$ is a positive definite matrix.

In fact one can find this factorization (also called polar factorization) from the singular value decomposition of A

$$A = U^\mathsf{T}\Sigma V, \quad U, V \in \mathcal{O}(n),\ \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n),$$

where $\sigma_i \geqslant 0$ are the singular values of A. Then,

$$A = XP \quad \text{with} \quad X = U^\mathsf{T}\Sigma U,\ P = U^\mathsf{T}V.$$

The singular values of A are at the same time the lengths of the semiaxis of the ellipsoid $\mathcal{E}_{A,c}$.

The volume of the ellipsoid equals

$$\operatorname{vol} \mathcal{E}_{A,c} = \det A \operatorname{vol} B \quad \text{where} \quad \operatorname{vol} B = \frac{\pi^{n/2}}{\Gamma(n/2+1)}.$$

2. We can define ellipsoids implicitly by a strictly convex quadratic inequality: Let $A \in \mathcal{S}^n_{\succ 0}$ be a positive definite matrix, then

$$\begin{aligned} \mathcal{E}_{A,c} &= \{Ax + c : \|x\| \leqslant 1\} \\ &= \{x + c : \|A^{-1}x\| \leqslant 1\} \\ &= \left\{x \in \mathbb{R}^n : (x-c)^{\mathsf{T}} A^{-2}(x-c) \leqslant 1\right\}. \end{aligned}$$

Ellipsoids are important geometric objects partially due to their simple descriptions. They can be used for instance to approximate other more complicated convex sets. In this lecture we will use ellipsoids to approximate polytopes. In particular we will answer the questions:

- Inner approximation: How can we determine an ellipsoid contained in a polytope which has largest volume?
- Outer approximation: How can we determine an ellipsoid containing a polytope which has smallest volume?
- Can we estimate the quality of this inner and of this outer approximation?

9.1 Determinant maximization problems

To be able to maximize the volume of ellipsoids we want to maximize the determinant of positive definite matrices. In the next section we will see that the logarithm of the determinant is a concave function so that determinant maximization can be dealt with tools from convex optimization.

In fact one can reformulate determinant maximization problems as semidefinite programs but we will not do this here; dealing directly with determinant maximization problem is generally easier and more efficient.

Here we give the primal-dual pair of a determinant maximization problem together with the corresponding duality theory. All in all it is very similar to semidefinite programming, only the objective function is not linear.

The *primal determinant maximization problem* is defined as

$$\begin{aligned} \sup \quad & n + \langle C, X \rangle + \ln \det X \\ & X \in \mathcal{S}^n_{\succ 0} \\ & \langle A_j, X \rangle = b_j, \quad j = 1, \ldots, m, \end{aligned}$$

and its *dual* is

$$\begin{array}{ll} \inf & b^{\mathsf{T}} y - \ln\det\left(\sum_{j=1}^{m} y_j A_j - C\right) \\ & y \in \mathbb{R}^m \\ & \sum_{j=1}^{m} y_j A_j - C \in \mathcal{S}^n_{\succ 0} \end{array}$$

We have statements for weak duality and strong duality, which are very similar to Theorem 3.4.1. A slight difference is complementary slackness and the optimality criterion. The optimality criterion says: Suppose that X is feasible for the primal, y is feasible for the dual and the equality

$$X\left(\sum_{j=1}^{m} y_j A_j - C\right) = I_n$$

holds. Then X and y are both optimal (see Exercise 9.1 (a)).

In many cases it is useful to combine determinant maximization problems with linear conic programs. In these cases we want to work with the following primal-dual pair.

Primal:

$$\begin{array}{ll} \sup & n + \langle C, X\rangle + \ln\det X + c^{\mathsf{T}} x \\ & X \in \mathcal{S}^n_{\succ 0} \\ & x \in K \\ & \langle A_j, X\rangle + a_j^{\mathsf{T}} x = b_j, \quad j = 1, \ldots, m \end{array}$$

Dual:

$$\begin{array}{ll} \inf & b^{\mathsf{T}} y - \ln\det\left(\sum_{i=1}^{m} y_j A_j - C\right) \\ & y \in \mathbb{R}^m \\ & \sum_{j=1}^{m} y_j A_j - C \in \mathcal{S}^n_{\succ 0} \\ & \sum_{j=1}^{m} y_j a_j - c \in K^* \end{array}$$

9.2 Convex spectral functions

In this section we shall prove that the function $X \mapsto -\ln\det X$ is a (strictly) convex function. For this we will give two proofs. One simple adhoc proof and one which is conceptual. The second proof will characterize all convex functions on symmetric matrices which only depend on the eigenvalues.

9.2.1 Minkowski's determinant inequality

Theorem 9.2.1. *The function*

$$F : \mathcal{S}^n_{>0} \to \mathbb{R}, \quad X \mapsto -\ln\det X$$

is a strictly convex function on the set of positive definite matrices.

Proof. It suffices to show that the function $X \mapsto -\ln\det X$ is strictly convex on any line segment

$$[X, Y] = \{tX + (1-t)Y : t \in [0,1], X \neq Y\}$$

in $\mathcal{S}^n_{>0}$. Therefore, we compute the second derivative of the one-dimensional function $f(t) = -\ln\det(tX + (1-t)Y)$ and see that it is always strictly positive: From linear algebra we know that there is a matrix T with determinant 1 whose inverse simultaneously diagonalizes X and Y. Hence,

$$X = T^{\mathsf{T}} \operatorname{diag}(x_1, \ldots, x_n)T \quad \text{and} \quad Y = T^{\mathsf{T}} \operatorname{diag}(y_1, \ldots, y_n)T$$

and

$$\begin{aligned}
f(t) &= -\ln(y_1 + t(x_1 - y_1)) - \cdots - \ln(y_n + t(x_n - y_n)),\\
\frac{\partial f}{\partial t}(t) &= -\frac{x_1 - y_1}{y_1 + t(x_1 - y_1)} - \cdots - \frac{x_n - y_n}{y_n + t(x_n - y_n)},\\
\frac{\partial^2 f}{\partial t^2}(t) &= \left(\frac{x_1 - y_1}{y_1 + t(x_1 - y_1)}\right)^2 + \cdots + \left(\frac{x_n - y_n}{y_n + t(x_n - y_n)}\right)^2 > 0. \quad \square
\end{aligned}$$

With the same argument we can derive *Minkowski's determinental inequality*:

$$(\det(X + Y))^{1/n} \geqslant (\det X)^{1/n} + (\det Y)^{1/n}$$

which holds for all $X, Y \in \mathcal{S}^n_{>0}$.

Geometrically, this means that in the cone of positive semidefinite matrices, the set of matrices having determinant greater or equal than a given constant is a convex set.

9.2.2 Davis' characterization of convex spectral functions

The function $F(X) = -\ln\det X$ is an example of a convex spectral function.

Definition 9.2.2. *A* convex spectral function *is a convex function*

$$F : \mathcal{S}^n \to \mathbb{R} \cup \{\infty\}$$

where $F(X)$ only depends on the spectrum (the collection of the eigenvalues $\lambda_1(X)$, $\ldots$, $\lambda_n(X)$) of the matrix X. In other words, by the spectral theorem,

$$F(X) = f(AXA^{\mathsf{T}}) \quad \textit{for all } A \in \mathcal{O}(n).$$

Hence, there is a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ which defines F by the following equation

$$F(X) = f(\lambda_1(X), \ldots, \lambda_n(X)).$$

Note that this implies that the function f is *symmetric*, i.e. its value stays the same if we permute its n arguments; it is invariant under permutation of the variables.

In our example

$$f(\lambda_1, \ldots, \lambda_n) = \begin{cases} -\ln \prod_{i=1}^{n} \lambda_i & \text{if all } \lambda_i > 0, \\ \infty & \text{otherwise.} \end{cases}$$

The following theorem is due to Davis (1957). It gives a complete characterization of convex spectral functions.

Theorem 9.2.3. *A function $F : \mathcal{S}^n \to \mathbb{R} \cup \{\infty\}$ is a convex spectral function if and only if the function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ defined by*

$$F(X) = f(\lambda_1(X), \ldots, \lambda_n(X))$$

is symmetric and convex. In particular,

$$F(X) = \max_{A \in \mathcal{O}(n)} f((AXA^{\mathsf{T}})_{11}, \ldots, (AXA^{\mathsf{T}})_{nn})$$

holds.

Proof. One implication follows without any work.

Let F be a convex spectral function. Then f is symmetric by definition. It is convex since F and f "coincide" on diagonal matrices. Let $\Lambda = (\lambda_1, \ldots, \lambda_n)$, $M = (\mu_1, \ldots, \mu_n)$ and $t \in [0, 1]$ be given. Then

$$\begin{aligned} f(t\Lambda + (1-t)M) &= F(\operatorname{diag}(t\Lambda + (1-t)M)) \\ &= F(t \operatorname{diag}(\Lambda) + (1-t) \operatorname{diag}(M)) \\ &\leqslant tF(\operatorname{diag}(\Lambda)) + (1-t)F(\operatorname{diag}(M)) \\ &= tf(\Lambda) + (1-t)f(M). \end{aligned}$$

The proof of the other implication is more interesting. It is an application of Birkhoff's theorem (cf. Chapter 1.7.3), see also the geometric interpretation at the end of this section.

If we show that

$$F(X) = \max_{A \in \mathcal{O}(n)} f((AXA^{\mathsf{T}})_{11}, \ldots, (AXA^{\mathsf{T}})_{nn})$$

holds, then it follows that F is convex because it is a maximum of a family of convex functions.

For this consider the spectral decomposition of X

$$X = \sum_{j=1}^{n} \lambda_j u_j u_j^{\mathsf{T}}.$$

with orthonormal basis $u_1, \ldots, u_n$. If we assemble these vectors as row vectors in the orthogonal matrix A we see that

$$f((AXA^{\mathsf{T}})_{11}, \ldots, (AXA^{\mathsf{T}})_{nn}) = f(\lambda_1, \ldots, \lambda_n) = F(X)$$

holds. Thus,

$$F(X) \leqslant \max_{A \in \mathcal{O}(n)} f((AXA^{\mathsf{T}})_{11}, \ldots, (AXA^{\mathsf{T}})_{nn}).$$

The other inequality. For $A \in \mathcal{O}(n)$ define $Y = AXA^{\mathsf{T}}$. Then,

$$Y_{ii} = e_i^{\mathsf{T}} Y e_i = e_i^{\mathsf{T}} \left(\sum_{j=1}^{n} \lambda_j A u_j (A u_j)^{\mathsf{T}} \right) e_i = \sum_{j=1}^{n} \lambda_j ((Au_j)^{\mathsf{T}} e_i)^2.$$

Here is the trick: The matrix

$$S = (S_{ij})_{1 \leqslant i,j \leqslant n} \quad \text{with} \quad S_{ij} = ((Au_j)^{\mathsf{T}} e_i)^2$$

is doubly stochastic (why?). So by Birkhoff's theorem S is a convex combination of permutation matrices P^σ.

$$S = \sum_{\sigma \in S_n} \mu_\sigma P^\sigma, \quad \text{where} \quad \mu_\sigma \geqslant 0, \sum_{\sigma \in S_n} \mu_\sigma = 1.$$

Hence by the convexity and the symmetry of f, we have

$$\begin{aligned}
f(Y_{11}, \ldots, Y_{nn}) &= f\left(\sum_{\sigma \in S_n} \mu_\sigma \sum_{j=1}^{n} P_{1j}^\sigma \lambda_j, \ldots, \sum_{\sigma \in S_n} \mu_\sigma \sum_{j=1}^{n} P_{nj}^\sigma \lambda_j \right) \\
&\leqslant \sum_{\sigma \in S_n} \mu_\sigma f\left(\sum_{j=1}^{n} P_{1j}^\sigma \lambda_j, \ldots, \sum_{j=1}^{n} P_{nj}^\sigma \lambda_j \right) \\
&= \sum_{\sigma \in S_n} \mu_\sigma f(\lambda_{\sigma^{-1}(1)}, \ldots, \lambda_{\sigma^{-1}(n)}) \\
&= \sum_{\sigma \in S_n} \mu_\sigma f(\lambda_1, \ldots, \lambda_n) \\
&= f(\lambda_1, \ldots, \lambda_n) \\
&= F(X).
\end{aligned}$$

Hence, for all $A \in \mathcal{O}(n)$

$$f((AXA^{\mathsf{T}})_{11}, \ldots, (AXA^{\mathsf{T}})_{nn}) \leqslant F(X),$$

and the theorem is proved. □

We conclude this excursion with a geometric observation which unifies some of our previous considerations. If we project matrices in the Schur-Horn orbitope of X which is defined by

$$\operatorname{conv}\{AXA^{\mathsf{T}} : A \in \mathcal{O}(n)\}$$

on the diagonal elements, then we get the permutahedron given by the eigenvalues $\lambda_1, \ldots, \lambda_n$ of X which is defined as

$$\operatorname{conv}\{(\lambda_{\sigma(1)}, \ldots, \lambda_{\sigma(n)}) : \sigma \in S_n\}.$$

From this the Hoffman-Wielandt inequality and the characterization of Davis follow.

Another side remark: Davis' characterization together with Fan's theorem (Theorem 2.2.2) can be used to determine an explicit linear matrix inequality modeling the condition $F(X) \leqslant t$ for many functions F. See Ben-Tal, Nemirovski [3][Proposition 4.2.1] for the complete statement. A similar argument also works for functions depending only on singular values.

9.3 Approximating polytopes by ellipsoids

Now we are ready to describe how ellipsoids can be used to approximate polytopes.

Recall that one can represent a polytope in two ways. Either as a convex hull of finitely many points

$$P = \operatorname{conv}\{x_1, \ldots, x_N\} \in \mathbb{R}^n,$$

or as a bounded intersection of finitely many halfspaces

$$P = \{x \in \mathbb{R}^n : a_1^\mathsf{T} x \leqslant b_1, \ldots, a_m^\mathsf{T} x \leqslant b_m\}.$$

The first representation, also called the $\mathcal{V}$-representation, is an explicit parameterization whereas the second one, also called the $\mathcal{H}$-representation, is implicit. In general it is computationally demanding to transform one representation into the other.

9.3.1 Inner approximation

To formulate the condition that an ellipsoid $\mathcal{E}_{A,c}$ is contained in a polytope P we will use the explicit representation of the ellipsoid and the implicit representation of the polytope.

Proposition 9.3.1. *The ellipsoid*

$$\mathcal{E}_{A,c} = \{Ax + c : \|x\| \leqslant 1\}$$

is contained in the polytope

$$P = \{x \in \mathbb{R}^n : a_1^\mathsf{T} x \leqslant b_1, \ldots, a_m^\mathsf{T} x \leqslant b_m\}$$

if and only if the inequality

$$\|Aa_i\| \leqslant b_i - a_i^\mathsf{T} c$$

holds for all $i = 1, \ldots, m$.

Proof. We have

$$\begin{aligned} & a_i^{\mathsf{T}}(Ax+c) \leqslant b_i \quad \forall x \in \mathbb{R}^n \ : \ \|x\| \leqslant 1 \\ \iff & \max_{x:\|x\|\leqslant 1} (Aa_i)^{\mathsf{T}} x \leqslant b_i - a_i^{\mathsf{T}} c \\ \iff & \|Aa_i\| \leqslant b_i - a_i^{\mathsf{T}} c, \end{aligned}$$

because by the Cauchy-Schwarz inequality $\max_{x:\|x\|\leqslant 1}(Aa_i)^{\mathsf{T}}x = \|Aa_i\|$. □

The inequality $\|Aa_i\| \leqslant b_i - a_i^{\mathsf{T}} c$ can be directly modeled by a second order cone programming constraint or, using the Schur complement, by a semidefinite constraint.

9.3.2 Outer approximation

To formulate the condition that an ellipsoid $\mathcal{E}_{A,c}$ is containing a polytope P we will use the implicit representation of the ellipsoid and the explicit representation of the polytope.

Proposition 9.3.2. *The ellipsoid*

$$\mathcal{E}_{A^{-1/2},c} = \{x \in \mathbb{R}^n : (x-c)^{\mathsf{T}} A (x-c) \leqslant 1\}$$

contains the polytope

$$P = \operatorname{conv}\{x_1, \ldots, x_N\}$$

if and only if the matrix

$$\begin{pmatrix} s & d^{\mathsf{T}} \\ d & A \end{pmatrix}$$

is positive semidefinite with $Ac = d$ *and the inequality*

$$x_i^{\mathsf{T}} A x_i - 2 x_i^{\mathsf{T}} d + s \leqslant 1$$

holds for all $i = 1, \ldots, N$.

Proof. The point x_i lies in the ellipsoid $\mathcal{E}_{A^{-1/2},c}$ if and only if

$$\begin{aligned} & (x_i - c)^{\mathsf{T}} A (x_i - c) \leqslant 1 \\ \iff & x_i^{\mathsf{T}} A x_i - 2 x_i^{\mathsf{T}} A c + c^{\mathsf{T}} A c \leqslant 1 \\ \iff & x_i^{\mathsf{T}} A x_i - 2 x_i^{\mathsf{T}} d + d^{\mathsf{T}} A^{-1} d \leqslant 1 \\ \iff & x_i^{\mathsf{T}} A x_i - 2 x_i^{\mathsf{T}} d + s \leqslant 1, \quad \text{where } s \geqslant d^{\mathsf{T}} A^{-1} d. \end{aligned}$$

Because the matrix A is positive definite we can express $s \geqslant d^{\mathsf{T}} A^{-1} d$ using the Schur complement as

$$\begin{pmatrix} s & d^{\mathsf{T}} \\ d & A \end{pmatrix} \succeq 0.$$ □

The constraint

$$x_i^\mathsf{T} A x_i - 2x_i^\mathsf{T} d + s \leqslant 1$$

can be expressed by the linear matrix inequality

$$\left\langle \begin{pmatrix} 1 & x_i^\mathsf{T} \\ x_i & x_i x_i^\mathsf{T} \end{pmatrix}, \begin{pmatrix} s & d^\mathsf{T} \\ d & A \end{pmatrix} \right\rangle \leqslant 1.$$

9.4 The Löwner-John ellipsoids

Using Proposition 9.3.1 and Proposition 9.3.2 one can find the ellipsoid of largest volume contained in a polytope P as well as the ellipsoid of smallest volume containing P by solving determinant maximization problems. In both cases one maximizes the logarithm of the determinant of A. Because the logarithm of the determinant is a *strictly* concave function both optimization problems have a unique solution. The ellipsoids are called the *Löwner-John ellipsoids*[1] of P. Notation: $\mathcal{E}_{in}(P)$ for the ellipsoid giving the optimal inner approximation of P and $\mathcal{E}_{out}(P)$ for the ellipsoid giving the optimal outer approximation of P.

The following theorem can be traced back to John (1948). Historically, it is considered to be one of the first theorems involving an optimality condition for nonlinear optimization.

Theorem 9.4.1. *(a) Let P be a polytope. The Löwner-John ellipsoid $\mathcal{E}_{out}(P)$ is equal to the unit ball if and only if P is contained in the unit ball and there is are positive numbers $\lambda_1, \ldots, \lambda_N$ and vertices $x_1, \ldots, x_N$ of P having unit length so that*

$$\sum_{i=1}^{N} \lambda_i x_i = 0 \quad \textit{and} \quad \sum_{i=1}^{N} \lambda_i x_i x_i^\mathsf{T} = I_n$$

holds.

(b) Let P be a polytope. The Löwner-John ellipsoid $\mathcal{E}_{in}(P)$ is equal to the unit ball if and only if P is containing the unit ball and if there are unit vectors $x_1, \ldots, x_N$ on the boundary of P and there are positive numbers $\lambda_1, \ldots, \lambda_N$ so that

$$\sum_{i=1}^{N} \lambda_i x_i = 0 \quad \textit{and} \quad \sum_{i=1}^{N} \lambda_i x_i x_i^\mathsf{T} = I_n$$

holds.

Before we give the proof we comment on the optimality conditions. The first equality makes sure that not all the vectors $x_1, \ldots x_N$ lie on one side of the sphere. The second equality shows that the vectors behave similar to an

[1] In the literature the terminology seems to differ from author to author.

orthonormal basis in the sense that we can compute the inner product of two vectors x and y by

$$x^{\mathsf{T}} y = \sum_{i=1}^{N} \lambda_i (x_i^{\mathsf{T}} x)(x_i^{\mathsf{T}} y).$$

Proof. Statement (a) follows from the optimality conditions of the underlying determinant maximization problem. See Exercise 9.1 (b).

Statement (b) follows from (a) by polarity:

Let $C \subseteq \mathbb{R}^n$ be a convex body. Its *polar body* C^* is

$$C^* = \{x \in \mathbb{R}^n : x^{\mathsf{T}} y \leqslant 1 \text{ for all } y \in C\}.$$

The unit ball is self-polar, $B^* = B$. Furthermore, for every ellipsoid $\mathcal{E}$ we have

$$\operatorname{vol} \mathcal{E} \operatorname{vol} \mathcal{E}^* \geqslant (\operatorname{vol} B)^2,$$

because direct verification yields

$$(\mathcal{E}_{A,c})^* = \mathcal{E}_{A',c'} \quad \text{with} \quad A' = \left(\frac{A^{\mathsf{T}} A}{(1 + \frac{1}{4} c^{\mathsf{T}} (A^{\mathsf{T}} A)^{-1} c)} \right)^{-1/2}, \; c' = -\frac{1}{2} (A^{\mathsf{T}} A)^{-1} c,$$

and so

$$\det A \det A' \geqslant 1.$$

Let P be a polytope and assume that $\mathcal{E}_{in}(P) = B$. We will now prove by contradiction that $B = \mathcal{E}_{out}(P^*)$. For suppose not. Then the volume of $\mathcal{E}_{out}(P^*)$ is strictly smaller than the volume of B since $P^* \subseteq B$. However, by taking the polar again we have

$$\mathcal{E}_{out}(P^*)^* \subseteq P,$$

and $\operatorname{vol} \mathcal{E}_{out}(P^*)^* > \operatorname{vol}(B)$ a contradiction. So by (a) we have for vertices $x_1, \ldots, x_N$ of P, which are of unit length, the conditions

$$\sum_{i=1}^{N} \lambda_i x_i = 0 \quad \text{and} \quad \sum_{i=1}^{N} \lambda_i x_i x_i^{\mathsf{T}} = I_n,$$

for positive $\lambda_1, \ldots, \lambda_N$. Then the unit vectors x_i also lie on the boundary of the polytope P because

$$P = (P^*)^* = \{x \in \mathbb{R}^n : x_i^{\mathsf{T}} x \leqslant 1, i = 1, \ldots, N\}.$$

Now let

$$\mathcal{E}_{in}(P) = \{x \in \mathbb{R}^n : (x - c)^{\mathsf{T}} A^{-2} (x - c) \leqslant 1\}$$

be the optimal inner approximation of P. We want to derive from the optimality conditions that $\det A \leqslant 1$ holds.

First we realize that the second optimality condition implies that the equation $\sum_{i=1}^{N} \lambda_i = n$ holds; simply take the trace.

The points

$$y_i = c + (x_i^{\mathsf{T}} A^2 x_i)^{-1/2} A^2 x_i$$

lie in $\mathcal{E}_{in}(P)$ and so $y_i^{\mathsf{T}} x_i \leqslant 1$ holds because this inequality determines a supporting hyperplanes of P. Then,

$$n \geqslant \sum_{i=1}^{N} \lambda_i y_i^{\mathsf{T}} x_i = \sum_{i=1}^{N} \lambda_i (x_i^{\mathsf{T}} A^2 x_i)^{1/2}$$

where we used the first equality when simplifying $\sum_{i=1}^{N} \lambda_i c^{\mathsf{T}} x_i = 0$. The trace of A can be estimated by using the second equality

$$\langle A, I_n \rangle = \Big\langle A, \sum_{i=1}^{N} \lambda_i x_i x_i^{\mathsf{T}} \Big\rangle = \sum_{i=1}^{N} \lambda_i x_i^{\mathsf{T}} A x_i \leqslant \sum_{i=1}^{N} \lambda_i (x_i^{\mathsf{T}} A^2 x_i)^{1/2} \leqslant n,$$

where we used in the first inequality the spectral factorization of $A = P^{\mathsf{T}} D P$, with orthogonal matrix P and diagonal matrix D, together with the Cauchy-Schwarz inequality

$$(x_i^{\mathsf{T}} P^{\mathsf{T}} D)(P x_i) \leqslant (x_i^{\mathsf{T}} P^{\mathsf{T}} D^2 P x_i)^{1/2} ((P x_i)^{\mathsf{T}} P x_i)^{1/2} = (x_i^{\mathsf{T}} P^{\mathsf{T}} D^2 P x_i)^{1/2}.$$

Now we finish the proof by realizing that ($\ln x \leqslant x - 1$)

$$\ln \det A \leqslant \mathrm{Tr}(A) - n \leqslant 0,$$

and so $\det A \leqslant 1$. □

This optimality condition is helpful in surprisingly many situation. For example one can use them to prove an estimate on the quality of the inner and outer approximation.

Corollary 9.4.2. *Let $P \subseteq \mathbb{R}^n$ be an n-dimensional polytope, then there are invertible affine transformations T_{in} and T_{out} so that*

$$B = T_{in} \mathcal{E}_{in}(P) \subseteq T_{in} P \subseteq n T_{in} \mathcal{E}_{in}(P) = nB$$

and

$$\frac{1}{n} B = \frac{1}{n} T_{out} \mathcal{E}_{out}(P) \subseteq T_{out} P \subseteq T_{out} \mathcal{E}_{out}(P) = B$$

holds.

Proof. We only prove the first statement, the second follows again by polarity.

It is clear that we can map $\mathcal{E}_{in}(P)$ to the unit ball by an invertible affine transformation. So we can use the equations

$$\sum_{i=1}^{N} \lambda_i x_i = 0 \quad \text{and} \quad \sum_{i=1}^{N} \lambda_i x_i x_i^{\mathsf{T}} = I_n$$

to show $T_{in}P \subseteq n\mathcal{E}_{in}(P)$. By taking the trace on both sides of the second equations we also have

$$\sum_{i=1}^{N} \lambda_i = n.$$

The supporting hyperplane through the boundary point x_i of P is orthogonal the unit vector x_i (draw a figure). Hence,

$$B \subseteq P \subseteq Q = \{x \in \mathbb{R}^n : x_i^{\mathsf{T}} x \leqslant 1, i = 1, \ldots, N\}.$$

Let x be in Q, then because $x^{\mathsf{T}} x_i \in [-\|x\|, 1]$ we have

$$\begin{aligned} 0 &\leqslant \sum_{i=1}^{N} \lambda_i (1 - x^{\mathsf{T}} x_i)(\|x\| + x^{\mathsf{T}} x_i) \\ &= \|x\| \sum_{i=1}^{N} \lambda_i + (1 - \|x\|) \sum_{i=1}^{N} \lambda_i x^{\mathsf{T}} x_i - \sum_{i=1}^{N} \lambda_i (x^{\mathsf{T}} x_i)^2 \\ &= \|x\| n + 0 - \|x\|^2, \end{aligned}$$

and so $\|x\| \leqslant n$. □

If P is centrally symmetric, i.e. $P = -P$, then in the above inequalities n can be replaced by $\sqrt{n}$. See Exercise 9.1 (c).

Another nice mathematical application of the uniqueness Löwner-John ellipsoids is the following.

Proposition 9.4.3. *Let P be a polytope and consider the group G of all affine transformations which map P into itself. Then there is an affine transformation T so that TGT^{-1} is a subgroup of the orthogonal group.*

Proof. Since the volume is invariant under affine transformations with determinant equal to 1 or -1 (only those affine transformations can be in G) and since the Löwner-John ellipsoid is the unique maximum volume ellipsoid contained in a polytope we have

$$A\mathcal{E}_{in}(P) = \mathcal{E}_{in}(AP) = \mathcal{E}_{in}(P)$$

for all $A \in G$.

Let T be the affine transformation which maps the Löwner-John ellipsoid $\mathcal{E}_{in}(P)$ to the unit ball B. Then for every $A \in G$

$$TAT^{-1}B = TA\mathcal{E}_{in}(P) = T\mathcal{E}_{in}(P) = B.$$

So TAT^{-1} leaves the unit ball invariant, hence it is an orthogonal transformation. □

9.5 Further reading

Many examples of determinant maximization problems are in Vandenberghe, Boyd and Wu [5]. They treat matrix completion problems, risk-averse linear estimation, experimental design, maximum likelihood estimation of structured covariance matrices, and Gaussian channel capacity. Next to this, they also develop the duality theory and an interior point algorithm for determinant maximization problems.

For more information on convex spectral functions and general eigenvalue optimization problems the survey [4] by Lewis is a good start.

Many more examples of computing ellipsoidal approximations are in the book [3][Chapter 4.9], especially ellipsoidal approximations of unions and intersections of ellipsoids and approximating sums of ellipsoids.

The Löwner-John ellipsoid is an important and useful concept in geometry, optimization, and functional analysis. For instance, Lenstra's polynomial time algorithm for solving integer programs in fixed dimension is based on it (see LNMB course: Integer programming methods).

Another excellent and very elegant source on applications of the Löwner-John ellipsoid in geometry and functional analysis is by Ball [2]. He uses John's optimality criterion to give a reverse isoperimetric inequality (the ratio between surface and volume is maximized by cubes) and to prove Dvoretsky's theorem (high dimensional convex bodies have almost ellipsoidal slices). The proof of the second part of Theorem 9.4.1 (b) is from the beautiful note of Ball [1].

One general strategy when working with convex sets is to find an affine transformation of the convex set so that the unit ball and the convex set are as close as possible. Here the notion of closeness depends of course on the question. In many cases these affine transformations can be found by solving an optimization problem involving positive definite matrices.

9.6 Exercises

9.1** (a) Prove weak duality of determinant maximization problems: Let X be a solution of the primal and let y be a solution of the dual. Then,

$$b^{\mathsf{T}} y - \ln\det\left(\sum_{j=1}^{m} y_j A_j - C\right) - (n + \langle C, X\rangle + \ln\det X) \geqslant 0.$$

Hint: $\ln x \leqslant x - 1$.

(b) Prove Theorem 9.4.1 (a).

(c) Show the strengthening of Corollary 9.4.2

$$B = T_{in}\mathcal{E}_{in}(P) \subseteq T_{in}P \subseteq \sqrt{n}T_{in}\mathcal{E}_{in}(P) = \sqrt{n}B$$

in the case of centrally symmetric polytopes P.

(d) Find a polytope P for which the inclusion

$$B \subseteq P \subseteq nB$$

cannot be improved. Find a centrally symmetric polytope for which

$$B \subseteq P \subseteq \sqrt{n}B$$

cannot be improved.

9.2 (a) Show that the sum of the largest k eigenvalues of a symmetric matrix is a convex spectral function.

(b) True or false: The second largest eigenvalue of a symmetric matrix is a convex spectral function.

9.3 Compute the gradient of the function

$$F : \mathcal{S}^n_{\succ 0} \to \mathbb{R}, \quad X \mapsto -\ln\det X.$$

BIBLIOGRAPHY

[1] K.M. Ball, *Ellipsoids of maximal volume in convex bodies*, Geometriae Dedicata **41** (1992), 241–250.

[2] K.M. Ball, *An Elementary Introduction to Modern Convex Geometry*, pp. 1–58 in: Flavors of Geometry (S. Levy (ed.)), Cambridge University Press, 1997.

`http://library.msri.org/books/Book31/files/ball.pdf`

[3] A. Ben-Tal, A. Nemirovski, *Lectures on Modern Convex Optimization*, SIAM 2001.

[4] A.S. Lewis, *The mathematics of eigenvalue optimization*, Math. Program., Ser. B **97** (2003), 155–176.

`http://people.orie.cornell.edu/~aslewis/publications/03-mathematics.pdf`

[5] L. Vandenberghe, S. Boyd, S.-P. Wu, *Determinant Maximization with Linear Matrix Inequality Constraints*, SIAM Journal on Matrix Analysis and Applications **19** (1998), 499–533.

`http://www.stanford.edu/~boyd/papers/maxdet.html`

CHAPTER 10

EUCLIDEAN EMBEDDINGS: LOW DIMENSION

In many situations one is interested in finding solutions to semidefinite programs having a small rank. For instance, if the semidefinite program arises as relaxation of a combinatorial optimization problem (like max-cut or max clique), then its rank one solutions correspond to the solutions of the underlying combinatorial problem. Finding an embedding of a weighted graph in the Euclidean space of dimension d, or finding a sum of squares decomposition of a polynomial with d squares, amounts to finding a solution of rank at most d to some semidefinite program. As another example, the minimum dimension of an orthonormal representation of a graph $G = (V, E)$ (introduced in Chapter 6) is the minimum rank of a positive semidefinite matrix X satisfying $X_{ij} = 0$ for all non-edges.

This chapter is organized as follows. First we show some upper bounds on the rank of solutions to semidefinite programs. For this we have to look into the geometry of the faces of the cone of positive semidefinite matrices. Then we discuss several applications: Euclidean embeddings of weighted graphs, hidden convexity results for images of quadratic maps, and the S-lemma which deals with quadratic inequalities. We also discuss complexity issues related to the problem of determining the smallest possible rank of solutions to semidefinite programs.

10.1 Geometry of the positive semidefinite cone

10.1.1 Faces of convex sets

We begin with some preliminary facts about faces of convex sets which we will use to study the faces of the positive semidefinite cone $\mathcal{S}^n_{\succeq 0}$.

Let K be a convex set in $\mathbb{R}^n$. A set $F \subseteq K$ is called a *face* of K if for all $x \in F$ the following holds:

$$x = ty + (1-t)z \text{ with } t \in (0,1),\ y, z \in K \Longrightarrow y, z \in F.$$

Clearly any intersection of faces is again a face. Hence, for $x \in K$, the smallest face containing x is well defined (as the intersection of all the faces of K that contain x), let us denote it by $F_K(x)$.

A point $z \in \mathbb{R}^n$ is called a *perturbation* of $x \in K$ if $x \pm \epsilon z \in K$ for some $\epsilon > 0$; then the whole segment $[x - \epsilon z, x + \epsilon z]$ is contained in the face $F_K(x)$.

Lemma 10.1.1. *Given a convex set K and $x \in K$, let $F_K(x)$ be the smallest face of K containing x. The following properties hold.*

(i) *x belongs to the relative interior of $F_K(x)$.*

(ii) *$F_K(x)$ is the unique face of K containing x in its relative interior.*

Proof. (i) Assume for a contradiction that $x \notin \text{relint } F_K(x)$. Then, by applying the separation theorem from Theorem 1.3.8 (i), there exists a hyperplane

$$H_{c,\gamma} = \{y : c^\mathsf{T} y = \gamma\}$$

separating the two convex sets $\{x\}$ and $F_K(x)$ properly: There exist a nonzero vector $c \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$ such that

$$c^\mathsf{T} x \geqslant \gamma,\ c^\mathsf{T} y \leqslant \gamma\ \forall y \in F_K(x), \text{ and } F_K(x) \nsubseteq H_{c,\gamma}.$$

Then the set $F_K(x) \cap H_{c,\gamma}$ is a face of K, which contains x and is strictly contained in $F_K(x)$ (check it). This contradicts the fact that $F_K(x)$ is the smallest face containing x.

(ii) Let F be a face of K containing x in its relative interior. Then $F_K(x) \subseteq F$. To show the reverse inclusion, pick $y \in F$, $y \neq x$. As x lies in the relative interior of F, Lemma 1.2.1 implies that there exists a point $z \in F$ and a scalar $t \in (0,1)$ such that $x = ty + (1-t)z$. As $F_K(x)$ is a face, we deduce that $y, z \in F_K(x)$. This shows that $F \subseteq F_K(x)$. □

Hence, x lies in the relative interior of K precisely when $F_K(x) = K$ and x is an *extreme point* of K, i.e.,

$$x = ty + (1-t)z \text{ with } y, z \in K \text{ and } t \in (0,1) \Longrightarrow y = z = x,$$

precisely when $F_K(x) = \{x\}$. Recall that if K does not contain a line then it has at least one extreme point.

10.1.2 Faces of the positive semidefinite cone

Here we describe the faces of the positive semidefinite cone $\mathcal{S}^n_{\succeq 0}$. We show that each face of $\mathcal{S}^n_{\succeq 0}$ can be identified to a smaller semidefinite cone $\mathcal{S}^r_{\succeq 0}$ for some $0 \leqslant r \leqslant n$.

Proposition 10.1.2. *Let $A \in \mathcal{S}^n_{\succeq 0}$, $r = \mathrm{rank}(A)$, and let $F(A) = F_{\mathcal{S}^n_{\succeq 0}}(A)$ denote the smallest face of $\mathcal{S}^n_{\succeq 0}$ containing A. Let $u_1, \cdots, u_n$ be an orthonormal set of eigenvectors of A, where $u_1, \cdots, u_r$ correspond to its nonzero eigenvalues, and let U (resp., U_0) be the matrix with columns $u_1, \cdots, u_n$ (resp., $u_1, \cdots, u_r$). The map*

$$\begin{array}{rccc} \phi_A : & \mathcal{S}^r & \to & \mathcal{S}^n \\ & Z & \mapsto & U \begin{pmatrix} Z & 0 \\ 0 & 0 \end{pmatrix} U^\mathsf{T} = U_0 Z U_0^\mathsf{T} \end{array} \tag{10.1}$$

is a rank-preserving isometry, which identifies $F(A)$ and $\mathcal{S}^r_{\succeq 0}$:

$$F(A) = \phi(\mathcal{S}^r_{\succeq 0}) = \left\{ U \begin{pmatrix} Z & 0 \\ 0 & 0 \end{pmatrix} U^\mathsf{T} = U_0 Z U_0^\mathsf{T} : Z \in \mathcal{S}^r_{\succeq 0} \right\}.$$

Moreover, $F(A)$ is given by

$$F(A) = \{X \in \mathcal{S}^n_{\succeq 0} : \mathrm{Ker} X \supseteq \mathrm{Ker} A\} \tag{10.2}$$

and its dimension is equal to $\binom{r+1}{2}$.

Proof. Set $D = \mathrm{diag}(\lambda_1, \cdots, \lambda_r, 0, \cdots, 0) \in \mathcal{S}^n_{\succeq 0}$, $D_0 = \mathrm{diag}(\lambda_1, \cdots, \lambda_r) \in \mathcal{S}^r_{\succ 0}$, where λ_i is the eigenvalue for eigenvector u_i, $C = \mathrm{diag}(0, \cdots, 0, 1, \cdots, 1) \in \mathcal{S}^n_{\succeq 0}$, where the first r entries are 0 and the last $n - r$ entries are 1. Finally, set $Q = UCU^\mathsf{T} = \sum_{i=r+1}^n u_i u_i^\mathsf{T}$. Then, $A = UDU^\mathsf{T}$ and $\langle C, D \rangle = 0$. Moreover, $\langle Q, A \rangle = 0$, as the vectors $u_{r+1}, \cdots, u_n$ span the kernel of A.

As $Q \succeq 0$, the hyperplane

$$H = \{X \in \mathcal{S}^n : \langle Q, X \rangle = 0\}$$

is a supporting hyperplane for $\mathcal{S}^n_{\succeq 0}$ and the intersection

$$F = \mathcal{S}^n_{\succeq 0} \cap H = \{X \in \mathcal{S}^n_{\succeq 0} : \langle Q, X \rangle = 0\}$$

is a face of $\mathcal{S}^n_{\succeq 0}$ containing A. Moreover,

$$F = \{X \in \mathcal{S}^n_{\succeq 0} : \mathrm{Ker} X \supseteq \mathrm{Ker} A\}.$$

Indeed, the condition $\langle Q, X \rangle = 0$ reads $\sum_{i=r+1} u_i^\mathsf{T} X u_i = 0$. For $X \succeq 0$, $u_i^\mathsf{T} X u_i \geqslant 0$ for all i, so that $\langle Q, X \rangle = 0$ if and only if $u_i^\mathsf{T} X u_i = 0$ or, equivalently, $X u_i = 0$ for all $i \in \{r+1, \cdots, n\}$, i.e., $\mathrm{Ker} A \subseteq \mathrm{Ker} X$.

We now show that $F = F(A)$. In view of Lemma 10.1.1, it suffices to show that A lies in the relative interior of the face F.

For this, consider the linear bijection $X \mapsto Y = U^\mathsf{T} X U$. It maps $\mathcal{S}^n_{\succeq 0}$ onto itself, Q onto C, and A onto D, and the face F onto the face

$$F' = \{U^\mathsf{T} X U : X \in F\}.$$

Hence, F' contains D and F' is equal to

$$F' = \{Y \in \mathcal{S}^n_{\succeq 0} : \langle C, Y\rangle = 0\}.$$

Any matrix $Y \in F'$ has its last $n-r$ diagonal entries equal to 0 and thus it has the block form:

$$Y = \begin{pmatrix} Z & 0 \\ 0 & 0 \end{pmatrix} \quad \text{where } Z \in \mathcal{S}^r_{\succeq 0}.$$

Therefore, the faces F' and F are given by

$$F' = \left\{ \begin{pmatrix} Z & 0 \\ 0 & 0 \end{pmatrix} : Z \in \mathcal{S}^r_{\succeq 0} \right\},\ F = \left\{ U \begin{pmatrix} Z & 0 \\ 0 & 0 \end{pmatrix} U^\mathsf{T} : Z \in \mathcal{S}^r_{\succeq 0} \right\}.$$

As $D_0 \succ 0$, D_0 lies in the interior of $\mathcal{S}^r_{\succeq 0}$. This implies that D lies in the relative interior of F' and, in turn, that A belongs to the relative interior of F. Thus, $F = F(A)$.

Summarizing, we have shown that $F(A)$ can be identified with $\mathcal{S}^r_{\succeq 0}$ via the rank-preserving isometry:

$$\begin{array}{lcccl} Z & \mapsto & Y = \begin{pmatrix} Z & 0 \\ 0 & 0 \end{pmatrix} & \mapsto & X = UYU^\mathsf{T} \\ D_0 & \mapsto & D & \mapsto & A \\ \mathcal{S}^r_{\succeq 0} & \to & F' & \to & F(A) \end{array}$$

and the dimension of F is equal to $\dim \mathcal{S}^r_{\succeq 0} = \binom{r+1}{2}$. □

As a direct application, the possible dimensions for the faces of the cone $\mathcal{S}^n_{\succeq 0}$ are $\binom{r+1}{2}$ for $r = 0, 1, \cdots, n$. Moreover there is a one-to-one correspondence between the lattice of faces of $\mathcal{S}^n_{\succeq 0}$ and the lattice of subspaces of $\mathbb{R}^n$:

$$U \ \text{ subspace of } \mathbb{R}^n \mapsto F_U = \{X \in \mathcal{S}^n_{\succeq 0} : \mathrm{Ker} X \supseteq U\}, \tag{10.3}$$

with $U_1 \subseteq U_2 \iff F_{U_1} \supseteq F_{U_2}$.

10.1.3 Faces of spectrahedra

Consider an affine subpsace $\mathcal{A}$ in the space of symmetric matrices, of the form

$$\mathcal{A} = \{X \in \mathcal{S}^n : \langle A_j, X\rangle = b_j \ (j \in [m])\}, \tag{10.4}$$

where $A_1, \cdots, A_m$ are given symmetric matrices and $b_1, \cdots, b_m$ are given scalars. The *codimension* of $\mathcal{A}$ is

$$\text{codim } \mathcal{A} = \dim \mathcal{S}^n - \dim \mathcal{A} = \dim\langle A_1, \cdots, A_m\rangle.$$

If we intersect the cone of positive semidefinite matrices with the affine space $\mathcal{A}$, we obtain the convex set

$$K = \mathcal{S}^n_{\succeq 0} \cap \mathcal{A} = \{X \in \mathcal{S}^n : X \succeq 0,\ \langle A_j, X\rangle = b_j\ (j \in [m])\}. \tag{10.5}$$

This is the feasible region of a typical semidefinite program (in standard primal form). Such a convex set is called a *spectrahedron* – this name is in the analogy with *polyhedron*, which corresponds to the feasible region of a linear program and *spectra* reflects the fact that the definition involves spectral properties of matrices.

An example of a spectrahedron is the *elliptope*

$$\mathcal{E}_n = \{X \in \mathcal{S}^n_{\succeq 0} : X_{ii} = 1\ \forall i \in [n]\}, \tag{10.6}$$

which is the feasible region of the semidefinite relaxation for Max-Cut considered in earlier chapters.

As an application of the description of the faces of the positive semidefinite cone in Proposition 10.1.2, we can describe the faces of K.

Proposition 10.1.3. *Let K be the spectrahedron (15.2). Let $A \in K$, $r = \text{rank}(A)$, and let U, U_0 be as in Proposition 10.1.2. Define the affine space in $\mathcal{S}^r$:*

$$\mathcal{A}_A = \{Z \in \mathcal{S}^r : \langle U_0^\mathsf{T} A_j U_0, Z\rangle = b_j\ \forall j \in [m]\}, \tag{10.7}$$

and the corresponding linear space:

$$\mathcal{L}_A = \{Z \in \mathcal{S}^r : \langle U_0^\mathsf{T} A_j U_0, Z\rangle = 0\ \forall j \in [m]\}. \tag{10.8}$$

The map ϕ from (10.1) identifies $F_K(A)$ and $\mathcal{S}^r_{\succeq 0} \cap \mathcal{A}_A$: $F_K(A) = \phi(\mathcal{S}^r_{\succeq 0} \cap \mathcal{A}_A)$. Moreover, $F_K(A)$ is given by

$$F_K(A) = \{X \in K : \text{Ker} X \supseteq \text{Ker} A\} \tag{10.9}$$

and its dimension is equal to

$$\dim F_K(A) = \dim \mathcal{A}_A = \binom{r+1}{2} - \dim\langle U_0^\mathsf{T} A_j U_0 : j \in [m]\rangle. \tag{10.10}$$

Finally, a matrix $B \in \mathcal{S}^n$ is a perturbation of A if and only if $B \in U_0 \mathcal{L}_A U_0^\mathsf{T}$.

Proof. As $K = \mathcal{S}^n_{\succeq 0} \cap \mathcal{A}$, we have that $F_K(A) = F(A) \cap \mathcal{A}$, where $F(A)$ is the smallest face of $\mathcal{S}^n_{\succeq 0}$ containing A, and (10.9) follows from (10.2). If $X = \phi(Z)$ is the image of $Z \in \mathcal{S}^r$ under the map ϕ from (10.1) then

$$\langle A_j, X\rangle = \langle U^\mathsf{T} A_j U, U^\mathsf{T} X U\rangle = \left\langle U^\mathsf{T} A_j U, \begin{pmatrix} Z & 0 \\ 0 & 0 \end{pmatrix} \right\rangle = \langle U_0^\mathsf{T} A_j U_0, Z\rangle.$$

Therefore, the face $F_K(A)$ is the image of $\mathcal{S}^r_{\succeq 0} \cap \mathcal{A}_A$ under the map ϕ and its dimension is equal to $\dim \mathcal{A}_A$. Finally, B is a perturbation of A if and only if $A \pm \epsilon B \in F_K(A)$ for some $\epsilon > 0$, which is equivalent to $B \in U_0 \mathcal{L}_A U_0^\mathsf{T}$ using the description of $F_K(A)$. □

Corollary 10.1.4. *Let K be defined as in (15.2). Let $A \in K$ and $r = \text{rank}(A)$. If A is an extreme point of K then*

$$\binom{r+1}{2} \leqslant \text{codim } \mathcal{A} \leqslant m \tag{10.11}$$

In particular, K contains a matrix A whose rank r satisfies

$$r \leqslant \frac{-1+\sqrt{8m+1}}{2}. \tag{10.12}$$

Proof. If A is an extreme point of K then $\dim F_K(A) = 0$ and (15.15) follows directly from (10.10). As K contains no line, K has at least one extreme point. Now (10.12) follows directly from $\binom{r+1}{2} \leqslant m$ for any matrix A which is an extreme point of K. □

Remark 10.1.5. *The codimension of the affine space $\mathcal{A}_A$ can be expressed from any Cholesky decomposition: $A = WW^\mathsf{T}$, where $W \in \mathbb{R}^{n\times r}$, by*

$$\text{codim } \mathcal{A}_A = \dim\langle WA_jW^\mathsf{T} : \ j \in [m]\rangle.$$

Indeed, the matrix $P = W^\mathsf{T}U_0D_0^{-1}$ is nonsingular, since $P^\mathsf{T}P = D_0^{-1}$ using the fact that $U_0^\mathsf{T}U_0 = I_r$. Moreover, $WP = U_0$, and thus

$$\dim\langle W^\mathsf{T}A_jW : j \in [m]\rangle = \dim\langle P^\mathsf{T}W^\mathsf{T}A_jWP : j \in [m]\rangle = \dim\langle U_0^\mathsf{T}A_jU_0 : j \in [m]\rangle.$$

As an illustration, for the elliptope $K = \mathcal{E}_n$, if $A \in \mathcal{E}_n$ is the Gram matrix of vectors $\{a_1, \cdots, a_n\} \subseteq \mathbb{R}^k$, then codim $\mathcal{A}_A = \dim\langle a_1a_1^\mathsf{T}, \cdots, a_na_n^\mathsf{T}\rangle$.

As an illustration we discuss a bit the geometry of the elliptope $\mathcal{E}_n$. As a direct application of Corollary 10.1.4, we obtain the following bound for the rank of extreme points:

Corollary 10.1.6. *Any extreme point of $\mathcal{E}_n$ has rank r satisfying $\binom{r+1}{2} \leqslant n$.*

A matrix $X \in \mathcal{E}_n$ has rank 1 if and only if it is of the form $X = xx^\mathsf{T}$ for some $x \in \{\pm 1\}^n$. Such matrix is also called a *cut matrix* (since it corresponds to a cut in the complete graph K_n). There are 2^{n-1} distinct cut matrices. They are extreme points of $\mathcal{E}_n$ and any two of them form an edge (face of dimension 1) of $\mathcal{E}_n$. While for $n \leqslant 4$, these are the only faces of dimension 1, the elliptope $\mathcal{E}_n$ for $n \geqslant 5$ has faces of dimension 1 that are not an edge between two cut matrices. You will see an example in Exercise 10.3.

Figure 10.1 shows the elliptope $\mathcal{E}_3$ (more precisely, its bijective image in $\mathbb{R}^3$ obtained by taking the upper triangular part of X). Note the four corners, which correspond to the four cuts of the graph K_3. All the points on the boundary of $\mathcal{E}_3$ - except those lying on an edge between two of the four corners – are extreme points. For instance, the matrix

$$A = \begin{pmatrix} 1 & 0 & 1/\sqrt{2} \\ 0 & 1 & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} & 1 \end{pmatrix}$$

is an extreme point of $\mathcal{E}_3$ (check it), with rank $r = 2$.

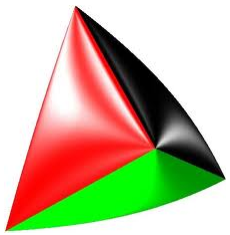


Figure 10.1: The elliptope $\mathcal{E}_3$

10.1.4 Finding an extreme point in a spectrahedron

In order to find a matrix A in a spectrahedron K whose rank satisfies (10.12), it suffices to find an extreme point A of K. Algorithmically this can be done as follows.

Suppose we have a matrix $A \in K$ with rank r. Observe that A is an extreme point of K precisely when the linear space $\mathcal{L}_A$ (in (10.8)) is reduced to the zero matrix. Assume that A is not an extreme point of K. Pick a nonzero matrix $C \in \mathcal{L}_A$, so that $B = U_0 C U_0^\mathsf{T}$ is a nonzero perturbation of A. Hence $A \pm tB \succeq 0$ for some $t > 0$. Moreover, at least one of the supremums: $\sup\{t > 0 : A + tB \succeq 0\}$ and $\sup\{t > 0 : A - tB \succeq 0\}$ is finite, since K contains no line. Say, the first supremum is finite, and compute the largest scalar $t > 0$ for which $A + tB \succeq 0$ (this is a semidefinite program). Then the matrix $A' = A + tB$ still belongs to the face $F_K(A)$, but it now lies on its border (by the maximality of t). Therefore, A' has a larger kernel: $\mathrm{Ker} A' \supset \mathrm{Ker} A$, and thus a smaller rank: $\mathrm{rank} A' \leqslant \mathrm{rank} A - 1$. Then iterate, replacing A by A', until finding an extreme point of K.

Therefore, one can find an extreme point of K by solving at most n semidefinite programs. However, finding the smallest possible rank of a matrix in K is a hard problem – see Proposition 10.2.4.

10.1.5 A refined bound on ranks of extreme points

The upper bound on the rank of an extreme point from Corollary 10.1.4 is tight – see Example 10.2.3 below. However, there is one special case when it can be sharpened, as we explain here. Consider again the affine space $\mathcal{A}$ from (10.4) and the spectrahedron $K = \mathcal{S}^n_{\succeq 0} \cap \mathcal{A}$. From Corollary 10.1.4, we know that any extreme point A of K has rank r satisfying

$$\binom{r+1}{2} \leqslant \mathrm{codim}\, \mathcal{A}.$$

Hence, $r \leqslant s+1$ if $\mathrm{codim}\, \mathcal{A} = \binom{s+2}{2}$. Under some assumptions, Barvinok shows that $r \leqslant s$ for at least one extreme point of K.

Proposition 10.1.7. *Assume that K is nonempty bounded and* codim $\mathcal{A} = \binom{s+2}{2}$ *for some integer $s \geqslant 1$ satisfying $n \geqslant s+2$. Then there exists $A \in K$ with* rank $A \leqslant s$.

The proof uses the following topological result.

Theorem 10.1.8. *Consider the projective space $\mathbf{P}^{n-1}$, consisting of all lines in $\mathbb{R}^n$ passing through the origin, and let $\mathbf{S}^{n-1}$ be the unit sphere in $\mathbb{R}^n$. For $n \geqslant 3$ there does not exist a continuous map $\Phi : \mathbf{S}^{n-1} \to \mathbf{P}^{n-1}$ such that $\Phi(x) \neq \Phi(y)$ for all distinct $x, y \in \mathbf{S}^{n-1}$.*

The following lemma deals with the case $n = s+2$, it is the core of the proof of Proposition 10.1.7.

Lemma 10.1.9. *Let $n = s+2$ with $s \geqslant 1$ and let $\mathcal{A} \subseteq \mathcal{S}^{s+2}$ be an affine space with* codim $\mathcal{A} = \binom{s+2}{2}$. *If $K = \mathcal{S}^{s+2}_{\geq 0} \cap \mathcal{A}$ is nonempty and bounded, then there is a matrix $A \in K$ with* rank $A \leqslant s$.

Proof. Assume first that $\mathcal{A} \cap \mathcal{S}^{s+2}_{>0} = \varnothing$. Then $\mathcal{A}$ lies in a hyperplane H supporting a proper face F of $\mathcal{S}^{s+2}_{\geq 0}$. (This can be checked using the separating theorem from Theorem 1.3.8 (i).) By Proposition 10.1.2, F can be identified with $\mathcal{S}^{t}_{\geq 0}$ for some $t \leqslant s+1$ and thus an extreme point of K has rank at most $t-1 \leqslant s$.

Suppose now that $\mathcal{A} \cap \mathcal{S}^{s+2}_{>0} \neq \varnothing$. By (10.10), $\dim K = \binom{s+3}{2} -$ codim $\mathcal{A} = s+2$. Hence, K is a $(s+2)$-dimensional compact convex set, whose boundary ∂K is (topologically) the sphere $\mathbf{S}^{s+1}$. We now show that the boundary of K contains a matrix with rank at most s.

Clearly every matrix in ∂K has rank at most $s+1$. Suppose for a contradiction that no matrix of ∂K has rank at most s. Then, each matrix $X \in \partial K$ has rank $s+1$ and thus its kernel Ker$X$ has dimension 1, it is a line though the origin. We can define a continuous map Φ from ∂K to $\mathbf{P}^{s+1}$ in the following way: For each matrix $X \in \partial K$, its image $\Phi(X)$ is the line Ker$X$. The map Φ is continuous (check it) from $\mathbf{S}^{s+1}$ to $\mathbf{P}^{s+1}$ with $s+1 \geqslant 2$. Hence, applying Theorem 10.1.8, we deduce that there are two distinct matrices $X, X' \in \partial K$ with the same kernel: Ker$X =$ KerX'. Hence X and X' are two distinct points lying in the same face of K: $F_K(X) = F_K(X')$. Then this face has an extreme point A, whose rank satisfies rank$A \leqslant$ rank$X - 1 \leqslant s$. □

We can now conclude the proof of Proposition 10.1.7.

Proof. (of Proposition 10.1.7). By Corollary 10.1.4 there exists a matrix $A \in K$ with rank $A \leqslant s+1$. Pick a vector space $U \subseteq$ Ker$A$ with codim $U = s+2$. By Proposition 10.1.2, there is a rank-preserving isometry between F_U and $\mathcal{S}^{s+2}_{\geq 0}$. Moreover, $A \in F_U \cap \mathcal{A}$. Hence the result follows by applying Lemma 10.1.9. □

Example 10.1.10. *Consider the three matrices*

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \; B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \; C = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

and the affine space

$$\mathcal{A} = \{X \in \mathcal{S}^2 : \langle A, X\rangle = 0,\ \langle B, X\rangle = 0,\ \langle C, X\rangle = 1\}.$$

Then $\mathcal{S}^2_{\geq 0} \cap \mathcal{A} = \{I\}$ *thus contains no rank 1 matrix, and* codim $\mathcal{A} = 3 = \binom{s+2}{2}$ *with* $s = 1$*. This example shows that the condition* $n \geqslant s + 2$ *cannot be omitted in Lemma 10.1.9.*

Example 10.2.3 below shows that the assumption that K *is bounded cannot be omitted as well.*

10.2 Applications

10.2.1 Euclidean realizations of graphs

The *graph realization problem* can be stated as follows. Suppose we are given a graph $G = (V = [n], E)$ together with nonnegative edge weights $w \in \mathbb{R}^E_+$, viewed as 'lengths' assigned to the edges. We say that (G, w) *is* d*-realizable* if one can place the nodes of G at points $v_1, \cdots, v_n \in \mathbb{R}^d$ in such a way that their Euclidean distances respect the given edge lengths:

$$\exists v_1, \cdots, v_n \in \mathbb{R}^d \ \ \|v_i - v_j\|^2 = w_{ij} \ \ \forall \{i, j\} \in E. \tag{10.13}$$

(We use here the squares of the Euclidean distances as this makes the notation easier). Moreover, (G, w) is *realizable* if it is d-realizable for some $d \geqslant 1$. In dimension 3, the problem of testing d-realizability arises naturally in robotics or computational chemistry (the given lengths represent some known distances between the atoms of a molecule and one wants to reconstruct the molecule from these partial data).

Testing whether a weighted graph is realizable amounts to testing feasibility of a semidefinite program:

Lemma 10.2.1. (G, w) *is realizable if and only if the following semidefinite program (in matrix variable* $X \in \mathcal{S}^n$*):*

$$X_{ii} + X_{jj} - 2X_{ij} = w_{ij}\ \forall \{i, j\} \in E,\ X \succeq 0 \tag{10.14}$$

has a feasible solution. Moreover, (G, w) *is* d*-realizable if and only if the system (10.14) has a solution of rank at most* d*.*

Proof. If $v_1, \cdots, v_n \in \mathbb{R}^d$ is a realization of (G, w), then their Gram matrix $X = (v_i^{\mathsf{T}} v_j)$ is a solution of rank at most d of (10.14). Conversely, if X is a solution of (10.14) of rank $\leqslant d$ and $v_1, \cdots, v_n \in \mathbb{R}^d$ is a Gram decomposition of X, then the v_i's form a d-realization of (G, w). □

As a direct application of Corollary 10.1.4, any realizable graph (G, w) is d-realizable in dimension d satisfying

$$\binom{d+1}{2} \leqslant |E|, \ \text{ i.e., } \ d \leqslant \frac{-1 + \sqrt{8|E| + 1}}{2}. \tag{10.15}$$

When $G = K_n$ is a complete graph, checking whether (K_n, w) is d-realizable amounts to checking whether a suitable matrix is positive semidefinite and computing its rank:

Lemma 10.2.2. *Consider the complete graph $G = K_n$ with edge weights w, and define the matrix $X \in \mathcal{S}^{n-1}$ by*

$$X_{ii} = w_{in} \ (i \in [n-1]), \ X_{ij} = \frac{w_{in} + w_{jn} - w_{ij}}{2} \ (i \neq j \in [n-1]).$$

Then, (K_n, w) is d-realizable if and only if $X \succeq 0$ and rank$X \leqslant d$.

Proof. The proof relies on the observation that if a set of vectors $v_1, \cdots, v_n \in \mathbb{R}^d$ satisfies (10.13), then one can translate it and thus assume without loss of generality that $v_n = 0$. □

Example 10.2.3. *Consider the complete graph $G = K_n$ with weights $w_{ij} = 1$ for all edges. Then (K_n, w) is $(n-1)$-realizable but it is not $(n-2)$-realizable (easy to check using Lemma 10.2.2).*

Hence, the upper bound (10.15) is tight on this example. This shows that the condition that K is bounded cannot be omitted in Proposition 10.1.7. (Note that the set of feasible solutions to the program (10.14) is indeed not bounded).

On the other hand, for any fixed $d \geqslant 1$, deciding whether a graph (G, w) is d-realizable is a hard problem. Therefore, deciding whether the semidefinite program (10.14) has a solution of rank at most d is a hard problem.

We show this for $d = 1$. Then there is a simple reduction from the *partition problem*: Decide whether a given sequence of integers $a_1, \cdots, a_n \in \mathbb{N}$ can be partitioned, i.e., whether there exists $\epsilon \in \{\pm 1\}^n$ such that $\epsilon_1 a_1 + \cdots + \epsilon_n a_n = 0$.

Proposition 10.2.4. *Given a graph (G, w) with integer lengths $w \in \mathbb{N}^E$, deciding whether (G, w) is 1-embeddable is an $\mathcal{NP}$-complete problem, already when G is restricted to be a circuit.*

Proof. Let $a_1, \cdots, a_n \in \mathbb{N}$ be an instance of the partition problem. Consider the circuit $G = C_n$ of length n, with edges $\{i, i+1\}$ for $i \in [n]$ (indices taken modulo n). Assign the length $w_{i,i+1} = a_{i+1}$ to edge $\{i, i+1\}$ for $i = 1, \cdots, n$. It is now an easy exercise to show that (C_n, w) is 1-realizable if and only if the sequence $(a_1, \cdots, a_n)$ can be partitioned.

Indeed, assume that $v_1, \cdots, v_{n-1}, v_n \in \mathbb{R}$ is a 1-realization of (C_n, w). Without loss of generality we may assume that $v_n = 0$. The condition $w_{n,1} = a_1 = |v_1|$ implies that $v_1 = \epsilon_1 a_1$ for some $\epsilon_1 \in \{\pm 1\}$. Next, for $i = 1, \cdots, n-1$, the conditions $w_{i,i+1} = a_{i+1} = |v_i - v_{i+1}|$ imply the existence of $\epsilon_2, \cdots, \epsilon_n \in \{\pm 1\}$ such that $v_{i+1} = v_i + \epsilon_{i+1} a_{i+1}$. This implies $0 = v_n = \epsilon_1 a_1 + \cdots + \epsilon_n a_n$ and thus the sequence $a_1, \cdots, a_n$ can be partitioned.

These arguments can be reversed to show the reverse implication. □

On the other hand:

Lemma 10.2.5. *If a circuit (C_n, w) is realizable, then it is 2-realizable.*

This can be shown (Exercise 10.1) using the following basic geometrical fact.

Lemma 10.2.6. *Let $u_1, \cdots, u_k \in \mathbb{R}^n$ and $v_1, \cdots, v_k \in \mathbb{R}^n$ two sets of vectors representing the same Euclidean distances, i.e., satisfying*

$$\|u_i - u_j\| = \|v_i - v_j\| \ \ \forall i, j \in [k].$$

Then there exists an orthogonal matrix $A \in \mathcal{O}(n)$ and a vector $a \in \mathbb{R}^n$ such that $v_i = Au_i + a$ for all $i \in [k]$.

But the above shows: Any realizable weighted circuit can be embedded in the line or in the plane, but deciding which one of these two possibilities holds is an $\mathcal{NP}$-complete problem!

10.2.2 Hidden convexity results for quadratic maps

As a direct application of Proposition 10.1.4, we obtain the following result for systems of two quadratic equations.

Proposition 10.2.7. *Consider two matrices $A, B \in \mathcal{S}^n$ and $a, b \in \mathbb{R}$. Then the system of two quadratic equations*

$$\sum_{i,j=1}^{n} A_{ij} x_i x_j = a, \quad \sum_{i,j=1}^{n} B_{ij} x_i x_j = b \tag{10.16}$$

has a real solution $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$ if and only if the system of two linear matrix equations

$$\langle A, X \rangle = a, \ \ \langle B, X \rangle = b \tag{10.17}$$

has a positive semidefinite solution $X \succeq 0$.

Proof. If x is a solution of (10.16), then $X = xx^{\mathsf{T}}$ is a solution of (10.17). Conversely, assume that the system (10.17) has a solution. Applying Corollary 10.1.4, we know that it has a solution of rank r satisfying $\binom{r+1}{2} \leqslant m = 2$, thus with $r \leqslant 1$. Now, if X has rank 1, it can be written in the form $X = xx^{\mathsf{T}}$, so that x is a solution of (10.16). □

This result does not extend to three equations: The affine space from Example 10.1.10 contains a positive semidefinite matrix, but none of rank 1. As we now observe, the above result can be reformulated as follows: The image of $\mathbb{R}^n$ under a quadratic map into $\mathbb{R}^2$ is a convex set.

Proposition 10.2.8. (Dines 1941) *Given two matrices $A, B \in \mathcal{S}^n$, the image of $\mathbb{R}^n$ under the quadratic map $q(x) = (x^{\mathsf{T}}Ax, x^{\mathsf{T}}Bx)$:*

$$\mathcal{Q} = \{(x^{\mathsf{T}}Ax, x^{\mathsf{T}}Bx) : x \in \mathbb{R}^n\}, \tag{10.18}$$

is a convex set in $\mathbb{R}^2$.

Proof. Set

$$\mathcal{Q}' = \{(\langle A, X\rangle, \langle B, X\rangle) \in \mathbb{R}^2 : X \in \mathcal{S}^n_{\geq 0}\}.$$

Clearly, $\mathcal{Q} \subseteq \mathcal{Q}'$ and $\mathcal{Q}'$ is convex. Thus it suffices to show equality: $\mathcal{Q} = \mathcal{Q}'$. For this, let $(a, b) \in \mathcal{Q}'$. Then the system (10.17) has a solution $X \succeq 0$. By Proposition 10.2.7, the system (10.16) too has a solution, and thus $(a, b) \in \mathcal{Q}$. □

While it is *not obvious from its definition* that the set $\mathcal{Q}$ is convex, it is *obvious from its definition* that the above set $\mathcal{Q}'$ is convex. For this reason, such a result is called a *hidden convexity result*.

Here is another hidden convexity result, showing that the image of the unit sphere $\mathbf{S}^{n-1}$ ($n \geqslant 3$) under a quadratic map in $\mathbb{R}^2$ is convex. We show it using the refined bound from Proposition 10.1.7.

Proposition 10.2.9. (Brickman 1961) *Let $n \geqslant 3$, $A, B \in \mathcal{S}^n$ and $a, b \in \mathbb{R}$. Then the image of the unit sphere under the quadratic map $q(x) = (x^\mathsf{T} Ax, x^\mathsf{T} Bx)$:*

$$\mathcal{C} = \{(x^\mathsf{T} Ax, x^\mathsf{T} Bx) : \sum_{i=1}^n x_i^2 = 1\}$$

is a convex set in $\mathbb{R}^2$.

Proof. It suffices to show that, if the set

$$K = \{X \in \mathcal{S}^n_{\geq 0} : \langle A, X\rangle = a,\ \langle B, X\rangle = b,\ \mathrm{Tr}(X) = 1\}$$

is not empty then it contains a matrix of rank 1. Define the affine space

$$\mathcal{A} = \{X \in \mathcal{S}^n : \langle A, X\rangle = a,\ \langle B, X\rangle = b,\ \mathrm{Tr}(X) = 1\}.$$

Then the existence of a matrix of rank 1 in K follows from Corollary 10.1.4 if codim $\mathcal{A} \leqslant 2$, and from Proposition 10.1.7 if codim $\mathcal{A} = 3$ (as K is bounded, codim $\mathcal{A} = \binom{s+2}{3}$, $n \geqslant s + 2$ for $s = 1$). □

The assumption $n \geqslant 3$ cannot be omitted in Proposition 10.2.9: Consider the quadratic map q defined using the matrices A and B from Example 10.1.10. Then, $q(1, 0) = (1, 0)$, $q(0, 1) = (-1, 0)$, but $(0, 0)$ does not belong to the image of $\mathbf{S}^1$ under q.

We conclude with the following application of Proposition 10.2.9, which shows that the numerical range $R(M)$ of a complex matrix $M \in \mathbb{C}^{n\times n}$ is a convex subset of $\mathbb{C}$ (viewed as $\mathbb{R}^2$). Recall that the *numerical range* of M is

$$R(M) = \{z^* Mz = \sum_{i,j=1}^n \overline{z_i} M_{ij} z_i : z \in \mathbb{C}^n, \sum_{i=1}^n |z_i|^2 = 1\}.$$

Proposition 10.2.10. (Toeplitz-Hausdorff) *The numerical range of a complex matrix is convex.*

Proof. Write $z \in \mathbb{C}^n$ as $z = x + \mathbf{i}y$ where $x, y \in \mathbb{R}^n$, so that $\sum_i |z_i|^2 = \sum_i x_i^2 + y_i^2$. Define the quadratic map $q(x, y) = (q_1(x, y), q_2(x, y))$ by

$$z^* M z = q_1(x, y) + \mathbf{i} q_2(x, y).$$

Then, the numerical range of M is the image of the unit sphere S^{2n-1} under the map q, and the result follows from Proposition 10.2.9. □

10.2.3 The S-Lemma

In the preceding section we dealt with systems of quadratic equations. We now discuss systems of quadratic inequalities.

Recall Farkas' lemma for linear programming: If a system of linear inequalities:

$$\begin{cases} a_1^\mathsf{T} x \leqslant b_1 \\ \vdots \\ a_m^\mathsf{T} x \leqslant b_m \end{cases}$$

implies the linear inequality $c^\mathsf{T} x \leqslant d$, then there exist nonnegative scalars $\lambda_1, \cdots, \lambda_m \geqslant 0$ such that $c = \lambda_1 a_1 + \cdots + \lambda_m a_m$ and $\lambda_1 b_1 + \cdots + \lambda_m b_m \leqslant d$.

This type of inference rules does not extend to general nonlinear inequalities. However such an extension does hold in the case of quadratic polynomials, in the special case $m = 1$ (and under some strict feasibility assumption).

Theorem 10.2.11. (The homogeneous S-lemma) *Given matrices $A, B \in \mathcal{S}^n$, assume that $x^\mathsf{T} A x > 0$ for some $x \in \mathbb{R}^n$. The following assertions are equivalent.*

(i) $\{x \in \mathbb{R}^n : x^\mathsf{T} A x \geqslant 0\} \subseteq \{x \in \mathbb{R}^n : x^\mathsf{T} B x \geqslant 0\}$.

(ii) *There exists a scalar $\lambda \geqslant 0$ such that $B - \lambda A \succeq 0$.*

Proof. The implication (ii) $\Longrightarrow$ (i) is obvious. Now, assume (i) holds, we show (ii). For this consider the semidefinite program (P):

$$\inf\{\langle B, X\rangle : \langle A, X\rangle \geqslant 0,\ \mathrm{Tr}(X) = 1,\ X \succeq 0\}$$

and its dual (D):

$$\sup\{y : B - zA - yI \succeq 0,\ z \geqslant 0\}.$$

First we show that (P) is strictly feasible. By assumption, there exists a unit vector x for which $x^\mathsf{T} A x > 0$. If $\mathrm{Tr}(A) \geqslant 0$ then $X = xx^\mathsf{T}/2 + I/2n$ is a strictly feasible solution. Assume now that $\mathrm{Tr}(A) < 0$. Set $X = \alpha x x^\mathsf{T} + \beta I$, where we choose $\alpha \geqslant 0$, $\beta > 0$ in such a way that $1 = \mathrm{Tr}(X) = \alpha + \beta n$ and $0 < \langle A, X\rangle = \alpha x^\mathsf{T} A x + \beta \mathrm{Tr}(A)$, i.e.,

$$\frac{x^\mathsf{T} A x}{n x^\mathsf{T} A x - \mathrm{Tr}(A)} < \beta \leqslant \frac{1}{n}.$$

Then X is strictly feasible for (P).

Next we show that the optimum value of (P) is nonnegative. For this, consider a feasible solution X_0 of (P) and consider the set

$$K = \{X \in \mathcal{S}^n_{\geq 0} : \langle A, X\rangle = \langle A, X_0\rangle,\ \langle B, X\rangle = \langle B, X_0\rangle\}.$$

As $K \neq \varnothing$, applying Corollary 10.1.4, there is a matrix $X \in K$ with rank 1. Say $X = xx^{\mathsf{T}}$. Then, $x^{\mathsf{T}}Ax = \langle A, X_0\rangle \geqslant 0$ which, by assumption (i), implies $x^{\mathsf{T}}Bx \geqslant 0$, and thus $\langle B, X_0\rangle = x^{\mathsf{T}}Bx \geqslant 0$.

As (P) is bounded and strictly feasible, applying the duality theorem, we deduce that there is no duality gap and that the dual problem has an optimal solution (y, z) with $y, z \geqslant 0$. Therefore, $B - zA = (B - zA - yI) + yI \succeq 0$, thus showing (ii). □

This extends to non-homogeneous quadratic polynomials (Exercise 10.2):

Theorem 10.2.12. (The non-homogeneous S-lemma)
Let $f(x) = x^{\mathsf{T}}Ax + 2a^Tx + \alpha$ and $g(x) = x^{\mathsf{T}}Bx + 2b^{\mathsf{T}}x + \beta$ be two quadratic polynomials where $A, B \in \mathcal{S}^n$, $a, b \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$. Assume that $f(x) > 0$ for some $x \in \mathbb{R}^n$. The following assertions are equivalent.

(i) $\{x \in \mathbb{R}^n : f(x) \geqslant 0\} \subseteq \{x \in \mathbb{R}^n : g(x) \geqslant 0\}$.

(ii) *There exists a scalar $\lambda \geqslant 0$ such that* $\begin{pmatrix} \beta & b^{\mathsf{T}} \\ b & B \end{pmatrix} - \lambda \begin{pmatrix} \alpha & a^{\mathsf{T}} \\ a & A \end{pmatrix} \succeq 0$.

(iii) *There exist a nonnegative scalar λ and a polynomial $h(x)$ which is a sum of squares of polynomials such that $g = \lambda f + h$.*

10.3 Notes and further reading

Part of the material in this chapter can be found in the book of Barvinok [1]. In particular, the refined bound (from Section 10.1.5) on the rank of extreme points of a spectrahedron is due to Barvinok. Details about the geometry of the elliptope can be found in [3].

The structure of the d-realizable graphs has been studied by Belk and Connelly [2]. It turns out that the class of d-realizable graphs is closed under taking minors, and it can be characterized by finitely many forbidden minors. For $d \leqslant 3$ the forbidden minors are known: A graph G is 1-realizable if and only if it is a forest (no K_3-minor), G is 2-realizable if and only if it has no K_4-minor, and G is 3-realizable if and only if it does not contain K_5 and $K_{2,2,2}$ as a minor. (You will show some partial results in Exercise 10.1.) Saxe [5] has shown that testing whether a weighted graph is d-realizable is $\mathcal{NP}$-hard for any fixed d.

The S-lemma dates back to work of Jakubovich in the 1970s in control theory. There is a rich history and many links to classical results about quadratic systems of (in)equations (including the results of Dines and Brickman presented here), this is nicely exposed in the survey of Polik and Terlaky [4].

10.4 Exercises

10.1** A graph G is said to be *d-realizable* if, for any edge weights w, (G, w) is d-realizable whenever it is realizable. For instance, the complete graph K_n is $(n-1)$-realizable, but not $(n-2)$-realizable (Example 10.2.3).

(a) Given two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ such that $V_1 \cap V_2$ is a clique in G_1 and G_2, their *clique sum* is the graph $G = (V_1 \cup V_2, E_1 \cup E_2)$.

Show that if G_1 is d_1-realizable and G_2 is d_2-realizable, then G is d-realizable where $d = \max\{d_1, d_2\}$.

(b) Given a graph $G = (V, E)$ and an edge $e \in E$, $G\backslash e = (V, E\backslash\{e\})$ denotes the graph obtained by *deleting* the edge e in G.

Show that if G is d-realizable, then $G\backslash e$ is d-realizable.

(c) Given a graph $G = (V, E)$ and an edge $e = \{i_1, i_2\} \in E$, G/e denotes the graph obtained by *contracting* the edge e in G, which means: Identify the two nodes i_1 and i_2, i.e., replace them by a new node, called i_0, and replace any edge $\{i_1, j\} \in E$ by $\{i_0, j\}$ and any edge $\{i_2, j\} \in E$ by $\{i_0, j\}$.

Show that if G is d-realizable, then G/e is d-realizable.

(d) Show that the circuit C_n is 2-realizable, but not 1-realizable.

(e) Show that G is 1-realizable if and only if G is a forest (i.e., a disjoint union of trees).

(f) Show that $K_{2,2,2}$ is 4-realizable, but not 3-realizable.

NB: A *minor* of G is a graph that can be obtained from G by deleting and contracting edges and by deleting nodes. So the above shows that if G is d-realizable then any minor of G is d-realizable. Moreover, if G is 3-realizable then G has no K_5 and $K_{2,2,2}$ minor. The reverse implication holds but requires more work [2].

10.2** (a) Let $A, B, C \in \mathcal{S}^n$, $a, b, c \in \mathbb{R}$ and let

$$\mathcal{Q} = \{q(x) = (x^\mathsf{T} Ax, x^\mathsf{T} Bx, x^\mathsf{T} Cx) : x \in \mathbb{R}^n\} \subseteq \mathbb{R}^3$$

denote the image of $\mathbb{R}^n$ under the quadratic map q. Assume that $n \geqslant 3$ and that there exist $\alpha, \beta, \gamma \in \mathbb{R}$ such that $\alpha A + \beta B + \gamma C \succ 0$.

Show that the set $\mathcal{Q}$ is convex.

(b) Show Theorem 10.2.12.

10.3 (a) Consider the two cut matrices J (the all-ones matrix) and $X = xx^\mathsf{T}$ where $x \in \{\pm 1\}^n$, distinct from the all-ones vector. Show that the segment $F = [J, X]$ is a face of the elliptope $\mathcal{E}_n$.

(b) Consider the matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 1 & 0 & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 & 0 & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 & 1 & 1/2 \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} & 1/2 & 1 \end{pmatrix} \in \mathcal{E}_5.$$

What is the dimension of the face $F_{\mathcal{E}_5}(A)$? What are its extreme points?

10.4 Let p be a polynomial in two variables and with (even) degree d. Show that if p can be written as a sum of squares, then it can be written as a sum of at most $d + 1$ squares.

NB: For $d = 4$, Hilbert has shown that p can be written as sum of at most *three* squares but this is a difficult result.

BIBLIOGRAPHY

[1] A. Barvinok. *A Course in Convexity.* AMS, 2002.

[2] M. Belk and R. Connelly. Realizability of graphs. *Discrete and Computational Geometry*, **37**:125–137, 2007.

[3] M. Laurent and S. Poljak. On the facial structure of the set of correlation matrices. *SIAM Journal on Matrix Analysis,* **17**:530–547, 1996.

[4] I. Polik and T. Terlaky. A survey of the S-lemma. *SIAM Review*, **49(3)**: 371–418, 2007.

[5] J. B. Saxe. Embeddability of weighted graphs in k-space is strongly NP-hard. In *Proc. 17-th Allerton Conf. Comm. Control Comp.,* 480–489, 1979.

CHAPTER 11

EUCLIDEAN EMBEDDINGS: LOW DISTORTION

11.1 Motivation: Embeddings of finite metric spaces

Definition 11.1.1. *A* finite metric space *is a pair* (X, d) *where* X *is a finite set and where the function* $d : X \times X \to \mathbb{R}$ *defines a metric: For all* $x, y, z \in X$ *we have*

(non-negativity) $d(x, y) \geqslant 0$*, and* $d(x, y) = 0$ *if and only if* $x = y$*,*

(symmetry) $d(x, y) = d(y, x)$*,*

(triangle inequality) $d(x, z) \leqslant d(x, y) + d(y, z)$*.*

One important example is the *shortest path metric* of a connected graph $G = (V, E)$. There we measure the distance $d(x, y)$ between two vertices x, y in G by the length of a shortest path connecting x and y. Here the length of a path is the number of its edges.

In computational phylogenetics one frequently deals with genetic distance matrices. See Table 11.1.

To work with finite metric spaces one wants to perform data analysis or one wants to visualize them. For these tasks there are many geometric algorithms available which are based on the Euclidean metric but which are not available for arbitrary metric spaces. So it is an obvious method to map the points of the finite metric space into a Euclidean space, preferably one of low dimension.

A *Euclidean embedding* $f : X \to \mathbb{R}^n$ is an injective map from X to n-dimensional Euclidean space. We want to embed X *isometrically* into Euclidean

	Ban	E.Af.	W.Af.	San	Ind.	N.E.	Kor.	S.C.	Eng.	Aus.
Bantu	0									
E. Africa	658	0								
W. Africa	188	697	0							
San	94	776	885	0						
India	2202	1078	1748	1246	0					
Near East	1779	709	1454	880	229	0				
Korea	2668	1475	1807	1950	681	933	0			
S. China	2963	1664	1958	2231	847	983	498	0		
English	2288	1163	1487	1197	280	236	982	1152	0	
Australia	3272	2131	2694	2705	1176	1408	850	1081	1534	0

Table 11.1: Genetic distance matrix due to Cavalli-Sforza (1994). (What is "wrong" with this matrix?)

space $\mathbb{R}^n$, so that for all $x, y \in X$ we have

$$d(x,y) = \|f(x) - f(y)\| = \sqrt{\sum_{i=1}^{n} (f(x)_i - f(y)_i)^2},$$

where $f(x)_i$ denotes the i-th component of the vector $f(x) \in \mathbb{R}^n$.

There are two problems with isometric embeddings into Euclidean spaces: If we insist on finding an isometric embedding into Euclidean space with a fixed dimension n, independent of the cardinality of X, then finding such an embedding is a semidefinite optimization problem with a rank constraint; indeed an NP-hard problem. If we relax the rank constraint, then we are dealing with a semidefinite feasibility problem. However, in general it will not be feasible.

Example 11.1.2. *Consider for instance the shortest path metric of the star graph*

$$X = \{1,2,3,4\}, \text{ with } d(1,4) = d(2,4) = d(3,4) = 1, \text{ and } d(i,j) = 2, \text{ otherwise.}$$

To embed (X,d) isometrically into Euclidean space, one needs that each of the triplets $\{1,2,4\}$, $\{1,3,4\}$ and $\{2,3,4\}$ lie on a single line, which is impossible.

In this lecture we propose to use Euclidean embedding having low distortion instead of (non-existing) isometric Euclidean embeddings.

Definition 11.1.3. *Let (X,d) be a finite metric space and let $f : X \to \mathbb{R}^n$ be an embedding into Euclidean space. We define the* expansion*,* contraction *and the* distortion *of f by*

$$\text{expansion}(f) = \max_{x,y \in X} \frac{\|f(x) - f(y)\|}{d(x,y)}$$

$$\text{contraction}(f) = \max_{x,y \in X} \frac{d(x,y)}{\|f(x) - f(y)\|}$$

$$\text{distortion}(f) = \text{expansion}(f) \cdot \text{contraction}(f)$$

Definition 11.1.4. *The* optimal distortion *of* (X, d) *is given by*

$$c_2(X, d) = \min_{f: X \to \mathbb{R}^{|X|}} \text{distortion}(f).$$

In the case when (X, d) *is the shortest path metric of a graph* G *we write* $c_2(G)$.

11.2 Computing optimal Euclidean embeddings

Let (X, d) be a finite metric space with $X = \{x_1, \ldots, x_n\}$. Then we can find a Euclidean embedding of X which minimizes the distortion by solving a semidefinite optimization problem. For this let $f : X \to \mathbb{R}^n$ be an embedding. Then we can assume by scaling that $\text{contraction}(f) = 1$. So we have to minimize $\text{expansion}(f)$ to minimize the distortion. The following optimization problem does this:

$$\begin{aligned} \text{minimize} \quad & \gamma^2 \\ & \gamma \in \mathbb{R},\ f : X \to \mathbb{R}^n \\ & d(x_i, x_j)^2 \leqslant \|f(x_i) - f(x_j)\|^2 \leqslant \gamma^2 d(x_i, x_j)^2 \end{aligned}$$

By considering the inner product matrix $Z = (f(x_i)^\mathsf{T} f(x_j))_{1 \leqslant i,j \leqslant n}$, which is positive semidefinite, and by noting that

$$\|f(x_i) - f(x_j)\|^2 = Z_{ii} - 2Z_{ij} + Z_{jj} = \langle e_i e_i^\mathsf{T} + e_j e_j^\mathsf{T} - (e_i e_j^\mathsf{T} + e_j e_i^\mathsf{T}), Z \rangle$$

we get a semidefinite optimization problem

$$\begin{aligned} \text{minimize} \quad & \tau \\ & \tau \in \mathbb{R},\ Z \in \mathcal{S}^n_{\geq 0} \\ & \langle e_i e_i^\mathsf{T} + e_j e_j^\mathsf{T} - (e_i e_j^\mathsf{T} + e_j e_i^\mathsf{T}), Z \rangle \geqslant d(x_i, x_j)^2 \\ & \langle e_i e_i^\mathsf{T} + e_j e_j^\mathsf{T} - (e_i e_j^\mathsf{T} + e_j e_i^\mathsf{T}), Z \rangle \leqslant \tau d(x_i, x_j)^2 \end{aligned}$$

for which $\sqrt{\tau} = c_2(X, d)$ holds.

By using strong duality of conic programming (Exercise 11.1 (a)) we arrive at the following theorem.

Theorem 11.2.1. *The least distortion of a finite metric space* (X, d)*, with* $X = \{x_1, \ldots, x_n\}$*, into Euclidean space is given by*

$$c_2(X, d) = \max_{Y \in \mathcal{S}^n_{\geq 0}, Ye = 0} \sqrt{\frac{\sum_{ij: Y_{ij} > 0} Y_{ij} d(x_i, x_j)^2}{-\sum_{ij: Y_{ij} < 0} Y_{ij} d(x_i, x_j)^2}}.$$

The condition $Ye = 0$ *says that the all-ones vector* e *lies in the kernel of* Y.

We will use this theorem to find lower bounds for the optimal distortion embeddings of several graphs.

11.2.1 Least distortion embedding of the cube

To warm up we consider the graph of the r-dimensional unit cube $Q_r = (V_r, E_r)$ with $V_r = \{0,1\}^r$. Here two vertices are adjacent whenever their Euclidean distance equals 1. Clearly (check it), the usual Euclidean embedding has distortion $\sqrt{r}$. In fact, as the following theorem shows, one cannot improve it.

Theorem 11.2.2.

$$c_2(Q_r) = \sqrt{r}.$$

Proof. Define the matrix $Y \in \mathbb{R}^{V_r \times V_r}$ by

$$Y(i,j) = \begin{cases} -1 & \text{if } d(i,j) = 1, \\ r-1 & \text{if } i = j, \\ 1 & \text{if } d(i,j) = r, \\ 0 & \text{otherwise.} \end{cases}$$

It satisfies the properties of Theorem 11.2.1. We clearly have $Ye = 0$. The fact that Y is positive semidefinite follows from the fact that for $y \in \{0,1\}^r$ the vectors $f_y \in \mathbb{R}^{V_r}$ defined by $f_y(x) = (-1)^{x^\mathsf{T} y}$ form a basis of eigenvectors of Y. One directly verifies that the corresponding eigenvalues are nonnegative. To end the proof we only have to evaluate Y's objective value:

$$\sum_{ij:Y_{ij}>0} Y_{ij} d(i,j)^2 = 2^r r^2$$

and

$$-\sum_{ij:Y_{ij}<0} Y_{ij} d(i,j)^2 = 2^r r.$$

Hence,

$$c_2(Q_r) \geqslant \sqrt{\frac{2^r r^2}{2^r r}} = \sqrt{r}. \qquad \square$$

11.3 Corner stones of metric embeddings

11.3.1 Bourgain's theorem

Bourgain showed in 1985 that every finite metric space embeds into Euclidean space with low distortion. This theorem is according to Hoory, Linial, and Wigderson the "grand ancestor" of the area of metric embeddings.

Theorem 11.3.1. *There is a constant C so that any finite metric space (X,d) can be embedded into Euclidean space with distortion at most $C \log |X|$:*

$$c_2(X,d) = O(\log |X|).$$

In particular it shows that the optimal solution of the semidefinite optimization problem in Theorem 11.2.1 is bounded by $C \log |X|$. The proof is presented in Chapter 15.7 of the book by Matoušek [5]. However, currently, there is no proof known which is based on semidefinite optimization. In fact, Goemans [3] writes: "it would be nice to prove this result from semidefinite programming duality."

11.3.2 Johnson-Lindenstrauss flattening lemma

Another major result in the area of metric embeddings with many applications is by Johnson and Lindenstrauss from 1984. It says that one can reduce the dimension of Euclidean embeddings significantly.

Theorem 11.3.2. *Let (X, d) be a finite metric space which isometrically embeds into Euclidean space of dimension $|X|$. Then there is an embedding of X into a Euclidean space of dimension $O(\log |X|/\epsilon^2)$ with distortion at most $1 + \epsilon$.*

The construction behind the proof (see Theorem 15.2.1 in Matoušek [5]) is very simple: One uses a random linear projection onto a low dimensional subspace.

11.4 Embeddings of expanders

An expander is a graph which is sparse but at the same time highly connected. Expanders are remarkable graphs which have many applications in mathematics and computer science. In the last forty years they were subject of a huge amount of research.

Here we will use them to show that Bourgain's theorem is tight in the sense that the shortest path metric on expander graphs can only be embedded into Euclidean space with distortion $\Omega(\log n)$.

For this we start by defining the edge expansion ratio. Although this definition gives some intuition how expander graphs look like it is frequently much easier to work with expanders algebraically using spectral properties of their adjacency matrix. These spectral properties will then be useful for proving that expanders embed rather badly into Euclidean space.

11.4.1 Edge expansion

Let $G = (V, E)$ be a graph. We assume that in G every vertex has exactly d neighbors, i.e. that G is *d-regular*. Let $S \subseteq V$ be a subset of the vertices and let $\overline{S} = V \backslash S$ be its complement. The *edge boundary* of S is

$$\partial S = \{\{u, v\} \in E : u \in S, v \in \overline{S}\}.$$

For the edges which stay in S or $\overline{S}$ define

$$E(S) = \{\{u, v\} \in E : u, v \in S\}, \quad E(\overline{S}) = \{\{u, v\} \in E : u, v \in \overline{S}\}.$$

Definition 11.4.1. *The* edge expansion ratio *of a graph G is*

$$h(G) = \min_{S \subseteq V : |S| \leqslant |V|/2} \frac{|\partial S|}{|S|}$$

Definition 11.4.2. *Let $d \geqslant 3$ be an integer. A family of d-regular graphs $G_n = (V_n, E_n)$ with $|V_n| \to \infty$ when n tends to infinity is called a* family of d-regular expander graphs *if there exists $\epsilon > 0$ with $h(G_n) > \epsilon$.*

In the following two sections we will prove a fundamental inequality due to Dodziuk (1984) and independently Alon and Milman (1985) and Alon (1986). It relates the edge expansion ratio $h(G)$ of a d-regular graph with the *spectral gap* of a graph, the difference $d - \lambda_2$ between the largest and the second largest eigenvalue of its adjacency matrix

$$\boxed{\frac{d-\lambda_2}{2} \leqslant h(G) \leqslant \sqrt{2d(d-\lambda_2)}}$$

This shows that G_n is a family of d-regular expander graphs if and only if there exists an $\epsilon > 0$ so that $d - \lambda_2(G_n) > \epsilon$ for all n.

11.4.2 Large spectral gap implies high expansion

Theorem 11.4.3. *Let $G = (V, E)$ be a connected, d-regular graph. Let $\lambda_1 = d$ and λ_2 be the largest and the second largest eigenvalue of the adjacency matrix of G. Then,*

$$\frac{d-\lambda_2}{2} \leqslant h(G).$$

Proof. The largest eigenvalue of the adjacency matrix A of the d-regular graph G equals d and the corresponding eigenvector is the all-ones vector e (see Exercise 11.2). So the second largest eigenvalue λ_2 of A is given by

$$\lambda_2 = \max_{f \in \mathbb{R}^V \setminus \{0\}, f \perp e} \frac{f^\mathsf{T} A f}{f^\mathsf{T} f}$$

because of the Rayleigh principle. If we would find a vector f which is perpendicular to e so that

$$\frac{f^\mathsf{T} A f}{f^\mathsf{T} f} \geqslant d - 2h(G)$$

holds, then we would prove the desired inequality. Let $S \subseteq V$ be a set attaining the edge expansion ratio

$$h(G) = \frac{|\partial S|}{|S|}, \quad \text{with } |S| \leqslant |V|/2.$$

Define the vector

$$f = |\overline{S}|\chi^S - |S|\chi^{\overline{S}} \in \mathbb{R}^V$$

where $\chi^S \in \mathbb{R}^V$ denotes the characteristic vector of the set S. This vector is perpendicular to e. The denominator of the Rayleigh quotient equals

$$\begin{aligned} f^\mathsf{T} A f &= 2 \sum_{\{u,v\} \in E} f(u) f(v) \\ &= 2 \left(|E(S)||\overline{S}|^2 + |E(\overline{S})||S|^2 - |S||\overline{S}||\partial S| \right) \\ &= (d|S| - |\partial S|)|\overline{S}|^2 + (d|\overline{S}| - |\partial S|)|S|^2 - 2|S||\overline{S}||\partial S| \\ &= d(|\overline{S}| + |S|)|S||\overline{S}| - (|\overline{S}| + |S|)^2 |\partial S| \\ &= d|V||S||\overline{S}| - |V|^2 |\partial S|, \end{aligned}$$

where we first split the sum into $\{u,v\} \in E(S)$, $\{u,v\} \in E(\overline{S})$, and $\{u,v\} \in \partial S$, and then use the identities

$$d|S| = 2|E(S)| + |\partial S|, \quad d|\overline{S}| = 2|E(\overline{S})| + |\partial S|, \quad |V| = |S| + |\overline{S}|.$$

The numerator of the Rayleigh quotient equals

$$f^\mathsf{T} f = |\overline{S}|^2 |S| + |S|^2 |\overline{S}| = |S||\overline{S}|(|\overline{S}| + |S|) = |V||S||\overline{S}|.$$

Together,

$$\frac{f^\mathsf{T} A f}{f^\mathsf{T} f} = \frac{d|V||S||\overline{S}| - |V|^2 |\partial S|}{|V||S||\overline{S}|} = d - \frac{n|\partial S|}{|S||\overline{S}|} \geqslant d - 2h(G),$$

where we use that $h(G) = |\partial S|/|S|$ and $|\overline{S}| \geqslant |V|/2$. □

11.4.3 High expansion implies large spectral gap

Theorem 11.4.4. *Let $G = (V, E)$ be a connected, d-regular graph. Let $\lambda_1 = d$ and λ_2 be the largest and the second largest eigenvalue of the adjacency matrix of G. Then,*

$$h(G) \leqslant \sqrt{2d(d - \lambda_2)}.$$

Proof. Let g be an eigenvector of the adjacency matrix A of G corresponding to λ_2. Since g is perpendicular to the all-ones vector, the vector g has positive as well as negative entries. Define $f \in \mathbb{R}^V$ by

$$f(u) = \begin{cases} g(u) & \text{if } g(u) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Let $S = \{u \in V : f(u) \neq 0\}$ be the support of f. We may assume that S has at most $|V|/2$ vertices, otherwise we would replace the eigenvector g by its negative $-g$.

The theorem will follow once we prove the inequalities

$$\frac{h(G)^2}{2d} \leqslant \frac{f^\mathsf{T} L f}{f^\mathsf{T} f} \leqslant d - \lambda_2 \tag{11.1}$$

for the Laplacian matrix $L = dI - A$ of the d-regular graph G.

The upper bound in (11.1) is (relatively) easy: For $u \in S$ we have

$$\begin{aligned}(Lf)(u) &= df(u) - \sum_{v\in V:\{u,v\}\in E} f(v)\\ &= dg(u) - \sum_{v\in S:\{u,v\}\in E} g(v)\\ &\leqslant dg(u) - \sum_{v\in V:\{u,v\}\in E} g(v)\\ &= (d-\lambda_2)g(u).\end{aligned}$$

Because $f(u) = 0$ whenever $u \notin S$ we arrive at

$$f^{\mathsf{T}}Lf = \sum_{u\in V} f(u)(Lf)(u) \leqslant (d-\lambda_2)\sum_{u\in S} g(u)^2 = (d-\lambda_2)f^{\mathsf{T}}f.$$

The lower bounds in (11.1) is harder and needs more work and ingenuity.

Some preparation: Let us label the vertices of G by $1,\ldots,|V|$ so that

$$f(1) \geqslant f(2) \geqslant \ldots \geqslant f(|V|).$$

Direct the edges of the graph G (arbitrarily) and define $K \in \mathbb{R}^{V\times E}$ by

$$K(u,e) = \begin{cases} +1 & \text{if edge } e \text{ enters vertex } u,\\ -1 & \text{if edge } e \text{ exits vertex } u,\\ 0 & \text{otherwise.}\end{cases}$$

Then one has $L = KK^{\mathsf{T}}$. Define the quantity

$$B = \sum_{\{u,v\}\in E} |f(u)^2 - f(v)^2|.$$

We shall prove

$$h(G)f^{\mathsf{T}}f \leqslant B \leqslant \sqrt{2d}\sqrt{(Kf)^{\mathsf{T}}Kf}\sqrt{f^{\mathsf{T}}f}, \tag{11.2}$$

which implies the lower bound in (11.1) because $f^{\mathsf{T}}Lf = (Kf)^{\mathsf{T}}(Kf)$.

The upper bound in (11.2) follows from Cauchy-Schwarz

$$\begin{aligned}B &= \sum_{\{u,v\}\in E} |f(u)^2 - f(v)^2|\\ &= \sum_{\{u,v\}\in E} |f(u)+f(v)|\cdot|f(u)-f(v)|\\ &\leqslant \sqrt{\sum_{\{u,v\}\in E}(f(u)+f(v))^2}\cdot\sqrt{\sum_{\{u,v\}\in E}(f(u)-f(v))^2}\end{aligned}$$

and by

$$\sqrt{\sum_{\{u,v\}\in E}(f(u)-f(v))^2} = \sqrt{(Kf)^\mathsf{T} Kf}$$

as well by

$$\sqrt{\sum_{\{u,v\}\in E}(f(u)+f(v))^2} \leqslant \sqrt{2\sum_{\{u,v\}\in E}(f(u)^2+f(v)^2)} = \sqrt{2d\sum_{u\in V} f(u)^2} = \sqrt{2d} f^\mathsf{T} f.$$

The lower bound in (11.2) follows from the following calculation which uses telescopic summation and the ordering of the vertices of G:

$$\begin{aligned}
B &= \sum_{\{u,v\}\in E} |f(u)^2 - f(v)^2| \\
&= \sum_{\{u,v\}\in E, u<v} (f(u)^2 - f(v)^2) \\
&= \sum_{\{u,v\}\in E, u<v} \sum_{i=u}^{v-1} (f(i)^2 - f(i+1)^2) \\
&= \sum_{i=1}^{|V|-1} (f(i)^2 - f(i+1)^2)|\partial\{1,\ldots,i\}| \\
&= \sum_{i\in S} (f(i)^2 - f(i+1)^2)|\partial\{1,\ldots,i\}| \\
&\geqslant h(G)\sum_{i\in S}(f(i)^2 - f(i+1)^2)i \\
&= h(G)\sum_{i\in S}(f(i))^2 \\
&= h(G)\sqrt{f^\mathsf{T} f}.
\end{aligned}$$

Here we use the fact that $|S| \leqslant |V|/2$ and so $|\partial\{1,\ldots,i\}|/i \geqslant h(G)$ if $i \leqslant |V|/2$. Furthermore, notice that $f(i+1) = 0$ for $i = |S|$ when collapsing the telescopic sum. □

11.4.4 Low distortion embeddings of expander graphs

Theorem 11.4.5. *Let $d \geqslant 3$ be an integer and let $\epsilon > 0$ be a positive real. For every d-regular graph $G = (V, E)$ and $\lambda_2 \leqslant d - \epsilon$, we have*

$$c_2(G) \geqslant \sqrt{\frac{\epsilon}{2d}}\lfloor \log_d |V| \rfloor.$$

In particular, Bourgain's theorem is tight for families of d-regular expander graphs.

Proof. For simplicity we assume that $|V|$ is even.

Since G is d-regular, every vertex has $\leqslant d^r$ vertices at distance r. In particular if $r = \lfloor \log_d |V| \rfloor - 1$, then there are $\leqslant |V|/2$ vertices at distance r from any given vertex. Define the graph $H = (V, E_H)$ by connecting two vertices if their distance in G is $\geqslant \lfloor \log_d |V| \rfloor$. Then the minimal degree of H is $\geqslant |V|/2$. By a classical theorem of Dirac from 1952 we know that every graph on $|V| \geqslant 3$ vertices with minimum degree at least $|V|/2$ contains a Hamiltonian cycle; one can find the (simple) proof for instance as Theorem 10.1.1 in the book [1] by Diestel. Since $|V|$ is even, we derive that H has a perfect matching.

Let $B \in \mathcal{S}^V$ be the adjacency matrix of such a perfect matching. It is a permutation matrix of a permutation consisting out of $|V|/2$ disjoint transpositions. We denote the edges participating in the perfect matching by F.

Let $A \in \mathcal{S}^V$ be the adjacency matrix of G.

Define the matrix Y by

$$Y = dI - A + \frac{\epsilon}{2}(B - I),$$

and we want to show that Y satisfies the assumptions of Theorem 11.2.1.

It is easy to verify that $Ye = 0$ holds. The matrix Y is positive semidefinite because for every $x \in \mathbb{R}^V$ which is perpendicular to e we have the inequality

$$\begin{aligned}
x^{\mathsf{T}} Y x &= x^{\mathsf{T}}(dI - A + \frac{\epsilon}{2}(B - I))x \\
&\geqslant (d - \lambda_2)x^{\mathsf{T}}x + \frac{\epsilon}{2}x^{\mathsf{T}}(B - I)x \\
&\geqslant \epsilon x^{\mathsf{T}}x + \frac{\epsilon}{2}\sum_{\{u,v\}\in F}(2x(u)x(v) - x(u)^2 - x(v)^2) \\
&\geqslant \epsilon x^{\mathsf{T}}x - \frac{\epsilon}{2}2\sum_{\{u,v\}\in F}(x(u)^2 + x(v)^2) \\
&\geqslant \epsilon x^{\mathsf{T}}x - \epsilon x^{\mathsf{T}}x \\
&\geqslant 0.
\end{aligned}$$

To end the proof we only have to evaluate Y's objective value:

$$-\sum_{ij:Y_{ij}<0} Y_{ij} d(x_i, x_j)^2 = d|V|$$

and

$$\sum_{ij:Y_{ij}>0} Y_{ij} d(x_i, x_j)^2 \geqslant \frac{\epsilon}{2}|V|\lfloor \log_d |V| \rfloor^2.$$

Hence,

$$c_2(G) \geqslant \sqrt{\frac{\epsilon}{2d}}\lfloor \log_d |V| \rfloor.$$

□

11.4.5 Construction of a family of expander graphs

Explicit constructions of a family of expander graphs are very much non-trivial. An easy construction of a family of 3-regular expander graphs which nevertheless relies on a deep result in number theory (Selberg's $3/16$ theorem) is as follows: Let p be a prime. The vertex set of G_p is $\mathbb{Z}_p$ and a vertex x is connected to $x+1$, $x-1$ and x^{-1} where all operations are performed modulo p and where the inverse of 0 is defined to be 0.

11.5 Further reading

In this lecture we mostly followed the presentation of the price winning, very fascinating, survey article by Hoory, Linial, Wigderson [2] on expander graphs (especially Section 13). There the authors present the many, often surprising, connections of expanders with other parts of mathematics and computer science. The recent survey [4] by Lubotzky is fascinating too. It focuses on the deep algebraic side of expanders.

Much more on metric embeddings and its applications can be found in Chapter 15 of Matoušek's book on discrete geometry [5].

11.6 Exercises

11.1** (a) Prove Theorem 11.2.1.

(b) Show: Let $f : X \to \mathbb{R}^n$ be an optimal distortion embedding. If Y attains the optimum in Theorem 11.2.1 then $Y_{ij} > 0$ only for f's most contracted pairs i and j and $Y_{ij} < 0$ only for f's most expanded pairs i and j.

(c) Find an optimal distortion embedding of the Petersen graph (see Figure 6.1).

11.2 Let $G = (V, E)$ be a d-regular graph and let

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$$

be the eigenvalues of the adjacency matrix of G. Show that

(a) $\lambda_i \in [-d, d]$ for all $i = 1, \ldots, n$.

(b) G is connected if and only if $\lambda_1 > \lambda_2$.

(c) G is bipartite if and only if $\lambda_1 = -\lambda_n$.

(d) $\lambda_2^2 \geqslant d\frac{|V|-d}{|V|-1}$.

11.3 Let $G = (V, E)$ be a d-regular graph and let λ_2 the second largest eigenvalue of its adjacency matrix. Then for $S, T \subseteq V$ we have

$$\left| |\{\{u, v\} \in E : u \in S, v \in T\}| - d\frac{|S||T|}{|V|} \right| \leqslant \lambda_2\sqrt{|S||T|}.$$

11.4 Computer exercise: Compute the optimal distortion embedding of the semimetric in Table 11.1 and draw a random projection onto the two-dimensional Euclidean plane.

BIBLIOGRAPHY

[1] R. Diestel, *Graph Theory*, Springer, 2010.

`http://diestel-graph-theory.com/`

[2] S. Hoory, N. Linial, A. Wigderson, *Expander graphs and their applications*, Bull. Amer. Math. Soc. **43** (2006), 439–561.

`www.ams.org/bull/2006-43-04/S0273-0979-06-01126-8/`

[3] M.X. Goemans, *Semidefinite programming in combinatorial optimization*, Math. Program. **79** (1997), 143–161.

`http://math.mit.edu/~goemans/PAPERS/semidef-survey.ps`

[4] A. Lubotzky *Expander graphs in pure and applied mathematics*, Bull. Amer. Math. Soc. **49** (2012), 113–162.

`http://www.ams.org/journals/bull/2012-49-01/`
`S0273-0979-2011-01359-3/home.html`

[5] J. Matoušek, *Lectures on discrete geometry*, Springer 2002.

`http://kam.mff.cuni.cz/~matousek/dg-nmetr.ps.gz`

[6] P. Sarnak, *What is . . . an Expander?*, Notices of the AMS **51** (2004), 762–763.

`http://www.ams.org/notices/200407/what-is.pdf`

CHAPTER 12

PACKINGS ON THE SPHERE

Packing problems are fundamental in geometric optimization and coding theory: How densely can one pack given objects into a given container?

In this lecture the container will be the unit sphere

$$S^{n-1} = \{x \in \mathbb{R}^n : x \cdot x = 1\}$$

and the objects we want to pack are spherical caps of angle γ. The *spherical cap* with angle $\gamma \in [0, \pi]$ and center $x \in S^{n-1}$ is given by

$$C(x, \gamma) = \{y \in S^{n-1} : x \cdot y \geqslant \cos\gamma\}.$$

Its normalized volume equals (by integration with spherical coordinates)

$$w(\gamma) = \frac{\omega_{n-1}(S^{n-2})}{\omega_n(S^{n-1})} \int_{\cos\gamma}^{1} (1-u^2)^{(n-3)/2}\, du,$$

where $\omega_n(S^{n-1}) = (2\pi^{n/2})/\Gamma(n/2)$ is the surface area of the unit sphere. Two spherical caps $C(x_1, \gamma)$ and $C(x_2, \gamma)$ intersect in their topological interior if and only if the inner product of x_1 and x_2 lies in the half-open interval $(\cos(2\gamma), 1]$. Conversely we have

$$C(x_1, \gamma)^\circ \cap C(x_2, \gamma)^\circ = \varnothing \quad \Longleftrightarrow \quad -1 \leqslant x_1 \cdot x_2 \leqslant \cos(2\gamma).$$

A *packing* of spherical caps with angle γ, is a collection of any number of spherical caps with this angle and pairwise-disjoint topological interiors. Given the dimension n and the angle γ we define[1]

$$A(n, 2\gamma) = \max\{N : C(x_1, \gamma), \ldots, C(x_N, \gamma) \text{ is a packing in } S^{n-1}\}.$$

[1]Note here that we use 2γ in the definition of $A(n, 2\gamma)$ because we want to make the notation consistent with the common literature. There one emphasizes that 2γ is the angle between the centers of the spherical caps.

One particular case of packings of spherical caps has received a lot of attention over the last centuries.

In geometry, the *kissing number* τ_n is the maximum number of non-overlapping equally-sized spheres that can simultaneously touch a central sphere. It is easy to see that $\tau_n = A(n, \pi/3)$ because the points where the spheres touch the central sphere form the centers of a packing of spherical caps with angle $\pi/6$.

Today, the kissing number is only known for dimensions 1, 2, 3, 4, 8 and 24. It is easy to see that the kissing number in dimension 1 is 2, and in dimension 2 it is 6. The kissing number problem has a rich history. In 1694 Isaac Newton and David Gregory had a famous discussion about the kissing number in three dimensions. The story is that Gregory thought thirteen spheres could fit while Newton believed the limit was twelve. Note that the easy area argument, which proves $\tau_2 = 6$, only gives that

$$\tau_3 \leqslant \left\lfloor \frac{1}{w(\pi/3)} \right\rfloor = \left\lfloor \frac{4\pi}{2\pi(1-\cos(\pi/6))} \right\rfloor = \lfloor 14.92\ldots \rfloor = 14.$$

It took many years, until 1953, when Schütte and van der Waerden proved Newton right.

Figure 12.1: Construction of 12 kissing spheres. Image credit: Anja Traffas

In the 1970s advanced methods to determine upper bounds for the kissing number based on linear programming were introduced. Using these new techniques, the kissing number problem in dimension 8 and 24 was solved by Odlyzko, Sloane, and Levensthein. For four dimensions, however, the optimization bound is 25, while the exact kissing number is 24. In a celebrated work Oleg Musin proved this in 2003, see [3].

The goal of this lecture is to provide a proof of $\tau_8 = 240$.

12.1 α and ϑ for packing graphs

Many, often notoriously difficult, problems in combinatorics and geometry can be modeled as packing problems of graphs $G = (V, E)$ where the vertex set V can be an infinite or even a continuous set. All possible positions of the objects which we can use for the packing are vertices of a graph and we draw edges between two vertices whenever the two corresponding objects cannot be

simultaneously present in the packing because they overlap in their interior. Now every independent set in this conflict graph gives a valid packing.

For the problem of determining the optimal packing of spherical caps with angle γ, $A(n, 2\gamma)$, we define the packing graph $G(n, 2\gamma)$ with vertex set

$$V = S^{n-1} = \{x \in \mathbb{R}^n : x \cdot x = 1\},$$

and edge set

$$x \sim y \Longleftrightarrow x \cdot y \in (\cos(2\gamma), 1).$$

Then,

$$A(n, 2\gamma) = \alpha(G(n, 2\gamma)), \quad \text{and} \quad \tau_n = A(n, \pi/3) = \alpha(G(n, \pi/3))$$

Now it is an "obvious" strategy to compute the theta number for this graph in order to find upper bounds for the independence number $\alpha(G(n, 2\gamma))$.

To generalize the theta number for infinite graphs, we will need a notion of positive semidefinite *infinite matrices* because in the definition of the theta number we need matrices whose rows and columns are indexed by the vertex set of the graph.

This leads to positive semidefinite, continuous *Hilbert-Schmidt kernels*.

Definition 12.1.1. *A continuous function (called continuous Hilbert-Schmidt kernel)*

$$K : S^{n-1} \times S^{n-1} \to \mathbb{R}$$

is called symmetric *if $K(x, y) = K(y, x)$ holds for all $x, y \in S^{n-1}$. It is called* positive semidefinite *if for all N and all $x_1, \ldots, x_N \in S^{n-1}$ the symmetric $N \times N$ matrix*

$$\big(K(x_i, x_j)\big)_{1 \leqslant i,j \leqslant N} \geq 0$$

is positive semidefinite. We denote the cone of positive semidefinite continuous Hilbert-Schmidt kernels by $\mathcal{C}(S^{n-1} \times S^{n-1})_{\geq 0}$

We use this cone $\mathcal{C}(S^{n-1} \times S^{n-1})_{\geq 0}$ to define the theta prime number of the packing graph $G(n, 2\gamma)$:

$$\begin{aligned} \vartheta'(G(n, 2\gamma)) = \inf \quad & \lambda \\ & K \in \mathcal{C}(S^{n-1} \times S^{n-1})_{\geq 0} \\ & K(x, x) = \lambda - 1 \text{ for all } x \in S^{n-1} \\ & K(x, y) \leqslant -1 \text{ for all } \{x, y\} \notin E. \end{aligned}$$

We have $\{x, y\} \notin E$ whenever the spherical caps $C(x, \gamma)$ and $C(y, \gamma)$ do not intersect in their topological interior, i.e. whenever $x \cdot y \in [-1, \cos(2\gamma)]$. The definition ϑ' is similar to the dual formulations in Lemma 6.4.1. We use a prime to indicate that we replace the equality $K(x, y) = -1$ by the inequality $K(x, y) \leqslant -1$.

Similar to the finite case, ϑ' provides an upper bound for the independence number:

Theorem 12.1.2.

$$\alpha(G(n, 2\gamma)) \leqslant \vartheta'(G(n, 2\gamma))$$

Proof. Let $C \subseteq S^{n-1}$ be an independent set. Let K be a feasible solution of $\vartheta'(G(n, 2\gamma))$. Because K is positive semidefinite we have

$$\begin{aligned} 0 &\leqslant \sum_{x\in C}\sum_{y\in C} K(x,y) \\ &= \underbrace{\sum_{x\in C} K(x,x)}_{=|C|(\lambda-1)} + \underbrace{\sum_{x\neq y} K(x,y)}_{\leqslant(-1)(|C|^2-|C|)} \\ &\leqslant |C|(\lambda-1) - (|C|^2 - |C|) \end{aligned}$$

This implies $|C| \leqslant \lambda$, yielding the theorem. □

Note that if we are in the lucky case that $\alpha(G(n, 2\gamma)) = \vartheta'(G(n, 2\gamma))$, the inequalities in the proof of the theorem are tight. This can only happen when $K(x, y) = -1$ for $\{x, y\} \notin E$. We will use this observation later when we determine τ_8.

12.2 Symmetry reduction

Computing ϑ' does not seem to be easy since it is defined as an infinite-dimensional semidefinite program. However, the underlying graph is highly symmetric and so we can perform symmetry reduction, similar to the one in Chapter 6.6.

The automorphism group of the graph $G(n, 2\gamma)$ is the orthogonal group $\mathcal{O}(n)$ because for all $A \in \mathcal{O}(n)$ we have

$$Ax \cdot Ay = x \cdot y.$$

Furthermore the graph $G(n, 2\gamma)$ is vertex transitive because for every two points x and y on the unit sphere there is an orthogonal matrix mapping x to y. Even stronger it is *two-point homogeneous*, meaning that if $x, y, x', y' \in S^{n-1}$ are so that

$$x \cdot y = x' \cdot y',$$

then there is an $A \in \mathcal{O}(n)$ with $Ax = x'$, $Ay = y'$.

If K is a feasible solution for ϑ' with objective value $\lambda = K(x, x) + 1$ and if $A \in \mathcal{O}(n)$ is an orthogonal matrix then also

$$K^A(x, y) = K(Ax, Ay)$$

is a feasible solution for ϑ' with the same objective value. So we can symmetrize any feasible solution K of ϑ'

$$K'(x, y) = \int_{A\in\mathcal{O}(n)} K^A(x, y) d\mu(A),$$

where μ is the normalized Haar measure of the orthogonal group.

That means that we can restrict the optimization variable K to be a positive semidefinite continuous Hilbert-Schmidt kernel which is invariant under the orthogonal group, i.e.

$$K \in \mathcal{C}(S^{n-1} \times S^{n-1})^{\mathcal{O}(n)}_{\geq 0},$$

where

$$\begin{aligned}&\mathcal{C}(S^{n-1} \times S^{n-1})^{\mathcal{O}(n)}_{\geq 0}\\ &\quad = \{K \in \mathcal{C}(S^{n-1} \times S^{n-1}) : K^A(x,y) = K(Ax, Ay) = K(x,y) \text{ for all } A \in \mathcal{O}(n)\}.\end{aligned}$$

So we get

$$\begin{aligned}\vartheta'(G(n, 2\gamma)) = \inf \quad & \lambda \\ & K \in \mathcal{C}(S^{n-1} \times S^{n-1})^{\mathcal{O}(n)}_{\geq 0} \\ & K(x,x) = \lambda - 1 \text{ for all } x \in S^{n-1} \\ & K(x,y) \leqslant -1 \text{ for all } \{x,y\} \notin E.\end{aligned}$$

12.3 Schoenberg's theorem

Now the idea is to find an explicit characterization of the cone $\mathcal{C}(S^{n-1} \times S^{n-1})^{\mathcal{O}(n)}_{\geq 0}$. Such a characterization was proved by Schoenberg in 1941. He parameterized this cone by its extreme rays.

Theorem 12.3.1 (Schoenberg (1941))**.**

$$\mathcal{C}(S^{n-1} \times S^{n-1})^{\mathcal{O}(n)}_{\geq 0} = \left\{ \sum_{k=0}^{\infty} f_k E^n_k(x,y) : f_k \geqslant 0, \sum_{k=0}^{\infty} f_k < \infty \right\}, \qquad (12.1)$$

where

$$E^n_k(x,y) = P^n_k(x \cdot y),$$

and where P^n_k is a polynomial of degree k satisfying the orthogonality relation

$$\int_{-1}^{1} P^n_k(t) P^n_l(t) (1-t^2)^{\frac{n-3}{2}} dt = 0 \text{ if } k \neq l,$$

and where the polynomial P^n_k is normalized by $P^n_k(1) = 1$.

The equality in (12.1) should be interpreted as follows: A kernel K lies in $\mathcal{C}(S^{n-1} \times S^{n-1})^{\mathcal{O}(n)}$ if and only if there are nonnegative numbers $f_0, f_1, \ldots$ so that the series $\sum_{k=0}^{\infty} f_k$ converges and so that

$$K(x,y) = \sum_{k=0}^{\infty} f_k E^n_k(x,y)$$

holds. Here the right hand side converges absolutely and uniformly over $S^{n-1} \times S^{n-1}$.

For $n = 2$, P_k^2 are the Chebyshev polynomials (of the first kind). For larger n the polynomials belong to the family of Jacobi polynomials. The *Jacobi polynomials* with parameters (α, β) are orthogonal polynomials for the measure $(1-t)^\alpha(1+t)^\beta dt$ on the interval $[-1, 1]$. They form a complete orthogonal system of the space $L^2([-1,1], (1-t)^\alpha(1+t)^\beta dt)$. This space consists of all real-valued functions $f : [-1, 1] \to \mathbb{R}$ for which the integral

$$\int_{-1}^{1} f^2(t)(1-t)^\alpha(1+t)^\beta dt$$

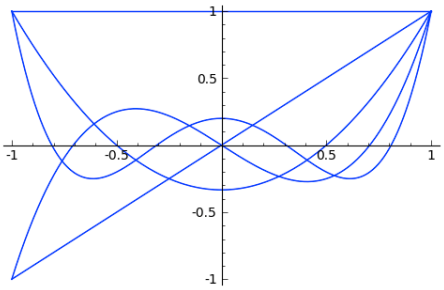
exists and is finite. We denote by $P_k^{(\alpha,\beta)}$ the *normalized Jacobi polynomial* of degree k with normalization $P_k^{(\alpha,\beta)}(1) = 1$. The first few normalized Jacobi polynomials with parameter (α, α) and $\alpha = (n-3)/2$ are

$$\begin{aligned} P_0^n(t) &= P_0^{(\alpha,\alpha)}(t) = 1, \\ P_1^n(t) &= P_1^{(\alpha,\alpha)}(t) = t, \\ P_2^n(t) &= P_2^{(\alpha,\alpha)}(t) = \frac{n}{n-1}t^2 - \frac{1}{n-1}. \end{aligned}$$

Much more information is known about these orthogonal polynomials. They are also known to many computer algebra systems.

```
sage: x = PolynomialRing(QQ, 'x').gen()
sage: n = 4
sage: a = (n-3)/2
sage: for k in range(0,5):
sage:     print(jacobi_P(k,a,a,x)/jacobi_P(k,a,a,1))

1
x
4/3*x^2 - 1/3
2*x^3 - x
16/5*x^4 - 12/5*x^2 + 1/5
```



12.4 Proof of Schoenberg's theorem

In this section we prove Theorem 12.3.1 in three steps. In the first two steps we derive some properties of the extreme rays E_k^n.

12.4.1 Orthogonality relation

The space of symmetric continuous Hilbert-Schmidt kernel is an inner product space, just like the space of symmetric matrices. The inner product between K and L is

$$\langle K, L\rangle = \int_{S^{n-1}} K(x,y)L(x,y)d\omega_n(x)d\omega_n(y).$$

Lemma 12.4.1. *We have the orthogonality relation $E_k^n \perp E_l^n$ whenever $k \neq l$.*

Proof. Since $E_k^n(x,y) = P_k^n(x \cdot y)$ and the integrals are invariant under $\mathcal{O}(n)$, we can take $x = N$, where N is the North Pole, and therefore,

$$\begin{aligned}\langle E_k^n, E_l^n\rangle &= \omega_n(S^{n-1})\int_{S^{n-1}} P_k^n(N\cdot y)P_l^n(N\cdot y)d\omega_n(y)\\ &= \omega_n(S^{n-1})\omega_{n-1}(S^{n-2})\int_{-1}^{1} P_k^n(t)P_l^n(t)(1-t^2)^{\frac{n-3}{2}}dt\\ &= 0,\end{aligned}$$

if $k \neq l$. □

12.4.2 Positive semidefiniteness

Lemma 12.4.2. *The E_k^n's are positive semidefinite.*

Proof. Let us consider the space of continuous functions $f : S^{n-1} \to \mathbb{R}$ with inner product

$$(f,g) = \int_{S^{n-1}} f(x)g(x)d\omega_n(x).$$

Let V_0 be the space of constant functions on S^{n-1} and, for $k \geqslant 1$, let V_k be the space of polynomial functions on S^{n-1} of degree k which are orthogonal to $V_0, V_1, \ldots, V_{k-1}$.

The key idea is to relate V_k to E_k^n.

Fix $x \in S^{n-1}$. Consider the evaluation map $f \mapsto f(x)$. This is a linear function on V_k. By the Riesz representation theorem[2], there is a unique $v_{k,x} \in V_k$ with

$$(v_{k,x}, f) = f(x).$$

Claim: $\alpha_k v_{k,x}(y) = E_k^n(x,y)$ for some $\alpha_k > 0$

[2]In fact it follows from basic linear algebra because V_k is of finite dimension.

Proof. Note that both sides are polynomials of the right degree. Also, $v_{k,x}(y)$ is invariant under rotations that leave x fixed: Let $A \in \mathcal{O}(n)$ such that $Ax = x$. Then,

$$(Av_{k,x})(y) = v_{k,x}(A^{-1}y) = v_{k,x}(y),$$

because we have

$$\begin{aligned}
(Av_{k,x}, f) &= (v_{k,x}, A^{-1}f) \\
&= f(Ax) \\
&= (v_{k,Ax}, f) \qquad \text{(by definition of } v_{k,\cdot}) \\
&= (v_{k,x}, f) \qquad (Ax = x)
\end{aligned}$$

and by uniqueness of $v_{k,x}$, it follows that $Av_{k,x} = v_{k,x}$. Thus $(x, y) \mapsto v_{k,x}(y)$ is purely a function of $x \cdot y$.

Also, for $k \neq l$, $v_{k,x} \perp v_{l,x}$ since $V_k \perp V_l$, thus they have the right orthogonality relations. Hence $E_k^n(x, y)$ and $v_{k,x}(y)$ are multiples of each other. Since we have

$$E_k^n(x, x) = 1 \quad \text{and} \quad v_{k,x}(x) = (v_{k,x}, v_{k,x}) > 0,$$

the claim follows. □

Now we are ready to show that E_k^n is positive semidefinite. Observe that $E_k^n(x, y) = \alpha_k v_{k,x}(y)$ and that $v_{k,x}(y) = (v_{k,y}, v_{k,x})$. Thus we have,

$$\begin{aligned}
&\int_{S^{n-1}} \int_{S^{n-1}} E_k^n(x, y) f(x) f(y) d\omega_n(x) d\omega_n(y) \\
= \;& \alpha_k \int_{S^{n-1}} \int_{S^{n-1}} (v_{k,y}, v_{k,x}) f(x) f(y) d\omega_n(x) d\omega_n(y) \\
= \;& \alpha_k \left(\int_{S^{n-1}} v_{k,x} f(x) d\omega_n(x), \int_{S^{n-1}} v_{k,y} f(y) d\omega_n(y) \right) \\
\geqslant \;& 0
\end{aligned}$$

as both the integrals in the last inner product are identical. It follows that E_k^n is positive semidefinite. □

12.4.3 End of proof

We first show that, if $f_0, f_1, \ldots$ are nonnegative numbers such that $\sum_{k=0}^{\infty} f_k$ converges, then the series $\sum_{k=0}^{\infty} f_k E_k^n(x, y)$ converges absolutely and uniformly for all $x, y \in S^{n-1}$.

By Lemma 12.4.2 E_k^n is positive semidefinite and so

$$|E_k^n(x, y)| \leqslant E_k^n(x, x) = P_k^n(1) = 1$$

for all $x, y \in S^{n-1}$ and so

$$\sum_{k=0}^{\infty} f_k E_k^n(x, y)$$

converges absolutely for all $x, y \in S^{n-1}$.

Now, for all $x, y \in S^{n-1}$ for all $m \in \mathbb{N}$ we have

$$\left| \sum_{k=m}^{\infty} f_k E_k^n(x,y) \right| \leqslant \sum_{k=m}^{\infty} f_k$$

and so the series also converges uniformly for all $x, y \in S^{n1}$.

With the above observation, if we are given nonnegative numbers $f_0, f_1, \ldots$ such that $\sum_{k=0}^{\infty} f_k$ converges, then the kernel

$$K(x,y) = \sum_{k=0}^{\infty} f_k E_k^n(x,y)$$

is continuous. From Lemma 12.4.2 it is also positive semidefinite, and so we showed the inclusion "$\supseteq$".

For the other inclusion "$\subseteq$" let $K : S^{n-1} \times S^{n-1} \to \mathbb{R}$ be a continuous, positive semidefinite, and invariant kernel. Kernel K is invariant, so let $h : [1,1] \to \mathbb{R}$ be the function such that $K(x,y) = h(x \cdot y)$ for all $x, y \in S^{n-1}$. The polynomials P_0^n, P_1^n form a complete orthogonal system of $L^2([-1,1], (1-t^2)^{(n-3)/2} dt)$ with convergence in the L^2-norm.

We first claim that the f_k are all nonnegative. To see this, recall the orthogonality relation from Lemma 12.4.1. First note that

$$\left\langle \sum_k f_k E_k^n, E_l^n \right\rangle \geqslant 0$$

since this is the inner product of two positive semidefinite kernels. Now by orthogonality of E_k^n's, we have

$$0 \leqslant \left\langle \sum_k f_k E_k^n, E_l^n \right\rangle = f_l \underbrace{\langle E_l^n, E_l^n \rangle}_{>0}$$

This is possible only if $f_l \geqslant 0$.

To finish, we show that the series $\sum_{k=0}^{\infty} f_k$ converges. To this end, consider for $m = 0, 1, \ldots$ the function

$$h_m(u) = h(u) - \sum_{k=0}^{m} f_k P_k^n(u) \quad \text{for all } u \in [1,1].$$

These are continuous functions. Moreover, since we have

$$h_m = \sum_{k=m+1}^{\infty} f_k P_k^n$$

in the sense of L^2 convergence, it follows that for each m the kernel $K_m(x,y) = h_m(x \cdot y)$ is positive semidefinite.

This implies in particular that $h_m(1) \geqslant 0$ for all m. But then we have

$$h(1) - \sum_{k=0}^{m} f_k = h(1) - \sum_{k=0}^{m} f_k P_k^m(1) = h_m(1) \geqslant 0$$

and we conclude that the series of nonnegative terms $\sum_{k=0}^{\infty} f_k$ converges to a number less than or equal to $h(1)$, as we wanted.

12.5 Delsarte's LP method

Using Schoenberg's theorem we can reformulate $\vartheta'(G(n, 2\gamma))$ where we use the nonnegative optimization variables $f_0, f_1, \ldots$

$$\begin{aligned} \vartheta'(G(n,2\gamma)) = \inf \quad & \lambda \\ & f_0, f_1, \ldots \geqslant 0 \\ & \sum_{k=0}^{\infty} f_k < \infty \\ & \sum_{k=0}^{\infty} f_k P_k^n(1) = \lambda - 1 \\ & \sum_{k} f_k P_k^n(t) \leqslant -1 \text{ for all } t \in [-1, \cos(2\gamma)] \end{aligned} \tag{12.2}$$

This problem has infinitely many variables. If we truncate the variables[3] we get the following bound:

$$\begin{aligned} \alpha(G(n,2\gamma)) \leqslant \vartheta'(G(n,2\gamma)) \leqslant \inf \quad & \lambda \\ & f_0, f_1, \ldots, f_d \geqslant 0, \\ & \sum_{k=1}^{d} f_k P_k^n(1) = \lambda - 1 \\ & \sum_{k=1}^{d} f_k P_k^n(t) \leqslant -1 \quad \forall t \in [-1, \cos(2\gamma)] \end{aligned}$$

Since this optimization problem is a linear program (with infinitely many constraints) it carries the name linear programming bound. These kind of linear programming bounds were first invented by Delsarte in 1973 in the context of error correcting codes and therefore they also carry the name "Delsarte's LP method".

Note that the infinitely many inequalities can be replaced by a finite dimensional semidefinite condition using sums of squares (see Chapter 2.7):

$$-1 - \sum_{k=1}^{d} f_k P_k^n(t) = p(t) - (t+1)(t - \cos(2\gamma))q(t)$$

[3]Formally we set $0 = f_{d+1} = f_{d+2} = \ldots$

where p and q are polynomials which can be written as sum of squares.

12.6 τ_8 equals 240

It so happens that for $n = 8$, $\alpha(G(8, \pi/3)) = \vartheta'(G(8, \pi/3)) = 240$. This result is due to Odlyzko, Sloane, and independently due to Levenshtein.

First, consider the set of 240 points in S^7 obtained by all possible permutations and sign-changes of the point

$$A = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0, 0, 0, 0\right)^{\mathsf{T}}$$

and all possible even sign-changes of the point

$$B = \left(\frac{1}{\sqrt{8}}, \frac{1}{\sqrt{8}}, \frac{1}{\sqrt{8}}, \frac{1}{\sqrt{8}}, \frac{1}{\sqrt{8}}, \frac{1}{\sqrt{8}}, \frac{1}{\sqrt{8}}\right)^{\mathsf{T}}$$

There are $\binom{8}{2}2^2 = 112$ points generated by A and $2^7 = 128$ points generated by B. All possible inner products for points from this set are

$$\left\{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\right\}.$$

In particular, note that there is no inner product between $\frac{1}{2}$ and 1. Thus, this is a valid kissing configuration. In fact, this configuration of points on the unit sphere is coming from the root system E_8 which has connections to many areas in mathematics and physics.

Now, taking hints from the formulation for $\vartheta'(G(8, \pi/3))$, we explicitly construct a kernel $K(x, y)$. Recall, $K(x, y) = -1$ if $\{x, y\} \notin E$. Also, recall that $K(x, y)$ was a function of the inner product $x \cdot y = t$ only. Now, consider the following polynomial

$$F(t) = -1 + \beta(t+1)\left(t + \frac{1}{2}\right)^2 t^2 \left(t - \frac{1}{2}\right)$$

Note that, $F(-1) = F(-1/2) = F(0) = f(1/2) = -1$ by construction. Also, $F(t) \leqslant -1$ for $t \in [-1, 1/2]$. Setting, $F(1) = \lambda - 1 = 240 - 1 = 239$, we get $\beta = \frac{320}{3}$.

Now, it can be verified (Exercise 12.1 (a)) that

$$F(t) = \sum_{k=0}^{6} f_k P_k^8(t), \quad f_k \geqslant 0. \tag{12.3}$$

Thus, $F(t)$ is a feasible point for the optimization problem (12.2).

Now by construction of the set of points, we know that $\alpha(G(8, \pi/3)) \geqslant 240$. By the construction of $F(t)$, we know that $\vartheta'(G(8), \pi/3) \leqslant 240$. Thus we have $\alpha = \vartheta'(G(8, \pi/3))$. Thus,

$$\tau_8 = 240.$$

12.7 Further reading

Schoenberg's result can be seen as a special case of Bochner's theorem which gives a similar statement for every compact, homogeneous space and which is based on the Peter-Weyl theorem. In [4] and [1] general techniques are presented which use Bochner's theorem to simplify semidefinite programs which are invariant under a group of symmetries.

A very readable introduction to the area of geometric packing problems and energy minimization is [2] from Henry Cohn.

12.8 Exercises

12.1** (a) Determine f_k in (12.3), completing the proof of $\tau_8 = 240$.

(b) Compute $\vartheta'(G(2, \pi/3))$.

(c) Determine $\alpha(G(n, \pi/4))$.

12.2** Consider 12 points $x_1, \ldots, x_{12}$ on the sphere S^2. What is the largest possible minimal angle between distinct points x_i and x_j with $i \neq j$?

12.3** Write a computer program for finding $\vartheta'(G(n), \pi/3)$ and produce a table for $n = 2, \ldots, 24$.

12.4 Determine $\alpha(G(24, \pi/3))$.

BIBLIOGRAPHY

[1] C. Bachoc, D.C. Gijswijt, A. Schrijver, and F. Vallentin, *Invariant semidefinite programs*, arXiv:1007.2905v2 [math.OC]

`http://arxiv.org/abs/1007.2905`

[2] H. Cohn, *Order and disorder in energy minimization*, arXiv:1003.3053v1 [math.MG]

`http://arxiv.org/abs/1003.3053`

[3] F. Pfender, G.M. Ziegler, *Kissing numbers, sphere packings and some unexpected proofs*, Notices Amer. Math. Soc. **51** (2004), 873–883.

`http://www.ams.org/notices/200408/fea-pfender.pdf`

[4] F. Vallentin, *Lecture notes: Semidefinite programs and harmonic analysis*, arXiv:0809.2017 [math.OC]

`http://arxiv.org/abs/0809.2017`

Part IV

Applications in algebra

CHAPTER 13

SUMS OF SQUARES OF POLYNOMIALS

In this chapter we return to sums of squares of polynomials, which we had already briefly introduced in Chapter 2. We address the following basic question: Given a subset $K \subseteq \mathbb{R}^n$ defined by finitely many polynomial inequalities, how can one certify that a polynomial p is nonnegative on K? This question is motivated by its relevance to the problem of minimizing p over K, to which we will return in the next two chapters. We collect a number of results from real algebraic geometry which give certificates for nonnegative (positive) polynomials on K in terms of sums of squares. We give a full proof for the representation result of Putinar, which we will use later for designing a hierarchy of semidefinite relaxations for polynomial optimization problems.

In this and the next two chapters we use the following notation. $\mathbb{R}[x_1, \ldots, x_n]$ (or simply $\mathbb{R}[x]$) denotes the ring of polynomials in n variables. A polynomial $p \in \mathbb{R}[x]$ can be written as $p = \sum_\alpha p_\alpha x^\alpha$, where $p_\alpha \in \mathbb{R}$ and x^α stands for the monomial $x_1^{\alpha_1} \cdots x_n^{\alpha_n}$. The sum is finite and the maximum value of $|\alpha| = \sum_{i=1}^n \alpha_i$ for which $p_\alpha \neq 0$ is the degree of p. For an integer d, $\mathbb{N}_d^n$ denotes the set of sequences $\alpha \in \mathbb{N}^n$ with $|\alpha| \leqslant d$, thus the exponents of the monomials of degree at most d. Moreover, $\mathbb{R}[x]_d$ denotes the vector space of all polynomials of degree at most d, its dimension is $s(n,d) = |\mathbb{N}_d^n| = \binom{n+d}{d}$ and the set $\{x^\alpha : \alpha \in \mathbb{N}^n, |\alpha| \leqslant d\}$ of monomials of degree at most d is its canonical base.

13.1 Sums of squares of polynomials

A polynomial p is said to be a *sum of squares*, abbreviated as p *is sos*, if p can be written as a sum of squares of polynomials. Σ denotes the set of all polynomials

that are sos. A fundamental property, already proved in Section 2.7, is that sums of squares of polynomials can be recognized using semidefinite programming.

Lemma 13.1.1. *Let $p \in \mathbb{R}[x]_{2d}$. Then p is sos if and only if the following semidefinite program in the matrix variable $Q \in \mathcal{S}^{s(n,d)}$ is feasible:*

$$Q \succeq 0, \quad \sum_{\substack{\beta,\gamma \in \mathbb{N}^n_d \\ \beta+\gamma=\alpha}} Q_{\beta,\gamma} = p_\alpha \quad \forall \alpha \in \mathbb{N}^n_{2d}. \tag{13.1}$$

13.1.1 Polynomial optimization

Why do we care about sums of squares?

Sums of squares are useful because they constitute a sufficient condition for nonnegative polynomials.

Example 13.1.2. *Consider the polynomial:*

$$f_n(x) = x_1^n + \cdots + x_n^n - n x_1 \cdots x_n.$$

One can show that f_n is a sum of squares for any even n, which permits to derive the arithmetic-geometric mean inequality:

$$\sqrt[n]{x_1 \cdots x_n} \leqslant \frac{x_1 + \cdots + x_n}{n} \tag{13.2}$$

for $x_1, \cdots, x_n \geqslant 0$ and any $n \geqslant 1$. (You will show this in Exercise 13.1).

As one can recognize sums of squares using semidefinite programming, sums of squares can be used to design tractable bounds for hard optimization problems of the form: *Compute the infimum $p_{\min}$ of a polynomial p over a subset $K \in \mathbb{R}^n$ defined by polynomial inequalities:*

$$K = \{x \in \mathbb{R}^n : g_1(x) \geqslant 0, \cdots, g_m(x) \geqslant 0\},$$

where $g_1, \cdots, g_m \in \mathbb{R}[x]$. Such optimization problem, where the objective and the constraints are polynomial functions, is called a *polynomial optimization problem*.

Define the set of nonnegative polynomials on K:

$$\mathcal{P}(K) = \{f \in \mathbb{R}[x] : f(x) \geqslant 0 \ \forall x \in K\}. \tag{13.3}$$

Clearly,

$$p_{\min} = \inf_{x \in K} p(x) = \sup\{\lambda : p - \lambda \in \mathcal{P}(K)\}. \tag{13.4}$$

Computing $p_{\min}$ is hard in general.

Example 13.1.3. *Given integers $a_1, \cdots, a_n \in \mathbb{N}$, consider the polynomial*

$$p(x) = \left(\sum_{i=1}^{n} a_i x_i \right)^2 + \sum_{i=1}^{n} (x_i^2 - 1)^2.$$

Then the infimum of p over $\mathbb{R}^n$ is equal to 0 if and only if the sequence $a_1, \cdots, a_n$ can be partitioned. So if one could compute the infimum over $\mathbb{R}^n$ of a quartic polynomial then one could solve the $\mathcal{N}P$-complete partition problem.

As another example, the stability number $\alpha(G)$ of a graph $G = (V, E)$ can be computed using any of the following two programs:

$$\alpha(G) = \max \left\{ \sum_{i \in V} x_i : x_i + x_j \leqslant 1 \ \forall \{i, j\} \in E, \ x_i^2 - x_i = 0 \ \forall i \in V \right\}, \quad (13.5)$$

$$\frac{1}{\alpha(G)} = \min \left\{ x^{\mathsf{T}}(A_G + I)x : \sum_{i \in V} x_i = 1, \ x \geqslant 0 \right\}, \quad (13.6)$$

where A_G is the adjacency matrix of G. The formulation (13.5) is due to Motzkin. This shows that polynomial optimization captures $\mathcal{N}P$-hard problems, as soon as the objective or the constraints are quadratic polynomials.

A natural idea is to replace the *hard positivity condition*: $p \in \mathcal{P}(K)$ by the *easier sos type condition*: $p \in \Sigma + g_1\Sigma + \cdots + g_m\Sigma$. This leads to defining the following parameter:

$$p_{\text{sos}} = \sup\{\lambda : p - \lambda \in \Sigma + g_1\Sigma + \cdots + g_m\Sigma\}. \quad (13.7)$$

As a direct application of Lemma 13.1.1, one can compute p_{sos} using semidefinite programming. For instance, when $K = \mathbb{R}^n$,

$$p_{\text{sos}} = p_0 + \sup \left\{ -Q_{00} : Q \succeq 0, \ p_\alpha = \sum_{\substack{\beta, \gamma \in \mathbb{N}^n_d \\ \beta + \gamma = \alpha}} Q_{\beta,\gamma}, \ \ \forall \alpha \in \mathbb{N}^n_{2d} \backslash \{0\} \right\}. \quad (13.8)$$

Clearly the inequality holds:

$$p_{\text{sos}} \leqslant p_{\min}. \quad (13.9)$$

In general the inequality is strict. However, when the set K is compact and satisfies an additional condition, equality holds. This follows from Putinar's theorem (Theorem 13.2.9), which claims that any polynomial positive on K belongs to $\Sigma + g_1\Sigma + \cdots + g_m\Sigma$. We will return to the polynomial optimization problem (14.1) and its sos relaxation (13.7) in the next chapters. In the remaining of this chapter we investigate sums of squares representations for positive polynomials and we will prove Putinar's theorem.

13.1.2 Hilbert's theorem

Hilbert has classified in 1888 the pairs (n, d) for which every nonnegative polynomial of degree d in n variables is a sum of squares of polynomials:

Theorem 13.1.4. *Every nonnegative n-variate polynomial of even degree d is a sum of squares if and only if $n = 1$, or $d = 2$, or $(n, d) = (2, 4)$.*

We saw earlier that nonnegative univariate polynomials are sos, the case $d = 2$ boils down to the fact that positive semidefinite matrices have a Cholesky factorization, but the last exceptional case $(n, d) = (2, 4)$ is difficult. For every pair $(n, d) \neq (2, 4)$ with $n \geqslant 2$ and even $d \geqslant 4$, there is an n-variate polynomial of degree d which is nonnegative over $\mathbb{R}^n$ but not sos. It is not difficult to see that it suffices to give such a polynomial for the two pairs $(n, d) = (2, 6), (3, 4)$.

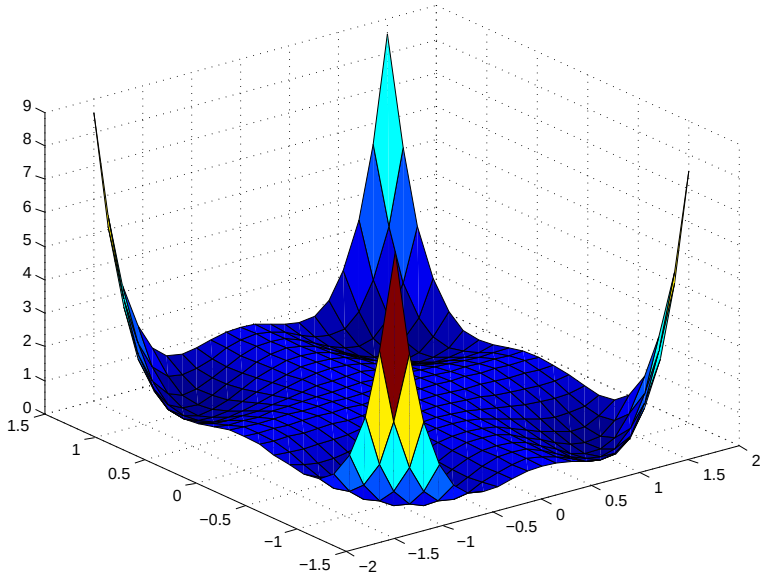


Figure 13.1: The Motzkin polynomial

Example 13.1.5. *Hilbert's proof for the 'only if' part of Theorem 13.1.4 was not constructive, the first concrete example of a nonnegative polynomial that is not sos is the following polynomial, for the case $(n, d) = (2, 6)$:*

$$p(x, y) = x^2y^2(x^2 + y^2 - 3) + 1,$$

constructed by Motzkin in 1967.

Proof that p is nonnegative on $\mathbb{R}^2$: If $x^2 + y^2 - 3 \geqslant 0$ then clearly $M(x, y) \geqslant 0$. Otherwise, set $z^2 = 3 - x^2 - y^2$ and use the arithmetic-geometric mean inequality: $\sqrt[3]{x^2y^2z^2} \leqslant (x^2 + y^2 + z^2)/3$ to deduce $M(x, y) \geqslant 0$.

To show that p is not sos, use brute force: Say $p = \sum_l s_l^2$ for some polynomials s_l of degree at most 3. As the coefficient of x^6 in p is 0, we see that the coefficient of x^3 in each s_l is 0; analogously, the coefficient of y^3 in s_l is 0. Then, as the coefficients of x^4 and y^4 in p are 0, we get that the coefficients of x^2 and y^2 in s_l are 0. After that we can conclude that the coefficients of x and y in s_l are 0. Finally, say $s_l = a_l x y^2 + b_l x^2 y + c_l xy + d_l$. Then the coefficient of x^2y^2 in p is equal to $-3 = \sum_l c_l^2$, yielding a contradiction.

In fact, the same argument shows that $p - \lambda$ is not sos for any scalar $\lambda \in \mathbb{R}$. Therefore, for the infimum of the Motzkin polynomial p over $\mathbb{R}^2$, the sos bound p_{sos} carries no information: $p_{\text{sos}} = -\infty$, while $p_{\text{min}} = 0$ is attained at $(\pm 1, \pm 1)$.

For the case $(n, d) = (3, 4)$, the Choi-Lam polynomial:

$$q(x, y, z) = 1 + x^2y^2 + y^2z^2 + x^2z^2 - 4xyz$$

is nonnegative (directly, using the arithmetic-geometric mean inequality) but not sos (direct inspection).

13.1.3 Are sums of squares a rare event?

A natural question is whether sums of squares abound or not within the cone of of nonnegative polynomials. It turns out hat the answer depends, whether we fix or let grow the number of variables and the degree.

On the one hand, if we fix the number of variables and allow the degree to grow, then every nonnegative polynomial p can be approximated by sums of squares obtained by adding a small high degree perturbation to p.

Theorem 13.1.6. *If $p \geqslant 0$ on $[-1, 1]^n$, then the following holds:*

$$\forall \epsilon > 0 \; \exists k \in \mathbb{N} \quad p + \epsilon \left(1 + \sum_{i=1}^{n} x_i^{2k}\right) \in \Sigma.$$

On the other hand, if we fix the degree and let the number of variables grow, then there are significantly more nonnegative polynomials than sums of squares: There exist universal constants $c, C > 0$ such that

$$c \cdot n^{(d-1)/2} \leqslant \left(\frac{\text{vol}(\hat{\mathcal{P}}_{n,2d})}{\text{vol}(\hat{\Sigma}_{n,2d})}\right)^{1/D} \leqslant C \cdot n^{(d-1)/2}. \tag{13.10}$$

Here $\hat{\mathcal{P}}_{n,2d}$ is the set of nonnegative homogeneous polynomials of degree $2d$ in n variables intersected with the hyperplane $H = \{p : \int_{\mathbf{S}^{n-1}} p(x)\mu(dx) = 1\}$. Analogously, $\hat{\Sigma}_{n,2d}$ is the set of homogeneous polynomials of degree $2d$ in n variables that are sums of squares, intersected by the same hyperplane H. Finally, $D = \binom{n+2d-1}{2d} - 1$ is the dimension of the ambient space.

13.1.4 Artin's theorem

Hilbert asked in 1900 the following question, known as *Hilbert's 17th problem: Is it true that every nonnegative polynomial on $\mathbb{R}^n$ is a sum of squares of* **rational** *functions?* Artin answered this question in the affirmative in 1927:

Theorem 13.1.7. (Artin's theorem) *A polynomial p is nonnegative on $\mathbb{R}^n$ if and only if $p = \sum_{j=1}^m \left(\frac{p_j}{q_j}\right)^2$ for some $p_j, q_j \in \mathbb{R}[x]$.*

This was a major breakthrough, which started the field of real algebraic geometry.

13.2 Positivstellensätze

We now turn to the study of nonnegative polynomials p on a basic closed semialgebraic set K, i.e., a set K of the form

$$K = \{x \in \mathbb{R}^n : g_1(x) \geqslant 0, \cdots, g_m(x) \geqslant 0\}, \tag{13.11}$$

where $g_1, \cdots, g_m \in \mathbb{R}[x]$. Set $g_0 = 1$. When the polynomials p, g_j are linear, Farkas' lemma implies:

$$p \geqslant 0 \text{ on } K \Longleftrightarrow p = \sum_{j=0}^m \lambda_j g_j \text{ for some scalars } \lambda_j \geqslant 0. \tag{13.12}$$

We will show the following result, due to Putinar: Assume that K is compact and satisfies the additional condition (13.17) below. Then

$$p > 0 \text{ on } K \Longrightarrow p = \sum_{j=0}^m s_j g_j \text{ for some polynomials } s_j \in \Sigma. \tag{13.13}$$

Of course, the following implication holds trivially:

$$p = \sum_{j=0}^m s_j g_j \text{ for some polynomials } s_j \in \Sigma \Longrightarrow p \geqslant 0 \text{ on } K.$$

However, this is not an equivalence, one needs a stronger assumption: strict positivity of p over K. Note the analogy between (13.12) and (13.13): While the variables in (13.12) are nonnegative scalars λ_i, the variables in (13.13) are sos polynomials s_i. A result of the form (13.13) is usually called a **Positivstellensatz**. This has historical reasons, the name originates from the analogy to the classical **Nullstellensatz** of Hilbert for the existence of *complex* roots:

Theorem 13.2.1. (Hilbert's Nullstellensatz) *Given $g_1, \cdots, g_m \in \mathbb{R}[x]$, define the complex variety, consisting of their common complex roots:*

$$V_{\mathbb{C}}(g_1, \cdots, g_m) = \{x \in \mathbb{C}^n : g_1(x) = 0, \cdots, g_m(x) = 0\}.$$

For a polynomial $p \in \mathbb{R}[x]$,

$$p = 0 \text{ on } V_{\mathbb{C}}(g_1, \cdots, g_m) \Longleftrightarrow p^k = \sum_{j=1}^{m} u_j g_j \text{ for some } u_j \in \mathbb{R}[x], k \in \mathbb{N}.$$

In particular, $V_{\mathbb{C}}(g_1, \cdots, g_m) = \varnothing \Longleftrightarrow 1 = \sum_{j=1}^{m} u_j g_j$ for some $u_j \in \mathbb{R}[x]$.

Checking a Nullstellensatz certificate: whether there exist polynomials u_j satisfying $p = \sum_j u_j h_j$, amounts to solving a linear program (after fixing a bound d on the degrees of the unknown u_j's). On the other hand, checking a certificate of the form: $p = \sum_j s_j g_j$ where the s_j's are sos, amounts to solving a semidefinite program (again, after fixing some bound d on the degrees of the unknown s_j's). In a nutshell, semidefinite programming is the key ingredient to deal with *real* elements while linear programming permits to deal with *complex* elements. We will return to this in the last chapter.

13.2.1 The univariate case

We consider here nonnegative univariate polynomials over a closed interval $K \subseteq \mathbb{R}$, thus of the form $K = [0, \infty)$ or $K = [-1, 1]$ (up to scaling). Then a full characterization is known, moreover with explicit degree bounds.

Theorem 13.2.2. (Pólya-Szegö) *Let p be a univariate polynomial of degree d. Then, $p \geqslant 0$ on $[0, \infty)$ if and only if $p = s_0 + s_1 x$ for some $s_0, s_1 \in \Sigma$ with $\deg(s_0) \leqslant d$ and $\deg(s_1) \leqslant d - 1$.*

Theorem 13.2.3. (Fekete, Markov-Lukácz) *Let p be a univariate polynomial of degree d. Assume that $p \geqslant 0$ on $[-1, 1]$.*

(i) *$p = s_0 + s_1(1 - x^2)$, where $s_0, s_1 \in \Sigma$, $\deg(s_0) \leqslant d + 1$ and $\deg(s_1) \leqslant d - 1$.*

(ii) *For d odd, $p = s_1(1+x)+s_2(1-x)$ where $s_1, s_2 \in \Sigma$, $\deg(s_1), \deg(s_2) \leqslant d - 1$.*

Note the two different representations in (i), (ii), depending on the choice of the polynomials describing the set $K = [-1, 1]$.

13.2.2 Krivine's Positivstellensatz

Here we state the Positivstellensatz of Krivine (1964), which characterizes nonnegative polynomials on an arbitrary basic closed semi-algebraic set K (with no compactness assumption). Let K be as in (13.11). Set $\mathbf{g} = (g_1, \cdots, g_m)$ and, for a set of indices $J \subseteq \{1, \cdots, m\}$, set $g_J = \prod_{j \in J} g_j$. The set

$$\mathbf{T}(\mathbf{g}) = \left\{ \sum_{J \subseteq [m]} s_J g_J : s_J \in \Sigma \right\} \tag{13.14}$$

is called the *preordering* generated by $\mathbf{g} = (g_1, \cdots, g_m)$. It consists of all weighted sums of the products g_J, weighted by sums of squares. Clearly, any polynomial in $\mathbf{T}(\mathbf{g})$ is nonnegative on K: $\mathbf{T}(\mathbf{g}) \subseteq \mathcal{P}(K)$.

Example 13.2.4. *Let $K = \{x \in \mathbb{R} : g = (1-x^2)^3 \geqslant 0\}$ and $p = 1 - x^2$. Then, p is nonnegative on K, but $p \notin \mathbf{T}(g)$ (check it). But, note that $pg = p^4$ (compare with item (ii) in the next theorem).*

Theorem 13.2.5. (Krivine's Positivstellensatz) *Let K be as in (13.11) and let $p \in \mathbb{R}[x]$. The following holds.*

(i) $p > 0$ *on* $K \iff pf = 1 + h$ *for some* $f, h \in \mathbf{T}(\mathbf{g})$.

(ii) $p \geqslant 0$ *on* $K \iff pf = p^{2k} + h$ *for some* $f, h \in \mathbf{T}(\mathbf{g})$ *and* $k \in \mathbb{N}$.

(iii) $p = 0$ *on* $K \iff -p^{2k} \in \mathbf{T}(\mathbf{g})$ *for some* $k \in \mathbb{N}$.

(iv) $K = \varnothing \iff -1 \in \mathbf{T}(\mathbf{g})$.

In (i)-(iv) above, there is one trivial implication. For example, it is clear that $-1 \in \mathbf{T}(\mathbf{g})$ implies $K = \varnothing$. And in (i)-(iii), the existence of a sos identity for p of the prescribed form implies the desired property for p.

Choosing $K = \mathbb{R}^n$ ($\mathbf{g} = 1$), we have $\mathbf{T}(\mathbf{g}) = \Sigma$ and thus (ii) implies Artin's theorem. Moreover, one can derive the following result, which characterizes the polynomials that vanish on the set of common *real* roots of a set of polynomials.

Theorem 13.2.6. (The Real Nullstellensatz) *Given $g_1, \cdots, g_m \in \mathbb{R}[x]$, define the real variety, consisting of their common real roots:*

$$V_{\mathbb{R}}(g_1, \cdots, g_m) = \{x \in \mathbb{R}^n : g_1(x) = 0, \cdots, g_m(x) = 0\}. \tag{13.15}$$

For a polynomial $p \in \mathbb{R}[x]$,

$$p = 0 \text{ on } V_{\mathbb{R}}(g_1, \cdots, g_m) \iff p^{2k} + s = \sum_{j=1}^{m} u_j g_j \text{ for some } s \in \Sigma, u_j \in \mathbb{R}[x], k \in \mathbb{N}.$$

In particular,

$$V_{\mathbb{R}}(g_1, \cdots, g_m) = \varnothing \iff -1 = s + \sum_{j=1}^{m} u_j g_j \text{ for some } s \in \Sigma, u_j \in \mathbb{R}[x].$$

The above result does not help us yet to tackle the polynomial optimization problem (14.1): Indeed, using (i), we can reformulate p_{sos} as

$$p_{\text{sos}} = \sup_{\lambda \in \mathbb{R}, f, g \in \mathbb{R}[x]} \{\lambda : (p - \lambda) f = 1 + g, f, g \in \mathbf{T}(\mathbf{g})\}.$$

However, this does not lead to a semidefinite program, because of the quadratic term λf where both λ and f are unknown. Of course, one could fix λ and solve the corresponding semidefinite program, and iterate using binary search on λ. However, there is an elegant, more efficient remedy: Using the refined representation results of Schmüdgen and Putinar in the next sections one can set up a simpler semidefinite program permmitting to search over the variable λ.

13.2.3 Schmüdgen's Positivstellensatz

When K is compact, Schmüdgen [7] proved the following simpler representation result for *positive* polynomials on K.

Theorem 13.2.7. (Schmüdgen's Positivstellensatz) *Assume K is compact. Then,*

$$p(x) > 0 \ \forall x \in K \Longrightarrow p \in \mathbf{T}(\mathbf{g}).$$

A drawback of a representation $\sum_J s_J g_J$ in the preordering $\mathbf{T}(\mathbf{g})$ is that it involves 2^m sos polynomials s_J, thus exponential in the number m of constraints defining K. Next we see how to get a representation of the form $\sum_j s_j g_j$, thus involving only a linear number of terms.

13.2.4 Putinar's Positivstellensatz

Under an additional (mild) assumption on the polynomials defining the set K, Putinar [5] showed the analogue of Schmüdgen's theorem, where the preordering $\mathbf{T}(\mathbf{g})$ is replaced by the following *quadratic module*:

$$\mathbf{M}(\mathbf{g}) = \left\{ \sum_{j=0}^{m} s_j g_j : s_j \in \Sigma \right\}. \tag{13.16}$$

First we describe this additional assumption. For this consider the following conditions on the polynomials g_j defining K:

$$\exists h \in \mathbf{M}(\mathbf{g}) \ \ \{x \in \mathbb{R}^n : h(x) \geqslant 0\} \text{ is compact}, \tag{13.17}$$

$$\exists N \in \mathbb{N} \ \ N - \sum_{i=1}^{n} x_i^2 \in \mathbf{M}(\mathbf{g}), \tag{13.18}$$

$$\forall f \in \mathbb{R}[x] \ \exists N \in \mathbb{N} \ \ N \pm f \in \mathbf{M}(\mathbf{g}). \tag{13.19}$$

Proposition 13.2.8. *The conditions (13.17), (13.18) and (13.19) are all equivalent. If any of them holds, the quadratic module $\mathbf{M}(\mathbf{g})$ is said to be* Archimedean.

Proof. The implications (13.19) $\Longrightarrow$ (13.18) $\Longrightarrow$ (13.17) are clear. Assume (13.17) holds and let $f \in \mathbb{R}[x]$. As the set $K_0 = \{x : h(x) \geqslant 0\}$ is compact, there exists $N \in \mathbb{N}$ such that $-N < f(x) < N$ over K_0. Hence, $N \pm f$ is positive on K. Applying Theorem 13.2.7, we deduce that $N \pm f \in \mathbf{T}(h) \subseteq \mathbf{M}(\mathbf{g})$. □

Clearly, (13.17) implies that K is compact. On the other hand, if K is compact, then it is contained in some ball $\{x \in \mathbb{R}^n : g_{m+1} = R^2 - \sum_{i=1}^n x_i^2 \geqslant 0\}$. Hence, if we know the radius R of a ball containing K, then it suffices to add the (redundant) ball constraint $g_{m+1}(x) \geqslant 0$ to the description of K so that the quadratic module $\mathbf{M}(\mathbf{g}')$ is now Archimedean, where $\mathbf{g}' = (\mathbf{g}, g_{m+1})$.

Theorem 13.2.9. (Putinar's Positivstellensatz) *Assume that the qudratic module* $\mathbf{M}(\mathbf{g})$ *is Archimedean (i.e., the* g_j*'s satisfy any of the equivalent conditions (13.17)-(13.19)). Then,*

$$p(x) > 0 \ \forall x \in K \Longrightarrow p \in \mathbf{M}(\mathbf{g}).$$

Example 13.2.10. *Consider the simplex* $K = \{x \in \mathbb{R}^n : x \geqslant 0, \ \sum_{i=1}^n x_i \leqslant 1\}$ *and the corresponding quadratic module* $M = \mathbf{M}(x_1, \cdots, x_n, 1 - \sum_{i=1}^n x_i)$*. Then* M *is Archimedean. To see it note that the polynomial* $n - \sum_i x_i^2 \in M$*. This follows from the following identities:*

- $1 - x_i = (1 - \sum_j x_j) + \sum_{j \neq i} x_j \in M.$
- $1 - x_i^2 = \frac{(1+x_i)(1-x_i^2)}{2} + \frac{(1+x_i)(1-x_i^2)}{2} = \frac{(1+x_i)^2}{2}(1 - x_i) + \frac{(1-x_i)^2}{2}(1 + x_i) \in M.$
- $n - \sum_i x_i^2 = \sum_i (1 - x_i^2) \in M.$

Example 13.2.11. *Consider the cube* $K = [01]^n = \{x \in \mathbb{R}^n : 0 \leqslant x_i \leqslant 1 \ \forall i \in [n]\}$ *and the corresponding quadratic module* $M = \mathbf{M}(x_1, 1-x_1, \cdots, x_n, 1-x_n)$*. Then* M *is Archimedean. Indeed, as in the previous example,* $1 - x_i^2 \in M$ *and thus* $n - \sum_i x_i^2 \in M$*.*

13.2.5 Proof of Putinar's Positivstellensatz

In this section we give a full proof for Theorem 13.2.9. The proof is elementary, combining some (sometimes ingenious) algebraic manipulations. We start with defining the notions of ideal and quadratic module in the ring $\mathbb{R}[x]$.

Definition 13.2.12. *A set* $I \subseteq \mathbb{R}[x]$ *is an* ideal *if* I *is closed under addition and multiplication by* $\mathbb{R}[x]$*:* $I + I \subseteq I$ *and* $\mathbb{R}[x] \cdot I \subseteq I$*.*

Definition 13.2.13. *A subset* $M \subseteq \mathbb{R}[x]$ *is a* quadratic module *if* $1 \in M$ *and* M *is closed under addition and multiplication by squares:* $M + M \subseteq M$ *and* $\Sigma \cdot M \subseteq M$*.* M *is said to be* proper *if* $M \neq \mathbb{R}[x]$ *or, equivalently, if* $-1 \notin M$*.*

Example 13.2.14. *Given polynomials* $g_1, \cdots, g_m$*,*

$$(g_1, \cdots, g_m) = \left\{ \sum_{j=1}^m u_j g_j : u_j \in \mathbb{R}[x] \right\}$$

is an ideal (the ideal generated by the g_j's*) and the set* $\mathbf{M}(\mathbf{g})$ *from (13.16) is a quadratic module (the* quadratic module generated by the g_j's*).*

We start with some technical lemmas.

Lemma 13.2.15. *If* $M \subseteq \mathbb{R}[x]$ *is a quadratic module, then* $I = M \cap (-M)$ *is an ideal.*

Proof. This follows from the fact that, for any $f \in \mathbb{R}[x]$ and $g \in I$, we have: $fg = \left(\frac{f+1}{2}\right)^2 g + \left(\frac{f-1}{2}\right)^2 (-g) \in I.$ □

Lemma 13.2.16. *Let $M \subseteq \mathbb{R}[x]$ be a maximal proper quadratic module. Then, $M \cup (-M) = \mathbb{R}[x]$.*

Proof. Assume $f \notin M \cup (-M)$. Each of the sets $M' = M + f\Sigma$ and $M'' = M - f\Sigma$ is a quadratic module, strictly containing M. By the maximality assumption on M, M' and M'' are not proper: $M' = M'' = \mathbb{R}[x]$. Hence:

$$-1 = g_1 + s_1 f, \ \ -1 = g_2 - s_2 f \quad \text{for some } g_1, g_2 \in M, \ s_1, s_2 \in \Sigma.$$

This implies: $-s_2 - s_1 = s_2(g_1 + s_1 f) + s_1(g_2 - s_2 f) = s_2 g_1 + s_1 g_2$ and thus $s_1, s_2 \in -M$. On the other hand, $s_1, s_2 \in \Sigma \subseteq M$. Therefore, $s_1, s_2 \in I = M \cap (-M)$. As I is an ideal (by Lemma 13.2.15), we get $s_1 f \in I \subseteq M$ and therefore $-1 = g_1 + s_1 f \in M$, contradicting M proper. □

Lemma 13.2.17. *Let M be a maximal proper quadratic module in $\mathbb{R}[x]$ and $I = M \cap (-M)$. Assume that M is Archimedean, i.e., satisfies:*

$$\forall f \in \mathbb{R}[x] \ \exists N \in \mathbb{N} \ N \pm f \in M.$$

Then, for any $f \in \mathbb{R}[x]$, there exists a (unique) scalar $a \in \mathbb{R}$ such that $f - a \in I$.

Proof. Define the sets

$$A = \{a \in \mathbb{R} : f - a \in M\}, \ B = \{b \in \mathbb{R} : b - f \in M\}.$$

As M is Archimedean, A, B are both non-empty. We show that $|A \cap B| = 1$. First observe that $a \leqslant b$ for any $a \in A$ and $b \in B$. For, if one would have $a > b$, then $b - a = (f - a) + (b - f)$ is a negative scalar in M, contradicting M proper. Let a_0 be the supremum of A and b_0 the infimum of B. Thus $a_0 \leqslant b_0$. Moreover, $a_0 = b_0$. For, if not, there is a scalar c such that $a_0 < c < b_0$. Then, $f - c \notin M \cup (-M)$, which contradicts Lemma 13.2.16.

We now show that $a_0 = b_0$ belongs to $A \cap B$, which implies that $A \cap B = \{a_0\}$ and thus concludes the proof. Suppose for a contradiction that $a_0 \notin A$, i.e., $f - a_0 \notin M$. Then the quadratic module $M' = M + (f - a_0)\Sigma$ is not proper: $M' = \mathbb{R}[x]$. Hence,

$$-1 = g + (f - a_0)s \quad \text{for some } g \in M, \ s \in \Sigma.$$

As M is Archimedean, there exists $N \in \mathbb{N}$ such that $N - s \in M$. Pick ϵ such that $0 < \epsilon < 1/N$. Then, $a_0 - \epsilon \in A$ and $f - (a_0 - \epsilon) = (f - a_0) + \epsilon \in M$ implies:

$$-1 + \epsilon s = g + (f - a_0 + \epsilon)s \in M.$$

Adding with $\epsilon(N - s) \in M$, we obtain:

$$-1 + \epsilon N = (-1 + \epsilon s) + \epsilon(N - s) \in M.$$

We reach a contradiction since $-1 + \epsilon N < 0$. □

Lemma 13.2.18. *Assume $p > 0$ on K. Then there exists $s \in \Sigma$ such that $sp - 1 \in \mathbf{M}(\mathbf{g})$.*

Proof. We need to show that the quadratic module $M_0 = \mathbf{M}(\mathbf{g}) - p\Sigma$ is not proper. Assume for a contradiction that M_0 is proper. We are going to construct $a \in K$ for which $p(a) \leqslant 0$, contradicting the assumption that p is positive on K. By Zorn's lemma[1] let M be a maximal proper quadratic module containing M_0. As $M \supseteq \mathbf{M}(\mathbf{g})$, M too is Archimedean. Applying Lemma 13.2.17 to M, we find some scalar $a_i \in \mathbb{R}$ for which

$$x_i - a_i \in I = M \cap (-M) \quad \forall i \in [n].$$

The a_i's make up a point $a \in \mathbb{R}^n$. As I is an ideal, this implies that

$$f - f(a) \in I \quad \forall f \in \mathbb{R}[x]. \tag{13.20}$$

Indeed, say $f = \sum_\alpha f_\alpha x^\alpha$, then $f - f(a) = \sum_\alpha f_\alpha(x^\alpha - a^\alpha)$. It suffices now to show that each $x^\alpha - a^\alpha$ belongs to I. We do this using induction on $|\alpha| \geqslant 0$. If $\alpha = 0$ there is nothing to prove. Otherwise, say $\alpha_1 \geqslant 1$ and write $\beta = \alpha - e_1$ so that $x^\alpha = x_1 x^\beta$ and $a^\alpha = a_1 a^\beta$. Then we have

$$x^\alpha - a^\alpha = x_1(x^\beta - a^\beta) + a^\beta(x_1 - a_1) \in I$$

since $x^\beta - a^\beta \in I$ (using induction) and $x_1 - a_1 \in I$.

Now we apply (13.20) to $f = g_j$ and we obtain that

$$g_j(a) = g_j - (g_j - g_j(a)) \in M$$

since $g_j \in \mathbf{M}(\mathbf{g}) \subseteq M$ and $g_j - g_j(a) \in -M$. As M is proper, we must have that $g_j(a) \geqslant 0$ for each j. This shows that $a \in K$. Finally,

$$-p(a) = (p - p(a)) - p \in M,$$

since $p - p(a) \in I \subseteq M$ and $-p \in M_0 \subseteq M$. Again, as M is proper, this implies that $-p(a) \geqslant 0$, yielding a contradiction because $p > 0$ on K. □

Lemma 13.2.19. *Assume $p > 0$ on K. Then there exist $N \in \mathbb{N}$ and $h \in \mathbf{M}(\mathbf{g})$ such that $N - h \in \Sigma$ and $hp - 1 \in \mathbf{M}(\mathbf{g})$.*

Proof. Choose s as in Lemma 13.2.18. Thus, $s \in \Sigma$ and $sp - 1 \in \mathbf{M}(\mathbf{g})$. As $\mathbf{M}(\mathbf{g})$ is Archimedean, we can find $k \in \mathbb{N}$ such that

$$2k - s,\ 2k - s^2 p - 1 \in \mathbf{M}(\mathbf{g}).$$

Set $h = s(2k - s)$ and $N = k^2$. Then, $h \in \mathbf{M}(\mathbf{g})$ and $N - h = (k - s)^2 \in \Sigma$. Moreover,

$$hp - 1 = s(2k - s)p - 1 = 2k(sp - 1) + (2k - s^2 p - 1) \in \mathbf{M}(\mathbf{g}),$$

since $sp - 1, 2k - s^2 p - 1 \in \mathbf{M}(\mathbf{g})$. □

[1] Zorn's lemma states the following: Let $(P, \leqslant)$ be a partially ordered set in which every chain (totally ordered subset) has an upper bound. Then P has a maximal element.

We can now show Theorem 13.2.9. Assume $p > 0$ on K. Let h and N satisfy the conclusion of Lemma 13.2.19 and $k \in \mathbb{N}$ such that $k + p \in \mathbf{M}(\mathbf{g})$. We may assume that $N > 0$. Note that:

$$\left(k - \frac{1}{N}\right) + p = \frac{1}{N}\left((N-h)(k+p) + (hp-1) + kh\right) \in \mathbf{M}(\mathbf{g}).$$

So what we have just shown is that $k + p \in \mathbf{M}(\mathbf{g})$ implies $(k - 1/N) + p \in \mathbf{M}(\mathbf{g})$. Iterating this (kN) times, we obtain that

$$p = \left(k - kN\frac{1}{N}\right) + p \in \mathbf{M}(\mathbf{g}).$$

This concludes the proof of Theorem 13.2.9.

13.3 Notes and further reading

Hilbert obtained the first fundamental results about the links between nonnegative polynomials and sums of squares. He posed in 1900 at the first International Congress of Mathematicians in Paris the following question, known as *Hilbert's 17th problem: Is it true that every nonnegative polynomial on $\mathbb{R}^n$ is a sum of squares of rational functions?* The solution of Artin in 1927 to Hilbert's 17th problem was a major breakthrough, which started the field of real algebraic geometry. Artin's proof works in the setting of formal real (ordered) fields. It combines understanding which elements are positive in any ordering of the field and using Tarksi's transfer principle which roughly states the following: *If $(F, \leqslant)$ is an ordered field extension of $\mathbb{R}$ which contains a solution $x \in F^n$ of a system of polynomial equations and inequalities with coefficients in $\mathbb{R}$, then this system also has a solution $x' \in \mathbb{R}^n$.* Tarski's transfer principle also plays a crucial role in the proof of the Positivstellensatz of Krivine (Theorem 13.2.5). The book of Marshall [3] contains the proofs of all the Positivstellensätze described in this chapter.

Reznick [6] gives a nice historical overview of results about positive polynomials and sums of squares. The idea of using sums of squares combined with the power of semidefinite programming in order to obtain tractable sufficient conditions for nonnegativity of polynomials goes back to the PhD thesis of Parrilo [4]. He exploits this idea to attack various problems from optimization and control theory. Lasserre and Netzer [2] showed that every nonnegative polynomial can be approximated by sums of squares of increasing degrees (Theorem 13.1.6). Blekherman [1] proved the inequalities (13.10) relating the volumes of the cones of sums of squares and of nonnegative polynomials.

13.4 Exercises

13.1. Given $a \in \mathbb{N}^n$ with $|a| = \sum_i a_i = 2d$, define the polynomial in n variables $x = (x_1, \cdots, x_n)$ and of degree $2d$:

$$F_{n,2d}(a,x) = \sum_{i=1}^{n} a_i x_i^{2d} - 2d \prod_{i=1}^{n} x_i^{a_i} = \sum_{i=1}^{n} a_i x_i^{2d} - 2dx^a.$$

(a) Let $a \in \mathbb{N}^n$ with $|a| = 2d$. Show that $a = b + c$ for some $b, c \in \mathbb{N}^n$, where $|b| = |c| = d$ and both $b_i, c_i > 0$ for at most one index $i \in [n]$.

(b) With a, b, c as in (a), show that

$$F_{n,2d}(a,x) = \frac{1}{2}(F_{n,2d}(2b,x) + F_{n,2d}(2c,x)) + d(x^b - x^c)^2.$$

(c) Show that, for any $a \in \mathbb{N}^n$ with $|a| = 2d$, the polynomial $F_{n,2d}(a,x)$ can be written as the sum of at most $3n-4$ squares.

(d) Show the arithmetic-geometric mean inequality (13.2).

13.2 (a) Show Theorem 13.2.2.

(b) For a univariate polynomial f of degree d define the following polynomial $G(f)$, known as its Goursat transform:

$$G(f)(x) = (1+x)^d f\left(\frac{1-x}{1+x}\right).$$

Show that $f \geqslant 0$ on $[-1,1]$ if and only if $G(f) \geqslant 0$ on $[0,\infty)$.

(c) Show Theorem 13.2.3.

13.3 Show the Real Nullstellensatz (Theorem 13.2.6) (you may use Theorem 13.2.5).

13.4 Let $G = (V,E)$ be a graph. The goal is to show Motzkin's formulation (13.6) for the stability number $\alpha(G)$. Set

$$\mu = \min\left\{x^{\mathsf{T}}(A_G + I)x : \sum_{i \in V} x_i = 1,\ x \geqslant 0\right\}. \tag{13.21}$$

(a) Show that $\mu \leqslant 1/\alpha(G)$.

(b) Let x be an optimal solution of the program (13.21), $S = \{i : x_i \neq 0\}$ denotes its support. Show that $\mu \geqslant 1/\alpha(G)$ if S is a stable set in G.

(c) Show that the program (13.21) has an optimal solution x whose support is a stable set. Conclude that (13.6) holds.

BIBLIOGRAPHY

[1] G. Blekherman. There are significantly more nonnegative polynomials than sums of squares. *Israel Journal of Mathematics* **153**:355–380, 2006.

[2] J.B. Lasserre and T. Netzer. SOS approximations of nonnegative polynomials via simple high degree perturbations. *Mathematische Zeitschrift* **256**:99–112, 2006.

[3] M. Marshall. *Positive Polynomials and Sums of Squares.* AMS, vol. 146, 2008.

[4] P.A. Parrilo. *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*. Ph.D. thesis, California Institute of Technology, 2000.

`http://thesis.library.caltech.edu/1647/1/Parrilo-Thesis.pdf`

[5] M. Putinar. Positive polynomials on compact sem-algebraic sets. *Indiana University Mathematics Journal* **42**:969–984, 1993.

[6] B. Reznick. Some concrete aspects of Hilbert's 17th problem. In *Real Algebraic Geometry and Ordered Structures*. C.N. Delzell and J.J. Madden (eds.), *Contemporary Mathematics* **253**:251–272, 2000.

[7] K. Schmüdgen. The K-moment problem for compact semi-algebraic sets. *Mathematische Annalen* **289**:203–206, 1991.

CHAPTER 14

POLYNOMIAL EQUATIONS AND MOMENT MATRICES

Consider the polynomial optimization problem:

$$p_{\min} = \inf_{x \in K} p(x), \tag{14.1}$$

which asks for the infimum $p_{\min}$ of a polynomial p over a basic closed semi-algebraic set K, of the form:

$$K = \{x \in \mathbb{R}^n : g_1(x) \geqslant 0, \cdots, g_m(x) \geqslant 0\} \tag{14.2}$$

where $g_1, \cdots, g_m \in \mathbb{R}[x]$. In the preceding chapter we defined a lower bound for $p_{\min}$ obtained by considering sums of squares of polynomials. Here we consider another approach, which will turn out to be dual to the sums of squares approach.

Say, $p = \sum_\alpha p_\alpha x^\alpha$, where there are only finitely many nonzero coefficients p_α and let $\boldsymbol{p} = (p_\alpha)_{\alpha \in \mathbb{N}^n}$ denote the vector of coefficients of p, so $p_\alpha = 0$ for all $|\alpha| > \deg(p)$. Moreover, let $[x]_\infty = (x^\alpha)_{\alpha \in \mathbb{N}^n}$ denote the vector consisting of all monomials x^α. Then, one can write:

$$p(x) = \sum_\alpha p_\alpha x^\alpha = \boldsymbol{p}^{\mathsf{T}} [x]_\infty.$$

We define the set $\mathcal{C}_\infty(K)$ as the convex hull of the vectors $[x]_\infty$ for $x \in K$:

$$\mathcal{C}_\infty(K) = \operatorname{conv}\{[x]_\infty : x \in K\}. \tag{14.3}$$

Let us introduce a new variable $y_\alpha = x^\alpha$ for each monomial. Then, using these variables $y = (y_\alpha)$ and the set $\mathcal{C}_\infty(K)$, we can reformulate problem (14.1)

equivalently as

$$p_{\min} = \inf_{x \in K} p(x) = \inf_{x \in K} \boldsymbol{p}^{\mathsf{T}}[x]_\infty = \inf_{y=(y_\alpha)_{\alpha \in \mathbb{N}^n}} \{\boldsymbol{p}^{\mathsf{T}} y : y \in \mathcal{C}_\infty(K)\}. \tag{14.4}$$

This leads naturally to the problem of understanding which sequences y belong to the set $\mathcal{C}_\infty(K)$. In this chapter we give a characterization for the set $\mathcal{C}_\infty(K)$, we will use it in the next chapter as a tool for deriving global optimal solutions to the polynomial optimization problem (14.1).

This chapter is organized as follows. We introduce some algebraic facts about polynomial ideals $I \subseteq \mathbb{R}[x]$ and their associated quotient spaces $\mathbb{R}[x]/I$, which we will need for the characterization of the set $\mathcal{C}_\infty(K)$. Using these tools we can also describe the so-called *eigenvalue method* for computing the complex solutions of a system of polynomial equations. This method also gives a useful tool to extract the global optimizers of problem (14.1). Then we give a characterization for the sequences y belonging to the set $\mathcal{C}_\infty(K)$, in terms of associated (moment) matrices required to be positive semidefinite.

14.1 The quotient algebra $\mathbb{R}[x]/I$

14.1.1 (Real) radical ideals and the (Real) Nullstellensatz

Here, $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ denotes the field of real or complex numbers. A set $I \subseteq \mathbb{K}[x]$ is an *ideal* if $I + I \subseteq I$ and $\mathbb{K}[x] \cdot I \subseteq I$. Given polynomials $h_1, \cdots, h_m$, the ideal *generated by the h_j's* is

$$I = (h_1, \cdots, h_m) = \left\{ \sum_{j=1}^m u_j h_j : u_j \in \mathbb{K}[x] \right\}.$$

A basic property of the polynomial ring $\mathbb{K}[x]$ is that it is Noetherian: every ideal admits a finite set of generators. Given a subset $V \subseteq \mathbb{C}$, the set

$$\mathcal{I}(V) = \{f \in \mathbb{K}[x] : f(x) = 0 \ \forall x \in V\}$$

is an ideal, called the *vanishing ideal* of V.

The *complex variety* of an ideal $I \subseteq \mathbb{K}[x]$ is

$$V_{\mathbb{C}}(I) = \{x \in \mathbb{C}^n : f(x) = 0 \ \forall f \in I\}$$

and its *real variety* is

$$V_{\mathbb{R}}(I) = \{x \in \mathbb{R}^n : f(x) = 0 \ \forall f \in I\} = V_{\mathbb{C}}(I) \cap \mathbb{R}^n.$$

The elements $x \in V_{\mathbb{C}}(I)$ are also called the common *roots* of the polynomials in I. Clearly, if $I = (h_1, \cdots, h_m)$ is generated by the h_j's, then $V_{\mathbb{C}}(I)$ is the set of common complex roots of the polynomials $h_1, \cdots, h_m$ and $V_{\mathbb{R}}(I)$ is their set of common real roots.

Given an ideal $I \subseteq \mathbb{K}[x]$, the set

$$\sqrt{I} = \{f \in \mathbb{K}[x] : f^m \in I \text{ for some } m \in \mathbb{N}\} \tag{14.5}$$

is an ideal (Exercise 14.1), called the *radical* of I. Clearly we have the inclusions:

$$I \subseteq \sqrt{I} \subseteq \mathcal{I}(V_{\mathbb{C}}(I)).$$

Consider, for instance, the ideal $I = (x^2)$ generated by the monomial x^2. Then, $V_{\mathbb{C}}(I) = \{0\}$. The polynomial x belongs to $\sqrt{I}$ and to $\mathcal{I}(V_{\mathbb{C}}(I))$, but x does not belong to I. Hilbert's Nullstellensatz states that both ideals $\sqrt{I}$ and $\mathcal{I}(V_{\mathbb{C}}(I))$ coincide:

Theorem 14.1.1. (Hilbert's Nullstellensatz) *For any ideal $I \subseteq \mathbb{K}[x]$, we have equality:*

$$\sqrt{I} = \mathcal{I}(V_{\mathbb{C}}(I)).$$

That is, a polynomial f vanishes at all $x \in V_{\mathbb{C}}(I)$ if and only if some power of f belongs to I.

The ideal I is said to be *radical* if $I = \sqrt{I}$ or, equivalently (in view of the Nullstellensatz), $I = \mathcal{I}(V_{\mathbb{C}}(I))$. For instance, the ideal $I = (x^2)$ is not radical. Note that 0 is a root with double multiplicity. Roughly speaking, an ideal is radical when all roots $x \in V_{\mathbb{C}}(I)$ have single multiplicity, but we will not go into details about multiplicities of roots.

Given an ideal $I \subseteq \mathbb{R}[x]$, the set

$$\sqrt[\mathbb{R}]{I} = \{f \in \mathbb{R}[x] : f^{2m} + s \in I \quad \text{for some } m \in \mathbb{N}, s \in \Sigma\} \tag{14.6}$$

is an ideal in $\mathbb{R}[x]$ (Exercise 14.1), called the *real radical* of I. Clearly we have the inclusions:

$$I \subseteq \sqrt[\mathbb{R}]{I} \subseteq \mathcal{I}(V_{\mathbb{R}}(I)).$$

As an example, consider the ideal $I = (x^2+y^2) \subseteq \mathbb{R}[x, y]$. Then, $V_{\mathbb{R}}(I) = \{(0, 0)\}$ while $V_{\mathbb{C}}(I) = \{(x, \pm\mathbf{i}x) : x \in \mathbb{C}\}$. Both polynomials x and y belong to $\sqrt[\mathbb{R}]{I}$ and to $\mathcal{I}(V_{\mathbb{R}}(I))$. The Real Nulstellensatz states that both ideals $\sqrt[\mathbb{R}]{I}$ and $\mathcal{I}(V_{\mathbb{R}}(I))$ coincide.

Theorem 14.1.2. (The Real Nullstellensatz*) For any ideal $I \subseteq \mathbb{R}[x]$,*

$$\sqrt[\mathbb{R}]{I} = \mathcal{I}(V_{\mathbb{R}}(I)).$$

That is, a polynomial $f \in \mathbb{R}[x]$ vanishes at all common real roots of I if and only if the sum of an even power of f and of a sum of squares belongs to I.

We will use the following characterization of (real) radical ideals (see Exercise 14.2).

Lemma 14.1.3.

(i) *An ideal $I \subseteq \mathbb{K}[x]$ is radical (i.e., $\sqrt{I} = I$) if and only if*

$$\forall f \in \mathbb{K}[x] \;\; f^2 \in I \Longrightarrow f \in I.$$

(ii) *An ideal $I \subseteq \mathbb{R}[x]$ is real radical (i.e., $\sqrt[\mathbb{R}]{I} = I$) if and only if*

$$\forall f_1, \cdots, f_m \in \mathbb{R}[x] \quad f_1^2 + \cdots + f_m^2 \in I \Longrightarrow f_1, \cdots, f_m \in I.$$

It is good to realize that, if V is a complex variety, i.e., if $V = V_{\mathbb{C}}(I)$ for some ideal I, then $V_{\mathbb{C}}(\mathcal{I}(V)) = V$. Indeed, the inclusion $V_{\mathbb{C}}(I) \subseteq V_{\mathbb{C}}(\mathcal{I}(V_{\mathbb{C}}(I)))$ is clear. Moreover, if $v \notin V_{\mathbb{C}}(I)$, then there is a polynomial $f \in I \subseteq \mathcal{I}(V_{\mathbb{C}}(I))$ such that $f(v) \neq 0$, thus showing $v \notin V_{\mathbb{C}}(\mathcal{I}(V_{\mathbb{C}}(I)))$.

However, the inclusion $V \subseteq V_{\mathbb{C}}(\mathcal{I}(V))$ can be strict if V is not a complex variety. For example, for $V = \mathbb{C}\backslash\{0\} \subseteq \mathbb{C}$, $\mathcal{I}(V) = \{0\}$, since the zero polynomial is the only polynomial vanishing at all elements of V. Hence, $V_{\mathbb{C}}(\mathcal{I}(V)) = \mathbb{C}$ contains strictly V.

For any ideal I, we have the inclusions:

$$I \subseteq \mathcal{I}(V_{\mathbb{C}}(I)) \subseteq \mathcal{I}(V_{\mathbb{R}}(I)),$$

with equality throughout if I is real radical. Yet this does not imply in general that $V_{\mathbb{C}}(I) = V_{\mathbb{R}}(I)$, i.e., that all roots are real. As an example illustrating this, consider e.g. the ideal $I = (x - y) \subseteq \mathbb{R}[x, y]$; then I is real radical, but $V_{\mathbb{R}}(I) \subset V_{\mathbb{C}}(I)$. However, equality holds if $V_{\mathbb{R}}(I)$ is finite.

Lemma 14.1.4. *If $I \subseteq \mathbb{R}[x]$ is a real radical ideal, with finite real variety: $|V_{\mathbb{R}}(I)| < \infty$, then $V_{\mathbb{C}}(I) = V_{\mathbb{R}}(I)$.*

Proof. By assumption, equality: $\mathcal{I}(V_{\mathbb{R}}(I)) = \mathcal{I}(V_{\mathbb{C}}(I))$ holds. Hence these two ideals have the same complex variety: $V_{\mathbb{C}}(\mathcal{I}(V_{\mathbb{R}}(I))) = V_{\mathbb{C}}(\mathcal{I}(V_{\mathbb{C}}(I)))$. This implies equality $V_{\mathbb{R}}(I) = V_{\mathbb{C}}(I)$, since $V_{\mathbb{R}}(I)$ is a complex variety (as it is finite, see Exercise 14.3) and $V_{\mathbb{C}}(I)$ too is a complex variety (by definition). □

14.1.2 The dimension of the quotient algebra $\mathbb{K}[x]/I$

Let I be an ideal in $\mathbb{K}[x]$. We define the quotient space $\mathcal{A} = \mathbb{K}[x]/I$, whose elements are the cosets

$$[f] = f + I = \{f + q : q \in I\}$$

for $f \in \mathbb{K}[x]$. Then $\mathcal{A}$ is an algebra with addition: $[f] + [g] = [f + g]$, scalar multiplication $\lambda[f] = [\lambda f]$, and multiplication $[f][g] = [fg]$, for $f, g \in \mathbb{K}[x]$ and $\lambda \in \mathbb{K}$. These operations are well defined. Indeed, if $[f] = [f']$ and $[g] = [g']$, i.e., f', g' are other representatives in the cosets $[f], [g]$, respectively, so that $f - f',\ g - g' \in I$, then

$$(f' + g') - (f + g) \in I, \lambda f' - \lambda f \in I,\ f'g' - fg = (f' - g')g' + f(g' - g) \in I.$$

As we now see, the dimension of the quotient space $\mathcal{A}$ is related to the cardinality of the complex variety $V_{\mathbb{C}}(I)$.

Theorem 14.1.5. *Let $I \subseteq \mathbb{K}[x]$ be an ideal and let $\mathcal{A} = \mathbb{K}[x]/I$ be the associated quotient space.*

(i) $\dim \mathcal{A} < \infty$ *if and only if* $|V_{\mathbb{C}}(I)| < \infty$.

(ii) *Assume* $|V_{\mathbb{C}}(I) < \infty$. *Then* $|V_{\mathbb{C}}(I)| \leqslant \dim \mathcal{A}$, *with equality if and only if the ideal I is radical (i.e., $I = \sqrt{I}$).*

Remark 14.1.6. *Let I be an ideal in $\mathbb{R}[x]$. Then the set $I + \mathbf{i}I = \{f + \mathbf{i}g : f, g \in I\}$ is an ideal in $\mathbb{C}[x]$ and it is easy to check that the two quotient spaces $\mathbb{R}[x]/I$ and $\mathbb{C}[x]/(I + \mathbf{i}I)$ have the same dimension. Hence, in order to compute the dimension of $\mathbb{R}[x]/I$, we can as well deal with the corresponding ideal $I + \mathbf{i}I$ in the complex polynomial ring.*

For the proof of Theorem 14.1.5, it is useful to have the following construction of interpolation polynomials.

Lemma 14.1.7. *Let $V \subseteq \mathbb{K}^n$ be a finite set. There exist polynomials $p_v \in \mathbb{K}[x]$ for $v \in V$ satisfying the following property:*

$$p_v(u) = \delta_{u,v} \ \forall u, v \in V.$$

They are called interpolation polynomials *at the points of V. Then, for any polynomial $f \in \mathbb{K}[x]$,*

$$f - \sum_{v \in V_{\mathbb{C}}(I)} f(v) p_v \in \mathcal{I}(V_{\mathbb{C}}(I)). \tag{14.7}$$

Proof. Fix $v \in V$. For any $u \in V \backslash \{v\}$, let i_u be a coordinate where v and u differ, i.e., $v_{i_u} \neq u_{iu}$. Then define the polynomial p_v by

$$p_v = \prod_{u \in V \backslash \{v\}} \frac{x_{i_u} - u_{i_u}}{v_{i_u} - u_{i_u}}.$$

Clearly, $p_v(v) = 1$ and $p_v(u) = 0$ if $u \in V$, $u \neq v$. By construction the polynomial in (14.7) vanishes at all $v \in V_{\mathbb{C}}(I)$ and thus belongs to $\mathcal{I}(V_{\mathbb{C}}(I))$. □

Example 14.1.8. *Say, $V = \{(0,0), (1,0), (0,2)\} \subseteq \mathbb{R}^2$. Then the polynomials $p_{(0,0)} = (x_1 - 1)(x_2 - 2)/2$, $p_{(1,0)} = x_1^2$ and $p_{(0,2)} = x_2(1 - x_1)/2$ are interpolation polynomials at the points of V.*

Lemma 14.1.9. *Let I be an ideal in $\mathbb{C}[x]$ and $\mathcal{A} = \mathbb{C}[x]/I$. Assume $V_{\mathbb{C}}(I)$ is finite, let p_v ($v \in V_{\mathbb{C}}(I)$) be interpolation polynomials at the points of $V_{\mathbb{C}}(I)$, and let*

$$\mathcal{L} = \{[p_v] : v \in V_{\mathbb{C}}(I)\}$$

be the corresponding set of cosets in $\mathcal{A}$. Then,

(i) *$\mathcal{L}$ is linearly independent in $\mathcal{A}$.*

(ii) *$\mathcal{L}$ generates the vector space $\mathbb{C}[x]/\mathcal{I}(V_{\mathbb{C}}(I))$.*

(iii) *If I is radical, then $\mathcal{L}$ is a basis of $\mathcal{A}$ and* $\dim \mathcal{A} = |V_{\mathbb{C}}(I)|$.

Proof. (i) Assume that $\sum_{v \in V_{\mathbb{C}}(I)} \lambda_v [p_v] = 0$ for some scalars λ_v. That is, the polynomial $f = \sum_{v \in V_{\mathbb{C}}(I)} \lambda_v p_v$ belongs to I. By evaluating the polynomial f at each $v \in V_{\mathbb{C}}(I)$ and using the fact that $p_v(v) = 1$ and $p_v(u) = 0$ if $u \in V_{\mathbb{C}}(I) \backslash \{v\}$, we deduce that $\lambda_v = 0$ for all v. This shows that $\mathcal{L}$ is linearly independent in $\mathcal{A}$.

(ii) Relation (14.7) implies directly that $\mathcal{L}$ is generating in $\mathbb{K}[x]/\mathcal{I}(V_{\mathbb{C}}(I))$.

(iii) Assume that I is radical and thus $I = \mathcal{I}(V_{\mathbb{C}}(I))$ (by the Nullstellensatz). Then, $\mathcal{L}$ is linearly independent and generating in $\mathcal{A}$ and thus a basis of $\mathcal{A}$. □

Proof. (of Theorem 14.1.5). In view of Remark 14.1.6, we may assume $\mathbb{K} = \mathbb{C}$. (i) Assume first that $\dim \mathcal{A} = k < \infty$, we show that $|V_{\mathbb{C}}(I)| < \infty$. For this, pick a variable x_i and consider the $k+1$ cosets $[1], [x_i], \cdots, [x_i^k]$. Then they are linearly dependent in $\mathcal{A}$ and thus there exist scalars λ_h $(0 \leqslant h \leqslant k)$ (not all zero) for which the (univariate) polynomial $f = \sum_{h=0}^{k} \lambda_h x_i^h$ is a nonzero polynomial belonging to I. As f is univariate, it has finitely many roots. This implies that the i-th coordinates of the points $v \in V_{\mathbb{C}}(I)$ take only finitely many values. As this holds for all coordinates we deduce that $V_{\mathbb{C}}(I)$ is finite.

Assume now that $|V_{\mathbb{C}}(I)| < \infty$, we show that $\dim \mathcal{A} < \infty$. For this, assume that the i-th coordinates of the points $v \in V_{\mathbb{C}}(I)$ take k distinct values: $a_1, \cdots, a_k \in \mathbb{C}$. Then the polynomial $f = (x_i - a_1) \cdots (x_i - a_k)$ vanishes at all $v \in V_{\mathbb{C}}(I)$. Applying the Nullstellensatz, $f^m \in I$ for some integer $m \in \mathbb{N}$. This implies that there is a linear dependency among the cosets $[1], [x_i], \cdots, [x_i^{mk}]$. Therefore, there exists an integer n_i for which $[x_i^{n_i}]$ lies in the linear span of $\{[x_i^h] : 0 \leqslant h \leqslant n_i - 1\}$. From this one can easily derive that the set $\{[x^\alpha] : 0 \leqslant \alpha_i \leqslant n_i - 1, i \in [n]\}$ generates the vector space $\mathcal{A}$, thus showing that $\dim \mathcal{A} < \infty$.

(ii) Assume $V_{\mathbb{C}}(I)$ is finite. If I is radical then equality $\dim \mathcal{A} = |V_{\mathbb{C}}(I)|$ follows from Lemma 14.1.9 (iii). Assume now that I is not radical and let $f \in \sqrt{I} \backslash I$. If p_v $(v \in V_{\mathbb{C}}(I))$ are interpolation polynomials at the points of $V_{\mathbb{C}}(I)$, one can easily verify that the system $\{[p_v] : v \in V_{\mathbb{C}}(I)\} \cup \{[f]\}$ is linearly independent in $\mathcal{A}$, so that $\dim \mathcal{A} > |V_{\mathbb{C}}(I)|$. □

14.1.3 The eigenvalue method for complex roots

A basic, fundamental problem in mathematics and many areas of applications is how to solve a system of polynomial equations: $h_1(x) = 0, \cdots, h_m(x) = 0$. In other words, how to compute the complex variety of the ideal $I = (h_1, \cdots, h_m)$. Here we assume that $I \subseteq \mathbb{K}[x]$ is an ideal which has *finitely many complex roots*: $|V_{\mathbb{C}}(I)| < \infty$. We now describe a well known method for finding the elements of $V_{\mathbb{C}}(I)$, which is based on computing the eigenvalues of a suitable linear map on the algebra $\mathcal{A} = \mathbb{K}[x]/I$.

Namely, given an arbitrary polynomial $h \in \mathbb{K}[x]$, we consider the following 'multiplication by h' linear map:

$$\begin{array}{rccc} m_h: & \mathcal{A} & \to & \mathcal{A} \\ & [f] & \mapsto & [fh]. \end{array} \tag{14.8}$$

As $V_{\mathbb{C}}(I)$ is finite we known from Theorem 14.1.5 that the vector space $\mathcal{A}$ has finite dimension. Say, $N = \dim \mathcal{A}$, then $N \geqslant |V_{\mathbb{C}}(I)|$, with equality if I is radical (by Theorem 14.1.5).

Let us choose a set of cosets $\mathcal{B} = \{[b_1], \cdots, [b_N]\}$ forming a basis of $\mathcal{A}$ and let M_h denote the matrix of m_h with respect to the base $\mathcal{B}$ (which not symmetric in general). Then, for $v \in V_{\mathbb{C}}(I)$, we define the vector $[v]_{\mathcal{B}} = (b_j(v))_{j=1}^N$ whose entries are the evaluations at v of the polynomials in $\mathcal{B}$.

Lemma 14.1.10. *The vectors $\{[v]_{\mathcal{B}} : v \in V_{\mathbb{C}}(I)\}$ are linearly independent.*

Proof. Assume $\sum_{v \in V_{\mathbb{C}}(I)} \lambda_v [v]_{\mathcal{B}} = 0$ for some scalars λ_v, i.e., $\sum_{v \in V_{\mathbb{C}}(I)} \lambda_v b_j(v) = 0$ for all $j \in [N]$. As $\mathcal{B}$ is a base of $\mathcal{A}$, this implies that $\sum_{v \in V_{\mathbb{C}}(I)} \lambda_v f(v) = 0$ for any $f \in \mathbb{K}[x]$ (check it). Applying this to the polynomial $f = p_v$, we obtain that $\lambda_v = 0$ for all $v \in V_{\mathbb{C}}(I)$. □

As we now show, the matrix M_h carries out useful information about the elements of $V_{\mathbb{C}}(I)$: its eigenvalues are the evaluations $h(v)$ of h at the points $v \in V_{\mathbb{C}}(I)$ and its left eigenvectors are the vectors $[v]_{\mathcal{B}}$.

Theorem 14.1.11. *Let $h \in \mathbb{K}[x]$, let $I \subseteq \mathbb{K}[x]$ be an ideal with $|V_{\mathbb{C}}(I)| < \infty$, and let m_h be the linear map from (14.8).*

(i) *Let $\mathcal{B}$ be a base of $\mathcal{A}$ and let M_h be the matrix of m_h in the base $\mathcal{B}$. Then, for each $v \in V_{\mathbb{C}}(I)$, the vector $[v]_{\mathcal{B}}$ is a left eigenvector of M_h with eigenvalue $h(v)$, i.e.,*

$$M_h^{\mathsf{T}} [v]_{\mathcal{B}} = h(v)[v]_{\mathcal{B}}. \tag{14.9}$$

(ii) *The set $\{h(v) : v \in V_{\mathbb{C}}(I)\}$ is the set of eigenvalues of m_h.*

(iii) *Assume that I is radical and let p_v ($v \in V_{\mathbb{C}}(I)$) be interpolation polynomials at the points of $V_{\mathbb{C}}(I)$. Then,*

$$m_h([p_u]) = h(u)[p_u]$$

for all $u \in V_{\mathbb{C}}(I)$. Therefore, the matrix of m_h in the base $\{[p_v] : v \in V_{\mathbb{C}}(I)\}$ is a diagonal matrix with $h(v)$ ($v \in V_{\mathbb{C}}(I)$) as diagonal entries.

Proof. (i) Say, $M_h = (a_{ij})_{i,j=1}^N$, so that

$$[hb_j] = \sum_{i=1}^N a_{ij}[b_i], \quad \text{i.e., } hb_j - \sum_{i=1}^N a_{ij} b_i \in I.$$

Evaluating the above polynomial at $v \in V_{\mathbb{C}}(I)$ gives directly relation (14.9).

(ii) By (i), we already know that each scalar $h(v)$ is an eigenvalue of M_h^T and thus of m_h. We now show that the scalars $h(v)$ ($v \in V_\mathbb{C}(I)$) are the *only* eigenvalues of m_h. For this, let $\lambda \notin \{h(v) : v \in V_\mathbb{C}(I)\}$, we show that λ is not an eigenvalue of m_h. Let J denote the ideal generated by $I \cup \{h - \lambda\}$. Then, $V_\mathbb{C}(J) = \varnothing$. Applying the Nullstellensatz, we obtain that $1 \in J$ and thus $1 - u(h - \lambda) \in I$ for some $u \in \mathbb{K}[x]$. It suffices now to observe that the latter implies that $m_u(m_h - \lambda\mathsf{id}) = \mathsf{id}$, where id is the identity map from $\mathcal{A}$ to $\mathcal{A}$. But then $m_h - \lambda\mathsf{id}$ is nonsingular, which implies that λ is not an eigenvalue of m_h.

(iii) Assume that I is radical and let $\{p_v : v \in V_\mathbb{C}(I)\}$ be interpolation polynomials. Using relation (14.7), we obtain that $m_h([f]) = \sum_{v \in V_\mathbb{C}(I)} f(v)h(v)[p_v]$ for any polynomial f. In particular, $m_h([p_v]) = h(v)[p_v]$. □

Here is a simple strategy on how to use the above result in order to compute the points $v \in V_\mathbb{C}(I)$. Assume that the ideal I is radical (this will be the case in our application to polynomial optimization) and suppose that we have a polynomial h for which the values $h(v)$ ($v \in V_\mathbb{C}(I)$) are pairwise distinct (e.g. pick a linear polynomial h with random coefficients). Suppose also that we know a base $\mathcal{B}$ of $\mathcal{A}$ and that we know the matrix M_h of m_h in this base. We know from Theorem 14.1.11 that M_h has $N = |V_\mathbb{C}(I)|$ distinct eigenvalues so that each eigenspace has dimension 1. Hence, by computing the eigenvectors of M_h^T, we can recover the vectors $[v]_\mathcal{B} = (b_j(v))_{j=1}^N$ (up to scaling). In order to compute the i-th coordinate v_i of v, just express the coset $[x_i]$ in the base $\mathcal{B}$: If $[x_i] = \sum_{j=1}^N c_{ij}[b_j]$ for some scalars c_{ij}, then $v_i = \sum_{j=1}^N c_{ij} b_j(v)$.

Example 14.1.12. *Let $I = (x^3 - 6x^2 + 11x - 6)$ be the ideal generated by the polynomial $x^3 - 6x^2 + 11x - 6 = (x-1)(x-2)(x-3)$ (univariate case). Then, $V_\mathbb{C}(I) = \{1, 2, 3\}$ and $\mathcal{B} = \{[1], [x], [x^2]\}$ is a base of $\mathcal{A} = \mathbb{R}[x]/I$. With respect to this base $\mathcal{B}$, the matrix of the multiplication operator by x is*

$$M_x = \begin{array}{c} \\ [1] \\ [x] \\ [x^2] \end{array} \begin{array}{c} \begin{array}{ccc} [x] & [x^2] & [x^3] \end{array} \\ \begin{pmatrix} 0 & 0 & 6 \\ 1 & 0 & -11 \\ 0 & 1 & 6 \end{pmatrix} \end{array}$$

(built using the relation $[x^3] = 6[1] - 11[x] + 6[x^2]$). It is an easy exercise to verify that M_x^T has three eigenvectors: $(1, 1, 1)$ with eigenvalue $\lambda = 1$, $(1, 2, 4)$ with eigenvalue $\lambda = 2$, and $(1, 3, 9)$ with eigenvalue $\lambda = 3$. Thus the eigenvectors are indeed of the form $[v]_\mathcal{B} = (1, v, v^2)$ for $v \in \{1, 2, 3\}$.

The polynomials $p_1 = (x-2)(x-3)/2$, $p_2 = -(x-1)(x-3)$ and $p_3 = (x-1)(x-2)/2$ are interpolation polynomials at the roots $v = 1, 2, 3$. Note that the matrix of m_x with respect to the base $\{[p_1], [p_2], [p_3]\}$ is

$$\begin{array}{c} \\ [p_1] \\ [p_2] \\ [p_3] \end{array} \begin{array}{c} \begin{array}{ccc} [xp_1] & [xp_2] & [xp_3] \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}, \end{array}$$

thus indeed a diagonal matrix with the values $v = 1, 2, 3$ as diagonal entries.

Finally, we indicate how to compute the number of real roots using the multiplication operators. This is a classical result, going back to work of Hermite in the univariate case. You will prove it in Exercise 14.4 for radical ideals.

Theorem 14.1.13. *Let I be an ideal in $\mathbb{R}[x]$ with $|V_{\mathbb{C}}(I)| < \infty$. Define the Hermite quadratic form:*

$$\begin{array}{rccc} \mathcal{H}: & \mathbb{R}[x]/I \times \mathbb{R}[x]/I & \to & \mathbb{R} \\ & ([f],[g]) & \mapsto & \mathrm{Tr}(m_{fg}), \end{array} \tag{14.10}$$

where $\mathrm{Tr}(m_{fg})$ denotes the trace of the multiplication operator by fg. Let $\sigma_+(\mathcal{H})$ (resp., $\sigma_-(\mathcal{H})$) denote the number of positive eigenvalues (resp., negative eigenvalues) of $\mathcal{H}$. Then, the rank of $\mathcal{H}$ is equal to $|V_{\mathbb{C}}(I)|$ and

$$\sigma_+(\mathcal{H}) - \sigma_-(\mathcal{H}) = |V_{\mathbb{R}}(I)|.$$

14.2 Characterizing the set $\mathcal{C}_\infty(K)$

Our goal in this section is to characterize the set $\mathcal{C}_\infty(K)$ from (14.3). We need one more ingredient: moment matrices.

14.2.1 Moment matrices

Let $y = (y_\alpha)_{\alpha \in \mathbb{N}^n}$ be a sequence of real numbers indexed by $\mathbb{N}^n$. It is convenient to introduce the corresponding linear functional L on the polynomial ring:

$$\begin{array}{rccc} L: & \mathbb{R}[x] & \to & \mathbb{R} \\ & x^\alpha & \mapsto & L(x^\alpha) = y_\alpha \\ & f = \sum_\alpha f_\alpha x^\alpha & \mapsto & L(f) = \sum_\alpha f_\alpha y_\alpha. \end{array} \tag{14.11}$$

Consider first the case when $y = [v]_\infty$ for some $v \in \mathbb{R}^n$. Then, L is the *evaluation at v* (denoted as L_v) since $L(f) = \sum_\alpha f_\alpha v^\alpha = f(v)$ for $f \in \mathbb{R}[x]$. Moreover, the matrix yy^{T} has a special structure: its (α,β)-th entry is equal to $v^\alpha v^\beta = v^{\alpha+\beta} = y_{\alpha+\beta}$, thus depending only on the sum of the indices α and β. This observation motivates the following definition.

Definition 14.2.1. *Given a sequence $y = (y_\alpha)_{\alpha \in \mathbb{N}^n}$ of real numbers, its* moment matrix *is the real symmetric (infinite) matrix indexed by $\mathbb{N}^n$, defined by*

$$M(y) = (y_{\alpha+\beta})_{\alpha,\beta \in \mathbb{N}^n}.$$

Next we observe that nonnegativity of L on the cone Σ of sums of squares can be reformulated in terms of positive semidefiniteness of the moment matrix $M(y)$.

Lemma 14.2.2. *Let $y = (y_\alpha)_{\alpha \in \mathbb{N}^n}$ be a sequence of real numbers and let L be the associated linear functional from (14.11). For any polynomials $f, g \in \mathbb{R}[x]$:*

$$L(f^2) = \boldsymbol{f}^{\mathsf{T}} M(y) \boldsymbol{f}, \quad L(gf^2) = \boldsymbol{f}^{\mathsf{T}} M(g * y) \boldsymbol{f},$$

where $g * y \in \mathbb{R}^{\mathbb{N}^n}$ is the new sequence with α-th entry

$$(g * y)_\alpha = L(gx^\alpha) = \sum_\gamma g_\gamma y_{\alpha+\gamma} \quad \forall \alpha \in \mathbb{N}^n.$$

Therefore, $L \geqslant 0$ on Σ if and only if $M(y) \succeq 0$, and $L \geqslant 0$ on $g\Sigma$ if and only if $M(g * y) \succeq 0$.

Proof. For $f = \sum_\alpha f_\alpha x^\alpha$, $g = \sum_\gamma g_\gamma x^\gamma$, we have:

$$L(f^2) = L\left(\sum_{\alpha,\beta} f_\alpha f_\beta x^{\alpha+\beta}\right) = \sum_{\alpha,\beta} f_\alpha f_\beta y_{\alpha+\beta} = \sum_{\alpha,\beta} f_\alpha f_\beta M(y)_{\alpha,\beta} = \boldsymbol{f}^{\mathsf{T}} M(y) \boldsymbol{f},$$

$$L(gf^2) = L\left(\sum_{\alpha,\beta,\gamma} f_\alpha f_\beta g_\gamma x^{\alpha+\beta+\gamma}\right) = \sum_{\alpha,\beta} f_\alpha f_\beta L(gx^\gamma) = \boldsymbol{f}^{\mathsf{T}} M(g * y) \boldsymbol{f}.$$

These two identities give directly the result of the lemma. □

Next we observe that the kernel of $M(y)$ can be seen as an ideal of $\mathbb{R}[x]$, which is real radical when $M(y) \succeq 0$. This observation will play a cucial role in the characterization of the set $\mathcal{C}_\infty(K)$ in the next section.

Lemma 14.2.3. *Let $y = (y_\alpha)_{\alpha \in \mathbb{N}^n}$ be a sequence of real numbers and let L be the associated linear functional from (14.11). Set*

$$I = \{f \in \mathbb{R}[x] : L(fh) = 0 \ \forall h \in \mathbb{R}[x]\}. \tag{14.12}$$

(i) *A polynomial f belongs to I if and only if its coefficient vector $\boldsymbol{f}$ belongs to the kernel of $M(y)$.*

(ii) *I is an ideal in $\mathbb{R}[x]$.*

(iii) *If $M(y) \succeq 0$ then the ideal I is real radical.*

Proof. (i), (ii): Direct verification.
(iii) Using Lemma 14.2.2 and the fact that $M(y) \succeq 0$, the following holds for any polynomial f:

$$L(f^2) = \boldsymbol{f}^{\mathsf{T}} M(y) \boldsymbol{f} \geqslant 0 \text{ and } L(f^2) = 0 \Longrightarrow M(y)\boldsymbol{f} = 0 \Longrightarrow f \in I.$$

We now show that I is real radical, using the characterization from Lemma 14.1.3: Assume that $\sum_i f_i^2 \in I$. Then, $0 = L(\sum_i f_i^2) = \sum_i L(f_i^2)$ and thus $L(f_i^2) = 0$, which in turn implies that $f_i \in I$ for all i. □

14.2.2 Finite rank positive semidefinite moment matrices

We can now characterize the sequences belonging to the set $\mathcal{C}_\infty(K)$, in terms of positivity and rank conditions on their moment matrices.

Theorem 14.2.4. *Let K be the set from (15.2). Let $y = (y_\alpha)_{\alpha\in\mathbb{N}^n}$ be a sequence of real numbers and let L be the linear functional from (14.11). The following assertions are equivalent.*

(i) $y \in \mathcal{C}_\infty(K)$*, i.e.,* $y = \sum_{i=1}^r \lambda_i [v_i]_\infty$ *for some scalars* $\lambda_i > 0$ *and* $v_i \in K$.

(ii) rank $M(y) < \infty$, $M(y) \succeq 0$ *and* $M(g_j * y) \succeq 0$ *for* $j \in [m]$.

(iii) rank $M(y) < \infty$ *and* $L \geqslant 0$ *on* $\Sigma + g_1\Sigma + \cdots + g_m\Sigma$.

Proof. Assume that (i) holds. Then, $M(y) = \sum_{i=1}^r \lambda_i M([v_i]_\infty)$ is positive semidefinite (since $M([v_i]_\infty) \succeq 0$ for each i) and $M(y)$ has finite rank. For $i \in [r]$ and $j \in [m]$, we have that $g_j * [v_i]_\infty = g_j(v_i)[v_i]_\infty$ with $g_j(v_i) \geqslant 0$. Therefore, $M(g_j * y) = \sum_{i=1}^r \lambda_i g_j(v_i) M([v_i]_\infty)$ is positive semidefinite. This shows (ii).

The equivalence of (ii) and (iii) follows directly from Lemma 14.2.2.

We now show the implication (ii) $\Longrightarrow$ (i). Assume that rank $M(y) = r < \infty$, $M(y) \succeq 0$, $M(g_j * y) \succeq 0$ for $j \in [m]$; we show (i). Let L be the linear functional from (14.11) and let I be the set from (14.12). By Lemma 15.3.1, we know that I is a real radical ideal in $\mathbb{R}[x]$. First we claim that

$$\dim \mathbb{R}[x]/I = r.$$

This follows directly from the fact that a set of columns $\{C_1, \cdots, C_s\}$ of $M(y)$, indexed (say) by $\{\alpha_1, \cdots, \alpha_s\} \subseteq \mathbb{N}^n$, is linearly independent if and only if the corresponding cosets of monomials $\{[x^{\alpha_1}], \cdots, [x^{\alpha_s}]\}$ is linearly independent in $\mathbb{R}[x]/I$.

As $\dim \mathbb{R}[x]/I = r < \infty$, we deduce using Lemma 14.1.9 that $|V_\mathbb{C}(I)| < \infty$; moreover, $|V_\mathbb{C}(I)| = \dim \mathbb{R}[x]/I = r$ since I is real radical (and thus radical). Furthermore, using Lemma 14.1.4, we deduce that $V_\mathbb{R}(I) = V_\mathbb{C}(I)$. Say,

$$V_\mathbb{C}(I) = \{v_1, \cdots, v_r\} \subseteq \mathbb{R}^n.$$

Let $p_{v_1}, \cdots, p_{v_r} \in \mathbb{R}[x]$ be interpolation polynomials at the v_i's. We next claim that

$$L = \sum_{i=1}^r L(p_{v_i}) L_{v_i}, \;\; \text{i.e.,} \;\; y = \sum_{i=1}^r L(p_{v_i})[v_i]_\infty, \tag{14.13}$$

where L_{v_i} is the evaluation at v_i. As both L and $L' = \sum_{i=1}^r L(p_{v_i}) L_{v_i}$ vanish at all polynomials in I, in order to show that $L = L'$, it suffices to show that L and L' coincide at all elements of a given base of $\mathbb{R}[x]/I$. Now, by Lemma 14.1.9, we know that the set $\{[p_{v_1}], \cdots, [p_{v_r}]\}$ is a base of $\mathbb{R}[x]/I$ and it is indeed true that $L'(p_{v_i}) = L(p_{v_i})$ for all i. Thus (14.13) holds.

Next, we claim that

$$L(p_{v_i}) > 0 \;\; \text{for all } i \in [r].$$

Indeed, $L(p_{v_i}) = L(p_{v_i}^2)$, since $p_{v_i} - p_{v_i}^2 \in I$ (as it vanishes at all points of $V_{\mathbb{C}}(I)$ and I is radical). Therefore, $L(p_{v_i}) \geqslant 0$ (since $M(y) \succeq 0$). Moreover, $L(p_{v_i}) \neq 0$ since, otherwise, the rank of $M(y)$ would be smaller than r.

Remains to show that $v_1, \cdots, v_r$ belong to the set K, i.e., that $g_j(v_i) \geqslant 0$ for all $j \in [m]$, $i \in [r]$. For this, we use the fact that $L(g_j p_{v_i}^2) \geqslant 0$, since $M(g_j * y) \succeq 0$. Indeed, using (14.13), we get:

$$L(g_j p_{v_i}^2) = g_j(v_i) L(p_{v_i}).$$

By assumption, $L(g_j p_{v_i}^2) \geqslant 0$ and we just showed that $L(p_{v_i}) > 0$. This implies that $g_j(v_i) \geqslant 0$, as desired, and the proof is complete. □

14.2.3 Moment relaxation for polynomial optimization

Let us return to the polynomial optimization problem (14.1). In Chapter 13, we defined the lower bound $p_{\text{sos}} \leqslant p_{\text{min}}$, obtained by considering sums of squares decompositions in the quadratic module $M(\mathbf{g}) = \Sigma + g_1\Sigma + \cdots + g_m\Sigma$:

$$p_{\text{sos}} = \sup\{\lambda : p - \lambda \in M(\mathbf{g}) = \Sigma + g_1\Sigma + \cdots g_m\Sigma\}. \tag{14.14}$$

Based on the discussion in the preceding section, we can also define the following lower bound for p_{min}:

$$p_{\text{mom}} = \inf\{\boldsymbol{p}^{\mathsf{T}} y : y_0 = 1,\ M(y) \succeq 0,\ M(g_j * y) \succeq 0\ (j \in [m])\} \tag{14.15}$$

These two bounds are 'dual' to each other, since the positivity conditions in (14.15) mean that the corresponding linear functional L is nonnegative on $M(\mathbf{g})$. We have the following inequalities:

Lemma 14.2.5. *We have:* $p_{\text{sos}} \leqslant p_{\text{mom}} \leqslant p_{\text{min}}$.

Proof. The inequality $p_{\text{sos}} \leqslant p_{\text{mom}}$ is 'weak duality': Let λ be feasible for (14.14) and let y be feasible for (14.15) with associated linear functional L. Then, $p - \lambda \in M(\mathbf{g})$, $L(1) = 1$ and $L \geqslant 0$ on $M(\mathbf{g})$. Therefore, $L(p-\lambda) = L(p) - \lambda \geqslant 0$ implies $\boldsymbol{p}^{\mathsf{T}} y = L(p) \geqslant \lambda$ and thus $p_{\text{mom}} \geqslant p_{\text{sos}}$.

The inequality $p_{\text{mom}} \leqslant p_{\text{min}}$ follows from the fact that, for each $v \in K$, $y = [v]_\infty$ is feasible for (14.15) with value $p(v)$. □

We saw in the preceding chapter that $p_{\text{sos}} = p_{\text{min}} = p_{\text{mom}}$ if K is compact and if moreover the quadratic module $M(\mathbf{g})$ is Archimedean.

On the other hand, it follows from Theorem 14.2.4 that $p_{\text{mom}} = p_{\text{min}}$ if the program (14.15) has an optimal solution y for which $M(y)$ has finite rank.

In the next chapter we will consider hierarchies of semidefinite programming relaxations for problem (14.1) obtained by adding degree constraints to the programs (14.14) and (14.15), and we will use the results of Theorems 14.1.11 and 14.2.4 for giving a procedure to find global optimizers of problem (14.1).

14.3 Notes and further reading

The terminology of 'moment matrix' which we have used for the matrix $M(y)$ is motivated by the relevance of these matrices to the classical moment problem. Recall that, given a (positive Borel) measure μ on a subset $K \subseteq \mathbb{R}^n$, the quantity $y_\alpha = \int_K x^\alpha d\mu(x)$ is called its *moment of order* α. The *K-moment problem* asks to characterize the sequences $y \in \mathbb{R}^{\mathbb{N}^n}$ which are the sequence of moments of some measure μ supported by K.

In the special case when μ is a finite atomic measure, i.e., when μ is supported by finitely many points of K, then its sequence of moments is of the form $y = \sum_{i=1}^r \lambda_i [v_i]_\infty$ for some positive scalars λ_i and some $v_i \in K$. In other words, the set $\mathcal{C}_\infty(K)$ corresponds to the set of sequences of moments of finite atomic measures on K. Moreover, the closure of the set $\mathcal{C}_\infty(K)$ is the set of sequences of moments of an arbitrary measure on K. Hence, Theorem 14.2.4 characterizes which sequences admit a finite atomic measure on K, when K is a basic closed semi-algebraic set, in terms of positivity and finite rank conditions on the sequence y. This result is due to Curto and Fialkow [1]. (When the condition rank $M(y) < \infty$ holds, Curto and Fialkow speak of *flat data*). The proof of [1] uses tools from functional analysis, the simpler algebraic proof given here is based on [4] (see also [4]).

We refer to the books of Cox, Little and O'Shea [1, 2] for further reading about ideals and varieties (and, in particular, about multiplication operators in the quotient space $\mathbb{R}[x]/I$).

14.4 Exercises

14.1 Recall the definitions (14.5) and (14.6) for $\sqrt{I}$ and $\sqrt[\mathbb{R}]{I}$.

(a) Show that the radical $\sqrt{I}$ of an ideal $I \subseteq \mathbb{C}[x]$ is an ideal.

(b) Show that the real radical $\sqrt[\mathbb{R}]{I}$ of an ideal $I \subseteq \mathbb{R}[x]$ is an ideal.

14.2 Show Lemma 14.1.3.

14.3 (a) Let I and J be two ideals in $\mathbb{C}[x]$. Show that $I \cap J$ is an ideal and that $V_\mathbb{C}(I \cap J) = V_\mathbb{C}(I) \cup V_\mathbb{C}(J)$.

(b) Given $v \in \mathbb{C}^n$, show that the set $\{v\}$ is a complex variety.

(b) Show that any finite set $V \subseteq \mathbb{C}^n$ is a complex variety.

14.4** The goal is to show Theorem 14.1.13 in the radical case.

Let I be a radical ideal in $\mathbb{R}[x]$ with $N = |V_\mathbb{C}(I)| = \dim \mathbb{R}[x]/I < \infty$. Let $\mathcal{B} = \{[b_1], \cdots, [b_N]\}$ be a base of $\mathcal{A} = \mathbb{R}[x]/I$ and, for any $h \in \mathbb{R}[x]$, let M_h denote the matrix of the multiplication by h in the base $\mathcal{B}$. Then, the matrix of the Hermite quadratic form (14.10) in the base $\mathcal{B}$ is the real symmetric matrix $H = (H_{ij})_{i,=1}^N$ with entries $H_{ij} = \mathrm{Tr}(M_{b_i b_j})$. Finally,

$\sigma_+(H)$, $\sigma_-(H)$ denote, respectively, the numbers of positive and negative eigenvalues of H.

(a) Show that $H = \sum_{v \in V_{\mathbb{C}}(I)} [v]_{\mathcal{B}} [v]_{\mathcal{B}}^{\mathsf{T}}$ and rank$(H) = |V_{\mathbb{C}}(I)|$.

(b) Show that $V_{\mathbb{C}}(I)$ can be partitioned into $V_{\mathbb{R}}(I) \cup T \cup \overline{T}$, where $\overline{T}$ is the set of complex conjugates of the elements of T.

(c) Show that $H = P - Q$ for some matrices P, Q such that $P, Q \succeq 0$, rank$(P) = |V_{\mathbb{R}}(I)| + |T|$ and rank$(Q) = |T|$.

(d) Show that $H = A - B$ for some matrices A, B such that $A, B \succeq 0$, $AB = BA = 0$, rank$(A) = \sigma_+(H)$ and rank$(B) = \sigma_-(H)$.

(e) Show that $\sigma_+(H) = |V_{\mathbb{R}}(I)| + |T|$ and $\sigma_-(H) = |T|$.

BIBLIOGRAPHY

[1] D.A. Cox, J.B. Little and D. O'Shea. *Ideals, Varieties and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*, Springer, 1997.

[2] D.A. Cox, J.B. Little and D. O'Shea. *Using Algebraic Geometry*, Springer, 1998.

[3] R. Curto and L. Fialkow. Solution of the truncated complex moment problem for flat data. *Memoirs of the AMS* **119**(568), 1996.

[4] M. Laurent. Revisiting two theorems of Curto and Fialkow on moment matrices. *Proceedings of the AMS* **133(10)**:2965–2976, 2005.

[5] M. Laurent. Sums of squares, moment matrices and optimization over polynomials. `http://homepages.cwi.nl/~monique/files/moment-ima-update-new.pdf`

CHAPTER 15

POLYNOMIAL OPTIMIZATION AND REAL ROOTS

We return to the polynomial optimization problem:

$$p_{\min} = \inf_{x \in K} p(x), \tag{15.1}$$

where K is defined by polynomial inequalities:

$$K = \{x \in \mathbb{R}^n : g_1(x) \geqslant 0, \cdots, g_m(x) \geqslant 0\} \tag{15.2}$$

with $p, g_1, \cdots, g_m \in \mathbb{R}[x]$. Throughout we set $g_0 = 1$. In the previous chapters we have introduced the two parameters:

$$p_{\text{sos}} = \sup \left\{ \lambda : p - \lambda \in M(\mathbf{g}) = \sum_{j=0}^{m} g_j \Sigma \right\},$$

$$p_{\text{mom}} = \inf\{L(p) : L \text{ linear function on } \mathbb{R}[x],\ L(1) = 1,\ L \geqslant 0 \text{ on } M(\mathbf{g})\},$$

which satisfy the inequalities:

$$p_{\text{sos}} \leqslant p_{\text{mom}} \leqslant p_{\min}.$$

Both parameters can be reformulated using positive semidefinite matrices. However these matrices are infinite (indexed by $\mathbb{N}^n$), since there is a priori no degree bound on the polynomials s_j entering a decomposition: $p - \lambda = \sum_j s_j g_j$ in $M(\mathbf{g})$, and since L is a linear function on $\mathbb{R}[x]$ which is infinite dimensional. Hence, it is not clear how to compute the parameters p_{mom} and p_{sos}. In this chapter, we consider hierarchies of approximations for problem (15.1) obtained by adding degree bounds to the programs defining p_{sos} and p_{mom}.

Given an integer t, recall that $\mathbb{R}[x]_t$ denotes the set of polynomials of degree at most t. We set $\Sigma_{2t} = \Sigma \cap \mathbb{R}[x]_{2t}$ and we define the *truncated (at degree $2t$) quadratic module*:

$$M(\mathbf{g})_{2t} = \left\{ \sum_{j=0}^{m} g_j s_j : s_j \in \Sigma,\ \deg(s_j g_j) \leqslant 2t \ \ (j = 0, 1, \cdots, m) \right\},$$

which consists of the elements $\sum_j s_j g_j$ of the quadratic module $M(\mathbf{g})$ where all summands have degree at most $2t$. Then, we define the bounds:

$$p_{\text{sos},t} = \sup\{\lambda : p - \lambda \in M(\mathbf{g})_{2t}\}, \tag{15.3}$$

$$p_{\text{mom},t} = \inf\{L(p) : L \text{ linear function on } \mathbb{R}[x]_{2t},\ L(1) = 1,\ L \geqslant 0 \text{ on } M(\mathbf{g})_{2t}\}. \tag{15.4}$$

Lemma 15.0.1. *For any integer t, $p_{\text{sos},t} \leqslant p_{\text{mom},t} \leqslant p_{\min}$.*

Proof. Let L be feasible for (15.4) and let λ be feasible for (15.3). Then, we have: $0 \leqslant L(p - \lambda) = L(p) - \lambda$. This implies that $p_{\text{sos},t} \leqslant p_{\text{mom},t}$.

Given $v \in K$, let L be the evaluation at v; that is, L is the linear function on $\mathbb{R}[x]_{2t}$ defined by $L(f) = f(v)$ for $f \in \mathbb{R}[x]_{2t}$. Then, L is feasible for the program (15.4) with objective value $L(p) = p(v)$. This implies: $p_{\text{mom},t} \leqslant p(v)$. As this holds for all $v \in K$, we deduce that $p_{\text{mom},t} \leqslant p_{\min}$. □

In this chapter we investigate some properties of these hierarchies of bounds:

1. **Duality:** The bounds $p_{\text{sos},t}$ and $p_{\text{mom},t}$ are defined by dual semidefinite programs.
2. **Asymptotic convergence:** Both bounds converge to $p_{\min}$, when $M(\mathbf{g})$ is Archimedean.
3. **Optimality certificate and global minimizers:** When (15.4) has an optimal solution satisfying a special rank condition, the bound $p_{\text{mom},t}$ is exact and one can compute global minimizers of the problem (15.1).
4. Application to computing **real roots** of polynomial equations.

15.1 Duality

We now indicate how to reformulate the programs (15.3) and (15.4) as semidefinite programs and to check that they are in fact *dual* semidefinite programs.

The following is the truncated analogue of what we did in Section 14.2 (for linear functions L on $\mathbb{R}[x]$ and sequences $y \in \mathbb{R}^{\mathbb{N}^n}$). Any linear function L on $\mathbb{R}[x]_{2t}$ is completely specified by the sequence of real numbers $y = (y_\alpha)_{\alpha \in \mathbb{N}^n_{2t}}$, where $y_\alpha = L(x^\alpha)$. Then we define the corresponding *truncated (at order t) moment matrix*:

$$M_t(y) = (y_{\alpha+\beta})_{\alpha,\beta \in \mathbb{N}^n_t},$$

indexed by $\mathbb{N}^n_t$. One can easily check that:

$$L \geqslant 0 \text{ on } \Sigma \cap \mathbb{R}[x]_{2t} \Longleftrightarrow M_t(y) \succeq 0.$$

Analogously,

$$L \geqslant 0 \text{ on } \{sg : s \in \Sigma,\ \deg(sg) \leqslant 2t\} \Longleftrightarrow M_{t-d_g}(g * y) \succeq 0,$$

after setting $d_g := \lceil \deg(g)/2 \rceil$ and where $g * y$ is the sequence indexed by $\mathbb{N}^n_{t-d_g}$ with $(g*y)_\alpha = L(x^\alpha g) = \sum_\gamma g_\gamma y_{\alpha+\gamma}$ (which is well defined if $|\alpha| \leqslant 2(t-d_g)$ as then $|\alpha+\gamma| \leqslant 2(t-d_g) + \deg(g) \leqslant 2t$). Therefore, the program (15.4) can be equivalently reformulated as:

$$p_{\mathrm{mom},t} = \inf_{y \in \mathbb{N}^n_{2t}} \{\boldsymbol{p}^{\mathsf{T}} y : y_0 = 1,\ M_t(y) \succeq 0,\ M_{t-d_{g_j}}(g_j * y) \succeq 0\ (j = 1, \cdots, m)\}. \tag{15.5}$$

We now explicit the fact that the dual semidefinite program of (15.5) coincides with (15.3); we do this only in the unconstrained case: $K = \mathbb{R}^n$ (i.e., with no constraints $g_j \geqslant 0$) in order to avoid tedious notational details. For $\gamma \in \mathbb{N}^n_{2t}$ let $A_{t,\gamma}$ denote the 0/1 matrix indexed by $\mathbb{N}^n_t$ with (α,β)-th entry $A_{t,\gamma}(\alpha,\beta) = 1$ when $\alpha + \beta = \gamma$ and 0 otherwise. Note that

$$M_t(y) = \sum_{\gamma \in \mathbb{N}^n_{2t}} y_\gamma A_{t,\gamma} \text{ and } \sum_{\gamma \in \mathbb{N}^n_{2t}} x^\gamma A_{t,\gamma} = [x]_t [x]_t^{\mathsf{T}} \tag{15.6}$$

after setting $[x]_t = (x^\alpha)_{\alpha \in \mathbb{N}^n_t}$.

Lemma 15.1.1. *The programs:*

$$\sup\{\lambda : p - \lambda \in \Sigma \cap \mathbb{R}[x]_{2t}\}, \tag{15.7}$$

and

$$\inf_{y \in \mathbb{R}^{\mathbb{N}^n_{2t}}} \{\boldsymbol{p}^{\mathsf{T}} y : y_0 = 1,\ M_t(y) \succeq 0\} \tag{15.8}$$

are dual semidefinite programs.

Proof. Using (15.6), we can express (15.8) as the following semidefinite program (in standard dual form):

$$p_0 + \inf \left\{ \sum_{\gamma \in \mathbb{N}^n_{2t} \setminus \{0\}} p_\gamma y_\gamma : A_{t,0} + \sum_{\gamma \in \mathbb{N}^n_{2t} \setminus \{0\}} y_\gamma A_{t,\gamma} \succeq 0 \right\}. \tag{15.9}$$

Next we express (15.7) as a semidefinite program (in standard primal form). For this, we use the fact that $p - \lambda \in \Sigma \cap \mathbb{R}[x]_{2t}$ if and only if there exists a positive semidefinite matrix Q indexed by $\mathbb{N}^n_t$ such that $p - \lambda = [x]_t^{\mathsf{T}} Q [x]_t$. Rewrite: $[x]_t^{\mathsf{T}} Q [x]_t = \langle Q, [x]_t [x]_t^{\mathsf{T}} \rangle = \sum_{\gamma \in \mathbb{N}^n_{2t}} \langle A_{t,\gamma}, Q \rangle x^\gamma$ (using (15.6)). Therefore, (15.7) is equivalent to

$$p_0 + \sup\{-\langle A_{t,0}, Q\rangle : \langle A_{t,\gamma}, Q\rangle = p_\gamma\ (\gamma \in \mathbb{N}^n_{2t} \setminus \{0\}),\ Q \succeq 0\}. \tag{15.10}$$

It is now clear that the programs (15.9) and (15.10) are dual semidefinite programs. □

15.2 Convergence

Theorem 15.2.1. *Assume that $M(\mathbf{g})$ is Archimedean (i.e., there exists a polynomial $f \in M(\mathbf{g})$ for which the set $\{x \in \mathbb{R}^n : f(x) \geqslant 0\}$ is compact). Then, the bounds $p_{\text{mom},t}$ and $p_{\text{sos},t}$ converge to $p_{\min}$ as $t \to \infty$.*

Proof. Pick $\epsilon > 0$. Then the polynomial $p - p_{\min} + \epsilon$ is strictly positive on K. As $M(\mathbf{g})$ is Archimedean, we can apply Putinar's theorem (Theorem 13.2.9) and deduce that $p - p_{\min} + \epsilon \in M(\mathbf{g})$. Hence, there exists $t \in \mathbb{N}$ such that $p - p_{\min} + \epsilon \in M(\mathbf{g})_{2t}$ and thus $p_{\min} - \epsilon \leqslant p_{\text{sos},t}$. Therefore, $\lim_{t\to\infty} p_{\text{sos},t} = p_{\min}$. Since, by Lemma 15.0.1, $p_{\text{sos},t} \leqslant p_{\text{mom},t} \leqslant p_{\min}$ for all t, we deduce: $\lim_{t\to\infty} p_{\text{mom},t} = p_{\min}$. □

15.3 Flat extensions of moment matrices

We state here a technical result about moment matrices which will be useful for establishing an optimality certificate for the moment bounds $p_{\text{mom},t}$. Roughly speaking, this result permits to extend a truncated sequence $y \in \mathbb{R}^{\mathbb{N}^n_{2s}}$ satisfying a rank condition (see (15.12) below) to an infinite sequence $\tilde{y} \in \mathbb{R}^{\mathbb{N}^n}$ whose moment matrix $M(\tilde{y})$ has the same rank as $M(y)$, to which we can then apply the result from Theorem 14.2.4.

We recall that we can view the kernel of a moment matrix as a set of polynomials, after identifying a polynomial f with its vector of coefficients $\boldsymbol{f}$. If y is a sequence in $\mathbb{R}^{\mathbb{N}^n_{2s}}$ and L is the associated linear function on $\mathbb{R}[x]_{2s}$, then

$$\boldsymbol{f} \in \ker M_s(y) \iff L(fg) = 0 \ \forall g \in \mathbb{R}[x]_s; \tag{15.11}$$

from now on we abuse notation and also write '$f \in \ker M_s(y)$'. We also recall that the kernel of an infinite moment matrix $M(\tilde{y})$ corresponds to an ideal I in $\mathbb{R}[x]$ (Lemma 14.2.3). The following simple result about kernels of matrices is useful (check it).

Lemma 15.3.1. *Let X be a symmetric matrix with block form*

$$X = \begin{pmatrix} A & B \\ B^{\mathsf{T}} & C \end{pmatrix}.$$

Assume that we are in one of the following two situations: (i) $\text{rank} X = \text{rank} A$ *(then one says that X is a* flat extension *of A), or (ii) $X \succeq 0$. Then the following holds:*

$$x \in \ker A \iff x \in \ker B^{\mathsf{T}} \iff (x^{\mathsf{T}}, 0)^{\mathsf{T}} \in \ker X.$$

As an application we obtain the following result showing that the kernel of a truncated moment matrix behaves like a 'truncated ideal'.

Lemma 15.3.2. *Given a sequence $y \in \mathbb{R}^{\mathbb{N}^n_{2s}}$ consider its moment matrices $M_s(y)$ and $M_{s-1}(y)$. Clearly $M_{s-1}(y)$ is a principal submatrix of $M_s(y)$. Assume that we*

are in one of the following two situations: (i) $\text{rank}M_s(y) = \text{rank}M_{s-1}(y)$*, or (ii)* $M_s(y) \succeq 0$*. Given polynomials* $f, g \in \mathbb{R}[x]$*, the following holds:*

$$f \in \ker M_s(y),\ \deg(fg) \leqslant s-1 \Longrightarrow fg \in \ker M_s(y).$$

Proof. Let L be the linear function on $\mathbb{R}[x]_{2s}$ associated to y. A first observation is that it suffices to show the result when g has degree 1, say $g = x_i$ (then the general result follows by iterating this special case). A second observation is that it suffices to show that fg belongs to the kernel of $M_{s-1}(y)$ (then fg also belongs to the kernel of $M_s(y)$, in view of Lemma 15.3.1). So, pick a polynomial u of degree at most $s-1$ and let us show that $L((fx_i)u) = 0$. But this follows from the fact that $f \in \ker M_s(y)$ since $\deg(x_i u) \leqslant s$ (recall (15.11)). □

Theorem 15.3.3. *Given a sequence* $y \in \mathbb{R}^{\mathbb{N}^n_{2s}}$*, consider its moment matrices* $M_s(y)$ *and* $M_{s-1}(y)$*. Assume that*

$$\text{rank } M_s(y) = \text{rank } M_{s-1}(y). \tag{15.12}$$

Then, one can extend y *to a sequence* $\tilde{y} \in \mathbb{R}^{\mathbb{N}^n}$ *satisfying:*

$$\text{rank } M(\tilde{y}) = \text{rank } M_s(y). \tag{15.13}$$

Let I *be the ideal in* $\mathbb{R}[x]$ *corresponding to the kernel of* $M(\tilde{y})$*. The following properties hold:*

(i) *If* $\{\alpha_1, \cdots, \alpha_r\} \subseteq \mathbb{N}^n_{s-1}$ *indexes a maximum linearly independent set of columns of* $M_{s-1}(y)$*, then the set* $\{[x^{\alpha_1}], \cdots, [x^{\alpha_r}]\} \subseteq \mathbb{R}[x]/I$ *is a base of* $\mathbb{R}[x]/I$*.*

(ii) *The ideal* I *is generated by the polynomials in* $\ker M_s(y)$*:* $I = (\ker M_s(y))$*.*

Proof. The first part of the proof consists of constructing the sequence $\tilde{y}$ satisfying (15.13). It is based on Lemma 15.3.2; the details are elementary but technical, so we omit them. (You will show the case $n = 1$ in Exercise 15.1).

(i) If the set $\{\alpha_1, \cdots, \alpha_r\}$ indexes a maximum set of linearly independent columns of $M_{s-1}(y)$ then, as $\text{rank}M(\tilde{y}) = \text{rank}M_{s-1}(y)$, it also indexes a maximum set of linearly independent columns of $M(\tilde{y})$. This implies that the set $\{[x^{\alpha_1}], \cdots, [x^{\alpha_r}]\}$ is a base of $\mathbb{R}[x]/I$.

(ii) As $\text{rank}M(\tilde{y}) = \text{rank}M_s(y)$, we have the inclusion: $\ker M_s(y) \subseteq \ker M(\tilde{y})$ (recall Lemma 15.3.1). Thus the ideal generated by $\ker M_s(y)$ is contained in the ideal $\ker M(\tilde{y})$:

$$(\ker M_s(y)) \subseteq \ker M(\tilde{y}).$$

Set $\mathcal{M} = \{x^{\alpha_1}, \cdots, x^{\alpha_r}\}$ where the α_i's are as in (i), and let $\langle \mathcal{M} \rangle$ denote the linear span of $\mathcal{M}$ (whose elements are the polynomials $\sum_i \lambda_i x^{\alpha_i}$ where $\lambda_i \in \mathbb{R}$). Then, $\langle \mathcal{M} \rangle \cap \ker M(\tilde{y}) = \{0\}$ (by (i)). We claim that

$$\mathbb{R}[x] = \langle \mathcal{M} \rangle + (\ker M_s(y)).$$

For this, one can show using induction on its degree that each monomial x^α can be written as $x^\alpha = p + q$ where p lies in the span of $\mathcal{M}$ and q lies in the ideal generated by $\ker M_s(y)$ (check it). Now, let $f \in \ker M(\tilde{y})$. Applying the above to f, we can write $f = p + q$ where $p \in \langle \mathcal{M} \rangle$ and $q \in (\ker M_s(y))$. This implies that $p = f - q \in \langle \mathcal{M} \rangle \cap \ker M(\tilde{y}) = \{0\}$ and thus $f = p \in (\ker M_s(y))$. □

15.4 Optimality certificate and global minimizers

Let $K_p^* = \{x \in K : p(x) = p_{\min}\}$ denote the set (possibly empty) of global minimizers of the polynomial p over K. We also set

$$d_K = \max\{d_{g_1}, \cdots, d_{g_m}\}, \quad \text{where } d_f = \lceil \deg(f)/2 \rceil \text{ for } f \in \mathbb{R}[x]. \tag{15.14}$$

Theorem 15.4.1. *Let L be an optimal solution to the program (15.4) and let $y = (L(x^\alpha)) \in \mathbb{R}^{\mathbb{N}^n_{2t}}$ be the corresponding sequence. Asssume that y satisfies the rank condition:*

$$\text{rank } M_s(y) = \text{rank } M_{s-d_K}(y) \tag{15.15}$$

for some integer s satisfying $\max\{d_p, d_K\} \leqslant s \leqslant t$. Then the following properties hold:

(i) *The relaxation (15.4) is exact: $p_{\text{mom},t} = p_{\min}$.*

(ii) *The common roots to the polynomials in $\ker M_s(y)$ are all real and they are global minimizers: $V_{\mathbb{C}}(\ker M_s(y)) \subseteq K_p^*$.*

(iii) *If L is an optimal solution of (15.4) for which the matrix $M_t(y)$ has maximum possible rank, then $V_{\mathbb{C}}(\ker M_s(y)) = K_p^*$.*

Proof. As y satisfies the rank condition (15.15), we can apply Theorem 15.3.3: There exists a sequence $\tilde{y} \in \mathbb{R}^{\mathbb{N}^n}$ extending the subsequence $(y_\alpha)_{|\alpha| \leqslant 2s}$ of y and satisfying rank $M(\tilde{y}) = \text{rank } M_s(y) =: r$. Thus, $\tilde{y}_\alpha = y_\alpha$ if $|\alpha| \leqslant 2s$, but it could be that $\tilde{y}$ and y differ at entries indexed by monomials of degree higher than $2s$, these entries of y will be irrelevant in the rest of the proof. Let I be the ideal corresponding to the kernel of $M(\tilde{y})$. By Theorem 15.3.3, I is generated by $\ker M_s(y)$ and thus $V_{\mathbb{C}}(I) = V_{\mathbb{C}}(\ker M_s(y))$. As $M(\tilde{y})$ is positive semidefinite with finite rank r, we can apply Theorem 14.2.4 (and its proof): We deduce that

$$V_{\mathbb{C}}(I) = \{v_1, \cdots, v_r\} \subseteq \mathbb{R}^n$$

and

$$\tilde{y} = \sum_{i=1}^r \lambda_i [v_i]_\infty \text{ where } \lambda_i > 0 \text{ and } \sum_{i=1}^r \lambda_i = 1.$$

Taking the projection onto the subspace $\mathbb{R}^{\mathbb{N}^n_{2s}}$, we obtain:

$$(y_\alpha)_{\alpha \in \mathbb{N}^n_{2s}} = \sum_{i=1}^r \lambda_i [v_i]_{2s} \text{ where } \lambda_i > 0 \text{ and } \sum_{i=1}^r \lambda_i = 1. \tag{15.16}$$

In other words, the restriction of the linear map L to the subspace $\mathbb{R}[x]_{2s}$ is the convex combination $\sum_{i=1}^r \lambda_i L_{v_i}$ of evaluations at the points of $V_{\mathbb{C}}(I)$. Moreover, let $\{\alpha_1, \cdots, \alpha_r\} \subseteq \mathbb{N}^n_{s-d_K}$ index a maximum linearly independent set of columns of $M_{s-d_K}(y)$, so that the set $\mathcal{B} = \{[x^{\alpha_1}], \cdots, [x^{\alpha_r}]\}$ is a base of $\mathbb{R}[x]/I$ (by Theorem 15.3.3).

First we claim that we can choose interpolation polynomials p_{v_i} at the points of $V_{\mathbb{C}}(I)$ with $\deg(p_{v_i}) \leqslant s - d_K$. Indeed, if p_{v_i} are arbitrary interpolation polynomials then, using the base $\mathcal{B}$, write $p_{v_i} = f_i + g_i$ where $g_i \in I$ and f_i lies in the linear span of the monomials $x^{\alpha_1}, \cdots, x^{\alpha_r}$. Thus the f_i's are again interpolation polynomials but now with degree at most $s - d_K$.

Next we claim that $v_1, \cdots, v_r$ belong to the set K. To see this, we use the fact that $L \geqslant 0$ on $(g_j\Sigma) \cap \mathbb{R}[x]_{2t}$ for all $j \in [m]$. As $\deg(p_{v_i}) \leqslant s - d_K$, we have: $\deg(g_j p_{v_i}^2) \leqslant \deg(g_j) + 2(s - d_K) \leqslant 2s$, and thus we can compute $L(g_j p_{v_i}^2)$ using (15.16) and obtain that $L(g_j p_{v_i}^2) = g_j(v_i)\lambda_i \geqslant 0$. This gives $g_j(v_i) \geqslant 0$ for all j and thus $v_i \in K$.

As $\deg(p) \leqslant 2s$, we can also evaluate $L(p)$ using (15.16): we obtain that $L(p) = \sum_{i=1}^r \lambda_i p(v_i) \geqslant p_{\min}$, since $p(v_i) \geqslant p_{\min}$ as $v_i \in K$. This gives the inequality: $p_{\text{mom},t} \geqslant p_{\min}$. The reverse inequality holds always (Lemma 15.0.1). Thus (i) holds: $p_{\text{mom},t} = p_{\min}$. In turn, this implies that $p(v_i) = p_{\min}$ for all i, which shows (ii): $\{v_1, \cdots, v_r\} \subseteq K_p^*$.

Assume now that rank$M_t(y)$ is maximum among all optimal solutions of (15.4). In other words, y lies in the relative interior of the face of the feasible region of (15.4) consisting of all optimal solutions. Therefore, for any other optimal solution y', we have that $\ker M_t(y) \subseteq \ker M_t(y')$. Consider a global minimizer $v \in K_p^*$ of p over K and the corresponding optimal solution $y' = [v]_{2t}$ of (15.4). The inclusion $\ker M_t(y) \subseteq \ker M_t(y')$ implies that any polynomial in $\ker M_t(y)$ vanishes at v. Therefore, $\ker M_s(y) \subseteq \mathcal{I}(K_p^*)$ and thus $I = (\ker M_s(y)) \subseteq \mathcal{I}(K_p^*)$. In turn, this implies the inclusions:

$$K_p^* \subseteq V_{\mathbb{C}}(\mathcal{I}(K_p^*)) \subseteq V_{\mathbb{C}}(I) = \{v_1, \cdots, v_r\}.$$

Thus (iii) holds and the proof is complete. □

Under the assumptions of Theorem 15.4.1, we can apply the eigenvalue method described in Section 14.1.3 for computing the points in the variety $V_{\mathbb{C}}(\ker M_s(y))$. Indeed, all the information that we need is contained in the matrix $M_s(y)$. Recall that what we need in order to recover $V_{\mathbb{C}}(I)$ is an explicit base $\mathcal{B}$ of the quotient space $\mathbb{R}[x]/I$ and the matrix in the base $\mathcal{B}$ of some multiplication operator in $\mathbb{R}[x]/I$, where $I = (\ker M_s(y))$.

First of all, if we choose $\{\alpha_1, \cdots, \alpha_r\} \subseteq \mathbb{N}^n_{s-d_K}$ indexing a maximum linearly independent set of columns of $M_{s-1}(y)$, then the set $\mathcal{B} = \{[x^{\alpha_1}], \cdots, [x^{\alpha_r}]\}$ of corresponding cosets in $\mathbb{R}[x]/I$ is a base of $\mathbb{R}[x]/I$. For any variable x_k, we now observe that it is easy to build the matrix M_{x_k} of the 'multiplication by x_k' in the base $\mathcal{B}$, using the moment matrix $M_s(y)$. Indeed, for any $j \in [r]$, as $\deg(x_k x^{\alpha_j}) \leqslant s$, we can compute the linear dependency among the columns of $M_s(y)$ indexed by the monomials $x_k x^{\alpha_j}, x^{\alpha_1}, \cdots, x^{\alpha_r}$. In this way, we obtain a polynomial in the kernel of $M_s(y)$ (thus in I) which directly gives the j-th column of the matrix M_{x_k}.

Finally, we point out that it is a property of most interior-point algorithms that they return an optimal solution in the relative interior of the optimal face, thus a point satisfying the assumption of (iii). In conclusion, if we have an optimal solution of the moment relaxation (15.4) satisfying the rank condition

(15.15), then we can (numerically) compute all the global optimizers of problem (15.1).

15.5 Real solutions of polynomial equations

Consider now the problem of computing all real roots to a system of polynomial equations:

$$h_1(x) = 0, \cdots, h_m(x) = 0$$

where $h_1, \cdots, h_m \in \mathbb{R}[x]$. In other words, with I denoting the ideal generated by the h_j's, this is the problem of computing the real variety $V_{\mathbb{R}}(I)$ of I. We address this question in the case when $V_{\mathbb{R}}(I)$ is finite.

Of course, if the complex variety $V_{\mathbb{C}}(I)$ of I is finite, then we can just apply the eigenvalue method presented in Chapter 14 to compute $V_{\mathbb{C}}(I)$ (then select the real elements). However, it can be that $V_{\mathbb{R}}(I)$ is finite while $V_{\mathbb{C}}(I)$ is infinite. As a trivial such example, consider the ideal generated by the polynomial $x_1^2+x_2^2$ in two variables, to which we come back in Example 15.5.2 below. In that case we cannot apply directly the eigenvalue method. However we can apply it indirectly: Indeed, we can view the problem of computing $V_{\mathbb{R}}(I)$ as an instance of polynomial optimization problem to which we can then apply the results of the preceding section. Namely, consider the problem of minimizing the constant polynomial $p = 0$ over the set

$$K = \{x \in \mathbb{R}^n : h_j(x) \geqslant 0, -h_j(x) \geqslant 0 \ \forall j \in [m]\}.$$

Then, $K = V_{\mathbb{R}}(I)$ coincides with the set of global minimizers of $p = 0$ over K.

As before, we consider the moment relaxations (15.4). Now, any feasible solution L is an optimal solution of (15.4). Hence, by Theorem 15.4.1, if the rank condition (15.15) holds, then we can compute all points in $V_{\mathbb{R}}(I)$. We now show that it is indeed the case that, for t large enough, the rank condition (15.15) will be satisfied.

Theorem 15.5.1. *Let $h_1, \cdots, h_m \in \mathbb{R}[x]$ be polynomials having finitely many real roots. Set $d_K = \max_j \lceil \deg(h_j)/2 \rceil$. For $t \in \mathbb{N}$, let $\mathcal{F}_t$ denote the set of sequences $y \in \mathbb{R}^{\mathbb{N}^n_{2t}}$ whose associated linear function L on $\mathbb{R}[x]_{2t}$ satisfies the conditions:*

$$L(1) = 1, \ L \geqslant 0 \ \text{ on } \Sigma_{2t}, \ L(uh_j) = 0 \ \forall j \in [m] \ \forall u \in \mathbb{R}[x] \text{ with } \deg(uh_j) \leqslant 2t. \tag{15.17}$$

Then, there exist integers t_0 and s such that $d_K \leqslant s \leqslant t_0$ and the following rank condition holds:

$$\text{rank} M_s(y) = \text{rank} M_{s-d_K}(y) \ \forall y \in \mathcal{F}_t \ \forall t \geqslant t_0. \tag{15.18}$$

Moreover, $\sqrt[\mathbb{R}]{I} = (\ker M_s(y))$ if $y \in \mathcal{F}_t$ has maximum possible rank.

Proof. The goal is to show that if we choose t large enough, the the kernel of $M_t(y)$ contains sufficiently many polynomials permitting to show the rank

condition (15.18). Here y is an arbitrary feasible solution in $\mathcal{F}_t$ and L is its corresponding linear function on $\mathbb{R}[x]_{2t}$. We assume that $t \geqslant \max_j \deg(h_j)$. Then,

$$h_j \in \ker M_t(y) \ \forall j \in [m] \tag{15.19}$$

(since then $L(h_j^2) = 0$).

Now we choose a 'nice' set of polynomials $\{f_1, \cdots, f_L\}$ generating $\sqrt[\mathbb{R}]{I}$, the real radical ideal of the ideal I; namely, one for which we can claim the following degree bounds:

$$\forall f \in \sqrt[\mathbb{R}]{I} \ \ f = \sum_{l=1}^{L} u_l f_l \quad \text{for some } u_l \in \mathbb{R}[x] \ \text{ with } \deg(u_l f_l) \leqslant \deg(f). \tag{15.20}$$

(That such a nice set of generators exists follows from the theory of Gröbner bases.) Next we claim:

$$\exists t_1 \in \mathbb{N} \ \ f_1, \cdots, f_L \in \ker M_t(y) \ \text{ for any } t \geqslant t_1. \tag{15.21}$$

Fix $l \in [L]$. Applying the Real Nullstellensatz, we know that there exist polynomials p_i and u_j and an integer N (which, for convenience, we can choose to be a power of 2) satisfying the following identity:

$$f_l^N + \sum_i p_i^2 = \sum_{j=1}^{m} u_j h_j.$$

If t is large enough, then L vanishes at each $u_j h_j$ (since $h_j \in \ker M_t(y)$ and apply Lemma 15.3.2). Hence L vanishes at the polynomial $f_l^N + \sum_i p_i^2$. As L is nonnegative on Σ_{2t}, we deduce that $L(f_l^N) = 0$. Now an easy induction permits to show that $L(f_l^2) = 0$ (this is where choosing N a power of 2 was helpful) and thus $f_l \in \ker M_t(y)$.

By assumption, the set $V_{\mathbb{R}}(I)$ is finite. Therefore, the quotient space $\mathbb{R}[x]/\sqrt[\mathbb{R}]{I}$ has finite dimension (Theorem 14.1.5). Let $\mathcal{M} = \{b_1, \cdots, b_r\}$ be a set of polynomials whose cosets form a base of the quotient space $\mathbb{R}[x]/\sqrt[\mathbb{R}]{I}$. Let d_0 denote the maximum degree of the polynomials in $\mathcal{M}$ and set

$$t_2 = \max\{t_1, d_0 + d_K\}.$$

Pick any monomial x^α of degree at most t_2. We can write:

$$x^\alpha = p^{(\alpha)} + q^{(\alpha)}, \quad \text{with } \ q^{(\alpha)} = \sum_{l=1}^{L} u_l^{(\alpha)} f_l, \tag{15.22}$$

where $p^{(\alpha)}$ lies in the span of $\mathcal{M}$ and thus has degree at most d_0, and each term $u_l^{(\alpha)} f_l$ has degree at most $\max\{|\alpha|, d_0\} \leqslant t_2$. Here we have used the fact that $\{[b_1], \cdots, [b_r]\}$ is a base of $\mathbb{R}[x]/\sqrt[\mathbb{R}]{I}$, combined with the property (15.20) of the generators f_l of $\sqrt[\mathbb{R}]{I}$.

We can now conclude the proof: We show that, if $t \geqslant t_0 := t_2 + 1$, then the rank condition (15.18) holds with $s = t_2$. For this pick a monomial x^α of degree at most t_2, so that (15.22) holds. As $\deg(u_l^{(\alpha)} f_l) \leqslant t_2 \leqslant t - 1$ and $f_l \in \ker M_t(y)$ (by (15.20)), we obtain that $u_l^{(\alpha)} f_l \in \ker M_t(y)$ (use Lemma 15.3.2). Therefore, the polynomial $x^\alpha - p^{(\alpha)}$ belongs to the kernel of $M_t(y)$. As the degree of $p^{(\alpha)}$ is at most $d_0 \leqslant t_2 - d_K$, we can conclude that $\text{rank} M_{t_2 - d_K}(y) = \text{rank} M_{t_2}(y)$.

Finally, the equality $\sqrt[\mathbb{R}]{I} = (\ker M_{t_2}(y)))$ follows from Theorem 15.4.1 (iii). □

Example 15.5.2. *Let I be the ideal generated by the polynomial $x_1^2 + x_2^2$. Clearly, $V_\mathbb{R}(I) = \{(0,0)\}$ and $\sqrt[\mathbb{R}]{I} = (x_1, x_2)$ is generated by the two monomials x_1 and x_2. Let us see how we can find this again by applying the above result.*

For this, let L be a feasible solution in the set $\mathcal{F}_t$ defined by (15.17) for $t = 1$. Then, we have that $L(x_1^2), L(x_2^2) \geqslant 0$ and $L(x_1^2 + x_2^2) = 0$. This implies: $L(x_1^2) = L(x_2^2) = 0$ and thus $L(x_1) = L(x_2) = L(x_1 x_2) = 0$. Hence the moment matrix $M_1(y)$ has the form:

$$M_1(y) = \begin{array}{c} \\ 1 \\ x_1 \\ x_2 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & x_1 & x_2 \end{array} \\ \begin{pmatrix} 1 & y_{10} & y_{01} \\ y_{10} & y_{20} & y_{11} \\ y_{01} & y_{11} & y_{02} \end{pmatrix} \end{array} = \begin{pmatrix} .1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Therefore, $\text{rank} M_1(y) = \text{rank} M_0(y)$*, x_1, x_2 belong to the kernel of $M_1(y)$, and we find that* $\ker M_1(y)$ *generates $\sqrt[\mathbb{R}]{I}$.*

As an exercise, check what happens when I is the ideal generated by $(x_1^2 + x_2)^2$. When does the rank condition holds?

15.6 Notes and further reading

The flat extension theorem (Theorem 15.3.3) was proved by Curto and Fialkow [1] (this result and some extensions are exposed in the survey [4]).

The moment approach to polynomial optimization presented in this chapter was introduced by Lasserre [3]. Lasserre realized the relevance of the results of Curto and Fialkow [1] for optimization, in particular, that their flat extension theorem yields an optimality certificate and together with Henrion he adapted the eigenvalue method to compute global optimizers. Having such a stopping criterium and being able to compute global optimizers is a remarkable property of this 'moment based' approach. It has been implemented in the software GloptiPoly, the most recent version can be found at [2]. The application to computing real roots (and real radical ideals) has been developed by Lasserre, Laurent and Rostalski, see the survey [5].

15.7 Exercises

15.1. Given an integer $s \geqslant 1$, consider a sequence $y = (y_0, y_1, \cdots, y_{2s}) \in \mathbb{R}^{2s+1}$ and its moment matrix $M_s(y)$ of order s. Assume that the rank condition holds:

$$\text{rank} M_s(y) = \text{rank} M_{s-1}(y).$$

(a) Show that one can find scalars $a, b \in \mathbb{R}$ for which the extended sequence $\tilde{y} = (y_0, y_1, \cdots, y_{2s}, a, b)$ satisfies:

$$\text{rank} M_{s+1}(\tilde{y}) = \text{rank} M_s(y).$$

(b) Show that one can find an (infinite) extension

$$\tilde{y} = (y_0, y_1, \cdots, y_{2s}, \tilde{y}_{2s+1}, \tilde{y}_{2s+2}, \cdots) \in \mathbb{R}^{\mathbb{N}}$$

satisfying

$$\text{rank} M(\tilde{y}) = \text{rank} M_s(y).$$

This shows the flat extension theorem (Theorem 15.3.3) in the univariate case $n = 1$.

15.2 Consider the problem of computing $p_{\min} = \inf_{x \in K} p(x)$, where $p = x_1 x_2$ and

$$K = \{x \in \mathbb{R}^2 : -x_2^2 \geqslant 0,\ 1 + x_1 \geqslant 0,\ 1 - x_1 \geqslant 0\}.$$

(a) Show that, at order $t = 1$, $p_{\text{mom},1} = p_{\min} = 0$ and $p_{\text{sos},1} = -\infty$.

(b) At order $t = 2$, what is the value of $p_{\text{sos},2}$?

BIBLIOGRAPHY

[1] R. Curto and L. Fialkow. Solution of the truncated complex moment problem for flat data. *Memoirs of the AMS* **119**(568), 1996.

[2] D. Henrion, J. B. Lasserre, J. Loefberg. GloptiPoly 3: moments, optimization and semidefinite programming. *Optimization Methods and Software* **24(4-5)**:761-779, 2009.

`http://homepages.laas.fr/henrion/software/gloptipoly/`

[3] J.B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization* **11**:796–817, 2001.

[4] M. Laurent. Sums of squares, moment matrices and optimization over polynomials. `http://homepages.cwi.nl/~monique/files/moment-ima-update-new.pdf`

[5] M. Laurent and P. Rostalski. The approach of moments for polynomial equations. Chapter in *Handbook on Semidefinite, Cone and Polynomial Optimization*, Springer, 2012.

`http://homepages.cwi.nl/~monique/files/Handbook-SDP.pdf`