Semidefinite Optimization

These Lecture Notes are based on material developed by M. Laurent and F. Vallentin

Monique Laurent

Centrum Wiskunde & Informatica (CWI), Amsterdam & Tilburg University
Science Park 123, 1098 XG Amsterdam

monique@cwi.nl

Frank Vallentin

Universität zu Kóln
Weyertal 86-90, D-50923 Köln, Germany

frank.vallentin@uni-koeln.de

April 25, 2016

CONTENTS

CHAPTER 1

PRELIMINARIES: 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

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

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 $\mathbb{R}^n \setminus A$ 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 $\mathbb{R}^n \setminus A$. 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).

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

If the dimension of $A \subseteq \mathbb{R}^n$ is strictly smaller than n, then A does not have an interior, $\operatorname{int} A = \emptyset$. 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

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

Convex sets are closed under *convex combinations*:

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

Throughout we set $\mathbb{R}_{\ge 0} = \{\lambda \in \mathbb{R} : \lambda \ge 0\}$. The convex hull of $A \subseteq \mathbb{R}^n$ is the smallest convex set containing A. It is

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

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

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.

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 $\operatorname{int} C = \emptyset$ 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 $c \in \mathbb{R}^n \setminus \{0\}$ 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*

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

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

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

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 a point $x \in A$ 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.

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 $x \in \mathbb{R}^n \setminus C$ be a point outside of C. Then there exists a unique point $\pi_C(x)$ in C which is closest to x. Moreover, $\pi_C(x) \in \partial C$.*

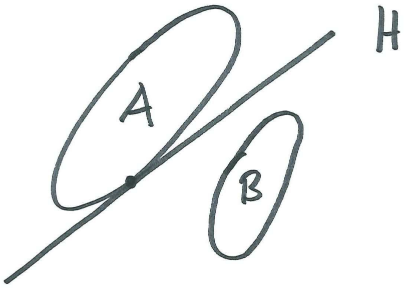


Figure 1.1: The hyperplane H supports A and separates A and B.

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 $y \in C'$. 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

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

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,*

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

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 $y_i \in \mathbb{R}^n \setminus C$ 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

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

By intersecting the line $\operatorname{aff}\{y_i, \pi_C(y_i)\}$ 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. □

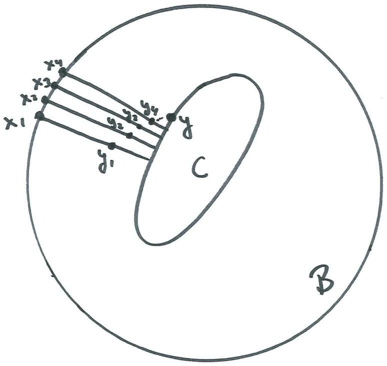


Figure 1.2: The construction which proves Lemma 1.3.3.

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 $x \in \mathbb{R}^n \setminus C$ 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 $\{y : c^{\mathsf{T}} y \leq c^{\mathsf{T}} \pi_C(x)\}$. 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. □

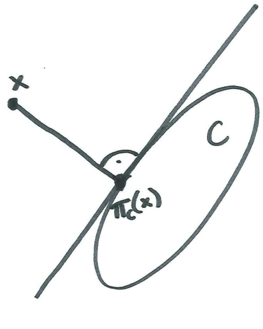


Figure 1.3: A separating hyperplane constructed using π_C.

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 $x \in \mathbb{R}^n \setminus C$ 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 $x \not\in \overline{C}$, 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.,

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

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

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

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

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

Define

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

Then,

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

and we want to show that

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

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 $\beta \leq c^{\mathsf{T}} x < c^{\mathsf{T}} y$. 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 $C \cap D = \emptyset$. 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

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

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 $\operatorname{int} C \neq \emptyset$. 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. □

An easy, but very useful application, is that if one maximizes a convex function on a compact convex set then the maximum is attained at an extreme point.

Lemma 1.4.2. *Let $C \subseteq \mathbb{R}^n$ be a compact convex set and let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a continuous convex function. Then, the maximization program $\max_{x \in X} f(x)$ has a maximizer which is an extreme point of C. This applies in particular to maximizing (or minimizing) a linear function over C.*

Proof. Let $x^* \in C$ be a maximizer of f over C (which exists by continuity of f over C compact). By Theorem 1.4.1, x^* is a convex combination of extreme points $x_1, \ldots, x_N$ of C, i.e. $x^* = \sum_{i=1}^N \lambda_i x_i$ where $\lambda_i > 0$ with $\sum_{i=1}^N \lambda_i = 1$. Then, as f is convex, $f(x^*) \leq \sum_{i=1}^N \lambda_i f(x_i)$. Moreover, $f(x_i) \leq f(x^*)$ by maximality of x^*. This implies $f(x_i) = f(x^*)$ for each i and thus any x_i is a maximizer as well. □

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:*

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

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.

The *dual* of a cone $K \subseteq \mathbb{R}^n$ is defined as

$$K^* = \{y \in \mathbb{R}^n : x^{\mathsf{T}} y \geq 0\ \forall x \in K\}.$$

The set K^* is a closed convex cone.

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

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

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:

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

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 moreover it should be full-dimensional, i.e. have non-empty interior. Then, we define strict inequalities by:

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

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 $x \in \mathbb{R}^n \setminus C$ 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*

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

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

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

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

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

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

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

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

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

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: $x_i \geq y_i$ 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

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

Sometimes it is also called the *ice cream cone* (make a drawing of $\mathcal{L}^3$ to convince yourself) or the *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 space $\mathcal{S}^n$ of symmetric $n\times n$ matrices, which can be seen as the $(n(n+1)/2)$-dimensional Euclidean space, equipped with the trace inner product: for two matrices $X, Y \in \mathbb{R}^{n\times n}$,

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

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 of the matrix X. We will come back to the trace iner product in Section 1.7.2 below.

The *cone of semidefinite matrices* is

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

where a matrix X is *positive semidefinite* if

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

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:

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

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

Throughout we let I_n denote the identity matrix and J_n denotes the all-ones matrix. Sometimes the index n may be omitted if there is no ambiguity on the size of the matrices.

A matrix $P \in \mathbb{R}^{n \times n}$ 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.

For a matrix $X \in \mathbb{R}^{n \times n}$, a nonzero vector $u \in \mathbb{R}^n$ is an *eigenvector* of X if there exists a scalar $\lambda \in \mathbb{R}$ such that $Xu = \lambda u$, then λ is the *eigenvalue* of X for the eigenvector u. A fundamental property of real symmetric matrices is that they admit a set of eigenvectors $\{u_1, \ldots, u_n\}$ forming an orthonormal basis of $\mathbb{R}^n$. This is the *spectral decomposition theorem*, one of the most important theorems about real symmetric matrices.

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

$$X = \sum_{i=1}^{n} \lambda_i u_i u_i^{\mathsf{T}}, \tag{1.1}$$

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: $x^T X x \geq 0$ 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 $\lambda_i \geq 0$.

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

(4) There exist vectors $v_1, \dots, v_n \in \mathbb{R}^k$ (for some $k \geq 1$) 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.

Proof. (1) $\Longrightarrow$ (2): By assumption, $u_i^\mathsf{T} X u_i \geq 0$ for all $i \in [n]$. On the other hand, $Xu_i = \lambda_i u_i$ implies $u_i^\mathsf{T} X u_i = \lambda_i \|u_i\|^2 = \lambda_i$, and thus $\lambda_i \geq 0$ for all i.

(2) $\Longrightarrow$ (3): By assumption, X has a decomposition (1.1) where all scalars λ_i are nonnegative. Define the matrix $L \in \mathbb{R}^{n\times n}$ whose i-th column is the vector $\sqrt{\lambda_i} u_i$. Then $X = LL^\mathsf{T}$ holds.

(3) $\Longrightarrow$ (4): Assume $X = LL^\mathsf{T}$ where $L \in \mathbb{R}^{n\times k}$. Let $v_i \in \mathbb{R}^k$ denote the i-th row of L. The equality $X = LL^\mathsf{T}$ gives directly that $X_{ij} = v_i^\mathsf{T} v_j$ for all $i, j \in [n]$.

(4) $\Longrightarrow$ (1): Assume $X = (v_i^\mathsf{T} v_j)_{i,j=1}^n$ for some vectors $v_1, \dots, v_n \in \mathbb{R}^k$ and some $k \geq 1$. Let $x = (x_1, \dots, x_n)^\mathsf{T} \in \mathbb{R}^n$. Then, $x^\mathsf{T} X x = \sum_{i,j=1}^n x_i x_j X_{ij} = \sum_{i,j=1}^n x_i x_j v_i^\mathsf{T} v_j = \| \sum_{i=1}^n x_i v_i \|^2$ is thus nonnegative. This shows that $X \succeq 0$.

The equivalence (1) $\Longleftrightarrow$ (5) can be found in any standard Linear Algebra textbook (and will not be used here). □

The above characterization 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 $x \in \mathbb{R}^n \setminus \{0\}$, (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).

Observe that, if X is a diagonal matrix (i.e., $X_{ij} = 0$ for all $i \neq j \in [n]$), then $X \succeq 0$ if and only if all its diagonal entries are nonnegative: $X_{ii} \geq 0$ for all $i \in [n]$. Moreover, $X \succ 0$ if and only if $X_{ii} > 0$ for all $i \in [n]$.

The positive semidefinite cone set $\mathcal{S}^n_{\succeq 0}$ is the set of all positive semidefinite matrices in $\mathcal{S}^n$. It is a pointed, closed, convex, full-dimensional cone in $\mathcal{S}^n$. Moreover, by Theorem 1.7.2(2), it is generated by rank one matrices, i.e.

$$\mathcal{S}^n_{\succeq 0} = \operatorname{cone}\{xx^\mathsf{T} : x \in \mathbb{R}^n\}. \tag{1.2}$$

The matrices lying in the interior of the cone $\mathcal{S}^n_{\succeq 0}$ are precisely the positive definite matrices (Exercise 1.9).

1.7.2 The trace inner product

The *trace* of an $n \times n$ matrix A is defined as $\text{Tr}(A) = \sum_{i=1}^n A_{ii}$. The trace is a linear mapping: $\text{Tr}(\lambda A) = \lambda \text{Tr}(A)$ and $\text{Tr}(A + B) = \text{Tr}(A) + \text{Tr}(B)$. Moreover the trace satisfies the following properties:

$$\text{Tr}(A) = \text{Tr}(A^{\mathsf{T}}),\ \text{Tr}(AB) = \text{Tr}(BA),\ \text{Tr}(uu^{\mathsf{T}}) = u^{\mathsf{T}}u = \|u\|^2\ \forall u \in \mathbb{R}^n. \quad (1.3)$$

Using the fact that $\text{Tr}(uu^{\mathsf{T}}) = 1$ for any unit vector u, combined with the relation (1.1), we deduce that the trace of a symmetric matrix is equal to the sum of its eigenvalues.

Lemma 1.7.3. *If $A \in \mathcal{S}^n$ has eigenvalues $\lambda_1, \dots, \lambda_n$, then $\text{Tr}(A) = \lambda_1 + \ldots + \lambda_n$.*

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 $\mathbb{R}^{n\times n}$ obtained by setting $\|A\| = \sqrt{\langle A, A\rangle} = \sqrt{\sum_{i,j=1}^{n} A_{ij}^2}$. For a vector $x \in \mathbb{R}^n$ we have $x^{\mathsf{T}}Ax = \langle A, xx^{\mathsf{T}}\rangle$. The following property is useful to know:

Lemma 1.7.4. *Let $A, B \in \mathcal{S}^n$ and $P \in \mathcal{O}(n)$. Then, $\langle A, B\rangle = \langle PAP^{\mathsf{T}}, PBP^{\mathsf{T}}\rangle$.*

Proof. Indeed, $\langle PAP^{\mathsf{T}}, PBP^{\mathsf{T}}\rangle$ is equal to

$$\text{Tr}(PAP^{\mathsf{T}}PBP^{\mathsf{T}}) = \text{Tr}(PABP^{\mathsf{T}}) = \text{Tr}(ABP^{\mathsf{T}}P) = \text{Tr}(AB) = \langle A, B\rangle,$$

where we have used the fact that $PP^{\mathsf{T}} = P^{\mathsf{T}}P = I_n$ and the commutativity rule from (1.3). □

Positive semidefinite matrices satisfy the following fundamental property:

Lemma 1.7.5. *For a symmetric matrix $A \in \mathcal{S}^n$,*

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

In other words, the cone $\mathcal{S}^n_{\succeq 0}$ is self-dual, i.e., it coincides with its dual cone:

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

Proof. The proof is based on the fact that $\mathcal{S}^n_{\succeq 0}$ is generated by rank 1 matrices (see (1.2)). Indeed, if $A \succeq 0$ then $\langle A, xx^{\mathsf{T}}\rangle \geq 0$ for all $x \in \mathbb{R}^n$ implies that $\langle A, B\rangle \geq 0$ for all $B \in \mathcal{S}^n_{\succeq 0}$. Conversely, if $\langle A, B\rangle \geq 0$ for all $B \in \mathcal{S}^n_{\succeq 0}$, then for $B = xx^{\mathsf{T}}$ we obtain that $x^{\mathsf{T}}Ax \geq 0$, which shows $A \succeq 0$. □

1.7.3 Basic operations

We recall some basic operations about positive semidefinite matrices. The proofs of the following Lemmas 1.7.6, 1.7.7 and 1.7.8 are easy and left as an exercise.

Given a matrix $X \in \mathcal{S}^n$ and a subset $I \subseteq [n]$ of its index set, the matrix $X[I] = (X_{ij})_{i,j\in I}$ is the principal submatrix of X indexed by I.

Lemma 1.7.6. *If $X \succeq 0$ then every principal submatrix of X is positive semidefinite.*

Moreover, any matrix congruent to $X \succeq 0$ (i.e., of the form PXP^{T} where P is non-singular) is positive semidefinite:

Lemma 1.7.7. *Let $X \in \mathcal{S}^n$ and let $P \in \mathbb{R}^{n\times n}$ be a non-singular matrix (i.e., P is invertible). Then,*

$$X \succeq 0 \Longleftrightarrow PXP^{\mathsf{T}} \succeq 0.$$

Given two matrices $A \in \mathcal{S}^n$ and $B \in \mathcal{S}^m$, we define the following block-diagonal matrix $A \oplus B \in \mathcal{S}^{n+m}$:

$$A \oplus B = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}. \tag{1.4}$$

Lemma 1.7.8. *For $A \in \mathcal{S}^n$ and $B \in \mathcal{S}^m$, we have:*

$$A \oplus B \succeq 0 \Longleftrightarrow A \succeq 0 \ \textit{ and } \ B \succeq 0.$$

From an algorithmic point of view it is more economical to deal with positive semidefinite matrices in block-form like (1.4).

If we have a set $\mathcal{A}$ of matrices that pairwise commute, then it is a fundamental result of linear algebra 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 $P^{\mathsf{T}}XP$ are diagonal for all $X \in \mathcal{A}$.

We now introduce the following notion of *Schur complement*, which can be very useful for showing positive semidefiniteness.

Definition 1.7.9. (Schur complement) *Consider a symmetric matrix X in block form*

$$X = \begin{pmatrix} A & B \\ B^{\mathsf{T}} & C \end{pmatrix}, \tag{1.5}$$

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 1.7.10. *Let $X \in \mathcal{S}^n$ be in block form (1.5) where A is non-singular. Then,*

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

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 $(P^{-1})^{\mathsf{T}}XP^{-1} \succeq 0$ (use Lemma 1.7.7) which is thus equivalent to $A \succeq 0$ and $C - B^{\mathsf{T}}A^{-1}B \succeq 0$ (use Lemma 1.7.8). □

1.7.4 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 $C = A \otimes B \in \mathbb{R}^{np\times mq}$ with entries

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

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 $i \in [n], j \in [m]$. As an example, the matrix $I_2 \otimes J_3$ takes the form:

$$\begin{pmatrix} I_2 & I_2 & I_2 \\ I_2 & I_2 & I_2 \\ I_2 & I_2 & I_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

or after permuting rows and columns, the form:

$$\begin{pmatrix} J_3 & 0 \\ 0 & J_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

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 $i \in [n], j \in [m]$. For an integer $k \geq 1$, $A^{\circ k} = A \circ A \circ \ldots \circ A$ (with k terms) is the matrix with (i, j)-th entry $(A_{ij})^k$, the k-th power of A_{ij}.

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,
$$A, B \succeq 0 \Longrightarrow A \otimes B \succeq 0 \text{ and } A \circ B \succeq 0,$$
$$A \succeq 0 \Longrightarrow A^{\circ k} = ((A_{ij})^k)_{i,j=1}^n \succeq 0\ \forall k \in \mathbb{N}.$$

1.7.5 Properties of the kernel

The *kernel* of a matrix $X \in \mathcal{S}^n$ is the subspace $\ker X = \{x \in \mathbb{R}^n : Xx = 0\}$. Here is a first useful property of the kernel of a positive semidefinite matrix.

Lemma 1.7.11. *Assume that $X \in \mathcal{S}^n$ is positive semidefinite and let $x \in \mathbb{R}^n$. Then,*

$$Xx = 0 \Longleftrightarrow x^{\mathsf{T}} Xx = 0.$$

Proof. The 'only if' part is clear. Conversely, decompose $x = \sum_{i=1}^n x_i u_i$ in the orthonormal basis of eigenvectors of X and let $\lambda_1, \ldots, \lambda_n$ be the corresponding eigenvalues of X. Then, $x^{\mathsf{T}} Xx = \sum_{i=1}^n \lambda_i x_i^2$. Hence, $0 = x^{\mathsf{T}} Xx$ gives $0 = \sum_{i=1}^n \lambda_i x_i^2$ and thus $x_i = 0$ for each i for which $\lambda_i > 0$. This shows that x is a linear combination of the eigenvectors u_i with eigenvalue $\lambda_i = 0$, and thus $Xx = 0$. □

As an example of application, we get the following fact:

Lemma 1.7.12. *Let $X = L^{\mathsf{T}} L \in \mathcal{S}^n$ where $L \in \mathbb{R}^{k \times n}$. Then, $\ker X = \ker L$ and thus* rank(X) = rank(L) *($\leq \min\{k, n\}$).*

Clearly, $X \succeq 0$ implies $X_{ii} \geq 0$ for all i, because $X_{ii} = e_i^{\mathsf{T}} X e_i$, where e_i denotes the i-th standard unit vector (with all zero coordinates except 1 at the i-th position). Moreover, if $X_{ii} = 0$ then the whole i-row and column are identically zero. This follows e.g. from the following property:

Lemma 1.7.13. *Let $X \in \mathcal{S}^n$ with the block-form:*

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

where $A \in \mathcal{S}^b$, $B \in \mathbb{R}^{p \times q}$, $C \in \mathcal{S}^q$ and $n = p + q$. Given a vector $y \in \mathbb{R}^p$, define the vector $x \in \mathbb{R}^n$ defined as $x = (y, 0, \ldots, 0)$. Then,

$$Ay = 0 \Longrightarrow Xx = 0.$$

Proof. We have: $x^{\mathsf{T}} Xx = y^{\mathsf{T}} Ay = 0$ which, by Lemma 9.3.1, implies that $Xx = 0$. □

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[1] 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.[2]

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.

[1] 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.

[2] 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

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

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 $C \neq \emptyset$, then $\mathrm{relint}\, C \neq \emptyset$

1.2. Give a proof for the following statement:

Let $C \subseteq \mathbb{R}^n$ be a closed convex set and let $x \in \mathbb{R}^n \setminus C$ a point lying outside of C. A separating hyperplane H is defined in Lemma 1.3.4. Consider a point y on the line $\operatorname{aff}\{x, \pi_C(x)\}$ 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,

$$\overline{\operatorname{conv} A} = \operatorname{conv} \overline{A}.$$

(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

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

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

(*) What happens for the other p?

1.5. Consider a subset $S \subseteq \mathbb{R}^n_{\geq 0}$. Then, S is said to be *down-monotone* in $\mathbb{R}^n_{\geq 0}$ if for each $x \in S$ all vectors $y \in \mathbb{R}^n_{\geq 0}$ with $0 \leq y \leq x$ belong to S. Moreover, its *antiblocker* abl(S) is defined as

$$\text{abl}(S) = \{ y \in \mathbb{R}^n_{\geq 0} : y^\mathsf{T} x \leq 1 \ \forall x \in S \}.$$

Show: abl(abl$(S)) = S$ if and only if S is nonempty, closed, convex and down-monotone in $\mathbb{R}^n_{\geq 0}$.

1.6. Let P and Q be polyhedra in $\mathbb{R}^n$ such that $P \subseteq Q$.

(a) Show: $P = Q$ if and only if the following equality holds for all weights $w \in \mathbb{R}^n$:

$$\max_{x \in P} w^\mathsf{T} x = \max_{x \in Q} w^\mathsf{T} x. \tag{1.6}$$

(b) Assume that $P \subseteq Q \subseteq \mathbb{R}^n_{\geq 0}$ are down-monotone in $\mathbb{R}^n_{\geq 0}$.
Show: $P = Q$ if and only if (1.6) holds for all nonnegative weights $w \in \mathbb{R}^n_{\geq 0}$.

(c) Show that in (a),(b) it suffices to show that (1.6) holds for all integer valued weights w.

1.7 Given an integer $k \in [n]$ consider the polyhedron

$$P = \{ x \in [0,1]^n : x_1 + \ldots + x_n = k \}.$$

(a) Show: $P = \mathrm{conv}(P \cap \{0,1\}^n)$.

(b) Show that each point $x \in P \cap \{0,1\}^n$ is an extreme point of P.

1.8. Consider the set of matrices:

$$\mathcal{D}_n = \{X \in \mathbb{R}^{n\times n} : Xe = e, X^{\mathsf{T}}e = e, X \geq 0\},$$

where e is the all-ones vector. Matrices in $\mathcal{D}_n$ are called *doubly stochastic* and $0/1$-valued matrices in $\mathcal{D}_n$ are *permutation matrices* (as they correspond to the permutations of $[n]$).
Show: [Birkhoff's theorem] $\mathcal{D}_n = \mathrm{conv}(\mathcal{D}_n \cap \{0,1\}^{n\times n})$.

1.9. Define the matrices $F_{ij}, G_{ij} \in \mathcal{S}^n$ for $1 \leq i < j \leq n$, where G_{ij} has entries 1 at positions $(i,i), (j,j), (i,j)$ and (j,i) and entries 0 elsewhere; F_{ij} has entries 1 at positions (i,i) and (j,j), entries -1 at positions (i,j) and (j,i), and entries 0 at all other positions.

(a) Show: $F_{ij}, G_{ij} \succeq 0$.

(b) Assume that $X \in \mathcal{S}^n$ satisfies the conditions:

$$X_{ii} \geq \sum_{j\in[n]:j\neq i} |X_{ij}| \ \text{ for all } i \in [n].$$

(Then X is said to be *diagonally dominant*.)
Show: $X \succeq 0$.

1.10. (a) Show that the identity matrix I_n lies in the interior of the positive semidefinite cone $\mathcal{S}^n_{\succeq 0}$.

(b) Show that a positive semidefinite matrix A lies in the interior of $\mathcal{S}^n_{\succeq 0}$ if and only if A is positive definite.

1.11. Given $x_1, \dots, x_n \in \mathbb{R}$, consider the following $(n+1)\times(n+1)$ matrix:

$$X = \begin{pmatrix} 1 & x_1 & \dots & x_n \\ x_1 & x_1 & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ x_n & 0 & 0 & x_n \end{pmatrix}.$$

That is, X is indexed by $\{0, 1, \dots, n\}$, with entries $X_{00} = 1$, $X_{0i} = X_{i0} = X_{ii} = x_i$ for $i \in [n]$, and all other entries are equal to 0.
Show: $X \succeq 0$ if and only if $x_i \geq 0$ for all $i \in [n]$ and $\sum_{i=1}^n x_i \leq 1$.

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 and we give some basic properties. Moreover, we present several problems that can be modeled as instances of semidefinite programs, arising from optimization, geometry and algebra. to which we will come back in later chapters.

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*: $\langle A, B \rangle = \text{Tr}(A^{\mathsf{T}}B) = \sum_{i,j=1}^n A_{ij}B_{ij}$ for two matrices $A, B \in \mathbb{R}^{n \times n}$. Here $\text{Tr}(A) = \langle I_n, A \rangle = \sum_{i=1}^n A_{ii}$ denotes the trace of A. Recall that $\text{Tr}(AB) = \text{Tr}(BA)$; 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 $\langle X, Y \rangle \geq 0$ for all $Y \in \mathcal{S}^n_{\succeq 0}$.

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*

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

obtained by intersecting the positive semidefinite cone $\mathcal{S}_{\succeq 0}$ 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, we have: $p^* \in \mathbb{R} \cup \{\pm\infty\}$, with $p^* = -\infty$ if the problem (2.1) is infeasible (i.e., $\mathcal{F} = \emptyset$) 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 $\mathcal{F}_a = \emptyset$ (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 $+\infty$. 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):

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

Indeed, let x denote the vector consisting of the diagonal entries of the matrix X, so that $x \geq 0$ 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:

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

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

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

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

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

The rest of the lemma follows by direct verification. $\square$

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:

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

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

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

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

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

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 $[-1, 1]$) and closed. Hence, in program (2.5), we maximize the continuous function $\langle C, X\rangle$ on a compact set and thus the supremum is attained. Moreover, the program (2.5) is strictly feasible (since the positive definite matrix I_n/n is feasible), hence the infimum is attained in the dual program (2.6) and 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:*

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

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

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 $\lambda_1 \geq \ldots \geq \lambda_n$. Then the sum of its k largest eigenvalues: $\lambda_1 + \ldots + \lambda_k$ is equal to the optimal value of any of the following two programs:*

$$\mu_1 := \max_{Y \in \mathbb{R}^{n \times k}} \left\{\langle C, YY^{\mathsf{T}}\rangle : Y^{\mathsf{T}}Y = I_k\right\}, \tag{2.7}$$

$$\mu_2 := \max_{X \in \mathcal{S}^n} \{\langle C, X\rangle : \mathrm{Tr}(X) = k,\ I_n \succeq X \succeq 0\}. \tag{2.8}$$

That is, $\lambda_1 + \ldots + \lambda_k = \mu_1 = \mu_2$.

The proof will use the fact that the extreme points of the polytope

$$\mathcal{P} = \{x \in [0, 1]^n : e^{\mathsf{T}}x = k\} \tag{2.9}$$

are the points $x \in \mathcal{P} \cap \{0, 1\}^n$. (Recall Exercise 1.7).

Proof. Let $u_1, \ldots, u_n$ denote the eigenvectors corresponding to the eigenvalues $\lambda_1. \ldots, \lambda_n$ of C, let U denote the matrix with columns $u_1, \ldots, u_n$ which is thus an orthogonal matrix, and let D denote the diagonal matrix with entries $\lambda_1, \ldots, \lambda_n$. Thus we have: $C = UDU^{\mathsf{T}}$.

The proof is in three steps and consists of showing each of the following three inequalities: $\lambda_1 + \ldots + \lambda_k \leq \mu_1 \leq \mu_2 \leq \lambda_1 + \ldots + \lambda_k$.

Step 1: $\lambda_1 + \ldots + \lambda_k \leq \mu_1$: Consider the matrix Y with columns $u_1, \ldots, u_k$, then Y is feasible for the program (2.7) with value $\langle C, YY^{\mathsf{T}}\rangle = \lambda_1 + \ldots + \lambda_k$.

Step 2: $\mu_1 \leq \mu_2$: Let Y be feasible for the program (2.7) and set $X = YY^{\mathsf{T}}$, then X is feasible for the program (2.8).

Step 3: $\mu_2 \leq \lambda_1 + \ldots + \lambda_k$: This is the most interesting part of the proof. A first key observation is that the program (2.8) is equivalent to the same program where we replace C by the diagonal matrix D (containing its eigenvalues). Indeed, using the spectral decomposition $C = UDU^{\mathsf{T}}$, we have: $\langle C, X\rangle = \text{Tr}(CX) = \text{Tr}(UDU^{\mathsf{T}}X) = \text{Tr}(DU^{\mathsf{T}}XU) = \text{Tr}(DZ)$, where the matrix $Z = U^{\mathsf{T}}XU$ is again feasible for (2.8). Therefore we obtain:

$$\mu_2 = \max_{Z \in \mathcal{S}^n} \left\{ \langle D, Z\rangle : \text{Tr}(Z) = k,\ I_n \succeq Z \succeq 0 \right\}.$$

Now let $z = (Z_{ii})_{i=1}^n$ denote the vector containing the diagonal entries of Z. The condition: $I \succeq Z \succeq 0$ implies that $z \in [0,1]^n$. Moreover, the condition: $\text{Tr}(Z) = k$ implies $e^{\mathsf{T}}z = k$ and we have: $\text{Tr}(DZ) = \sum_{i=1}^n \lambda_i z_i$. Hence the vector z lies in the polytope $\mathcal{P}$ from (2.9) and we obtain: $\mu_2 \leq \max_{z\in\mathcal{P}} \sum_{i=1}^n \lambda_i z_i$. Now recall that the maximum of the linear function $\sum_{i=1}^n \lambda_i z_i$ is attained at an extreme point of $\mathcal{P}$. As recalled above, the extreme points of $\mathcal{P}$ are the 0/1 valued vectors with exactly k ones. From this follows immediately that the maximum value of $\sum_{i=1}^n \lambda_i z_i$ taken over $\mathcal{P}$ is equal to $\lambda_1 + \ldots + \lambda_k$. Thus we have shown the last inequality: $\mu_2 \leq \lambda_1 + \ldots + \lambda_k$ and this concludes the proof. □

As an application, we obtain that the feasible region of the program (2.8) is equal to the convex hull of the feasible region of the program (2.7). That is,

$$\{X \in \mathcal{S}^n : I_n \succeq X \succeq 0,\ \text{Tr}(X) = k\} = \text{conv}\{YY^{\mathsf{T}} : Y \in \mathbb{R}^{n\times k},\ Y^{\mathsf{T}}Y = I_k\}.$$

2.3 Hoffman-Wielandt inequality and quadratic assignment

In this section we consider the following optimization problem over the set of orthogonal matrices:

$$\text{OPT}(A,B) = \min\left\{\text{Tr}(AXBX^{\mathsf{T}}) : X \in \mathcal{O}(n)\right\}, \tag{2.10}$$

where $A, B \in \mathcal{S}^n$ are two given symmetric matrices. We will indicate below its relation to the quadratic assignment problem.

Quite surprisingly, it turns out that the optimal value of the program (2.10) can be expressed in a closed form in terms of the eigenvalues of A and B. This gives the nice inequality (2.12) about interlacing of eigenvalues of two matrices, due to Hoffman-Wielandt (1953). Moreover, the program (2.10) has an equivalent reformulation as a semidefinite program given in (2.11).

Theorem 2.3.1. *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: $\alpha_1 \leq \ldots \leq \alpha_n$ and $\beta_1 \geq \ldots \geq \beta_n$.*

The program (2.10) is equivalent to the following semidefinite program:

$$\max_{S,T\in\mathcal{S}^n} \{\text{Tr}(S)+\text{Tr}(T) : A\otimes B - I_n\otimes T - S\otimes I_n \succeq 0\}. \tag{2.11}$$

and its optimum value is equal to

$$\text{OPT}(A,B)=\sum_{i=1}^n \alpha_i\beta_i.$$

In particular, the following inequality holds:

$$\text{Tr}(AB)\geq \sum_{i=1}^n \alpha_i\beta_i. \tag{2.12}$$

For the proof we will use an intermediary result about doubly-stochastic matrices. Recall that a matrix $X\in\mathbb{R}^{n\times n}$ is *doubly stochastic* if X is nonnegative and has all its row and column sums equal to 1. So the polyhedron

$$\mathcal{D}_n=\left\{X\in\mathbb{R}^{n\times n}_{\geq 0} : \sum_{i=1}^n X_{ij}=1\ \forall j\in[n],\ \sum_{j=1}^n X_{ij}=1\ \forall i\in[n]\right\}$$

is the set of all doubly stochastic matrices. Given a permutation σ of $[n]$ one can represent it by the corresponding *permutation matrix* $X(\sigma)$ with entries $X(\sigma)_{i\sigma(i)}=1$ for all $i\in[n]$ and all other entries are equal to 0. Hence $0/1$ valued doubly stochastic matrices are precisely the permutation matrices. Moreover, the well known theorem of Birkhoff shows that the set of doubly stochastic matrices is equal to the convex hull of the set of permutation matrices.

Theorem 2.3.2 (Birkhoff's theorem). $\mathcal{D}_n=\text{conv}\{X(\sigma) : \sigma \textit{ is a permutation of } [n]\}$.

Lemma 2.3.3. *Given scalars $\alpha_1,\ldots,\alpha_n$ and $\beta_1,\ldots,\beta_n$ ordered as $\alpha_1\leq\ldots\leq\alpha_n$ and $\beta_1\geq\ldots\geq\beta_n$, consider the following linear program:*

$$\max_{x,y\in\mathbb{R}^n}\left\{\sum_{i=1}^n x_i+\sum_{j=1}^n y_j : \alpha_i\beta_j - x_i - y_j\geq 0\ \forall i,j\in[n]\right\} \tag{2.13}$$

and its dual linear program:

$$\min_{Z\in\mathbb{R}^{n\times n}}\left\{\sum_{i,j=1}^n \alpha_i\beta_j Z_{ij} : \sum_{i=1}^n Z_{ij}=1\ \forall j\in[n],\ \sum_{j=1}^n Z_{ij}=1\ \forall i\in[n],\ Z\geq 0\right\}. \tag{2.14}$$

The optimum value of (2.13) and (2.14) is equal to $\sum_{i=1}^n \alpha_i\beta_i$.

Proof. The feasible region of the program (2.14) is the set $\mathcal{D}_n$of doubly stochastic matrices and, by the above mentioned result of Birkhoff, it is equal to the convex hull of permutation matrices. As the minimum value of (2.14) is attained at an extreme point of $\mathcal{D}_n$ (i.e., at a permutation matrix), it is equal to the minimum value of $\sum_{i=1} \alpha_i \beta_{\sigma(i)}$ 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 (2.14) (and thus of (2.13)) is equal to $\sum_{i=1}^n \alpha_i \beta_i$. □

Proof. (*of Theorem 2.3.1*) The first step in the proof consists of replacing the program (2.10) by an equivalent program where the matrices A and B are diagonal. For this, 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). For $X \in \mathcal{O}(n)$, we have $Y := P^{\mathsf{T}}XQ \in \mathcal{O}(n)$ and $\text{Tr}(AXBX^{\mathsf{T}}) = \text{Tr}(DYEY^{\mathsf{T}})$. Hence the optimization problem (2.10) is equivalent to the program:

$$\text{OPT}(D, E) = \min\{\text{Tr}(DXEX^{\mathsf{T}}) : X \in \mathcal{O}(n)\}. \tag{2.15}$$

That is,

$$\text{OPT}(A, B) = \text{OPT}(D, E).$$

The next step is to show that the program (2.15) has the same optimum value as the linear program (2.14). For this, pick $X \in \mathcal{O}(n)$ and consider the matrix $Z = ((X_{ij})^2)_{i,j=1}^n$ which is doubly-stochastic (since X is orthogonal). Moreover, since

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

it follows that $\text{Tr}(DXEX^{\mathsf{T}})$ is at least the minimum value of the program (2.14) and thus the minimum value of (2.15) is at least the minimum value of (2.14). By Lemma 2.3.3, the minimum value of (2.14) is equal to $\sum_{i=1}^n \alpha_i \beta_i$. So we can already conclude that $\text{OPT}(D, E) \geq \sum_{i=1}^n \alpha_i \beta_i$. The reverse inequality follows by selecting the orthogonal matrix $X = I_n$ as feasible solution of (2.15), so that $\text{OPT}(D, E) \leq \text{Tr}(DE) = \sum_{i=1}^n \alpha_i \beta_i$. Hence we have shown that $\text{OPT}(D, E) = \sum_{i=1}^n \alpha_i \beta_i$ and thus $\text{OPT}(A, B) = \sum_{i=1}^n \alpha_i \beta_i$.

We now show that the semidefinite program:

$$\min_{S', T' \in \mathcal{S}^n} \{\text{Tr}(S') + \text{Tr}(T') : E \otimes F - I_n \otimes T' - S' \otimes I_n \succeq 0\} \tag{2.16}$$

is equivalent to the program (2.11). Indeed, using the relation

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

and the fact that $P \otimes Q$ is orthogonal, we see that S, T is feasible for (2.11) if and only if $S' = PSP^{\mathsf{T}}$, $T' = QTQ^{\mathsf{T}}$ is feasible for (2.16) and moreover we have $\text{Tr}(S) + \text{Tr}(T) = \text{Tr}(S') + \text{Tr}(T')$.

Finally we show that the program (2.16) has the same optimum value as the linear program (2.13). For this, first observe that in the program (2.16) we may assume without loss of generality that the matrices S' and T' are diagonal. Indeed, if we define the vectors $x = \mathrm{diag}(S')$ and $y = \mathrm{diag}(T')$, we see that, since $E \otimes F$ is diagonal, the diagonal matrices $S'' = \mathrm{Diag}(x)$ and $T'' = \mathrm{Diag}(y)$ are still feasible for (2.16) with the same objective value: $\mathrm{Tr}(S') + \mathrm{Tr}(T') = \mathrm{Tr}(S'') + \mathrm{Tr}(T'')$. Now, the program (2.16) with the additionnal condition that S', T' are diagonal matrices can be rewritten as the linear program (2.13), since the matrix $E \otimes F - I_n \otimes T' - S' \otimes I_n$ is diagonal with diagonal entries $\alpha_i\beta_j - x_i - y_j$ for $i, j \in [n]$. From this we can conclude that the maximum value of the program (2.16) is equal to the maximum value of (2.13) (and thus of (2.14)). In turn we can conclude that the program (2.16) (and thus (2.11)) has the same optimum value as the program (2.10), which finishes the proof. □

The result of Theorem 2.3.1 can be used to give an explicit lower bound for the following *quadratic assignment problem* (QAP):

$$\mathrm{QAP}(A, B) = \min \left\{ \sum_{i,j=1}^{n} A_{ij} B_{\sigma(i)\sigma(j)} : \sigma \text{ is a permutation of } [n] \right\}. \tag{2.17}$$

The QAP problem models e.g. the following facility location problem, where one wants to allocate n facilities to n locations at the lowest possible total cost. The cost of allocating facilities i and j to the respective locations $\sigma(i)$ and $\sigma(j)$ is then $A_{ij}B_{\sigma(i)\sigma(j)}$ (A_{ij} is the 'flow' cost between the facilities i and j, and B_{hk} is the 'distance' between the locations h and k). Or think of the *campus building problem*, where one needs to locate n buildings at n locations, A_{ij} represents the traffic intensity between buildings i and j, and B_{hk} is the distance between locations h and k.

As QAP is NP-hard one needs to find good tractable lower bounds for it. For this observe first that problem (2.17) can be reformulated as the following optimization problem over the set of permutation matrices:

$$\mathrm{QAP}(A, B) = \min\{\mathrm{Tr}(AXBX^{\mathsf{T}}) : X \text{ is a permutation matrix}\}$$

(because $\sum_{i,j=1}^{n} A_{ij}B_{\sigma(i)\sigma(j)} = \mathrm{Tr}(AXBX^{\mathsf{T}})$ if $X = X(\sigma)$). Then, observe that a matrix X is a permutation matrix if and only if it is doubly stochastic and orthogonal (Exercise 2.6). Hence, if in program (2.17) we relax the condition that X be a permutation matrix by the condition that X be orthogonal we obtain program (2.10). This shows:

$$\mathrm{QAP}(A, B) \geq \mathrm{OPT}(A, B)$$

and the next result.

Theorem 2.3.4. *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: $\alpha_1 \leq \ldots \leq \alpha_n$ and $\beta_1 \geq \ldots \geq \beta_n$. Then,*

$$\mathrm{QAP}(A, B) \geq \mathrm{OPT}(A, B) = \sum_{i=1}^{n} \alpha_i \beta_i.$$

2.4 Convex quadratic constraints

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

$$x^{\mathsf{T}}Ax \leq b^{\mathsf{T}}x + c, \tag{2.18}$$

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

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

Proof. The equivalence follows as a direct application of Lemma 1.7.10: 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:

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

as well as its homogenization:

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

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*), which we briefly introduced in the previous chapter and 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 2.4.2. *Given $c \in \mathbb{R}^n$, the following holds:*

$$\begin{aligned} \min_{\|x\|\leq 1} c^{\mathsf{T}}x &= \min_{x\in\mathbb{R}^n} c^{\mathsf{T}}x \;\; \textit{s.t.} \;\; \begin{pmatrix} 1 & x^{\mathsf{T}} \\ x & I_n \end{pmatrix} \succeq 0 \\ &= \max_{X\in\mathcal{S}^{n+1}} -\mathrm{Tr}(X) \;\; \textit{s.t.} \;\; 2X_{0i} = c_i \; (i \in [n]), \; X \succeq 0. \end{aligned} \tag{2.19}$$

Proof. Apply Lemma 2.4.1 combined with the duality theorem (Theorem 2.1.2). □

2.5 Robust optimization

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

Consider the following linear programming problem:

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

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

$$\max\{c^{\mathsf{T}}x : a(z)^{\mathsf{T}}x \geq b(z)\ \forall z \in \mathcal{Z}\}. \tag{2.20}$$

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

$$a(z) = a_0 + \sum_{j=1}^{k} \zeta_j a_j,\ b(z) = b_0 + \sum_{j=1}^{k} \zeta_j b_j \tag{2.21}$$

where $a_j \in \mathbb{R}^n$ and $b_j \in \mathbb{R}$ are known. Then the robust optimization problem (2.20) can be reformulated as a semidefinite programming problem involving the variable $x \in \mathbb{R}^n$ and a new matrix variable $Z \in \mathcal{S}^k_{\succeq 0}$. The proof relies on the result from Corollary 2.4.2, where we made use in a crucial manner of the duality theory for semidefinite programming, for showing the equivalence of both problems in (2.19).

Theorem 2.5.1. *Suppose that the functions $a(z)$ and $b(z)$ are given by (2.21) and that $\mathcal{Z} = \{z \in \mathbb{R}^k : \|z\| \leq 1\}$. Then problem (2.20) is equivalent to the problem:*

$$\begin{aligned} \min_{x \in \mathbb{R}^n, Z \in \mathcal{S}^{k+1}} c^{\mathsf{T}}x \ \textit{ such that } \quad & a_j^{\mathsf{T}}x - 2Z_{0j} = b_j \ (j \in [k]) \\ & a_0^{\mathsf{T}}x - \mathrm{Tr}(Z) \geq b_0,\ Z \succeq 0. \end{aligned} \tag{2.22}$$

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: $a(z)^{\mathsf{T}}x \geq b(z)$ $\forall z \in Z$ can be rewritten as

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

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

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

Now the above problem fits precisely within the setting considered in Corollary 2.4.2. Hence, we can rewrite it using the second formulation in (2.19) – the one in maximization form – as

$$p^* = \max_{Z \in \mathcal{S}^{k+1}} \{-\mathrm{Tr}(Z) : 2Z_{0j} = \alpha_j \ (j \in [k]), Z \succeq 0\}.$$

So, in problem (2.20), we can substitute the condition: $a(z)^{\mathsf{T}} x \geq b(z)\ \forall z \in \mathcal{Z}$ by the condition:

$$\exists Z \in \mathcal{S}^{k+1}_{\succeq 0} \ \text{ s.t. } -\mathrm{Tr}(Z) \geq -\alpha_0,\ 2Z_{0j} = \alpha_j \ (j \in [k]).$$

The crucial fact here is that the quantifier "$\forall z$" has been replaced by the existential quantifier "$\exists Z$". As problem (2.20) is a maximization problem in x, it is equivalent to the following maximization problem in the variables x and Z:

$$\max_{x \in \mathbb{R}^n, Z \in \mathcal{S}^{k+1}} \left\{ c^{\mathsf{T}} x : a_0^{\mathsf{T}} x - \mathrm{Tr}(Z) \geq b_0,\ a_j^{\mathsf{T}} x - 2Z_{0j} = b_j \ (j \in [k]) \right\}$$

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

2.6 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 later chapters. For now we illustrate the main idea on the following two examples: finding a maximum independent set and a maximum cut in a graph.

2.6.1 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 $\mathcal{NP}$-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

$$\vartheta(G) = \max_{X \in \mathcal{S}^n} \{\langle J, X \rangle : \mathrm{Tr}(X) = 1,\ X_{ij} = 0 \ (\{i,j\} \in E),\ X \succeq 0\}, \tag{2.23}$$

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,

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

The dual semidefinite program reads

$$\min_{y \in \mathbb{R}^E} \left\{ \sum_{ij \in E} y_{ij} : \sum_{ij \in E} y_{ij} E_{ij} - J \succeq 0 \right\}, \tag{2.24}$$

and its optimum value is equal to $\vartheta(G)$ (because (4.16) 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.

2.6.2 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 $(S, V \setminus S)$, 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 $i \in V \setminus S$. 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

$$\text{max-cut} = \max_{x \in \mathbb{R}^n} \left\{ \sum_{\{i,j\} \in E} (1 - x_i x_j)/2 : x \in \{\pm 1\}^n \right\}. \tag{2.25}$$

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:

$$\text{sdp} = \max_{X \in \mathcal{S}^n} \left\{ \sum_{\{i,j\} \in E} (1 - X_{ij})/2 : X \succeq 0,\ X_{ii} = 1 \ (i \in [n]) \right\}. \tag{2.26}$$

As we will see later this semidefinite program provides a very good approximation for the max-cut problem: sdp $\leq 1.13 \cdot$ 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

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

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 (2.26) as

$$\text{sdp } = \max \left\{ \frac{1}{4} \langle L_G, X \rangle : X \succeq 0, \; X_{ii} = 1 \; (i \in [n]) \right\}. \tag{2.27}$$

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

$$\text{opt } = \max\{x^T A x : \|x\|_\infty \leq 1\}. \tag{2.28}$$

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 (5.11) is equivalent to

$$\text{opt} = \max\{x^T A x : x \in \{\pm 1\}^n\}. \tag{2.29}$$

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

Note that if we would replace in (5.11) 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 (2.29) (and thus of (5.11)):

$$\text{sdp } = \max\{\langle A, X \rangle : X \succeq 0, \; X_{ii} = 1 \; \forall i \in [n]\}. \tag{2.30}$$

We will see later that this semidefinite program too gives a good approximation of the quadratic problem (5.11): sdp $\leq \frac{\pi}{2}$ opt .

2.7 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 (e.g., 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, that we can view as the set of 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 $d \in \mathbb{R}^E_{\geq 0}$, 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 $k = 1, 2$, or 3 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* $k \geq 1$, then the problem can be cast as the problem of deciding feasibility of a semidefinite program.

Lemma 2.7.1. *Given $d \in \mathbb{R}^E_{\geq 0}$, there exist vectors $u_1, \ldots, u_n \in \mathbb{R}^k$ (for some $k \geq 1$) if and only if the following semidefinite program is feasible:*

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

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 $k \geq 1$), 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.

2.8 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

$$p = \sum_{\alpha} p_\alpha x^\alpha = \boldsymbol{p}^\mathsf{T}[x]_d. \qquad (2.31)$$

Definition 2.8.1. *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.*

As an example, consider the polynomial $p = 3x_1^2 + x_2^2 - 2x_1x_2 - 2x_1 + 4 = \boldsymbol{p}^\mathsf{T}[x]_2$, where $[x]_2 = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)^\mathsf{T}$ and $\boldsymbol{p} = (4, -2, 0, 3, -2, 1)^\mathsf{T}$. Then p is SOS since $p = (x_1 - x_2)^2 + (x_1 - 2)^2 + x_1^2$.

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 (2.31):

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

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

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

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 $p(x) - \lambda \geq 0$ 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: $p^* \leq p^{\min}$.

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.

2.9 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 2.3.1 and applications 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 2.5.1 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 [10]. 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 [5] 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 (2.26) 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.

2.10 Exercises

2.1. (a) Show that the dual SDP of the program (2.8) can be formulated as the following SDP:

$$\min_{z\in\mathbb{R}, Z\in\mathcal{S}^n} \left\{ kz + \sum_{i=1}^n Z_{ii} : Z \succeq 0,\ -C + zI + Z \succeq 0 \right\}.$$

(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., $p(x) \geq 0\ \forall x \in \mathbb{R}$.

(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 (2.27) 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?

2.5. Let $G = (V = [n], E)$ be a graph and let $L_G \in \mathcal{S}^n$ be its Laplacian matrix, whose eigenvalues are denoted $\lambda_1 \leq \lambda_2 \leq \ldots, \leq \lambda_n$.

(a) Show that L_G is positive semidefinite.

(b) Show: If G is connected then the kernel of L_G has dimension 1.

(c) Show: The dimension of the kernel of L_G is equal to the number of connected components of G.

(d) Show: $\lambda_2 > 0$ if and only if G is connected.

2.6. Show that the following assertions are equivalent for a matrix $X \in \mathbb{R}^{n \times n}$:

(1) X is a permutation matrix.

(2) X is an orthogonal matrix and X is doubly stochastic.

(3) X is doubly stochastic and $\|X\| = \sqrt{n}$.

(4) X is doubly stochastic with entries in $\{0, 1\}$.

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

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

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*

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

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

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

The *equality constraints* are given by vectors $a_j \in \mathbb{R}^n \setminus \{0\}$ 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

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

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 $\mathbb{R}^n_{\geq 0}$ – 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

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

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

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

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 $\lambda_1, \ldots, \lambda_N \in \mathbb{R}_{\geq 0}$ 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*

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

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*

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

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 $\sum_{j=1}^{m} y_j a_j - c \in \operatorname{int} K^*$.

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 $x \in \mathbb{R}^n \setminus K$. 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 $c \in \mathbb{R}^n \setminus \{0\}$. Say, $K \subseteq H^+ = \{z : c^{\mathsf{T}}z \geq 0\}$ 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 $\mathbb{R}^n_{\geq 0}$ is a *linear program*. We write a primal linear program (in standard form) as

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

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

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

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 $\mathcal{L}^{n+1}$:

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

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

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

3.3.3 Semidefinite programming (SDP)

A conic program where K is the cone of semidefinite matrices $\mathcal{S}^n_{\succeq 0}$ 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,*

   $$c^{\mathsf{T}} x \leq b^{\mathsf{T}} y.$$

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

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*

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

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*

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

   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

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

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 $c^{\mathsf{T}} x^* \geq d^*$. Then, weak duality implies $p^* = d^*$ and hence x^* is a maximizer of the primal conic program.

Consider the set

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

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

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

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 $y^{\mathsf{T}} b \leq d^*$. 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 $(y_1 + \epsilon) b_1 + \sum_{j=2}^{m} y_j b_j < y^{\mathsf{T}} b \leq d^*$. In other words, the vector $\tilde{y} = (y_1 + \epsilon, y_2, \ldots, y_m)^{\mathsf{T}}$ is dual feasible with $b^{\mathsf{T}} \tilde{y} < d^*$. 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

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

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

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

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 λy lies in the convex cone K^*. Making λ extremely large drives $x^{\mathsf{T}} \lambda y$ 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 $x^{\mathsf{T}} c \geq \mu d^*$.

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), $\sup_{z \in M} x^{\mathsf{T}} z \leq 0$. In other words, by the definition of M, for any $y \in \mathbb{R}^m$,

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

or, equivalently,

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

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

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

It suffices now to verify that μ is positive. Indeed suppose that $\mu = 0$. Then, on the one hand, we have that $x^{\mathsf{T}} c \geq 0$. 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 $x \in K \setminus \{0\}$ (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 $c^{\mathsf{T}} x^* \geq d^*$. □

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

$$d^* = \inf\left\{y_2 : y_1A_1 + y_2A_2 - C = \begin{pmatrix}y_1 + 1 & 0 & y_2\\ 0 & y_2 + 1 & 0\\ y_2 & 0 & 0\end{pmatrix} \succeq 0\right\}.$$

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: $d^* - p^* = 1$.

3.5.3 Both primal and dual infeasible

Consider the semidefinite program

$$p^* = \sup\left\{\left\langle \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, X \right\rangle : X \succeq 0,\ \left\langle \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, X \right\rangle = 0,\ \left\langle \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix}, X \right\rangle = 1\right\}$$

and its dual

$$d^* = \inf\left\{ y_2 : y_1 \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} + y_2 \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \succeq 0 \right\}.$$

Both programs are infeasible, so that $-\infty = p^* < d^* = +\infty$.

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

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

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 $A^{\mathsf{T}} y \geq 0$ 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 $Ax = b$, $x \geq 0$ has a solution,

(2) Or the linear system $A^{\mathsf{T}} y \geq 0$, $b^{\mathsf{T}} y < 0$ 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:

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

$$y_1E_{11} + y_2E_{12} \succeq 0,\ y_2 < 0, \tag{3.6}$$

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 $a_1^{\mathsf{T}}, \dots, a_m^{\mathsf{T}}$ 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 $\sum_{j=1}^m y_j a_j = A^{\mathsf{T}}y \in K^* \setminus \{0\}, b^{\mathsf{T}}y \leq 0$.

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

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

so that $L = \mathcal{L} + x_0$. By assumption, $L \cap \operatorname{int} K = \emptyset$. 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

$$\forall x \in K : c^{\mathsf{T}}x \geq \beta \text{ and } \forall x \in L : c^{\mathsf{T}}x \leq \beta.$$

Then $\beta \leq 0$ (as $0 \in K$) and $c \in K^*$ (as $c^{\mathsf{T}}tx \geq \beta$ for all $x \in K$ and $t > 0$, which implies that $c^{\mathsf{T}}x \geq 0$). Moreover, for any $x \in \mathcal{L}$ and any scalar $t \in \mathbb{R}$, we have that $c^{\mathsf{T}}(tx + x_0) \leq \beta$ 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 $A^{\mathsf{T}}y \in K^* \setminus \{0\}$. Finally, $y^{\mathsf{T}}b = y^{\mathsf{T}}Ax_0 = c^{\mathsf{T}}x_0 \leq \beta \leq 0$ (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 $y_2 \leq 0$ 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 $\|Ax - b\| \le \epsilon$.*

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, i.e., there does not exist $x \in K$ such that $Ax = b$. 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 $\sum_{j=1}^m y_j a_j - c \in K^*$.

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

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

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

Hence, $b^{\mathsf{T}}y < 0$ implies $c^{\mathsf{T}}y \le 0$. This implies that $c = \lambda b$ for some positive λ and, up to rescaling, we can assume that $c = b$. Therefore,

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

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

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

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)

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

Clearly the primal (3.8) is strictly feasible and $d^* \ge 0$ (since $y = 0$ is feasible). Moreover, $d^* \le 0$ by (4.26). 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* $\|(\sum_{j=1}^m y_j a_j - c) - z\| \leq \epsilon$*.*

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

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

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 (5.4) 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 $x_n \geq 2^{2^n}$. 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 [1]. 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 §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

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

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

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 $\chi^S \in \{0, 1\}^S$, whose i-th entry is 1 if $1 \in S$ and 0 otherwise. As before, e denotes the all-ones vector.

4.1 Preliminaries on graphs

4.1.1 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 $\alpha(G) \geq k$ is an $\mathcal{N}P$-complete problem.

Given an integer $k \geq 1$, 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 $\mathcal{N}P$-complete problem to decide whether a graph is k-colorable. In fact, it is $\mathcal{N}P$-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.

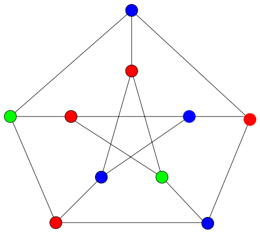


Figure 4.1: The Petersen graph has $\alpha(G) = 4$, $\omega(G) = 2$ and $\chi(G) = 3$

Clearly, any two nodes in a clique of G must receive distinct colors. Therefore, for any graph, the following inequality holds:

$$\omega(G) \leq \chi(G). \tag{4.1}$$

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} ($n \geq 2$), $\omega(C_{2n+1}) = 2$ 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 4.2.

4.1.2 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

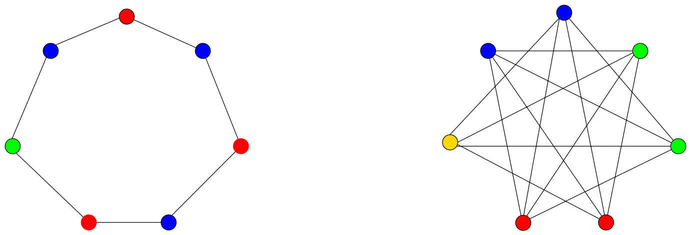


Figure 4.2: For C_7 and its complement $\overline{C_7}$: $\omega(C_7) = 2$, $\chi(C_7) = 3$, $\omega(\overline{C_7}) = \alpha(C_7) = 3$, $\chi(\overline{C_7}) = 4$

$\hat{G}$ with $\omega(\hat{G}) = \chi(\hat{G})$, simply by adding to G a set of $\chi(G)$ new nodes forming a clique. This justifies the following definition, introduced by C. Berge in the early sixties, which makes the problem well posed.

Definition 4.1.1. *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 (the min-max relation $\omega(G) = \chi(G) \leq 2$ is clear) and complement of line graphs of bipartite graphs are perfect (then the min-max relation claims that in a bipartite graph the maximum cardinality of a matching is equal to the minimum cardinality of a vertex cover, which is true by a theorem of König).

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 in 1961 that *all* perfect graphs arise in this way. Resolving this conjecture has haunted generations of graph theorists. It was finally settled in 2006 by Chudnovsky, Robertson, Seymour and Thomas who proved the following result, known as the *strong perfect graph theorem*:

Theorem 4.1.2. (The strong perfect graph theorem)*[2] 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 4.1.3. (The perfect graph theorem) *If G is a perfect graph, then its complement $\overline{G}$ too is a perfect graph.*

We give a direct proof of Theorem 4.1.3 in the next section and we will mention later some other, more geometric, characterizations of perfect graphs (see, e.g., Theorem 4.2.4).

4.1.3 The perfect graph theorem

Lovász [9] proved the following result, which implies the perfect graph theorem (Theorem 4.1.3). The proof given below follows the elegant linear-algebraic argument of Gasparian [4].

Theorem 4.1.4. *A graph G is perfect if and only if $|V(G')| \leq \alpha(G')\omega(G')$ for each induced subgraph G' of G.*

Proof. Necessity is easy: Assume that G is perfect and let G' be an induced subgraph of G. Then $\chi(G') = \omega(G')$ and thus $V(G')$ can be covered by $\omega(G')$ stable sets, which implies that $|V(G')| \leq \omega(G')\alpha(G')$.

To show sufficiency, assume for a contradiction that there exists a graph G which satisfies the condition but is not perfect; choose such a graph with $|V(G)|$ minimal. Then, $n \leq \alpha(G)\omega(G)$, $\omega(G) < \chi(G)$ and $\omega(G') = \chi(G')$ for each induced subgraph $G' \neq G$ of G. Set $\omega = \omega(G)$ and $\alpha = \alpha(G)$ for simplicity. Our first claim is:

Claim 1: There exist $\alpha\omega + 1$ stable sets $S_0, \ldots, S_{\alpha\omega}$ such that each vertex of G is covered by exactly α of them.

Proof of the claim: Let S_0 be a stable set of size α in G. For each node $v \in S_0$, as $G \setminus v$ is perfect (by the minimality assumption on G), $\chi(G \setminus v) = \omega(G \setminus v) \leq \omega$. Hence, $V \setminus \{v\}$ can be partitioned into ω stable sets. In this way we obtain a collection of $\alpha\omega$ stable sets which together with S_0 satisfy the claim. □

Our next claim is:

Claim 2: For each $i = 0, 1, \ldots, \alpha\omega$, there exists a clique K_i of size ω such that $K_i \cap S_i = \emptyset$ and $K_i \cap S_j \neq \emptyset$ for $j \neq i$.

Proof of the claim: For each $i = 0, 1, \ldots, \alpha\omega$, as $G \setminus S_i$ is perfect we have that $\chi(G \setminus S_i) = \omega(G \setminus S_i) \leq \omega$. This implies that $\chi(G \setminus S_i) = \omega$ since, if $\chi(G \setminus S_i) \leq \omega - 1$, then one could color G with ω colors, contradicting our assumption on G. Hence there exists a clique K_i disjoint from S_i and with $|K_i| = \omega$. Moreover K_i meets all the other $\alpha\omega$ stable sets S_j for $j \neq i$. This follows from the fact that each of the ω elements of K_i belongs to α stable sets among the S_j's (Claim 1) and these $\omega\alpha$ sets are pairwise distinct. □

We can now conclude the proof. Define the matrices $M, N \in \mathbb{R}^{n \times (\alpha\omega+1)}$, whose columns are $\chi^{S_0}, \ldots, \chi^{S_{\alpha\omega}}$ (the incidence vectors of the stable sets S_i), and the vectors $\chi^{K_0}, \ldots, \chi^{\alpha\omega+1}$ (the incidence vectors of the cliques K_i), respectively. By Claim 2, we have that $M^{\mathsf{T}}N = J - I$ (where J is the all-ones matrix and I the identity). As $J - I$ is nonsingular, we obtain that rank($M^{\mathsf{T}}N$) = rank($J - I$) = $\alpha\omega + 1$. On the other hand, rank($M^{\mathsf{T}}N$) $\leq$ rank$N \leq n$. Thus we obtain that $n \geq \alpha\omega + 1$, contradicting our assumption on G. □

4.2 Linear programming bounds

4.2.1 Fractional stable sets and colorings

Let ST(G) denote the polytope in $\mathbb{R}^V$ defined as the convex hull of the characteristic vectors of the stable sets of G:

$$\text{ST}(G) = \text{conv}\{\chi^S : S \subseteq V,\ S \text{ is a stable set in } G\},$$

called the *stable set polytope* of G. Hence, computing $\alpha(G)$ is linear optimization over the stable set polytope:

$$\alpha(G) = \max\{e^{\mathsf{T}}x : x \in \text{ST}(G)\}.$$

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

$$\text{ST}(G) = \{x \in \mathbb{R}^V : Ax \leq b\}$$

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*

$$x(C) = \sum_{i \in C} x_i \leq 1 \tag{4.2}$$

is valid for ST(G): any stable set can contain at most one vertex from the clique C. The clique inequalities define the polytope

$$\text{QST}(G) = \left\{x \in \mathbb{R}^V : x \geq 0,\ x(C) \leq 1\ \forall C \text{ clique of } G\right\}. \tag{4.3}$$

Cleary, QST(G) is a relaxation of the stable set polytope:

$$\text{ST}(G) \subseteq \text{QST}(G). \tag{4.4}$$

Maximizing the linear function $e^{\mathsf{T}}x$ over the polytope QST(G) gives the parameter

$$\alpha^*(G) = \max\{e^{\mathsf{T}}x : x \in \text{QST}(G)\}, \tag{4.5}$$

known as the *fractional stability number* of G. Analogously, $\chi^*(G)$ denotes the *fractional coloring number* of G, defined by the following linear program:

$$\chi^*(G) = \min\left\{\sum_{S \text{ stable in } G} \lambda_S : \sum_{S \text{ stable in } G} \lambda_S \chi^S = e,\ \lambda_S \geq 0\ \forall S \text{ stable in } G\right\}. \tag{4.6}$$

If we add the constraint that all λ_S should be integral then we obtain the coloring number of G. Thus, $\chi^*(G) \leq \chi(G)$. In fact the fractional stability number of G coincides with the fractional coloring number of its complement: $\alpha^*(G) = \chi^*(\overline{G})$, and it is nested between $\alpha(G)$ and $\chi(\overline{G})$.

Lemma 4.2.1. *For any graph G, we have*

$$\alpha(G) \leq \alpha^*(G) = \chi^*(\overline{G}) \leq \chi(\overline{G}), \tag{4.7}$$

where $\chi^*(\overline{G})$ is the optimum value of the linear program:

$$\min \left\{ \sum_{C \text{ clique of } G} y_C : \sum_{C \text{ clique of } G} y_C \chi^C = e, \; y_C \geq 0 \; \forall C \text{ clique of } G \right\}. \tag{4.8}$$

Proof. The inequality $\alpha(G) \leq \alpha^*(G)$ in (4.7) follows from the inclusion (4.4) and the inequality $\chi^*(\overline{G}) \leq \chi(\overline{G})$ was observed above. We now show that $\alpha^*(G) = \chi^*(\overline{G})$. For this, we first observe that in the linear program (4.5) the condition $x \geq 0$ can be removed without changing the optimal value; that is,

$$\alpha^*(G) = \max\{e^{\mathsf{T}} x : x(C) \leq 1 \; \forall C \text{ clique of } G\} \tag{4.9}$$

(check it). Now, it suffices to observe that the dual LP of the above linear program (4.9) coincides with the linear program (4.8). □

For instance, for an odd circuit C_{2n+1} $(n \geq 2)$, $\alpha^*(C_{2n+1}) = \frac{2n+1}{2}$ (check it) lies strictly between $\alpha(C_{2n+1}) = n$ and $\chi(\overline{C_{2n+1}}) = n + 1$.

When G is a perfect graph, equality holds throughout in relation (4.7). As we see in the next section, there is a natural extension of this result to weighted graphs, which permits to show the equality ST(G) = QST(G) when G is a perfect graph. Moreover, it turns out that this geometric property characterizes perfect graphs.

4.2.2 Polyhedral characterization of perfect graphs

For any graph G, the fractional stable set polytope is a linear relaxation of the stable set polytope: ST$(G) \subseteq$ QST(G). Here we show a geometric characterization of perfect graphs: G is perfect if and only if both polytopes coincide: ST(G) = QST(G).

The following operation of *duplicating a node* will be useful. Let $G = (V, E)$ be a graph and let $v \in V$. Add to G a new node, say v', which is adjacent to v and to all neighbours of v in G. In this way we obtain a new graph H, which we say is obtained from G by *duplicating* v. Repeated duplicating is called *replicating*.

Lemma 4.2.2. *Let H arise from G by duplicating a node. If G is perfect then H too is perfect.*

Proof. First we show that $\alpha(H) = \chi(\overline{H})$ if H arises from G by duplicating node v. Indeed, by construction, $\alpha(H) = \alpha(G)$, which is equal to $\chi(\overline{G})$ since G is perfect. Now, if $C_1, \ldots, C_t$ are cliques in G that cover V with (say) $v \in C_1$, then $C_1 \cup \{v'\}, \ldots, C_t$ are cliques in H covering $V(H)$. This shows that $\chi(\overline{G}) = \chi(\overline{H})$, which implies that $\alpha(H) = \chi(\overline{H})$.

From this we can conclude that, if H arises from G by duplicating a node v, then $\alpha(H') = \chi(\overline{H'})$ for any induced subgraph H' of H, using induction on the number of nodes of G. Indeed, either H' is an induced subgraph of G (if H' does not contain both v and v'), or H' is obtained by duplicating v in an induced subgraph of G; in both cases we have that $\alpha(H') = \chi(\overline{H'})$.

Hence, if H arises by duplicating a node in a perfect graph G, then $\overline{H}$ is perfect which, by Theorem 4.1.3, implies that H is perfect. □

Given node weights $w \in \mathbb{R}^V_+$, we define the following weighted analogues of the (fractional) stability and chromatic numbers:

$$\alpha(G, w) = \max_{x \in \mathrm{ST}(G)} w^{\mathsf{T}} x,$$

$$\alpha^*(G, w) = \max_{x \in \mathrm{QST}(G)} w^{\mathsf{T}} x,$$

$$\chi^*(\overline{G}, w) = \min_y \left\{ \sum_{C \text{ clique of } G} y_C : \sum_{C \text{ clique of } G} y_C \chi^C = w, \ y_C \geq 0 \ \forall C \text{ clique of } G \right\},$$

$$\chi(\overline{G}, w) = \min_y \left\{ \sum_{C \text{ clique of } G} y_C : \sum_{C \text{ clique of } G} y_C \chi^C = w, \ y_C \in \mathbb{Z}, \ y_C \geq 0 \ \forall C \text{ clique of } G \right\}.$$

When w is the all-ones weight function, we find again $\alpha(G)$, $\alpha^*(G)$, $\chi^*(\overline{G})$ and $\chi(\overline{G})$, respectively. The following analogue of (4.7) holds for arbitrary node weights:

$$\alpha(G, w) \leq \alpha^*(G, w) = \chi^*(\overline{G}, w) \leq \chi(\overline{G}, w). \tag{4.10}$$

Lemma 4.2.3. *Let G be a perfect graph and let $w \in \mathbb{Z}^V_{\geq 0}$ be nonnegative integer node weights. Then, $\alpha(\overline{G}, w) = \chi(G, w)$.*

Proof. Let H denote the graph obtained from G by duplicating node i w_i times if $w_i \geq 1$ and deleting node i if $w_i = 0$. Then, by construction, $\alpha(\overline{G}, w) = \omega(H)$, which is equal to $\chi(H)$ since H is perfect (by Lemma 4.2.2). Say, $\tilde{S}_1, \ldots, \tilde{S}_t$ are $t = \chi(H)$ stable sets in H partitioning $V(H)$. Each stable set $\tilde{S}_k$ corresponds to a stable set S_k in G (since $\tilde{S}_k$ contains at most one of the w_i copies of each node i of G). Now, these stable sets $S_1, \ldots, S_t$ have the property that each node i of G belongs to exactly w_i of them, which shows that $\chi(G, w) \leq t = \chi(H)$. This implies that $\chi(G, w) \leq \chi(H) = \alpha(\overline{G}, w)$, giving equality $\chi(G, w) = \alpha(\overline{G}, w)$. □

We can now show the following geometric characterization of perfect graphs, due to Chvátal [1]. In the proof we will use the fact that $\mathrm{ST}(G) \subseteq \mathrm{QST}(G)$ are down-monotone polytopes in $\mathbb{R}^n_{\geq 0}$ (and the properties from Exercise 1.6). Recall that a polytope $P \subseteq \mathbb{R}^n_{\geq 0}$ is *down-monotone* if $x \in P$ and $0 \leq y \leq x$ (coordinate-wise) implies $y \in P$.

Theorem 4.2.4. *[1] A graph G is perfect if and only if* $\mathrm{ST}(G) = \mathrm{QST}(G)$.

Proof. First assume that G is perfect, we show that $\mathrm{ST}(G) = \mathrm{QST}(G)$. As $\mathrm{ST}(G) \subseteq \mathrm{QST}(G)$ are down-monotone in $\mathbb{R}^V_{\geq 0}$, we can use the following property shown in Exercise 1.6: To show equality $\mathrm{ST}(G) = \mathrm{QST}(G)$ it suffices to show that $\alpha(G, w) = \alpha^*(G, w)$ for all $w \in \mathbb{Z}^V_{\geq 0}$; now the latter property follows from Lemma 4.2.3 (applied to $\overline{G}$).

Conversely, assume that $\mathrm{ST}(G) = \mathrm{QST}(G)$ and that G is not perfect. Pick a minimal subset $U \subseteq V$ for which the subgraph G' of G induced by U satisfies $\alpha(G') < \chi(\overline{G'})$. Setting $w = \chi^U$, we have that $\alpha(G') = \alpha(G, w)$ which, by assumption, is equal to $\max_{x \in \mathrm{QST}(G)} w^{\mathsf{T}} x = \alpha^*(G, w)$. Consider the dual of the linear program defining $\alpha^*(G, w)$ with an optimal solution $y = (y_C)$. Pick a clique C of G for which $y_C > 0$, then C is a nonempty subset of U. Moreover, using complementary slackness, we deduce that $x(C) = 1$ for any optimal solution $x \in \mathrm{QST}(G)$ and thus $|C \cap S| = 1$ for any maximum cardinality stable set $S \subseteq U$. Let G'' denote the subgraph of G induced by $U \setminus C$. Then, $\alpha(G'') \leq \alpha(G') - 1 < \chi(\overline{G'}) - 1 \leq \chi(\overline{G''})$, which contradicts the minimality assumption made on U. □

When G is a perfect graph, equality $\mathrm{ST}(G) = \mathrm{QST}(G)$ holds. Hence an explicit linear inequality description is known for its stable set polytope, given by the clique inequalities. However, it is not clear how to use this information in order to give an efficient algorithm for optimizing over the stable set polytope of a perfect graph. As we see later in Section 4.5.1 there is yet another description of $\mathrm{ST}(G)$ – in terms of semidefinite programming, using the theta body $\mathrm{TH}(G)$ – that will allow to give an efficient algorithm.

4.3 Semidefinite programming bounds

4.3.1 The theta number

Definition 4.3.1. *Given a graph $G = (V, E)$, consider the following semidefinite program*

$$\max_{X \in \mathcal{S}^n} \{\langle J, X\rangle : \mathrm{Tr}(X) = 1,\ X_{ij} = 0\ \forall \{i,j\} \in E,\ X \succeq 0\}. \tag{4.11}$$

Its optimal value is denoted as $\vartheta(G)$, and called the theta number *of G.*

This parameter was introduced by Lovász [10]. He proved the following simple, but crucial result – called the Sandwich Theorem by Knuth [6] – 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 4.3.2. (Lovász' sandwich theorem) *For any graph G, we have that*

$$\alpha(G) \leq \vartheta(G) \leq \chi(\overline{G}).$$

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 (4.11) with objective value $\langle J, X\rangle = |S|$ (check it). This shows the inequality $\alpha(G) \le \vartheta(G)$.

Now, consider a matrix X feasible for the program (4.11) and a partition of V into k cliques: $V = C_1 \cup \cdots \cup C_k$. Our goal is now to show that $\langle J, X\rangle \le k$, which will imply $\vartheta(G) \le \chi(\overline{G})$. 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

$$0 \le \langle X, Y\rangle = k^2\mathrm{Tr}(X) - k\langle J, X\rangle$$

and thus $\langle J, X\rangle \le k\ \mathrm{Tr}(X) = k$. □

An alternative argument for the inequality $\vartheta(G) \le \chi(\overline{G})$, showing an even more transparent link to coverings by cliques, will be given later in the paragraph after the proof of Lemma 4.4.2.

4.3.2 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 4.3.2, $\alpha(G) = \chi(\overline{G}) = \vartheta(G)$ can be computed by solving the semidefinite program (4.11), 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 4.1.3):

Theorem 4.3.3. *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 $\alpha(G_n \backslash v_i) \leq \alpha(G_{i-1} \backslash v_i)$ 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 node weights $w \in \mathbb{Z}^V_{\geq 0}$, the above algorithm can also be used to find a stable set S of maximum weight $w(S)$. For this, construct the new graph G' in the following way: Duplicate each node $i \in V$ w_i times, i.e., replace 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. By Lemma 4.2.2, the graph G' is perfect. Moreover, $\alpha(G')$ is equal to the maximum weight $w(S)$ of a stable set S in G. From this it follows that, if the weights w_i are bounded by a polynomial in n, then one can compute $\alpha(G, w)$ in polynomial time. (More generally, one can compute $\alpha(G, w)$ in polynomial time, e.g. by optimizing the linear function $w^{\mathsf{T}} x$ over the theta body TH(G), introduced in Section 4.5.1 below.)

4.3.3 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.*

First of all, such a stable set S exists: in a $\omega(G)$-coloring, any color class S must meet all maximum cliques, since $\omega(G \setminus S) = \chi(G \setminus S) = \omega(G) - 1$.

Now, if we have found such a stable set S, then one can recursively color $G \backslash S$ with $\omega(G \backslash S) = \omega(G) - 1$ colors (in polynomial time), and thus one obtains a coloring of G with $\omega(G)$ colors.

The algorithm goes as follows: For $t \geq 1$, we 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 $w = \sum_{i=1}^t \chi^{C_i} \in \mathbb{Z}^V_{\geq 0}$. As G is perfect, we know that $\alpha(G, w) = \chi(\overline{G}, w)$, which in turn is equal to t. (Indeed, $\chi(\overline{G}, w) \leq t$ follows from the definition of w. Moreover, if $y = (y_C)$ is feasible for the program defining $\chi(\overline{G}, w)$ then, on the one hand, $w^{\mathsf{T}} e = \sum_C y_C |C| \leq \sum_C y_C \omega(G)$ and, on the other hand, $w^{\mathsf{T}} e = t\omega(G)$, thus implying $t \leq \chi(\overline{G}, w)$.) Now we compute a stable set S having maximum possible weight $w(S)$. Hence, $w(S) = t$ and thus S meets each of the cliques $C_1, \cdots, C_t$.

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 $\chi^S \in L_t \setminus L_{t+1}$ for the set S constructed in the first step, and thus the dimension decreases at least by 1 at each iteration.

4.4 Other formulations of the theta number

4.4.1 Dual formulation

We now give several equivalent formulations for the theta number obtained by applying semidefinite programming duality and some further elementary manipulations.

Lemma 4.4.1. *The theta number can be expressed by any of the following programs:*

$$\vartheta(G) = \min_{t \in \mathbb{R}, A \in \mathcal{S}^n} \{t : tI + A - J \succeq 0,\ A_{ij} = 0\ (i = j \text{ or } \{i,j\} \in \overline{E})\}, \tag{4.12}$$

$$\vartheta(G) = \min_{t \in \mathbb{R}, B \in \mathcal{S}^n} \left\{t : tI - B \succeq 0,\ B_{ij} = 1\ (i = j \text{ or } \{i,j\} \in \overline{E})\right\}, \tag{4.13}$$

$$\vartheta(G) = \min_{t \in \mathbb{R}, C \in \mathcal{S}^n} \{t : C - J \succeq 0,\ C_{ii} = t\ (i \in V),\ C_{ij} = 0\ (\{i,j\} \in \overline{E})\}, \tag{4.14}$$

$$\vartheta(G) = \min_{B \in \mathcal{S}^n} \left\{\lambda_{\max}(B) : B_{ij} = 1\ (i = j \text{ or } \{i,j\} \in \overline{E})\right\}. \tag{4.15}$$

Proof. First we build the dual of the semidefinite program (4.11), which reads:

$$\min_{t\in\mathbb{R},y\in\mathbb{R}^E}\left\{t : tI + \sum_{\{i,j\}\in E} y_{ij}E_{ij} - J \succeq 0\right\}. \tag{4.16}$$

As both programs (4.11) and (4.16) are strictly feasible, there is no duality gap: the optimal value of (4.16) 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 (4.16), it follows that the program (4.16) is equivalent to (4.12), (4.13) and (4.14). Finally the formulation (4.15) follows directly from (4.13) after recalling that $\lambda_{\max}(B)$ is the smallest scalar t for which $tI - B \succeq 0$. □

4.4.2 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 4.4.2. *The theta number $\vartheta(G)$ is equal to the optimal value of the following semidefinite program:*

$$\min_{Z\in\mathcal{S}^{n+1}}\{Z_{00} : Z \succeq 0,\ Z_{0i} = Z_{ii} = 1\ (i \in V),\ Z_{ij} = 0\ (\{i,j\} \in \overline{E})\}. \tag{4.17}$$

Proof. We show that the two semidefinite programs in (4.12) and (4.17) 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 (4.12), then Z is feasible for (4.17) with same objective value: $Z_{00} = t$. The construction can be reversed: if Z is feasible for (4.17), then one can construct (t, A) feasible for (4.12) with $t = Z_{00}$. Hence both programs are equivalent. □

From the formulation (4.17), the link of the theta number to the (fractional) chromatic number is even more transparent.

Lemma 4.4.3. *For any graph G, we have that $\vartheta(G) \leq \chi^*(\overline{G})$.*

Proof. Let $y = (y_C)$ be feasible for the linear program (4.8) defining $\chi^*(\overline{G})$. For each clique C define the vector $z_C = (1\ \chi^C) \in \mathbb{R}^{1+n}$, obtained by appending an entry equal to 1 to the characteristic vector of C. Define the matrix $Z = \sum_{C \text{ clique of } G} y_C z_C z_C^{\mathsf{T}}$. One can verify that Z is feasible for the program (4.17) with objective value $Z_{00} = \sum_C y_C$ (check it). This shows $\vartheta(G) \leq \chi^*(\overline{G})$. □

Applying duality to the semidefinite program (4.17), we obtain[1] the following formulation for $\vartheta(G)$.

Lemma 4.4.4. *The theta number $\vartheta(G)$ is equal to the optimal value of the following semidefinite program:*

$$\max_{Y\in\mathcal{S}^{n+1}}\left\{\sum_{i\in V}Y_{ii} : Y\succeq 0,\ Y_{00}=1,\ Y_{0i}=Y_{ii}\ (i\in V),\ Y_{ij}=0\ (\{i,j\}\in E)\right\}. \tag{4.18}$$

Proof. First we write the program (4.17) in standard form, using the elementary matrices E_{ij} (with entries 1 at positions (i,j) and (j,i) and 0 elsewhere):

$$\inf\{\langle E_{00},Z\rangle : \langle E_{ii},Z\rangle=1,\ \langle E_{0i},Z\rangle=2\ (i\in V),\ \langle E_{ij},Z\rangle=0\ (\{i,j\}\in\overline{E}),\ Z\succeq 0\}.$$

Next we write the dual of this sdp:

$$\sup\left\{\sum_{i\in V}y_i+2z_i : Y=E_{00}-\sum_{i\in V}y_iE_{ii}+z_iE_{0i}+\sum_{\{i,j\}\in\overline{E}}u_{ij}E_{ij}\succeq 0\right\}.$$

Observe now that the matrix $Y\in\mathcal{S}^{n+1}$ occurring in this program can be equivalently characterized by the conditions: $Y_{00}=1$, $Y_{ij}=0$ if $\{i,j\}\in E$ and $Y\succeq 0$. Moreover the objective function reads: $\sum_{i\in V}y_i+2z_i=-\big(\sum_{i\in V}Y_{ii}+2Y_{0i}\big)$. Therefore the dual can be equivalently reformulated as

$$\max\left\{-\left(\sum_{i\in V}Y_{ii}+2Y_{0i}\right) : Y\succeq 0,\ Y_{00}=1,\ Y_{ij}=0\ (\{i,j\}\in E)\right\}. \tag{4.19}$$

As (4.17) is strictly feasible (check it) there is no duality gap, the optimal value of (4.19) is attained and it is equal to $\vartheta(G)$.

Let Y be an optimal solution of (4.19). We claim that $Y_{0i}+Y_{ii}=0$ for all $i\in V$. Indeed, assume that $Y_{0i}+Y_{ii}\neq 0$ for some $i\in V$, so that $Y_{ii}\neq 0$. We construct a new matrix Y' feasible for (4.19) having a larger objective value than Y, thus contradicting the optimality of Y. If $Y_{0i}\geq 0$, then we let Y' be obtained from Y by setting to 0 all the entries at the positions $(i,0)$ and (i,j) for $j\in[n]$, which has a larger objective value since $Y_{ii}+2Y_{0i}>0$. Assume now $Y_{0i}<0$. Then set $\lambda=-Y_{0i}/Y_{ii}>0$ and let Y' be obtained from Y by multiplying its i-th row and column by λ. Then, $Y'_{ii}=\lambda^2Y_{ii}=Y_{0i}^2/Y_{ii}$, $Y'_{0i}=\lambda Y_{0i}=-Y'_{ii}$, and Y' has a larger objective value than Y since $-Y'_{ii}-2Y'_{0i}=Y_{0i}^2/Y_{ii}>-Y_{ii}-2Y_{0i}$.

Therefore, we can add w.l.o.g. the condition $Y_{0i}=-Y_{ii}$ ($i\in V$) to (4.19), so that its objective function can be replaced by $\sum_{i\in V}Y_{ii}$. Finally, in order to get the program (4.18), it suffices to observe that one can change the signs on

[1] Of course there is more than one road leading to Rome: one can also show directly the equivalence of the two programs (4.11) and (4.18).

the first row and column of Y (indexed by the index 0). In this way we obtain a matrix $\tilde{Y}$ such that $\tilde{Y}_{0i} = -Y_{0i}$ for all i and $\tilde{Y}_{ij} = Y_{ij}$ at all other positions. Thus $\tilde{Y}$ now satisfies the conditions $\tilde{Y}_{ii} = \tilde{Y}_{0i}$ for $i \in V$ and it is an optimal solution of (4.18). □

4.5 Geometric properties of the theta number

In this section we introduce the theta body TH(G). This is a semidefinite relaxation of the stable set polytope ST(G) tighter that its linear relaxation QST(G), which provides another more geometric formulation for the theta number as well as geometric characterizations of perfect graphs.

4.5.1 The theta body 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 $\{0\} \cup V$:

$$\mathcal{M}_G = \{Y \in \mathcal{S}^{n+1} : Y_{00} = 1,\ Y_{0i} = Y_{ii}\ (i \in V),\ Y \succeq 0\}, \tag{4.20}$$

which is thus the feasible region of the semidefinite program (4.18). Now let TH(G) denote the convex set obtained by projecting the set $\mathcal{M}_G$ onto the subspace $\mathbb{R}^V$ of the diagonal entries:

$$\text{TH}(G) = \{x \in \mathbb{R}^V : \exists Y \in \mathcal{M}_G \text{ such that } x_i = Y_{ii}\ \forall i \in V\}, \tag{4.21}$$

called the *theta body* of G. It turns out that TH(G) is nested between ST(G) and QST(G).

Lemma 4.5.1. *For any graph G, we have that* ST$(G) \subseteq$ TH$(G) \subseteq$ QST(G).

Proof. The inclusion ST$(G) \subseteq$ TH(G) follows from the fact that the characteristic vector of any stable set S lies in TH(G). To see this, define the vector $y = (1\ \chi^S) \in \mathbb{R}^{n+1}$ obtained by adding an entry equal to 1 to the characteristic vector of S, and define the matrix $Y = yy^{\mathsf{T}} \in \mathcal{S}^{n+1}$. Then $Y \in \mathcal{M}_G$ and $\chi^S = (Y_{ii})_{i \in V}$, which shows that $\chi^S \in$ TH(G).

We now show the inclusion TH$(G) \subseteq$ QST(G). For this pick a vector $x \in$ TH(G) and a clique C of G; we show that $x(C) \leq 1$. 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

$$Y_C = \begin{pmatrix} 1 & x_C^{\mathsf{T}} \\ x_C & \text{Diag}(x_C) \end{pmatrix},$$

where we set $x_C = (x_i)_{i \in C}$. Now, $Y_C \succeq 0$ implies that Diag$(x_C) - x_C x_C^{\mathsf{T}} \succeq 0$ (taking a Schur complement). This in turn implies: $e^{\mathsf{T}}(\text{Diag}(x_C) - x_C x_C^{\mathsf{T}})e \geq 0$, which can be rewritten as $x(C) - (x(C))^2 \geq 0$, giving $x(C) \leq 1$. □

In view of Lemma 4.4.4, maximizing the all-ones objective function over TH(G) gives the theta number:

$$\vartheta(G) = \max_{x \in \mathbb{R}^V} \{e^{\mathsf{T}} x : x \in \text{TH}(G)\}.$$

As maximizing $e^{\mathsf{T}}x$ over QST(G) gives the LP bound $\alpha^*(G)$, Lemma 4.5.1 implies directly that the SDP bound $\vartheta(G)$ dominates the LP bound $\alpha^*(G)$:

Corollary 4.5.2. *For any graph G, we have that $\alpha(G) \le \vartheta(G) \le \alpha^*(G)$.*

Combining the inclusion from Lemma 4.5.1 with Theorem 4.2.4, we deduce that TH(G) = ST(G) = QST(G) for perfect graphs. As we will see in Theorem 4.5.9 below it turns out that these equalities characterize perfect graphs.

4.5.2 Orthonormal representations of graphs

We introduce orthonormal representations of a graph G, which will be used in the next section to give further geometric descriptrions of the theta body TH(G).

Definition 4.5.3. *An* orthonormal representation *of G, abbreviated as* ONR*, consists of a set of unit vectors $\{u_1, \dots, u_n\} \subseteq \mathbb{R}^d$ (for some $d \ge 1$) satisfying*

$$u_i^{\mathsf{T}} u_j = 0 \;\; \forall \{i,j\} \in \overline{E}.$$

Note that the smallest integer d for which there exists an orthonormal representation of G is upper bounded by $\chi(\overline{G})$ (check it). Moreover, if C is a clique in G and the u_i's form an ONR of G of dimension d, then the vectors u_i labeling the nodes of C are pairwise orthogonal, which implies that $d \ge \omega(G)$. It turns out that the stronger lower bound: $d \ge \vartheta(G)$ holds.

Lemma 4.5.4. *The minimum dimension d of an orthonormal representation of a graph G satisfies: $\vartheta(G) \le d$.*

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 (4.17) defining $\vartheta(G)$ (check it) with $Z_{00} = d$. This gives $\vartheta(G) \le d$. □

4.5.3 Geometric properties of the theta body

There is a beautiful relationship between the theta bodies of a graph G and of its complementary graph $\overline{G}$:

Theorem 4.5.5. *For any graph G,*

$$\text{TH}(G) = \{x \in \mathbb{R}^V_{\ge 0} : x^T z \le 1 \; \forall z \in \text{TH}(\overline{G})\}.$$

In other words, we know an explicit linear inequality description of TH(G); moreover, the normal vectors to the supporting hyperplanes of TH(G) are precisely the elements of TH($\overline{G}$). One inclusion is easy:

Lemma 4.5.6. *If $x \in \mathrm{TH}(G)$ and $z \in \mathrm{TH}(\overline{G})$ then $x^{\mathsf{T}}z \leq 1$.*

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 $\langle Y, Z' \rangle \geq 0$ as $Y, Z' \succeq 0$. Moreover, $\langle Y, Z' \rangle = 1 - x^{\mathsf{T}}z$ (check it), thus giving $x^{\mathsf{T}}z \leq 1$. □

Next we observe how the elements of $\mathrm{TH}(G)$ can be expressed in terms of orthonormal representations of $\overline{G}$.

Lemma 4.5.7. *For $x \in \mathbb{R}^V_{\geq 0}$, $x \in \mathrm{TH}(G)$ 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 4.5.5 we use the following result, which characterizes which partially specified matrices can be completed to a positive semidefinite matrix – this will be proved in Exercise 6.1.

Proposition 4.5.8. *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*

$$K_a = \{Y \in \mathcal{S}^W : Y \succeq 0,\ Y_{ij} = a_{ij}\ \forall i = j \in W \text{ and } \{i,j\} \in F\} \tag{4.22}$$

(consisting of all possible positive semidefinite completions of a) and the cone

$$\mathcal{C}_H = \{Z \in \mathcal{S}^W : Z \succeq 0,\ Z_{ij} = 0\ \forall \{i,j\} \in \overline{F}\} \tag{4.23}$$

(consisting of all positive semidefinite matrices supported by the graph H). Then, $K_a \neq \emptyset$ if and only if

$$\sum_{i \in W} a_{ii}Z_{ii} + 2 \sum_{\{i,j\} \in F} a_{ij}Z_{ij} \geq 0\ \ \forall Z \in \mathcal{C}_H. \tag{4.24}$$

Proof. (*of Theorem 4.5.5*). Let $x \in \mathbb{R}^V_{\geq 0}$ such that $x^{\mathsf{T}}z \leq 1$ 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 4.5.8, where the graph H is G with an additional node 0 adjacent to all $i \in V$. Hence it suffices now to show that $\langle Y, Z\rangle \geq 0$ for all $Z \in \mathcal{S}_{\succeq 0}^{\{0\}\cup V}$ 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 4.5.7 (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 TH($\overline{G}$) and thus $x^{\mathsf{T}} z \leq 1$ by assumption. Therefore, $\langle Y, Z\rangle$ is equal to

$$1 + 2\sum_{i\in V} x_i w_0^{\mathsf{T}} w_i + \sum_{i\in V} x_i \|w_i\|^2 \geq \sum_{i\in V} x_i \left(\frac{(w_0^{\mathsf{T}} w_i)^2}{\|w_i\|^2} + 2 w_0^{\mathsf{T}} w_i + \|w_i\|^2 \right)$$

$$= \sum_{i\in V} x_i \left(\frac{w_0^{\mathsf{T}} w_i}{\|w_i\|} + \|w_i\| \right)^2 \geq 0.$$

□

4.5.4 Geometric characterizations of perfect graphs

We can now prove the following geometric characterization of perfect graphs, which strengthens the polyhedral characterization of Theorem 4.2.4.

Theorem 4.5.9. *For any graph G the following assertions are equivalent.*

1. *G is perfect.*
2. TH(G) = ST(G)
3. TH(G) = QST(G).
4. TH(G) *is a polytope.*

We start with the following observations which will be useful for the proof. Recall that the *antiblocker* of a set $P \subseteq \mathbb{R}^n_{\geq 0}$ is defined as

$$\text{abl}(P) = \{y \in \mathbb{R}^n_{\geq 0} : y^{\mathsf{T}} x \leq 1 \ \forall x \in P\}.$$

We will use the following property, shown in Exercise 1.6: If $P \subseteq \mathbb{R}^n_{\geq 0}$ is a down-monotone polytope in $\mathbb{R}^n_{\geq 0}$ then $P = \text{abl}(\text{abl}(P))$.

Using this notion of antiblocker, we see that Theorem 4.5.5 shows that TH(G) is the antiblocker of TH($\overline{G}$): TH(G) = abl(TH($\overline{G}$)) and, analogously, TH($\overline{G}$) = abl(TH(G)). Moreover, by its definition, QST(G) is the antiblocker of ST($\overline{G}$): QST(G) = abl(ST($\overline{G}$)). This implies the equalities

$$\text{abl}(\text{QST}(G)) = \text{abl}(\text{abl}(\text{ST}(\overline{G}))) = \text{ST}(\overline{G})$$

and thus

$$\text{ST}(G) = \text{abl}(\text{QST}(\overline{G})). \tag{4.25}$$

We now show that if TH(G) is a polytope then it coincides with QST(G), which is the main ingredient in the proof of Theorem 4.5.9. As any polytope is equal to the solution set of its facet defining inequalities, it suffices to show that the only inequalities that define facets of TH(G) are the nonnegativity conditions and the clique inequalities.

Lemma 4.5.10. *Let $a \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$. If the inequality $a^\mathsf{T} x \le \alpha$ defines a facet of* TH(G) *then it is a multiple of a nonnegativity condition $x_i \ge 0$ for some $i \in V$ or of a clique inequality $x(C) \le 1$ for some clique C of G.*

Proof. Let $F = \{x \in \text{TH}(G) : a^\mathsf{T} x = \alpha\}$ be the facet of TH(G) defined by the inequality $a^\mathsf{T} x \le \alpha$. Pick a point z in the relative interior of F, thus z lies on the boundary of TH(G). We use the description of TH(G) from Theorem 4.5.5. If $z_i = 0$ for some $i \in V$, then the inequality $a^\mathsf{T} x \le \alpha$ is equivalent to the nonnegativity condition $x_i \ge 0$. Suppose now that $z^\mathsf{T} y = 1$ for some $y \in \text{TH}(\overline{G})$. In view of Lemma 4.5.7, $y = ((c^\mathsf{T} u_i)^2)_{i=1}^n$ for some unit vectors $c, u_1, \ldots, u_n \in \mathbb{R}^k$ forming an orthonormal representation of G, i.e., satisfying $u_i^\mathsf{T} u_j = 0$ for $\{i,j\} \in \overline{E}$. Then the inequality $a^\mathsf{T} x \le \alpha$ is equivalent to $\sum_{i=1}^n (c^\mathsf{T} u_i)^2 x_i \le 1$, i.e., up to scaling we may assume that $\alpha = 1$ and $a_i = (c^\mathsf{T} u_i)^2$ for all $i \in V$. We claim that

$$c = \sum_{i=1}^n (c^\mathsf{T} u_i) x_i u_i \quad \text{for all } x \in F. \tag{4.26}$$

Indeed, for any unit vector $d \in \mathbb{R}^k$, the vector $((d^\mathsf{T} u_i)^2)_{i=1}^n$ belongs to TH$(\overline{G})$ and thus $\sum_{i=1}^n (d_i^u)^2 x_i \le 1$ for all $x \in F$. In other words, the maximum of the quadratic form $d^\mathsf{T}(\sum_{i=1}^n x_i u_i u_i^\mathsf{T})d$ taken over all unit vectors $d \in \mathbb{R}^k$ is equal to 1 and is attained at $d = c$. This shows that c is an eigenvector of the matrix $\sum_{i=1}^n x_i u_i u_i^\mathsf{T}$ for the eigenvalue 1, and thus equality $(\sum_{i=1}^n x_i u_i u_i^\mathsf{T})c = c$ holds, which gives (4.26).

From (4.26) we deduce that each equation $\sum_{i=1}^n (u_i^\mathsf{T} c) x_i (u_i)_j = c_j$ is a scalar multiple of $\sum_{i=1}^n (u_i^\mathsf{T} c)^2 = 1$. This implies that $(u_i^\mathsf{T} c)(u_i)_j = c_j (u_i^\mathsf{T} c)^2$ for all $i, j \in [n]$. Hence, $u_i^\mathsf{T} c \ne 0$ implies $u_i = (u_i^\mathsf{T} c)c$, thus $u_i = \pm c$ (since u_i and c are both unit vectors) and without loss of generality $u_i = c$. Set $C = \{i \in V : u_i = c\}$, so that the inequality $\sum_{i=1}^n (c^\mathsf{T} u_i)^2 x_i \le 1$ reads $\sum_{i \in C} x_i \le 1$. Finally we now observe that C is a clique in G, since $i \ne j \in C$ implies $u_i^\mathsf{T} u_j = c^\mathsf{T} c = 1$ and thus $\{i,j\} \in E$. This concludes the proof. □

Proof. (of Theorem 4.5.9). By Theorem 4.2.4 we know that G is perfect if and only if QST(G) = ST(G). Moreover, by Lemma 4.5.1, we have the inclusion ST$(G) \subseteq$ TH$(G) \subseteq$ QST(G). Hence, in order to show the theorem it suffices to show that G is perfect if and only if TH(G) is a polytope.

There remains only to show the 'if' part. Assume that TH(G) is a polytope. Then TH$(\overline{G})$ too is a polytope since TH$(\overline{G})$ = abl(TH(G)). Therefore, by Lemma 4.5.10 (applied to $\overline{G}$), TH$(\overline{G})$ = QST$(\overline{G})$. Taking the antiblocker of both sides (and using (4.25)), we obtain that TH(G) = abl(TH$(\overline{G})$) = abl(QST$(\overline{G})$) = ST(G). □

4.6 Bounding the Shannon capacity

The theta number was introduced by Lovász [10] 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 4.6.1. (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, $\alpha(G \cdot H) \geq \alpha(G)\alpha(H)$. Let G^k denote the strong product of k copies of G. For any integers $k, m \in \mathbb{N}$ we have that

$$\alpha(G^{k+m}) \geq \alpha(G^k)\alpha(G^m)$$

and thus $\alpha(G^k) \geq (\alpha(G))^k$. Consider the parameter

$$\Theta(G) := \sup_{k\geq 1} \sqrt[k]{\alpha(G^k)}, \tag{4.27}$$

called the *Shannon capacity* of the graph G. Using Fekete's lemma[2] one can verify that $\Theta(G) = \lim_{k\to\infty} \sqrt[k]{\alpha(G^k)}$.

The parameter $\Theta(G)$ was introduced by Shannon in 1956. 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 $\alpha((C_5)^2) \geq 5$. 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 $\Theta(C_5) \geq \sqrt{5}$.

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 [10] showed how to use the theta number in order to upper bound the Shannon capacity: Lovász showed

[2] Consider a sequence $(a_k)_k$ of positive real numbers satisfying: $a_{k+m} \geq a_k + a_m$ for $k, m \in \mathbb{N}$. Fekete's lemma claims that $\lim_{k\to\infty} a_k/k = \sup_{k\in\mathbb{N}} a_k/k$. Then apply Fekete's lemma to the sequence $a_k = \log \alpha(G^k)$.

that $\Theta(G) \leq \vartheta(G)$ 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 4.7.6), the exact value of the Shannon capacity of C_{2n+1} is not known, already for C_7.

Theorem 4.6.2. *For any graph G, we have that $\Theta(G) \leq \vartheta(G)$.*

The proof is based on the multiplicative property of the theta number from Lemma 4.6.3 – 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, $\alpha(G^k) \leq \vartheta(G^k) = (\vartheta(G))^k$ implies $\sqrt[k]{\alpha(G^k)} \leq \vartheta(G)$ and thus $\Theta(G) \leq \vartheta(G)$.

Lemma 4.6.3. *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 C_5: $\Theta(C_5) = \sqrt{5}$. Indeed, $\Theta(C_5) \geq \sqrt{\alpha(C_5^2)} \geq \sqrt{5}$ and $\Theta(C_5) \leq \vartheta(C_5)$, with $\vartheta(C_5) = \sqrt{5}$ as we will see below in relation (4.29).

4.7 The theta number for vertex-transitive graphs

The following inequalities relate the stability number and the (fractional) coloring number of a graph:

$$|V| \leq \alpha(G)\chi^*(G) \leq \alpha(G)\chi(G).$$

(Check it.) First we mention the following analogous inequality relating the theta numbers of G and its complement $\overline{G}$.

Proposition 4.7.1. *For any graph $G = (V, E)$, we have that $\vartheta(G)\vartheta(\overline{G}) \geq |V|$.*

Proof. Using the formulation of the theta number from (4.14), 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 $\{i,j\} \in E$. Combining the inequalities $\langle C - J, J\rangle \geq 0$, $\langle C' - J, J\rangle \geq 0$ and $\langle C - J, C' - J\rangle \geq 0$ with the identity $\langle C, C'\rangle = n\vartheta(G)\vartheta(\overline{G})$, we get 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., $\{i, j\} \in E$ if and only if $\{\sigma(i), \sigma(j)\} \in E$. Then the set 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 $\sigma \in$ 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 (i,j)th entry $P_\sigma(i,j) = 1$ if $j = \sigma(i)$ and 0 otherwise), 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 (4.11) defining the theta number $\vartheta(G)$. This is an easy, but very useful fact.It follows from the fact that the matrices entering the program (4.11) (the all-ones matrix J, the identity I and the elementary matrices E_{ij} for edges $\{i,j\} \in E$) are invariant under action of Aut(G).

Lemma 4.7.2. *If X is feasible for the program (4.11) and σ is an automorphism of G, then $\sigma(X)$ is again feasible for (4.11), moreover with the same objective value as X.*

Proof. Directly from the fact that $\langle J, \sigma(X)\rangle = \langle J, X\rangle$, Tr$(\sigma(X))$ = 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 4.7.3. *The program (4.11) has an optimal solution X^* which is invariant under action of the automorphism group of G, i.e., satisfies $\sigma(X^*) = X^*$ for all $\sigma \in$ Aut(G).*

Proof. Let X be an optimal solution of (4.11). By Lemma 4.7.2, $\sigma(X)$ is again an optimal solution for each $\sigma \in$ Aut(G). Define the matrix

$$X^* = \frac{1}{|\text{Aut}(G)|} \sum_{\sigma\in\text{Aut}(G)} \sigma(X),$$

obtained by averaging over all matrices $\sigma(X)$ for $\sigma \in$ Aut(G). As the set of optimal solutions of (4.11) is convex, X^* is still an optimal solution of (4.11). Moreover, by construction, X^* is invariant under action of Aut(G). □

Corollary 4.7.4. *If G is a vertex-transitive graph then the program (4.11) 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 4.7.3, there is an optimal solution X^* which is invariant under action of Aut(G). As G is vertex-transitive, all diagonal entries of X^* are equal. Indeed, let $i,j \in V$ and $\sigma \in$ Aut(G) such that $\sigma(i) = j$. Then, $X^*_{jj} = X^*_{\sigma(i)\sigma(i)} = X^*_{ii}$. As Tr$(X^*) = 1$ we must have $X^*_{ii} = 1/n$ for all i. Moreover, $\sum_{k\in V} X^*_{jk} = \sum_{k\in V} X^*_{\sigma(i)k} = \sum_{h\in V} X^*_{\sigma(i)\sigma(h)} = \sum_{h\in V} X^*_{ih}$, which shows that $X^*e = \lambda e$ for some scalar λ. Combining with the condition $\langle J, X^*\rangle = \vartheta(G)$ we obtain that $\lambda = \frac{\vartheta(G)}{n}$. □

Proposition 4.7.5. *If G is a vertex-transitive graph, then $\vartheta(G)\vartheta(\overline{G}) = |V|$.*

Proof. By Corollary 4.7.4, there is an optimal solution X^* of the program (4.11) 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 (4.14) defining $\vartheta(\overline{G})$, which implies $\vartheta(\overline{G}) \leq n/\vartheta(G)$. Combining with Proposition 4.7.1 we get the equality $\vartheta(G)\vartheta(\overline{G}) = |V|$. □

For instance, the cycle C_n is vertex-transitive, so that

$$\vartheta(C_n)\vartheta(\overline{C_n}) = n. \tag{4.28}$$

In particular, as C_5 is isomorphic to $\overline{C_5}$, we deduce that

$$\vartheta(C_5) = \sqrt{5}. \tag{4.29}$$

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 4.7.6. *For any odd $n \geq 3$,*

$$\vartheta(C_n) = \frac{n\cos(\pi/n)}{1+\cos(\pi/n)} \quad \textit{and} \quad \vartheta(\overline{C_n}) = \frac{1+\cos(\pi/n)}{\cos(\pi/n)}.$$

Proof. As $\vartheta(C_n)\vartheta(\overline{C_n}) = n$, it suffices to compute $\vartheta(C_n)$. We use the formulation (4.15). 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

$$n + 2x \ \text{ and } x(\omega^k + \omega^{-k}) \ \text{ for } k = 1, \cdots, n-1. \tag{4.30}$$

We minimize the maximum of the values in (4.30) when choosing x such that

$$n + 2x = -2x\cos(\pi/n)$$

(check it). This gives $\vartheta(C_n) = \lambda_{\max}(B) = -2x\cos(\pi/n) = \frac{n\cos(\pi/n)}{1+\cos(\pi/n)}$. □

As another application, one can compute the Shannon capacity of any graph G which is vertex-transitive and self-complementary (like C_5).

Theorem 4.7.7. *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 $\Theta(G \cdot \overline{G}) \geq \alpha(G \cdot \overline{G}) \geq |V|$, since the set of diagonal pairs $\{(i,i) : i \in V\}$ is stable in $G \cdot \overline{G}$. The reverse inequality follows from Proposition 4.7.5 combined with Lemma 4.6.3: $\Theta(G \cdot \overline{G}) \leq \vartheta(G \cdot \overline{G}) = \vartheta(G)\vartheta(\overline{G}) = |V|$. Therefore, $\Theta(G \cdot \overline{G}) = |V|$.

If moreover G is isomorphic to $\overline{G}$ then $\vartheta(G) = \sqrt{|V|}$ and thus $\Theta(G) \leq \vartheta(G) = \sqrt{|V|}$. On the other hand, $|V| = \Theta(G \cdot \overline{G}) = \Theta(G^2) \leq (\Theta(G))^2$ (check it), which implies: $\Theta(G) \geq \sqrt{|V|}$ and thus equality: $\Theta(G) = \sqrt{|V|}$. □

4.8 Application to Hamming graphs: Delsarte LP bound for codes

A *binary code of length* n is a subset C of the set $V = \{0,1\}^n$ of binary sequences (aka *words*) of length n. Given two words $u, v \in V$, their *Hamming distance* $d_H(u,v)$ is the number of positions $i \in [n]$ such that $u_i \neq v_i$. The *Hamming weight* $|u|$ of a word $u \in V$ is its number of nonzero coordinates: $|u| = d_H(u,0)$.

Given an integer $d \in [n]$, one says that C has *minimum distance* d if any two distinct words of C have Hamming distance at least d. A fundamental problem in coding theory is to compute the maximum cardinality $A(n,d)$ of a code of length n with minimum distance d. This is the maximum number of messages of length n that can correctly be decoded if after transmission at most $(d-1)/2$ bits can be erroneously transmitted in each word of C.

Computing $A(n,d)$ is in fact an instance of the maximum stable set problem. Indeed, let $G(n,d)$ denote the graph with vertex set $V = \{0,1\}^n$ and with an edge $\{u,v\}$ if $d_H(u,v) \leq d-1$, called a *Hamming graph*. Then, a code $C \subseteq V$ has minimum distance d if and only if C is a stable set in $G(n,d)$ and thus $A(n,d) = \alpha(G(n,d))$.

A natural idea for getting an upper bound for $A(n,d)$ is to use the theta number $\vartheta(G)$, or its strengthening $\vartheta'(G)$ obtained by adding nonnegativity conditions on the entries of the matrix variable:

$$\vartheta'(G) = \{\max\langle J, X\rangle : \text{Tr}(X) = 1,\ X_{uv} = 0\ (\{u,v\} \in E),\ X \geq 0,\ X \succeq 0\}. \tag{4.31}$$

Computing the paramater $\vartheta(G(n,d))$ or $\vartheta'(G(n,d))$ is apparently a difficult problem. Indeed the graph $G(n,d)$ has 2^n vertices and thus the matrix X in the above semidefinite program has size 2^n. However, using the fact that the Hamming graph has a large automorphism group, one can simplify the above semidefinite program and in fact reformulate it as an equivalent linear program with only $n+1$ variables and linear constraints. This is thus an enormous gain in complexity, thanks to which one can compute the parameter $\vartheta'(G(n,d))$ for large values of n.

In a nutshell this symmetry reduction is possible because the symmetric matrices that are invariant under action of the automorphism group of the Hamming graph form a commutative algebra, so that they can all be diagonalized simultaneously, by the same orthogonal basis. In what follows we explain how to derive the linear program equivalent to (4.31).

Automorphisms of the Hamming graph.

Any permutation π of $[n]$ induces an automorphism of $G(n,d)$ by setting

$$\sigma(v) = (v_{\sigma(1)}, \ldots, v_{\sigma(n)}) \in \{0,1\}^n \text{ for } v \in \{0,1\}^n.$$

Moreover, any $a \in \{0,1\}^n$ induces an automorphism s_a of $G(n,d)$ by setting

$$s_a(v) = v \oplus a \text{ for } v \in \{0,1\}^n.$$

Here we use addition moduo 2 in $\{0,1\}^n$: $u \oplus v = (u_i \oplus v_i)_{i=1}^n$, setting $0 \oplus 0 = 1 \oplus 1 = 0$ and $1 \oplus 0 = 0 \oplus 1 = 1$. Thus $d_H(u,v) = |u \oplus v|$. The set

$$\mathcal{G}_n = \{\pi s_a : \pi \in \text{Sym}(n),\ a \in \{0,1\}^n\}$$

is a group (check it), which is contained in the automorphism group of $G(n,d)$. Note, for instance, that $\pi s_a = s_{\pi(a)}\pi$ (check it).

The graph $G(n,d)$ is vertex-transitive under action of $\mathcal{G}_n$ (since, for any two vertices $u,v \in V$, the map $s_{u\oplus v}$ maps u to v). Moreover, given words $u,v,u',v' \in V$, there exists $\sigma \in \mathcal{G}_n$ such that $\sigma(u) = u'$ and $\sigma(v) = v'$ if and only if $d_H(u,v) = d_H(u',v')$ (check it).

Invariant matrices under action of $\mathcal{G}_n$.

Let $\mathcal{B}_n$ denote the set of matrices indexed by $\{0,1\}^n$ that are invariant under action of $\mathcal{G}_n$. That is, $X \in \mathcal{B}_n$ if it satisfies: $X(u,v) = X(\sigma(u),\sigma(v))$ for all $u,v \in V$ and $\sigma \in \mathcal{G}_n$ or, equivalently, if each entry $X(u,v)$ depends only on the value of the Hamming distance $d_H(u,v)$. For $k \in \{0,1,\ldots,n\}$ let M_k denote the matrix indexed by V with entries $M_k(u,v) = 1$ if $d_H(u,v) = k$ and $M_k(u,v) = 0$ otherwise. Then the matrices $M_0, M_1, \ldots, M_n$ form a basis of the vector space $\mathcal{B}_n$, and $\mathcal{B}_n$ has dimension $n+1$. Moreover, $\mathcal{B}_n$ is a commutative algebra (this will be clear from Lemma 4.8.1), known as the *Bose-Mesner algebra*. It will be convenient to use another basis in order to describe positive semidefinite matrices in $\mathcal{B}_n$.

Given $a \in V = \{0,1\}^n$, define the vector $C_a \in \{\pm 1\}^V$ defined by

$$C_a = ((-1)^{a^\mathsf{T} v})_{v\in V} \quad \text{for } a \in V. \tag{4.32}$$

Next define the $V \times V$ matrices $B_0, B_1, \ldots, B_n$ by

$$B_k = \sum_{a\in V:|a|=k} C_a C_a^\mathsf{T} \quad \text{for } k = 0,1,\ldots,n. \tag{4.33}$$

Lemma 4.8.1. *(i) The vectors C_a ($a \in V$) are pairwise orthogonal.*

(ii) The matrices $B_0, B_1, \ldots, B_n$ are pairwise orthogonal: $B_h B_k = 2^n B_k \delta_{h,k}$.

(iii) $B_0 = J$, $\text{Tr}(B_k) = 2^n \binom{n}{k}$ for $0 \le k \le n$, and $\langle J, B_k\rangle = 0$ for $1 \le k \le n$.

(iv) For any k and $u,v \in V$, $B_k(u,v) = P_n^k(d_H(u,v))$, where $P_n^k(t)$ is the Krawtchouk polynomial *which, at any integer $t = 0,1,\ldots,n$, is given by*

$$P_n^k(t) = \sum_{i=0}^{k} (-1)^i \binom{t}{i}\binom{n-t}{k-i}. \tag{4.34}$$

(v) The set $\{B_0,\ldots,B_n\}$ is a basis of $\mathcal{B}_n$ and $\mathcal{B}_n$ is a commutative algebra.

Proof. (i) is direct verification (check it) and then (ii),(iii) follow easily.

(iv) Set $t = d_H(u,v)$ and $Z = \{i \in [n] : u_i \neq v_i\}$ with $t = |Z|$. Moreover for each $a \in V$ define the set $A = \{i \in [n] : a_i = 1\}$. Then, we find that $B_k(u,v) = \sum_{A\subseteq[n]:|A|=k}(-1)^{|A\cap Z|} = \sum_{i=0}^{t}(-1)^i\binom{t}{i}\binom{n-t}{k-i} = P_n^k(t)$.

(v) follows from (ii) and (iv). □

Note that the vectors $\{C_a : a \in V = \{0,1\}^n\}$ form an orthogonal basis of $\mathbb{R}^V$, and they are the common eigenvectors to all matrices B_k and thus to all matrices in the Bose-Mesner algebra $\mathcal{B}_n$.

Lemma 4.8.2. *Let* $X = \sum_{k=0}^{n} x_k B_k \in \mathcal{B}_n$ *Then,* $X \succeq 0 \Longleftrightarrow x_0, x_1, \ldots, x_n \geq 0.$

Proof. The claim follows from the fact that the B_k's are positive semidefinite and pairwise orthogonal. Indeed, $X \succeq 0$ if all x_k's are nonnegative. Conversely, if $X \succeq 0$ then $0 \leq \langle X, B_k\rangle = x_k\langle B_k, B_k\rangle$, implying $x_k \geq 0$. □

Delsarte linear programming bound for $A(n,d)$

Using the above facts we can now formulate the parameter $\vartheta'(G(n,d))$ as the optimum value of a linear program. This linear program (4.35) provides an upper bound for $A(n,d)$, which was first discovered by Delsarte [3]; that this bound coincides with the theta number $\vartheta'(G(n,d))$ was proved by Schrijver [12].

Theorem 4.8.3. *(Delsarte LP bound for* $A(n,d) = \alpha(G(n,d))$*) The parameter* $\vartheta'(G(n,d))$ *can be computed with the following linear program:*

$$\begin{array}{lll} \max_{x_0,\ldots,x_n\in\mathbb{R}} \ 2^{2n}x_0 \ \ \textit{s.t.} & \sum_{k=0}^{n} x_k\binom{n}{k} = 2^{-n}, & \\ & \sum_{k=0}^{n} x_k P_n^k(t) = 0 & \textit{for } t = 1,\ldots,d-1, \\ & \sum_{k=0}^{n} x_k P_n^k(t) \geq 0 & \textit{for } t = d,\ldots,n, \\ & x_k \geq 0 & \textit{for } k = 0,1,\ldots,n, \end{array} \tag{4.35}$$

where $P_n^k(t)$ *is the Krawtchouk poynomial in (4.34).*

Proof. In the formulation (4.31) of $\vartheta'(G)$ we may assume that the variable X belongs to $\mathcal{B}_n$, i.e., $X = \sum_{k=0}^{n} x_k B_k$ for some scalars $x_0,\ldots,x_n \in \mathbb{R}$. It now suffices to rewrite the constraints on X as constraints on the x_k's. Using Lemma 4.8.1, we find: $\langle J, X\rangle = 2^{2n}x_0$, $1 = \mathrm{Tr}(X) = \sum_{k=0}^{n} 2^n\binom{n}{k}x_k$, and $X(u,v) = \sum_{k=0}^{n} x_k P_n^k(t)$ if $d_H(u,v) = t$. Finally the condition $X \succeq 0$ gives $x \geq 0$. □

4.9 Lasserre hierarchy of semidefinite bounds

A first easy way of getting a stronger bound toward $\alpha(G)$ is by adding nonnegativity constraints to the formulation of $\vartheta(G)$. In this way we get the parameter $\vartheta'(G)$ in (4.31), which satisfies $\alpha(G) \leq \vartheta'(G) \leq \vartheta(G)$.

There is a more systematic way of constructing stronger and stronger bounds for $\alpha(G)$. The idea is to start from the formulation of $\vartheta(G)$ from (4.18) and to observe that the matrix variable is indexed by all nodes together with an additional index 0. More generally, we can define a hierarchy of upper bounds for $\alpha(G)$ that are obtained by optimizing over a matrix variable indexed by all products of at most t (distinct) variables, for increasing values of t.

The idea is simple and consists of 'lifting' the problem into higher dimension by adding new variables. Given a set $S \subseteq V$, let $x = \chi^S \in \{0,1\}^n$ denote its characteristic vector and, for $t \in [n]$, define the vector

$$[x]_t = (1, x_1, \ldots, x_t, x_1x_2, \ldots, x_{n-1}x_n, \ldots, x_1x_2\cdots x_t, \ldots, x_{n-t+1}\cdots x_n) \in \mathbb{R}^{\mathcal{P}_t(n)}$$

consisting of all products of at most t distinct x_i's (listed in sone order). Here we let $\mathcal{P}_t(n)$ denote the collection of all subsets $I \subseteq [n]$ with $|I| \leq t$. For instance, $[x]_1 = (1, x_1, \ldots, x_n)$ and $[x]_n$ contains all 2^n possible products of distinct x_i's.

Next we consider the matrix $Y = [x]_t[x]_t^{\mathsf{T}}$ which, by construction, is positive semidefinite and satisfies the following linear conditions:

$$Y(I,J) = Y(I',J') \text{ if } I \cup J = I' \cup J'$$

and $Y_{\emptyset,\emptyset} = 1$. This motivates the following definition.

Definition 4.9.1. *Given an integer $0 \leq t \leq n$ and a vector $y = (y_I) \in \mathbb{R}^{\mathcal{P}_{2t}(n)}$, let $M_t(y)$ denote the symmetric matrix indexed by $\mathcal{P}_t(n)$, with (I,J)th entry $y_{I\cup J}$ for $I, J \in \mathcal{P}_t(n)$. $M_t(y)$ is called the* moment matrix of order t *of y.*

Example 4.9.2. *As an example, for $n = 2$, the matrices $M_1(y)$ and $M_2(y)$ have the form*

$$M_1(y) = \begin{array}{c|ccc} & \emptyset & 1 & 2 \\ \hline \emptyset & y_\emptyset & y_1 & y_2 \\ 1 & y_1 & y_1 & y_{12} \\ 2 & y_2 & y_{12} & y_2 \end{array}, \quad M_2(y) = \begin{array}{c|cccc} & \emptyset & 1 & 2 & 12 \\ \hline \emptyset & y_\emptyset & y_1 & y_2 & y_{12} \\ 1 & y_1 & y_1 & y_{12} & y_{12} \\ 2 & y_2 & y_{12} & y_2 & y_{12} \\ 12 & y_{12} & y_{12} & y_{12} & y_{12} \end{array}$$

Note that $M_1(y)$ corresponds to the matrix variable in the formulation (4.18) of $\vartheta(G)$. Moreover, $M_1(y)$ occurs as a principal submatrix of $M_2(y)$.

We can now formulate new upper bounds for the stability number.

Definition 4.9.3. *For any integer $1 \leq t \leq n$, define the parameter*

$$\text{las}_t(G) = \max_{y \in \mathbb{R}^{\mathcal{P}_{2t}(n)}} \{\sum_{i=1}^n y_i : y_\emptyset = 1,\ y_{ij} = 0 \ (\{i,j\} \in E),\ M_t(y) \succeq 0\}, \quad (4.36)$$

known as the Lasserre bound of order t.

Lemma 4.9.4. *For each $1 \leq t \leq n$, we have that $\alpha(G) \leq \text{las}_t(G)$. Moreover, $\text{las}_{t+1}(G) \leq \text{las}_t(G)$.*

Proof. Let $x = \chi^S$ where S is a stable set of G, and let $y = [x]_t$. Then the moment matrix $M_t(y)$ is feasible for the program (4.36) with value $\sum_{i=1}^n y_i = |S|$, which shows $|S| \leq \mathrm{las}_t(G)$ and thus $\alpha(G) \leq \mathrm{las}_t(G)$.

The inequality $\mathrm{las}_{t+1}(G) \leq \mathrm{las}_t(G)$ follows from the fact that $M_t(y)$ occurs as a principal submatrix of $M_{t+1}(y)$. □

Some further observations:
• For $t = 1$, the Lasserrre bound is simply the theta number: $\mathrm{las}_1(G) = \vartheta(G)$.
• For $t = 2$, the Lasserre bound improves $\vartheta'(G)$: $\mathrm{las}_2(G) \leq \vartheta'(G)$. This is because the condition $M_2(y) \succeq 0$ implies that all entries y_{ij} are nonnegative (as y_{ij} occurs a diagonal entry of $M_2(y)$).
• The bounds form a hierarchy of stronger and stronger bounds:

$$\alpha(G) \leq \mathrm{las}_n(G) \leq \ldots \leq \mathrm{las}_t(G) \leq \ldots \leq \mathrm{las}_2(G) \leq \mathrm{las}_1(G) = \vartheta(G).$$

It turns out that, at order $t = \alpha(G)$, the Lasserre bound is exact: $\mathrm{las}_t(G) = \alpha(G)$.

Theorem 4.9.5. *For any graph G, $\mathrm{las}_t(G) = \alpha(G)$ for any $t \geq \alpha(G)$.*

In the rest of the section we prove this result.

Characterizing positive semidefinite full moment matrices $M_n(y)$.

In a first step we characterize the vectors $y = (y_I)_{I \subseteq [n]}$ whose moment matrix $M_n(y)$ is positive semidefinite.

For this we use the $2^n \times 2^n$ matrix Z_n, whose columns are the vectors $[x]_n$ for $x \in \{0,1\}^n$. Alternatively, Z_n is the matrix indexed by $\mathcal{P}_n(n)$, with entries $Z_n(I,J) = 1$ if $I \subseteq J$ and $Z_n(I,J) = 0$ otherwise. Its inverse matrix Z_n^1 is defined by $Z_n^{-1}(I,J) = (-1)^{|J \setminus I|}$ if $I \subseteq J$ and 0 otherwise. (Check it). The matrix Z_n is known as the *Zeta matrix* of the lattice $\mathcal{P}_n(n)$ (all subsets of the set $[n]$, ordered by set inclusion) and its inverse Z_n^{-1} as its Möbius matrix (cf., e.g., [11]).

Example 4.9.6. *For $n = 2$ we have:*

$$Z_2 = \begin{array}{c} \\ \emptyset \\ 1 \\ 2 \\ 12 \end{array} \!\!\begin{array}{c} \begin{array}{cccc} \emptyset & 1 & 2 & 12 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}, \quad Z_2^{-1} = \begin{array}{c} \\ \emptyset \\ 1 \\ 2 \\ 12 \end{array} \!\!\begin{array}{c} \begin{array}{cccc} \emptyset & 1 & 2 & 12 \end{array} \\ \begin{pmatrix} 1 & -1 & -1 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Lemma 4.9.7. *Let $y \in \mathbb{R}^{\mathcal{P}_n(n)}$ and set $\lambda = Z_n^{-1} y$. Then,*

$$M_n(y) = Z_n \mathrm{Diag}(\lambda) Z_n^{\mathsf{T}}.$$

Proof. Pick $I, J \subseteq [n]$. We show that the (I,J)th entry of $Z_n \mathrm{Diag}(\lambda) Z_n^{\mathsf{T}}$ is equal to $y_{I \cup J}$. This is direct verification:

$$(Z_n \mathrm{Diag}(\lambda) Z_n^{\mathsf{T}})_{I,J} = \sum_{K : I \subseteq K} (\mathrm{Diag}(\lambda) Z_n^{\mathsf{T}})_{K,J} = \sum_{K : I \cup J \subseteq K} (Z_n^{-1} y)_K$$

$$= \sum_{K:I\cup J\subseteq K} \sum_{H:K\subseteq H} (-1)^{|H\setminus K|} y_H \;=\; \sum_{H:I\cup J\subseteq H} y_H \sum_{K:I\cup J\subseteq K\subseteq H} (-1)^{|H\setminus K|},$$

which is equal to $y_{I\cup J}$, since the inner summation $\sum_{K:I\cup J\subseteq K\subseteq H}(-1)^{|H\setminus K|}$ is equal to zero whenever $H \neq I \cup J$. □

Corollary 4.9.8. *Let $y \in \mathbb{R}^{\mathcal{P}_n(n)}$. The following assertions are equivalent.*

(i) $M_n(y) \succeq 0$.

(ii) $Z_n^{-1} y \geq 0$.

(iii) *y is a conic combination of the vectors $[x]_n$ for $x \in \{0,1\}^n$, i.e., $y = \sum_{x\in\{0,1\}^n} \lambda_x [x]_n$ for some nonnegative scalars λ_x.*

Example 4.9.9. *Let $n = 2$ and consider a vector $y = (y_\emptyset, y_1, y_2, y_{12})$. Then, y can be written as the following linear combination of the vectors $[x]_2$ for $x \in \{0,1\}^2$:*

$$y = (y_\emptyset - y_1 - y_2 + y_{12})[0]_2 + (y_1 - y_{12})[e_1]_2 + (y_2 - y_{12})[e_2]_2 + y_{12}[e_1 + e_2]_2$$

(setting $e_1 = (1,0)$ and $e_2 = (0,1)$). Therefore, we see that this is indeed a conic combination if and only if any of the following equivalent conditions holds:

$$M_2(y) = \begin{array}{c} \\ \emptyset \\ 1 \\ 2 \\ 12 \end{array} \!\!\begin{array}{c} \begin{array}{cccc} \emptyset & 1 & 2 & 12 \end{array} \\ \begin{pmatrix} y_0 & y_1 & y_2 & y_{12} \\ y_1 & y_1 & y_{12} & y_{12} \\ y_2 & y_{12} & y_2 & y_{12} \\ y_{12} & y_{12} & y_{12} & y_{12} \end{pmatrix} \end{array} \succeq 0 \iff \begin{cases} y_\emptyset - y_1 - y_2 + y_{12} \geq 0 \\ y_1 - y_{12} \geq 0 \\ y_2 - y_{12} \geq 0 \\ y_{12} \geq 0. \end{cases}$$

Canonical lifted representation for $0-1$ polytopes

We sketch here the significance of the above results about moment matrices for discrete optimization. A fundamental question is whether one can optimize efficiently a linear objective function over a given set $\mathcal{X} \subseteq \{0,1\}^n$. Think for instance of the traveling salesman problem, in which case $\mathcal{X}$ is the set of the incidence vectors of all Hamiltonian cycles in a graph, or of the maximum stable set problem considered here, in which case $\mathcal{X}$ is the set of incidence vectors of the stable sets in a graph. The classical so-called polyhedral approach is to consider the polytope $P = \mathrm{conv}(\mathcal{X})$, defined as the convex hull of all vectors in $\mathcal{X}$. Then the question boils down to finding the linear inequality description of P (or at least a part of it). It turns out that this question gets a simpler answer if we 'lift' the problem into higher dimension and allow the use of additionnal variables.

Define the polytope $\mathcal{P} = \mathrm{conv}([x]_n : x \in \mathcal{X})$. Let π denote the projection from the space $\mathbb{R}^{\mathcal{P}_n(n)}$ to the space $\mathbb{R}^n$ where, for a vector $y = (y_I)_{I\subseteq[n]}$, $\pi(y) = (y_1, \ldots, y_n)$ denotes its projection onto the coordinates indexed by the n singleton subsets of $[n]$. Then, by construction, we have that $P = \pi(\mathcal{P})$. As we now indicate the results from the previous subsection show that the lifted polytope $\mathcal{P}$ admits a simple explicit description.

Indeed, for a vector $y \in \mathbb{R}^{\mathcal{P}_n(n)}$, the trivial identity $y = Z_n(Z_n^{-1}y)$ shows that $y \in \mathcal{P}$ if and only if it satisfies the following conditions:

$$Z^{-1}y \geq 0, \ (Z^{-1}y)_x = 0 \ \forall x \notin \mathcal{X}, \ e^T Z^{-1}y = 1. \tag{4.37}$$

The first condition says that y is a conic combination of vectors $[x]_n$ (for $x \in \{0,1\}^n$), the second one says that only vectors $[x]_n$ for $x \in \mathcal{X}$ are used in this conic combination, the last one says that the conic combination is in fact a convex combination and it can be equivalently written as $y_\emptyset = 1$. Hence (9.7) gives an explicit linear inequality description for the polytope $\mathcal{P}$. Moreover, using Corollary 4.9.8, we can replace the condition $Z_n^{-1}y \geq 0$ by the condition $M_n(y) \succeq 0$. In this way we get a description of $\mathcal{P}$ involving positive semidefiniteness of the full moment matrix $M_n(y)$.

So what this says is that any polytope P with $0-1$ vertices can be obtained as projection of a polytope $\mathcal{P}$ admitting a simple explicit description. The price to pay however is $\mathcal{P}$ "lives" in a 2^n-dimensional space, thus exponentially large with respect to the dimension n of the ambiant space of the polytope P. Nevertheless this perspective leads naturally to hierarchies of semidefinite relaxations for P, obtained by considering only truncated parts $M_t(y)$ of $M_n(y)$ for growing orders t. We refer to [7, 8] for a detailed treatment, also about the links to other ift-and-project techniques used in combinatorial optimization.

This idea of 'lifting' a problem by adding new variables is widely used in optimization and it can sometimes lead to a huge efficiency gain. As a simple illustrating example consider the ℓ_1-ball $B = \{x \in \mathbb{R}^n : |x_1|, \ldots, |x_n| \leq 1\}$. The explicit linear inequality description of B requires the following 2^n inequalities: $\sum_{i=1}^n a_i x_i \leq 1$ for all $a \in \{\pm 1\}^n$. On the other hand, if we allow n additional variables we can describe B using only $3n$ linear inequalities. Namely, define the polytope

$$Q = \{(x,y) \in \mathbb{R}^n \times \mathbb{R}^n : x_i \leq y_i, \ -x_i \leq y_i, \ y_i \leq 1 \ (i \in [n]\}.$$

Then B coincides with the projection of Q onto the x-subspace.

Convergence of the Lasserre hierarchy to $\alpha(G)$.

We can now conclude the proof of Theorem 4.9.5, showing that the Lasserre relaxation solves the maximum stable set problem at any order $t \geq \alpha(G)$. It follows through the following claims. We let $\mathcal{S}_G$ denote the set of all characteristic vectors of the stable sets of G (so we consider here the set $\mathcal{X} = \mathcal{S}_G$).

Lemma 4.9.10. *Assume $M_t(y) \succeq 0$ and $y_{ij} = 0$ for all edges $\{i,j\} \in E$. Then $y_I = 0$ if I contains an edge and $|I| \leq 2t$.*

Proof. Exercise. □

Lemma 4.9.11. *Assume $y_0 = 1$, $M_n(y) \succeq 0$ and $y_{ij} = 0$ for all edges $\{i,j\} \in E$. Then y is a convex combination of vectors $[x]_n$ for $x \in \mathcal{S}_G$.*

Proof. By Corollary 4.9.8, $y = \sum_{x\in\{0,1\}^n} \lambda_x [x]_n$, with all $\lambda_x \geq 0$. It suffices now to observe that $\lambda_x > 0$ implies $x \in \mathcal{S}(G)$, which follows directly from the fact that $0 = y_{ij} = \sum_x \lambda_x x_i x_j$ for all edges $\{i,j\} \in E$. □

Lemma 4.9.12. *Let $t \geq \alpha(G)$. Assume $y \in \mathbb{R}^{\mathcal{P}_{2t}(n)}$ satisfies $M_t(y) \succeq 0$, $y_{ij} = 0$ for all edges $\{i,j\} \in E$ and $y_0 = 1$. Then, $\sum_{i=1}^n y_i \leq \alpha(G)$ and thus $\mathrm{las}_t(G) \leq \alpha(G)$.*

Proof. We extend the vector y to a vector $\tilde{y} \in \mathcal{P}_n(n)$ by setting $\tilde{y}_I = y_I$ if $|I| \leq 2t$ and $\tilde{y}_I = 0$ if $|I| > 2t$. We claim that the matrix $M_n(\tilde{y})$ has the block-form:

$$M_n(\tilde{y}) = \begin{pmatrix} M_t(y) & 0 \\ 0 & 0 \end{pmatrix}.$$

Indeed, by construction, $\tilde{y}_{I\cup J} = y_{I\cup J}$ if both I and J have cardinality at most t. Otherwise, say $|I| > t$. If $|I \cup J| \leq 2t$, then $\tilde{y}_{I\cup J} = y_{I\cup J} = 0$ by Lemma 4.9.10 (since I is not stable). If $|I \cup J| > 2t$, then $\tilde{y}_{I\cup J} = 0$ by construction.

Hence $M_t(y) \succeq 0$ implies $M_n(\tilde{y}) \succeq 0$. Using Lemma 4.9.11, we can conclude that $\tilde{y}$ is a convex combination of vectors $[x]_n$ for $x \in \mathcal{S}_G$. By projecting onto the positions indexed by $1, 2, \ldots, n$, we get that the vector $(y_1, \ldots, y_n) = (\tilde{y}_1, \ldots, \tilde{y}_n)$ is a convex combination of characteristic vectors of stable sets of G, and thus this implies $\sum_{i=1}^n y_i \leq \alpha(G)$. □

4.10 Further reading

In his seminal paper [10], 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 [6] and a detailed treatment about the material about the theta body TH(G) can be found in Chapter 9 of Grötschel, Lovász and Schrijver [5].

The Lasserre hierarchy of semidefinite bounds for $\alpha(G)$ is based on the work of Lasserre [7] and Laurent [8]. As explained there this type of hierarchies extends to arbitrary 0/1 polynomial optimization problems.

4.11 Exercises

4.1 Show the result of Proposition 4.5.8.

4.2 The goal is to show the result of Lemma 4.6.3 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 $\vartheta(G \cdot H) \geq \vartheta(G)\vartheta(H)$.

(b) Show that $\vartheta(G \cdot H) \leq \vartheta(G)\vartheta(H)$.

Hint: Use the primal formulation (4.11) for (a), and the dual formulation (4.12) for (b), and think of using Kronecker products of matrices in order to build feasible solutions.

4.3 Given a graph $G = (V = [n], E)$, a symmetric matrix $B \in \mathcal{S}^n$ is said to *fit* G if it has non-zero diagonal entries and zero entries at the off-diagonal positions corresponding to non-edges of G. Consider the parameter $R(G)$, defined as the smallest possible rank of a matrix B which fits G, i.e.,

$$R(G) = \min \ \text{rank}(B) \text{ such that } B_{ii} \neq 0 \ (i \in V), B_{ij} = 0 \ (\{i, j\} \in \overline{E}).$$

(a) Show that $R(G) \leq \chi(\overline{G})$.

(b) Show that $R(G) \geq \alpha(G)$.

(c) Show that $R(G) \geq \Theta(G)$.

(This upper bound on the Shannon capacity is due to W. Haemers.)

4.4 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 (i.e., $u_1, \ldots, u_n$ are unit vectors satisfying $u_i^{\mathsf{T}} u_j = 0$ for all pairs $\{i, j\} \in \overline{E}$).

Show: $\vartheta(G) = \vartheta_1(G)$.

Hint for the inequality $\vartheta(G) \leq \vartheta_1(G)$*:* Use the dual formulation of $\vartheta(G)$ from Lemma 4.4.1 and the matrix $M = (v_i^{\mathsf{T}} v_j)_{i,j=1}^n$, where $v_i = c - \frac{u_i}{c^{\mathsf{T}} u_i}$ for $i \in [n]$.

Hint for the inequality $\vartheta_1(G) \leq \vartheta(G)$: Use an optimal solution $X = tI - B$ of the dual formulation for $\vartheta(G)$, written as the Gram matrix of vectors $x_1, \ldots, x_n$. Show that there exists a nonzero vector c which is orthogonal to $x_1, \ldots, x_n$, and consider the vectors $u_i = \frac{c + x_i}{\sqrt{t}}$.

4.5 Show: $\vartheta(C_5) \leq \sqrt{5}$, using the formulation of Exercise 4.3 for the theta number.

Hint: Consider the following vectors $c, u_1, \ldots, u_5 \in \mathbb{R}^3$: $c = (0, 0, 1)$, $u_k = (s\cos(2k\pi/5), s\sin(2k\pi/5), t)$ for $k = 1, 2, 3, 4, 5$, where the scalars

$s, t \in \mathbb{R}$ are chosen in such a way that $u_1, \ldots, u_5$ form an orthonormal representation of C_5. Recall $\cos(2\pi/5) = \frac{\sqrt{5}-1}{4}$.

(This is the original proof of Lovász [10], known as the *umbrella construction*.)

4.6 Let $G = \overline{C_{2n+1}}$ be the complement of the odd cycle $(1, 2, \ldots, 2n+1)$. Consider a vector $y = (y_I)_{I \subseteq V, |I| \leq 4}$ which satisfies the conditions of the Lasserre relaxation of order 2. That is, $M_2(y) \succeq 0$, $y_{ij} = 0$ for all edges $\{i, j\} \in E(G)$, and $y_\emptyset = 1$.

Show: $\sum_{i \in V(G)} y_i \leq 2$.

BIBLIOGRAPHY

[1] V. Chvátal. On certain polytopes associated with graphs. *Journal of Combinatorial Theory, Series B* **18**:138–154, 1975.

[2] M. Chudnovsky, N. Robertson, P. Seymour, and R. Thomas. The strong perfect graph theorem, *Annals of Mathematics* **164 (1):** 51–229, 2006.

[3] P. Delsarte. *An algebraic approach to the association schemes of coding theory*. Philips Research Reports Supplements 1973, N. 10, Philips Research Laboratories, Eindhoven.

[4] G.S. Gasparian. Minimal imperfect graphs: a simple approach. *Combinatorica*, **16**:209–212, 1996.

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

[6] D.E. Knuth. The Sandwich Theorem. *The Electronic Journal of Combinatorics* **1, A1**, 1994.

[7] J.B. Lasserre. An explicit exact SDP relaxation for nonlinear $0-1$ programs In K. AARDAL AND A.M.H. GERARDS (EDS.), Lecture Notes in Computer Science **2081**:293–303, 2001.

[8] M. Laurent, A comparison of the Sherali-Adams, Lovász-Schrijver and Lasserre relaxations for 0-1 programming, *Mathematics of Operations Research* **28**(3):470–496, 2003.

[9] L. Lovász. A characterization of perfect graphs. *Journal of Combinatorial Theory, Series B* **13**:95–98, 1972.

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

[11] L. Lovász and A. Schrijver. Cones of matrices and set-functions and 0 ? 1 optimization. *SIAM Journal on Optimization* **1**:166–190, 1991.

[12] A. Schrijver. A comparison of the Delsarte and Lovász bounds. *IEEE Transactions on Information Theory*, IT-25: 425–429, 1979.

CHAPTER 5

APPROXIMATING THE MAX CUT PROBLEM

5.1 Introduction

5.1.1 The MAX CUT problem

The maximum cut problem (MAX CUT) is the following problem in combinatorial optimization. Let $G = (V, E)$ be a graph and let $w = (w_{ij}) \in \mathbb{R}^E_+$ be nonnegative weights assigned to the edges. Given a subset $S \subseteq V$, the *cut* $\delta_G(S)$ consists of the edges $\{u, v\} \in E$ having exactly one endnode in S, i.e., with $|\{i, j\} \cap S| = 1$. In other words, $\delta_G(S)$ consists of the edges that are *cut* by the partition $(S, \overline{S} = V \setminus S)$ of V. The cut $\delta_G(S)$ is called *trivial* if $S = \emptyset$ or V (in which case it is empty). Then the weight of the cut $\delta_G(S)$ is $w(\delta_G(S)) = \sum_{\{i,j\}\in\delta_G(S)} w_{ij}$ and the MAX CUT problem asks for a cut of maximum weight, i.e., compute

$$\text{mc}(G, w) = \max_{S\subseteq V} w(\delta_G(S)).$$

It is sometimes convenient to extend the weight function $w \in \mathbb{R}^E$ to all pairs of nodes of V, by setting $w_{ij} = 0$ if $\{i, j\}$ is not an edge of G. Given disjoint subsets $S, T \subseteq V$, it is also convenient to use the following notation:

$$w(S, T) = \sum_{i\in S, j\in T} w_{ij}.$$

Thus,

$$w(S, \overline{S}) = w(\delta_G(S)) \text{ for all } S \subseteq V.$$

To state its complexity, we formulate MAX CUT as a decision problem:

MAX CUT: *Given a graph $G = (V, E)$, edge weights $w \in \mathbb{Z}_+^E$ and an integer $k \in \mathbb{N}$, decide whether there exists a cut of weight at least k.*

It is well known that MAX CUT is an NP-complete problem. In fact, MAX CUT is one of Karp's 21 NP-complete problems. So unless the complexity classes P and NP coincide there is no efficient polynomial-time algorithm which solves MAX CUT exactly. We give here a reduction of MAX CUT from the PARTITION problem, defined below, which is one the first six basic NP-complete problems in Garey and Johnson [3]:

PARTITION: *Given natural numbers $a_1, \dots, a_n \in \mathbb{N}$, decide whether there exists a subset $S \subseteq [n]$ such that $\sum_{i \in S} a_i = \sum_{i \notin S} a_i$.*

Theorem 5.1.1. *The MAX CUT problem is NP-complete.*

Proof. It is clear that MAX CUT to the class NP. We now show a reduction from PARTITION. Let $a_1, \dots, a_n \in \mathbb{N}$ be given. Construct the following weights $w_{ij} = a_i a_j$ for the edges of the complete graph K_n. Set $\sigma = \sum_{i=1}^n a_i$ and $k = \sigma^2/4$. For any subset $S \subseteq [n]$, set $a(S) = \sum_{i \in S} a_i$. Then, we have

$$w(S, \overline{S}) = \sum_{i \in S, j \in \overline{S}} w_{ij} = \sum_{i \in S, j \in \overline{S}} a_i a_j = (\sum_{i \in S} a_i)(\sum_{j \in \overline{S}} a_j) = a(S)(\sigma - a(S)) \leq \sigma^2/4,$$

with equality if and only if $a(S) = \sigma/2$ or, equivalently, $a(S) = a(\overline{S})$. From this it follows that there is a cut of weight at least k if and only if the sequence $a_1, \dots, a_n$ can be partitioned. This concludes the proof. □

This hardness result for MAX CUT is in sharp contrast to the situation of the MIN CUT problem, which asks for a *nontrivial* cut of minimum weight, i.e., to compute

$$\min_{S \subseteq V : S \neq \emptyset, V} w(S, \overline{S}).$$

(For MIN CUT the weights of edges are usually called *capacities* and they also assumed to be nonnegative). It is well known that the MIN CUT problem can be solved in polynomial time (together with its dual MAX FLOW problem), using the Ford-Fulkerson algorithm. Specifically, the Ford-Fulkerson algorithm permits to find in polynomial time a minimum cut $(S, \overline{S})$ separating a given source s and a given sink t, i.e., with $s \in S$ and $t \in \overline{S}$. Thus a minimum weight nontrivial cut can be obtained by applying this algorithm $|V|$ times, fixing any $s \in V$ and letting t vary over all nodes of $V \setminus \{s\}$. Details can be found in Chapter 4 of the Lecture Notes [7].

Even stronger, Håstad in 2001 showed that it is NP-hard to approximate MAX CUT within a factor of $\frac{16}{17} \sim 0.941$.

On the positive side, one can compute a 0.878-approximation of MAX CUT in polynomial time, using semidefinite programming. This algorithm, due to

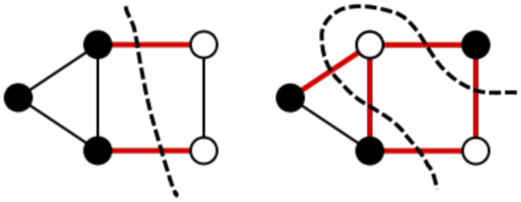


Figure 5.1: Minimum and maximum weight cuts

Goemans and Williamson [4], is one of the most influential approximation algorithms which are based on semidefinite programming. We will explain this result in detail in Section 5.2.1.

Before doing that we recall some results for MAX CUT based on using linear programming.

5.1.2 Linear programming relaxation

In order to define linear programming bounds for MAX CUT, one needs to find some linear inequalities that are satisfied by all cuts of G, i.e., some valid inequalities for the cut polytope of G. Large classes of such inequalities are known (cf. e.g. [2] for an overview and references).

We now present some simple but important valid inequalities for the cut polytope of the complete graph K_n, which is denoted as CUT_n, and defined as the convex hull of the incidence vectors of the cuts of K_n:

$$\text{CUT}_n = \text{conv}\{\chi^{\delta_{K_n}(S)} : S \subseteq [n]\}.$$

For instance, for $n = 2$, $\text{CUT}_n = [0, 1]$ and, for $n = 3$, CUT_3 is a simplex in $\mathbb{R}^3$ (indexed by the edges of K_3 ordered as $\{1, 2\}, \{1, 3\}, \{2, 3\}$) with as vertices the incidence vectors of the four cuts $(S, \overline{S})$ of K_3: $(0,\ 0,\ 0)$, $(1,\ 1,\ 0)$, $(1,\ 0,\ 1)$, and $(0\ 1\ 1)$ (for $S = \emptyset, \{1\}, \{2\}$ and $\{3\}$, respectively).

As a first easy observation it is important to realize that in order to compute the maximum cut $\text{mc}(G, w)$ in a weighted graph G on n nodes, one can as well deal with the complete graph K_n. Indeed, any cut $\delta_G(S)$ of G can be obtained from the corresponding cut $\delta_{K_n}(S)$ of K_n, simply by ignoring the pairs that are not edges of G, in other words, by projecting onto the edge set of G. Hence one can reformulate any maximum cut problem as a linear optimization problem over the cut polytope of K_n:

$$\text{mc}(G, w) = \max_{x \in \text{CUT}_n} \sum_{\{i,j\} \in E} w_{ij} x_{ij};$$

the graph G is taken into account by the objective function of this LP.

The following *triangle inequalities* are valid for the cut polytope CUT_n:

$$x_{ij} - x_{ik} - x_{jk} \leq 0,\ x_{ij} + x_{jk} + x_{jk} \leq 2, \tag{5.1}$$

for all distinct $i, j, k \in [n]$. This is easy to see, just verify that these inequalities hold when x is equal to the incidence vector of a cut. The triangle inequalities (5.1) imply the following bounds (check it):

$$0 \leq x_{ij} \leq 1 \tag{5.2}$$

on the variables. Let MET$_n$ denote the polytope in $\mathbb{R}^{E(K_n)}$ defined by the triangle inequalities (5.1). Thus, MET$_n$ is a linear relaxation of CUT$_n$, tighter than the trivial relaxation by the unit hypercube:

$$\text{CUT}_n \subseteq \text{MET}_n \subseteq [0,1]^{E(K_n)}.$$

It is known that equality holds for $n \leq 4$, but the inclusion is strict for $n \geq 5$. Indeed, the inequality:

$$\sum_{1 \leq i < j \leq 5} x_{ij} \leq 6 \tag{5.3}$$

is valid for CUT$_5$ (as any cut of K_5 has cardinality 0, 4 or 6), but it is not valid for MET$_5$. For instance, the vector $(2/3, \ldots, 2/3) \in \mathbb{R}^{10}$ belongs to MET$_5$ but it violates the inequality (5.3) (since $10.2/3 > 6$).

We can define the following linear programming bound:

$$\text{lp}(G, w) = \max \left\{ \sum_{\{i,j\} \in E(G)} w_{ij} x_{ij} : x \in \text{MET}_n \right\} \tag{5.4}$$

for the maximum cut:

$$\text{mc}(G, w) \leq \text{lp}(G, w).$$

The graphs for which this bound is tight have been characterized by Barahona [1]:

Theorem 5.1.2. *Let G be a graph. Then,* $\text{mc}(G, w) = \text{lp}(G, w)$ *for all weight functions $w \in \mathbb{R}^E$ if and only if the graph G has no K_5 minor.*

In particular, if G is a planar graph, then $\text{mc}(G, w) = \text{lp}(G, w)$ so that the maximum cut can be computed in polynomial time using linear programming.

A natural question is how good the LP bound is for general graphs. Here are some easy bounds.

Lemma 5.1.3. *Let G be a graph with nonnegative weights w. The following holds.*

(i) $\text{mc}(G, w) \leq \text{lp}(G, w) \leq w(E)$.

(ii) $\text{mc}(G, w) \geq w(E)/2$.

Proof. (i) follows from the fact that $\text{MET}_n \subseteq [0,1]^{E(K_n)}$ and $w \geq 0$. For (ii) pick $S \subseteq V$ for which $(S, \overline{S})$ is a cut of maximum weight: $w(S, \overline{S}) = \text{mc}(G, w)$. Thus if we move one node $i \in S$ to $\overline{S}$, or if we move one node $j \in \overline{S}$ to S, then we obtain another cut whose weight is at most $w(S, \overline{S})$. This gives:

$$w(S \setminus \{i\}, \overline{S} \cup \{i\}) - w(S, \overline{S}) = w(S \setminus \{i\}, \{i\}) - w(\{i\}, \overline{S}) \geq 0,$$

$$w(S \cup \{j\}, \overline{S} \setminus \{j\}) - w(S, \overline{S}) = w(\{j\}, \overline{S} \setminus \{j\}) - w(S, \{j\}) \geq 0.$$

Summing the first relation over $i \in S$ and using the fact that $2w(E(S)) = \sum_{i \in S} w(S \setminus \{i\}, \{i\})$, where $E(S)$ is the set of edges contained in S, and the fact that $\sum_{i \in S} w(\{i\}, \overline{S}) = w(S, \overline{S})$, we obtain:

$$2w(E(S)) \geq w(S, \overline{S}).$$

Analogously, summing over $j \in \overline{S}$, we obtain:

$$2w(E(\overline{S})) \geq w(S, \overline{S}).$$

Summing these two relations yields: $w(E(S)) + w(E(\overline{S})) \geq w(S, \overline{S})$. Now adding $w(S, \overline{S})$ to both sides implies: $w(E) \geq 2w(S, \overline{S}) = 2\text{mc}(G, w)$, which shows (ii). □

As an application of Lemma 5.1.3, we obtain that

$$\frac{1}{2} \leq \frac{\text{mc}(G, w)}{\text{lp}(G, w)} \leq 1 \quad \text{for all nonnegative weights } w \geq 0.$$

It turns out that there are graphs for which the ratio $mc(G, w)/\text{lp}(G, w)$ can be arbitrarily close to 1/2 [6]. This means that for these graphs, the metric polytope does not provide a better approximation of the cut polytope than its trivial relaxation by the hypercube $[0, 1]^E$.

We now provide another argument for the lower bound $\text{mc}(G, w) \geq w(E)/2$. This argument is probabilistic and based on the following simple randomized algorithm: Construct a random partition $(S, \overline{S})$ of V by assigning, independently, with probability 1/2, each node $i \in V$ to either side of the partition. Then the probability that an edge $\{i, j\}$ is cut by the partition is equal to

$$\mathbb{P}(\{i, j\} \text{ is cut}) = \mathbb{P}(i \in S, j \in \overline{S}) + \mathbb{P}(i \in \overline{S}, j \in S) = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}.$$

Hence, the expected weight of the cut produced by this random partition is equal to

$$\mathbb{E}(w(S, \overline{S})) = \sum_{\{i,j\} \in E} w_{ij} \mathbb{P}(\{i, j\} \text{ is cut}) = \sum_{\{i,j\} \in E} w_{ij} \frac{1}{2} = \frac{w(E)}{2}.$$

Here we have used the linearity of the expectation.

In the next section, we will see another probabilistic argument, due to Goemans and Williamson, which permits to construct a much better random cut. Namely we will get a random cut whose expected weight satisfies:

$$E(w(S, \overline{S})) \geq 0.878 \cdot w(E),$$

thus improving the above factor 0.5. The crucial tool will be to use a semidefinite relaxation for MAX CUT combined with a simple, but ingenious randomized “hyperplane rounding” technique.

5.2 The algorithm of Goemans and Williamson

5.2.1 Semidefinite programming 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. With every vertex $i \in V$, we associate a binary variable $x_i \in \{-1, +1\}$ which indicates whether i lies in S or in $\overline{S}$, say, $i \in S$ if $x_i = -1$ and $i \in \overline{S}$ if $x_i = +1$. We model the binary constraint $x_i \in \{-1, +1\}$ as a quadratic equality constraint

$$x_i^2 = 1 \text{ for } i \in V.$$

For two vertices $i, j \in V$ we have

$$1 - x_i x_j \in \{0, 2\}.$$

This value equals to 0 if i and j lie on the same side of the cut $(S, \overline{S})$ and the value equals to 2 if i and j lie on different sides of the cut. Hence, one can express the weight of the cut $(S, \overline{S})$ by

$$w(S, \overline{S}) = \sum_{\{i,j\} \in E} w_{ij} \frac{1 - x_i x_j}{2}.$$

Now, the MAX CUT problem can be equivalently formulated as

$$\text{mc}(G, w) = \max \left\{ \frac{1}{2} \sum_{\{i,j\} \in E} w_{ij}(1 - x_i x_j) : x_i^2 = 1 \ \forall i \in V \right\}. \tag{5.5}$$

Next, we introduce a matrix variable $X = (x_{ij}) \in \mathcal{S}^n$, whose entries x_{ij} model the pairwise products $x_i x_j$. Then, as the matrix $(x_i x_j)_{i,j=1}^n = xx^{\mathsf{T}}$ is positive semidefinite, we can require the condition that X should be positive semidefinite. Moreover, the constraints $x_i^2 = 1$ give the constraints $X_{ii} = 1$ for all $i \in [n]$. Therefore we can formulate the following semidefinite programming relaxation:

$$\text{sdp}(G, w) = \max \left\{ \frac{1}{2} \sum_{\{i,j\} \in E} w_{ij}(1 - X_{ij}) : X \succeq 0, \ X_{ii} = 1 \ \forall i \in [n] \right\}. \tag{5.6}$$

By construction, we have:

$$\text{mc}(G, w) \leq \text{sdp}(G, w). \tag{5.7}$$

The feasible region of the above semidefinite program is the convex (non-polyhedral) set

$$\mathcal{E}_n = \{X \in \mathcal{S}^n : X \succeq 0,\ X_{ii} = 1\ \forall i \in [n]\},$$

called the *elliptope* (and its members are known as *correlation matrices*). One can visualize the elliptope $\mathcal{E}_3$. Indeed, for a 3×3 symmetric matrix X with an all-ones diagonal, we have:

$$X = \begin{pmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{pmatrix} \succeq 0 \Longleftrightarrow 1 + 2xyz - x^2 - y^2 - z^2 \geq 0,\ x, y, z \in [-1, 1],$$

which expresses the fact that the determinant of X is nonnegative as well as the three 2×2 principal subdeterminants. The following Figure 5.2.1 visualizes the set of triples (x, y, z) for which $X \in \mathcal{E}_3$. Notice that the elliptope $\mathcal{E}_3$ looks like an "inflated" tetrahedron, while the underlying tetrahedron corresponds to the linear relaxation MET_3.

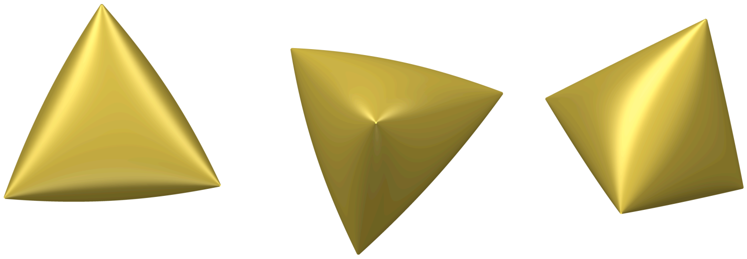


Figure 5.2: Views on the convex set $\mathcal{E}_3$ behind the semidefinite relaxation.

5.2.2 The Goemans-Williamson algorithm

Goemans and Williamson [4] show the following result for the semidefinite programming bound $\mathrm{sdp}(G, w)$.

Theorem 5.2.1. *Given a graph G with nonnegative edge weights w, the following inequalities hold:*

$$\mathrm{sdp}(G, w) \geq \mathrm{mc}(G, w) \geq 0.878 \cdot \mathrm{sdp}(G, w).$$

The proof is algorithmic and it gives an approximation algorithm which approximates the MAX CUT problem within a ratio of 0.878. The Goemans-Williamson algorithm has five steps:

1. Solve the semidefinite program (5.6); let X be an optimal solution, so that $\mathrm{sdp}(G, w) = \sum_{\{i,j\} \in E} w_{ij}(1 - X_{ij})/2$.

2. Perform a Cholesky decomposition of X to find unit vectors $v_i \in \mathbb{R}^{|V|-1}$ for $i \in V$, so that $X = (v_i^{\mathsf{T}} v_j)_{i,j \in V}$.

3. Choose a random unit vector $r \in \mathbb{R}^{|V|-1}$, according to the rotationally invariant probability distribution on the unit sphere.

4. Define a cut $(S, \overline{S})$ by setting $x_i = \operatorname{sign}(v_i^{\mathsf{T}} r)$ for all $i \in V$. That is, $i \in S$ if and only if $\operatorname{sign}(v_i^{\mathsf{T}} r) \leq 0$.

5. Check whether $\sum_{\{i,j\} \in E} w_{ij}(1 - x_i x_j)/2 \geq 0.878 \cdot \text{sdp}(G, w)$. If not, go to step 3.

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 $0.878 \cdot \text{sdp}(G, w)$ 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.

The following lemma (also known as *Grothendieck’s identity*, since it came up in work of Grothendieck in the 50’s, however in the different context of functional analysis) is the key to the proof of Theorem 5.2.1.

Lemma 5.2.2. *Let $u, v \in \mathbb{R}^d$ (for some $d \geq 1$) be unit vectors and let $r \in \mathbb{R}^d$ be a random unit vector chosen according to the rotationally invariant probability distribution on the unit sphere. The following holds.*

(i) The probability that $\operatorname{sign}(u^{\mathsf{T}} r) \neq \operatorname{sign}(v^{\mathsf{T}} r)$ is equal to

$$\mathbb{P}(\operatorname{sign}(u^{\mathsf{T}} r) \neq \operatorname{sign}(v^{\mathsf{T}} r)) = \frac{\arccos(u^{\mathsf{T}} v)}{\pi}. \tag{5.8}$$

(ii) The expectation of the random variable $\operatorname{sign}(u^{\mathsf{T}} r)\operatorname{sign}(v^{\mathsf{T}} r) \in \{-1, +1\}$ is equal to

$$\mathbb{E}[\operatorname{sign}(u^{\mathsf{T}} r)\operatorname{sign}(v^{\mathsf{T}} r)] = \frac{2}{\pi}\arcsin(u^{\mathsf{T}} v). \tag{5.9}$$

Proof. (i) Since the probability distribution from which we sample the unit vector r is rotationally invariant we can assume that u, v and r lie in a common plane. Hence we can assume that they lie on a unit circle in $\mathbb{R}^2$ and that r is chosen according to the uniform distribution on this circle. Then the probability that $\operatorname{sign}(u^{\mathsf{T}} r) \neq \operatorname{sign}(v^{\mathsf{T}} r)$ depends only on the angle between u and v. Using a figure (draw one!) it is easy to see that

$$\mathbb{P}[\operatorname{sign}(u^{\mathsf{T}} r) \neq \operatorname{sign}(v^{\mathsf{T}} r)] = 2 \cdot \frac{1}{2\pi}\arccos(u^{\mathsf{T}} v) = \frac{1}{\pi}\arccos(u^{\mathsf{T}} v).$$

(ii) By definition, the expectation $\mathbb{E}[\text{sign}(u^\mathsf{T} r)\,\text{sign}(v^\mathsf{T} r)]$ can be computed as

$$(+1)\cdot\mathbb{P}[\text{sign}(u^\mathsf{T} r) = \text{sign}(v^\mathsf{T} r)] + (-1)\cdot\mathbb{P}[\text{sign}(u^\mathsf{T} r) \neq \text{sign}(v^\mathsf{T} r)]$$

$$= 1 - 2\cdot\mathbb{P}[\text{sign}(u^\mathsf{T} r) \neq \text{sign}(v^\mathsf{T} r)] = 1 - 2\cdot\tfrac{\arccos(u^\mathsf{T} v)}{\pi},$$

where we have used (i) for the last equality. Now use the trigonometric identity

$$\arcsin t + \arccos t = \frac{\pi}{2},$$

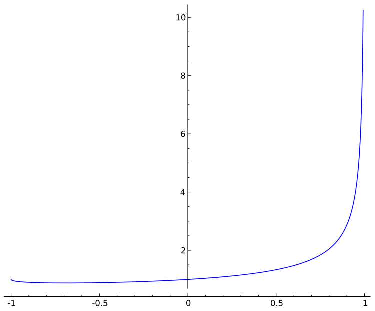
to conclude the proof of (ii). □

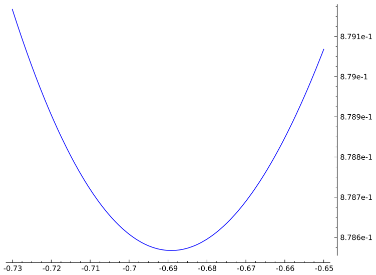
Using elementary univariate calculus one can show the following fact.

Lemma 5.2.3. *For all $t \in [-1, 1)]$, the following inequality holds:*

$$\frac{2}{\pi}\frac{\arccos t}{1-t} \geq 0.878. \tag{5.10}$$

One can also "see" this on the following plots of the function in (5.10), where t varies in $[-1, 1)$ in the first plot and in $[-0.73, -0.62]$ in the second plot.





Proof. (of Theorem 5.2.1) Let X be the optimal solution of the semidefinite program (5.6) and let $v_1, \ldots, v_n$ be unit vectors such that $X = (v_i^\mathsf{T} v_j)_{i,j=1}^n$, as in Steps 1,2 of the GW algorithm. Let $(S, \overline{S})$ be the random partition of V, as in Steps 3,4 of the algorithm. We now use Lemma 5.2.2(i) to compute the expected value of the cut $(S, \overline{S})$:

$$\mathbb{E}(w(S,\overline{S})) = \textstyle\sum_{\{i,j\}\in E} w_{ij}\mathbb{P}(\{i,j\} \text{ is cut}) = \sum_{\{i,j\}\in E} w_{ij}\mathbb{P}(x_i \neq x_j)$$

$$= \textstyle\sum_{\{i,j\}\in E} w_{ij}\mathbb{P}(\text{sign}(v_i^\mathsf{T} r) \neq \text{sign}(v_j^\mathsf{T} r)) = \sum_{\{i,j\}\in E} w_{ij}\,\frac{\arccos(v_i^\mathsf{T} v_j)}{\pi}$$

$$= \textstyle\sum_{\{i,j\}\in E} w_{ij}\left(\frac{1-v_i^\mathsf{T} v_j}{2}\right)\cdot\left(\frac{2}{\pi}\frac{\arccos(v_i^\mathsf{T} v_j)}{1-v_i^\mathsf{T} v_j}\right).$$

By Lemma 5.2.3, each term $\frac{2}{\pi}\frac{\arccos(v_i^\mathsf{T} v_j)}{1-v_i^\mathsf{T} v_j}$ can be lower bounded by the constant 0.878. Since all weights are nonnegative, each term $w_{ij}(1-v_i^\mathsf{T} v_j)$ is nonnegative.

Therefore, we can lower bound $\mathbb{E}(w(S,\overline{S}))$ in the following way:

$$\mathbb{E}(w(S,\overline{S})) \geq 0.878 \cdot \sum_{\{i,j\}\in E} w_{ij} \left(\frac{1 - v_i^{\mathsf{T}} v_j}{2}\right).$$

Now we recognize that the objective value $\text{sdp}(G,w)$ of the semidefinite program is appearing in the right hand side and we obtain:

$$\mathbb{E}(w(S,\overline{S})) \geq 0.878 \cdot \sum_{\{i,j\}\in E} w_{ij} \left(\frac{1 - v_i^{\mathsf{T}} v_j}{2}\right) = 0.878 \cdot \text{sdp}(G,w).$$

Finally, it is clear that the maximum weight of a cut is at least the expected value of the random cut $(S,\overline{S})$:

$$\text{mc}(G,w) \geq \mathbb{E}(w(S,\overline{S})).$$

Putting things together we can conclude that

$$\text{mc}(G,w) \geq \mathbb{E}(w(S,\overline{S})) \geq 0.878 \cdot \text{sdp}(G,w).$$

This concludes the proof, since the other inequality $\text{mc}(G,w) \leq \text{sdp}(G,w)$ holds by (5.7). □

5.2.3 Remarks on the algorithm

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.

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.

Finally we mention that 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.

5.3 Extensions

5.3.1 Reformulating MAX CUT using the Laplacian matrix

Given a graph G with edge weights w, its *Laplacian matrix* L_w is the symmetric $n \times n$ matrix with entries:

$$(L_w)_{ii} = \sum_{j:\{i,j\} \in E} w_{ij} \ (i \in [n]),$$

$$(L_w)_{ij} = -w_{ij} \ (\{i,j\} \in E), \ (L_w)_{ij} = 0 \ (i \neq j, \{i,j\} \notin E).$$

The following can be checked (Exercise 4.2).

Lemma 5.3.1. *The following properties hold for the Laplacian matrix L_w:*

(i) For any vector $x \in \{\pm 1\}^n$, $\frac{1}{4}x^{\mathsf{T}} L_w x = \frac{1}{2}\sum_{\{i,j\}\in E} w_{ij}(1 - x_i x_j)$.

(ii) For any nonnegative edge weights $w \geq 0$, $L_w \succeq 0$.

This permits to reformulate the quadratic formulation (5.5) of MAX CUT as

$$\text{mc}(G,w) = \max\left\{\frac{1}{4}x^{\mathsf{T}} L_w x : x_i^2 = 1 \ \forall i \in V\right\}$$

and its semidefinite relaxation (5.6) as

$$\text{sdp}(G,w) = \max\left\{\frac{1}{4}\langle L_w, X\rangle : X \succeq 0, \ X_{ii} = 1 \ \forall i \in V\right\}.$$

A property of the above programs is that the matrix $L_w/4$ occurring in the objective function is positive semidefinite. In the next section we consider general quadratic programs, where L_w is replaced by an arbitrary positive semidefinite matrix A. Then one can still show an approximation algorithm, however with performance ration $\frac{2}{\pi} \sim 0.636$, thus weaker than the 0.878 ratio of Goemans and Williamson for the case when $A = L_w$ for some $w \geq 0$.

5.3.2 Nesterov's approximation algorithm

Nesterov [5] considers the class of quadratic problems:

$$\text{qp}(A) = \max\left\{\sum_{i,j=1}^{n} A_{ij} x_i x_j : x_i^2 = 1 \ \forall i \in [n]\right\}, \tag{5.11}$$

where $A \in \mathcal{S}^n$ is a symmetric matrix. (Thus, qp$(A) = $ mc(G,w) for $A = L_w/4$). Analogously define the semidefinite programming relaxation:

$$\text{sdp}(A) = \max\left\{\langle A, X\rangle : X \succeq 0, \ X_{ii} = 1 \ \forall i \in [n]\right\}. \tag{5.12}$$

The following inequality holds:

$$\text{qp}(A) \leq \text{sdp}(A)$$

for any symmetric matrix A. In the special case when A is positive semidefinite, Nesterov shows that $\text{sdp}(A)$ is a $\frac{2}{\pi}$-approximation for $\text{qp}(A)$. The proof is based on the same rounding technique of Goemans-Williamson, but the analysis is different. It relies on the following property of the function $\arcsin t$: There exist positive scalars $a_k > 0$ $(k \geq 0)$ such that

$$\arcsin t = t + \sum_{k \geq 0} a_k t^{2k+1} \text{ for all } t \in [-1, 1]. \tag{5.13}$$

Based on this one can show the following result.

Lemma 5.3.2. *Given a matrix $X = (x_{ij}) \in \mathcal{S}^n$, define the new matrix*

$$\tilde{X} = (\arcsin X_{ij} - X_{ij})_{i,j=1}^n,$$

whose entries are the images of the entries of X under the map $t \mapsto \arcsin t - t$. Then, $X \succeq 0$ implies $\tilde{X} \succeq 0$.

Proof. The proof uses the following fact: If $X = (x_{ij})_{i,j=1}^n$ is positive semidefinite then, for any integer $k \geq 1$, the matrix $(X_{ij}^k)_{i,j=1}^n$ (whose entries are the k-th powers of the entries of X) is positive semidefinite as well. (Recall Section 1.2.2 of Chapter 1.)
Using this fact, the form of the series decomposition (5.13), and taking limits, implies the result of the lemma. □

Theorem 5.3.3. *Assume A is a positive semidefinite matrix. Then,*

$$\text{sdp}(A) \geq \text{qp}(A) \geq \frac{2}{\pi}\text{sdp}(A).$$

Proof. Let X be an optimal solution of the semidefinite program (5.12) and let $v_1, \ldots, v_n$ be unit vectors such that $X = (v_i^\mathsf{T} v_j)_{i,j=1}^n$ (as in Steps 1,2 of the GW algorithm). Pick a random unit vector r and set $x_i = \text{sign}(v_i^\mathsf{T} r)$ for $i \in V$ (as in Steps 3,4 of the GW algorithm). We now use Lemma 5.2.2(ii) to compute the expected value of $\sum_{i,j=1}^n A_{ij} x_i x_j$:

$$\begin{aligned}
&\mathbb{E}(\textstyle\sum_{i,j=1}^n A_{ij} x_i x_j) = \sum_{i,j=1}^n A_{ij} \mathbb{E}(x_i x_j) \\
&= \tfrac{2}{\pi} \textstyle\sum_{i,j=1}^n A_{ij} \arcsin(v_i^\mathsf{T} v_j) = \tfrac{2}{\pi} \sum_{i,j=1}^n A_{ij} \arcsin X_{ij} \\
&= \tfrac{2}{\pi} \left(\textstyle\sum_{i,j=1}^n A_{ij} X_{ij} + \sum_{i,j=1}^n A_{ij} (\arcsin X_{ij} - X_{ij}) \right).
\end{aligned}$$

By Lemma 5.3.2, the second term is equal to $\langle A, \tilde{X} \rangle \geq 0$, since $\tilde{X} \succeq 0$. Moreover, we recognize in the first term the objective value of the semidefinite program

(5.12). Combining these facts, we obtain:

$$\mathbb{E}(\sum_{i,j=1}^{n} A_{ij}x_ix_j) \geq \frac{2}{\pi}\text{sdp}(A).$$

On the other hand, it is clear that

$$\text{qp}(A) \geq \mathbb{E}(\sum_{i,j=1}^{n} A_{ij}x_ix_j).$$

This concludes the proof. □

5.3.3 Quadratic programs modeling MAX 2SAT

Here we consider another class of quadratic programs, of the form:

$$\text{qp}(a,b) = \max\left\{\sum_{ij\in E_1} a_{ij}(1-x_ix_j) + \sum_{ij\in E_2} b_{ij}(1+x_ix_j) : x\in\{\pm1\}^n\right\}, \tag{5.14}$$

where $a_{ij}, b_{ij} \geq 0$ for all ij. Write the semidefinite relaxation:

$$\text{sdp}(a,b) = \max\left\{\sum_{ij\in E_1} a_{ij}(1-X_{ij}) + \sum_{ij\in E_2} b_{ij}(1+X_{ij}) : X\succeq 0,\ X_{ii}=1\ \forall i\in[n]\right\}. \tag{5.15}$$

Goemans and Williamson [4] show that the same approximation result holds as for MAX CUT:

Theorem 5.3.4. *Assume that $a, b \geq 0$. Then,*

$$\text{sdp}(a,b) \geq \text{qp}(a,b) \geq 0.878\cdot\text{sdp}(a,b).$$

In the proof we will use the following variation of Lemma 5.2.3.

Lemma 5.3.5. *For any $z\in[-1,1]$, the following inequality holds:*

$$\frac{2}{\pi}\frac{\pi-\arccos z}{1+z} \geq 0.878.$$

Proof. Set $t=-z\in[-1,1]$. Using the identity $\arccos(-t)=\pi-\arccos t$ and applying (5.10), we get: $\frac{2}{\pi}\frac{\pi-\arccos z}{1+z} = \frac{2}{\pi}\frac{\arccos t}{1-t} \geq 0.878$. □

Proof. (of Theorem 5.3.4) We apply the GW algorithm: Let $X=(v_i^{\mathsf{T}}v_j)$ be an optimal solution of (5.15). Pick a random unit vector r and set $x_i=\text{sign}(v_i^{\mathsf{T}}r)$ for $i\in[n]$. Using the fact that $\mathbb{E}(x_ix_j)=1-2\cdot\mathbb{P}(x_i\neq x_j)=1-2\cdot\frac{\arccos(v_i^{\mathsf{T}}v_j)}{\pi}$, we can compute the expected value of the quadratic objective of (5.14) evaluated at x:

$$\mathbb{E}\left(\sum_{ij\in E_1} a_{ij}(1-x_ix_j)+\sum_{ij\in E_2} b_{ij}(1+x_ix_j)\right)$$

$$= 2\cdot\sum_{ij\in E_1} a_{ij}\tfrac{\arccos(v_i^\mathsf{T}v_j)}{\pi} + 2\cdot\sum_{ij\in E_2} b_{ij}\left(1-\tfrac{\arccos(v_i^\mathsf{T}v_j)}{\pi}\right)$$

$$= \sum_{ij\in E_1} \underbrace{a_{ij}(1-v_i^\mathsf{T}v_j)}_{\geq 0}\underbrace{\frac{2}{\pi}\frac{\arccos(v_i^\mathsf{T}v_j)}{1-v_i^\mathsf{T}v_j}}_{\geq 0.878} + \sum_{ij\in E_2} \underbrace{b_{ij}(1+v_i^\mathsf{T}v_j)}_{\geq 0}\underbrace{\frac{2}{\pi}\frac{\pi-\arccos(v_i^\mathsf{T}v_j)}{1+v_i^\mathsf{T}v_j}}_{\geq 0.878}$$

$$\geq 0.878\cdot \mathrm{sdp}(a,b).$$

Here we have used Lemmas 5.2.3 and 5.3.5. From this we can conclude that $\mathrm{qp}(a,b)\geq 0.878\cdot\mathrm{sdp}(a,b)$. □

In the next section we indicate how to use the quadratic program (5.14) in order to formulate MAX 2 SAT.

5.3.4 Approximating MAX 2-SAT

An instance of MAX SAT is given by a collection of Boolean clauses $C_1,\ldots,C_m$, where each clause C_j is a disjunction of literals, drawn from a set of variables $\{z_1,\ldots,z_n\}$. A literal is a variable z_i or its negation $\overline{z}_i$. Moreover there is a weight w_j attached to each clause C_j. The MAX SAT problem asks for an assignment of truth values to the variables $z_1,\ldots,z_n$ that maximizes the total weight of the clauses that are satisfied. MAX 2SAT consists of the instances of MAX SAT where each clause has at most two literals. It is an NP-complete problem [3] and analogously to MAX CUT it is also hard to approximate.

Goemans and Williamson show that their randomized algorithm for MAX CUT also applies to MAX 2SAT and yields again a 0.878-approximation algorithm. Prior to their result, the best approximation was 3/4, due to Yannakakis (1994).

To show this it suffices to model MAX 2SAT as a quadratic program of the form (5.14). We now indicate how to do this. We introduce a variable $x_i\in\{\pm1\}$ for each variable z_i of the SAT instance. We also introduce an additional variable $x_0\in\{\pm1\}$ which is used as follows: z_i is true if $x_i=x_0$ and false otherwise.

Given a clause C, define its value $v(C)$ to be 1 if the clause C is true and 0 otherwise. Thus,

$$v(z_i)=\frac{1+x_0x_i}{2},\ v(\overline{z}_i)=1-v(z_i)=\frac{1-x_0x_i}{2}.$$

Based on this one can now express $v(C)$ for a clause with two literals:

$$v(z_i\vee z_j)=1-v(\overline{z}_i\wedge\overline{z}_j)=1-v(\overline{z}_i)v(\overline{z}_j)=1-\tfrac{1-x_0x_i}{2}\tfrac{1-x_0x_j}{2}$$

$$=\tfrac{1+x_0x_i}{4}+\tfrac{1+x_0x_j}{4}+\tfrac{1-x_ix_j}{4}.$$

Analogously, one can express $v(z_i \vee \overline{z}_j)$ and $v(\overline{z}_i \vee \overline{z}_j)$, by replacing x_i by $-x_i$ when z_i is negated. In all cases we see that $v(C)$ is a linear combination of terms of the form $1 + x_i x_j$ and $1 - x_i x_j$ with *nonnegative* coefficients.

Now MAX 2SAT can be modelled as

$$\max\{\sum_{j=1}^{m} w_j v(C_j) : x_1^2 = \ldots = x_n^2 = 1\}.$$

This quadratic program is of the form (5.14). Hence Theorem 5.3.4 applies. Therefore, the approximation algorithm of Goemans and Williamson gives a 0.878 approximation for MAX 2SAT.

5.4 Further reading and remarks

We start with an anecdote. About the finding of the approximation ratio 0.878, 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.

For their work [4], Goemans and Williamson won in 2000 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.

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 apart or is this value 0.878, which maybe looks strange at first sight, only an artefact of our analysis? It turns out that the value is optimal. In 2002 Feige and Schechtmann gave an infinite family of graphs for which the ratio mc/sdp converges to exactly $0.878\ldots$. This proof uses a lot of nice mathematics (continuous graphs, Voronoi regions, isoperimetric inequality) and it is explained in detail in the Chapter 8 of the book *Approximation Algorithms and Semidefinite Programming* of Gärtner and Matoušek.

In 2007, Khot, Kindler, Mossel, O'Donnell showed that the algorithm of Goemans and Williamson 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 unless $\mathrm{P} = \mathrm{NP}$. Currently, the validity and the implications of the unique games conjecture are under heavy investigation. The book of Gärtner and Matoušek also contains an introduction to the unique games conjecture.

5.5 Exercises

5.1 The goal of this exercise is to show that the maximum weight stable set problem can be formulated as an instance of the maximum cut problem.

Let $G = (V, E)$ be a graph with node weights $c \in \mathbb{R}^V_+$. Define the new graph $G' = (V', E')$ with node set $V' = V \cup \{0\}$, with edge set $E' = E \cup \{\{0, i\} : i \in V\}$, and with edge weights $w \in \mathbb{R}^{E'}_+$ defined by

$$w_{0i} = c_i - \deg_G(i)M \text{ for } i \in V, \text{ and } w_{ij} = M \text{ for } \{i,j\} \in E.$$

Here, $\deg_G(i)$ denotes the degree of node i in G, and M is a constant to be determined.

(a) Let $S \subseteq V$. Show: $w(S, V' \setminus S) = c(S) - 2M|E(S)|$.

(b) Show: If M is sufficiently large, then $S \subseteq V$ is a stable set of maximum weight in (G, c) if and only if $(S, V' \setminus S)$ is a cut of maximum weight in (G', w).

Give an explicit value of M for which the above holds.

5.2 Let $G = (V = [n], E)$ be a graph with edge weights $w \in \mathbb{R}^E$. Define the Laplacian matrix $L_w \in \mathcal{S}^n$ by: $L_{ii} = \sum_{j \in V : \{i,j\} \in E} w_{ij}$ for $i \in V$, $L_{ij} = -w_{ij}$ if $\{i, j\} \in E$, and $L_{ij} = 0$ otherwise.

(a) Show: $x^{\mathsf{T}} L_w x = 2 \cdot \sum_{\{i,j\} \in E} w_{ij}(1 - x_i x_j)$ for any vector $x \in \{\pm 1\}^n$.

(b) Show: If $w \geq 0$ then $L_w \succeq 0$.

(c) Given an example of weights w for which L_w is not positive semidefinite.

5.3 Let $G = (V = [n], E)$ be a graph and let $w \in \mathbb{R}^E_+$ be nonnegative edge weights.

(a) Show the following reformulation for the MAX CUT problem:

$$\mathrm{mc}(G, w) = \max \left\{ \sum_{\{i,j\} \in E} w_{ij} \frac{\arccos(v_i^{\mathsf{T}} v_j)}{\pi} : v_1, \ldots, v_n \text{ unit vectors in } \mathbb{R}^n \right\}.$$

Hint: Use the analysis of the Goemans-Williamson algorithm.

(b) Let $v_1, \ldots, v_7$ be unit vectors. Show:

$$\sum_{1 \leq i < j \leq 7} \arccos(v_i^{\mathsf{T}} v_j) \leq 12\pi.$$

5.4 For a matrix $A \in \mathbb{R}^{m \times n}$ we define the following quantities:

$$f(A) = \max_{I \subseteq [m], J \subseteq [n]} |\sum_{i \in I} \sum_{j \in J} A_{ij}|,$$

called the cut norm of A, and

$$g(A) = \max \left\{ \sum_{i \in [m]} \sum_{j \in [n]} A_{ij} x_i y_j : x_1, \ldots, x_m, y_1, \ldots, y_n \in \{\pm 1\} \right\}.$$

(a) Show: $f(A) \leq g(A) \leq 4f(A)$.

(b) Assume that all row sums and all column sums of A are equal to 0. Show: $g(A) = 4f(A)$.

(c) Formulate a semidefinite programming relaxation for $g(A)$.

(d) Show:

$$g(A) = \max \left\{ \sum_{i \in [m]} \sum_{j \in [n]} A_{ij} x_i y_j : x_1, \ldots, x_m, y_1, \ldots, y_n \in [-1, 1] \right\}.$$

(e) Assume that A is a symmetric positive semidefinite $n \times n$ matrix. Show:

$$g(A) = \max \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j : x_1, \ldots, x_n \in \{\pm 1\} \right\}.$$

(f) Show that the maximum cut problem in a graph $G = ([n], E)$ with nonnegative edge weights can be formulated as an instance of computing the cut norm $f(A)$ of some matrix A.

5.5 Let $G = C_5$ denote the cycle on 5 nodes. Compute the Goemans-Williamson semidefinite relaxation for max-cut (where all edge weights are taken equal to 1):

$$\text{sdp}(C_5) = \max\left\{\frac{1}{2}\sum_{i=1}^{5}(1 - X_{i,i+1}) : X \in \mathcal{S}^5, X \succeq 0, X_{ii} = 1\ \forall i \in [5]\right\}.$$

How does the ratio $\frac{\text{mc}(C_5)}{\text{sdp}(C_5)}$ compare to the GW ratio 0.878?

5.6 Given a vector $y = (y_{ij,ijkl})$ indexed by all subsets of $V = \{1, \ldots, n\}$ of cardinality 2 or 4, consider the symmetric matrix $M(y)$ of size $1 + \binom{n}{2}$ indexed by all pairs of V and an additional index denoted 0, with entries:

$$M(y)_{0,0} = M(y)_{ij,ij} = 1,\ M(y)_{ij,ik} = y_{jk},\ M(y)_{ij,kl} = y_{ijkl}$$

for all distinct $i, j, k, l \in V$. (Here y_{ij}, y_{ijkl} denote the cooordinates of y indexed by the subsets $\{i, j\}$, $\{i, j, k, l\}$, resp.).

Given a graph $G = (V, E)$ with edge weights w, consider the SDP:

$$\text{sdp}_2(G, w) = \max_{y} \sum_{ij \in E} w_{ij}\frac{1 - y_{ij}}{2} \ \text{ s.t. } M(y) \succeq 0. \tag{5.16}$$

(a) Show that (5.16) is a relaxation of max-cut: $\text{mc}(G, w) \leq \text{sdp}_2(G, w)$.

(b) Show that the bound given by (5.16) is at least as good as the basic bound $\text{sdp}(G, w)$ of (5.6): $\text{sdp}_2(G, w) \leq \text{sdp}(G, w)$.

(c) Show that $M(y) \succeq 0$ implies the following (triangle) inequalities:

$$y_{ij} + y_{ik} + y_{jk} \geq -1,\ y_{ij} - y_{ik} - y_{jk} \geq -1$$

for all distinct $i, j, k \in V$.

BIBLIOGRAPHY

[1] F. Barahona. The max-cut problem in graphs not contractible to K_5. *Operations Research Letters* **2**:107–111, 1983.

[2] M. Deza and M. Laurent. *Geometry of Cuts and Metrics.* Springer, 1997.

[3] M.R. Garey and D.S. Johnson. *Computers and Intractability - A guide to the Theory of NP-Completeness*. Freeman, 1979.

[4] M.X. Goemans, D.P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. ACM **42**:1115–1145, 1995.

[5] Y. Nesterov. Quality of semidefinite relaxation for nonconvex quadratic optimization. *CORE Discussion Paper*, Number 9719, 1997.

[6] S. Poljak and Z. Tuza. The expected relative error of the polyhedral approximation of the max-cut problem. *Operations Research Letters* **16**:191–1998, 1994.

[7] A. Schrijver. *A Course in Combinatorial Optimization*. Lecture Notes. Available at `http://homepages.cwi.nl/~lex/files/dict.pdf`

CHAPTER 6

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 to 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 $G = (V, E)$ is the minimum rank of a positive semidefinite matrix X with nonzero diagonal entries 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.

6.1 Geometry of the positive semidefinite cone

6.1.1 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 $F_K(x)$. The set $P_K(x)$ of perturbations of $x \in K$ is a linear space, whose dimension is equal to the dimension of the face $F_K(x)$.

Lemma 6.1.1. *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

$$c^{\mathsf{T}} x \geq \gamma,\ c^{\mathsf{T}} y \leq \gamma\ \forall y \in F_K(x), \text{ and } F_K(x) \not\subseteq H_{c,\gamma}.$$

We may assume that $\gamma = c^{\mathsf{T}} x$. 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.

6.1.2 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 $0 \leq r \leq n$.

Proposition 6.1.2. *Let $A \in \mathcal{S}^n_{\succeq 0}$, $r = \text{rank}(A)$, 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*

$$\begin{array}{rcl} \phi_A : \ \mathcal{S}^r & \to & \mathcal{S}^n \\ Z & \mapsto & U \begin{pmatrix} Z & 0 \\ 0 & 0 \end{pmatrix} U^{\mathsf{T}} = U_0 Z U_0^{\mathsf{T}} \end{array} \tag{6.1}$$

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

$$F(A) = \{X \in \mathcal{S}^n_{\succeq 0} : \text{Ker}X \supseteq \text{Ker}A\} \tag{6.2}$$

and its dimension is equal to $\binom{r+1}{2}$.

Proof. Set $D = \text{diag}(\lambda_1, \cdots, \lambda_r, 0, \cdots, 0) \in \mathcal{S}^n_{\succeq 0}$, $D_0 = \text{diag}(\lambda_1, \cdots, \lambda_r) \in \mathcal{S}^r_{\succ 0}$, where λ_i is the eigenvalue for eigenvector u_i, $\Delta = \text{diag}(0, \cdots, 0, 1, \cdots, 1) \in \mathcal{S}^n_{\succeq 0}$, where the first r entries are 0 and the last $n - r$ entries are 1. Finally, set $Q = U \Delta U^{\mathsf{T}} = \sum_{i=r+1}^n u_i u_i^{\mathsf{T}}$. Then, $A = UDU^{\mathsf{T}}$ and $\langle \Delta, D \rangle = 0$. 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. We claim that

$$F = \{X \in \mathcal{S}^n_{\succeq 0} : \text{Ker}X \supseteq \text{Ker}A\}.$$

Indeed, the condition $\langle Q, X \rangle = 0$ reads $\sum_{i=r+1} u_i^{\mathsf{T}} X u_i = 0$. For $X \succeq 0$, $u_i^{\mathsf{T}} X u_i \geq 0$ 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., $\text{Ker}A \subseteq \text{Ker}X$. Moreover,

$$F = \phi_A(\mathcal{S}^r_{\succeq 0}).$$

For this, consider $X \in \mathcal{S}^n$ written as $X = UYU^{\mathsf{T}}$ where $Y \in \mathcal{S}^n$. Then, $X \succeq 0$ if and only if $Y \succeq 0$. Moreover, $\langle Q, X\rangle = 0$ if and only if $\langle \Delta, Y\rangle = 0$ or, equivalently, $Y = \phi_A(Z)$ for some $Z \in \mathcal{S}^r$. Summarizing, $X \in F$ if and only if $X = \phi_A(Z)$ for some $Z \in \mathcal{S}^r_{\succeq 0}$.

We now show that $F = F(A)$. In view of Lemma 6.1.1, it suffices to show that A lies in the relative interior of the face F. We use Lemma 1.2.1: let $X \in F$, we show that there exist $X' \in F$ and a scalar $t \in (0,1)$ such that $A = tX + (1-t)X'$. As we just saw above, $X = \phi_A(Z)$ for some $Z \in \mathcal{S}^r_{\succeq 0}$. As Λ_0 is an interior point of $\mathcal{S}^r_{\succeq 0}$, there exists $Z' \in \mathcal{S}^r_{\succeq 0}$ and $t \in (0,1)$ such that $\Lambda_0 = tZ + (1-t)Z'$. Then, $X' = \phi_A(Z') \in F$ and $A = tX + (1-t)X'$, as required.

Summarizing, we have shown that $F(A)$ can be identified with $\mathcal{S}^r_{\succeq 0}$ via the rank-preserving isometry:

$$\begin{array}{lclcl} Z & \mapsto & Y = \begin{pmatrix} Z & 0 \\ 0 & 0 \end{pmatrix} & \mapsto & X = UYU^{\mathsf{T}} \\ D_0 & \mapsto & D & \mapsto & A \\ \mathcal{S}^r_{\succeq 0} & \to & F' & \to & F(A) \end{array}$$

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 $\mathbb{R}^n$:

$$U \text{ subspace of } \mathbb{R}^n \mapsto F_U = \{X \in \mathcal{S}^n_{\succeq 0} : \text{Ker}X \supseteq U\}, \tag{6.3}$$

with $U_1 \subseteq U_2 \Longleftrightarrow F_{U_1} \supseteq F_{U_2}$.

6.1.3 Faces of spectrahedra

Consider an affine subpsace $\mathcal{A}$ in the space of symmetric matrices, of the form

$$\mathcal{A} = \{X \in \mathcal{S}^n : \langle A_j, X\rangle = b_j \ (j \in [m])\}, \tag{6.4}$$

where $A_1, \cdots, A_m$ are given symmetric matrices and $b_1, \cdots, b_m$ are given scalars. Assume that $\mathcal{A}$ is not empty. The *codimension* of $\mathcal{A}$ is

$$\text{codim } \mathcal{A} = \dim \mathcal{S}^n - \dim \mathcal{A} = \dim\langle A_1, \cdots, A_m\rangle,$$

where $\langle A_1, \cdots, A_m\rangle$ denotes the linear subspace of $\mathcal{S}^n$ spanned by $\{A_1, \ldots, A_m\}$.

If we intersect the cone of positive semidefinite matrices with the affine space $\mathcal{A}$, we obtain the convex set

$$K = \mathcal{S}^n_{\succeq 0} \cap \mathcal{A} = \{X \in \mathcal{S}^n : X \succeq 0, \ \langle A_j, X\rangle = b_j \ (j \in [m])\}. \tag{6.5}$$

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 spectrahedron is the *elliptope*

$$\mathcal{E}_n = \{X \in \mathcal{S}^n_{\succeq 0} : X_{ii} = 1 \ \forall i \in [n]\}, \tag{6.6}$$

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 6.1.2, we can describe the faces of K.

Proposition 6.1.3. *Let K be the spectrahedron (9.2). Let $A \in K$, $r = \text{rank}(A)$, and let U, U_0 be as in Proposition 6.1.2. Define the affine space in $\mathcal{S}^r$:*

$$\mathcal{A}_A = \{Z \in \mathcal{S}^r : \langle U_0^\mathsf{T} A_j U_0, Z\rangle = b_j \ \forall j \in [m]\}, \tag{6.7}$$

and the corresponding linear space:

$$\mathcal{L}_A = \{Z \in \mathcal{S}^r : \langle U_0^\mathsf{T} A_j U_0, Z\rangle = 0 \ \forall j \in [m]\}. \tag{6.8}$$

The map ϕ_A from (6.1) identifies $F_K(A)$ and $\mathcal{S}^r_{\succeq 0} \cap \mathcal{A}_A$:

$$F_K(A) = \phi_A(\mathcal{S}^r_{\succeq 0} \cap \mathcal{A}_A)$$

and the set of perturbations of $A \in K$ is

$$P_K(A) = \phi_A(\mathcal{L}_A).$$

Moreover, $F_K(A)$ is given by

$$F_K(A) = \{X \in K : \text{Ker}X \supseteq \text{Ker}A\} \tag{6.9}$$

and its dimension is equal to

$$\dim F_K(A) = \dim \mathcal{A}_A = \binom{r+1}{2} - \dim\langle U_0^\mathsf{T} A_j U_0 : j \in [m]\rangle. \tag{6.10}$$

Proof. Recall that $K = \mathcal{S}^n_{\succeq 0} \cap \mathcal{A}$ and that $F(A)$ denotes the smallest face of $\mathcal{S}^n_{\succeq 0}$ containing A. One can verify that the set $F(A) \cap \mathcal{A}$ contains A in its relative interior and thus we have that $F_K(A) = F(A) \cap \mathcal{A}$. Hence (6.9) follows from (6.2).

If $X = \phi_A(Z)$ is the image of $Z \in \mathcal{S}^r$ under the map ϕ_A from (6.1), 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_A, i.e., $F_K(A) = \phi_A(\mathcal{S}^r_{\succeq 0} \cap \mathcal{A}_A)$. Moreover, a matrix $B \in \mathcal{S}^n$ is a perturbation of A if and only if $A \pm \epsilon B \in K$ for some $\epsilon > 0$, which is equivalent to $B \in U_0 \mathcal{L}_A U_0^\mathsf{T}$, i.e., $B \in \phi_A(\mathcal{L}_A)$. Therefore, we find that $P_K(A) = \phi_A(\mathcal{L}_A)$, and thus the dimension of $F_K(A)$ is equal to $\dim P_K(A) = \dim \phi_A(\mathcal{L}_A) = \dim \mathcal{A}_A$, which gives (6.10). □

Corollary 6.1.4. *Let K be defined as in (9.2). Let $A \in K$ and $r = \text{rank}(A)$. If A is an extreme point of K then*

$$\binom{r+1}{2} \leq \text{codim } \mathcal{A} \leq m \tag{6.11}$$

In particular, K contains a matrix A whose rank r satisfies

$$r \leq \frac{-1+\sqrt{8m+1}}{2}. \tag{6.12}$$

Proof. If A is an extreme point of K then $\dim F_K(A) = 0$. Then, by (6.10), $\binom{r+1}{2} = \text{codim}\mathcal{A}_A \leq \text{codim}\mathcal{A}$ (since, for any $J \subseteq [m]$, $\{U_0^\mathsf{T} A_j U_0 : j \in J\}$ linearly independent implies that $\{A_j : j \in J\}$ too is linearly independent). As $\text{codim}\mathcal{A} \leq m$, then (9.15) follows from (6.10).

As K contains no line, K has at least one extreme point. Now (6.12) follows directly from $\binom{r+1}{2} \leq m$ for any matrix A which is an extreme point of K. □

Remark 6.1.5. *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*

$$\text{codim } \mathcal{A}_A = \dim\langle W^\mathsf{T} A_j W : \ j \in [m]\rangle.$$

Indeed, the matrix $P = W^\mathsf{T} U_0 D_0^{-1}$ 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

$$\dim\langle W^\mathsf{T} A_j W : j \in [m]\rangle = \dim\langle P^\mathsf{T} W^\mathsf{T} A_j W P : j \in [m]\rangle = \dim\langle U_0^\mathsf{T} A_j U_0 : j \in [m]\rangle.$$

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 $\mathcal{A}_A = \dim\langle a_1 a_1^\mathsf{T}, \cdots, a_n a_n^\mathsf{T}\rangle$.

As an illustration we discuss a bit the geometry of the elliptope $\mathcal{E}_n$. As a direct application of Corollary 6.1.4, we obtain the following bound for the rank of extreme points:

Corollary 6.1.6. *Any extreme point of $\mathcal{E}_n$ has rank r satisfying $\binom{r+1}{2} \leq n$.*

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 $n \leq 4$, these are the only faces of dimension 1, the elliptope $\mathcal{E}_n$ for $n \geq 5$ has faces of dimension 1 that are not an edge between two cut matrices. You will see an example in Exercise 10.3.

Figure 6.1 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

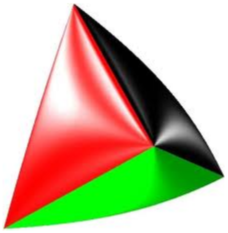


Figure 6.1: The elliptope $\mathcal{E}_3$

$\mathcal{E}_3$ - 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 $r = 2$.

6.1.4 Finding an extreme point in a spectrahedron

In order to find a matrix A in a spectrahedron K whose rank satisfies (6.12), 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 (6.8)) 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: Ker$A' \supset$ Ker$A$, and thus a smaller rank: rank$A' \leq$ 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 6.2.4.

6.1.5 A refined bound on ranks of extreme points

The upper bound on the rank of an extreme point from Corollary 6.1.4 is tight – see Example 6.2.3 below. However, there is one special case when it can be sharpened, as we explain here. Consider again the affine space $\mathcal{A}$ from (6.4) and the spectrahedron $K = \mathcal{S}^n_{\succeq 0} \cap \mathcal{A}$. From Corollary 6.1.4, we know that every extreme point A of K has rank r satisfying

$$\binom{r+1}{2} \leq \text{codim } \mathcal{A}.$$

Hence, $r \leq s+1$ if codim $\mathcal{A} = \binom{s+2}{2}$. Under some assumptions, Barvinok shows that $r \leq s$ for at least one extreme point of K.

Proposition 6.1.7. *Assume that K is nonempty bounded and* codim $\mathcal{A} = \binom{s+2}{2}$ *for some integer $s \geq 1$ satisfying $n \geq s+2$. Then there exists $A \in K$ with rank* rank $A \leq s$.

The proof uses the following topological result.

Theorem 6.1.8. *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 $n \geq 3$ 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 6.1.7.

Lemma 6.1.9. *Let $n = s+2$ with $s \geq 1$ and let $\mathcal{A} \subseteq \mathcal{S}^{s+2}$ be an affine space with* codim $\mathcal{A} = \binom{s+2}{2}$. *If $K = \mathcal{S}^{s+2}_{\succeq 0} \cap \mathcal{A}$ is nonempty and bounded, then there is a matrix $A \in K$ with* rank $A \leq s$.

Proof. Assume first that $\mathcal{A} \cap \mathcal{S}^{s+2}_{\succ 0} = \emptyset$. Then $\mathcal{A}$ lies in a hyperplane H supporting a proper face F of $\mathcal{S}^{s+2}_{\succeq 0}$. (This can be checked using the separating theorem from Theorem 1.3.8 (i).) By Proposition 6.1.2, F can be identified with $\mathcal{S}^t_{\succeq 0}$ for some $t \leq s+1$ and thus an extreme point of K has rank at most $t-1 \leq s$.

Suppose now that $\mathcal{A} \cap \mathcal{S}^{s+2}_{\succ 0} \neq \emptyset$. By (6.10), $\dim K = \binom{s+3}{2} - \text{codim } \mathcal{A} = s+2$. 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 $s+1 \geq 2$. Hence, applying Theorem 6.1.8, 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$A \leq$ rank$X - 1 \leq s$. □

We can now conclude the proof of Proposition 6.1.7.

Proof. (of Proposition 6.1.7). By Corollary 6.1.4 there exists a matrix $A \in K$ with rank $A \leq s+1$. Pick a vector space $U \subseteq \mathrm{Ker}A$ with codim $U = s+2$. By Proposition 6.1.2, there is a rank-preserving isometry between F_U and $\mathcal{S}^{s+2}_{\succeq 0}$. Moreover, $A \in F_U \cap \mathcal{A}$. Hence the result follows by applying Lemma 6.1.9. □

Example 6.1.10. *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 $\mathcal{S}^2_{\succeq 0} \cap \mathcal{A} = \{I\}$ thus contains no rank 1 matrix, and codim $\mathcal{A} = 3 = \binom{s+2}{2}$ *with $s = 1$. This example shows that the condition $n \geq s+2$ cannot be omitted in Lemma 6.1.9.*

Example 6.2.3 below shows that the assumption that K is bounded cannot be omitted as well.

6.2 Applications

6.2.1 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:

$$\exists v_1, \cdots, v_n \in \mathbb{R}^d \ \ \|v_i - v_j\|^2 = w_{ij} \ \ \forall \{i, j\} \in E. \tag{6.13}$$

(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 $d \geq 1$. 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 6.2.1. *(G, w) is realizable if and only if the following semidefinite program (in matrix variable $X \in \mathcal{S}^n$):*

$$X_{ii} + X_{jj} - 2X_{ij} = w_{ij}\ \forall \{i, j\} \in E,\ X \succeq 0 \tag{6.14}$$

has a feasible solution. Moreover, (G, w) is d-realizable if and only if the system (6.14) 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 (6.14). Conversely, if X is a solution of (6.14) of rank 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 6.1.4, any realizable graph (G, w) is d-realizable in dimension d satisfying

$$\binom{d+1}{2} \leq |E|, \text{ i.e., } d \leq \frac{-1 + \sqrt{8|E| + 1}}{2}. \tag{6.15}$$

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 6.2.2. *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 $\text{rank} X \leq d$*.*

Proof. The proof relies on the observation that if a set of vectors $v_1, \cdots, v_n \in \mathbb{R}^d$ satisfies (6.13), then one can translate it and thus assume without loss of generality that $v_n = 0$. □

Example 6.2.3. *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 6.2.2).*

Hence, the upper bound (6.15) is tight on this example. This shows that the condition that K is bounded cannot be omitted in Proposition 6.1.7. (Note that the set of feasible solutions to the program (6.14) is indeed not bounded).

On the other hand, for any fixed $d \geq 1$, deciding whether a graph (G, w) is d-realizable is a hard problem. Therefore, deciding whether the semidefinite program (6.14) 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 following *partition problem*, well known to be NP-complete: 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 6.2.4. *Given a graph (G, w) with integer lengths $w \in \mathbb{N}^E$, deciding whether (G, w) is 1-realizable 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 6.2.5. *If a circuit (C_n, w) is realizable, then it is 2-realizable.*

This can be shown (Exercise 6.1) using the following basic geometrical fact.

Lemma 6.2.6. *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 $i \in [k]$.

Hence what the above shows is that 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 $\mathcal{N}P$-complete problem!

6.2.2 Hidden convexity results for quadratic maps

As a direct application of Proposition 6.1.4, we obtain the following result for systems of two quadratic equations.

Proposition 6.2.7. *Consider two matrices $A, B \in \mathcal{S}^n$ and $a, b \in \mathbb{R}$. Then the system of two quadratic equations*

$$\sum_{i,j=1}^{n} A_{ij} x_i x_j = a, \ \ \sum_{i,j=1}^{n} B_{ij} x_i x_j = b \tag{6.16}$$

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

$$\langle A, X \rangle = a, \ \ \langle B, X \rangle = b \tag{6.17}$$

has a positive semidefinite solution $X \succeq 0$.

Proof. If x is a solution of (6.16), then $X = xx^{\mathsf{T}}$ is a solution of (6.17). Conversely, assume that the system (6.17) has a solution. Applying Corollary 6.1.4, we know that it has a solution of rank r satisfying $\binom{r+1}{2} \leq m = 2$, thus with $r \leq 1$. Now, if X has rank 1, it can be written in the form $X = xx^{\mathsf{T}}$, so that x is a solution of (6.16). □

This result does not extend to three equations: The affine space from Example 6.1.10 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 6.2.8. (Dines 1941) *Given two matrices $A, B \in \mathcal{S}^n$, the image of $\mathbb{R}^n$ under the quadratic map $q(x) = (x^\mathsf{T} Ax, x^\mathsf{T} Bx)$:*

$$\mathcal{Q} = \{(x^\mathsf{T} Ax, x^\mathsf{T} Bx) : x \in \mathbb{R}^n\}, \tag{6.18}$$

is a convex set in $\mathbb{R}^2$.

Proof. Set

$$\mathcal{Q}' = \{(\langle A, X\rangle, \langle B, X\rangle) \in \mathbb{R}^2 : X \in \mathcal{S}^n_{\succeq 0}\}.$$

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 (6.17) has a solution $X \succeq 0$. By Proposition 6.2.7, the system (6.16) 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 $\mathbf{S}^{n-1}$ ($n \geq 3$) under a quadratic map in $\mathbb{R}^2$ is convex. We show it using the refined bound from Proposition 6.1.7.

Proposition 6.2.9. (Brinkman 1961) *Let $n \geq 3$ and $A, B \in \mathcal{S}^n$. 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

$$K = \{X \in \mathcal{S}^n_{\succeq 0} : \langle A, X\rangle = a,\ \langle B, X\rangle = b,\ \mathrm{Tr}(X) = 1\}$$

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 6.1.4 if $\operatorname{codim} \mathcal{A} \leq 2$, and from Proposition 6.1.7 if $\operatorname{codim} \mathcal{A} = 3$ (as K is bounded, $\operatorname{codim} \mathcal{A} = \binom{s+2}{3}$, $n \geq s + 2$ for $s = 1$). □

The assumption $n \geq 3$ cannot be omitted in Proposition 6.2.9: Consider the quadratic map q defined using the matrices A and B from Example 6.1.10. 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 6.2.9, 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

$$R(M) = \{z^* M z = \sum_{i,j=1}^{n} \overline{z_i} M_{ij} z_i : z \in \mathbb{C}^n, \sum_{i=1}^{n} |z_i|^2 = 1\}.$$

Proposition 6.2.10. (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 6.2.9. □

6.2.3 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:

$$\begin{cases} a_1^{\mathsf{T}} x \le b_1 \\ \vdots \\ a_m^{\mathsf{T}} x \le b_m \end{cases}$$

implies the linear inequality $c^{\mathsf{T}} x \le d$, then there exist nonnegative scalars $\lambda_1, \cdots, \lambda_m \ge 0$ such that $c = \lambda_1 a_1 + \cdots + \lambda_m a_m$ and $\lambda_1 b_1 + \cdots + \lambda_m b_m \le d$.

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 6.2.11. (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) $\{x \in \mathbb{R}^n : x^{\mathsf{T}} A x \ge 0\} \subseteq \{x \in \mathbb{R}^n : x^{\mathsf{T}} B x \ge 0\}$.

(ii) *There exists a scalar $\lambda \ge 0$ 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:

$$\inf\{\langle B, X\rangle : \langle A, X\rangle \ge 0,\ \mathrm{Tr}(X) = 1,\ X \succeq 0\} \tag{P}$$

and its dual:

$$\sup\{y : B - zA - yI \succeq 0,\ z \ge 0\}. \tag{D}$$

First we show that (P) is strictly feasible. By assumption, there exists a unit vector x for which $x^{\mathsf{T}}Ax > 0$. If $\mathrm{Tr}(A) \geq 0$ then $X = xx^{\mathsf{T}}/2 + I/2n$ is a strictly feasible solution. Assume now that $\mathrm{Tr}(A) < 0$. Set $X = \alpha xx^{\mathsf{T}} + \beta I$, where we choose $\alpha \geq 0$, $\beta > 0$ in such a way that $1 = \mathrm{Tr}(X) = \alpha + \beta n$ and $0 < \langle A, X\rangle = \alpha x^{\mathsf{T}}Ax + \beta\mathrm{Tr}(A)$, i.e.,

$$0 < \beta < \min\left\{\frac{1}{n}, \frac{x^{\mathsf{T}}Ax}{nx^{\mathsf{T}}Ax - \mathrm{Tr}(A)}\right\}.$$

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

$$K = \{X \in \mathcal{S}^n_{\succeq 0} : \langle A, X\rangle = \langle A, X_0\rangle,\ \langle B, X\rangle = \langle B, X_0\rangle\}.$$

As $K \neq \emptyset$, applying Corollary 6.1.4, there is a matrix $X \in K$ with rank 1. Say $X = xx^{\mathsf{T}}$. Then, $x^{\mathsf{T}}Ax = \langle A, X_0\rangle \geq 0$ which, by assumption (i), implies $x^{\mathsf{T}}Bx \geq 0$, and thus $\langle B, X_0\rangle = x^{\mathsf{T}}Bx \geq 0$.

As (P) is bounded and strictly feasible, applying the strong duality theorem, we deduce that there is no duality gap and that the dual problem has an optimal solution (y, z) with $y, z \geq 0$. Therefore, $B - zA = (B - zA - yI) + yI \succeq 0$, thus showing (ii). □

This extends to non-homogeneous quadratic polynomials (Exercise 6.5):

Theorem 6.2.12. (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) $\{x \in \mathbb{R}^n : f(x) \geq 0\} \subseteq \{x \in \mathbb{R}^n : g(x) \geq 0\}$.

(ii) *There exists a scalar $\lambda \geq 0$ 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 $g = \lambda f + h$.*

6.3 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 6.1.5) 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 $d \leq 3$

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 $\mathcal{N}P$-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 Brinkman presented here), this is nicely exposed in the survey of Polik and Terlaky [4].

6.4 Exercises

6.1 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 6.2.3).

(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 $e \in E$, $G\backslash e = (V, E \setminus \{e\})$ 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.

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.

Belk and Connelly [2] show that $K_{2,2,2}$ is not 3-realizable, and that a graph G is 3-realizable if and only if G has no K_5 and $K_{2,2,2}$ minor. (The 'if part' requires quite some work.)

6.2 Let $A, B, C \in \mathcal{S}^n$ 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 $n \geq 3$ 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.

Hint: Use Proposition 6.1.7.

6.3 (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?

6.4 Let p be 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.

6.5 Show the result of Theorem 6.2.12.

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 7

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, $\mathbb{N}^n_d$ denotes the set of sequences $\alpha \in \mathbb{N}^n$ with $|\alpha| \leq d$, 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 $s(n,d) = |\mathbb{N}^n_d| = \binom{n+d}{d}$ and the set $\{x^\alpha : \alpha \in \mathbb{N}^n, |\alpha| \leq d\}$ of monomials of degree at most d is its canonical base.

7.1 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 7.1.1. *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:*

$$Q \succeq 0, \quad \sum_{\substack{\beta,\gamma \in \mathbb{N}^n_d \\ \beta+\gamma=\alpha}} Q_{\beta,\gamma} = p_\alpha \quad \forall \alpha \in \mathbb{N}^n_{2d}. \tag{7.1}$$

7.1.1 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 7.1.2. *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:

$$\sqrt[n]{x_1 \cdots x_n} \leq \frac{x_1 + \cdots + x_n}{n} \tag{7.2}$$

for $x_1, \cdots, x_n \geq 0$ and any $n \geq 1$. (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:*

$$K = \{x \in \mathbb{R}^n : g_1(x) \geq 0, \cdots, g_m(x) \geq 0\},$$

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:

$$\mathcal{P}(K) = \{f \in \mathbb{R}[x] : f(x) \geq 0 \ \forall x \in K\}. \tag{7.3}$$

Clearly,

$$p_{\min} = \inf_{x \in K} p(x) = \sup\{\lambda : p - \lambda \in \mathcal{P}(K)\}. \tag{7.4}$$

Computing $p_{\min}$ is hard in general.

Example 7.1.3. *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:

$$\alpha(G) = \max\left\{\sum_{i \in V} x_i : x_i + x_j \leq 1 \ \forall \{i,j\} \in E,\ x_i^2 - x_i = 0 \ \forall i \in V\right\}, \quad (7.5)$$

$$\frac{1}{\alpha(G)} = \min\left\{x^{\mathsf{T}}(A_G + I)x : \sum_{i \in V} x_i = 1,\ x \geq 0\right\}, \quad (7.6)$$

where A_G is the adjacency matrix of G. The formulation (7.6) 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:

$$p_{\mathsf{sos}} = \sup\{\lambda : p - \lambda \in \Sigma + g_1\Sigma + \cdots + g_m\Sigma\}. \quad (7.7)$$

As a direct application of Lemma 7.1.1, one can compute p_{sos} using semidefinite programming. For instance, when $K = \mathbb{R}^n$,

$$p_{\mathsf{sos}} = p_0 + \sup\left\{-Q_{00} : Q \succeq 0,\ p_\alpha = \sum_{\substack{\beta,\gamma \in \mathbb{N}_d^n \\ \beta+\gamma=\alpha}} Q_{\beta,\gamma},\ \ \forall \alpha \in \mathbb{N}_{2d}^n \setminus \{0\}\right\}. \quad (7.8)$$

Clearly the inequality holds:

$$p_{\mathsf{sos}} \leq p_{\mathsf{min}}. \quad (7.9)$$

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 7.2.9), 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 (8.1) and its sos relaxation (7.7) 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.

7.1.2 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 7.1.4. *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 $n \geq 2$ and even $d \geq 4$, 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 $(n, d) = (2, 6)$, $(3, 4)$.

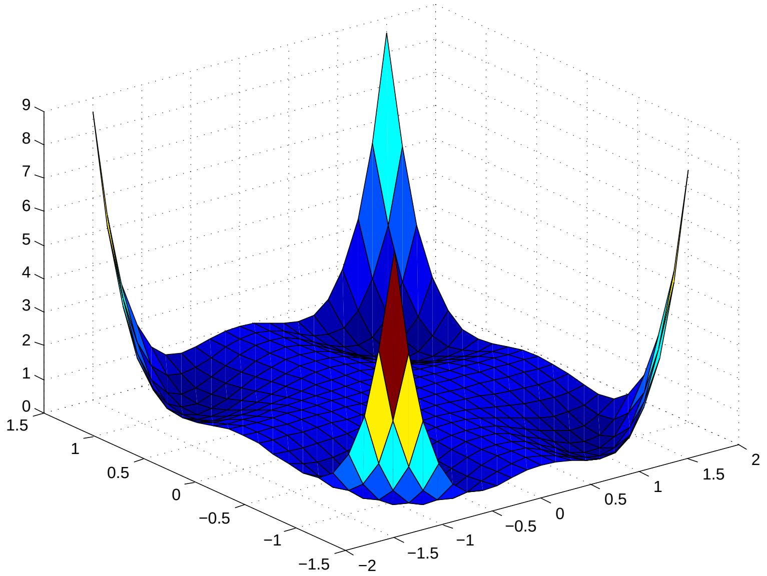


Figure 7.1: The Motzkin polynomial

Example 7.1.5. *Hilbert's proof for the 'only if' part of Theorem 7.1.4 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.

To see that p is nonnegative on $\mathbb{R}^2$, one can use the arithmetic-geometric mean inequality: $\frac{a+b+c}{3} \geq \sqrt[3]{abc}$, applied to $a = x^4y^2$, $b = x^2y^4$ and $c = 1$.

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, as the coefficients of x^2 and y^2 in p are 0, we can conclude that the coefficients of x and y in s_l are 0. Finally, say $s_l = a_l xy^2 + b_l x^2 y + c_l xy + d_l$. 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).

7.1.3 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 7.1.6. *If $p \geq 0$ 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

$$c \cdot n^{(d-1)/2} \leq \left(\frac{\text{vol}(\hat{\mathcal{P}}_{n,2d})}{\text{vol}(\hat{\Sigma}_{n,2d})} \right)^{1/D} \leq C \cdot n^{(d-1)/2}. \tag{7.10}$$

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.

7.1.4 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 7.1.7. (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.

7.2 Positivstellensätze

We now turn to the study of nonnegative polynomials p on a basic closed semi-algebraic set K, i.e., a set K of the form

$$K = \{x \in \mathbb{R}^n : g_1(x) \geq 0, \cdots, g_m(x) \geq 0\}, \tag{7.11}$$

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:

$$p \geq 0 \text{ on } K \iff p = \sum_{j=0}^{m} \lambda_j g_j \text{ for some scalars } \lambda_j \geq 0. \tag{7.12}$$

We will show the following result, due to Putinar: Assume that K is compact and satisfies the additional condition (7.17) below. Then

$$p > 0 \text{ on } K \Longrightarrow p = \sum_{j=0}^{m} s_j g_j \text{ for some polynomials } s_j \in \Sigma. \tag{7.13}$$

Of course, the following implication holds trivially:

$$p = \sum_{j=0}^{m} s_j g_j \text{ for some polynomials } s_j \in \Sigma \Longrightarrow p \geq 0 \text{ on } K.$$

However, this is not an equivalence, one needs a stronger assumption: strict positivity of p over K. Note the analogy between (7.12) and (7.13): While the variables in (7.12) are nonnegative scalars λ_i, the variables in (7.13) are sos polynomials s_i. A result of the form (7.13) 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 7.2.1. (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, $V_{\mathbb{C}}(g_1, \cdots, g_m) = \emptyset \Longleftrightarrow 1 = \sum_{j=1}^{m} u_j g_j$ 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.

7.2.1 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 7.2.2. (Pólya-Szegö) *Let p be a univariate polynomial of degree d. Then, $p \geq 0$ on $[0, \infty)$ if and only if $p = s_0 + s_1 x$ for some $s_0, s_1 \in \Sigma$ with $\deg(s_0) \leq d$ and $\deg(s_1) \leq d - 1$.*

Theorem 7.2.3. (Fekete, Markov-Lukácz) *Let p be a univariate polynomial of degree d. Assume that $p \geq 0$ on $[-1, 1]$.*

(i) *$p = s_0 + s_1(1 - x^2)$, where $s_0, s_1 \in \Sigma$, $\deg(s_0) \leq d + 1$ and $\deg(s_1) \leq d - 1$.*

(ii) *For d odd, $p = s_1(1+x) + s_2(1-x)$ where $s_1, s_2 \in \Sigma$, $\deg(s_1), \deg(s_2) \leq d - 1$.*

Note the two different representations in (i), (ii), depending on the choice of the polynomials describing the set $K = [-1, 1]$.

7.2.2 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 (7.11). 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

$$\mathbf{T}(\mathbf{g}) = \left\{ \sum_{J \subseteq [m]} s_J g_J : s_J \in \Sigma \right\} \tag{7.14}$$

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 7.2.4. *Let $K = \{x \in \mathbb{R} : g = (1-x^2)^3 \geq 0\}$ 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 7.2.5. (Krivine's Positivstellensatz) *Let K be as in (7.11) and let $p \in \mathbb{R}[x]$. The following holds.*

(i) $p > 0$ *on* $K \Longleftrightarrow pf = 1 + h$ *for some* $f, h \in \mathbf{T}(\mathbf{g})$.

(ii) $p \geq 0$ *on* $K \Longleftrightarrow pf = p^{2k} + h$ *for some* $f, h \in \mathbf{T}(\mathbf{g})$ *and* $k \in \mathbb{N}$.

(iii) $p = 0$ *on* $K \Longleftrightarrow -p^{2k} \in \mathbf{T}(\mathbf{g})$ *for some* $k \in \mathbb{N}$.

(iv) $K = \emptyset \Longleftrightarrow -1 \in \mathbf{T}(\mathbf{g})$.

In (i)-(iv) above, there is one trivial implication. For example, it is clear that $-1 \in \mathbf{T}(\mathbf{g})$ implies $K = \emptyset$. 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 7.2.6. (The Real Nullstellensatz) *Given $g_1, \cdots, g_m \in \mathbb{R}[x]$, define the real variety, consisting of their common real roots:*

$$V_{\mathbb{R}}(g_1, \cdots, g_m) = \{x \in \mathbb{R}^n : g_1(x) = 0, \cdots, g_m(x) = 0\}. \tag{7.15}$$

For a polynomial $p \in \mathbb{R}[x]$,

$$p = 0 \text{ on } V_{\mathbb{R}}(g_1, \cdots, g_m) \Longleftrightarrow p^{2k} + s = \sum_{j=1}^{m} u_j g_j \text{ for some } s \in \Sigma, u_j \in \mathbb{R}[x], k \in \mathbb{N}.$$

In particular,

$$V_{\mathbb{R}}(g_1, \cdots, g_m) = \emptyset \Longleftrightarrow -1 = s + \sum_{j=1}^{m} u_j g_j \text{ for some } s \in \Sigma, u_j \in \mathbb{R}[x].$$

The above result does not help us yet to tackle the polynomial optimization problem (8.1): 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.

7.2.3 Schmüdgen's Positivstellensatz

When K is compact, Schmüdgen [7] proved the following simpler representation result for *positive* polynomials on K.

Theorem 7.2.7. (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.

7.2.4 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*:

$$\mathbf{M}(\mathbf{g}) = \left\{ \sum_{j=0}^{m} s_j g_j : s_j \in \Sigma \right\}. \tag{7.16}$$

First we describe this additional assumption. For this consider the following conditions on the polynomials g_j defining K:

$$\exists h \in \mathbf{M}(\mathbf{g}) \;\; \{x \in \mathbb{R}^n : h(x) \geq 0\} \text{ is compact,} \tag{7.17}$$

$$\exists N \in \mathbb{N} \;\; N - \sum_{i=1}^{n} x_i^2 \in \mathbf{M}(\mathbf{g}), \tag{7.18}$$

$$\forall f \in \mathbb{R}[x] \, \exists N \in \mathbb{N} \;\; N \pm f \in \mathbf{M}(\mathbf{g}). \tag{7.19}$$

Proposition 7.2.8. *The conditions (7.17), (7.18) and (7.19) are all equivalent. If any of them holds, the quadratic module $\mathbf{M}(\mathbf{g})$ is said to be* Archimedean.

Proof. The implications (7.19) $\Longrightarrow$ (7.18) $\Longrightarrow$ (7.17) are clear. Assume (7.17) holds and let $f \in \mathbb{R}[x]$. As the set $K_0 = \{x : h(x) \geq 0\}$ 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 7.2.7, we deduce that $N \pm f \in \mathbf{T}(h) \subseteq \mathbf{M}(\mathbf{g})$. □

Clearly, (7.17) implies that K is compact. On the other hand, if K is compact, then it is contained in some ball $\{x \in \mathbb{R}^n : g_{m+1} = R^2 - \sum_{i=1}^n x_i^2 \geq 0\}$. Hence, if we know the radius R of a ball containing K, then it suffices to add the (redundant) ball constraint $g_{m+1}(x) \geq 0$ 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 7.2.9. (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 (7.17)-(7.19)). Then,*

$$p(x) > 0 \; \forall x \in K \Longrightarrow p \in \mathbf{M}(\mathbf{g}).$$

Example 7.2.10. *Consider the simplex $K = \{x \in \mathbb{R}^n : x \geq 0, \sum_{i=1}^n x_i \leq 1\}$ 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 7.2.11. *Consider the cube $K = [01]^n = \{x \in \mathbb{R}^n : 0 \leq x_i \leq 1 \; \forall i \in [n]\}$ 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 $n - \sum_i x_i^2 \in M$.*

7.2.5 Proof of Putinar's Positivstellensatz

In this section we give a full proof for Theorem 7.2.9. 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 7.2.12. *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 7.2.13. *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 7.2.14. *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 (7.16) is a quadratic module (the* quadratic module generated by the g_j's*).*

We start with some technical lemmas.

Lemma 7.2.15. *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 7.2.16. *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 7.2.15), we get $s_1 f \in I \subseteq M$ and therefore $-1 = g_1 + s_1 f \in M$, contradicting M proper. □

Lemma 7.2.17. *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 $a \leq b$ 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 $a_0 \leq b_0$. 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 7.2.16.

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 $N - s \in M$. As $a_0 = \sup A$, there exists ϵ such that $0 < \epsilon < 1/N$ and $a_0 - \epsilon \in A$. Then, $f - (a_0 - \epsilon) = (f - a_0) + \epsilon \in M$ and thus

$$-1 + \epsilon s = g + (f - a_0 + \epsilon)s \in M.$$

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 7.2.18. *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 $p(a) \leq 0$, 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 7.2.17 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 constitute a vector $a \in \mathbb{R}^n$. As I is an ideal, this implies that

$$f - f(a) \in I \quad \forall f \in \mathbb{R}[x]. \tag{7.20}$$

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 $|\alpha| \geq 0$. If $\alpha = 0$ there is nothing to prove. Otherwise, say $\alpha_1 \geq 1$ 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 (7.20) to each of the polynomials $f = g_j$ defining K 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 $g_j(a) \geq 0$ 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 $-p(a) \geq 0$, yielding a contradiction because $p > 0$ on K. □

Lemma 7.2.19. *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 7.2.18. 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

$$2k - s,\ 2k - s^2p - 1 \in \mathbf{M}(\mathbf{g}).$$

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,

$$hp - 1 = s(2k - s)p - 1 = 2k(sp - 1) + (2k - s^2p - 1) \in \mathbf{M}(\mathbf{g}),$$

since $sp - 1, 2k - s^2p - 1 \in \mathbf{M}(\mathbf{g})$. □

[1] Zorn's lemma states the following: Let $(P, \leq)$ 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 7.2.9. Assume $p > 0$ on K. Let h and N satisfy the conclusion of Lemma 7.2.19. We may assume that $N > 0$. Moreover let $k \in \mathbb{N}$ such that $k + p \in \mathbf{M}(\mathbf{g})$ (such k exists since $\mathbf{M}(\mathbf{g})$ is Archimedean). Then,

$$\left(k - \frac{1}{N}\right) + p = \frac{1}{N}\left((N - h)(k + p) + (hp - 1) + kh\right) \in \mathbf{M}(\mathbf{g}).$$

So what we have just shown is that

$$k + p \in \mathbf{M}(\mathbf{g}) \Longrightarrow (k - 1/N) + p \in \mathbf{M}(\mathbf{g}).$$

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 7.2.9.

7.3 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 $(F, \leq)$ 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 7.2.5). 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 7.1.6). Blekherman [1] proved the inequalities (7.10) relating the volumes of the cones of sums of squares and of nonnegative polynomials.

7.4 Exercises

7.1. Let $f(x_1, \ldots, x_n) = \sum_{\alpha:|\alpha|\leq 2d} f_\alpha x^\alpha$ be an n-variate polynomial of degree $2d$ and let $F(x_1, \ldots, x_n, t) = \sum_{\alpha:|\alpha|\leq 2d} f_\alpha x^\alpha t^{2d-|\alpha|}$ be the corresponding homogeneous $(n+1)$-variate polynomial (in the $n+1$ variables $x_1, \ldots, x_n, t$).

(a) Show: $f(x) \geq 0$ for all $x \in \mathbb{R}^n \iff F(x,t) \geq 0$ for all $(x,t) \in \mathbb{R}^{n+1}$.

(b) Show: f is a sum of squares of polynomials in $\mathbb{R}[x_1, \ldots, x_n] \iff F$ is a sum of squares of polynomials in $\mathbb{R}[x_1, \ldots, x_n, t]$.

7.2. 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) Show: For $n = 2$, $F_{n,2d}(a,x)$ is sum of two squares of polynomials.

(b) 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 $i \in [n]$.

(c) With a, b, c as in (b), show that

$$F_{n,2d}(a,x) = \frac{1}{2}(F_{n,2d}(2b,x) + F_{n,2d}(2c,x)) + d(x^b - x^c)^2.$$

(d) 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.

(e) Show the arithmetic-geometric mean inequality (7.2) for any $n \in \mathbb{N}$.

7.3 Show Theorem 7.2.2: A univariate polynomial p of degree d is nonnegative on $[0, \infty)$ if and only if $p = s_0 + s_1 x$ for some $s_0, s_1 \in \Sigma$ with $\deg(s_0), \deg(s_1 x) \leq d$.

7.4 For a univariate polynomial f of degree d define the following univariate polynomial $G(f)$, known as its Goursat transform:

$$G(f)(x) = (1+x)^d f\left(\frac{1-x}{1+x}\right).$$

(a) Show that $f \geq 0$ on $[-1, 1]$ if and only if $G(f) \geq 0$ on $[0, \infty)$.

(b) Show Theorem 7.2.3.

7.5 Show the Real Nullstellensatz (Theorem 7.2.6) (you may use Theorem 7.2.5).

7.6 Let $G = (V, E)$ be a graph. The goal is to show Motzkin's formulation (7.6) for the stability number $\alpha(G)$. Set

$$\mu = \min \left\{ x^{\mathsf{T}} (A_G + I) x : \sum_{i \in V} x_i = 1, \ x \geq 0 \right\}. \tag{7.21}$$

(a) Show that $\mu \leq 1/\alpha(G)$.

(b) Let x be an optimal solution of the program (7.21), $S = \{i : x_i \neq 0\}$ denotes its support. Show that $\mu \geq 1/\alpha(G)$ if S is a stable set in G.

(c) Show that the program (7.21) has an optimal solution x whose support is a stable set. Conclude that (7.6) 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 8

POLYNOMIAL EQUATIONS AND MOMENT MATRICES

Consider the polynomial optimization problem:

$$p_{\min} = \inf_{x \in K} p(x), \tag{8.1}$$

which asks for the infimum $p_{\min}$ of a polynomial p over a basic closed semi-algebraic set K, of the form:

$$K = \{x \in \mathbb{R}^n : g_1(x) \geq 0, \cdots, g_m(x) \geq 0\} \tag{8.2}$$

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 $x \in K$:

$$\mathcal{C}_\infty(K) = \operatorname{conv}\{[x]_\infty : x \in K\}. \tag{8.3}$$

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 (8.1) equivalently as

$$p_{\min} = \inf_{x\in K} p(x) = \inf_{x\in K} \boldsymbol{p}^{\mathsf{T}}[x]_\infty = \inf_{y=(y_\alpha)_{\alpha\in\mathbb{N}^n}} \{\boldsymbol{p}^{\mathsf{T}} y : y \in \mathcal{C}_\infty(K)\}. \tag{8.4}$$

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 (8.1).

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 (8.1). 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.

8.1 The quotient algebra $\mathbb{R}[x]/I$

8.1.1 (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

$$\sqrt{I} = \{f \in \mathbb{K}[x] : f^m \in I \text{ for some } m \in \mathbb{N}\} \tag{8.5}$$

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 8.1.1. (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

$$\sqrt[\mathbb{R}]{I} = \{f \in \mathbb{R}[x] : f^{2m} + s \in I \quad \text{for some } m \in \mathbb{N}, s \in \Sigma\} \tag{8.6}$$

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 8.1.2. (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 8.1.3.

(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*

$$\forall f_1, \cdots, f_m \in \mathbb{R}[x] \;\; f_1^2 + \cdots + f_m^2 \in I \Longrightarrow f_1, \cdots, f_m \in I.$$

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 $V = \mathbb{C} \setminus \{0\} \subseteq \mathbb{C}$, $\mathcal{I}(V) = \{0\}$, 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 8.1.4. *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). □

8.1.2 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 8.1.5. *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 $|V_{\mathbb{C}}(I)| \leq \dim \mathcal{A}$, with equality if and only if the ideal I is radical (i.e., $I = \sqrt{I}$).*

Remark 8.1.6. *Let I be an ideal in $\mathbb{R}[x]$. Then the set*

$$I_{\mathbb{C}} := I + \mathbf{i}I = \{f + \mathbf{i}g : f, g \in I\}$$

is an ideal in $\mathbb{C}[x]$. Moreover, the two quotient spaces $\mathbb{R}[x]/I$ and $\mathbb{C}[x]/I_{\mathbb{C}}$ have the same dimension. Indeed, if $f_1, \ldots, f_r \in \mathbb{R}[x]$ are real polynomials whose cosets in $\mathbb{R}[x]/I$ form a basis of $\mathbb{R}[x]/I$, then their cosets in $\mathbb{C}[x]/I_{\mathbb{C}}$ form a basis of $\mathbb{C}[x]/I_{\mathbb{C}}$. Hence, in order to compute the dimension of $\mathbb{R}[x]/I$, we can as well deal with the corresponding ideal $I_{\mathbb{C}} = I + \mathbf{i}I$ in the complex polynomial ring.

For the proof of Theorem 8.1.5, it is useful to have the following construction of interpolation polynomials.

Lemma 8.1.7. *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 $f \in \mathbb{K}[x]$,*

$$f - \sum_{v \in V_{\mathbb{C}}(I)} f(v) p_v \in \mathcal{I}(V_{\mathbb{C}}(I)). \tag{8.7}$$

Proof. Fix $v \in V$. For any $u \in V \setminus \{v\}$, 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

$$p_v = \prod_{u \in V \setminus \{v\}} \frac{x_{i_u} - u_{i_u}}{v_{i_u} - u_{i_u}}.$$

Clearly, $p_v(v) = 1$ and $p_v(u) = 0$ if $u \in V$, $u \neq v$. By construction the polynomial in (8.7) vanishes at all $v \in V_{\mathbb{C}}(I)$ and thus belongs to $\mathcal{I}(V_{\mathbb{C}}(I))$. □

Example 8.1.8. *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 8.1.9. *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 $u \in V_{\mathbb{C}}(I) \setminus \{v\}$, we deduce that $\lambda_v = 0$ for all v. This shows that $\mathcal{L}$ is linearly independent in $\mathcal{A}$.

(ii) Relation (8.7) 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 8.1.5). In view of Remark 8.1.6, 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 $(0 \leq h \leq k)$ (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 $\{[x_i^h] : 0 \leq h \leq n_i - 1\}$. From this one can easily derive that the set $\{[x^\alpha] : 0 \leq \alpha_i \leq n_i - 1, i \in [n]\}$ generates the vector space $\mathcal{A}$, thus showing that $\dim \mathcal{A} < \infty$.

(ii) Assume $|V_{\mathbb{C}}(I)| < \infty$. Lemma 8.1.9 (i) shows that $|V_{\mathbb{C}}(I)| \leq \dim \mathcal{A}$. If I is radical then the equality $\dim \mathcal{A} = |V_{\mathbb{C}}(I)|$ follows from Lemma 8.1.9 (iii). Assume now that I is not radical and let $f \in \sqrt{I} \setminus I$. 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 $\dim \mathcal{A} \geq |V_{\mathbb{C}}(I)| + 1$. □

8.1.3 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:

$$\begin{array}{rccc} m_h : & \mathcal{A} & \to & \mathcal{A} \\ & [f] & \mapsto & [fh]. \end{array} \tag{8.8}$$

As $V_{\mathbb{C}}(I)$ is finite we known from Theorem 8.1.5 that the vector space $\mathcal{A}$ has finite dimension. Say, $N = \dim \mathcal{A}$, then $N \geq |V_{\mathbb{C}}(I)|$, with equality if I is radical (by Theorem 8.1.5).

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 8.1.10. *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 8.1.11. *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 (8.8).*

(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.,*

$$M_h^{\mathsf{T}} [v]_{\mathcal{B}} = h(v)[v]_{\mathcal{B}}. \tag{8.9}$$

(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

$$[hb_j] = \sum_{i=1}^N a_{ij}[b_i], \text{ i.e., } hb_j - \sum_{i=1}^N a_{ij} b_i \in I.$$

Evaluating the above polynomial at $v \in V_{\mathbb{C}}(I)$ gives directly relation (8.9).

(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, $V_{\mathbb{C}}(J) = \emptyset$. 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 $m_u(m_h - \lambda \mathrm{id}) = \mathrm{id}$, where id is the identity map from $\mathcal{A}$ to $\mathcal{A}$. But then $m_h - \lambda \mathrm{id}$ 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 (8.7), 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 8.1.11 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 8.1.12. *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*

$$M_x = \begin{array}{c} \\ \begin{matrix} [1] \\ [x] \\ [x^2] \end{matrix} \end{array} \begin{array}{c} \begin{matrix} [x] & [x^2] & [x^3] \end{matrix} \\ \begin{pmatrix} 0 & 0 & 6 \\ 1 & 0 & -11 \\ 0 & 1 & 6 \end{pmatrix} \end{array}$$

(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

$$\begin{array}{c} \\ \begin{matrix} [p_1] \\ [p_2] \\ [p_3] \end{matrix} \end{array} \begin{array}{c} \begin{matrix} [xp_1] & [xp_2] & [xp_3] \end{matrix} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \end{array},$$

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 8.1.13. *Let I be an ideal in $\mathbb{R}[x]$ with $|V_{\mathbb{C}}(I)| < \infty$. Define the Hermite quadratic form:*

$$\begin{array}{rlcl} \mathcal{H}: & \mathbb{R}[x]/I \times \mathbb{R}[x]/I & \rightarrow & \mathbb{R} \\ & ([f],[g]) & \mapsto & \mathrm{Tr}(m_{fg}), \end{array} \tag{8.10}$$

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

$$\sigma_+(\mathcal{H}) - \sigma_-(\mathcal{H}) = |V_{\mathbb{R}}(I)|.$$

8.2 Characterizing the set $\mathcal{C}_\infty(K)$

Our goal in this section is to characterize the set $\mathcal{C}_\infty(K)$ from (8.3). We need one more ingredient: moment matrices.

8.2.1 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:

$$\begin{array}{rrcl} L: & \mathbb{R}[x] & \rightarrow & \mathbb{R} \\ & x^\alpha & \mapsto & L(x^\alpha) = y_\alpha \\ & f = \sum_\alpha f_\alpha x^\alpha & \mapsto & L(f) = \sum_\alpha f_\alpha y_\alpha. \end{array} \tag{8.11}$$

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 8.2.1. *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 8.2.2. *Let $y = (y_\alpha)_{\alpha\in\mathbb{N}^n}$ be a sequence of real numbers and let L be the associated linear functional from (8.11). For any polynomials $f, g \in \mathbb{R}[x]$:*

$$L(f^2) = \boldsymbol{f}^\mathsf{T} M(y)\boldsymbol{f}, \;\; L(gf^2) = \boldsymbol{f}^\mathsf{T} M(g * y)\boldsymbol{f},$$

where $g * y \in \mathbb{R}^{\mathbb{N}^n}$ is the new sequence with α-th entry

$$(g * y)_\alpha = L(gx^\alpha) = \sum_\gamma g_\gamma y_{\alpha+\gamma} \;\; \forall \alpha \in \mathbb{N}^n.$$

Therefore, $L \geq 0$ on Σ if and only if $M(y) \succeq 0$, and $L \geq 0$ 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 8.2.3. *Let $y = (y_\alpha)_{\alpha\in\mathbb{N}^n}$ be a sequence of real numbers and let L be the associated linear functional from (8.11). Set*

$$I = \{f \in \mathbb{R}[x] : L(fh) = 0 \;\forall h \in \mathbb{R}[x]\}. \tag{8.12}$$

(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 8.2.2 and the fact that $M(y) \succeq 0$, the following holds for any polynomial f:

$$L(f^2) = \boldsymbol{f}^\mathsf{T} M(y)\boldsymbol{f} \geq 0 \;\text{ and } L(f^2) = 0 \Longrightarrow M(y)\boldsymbol{f} = 0 \Longrightarrow f \in I.$$

We now show that I is real radical, using the characterization from Lemma 8.1.3: 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. □

8.2.2 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 8.2.4. *Let K be the set from (9.2). Let $y = (y_\alpha)_{\alpha\in\mathbb{N}^n}$ be a sequence of real numbers and let L be the linear functional from (8.11). 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 $v_1, \dots, v_r \in K$ and for some scalars $\lambda_1, \dots, \lambda_r > 0$ with $\sum_{i=1}^r \lambda_i = 1$.*

(ii) rank *$M(y) < \infty$, $M(y) \succeq 0$, $M(g_j * y) \succeq 0$ for $j \in [m]$, and $y_0 = 1$.*

(iii) rank *$M(y) < \infty$, $L \geq 0$ on $\Sigma + g_1\Sigma + \cdots + g_m\Sigma$, and $L(1) = 1$.*

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 $g_j(v_i) \geq 0$. 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 8.2.2.

We now show the implication (ii) $\Longrightarrow$ (i), which is the core of Theorem 8.2.4. 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 (8.11) and let I be the set from (8.12). By Lemma 9.3.1, 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 8.1.9 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 8.1.4, 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

$$L = \sum_{i=1}^r L(p_{v_i}) L_{v_i}, \text{ i.e., } y = \sum_{i=1}^r L(p_{v_i})[v_i]_\infty, \tag{8.13}$$

where L_{v_i} is the evaluation at v_i. For this, set $L' = \sum_{i=1}^r L(p_{v_i})L_{v_i}$; we show that $L = L'$. As both L and L' 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 8.1.9, 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 (8.13) 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, $L(p_{v_i}) \geq 0$ (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 $g_j(v_i) \geq 0$ for all $j \in [m]$, $i \in [r]$. For this, we use the fact that $L(g_j p_{v_i}^2) \geq 0$, since $M(g_j * y) \succeq 0$. Indeed, using (8.13), we get:

$$L(g_j p_{v_i}^2) = g_j(v_i) L(p_{v_i}).$$

By assumption, $L(g_j p_{v_i}^2) \geq 0$ and we just showed that $L(p_{v_i}) > 0$. This implies that $g_j(v_i) \geq 0$, as desired, and the proof is complete. □

8.2.3 Moment relaxation for polynomial optimization

Let us return to the polynomial optimization problem (8.1). In Chapter 13, we defined the lower bound $p_{\text{sos}} \leq p_{\text{min}}$, obtained by considering sums of squares decompositions in the quadratic module $M(\mathbf{g}) = \Sigma + g_1\Sigma + \cdots + g_m\Sigma$:

$$p_{\text{sos}} = \sup\{\lambda : p - \lambda \in M(\mathbf{g}) = \Sigma + g_1\Sigma + \cdots g_m\Sigma\}. \tag{8.14}$$

Based on the discussion in the preceding section, we can also define the following lower bound for p_{min}:

$$p_{\text{mom}} = \inf\{\boldsymbol{p}^\mathsf{T} y : y_0 = 1,\ M(y) \succeq 0,\ M(g_j * y) \succeq 0\ (j \in [m])\} \tag{8.15}$$

These two bounds are ‘dual' to each other, since the positivity conditions in (8.15) mean that the corresponding linear functional L is nonnegative on $M(\mathbf{g})$. We have the following inequalities:

Lemma 8.2.5. *We have:* $p_{\text{sos}} \leq p_{\text{mom}} \leq p_{\text{min}}$.

Proof. The inequality $p_{\text{sos}} \leq p_{\text{mom}}$ is ‘weak duality': Let λ be feasible for (8.14) and let y be feasible for (8.15) with associated linear functional L. Then, $p-\lambda \in M(\mathbf{g})$, $L(1) = 1$ and $L \geq 0$ on $M(\mathbf{g})$. Therefore, $L(p-\lambda) = L(p) - \lambda \geq 0$ implies $\boldsymbol{p}^\mathsf{T} y = L(p) \geq \lambda$ and thus $p_{\text{mom}} \geq p_{\text{sos}}$.

The inequality $p_{\text{mom}} \leq p_{\text{min}}$ follows from the fact that, for each $v \in K$, $y = [v]_\infty$ is feasible for (8.15) 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 8.2.4 that $p_{\text{mom}} = p_{\text{min}}$ if the program (8.15) 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 (8.1) obtained by adding degree constraints to the programs (8.14) and (8.15), and we will use the results of Theorems 8.1.11 and 8.2.4 for giving a procedure to find global optimizers of problem (8.1).

8.3 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 8.2.4 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 [5]).

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 $\mathbb{R}[x]/I$).

8.4 Exercises

8.1 Recall the definitions (8.5) and (8.6) 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.

8.2 Show Lemma 8.1.3.

8.3 (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.

(c) Show that any finite set $V \subseteq \mathbb{C}^n$ is a complex variety.

8.4 The goal is to show Theorem 8.1.13 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 (8.10) in the base $\mathcal{B}$ is the real symmetric matrix $H = (H_{ij})_{i,=1}^N$ with entries $H_{ij} = \text{Tr}(M_{b_i b_j})$. 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 $\text{rank}(H) = |V_{\mathbb{C}}(I)|$.

(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 $P, Q \succeq 0$, $\text{rank}(P) = |V_{\mathbb{R}}(I)| + |T|$ and $\text{rank}(Q) = |T|$.

(d) Show that $H = A - B$ for some matrices A, B such that $A, B \succeq 0$, $AB = BA = 0$, $\text{rank}(A) = \sigma_+(H)$ and $\text{rank}(B) = \sigma_-(H)$.

(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. In *Emerging Applications of Algebraic Geometry*, Vol. 149 of IMA Volumes in Mathematics and its Applications, M. Putinar and S. Sullivant (eds.), Springer, pages 157-270, 2009. Available at `http://homepages.cwi.nl/~monique/files/moment-ima-update-new.pdf`

CHAPTER 9

POLYNOMIAL OPTIMIZATION AND REAL ROOTS

We return to the polynomial optimization problem:

$$p_{\text{min}} = \inf_{x \in K} p(x), \tag{9.1}$$

where K is defined by polynomial inequalities:

$$K = \{x \in \mathbb{R}^n : g_1(x) \geq 0, \cdots, g_m(x) \geq 0\} \tag{9.2}$$

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:

$$p_{\text{sos}} = \sup \left\{ \lambda : p - \lambda \in M(\mathbf{g}) = \sum_{j=0}^{m} g_j \Sigma \right\},$$

$$p_{\text{mom}} = \inf\{L(p) : L \text{ linear function on } \mathbb{R}[x],\ L(1) = 1,\ L \geq 0 \text{ on } M(\mathbf{g})\},$$

which satisfy the inequalities:

$$p_{\text{sos}} \leq p_{\text{mom}} \leq p_{\text{min}}.$$

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 (9.1) 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*:

$$M(\mathbf{g})_{2t} = \left\{ \sum_{j=0}^{m} g_j s_j : s_j \in \Sigma,\ \deg(s_j g_j) \leq 2t\ \ (j = 0, 1, \cdots, m) \right\},$$

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:

$$p_{\text{sos},t} = \sup\{\lambda : p - \lambda \in M(\mathbf{g})_{2t}\}, \tag{9.3}$$

$$p_{\text{mom},t} = \inf\{L(p) : L \text{ linear function on } \mathbb{R}[x]_{2t},\ L(1) = 1,\ L \geq 0 \text{ on } M(\mathbf{g})_{2t}\}. \tag{9.4}$$

Lemma 9.0.1. *For any integer t, $p_{\text{sos},t} \leq p_{\text{mom},t} \leq p_{\text{min}}$.*

Proof. Let L be feasible for (9.4) and let λ be feasible for (9.3). Then, we have: $0 \leq L(p - \lambda) = L(p) - \lambda$. This implies that $p_{\text{sos},t} \leq p_{\text{mom},t}$.

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 (9.4) with objective value $L(p) = p(v)$. This implies: $p_{\text{mom},t} \leq p(v)$. As this holds for all $v \in K$, we deduce that $p_{\text{mom},t} \leq p_{\text{min}}$. □

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 (9.4) 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 (9.1).
4. Application to computing **real roots** of polynomial equations.

9.1 Duality

We now indicate how to reformulate the programs (9.3) and (9.4) 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:

$$L \geq 0 \text{ on } \Sigma \cap \mathbb{R}[x]_{2t} \Longleftrightarrow M_t(y) \succeq 0.$$

Analogously,

$$L \geq 0 \text{ on } \{sg : s \in \Sigma,\ \deg(sg) \leq 2t\} \Longleftrightarrow M_{t-d_g}(g * y) \succeq 0,$$

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 $|\alpha| \leq 2(t-d_g)$ as then $|\alpha + \gamma| \leq 2(t - d_g) + \deg(g) \leq 2t$). Therefore, the program (9.4) can be equivalently reformulated as:

$$p_{\mathrm{mom},t} = \inf_{y \in \mathbb{N}^n_{2t}} \{\boldsymbol{p}^\mathsf{T} y : y_0 = 1,\ M_t(y) \succeq 0,\ M_{t-d_{g_j}}(g_j * y) \succeq 0 \ (j = 1, \cdots, m)\}. \tag{9.5}$$

We now explicit the fact that the dual semidefinite program of (9.5) coincides with (9.3); we do this only in the unconstrained case: $K = \mathbb{R}^n$ (i.e., with no constraints $g_j \geq 0$) 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

$$M_t(y) = \sum_{\gamma \in \mathbb{N}^n_{2t}} y_\gamma A_{t,\gamma} \text{ and } \sum_{\gamma \in \mathbb{N}^n_{2t}} x^\gamma A_{t,\gamma} = [x]_t [x]_t^\mathsf{T} \tag{9.6}$$

after setting $[x]_t = (x^\alpha)_{\alpha \in \mathbb{N}^n_t}$.

Lemma 9.1.1. *The programs:*

$$\sup\{\lambda : p - \lambda \in \Sigma \cap \mathbb{R}[x]_{2t}\}, \tag{9.7}$$

and

$$\inf_{y \in \mathbb{R}^{\mathbb{N}^n_{2t}}} \{\boldsymbol{p}^\mathsf{T} y : y_0 = 1,\ M_t(y) \succeq 0\} \tag{9.8}$$

are dual semidefinite programs.

Proof. Using (9.6), we can express (9.8) as the following semidefinite program (in standard dual form):

$$p_0 + \inf \left\{ \sum_{\gamma \in \mathbb{N}^n_{2t} \setminus \{0\}} p_\gamma y_\gamma : A_{t,0} + \sum_{\gamma \in \mathbb{N}^n_{2t} \setminus \{0\}} y_\gamma A_{t,\gamma} \succeq 0 \right\}. \tag{9.9}$$

Next we express (9.7) 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 (9.6)). Therefore, (9.7) is equivalent to

$$p_0 + \sup\left\{ -\langle A_{t,0}, Q \rangle : \langle A_{t,\gamma}, Q \rangle = p_\gamma \ (\gamma \in \mathbb{N}^n_{2t} \setminus \{0\}),\ Q \succeq 0 \right\}. \tag{9.10}$$

It is now clear that the programs (9.9) and (9.10) are dual semidefinite programs. □

9.2 Convergence

Theorem 9.2.1. *Assume that $M(\mathbf{g})$ is Archimedean (i.e., there exists a polynomial $f \in M(\mathbf{g})$ for which the set $\{x \in \mathbb{R}^n : f(x) \geq 0\}$ 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 $p_{\min} - \epsilon \leq p_{\text{sos},t}$. Therefore, $\lim_{t\to\infty} p_{\text{sos},t} = p_{\min}$. Since, by Lemma 9.0.1, $p_{\text{sos},t} \leq p_{\text{mom},t} \leq p_{\min}$ for all t, we deduce: $\lim_{t\to\infty} p_{\text{mom},t} = p_{\min}$. □

9.3 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 (9.12) 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

$$\boldsymbol{f} \in \ker M_s(y) \Longleftrightarrow L(fg) = 0 \ \forall g \in \mathbb{R}[x]_s; \tag{9.11}$$

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 9.3.1. *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:*

$$x \in \ker A \Longleftrightarrow x \in \ker B^{\mathsf{T}} \Longleftrightarrow (x^{\mathsf{T}}, 0)^{\mathsf{T}} \in \ker X.$$

As an application we obtain the following result showing that the kernel of a truncated moment matrix behaves like a 'truncated ideal'.

Lemma 9.3.2. *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)* rank$M_s(y)$ = rank$M_{s-1}(y)$*, or (ii)* $M_s(y) \succeq 0$*. Given polynomials* $f, g \in \mathbb{R}[x]$*, the following holds:*

$$f \in \ker M_s(y),\ \deg(fg) \leq s-1 \Longrightarrow fg \in \ker M_s(y).$$

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 9.3.1). 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 $\deg(x_i u) \leq s$ (recall (9.11)). □

Theorem 9.3.3. *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*

$$\text{rank } M_s(y) = \text{rank } M_{s-1}(y). \tag{9.12}$$

Then, one can extend y *to a sequence* $\tilde{y} \in \mathbb{R}^{\mathbb{N}^n}$ *satisfying:*

$$\text{rank } M(\tilde{y}) = \text{rank } M_s(y). \tag{9.13}$$

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 (9.13). It is based on Lemma 9.3.2; 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 rank$M(\tilde{y})$ = 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 rank$M(\tilde{y})$ = rank$M_s(y)$, we have the inclusion: $\ker M_s(y) \subseteq \ker M(\tilde{y})$ (recall Lemma 9.3.1). 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 $f = p \in (\ker M_s(y))$. □

9.4 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

$$d_K = \max\{d_{g_1}, \cdots, d_{g_m}\}, \quad \text{where } d_f = \lceil \deg(f)/2 \rceil \text{ for } f \in \mathbb{R}[x]. \qquad (9.14)$$

Theorem 9.4.1. *Let L be an optimal solution to the program (9.4) and let $y = (L(x^\alpha)) \in \mathbb{R}^{\mathbb{N}^n_{2t}}$ be the corresponding sequence. Asssume that y satisfies the rank condition:*

$$\text{rank } M_s(y) = \text{rank } M_{s-d_K}(y) \qquad (9.15)$$

for some integer s satisfying $\max\{d_p, d_K\} \leq s \leq t$. Then the following properties hold:

(i) *The relaxation (9.4) 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 (9.4) 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 (9.15), we can apply Theorem 9.3.3: There exists a sequence $\tilde{y} \in \mathbb{R}^{\mathbb{N}^n}$ extending the subsequence $(y_\alpha)_{|\alpha| \leq 2s}$ of y and satisfying rank $M(\tilde{y}) =$ rank $M_s(y) =: r$. Thus, $\tilde{y}_\alpha = y_\alpha$ if $|\alpha| \leq 2s$, 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 9.3.3, 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:

$$(y_\alpha)_{\alpha \in \mathbb{N}^n_{2s}} = \sum_{i=1}^r \lambda_i [v_i]_{2s} \text{ where } \lambda_i > 0 \text{ and } \sum_{i=1}^r \lambda_i = 1. \qquad (9.16)$$

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 9.3.3).

First we claim that we can choose interpolation polynomials p_{v_i} at the points of $V_{\mathbb{C}}(I)$ with $\deg(p_{v_i}) \leq s - d_K$. 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 $L \geq 0$ on $(g_j\Sigma) \cap \mathbb{R}[x]_{2t}$ for all $j \in [m]$. As $\deg(p_{v_i}) \leq s - d_K$, we have: $\deg(g_j p_{v_i}^2) \leq \deg(g_j) + 2(s - d_K) \leq 2s$, and thus we can compute $L(g_j p_{v_i}^2)$ using (9.16) and obtain that $L(g_j p_{v_i}^2) = g_j(v_i)\lambda_i \geq 0$. This gives $g_j(v_i) \geq 0$ for all j and thus $v_i \in K$.

As $\deg(p) \leq 2s$, we can also evaluate $L(p)$ using (9.16): we obtain that $L(p) = \sum_{i=1}^r \lambda_i p(v_i) \geq p_{\text{min}}$, since $p(v_i) \geq p_{\text{min}}$ as $v_i \in K$. This gives the inequality: $p_{\text{mom},t} \geq p_{\text{min}}$. The reverse inequality holds always (Lemma 9.0.1). Thus (i) holds: $p_{\text{mom},t} = p_{\text{min}}$. In turn, this implies that $p(v_i) = p_{\text{min}}$ 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 (9.4). In other words, y lies in the relative interior of the face of the feasible region of (9.4) 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 (9.4). 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 9.4.1, 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 $\deg(x_k x^{\alpha_j}) \leq s$, 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 (9.4) satisfying the rank condition

(9.15), then we can (numerically) compute all the global optimizers of problem (9.1).

9.5 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 9.5.2 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

$$K = \{x \in \mathbb{R}^n : h_j(x) \geq 0, -h_j(x) \geq 0 \; \forall j \in [m]\}.$$

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 (9.4). Now, any feasible solution L is an optimal solution of (9.4). Hence, by Theorem 9.4.1, if the rank condition (9.15) 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 (9.15) will be satisfied.

Theorem 9.5.1. *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:*

$$L(1) = 1, \; L \geq 0 \; \text{ on } \Sigma_{2t}, \; L(uh_j) = 0 \; \forall j \in [m] \; \forall u \in \mathbb{R}[x] \text{ with } \deg(uh_j) \leq 2t. \tag{9.17}$$

Then, there exist integers t_0 and s such that $d_K \leq s \leq t_0$ and the following rank condition holds:

$$\text{rank} M_s(y) = \text{rank} M_{s-d_K}(y) \; \forall y \in \mathcal{F}_t \; \forall t \geq t_0. \tag{9.18}$$

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 (9.18). 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 $t \geq \max_j \deg(h_j)$. Then,

$$h_j \in \ker M_t(y) \; \forall j \in [m] \tag{9.19}$$

(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:

$$\forall f \in \sqrt[\mathbb{R}]{I} \;\; f = \sum_{l=1}^{L} u_l f_l \;\; \text{for some } u_l \in \mathbb{R}[x] \;\; \text{with } \deg(u_l f_l) \leq \deg(f). \tag{9.20}$$

(That such a nice set of generators exists follows from the theory of Gröbner bases.) Next we claim:

$$\exists t_1 \in \mathbb{N} \;\; f_1, \cdots, f_L \in \ker M_t(y) \;\; \text{for any } t \geq t_1. \tag{9.21}$$

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 9.3.2). 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:

$$x^\alpha = p^{(\alpha)} + q^{(\alpha)}, \quad \text{with } q^{(\alpha)} = \sum_{l=1}^{L} u_l^{(\alpha)} f_l, \tag{9.22}$$

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 $\max\{|\alpha|, d_0\} \leq t_2$. 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 (9.20) of the generators f_l of $\sqrt[\mathbb{R}]{I}$.

We can now conclude the proof: We show that, if $t \geq t_0 := t_2 + 1$, then the rank condition (9.18) holds with $s = t_2$. For this pick a monomial x^α of degree at most t_2, so that (9.22) holds. As $\deg(u_l^{(\alpha)} f_l) \leq t_2 \leq t - 1$ and $f_l \in \ker M_t(y)$ (by (9.20)), we obtain that $u_l^{(\alpha)} f_l \in \ker M_t(y)$ (use Lemma 9.3.2). Therefore, the polynomial $x^\alpha - p^{(\alpha)}$ belongs to the kernel of $M_t(y)$. As the degree of $p^{(\alpha)}$ is at most $d_0 \leq t_2 - d_K$, we can conclude that rank$M_{t_2 - d_K}(y) =$ rank$M_{t_2}(y)$.

Finally, the equality $\sqrt[\mathbb{R}]{I} = (\ker M_{t_2}(y)))$ follows from Theorem 9.4.1 (iii). □

Example 9.5.2. *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 (9.17) for $t = 1$. Then, we have that $L(x_1^2), L(x_2^2) \geq 0$ and $L(x_1^2 + x_2^2) = 0$. This implies: $L(x_1^2) = L(x_2^2) = 0$ and thus $L(x_1) = L(x_2) = L(x_1x_2) = 0$. 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, rank$M_1(y) =$ 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?

9.6 Notes and further reading

The flat extension theorem (Theorem 9.3.3) 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].

Other implementations of the sums of squares vs. moment approach for polynomial optimization include
- YALMIP:
`http://users.isy.liu.se/johanl/yalmip/pmwiki.php?n=Main.HomePage,`

- SOSTOOLS: `http://www.cds.caltech.edu/sostools/`,

- SparsePOP, for polynomial optimization problems with sparsity pattern: `http://www.is.titech.ac.jp/~kojima/SparsePOP/`.

9.7 Exercises

9.1. Given an integer $s \geq 1$, 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 9.3.3) in the univariate case $n = 1$.

9.2 Consider the problem of computing $p_{\text{min}} = \inf_{x \in K} p(x)$, where $p = x_1 x_2$ and

$$K = \{x \in \mathbb{R}^2 : -x_2^2 \geq 0,\ 1 + x_1 \geq 0,\ 1 - x_1 \geq 0\}.$$

(a) Show that, at order $t = 1$, $p_{\text{mom},1} = p_{\text{min}} = 0$ 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. In *Emerging Applications of Algebraic Geometry*, Vol. 149 of IMA Volumes in Mathematics and its Applications, M. Putinar and S. Sullivant (eds.), Springer, pages 157-270, 2009. Available at `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 2 in *Handbook on Semidefinite, Cone and Polynomial Optimization*, M. Anjos and J.B. Lasserre (eds.), International Series in Operations Research & Management Science, Springer, 2012, Volume 166, Part 1, 25-60. Available at `http://homepages.cwi.nl/~monique/files/Handbook-SDP.pdf`