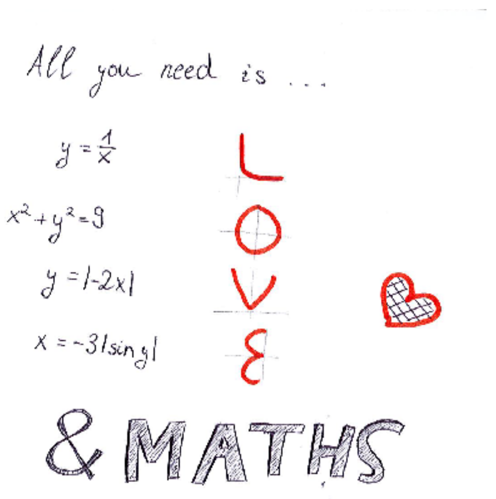
Polynomial Optimization

G. Averkov
(with edits by Matthias Schymura)

September 16, 2022

OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

DFG-Graduiertenkolleg
MATHEMATISCHE
KOMPLEXITÄTSREDUKTION



Contents

Abberviations

LMI	linear matrix inequality
LP	linear program or linear programming
NLP	non-linear program or non-linear programming
POP	polynomial optimization program or polynomial optimization
PSD	(symmetric) positive semidefinite
SDP	(linear) semidefinite optimization program or semidefinite programming
SOS	sum of squares

About these notes

In July 2017, Prof. Sebastian Sager and I held a course aiming at presenting some of the trends in modern optimization. This course is one of the teaching activities of the Research Training Network Mathematical Aspects of Complexity Reduction, supported by the German Research Foundation. Sebastian presented various aspects of linear, non-linear, mixed-integer and global optimization and also addressed optimization in infinite-dimensional spaces (in systems and control theory).

In my part of the course, I explained the reduction of polynomial optimization problems to semidefinite optimization via the sum-of-squares approach, and explained how the interior point methods can be used to solve semidefinite problems based on the duality of semidefinite programming. This manuscript presents the notes, on which my part of the course was based.

Currently, there are several regular sources, from which you can learn about the topic: a long expository article of Monique Laurent [Lau09], the book of Jean B. Lasserre [Las15], which is the standard reference book in polynomial optimization and the book of Murray Marshall [Mar08a] that contains the necessary theoretical background from real algebra and ends with a short discussion of polynomial optimization. There are basic dual approaches to reduce polynomial problems to semidefinite programming: one can use sum-of-squares relaxations and truncated-moment relaxations. In my notes, I discuss truncated-moment relaxations very briefly and towards the end of the notes. The line of thought is as follows:

1. To obtain lower bounds we need algebraic certificates for positivity; Sums of squares of polynomials turn out to be a very nice tractable certificate that can be expressed in terms of semidefinite optimization. Unfortunately, one cannot use sums of squares of polynomials to always certify positivity.

2. One can always use sums of squares of rational functions, which shows that positivity is intrinsically related to sums of squares.

3. On compact semialgebraic sets, positivity *can* be expressed through sums of squares of polynomials.

4. For solving semidefinite problems, one needs to understand duality of semidefinite programming, which is a special case of conic duality.

5. Semidefinite programs can be solved using interior point methods.

Each of the above points corresponds to a chapter of these notes. Discussing point 3 I rely on my short exposition article [Ave13] on elementary approaches to denominator-free positivstellensätze, which collects various pieces of information that were spread all over the literature.

I made no attempt to give an extensive exposition of the background literature (though I plan to extend the literature list in the future).

A second iteration of the compact course based on these notes was held by Maximilian Merkert in 2020, whom I would like to thank for implementing some corrections compared to the initial version.

If you happen to spot any issues (inconsistencies etc.), please let me know.

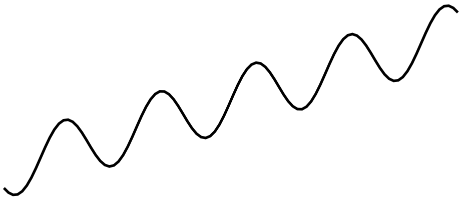
0 Introduction

0.1 Global non-linear optimization

- Convex and linear optimization problems are computationally tractable. This is confirmed in theory and practice! One very special thing about convex optimization is that there is no difference between local and global optimality. So, for checking global optimality, local information is enough.

- In applications, one frequently needs to solve non-linear problems. When one says *non-linear* one usually means *non-convex problems*, because in optimization the crucial difference is not between linear and non-linear but between convex and non-convex.

- Originally, one is usually interested in *global non-linear optimization.* Of course, when the underlying global optimization problem is hard, one can try to find a locally optimal solution rather than solving the problem globally. However, if there are lots of local solutions of all possible qualities, why should a 'random' local solution be good? See figure: if you pick any locally optimal solution, you will get as good as any value of the objective function. So, why is a locally optimal solution any better than just any solution?



- So, if we are really interested in solving an original problem, we usually have to deal with global non-linear programming (NLP).

- *General* global NLP is extremely hard. It's a *huge* class of problems. Depending on how you define it, not algorithmically solvable. Consider the feasibility problem

$$\begin{aligned} \sin(\pi x_1) = \ldots = \sin(\pi x_n) &= 0 \quad \text{(Integrality constraints } x_1, \ldots, x_n \in \mathbb{Z}) \\ p(x_1, \ldots, x_n) &= 0 \quad \text{(Polynomial equality)} \end{aligned}$$

  It is not algorithmically solvable (see Matyasievich's answer to Hilbert's 10th problem). If you do not like the feasibility formulation you can reword the problem equivalently as an optimization problem. Just ask whether the optimal value of the minimization problem

$$\inf \left\{ p(x_1, \ldots, x_n)^2 + \sum_{i=1}^{n} \sin(\pi x_i)^2 \; : \; x = (x_1, \ldots, x_n) \in \mathbb{R}^n \right\}$$

  is 0 and if this value is attained for some $x \in \mathbb{R}^n$.

- Take a look at the following chain of complexity classes

$$\mathrm{P} \subseteq \mathrm{NP} \subseteq \mathrm{PSPACE} \subseteq \mathrm{EXP} \subseteq \mathrm{DECIDABLE}$$

It is conjectured that all inclusions are strict and the inclusion P $\subseteq$ EXP is known to be strict. As we have seen, global NLP is not decidable and so it is easy to imagine that subclasses of global NLP are spread all over this chain. So, they cover lots of extremely difficult problems, including the NP-complete problems, which researchers have already been working on for decades.

Complexity theory tells us that there is no way of finding a *universal* efficient approach to very hard problems. Still, we can and will develop some quite general theory which can help us to do appropriate algorithmic choices for our concrete problems we want to solve.

- Currently, the global NLP community seems to be split into two sub-communities:
  - One can use
    - branching,
    - introducing intermediate variables to keep track of the intermediate results in the expression digraphs of the underlying functions, or
    - basic outer convexification (McCormick, α-relaxations et al.)

    Such techniques are available in solvers like Baron.
  - Alternatively, one can use sums-of-squares which has also been implemented.

0.2 Polynomial optimization

I what follows, I am going to use POP to abbreviate either *polynomial optimization* or a *polynomial optimization problem*.

POP is the class of problems of dealing with optimization of

- a polynomial objective function
- in the presence of finitely many real-valued decision variables and
- under finitely many polynomial equality/inequality constraints

Direct usage of equality constraints can be avoided, as we can replace $p(x) = 0$ by $p(x) \geq 0$ and $-p(x) \geq 0$. When we talk about inequality constraints, we mean the non-strict inequalities by default, but it can easily be seen that strict inequalities can be easily modelled via non-strict ones in lifting in the context of POP. Indeed $p(x) > 0$ can be expressed as $p(x)y = 1$ and $y \geq 0$ using an additional variable y.

We'll see how POPs can be converted to so-called semidefinite problems (SDPs). Some more specific situations can also be converted to linear problems (LPs). For LPs and SDPs various efficient solution methods can be employed.

General POP involves a lot of kinds of problems as special cases, including the following NP-hard) problems:

- Binary integer linear programming (so there is a connection to *combinatorial optimization*)
- Inequality constrained quadratic programming (and so there is a connection to classical numerical *nonlinear optimization* dealing with iterative methods based on first and second derivatives and Taylor expansions).

Since on a compact set, every continuous function can be approximated by a polynomial (Stone-Weierstrass theorem) with any given accuracy, in some sense, POP is dense in NLP. Apart from that, polynomials provide a very natural ‘modeling language’, because arithmetics over reals (multiplication and addition) is one of the basic things a computer can do.

There exist results showing that the decision version of polynomial optimization is in EXP (see [BGHED14]), so polynomial optimization (in its decision form) is somewhere between NP-hard and EXP (but I do not know its precise complexity).

0.3 Prerequisites

The more you know about the following topics, the easier it would be for you to follow the discussion.

Linear algebra. Vector spaces, Euclidean spaces, linear maps.

Convexity. Convex sets and cones, polyhedra, separations theorems, faces and extreme points. In the standard mathematics curriculum, convexity is usually integrated into introductory optimization courses.

Linear optimization. Duality, understanding the geometry of the simplex method.

Analysis. Basic knowledge.

Algebra. A bit of experience with groups, rings (and ideals), and fields.

0.4 Notation

Set relations:		
	$\subseteq$	regular inclusion
	$\subsetneq$	proper inclusion
Sets of numbers:		
	$\mathbb{N}$	positive integers (called natural numbers in these notes)
	$\mathbb{Z}_+$	non-negative integers
	$\mathbb{Q}$	rational numbers
	$\mathbb{R}$	real numbers
	$\mathbb{C}$	complex numbers
	$\mathbb{R}_+$	non-negative reals
	$\mathbb{R}_{\geq 0}$	non-negative reals (another notation)
	$\mathbb{R}_{>0}$	positive reals
	$[m]$	natural numbers not greater than $m \in \mathbb{Z}_+$
Operations for sets:		
	int	interior
	conv	convex hull
	cone	*convex* conic hull
	lin	linear hull
	aff	affine hull

We use 0 to denote the zero element; it can be zero value, zero vector or a zero matrix, depending on the context.

Throughout, $n \in \mathbb{N}$ is the dimension of the ambient space, which is usually $\mathbb{R}^n$. The elements of $\mathbb{R}^n$ are viewed as columns, but we tend to omit transposition to simplify the expressions. For example, we write $(1,2,3) \in \mathbb{R}^3$ rather than $(1,2,3)^\top \in \mathbb{R}^3$. If I is a set, then $\mathbb{R}^I$ is the set of all functions from I to $\mathbb{R}$ and it can also be viewed (and written) as a vector indexed by elements of I. So, we can write for example $x = (x_i)_{i \in I} \in \mathbb{R}^I$. With this point of view, $\mathbb{R}^n$ is nothing but a special case of $\mathbb{R}^I$ with $I = [n]$. Please note that we do not insist x_i to be the default notation for the i-component, because in cases we do not work with components of vectors, we prefer to use lower indexing for vectors.

For $x = (x_i)_{i \in [n]}$ and $y = (y_i)_{i \in [n]}$, we introduce the standard scalar product $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$ of x and y and the respective Euclidean norm $\|x\| := \sqrt{\langle x, x \rangle}$ of x. We also use the notation $\langle \cdot\,, \cdot \rangle$ and $\| \cdot \|$ to denote the scalar product and the respective norm of other Euclidean spaces appearing in our discussion. The scalar product gives rise to the orthogonality relation and we use the notation $X^\perp$ to denote the orthogonal complement of X (which is the set of all vectors orthogonal to every vector from X).

By I_n we denote the identity matrix of size n, and if the size is clear from the context we omit the subscript n. If A is a matrix, then $A^\top$ denotes the transpose of A. When we say that a matrix A is PSD we always mean symmetric PSD.

Asymptotic notation for $f, g : X \to \mathbb{R}$, for $x \to x^*$:

$$\begin{array}{ll}
f = O(g) & \limsup_{x \to x^*} \frac{|f(x)|}{|g(x)|} < +\infty \\
f = \Omega(g) & \liminf_{x \to x^*} \frac{|f(x)|}{|g(x)|} < +\infty \\
f = \Theta(g) & f = O(g) \text{ and } f = \Omega(g) \\
f = o(g) & \lim_{x \to x^*} \frac{|f(x)|}{|g(x)|} = 0.
\end{array}$$

By $\mathbb{R}[X_1, \ldots, X_n]$ we denote the set of polynomials in variables $X_1, \ldots, X_n$ with coefficients in $\mathbb{R}$. Instead of $\mathbb{R}$ one can use other sets, so one can consider other sets of polynomials, but we'll mostly work with $\mathbb{R}[X_1, \ldots, X_n]$. A warning for those, who don't have much experience with algebra: polynomials and functions are not quite the same thing. Rather, polynomials are formal expressions, and their variables $X_1, \ldots, X_n$ are just formal symbols (also called *indeterminates*). So, a symbolic variable X in the ring $\mathbb{R}[X]$ of univariate polynomials with real coefficients is not equal to any element of $\mathbb{R}$, because X is just another element of $\mathbb{R}[X]$ (as any element of $\mathbb{R}$). To illustrate this, consider the following code in sage which introduces a symbol variable and compares it to the number 3:

```
x = var(’x’)
if x!=3:
   print "Not equal"
else:
   print "Equal"
```

You can try this code at `http://sagecell.sagemath.org/`. Another way to explain this is by means of coefficients. The element 3 of $\mathbb{R}[X]$ is a degree one polynomial, that is $3 = 3 + 0X + 0X^2 + \ldots$, and its coefficients are $3, 0, 0, \ldots$. On the other hand X is another polynomial $X = 0 + 1X + 0X^2 + \cdots$ and its coefficients are $0, 1, 0, 0, \ldots$. Comparison of polynomials is defined through comparision of coefficients and so 3 and X are not equal, because their respective coefficients are not all equal.

$\mathbb{R}(X_1, \ldots, X_n)$ denotes the set of rational functions in $X_1, \ldots, X_n$ with coefficients in $\mathbb{R}$. Even though rational functions are called functions, they are not functions but formal quotients f/g with $f, g \in \mathbb{R}[X_1, \ldots, X_n]$ and $g \neq 0$. By definition, two such functions f_1/g_1 and f_2/g_2 are equal if the polynomial equality $f_1 g_2 = f_2 g_1$ holds.

For various kinds of structures X and Y, we use notations like $X + Y = \{x + y : x \in X, y \in Y\}$, $X - Y = \{x - y : x \in X, y \in Y\}$, $aX = \{ax : x \in X\}$ etc. This allows us to express things in a concise form. For example, if R is a ring, then an ideal I of R can be defined as an additive subgroup of R satisfying $RI \subseteq I$, where we use the notation $XY = \{xy : x \in X, y \in Y\}$.

0.5 Numbering

Theorems, Lemmas, Propositions, Remarks, Exercises and Examples are all numbered based on a common counter. In my opinion, this simplifies searching.

0.6 Software

The following software will or may be useful:

Name	**Description**	**Used**
MATLAB	proprietary numerical computing environment	yes
GNU Octave	free alternative for MATLAB	no
AMPL	modelling language for optimization problems	yes
SeDuMi	MATLAB package for solving semidefinite problems	yes
SOSTOOLS	sums-of-squares based optimization toolbox for MATLAB	yes
GloptyPoly	sums-of-squares and moment-relaxation based optimization toolbox for MATLAB	no
BARON software	global solver for non-convex problems	yes

One could have listed more software.

1 Non-negativity and sums of squares

In this chapter, $n \in \mathbb{N}$ and

$$X = (X_1, \dots, X_n)$$

is a tuple of n indeterminates. In particular, if $n = 1$, then X is an indeterminate. The aim is to present the relationship between real algebra, polynomial optimization and semidefinite optimization.

1.1 A Topic of real algebra

In contrast to classical algebra over complex numbers (or more generally, over algebraically closed fields), in real algebra one tries to understand positivity and non-negativity of polynomials and other algebraic objects (one just cannot define the non-negativity concept over complex numbers). So, we have to work with reals (or ordered fields, more generally).

Let $f \in \mathbb{R}[X] = \mathbb{R}[X_1, \dots, X_n]$. For $K \subseteq \mathbb{R}^n$, we write

- $f \geq 0$ on K if $f(x) \geq 0$ for every $x \in K$ and
- $f > 0$ on K if $f(x) > 0$ for every $x \in K$.

We'll frequently write polynomials $f \in \mathbb{R}[X]$ using coefficients $c_\alpha \in \mathbb{R}$ where $\alpha \in \mathbb{Z}_+^n$. For monomials we use the notation

$$X^\alpha := X_1^{\alpha_1} \cdots X_n^{\alpha_n} \qquad (\alpha \in \mathbb{Z}_+^n),$$

so that we can write f in the form

$$f = \sum_{\alpha \in \mathbb{Z}_+^n} c_\alpha X^\alpha$$

where all but finitely many c_α's are zero. We call $\alpha \in \mathbb{Z}_+^n$ an *exponent vector* or a *multi-index* and we use the notation

$$|\alpha| := \alpha_1 + \cdots + \alpha_n$$

to denote the degree of the monomial X^α. For $n \in \mathbb{N}$ and $d \in \mathbb{Z}_+$, we introduce the notation

$$E_d^n := \left\{ \alpha \in \mathbb{Z}_+^n \, : \, |\alpha| \leq d \right\}$$

for the set of multi-indices of degree at most d. Thus, if $\deg(f) \leq d$, we can write f as $f = \sum_{\alpha \in E_d^n} c_\alpha X^\alpha$.

Exercise 1.1. *Determine the cardinality* $|E_d^n|$ *of* E_d^n*.*

Solution. We are looking for the number of ordered non-negative integer solutions of the inequality $\alpha_1 + \cdots + \alpha_n \leq d$. Adding $\alpha_{n+1} \in \mathbb{Z}_+$ we switch to looking for the number of non-negative integer solutions of the equality $\alpha_1 + \cdots + \alpha_{n+1} = d$ in unknowns $\alpha_1, \dots, \alpha_{n+1}$.

Substituting, $\alpha_i + 1 =: \beta_i$, we switch to looking for the number of positive integer solutions of the equation $\beta_1 + \cdots + \beta_{n+1} = d + n + 1$. Now imagine you've got $d + n + 1$ objects, which are lined up, like in this example with $d + n + 1 = 8$ objects

□ □ □ □ □ □ □ □

Now, there are $d+n$ spaces between the objects and by marking n of the spaces as the separating spaces you can split your objects into $n+1$ non-empty groups of consecutive objects. If say, $n=3$ the following marks

$$\square\,\square\cdot\square\,\square\cdot\square\cdot\square\,\square\,\square$$

split the objects in four groups with, respectively, 2, 2, 1 and 3 objects. So, each such splitting corresponds to a unique choice of $\beta_1,\ldots,\beta_{n+1}$ and, vice versa, each choice of $\beta_1,\ldots,\beta_{n+1}$ gives a unique splitting. In our example, we get $\beta_1=2,\beta_2=2,\beta_3=1,\beta_4=3$. The number of splittings is clearly $\binom{n+d}{n}$ since we choose n spaces out of $n+d$. This shows $|E_d^n|=\binom{n+d}{n}$ and since the binomial coefficients are symmetric we also have $|E_d^n|=\binom{n+d}{(n+d)-n}=\binom{n+d}{d}$. □

1.2 Unconstrained polynomial optimization

For the time being, let's restrict our attention to unconstrained polynomial optimization. It is a sufficiently large class, on the one hand, and it is a convenient class to explain the basic ideas of approaching POP via SOS, on the other hand.

We are given a polynomial objective $f\in\mathbb{R}[X]$ and want to solve

$$\inf\{f(x)\,:\,x\in\mathbb{R}^n\}\,. \tag{1.1}$$

Solving the problem means, as usual, to determine the optimal value which can be finite, $-\infty$ or $+\infty$ and, if there exists an $x^*\in\mathbb{R}^n$, at which the optimal value is attained, one such x^* should be computed. Currently, we do not go into the subtleties on what it actually means to compute x^*. So, the statement of the problem is a bit vague, but that's not really disturbing, because we are not going to do computational-complexity studies so far.

Exercise 1.2. *Show that for $n=1$, whenever the infimum in* (1.1) *is finite, it is attained at some $x\in\mathbb{R}^n$ and that for every $n\geq 2$, there exists a polynomial f such that the optimal value of* (1.1) *is finite but not attained at any $x\in\mathbb{R}^n$.*

Solution. As for $n=1$, we see that for $|x|\to\infty$, the dominating term will be the one for the monomial of the highest degree. The degree must be even, say $2d$, for otherwise, the polynomial wouldn't be bounded from below. The coefficient at $2d$ should be positive by the same reasons. So we see that $f(x)>0$ if x is sufficiently large. Thus, (1.1) can be turned into the optimization of a polynomial over a closed segment. Since polynomial functions are continuous, we are done.

As for $n\geq 2$, of course it is sufficient to deal with the case $n=2$. Consider $f(X_1,X_2):=X_2^2+(X_1X_2-1)^2$. The infimum is zero, because f is nonnegative on $\mathbb{R}^2$ and $f(t,1/t)=t^2\to 0$ as $t\to 0$. Let $x_1,x_2\in\mathbb{R}$. If $x_2\neq 0$, then $f(x_1,x_2)\geq x_2^2>0$. If $x_2=0$, then $f(x_1,x_2)\geq 1>0$. □

1.3 Sum-of-squares relaxation from the dual formulation

Approaches to solving (1.1) (and other polynomial problems) can be roughly split into heuristic and non-heuristic ones. For the former, a choice of $x^*\in\mathbb{R}^n$ is suggested, but one does not say whether the chosen x is good (no guarantees). For the latter, one does suggest x^* and gives a lower bound y on the optimal value, so that by comparing $f(x^*)$ and y one can see how good x^* actually is. Heuristic approaches to solving general problems is a very popular topic (machine learning and neural networks etc.) In this course, we are interested in non-heuristic approaches. So, anyway we need an approach to derive lower bounds, and that's what we're actually going to start with. Let's formally dualize our problem (1.1) to

$$\sup \{y \in \mathbb{R} \, : \, f - y \geq 0 \quad \text{on } \mathbb{R}^n\} \, . \tag{1.2}$$

This formulation just tells us that we are interested in lower bounds (that's why 'formally dualize'), because the formulation does not yet tell us how to actually find such bounds. While a non-negativity condition for a polynomial is hard to check, there is a stronger condition that turns out to be easy to test. We call $g \in \mathbb{R}[X]$ sum of squares (SOS), if $g = g_1^2 + \cdots + g_k^2$ for finitely many polynomials $g_1, \ldots, g_k \in \mathbb{R}[X]$. If we succeeded to write g as above, we had an algebraic evidence of g being non-negative. So, by strengthening the non-negativity constraint in (1.2) by an SOS constraint, our supremum becomes smaller (in general) and we arrive at what is called an SOS relaxation of (1.1).

$$\sup \{y \in \mathbb{R} \, : \, f - y \quad \text{SOS}\} \, . \tag{1.3}$$

The term relaxation should be understood in the sense that by solving (1.3) we obtain a *lower* bound on our original problem (1.1).

In what follows we've got two things to take care of: we need to figure out how we could solve (1.3), and it would be good to understand how good the lower bounds on (1.1) obtained from (1.3) are.

1.4 SOS-relaxations and semidefinite optimization

For $k \in \mathbb{N}$, let $\mathcal{S}^k$ be the vector space of symmetric matrices $A \in \mathbb{R}^{k \times k}$ of size $k \times k$ and let $\mathcal{S}^k_+$ be the closed convex cone of psd matrices in this space. The constraint '$Z \in \mathcal{S}^k_+$' can be viewed as a generalization of non-negativity to the world of matrices. For $k = 1$, we just have a regular non-negativity constraint for a real variable.

A *(linear) semidefinite problem* (SDP, for short) is a problem with the following properties:

- Linear objective
- Finitely many variables, which can be real variables ranging in $\mathbb{R}$ or matrix variables in spaces $\mathcal{S}^k$ with $k \in \mathbb{N}$
- Finitely many constraints of the form

$$A_0 + x_1 A_1 + \cdots + x_n A_n \in \mathcal{S}^k_+,$$

  with $A_0, \ldots, A_n \in \mathcal{S}^k$. The latter is called a *linear matrix inequality* (LMI) of size k for real variables $x_1, \ldots, x_n \in \mathbb{R}$. The matrices $A_0, \ldots, A_n$ are the coefficients of the LMI. (Note that if $A_0, \ldots, A_n$ are diagonal matrices, we just have a system of k linear inequalities).
- Constraints of the form

$$Z \in \mathcal{S}^k_+$$

  with $k \in \mathbb{N}$. The latter is called a PSD-constraint on a matrix-variable Z.

- Linear equality constraints involving the real variables and/or entries of the matrix variables.

We'll see that (1.3) can be formulated as a semidefinite problem. It is enough to derive the following

Proposition 1.3. *Let $d \in \mathbb{Z}_+$. Let $f = \sum_{\gamma \in E_{2d}^n} c_\gamma X^\gamma \in \mathbb{R}[X]$ be a polynomial of degree at most $2d$. Then the following conditions are equivalent:*

(i) The polynomial f is SOS.

(ii) There exists a symmetric psd matrix $Z := (z_{\alpha,\beta})_{\alpha,\beta \in E_d^n}$ satisfying the linear equations

$$\sum_{\alpha,\beta \in E_d^n \,:\, \alpha+\beta=\gamma} z_{\alpha,\beta} = c_\gamma \qquad \forall\ \gamma \in E_{2d}^n \tag{1.4}$$

Proof. We introduce the vector of all monomials of degree at most d:

$$m(X) := (X^\alpha)_{\alpha \in E_d^n}.$$

If $f = f_1^2 + \cdots + f_r^2$ for some $f_1, \ldots, f_r \in \mathbb{R}[X]$, then all $f_1, \ldots, f_r$ are of degree at most d (this will be justified below, in Proposition 1.7). For each f_j we introduce the vector $u_j \in \mathbb{R}^{E_d^n}$ of its coefficients, so that we can write f_j as

$$f_j = \langle m(X), u_j \rangle = m(X)^\top u_j = u_j^\top m(X).$$

Thus, $f_j^2 = \langle m(X), u_j \rangle^2 = m(X)^\top u_j u_j^\top m(X)$ and $f = f_1^2 + \cdots + f_r^2$ can be rewritten as

$$f = m(X)^\top \underbrace{(u_1 u_1^\top + \cdots + u_r u_r^\top)}_{=:Z} m(X).$$

For each j, the matrix $u_j u_j^\top$ is psd (of rank at most 1). Hence, the sum $Z = (z_{\alpha,\beta})_{\alpha,\beta \in E_d^n}$ is psd, too. We arrive at the representation

$$f(X) = m(X)^\top Z m(X).$$

The latter can be described as a system of linear equations in the coefficients of Z, and this system is written explicitly as (1.4).

Conversely, if (1.4) is fulfilled, then we have $f(X) = m(X)^\top Z m(X)$. Since Z is psd, we can write it as $Z = \sum_{j=1}^r u_j u_j^\top$ for some finitely many vectors $u_1, \ldots, u_r \in \mathbb{R}^{E_d^n}$. This yields $f = f_1^2 + \cdots + f_j^2$ for $f_j = \langle m(X), u_j \rangle$. □

In the last step of the previous proof, we used a fact from linear algebra, which we formulate as an exercise

Exercise 1.4. *Show that if a matrix $A \in \mathcal{S}^k$ is psd, then it can be written as $A = u_1 u_1^\top + \cdots + u_r u_r^\top$ for some finitely many vectors $u_1, \ldots, u_r \in \mathbb{R}^k$. Can the choice of r be bounded in terms of k?*

Solution. We use $r = k$.

Showing existence: This solution is based on the spectral theory of symmetric matrices. Since A is symmetric, there exists an orthonormal basis $v_1, \ldots, v_k$ consisting of eigenvectors of A. Let $\lambda_1, \ldots, \lambda_k$ be the respective eigenvalues. Since A is psd, the eigenvalues are non-negative. Let $u_j = \sqrt{\lambda_j} v_j$. It suffices to check that $Av_j = (u_1 u_1^\top + \cdots + u_k u_k^\top) v_j$ for every $j \in [k]$. The left as well as the right hand side is $\lambda_j v_j$. Clearly, if A is of rank k, we cannot choose a smaller r, since the rank of $u_1 u_1^\top + \cdots + u_r u_r^\top$ is at most r.

Computing the decomposition: There is a Gauss-method-like approach to diagonalizing a given quadratic form with $O(k^3)$ arithmetic operations. This would definitely do.

Every psd matrix A has a Cholesky factorization $A = LL^\top$, where L is a lower triangular matrix. For us, it is not of primary importance that L is lower triangular (any matrix L with $A = LL^\top$ would do). Choosing $u_1, \ldots, u_k$ to be the columns of L we get a desired representation of A. Note that, on the algorithmic side (when one really wants to find a decomposition $A = LL^\top$ efficiently), one usually presents how to efficiently compute a Cholesky decomposition of positive definite matrices. For positive semidefinite matrices, the existence of such a decomposition can be shown as follows.

Every decomposition $A = UU^\top$, where U is arbitrary, gives a decomposition $A = LL^\top$, because $U^\top$ has a QR-factorization.

On the level of software, I've tried out the function **chol** in Matlab but it does not seem to accept matrices of non-full rank (there has been a respective error message). What one can do to compute a desired decomposition within a few lines of code is taking a square root of the matrix using **sqrtm**. Here is the code illustrating the two possibilities:

```
U=rand(3,3);
A=U*U'
R=chol(A)
R'*R
V=sqrtm(A)
V'*V
```

The code generates a random positive semidefinite matrix and uses the two approaches to get the decomposition. So, both $R^\top R$ and $V^\top V$ coincide with A (in the numerical sense). □

With Proposition 1.3 we can easily convert the SOS-relaxation to an SDP.

Corollary 1.5. *Let $f \in \mathbb{R}[X]$ be of degree at most $2d$, where $d \in \mathbb{Z}_+$. Then (1.3) is an SDP of the form*

$$\sup \left\{ y \in \mathbb{R} \,:\, Z \text{ psd},\ m(X)^\top Z m(X) + y = f(X) \right\}. \tag{1.5}$$

where $m(X) := (X^\alpha)_{\alpha \in E_d^n}$ and the decision-variable Z is a $k \times k$ symmetric matrix with $k = |E_d^n|$.

The linear equality system

$$m(X)^\top Z m(X) + y = f(X)$$

in unknowns Z and y can be written explicitly, similarly to the system (1.4), but one can use also the above concise form. Already from this form it is clear that this is a linear system: $f(X)$ is the right hand side of the system and Z and y occur linearly in the left hand side.

The total number of variables of the latter system is $k^2 + 1$, where $k = \binom{n+d}{d}$ and the number of equations is $|E^n_{2d}| = \binom{n+2d}{2d}$.

1.5 Employing Newton polytopes to reduce the size of an SDP

It turns out that, by taking more care to what our choice of f actually is, the size of (1.5) can be reduced. For a polynomial $f = \sum_\alpha c_\alpha X^\alpha \in \mathbb{R}[X]$ the set

$$\mathrm{Newt}(f) := \mathrm{conv}\,\{\alpha \,:\, c_\alpha \neq 0\}$$

is called the *Newton polytope* of f. We want to find out how the Newton polytope of a sum of squares looks like.

Lemma 1.6. *For every $f \in \mathbb{R}[X]$ one has $\mathrm{Newt}(f^2) = 2\,\mathrm{Newt}(f)$.*

Proof. We assume $f \neq 0$. If $f = \sum_{\alpha\in E} c_\alpha X^\alpha$ with $c_\alpha \neq 0$ for all $\alpha \in E$, then $\mathrm{Newt}(f) = \mathrm{conv}(E)$. We've got $f^2 = \sum_{\alpha,\beta\in E} c_\alpha c_\beta X^{\alpha+\beta}$. This shows that $\mathrm{Newt}(f^2) \subseteq \mathrm{conv}(E+E) = \mathrm{conv}(E) + \mathrm{conv}(E) = 2\,\mathrm{Newt}(f)$. To see the converse let α be a vertex of $\mathrm{Newt}(f)$. Then, for all $\beta, \gamma \in E$, whenever one has $\alpha = (\beta+\gamma)/2$, one must have $\beta = \gamma = \alpha$. In other words, one can have $2\alpha = \beta + \gamma$ for $\beta, \gamma \in E$ if and only if $\beta = \gamma = \alpha$. This shows that $X^{2\alpha}$ occurs in f^2 with coefficient $c_\alpha^2 \neq 0$. Thus, we also have the converse inclusion $2\,\mathrm{Newt}(f) \subseteq \mathrm{Newt}(f^2)$. □

The latter can be generalized to

Proposition 1.7. *Let $f_1, \ldots, f_r \in \mathbb{R}[X]$ and let $f = f_1^2 + \cdots + f_r^2$. Then, we have*

$$\mathrm{Newt}(f) = 2\,\mathrm{conv}\left(\bigcup_{j=1}^r \mathrm{Newt}(f_j)\right),$$

and, in particular,

$$\deg(f) = 2\max\{\deg(f_1), \ldots, \deg(f_r)\}.$$

Proof. The argument is inspired by Proposition 1.1 of [CPSV16] (which is a somewhat weaker formulation of this proposition). It seems that the idea of looking at the Newton polytope for non-negativity questions goes back to Reznick [Rez78] (I do not have access to this paper).

Without loss of generality let all f_j be non-zero polynomials. The inclusion $\subseteq$ follows by observing $\mathrm{Newt}(f_1^2 + \cdots + f_r^2) \subseteq \mathrm{conv}\left(\bigcup_{j=1}^r \mathrm{Newt}(f_j^2)\right)$ and using Lemma 1.6.

To see the converse consider an arbitrary vertex α of $P := \mathrm{conv}\left(\bigcup_{j=1}^r \mathrm{Newt}(f_j)\right)$. Then, for every $\beta, \gamma \in P \cap \mathbb{Z}^n$, the condition $\alpha = \frac{1}{2}(\beta+\gamma)$ implies $\beta = \gamma = \alpha$. In other words the condition $2\alpha = \beta + \gamma$ for $\beta, \gamma \in P \cap \mathbb{Z}^n$ implies $\beta = \gamma = \alpha$. Each f_j can be written as $f_j = \sum_{\beta\in P\cap\mathbb{Z}^n} c_{j,\beta} X^\beta$. From the previous condition, we see that

$X^{2\alpha}$ occurs in $f_1^2 + \cdots + f_r^2$ with the coefficient $\sum_{j=1}^r c_{j,\alpha}^2$. Since α is a vertex of one of the polytopes $\mathrm{Newt}(f_j)$, at least one square in the previous sum is non-zero. We thus conclude that $2\alpha \in \mathrm{Newt}(f_1^2 + \cdots + f_r^2)$.

The equality for the degrees is an obvious consequence of the equality for the Newton polytopes (the information about the degree is saved in the Newton polytope). □

For (1.5) to be feasible, we need to have $f - y > 0$ on $\mathbb{R}^n$ for at least one $y \in \mathbb{R}$. With such choice $\mathrm{Newt}(f - y) = \mathrm{conv}(\mathrm{Newt}(f) \cup \{0\})$. So, by Proposition 1.7, for feasibility it is necessary that $\mathrm{conv}(\mathrm{Newt}(f) \cup \{0\})$ is of the form $2P$, where P is a polytope with integer vertices (so called integer polytope). Proposition 1.7 also shows that in that case, we can use $m(X)$ of the form $m(X) = (X^\alpha)_{\alpha \in P \cap \mathbb{Z}^n}$. So, rather than using exponent vectors from the whole set E_n^d, we can restrict ourselves to $P \cap \mathbb{Z}^n$ (a subset of E_n^d), so that the size of the SDP (1.5) gets reduced.

Exercise 1.8. *Consider*

$$f = 2 + X_1^2 + X_1^2 X_2^4 - 4X_1X_2$$

(a) Show that the two-variate polynomial is SOS.

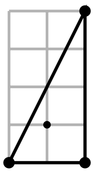
(b) Determine a vector $m(X)$ of monomials and a PSD matrix Z with

$$f = m(X)^\top Z m(X).$$

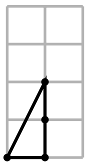
(c) Describe, for your choice of $m(X)$, all PSD matrices Z satisfying

$$f = m(X)^\top Z m(X).$$

Solution. Parts of this exercise can be solved just by guessing an SOS decomposition of f, which is a valid approach to solve exercises and it is quite feasible here because the polynomial f is not too complicated. Anyway, we want to follow up the systematic approach we've just developed and see how the theory works for a concrete example. Look at the exponent vectors and the Newton polytope:



The necessary condition for SOS is fulfilled. It's twice a lattice polytope and the coefficient corresponding to the vertices are positive. Consider the half of the Newton polytope:



So, we know that we can choose $m(X)$ consisting of monomials with the vector exponents $00, 10, 11, 12$. So, we'll need to come up with a matrix Z whose rows and columns are indexed by $00, 10, 11, 12$, and there will be linear equalities for the entries of Z derived from the condition $f = m(X)^\top Z m(X)$. Let's see what results we obtain when we add two vectors from the list $00, 10, 11, 12$:

$$\begin{aligned}
(0,0) &= (0,0) + (0,0)\\
(1,0) &= (0,0) + (1,0)\\
(1,1) &= (0,0) + (1,1)\\
(1,2) &= (0,0) + (1,2)\\
(2,0) &= (1,0) + (1,0)\\
(2,1) &= (1,0) + (1,1)\\
(2,2) &= (1,1) + (1,1) = (1,0) + (1,2)\\
(2,3) &= (1,1) + (1,2)\\
(2,4) &= (1,2) + (1,2)
\end{aligned}$$

(the above equalities can also be nicely illustrated in a picture showing $c = \frac{1}{2}(a+b)$ with a, b being even integer vectors). So, we see that almost all entries of Z are actually determined uniquely by f. The only entries, where we cannot guarantee uniqueness are $z_{11,11}$ and $z_{10,12}$. So, we will denote $z_{10,12}$ by $-t$ and then we see that one needs to have $z_{11,11} = 2t$, because of the relation $z_{11,11} + 2z_{10,12} = 0$.

$$Z := \begin{array}{c} \\ 00 \\ 10 \\ 11 \\ 12 \end{array} \begin{array}{c} \begin{array}{cccc} 00 & 10 & 11 & 12 \end{array} \\ \begin{pmatrix} 2 & 0 & -2 & 0 \\ 0 & 1 & 0 & -t \\ -2 & 0 & 2t & 0 \\ 0 & -t & 0 & 1 \end{pmatrix} \end{array}$$

Our matrix follows a checkerboard template. If we order the vector monomials differently, we'll get a block structure:

$$Z := \begin{array}{c} \\ 00 \\ 11 \\ 10 \\ 12 \end{array} \begin{array}{c} \begin{array}{cccc} 00 & 11 & 10 & 12 \end{array} \\ \begin{pmatrix} 2 & -2 & 0 & 0 \\ -2 & 2t & 0 & 0 \\ 0 & 0 & 1 & -t \\ 0 & 0 & -t & 1 \end{pmatrix} \end{array}$$

For the positive semidefiniteness both blocks should be positive semidefinite. The positive semidefiniteness of 2 × 2 blocks can be expressed through minors easily. And so, we easily convince ourselves that the matrix is PSD only for one choice of t, which is $t = 1$. That is, our Z turns out to be unique

$$Z := \begin{array}{c} \\ 00 \\ 11 \\ 10 \\ 12 \end{array} \begin{array}{c} \begin{array}{cccc} 00 & 11 & 10 & 12 \end{array} \\ \begin{pmatrix} 2 & -2 & 0 & 0 \\ -2 & 2 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix} \end{array}$$

The two blocks are rank one matrices. So, they can be written as a product of a vector times its transposed. Out of this decomposition, we'll come to a respective decomposition of the whole Z.

$$Z := 2 \cdot \begin{array}{c} 00 \\ 11 \\ 10 \\ 12 \end{array} \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix} \begin{array}{c} \begin{array}{cccc} 00 & 11 & 10 & 12 \end{array} \\ \begin{pmatrix} 1 & -1 & 0 & 0 \end{pmatrix} \end{array} + \begin{array}{c} 00 \\ 11 \\ 10 \\ 12 \end{array} \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix} \begin{array}{c} \begin{array}{cccc} 00 & 11 & 10 & 12 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & -1 \end{pmatrix} \end{array}$$

Now, we can write down the SOS decomposition of f

$$f = 2(1 - X_1X_2)^2 + (X_1 - X_1X_2^2)^2.$$

□

1.6 Unconstrained univariate POP gets reduced to SDP

One of the simplest possible cases is the case $n = 1$. By the following exercise, we show that univariate global polynomial optimization can be reduced to semidefinite optimization. Some might think that univariate POP is trivial. This is not quite true, there are a number of publications on the univariate case; see related publications [SM16, MS11, MS09] on computing the real roots of real univariate polynomials.

Exercise 1.9. *Show the following:*

(a) A univariate polynomial $f \in \mathbb{R}[X]$ is non-negative if and only if f is SOS.

(b) If a univariate polynomial $f \in \mathbb{R}[X]$ is non-negative, it is a sum of at most two squares.

Solution. For (a), we need to show that every non-negative polynomial $f \in \mathbb{R}[X]$ is SOS (the converse is clear) and (b) is a refinement of (a). So, we only prove (b). Assume $f \neq 0$. The coefficient at the highest degree monomial of f is strictly positive (because $f(x) \to +\infty$ as $x \to +\infty$). After rescaling, we can assume that the coefficient at the highest order monomial of f is 1. Since f is real, the complex roots of f come in conjugate pairs. Thus, from the complex factorization of f, we can derive the representation

$$f = \prod_{s \in S} (X - a_s)^{k_s} \prod_{j \in J} ((X - b_j)^2 + c_j^2)^{l_j}$$

where S and J are finite index sets, $a_s, b_j \in \mathbb{R}, c_j \in \mathbb{R} \setminus \{0\}$ and $k_s, l_j \in \mathbb{N}$. Clearly, each k_s is even (as otherwise f would change sign at the root a_s). Thus the product over $s \in S$ is a square of a polynomial and thus also a sum of two squares (add the trivial square 0^2). Each term $(X - b_j)^2 + c_j^2$ is a sum of squares of two polynomials $(X - b_j)$ and c_j. Thus, f can be written as $f = (f_1^2 + g_1^2) \cdots (f_t^2 + g_t^2)$ for some finitely many polynomials $f_1, \ldots, f_t, g_1, \ldots, g_t \in \mathbb{R}[X]$. We can now modify the above representation iteratively, decreasing t by one in each iteration. For this, use the following formula which shows that the product of sums of two squares can be converted into a sum of two squares

$$(a^2+b^2)(c^2+d^2)=(ac-bd)^2+(ad+bc)^2,$$

valid for all a,b,c,d. To convince yourself in the validity of the formula you may introduce two complex numbers $z=a+ib$ and $w=c+id$, where $i=\sqrt{-1}$. Then you can easily see that the formulas is just the equality $|z|^2|w|^2=|zw|^2$ involving the absolute values of z,w and zw. □

1.7 The case of degree two gets reduced to linear algebra

Lemma 1.10. *Let $f\in\mathbb{R}[X_1,\dots,X_n]$ be a polynomial of degree at most $2d$. Then*

(a) $g:=X_0^{2d}f(X_1/X_0,\dots,X_n/X_0)\in\mathbb{R}[X_0,\dots,X_n]$ is a homogeneous polynomial of degree at most $2d$ with $g(1,X_1,\dots,X_n)=f(X_1,\dots,X_n)$.

(b) $f\ge 0$ on $\mathbb{R}^n\Leftrightarrow g\ge 0$ on $\mathbb{R}^{n+1}\Leftrightarrow g\ge 0$ on a sphere in $\mathbb{R}^{n+1}$ with the center at the origin.

(c) f is SOS $\Leftrightarrow g$ is SOS of homogeneous polynomials.

Proof. (a) and (b) are quite clear.

(c): if $f=f_1^2+\cdots+f_r^2$, then by Proposition 1.7 the degrees of f_i's are at most d. Then $g=g_1^2+\cdots+g_r^2$, where $g_i=X_0^d f_i(X_1/X_0,\dots,X_n/X_0)$ is a homogeneous polynomial of degree d. Conversely, if $g=g_1^2+\cdots+g_r^2$ for some homogeneous polynomials $g_1,\dots,g_r$, then $f=f_1^2+\cdots+f_r^2$ with $f_i=g_i(1,X_1,\dots,X_n)$. □

Quadratic optimization is a classical topic in numerical linear algebra. It is well-known there that solving a linear system $Ax=b$ for a symmetric positive definite matrix A is equivalent to the minimization of $f(x):=\frac{1}{2}\langle Ax,x\rangle-\langle b,x\rangle$, and that there are methods employing this equivalence (e.g., conjugate gradients).

Exercise 1.11. *If $f\in\mathbb{R}[X]$ is of degree two, then f is non-negative if and only if f is SOS.*

Solution. Using an additional variable X_0, we can turn $f(X_1,\dots,X_n)$ into a homogeneous degree-two polynomial $h(X_0,X_1,\dots,X_n)=X_0^2f(X_1/X_0,\dots,X_n/X_0)$.

If f is non-negative, then its homogenization h, too, is non-negative. We can write h as

$$h=\bar{X}^\top A\bar{X},$$

where $A\in\mathcal{S}^{n+1}$ and $\bar{X}=(X_0,X_1,\dots,X_n)^\top$. Since h is non-negative, A is psd. Writing A as a sum of rank-one psd matrices (see Exercise 1.4), we get the claim. □

1.8 Non-negativity vs. SOS: multivariate counterexamples

It turns out that for every $n\ge 2$, there exist non-negative n-variate polynomials which are not SOS.

Proposition 1.12. *For the two-variate degree-six polynomial*

$$f = 1 - 3X_1^2X_2^2 + X_1^2X_2^4 + X_1^4X_2^2$$

the following hold:

(a) f is non-negative.

(b) f attains the value 0.

(c) f is not SOS.

(d) $f - y$ is not SOS for every $y \in \mathbb{R}$.

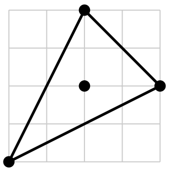
Proof. (a): The inequality for the arithmetic geometric mean of three values yields

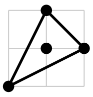
$$\frac{1 + x_1^2x_2^4 + x_1^4x_2^4}{3} \geq \sqrt[3]{1(x_1^2x_2^4)(x_1^4x_2^2)} = x_1^2x_2^2$$

for all $x_1, x_2 \in \mathbb{R}$. This shows that $f \geq 0$ on $\mathbb{R}^2$.

(b): When $x_1, x_2 \in \{-1, 1\}$, the value 0 is attained, because we take the arithmetic and the geometric mean of three ones.

(c): We assume the contrary, $f = f_1^2 + \cdots + f_r^2$ for some polynomials $f_1, \ldots, f_r \in \mathbb{R}[X]$ and get a contradiction using Proposition 1.7. The Newton polytope of f is $2P$, where P is a triangle with vertices $(0,0), (1,2)$ and $(2,1)$.


Newt(f)


P

There are not so many integer points in this triangle: apart from the vertices, it is only the point $(1,1)$. Since $\mathrm{Newt}(f_j) \subseteq P$ for every $j \in [r]$, we get

$$f_j = \sum_{\alpha \in E} c_{j,\alpha} X^\alpha,$$

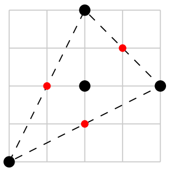
where

$$E := \{(0,0), (1,2), (2,1), (1,1)\}.$$

For the sum of squares we have

$$f = \sum_{j=1}^r f_j^2 = \sum_{j=1}^r \sum_{(\alpha,\beta)\in E^2} c_{j,\alpha}c_{j,\beta}X^{\alpha+\beta}.$$

Thus, if we are interested in expressing a coefficient at the monomial X^γ of the polynomial through the coefficients of the f_j's, we need to check for representations $\gamma = \alpha + \beta$ with $\alpha, \beta \in E$. It is convenient to view the latter equation as an equation for the midpoints of segments with points in $2E$, by rewriting it as $\gamma = \frac{1}{2}(2\alpha + 2\beta)$, where $2\alpha, 2\beta \in E$.



$2E$

One immediately sees that the only way to obtain the representation $(2,2) = \frac{1}{2}(2\alpha + 2\beta)$ is by taking both α and β equal to $(1,1)$. This shows that $\sum_{j=1}^{r} c_{j,(1,1)}^2$ is the coefficient of f at the monomial X^γ with $\gamma = (2,2)$. But this coefficient is -3, and we arrive at $-3 \geq 0$, which is a contradiction.

(d): The proof of (c) can be used without any changes if $y \neq 1$. If $y = 1$, then the Newton polytope of f gets different, because the constant term disappears. In this case, however we see that $f - y$ is not SOS, because f attains negative values. We get $f - 1 = X_1^2 X_2^2(-3 + X_1^2 + X_2^2)$. So $f(x_1, x_2) < 0$ if $x_1, x_2 \in \mathbb{R} \setminus \{0\}$ and the distance of (x_1, x_2) to $(0,0)$ is strictly less than $\sqrt{3}$. □

- Proposition 1.12(c) was originally proved without any explicit use of Newton polytopes; see, for example, [Mar08b]. The use of Newton polytopes is helpful as it makes the proof idea very clear.

- Another useful thing we learn from the proof that uses Newton polytopes is that there is a generalization of the notion of vertex in the world of lattice polytopes (or one can call the world of lattice convex sets, if you like). What we actually proved is the following: if f is SOS and $2P = \mathrm{Newt}(f)$, then we can define $E = P \cap \mathbb{Z}^n$ and consider the set $2E$. The set $2E$ is the set of all points of the lattice $(2\mathbb{Z})^n$ that belong to $\mathrm{Newt}(f)$. This set $2E$ has 'vertices' (which are the vertices of $\mathrm{Newt}(f)$) and 'generalized vertices', which are points γ that cannot be written as $\gamma = \frac{1}{2}(2\alpha + 2\beta)$ with $\alpha, \beta \in E$ and $\alpha \neq \beta$. We have essentially shown that if f is SOS and γ is a generalized vertex of $2E$, then the coefficient of f at X^γ must be non-negative.

Our SDP approach to lower-bounding f does not work at all for f in Proposition 1.12, even though f is pretty simple: only two variables and degree six. This f is called the *Motzkin polynomial.* It was discovered by Motzkin. Hilbert was the first to show (in 1888) that SOS is not always equivalent to non-negativity but he didn't give any explicit examples. An explicit example (Motzkin polynomial) was discovered much later, in 1965.

If we allow three variables, we can find a similar polynomial of degree four.

Exercise 1.13. *Show that the following three-variate degree-four polynomial*

$$f = 1 + X_1^2 X_2^2 + X_2^2 X_3^2 + X_1^2 X_3^2 - 4X_1 X_2 X_3$$

is non-negative, but not SOS.

Solution. To see non-negativity, one can again use the inequality for the geometric and the arithmetic mean: observe that for the four exponent vectors

$$(0,0,0), (2,2,0), (0,2,2), (2,0,2), (1,1,1)$$

the last one is the arithmetic mean of the remaining ones. For showing that f is not SOS, we can use the Newton polytope of f. It can be represented as $2P$, with $P = \operatorname{conv}((0,0,0),(1,1,0),(0,1,1),(1,0,1))$. Here, P is a simplex, and apart from its vertices, there are no other integer points in P. If we could write f as $f = f_1^2 + \cdots + f_r^2$, then we had $\operatorname{Newt}(f_j) \subseteq P$. Due to the observation about P, f_j^2 does not contain the monomial $X_1X_2X_3$. But f does contain this monomial, so we get a contradiction. □

The polynomial from the previous exercise is called the *Choi-Lam-polynomial.*

Here are two exercises that illustrate that passing from the dual problem (1.2) to its SOS-relaxation (1.3) one can get a finite positive gap.

Exercise 1.14. *Consider the homogenization*

$$h(X_1, X_2, X_3) := X_3^6 - 3X_1^2X_2^2X_3^2 + X_1^2X_2^4 + X_1^4X_2^2$$

of the Motzkin polynomial (it is called the Motzkin form*). Show that*

(a) $f = h(X_1, 1, X_3)$ *is non-negative,*

(b) f *is not SOS, but*

(c) $h(X_1, 1, X_3) + c$ *is SOS for some* $c \in \mathbb{R}$*.*

Solution. The examples can be found in [Par03, Example 7.2].

(a) Since the Motzkin polynomial $h(X_1, X_2, 1)$ is non-negative, also by Lemma 1.10, $h(X_1, X_2, X_3)$ is non-negative. Since h is non-negative, then also f is non-negative.

(b) If f were SOS, then by homogenization of the SOS-representations, we'd get that h is SOS. But then also the Motzkin polynomial $h(X_1, X_2, 1)$ would be SOS, which is a contradiction.

(c) We have
$$f = X_3^6 - 3X_1^2X_3^2 + X_1^2 + X_1^4.$$
The disturbing term here is $-3X_1^2X_3^2$, because its coefficient is negative. We get rid of this term using
$$X_1^4 - 3X_1^2X_3^2 = (X_1^2 - \frac{3}{2}X_3^2)^2 - \frac{9}{4}X_3^4.$$
Thus, we arrive at
$$f = \left(X_3^6 - \frac{9}{4}X_3^4\right) + (X_1^2 - \frac{3}{2}X_3^2)^2 + X_1^2$$
The second and third summands are squares. The first summand depends only on X_3 and if we add a sufficiently large constant c, it becomes non-negative and so SOS. One can for example check that $c = 2$ will do (to see that $t^6 - \frac{9}{4}t^4 + 2 \geq 0$ for all $t \in \mathbb{R}$ one can distinguish between $|t| \leq 1$ and $|t| \geq 1$ and estimate t^6 by t^4 in the latter case). □

Exercise 1.15. *Consider the Motzkin form h from Exercise 1.14 and the three-variate degree-twelve polynomial*

$$f = (h + 1)^2.$$

By construction, f is SOS (in fact, f is a square). Show that $f - 1$ is non-negative but not SOS.

Solution. The construction (and the solution) was communicated to me by Claus Scheiderer. Observe that h is not SOS. In fact, if h were SOS, then also the Motzkin polynomial $h(X_1, X_2, 1)$ would be SOS, which is a contradiction. One has $f - 1 = (h + 1)^2 - 1 = h(h + 2)$, which implies that f is non-negative. To see that $f - 1$ is not SOS, note that $f - 1 = 2h + h^2$, where h is homogeneous of degree 6 and h^2 homogeneous of degree 12. Assume the contrary, $f - 1 = 2h + h^2 = \sum_{j=1}^{r} g_j^2 + \cdots + g_r^2$ for some $g_1, \ldots, g_r \in \mathbb{R}[X]$. We introduce another indeterminate Y and evaluate the above equality at $(Y X_1, Y X_2, Y X_3)$. Taking into account the homogeneity we get

$$Y^6 2h(X_1, X_2, X_3) + Y^{12} h(X_1, X_2, X_3) = \sum_{j=1}^{r} g_j(Y X_1, Y X_2, Y X_3)^2.$$

Now we can view the left and the right hand sides as elements of $\mathbb{R}[X_1, X_2, X_3][Y]$ (univariate polynomials in Y with polynomials in X_1, X_2, X_3 as coefficients). Since the left hand side involves only monomials Y^6 and Y^{12}, each $g_j(Y X_1, Y X_2, Y X_3)$ can be written as $Y^3 h_j(X_1, X_2, X_3)$ plus higher order terms (for monomials Y^4, Y^5, Y^6). Comparing the coefficients at Y^6 we arrive at

$$2h(X_1, X_2, X_3) = \sum_{j=1}^{r} h_j(X_1, X_2, X_3)^2.$$

This contradicts the fact that h is not SOS. □

Also other examples can be generated using the above template (using the Choi-Lam-polynomial, we can construct a four-variate polynomial of degree 8 with similar properties).

1.9 Equivalence of non-negativity and SOS for two-variate polynomials of degree four

There is yet another special situation, where both the degree and the dimensions are fixed to be some specific values, in which non-negativity is equivalent to SOS. If, unlike algebraists, you are not very interested in special degrees and dimensions, you can skip this subsection or read only the formulation of Theorem 1.17 below.

We show that for two-variate polynomials of degree four, SOS is equivalent to non-negativity. The result goes back to Hilbert, but here we follow a different proof recently suggested in the literature.

Lemma 1.16. *Let $f \in \mathbb{R}[X]$ be a univariate non-negative polynomial and let $q \in \mathbb{R}[X]$ be a strictly positive quadratic univariate polynomial. Then there exist polynomials $\eta, \xi \in \mathbb{R}[X]$ such that*

$$f = \eta^2 + q\xi^2.$$

Proof. Changing coordinates and rescaling, we can assume that $q = X^2 + 1$. Since f is non-negative and non-zero, its degree is at least two.

If the degree of f is two, we can write f as $f = (X+a)^2 + b^2$ for some $a, b \in \mathbb{R}$. In this case ξ^2 must be a constant, an we show that an appropriate constant can be chosen. We want the constant ξ to be chosen in such a way that $f - q\xi^2$ is a square of a polynomial. Let's compute the coefficients of $f - q\xi^2$:

$$\begin{aligned} f - q\xi^2 &= (X+a)^2 + b^2 - (X^2+1)\xi^2 \\ &= (1-\xi^2)X^2 + 2aX + a^2 + b^2 - \xi^2. \end{aligned}$$

For $f - q\xi^2$ to be a square of a real polynomial, it is necessary that the coefficient at X^2 is non-negative. Thus, we should look for ξ satisfying $\xi^2 \le 1$. Under this condition, the polynomial $f - q\xi^2$ is a square of a linear polynomial if both its roots coincide, which can be expressed as the discriminant being equal to zero. The discriminant of $f - q\xi^2$ is

$$\Delta := 4a^2 - 4(1-\xi^2)(a^2+b^2-\xi^2).$$

As ξ^2 moves from 0 to 1, the discriminant Δ gets changed from $-4b^2$ to $4a^2$. So there is a choice of ξ^2 with $\Delta = 0$.

If the polynomial f is of degree larger than two, we use induction. Factorize f as $f = f_1 f_2$, where f_1 and f_2 are non-constant non-negative polynomials. Such a factorization exists, because f is a product of linear terms and quadratic polynomials (by the fundamental theorem of algebra).

Using the induction assumption for f_1 and f_2, write each of them in the desired form $f_j = \eta_j^2 + q\xi_j^2$. This gives

$$f = (\eta_1^2 + q\xi_1^2)(\eta_2^2 + q\xi_2^2).$$

We will convert the expression above to the desired form $f = \eta^2 + q\xi^2$. This can be done just by defining the right η and ξ and leaving the comparison of the two expressions as a routine verification. But it would probably be better to have a derivation that is easy to track. So, we'll be using a formal root $\sqrt{q}$ of q (which can be rigorously introduced algebraically) and we will use the root of -1, defined by $i^2 = -1$. With this additional objects, we can essentially use the formula $|z_1 z_2| = |z_1| \cdot |z_2|$ for the absolute value of the product of complex numbers $z_1, z_2 \in \mathbb{C}$ as follows:

$$\begin{aligned} f &= \left(\eta_1^2 + (\sqrt{q}\xi_1)^2\right)\left(\eta_2^2 + (\sqrt{q}\xi_2)^2\right) \\ &= \left|\eta_1 + i\sqrt{q}\xi_1\right|^2 \left|\eta_2 + i\sqrt{q}\xi_2\right|^2 \\ &= \left|(\eta_1 + i\sqrt{q}\xi_1)(\eta_2 + i\sqrt{q}\xi_2)\right| \\ &= \left|\eta_1\eta_2 - q\xi_1\xi_2 + i(\sqrt{q}\xi_1\eta_2 + \sqrt{q}\xi_2\eta_1)\right| \\ &= (\underbrace{\eta_1\eta_2 - q\xi_1\xi_2}_{=:\eta})^2 + q(\underbrace{\xi_1\eta_2 + \xi_2\eta_1}_{=:\xi})^2. \end{aligned}$$

This gives a desired expression $f = \eta^2 + q\xi^2$. In case, one does not believe that the intermediate steps in this calculation have a rigorous mathematical meaning, one could just check that

$$(\eta_1^2 + q\xi_1^2)(\eta_2^2 + q\xi_2^2) = (\eta_1\eta_2 - q\xi_1\xi_2)^2 + q(\xi_1\eta_2 + \xi_2\eta_1)^2,$$

is true, which is nothing but checking a polynomial identity. □

Having this lemma, we can now prove the following.

Theorem 1.17 (Hilbert)**.** *Every non-negative two-variate polynomial of degree at most four is SOS.*

Proof. For the case $n = 2, d = 4$, the following proof can be given, which is obtained by simplifying the proof in [PS12] (see also a related proof in [BCR98, Prop. 6.3.4]).

We'll work with homogeneous polynomials in three indeterminates here and we prefer to denote these indeterminates as X, Y, Z in this proof.

A non-negative two-variate polynomial $g \in \mathbb{R}[X, Y]$ of degree at most four can be homogenized to the three-variate non-negative polynomial $f \in \mathbb{R}[X, Y, Z]$ of degree 4 via

$$f(X, Y, Z) := Z^4 g(X/Z, Y/Z).$$

We have seen in Lemma 1.10, that whenever $g \geq 0$ on $\mathbb{R}^2$, the homogenization f is non-negative on $\mathbb{R}^3$. Using this operation, we replace the study of arbitrary non-negative polynomials of degree at most 4 in 2 indeterminates by a study of homogeneous non-negative polynomials of degree exactly 4 in 3 indeterminates. Let's denote by C the cone of these polynomials

$$C := \left\{ f \in \mathbb{R}[X, Y, Z] \, : \, f \text{ homogeneous}, f \geq 0 \text{ on } \mathbb{R}^3, \ \deg(f) = 4 \right\}.$$

Clearly, the cone is finite-dimensional, convex and closed. It is also not hard to see that the cone is pointed. To see this, we need to consider $f \in C \cap (-C)$. Such f satisfies $f = 0$ on $\mathbb{R}^3$. But, since our underlying domain of coefficients is $\mathbb{R}$, the latter implies that f is a zero polynomial. Let's recall that a zero polynomial is a polynomial, whose all coefficients are zero. Thus, out of $f(x, y, z) = 0$ for all $(x, y, z) \in \mathbb{R}^3$ one needs to conclude that all coefficients of f are zero. This is not particularly difficult but still requires a small argument (we leave the verification as an exercise).

We know from convexity theory that every vector in a closed pointed convex cone is a sum of finitely many vectors that lie on the extremal rays of the cone. Thus, it suffices to show the assertion for f lying on the extremal ray of C. Consider the unit sphere

$$S := \left\{ (x, y, z) \in \mathbb{R}^3 \, : \, x^2 + y^2 + z^2 = 1 \right\}.$$

Due to the homogeneity, the condition $f \geq 0$ on $\mathbb{R}^3$ for homogeneous polynomials can be expressed as $f \geq 0$ on S. If a homogeneous polynomial f is strictly positive on S, then f would lie in the interior of C, as one would have $f + h \geq 0$ on S for a homogeneous polynomial h with sufficiently small coefficients. Polynomials in the interior of C are not contained in the extremal rays of C. So, we can assume that $f \in C$ is equal to zero for at least one $(x, y, z) \in S$. Applying a rotation of $\mathbb{R}^3$ around $(0, 0, 0)$ we can assume that $f(0, 0, 1) = 0$.

We can interpret $\mathbb{R}[X, Y, Z]$ as $\mathbb{R}[X, Y][Z]$ and correspondingly write f as

$$f(X, Y, Z) = f_0 Z^4 + f_1 Z^3 + f_2 Z^2 + f_3 Z + f_4,$$

where $f_j \in \mathbb{R}[X, Y]$ is homogeneous of degree j, or the zero polynomial. Since $f(0, 0, 1) = 0$, we get $f_0 = 0$. But then we can also see that $f_1 = 0$. A rigorous way to see this is as follows: We have

$$0 \leq z^{-3} f(x, y, z) = f_1(x, y) + f_2(x, y) z^{-1} + f_3 z^{-2} \to f_1(x, y), \qquad \text{as } z \to +\infty.$$

This shows $f_1(x,y) \geq 0$, where f_1 is a homogeneous polynomial f_1 of degree 1 or equal to 0. This implies $f_1 = 0$.

Thus, we get

$$f(X,Y,Z) = f_2(X,Y)Z^2 + f_3(X,Y)Z + f_4(X,Y).$$

That is, out of the assumption $f(0,0,1)$ we came to the conclusion that the degree of f with respect to Z is at most two. We proceed by looking at the properties of the 'coefficient polynomials' f_2 and f_3 and f_4. Clearly, $f_4 \geq 0$ on $\mathbb{R}^2$ since $f_4(X,Y) = f(X,Y,0)$. On the other hand we have $f_2 \geq 0$ on $\mathbb{R}^2$, which can be shown using an argument similar to the one that we used to show $f_1 \geq 0$ above. We distinguish cases according to the properties of the quadratic form $f_2(X,Y)$.

Case 1: $f_2(X,Y) = 0$. Then $f(X,Y,Z) = f_3(X,Y)Z + f_4(X,Y)$ and we must have $f_3 = 0$. Indeed, if we did have $f_3 \geq 0$, then there would be a point $(x,y) \in \mathbb{R}^2$ with $f_3(x,y) \neq 0$. Then $f(x,y,Z)$ is a polynomial of degree one, and we know that a polynomial of degree one cannot be non-negative, which is a contradiction. Hence $f(X,Y,Z) = f_4(X,Y)$. Since $f_4(X,Y) \geq 0$ on $\mathbb{R}^2$, we conclude that $f_4(X,1) \geq 0$ is non-negative on $\mathbb{R}$. In view of Exercise 1.9, we can write $f_4(X,1)$ as a sum of squares $f_4(X,1) = \sum_{j=1}^r g_j(X)^2$ of polynomials g_j of degree at most two. Homogenizing this we have

$$f(X,Y,Z) = f_4(X,Y) = Y^4 f_4(X/Y,1) = \sum_{j=1}^r (Y^2 g_j(X/Y))^2,$$

where $Y^2 g_j(X/Y)$ is a homogeneous polynomial of degree two.

Case 2: The set of zeros of f_2 is a one-dimensional linear space. Then $f_2 = l^2$ for some non-zero linear form $l \in \mathbb{R}[X,Y]$.

$$f = l^2 Z^2 + f_3 Z + f_4.$$

Whenever $(x,y) \in \mathbb{R}^2$ is such that $l(x,y) = 0$ holds we get $f = f_3 Z + f_4$. But then $f_3(x,y) = 0$ since otherwise, we get a contradiction to $f \geq 0$ in the same way as we did above. It follows that $l(x,y) = 0$ implies $f_3(x,y) = 0$. This means that l is a factor of f_3. Thus, we can write $f_3 = 2lg_2$ for some quadratic polynomial $g_2 \in \mathbb{R}[X,Y]$. Thus,

$$f = (lZ)^2 + 2g_2(lZ) + f_4 = (lZ + g_2)^2 + f_4 - g_2^2.$$

Clearly, $f_4 - g_2^2$ is non-negative and so SOS (just consider $f(x,y,z)$ for $z = -g_2(x,y)/l(x,y)$ with $l(x,y) \neq 0$). It follows that f is SOS.

Case 3: The set of zeros of f_2 is $\{(0,0)\}$. Then

$$f_2 = l_1^2 + l_2^2 \tag{1.6}$$

for some non-trivial linear forms $l_1, l_2 \in \mathbb{R}[X,Y]$. We have a polynomial f which is quadratic in Z, and we want to extract a full square with respect to Z. For doing this in terms of polynomial coefficients we need to multiply f with an appropriate polynomial:

$$4f_2 f = (2f_2 Z)^2 + 2f_3(2f_2 Z) + 4f_2 f_4 = (2f_2 Z + f_3)^2 + 4f_2 f_4 - f_3^2.$$

As above, we see that $4f_2f_4 - f_3^2$ is non-negative. Thus, by Lemma 1.16, there exist bivariate polynomials $\xi(x, y)$ and $\eta(x, y)$ of degrees 2 and 3 respectively such that

$$\eta^2 + \xi^2 f_2 = 4f_2f_4 - f_3^2.$$

This is equivalent to

$$\eta^2 + f_3^2 = f_2(4f_4 - \xi^2). \tag{1.7}$$

Using (1.6), this can be written as

$$(\eta + if_3)(\eta - if_3) = (l_1 + il_2)(l_1 - il_2)(4f_4 - \xi^2),$$

where $i^2 = -1$. The polynomials $l_1 \pm il_2$ have degree one and so they are prime factors. Without loss of generality, we can assume that $l_1 + il_2$ divides $\eta + if_3$. Hence, f_2 divides

$$(\eta + if_3)(l_1 - il_2) = (\eta l_1 + f_3l_2) + i(f_3l_1 - \eta l_2).$$

Since f_2 has real coefficients, f_2 divides both real and the imaginary part of the latter polynomial. So, the following two fractions are polynomials

$$h_1 := \frac{f_3l_1 - \eta l_2}{2f_2} \qquad\qquad h_2 := \frac{\eta l_1 + f_3l_2}{2f_2}.$$

By definition of h_1 and h_2, we have

$$\begin{aligned} h_1^2 + h_2^2 &= \frac{1}{4f_2^2}\Big((\mathrm{re}((\eta + if_3)(l_1 - il_2))^2 + (\mathrm{im}((\eta + if_3)(l_1 - il_2))^2\Big) \\ &= \frac{1}{4f_2^2}|(\eta + if_3)(l_1 - il_2)|^2 \\ &= \frac{1}{4f_2^2}|\eta + if_3|^2 \cdot |l_1 - il_2|^2 \\ &= \frac{(\eta^2 + f_3^2)(l_1^2 + l_2^2)}{4f_2^2} \\ &= \frac{\eta^2 + f_3^2}{4f_2} \\ &= f_4 - \frac{1}{4}\xi^2. \end{aligned}$$

This gives

$$h_1^2 + h_2^2 = f_4 - \frac{1}{4}\xi^2. \tag{1.8}$$

Moreover, from the definition of h_1 and h_2 we also have

$$h_1l_1 + h_2l_2 = \frac{f_3(l_1^2 + l_2^2)}{2f_2} = \frac{1}{2}f_3.$$

Consequently,

$$\begin{aligned} (\xi/2)^2 + (h_1 + l_1Z)^2 + (h_2 + l_2Z)^2 &\overset{(1.8)}{=} f_4 - h_1^2 - h_2^2 + (h_1 + l_1Z)^2 + (h_2 + l_2Z)^2 \\ &= f_4 + 2(\underbrace{h_1l_1 + h_2l_2}_{=\frac{1}{2}f_3})Z + (\underbrace{l_1^2 + l_2^2}_{f_2})Z^2 \\ &= f_2Z^2 + f_3Z + f_4 \\ &= f. \end{aligned}$$

□

Exercise 1.18. *Show that $f \in \mathbb{R}[X]$ is a zero polynomial if and only if $f = 0$ on $\mathbb{R}^n$.*

1.10 Non-negativity vs. SOS: Summary

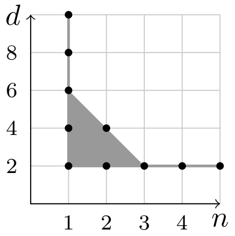
Theorem 1.19. *Let $d \in \mathbb{N}$ and $n \in \mathbb{N}$. Every n-variate non-negative polynomial of degree at most $2d$ is SOS if and only if one of the following conditions is fulfilled:*

(a) $n = 1$, *or*

(b) $d = 1$, *or*

(c) $n = 2, d = 2$.

Proof. See results and examples of this chapter (Exercise 1.9, Exercise 1.11, Proposition 1.12, Exercise 1.13 and Theorem 1.17). □



Remark 1.20. Similar characterizations were also obtained for symmetric polynomials (invariant up to permutation of variables) and for symmetric polynomials even in each variable; see [CL77] and [GKR16].

2 Stellensätze with denominators and Hilbert's 17th problem

In this chapter, too, we deal with $n \in \mathbb{N}$ and indeterminates $X = (X_1, \dots, X_n)$.

2.1 Hilbert's 17th problem

Non-negativity is not equivalent to sos, but can one still certify non-negativity of polynomials algebraically in a different way? This already concerned Hilbert, who included the following problem on his famous list of the 23 problems from the year 1900: is every non-negative polynomial in $\mathbb{R}[X]$ a sum of squares of rational functions?

Recall that we define the field $\mathbb{R}(X)$ of formal quotients f/g with $f, g \in \mathbb{R}[X]$ and $g \neq 0$ with the standard multiplication and addition.

Artin gave a complete positive solution of Hilbert's problem in 1927, and here we present a 'modern version' of Artin's solution. It will turn out that as a byproduct we'll be able to derive a number of Stellensätze, which characterize non-negativity, positivity and equality to zero of a polynomial on a so-called basic closed semialgebraic set.

2.2 Ordered fields

We call a subset P of a field F a *preordering* of the field F if P is closed under addition and multiplication, and every square is an element of P. That is $x + y \in P$ and $xy \in P$ for all $x, y \in P$ and $x^2 \in P$ for every $x \in F$. By $\sum F^2$ we denote the set of all sums of squares of elements of F. Clearly, one has $\sum F^2 \subseteq P$ for every preordering P and $\sum F^2$ itself is a preordering.

If F is a subset of a ring, we'll use the notation $\sum F^2$ for the set of all sums of squares of elements from F. For example, $\sum \mathbb{R}[X]^2$ is the set of all sos-polynomials in variables X with coefficients in $\mathbb{R}$ and $\sum \mathbb{R}(X)^2$ is the set of all sums of squares of rational functions in variables X.

We call a subset P of F an *ordering* of F if P is closed under addition and multiplication and, furthermore, the equalities $P \cup (-P) = F$ and $P \cap (-P) = \{0\}$ hold.

Exercise 2.1. *Show that every ordering is a preordering. For this, it suffices to check that if P is an ordering of the field F, then every square x^2 with $x \in F$ belongs to P.*

Solution. One has either $1 \in P$ or $-1 \in P$. If we had $-1 \in P$, then we also had $(-1)(-1) = 1 \in P$, which is a contradiction. This shows that $1 \in P$. If $x \in F$ and $x \neq 0$, then we either have $x \in F$ or $-x \in F$. In the former case $x^2 \in P$, because x^2 is a square of x and in the latter case $x^2 \in P$, because x^2 is a square of $-x$. □

Every ordering P of F defines a total-order relation $\leq$ on F given by $x \leq y$ if and only if $y - x \in P$. A field F equipped with P (and the total-order relation $\leq$ arising from P) is called an ordered field. For an ordered field, one can also introduce $\geq, >$ and $<$ in a natural way.

Example 2.2. *Let $n = 1$ and consider the field $\mathbb{R}(X)$ of univariate rational functions. We order the subfield $\mathbb{R}$ of $\mathbb{R}(X)$ in a standard way ($\mathbb{R}_+$ is the standard ordering of $\mathbb{R}$). Let's order the whole $\mathbb{R}(X)$ by claiming that $X \geq r$ for every $r \in \mathbb{R}$. Once this is required, it is clear how the rest of $\mathbb{R}(X)$ gets ordered. For example, we have $X^2 \geq X$ and $X^3 \geq X$. We also have $1/X < r$ for every $r \in \mathbb{R}$ with $r > 0$. In this ordering X is infinitely large, X^2 is even larger, $1/X$ is infinitely small positive, $1/X^2$ is even smaller etc.*

Exercise 2.3. *The field $\mathbb{R}(X)$ in the above examples can be ordered in more than one way.*

(a) Try to find other orderings that extend the ordering of $\mathbb{R}$.

(b) Can you describe all such orderings?

Solution. (a): Just by interchanging the roles of X and $1/X$, we get another ordering. In this ordering, $1/X$ is infinitely large. One can also require $-X$ to be infinitely large or $-1/X$ to be infinitely large.

(b): Yet another general possibility (that covers the case of $1/X$ or $-1/X$ being infinitely large) is to fix a number $a \in \mathbb{R}$ and require $X - a$ to be infinitely small positive or infinitely small negative. It is not very hard to convince oneself that the above suggestions cover all possible orderings. In fact, $\mathbb{R}$ is already ordered, and X should occupy some place with respect to the real numbers. We can put it either behind all real numbers or before all real numbers, or fix $a \in \mathbb{R}$ and put X before a or behind a (infinitely close to a). □

We call a field R *real closed* if R is not algebraically closed but $R[\sqrt{-1}]$ is an algebraically closed field. That is, in R we only miss an imaginary unit for representing the roots of polynomial equations. The abstract field $R[\sqrt{-1}]$ is in the same relation to the abstract field R as the field $\mathbb{C}$ of complex numbers to the field $\mathbb{R}$ of real numbers.

Exercise 2.4. *Show the following. For a real closed field R, the set $\sum R^2$ is the unique ordering of R. Even more specifically, an element of R is non-negative if and only if it is a square x^2 with $x \in R$.*

Solution. Let $i := \sqrt{-1}$. Consider an arbitrary ordering of R and let's denote the corresponding order relation (as usual) by $\leq$. If $a \in R \setminus \{0\}$ satisfies $a \geq 0$, then the equation $\lambda^2 = a$ has two roots in $R[i]$. These roots are either of the form $\pm b$ with $b \in R \setminus \{0\}$ or of the form $\pm ib$ with $b \in R \setminus \{0\}$. In the first case $a = b^2$. In the second case $a = -b^2$. Since $b^2 \geq 0$, we see that in the former case $a \geq 0$ and in the latter case $a \leq 0$. Thus, the whole R gets decomposed into squares and 'minus squares' (that intersect at the element 0). This shows that squares are all non-negative elements with respect to our ordering. □

Theorem 2.5. *For every ordered field F there exists a (unique) real closed extension of R. This means R is an extension of the field F, R is a real closed field and the set of all elements in R which are non-negative in R and belong to F is exactly the set of all elements of F that are non-negative in F.*

Proof. The proof is contained in [BCR98] (one of the first chapters). □

2.3 Quantifier elimination and Tarski's transfer principle

Let R be a real closed field (for example, $R = \mathbb{R}$). We start with some terminology

- A *boolean formula* is a formula based on and,or,not operations and involving boolean variables (true/false-variables).
- A quantifier-free formula F is obtained by plugging in relations of the form $f(X) = 0, f(X) > 0, f(X) \geq 0$ etc. for $f \in R[X]$ into a boolean formula. So, a quantifier formula depends on indeterminates. If the underlying polynomials have rational coefficients (that is $f \in \mathbb{Q}[X]$), we say that the formula has rational coefficients.
- A first-order formula is a formula, in which some of the variables are quantified. It has the form
$$F(X) := \mathcal{Q}_1 Y_1 \dots \mathcal{Q}_k Y_k \ \ G(X_1, \dots, X_n, Y_1, \dots, Y_k),$$
where a G is a quantifier free-formula and $\mathcal{Q}_1, \dots, \mathcal{Q}_k \in \{\forall, \exists\}$ are quantifiers.
- Two formulas $F_1(X)$ and $F_2(X)$ are said to be equivalent over R if $F_1(x) = F_2(x)$ for all $x \in R^n$.

Theorem 2.6 (Tarski-Seidenberg quantifier elimination)**.** *Let R be a real closed field. Let F be a first-order formula with rational coefficients. Then there exists an algorithm that constructs a quantifier-free formula G with rational coefficients equivalent to F. The algorithm depends only on F and not on the choice of the underlying real closed field R.*

About the proof. The proof is not particularly hard but a bit long. It proceeds by induction, and one basic thing one should learn to do is counting real roots of a polynomial in an interval (to start the induction). A complete proof can be found in [BCR98, Th. 1.4.6]. If abstraction is not your favorite activity, one can ready the proof setting $R = \mathbb{R}$, but the point is that for every real closed field R the proof works just the same, and this will turn out to be a *crucial* observation when it comes to proving Stellensätze. □

Sets defined by a first-order formula are called *semialgebraic*, these are the sets of the form $\{x \in R^n : F(x) = 0\}$, where $F(X)$ is a first order formula with free variables $X = (X_1, \dots, X_n)$. Semialgebraic sets defined by a system of non-strict resp. strict polynomial inequalities are called *basic closed* resp. *basic open* .

Exercise 2.7. *Let $k, d \in \mathbb{N}$.*

(a) Is $\mathcal{S}_+^k$ semialgebraic? Is it basic closed semialgebraic?

(b) Is $\operatorname{int}(\mathcal{S}_+^k)$ *semialgebraic? Basic open semialgebraic?*

(c) Is the set of all non-negative n-variate polynomials of degree at most $2d$ semialgebraic?

(d) Is the set of all positive n-variate polynomials of degree at most $2d$ semialgebraic?

Remark 2.8 (Quantifier elimination algorithms)**.** Currently the best quantifier elimination algorithms can carry out the quantifier elimination procedure in $O((md)^{O(n^{4(t+1)})})$ arithmetic operations, where m is the number of polynomials involved, d a degree bound and t is the number of quantifier alternations. In particular, if there are no alternations (say, all the quantifiers are existential), the time bound is exponential. In principle, one can solve polynomial optimization problems using quantifier elimination in exponential time (though, no one would dare to do that on large problems). One reason for the running time being so high, is that in certain situation the output (quantifier-free formula equivalent to the given one) is exponentially large.

Remark 2.9 (Fourier-Motzkin elimination)**.** Fourier-Motzkin elimination is a special case of quantifier elimination occurring in the theory of polyhedra and linear programming. We have a system of affine linear inequalities of the form $f_1(x,y) \geq 0, \ldots, f_m(x,y) \geq 0$ in variables $x \in \mathbb{R}^n$ and $y \in \mathbb{R}$ and we want to write the condition that there exists y satisfying $f_1(x,y) \geq 0, \ldots, f_m(x,y) \geq 0$ in terms of x-Variables only. This is possible. The inequalities can be written in the form $y \leq u_i(x)$, $y \geq l_j(x)$ and $g_k(x) \geq 0$ with $i \in I$, $j \in J$ and $k \in K$. That is, each inequality either provides an upper bound on y, or a lower bound on y, or is independent of y. Now, the existence of y can be written as the system $\max_{j \in J} l_j(x) \leq y \leq \min_{i \in I} u_i(x)$, $g_k(x) \geq 0$ for every k. The system can be reformulated without any use of y as $l_j(x) \leq u_i(x), g_k(x) \geq 0$ for all $i \in I, j \in J, k \in K$. One can see that out of m original inequalities one gets up to $\Theta(m^2)$ inequalities after elimination. There exists examples that show that this cannot be avoided. The situation with a general quantifier elimination is pretty similar.

Just to get a feeling how quantifier elimination can be carried out, let us consider a particular example.

Exercise 2.10. *Compute a quantifier-free formula equivalent to*

$$F(p,q) := \exists x \ \big[\ -1 \leq x \leq 1 \ \textit{and} \ x^2 + px + q = 0 \ \big]$$

Draw a sketch of the respective semialgebraic set $S := \big\{(p,q) \in \mathbb{R}^2 \ : \ F(p,q) = 0\big\}$.

Solution. The condition $F(p,q)$ just tells us that the quadratic polynomial $f(X) := X^2 + pX + q$ given by coefficients p and q has a root in the segment $[-1,1]$. Let's figure out how we could characterize this condition in p and q directly, without any use of quantifiers. If the the signs of f at the endpoints of $[-1,1]$ are different, we'll get a root in $[-1,1]$ (by the intermediate value theorem from analysis). So, we'll have a root if $f(-1) \geq 0$ and $f(1) \geq 0$ or $f(-1) \leq 0$ and $f(1) \geq 0$. If $f(-1) < 0$ and $f(1) < 0$, we'll have no roots in $[-1,1]$, because f is convex. So, we are left with the case $f(-1) \geq 0, f(1) \geq 0$. Note that the global optimum of f is attained at $x = -p/2$. Thus, $x = -p/2$ is in $[-1,1]$ and $f(-p/2) \leq 0$, then f has a root in $[-1,1]$.

To sum up, if at least one of the following cases occurs, f has a root in $[-1,1]$:

(a) $f(-1) \leq 0, f(1) \geq 0$ or

(b) $f(-1) \geq 0, f(1) \leq 0$ or

(c) $f(-1) \geq 0, f(1) \geq 0$, $-1 \leq -p/2 \leq 1$, $f(-p/2) \leq 0$.

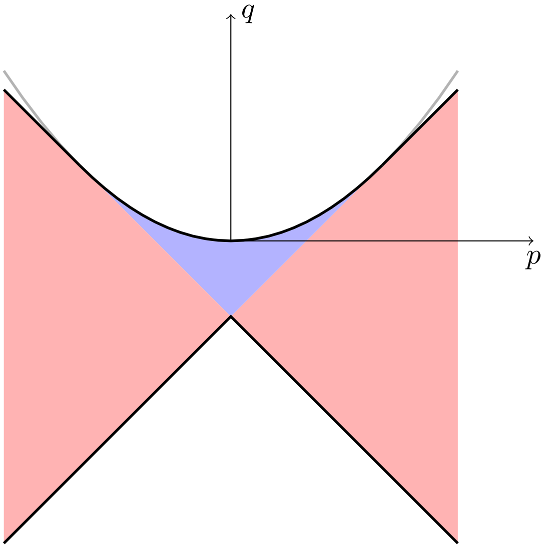
Converse, if neither of the above cases occurs, we must have $f(-1) < 0, f(-1) < 0$ or $f(-1) > 0, f(1) > 0$. In the former situation, f has no roots in $[-1,1]$, which follows from convexity of f. In the latter situation, if $-p/2$ lies outside $[-1,1]$, then f has no roots in $[-1,1]$ (because f either grows or falls in $[-1,1]$ and remains strictly positive on the whole segment). If $-p/2 \in [-1,1]$, we have $f(-p/2) > 0$ (for, otherwise, the third condition would be fulfilled), which shows that f is strictly positive on $\mathbb{R}$ and thus also on $[-1,1]$. Thus, we have characterized $F(p,q)$ by the above three conditions. In terms of p and q the conditions can be formulated as

(A) $1 - p + q \leq 0, 1 + p + q \geq 0$ or

(B) $1 - p + q \geq 0, 1 + p + q \leq 0$

(C) $1 - p + q \geq 0, 1 + p + q \geq 0$, $-1 \leq -p/2 \leq 1$, $p^2 - 4q \geq 0$.

Here is a picture of the semialgebraic set described by $F(p,q)$:



This exercise illustrates the following situation. Assume that you are optimizing a function $g(p,q)$ subject to constraints $-1 \leq x \leq 1$ and $x^2 + px + q = 0$. Even though, it looks as if your feasible set is nice (it is just a piece of a surface described by a simple equation $x^2 + px + q = 0$), projecting out x and writing your constraints without x, you see that you are optimizing g over a weird semialgebraic subset of $\mathbb{R}^2$ (which is not basic: it cannot be described by a system of polynomial inequalities). This is in contrast to the situation in linear optimization. In linear optimization, a projection of a polyhedron is a polyhedron again. This exercise shows that a projection of a basic semialgebraic set is not necessarily basic. □

As a consequence of Tarski's principle, we obtain the following crucial corollary. It says that whenever a first-order formula over $\mathbb{R}$ is satisfiable over a bigger real closed field R, then it is also satisfiable over our original smaller field $\mathbb{R}$.

Corollary 2.11 (Tarski's transfer principle)**.** *Let R be a real ordered field, which is an ordered extension of the field $\mathbb{R}$. Let $F(X)$ be a first-order formula with coefficients in $\mathbb{R}$. If there exists $x^* \in R^n$ such that $F(x^*)$ is fulfilled, then there exists also $x' \in \mathbb{R}^n$ such that $F(x')$ is fulfilled.*

Proof. The proof is borrowed from [BCR98, Cor. 1.4.7]. It is known that every real ordered field can be extended to a real closed field (see [BCR98]). So, without loss of generality we assume that R is real closed.

Let $Y = (Y_1, \dots, Y_k)$ be the quantified variables used in $F(X)$. Each polynomial $f \in \mathbb{R}[X, Y]$ involved in $F(X)$ can be written as $f(X, Y) = g(X, Y, a)$ where $a \in \mathbb{R}^m$ is independent of f and where $g \in \mathbb{Q}[X, Y, Z]$ (a polynomial with rational coefficients) and $Z = (Z_1, \dots, Z_m)$ are additional indeterminates. For example, if $f(X,Y) = \sqrt{2}X^2 + \sqrt{3}(X - Y) + \sqrt{2}Y^2$ we can introduce $g(X, Y, Z_1, Z_2) = Z_1X^2 + Z_2(X - Y) + Z_1Y^2 \in \mathbb{Q}[X, Y, Z_1, Z_2]$ with $f(X, Y, a) = f(X, Y)$ for $a = (\sqrt{2}, \sqrt{3})$.

That is, we put all the real coefficients of all the polynomials involved in $F(X)$ into a vector a. This gives rise to a first-order formula $G(X, Z)$ such that $G(X, a) = F(X)$. By Tarski-Seidenberg, $\exists X : G(X, Z)$ can be turned into a quantifier-free form. So, there exists a quantifier-free formula $H(Z)$ with rational coefficients equivalent to $\exists X : G(X, Z)$ over every real closed field. We plug in $Z = a$ and see that $\exists X : G(X, a)$ is equivalent to $H(a)$, where $H(a)$ is independent of X and so $H(a)$ is either true or false. By assumption $\exists X : G(X, a) = \exists X : F(X)$ is true over R. So, $H(a)$ is true and so $\exists X : G(X, a) = \exists X : F(X)$ is true over every real closed field containing $\mathbb{R}$ including $\mathbb{R}$ itself. This gives the assertion. □

2.4 Solution of Hilbert's 17th problem

Lemma 2.12 (Serre 1947)**.** *Let F be field of zero characteristic (this means that $\underbrace{1 + \cdots + 1}_{k} \neq 0$ in F for every $k \in \mathbb{N}$). Let T be a preordering of F and let $f \in F \setminus T$. The inclusion-maximal preordering P with the properties $T \subseteq P$ and $f \notin P$ is an ordering.*

Proof. Note that the existence of P follows from Zorn's lemma (and is by this based on the axiom of choice).

We recall that a preodering is a set closed under addition, multiplication and taking squares. We also shortly mention that a preordering is closed under divisions, too. If both a and b are in the preordering and b is not zero, then $a/b = ab(1/b)^2$ and so, we see that immediately.

We first observe that $-1 \notin P$. In fact, if we had $-1 \in P$, then P is going to be the whole field F, which is a contradiction. To see this, it suffices to observe that every element of F is a difference of two squares. Just use the simple identity $4x = (1 + x)^2 - (1 - x)^2$ and divide it by four (division by four requires $1 + 1 \neq 0$ in F – this is the reason of having the assumption on the characteristic of the field F). All squares are in the preordering just by definition. If -1 is in P, then also the minus squares are in the preordering. But then also the differences of the squares are in the preordering and we get $P = F$, which contradicts $f \notin P$.

We also note that $-f \in P$. If we had $-f \notin P$, we would get that the set $P - fP$ containing both P and $-f$ is a preordering. In fact, $P - fP$ is closed under addition, because P is closed under addition, $P - fP$ contains all squares, because P contains all squares and $P - fP$ is closed under multiplication, which can be seen by expanding the brackets in $(P - fP)(P - fP)$ and using the fact that P is a preordering.

Furthermore, since P does not contain f, also the set $P - fP$ does not contain f. For if $f \in P - fP$, then we had $f = a - fb$ with $a, b \in P$, which gives $(1+b)f = a$. We have $b+1 \neq 0$, since otherwise $-1 = b \in P$ would be an element of P. Consequently, $f = a/(1+b) \in P$, which is a contradiction.

Furthermore, we observe that for every $g \in F$, one has $g \in P$ or $-g \in P$, which means $P \cup (-P) = F$. If, say, $g \notin P$, we'll show that $-g \in P$. Consider the set $P + gP$. As above, $P + gP$ is a preordering containing P, and the containment is proper because $g \in P + gP$ and $g \notin P$. By the inclusion-maximality of P under the property $f \notin P$, we must have $f \in P + gP$, which means that $f = a + gb$ holds for some $a, b \in P$. Then $-bg = a + (-f) \in P$, because $a \in P$ and we have seen that $-f$ belongs to P, too. One cannot have $a - f = 0$, since otherwise $f = a \in P$. This implies $gb = f - a \neq 0$. Thus, $b \neq 0$ and by this $-g = (a-f)/b \in P$.

It remains to show that $P \cap (-P) = \{0\}$ holds. Let $g \in P \cap (-P)$, that is $g, -g \in P$. Then $g = 0$, for otherwise, we had $-1 = g(-g)(1/g)^2 \in P$, which contradicts $-1 \notin P$, which we have derived above. □

Theorem 2.13 (Artin 1927). *Let $f \in \mathbb{R}[X]$ be a non-negative polynomial. Then $f \in \sum \mathbb{R}(X)^2$ (that is, f is a sum of squares of rational functions).*

Proof. Assume to the contrary that $f \notin \sum \mathbb{R}(X)^2$. Applying the previous lemma to the preordering $\sum \mathbb{R}(X)^2$ of $\mathbb{R}(X)$, we see that there exists an ordering P on $\mathbb{R}(X)$ with respect to which f is negative. We fix $R := \mathbb{R}(X)$ and endow R with the ordering P (this can be viewed as follows: P is a kind of evaluation of rational functions at an 'abstract point' and by construction f is strictly negative at this abstract point). By construction, there exists $(x_1, \ldots, x_n) \in R^n$ such that $f(x_1, \ldots, x_n) < 0$. Indeed $X_1, \ldots, X_n$ are rational functions and by this elements of R and so $f(x_1, \ldots, x_n) < 0$ holds for $(x_1, \ldots, x_n) \in R^n$ with $x_1 = X_1, \ldots, x_n = X_n$. Since R is a field extending $\mathbb{R}$ and since the coefficients of the polynomial f are in $\mathbb{R}$, Tarski's principle shows that there also exists $(x_1, \ldots, x_n)$ belonging to $\mathbb{R}^n$ such that $f(x_1, \ldots, x_n) < 0$. This contradicts the non-negativity of f. Consequently, $f \in \sum \mathbb{R}(X)^2$. □

Remark 2.14. The previous proof is not constructive, so it does not provide a method to write a non-negative polynomial as a sum of squares of rational functions. E.g., from the proof, it is not clear *how* the Motzkin polynomial (or other non-sos-polynomials) can be represented as a sum of squares of rational functions.

Remark 2.15. It is known that 2^n squares suffice. For $n = 1$, we have seen this in our proof. The case $n = 2$ is due to Hilbert (1893). The general case is due to Pfister (1965); see [BCR98].

2.5 An algebraic certificate of non-negativity

Now, let's consider a more general setting, in which we want to certify whether a polynomial f is positive or non-negative on a basic closed semi-algebraic set K. The solution of Hilbert's 17th problems gives a certificate for $K = \mathbb{R}^n$ and, using similar principles, we can also do the case of a more general K.

Given s polynomials $g = (g_1, \ldots, g_s) \in \mathbb{R}[X]^s$, we introduce the basic closed set

$$\{g \geq 0\} := \{x \in \mathbb{R}^n \,:\, g(x) \geq 0\}$$

We'll also use the longer notation

$$\{g_1 \geq 0, \ldots, g_s \geq 0\} := \{x \in \mathbb{R}^n \,:\, g_1(x) \geq 0, \ldots, g_s(x) \geq 0\}.$$

and we'll use an analogous notation for strict inequalities and equalities. With g we also associate the so-called preordering generated by g

$$\begin{aligned}\mathcal{P}(g) &:= \mathcal{P}(g_1, \dots, g_s) \\ &:= \left\{ \sum_{e:=(e_1,\dots,e_s)\in\{0,1\}^s} \sigma_e g_1^{e_1} \dots g_s^{e_s} \ : \ \sigma_e \text{ sos for every } e \in \{0,1\}^s \right\}.\end{aligned}$$

For example, for $s = 2$, $\mathcal{P}(g_1, g_2)$ consists of polynomials $\sigma_{0,0} + \sigma_{1,0}g_1 + \sigma_{0,1}g_2 + \sigma_{1,1}g_1g_2$ where $\sigma_{0,0}, \sigma_{1,0}, \sigma_{1,0}, \sigma_{1,1}$ are sos.

Clearly, $\mathcal{P}(g)$ is indeed a preordering and it is the inclusion-minimal preordering containing all the polynomials $g_1, \dots, g_s$. Each $f \in \mathcal{P}(g)$ is non-negative on $\{g \geq 0\}$. Furthermore, we can recover $\{g \geq 0\}$ from $\mathcal{P}(g)$ since we obviously have

$$\{g \geq 0\} = \{x \in \mathbb{R}^n \ : \ f(x) \geq 0 \text{ for all } f \in \mathcal{P}(g)\}.$$

The latter equality allows us to view $\mathcal{P}(g)$ as a kind of dual object associated to the semi-algebraic set $\{g \geq 0\}$.

While the condition that $f \geq 0$ on $\{g \geq 0\}$ is a 'geometric' non-negativity condition (f non-negative on a semi-algebraic) set, $f \in \mathcal{P}(g)$ is a sort of 'algebraic' non-negativity condition. Though it is not true that the two conditions are equivalent, there is a strong relationship and, for understanding the ideas of the following proofs, it is nice to keep in mind the analogy between the two notions.

We also introduce the set

$$\mathcal{P}^0(g) := \mathcal{P}(g) \cap (-\mathcal{P}(g))$$

- A representation of f as an element of $\mathcal{P}(g)$ provides an algebraic evidence for $f \geq 0$ on $\{g \geq 0\}$.
- A representation of f as an element of $\mathcal{P}^0(g)$ provides an algebraic evidence for $f = 0$ on $\{g \geq 0\}$.

So, it is nice to think of $f \in \mathcal{P}(g)$ and $f \in \mathcal{P}^0(g)$ as some kind of '≥ 0' and '$= 0$' conditions.

We recall that an ideal I in a commutative ring $(R, +, \cdot)$ is an additive subgroup of R with the property $IR \subseteq I$.

Lemma 2.16. *Let $n \in \mathbb{N}$. Let $g = (g_1, \dots, g_s) \in \mathbb{R}[X]^s$, let $P := \mathcal{P}(g)$ and $I := \mathcal{P}^0(g)$. Then the following hold:*

(a) I is an ideal of $\mathbb{R}[X]$.

(b) If I is proper (that is $I \neq \mathbb{R}[X]$), then $-1 \notin P$ (analogy: if there is an element not equal to zero 'in the abstract sense', the constants like -1 are not non-negative 'in the abstract sense').

(c) If $p, q \in P$ and $p + q \in I$, then $p, q \in I$ (analogy: if the sum of non-negative elements is zero, then all the summands are zero).

Proof. (a): Since P is closed under addition, I too is closed under addition. Since $0 \in P$ we have $0 \in I$ and it is clear that $I + I = I$. For every $f \in I$ also $-f \in I$. So $(I, +)$ is a subgroup of $\mathbb{R}[X]$. Let $q \in I = P \cap (-P)$. As we did above, we use a formula telling us that every element of a $\mathbb{R}[X]$ is a difference of two squares.

For every $p \in \mathbb{R}[X]$, one has $4p = (1+p)^2 - (1-p)^2$. Since $(1+p)^2 q \in P \cap (-P)$ and $(1-p)^2 q \in P \cap (-P)$, we conclude that $4pq \in P \cap (-P)$. Dividing by 4, we get $pq \in P \cap (-P)$. Here is a bit more detailed explanation how we can get the conclusion: From $q \in P$ and $-q \in P$, we get $(1+p)^2 q \in P$ and $-(1-p)^2 q \in P$, respectively. This yields $(1+p)^2 q - (1-p)^2 q \in P$, meaning $4pq \in P$. Dividing by 4, we get $pq \in P$. Completely analogously, we also get $pq \in -P$.

(b): If $-1 \in P$, then using $4p = (1+p)^2 - (1-p)^2$, we conclude that $P = \mathbb{R}[X]$. Hence $I = P \cap (-P) = \mathbb{R}[X]$, which is a contradiction.

(c): Since, $p, q \in P$, we have $p + q \in P$. So we just need to show $-p, -q \in P$. Indeed, showing this, we get $-(p+q) \in P$ and conclude $p + q \in P \cap (-P) = I$. By $p + q \in I$ we have $-p - q \in P$. So $-p \in q + P \subseteq P$ and $-q \in p + P \subseteq P$. □

An ideal $\mathcal{P}^0(g)$ should help us certify the condition $f = 0$ on $\{g \geq 0\}$. However, $\mathcal{P}^0(g)$ may have an undesired property of being non-prime. Here is just a small example with $g = (g_1, g_2)$, $g_1 = X_1X_2, g_2 = -X_1X_2$. In this case, the set $\{g \geq 0\} = \{X_1X_2 = 0\}$ is just the union of two coordinate axes. The polynomial X_1 does not belong to $\mathcal{P}(g)$, because X_1 is not non-negative on $\{g \geq 0\}$. Analogously, the polynomial X_2 does not belong to $\mathcal{P}(g)$ because X_2 is not non-negative on $\{g \geq 0\}$. It is clear that the product X_1X_2 belongs to $\mathcal{P}^0(g)$. So, we conclude that neither X_1 nor X_2 belongs to $\mathcal{P}^0(g)$ but their product X_1X_2 does belong to $\mathcal{P}^0(g)$. This shows that the ideal $\mathcal{P}^0(g)$ is not prime. The reason of $\mathcal{P}^0(g)$ being non-prime is the fact that the variety $\{X_1X_2 = 0\}$ is not irreducible.

In the proof of our positivstellensatz we'll need a step of enlarging $\mathcal{P}^0(g)$ to a prime ideal (which dually can be viewed as picking an irreducible component in a variety).

Exercise 2.17. *Show that, if J is a prime ideal and $m \in \mathbb{N}$, then $p^m \in J$ implies $p \in J$.*

Solution. For $m = 1$ this is clear. If $m \geq 2$, then $pp^{m-1} \in J$ and since J is prime, p or p^{m-1} is in J. In the former case, we get the desired assertion. In the latter case, by induction, we deduce that $p \in J$. □

Lemma 2.18. *In the notation of Lemma 2.16, we have*

(a) If I is a proper subset of $\mathbb{R}[X]$ and if J is a minimal prime ideal containing I, then, for all $p, q \in P$ satisfying $p, q \in P$ and $p + q \in J$, one has $p, q \in J$.

(b) For the ideal J in (a) and the field F of fractions over the integral domain $\mathbb{R}[X]/J$, the set

$$P' := \left\{ \sum_{e \in \{0,1\}^s} \sigma_e g^e \; : \; \sigma_e \in \sum F^2 \right\}.$$

is a proper preordering of F.

Proof. See [Mar08b, Prop. 2.1.7]. (a): It is known that there exists an inclusion-minimal prime ideal $J \subseteq \mathbb{R}[X]$ with $I \subseteq J$ (follows by applying Zorn's lemma to the family of all prime ideals that contain I as a subset). It is also known that in this case for every $p \in J$, there exists an integer $m \geq 0$ and an element $q \in \mathbb{R}[X] \setminus J$ with $p^m q \in I$; see any basic book in commutative algebra, for example, [AM16, Prop. 1.14]. Here is an illustration in the context of number theory of how this works. The ideal $12\mathbb{Z}$ is contained in two inclusion-minimal prime ideals $2\mathbb{Z}$ and $3\mathbb{Z}$. If we raise an element of $2\mathbb{Z}$ to the power two and then multiply with $3 \notin 2\mathbb{Z}$, we get an element of $12\mathbb{Z}$. An illustration in terms of polynomial rings is as follows. The ideal $X_1^4 X_2 \mathbb{R}[X_1, X_2]$ is contained in two prime ideals $X_1\mathbb{R}[X_1, X_2]$ and $X_2\mathbb{R}[X_1, X_2]$. Taking a fourth power of an element of $X_1\mathbb{R}[X_1, X_2]$ and multiplying it with $X_2 \notin X_1\mathbb{R}[X_1, X_2]$, we get an element of $X_1^4 X_2\mathbb{R}[X_1, X_2]$.

So, if $p + q \in J$, then $(p + q)^n u \in I$ holds for some $n \geq 0$ and $u \notin J$. This gives $u^2(p + q)^n \in I$. We can always assume that n is odd. If n is even, just increase n by one, by multiplying with $(p + q)$. Using binomial expansion we can write $u^2(p + q)^n$ as a non-negative linear combination of the terms $u^2p^iq^{n-i}$ with $i \in \{0, \dots, n\}$. Depending on whether i is odd or even, we can write the latter term as a square multiplied by p or a square multiplied by q. So we conclude that $u^2(p+q)^n$ is an element of I which can be written as a sum of terms belonging to P. By Lemma 2.16(c), every term $u^2p^iq^{n-i}$ belongs to I. In particular also $u^2p^n \in I$. Since $I \subseteq J$, J is prime and $u \notin J$, we get $p \in J$. Indeed, $u^2 \notin J$ since $u \notin J$ and J is prime (see Exercise 2.17). Then from $u^2 \notin J$ and $u^2p^n \in J$, we deduce that $p^n \in J$ (since J is prime). From p^n we deduce $p \in J$ (see Exercise 2.17). Since p and q play the same roles, we also get $q \in J$.

(b): This is related to [Mar08b, Prop. 2.1.6]. Assume the contrary, that is, $P' = F$. Then $-1 \in P'$. This means

$$-1 = \sum_{e \in \{0,1\}^s} \sigma_e g^e$$

for some $\sigma_e \in \sum F^2$, $g^e := g_1^{e_1} \cdots g_s^{e_s}$ and $e = (e_1, \dots, e_s)$. Note that F is a field of fractions over the integral domain $\mathbb{R}[X]/J$. The terms σ_e are sums of squares of fractions of elements of $\mathbb{R}[X]/J$. So, setting f to be the product of all the denominators in the mentioned fractions, we get

$$-f^2 = \sum_{e \in \{0,1\}^s} \sigma_e g^e$$

for some $\sigma_e \in \sum(\mathbb{R}[X]/J)^2$ and $f \in R[X]/J$ with $f \neq 0$ in $\mathbb{R}[X]/J$. The latter identity can be written as an identity modulo J. So, coming from $\mathbb{R}[X]/J$ to $\mathbb{R}[X]$, this gives

$$-f^2 \in \underbrace{\sum_{e \in \{0,1\}^s} \sigma_e g_e}_{=:h} + J.$$

for some $f \in \mathbb{R}[X]$ and $f \notin J$ and some choices $\sigma_e \in \sum \mathbb{R}[X]^2$.

We have thus shown $h + f^2 \in J$. Here, $h \in P$ by construction, and $f^2 \in P$, because f^2 is a square. By (a), we conclude $f^2 \in J$. Since J is prime, we get $f \in J$, which is a contradiction. □

2.6 A Farkas-type lemma for polynomial inequalities

Theorem 2.19 (Certifying infeasibility of a polynomial system; a Farkas Lemma for POP)**.** *Let $g = (g_1, \dots, g_s) \in \mathbb{R}[X]^s$. Then $\{g \geq 0\} = \emptyset$ if and only if $-1 \in \mathcal{P}(g)$.*

Proof. The sufficiency is clear, so we need to prove the necessity. Assume $-1 \notin \mathcal{P}(g)$. We will prove that $\{g \geq 0\}$ is non-empty. Consider the ideal $I := \mathcal{P}^0(g)$. Since $\mathcal{P}(g) \supseteq I$, we have $-1 \notin I$, that is, I is a proper subset of $\mathbb{R}[X]$. There exists a prime ideal J containing I as a subset and being minimal with respect to inclusion with respect to this property. Let F be the field of fractions of $\mathbb{R}[X]/J$. The preordering P' from Lemma 2.18 is a proper subset of F and so, using Lemma 2.12, we can find an ordering P with $P \supseteq P'$. With the ordering P, the field F becomes an ordered field. Consider the natural homomorphisms $f \mapsto f + J \mapsto \frac{f+J}{1}$ on the spaces $\mathbb{R}[X] \to \mathbb{R}[X]/J \to F$. These homomorphisms send $X_1, \dots, X_n \in \mathbb{R}[X]$ to some $x_1, \dots, x_n \in F$. By construction, we have $g_1(x_1, \dots, x_n) \geq 0, \dots g_s(x_1, \dots, x_n) \geq 0$ for the constructed $(x_1, \dots, x_n) \in F^n$ (with respect to the ordering P we've fixed for F). Thus, we found a solution of the system $g \geq 0$ in the space F^n. By Tarski-Seideberg transfer principle (Corollary 2.11), there exists also a solution in $\mathbb{R}^n$. □

Remark 2.20. A representation of -1 as an element of $\mathcal{P}(g)$ is an infeasibility certificate for the system $g_1(x) \geq 0, \dots, g_s(x) \geq 0$. This theorem does not provide a direction control of the size of such a certificate, as there is nothing mentioned about the degree of the sos-polynomials σ_e in the representation $-1 = \sum_{e \in \{0,1\}^s} \sigma_e g^e$. Tools from real algebra allow to show that such a certificate is realizable with degrees of the polynomials σ_e bounded in terms of n and the degrees of g_i's only, but it is likely that such degree bounds for σ_e are necessarily very huge.

2.7 Stellensätze with denominators

The following theorem contains a Positivstellensatz (positive-locus theorem), a Nichtnegativstellensatz (non-negative-locus theorem) and a Nullstellensatz (zero-locus theorem).

Theorem 2.21 (Krivine 1964, Stengle 1974; see [Mar08b, Thm. 2.2.1])**.** *Let $g = (g_1, \dots, g_s) \in \mathbb{R}[X]^s$, let $K = \{g \geq 0\}$ and $P = \mathcal{P}(g)$. Then for $f \in \mathbb{R}[X]$ the following conditions hold.*

(a) $f > 0$ on K if and only if there exist $p, q \in P$ with $pf = 1 + q$.

(b) $f \geq 0$ on K if and only if there exists an integer $m \geq 0$ and $p, q \in P$ such that $pf = f^{2m} + q$.

(c) $f = 0$ on K if and only if there exists an integer $m \geq 0$ such that $-f^{2m} \in P$.

Proof. (a): The sufficiency is easy: if $pf = 1 + q$, then there exists no $x \in K$ with $f(x) \leq 0$: if $g(x) \geq 0$ and $f(x) \leq 0$, then $p(x)f(x) = 1 + q(x)$ yields a contradiction, because the left hand side is ≤ 0 and the right hand side is > 0. Conversely, if $f > 0$ on K, then $\{g \geq 0, -f \geq 0\}$ is empty and so by the Theorem 2.19 (the 'polynomial Farkas-lemma'), we know that $-1 \in \mathcal{P}(g, -f)$. Clearly, $\mathcal{P}(g, -f) = \mathcal{P}(g) - f\mathcal{P}(g) = P - fP$ and so we get $-1 = q - fp$ for some $q, p \in P$, which is exactly what we need to derive.

(b): Sufficiency : if $pf = f^{2m} + q$ then there exists no $x \in K$ with $f(x) < 0$: if we had $g(x) \geq 0$ and $f(x) < 0$, then $f^{2m}(x) + q(x) > 0$ and $p(x)f(x) \leq 0$, which is a contradiction. Conversely, assume $f \geq 0$ on K. We assume $f \neq 0$ as otherwise the assertion is clear. We'll get to the situation of (a) using lifting. Condition $f \geq 0$ on $\{g \geq 0\}$ can be phrased as $f > 0$ on $\{g \geq 0, f \neq 0\}$. We can rewrite $f(X) \neq 0$ as $f(X)Y = 1$ using an addition indeterminate Y. This gives $f > 0$ on $\{g(X) \geq 0, f(X)Y = 1\} \subseteq \mathbb{R}^{n+1}$. Rewriting the equality condition through inequality gives $f > 0$ on $\{g(X) \geq 0, 1 - f(X)Y \geq 0, f(X)Y - 1 \geq 0\}$. By (a), we get $p(X,Y)f(X) = 1 + q(X,Y)$ with $p, q \in \mathcal{P}(g, 1 - f(X)Y, -1 + f(X)Y)$. Substituting $Y = 1/f(X)$. This gives $p(X, 1/f(X))f(X) = 1 + q(X, 1/f(X))$. Multiplying through by a sufficiently large power $f(X)^{2m}$ of $f(X)$ clears the denominators and we get the desired assertion.

(c): Sufficiency: if $-f^{2m} \in P$, then there exists no $x \in K$ and $f(x) \neq 0$: indeed, for such x we would have $-f^{2m}(x) < 0$, but the condition $-f^{2m} \in P$ tells us that one has $-f^{2m}(x) \geq 0$, which is a contradiction. Conversely, if we have $f = 0$ on K, then $f \geq 0$ on K and $f \leq 0$ on K and we can use the argument of (b). We have $p_1 f = f^{2m} + q_1$ and $p_2(-f) = f^{2m} + q_2$ for some $p_1, p_2, q_1, q_2 \in P$ (using the same m in both relations is possible, as we can see from the proof of (b)). Multiplication of both equalities yields $-p_1 p_2 f^2 = f^{4m} + (q_1 + q_2) f^{2m} + q_1 q_2$. So we see that $-f^{4m} \in P$. □

2.8 Remarks on Stellensätze

(a) Positive solution of Hilbert's 17th problem says that the condition $f(X) \geq 0$ can be characterized algebraically as $f \in \sum \mathbb{R}(X)^2$. Theorem 2.21(b) gives yet another certificate: $f(X) \geq 0$ iff $pf = f^{2m} + q$ for some $p, q \in \sum \mathbb{R}[X]^2$ and some integer $m \geq 0$. It turns out that the certificate of (b) is in a way stronger, as we can convert it to the $\sum \mathbb{R}(X)^2$-certificate as follows. Assume $f \neq 0$ (as otherwise, things are clear). In this case p is not a zero polynomial as well, since the right hand side $f^{2m} + q$ is not a zero polynomial. So we can rewrite $pf = f^{2m} + q$ as $f = \frac{1}{p}(f^{2m} + q)$ and, rewriting this as $f = (\frac{1}{p})^2 p(f^{2m} + q)$, we see that $f \in \sum \mathbb{R}(X)^2$.

(b) What happens when we consider equality constraints, say $g = 0$ rather than $g \geq 0$? Rewriting $g = 0$ as $g \geq 0, -g \geq 0$, we can use Theorem 2.21. A polynomial f is positive on $\{g = 0\}$ if and only if $pf = 1 + q$, where $p, q \in \mathcal{P}(g, -g)$. Note however that every polynomial is a difference of two squares. This shows that $\mathcal{P}(g, -g) = \sum \mathbb{R}[X]^2 + I$, where

$$I = \mathbb{R}[X]g_1 + \cdots + \mathbb{R}[X]g_s$$

is the so-called ideal generated by $g_1, \ldots, g_s$. Thus, working with equality constraints we have a particularly simple certificate of positivity. This observation can be extended to non-negativity condition and to the case where both inequality and equality constraints are present.

(c) In the classical algebraic geometry (over an algebraic field like complex numbers), one would try to describe polynomials that vanish on an $\{g = 0\} = \{g_1 = 0, \ldots, g_s = 0\}$. We want can characterize $f = 0$ on $\{g = 0\}$ algebraically. By

Theorem 2.21(c), we've got $-f^{2m} \in \mathcal{P}(-g, g)$. Since $\mathcal{P}(g, -g) = \sum \mathbb{R}[X]^2 + I$, with I as above, we have the certificate $-f^{2m} = \sum \mathbb{R}[X]^2 + I$ for the condition $f = 0$ on $\{g = 0\}$. This is analogous to the Nullstellensatz from classical algebraic geometry.

(d) Of course, one may wonder whether one really needs denominators. Our intention was to describe the preordering

$$\tilde{\mathcal{P}}(g) := \{f \in \mathbb{R}[X] \, : \, f(x) \geq 0 \text{ on } \{g \geq 0\}\}$$

and we would prefer a simple description if possible. Theorem 2.21 gives one such description, which is quite involved: the certificate $pf = f^{2m} + q$ involves two polynomials p and q and an integer $m \geq 0$. So, it would be nice if we had, say $\tilde{\mathcal{P}}(g) = \mathcal{P}(g)$. Unfortunately, this equality doesn't always hold. We have seen, this equality does not hold for every $n \geq 2$ if we have no constraints. It's known that when $n \geq 3$ and $\{g \geq 0\}$ has non-empty interior, we've got just the same situation, namely $\tilde{cP}(g) \neq \mathcal{P}(g)$. This means that characterizing non-negativity on $\{g \geq 0\}$ algebraically is not an easy task. Strict positivity is characterized in a nicer way. Computationally, dealing with strict positivity rather than non-negativity we do not lose much. Apart from that, for posivitity, in a number of cases there are denominator-free certificates (this is the topic of the following chapter).

(e) One serious difference to the Farkas lemmas from linear programming is that the degree of sos-polynomials in our certificates can be extremely large. The maximum degree has not been estimated exactly so far. But there is an extremely huge upper bound that has recently been established in [LPR14]. It was shown that if a polynomial f of degree d is non-negative on $\mathbb{R}^n$, then $qf = p$, where q, p are sos-polynomials of degree at most

$$2^{2^{2^{d^{4^n}}}} .$$

I do not know, up to what extent the bound can be improved.

(f) Stellensätze could have also been derived with respect to an arbitrary real closed field. This has imporatant quantitative consequences in the spirit of the previous remark. Real algebraic geometry contains results that give us as a corollary of the latter fact that the degrees of the polynomials $p, q \in \mathcal{P}(g)$ involved in the certificates can be bounded in terms of n, and the degrees of f and $g_1, \ldots, g_s$. However, the theory does not tell us what the bounds exactly are.

3 Positivstellensätze without denominators

Let $X = (X_1, \ldots, X_n)$ be indeterminates and $n \in \mathbb{N}$. Consider the basic semi-algebraic set $\{g \geq 0\}$ given by $g = (g_1, \ldots, g_s) \in \mathbb{R}[X]^s$. The template questions of this chapter are: Does the condition $f > 0$ on $\{g \geq 0\}$ imply that $f \in \mathcal{P}(g)$? Can one use another set rather than $\mathcal{P}(g)$ in such a context? We'll see positive answers in a number of situations and provide links to linear and semidefinite programming. The presentation in this chapter is based on [Ave13].

3.1 Affine version of Farkas lemma

Let's recall that $\operatorname{cone}(X)$ is the convex conic hull of X. One of the basic Nichtnegativstellensätze without denominators is the Farkas lemma. Its affine version can be formulated as follows. We'll say that a polynomial is linear if it is of degree at most one.

Lemma 3.1 (Affine version of Farkas lemma)**.** *Let $f, a_1, \ldots, a_m \in \mathbb{R}[X]$ be an n-variate linear polynomials, and let the polyhedron $K = \{a_1 \geq 0, \ldots a_m \geq 0\}$ be non-empty. Then the following conditions are equivalent:*

(i) $f \geq 0$ *on* K

(ii) $f \in \operatorname{cone}(1, a_1, \ldots, a_m)$.

Proof. Exercise □

3.2 Pólya: a positivstellensatz on a simplex

The templates for the following results will be: if $f > 0$ on a compact set K, then there exists some algebraic representation of f providing the evidence that $f \geq 0$ on K. So, you see that it is not a characterization, but an implication: you require strict positivity and have the evidence for non-negativity only. Such results can of course be converted into a characterization of positivity. It suffices to invoke the above implication for $f - \varepsilon$ in place of f, where $\varepsilon > 0$ is small enough so that $f - \varepsilon$ is still positive on K.

The following theorem is about homogeneous polynomials, which are positive on the standard simplex. For dealing with exponent vectors of homogeneous polynomials, we introduce the notation

$$\overline{E}_d^n := \left\{\alpha \in \mathbb{Z}_+^n \,:\, |\alpha| = d\right\}.$$

We also use the notation $\alpha! = \alpha_1! \cdot \ldots \cdot \alpha_n!$ for $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{Z}_+$.

Theorem 3.2 (Pólya 1928)**.** *Let $f \in \mathbb{R}[X]$ be a homogeneous polynomial, with $f > 0$ on the simplex*

$$\Delta := \left\{(x_1, \ldots, x_n) \in \mathbb{R}_+^n \,:\, x_1 + \cdots + x_n = 1\right\}.$$

Then there exists $N \in \mathbb{Z}_+$ such that all coefficients of $(X_1 + \cdots + X_n)^N f(X)$ are non-negative.

Proof. We write f as $\sum_{\alpha \in \overline{E}^n_d} c_\alpha X^\alpha$. The expression $(X_1 + \cdots + X_n)^N$ can be expanded so that we get

$$\begin{aligned} g &:= (X_1 + \cdots + X_n)^N f \\ &= \sum_{\beta \in \overline{E}^n_N} \frac{N!}{\beta!} X^\beta \sum_{\alpha \in \overline{E}^n_d} c_\alpha X^\alpha \\ &= \sum_{\alpha \in \overline{E}^n_d, \beta \in \overline{E}^n_N} c_\alpha \frac{N!}{\beta!} X^{\alpha+\beta}. \end{aligned}$$

We see that g is a homogeneous polynomial of degree $d + N$:

$$g = \sum_{\gamma \in \overline{E}^n_{d+N}} A_\gamma X^\gamma$$

whose coefficients can be expressed through the coefficients of f by:

$$A_\gamma = \sum_{\alpha \in \overline{E}^n_d \,:\, \alpha \le \gamma} c_\alpha \frac{N!}{(\gamma - \alpha)!} \tag{3.1}$$

We want to analyze the behavior of A_γ for growing N. To this end, we factor out an appropriate term in the above sum describing A_γ:

$$A_\gamma = \frac{N!(N+d)^n}{\gamma!} \sum_{\alpha \in \overline{E}^n_d : \alpha \le \gamma} c_\alpha \frac{\gamma!}{(\gamma - \alpha)!(N+d)^n}$$

In the latter sum, the coefficients c_α are multiplied by values depending on N. In order to see how these values behave for $N \to \infty$, we express them differently. Since both $\gamma!$ and $(\gamma - \alpha)!$ are products of n factorials and since $|\alpha| = d$, we get

$$\frac{\gamma!}{(\gamma - \alpha)!(N+d)^n} = \prod_{i=1}^n \frac{\gamma_i!}{(\gamma_i - \alpha_i)!(N+d)^{\alpha_i}}$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)$ and $\gamma = (\gamma_1, \ldots, \gamma_n)$. The quotient $\gamma!/(\gamma_i - \alpha_i)!$ is the product of α_i many values $\gamma_i - \alpha_i + 1, \ldots, \gamma_i$. Hence

$$\frac{\gamma_i!}{(\gamma_i - \alpha_i)!(N+d)^{\alpha_i}} = \left(\frac{\gamma_i - \alpha_i + 1}{N+d}\right) \cdots \left(\frac{\gamma_i}{N+d}\right)$$

Summarizing this gives

$$A_\gamma = \frac{N!(N+d)^d}{\gamma!} \sum_{\alpha \in \overline{E}^n_d : \alpha \le \gamma} c_\alpha \prod_{i=1}^n \left(\frac{\gamma_i - \alpha_i + 1}{N+d}\right) \cdots \left(\frac{\gamma_i}{N+d}\right).$$

We can discard the condition $\alpha \le \gamma$ and let the sum go over all $\alpha \in \overline{E}^n_d$, for if $\alpha_i > \gamma_i$ for some $i \in [n]$, then the respective term in the product occurring in the latter representation of A_γ is equal to zero.

In this context, it will be convenient to use the notation

$$(x)^m_t := x(x-t) \cdots x(x-(m-1)t)$$

Note that for $t \to 0$, $(x)_t^m \to x^m$, so for small t, the value $(x)_t^m$ is a perturbed version of x^m. In this notation A_γ can be represented as

$$A_\gamma = \frac{N!(N+d)^d}{\gamma!} \sum_{\alpha \in \overline{E}_d^n} c_\alpha \left(\frac{\gamma_1}{N+d}\right)_{1/(N+d)}^{\alpha_1} \cdots \left(\frac{\gamma_n}{N+d}\right)_{1/(N+d)}^{\alpha_n}$$

Since one has $|\gamma| = N + d$, we get $\gamma/(N+d) \in \Delta$. Consider the polynomial

$$f_t(X) = \sum_{\alpha \in \overline{E}_d^n} c_\alpha (X_1)_t^{\alpha_1} \cdots (X_n)_t^{\alpha_n}.$$

The polynomial function $f_t(x)$ converges to $f(x)$ uniformly on Δ as $t \to 0$ (recall that the uniform and the pointwise convergence are equivalent for continuous functions on a compact set). Thus, $f_t(x) > 0$ for all $x \in \Delta$ if t is small enough. Up to a positive multiple, our A_γ coincides with $f_t(\gamma/(N+d))$ for $t = \frac{1}{N+d}$. It follows that $A_\gamma > 0$ for all γ, if N is large enough. □

Remark 3.3. Let's denote by $\overline{P}_{n,d}(\Delta)$ the closed convex cone of all homogeneous n-variate polynomials of degree d that are non-negative on the simplex Δ. Pólyas theorem provides a sequence of polyhedral cones $\overline{P}_{n,d}^N(\Delta)$ approximating the cone arbitrarily well, as $N \to \infty$. In fact, the condition that all coefficients of $(X_1 + \cdots + X_n)^N f$ are non-negative is a system of linear inequalities for the coefficients of f. The polynomial f occurs linearly in the expression $(X_1 + \cdots + X_n)^N f$, or to put it a bit differently, the map $f \mapsto (X_1 + \cdots + X_n)^N f$ is linear. For $N = 0$, the condition is just that all coefficients of f are non-negative. Thus, for $N = 0$, the set $\overline{P}_{n,d}^N(\Delta)$ is just a simplicial cone. For large N explicit inequalities on the coefficients of f can be found in the proof. By (3.1), the inequalities are

$$\sum_{\alpha \in \overline{E}_d^n \,:\, \alpha \le \gamma} c_\alpha \frac{1}{(\gamma - \alpha)!} \ge 0 \qquad\qquad \forall \gamma \in \overline{E}_{N+d}^n.$$

Remark 3.4. Quantitative aspects of Pólya's theorem (how large is N, depending on f) were investigated by Reznick & Powers [PR01]. They give a choice of N depending on d, n and $\min_{x \in \Delta} f(x)$. Their paper also contains a complete proof of Pólya's theorem.

3.3 Pólyas theorem and copositivity

A symmetric matrix $A \in \mathcal{S}^n$ is called co-positive if $\langle x, Ax \rangle \ge 0$ holds for all $x \in \mathbb{R}_{\ge 0}^n$. This means the quadratic polynomial $f(x) = \langle x, Ax \rangle$ is non-negative on the standard simplex Δ. If $\langle x, Ax \rangle > 0$ holds for all $x \neq 0$, then A is called strictly copositive and the corresponding quadratic polynomial $f(x)$ is strictly positive on the standard simplex. Pólyas theorem addresses strict positivity as a special case. Checking whether a matrix is copositive (strictly copositive) is hard. This can also be seen from the fact that a number of NP-hard problems can be formulated as conic problems over the copositive cone (conic problems will be the topic of Chapter 4; roughly, a conic program is a problem of optimization of a linear objective function over an affine slice of a give cone). See also the survey [Dür10]. Copositive programming is a special case of inequality constrained quadratic programming.

Example 3.5. *To illustrate Remark 3.3, we'll consider approximations of the copositive cone for $n = 2$. We consider a quadratic form $f(x, y)$ given as $f(x, y) = ax^2 + bxy + cy^2$. So, in this case $\overline{P}_{n,d}(\Delta) = \overline{P}_{2,2}(\Delta)$ is a three-dimensional cone. It will be more convenient for us to look at the slices of the cone and so we fix a normalization $2a + 2b + c = 2$ of coefficients of f, which corresponds to slicing $\overline{P}_{2,2}(\Delta)$ by a hyperplane (the slice by the hyperplane $a + b + c = 1$ is unbounded, see Exercise 3.6 below). It is known and not very difficult to see that the cone $\overline{P}_{2,2}(\Delta)$ is generated by the polynomial xy and the psd quadratic forms (that is, each $f \in \overline{P}_{2,2}(\Delta)$ is $f = \alpha g + \beta xy$, where $\alpha, \beta \geq 0$ and g is psd). Note that f is psd if and only if $a \geq 0, c \geq 0, ac - (\frac{b}{2})^2 \geq 0$ holds (this is a basic fact in linear algebra). We look at the the equality $ac - (\frac{b}{2})^2 = 0$ which, in view of $2a + 2c + b = 2$, can be formulated as $ac - (1 - (a + c))^2 = 0$. This equation describes an ellipse tangent to the a and c axes and lying in the non-negative orthant.*

The best is to write the condition in $a+c$ and $a-c$. Using $\xi = a+c$ and $\eta = a-c$ and multiplying by 4, this can be formulated as

$$\xi^2 - \eta^2 - 4(1 - \xi)^2 = 0.$$

This is

$$-3\xi^2 + 8\xi - 4 - \eta^2 = 0$$

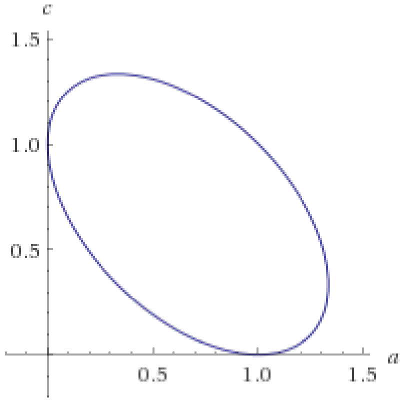
which gives

$$3\xi^2 - 8\xi + 4 + \eta^2 = 0$$

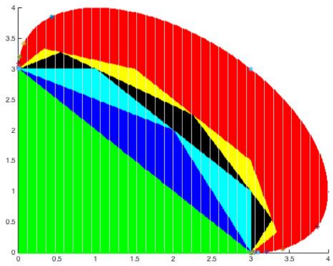
and can be written as

$$3(\xi - 4/3)^2 + \eta^2 = \frac{4}{3}.$$

So we get an equation of the ellipse (that is, the boundary of the psd forms sliced by $2a + 2c + b = 2$ is an ellipse). We remark that from $ac - (1 - (a + c))^2 = 0$ one gets $a \geq 0$ and $c \geq 0$. Furthermore both $a = 0$ and $c = 0$ are possible (by choosing $a = 0, c = 1$ and $a = 1, c = 0$). Here is a picture of it generated with WolframAlpha:



Here is a picture, generated by Benjamin Peters using Matlab that shows how (the slices of) $\overline{P}^N_{2,2}(\Delta)$ approximate (the slice of) $\overline{P}_{2,2}(\Delta)$.



As N grows, $\overline{P}_{2,2}^N(\Delta)$ is growing larger and larger so that every interior point of the red region is eventually covered. But for some of these points it takes very long to be covered.

Exercise 3.6. *Show that the slice of*

$$\overline{P}_{2,2}(\Delta) := \left\{(a,b,c) \in \mathbb{R}^3 \,:\, f := ax^2 + bxy + cy^2 \geq 0 \text{ on } \Delta\right\}$$

by the hyperplane $a + b + c = 1$ is unbounded.

Solution. Clearly, the slice is non-empty, because all (a,b,c) with $a+b+c = 1$ and $a,b,c \geq 0$ are in the slice. It suffices to find $(a',b',c') \in \overline{P}_{2,2}(\Delta)$ with $a'+b'+c' = 0$. Adding arbitrary non-negative multiples of this to any point of the slice we do not fall out of the slice. This shows that a slice contains a ray and so is unbounded. As we know from the convexity theorem, unboundedness of a closed convex set can always be certified by providing a ray that is contained in the set. Finding such an element is easy: $(x-y)^2 = x^2 - 2xy + y^2$ is non-negative ant it yields $a' = 1, b' = -2$ and $c' = 1$. Done! □

Exercise 3.7. *Show that $\overline{P}_{2,2}(\Delta)$ is the convex conic hull of all psd quadratic forms $q(x,y)$ and the polynomials xy. (Sketch of the solution: clearly, the mentioned convex conic hull is a subset of $\overline{P}_{2,2}(\Delta)$. To prove the converse, it suffices to show that every point on an extremal ray of $\overline{P}_{2,2}(\Delta)$ is psd quadratic form or coincides with xy up to a multiple. Note that $\overline{P}_{2,2}(\Delta)$ is pointed and full-dimensional. Every point f on the extremal ray is on the boundary ans so has a zero in Δ. If it has a zero in the relative interior of Δ it is a square and so a psd form. If it has a zero in an endpoint of Δ, there are the following cases. If both endpoints are zeros of f, then f coincides with xy up to a non-negative multiple. If only one endpoint is zero, then we tweak on f so that it remains non-negative on Δ using x^2 or y^2. This shows that f is not extremal.)*

3.4 Handelman: a positivstellesatz on a polytope

Pólya's theorem can be used to derive a positivstellensatz on polytopes.

For $a_1, \ldots, a_m \in \mathbb{R}[X]$ and $a = (a_1, \ldots, a_m)$ we introduce the semiring generated by a as

$$\mathcal{S}(a) := \left\{ \sum_{e=(e_1,\ldots,e_m)\in E} c_e a_1^{e_1} \cdots a_m^{e_m} \,:\, c_e \in \mathbb{R}_{\geq 0} \;\forall e \in E,\; E \subseteq \mathbb{Z}_+^m \text{ finite} \right\}.$$

The set $\mathcal{S}(a)$ consists of conic combinations of products of powers of $a_1, \ldots, a_m$. It is clear, that representation of f as an element of $\mathcal{S}(a)$ gives a certificate for non-negativity of f on $\{a \geq 0\}$.

Remark 3.8 (Pólya on general simplices)**.** Pólya's theorem applied in a straightforward way yields a positivstellensatz on an arbitrary simplex. If $a_1, \ldots, a_{n+1}$ are linear polynomials and $S := \{a_1 \geq 0, \ldots, a_{n+1} \geq 0\}$ is an n-dimensional simplex, then every polynomial f strictly positive on S belongs to $\mathcal{S}(a)$. The sketch of the argument is as follows. Up to rescaling, one can assume $a_1 + \cdots + a_{n+1} = 1$. The original variables $X_1, \ldots, X_n$ can be expressed through $a_1, \ldots, a_{n+1}$, which will become our 'new variables', and if f has degree at most d, we can replace each monomial in f by a homogeneous polynomial of degree d in 'new variables' $a_1, \ldots, a_{n+1}$. After this, applying Pólya's theorem, we get a desired conclusion.

If $f \in \mathbb{R}[X]$ is a polynomial of degree at most d and l is a linear homogeneous polynomial, then $g = l^d f(X/l)$ is polynomial, too. Moreover, g is a homogeneous polynomial and we call it a *homogenization* of f with respect to l.

Let's see an example for how this kind of homogenization works. If say $f = X_1 + X_1X_2 + X_1X_2X_3^2$ and $l = X_1 + X_2 + X_3$, then the degree of f is 4. There is a monomial of degree one and a monomial of degree two. We multiply each of these monomials by an appropriate power of l to get the homogeneous polynomial

$$g = X_1 l^3 + X_1X_2 l^2 + X_1X_2X_3^2.$$

Theorem 3.9 (Positvstellensatz of Handelman)**.** *Let $a = (a_1, \ldots, a_m) \in \mathbb{R}[X]^m$ be polynomials of degree at most one and let the polyhedron $S := \{a \geq 0\}$ defined by the inequalities $a \geq 0$ be non-empty and bounded. Then every polynomial $f \in \mathbb{R}[X]$ strictly positive on S belongs to $\mathcal{S}(a)$.*

Proof. Without loss of generality, we can assume that $S \subseteq \mathbb{R}^n_{\geq 0}$ (because we can translate S into the non-negative orthant $\mathbb{R}^n_{\geq 0}$, and translation is just an affine change of coordinates in the space $\mathbb{R}^n$).

Every polynomial is bounded on S so that the polynomial $X_1 + \cdots + X_n + a_1(X) + \cdots + a_m(X)$ has an upper bound $t \in \mathbb{R}_{\geq 0}$ on S. Let

$$q = t - (X_1 + \cdots + X_N) - (a_1(X) + \cdots + a_m(X)).$$

By construction $q \geq 0$ on S. We introduce new indeterminates $Y = (Y_1, \ldots, Y_m)$ and Z and the polynomials

$$\sigma = \frac{1}{t}(X_1 + \cdots + X_n + Y_1 + \cdots + Y_m + Z)$$

and

$$g = f(X) + c\sum_{i=1}^{m}(a_i(X) - Y_i)^2$$

in these indeterminates. Here, $c \in \mathbb{R}_{>0}$ is a constant that will be fixed in what follows.

Let Δ be the simplex given by

$$\Delta := \left\{(x, y, z) \in \mathbb{R}^{n+m+1}_{\geq 0} \,:\, \sigma(x, y, z) = 1\right\}.$$

The Δ is up to rescaling a standard simplex (and so it is as good as a standard simplex for our purposes). Let A be a subset of Δ given by

$$A = \{(x, y, z) \in \Delta \,:\, a_i(x) = y_i \ \forall i \in [m]\}$$

The polynomials f and g coincide on A and so $f > 0$ on A. Above, we've lifted our f to a polynomial g in a larger space and in that large space, the positivity of f on the original set S is translated to the positivity of g on A.

Since $f > 0$ on A, f is also positive on a small compact neighborhood of A. In Δ and outside this small neighborhood, g can be made arbitrarily large, by choosing $c > 0$ sufficiently large. Indeed, if we are away from A, the sum occurring in the definition of g can be bounded from below by a positive constant and so, choosing the factor c in the definition of c sufficiently large we can ensure that $g > 0$ on the whole Δ. Clearly, one can construct a homogeneous polynomial g_0 such that $g = g_0$ on Δ: just homogenize g with respect to σ, by setting $g_0 := \sigma^{\deg g} g(X/\sigma, Y/\sigma, Z/\sigma)$.

Applying Pólya's theorem to g_0, we get that for some large enough integer $N \in \mathbb{Z}_+$ all coefficients of $\sigma^N(X, Y, Z)g_0(X, Y, Z)$ are non-negative. Substituting $Z = t - (X_1 + \cdots + X_n) - (Y_1 + \cdots + Y_m)$, the polynomial σ is turned to 1. Subsequent substitutions $Y_i = a_i(X)$, turn the polynomial g to f. We thus, conclude that $f = g_0(X, a_1(X), \ldots, a_m(X), q(X))$ is in the semiring

$$\mathcal{S}(X_1, \ldots, X_m, a_1(X), \ldots, a_m(X), q(X)).$$

Note that $X_1, \ldots, X_m, q(X)$ do not contribute to this semiring, because by Farkas' lemma (Lemma 3.1), these polynomials of degree at most 1 are in $\operatorname{cone}(1, a_1, \ldots, a_m)$. This shows that the latter semiring coincides with $\mathcal{S}(a)$. So, the assertion follows. □

3.5 Handelman's theorem and linear programming

Let's borrow the notation $a = (a_1, \ldots, a_m)$ and S of Handelman's theorem and let $f \in \mathbb{R}[X]$ be arbitrary. In view of Handelman's theorem, the polynomial optimization problem

$$\min \{f(x) \,:\, a \geq 0\}$$

over the polytop $S = \{a \geq 0\}$ has the dual formulation

$$\max \{y \in \mathbb{R} \,:\, f - y \geq 0 \text{ on } S\} = \sup \{y \in \mathbb{R} \,:\, f - y \in \mathcal{S}(a)\}.$$

We really need to pass to supremum, because Handelman's theorem is a positivstellensatz (it is about 'strict' positivity). Now we can modify this problem by replacing the whole $\mathcal{S}(a)$ by its truncated version

$$\mathcal{S}_E(a) := \left\{ \sum_{e \in E} c_e a^{e_1} \cdots a^{e_m} \,:\, c_e \geq 0 \ \forall e \in E \right\}$$

where $E \subseteq \mathbb{Z}_+^m$ is finite. The problems

$$\max \{y \,:\, f - y \in \mathcal{S}_E(a)\} \tag{3.2}$$

give lower bounds on the original minimization problem. In view of Handelman's theorem, the hierarchy of such problems with, say, $E = E_d^n$ and $d \in \mathbb{N}$ gives optimal values that converge to the optimal value of the original problem, as d grows.

Unfortunately, the convergence may be quite slow (because Handelman is based on Pólya and so it inherits the convergence problems of Pólya).

Let's also note that (3.2) is a linear problem. The objective is linear (it's just one variable y). The constraint $f - y \in \mathcal{S}_E(a)$ is the linear equality system

$$y + \sum_{e=(e_1,\ldots,e_m)\in E} c_e a_1^{e_1} \cdots a_m^{e_m} = f$$

for the decision variable $y \in \mathbb{R}$ and the non-negative decision variables $c_e \in \mathbb{R}_{\geq 0}$. Indeed, c_e's and y occur linearly in this expression. The polynomial f is the right hand side.

Remark 3.10. Applicability of Handelman-based approaches (including analysis of convergence etc.) has been discussed in several sources.

3.6 Quadratic modules and constrained polynomial optimization

If $a = (a_1, \ldots, a_s) \in \mathbb{R}[X]^m$ are arbitrary polynomials, then the set

$$\mathcal{M}(a) := \{g_0 + g_1 a_1 + \cdots + g_s a_s \,:\, g_0, \ldots, g_s \text{ sos}\}$$

is called the quadratic module generated by $a = (a_1, \ldots, a_s)$. Thus, the general constrained polynomial optimization problem

$$\inf \{f(x) \,:\, x \in \mathbb{R}^n, a_1(x) \geq 0, \ldots, a_s(x) \geq 0\} \tag{3.3}$$

over the set $\{a \geq 0\}$ can be relaxed to

$$\sup \{y \in \mathbb{R} \,:\, f - y \in \mathcal{M}(a)\}. \tag{3.4}$$

In fact, the supremum is a lower bound on the infimum, because the elements of $\mathcal{M}(a)$ are non-negative on $\{a \geq 0\}$. The condition $f - y \in \mathcal{M}(a)$ can be formulated as a semidefinite constraint. For this purpose it will be convenient to introduce the family $\mathcal{S}^\infty$ of infinite-size of symmetric matrices with only finitely many non-zero entries. For such matrices A the psd property can be defined in a usual way. So, we also have a psd cone $\mathcal{S}_+^\infty$ in that space. Let

$$m(X) = (X^\alpha)_{\alpha \in \mathbb{Z}_+^n}$$

be the infinite vector containing all possible monomials X^α.

A polynomial g is sos if and only if $g = m(X)^\top Z m(X)$ where $Z = (z_{\alpha,\beta})_{\alpha,\beta\in\mathbb{Z}_+^n}$ is psd. Thus, the condition $f - y \in \mathcal{M}(a)$ can be written as

$$f(X) - y = m(X)^\top Z_0 m(X) + \sum_{j=1}^s a_j(X) m(X)^\top Z_j m(X), \tag{3.5}$$

where $Z_0, \ldots, Z_s \in \mathcal{S}_+^\infty$ are infinite psd matrices, whose entries are indexed by vector exponents $\alpha = (\alpha_1, \ldots, \alpha_k) \in \mathbb{Z}_+^n$. Thus, our relaxed problem (3.4) is an infinte sdp

$$\sup \left\{ y \in \mathbb{R} \,:\, Z_0, \ldots, Z_s \in \mathcal{S}_+^\infty,\ y + m(X)^\top Z_0 m(X) + \sum_{j=1}^s a_j(X) m(X)^\top Z_j m(X) = f(X) \right\} \tag{3.6}$$

The latter problem can be modified to a finite problem, since $\mathcal{S}_+^\infty$ essentially contains $\mathcal{S}_+^k$ for every k. That is, the latter problem can be truncated to the problem

$$\sup \left\{ y \in \mathbb{R} \,:\, Z_0, \ldots, Z_s \text{ sdp},\ y + m_d(X)^\top Z_0 m_d(X) + \sum_{j=1}^{s} a_j(X) m_d(X)^\top Z_j m_d(X) = f(X) \right\}, \tag{3.7}$$

where $m_d(X) = (X^\alpha)_{\alpha \in E_d^n}$ and $d \in \mathbb{N}$. It is clear that the optimal values of (3.7) are lower bounds for the optimal (3.6) and one can approximate the optimum of (3.6) arbitrarily well by choosing d sufficiently large.

In what follows we try to establish the cases of equivalence of (3.3) and (3.4) by deriving positivstellensätze that are based on the quadratic modules.

Remark 3.11 (What kind of certificate should we prefer?)**.** So far, we've seen several sets used for certifying positivity and leading to various relaxations of polynomial optimization. The sets are

- the semiring $\mathcal{S}(a)$ (used in Handelman's theorem when a_j's are linear).
- the quadratic module $\mathcal{M}(a)$ (introduced above).
- the preordering $\mathcal{P}(a)$ (can also be used just in the same way as $\mathcal{M}(a)$ was used above).

Which of them can one use? Which of them are the better ones? As for the first question, there is a guarantee of convergence of the respective hierarchies if we have established a respective stellensatz (though, convergence may be really slow). As for the second question, comparison of such sets is not an obvious matter. Assume that $f > 0$ on $\{a \geq 0\}$ and we can certify non-negativity of f on $\{a \geq 0\}$ by representing it as an element of $\mathcal{M}(a)$ and also by representing it as an element of $\mathcal{P}(a)$. Will the representation as an element of $\mathcal{M}(a)$ be shorter? This is not really clear! Note that both representations would involve sos-polynomials and we do not know much about their degrees.

3.7 Positivstellensätze involving quadratic modules

In this section, well see that quadratic modules yield positivstellensätze on bounded sets, if we add an additional special polynomial to the generators of the underlying quadratic module. As a consequence, we'll also see that the assertion of Handelman's theorem remains true if we replace $\mathcal{S}(a)$ by $\mathcal{M}(a)$.

Exercise 3.12. *Show the following: Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ be indeterminates. Let E_+^n and E_-^n, respectively, be the set of all vectors $e \in \{-1, 1\}^n$ with an even resp. odd number of entries equal to -1. Then*

$$X_1 \cdot \ldots \cdot X_n \pm Y_1 \cdot \ldots \cdot Y_n = \frac{1}{2^{n-1}} \sum_{e \in E_\pm^n} \prod_{i=1}^{n} (X_i + e_i Y_i).$$

In particular, $X_1 \cdot \ldots \cdot X_n \pm Y_1 \cdot \ldots \cdot Y_n$ belong to the semiring generated by $X_1 + Y_1, \ldots, X_n + Y_n, X_1 - Y_1, \ldots, X_n - Y_n$.

Solution. Induction on n. □

Lemma 3.13. *Let $f = \sum_\alpha c_\alpha X^\alpha \in \mathbb{R}[X] = \mathbb{R}[X_1, \dots, X_n]$. Let $\rho > 0$. Then there exists $t \in \mathbb{R}$ such that $t + f$ and $t - f$ belong to the preordering generated by $\rho - \|X\|^2 = \rho - (X_1^2 + \cdots + X_n^2)$.*

Proof. We use $t = \sum_\alpha |c_\alpha|(\rho + 1)^{|\alpha|}$. Since the definition of t does not depend on changing f to $-f$ (because $|c_\alpha| = |-c_\alpha|$), it suffices to check the assertion for $t+f$. We have

$$t + f = \sum_\alpha |c_\alpha|((\rho+1)^{|\alpha|} + \operatorname{sign}(c_\alpha)X^\alpha).$$

Let $\alpha \neq 0$. Applying Exercise 3.12 to $(\rho + 1)^{|\alpha|} \pm X^\alpha$, we conclude that $t + f$ is in the semiring generated by all $\rho + 1 \pm X_i$ with $i \in [n]$. Now, it suffices to see that $\rho + 1 \pm X_i$ is in the preordering generated by $\rho - \|X\|^2$. Due to the symmetry we can assume $i = 1$. Since changing X_1 to $-X_1$ does not change $\|X\|^2$, it suffices to look at $\rho + 1 + X_1$. We have

$$\rho + 1 + X_1 = \frac{1}{2}\big((\rho + 1) + (1 + X_1)^2 + X_2^2 + \cdots + X_n^2 + (\rho - \|X\|^2\big)$$

□

The following lemma will be applied in the case $\{a \geq 0\} \subseteq B$ by choosing B to be a large box.

Lemma 3.14. *Let $a = (a_1, \dots, a_s) \in \mathbb{R}[X]^s$ and let $f \in \mathbb{R}[X]$ be strictly positive on $\{a \geq 0\}$. Let $B \subseteq \mathbb{R}^n$ be compact. Then there exists $g \in \mathcal{M}(a)$ such that $f - g$ is strictly positive on B.*

Proof. The set $a(B) := \{a(x) : x \in B\}$ is compact and so there exists γ with $a(B) \subseteq (-\infty, 2\gamma]^s$. Consider $T := \{x \in B : f(x) \leq 0\}$. Since $f > 0$ on $\{a \geq 0\}$, we get $a(T) \cap [-2\varepsilon, 2\gamma]^s = \emptyset$ for a sufficiently small $\varepsilon > 0$. This can also be formulated as the inequality $f(x) > 0$ being fulfilled for all $x \in B$ satisfying $a_i(x) \geq -2\varepsilon$ for all $i \in [s]$. We now introduce the univariate polynomial $h(t) := t\left(\frac{t-\gamma}{\gamma+\varepsilon}\right)^{2N}$, where $N \in \mathbb{N}$. On $[0, 2\gamma]$, the polynomial is small for large N. In fact, on $[0, 2\gamma]$, we have $0 \leq h(t) \leq \gamma(\gamma/(\gamma + \varepsilon))^{2N} =: c(N)$. On the other hand when $t \leq -2\varepsilon$, $h(t)$ is negative and $|h(t)|$ is big: we have $-h(t) \geq 2\varepsilon((\gamma + 2\varepsilon)/(\gamma + \varepsilon))^{2N} =: C(N)$. We plug a_j's into h and take the sum, obtaining $g(X) := \sum_{j=1}^s h(a_j(X))$. For $x \in B$, in the case $a_j(x) \geq -2\varepsilon$ for all $j \in [s]$, the terms $h(a_j(x))$ with $a_j(x) \geq 0$ are small and the terms $h(a_j(x))$ with $a_j(x) \leq 0$ are negative. Thus, if N is large enough, we get $f(x) - g(x) > 0$. Whenever there exists $a_j(x) \leq -2\varepsilon$, the term $h(a_j(x))$ is large, while the terms $h(a_j(x))$ for $a_j(x) \geq -2\varepsilon$ are either negative or small. Hence $f(x) - g(x) > 0$ also in the case of having j with $a_j(x) \leq -2\varepsilon$. By construction, $g(x) \in \mathcal{M}(a)$. □

The following theorem shows that the relaxations in Section 3.6 can be applied for bounded feasible sets $\{a \geq 0\}$ provided that we know a ball containing $\{a \geq 0\}$ and if we add a special polynomial as one of the generators of the quadratic module.

Theorem 3.15. *Let $a := (a_1, \dots, a_s) \in \mathbb{R}[X]^s$. Let $\{a \geq 0\}$ be bounded and let $\rho \in \mathbb{R}_{>0}$ be such that $\rho - \|X\|^2$ is strictly positive on $\{a \geq 0\}$. Then every polynomial $f \in \mathbb{R}[X]$ which is strictly positive on $\{a \geq 0\}$ belongs to $\mathcal{M}(a, \rho - \|X\|^2)$.*

Proof. Fix linear polynomials $l_1, \ldots, l_r$ such that $\{l_1 \geq 0, \ldots, l_r \geq 0\}$ is non-empty and bounded (for example, it can be $[0,1]^n$). By Lemma 3.13, $t + l_1, \ldots, t + l_r$ belong to the preordering generated by $\rho - \|X\|^2$ if t is sufficiently large. The set $B := \{t + l_1 \geq 0, \ldots, t + l_r \geq 0\}$ is compact (this is easy to see for concrete choices of $l_1, \ldots, l_r$ and can be derived by observing that the recession cones of $\{l_1 \geq 0, \ldots, l_r \geq 0\}$ and B are the same, in general). By Lemma 3.14, there exists $g \in \mathcal{M}(a)$, such that $f - g$ is strictly positive on B. By Handelman, $f - g$ is in the semiring generated by $t + l_1, \ldots, t + l_r$. But every element of the semiring is in the preordering $\mathcal{P}(\rho - \|X\|^2)$, and, since it is generated by a single element, also in the quadratic module $\mathcal{M}(\rho - \|X\|^2)$. Thus, we get the assertion. □

Having derived the latter theorem, we may still be a little dissatisfied, because the aesthetics of the assertion is disturbed by the additional special polynomial $\rho - \|X\|^2$. So, we may wonder whether this polynomial is really necessary. In the case of linear polynomials $a_1, \ldots, a_s$, we'll get a nice assertion without $\rho - \|X\|^2$, but then we cannot always get rid of $\rho - \|X\|^2$.

Theorem 3.16 (Jacobi & Prestel)**.** *Let $a_1, \ldots, a_s$ be polynomials of degree one, such that $S := \{a_1 \geq 0, \ldots, a_s \geq 0\}$ is a bounded non-empty polyhedron. Let $f \in \mathbb{R}[X]$ be strictly positive on S. Then $f \in \mathcal{M}(a_1, \ldots, a_k)$.*

Proof. By rescaling, without loss of generality, we can assume $S \subseteq [-1,1]^n$. By Theorem 3.15, $f \in \mathcal{M}(a_1, \ldots, a_k, n + 1 - \|X\|^2)$. We can write $n + 1 - \|X\|^2$ as

$$n + 1 - \|X\|^2 = 1 + \frac{1}{2}\sum_{i=1}^{d}((1 + X_i)^2(1 - X_i) + (1 - X_i)^2(1 + X_i)),$$

which shows that $n+1-\|X\|^2 \in \mathcal{M}(1-X_1, \ldots, 1-X_n, 1+X_1, \ldots, 1+X_n)$. By the affine Farkas lemma, all $1 \pm X_i$ belong to the cone generated by $1, a_1, \ldots, a_s \in \mathbb{R}[X]$. Thus, we have shown that $f \in \mathcal{M}(a_1, \ldots, a_s)$. □

3.8 Schmüdgen: Positivstellensatz on a general compact semialgebraic set

Schmüdgen derived a positivstellensatz without any additional polynomials $\rho - \|X\|^2$ based on the preordering $\mathcal{P}(a)$ rather than the quadratic module $\mathcal{M}(a)$. The approach is to use positivstellensätze with denominator for $\rho - \|X\|^2$ in combination with a trick to remove denominators (Wörmann).

Lemma 3.17 (Wörmann's trick)**.** *Let $h \in \mathbb{R}[X]$ and $\rho \in \mathbb{R}_{>0}$. Then there exists a $\rho' \in \mathbb{R}_{>0}$ such that $\rho' - \|X\|^2 \in \mathcal{M}(h, (1+h)(\rho - \|X\|^2))$*

Proof. By Lemma 3.13 there exists $t \in \mathbb{R}_{>0}$ such that $t - h$ is in the preordering generated by $\rho - \|X\|^2$. Let's choose $\rho' := \rho(1+t/2)^2$ (so, this is a larger value than ρ. We'll need to see that $\rho(1 + t/2)^2 - \|X\|^2$ is in the desired quadratic module. Since $t - h$ is in the preordering generated by $\rho - \|X\|^2$, the product $(1 + h)(t - h)$ is in $\mathcal{M}(h, (1 + h)(\rho - \|X\|^2))$. Then the whole trick is to find a right expression (because there is a lot of freedom here), which involves $(1 + h)(t - h)$ and other 'allowed terms'. The expression is

$$\rho(1 + t/2)^2 - \|X\|^2 = (1 + h)(\rho - \|X\|^2) + h\|X\|^2 + \rho(1 + h)(t - h) + \rho(t/2 - h)^2.$$

Let's check that the representation is correct. The coefficient at $\|X\|^2$ is correct. So, let us just take this part away (basically, set $\|X\|^2 = 0$). With this change, we'll see that ρ is a linear factor on both sides. So, we can cancel ρ (basically, set $\rho = 1$). Having done that, we end up with the equality $1 + h + (1 + h)(t - h) + (t/2 - h)^2 = (1 + t/2)^2$. The right hand side doesn't involve h, while the left hand side does. The left hand side is $1 + h + t + ht - h - h^2 + t^2/4 - th + h^2 = 1 + t + t^2/4 = (1 + t/2)^2$. Done! □

Theorem 3.18 (Schmüdgen's positivstellensatz)**.** *Let $a = (a_1, \dots, a_s) \in \mathbb{R}[X]^s$. Let $\{a \ge 0\}$ be non-empty and compact and let $f \in \mathbb{R}[X]$ be strictly positive on $\{a \ge 0\}$. Then $f \in \mathcal{P}(a)$.*

Proof. If $\rho \in \mathbb{R}_{>0}$ is large enough, $\rho - \|X\|^2$ is strictly positive on $\{a \ge 0\}$. By Positivstellensatz with denominators (see Theorem 2.21(b)), $\rho - \|X\|^2 = (1+g)/(1+h)$, where g, h are in the preordering generated by a (the possibility to choose $1 + h$ rather than $1 + h$ is formulated as an exercise below). The latter means that $(1 + h)(\rho - \|X\|^2)$ is in the preordering generated by a. Then, by (3.17), there exists $\rho' \in \mathbb{R}_{>0}$ such that $\rho' - \|X\|^2$ is in the preordering generated by a. By Theorem 3.15, f is in the quadratic module generated by $\rho' - \|X\|^2$ and a. This gives the assertion. □

Exercise 3.19. *Show that if a polynomial f is positive on $\{a \ge 0\}$, then $(1 + h)f = 1 + g$ holds for some $g, h \in \mathcal{P}(a)$. The difference to the Positivstellensatz, we formulated before is that we use $1 + h$ rather than h. (Hint: use the polynomial version of Farkas lemma; if -1 is in $\mathcal{P}(a)$, then every polynomial is in $\mathcal{P}(a)$).*

3.9 Putinar: cases where quadratic modules certify positivity

The following theorem tells us when we can use $\mathcal{M}(a)$ without any additional generators.

Theorem 3.20 (Putinar)**.** *Let $a = (a_1, \dots, a_s) \in \mathbb{R}[X]^s$. If there exists $g \in \mathcal{M}(a)$ such that $\{g \ge 0\}$ is bounded, then every polynomial $f \in \mathbb{R}[X]$ strictly positive on $\{a \ge 0\}$ belongs to $\mathcal{M}(a)$.*

Proof. By Lemma 3.14, there exists $h \in \mathcal{M}(a)$ with $f - h > 0$ on $\{g \ge 0\}$. By Theorem 3.18, $f-h$ is in the preordering generated by g. This gives the assertion. □

Here is an example presenting a set, for which $\{g \ge 0\}$ is unbounded for every $g \in \mathcal{M}(a)$, though $\{a \ge 0\}$ is bounded.

Example 3.21 (Jacobi-Prestel counterexample)**.** *Consider the compact basic closed semialgebraic set $K := \{X_1 - 1/2 \ge 0, X_2 - 1/2 \ge 0, 1 - X_1X_2 \ge 0\}$. For this, region it is known that $\rho - \|X\|^2 \notin \mathcal{M}(X_1 - 1/2, X_2 - 1/2, 1 - X_1X_2)$ for every $\rho \in \mathbb{R}$. In particular, in view of Putinar's theorem, this shows that $f \notin \mathcal{M}(X_1 - 1/2, X_2 - 1/2, 1 - X_1X_2)$ for all $f \in \mathbb{R}[X]$ strictly positive on K with the property that $\{f \ge 0\}$ is bounded. In terms of optimization, quadratic-module relaxation of polynomial optimization under constraints $X_1 - 1/2 \ge 0, X_2 - 1/2 \ge 0, 1 - X_1X_2 \ge 0$ almost never gives any interesting bounds.*

4 Duality for conic and semidefinite optimization

This chapter is a preparation to describing interior-point methods for SDP. Interior-point methods for SDP rely on SDP duality. Since SDP duality is a special case of conic duality, we start with the conic duality and then specialize it to the SDP case.

The chapter was motivated by various literature sources including [ML12] and [GM12, Chapter 4]. When discussing general conic programming duality, we'll work in the space $\mathbb{R}^n$ with $n \in \mathbb{N}$. We'll use the notation

$$H^{\leq}(u,\beta) := \{x \in \mathbb{R}^n \,:\, \langle u, x\rangle \leq \beta\}$$

for $u \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$. If $u \neq 0$, such $H^{\leq}(u,\beta)$ is a closed half-space. An analogous notation can also be used for the relations $=, \geq, <, >$, giving also notation for hyperplanes and open half-spaces.

4.1 Every nonlinear problem is a conic problem (in principle)

A conic problem is a problem of optimizing a linear function $\langle c, x\rangle$ under constraints of the form $b_i - A_i x \in L_i$ for all $i \in [s]$, where $b_1, \ldots, b_s$ are vectors, $A_1, \ldots, A_s$ matrices and $L_1, \ldots, L_s$ are closed convex cones. If L is a closed convex cone, the constraint $x \in L$ is a kind of 'generalized non-negativity' and is a special case of the constraint $b - Ax \in L$ with $b = 0$ and $A = -I$.

It is clear that LP in any of its forms, say, in the form of optimizing $\langle c, x\rangle$ subject to $Ax = b$ and $x \geq 0$ is a conic problem. For this form one can use the cones $\{0\}$ and $\mathbb{R}^n_+$ and write the constraints as $b - Ax \in \{0\}$ and $x \in \mathbb{R}^n_+$. So, if we use the non-negative orthant in the definition of conic problem, we end up with a linear problem.

Are there any other interesting cones around? Yes, there are a few other cones which are good in terms of efficiency and are useful for modeling various situations. Consider for example the facility location problem of minimizing $f(x) := \sum_{i=1}^s \|x - p_i\|$, which is the total sum of distances to given facilities $p_1, \ldots, p_s \in \mathbb{R}^n$. We can write the problem as the problem of minimizing $\sum_{i=1}^s y_j$ for with $y_1, \ldots, y_s \in \mathbb{R}$, $x \in \mathbb{R}^n$ and $y_i \leq \|x - p_i\|$ for all i. If we define the second order cone $\mathrm{SOC}_n := \{(x,y) \,:\, y \geq \|x\|\}$, then our problem can be written using the conic constraints $(x - p_i, y_i) \in \mathrm{SOC}_n$ for $i \in [s]$.

A very interesting case of conic programming we need in this course is semidefinite programming, arising from the cones $\mathcal{S}^k_+$ of psd matrices.

Conic programming is a special case of convex programming. Interestingly, from the perspective of conic programming, we can pretend as every non-linear problem were convex. Consider, for example, the problem $\inf_{x \in \mathbb{R}^n} f(x)$ of constrained optimization of a polynomial $f \in \mathbb{R}[x]$ in n variables of degree at most d. Let $K_{n,d}$ be the closed convex cone of all non-negative polynomials of degree at most d in n variables. Then our problem can be rewritten as $\sup\{y \in \mathbb{R} \,:\, f - y \in K_{n,d}\}$. This is a conic problem with the decision variable y and the conic constraint $f - y \in K_{n,d}$ for this unknown. Of course, there is a catch here: our problem did not get simpler by transformation into the conic form. The problem remains as difficult as it was. The complexity of the problem is now hidden in the cone $K_{n,d}$, which is a complicated object (checking if a given polynomial belongs to $K_{n,d}$ is hard). Nevertheless, such conic reformulations *can* help computationally. The approach is to express problems using cones and then try to understand the structure of the underlying cones and exploit it in computations.

4.2 Duality operations for convex sets

With each set $X \subseteq \mathbb{R}^n$ we associate the *polar set*

$$X^\circ = \{y \in \mathbb{R}^n \,:\, \langle y, x\rangle \leq 1 \;\forall x \in X\}$$

and the *dual cone*

$$X^* = \{y \in \mathbb{R}^n \,:\, \langle y, x\rangle \geq 0 \;\forall x \in X\}$$

In what follows we will need X^*, but it is not a bad idea to see X° at least once. The set X° is a closed, convex set containing 0 and X^* is a closed convex cone. The following is a main duality statement for the above duality operations.

Proposition 4.1. *For every non-empty set $X \subseteq \mathbb{R}^n$, one has*

$$\begin{aligned}(X^\circ)^\circ &= \mathrm{cl}(\mathrm{conv}(X \cup \{0\})),\\ (X^*)^* &= \mathrm{cl}(\mathrm{cone}(X)).\end{aligned}$$

Proof. (a): We first 'spell out' what $(X^\circ)^\circ$ actually is:

$$\begin{aligned}& x' \in (X^\circ)^\circ\\ \Leftrightarrow\qquad & \big\langle x', y\big\rangle \leq 1 \text{ for all } y \in X^\circ\\ \Leftrightarrow\qquad & \big\langle x', y\big\rangle \leq 1 \text{ for all } y \text{ with } \langle y, x\rangle \leq 1 \forall x \in X.\end{aligned}$$

This shows that $(X^\circ)^\circ$ is the intersection of all half-spaces of the form $H^{\leq}_{y,1}$ containing the set X. Such subspaces definitely contain 0. So, we see that $X \cup \{0\} \subseteq (X^\circ)^\circ$ holds. Since polar sets are always closed convex sets, we even get $\mathrm{cl}(\mathrm{conv}(X \cup \{0\}) \subseteq (X^\circ)^\circ$, which is the 'easy' inclusion. The converse inclusion relies on separation theorems from the theory of convex sets.

If a point x' does not belong to $\mathrm{cl}(\mathrm{conv}(X \cup \{0\}))$, there exists a hyperplane separation of this point from our set in the following sense: there exists a vector u and values α and β with $\langle u, x'\rangle \geq \alpha > \beta \geq \langle u, x\rangle$ for all $x \in X \cup \{0\}$. In particular one has $\alpha > \beta \geq 0$. For the vector $v := 2u/(\alpha+\beta)$ we have $\langle v, x'\rangle > 1$ and $\langle v, x\rangle \leq 1$ for all $x \in X \cup \{0\}$. This shows, $x' \notin H^{\leq}_{v,1}$ and $X \cup \{0\} \subseteq H^{\leq}_{v,1}$. Thus, $x' \notin (X^\circ)^\circ$.

(b): The proof is analogous to (a). We 'spell out' what $(X^*)^*$ means:

$$\begin{aligned}& x' \in (X^*)^* & &\\ \Leftrightarrow\qquad & \big\langle x', y\big\rangle \geq 0 & & \text{für alle } y \in X^*\\ \Leftrightarrow\qquad & \big\langle x', y\big\rangle \geq 0 & & \text{für alle } y \text{ mit } \;\langle y, x\rangle \geq 0 \;\forall x \in X\end{aligned}$$

So $(X^*)^*$ is the intersection of all half-spaces $H^{\geq}_{y,0}$ given by a vector $y \neq 0$ that contain X as a subset. Following the same steps as in (a), this shows $\mathrm{cl}(\mathrm{cone}(X)) \subseteq (X^*)^*$. Conversely: if a point x' is not in $\mathrm{cl}(\mathrm{conv}(X))$, then by separation theorems for convex cones there exists a vector u with $\langle u, x'\rangle < 0$ and $\langle u, x\rangle \geq 0$ for all $x \in \mathrm{cl}(\mathrm{conv}(X))$. That is, x' is not in $H^{\geq}_{u,0}$, while $H^{\geq}_{u,0}$ contains X as a subset. This gives $x' \notin (X^*)^*$. □

4.3 Farkas lemmas in conic optimization

Farkas lemmas in linear programming involve conditions like $Ax \le b$ or $x \ge 0$. Recall that Farkas lemmas provide a certificate for infeasiblity. The template for Farkas lemmas is: a system of conditions is infeasible if and only if there is some data (certificate) that can be used to make the infeasiblity algebraically apparent. For conic optimization, we deal with conic constraints of the forms $b - Ax \in K$ and $x \in K$, where K is an arbitrary closed cone (choosing K to be a non-negative orthant we get back the linear programming duality). The duality of linear programming is perfect in the sense that it gives a complete characterization of infeasiblity. For conic-programming duality the situation is quite good, too, but not really perfect. There is a necessary and a sufficient condition for infeasiblity that look very similar but do not quite match.

In one step of the proof we use the following fact

Exercise 4.2. *Show the following. If $A \in \mathbb{R}^{m \times n}$, where $m, n \in \mathbb{N}$, then the image* $\operatorname{im}(A)$ *of A and the kernel* $\ker(A^\top)$ *are related by* $\operatorname{im}(A) = \ker(A^\top)^\perp$.

The formula $\operatorname{im}(A) = \ker(A^\top)^\perp$ is one possible way to express the duality of linear algebra.

Theorem 4.3 (Farkas-lemma for the system $x \in K, Ax = b$)**.** *Let $K \subseteq \mathbb{R}^n$ be a closed convex cone, let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Then implications (1) ⇒ (2) ⇒ (3) hold for conditions (1), (2) and (3) described as follows:*

(1) There exists an $y \in \mathbb{R}^m$ with $A^\top y \in K^*$ and $\langle y, b\rangle < 0$.

(2) There exists no $x \in K$ with $Ax = b$.

(3) There exists $y \in \mathbb{R}^m$ with $A^\top y \in K^*$ and $\langle y, b\rangle \le 0$, where $A^\top y$ or $\langle y, b\rangle$ is not 0.

In other words, (1) is a sufficient condition for infeasiblity of the system $x \in K, Ax = b$, while (3) is a necessary one.

Proof. We introduce the notation $X := \{x \in \mathbb{R}^n : Ax = b\}$.

(1) ⇒ (2): Assume to the contrary, that there exists an $x \in K \cap X$. Then $\langle y, b\rangle = \langle y, Ax\rangle = \langle A^\top y, x\rangle \ge 0$, which contradicts $\langle y, b\rangle < 0$.

(2) ⇒ (3): In the degenerate case $X = \emptyset$ we rely on linear algebra. One has $b \notin \operatorname{im}(A)$. Because of $\operatorname{im}(A) = \ker(A^\top)^\perp$ there exists y with $A^\top y = 0$ and $\langle y, b\rangle \ne 0$. Possibly, replacing y by $-y$, we can assume $\langle y, b\rangle < 0$. Thus, y is the desired vector.

In the case $X \ne \emptyset$, we use separation theorems for the sets K and X satisfying $K \cap X = \emptyset$. There exist a vector $u \in \mathbb{R}^n \setminus \{0\}$ and a scalar α with $\langle u, x\rangle \ge \alpha$ for all $x \in K$ and $\langle u, x\rangle \le \alpha$ for all $x \in X$.

Since K is a cone, the inequality $\langle u, x\rangle \ge \alpha$ for all $x \in K$ implies $\langle u, x\rangle \ge 0$ for all $x \in K$. That is $u \in K^*$. For $x = 0$ the inequality $\langle u, x\rangle \ge \alpha$ yields $0 \ge \alpha$. Thus, $\langle u, x\rangle \le \alpha \le 0$ holds for all $x \in X$. The function $x \mapsto \langle u, x\rangle$ is a bounded affine function on the affine space X. So, the function is constant on X. Then $\langle u, x'\rangle = \langle u, x''\rangle$ for all $x', x'' \in X$. Hence $\langle u, x\rangle = 0$ for $X - X = \ker(A)$. Hence $u \in \ker(A)^\perp$.

Since $\ker(A)^\perp = \operatorname{im}(A^\top)$, one has $u = A^\top y$ for some y. Thus, $\langle u, x\rangle = \langle A^\top y, x\rangle = \langle y, Ax\rangle = \langle y, b\rangle \le \alpha \le 0$ for all $x \in X$. The vector $u = A^\top y$ is not 0. □

4.4 Duality of conic optimization problems

We'll extend the LP-duality corresponding to the case $K = \mathbb{R}^n_+$ to general closed cones K. The weak duality extends easily, for the extension of the strong duality we need a kind of assumption that excludes degenerate situations.

Theorem 4.4. *Let $K \subseteq \mathbb{R}^n$ be closed convex cone, let $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$ and consider the following conic problems*

$$\alpha = \inf \{\langle c, x\rangle \,:\, x \in K,\ Ax - b = 0\} \tag{4.1}$$

$$\beta = \sup \left\{\langle y, b\rangle \,:\, y \in \mathbb{R}^m,\ c - A^\top y \in K^*\right\}. \tag{4.2}$$

Then the following hold:

(a) If both (4.1) and (4.2) have feasible solutions, then $\alpha \geq \beta$.

(b) If (4.1) is unbounded ($\alpha = -\infty$), then (4.2) is infeasible ($\beta = -\infty$)

(c) If (4.2) is unbounded ($\beta = +\infty$), then (4.1) is infeasible ($\alpha = +\infty$).

(d) If there exists $x' \in \operatorname{int}(K)$ with $Ax' - b = 0$, then one has equality $\alpha = \beta$.

Proof. (a): For every feasible solution x of (4.1) and every feasible solution y of (4.2) one has $\langle Ax - b, y\rangle = 0$ because $Ax - b = 0$. On the other hand, one has $\left\langle c - A^\top y, x\right\rangle \geq 0$ by $x \in K$ and $c - A^\top y \in K^*$. Thus, we get

$$\langle c, x\rangle \geq \left\langle A^\top y, x\right\rangle = \langle Ax, y\rangle = \langle b, y\rangle .$$

(b) and (c) follow from (a).

(d): The existence of x' implies $\alpha < \infty$. In the degenerate case $\alpha = -\infty$, the equality follows from (b). We consider the case $-\infty < \alpha < \infty$. Let $X := \{x \in \mathbb{R}^n \,:\, Ax = b\}$.

Fix an arbitrary $\varepsilon > 0$. By the definition of α there exists no x with $x \in K$, $\langle c, x\rangle = \alpha - \varepsilon$ and $Ax = b$. Application of the implication (2) ⇒ (3) of Theorem 4.3 (Farkas-Lemma) to the system

$$x \in K, \qquad \begin{pmatrix} -A \\ c^\top \end{pmatrix} x = \begin{pmatrix} -b \\ \alpha - \varepsilon \end{pmatrix}$$

yields the existence of $y \in \mathbb{R}^m$ and $\mu \in \mathbb{R}$ with $\mu c - A^\top y \in K^*$ and $\mu(\alpha - \varepsilon) - \langle y, b\rangle \leq 0$, where $\mu c - A^\top y$ or $\mu(\alpha - \varepsilon) - \langle y, b\rangle$ is not 0. We'll produce a feasible solution of the dual problem out of μ and y. Since the dual solution will have to be linked to the primal one, we derive the following inequalities:

$$\mu \left\langle c, x'\right\rangle - \langle y, b\rangle = \mu \left\langle c, x'\right\rangle - \left\langle y, Ax'\right\rangle = \left\langle \underbrace{\mu c - A^\top y}_{\in K^*}, \underbrace{x'}_{\in K} \right\rangle \geq 0, \tag{4.3}$$

So, we get

$$\mu(\alpha - \varepsilon) \leq \langle y, b\rangle \leq \mu \left\langle c, x'\right\rangle \tag{4.4}$$

By the implication (2) ⇒ (3) of Theorem 4.3 one has $\mu c - A^\top y \neq 0$ or $\mu(\alpha - \varepsilon) - \langle y, b\rangle < 0$. In the former case, the inequality in (4.3) is strict, because x' lies in

$\operatorname{int}(K)$. In the latter case the inequality $\mu(\alpha - \varepsilon) < \langle y, b\rangle$ is strict. This yields that at least one of the two inequalities in (4.4) is strict. We get $\mu(\alpha - \varepsilon) < \mu \langle c, x'\rangle$.

Taking into account that $\langle c, x'\rangle > \alpha - \varepsilon$ holds for all $x \in K \cap X$, we get $\mu > 0$. Now, wlog we can assume $\mu = 1$, because we can rescale y to y/μ and μ to $\mu/\mu = 1$. After this change we get $c - A^\top y \in K^*$ and $\langle y, b\rangle \geq \alpha - \varepsilon$. So, for an arbitrary $\varepsilon > 0$ we have found a feasible solution of (4.2), which satisfies $\langle y, b\rangle \geq \alpha - \varepsilon$. Thus, $\beta \geq \alpha - \varepsilon$ for every $\varepsilon > 0$, and we get the inequality $\beta \geq \alpha$. □

The conic problems (4.1) and (4.2) are called *dual* to each other. The assertions (a)-(c) are called *weak duality*, and the assertion (d) is called *strong duality*.

Exercise 4.5. *Show that, in the notation of Theorem 4.4, the following hold:*

(a) If x and y are feasible solutions of (4.1) *and* (4.2)*, respectively and $\left\langle c - A^\top y, x\right\rangle = 0$ holds, then x and y are optimal for* (4.1) *and* (4.2)*, respectively.*

(b) If $\alpha = \beta$ and (4.1) *and* (4.2) *have optimal solutions x^* and y^*, respectively, then $\left\langle c - A^\top y^*, x^*\right\rangle = 0$ holds.*

Solution. To get both assertions, just look at the derivation of the weak duality

$$\langle b, y\rangle = \langle Ax, y\rangle = \left\langle x, A^\top y\right\rangle \leq \langle x, c\rangle .$$

□

4.5 Duality for more general formulations

In conic optimization, similarly to linear programming, we can formulate duality for different or more general formulations.

We'll need the following simple observation.

Exercise 4.6. *Let $n_1, \ldots, n_t \in \mathbb{N}$ and $t \in \mathbb{N}$. Consider, for every $i \in [t]$, the set $X_i \subseteq \mathbb{R}^{n_i}$ with $0 \in X_i$. Then*

$$(X_1 \times \ldots \times X_t)^* = X_1^* \times \ldots \times X_t^*.$$

Solution. $(y_1, \ldots, y_t) \in (X_1 \times \ldots \times X_t)^*$ means $\langle (y_1, \ldots, y_t), (x_1, \ldots, x_t)\rangle \geq 0$ for all $(x_1, \ldots, x_t) \in X_1 \times \ldots \times X_t$. Plugging in 0 for all x_i but one, we get $\langle y_i, x_i\rangle \geq 0$ for all $i \in [t]$ and $x_i \in X_i$. This shows $y_i \in X_i^*$ for all $i \in [k]$. The converse inclusion $X_1^* \times \ldots \times X_t^* \subseteq (X_1 \times \ldots \times X_t)^*$ also follows directly: a vector $(y_1, \ldots, y_t)$ belonging to the left hand side satisfies $\langle y_i, x_i\rangle \geq 0$ for all $i \in [t]$ and all $x_i \in X_i$. This yields $(y_1, \ldots, y_t) \in (X_1 \times \ldots \times X_t)^*$. □

Exercise 4.7. *Show the following. Let $m, n \in \mathbb{N}$, let $K \subseteq \mathbb{R}^n$ and $L \subseteq \mathbb{R}^m$ be closed convex cones and let $A \in \mathbb{R}^{m\times n}, c \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$. Consider the problems*

$$\alpha := \inf \{\langle c, x\rangle \,:\, x \in K,\ Ax - b \in L\} \tag{4.5}$$

and

$$\beta := \sup \left\{\langle y, b\rangle \,:\, y \in L^*,\ c - A^\top y \in K^*\right\}. \tag{4.6}$$

Then the following hold:

(a) The problems satisfy weak duality.

(b) If there exists an x' with $x' \in \operatorname{int}(K)$ and $Ax - b \in \operatorname{int}(L)$, then strong duality holds.

(c) If there exists an y' with $y' \in \operatorname{int}(L^*)$ and $c - A^\top y \in \operatorname{int}(K^*)$, then strong duality holds.

Solution. (a): We reformulate the minimization problem as follows:

$$\begin{aligned}\alpha &= \inf \{\langle c, x\rangle \,:\, (x,u) \in K \times L,\ Ax - u = b\} \\ &= \inf \left\{ \left\langle \begin{pmatrix} c \\ 0 \end{pmatrix}, \begin{pmatrix} x \\ u \end{pmatrix} \right\rangle \,:\, (x,u) \in K \times L,\ \begin{pmatrix} A & -I \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} = b \right\}\end{aligned}$$

Application of Theorem 4.4 yields the dual problem

$$\sup \left\{ \langle y, b\rangle \,:\, \begin{pmatrix} c \\ 0 \end{pmatrix} - \begin{pmatrix} A^\top \\ -I \end{pmatrix} y \in (K \times L)^* \right\}.$$

Application of Exercise 4.6 yields $(K \times L)^* = K^* \times L^*$. So, the latter problem can be reformulated as (4.6).

(b): From the proof of (a) and in view of Theorem 4.4, one can see that the strong duality holds whenever there exists $(x', u') \in \operatorname{int}(K \times L) = \operatorname{int}(K) \times \operatorname{int}(L)$ with $Ax' - u' = b$.

(c): We reformulate (4.6) as a minimization problem:

$$-\beta = \inf \left\{ \langle y, -b\rangle \,:\, y \in L^*,\ (-A^\top)y - (-c) \in K^* \right\}.$$

Application of (b) to this minimization problem yields

$$-\beta = \sup \left\{ \langle x, -c\rangle \,:\, x \in (K^*)^*,\ (-b) - ((-A)^\top)^\top x \in (L^*)^* \right\}.$$

In view of $(K^*)^* = K$ and $(L^*)^* = L$ we get

$$-\beta = \sup \{- \langle c, x\rangle \,:\, x \in K,\ Ax - b \in L\} = -\alpha.$$

So, $\alpha = \beta$. □

Remark 4.8. Now, having deduced the more general form of duality, we can 'update' Theorem 4.4 that we had formulated before. In fact, there exists another case of strong duality for problems discussed in Theorem 4.4, namely, when there exists $y' \in \mathbb{R}^m$ with $c - A^\top y' \in \operatorname{int}(K^*)$. Observe, that the whole space $\mathbb{R}^m$ is a convex cone whose dual cone is $(\mathbb{R}^m)^* = \{0\}$. Thus, what we do is just use the corollary for $L = \mathbb{R}^n$.

We conclude this section by formulating a very general form of conic duality with an arbitrary number of conic constraints.

Exercise 4.9. *Show the following. Let $t, s \in \mathbb{N}$ and consider the problems*

$$\alpha = \inf \left\{ \sum_{i=1}^t \langle c_i, x_i\rangle \,:\, x_i \in K_i \ \forall i \in [k],\ \sum_{j=1}^t A_{i,j} x_j - b_i \in L_i \ \forall i \in [s] \right\}$$

and

$$\beta = \sup \left\{ \sum_{i=1}^{s} \langle b_i, y_i \rangle \,:\, y_i \in L_i^* \ \forall i \in [s],\ c_i - \sum_{j=1}^{s} A_{j,i}^\top y_j \in K_i^* \ \forall i \in [t] \right\}$$

based on closed convex cones $K_1, \ldots, K_t, L_1, \ldots, L_s$, Matrices $A_{i,j}$ with $i \in [s], j \in [t]$ and vectors $c_1, \ldots, c_t, b_1, \ldots, b_s$. Then:

(a) The weak duality holds for this pair of problems.

(b) If the minimization problem has solution $(x_1', \ldots, x_t')$ with $x_i' \in \operatorname{int}(K_i)$ for all $i \in [t]$ and $\sum_{j=1}^t A_{i,j} x_j' - b_i \in \operatorname{int}(L_i)$ for all $i \in [s]$, then the strong duality holds.

(c) If the maximization problem has a solution $(y_1', \ldots, y_s')$ with $y_i \in \operatorname{int}(L_i^*)$ for all $i \in [s]$ and $c_i - \sum_{j=1}^s A_{j,i}^\top y_j \in \operatorname{int}(K_i^*)$, then the strong duality holds.

Solution. Let $x = (x_1, \ldots, x_k)$. Let $c = (c_1, \ldots, c_t)$, $b = (b_1, \ldots, b_t)$ and let A be the matrix with blocks $A_{i,j}$. Then our problems can be formulated concisely as

$$\inf \{ \langle c, x \rangle \,:\, x \in K_1 \times \ldots \times K_t,\ Ax - b \in L_1 \times \ldots \times L_s \}$$

and

$$\beta = \sup \left\{ \langle b, y \rangle \,:\, y \in L_1^* \times \ldots \times L_t^*,\ c - A^\top y \in K_1^* \times \ldots \times K_t^* \right\}$$

The assertions now follow from Exercise 4.9 and Exercise 4.6. □

4.6 Semidefinite optimization

Semidefinite optimization was introduced in Section 1.4. Here, we are able to describe semidefinite problems concisely as a special conic problem. We work with linear spaces $\mathcal{S}^k$ and with the cones $\mathcal{S}_+^k$ in these spaces. For being able to write systems of linear inequalities concisely it is nice to have a scalar product fixed in $\mathcal{S}^k$. The space $\mathcal{S}^k$ is a subspace of $\mathbb{R}^{k \times k}$ (all $k \times k$ matrices, not necessarily symmetric), and in $\mathcal{S}^k$ one can borrow a scalar product from $\mathbb{R}^{k \times k}$. For $\mathbb{R}^{k \times k}$, a very natural scalar product is the standard one: we can think of $\mathbb{R}^{k \times k}$ as a vector with k^2 components and define the scalar product as the standard scalar product of such vectors. It turns out that this scalar product can be expressed nicely using the trace operation: It is straightforward to check that for $A = (a_{ij}), B = (b_{ij}) \in \mathbb{R}^{k \times k}$, one has

$$\langle A, B \rangle := \sum_{i,j} a_{ij} b_{ij} = \operatorname{tr}(AB^\top) = \operatorname{tr}(A^\top B).$$

Of course, if our matrices are symmetric we need not transpose so that we get the formula $\langle A, B \rangle = \operatorname{tr}(AB)$ for $A, B \in \mathcal{S}^k$.

Geometrically, SDP can be formulated as an optimization of a linear function over an affine slice of $\mathcal{S}_+^k$. That is, one optimizes a linear objective over $\mathcal{S}_+^k \cap H$, where $H \subseteq \mathcal{S}^k$ is an affine space. There are two ways to represent $H \cap \mathcal{S}_+^k$ algebraically: one can parameterize H or one can describe H by a system of linear inequalities. As a consequence, there are the following two basic formulations of SDP.

If we use equalities to describe H, we get a formulation

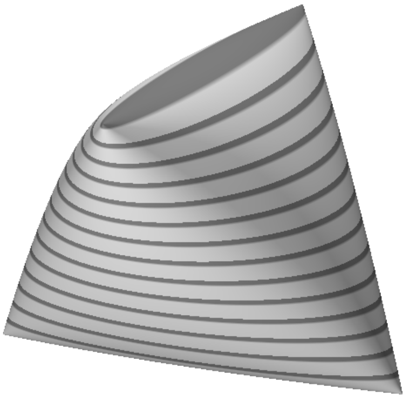
$$\inf \left\{ \langle C, X \rangle \,:\, X \in \mathcal{S}_+^k,\ \langle A_i, X \rangle = b_i \ \forall i \in [m] \right\},$$

where $C, A_1, \ldots, A_m \in \mathcal{S}^k$ and $b_1, \ldots, b_m \in \mathbb{R}$. If we parameterize H we end up with

$$\inf \left\{ \langle c, x \rangle \,:\, x \in \mathbb{R}^n, A(x) \in \mathcal{S}_+^k \right\},$$

with $c \in \mathbb{R}^n$, where $A : \mathbb{R}^n \to \mathcal{S}^k$ is affine map, meaning $A(x_1, \ldots, x_n) = A_0 + A_1 x_1 + \cdots + A_n x_n$ for some $A_0, \ldots, A_n \in \mathcal{S}^k$. The condition $A(x) \in \mathcal{S}_+^k$ is a *linear matrix inequality* for x, which we have already seen in Section 1.4 (abbreviated as LMI). The expression $A(x)$ can be viewed as an affine function in x with matrix coefficients or as symmetric matrices whose entries are affine functions in x.

LMIs are generalizations of linear inequalities (if the off-diagonal entries of $A(x)$ are identically equal to zero, we merely have a system of k linear inequalities). Here is a picture illustrating the geometry of sets described by LMIs (such sets are called spectrahedra):



$$\begin{pmatrix} 1 & x_1 & x_2 \\ x_1 & 1 & x_3 \\ x_2 & x_3 & 1 \end{pmatrix} \text{ psd}$$

The first thing we wish to do is discussing duality for SDP. We would like to use the results from the previous sections. While in the previous sections we worked with the duality with respect to the standard scalar product in $\mathbb{R}^n$, it is clear that the duality could have been introduced with respect to any scalar product and the results would still be the same. So, we can introduce the dual cone $(\mathcal{S}_+^k)^*$ in the space $\mathcal{S}^k$ just by the same formula that we used to introduce dual cones in $\mathbb{R}^n$.

It turns out that $\mathcal{S}_+^k$ is self-dual.

Exercise 4.10. *For every $k \in \mathbb{N}$ one has $(\mathcal{S}_+^k)^* = \mathcal{S}_+^k$ (the cone $\mathcal{S}_+^k$ is self-dual).*

Solution. For $A \in \mathcal{S}^k$ the equality says that the condition $A \in \mathcal{S}_+^k$ is equivalent to $\langle A, B \rangle \geq 0 \; \forall B \in \mathcal{S}_+^k$. We check this equivalence.

If $A \in \mathcal{S}_+^k$, then $\langle Ax, x \rangle \geq 0$ holds for all x. The latter can be turned into the form $\left\langle A, xx^\top \right\rangle \geq 0$. Note that $xx^\top$ is a symmetric (psd) matrix of rank 1 (for $x \neq 0$). The cone $\mathcal{S}_+^k$ is the conic hull of all such matrices (see Exercise 1.4). So, we get $\langle A, B \rangle \geq 0$ for all $B \in \mathcal{S}_+^k$. Conversely: if $\langle A, B \rangle \geq 0$ holds for every $B \in \mathcal{S}_+^k$, then choosing $B = xx^\top$, we get $\left\langle A, xx^\top \right\rangle \geq 0$ for all $x \in \mathbb{R}^n$. This gives $\langle Ax, x \rangle \geq 0$ for all x and shows $A \in \mathcal{S}_+^k$. □

Exercise 4.11. *Formulate the problem of computing the largest eigenvalue of a symmetric matrix as an SDP.*

Solution. For $A \in \mathcal{S}^k$ consider the matrix $tI - A$. It is easy to see that λ is an eigenvalue of A if and only if $t - \lambda$ is an eigenvalue of $tI - A$. A matrix A is psd if and only if all its eigenvalues are non-negative. So, we arrive at the SDP

$$\min \left\{ t \in \mathbb{R} \,:\, tI - A \in \mathcal{S}^k_+ \right\}$$

with one decision variable $t \in \mathbb{R}$. This is an SDP in one unknown $t \in \mathbb{R}$. In the same way, we can see that computation of the spectral norm of a symmetric matrix is again an SDP (in one unknown). □

Remark 4.12. Many researches are very excited about SDP. Why? It has turned out that a large number of problems can be formulated (exactly or approximately) as SDP. In particular, we have seen that polynomial problems can be formulated approximately as SDPs. Apart from that, SDPs arise in control theory, combinatorial optimization, statistics, probability etc. Yet another reasons for excitement is that SDP leads to interesting mathematical problems involving algebra, convexity and algorithms.

4.7 Duality for SDP

Using the previous section and conic duality, we can derive duality for SDPs. For duality in the space $\mathbb{R}^n$ we used the standard scalar product in $\mathbb{R}^n$, and the matrix A (of the left hand side of the underlying system of constraints) in the primal problem has been turned to the transposed matrix $A^\top$ in the dual problem. Now, we work in $\mathcal{S}^k$, which is a different Euclidean space, and so instead of using $A^\top$ we need to use an abstract analogue of it. The abstract transpose is called the adjoint operator. Actually, the adjoint operator is just another way of writing transpose (it's not more general in its essence, it's just a more general way of writing things down, that does not rely too much on the components). Let me shortly recall how this works (this material is typically presented in linear algebra). We've got two finite-dimensional Euclidean spaces V and W over $\mathbb{R}$. The spaces have scalar products $\langle \cdot, \cdot \rangle_V$ and $\langle \cdot, \cdot \rangle_W$ (frequently, one would just omit the subscript, as it is usually clear which of the scalar products is meant).

Consider a linear map $A : V \to W$ and a vector $b \in W$. With this data we can define an 'abstract' linear system $A(x) = b$ in the unknown $x \in V$. An analogue of transpose matrix is the so-called *adjoint operator* or *adjoint map* $A^* : W \to V$, defined as the unique linear map satisfying $\langle A(x), y \rangle_W = \langle x, A^*(y) \rangle_V$.

In this abstract setting the dual pair of conic programs looks as follows

$$\inf \{ \langle c, x \rangle \,:\, x \in K,\ A(x) - b \in L \}$$

and

$$\sup \{ \langle y, b \rangle \,:\, y \in L^*,\ c - A^*(y) \in K^* \}$$

where K is a closed convex cone in V and L a closed convex cone in W.

We just need to specify what we get in the case of SDPs in the above formulas. For example, let's determine the dual of the SDP in the form

$$\inf \left\{ \langle c, x \rangle \,:\, x_1 A_1 + \ldots + x_n A_n - B \in \mathcal{S}^k_+ \right\}$$

We can write it as

$$\inf \left\{ \langle c, x \rangle \,:\, x \in \mathbb{R}^n, A(x) - B \in \mathcal{S}^k_+ \right\}$$

where $A : \mathbb{R}^n \to \mathcal{S}^k$ is the linear map $A(x) = A_1x_1 + \cdots + A_nx_n$. To get the dual we need to determine $A^* : \mathcal{S}^k \to \mathbb{R}^n$. Just use the equality that defines the adjoint map:

$$\langle A(x), Y\rangle = \langle x, A^*(Y)\rangle \qquad \forall x \in \mathbb{R}^n\ \forall Y \in \mathcal{S}^k.$$

This can be spelled out as

$$\langle A_1x_1 + \ldots + A_nx_n, Y\rangle = \langle (x_1, \ldots, x_n), A^*(Y)\rangle \qquad \forall x_1, \ldots, x_n \in \mathbb{R}\ \forall Y \in \mathcal{S}^k.$$

Comparing coefficients for $x_1, \ldots, x_n$ gives the formula

$$A^*(Y) = (\langle A_1, Y\rangle, \ldots, \langle A_n, Y\rangle)$$

for the adjoint map.

Consequently, our dual problem is

$$\sup\left\{\langle B, Y\rangle \,:\, Y \in \mathcal{S}^k_+,\ c_i = \langle A_i, Y\rangle\ \forall i \in [n]\right\}$$

Remark 4.13. If you know the mnemonic rules to get a dual LP, you see that these rules carry over to the case of SDP in some very natural way. The mnemonic rule is about what corresponds to what in the primal-dual pair:

- variables ↔ constraints
- unconstrained variables↔ equality constraints
- non-negative variables ↔ inequality constraints;
- right hand sides ↔ coefficients of the objective functions.

We see that the rule remains valid for SDP. Our primal problem had $x_1, \ldots, x_n$ unconstrained real variables and one SDP constraint. The sdp constraint generates a psd-constrained variable Y in the dual. The n variables $x_1, \ldots, x_n$ generate the n equality constraints in the dual. Note that our primal/dual pair of SDP illustrates the following:

- real variables ↔ LP constraints (= linear (in)equalities)
- matrix variables ↔ SDP constraints (= LMIs).

4.8 SDP duality applied to polynomial optimization

Let's now apply the SDP duality to the SDP problem we derived in Section 1.4. Let me remind what we did. We wanted to solve the global POP

$$\inf\{f(x) \,:\, x \in \mathbb{R}^n\}, \tag{4.7}$$

where $f(X) = \sum_{\alpha \in E^n_{2d}} c_\alpha X^\alpha \in \mathbb{R}[X] = \mathbb{R}[X_1, \ldots, X_n]$. We introduce a formal dual problem $\sup\{y \,:\, f - y \geq 0 \text{ on } \mathbb{R}^n\}$ (the problem to find the lower bounds) and then replaced this problem by the simpler problem

$$\sup\{y \in \mathbb{R} \,:\, f - y \quad \text{sos}\},$$

which turned out to be the SDP

$$\sup\Big\{y \in \mathbb{R} \,:\, y + m(X)^\top Z m(X) = f(X),\ Z \text{ psd}\Big\}$$

where $m(X) = (X^\alpha)_{\alpha \in E_d^n}$ is the vector of monomials of degree at most d. So, the above problem is a problem with decision variables y and Z. To dualize the problem, we write the equality constraint $y + m(X)^\top Z m(X) = f(X)$ in coordinate form. Let $Z = (z_{\alpha,\beta})_{\alpha,\beta \in E_d^n}$. Then the equality $y + m(X)^\top Z m(X) = f(X)$ is

$$y + \sum_{\alpha,\beta \in E_d^n} z_{\alpha,\beta} X^{\alpha+\beta} = \sum_{\gamma \in E_{2d}^n} c_\gamma X^\gamma.$$

Comparing coefficients we get

$$y + z_{0,0} = c_0$$

and

$$\sum_{\alpha,\beta \in E_d^n:\ \alpha+\beta=\gamma} z_{\alpha,\beta} = c_\gamma \qquad\qquad \forall \gamma \in E_{2d}^n \setminus \{0\}$$

The latter can be written as equalities

$$y + \langle A_0, Z\rangle = c_0$$

and

$$\langle A_\gamma, Z\rangle = c_\gamma \qquad\qquad \forall \gamma \in E_{2d}^n \setminus \{0\},$$

where

$$A_\gamma := (\delta_{\alpha+\beta,\gamma})_{\alpha,\beta \in E_d^n} \qquad\qquad \forall \gamma \in E_{2d}^n$$

are symmetric matrices (we use the Kronecker delta notation: $\delta_{s,t} = 1$ if $s = t$ and $\delta_{s,t} = 0$ otherwise). Thus, our problem can now be written as

$$\sup\left\{y \in \mathbb{R} \,:\, Z \text{ psd},\ y + \langle Z, A_0\rangle = c_0,\ \langle A_\gamma, Z\rangle = c_\gamma \text{ for all } \gamma \in E_{2d}^n \setminus \{0\}\right\}. \quad (4.8)$$

Problem (4.8) can be dualized according to the principles we've discussed in previous sections. The psd matrix variable Z yields a psd constraint. The real variable y yields a linear equality. Each linear equality constraint for $\gamma \in E_{2d}^n$ yields a real variable, which we denote by v_γ.

The dual is the following problem

$$\inf\left\{\sum_{\gamma \in E_{2d}^n} v_\gamma c_\gamma \,:\, v_\gamma \in \mathbb{R}\ \forall \gamma \in E_{2d}^n,\ \sum_{\gamma \in E_{2d}^n} v_\gamma A_\gamma \text{ psd},\ v_0 = 1\right\}$$

Taking into account how A_γ were defined we see that

$$\sum_{\gamma \in E_{2d}^n} v_\gamma A_\gamma = \left(\sum_{\gamma \in E_{2d}^n} v_\gamma \delta_{\alpha+\beta,\gamma}\right)_{\alpha,\beta \in E_d^n} = \Big(v_{\alpha+\beta}\Big)_{\alpha,\beta \in E_d^n}.$$

So, the dual SDP gets the form

$$\inf\left\{\sum_{\gamma \in E_{2d}^n} v_\gamma c_\gamma \,:\, v_\gamma \in \mathbb{R}\ \forall \gamma,\ \Big(v_{\alpha+\beta}\Big)_{\alpha,\beta \in E_d^n} \text{ psd},\ v_0 = 1\right\}.$$

Remark 4.14. The problem (4.8) was established to derive lower bounds on (4.7), but (4.8) does not directly suggest any choice of $x \in \mathbb{R}^n$ that may be good. It turns out that an x can be chosen from the dual problem. It is known that, under certain assumptions, $x = (x_1, \ldots, x_n)$ with

$$x_1 := v_{(1,0,\ldots,0)}, \qquad x_2 := v_{(0,1,0\ldots,0)}, \qquad \ldots \qquad x_n := v_{(0,\ldots,0,1)}$$

is a good choice (see works of Marshall, Schweighofer, Lasserre, Laurent et al.).

4.9 Truncated moment problem

There is another way to arrive at the dual problem we've seen in the last section.

We show how the approach works for the unconstrained POP only, with the constrained case being similar.

$$\inf \{f(x) \,:\, x \in \mathbb{R}^n\} \tag{4.9}$$

with $f \in \mathbb{R}[X] = R[X_1, \ldots, X_n]$. Let $c = (c_\alpha)_{\alpha \in E^n_{2d}}$ be the vector of coefficients of f and $m(X) = \{X^\alpha \,:\, \alpha \in E^n_{2d}\}$ the vector of all respective monomials. Then the problem can be written as

$$\inf \{\langle c, m(x)\rangle \,:\, x \in \mathbb{R}^n\}. \tag{4.10}$$

This can be viewed as a linear optimization over the set $m(\mathbb{R}^n) := \{m(x) \,:\, x \in \mathbb{R}^n\}$. Thus, the problem is

$$\inf \{\langle c, v\rangle \,:\, v \in m(\mathbb{R}^n)\}. \tag{4.11}$$

If we enlarge $m(\mathbb{R}^n)$ to $\mathrm{conv}(m(\mathbb{R}^n))$ or even to $\mathrm{cl}(\mathrm{conv}(m(\mathbb{R}^n)))$, the problem remains the same. This way we arrive at

$$\inf \{\langle c, v\rangle \,:\, v \in M_{n,2d}\}. \tag{4.12}$$

In general, $M_{n,2d}$ is hard to describe, but at least we can relax it to a set that we can describe with SDP. By Caratheodory's theorem, every element of $v \in \mathrm{conv}(m(\mathbb{R}^n))$ is the convex combination of at most $\dim(m(\mathbb{R}^n)) + 1$ points of $m(\mathbb{R}^n)$. This convex combination can be written as integration

$$v = v^\mu := \int m(x)\mu(dx).$$

with respect to a finitely supported probability measure μ on $\mathbb{R}^n$. Using the definition of Lebesgue integrability, one can show that if μ is any probability measure on $\mathbb{R}^n$, for which the above integral exists, the respective vector v^μ belongs to $M_{n,2d} = \mathrm{cl}(\mathrm{conv}(M_{n,2d}))$; see [Rud76, Definitions 11.21, 11.22]. Note that v^μ is called the vector of moments of μ of order at most $2d$. The problem of characterizing whether a given collection of numbers associated to the multi-indices $\alpha \in \mathbb{Z}^n_+$ is a collection of moments of a probability measure supported in a given set S is a well-known problem in probability and it is known as the problem of moments. So, what special properties does the vector v_μ have? Consider an arbitrary vector $y = (y_\alpha)_{\alpha \in E^n_d} \in \mathbb{R}^{\binom{n+d}{n}}$. Of course $v^\mu_0 = 1$, since μ is a probability measure.

Furthermore, for the matrix $A(v^\mu) = (v^\mu_{\alpha+\beta})_{\alpha,\beta\in E^n_d}$ one has

$$\begin{aligned} y^\top A(v^\mu)y &= \int x^{\alpha+\beta} y_\alpha y_\beta \mu(dx) \\ &= \left(\int x^\alpha y_\alpha \mu(dx)\right)^2 \\ &\geq 0. \end{aligned}$$

That is, $A(v^\mu)$ is psd and (4.12) can be relaxed to

$$\inf\left\{\langle c, v\rangle \,:\, v_0 = 1,\ A(v) \text{ psd}\right\}. \tag{4.13}$$

This is an SDP, because the condition that $A(v)$ is psd is a linear matrix inequality. The matrix $A(v)$ is called a localizing matrix. This is just the same problem we obtained in the previous section through dualization of the SOS relaxation.

In the constrained case, the arguments are quite similar and one also gets similar inequalities. Just to give one example. Let μ be a measure supported on a set for which the inequality $g \geq 0$ is fulfilled, where $g = \sum_\alpha g_\alpha X^\alpha$ is a polynomial, and assume that v^μ is well-defined. Then, for every $y = (y_\beta)_\beta$ one has

$$\int g(x) \left(\sum_\beta x^\beta\right)^2 \mu(dx) \geq 0$$

which again turns out to be a condition $y^\top A_g(v^\mu)y$ for some matrix A_g. Thus we arrive at linear matrix inequalities (of a bit more general form).

The above considerations show that the problem of moments and the positivstellensätze and nichtnegativstellensätze are closely linked. By characterizing non-negativity of a polynomial on $\mathbb{R}, [0, +\infty), [0, 1]$ in terms of sos, we arrive at a characterization of sequences of moments of probability measures supported by the respective sets. All these are classical results in measure theory and probability.

4.10 Peculiarities of SDP

In contrast to LP, there are difficulties in solving SDP in exact arithmetics, as the following two exercises indicate.

Exercise 4.15. *Show that even if the input data of an SDP (coefficients of vectors and matrices) is rational, the optimal solution need not be rational.*

Solution. One can formulate the problem $\min\big\{y \,:\, y \geq x^2,\ y \geq 1 - x\big\}$ as an sdp. The minimum is attained for a positive $x > 0$ with $x^2 = 1 - x$, which is not a rational number. □

Exercise 4.16. *Even if the input data and the optimal solution of an SDP are rational, the encoding size of the output can be huge (the encoding size is the number of one needs to write down the solution, if one encodes rational values through numerator and denominator and the numerators and denominators are encoded in binary system). By huge, we mean exponential in the size of the input.*

Solution. Consider the constraints $x_j \geq x_{j-1}^2$ for $j > 1$, for variables $x_1, \ldots, x_n$ with the constraint $x_1 \geq \frac{1}{2}$. Minimizing x_n under these constraints gives the optimal solution $(x_1, \ldots, x_n) = (\frac{1}{2}, \frac{1}{2^2}, \ldots, \frac{1}{2^{2^n}})$. So for encoding x_n (in the standard form) one would need exponentially many bits. □

Exercise 4.17. *Show that, in contrast to LP, for SDP the finite optimum value (infimum/supremum) is not necessarily attained.*

Solution. To get examples confirming this, it helps to know that the second-order-cone programming is a special case of SDP (this is the topic of Section 4.11 below). So, also optimizing over conic sections is a special case of SDP. Two-dimensional conic sections are parabola or hyperbola or two lines. And the hyperbola is what we can use. If say, $x_1, x_2 \geq 0$ and $x_1 x_2 \geq 1$, then we cannot attain $x_1 = 0$ but can come arbitrarily close to 0. Modeling $x_1 x_2 \geq 1$ as an sdp constraint is not a problem. □

Exercise 4.18. *In degenerate situations, one can have a positive duality gap. Consider for example the problem*

$$\inf \left\{ x \,:\, \begin{pmatrix} 0 & x \\ x & y \end{pmatrix} \; psd, x \geq -1 \right\}$$

What is the optimal value? What is the optimal value of the dual?

Solution. I think, the example goes back to Lovász. The optimal value of this problem is 0. Let's establish the dual. The psd constraint produces a matrix variable $Z \in \mathcal{S}_2^+$ and the linear constraint the scalar variable $u \geq 0$. The objective function is $-u$. Since x, y are unconstrained variables, there will be two equality constraints on the entries of Z.

$$\max \left\{ -u \,:\, u \in \mathbb{R}_+,\; Z \in \mathcal{S}_2^+, \left\langle Z, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\rangle + u = 1, \left\langle Z, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\rangle = 0 \right\}$$

Written in a bit simpler notation

$$\max \left\{ -u \,:\, u \geq 0, \begin{pmatrix} z_{11} & z_{12} \\ z_{12} & 0 \end{pmatrix} \text{ sdp},\; 2z_{12} + u = 1 \right\}$$

The sdp constraint gives $z_{12} = 0$, so that for every feasible solution we get $u = 1$. Thus, the optimal solution is -1, which means that we have a positive finite duality gap. □

4.11 Detour: Second-order cone programming

This subsection is not necessary for discussing polynomial optimization. It can be read if you want to learn more about the connection of semidefinite optimization to other classes of optimization problems.

Second-order cone programming (second order cone problem) (SOCP for short) is more general than LP and less general than SDP. SOCP is the conic programming with respect to the second-order cones

$$\mathrm{SOC}_n := \{(x, y) \,:\, x \in \mathbb{R}^n, u \geq \|x\|\} \subseteq \mathbb{R}^{n+1}.$$

In particular, $\mathrm{SOC}_0 = \mathbb{R}_+$. Thus, LP is indeed a special case.

SOCPs can be written in the following form

$$\inf \{\langle c, x\rangle \,:\, \langle a_i, x\rangle + \beta_i \geq \|A_i x + b_i\| \ \forall i \in [t]\},$$

where $\beta_1, \ldots, \beta_t$ are scalars, $c, a_1, \ldots, a_t, b_1, \ldots, b_t$ are vectors $A_1, \ldots, A_t$ are matrices. Also, from this formulation we see that LP is just the case $A_i = 0$ and $b_i = 0$.

Exercise 4.19. *Show that SOCP is a special case of SDP. It suffices to write $y \geq \|x\|$ as an LMI.*

Solution. We write $y \geq \|x\|$ as $y \geq 0$ and $y^2 \geq x^\top x$. This condition can be written as the LMI $A(x, y) \in \mathcal{S}_+^{n+1}$ with

$$A(x, y) = \begin{pmatrix} yI & x \\ x^\top & y \end{pmatrix}.$$

Let's take a look at the condition $A(x, y) \in \mathcal{S}_+^{n+1}$. It turns out to be useful to consider

$$\det A(x, 1) = \det \begin{pmatrix} I & x \\ x^\top & 1 \end{pmatrix}$$

The matrix $A(x, 1)$ can be brought to a lower triangular form by subtracting a linear combination of the n first columns from the last one. The lower triangular matrix will have $1, \ldots, 1, 1 - \|x\|^2$ on the diagonal. Thus $\det A(x, 1) = 1 - \|x\|^2$. Viewing the determinant as an element of $\mathbb{R}[x, y] = \mathbb{R}[x_1, \ldots, x_n, y]$, we compute

$$\det A(x, y) = \det(yA(x/y, 1)) = y^{n+1}(1 - \frac{1}{y^2}\|x\|^2) = y^{n-1}(y^2 - \|x\|^2).$$

It follows that $\det A(x, y - \lambda) = (y - \lambda)^{n-1}((y - \lambda)^2 - \|x\|^2)$. This shows that the eigenvalues of $A(x, y)$ are

$\lambda = y$ and $\lambda = y \pm \|x\|$. Consequently, $A(x, y)$ is psd if and only if $y \geq \|x\|$ is fulfilled. □

Remark 4.20. Of course, when it comes to solving SOCP it is better to use solvers tailored to SOCP rather than reducing SOCP to SDP. Nevertheless, the above reduction shows what kind of problem classes can be found in SDP (which is important to understand).

4.12 Detour: Linear stochastic optimization

This subsection, too, is optional. It sort of continues the discussion started in the previous subsection.

Stochastic versions of non-stochastic optimization problems arise by allowing random coefficients (in the constraints and/or the objective function). It may not be immediately clear what it means to solve a stochastic problem. If constraints are stochastic, one may want to satisfy each of them with a certain probability. The following exercise show that stochastic linear programming can be connected to (non-stochastic) SOCP.

Exercise 4.21. *We consider a linear program with the non-stochastic objective* $\langle c, x\rangle$*, non-stochastic rand hand sides and stochastic left hand sides. The problem is*

$$\inf \{\langle c, x\rangle \, : \, \mathrm{Prob}[\langle a_i, x\rangle \leq b_i] \geq p \ \forall i \in [m]\}$$

with independent Gaussian vectors $a_i \sim N(\overline{a}_i, C_i)$ *having expectations* $\bar{a}_i \in \mathbb{R}^n$ *and covariance matrices* $C_i \in \mathcal{S}^n_+$*. The value* $p \in (0, 1)$ *is a certain threshold (tolerance). Show that the above problem can be formulated as SOCP.*

Solution. Let's spell out one such constraint (we omit indices to have a simpler notation)

$$\mathrm{Prob}[\langle a, x\rangle \leq b] \geq p \tag{4.14}$$

with $a \in N(\overline{a}, C)$, $\overline{a} \in \mathbb{R}^n$ and $C \in \mathcal{S}^n_+$. We can represent a as $a = C^{1/2}\xi + \bar{a}$ with $\xi = (\xi_1, \ldots, \xi_n)$ and iid $\xi_1, \ldots, \xi_n \in N(0, 1)$. The condition $\langle a, x\rangle \leq b$ can be formulated as $\left\langle \xi, C^{1/2}x\right\rangle \leq b - \langle \overline{a}, x\rangle$. If $C^{1/2}x = 0$ the condition is deterministic, it is just $b - \langle \overline{a}, x\rangle \geq 0$. Otherwise, we rewrite it as $\eta := \left\langle \xi, \frac{C^{1/2}x}{\|C^{1/2}x\|} \right\rangle \leq \frac{b-\langle \overline{a},x\rangle}{\|C^{1/2}x\|}$. Clearly, $\eta \sim N(0, 1)$. Thus, the condition can be written as

$$\frac{b - \langle \overline{a}, x\rangle}{\|C^{1/2}x\|} \geq \Phi^{-1}(p)$$

using the distribution function $\Phi : \mathbb{R} \to [0, 1]$ of $N(0, 1)$. We have thus shown that (4.14) is a conic constraint

$$b - \langle \overline{a}, x\rangle \geq \Phi^{-1}(p)\|C^{1/2}x\|,$$

In the case $C = 0$, we just get a linear constraint. □

5 Interior-point methods for linear and semidefinite optimization

In principle, one could start directly with the discussion of interior-point methods for SDP, to shorten the presentation. However, it is easier to understand the technicalities of the SDP case, if one got the idea of interior-point methods for LP first. So, we first describe general ideas, then discuss LP and then go over to SDP.

5.1 Ways to solve constrained convex problems: a very short synopsis

We are talking about sufficiently smooth problems here. Unconstrained problems can be solved by first and second order methods (gradient descent, Newton and the many generalizations and ramifications thereof). To solve constrained problems, one basic approach is to approximate the original problem by an unconstrained problem by 'deforming' the objective functions. So, we replace the original problem by a so-called auxiliary problem (with an auxiliary objective function), which is unconstrained (or as good as unconstrained). One can distinguish between penalty and barrier methods. In barrier methods we end up with optimizing a function over an open set (which is, more or less, unconstrained optimization). In penalty methods, the auxiliary function is defined on the whole space and the optimal solution for the auxiliary problem may not be feasible (but is hopefully almost feasible) for the original problem.

Yet, another method is the ellipsoid method. In principle, it can be applied for LP, SDP and other problems and it can be used to show solvability in polynomial time, but by now it has only been of theoretical importance.

For solving SDP we are free to choose any of the methods listed above; however, so far the interior-point methods have been the most successful ones. To develop an interior-point method, we need to decide which barrier we could use and then to figure out how to compute the derivatives of the modified objective function.

5.2 Central path for LP with constraints $Ax \leq b$

Consider the system $Ax \leq b$ which we can write as $\langle a_i, x\rangle \leq b_i$, so $a_1, \dots, a_m$ are the rows of A. We assume that the polyhedron P defined by $Ax \leq b$ has non-empty interior. We assume that none of the inequalities are trivial ones having the form $\langle 0, x\rangle \leq 0$. In this case the interior of P is described by $Ax < b$. We introduce the objective $f(x) = \langle c, x\rangle$. Our problem is $\inf\{f(x) : x \in P\}$. For every $\mu \geq 0$ consider the auxiliary objective function

$$f_\mu(x) = \langle c, x\rangle - \mu \sum_{i=1}^m \ln(b_i - \langle a_i, x\rangle)$$

The sum of the logarithms is called a barrier function. For each given $\mu > 0$, we get the auxiliary problem $\min\{f_\mu(x) : x \in \operatorname{int}(P)\}$.

Exercise 5.1. *Show that, if the above assumptions on P are fulfilled, and P is bounded, then f_μ is strictly convex on* $\operatorname{int}(P)$*. Show that f_μ attains its infimum on* $\operatorname{int}(P)$*.*

Thus, if P is bounded, the auxiliary problem has a unique optimal solution $x^*(\mu)$. The map $\mu \mapsto x^*(\mu)$ is a parameterization of a curve (the so-called *central path*).

The idea of central path methods is to start with $x^*(\mu)$ for some $\mu > 0$ and then decrease μ until a desired accuracy is reached. If $\mu_2 \geq \mu_1$ and $x^*(\mu_2)$ is given we can use $x^*(\mu_2)$ as an approximation for $x^*(\mu_1)$, so that one can invoke various kinds of iterative methods (most notably Newton and damped Newton) that would start with $x^*(\mu_2)$ to determine $x^*(\mu_1)$. In this way, we can gradually decrease μ reaching arbitrarily small values and making our auxiliary problem arbitrarily close to the original one.

5.3 Central path for LP with constraints $Ax = b, x \geq 0$

Here we consider the problem in the form

$$\min \{\langle c, x\rangle \,:\, Ax = b, x \geq 0\}.$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $x = (x_1, \ldots, x_n)$ is the vector of decision variables. We assume that there exists $x > 0$ with $Ax = b$. In that case we can introduce the auxiliary problem $\inf \{f_\mu(x) \,:\, Ax = b, x > 0\}$ with the auxiliary objective

$$f_\mu(x) = \langle c, x\rangle - \mu \sum_{j=1}^{n} \ln x_j$$

where $\mu > 0$ is a parameter. For the function f_μ defined on $\{x \,:\, Ax = b, x > 0\}$ one can establish analogous properties as for the function from the previous section (it is strictly convex and the optimum is attained, when $\{x \,:\, Ax = b, x \geq 0\}$ is bounded).

Since f_μ is convex, a necessary and a sufficient condition for $x > 0$ to be optimal for $f_\mu(x)$ is that $Ax = b$ is fulfilled and $\nabla f_\mu(x)$ is orthogonal to the space $Ax = b$. Since $\{x \,:\, Ax = b\}$ coincides with $\ker(A)$ up to translations, we get the condition $\nabla f_\mu(x) \in \ker(A)^\perp$. Since $\ker(A)^\perp = \operatorname{im}(A^\top)$, the condition can be rewritten as $\nabla f_\mu(x) \in \operatorname{im}(A^\top)$. The latter means that $\nabla f_\mu(x)$ is in the linear hull of the columns of A. The condition $\nabla f_\mu(x) \in \operatorname{im}(A^\top)$ is exactly the KKT condition, phrased in geometric terms. Noticing that $\nabla f_\mu(x) = c + \mu(1/x_1, \ldots, 1/x_n)$, we can now formulate the KKT conditions as

$$\begin{aligned} Ax &= b, \\ c + \mu(1/x_1, \ldots, 1/x_n) &= A^\top y \end{aligned}$$

where $y \in \mathbb{R}^m$. Introducing the vector $s = \mu(1/x_1, \ldots, 1/x_n)$, we can rewrite the latter conditions as

$$\begin{aligned} Ax &= b, && (5.1) \\ A^\top y - s &= c && (5.2) \\ (s_1 x_1, \ldots, s_n x_n) &= \mu(1, \ldots, 1) && (5.3) \\ x, s &> 0. && (5.4) \end{aligned}$$

All the equations are linear apart from the equations $s_i x_i = \mu$.

For solving the system (5.1)–(5.4) we can use the Newton method. If we have an approximate solution x, y, s, which satisfies $Ax = b$ and $A^\top y - s = c$ exactly and satisfies the conditions $x_i s_i = \mu$ approximately, we can try to determine a better solution in the form $x + \Delta x, y + \Delta y, s + \Delta s$. For $\Delta x, \Delta y$ and Δs, we obtain the conditions $A\Delta x = 0$, $A^\top \Delta y - \Delta s = 0$. By linearizing the condition $(s_i + \Delta s_i)(x_i + \Delta x_i) = \mu$ removing the quadratic term $\Delta s_i \Delta x_i$, we arrive at $s_i \Delta x_i + \Delta s_i x_i = \mu - s_i x_i$. The variables Δs_i can be eliminated, as we can express them via Δy. In this way we end up with a linear system of equalities in the unknowns $\Delta x, \Delta y$. The size of the system is $m + n$. If A has full row rank (which we can assume wlog), then the latter system has a unique solution). We would have to take care that the update produces a solution which is still positive (so, we are on the safe side if we use damped Newton).

Let's estimate how well the auxiliary problem approximates the original one. In view of the boundedness assumption LP duality gives us the equality

$$\alpha := \min \{\langle c, x\rangle \,:\, x \in \mathbb{R}^n,\ Ax = b, x \geq 0\} = \max \left\{\langle b, y\rangle \,:\, y \in \mathbb{R}^m,\ A^\top y \leq c\right\} \in \mathbb{R}.$$

of the optimal values of the primal and the dual problem. The dual problem can also be written using the slack vector $s \geq 0$ as $A^\top y + s = c, y \in \mathbb{R}^m s \in \mathbb{R}^n_+$. We can thus see now that this version of the central path method constructs a primal/dual pair of solutions. It turns out, these solutions get nearly optimal as $\mu \to 0$.

Indeed, repeating the derivation of the weak duality, we get the estimates

$$\alpha \geq \langle b, y\rangle = \langle Ax, y\rangle = \left\langle x, A^\top y\right\rangle = \langle x, c - s\rangle = \langle x, c\rangle - \langle x, s\rangle = \langle x, c\rangle - n\mu \geq \alpha - n\mu.$$

Hence

$$\alpha - n\mu \leq \langle b, y\rangle \leq \alpha$$

and

$$\alpha \leq \langle c, x\rangle \leq \alpha + n\mu.$$

So, our solutions are optimal up to the additive tolerance $n\mu$.

5.4 $\ln(\det(X))$ is a good barrier for SDP

A natural barrier we can employ for the SDP cone is $\ln(\det(X))$. Indeed, on the boundary of $\mathcal{S}^k_+$ the determinant gets zero, while in the interior of $\mathcal{S}^k_+$ the determinant is strictly positive. So, taking the logarithm, we create a function that goes to (minus) infinity on the boundary. This is exactly what we want.

For the auxiliary problem to be convex, the respective barrier needs to be a convex (or concave) function. We'll verify that $\ln(\det(X))$ is concave in X. Furthermore, in order to be able to apply first and second order methods for solving the auxiliary problem, we need to determine the first and the second derivatives of $\ln(\det(X))$. Note that the gradient and the Hesse matrix in an abstract Euclidean space can be defined analogously to the concrete space $\mathbb{R}^n$. That is, $\nabla f(x^*)$ satisfies $f(x) = f(x^*) + \langle \nabla f(x^*), x - x^*\rangle + o(\|x - x^*\|)$, where $o(x - x^*)$ is a function with $o(\|x - x^*\|)/\|x - x^*\| \to 0$ for $x \to x^*$ and the 'Hesse-Matrix' $\nabla^2 f(x^*)$ is a self-adjoint operator satisfying $f(x) = f(x^*) + \langle \nabla f(x^*), x - x^*\rangle + \frac{1}{2}\left\langle \nabla^2 f(x^*)(x - x^*), x - x^*\right\rangle + o(\|x - x^*\|^2)$.

Exercise 5.2 (Root of a psd matrix)**.** *Show that for every $X \in \mathcal{S}_+^k$, there exists a uniquely defined matrix $X^{1/2} \in \mathcal{S}_+^k$ satisfying $(X^{1/2})^2 = X$.*

Solution. We can write $\mathbb{R}^n = V_1 \oplus \ldots \oplus V_m$, where $V_1, \ldots, V_m$ are eigenspaces of X to the m pairwise distinct eigenvalues $\lambda_1, \ldots, \lambda_m \geq 0$. Thus, we can choose $X^{1/2}$ that sends $v \in V_i$ to $\lambda_i^{1/2} v$ on V_i. It remains to show uniqueness. Let $Y \in \mathcal{S}_+^k$ satisfy $Y^2 = X$. We consider eigenspaces $W_1, \ldots, W_s$ of Y to its eigenvalues $\mu_1, \ldots, \mu_s$. Because of $Y^2 = X$ we see that $W_1, \ldots, W_s$ are eigenspaces of X to the eigenvalues $\mu_1^2, \ldots, \mu_s^2$. So, we get $s = m$ and up to permutation of indices we must have $\mu_i^2 = \lambda_i$ for all $i \in [m]$. □

Exercise 5.3. *For all $A, B \in \mathbb{R}^{n \times n}$ one has* $\mathrm{tr}(AB) = \mathrm{tr}(BA)$*.*

Solution. Use the formula for the product of matrices and the formula for the trace (as the sum of diagonal entries) and see that the left and the right hand side are just the same. □

Proposition 5.4 (see, for example, Renegar, §1.2)**.** *The function $f : \mathrm{int}(\mathcal{S}_+^k) \to \mathbb{R}$ given by $f(X) := \ln(\det(X))$ is strictly concave and infinitely differentiable. Its gradient is*

$$\nabla f(X) = X^{-1}$$

and its 'Hesse-Matrix' is the operator

$$(\nabla^2 f(X))(U) = X^{-1} U X^{-1}$$

These two formulas can also be summarized as the second-order Taylor expansion

$$f(X + U) = f(X) + \langle X^{-1}, U \rangle + \frac{1}{2} \langle X^{-1} U X^{-1}, U \rangle + o(\|U\|^2), \qquad \text{as } U \to 0$$

for every $X \in \mathrm{int}(\mathcal{S}_+^k)$. (The norm $\|U\|$ is the norm with respect to the scalar product that we fixed in $\mathcal{S}^k$; it is the so-called Frobenius norm)

Proof. Note that $\det(X)$ is a polynomial function with respect to the components of X and it is strictly positive on $\mathrm{int}(S^k)$. The function ln is infinitely differentiable on $\mathbb{R}_{>0}$. So, f is infinitely differentiable.

Strict concavity of f. It suffices to check the strict concavity on each line. This means that we check the strict concavity of $\phi(t) = \ln \det(X + tU)$ for every $X \in \mathcal{S}_+^k \setminus \{0\}$ and every $U \in \mathcal{S}^k \setminus \{0\}$. We do this by showing that the second derivative is strictly negative

$$\begin{aligned}
\phi(t) &= \ln(\det(X^{1/2} X^{1/2} + tU)) \\
&= \ln\big(\det(X^{1/2}(I + tX^{-1/2} U X^{-1/2}) X^{1/2})\big) \\
&= \ln\big(\det(X^{1/2}) \det(I + tX^{-1/2} U X^{-1/2}) \det(X^{1/2})\big) \\
&= \ln\big(\det(X) \det(I + tX^{-1/2} U X^{-1/2})\big)
\end{aligned}$$

where $X^{-1/2} = (X^{1/2})^{-1}$. Let's use the notation $\tilde{U} := X^{-1/2} U X^{-1/2}$. This gives

$$\phi(t) = \ln(\det(X)) + \ln \det(I + t\tilde{U}).$$

It suffices to compute the derivative of $\ln\det(I + t\tilde{U})$. $\tilde{U}$ is symmetric, and we introduce its eigenvalues $\tilde{\lambda}_i$ (taking into account the multiplicities). Then $1+t\tilde{\lambda}_i$ are the eigenvalues of $I + t\tilde{U}$. Since $I + t\tilde{U}$ (for $X + tU \in \operatorname{int}(\mathcal{S}^k_+)$) is positive definite, we have $1 + t\tilde{\lambda}_i > 0$. Since the determinant of a symmetric matrix is a product of all the eigenvalues, we obtain

$$\phi'(t) = \sum_{i=1}^n \frac{\partial}{\partial t} \ln(1 + t\tilde{\lambda}_i) = \sum_{i=1}^n \frac{\tilde{\lambda}_i}{1 + t\tilde{\lambda}_i}.$$

and the second derivative is

$$\phi''(t) = -\sum_{i=1}^n \frac{\tilde{\lambda}_i^2}{(1 + t\tilde{\lambda}_i)^2}$$

We have $\phi''(t) \le 0$. We cannot have $\phi''(t) = 0$, because this would imply $\tilde{\lambda}_i = 0$ for all i. But then we'd have $\tilde{U} = 0$ and by this $U = 0$.

For computing the gradient of f we can use the above computation for $\phi'(t)$. We get

$$\phi'(0) = \sum_{i=1}^n \tilde{\lambda}_i = \operatorname{tr}(X^{-1/2}UX^{-1/2}) = \operatorname{tr}(X^{-1/2}X^{-1/2}U) = \operatorname{tr}(X^{-1}U) = \left\langle X^{-1}, U \right\rangle.$$

So, we get the 'directional' derivative

$$\frac{\partial f}{\partial U}(X) = \left\langle X^{-1}, U \right\rangle$$

which gives $\nabla f(X) = X^{-1}$.

For the derivation of the 'Hesse-Matrix' we can use the above equation for $\phi''(t)$. We get

$$\begin{aligned} \phi''(0) = -\sum_{i=1}^n \tilde{\lambda}_i^2 = \operatorname{tr}(\tilde{U}^2) &= \operatorname{tr}(X^{-1/2}UX^{-1/2}X^{-1/2}UX^{-1/2}) \\ &= \operatorname{tr}(X^{-1/2}UX^{-1}UX^{-1/2}) \\ &= \operatorname{tr}(UX^{-1}UX^{-1}) && \text{(by Exercise 5.3)} \\ &= \left\langle U, X^{-1}UX^{-1} \right\rangle. \end{aligned}$$

□

5.5 Interior-point methods for SDP

I rely more or less on Chapter 6 of [GM12].

Let's consider the SDP

$$\inf \left\{ \langle C, X \rangle \,:\, X \in \mathcal{S}^k_+,\ \langle A_i, X \rangle = b_i\ \forall i \in [m] \right\}.$$

We assume that the problem is strictly feasible, that is, there exists $X \in \operatorname{int}(\mathcal{S}^k_+)$ with $\langle A_i, X \rangle = b_i$ for all $i \in [m]$. We introduce a parameter $\mu > 0$ and introduce the auxiliary problem

$$\inf \left\{ \langle C, X \rangle - \mu \ln(\det(X)) \,:\, X \in \operatorname{int}(\mathcal{S}^k_+),\ \langle A_i, X \rangle = b_i\ \forall i \in [m] \right\}.$$

Our auxiliary objective function $f_\mu(X)$ is strictly convex and the the infimum is actually a minimum, because $f_\mu(X) \to \infty$ when X approaches the boundary of $\mathcal{S}^k_+$. Due to the strict convexity, the minimum is unique. By KKT-condition, the gradient of f_μ is orthogonal to the affine space $\{X \in \mathcal{S}^k : \langle A_i, X\rangle = b_i \ \forall i \in [m]\}$. Thus, we can write the gradient of f_μ as a linear combination of the 'normal vectors' A_i of the hyperplanes $\langle A_i, X\rangle = b_i$. So, we arrive at the condition

$$C - \mu X^{-1} = \sum_{i=1}^m y_i A_i$$

We introduce the matrix $S = \mu X^{-1}$, for which the identity $SX = \mu I$ holds. Summarizing, we arrive at the system

$$\begin{aligned} X, S \in \operatorname{int}(\mathcal{S}^k_+),\ y_1, \dots, y_m \in \mathbb{R} && (5.5)\\ \langle A_i, X\rangle = b_i \qquad \forall i \in [m] && (5.6)\\ S + \sum_{i=1}^m y_i A_i = C, && (5.7)\\ SX = \mu I. && (5.8) \end{aligned}$$

As in the case of LP, one can see that $y_1, \dots, y_m$ satisfying the above conditions is a feasible solution of the dual problem:

$$\sup \left\{ y_1 b_1 + \dots + y_m b_m \ : \ C - \sum_{i=1}^m y_i A_i \in \mathcal{S}^k_+ \right\}.$$

Note also that $SX = \mu I$ is a kind of modified complementary slackness.

Again, (5.5)–(5.8) can be solved using iterative methods (say, damped Newton). Assume that we have $X, S, y_1, \dots, y_m$ that satisfy all the above constraints exactly, except for the constraint $SX = \mu I$, which is only satisfied approximately.

For updating the current approximation $X, S, y_1, \dots, y_m$ to a new $X + \Delta X, S + \Delta S, y_1 + \Delta y_1, \dots, y_m + \Delta y_m$, we plug in the new approximation and linearize the last constraint with respect to the ΔS and ΔX:

$$\begin{aligned} \langle A_i, \Delta X\rangle = 0, \qquad \forall i \in [m] && (5.9)\\ \Delta S + \sum_{i=1}^m \Delta y_i A_i = 0, && (5.10)\\ \Delta S X + S \Delta X = \mu I - SX && (5.11) \end{aligned}$$

That is the point, where the theory of interior-point methods for SDP starts. We've got a system of linear equations. Solving such equations is linear algebra, but the size is large and we want to solve the system approximately.

Various approaches exist. One approach is to first give up the symmetry of ΔX and then do an update using $(\Delta X + \Delta X^\top)/2$ rather than ΔX. With this approach, from (5.11) we get

$$\Delta X = \mu S^{-1} - X - S^{-1} \Delta S X.$$

We plug in ΔS from (5.10) and get

$$\Delta X = \mu S^{-1} - X + \sum_{j=1}^{m} \Delta y_j S^{-1} A_j X.$$

Inserting this expression for ΔX into (5.9), which can be written as $\operatorname{tr}(A_i \Delta X) = 0$, we arrive at the equation

$$\sum_{j=1}^{m} \operatorname{tr}(A_i S^{-1} A_j X) \Delta y_j = \operatorname{tr}((X - \mu S^{-1}) A_i).$$

This is a $m \times m$ linear system. It can be shown that, if $A_1, \ldots, A_m$ are linearly independent vectors of the space $\mathcal{S}^k$, then the matrix of the system is symmetric positive definite (so that there exists a unique solution). The method of finding the non-symmetric ΔX and then symmetrizing works well in practice (though polynomial time bounds on the number of iterations are obtained using different, more involved methods, references can be found in [GM12]). Regarding the complexity: Determination of ΔX can be viewed as solving k linear equations of size $k \times k$. So, $O(k^4)$ operations. Solving the linear system for the $\Delta y_1, \ldots, \Delta y_m$ requires $O(m^3)$ operations. Note that m can be as large as $O(k^2)$. Thus, we end up with $O(m^3) = O(k^6)$. When m is not large, say $m = O(k)$, we still have $O(k^4)$ in one iteration of the Newton step. In the applications related to polynomial optimization, both m and k are big. So, one is forced to work with really huge matrices.

References

[AM16] M. F. Atiyah and I. G. Macdonald, *Introduction to commutative algebra*, economy ed., Addison-Wesley Series in Mathematics, Westview Press, Boulder, CO, 2016, For the 1969 original see [MR0242802]. MR 3525784

[Ave13] Gennadiy Averkov, *Constructive proofs of some positivstellensätze for compact semialgebraic subsets of* $\mathbb{R}^d$, Journal of Optimization Theory and Applications **158** (2013), no. 2, 410–418.

[BCR98] Jacek Bochnak, Michel Coste, and Marie-Françoise Roy, *Real algebraic geometry*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], vol. 36, Springer-Verlag, Berlin, 1998, Translated from the 1987 French original, Revised by the authors. MR 1659509

[BGHED14] Bernd Bank, Marc Giusti, Joos Heintz, and Mohab Safey El Din, *Intrinsic complexity estimates in polynomial optimization*, J. Complexity **30** (2014), no. 4, 430–443. MR 3212780

[CL77] Man Duen Choi and Tsit Yuen Lam, *An old question of hilbert*, Queen's papers in pure and applied mathematics **46** (1977), no. 385-405, 11.

[CPSV16] Lynn Chua, Daniel Plaumann, Rainer Sinn, and Cynthia Vinzant, *Gram spectrahedra*, arXiv preprint arXiv:1608.00234 (2016).

[Dür10] Mirjam Dür, *Copositive programming—a survey*, Recent advances in optimization and its applications in engineering **320** (2010).

[GKR16] Charu Goel, Salma Kuhlmann, and Bruce Reznick, *On the choi–lam analogue of hilbert's 1888 theorem for symmetric forms*, Linear Algebra and its Applications **496** (2016), 114–120.

[GM12] Bernd Gärtner and Jiří Matoušek, *Approximation algorithms and semidefinite programming*, Springer, Heidelberg, 2012. MR 3015090

[Las15] Jean Bernard Lasserre, *An introduction to polynomial and semialgebraic optimization*, Camb. Texts Appl. Math., Cambridge: Cambridge University Press, 2015 (English).

[Lau09] Monique Laurent, *Sums of squares, moment matrices and optimization over polynomials*, Emerging applications of algebraic geometry. Papers of the IMA workshops Optimization and control, January 16–20, 2007 and Applications in biology, dynamics, and statistics, March 5–9, 2007, held at IMA, Minneapolis, MN, USA, New York, NY: Springer, 2009, pp. 157–270 (English).

[LPR14] Henri Lombardi, Daniel Perrucci, and Marie-Françoise Roy, *An elementary recursive bound for effective positivstellensatz and hilbert 17-th problem*, arXiv preprint arXiv:1404.2338 (2014).

[Mar08a] Murray Marshall, *Positive polynomials and sums of squares*, Math. Surv. Monogr., vol. 146, Providence, RI: American Mathematical Society (AMS), 2008 (English).

[Mar08b] ______, *Positive polynomials and sums of squares*, Mathematical Surveys and Monographs, vol. 146, American Mathematical Society, Providence, RI, 2008. MR 2383959

[ML12] F. Vallentin M. Laurent, *Semidefinite optimization lecture notes*, 2012.

[MS09] Kurt Mehlhorn and Michael Sagraloff, *Isolating real roots of real polynomials*, ISSAC 2009—Proceedings of the 2009 International Symposium on Symbolic and Algebraic Computation, ACM, New York, 2009, pp. 247–254. MR 2742714

[MS11] ______, *A deterministic algorithm for isolating real roots of a real polynomial*, J. Symbolic Comput. **46** (2011), no. 1, 70–90. MR 2736359

[Par03] Pablo A. Parrilo, *Semidefinite programming relaxations for semialgebraic problems*, Math. Program. **96** (2003), no. 2, Ser. B, 293–320, Algebraic and geometric methods in discrete optimization. MR 1993050

[PR01] Victoria Powers and Bruce Reznick, *A new bound for Pólya's theorem with applications to polynomials positive on polyhedra*, J. Pure Appl. Algebra **164** (2001), no. 1-2, 221–229, Effective methods in algebraic geometry (Bath, 2000). MR 1854339

[PS12] Albrecht Pfister and Claus Scheiderer, *An elementary proof of Hilbert's theorem on ternary quartics*, J. Algebra **371** (2012), 1–25. MR 2975385

[Rez78] Bruce Reznick, *Extremal PSD forms with few terms*, Duke Math. J. **45** (1978), no. 2, 363–374. MR 0480338

[Rud76] Walter Rudin, *Principles of mathematical analysis*, third ed., McGraw-Hill Book Co., New York-Auckland-Düsseldorf, 1976, International Series in Pure and Applied Mathematics. MR 0385023

[SM16] Michael Sagraloff and Kurt Mehlhorn, *Computing real roots of real polynomials*, J. Symbolic Comput. **73** (2016), 46–86. MR 3385951