Positivity Certificates & Polynomial Optimization

Master's thesis in Mathematics

William H. Krill — 201407609

Advisor: Anders N. Jensen November 2020

Institut for Matematiske Fag
Aarhus Universitet

Abstract

We present different core results in real algebraic geometry, which all has to do with certifying non-negativity of real polynomial functions on semi-algebraic sets. After carefully introducing the necessary Artin-Schreier theory on ordered and real closed fields, we move on to give detailed proofs the theorems of Krivine and Stengle known as the Semi-algebraic (or Real) Nullstellensatz and the Positivstellensatz. We see how this in turn leads to Artin's solution to Hilbert's 17th problem. Giving an explicit formulation of the Semi-algebraic Nullstellensatz, we obtain a very short proof of the Positivstellensatz. Different improvements of the Positivstellensatz under certain stronger assumptions are presented. We also highlight the dual perspective from functional analysis by briefly discussing the Moment Problem.

Regarding the problem of computing certificates in practice, we illustrate how this can be done efficiently by techniques of semi-definite programming (SDP), a subfield of convex optimization. Among other applications we use SDP-based tools to obtain a simple computer assisted proof (in terms of a rational certificate) of a geometric inequality known as the Hadwiger-Finsler Inequality.

Finally we see how the results from real algebraic geometry and functional analysis, together with semidefinite programming, lead to an efficient and general algorithm for constrained polynomial optimization, known as Lasserre's Hierarchy. We implement the algorithm and apply it to different problems, including an approximation of the Max-Cut Problem from graph theory.

Contents

Introduction

Consider a real polynomial $f \in \mathbb{R}[\underline{X}] = \mathbb{R}[X_1, \ldots, X_n]$. A natural question to ask is whether or not f is globally non-negative. I.e. whether

$$f(x) \geq 0 \qquad \forall x \in \mathbb{R}^n, \tag{1}$$

when we see f as a polynomial function $f\colon \mathbb{R}^n \to \mathbb{R}$. A clear necessary condition is that the degree of f is even.[1]

Interestingly we also have an simple sufficient condition for f to be non-negative, namely if f can be written as a sum of squares (SOS) of polynomials $f_i \in \mathbb{R}[\underline{X}]$, i.e.

$$f = \sum_i f_i^2. \tag{2}$$

We call a relation of this form an *SOS-decomposition* of f.

If we can somehow find an SOS-decomposition of f we will have very little trouble convincing someone else that f *is* non-negative. One only have to understand a very basic property of the real numbers, namely that sums of squares are always non-negative. In that sense the SOS-decomposition serves as a so called *certificate* of non-negativity.

Some of the questions we will be concerned with in this thesis is:

- Under which conditions does an SOS-decomposition exist?
- How do we (efficiently) compute an SOS-decomposition if it exists?
- How do we certify non-negativity if no decomposition exists?
- More generally how can we certify non-negativity on a *subset* of $\mathbb{R}^n$?

[1] This is not difficult to see if f is univariate and if f is not univariate we can just restrict f to a sufficient line and obtain a univariate polynomial of same degree.

One of the reasons to care about the above questions is that if we can efficiently determine non-negativity of polynomials then we can also efficiently solve polynomial optimization problems. To see why this is, suppose we want to minimize $f \in \mathbb{R}[\underline{X}]$ over $\mathbb{R}^n$ and that the minimum

$$f^* = \inf\{f(x) \mid x \in \mathbb{R}^n\} > -\infty, \tag{3}$$

is finite. Then $f - a$ is globally non-negativ for an $a \in \mathbb{R}$ if and only if $a \leq f^*$ and hence we can obtain f^* as

$$f^* = \sup\{a \in \mathbb{R} \mid f - a \geq 0 \text{ on } \mathbb{R}^n\}, \tag{4}$$

which could e.g. be computed by bisection. We will discuss these matters in more detail later.

It turns out that in general it is computationally difficult (in fact NP-hard) to determine non-negativity of polynomials. However we can efficiently determine if a polynomial is a sum of squares. The main idea in polynomial optimization through sums of squares, so-called *SOS-programming*, is to replace the condition $f - a \geq 0$ in (4) by $f - a$ is a sum of squares. This will not always yield the true optimum f^* but instead give a lower bound (since we have reduced the feasible set).

We can use the same ideas to solve constrained optimization problems, i.e. minimize f on some subset $S \subseteq \mathbb{R}^n$. However this requires more sophisticated techniques as we will need "SOS-like" conditions for being non-negative on S – clearly any sum of squares being globally non-negative is also non-negative on S, but we need to broaden the class to get good lower bounds.

Acknowledgments

My greatest appreciation goes to my adviser, Anders Nedergaard Jensen, who has supported me throughout the process. He has been a priceless help, and I would like to thank him for all his good advice and questions. It has been a great experience.

I had the pleasure of writing some of my thesis in the summer period where the university was almost dead and most of my friends and family were spending their days at the beach. These long days at the office would have been fairly depressing if it had not been for Jonathan Ditlevsen and Lota Cedic, who were in the same situation and office as me. Thank you for pancakes, mandagssnaps, poetry readings, and walks in the park! Also special thanks to Jonathan for the careful readings and valuable comments. I owe you!

Thanks to Niels Lauritzen for pointing out some interesting references and providing me with Hagoromo chalk.

Finally, I would like to express my deepest gratitude to my family. For all their love and support – and for taking me to the beach when I needed it. To Ida for being Ida.

1 Sums of squares

In this chapter we examine the connection between polynomials which are sums of squares and the set of semidefinite matrices. We will also clarify the distinction between non-negative polynomials and SOS-polynomials by giving a short historical perspective on Hilbert's 17th problem.

1.1 Polynomials

We let k be a field and $\mathbb{N} = \{0, 1, 2, \ldots\}$ be the set of natural numbers.

Definition 1.1 (Monomial) *Given an $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$ and variables $X_1, \ldots, X_n$, we let $\underline{X}^\alpha = X_1^{\alpha_1} \cdots X_n^{\alpha_n}$ and call this a* monomial. *The* degree *of a monomial $\underline{X}^\alpha$ is defined to be* $\deg \underline{X}^\alpha = |\alpha| = \alpha_1 + \cdots + \alpha_n$.

The set of all monomials in n variables is denoted $\mathcal{T}^n = \{\underline{X}^\alpha \mid \alpha \in \mathbb{N}^n\}$ and the set of all monomials in n variables of degree at most d is denoted $\mathcal{T}_d^n = \{\underline{X}^\alpha \mid \alpha \in \mathbb{N}_d^n\}$ where $\mathbb{N}_d^n = \{\alpha \in \mathbb{N}^n \mid |\alpha| \leq d\}$. There are $|\mathcal{T}_d^n| = \binom{n+d}{n} = \binom{n+d}{d}$ monomials in n variables of degree at most d and $|\mathcal{T}_d^n \backslash \mathcal{T}_{d-1}^n| = \binom{n-d-1}{n-1}$ monomial of degree precisely d.

Definition 1.2 (Polynomial) *A polynomial is a finite linear combination of monomials*

$$f = \sum_{\alpha \in \mathbb{N}^n} b_\alpha \underline{X}^\alpha,$$

where $b_\alpha \in k$.

The support *of f is the set*

$$\operatorname{supp}(f) = \{\alpha \in \mathbb{N}^n \mid b_\alpha \neq 0\}.$$

The (total) degree of f is given by

$$\deg f = \max\{|\alpha| \mid \alpha \in \operatorname{supp}(f)\}, \tag{1.1}$$

with the convention $\deg 0 = -1$.

We denote by $k[\underline{X}] = k[X_1, \ldots, X_n]$ the set of polynomials in n variables and by $k[\underline{X}]_d = \{f \in k[\underline{X}] \mid \deg f \leq d\}$ the set polynomials of degree at most d. $k[\underline{X}]$ has a natural structure as a ring and can also be considered as an infinite dimensional vectorspace over k with $\mathcal{T}^n$ as a standard basis. Likewise $k[\underline{X}]_d$ is an $\binom{n+d}{d}$-dimensional vectorspace over k with $\mathcal{T}_d^n$ as a basis.

1.2 Semidefinite matrices

In this section ee collect the most important properties of positive-semidefinite matrices which we will need through the rest of the chapters.

The set of real symmetric n by n matrices is denoted $\mathcal{S}^n$. It is an $\frac{n(n+1)}{2}$-dimensional vector space over $\mathbb{R}$ which we can identify with $\mathbb{R}^{n(n+1)/2}$. We equip $\mathcal{S}^n$ with the Frobenius inner product $\langle X, Y \rangle := \operatorname{Tr}(XY) = \sum_{i,j} X_{ij} Y_{ij}$. A matrix $A \in \mathcal{S}^n$ is *positive semidefinite (psd)* if $x^T A x \geq 0$ for all $x \in \mathbb{R}^n$ and it is *positive definite (pd)* if $x^T A x > 0$ for all $x \in \mathbb{R}^n \backslash \{0\}$. The set of positive semidefinite n by n matrices is denoted $\mathcal{S}_+^n$ and the set of positive definite matrices is denoted $\mathcal{S}_{++}^n$.

We have the following characterization of the positive semidefinite matrices:

Proposition 1.3 *Let $G \in \mathcal{S}^n$. Then the following is equivalent:*

i) G is positive semidefinite.

ii) All eigenvalues of G are non-negative.

iii) $G = B^T B$ for some $B \in \mathbb{R}^{m \times n}$ where $m = \operatorname{rk} A \leq n$.

iv) All $2^n - 1$ principal minors[1] of G are non-negative.

v) $G + \epsilon I$ is positive definite for every $\epsilon > 0$, where $I = I_n$ denotes the identity matrix.

vi) $\langle G, H \rangle \geq 0$ for all $H \in \mathcal{S}_+^n$.

We direct the reader to [14, Cha. 9] and [4, Sec. 2.6.1] for proofs of these facts.

The condition *iv)* shows that the set of positive semidefinite matrices is a basic closed semi-algebraic set[2] defined by the $2^n - 1$ principal minors which

[1] The determinant of a sub-matrix of G with identical row and column indices.

[2] which we define in Section 2.2

(being determinants) are polynomials. We also note that the matrix B in the decomposition in *iii)* can be chosen to be square by simply extending B with $n - m$ zero-rows.

The space of psd matrices $\mathcal{S}^n_+$ has a nice geometric interpretation as a convex cone in $\mathcal{S}^n$.

Definition 1.4 *A set $\mathcal{C} \in \mathbb{R}^n$ is a* convex cone *if $x + y \in S$ and $\lambda x \in S$ for all $x, y \in S$ and $\lambda \geq 0$.*

A convex cone $\mathcal{C} \subseteq \mathbb{R}^n$ is pointed *if $\mathcal{C} \cap -\mathcal{C} = \{0\}$,* solid *if $\mathcal{C}$ has an interior point. $\mathcal{C}$ is* proper *if it is closed, solid, and pointed.*

Theorem 1.5 *$\mathcal{S}^n_+$ is a proper cone in $\mathcal{S}^n$.*

The interior of $\mathcal{S}^n_+$ is the positive definite matrices $\mathcal{S}^n_{++}$.

1.3 Sums of squares

Consider the $\mathcal{T}^n_d$-valued vector $z = (\underline{X}^\alpha)_{\alpha \in \mathbb{N}^n_d}$ and the square matrix $z^T z = (\underline{X}^{\beta+\gamma})_{\beta,\gamma \in \mathbb{N}^n_d}$ indexed by all the monomials in $\mathcal{T}^n_d$ (in some fixed order). Any polynomial $f = \sum_{\alpha \in \mathbb{N}^n_{2d}} b_\alpha \underline{X}^\alpha$ of degree $2d$ can be written in the form

$$f = z^T A z = \langle (\underline{X}^{\beta+\gamma})_{\beta,\gamma \in \mathbb{N}^n_d}, A \rangle, \tag{1.2}$$

for some symmetric matrix $A \in \mathcal{S}^{\mathbb{N}^n_d}$.[3] For instance we can define A elementwise by picking for each $\alpha \in \operatorname{supp}(f)$ two exponents $\beta, \gamma \in \mathbb{N}^n_d$ such $\alpha = \beta + \gamma$ and let $A_{\beta\gamma} = A_{\gamma\beta} = \frac{1}{2} b_\alpha$ if $\beta \neq \gamma$ and $A_{\beta,\beta} = b_\alpha$ otherwise. This is not the only possible choice. In fact by matching terms we see that the set of candidates A in the realtion $f = z^T A z$ is given by the affine subspace of $\mathcal{S}^{\mathbb{N}^n_d}$:

$$\mathcal{L}_f = \Big\{ A \in \mathcal{S}^{\mathbb{N}^n_d} \,\Big|\, \sum_{\beta+\gamma=\alpha} A_{\beta\gamma} = b_\alpha, \quad \forall \alpha \in \operatorname{supp}(f) \Big\}. \tag{1.3}$$

It turns out that f is a sum of squares if and only if f has a representation in terms of a positive semidefinite matrix G:

Lemma 1.6 *$f \in \mathbb{R}[\underline{X}]_{2d}$ is a sum of squares if and only if there exitsts a psd $G \in \mathcal{S}^{\mathbb{N}^n_d}_+$ such that*

$$f = z^T G z. \tag{1.4}$$

[3] When we use the superscript $\mathbb{N}^n_d$ we of course mean $|\mathbb{N}^n_d| = \binom{n+d}{d}$, but we prefer the former. For instance we will also use the notation $\mathbb{R}^{\mathbb{N}}$ and $\mathbb{R}^{\mathbb{N}^n}$ for infinite sequences and multi-sequences.

or equivalently

$$\mathcal{L}_f \cap \mathcal{S}_+^{\mathbb{N}_d^n} \neq \emptyset. \tag{1.5}$$

In other words

$$\sum \mathbb{R}[\underline{X}]_d^{(2)} = \{\langle (\underline{X}^{\beta+\gamma})_{\beta,\gamma\in\mathbb{N}_d^n}, A\rangle \mid G \in \mathcal{S}_+^{\mathbb{N}_d^n}\} \tag{1.6}$$

Proof Suppose $f = z^T G z$ for some psd G. Write $G = B^T B$ according to Proposition 1.3. Then

$$f = z^T G z = z^T B^T B z = \|Bz\|^2 = \sum_{i=1}^{m} [Bz]_i^2 \tag{1.7}$$

where $m = \operatorname{rk} G$.

On the other hand, if $f = \sum_{i=1}^m f_i^2$ then we can define $B \in \mathbb{R}^{m\times\mathbb{N}_d^n}$ by letting $[B]_{i\alpha}$ be the coefficient of $\underline{X}^\alpha$ in f_i. Then $[Bz]_i = f_i$ and (1.7) holds again with $G = B^T B$ which is psd by Proposition 1.3. □

A psd matrix G describing the SOS-decomposition of f as in (1.4) is called a *Gram matrix* for f.

What can be said about the polynomials f_i in a SOS-decomposition $f = \sum_{i=1}^m f_i^2$? If $\deg f = 2d$ then since no cancellation of leading terms can appear we must have $\deg f_i \leq d$. So f_i has at most $|\mathcal{T}_d^n| = \binom{n+d}{d}$ coefficients to be determined. When f is sparse meaning there are few terms compared to the maximum of $|\mathcal{T}_{2d}^n|$, it is helpful to look at the *Newton polytope*.

Definition 1.7 *The* Newton polytope $\mathrm{NP}(f)$ *of* f *is the convex hull of* $\mathrm{supp}(f)$ *as a subset of* $\mathbb{R}^n$.

The following theorem gives a nice restriction on the candidates in an SOS-decomposition which exploits sparsity.

Theorem 1.8 ([24]) *If* $f = \sum_{i=1}^m f_i^2$ *then* $\mathrm{NP}(f_i) \subseteq \frac{1}{2}\mathrm{NP}(f)$ *for* $i = 1, \ldots, m$. *In particular* $\mathrm{supp}(f_i) \subseteq \frac{1}{2}\mathrm{NP}(f) \cap \mathbb{Z}^n$.

Remark 1.9 According to Theorem 1.8 we can improve Lemma 1.6: since $\mathrm{supp}(f_i) \subseteq \frac{1}{2}\mathrm{NP}(f) \cap \mathbb{Z}^n$ we can use $\frac{1}{2}\mathrm{NP}(f) \cap \mathbb{Z}^n$ as our index set instead of $\mathcal{T}_d^n$. The first part of the proof is clearly not affected by this restriction. The other part follows by the fact that the column $[B^T]_\alpha$ is zero for any $\alpha \in \mathcal{T}_d^n \setminus \frac{1}{2}\mathrm{NP}(f)$ by construction. △

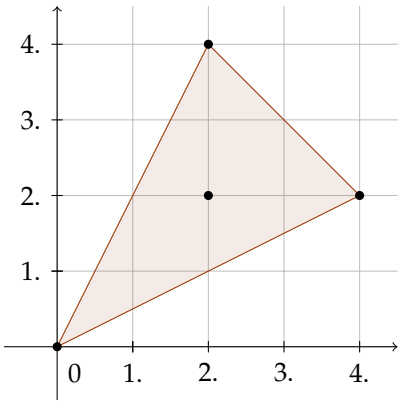

Figure 1.1: Newton polytope $\mathrm{NP}(f)$ of the polynomial $f = X^4Y^2 + X^2Y^4 - 3X^2Y^2 + 1$.

Example 1.10 Let

$$f = X_1^4 - 2X_1^2X_2X_3 - X_1^2 + X_2^2X_3^2 + 2X_2X_3 + 2 \in \mathbb{R}[X_1, X_2, X_3].$$

We wish to show that f is a sum of squares and hence positive by finding an SOS-decomposition $f = \sum_{i=1}^m f_i^2$.

Observing that f can be realized as a bivariate polynomial in the variables X_1^2 and X_2X_3, a first attempt could be to write

$$f = \begin{bmatrix} 1 \\ X_1^2 \\ X_2X_3 \end{bmatrix}^T \begin{bmatrix} 2 & -1/2 & 1 \\ -1/2 & 1 & -1 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ X_1^2 \\ X_2X_3 \end{bmatrix}. \tag{1.8}$$

Unfortunately this matrix is not psd since $\frac{1}{2}(3 - \sqrt{10}) < 0$ is a negative eigenvalue.

Inspecting the Newton polytope we see that

$$\tfrac{1}{2}\mathrm{NP}(f) \cap \mathbb{Z}^3 = \{(0,0,0), (1,0,0), (2,0,0), (0,1,1)\}.$$

According to Lemma 1.6 and Remark 1.9 we let $z = (1, X_1, X_1^2, X_2X_3)^T$. At first glance this does not seem like an improvement as we just get

$$f = \begin{bmatrix} 1 \\ X_1 \\ X_1^2 \\ X_2X_3 \end{bmatrix}^T \begin{bmatrix} 2 & 0 & -1/2 & 1 \\ 0 & 0 & 0 & 0 \\ -1/2 & 0 & 1 & -1 \\ 1 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ X_1 \\ X_1^2 \\ X_2X_3 \end{bmatrix} =: z^TAz. \tag{1.9}$$

This matrix A is neither psd (it has the same negative eigenvalue as before).

However by adding X_1 we have introduced some interdependence between the variables:

$$z_2^2 = (X_1)^2 = 1 \cdot (X_1^2) = z_1 z_3. \tag{1.10}$$

Letting $A_1 = 2E_{22} - E_{13} - E_{31}$ (where E_{ij} is an elementary matrix with 1 in entry (i, j) and 0 elsewhere) the relation (1.10) implies that

$$z^T(\lambda A_1)z = 0 \tag{1.11}$$

for any $\lambda \in \mathbb{R}$. We can then define the affine function $l\colon \mathbb{R} \to \mathcal{S}^4$ by:

$$l(\lambda) = A + \lambda A_1 = \begin{bmatrix} 2 & 0 & -1/2 - \lambda & 1 \\ 0 & 2\lambda & 0 & 0 \\ -1/2 - \lambda & 0 & 1 & -1 \\ 1 & 0 & -1 & 1 \end{bmatrix}$$

and obtain

$$f = z^T A z + z^T(\lambda A_1)z = z^T(A + \lambda A_1)z = z^T l(\lambda) z \tag{1.12}$$

for all $\lambda \in \mathbb{R}$.

Hence we can search for a psd matrix in the range of l which is a 1-dimensional affine subspace of $\mathcal{S}^4$. We will later see how semidefinite programming handles this effectively.

For this example we will take a more low level approach: Using a CAS tool[4] we can plot the four eigenvalues simultaneously as a function of λ. See Figure 1.2. This suggests that at $\lambda = 1/2$ all eigenvalues are non-negative and hence that $l(1/2)$ is psd. By direct computation we verify that this is the case. Now

$$l(1/2) = \begin{bmatrix} 2 & 0 & -1 & 1 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & -1 \\ 1 & 0 & -1 & 1 \end{bmatrix} = B^T B \tag{1.13}$$

where

$$B = \begin{bmatrix} \sqrt{2} & 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix}$$

From this and the proof of Lemma 1.6 we get the SOS-decomposition

$$f = 2\big(1 - \tfrac{1}{2}X_1^2 + \tfrac{1}{2}X_2X_3\big)^2 + X_1^2 + \tfrac{1}{2}\big(X_1^2 - X_2X_3\big)^2$$

which certifies that f is indeed non-negative. ○

[4]We used the SymPy and matplotlib packages in Python.

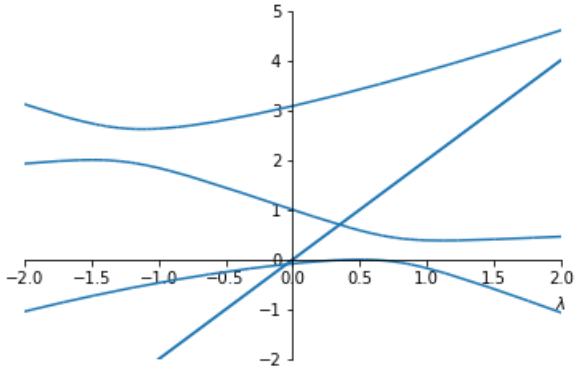

Figure 1.2: Eigenvalues of $l(\lambda)$ as a function of λ.

1.4 Hilbert's 17th problem

We denote by $P_{n,2d} \subseteq \mathbb{R}[\underline{X}]_{2d}$ the set of non-negative polynomials in n variables of degree at most $2d$. Similarly we denote by $\sum_{n,2d} \subseteq \mathbb{R}[\underline{X}]_{2d}$ the set of all polynomials which admits an SOS-decomposition, i.e. $\sum_{n,2d} = \sum \mathbb{R}[\underline{X}]_d^{(2)}$.

In 1888 Hilbert proved that $P_{n,2d} = \sum_{n,2d}$ if and only if either

- $n = 1$: f is univariate.
- $d = 2$: f is quadratic.
- $(n, 2d) = (2, 4)$: f is a bivariate quatric.

In all other cases Hilbert showed that there exist non-negative polynomials which are not sums of squares. This observation lead to a question which became one of his famous 23 problems for the 20th century published in 1900. Hilbert's 17th problem may be formulated:

- Suppose $f \in \mathbb{R}[\underline{X}]$ is non-negative on $\mathbb{R}^n$. Can f be written as a sum of squares of *rational functions*[5]?

An affirmative answer was given by Artin in 1927. In Chapter 2 we will see that this falls out as a special case of the much more general Krivine-Stengle Positivstellensatz.

[5] A rational function $\phi \in \mathbb{R}(\underline{X})$ is a function of the form $\frac{f}{g}$ where $f, g \in \mathbb{R}[\underline{X}]$ and g is not the zero polynomial.

It is surprising that a concrete example of a non-negative polynomial which was not a sum of squares of polynomials did not appear until 1967, more than 78 years after Hilbert's original proof.

Example 1.11 (Motzkin) The Motzkin polynomial $M \in \mathbb{R}[X,Y]_6$ has the form

$$M = X^4Y^2 + X^2Y^4 - 3X^2Y^2 + 1. \tag{1.14}$$

This is non-negative by the arithmetic-geometric inequality:

$$\frac{a_1 + \cdots + a_n}{n} \geq \sqrt[n]{a_1 \cdots a_n}, \quad \forall a_1, \ldots, a_n \geq 0. \tag{1.15}$$

For $x, y \in \mathbb{R}$ let $a = x^2$ and $b = y^2$ which are non-negative. Then

$$x^4y^2 + x^2y^4 + 1 = a^2b + ab^2 + 1 \geq 3\sqrt[3]{a^2bab^2} = 3ab = 3x^2y^2 \tag{1.16}$$

Thus $M(x,y) \geq 0$ for all $x, y \in \mathbb{R}$.

We will now show that M is not a sum of squares. Suppose for the sake of contradiction that $M = f_1^2 + \cdots + f_m^2$ for some $f_i \in \mathbb{R}[X,Y]$. Then

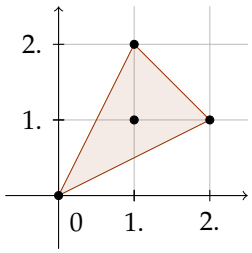

Figure 1.3: $\frac{1}{2}\mathrm{NP}(M)$

$$\mathrm{supp}(f_i) \subseteq \tfrac{1}{2}\mathrm{NP}(M) \cap \mathbb{Z}^2 \subseteq \{(0,0),(1,1),(1,2),(2,1)\}$$

by Theorem 1.8 and we can write $f_i = a_iX^2Y + b_iXY^2 + c_iXY + d_i$ for some $a_i, b_i, c_i, d_i \in \mathbb{R}$. Now observe that the coefficient of X^2Y^2 in $\sum_{i=1}^m f_i^2$ is $\sum_{i=1}^m c_i^2 \geq 0$. This is a contradiction as the coefficient of X^2Y^2 in M is negative. Hence M cannot be a sum of squares.

As a teaser we note that although M is not a sum of squares we have the following relation

$$\begin{aligned}(X^2+Y^2+1)M = {} & (X^2Y - Y)^2 + (XY^2 - X)^2 + (X^2Y^2 - 1)^2 \\ & + \tfrac{1}{4}(XY^3 - X^3Y)^2 + \tfrac{3}{4}(XY^3 + X^3Y - 2XY)^2.\end{aligned} \tag{1.17}$$

This is a purely syntactic and easily verifiable non-negativity certificate for M. For instance it yields (after a little manipulation as done in (2.75)) a representation of M as a sum of squares of rational functions. ○

Although it seems to require superhuman intuition to come up with the relation (1.17), we will later see how this search problem can be efficiently tackled by semidefinite programming (SDP), a very powerful and general tool in convex optimization.

The historical perspectives in this section and the examples on the Motzkin polynomial were based on the exposition [25].

2 Krivine-Stengle theorems

Hilbert's 17th problem which we discussed in last chapter has a natural generalization: Suppose $f \in \mathbb{R}[\underline{X}]$ is non-negative on a subset $S \subseteq \mathbb{R}^n$. Can we then give a certificate of this fact?

To be able to answer this we need some structure on S. We will only consider so called *basic closed semi-algebraic sets*: these take the form

$$\mathcal{W}(B) = \{x \in \mathbb{R}^n \mid g(x) \geq 0, \forall g \in B\} \tag{2.1}$$

where $B = \{g_1, \ldots, g_t\} \subseteq \mathbb{R}[\underline{X}]$ is finite. I.e. $\mathcal{W}(g_1, \ldots, g_t)$ is the common points of non-negativity of $g_1, \ldots, g_t$.

Associated to $\mathcal{W}(B)$ is the set

$$\begin{aligned} T[B] &= \sum_{\nu \in \{0,1\}^t} \Big(\sum \mathbb{R}[\underline{X}]^{(2)}\Big) g_1^{\nu_1} \cdots g_t^{\nu_t} \\ &= \Big\{ \sum_{\nu \in \{0,1\}^t} \sigma_\nu g_1^{\nu_1} \cdots g_t^{\nu_t} \;\Big|\; \sigma_\nu \in \sum \mathbb{R}[\underline{X}]^{(2)}, \forall \nu \in \{0,1\}^t \Big\} \end{aligned}$$

which is called the *preorder of* $\mathbb{R}[\underline{X}]$ *generated by* B. Polynomials in $T[B]$ are "obviously" non-negative on $\mathcal{W}(B)$ being sums of products of polynomials non-negative on $\mathcal{W}(B)$.

An answer to the above question is then given by a famous theorem of Stengle 1974 known as the Positivstellensatz [29]. Krivine is also credited this work. Apparently he proved an abstract version of this result already in 1964 [12], but it was unnoticed before Stengle's paper.

Theorem 2.1 (Positivstellensatz) *Let $S = \mathcal{W}(g_1, \ldots, g_m)$. Then $f \geq 0$ on S if and only if*

$$f^{2m} + p = qf \tag{2.2}$$

for some $m \geq 0$ and

$$p, q \in T[g_1, \ldots, g_m]. \tag{2.3}$$

A relation of the form (2.2) is a trivial sufficient condition for f being non-negative on S and in that way it serves as a *non-negativity certificate*: Suppose $f(x) < 0$ for some $x \in S$. Then the left-hand side is strictly positive since $p(x) \geq 0$ and $f^{2m}(x) = (f^m(x))^2 > 0$ but the right-hand side is non-positive. The hard part of the proof is to show that a non-negativity certificate always exists. After that comes the trouble with actually finding the certificates.

The Positivstellensatz falls out as a consequence of an even more general result known as the Semi-algebraic Nullstellensatz. An other corollary to this is the Real Nullstellensatz (Theorem 2.38) which is in some sense the real algebraic version of Hilbert's Nullstellensatz. In Section 2.4 we will see how the results apply to general systems of polynomial equations and inequalities. In Chapter 4 we will exploit the power of semidefinite programming to efficiently search for the certificates. This also lead to a general algorithm for polynomial optimization described in Chapter 5.

Although we are mostly interested in polynomials over $\mathbb{R}$ and $\mathbb{Q}$ we will need work with some quite abstract field theory to get the main results. In Section 2.1 we will cover some back ground theory on formally real and real closed fields and mention a fundamental result in real algebraic geometry known as the Transfer Principle. After this we will prove a general version of the Positivstellensatz in this abstract setting. This exposition will be based on Stengle's original paper [29].

2.1 Artin-Schreier theory

This section introduces the most basic definitions and results from the theory of real closed and formally real fields, also known as Artin-Schreier theory. We have used [1] and [23] as our main sources. The subsection on the Transer Principle is based on [1] and [17].

Orders and preorders

Definition 2.2 *A* partial order *on a set X is a binary $\leq$ relation satisfying*

$$x \leq y \text{ and } y \leq x \implies x = y, \tag{2.4}$$
$$x \leq y \text{ and } y \leq z \implies x \leq z, \tag{2.5}$$

for all $x, y \in X$. If further

$$x \leq y \text{ or } y \leq x \tag{2.6}$$

for all $x, y \in X$ then $\leq$ is called a total order *on X.*

Let $(X, \leq)$ be an ordered set. An element $x \in X$ is *maximal* if $x \leq y \implies x = y$ for all $y \in X$. An upper bound on a subset $Y \subseteq X$ is an element $z \in X$ such that $y \leq z$ for all $y \in Y$. A *chain* in $(X, \leq)$ is a totally ordered subset of $(X, \leq)$.

Quite a few results in the rest of this chapter depend on Zorn's Lemma.

Theorem 2.3 (Zorn's Lemma) *Let $(X, \leq)$ be a partially ordered set with the property that every chain in $(X, \leq)$ has an upper bound in S. Then $(X, \leq)$ has a maximal element.*

As Zorn's Lemma is equivalent to the Axiom of Choice under the Zermelo-Frankel axioms of set theory we will draw a line in the sand and omit the proof. We will always use the lemma in connection with families of subsets with an argument along the following lines: Our partially ordered set will be a family $\mathcal{F} = \{F \subseteq X \mid P(F)\} \subseteq 2^X$ of subsets $F \subseteq X$ satisfying some property $P(F)$ and the ordering on $\mathcal{F}$ will be inclusion. We will always construct $\mathcal{F}$ such that $\mathcal{F} \neq \emptyset$ and then show that $\bigcup_{F \in \mathcal{F}'} F \in \mathcal{F}$ for any chain in $\mathcal{F}$. Then by Zorn's Lemma $\mathcal{F}$ has a maximal element F', i.e. a subset $F' \subseteq X$ which satisfies P and has the property that if $F' \subsetneq F''$ then F'' does not satisfy P.

Definition 2.4 *An* ordered ring $(C, \leq)$ *is a commutative ring C together with a total order $\leq$ that satisfies*

$$x \leq y \implies x + z \leq y + z, \tag{2.7a}$$
$$0 \leq x \text{ and } 0 \leq y \implies 0 \leq xy, \tag{2.7b}$$

for all $x, y, z \in K$.

An ordered field $(K, \leq)$ *is an ordered ring where K is a field.*

In an ordered ring $(C, \leq)$ we write $a < b$ if $a \leq b$ and $a \neq b$ and we write $a \geq b$ (reps. $a > b$) if $b \leq a$ (resp. $b < a$). Sometimes when different ordered rings are in play we will write $\leq_C$ to avoid confusion.

Definition 2.5 *A* preorder $T \subseteq C$ *is a subset satisfying*

$$t_1, t_2 \in T \implies t_1 + t_2 \in T, \tag{2.8a}$$
$$t_1, t_2 \in T \implies t_1 t_2 \in T, \tag{2.8b}$$
$$c \in C \implies c^2 \in T. \tag{2.8c}$$

This is sometimes referred to as a cone.

A subset $P \subseteq C$ is called a proper cone[1] *if P is a preorder and $-1 \notin P$.*

[1] this is not to be confused with a proper cone in a vector space ...

The set $\sum C^{(2)}$ of sums of squares is a preorder. It is not necessarily a proper cone. For instance if $C = \mathbb{C}$ then $-1 = i^2 \in \sum \mathbb{C}^{(2)}$. Any preorder in C contains the sum of squares $\sum C^{(2)}$ which is in that sense the smallest preorder in C.

If $B \subseteq C$ then we denote by $T[B]$ the preorder *generated by* B. This is the smallest preorder in C containing B. For instance $T[\emptyset] = \sum C^{(2)}$. We are especially interested in *finitely generated preorders*, i.e. preoders of the form $T[g_1, \ldots, g_m]$ where $g_1, \ldots, g_t \in C$.

Lemma 2.6 *Let $g_1 \ldots, g_{t+1} \in C$. Then*

$$T[g_1, \ldots, g_{t+1}] = T[g_1, \ldots, g_t] + g_{t+1} T[g_1, \ldots, g_t]. \tag{2.9}$$

Proof $\supseteq$: clear since $T[g_1, \ldots, g_{t+1}]$ is closed under addition and multiplication.

$\subseteq$: It is clear that $T[g_1, \ldots, g_t] + g_{t+1} T[g_1, \ldots, g_t]$ contains $1, g_1, \ldots, g_{t+1}$, and c^2 for any $c \in C$. As $T[g_1, \ldots, g_{t+1}]$ is the smallest preorder containing $g_1, \ldots, g_{t+1}$ it remains to show that $T[g_1, \ldots, g_t] + g_{t+1} T[g_1, \ldots, g_t]$ is closed under addition and multiplication. Let $t_1, t_2, t_3, t_4 \in T[g_1, \ldots, g_t]$. Then

$$(t_1 + g_{t+1} t_2) + (t_3 + g_{t+1} t_4) = (t_1 + t_3) + g_{t+1}(t_2 + t_4), \tag{2.10}$$

$$(t_1 + g_{t+1} t_2)(t_3 + g_{t+1} t_4) = (t_1 t_3 + g_{t+1}^2 t_2 t_4) + g_{t+1}(t_1 t_4 + t_2 t_3). \tag{2.11}$$

□

Lemma 2.7 *Let $g_1, \ldots, g_t \in C$. Then*

$$T[g_1, \ldots, g_t] = \sum_{\nu \in \{0,1\}^t} \sum C^{(2)} g_1^{\nu_1} \cdots g_t^{\nu_t}. \tag{2.12}$$

That is, the elements of $T[g_1, \ldots, g_t]$ are of the form

$$\sum_{\nu \in \{0,1\}^t} (\sum_j c_{j\nu}^2) g_1^{\nu_1} \cdots g_m^{\nu_m} \tag{2.13}$$

where $c_{j\nu} \in C$.

Proof Since $T[\emptyset] = \sum C^{(2)}$ the result follows directly from Lemma 2.6 by induction in t. □

Let $(K, \leq)$ be an ordered field. The order relation $\leq$ gives rise to a set $K^+ = \{x \in K \mid x \geq 0\}$ wich is called the *positive cone* of $(K, \leq)$.

Remark 2.8 K^+ is a proper cone which satisfies $K = K^+ \cup -K^+$. It completely describes the order relation $\leq$ since $x \leq y$ iff $0 \leq y - x$ iff $y - x \in K^+$. Conversely any proper cone $P \subseteq K$ with $K = P \cup -P$ gives rise to an order on K by declaring $x \leq y$ iff $y - x \in P$. In that way a proper cone $P \subseteq K$ with $K = P \cup -P$ is sometimes simply called an *ordering*. △

Proposition 2.9 *Let P be a proper cone in a field F. Then F can be given an ordering such that $P \subseteq F^+$.*

Proof See [1, Prop. 2.14]. □

Formally real fields

Definition 2.10 *A field F is said to be* formally real *if*

$$\sum a_i^2 = 0 \implies a_i = 0 \tag{2.14}$$

for any $a_i \in F$.

The next theorem gives a characterization of formally real fields:

Theorem 2.11 *The following conditions are equivalent:*

i) F is formally real.

ii) $-1 \notin \sum F^{(2)}$.

iii) F can be ordered.

Proof See [1, Thm. 2.13] and the proof of Theorem 2.15 below. □

Example 2.12 $\mathbb{Q}$ and $\mathbb{R}$ are examples of formally real fields. They are ordered fields with their usual orderings. $\mathbb{C}$ however is not formally real since $-1 = i^2 \in \sum \mathbb{C}^{(2)}$. This reflects that $\mathbb{C}$ can not be ordered as a field. ○

A formally real field does not in general have a uniqe ordering. This allows for some freedom which is crucial in the proof of the Possitivstellensatz (see Lemma 2.42).

Example 2.13 The field $K = \mathbb{Q}(\sqrt{2})$ can be ordered in two ways:

We can declare $x \leq_K y$ iff $x \leq y$ with the ususal orderig on $\mathbb{R}$. This is clearly an ordering on K.

For any $x = a + \sqrt{2}b \in K$ we write $\overline{x} = a - \sqrt{2}b$ and we can declare $x \leq_K y$ iff $\overline{x} \leq \overline{y}$ in $\mathbb{R}$. This is also a field ordering: It is clearly a total ordering. One can check that $\bar{x} + \bar{y} = \overline{x + y}$ and $\bar{x}\bar{y} = \overline{xy}$.

Suppose $x, y, z \in K$. Then

$$\begin{aligned} x \leq_K y &\implies \overline{x} \leq \overline{y} \\ &\implies \overline{x+z} = \overline{x} + \overline{z} \leq \overline{y} + \overline{z} = \overline{y+z} \\ &\implies x + z \leq_K y + z. \end{aligned}$$

Likewise

$$\begin{aligned} 0 \leq_K x \text{ and } 0 \leq_K y &\implies 0 \leq \overline{x} \text{ and } 0 \leq \overline{y} && (2.15) \\ &\implies 0 \leq \overline{x}\overline{y} = \overline{xy} && (2.16) \\ &\implies 0 \leq_K xy. && (2.17) \end{aligned}$$

○

Let $(K_1, \leq_1)$ and $(K_2, \leq_2)$ be ordered fields. We say that $(K_2, \leq_2)$ extends $(K_1, \leq_1)$ if $K_1 \subseteq K_2$ and $K_1^+ \subseteq K_2^+$. If $(K_2, \leq_2)$ extends $(K_1, \leq_1)$ then the orders agree in the sense that

$$\begin{aligned} x \leq_1 y &\implies y - x \in K_1^+ \subseteq K_2^+ && (2.18) \\ &\implies x \leq_2 y. && (2.19) \end{aligned}$$

Sometimes we are interested in extending an ordered field to a bigger one. This motivates the following definition:

Definition 2.14 *Let $(K, \leq)$ be an ordered field and F be a field extending K. F is* formally real over K *if*

$$\sum_{i=1}^{N} r_i a_i^2 = 0 \implies a_1 = \cdots = a_N = 0 \tag{2.20}$$

for any $r_i \in K^+ \setminus \{0\}$ and $a_i \in F$, $i = 1, \ldots, N$.

In [29] it is noted but not proved that such a field can be given an ordering extending the ordering on K, i.e. such that $K^+ \subseteq F^+$. We state and prove this in the following lemma which is just a small adjustment of Theorem 2.11.

Theorem 2.15 *The following is equivalent:*

i) F is formally real over K.

ii) $-1 \notin \sum K^+ F^{(2)}$.

iii) F admits an ordering extending the ordering on K.

Proof $i) \Rightarrow ii)$: Suppose for the sake of contradiction that $-1 \in \sum K^+F^{(2)}$. Then we get

$$-1 = \sum_{i=1}^{N} r_i a_i^2 \implies 0 = 1^2 + \sum_{i=1}^{N} r_i a_i^2 \overset{(2.20)}{\implies} 1 = 0. \tag{2.21}$$

This is not possible in a field.

$ii) \Rightarrow iii)$: As $\sum K^+F^{(2)}$ is a proper cone Proposition 2.9 implies that there is an ordering on F such that $\sum K^+F^{(2)} \subseteq F^+$. But we also have $K^+ \subseteq \sum K^+F^{(2)}$ so the ordering on F extends the ordering on K.

$iii) \Rightarrow i)$: Let $a_i \in F$ and $r_i \in K^+\backslash\{0\}$, $i = 1, \ldots, N$. Suppose $a_j \neq 0$ for some $j \in \{1, \ldots, N\}$. We then get $0 < r_j a_j^2 \leq \sum_{i=1}^{N} r_i a_i^2$. Thus F is formally real over K. □

Example 2.16 $\mathbb{Q}(\sqrt{2})$ is clearly formally real over $\mathbb{Q}$. Both orderings considered in Example 2.13 extend the ordering on $\mathbb{Q}$. Other formally real extensions of $\mathbb{Q}$ include the real algebraic numbers[2] $\mathbb{R}_{\text{alg}}$ and the real numbers. ○

Real closed fields and real closures

The next notion is central and encapsulates most of the properties of the real numbers.

Definition 2.17 *A field R is said to be* real closed *if it is formally real and has no non-trivial algebraic*[3] *extension which is also formally real.*

The real closed fields has the following characterization [1, Thm. 2.17]:

Theorem 2.18 *R is real closed if and only if $R^+ = R^{(2)}$ and any polynomial $p \in R[X]$ of odd degree has a root in R.*

In particular we get

Corollary 2.19 *Any real closed field has a unique ordering.*

Proof By Remark 2.8 any order is determined by it positive cone. By Theorem 2.18 the positive cone is uniquely given as the set of squares in R. □

[2]The set of real numbers which are roots of polynomials with rational cefficients.

[3]A field extension $F_1 \subseteq F_2$ is algebraic if any element in F_2 is a root of a polynomial with coefficients in F_1.

Example 2.20 $\mathbb{R}$ and the real algebraic numbers $\mathbb{R}_{\text{alg}}$ are examples of real closed fields. $\mathbb{Q}$ is not real closed since $2 > 0$ is not a square in $\mathbb{Q}$. Also $\sqrt[3]{2} \notin \mathbb{Q}$ is the only root of $p(X) = X^3 - 2$. Likewise $\mathbb{Q}(\sqrt{2})$ is not real closed. This also follows from the fact that it admits multiple orderings. ○

We will next show that any formally real field can be extended to a real closed field.

Proposition 2.21 *Let F be a formally real field and Ω the algebraic closure of F. Then there is a real closed field R such that $F \subseteq R \subseteq \Omega$.*

Proof Let $\mathcal{F}$ be the family of all formally real fields contained in Ω and containing F. Clearly $F \in \mathcal{F}$ and the union of a chain in $\mathcal{F}$ is again in $\mathcal{F}$. Hence Zorns Lemma yields a maximal element R. We claim that R has no non-trivial algebraic extension which is formally real: If $R' \supset R$ is a proper algebraic field extension then it is contained in Ω. Maximality of R implies that $R' \notin \mathcal{F}$ and so R' cannot be formally real. Thus R is real closed. □

Example 2.20 illustrated that a formally real field is not neccesarily contained in a unique real closed field. We want a notion of the smallest real closed field extension (see Proposition 2.26).

Definition 2.22 *Let $(K, \leq_K)$ be an ordered field and R a field containg K. R is the* real closure of *K with respect to $\leq_K$ if*

i) R is real closed.

ii) R is algebraic over K.

iii) The ordering $\leq_R$ on R extends the ordering $\leq_K$ on K (i.e. any $r \geq_K 0$ in K is a square in R).

Example 2.23 $\mathbb{R}$ is not a real closure of $\mathbb{Q}$ although it extends the ordering on $\mathbb{Q}$. It fails to be algebraic over $\mathbb{Q}$ having trancendental numbers such as π and e. $\mathbb{R}_{\text{alg}}$ being algebraic over $\mathbb{Q}$ is indeed a real closure of $\mathbb{Q}$. ○

It turns out that any ordered field has a unique real closure.

Theorem 2.24 *Any ordered field $(K, \leq_K)$ has a unique[4] real closure R with respect to $\leq_K$.*

[4]up to isomorphism. I.e. if $(K_1, \leq_1)$ and $(K_2, \leq_2)$ are ordered fields with real closures R_1 and R_2 then any order isomorphism $\phi\colon K_1 \to K_2$ extends to an isomorphism $\bar{\phi}\colon R_1 \to R_2$.

We will only prove existence. We will follow the proof in [23, Thm. 15.8] and strengthen the conclusion a bit from "formally real" to "formally real over K" in our opinion making the induction step a bit more clear.

Lemma 2.25 *Let K be an ordered field, Ω an algebraic closure of K and E the algebraic field extension of K generated by the square roots $\sqrt{r}$ of positive elements $r > 0$.*

Then E is formally real over K.

Proof E is clearly a subfield of Ω. Assume $\sum r_i' a_i^2 = 0$ for some (finitely many) $a_i \in E$ and $r_i' > 0$. Then some finite extension $K(\sqrt{r_1}, \ldots, \sqrt{r_m})$ contains all the a_i's. It now suffices to prove that $K(\sqrt{r_1}, \ldots, \sqrt{r_m})$ is formally real over K, since then $a_i = 0$ for all i as wanted.

We proceed with induction in $m > 0$. Assume that $a_i \in K(\sqrt{r})$ for all i and some $r > 0$. We can then write $a_i = \alpha_i + \sqrt{r}\beta_i$. Now

$$0 = \sum r_i' a_i^2 = \sum r_i'(\alpha_i^2 + r\beta_i^2) + \sqrt{r}\sum 2r_i'\alpha_i\beta_i \tag{2.22}$$

implies $\sum r_i'(\alpha_i^2 + r\beta_i^2) = 0$ and, since K is formally real and $r_i' \geq 0$ we get that $\alpha_i^2 + r\beta_i^2 = 0$ and hence $\alpha_i = \beta_i = 0$ for all i. I.e. $a_i = 0$ for all i as wanted.

Let $r_1, \ldots, r_m >_K$ and assume that $\widetilde{K} = K(\sqrt{r_1}, \ldots, \sqrt{r_{m-1}})$ is formally real over K. Then $K(\sqrt{r_1}, \ldots, \sqrt{r_m}) = \widetilde{K}(\sqrt{r_m})$. Equip $\widetilde{K}$ with an ordering $\leq_{\widetilde{K}}$ that extends $\leq_K$. Then $r_m >_{\widetilde{K}} 0$ so the argument from above implies that $\widetilde{K}(\sqrt{r_m})$ is formally real over $\widetilde{K}$ and hence over K. □

Now for the proof of existence in Theorem 2.24:

Proof Let Ω and E be as in Lemma 2.25. As E is formally real and Ω is its algebraic closure (being the algebraic closure of K and hence of any algeraic extension of K), Proposition 2.21 yields a real closed field R such that $E \subseteq R \subseteq \Omega$.

We claim that R is a real closure of K. R is clearly algebraic over K being a subset of the algebraic closure Ω. To show that the ordering on R extends the ordering on K suppose $r \geq_K 0$. Then $\sqrt{r} \in E \subseteq R$ and hence $r = (\sqrt{r})^2 \geq_R 0$.□

The following result illustrates that the real closure of $(K, \leq)$ is indeed the smallest real closed field extending K. This seems to be a well known result, but we could not find it in the litterature, so the statement and proof is our own work. We will use it get a version of the Transfer Principle which is needen in the proof of Proposition 2.41.

Proposition 2.26 *Let $(K, \leq_K)$ be an ordered field and R its real closure. Then any real closed field R' extending $(K, \leq_K)$ also extends R. I.e. $R \subseteq R'$ or more precisely; R is isomorphic to a subset of R'.*

Proof We use Zorn's lemma to construct R: Consider the family

$$\mathcal{F} = \{F \mid K \subseteq F \subseteq R' \text{ and } F \text{ is an algebraic field extension of } K\}. \tag{2.23}$$

As $K \in \mathcal{F} \neq \emptyset$ and the union fo a chain in $\mathcal{F}$ is again in $\mathcal{F}$, we can choose a maximal element $R \in \mathcal{F}$. We claim that R is a real closure of $(K, \leq)$ contained in R'.

By construction $K \subseteq R \subseteq R'$ and R is algebraic over K.

We can equip R with the ordering of R' by declaring $x \leq_R y$ iff $x \leq_{R'} y$ for all $x, y \in \mathbb{R}$. As $R^+ = (R')^+ \cap R \supseteq K^+$ we see that $\leq_R$ extends $\leq_K$.

It only remains to show that R is real closed. For this we use the characterization Theorem 2.18. Suppose $r \geq_R 0$. Then $\sqrt{r} \in R'$ since $r \geq_{R'} 0$ and R' is real closed. Assume for the sake of contradiction that $\sqrt{r} \notin R$. Then $R(\sqrt{r}) \subseteq R'$ is a proper field extension of R. It is also algebraic over R since $\sqrt{r}$ is a root of $r - X^2 \in R[X]$. But then since R is algebraic over K transitivity of algebraic extensions ([11, Prop. 2.1.9]) implies that $R(\sqrt{r})$ is also algebraic over K. This contradicts maximality of R in $\mathcal{F}$. Hence $R^+ = R^{(2)}$.

Similarly if $p \in R[X]$ is a univariate polynomial of odd degree, then p has a root ζ in R'. If $\zeta \notin R$ then by same arguments as for $\sqrt{r}$ above $R(\zeta)$ will contradict minimality of R in $\mathcal{F}$. Hence $\zeta \in R$.

By Theorem 2.18 R is real closed and hence a real closure of K. By uniqueness of real closures any real closure of K is isomorphic to R. □

The Transfer Principle

One of the main reasons for generalizing to arbitrary real closed fields is that it allows one to find solutions to equations in a very useful way: Instead of searching for solutions in our ground field we can search in a bigger field, and if we can find a solution here, we get by pure magic the existence of a solution in the original field. This pure magic is known as the Transfer Principle. It is a deep result in the model theory of real closed fields and a consequence of a result known as the Tarski-Seidenberg Therorem, which states that any first order fomula in the language of real closed fields is equivalent to a quantifier free formula.

To clearify with an example, consider the first order formula

$$\exists x \in \mathbb{R}: \quad ax^2 + bx + c = 0 \tag{2.24}$$

with coefficients a, b, c in $\mathbb{Q}$ (or in some other ordered field $(K, \leq)$). This formula is equivalent to

$$(c = 0) \vee (a = 0 \wedge b \neq 0) \vee (a \neq 0 \wedge b^2 - 4ac \geq 0) \tag{2.25}$$

which is quantifier free. The derivation of such quntifier free formula can be done algorithmically in a way that only depends on the coefficients. In that way the formula remains true if $\mathbb{R}$ is exchanged by any real closed field extending $\mathbb{Q}$ (or $(K, \leq)$ in general). We will not go further into this subject but direct the reader to [1, Chapter 2] and [17, Appendix 1].

The following version of the Transfer Principle (which is a version found in [17]) is a special case of [1, Thm. 2.98] which avoids notions of first order logic:

Theorem 2.27 (Transfer Principle) *Suppose R_1 is a real closed field contained in the real closed field R_2 and consider the polynomial system*

$$\bigwedge_{i=1}^{m} f_i(\underline{X}) \rhd_i 0 \tag{2.26}$$

where $\rhd_i \in \{\geq, >, =, \neq\}$ and $f_i \in R[\underline{X}]$, $i = 1, \ldots, m$. Then (2.26) *has a solution in R_1^n if and only if* (2.26) *has a solution in R_2^n.*

Any solution to (2.26) in R_1^n is clearly also a soluiton in R_2^n. It is the other implication that is interesting: if we can find a solution to (2.26) in the potientially much larger space R_2^n then the Transfer Principle guarantees a solution in R_1^n as well but the theorem tells nothing about how this solution in R_1^n is related to the solution in R_2^n.

Using Proposition 2.26 we can get an other version which suits our purposes better (this version is also found in [17] but we have elaborated the proof with Proposition 2.26):

Theorem 2.28 (Transfer Principle) *Let R_1 and R_2 be real closed fields extending an ordered field $(K, \leq)$. Then the system*

$$\bigwedge_{i=1}^{m} f_i(\underline{X}) \rhd_i 0 \tag{2.27}$$

where $\rhd_i \in \{\geq, >, =, \neq\}$ and $f_i \in K[\underline{X}]$, $i = 1, \ldots, m$, has a solution in R_1^n if and only if it has a solution in R_2^n.

Proof Let R be the real closure of K with respect to $\leq_K$. By Proposition 2.26 R is (isomorphic to) a subset of both R_1 and R_2. By Theorem 2.27 the system (2.27) has a solution in R_1 iff it has a solution in R iff it has a solution in R_2. □

We will only need this specific version in the proof of Proposition 2.41.

Corollary 2.29 *Let $(K_1, \leq_1)$ and $(K_2, \leq_2)$ be ordered fields and R be real closed such that R and $(K_2, \leq_2)$ both extend $(K_1, \leq_2)$. Consider the system*

$$\bigwedge_{i=1}^{m} f_i(\underline{X}) \rhd_i 0 \tag{2.28}$$

where $\rhd_i \in \{\geq, >, =, \neq\}$ and $f_i \in K_1[\underline{X}]$, $i = 1, \ldots, m$. If (2.28) *has a solution in K_2^n then* (2.28) *also has a solution in R^n.*

Proof Let R' be the real closure of $(K_2, \leq_2)$ and apply Theorem 2.28. □

2.2 Semi-algebraic sets and their associated algebraic objects

In the rest of this chapter, unless anything else is specified, $(K, \leq)$ denotes an ordered field (e.g. $\mathbb{Q}$ with the usual ordering) and R a real closed field (not necessarily a real closure) extending $(K, \leq)$ (e.g. $\mathbb{R}$).

In (complex) algebraic geometry the objects of study are mainly polynomial ideals and varieties. In real algebraic geometry we work over an ordered field (which the complex numbers are not) and this gives rise to other geometric objects defined by polynomial inequalities – these are the *semi-algebraic sets*.

We define the objects of interest below:

Definition 2.30 *Let $X \subseteq R^n$. We define the* vanishing ideal *of X as*

$$\mathbb{I}(X) = \mathbb{I}_K(X) = \{p \in K[\underline{X}] \mid p(x) = 0, \forall x \in X\},$$

the ideal of all polynomials vanishing on X. We also define

$$\mathcal{A}(X) = \mathcal{A}_K(X) = \{p \in K[\underline{X}] \mid p(x) \geq 0, \forall x \in X\},$$

the subset of all polynomials non-negative on X.[5]

Let $B \subseteq K[\underline{X}]$. We then define the real variety *of B as*

$$\mathbb{V}(B) = \mathbb{V}_R(B) = \{x \in R^n \mid p(x) = 0, \forall p \in B\}$$

which is the common zeros of polynomials in B. And we define

$$\mathcal{W}(B) = \mathcal{W}_R(B) = \{x \in R^n \mid p(x) \geq 0, \forall p \in B\}, \tag{2.29}$$

the common points of non-negativity of polynomials in B.

[5]Note that this is a preorder in the polynomial ring $K[\underline{X}]$.

When $B = \{g_1, \ldots, g_t\}$ is finite the set $S = \mathcal{W}(B) = \mathcal{W}(g_1, \ldots, g_t)$ is called a *basic closed semi-algebraic set*.

In line with the introduction of this chapter we associate to (the defining polynomials of) a basic closed semi-algebraic set $S = \mathcal{W}(B)$ a certain set of polynomials which are obviously non-negative on S. This set is the preorder (see Definition 2.5)

$$T = T_+[B] := T[K^+ \cup B], \tag{2.30}$$

generated by K^+ (the non-negative elements in K) and $g_1, \ldots, g_m$.

By minor adjustments of the proof of Lemma 2.7 we have

$$T_+[B] = \sum_{\nu \in \{0,1\}^t} \big(\sum K^+ K[\underline{X}]^{(2)} \big) g_1^{\nu_1} \cdots g_m^{\nu_m}. \tag{2.31}$$

I.e any element $g \in T_+[B]$ can be written in the form

$$g = \sum_{\nu \in \{0,1\}^t} \big(\sum_j r_{j\nu} a_{j\nu}^2 \big) g_1^{\nu_1} \cdots g_m^{\nu_m} \tag{2.32}$$

where $r_{j\nu} \in K^+$ and $a_{j\nu} \in K[\underline{X}]$.

Remark 2.31 We note that when $K = R$ is real closed then $T_+[B] = T[B]$ because $R^+ = R^{(2)}$ by Theorem 2.18 so that $R^+ R[\underline{X}]^{(2)} = R[\underline{X}]^{(2)}$. This is not true in general. E.g. $2x^2 \in \mathbb{Q}^+\mathbb{Q}[\underline{X}]^+ \backslash \mathbb{Q}[\underline{X}]^{(2)}$ as $\sqrt{2} \notin \mathbb{Q}$. △

The polynomials in $T_+[B]$, being sums and products of non-negative polynomials, are clearly non-negative on S. The expressions of the form (2.32) serve as syntactic certificates of this fact. However not every polynomial non-negative on S can be expected to admit such a certificate, i.e. in general $T_+[B] \subsetneq \mathcal{A}(\mathcal{W}(B))$ is a strict subset. Also $T_+[B]$ does not depend on the geometry of $S = \mathcal{W}(B)$ alone but also on the defining polynomials B. This is illustrated in example below.

Example 2.32 We have seen that the Motzkin polynomial $M = X^4Y^2 + X^2Y^4 - 3X^2Y^2 + 1 \in \mathbb{R}[X, Y]$ is globally non-negative although it is not a sum of squares. Hence

$$\mathcal{W}(\emptyset) = \mathcal{W}(M) = \mathbb{R}^2,$$

while

$$\begin{aligned} T[\emptyset] &= \underbrace{\sum \mathbb{R}[X, Y]^{(2)}}_{M \notin} \\ &\subsetneq \underbrace{\sum \mathbb{R}[X, Y]^{(2)} + M \sum \mathbb{R}[X, Y]^{(2)}}_{M \in} = T[M] \\ &\subseteq \mathcal{A}(\mathbb{R}^2) = \mathcal{A}(\mathcal{W}(\emptyset)). \end{aligned}$$

○

The natural question is now: How can we represent the elements in $\mathcal{A}(\mathcal{W}(B))$ in a way that certifies non-negativity – this will be answered by the Positivstellensatz Theorem 2.48.

We need to introduce two important algebraic object before we move on.

Definition 2.33 *Let $I \subseteq C$ be an ideal in a commutative ring*[6] *C.*
The radical of *I is defined as*

$$\sqrt{I} = \{f \in C \mid f^m \in I \text{ for some } m > 0\} \tag{2.33}$$

I is said to be a radical ideal *if $I = \sqrt{I}$ is its own radical.*
*Let $T \subseteq C$ be a preorder in C. The T-*radical of *I is defined as*

$$\sqrt[T]{I} = \{f \in C \mid f^{2m} + a \in I \text{ for some } m > 0 \text{ and } a \in T\}. \tag{2.34}$$

*I is a T-*radical ideal *if $I = \sqrt[T]{I}$.*

T-radical ideals are in particular radical: Since if $f^m \in I$ for some $m > 0$ then $f^{2m} + 0 \in I$ so $f \in \sqrt[T]{I} = I$.

Lemma 2.34 *Let $I \subseteq C$ be an ideal in a commutative ring and $T \subseteq C$ a preorder in C. The following holds:*

i) $\sqrt{I}$ *is a radical ideal [5, Lem. 4.2.5].*

ii) $\sqrt[T]{I}$ *is a T-radical ideal [29, Lem. 1].*

The proofs does not use anything but smart algebraic manipulations (the proof that $\sqrt[T]{I}$ is closed under addition is particularly smart).

2.3 The Real- and Semi-algebraic Nullstellensatz

A fundamental result real agebraic geometry which we will prove in the next subsection is the so-called Semi-algebraic Nullstellensatz [29, Thm. 1] which describes the connection between a basic closed semi-algebraic set $S = \mathcal{W}(B)$ and the associated T-radical ideals $\sqrt[T]{I}$, where $T = T_+[B]$.

Theorem 2.35 (Semi-algebraic Nullstellensatz) *Let $I \subseteq K[\underline{X}]$ be an ideal, $S = \mathcal{W}(B)$ a basic closed semi-algebraic set and $T = T_+[B]$ the associated preorder. Then*

$$\mathbb{I}(\mathbb{V}_R(I) \cap S) = \sqrt[T]{I}. \tag{2.35}$$

[6] In any context here this will be a polynomial ring.

I.e. $\sqrt[T]{I}$ consists of the polynomials in $K[\underline{X}]$ which vanish on the intersection $\mathbb{V}_R(I) \cap S$. It is easy to see that any polynomial in $\sqrt[T]{I}$ vanish on $\mathbb{V}_R(I) \cap S$ but the other way around is not at all clear.

Remark 2.36 Recall that R was an arbitrary real closed field extending $(K, \leq_K)$. By definition the T-radical $\sqrt[T]{I}$ does not depend on this choice. It might seem strange that this is also the case for $\mathbb{I}(\mathbb{V}_R(I) \cap S)$, but in fact this is justified by the Transfer Principle: Suppose R' is another real closed extension of $(K, \leq)$. Write $I = \langle f_1, \ldots, f_s \rangle$. Then $h \in \mathbb{I}(\mathbb{V}_R(I) \cap S) \subseteq K[\underline{X}]$ if and only if the system

$$\begin{aligned} f_i &= 0, \quad i = 1, \ldots, s, \\ g_j &\geq 0, \quad j = 1, \ldots, t, \\ h &\neq 0, \end{aligned} \tag{2.36}$$

has *no* solution in R^n. By Theorem 2.28 (2.36) has a solution in R^n if and only (2.36) has a solution in $(R')^n$. △

Remark 2.37 As illustrated in Example 2.32 the preorder $T_+[B]$ associated to $S = \mathcal{W}(B)$ generally depends on the defining polynomials B and not on the geometry of S alone. The associated T-radical ideals however only depend on the geometry as the Semi-algebraic Nullstellensatz illustrates. Some authors thus prefer the notation $\sqrt[S]{I}$ over $\sqrt[T]{I}$. △

As a consequence of Theorem 2.35 we get the Real Nullstellensatz [29, Thm. 2]:

Theorem 2.38 (Real Nulstellensatz) *Let $I \subseteq K[\underline{X}]$ be an ideal. Then*

$$\mathbb{I}(\mathbb{V}_R(I)) = \sqrt[R]{I} \tag{2.37}$$

where $\sqrt[R]{I}$ is the real radical of I *given by*

$$\sqrt[R]{I} := \left\{ f \in K[\underline{X}] \;\middle|\; f^{2m} + \sum r_i g_i^2 \in I,\ r_i \geq_K 0,\ g_i \in K[\underline{X}],\ m > 0 \right\}. \tag{2.38}$$

Proof Observe that $\sqrt[R]{I} = \sqrt[T]{I}$ where $T = T_+[\emptyset] = \sum K^+ K[\underline{X}]^{(2)}$ and that $\mathcal{W}(\emptyset) = R^n$ so that $\mathbb{V}_R(I) = \mathbb{V}_R(I) \cap \mathcal{W}(\emptyset)$. Then apply Theorem 2.35. □

To put Theorem 2.38 more explicitly assume that f is a polynomial which vanish on $\mathbb{V}_R(I)$ where $I = \langle f_1, \ldots, f_s \rangle$. Then the Real Nullstellensatz guaranties a relation of the form

$$f^{2m} + \sum r_i g_i^2 = \sum h_i f_i \tag{2.39}$$

for some $m > 0$, $r_i >_K 0$ and $g_i, h_i \in K[\underline{X}]$. This relation certifies that $f \in \mathbb{I}(\mathbb{V}(I))$; assume namely that (2.39) holds. Then for any $x \in \mathbb{V}(I)$ we have

$$f^{2m}(x) + \sum r_i g_i^2(x) = \sum h_i(x) f_i(x) = 0 \implies f^m(x) = 0 \implies f(x) = 0,$$

using the fact that R is formally real over K and that any field is an integral domain.

The Real Nullstellensatz is in some sense the real algebraic version of Hilbert's (strong) Nullstellensatz [5, Thm. 4.2.1] from algebraic geometry:

Theorem 2.39 (Hilbert's Nullstellensatz) *Let $I \subseteq \mathbb{C}[\underline{X}]$ and $\mathbb{V}_{\mathbb{C}}(I) \subseteq \mathbb{C}^n$ be the corresponding* complex *variety. Then*

$$\mathbb{I}(\mathbb{V}_{\mathbb{C}}(I)) = \sqrt{I} \tag{2.40}$$

where $\sqrt{I}$ is the radical of I.

The situation is quite nicer in the complex case as the radical ideal $\sqrt{I}$ yields much simpler certificates. If f vanish on $\mathbb{V}_{\mathbb{C}}(I)$ where $I = \langle f_1, \ldots, f_s \rangle \subseteq \mathbb{C}[\underline{X}]$, then the relation from Hilbert's Nullstellensatz is simply

$$f^m = \sum h_i f_i \tag{2.41}$$

for some $m > 0$ and $h_i \in \mathbb{C}[\underline{X}]$.

Example 2.40 Let $I = \langle X^2 + 1 \rangle \subseteq \mathbb{R}[\underline{X}]$. Then $\mathbb{V}_{\mathbb{R}}(I) = \emptyset$ so $\mathbb{I}(\mathbb{V}_{\mathbb{R}}(I)) = \mathbb{R}[X] \ni X$. But $X^m \notin I$ for any $m > 0$. ○

Proof of the Semi-algebraic Nullstellensatz

In this section we prove Theorem 2.35 based on the original article of Stengle [29]. We have tried to be as explicit as possible especially concerning the arguments involving Artin-Schreier theory which we have introduced in Section 2.1.

The proof the right inclusion in Theorem 2.35 is the hard part. We will prove this in the special case when the ideal I is *prime*[7] and T-radical.

Proposition 2.41 *Let J be a prime T-radical ideal, $S = \mathcal{W}(B)$, and $T = T_+[B]$ where $B = \{g_1, \ldots, g_t\} \subseteq K[X]$. Then*

$$\mathbb{I}(\mathbb{V}(I) \cap S) \subseteq \sqrt[T]{J} = J \tag{2.42}$$

[7] I is prime if $ab \in I \implies a \in I \vee b \in I$ for all a, b in the containing ring.

The proof is a bit technical and we will need some preliminary work.

In the following we let $\iota : p \mapsto \overline{p}$ be the quotient map from $K[\underline{X}]$ to the quotient ring $\Gamma := K[\underline{X}]/J$. Γ is an integral domain as J is prime[8] so we may consider the field of fractions of Γ:

$$Q = \left\{ \frac{\bar{p}}{\bar{s}} \;\middle|\; \bar{p}, \bar{s} \in \Gamma, \bar{s} \neq 0 \right\} = \left\{ \frac{\bar{p}}{\bar{s}} \;\middle|\; p, s \in K[\underline{X}], s \notin J \right\}. \tag{2.43}$$

We then have the following embeddings

$$K \overset{\iota}{\hookrightarrow} \Gamma \overset{\tau}{\hookrightarrow} Q \tag{2.44}$$

where the first embedding is just the inclusion and $\tau(\bar{p}) = \frac{\bar{p}}{\bar{1}}$. Clearly ι and τ are homomorphisms. They are also injective: If $r \in K \backslash \{0\}$ then $\iota(r) = \bar{r} \neq 0$ since otherwise $r \in J$ and then $1 = r^{-1}r \in J$ which contradicts J being prime. Hence $\ker \iota = \{0\}$ and ι is injective. τ is injective since $\frac{\bar{p}}{\bar{1}} = \frac{\bar{q}}{\bar{1}}$ iff $\bar{p} = \bar{1}\bar{p} = \bar{1}\bar{q} = \bar{q}$.

We will simply write $\bar{f}$ and r instead of $\frac{\bar{f}}{\bar{1}}$ and $\frac{\bar{r}}{\bar{1}}$ when considering $\bar{f} \in \Gamma$ and $r \in K$ as elements in Q.

Lemma 2.42 *Let J be a prime T-radical ideal, where $T = T_+[g_1, \ldots, g_t]$. Then Q is formally real over $(K, \geq_K)$. Furthermore the ordering $\geq_Q$ that extends $\geq_K$ can be chosen such that $\bar{g}_i \geq_Q 0$ for all $i = 1, \ldots, t$.*

Proof We first show that Q is formally real over K. We will show that

$$\sum_{i=1}^{N} r_i q_i^2 = 0 \implies q_1 = \cdots = q_N = 0, \tag{2.45}$$

when $r_i >_K 0$ and $q_i \in Q$ for all $i = 1, \ldots, N$. In fact we will show the stronger condition:

$$\sum_{i=1}^{N} \alpha_i q_i^2 = 0 \implies q_1 = \cdots = q_N = 0, \tag{2.46}$$

when $\alpha_i \in \overline{T} \backslash \{0\} \supset \overline{K}^+ \backslash \{0\} \cong K^+ \backslash \{0\}$.

Assume $\sum_{i=1}^{N} \alpha_i q_i^2 = 0$ and write $\alpha_i = \bar{a}_i$ for an $a_i \in T \backslash \{0\}$ and $q_i = \frac{\bar{p}_i}{\bar{s}_i}$ where $p_i, s_i \in K[\underline{X}]$ and $s_i \notin J$. Clearing denominators we can WLOG assume $\bar{s}_1 = \cdots = \bar{s}_N = \bar{s}$ for a $s \in K[\underline{X}] \backslash J$. We now have

$$\sum_{i=1}^{N} \bar{a}_i \bar{p}_i^2 = \bar{s}^2 \sum_{i=1}^{N} \alpha_i q_i^2 = 0, \tag{2.47}$$

[8] $\bar{p}\bar{q} = 0$ implies $fg \in J$ and since J is prime $p \in J$ or $q \in J$

and thus $p := \sum_{i=1}^{N} a_i p_i^2 \in J$ since the quotient map is a ring homomorphism. For each $i = 1, \ldots, N$ we then see that

$$(a_i p_i^2)^2 + a_i' = p^2 \in J \quad \text{where} \quad a_i' = \sum_{\substack{1 \le j,k \le N \\ (j,k) \neq (i,i)}} a_j a_k p_j^2 p_k^2 \in T.$$

As J is T-radical it follows that $a_i f_i^2 \in J$. By assumption $\alpha_i \neq 0$ so $a_i \notin J$. As J is prime it follows that $p_i \in J$ and consequently $q_i = \bar{p}_i / \bar{s} = 0$ for all i. Thus Q is formally real over K.

By Theorem 2.15 there exists an ordering on Q extending the ordering on K. We will now argue that this ordering can be chosen such that additionally $g_i \geq 0$ for $i = 1, \ldots, t$. Let $\mathcal{Q}$ be the family of all field extensions $\widetilde{Q}$ of Q, satisfying (2.46) for all $q_i \in \widetilde{Q}$. Clearly $Q \in \mathcal{Q}$. Since the union of a chain in $\mathcal{Q}$ is again an element in $\mathcal{Q}$, Zorn's Lemma yields a maximal element Q' in $\mathcal{Q}$.

We now show that $\overline{T} \subseteq Q'^{(2)}$. Assume for the sake of contradiction that $\alpha \in \overline{T} \setminus \{0\}$ is not a square in Q'. Then $Q'(\sqrt{\alpha})$ is a nontrivial field extension of Q' and hence of Q. By maximality of Q' we get $Q'(\sqrt{\alpha}) \notin \mathcal{Q}$. Thus there exists some $\alpha_i \in \overline{T} \setminus \{0\}$ and $q_i + \sqrt{\alpha} q_i' \in Q'(\sqrt{\alpha})$ not all zero, such that

$$\begin{aligned} 0 &= \sum_i \alpha_i (q_i + \sqrt{\alpha} q_i')^2 \qquad (2.48) \\ &= \sum_i (\alpha_i q_i^2 + \alpha_i \alpha q_i'^2) + \sqrt{\alpha} \Big(\sum_i 2 \alpha_i q_i q_i' \Big). \end{aligned}$$

But $q + \sqrt{\alpha} q' = 0$ in $Q'(\sqrt{\alpha})$ holds if and only if $q = q' = 0$, so it follows that

$$\sum_i \alpha_i q_i^2 + \alpha_i \alpha q_i'^2 = 0 \qquad (2.49)$$

in Q'. Now, since $Q' \in \mathcal{F}$, (2.49) has only the trivial solution in Q' so $q_i = q_i' = 0$. Hence $q_i + \sqrt{\alpha} q_i' = 0$ for all i. This contradicts the non-triviality of the solution to (2.48).

As Q' is formally real over K, Q' admits an ordering extending the ordering on K. As $\bar{g}_i \in \overline{T}$ is a square in Q' for all i we conclude that $\bar{g}_i \geq 0$ in Q'. We finally restict the ordering on Q' to Q, i.e. for any $q \in Q$ we declare $q \geq_Q 0$ iff $q \geq_{Q'} 0$ as an element in Q'. This is the desired ordering on Q. □

We now prove Proposition 2.41. In [29] they apply Lang's Theorem [18, Thm. 1.1.2]. We found it more transparent and instructive to use the Transfer Principle (which is used to prove Lang's Theorem) directly.

Proof (of Proposition 2.41) Let $h \in K[\underline{X}] \backslash J$ and write $J = \langle f_1, \ldots, f_s \rangle$. We will show that $h \notin \mathbb{I}(\mathbb{V}_R(J) \cap \mathcal{W}(g_1, \ldots, g_t))$. This amounts to show that the polynomial system

$$\begin{aligned} f_i &= 0, \quad i = 1, \ldots, s, \\ g_j &\geq 0, \quad j = 1, \ldots, t, \\ h &\neq 0, \end{aligned} \tag{2.50}$$

has a solution in R^n.

Let $(Q, \leq_Q)$ be the ordered field from Lemma 2.42. By the simple version of the Transfer Principle (Corollary 2.29) it suffices to find a solution in Q^n that satisfies (2.50). Consider $\overline{X}_i = \tau(\iota(X_i)) \in Q$ which are just the monomials $X_i \in K[\underline{X}]$ embedded in Q, c.f. (2.44). Then

$$f_i(\overline{X}_1, \ldots, \overline{X}_n) = \bar{f}_i = 0 \tag{2.51}$$

in Q since $f_i \in J$ and the quotient map is a ring homomorphism. Similarly

$$g_j(\overline{X}_1, \ldots, \overline{X}_n) = \bar{g}_j \geq_Q 0 \tag{2.52}$$

by Lemma 2.42 and

$$h(\overline{X}_1, \ldots, \overline{X}_n) = \bar{h} \neq 0, \tag{2.53}$$

since $h \notin J$ by assumption.

Thus $(\overline{X}_1, \ldots, \overline{X}_n) \in Q^n$ is a solution to (2.50). This finishes the proof. □

In order to extend Proposition 2.41 to general ideals we have the following characterization of $\sqrt[T]{I}$ (which is similar to the characterization of radical ideals):

Proposition 2.43 *$\sqrt[T]{I}$ is the intersection of all prime T-radical ideals containing I.*

Proof Suppose $f \in \sqrt[T]{I}$ and let J be an arbitrary prime T-radical ideal containing I. We will show that $f \in J$. As $f \in \sqrt[T]{I}$ there is an $m > 0$ and an $a \in T$ such that $f^{2m} + a \in I \subseteq J$. J is a T-radical ideal so it follows that $f \in J$. This shows the inclusion to the right.

Suppose now $f \in K[\underline{X}] \backslash \sqrt[T]{I}$. We will find a prime T-radical ideal J containing I such that $f \notin J$. Let $\mathcal{I}$ be the set of all T-radical ideals that does not contain f^m for any $m > 0$. $\mathcal{I}$ is non-empty since it contains $\sqrt[T]{I}$: this is a T-radical ideal containing I by Lemma 2.34. Suppose for contradiction that $f^m \in \sqrt[T]{I}$ for some $m > 0$, then $f^{2m} + 0 \in \sqrt[T]{I}$ and hence $f \in \sqrt[T]{\sqrt[T]{I}} = \sqrt[T]{I}$ which is a contradiction. The union of a chain in $\mathcal{I}$ is again an element in $\mathcal{I}$ so Zorn's Lemma yields a maximal element $J \in \mathcal{I}$ w.r.t the partial order induced by inclusion.

We now show that J is a prime T-radical ideal containing I. Note that $I \subseteq \sqrt[T]{I} \subseteq J$ and J is T-radical being an element in $\mathcal{I}$. We only need to show that J is prime. Suppose $f_1, f_2 \in K[\underline{X}] \setminus J$. We will show that $f_1 f_2 \notin J$. Let $J_i = \langle J, f_i \rangle = \{g_i + b_i f_i \mid g_i \in J, b_i \in K[\underline{X}]\}$ be the ideal generated by f_i and the elements in J. We then have the strict inclusions

$$J \subsetneq J_i \subseteq \sqrt[T]{J_i}, \quad i = 1, 2, \tag{2.54}$$

and by maximality of J in $\mathcal{I}$ it follows that $f^{n_i} \in \sqrt[T]{J_i}$ for some $n_i > 0$, $i = 1, 2$. This in turn implies that $f^{2m_i} + a_i \in J_i$ for some $m_i > 0$ and $a_i \in T$, $i = 1, 2$. Thus we can write

$$f^{2m_i} + a_i = g_i + b_i f_i, \quad i = 1, 2, \tag{2.55}$$

for some $g_i \in J$ and $b_i \in K[\underline{X}]$. Muliplying the two identities in (2.55) we get

$$f^{2(m_1+m_2)} + a = g + b f_1 f_2, \tag{2.56}$$

where

$$\begin{aligned} a &= a_1 a_2 + a_1 f^{2m_2} + a_2 f^{2m_1} \in T, \\ b &= b_1 b_2 \in K[\underline{X}], \\ g &= g_1 g_2 + b_1 g_2 + b_2 g_1 \in J. \end{aligned}$$

If $f_1 f_2 \in J$ then (2.55) implies that $f^{2(m_1+m_2)} + a \in J$ and $f \in J$ since J is T-radical. This is a contradiction since $J \in \mathcal{I}$ so $f^m \notin J$ for any $m > 0$. We conclude that $f_1 f_2 \notin J$ so J is a prime ideal. □

Proof (of Theorem 2.35) The inclusion $\sqrt[T]{I} \subseteq \mathbb{I}(\mathbb{V}(I) \cap \mathcal{W}(B))$ is straight forward: Suppose that $f \in \sqrt[T]{I}$ and $x \in \mathbb{V}(I) \cap \mathcal{W}(B)$. We have that $f^{2m} + a \in I$ for some $m > 0$ and an $a \in T$. Thus $0 \leq f^{2m}(x) \leq f^{2m}(x) + a(x) = 0$ and so $f(x) = 0$. This shows that $f \in \mathbb{I}(\mathbb{V}(I) \cap \mathcal{W}(B))$.

The other inclusion follows from the preceding propositions: Let J be a prime T-radical ideal containing I. Then we have

$$\begin{aligned} I \subseteq J &\implies \mathbb{V}(J) \cap \mathcal{W}(B) \subseteq \mathbb{V}(I) \cap \mathcal{W}(B) \\ &\implies \mathbb{I}(\mathbb{V}(I) \cap \mathcal{W}(B)) \subseteq \mathbb{I}(\mathbb{V}(J) \cap \mathcal{W}(B)) \subseteq J \end{aligned}$$

where the last inclusion follows from Proposition 2.41. As J was arbitrarily chosen it follows from Proposition 2.43 that $\mathbb{I}(\mathbb{V}(I) \cap \mathcal{W}(B)) \subseteq \sqrt[T]{I}$. □

2.4 Infeasibility certificates for polynomial systems

We have presented the above results as they appear in [29]. We now aim at a more explicit formulation in terms of polynomial systems. The proofs in this section is our own work.

With the aim of arriving at Theorem 2.45 which is a variation appearing in [19, Thm. 4.4], we can use the preceding results to prove:

Corollary 2.44 *Consider the following system of polynomial equations and inequalities:*

$$\begin{aligned} f_i(x) &= 0, \quad i = 1, \dots, s \\ g_j(x) &\geq 0, \quad j = 1, \dots, t \end{aligned} \tag{2.57}$$

$f_i, g_j \in K[\underline{X}]$. *The system* (2.57) *has no solution in* R^n *if and only if there is a relation of the form*

$$1 + f + g = 0 \tag{2.58}$$

with $f \in \langle f_1, \dots, f_s \rangle$ *and* $g \in T_+[g_1, \dots, g_t]$.

Proof A relation of the form (2.58) is clearly a sufficient condition for the system (2.57) to be infeasible: If x is a solution to (2.57) then $f(x) = 0$ and $g(x) \geq 0$ for anly $f \in \langle f_1, \dots, f_s \rangle$ and $g \in T_+[g_1, \dots, g_t]$. This lead to the contradiction $0 = 1 + f(x) + g(x) = 1 + g(x) \geq 1$.

On the other hand. Suppose (2.57) is infeasible. Then $\mathbb{V}(I) \cap \mathcal{W}(B) = \emptyset$. Theorem 2.35 now implies $\sqrt[T]{I} = \mathbb{I}(\emptyset) = K[\underline{X}]$, in particular $1 \in \sqrt[T]{I}$. That is, $1 + g \in I$ for some $g \in T_+[g_1, \dots, g_t]$ and (2.58) follows by taking $f = -(1 + g)$. □

To be completely explicit (2.58) takes the form

$$1 + \sum_{i=1}^{s} a_i f_i + \sum_{\nu \in \{0,1\}^t} \Big(\sum_{j=1}^{u_\nu} r_{j\nu} b_{j\nu}^2 \Big) g_1^{\nu_1} \cdots g_t^{\nu_t} = 0 \tag{2.59}$$

for some $a_i, b_{j\nu} \in K[\underline{X}]$ and $r_{j\nu} \geq 0$ c.f. (2.13).

From a geometrical perspective systems of the form (2.57) correspond to basic closed semi-algebraic sets. In that way Corollary 2.44 guarantees certificates of emptiness of any basic closed semi-algebraic set.

We want to improve this result to any basic semi-algebraic set open or closed. This amounts to allow inequations in the polynomial system. The variation below appears in [19, Thm. 4.4] although the proof of necessity, which is the difficult part, is not presented there. The following proof is our own work:

Theorem 2.45 *Consider the following system of polynomial equations, inequalities, and inequations:*

$$f_i(x) = 0, \quad i = 1, \ldots, s, \tag{2.60a}$$
$$g_j(x) \geq 0, \quad j = 1, \ldots, t, \tag{2.60b}$$
$$h_k(x) \neq 0, \quad k = 1, \ldots, u, \tag{2.60c}$$

$f_i, g_j, h_k \in K[\underline{X}]$. *The system* (2.60) *has no solution in* R^n *if and only if there is a relation of the form*

$$f + g + h^{2m} = 0, \tag{2.61}$$

with $f \in I = \langle f_1, \ldots, f_s \rangle$, $g \in T = T_+[g_1, \ldots, g_t]$, $h = h_1 \cdots h_u$, *and* $m > 0$.

Proof The case $u = 0$ is covered by Corollary 2.44. So assume $u > 0$. The sufficiency is clear and proved similar to Corollary 2.44.

Suppose (2.60) has no solution. Let $B = \{g_1, \ldots, g_t\}$. Observe that

$$\mathbb{V}(I) \cap \mathcal{W}(B) \subseteq \mathbb{V}(\langle h \rangle). \tag{2.62}$$

If $x \in \mathbb{V}(I) \cap \mathcal{W}(B)$, then the equations (2.60a) and (2.60b) are satisfied. As the system (2.60) is infeasible there must exist a k such that $h_k(x) = 0$ and in particular $h(x) = 0$, so $x \in \mathbb{V}(\langle h \rangle)$, establishing (2.62).

Using Theorem 2.35 and the inclusion reversing property of the vanishing ideal we then get

$$h \in \mathbb{I}(\mathbb{V}(\langle h \rangle)) \subseteq \mathbb{I}(\mathbb{V}(I) \cap \mathcal{W}(B)) = \sqrt[T]{I}.$$

Thus $h^{2m} + g \in I$ for some $m > 0$ and $g \in T$. Finally, put $f = -(h^{2m} + g) \in I$. Then $f + g + h^2 = 0$ as wanted. □

Writing $h(x) > 0$ as $h(x) \geq 0 \wedge h(x) \neq 0$ we immediately get the following version similar to (but not equal to) what appears in [30]:

Corollary 2.46 *The following system of polynomial equations and inequalities*

$$f_i(x) = 0, \quad i = 1, \ldots, s \tag{2.63a}$$
$$g_j(x) \geq 0, \quad j = 1, \ldots, t \tag{2.63b}$$
$$h_k(x) > 0, \quad k = 1, \ldots, u \tag{2.63c}$$

is infeasible if and only if there exists a relation of the form

$$f + g + h^{2m} = 0 \tag{2.64}$$

where $f \in I = \langle f_1, \ldots, f_s \rangle$ *and* $g \in T[h_1, \ldots, h_u, g_1, \ldots, g_t]$, $h = h_1 \cdots h_u$, *and* $m > 0$.

Remark 2.47 In [30, Theorem 7.5] it is stated without proof that the system (2.63) is infeasible if and only if one can find a certificate of the form

$$f + g + h + \prod_{i=1}^{u} h_l^{v_l} = 0, \tag{2.65}$$

where $f \in I$, $g \in T[g_1, \ldots, g_t]$, and $h \in T[h_1, \ldots, h_u]$.

Notice that $T[g_1, \ldots, g_t, h_1, \ldots, h_u]$ which contains all mixed products is much larger than $T[g_1, \ldots, g_t] + T[h_1, \ldots, h_u]$. [30] only refer to [29] which is also our reference. It seems plausible that one could find an example of an infeasible system for which an infeasibility certificate would require mixed products. Unfortunately we have not succeeded in finding one. △

2.5 Positivstellensatz and Hilbert's 17th problem

Using Theorem 2.45 we get a very simple proof[9] of the Positivstellensatz [29, Thm. 3]. This proof is our own work. The relations on the right hand side below are clear certificates of non-negativity and positivity of f on S, so as before it is the implications to the right which are interesting.

Theorem 2.48 (Positivstellensatz) *Let $S = \mathcal{W}(g_1, \ldots, g_t)$ and $T = T_+[g_1, \ldots, g_t]$. Then*

i)

$$f \geq 0 \text{ on } S \iff f^{2m} + a_1 = f a_2 \text{ for some } a_1, a_2 \in T \text{ and } m > 0. \tag{2.66}$$

ii)

$$f > 0 \text{ on } S \iff 1 + a_1 = f a_2 \text{ for some } a_1, a_2 \in T. \tag{2.67}$$

Proof i): f is non-negative on S if and only if the system

$$\begin{aligned} g_i &\geq 0, \quad i = 1, \ldots, t, \\ -f &> 0, \end{aligned} \tag{2.68}$$

has no solution in R^n. By Corollary 2.46 this is equivalent to the existence of an $a \in T_+[-f, g_1, \ldots, g_t]$ and an $m > 0$ such that

$$a + f^{2m} = 0. \tag{2.69}$$

[9] compared to the original proof given in [29, Thm. 3]

Now

$$T_+[-f, g_1, \ldots, g_t] = T_+[g_1, \ldots, g_t] - fT_+[g_1, \ldots, g_t] \tag{2.70}$$

by Lemma 2.6 so the expression becomes

$$a_1 - fa_2 + f^{2m} = 0, \tag{2.71}$$

which is the desired relation.

ii): Similar to *i*) f is positive on S if and only if the system

$$\begin{aligned} g_i &\geq 0, \quad i = 1, \ldots, t, \\ -f &\geq 0, \end{aligned} \tag{2.72}$$

has no solution. The relation (2.67) then follows from Corollary 2.44 by same arguments as in part *i*). □

Remark 2.49 In [29, Thm. 3] the result is formulated

$$f \geq 0 \text{ on } S \iff f^{2m+1} + fa_1 = a_2 \text{ for some } a_1, a_2 \in T \text{ and } m > 0 \tag{2.73}$$

This is equivalent to (2.66): if $f^{2m+1} + a_1 f = a_2$ then multiplying through by f yields $f^{2m'} + a_1' = a_2 f$ where $m' = m + 1$ and $a_1' = f^2 a_1 \in T$. Similarly if $f^{2m} + a_1 = a_2 f$ then multiplying with f yields $f^{2m+1} + a_1 f = a_2'$ where $a_2' = f^2 a_2 \in T$. The characterization (2.66) will be useful in proving regularity of the rational functions in the strong version of Artin's Theorem below. △

Hilbert's 17th problem

We can now give an affirmative answer to Hilbert's 17th problem which we discussed in the introduction. This is also known as Artin's Theorem.

Here we actually prove something slightly stronger: in case of a strictly positive polynomial function we guarantee *regular* rational functions in the decomposition. Here a rational function $\phi \in R(x)$ is *regular* if it can be represented as $\frac{p}{q}$ with $q(x) \neq 0$ for all $x \in R^n$.

This result is mentioned in [29] but the proof is omitted there. We have carried out the argument below.

Corollary 2.50 ((Strong) Artin's Theorem) *Let $f \in R[\underline{X}]$.*

i) If $f \geq 0$ on R^n then f is a sum of squares of rational functions.

ii) If $f > 0$ on R^n then f is a sum of squares of regular rational functions.

Proof i) Let $fR[\underline{X}]\backslash\{0\}$. If f is non-negative on $R^n = \mathcal{W}(\emptyset)$ then by Theorem 2.48 *i)* we then get a relation

$$f^{2m} + a_1 = a_2 f \tag{2.74}$$

where $a_1, a_2 \in T[\emptyset] = \sum R[\underline{X}]^{(2)}$ are sums of squares. As f (and hence the l.h.s of (2.74)) is not identically zero we see that a_2 is not identically zero and f can be isolated. Thus

$$f = \frac{f^{2m} + a_1}{a_2} = \frac{f^{2m} + \sum_i b_i^2}{\sum_k c_k^2} = \sum_j \Big(\frac{f^m c_j}{\sum_k c_k^2}\Big)^2 + \sum_{i,j} \Big(\frac{b_i c_j}{\sum_k c_k^2}\Big)^2, \tag{2.75}$$

which is a sum of squares of rational functions.

ii): The arguments are similar when f is globally positive. We use Theorem 2.48 *ii)* and get

$$f = \frac{1 + a_1}{a_2} = \sum_j \Big(\frac{c_j}{\sum_k c_k^2}\Big)^2 + \sum_{i,j} \Big(\frac{b_i c_j}{\sum_k c_k^2}\Big)^2 \tag{2.76}$$

where $a_2(x) \neq 0$ for all $x \in R^n$ since $1 + a_1 > 0$. □

3 Simplifying certificates in the compact case

In this chapter we will study different improvements of the general Positivstellensatz Theorem 2.48 in case we have some additional structure on the basic closed semi-algebraic set $S = \mathcal{W}(g_1, \ldots, g_m)$. We will later see how these results lead to efficient and general optimization techniques when we study the Lasserre Hierachy.

The results in this chapter are less abstract and does not work for general real closed fields. So from now on we will only work with the usual polynomial ring $\mathbb{R}[\underline{X}] = \mathbb{R}[X_1, \ldots, X_n]$ over the real numbers.

Our main reference for Putinar's Positivstellensatz is the paper [28]. Most of the material in Section 3.2 is based on exercises found in [22]. We have used [17] as our reference in Section 3.3 and also partly in Section 3.4.

3.1 Putinar's Positivstellensatz

We start by introducing an algebraic structure which generalizes the preorder introduced in Chapter 2:

Definition 3.1 *A* quadratic module $M \subseteq \mathbb{R}[\underline{X}]$[1] *is a subset satisfying*

$$m_1, m_2 \in M \implies m_1 + m_2 \in M, \tag{3.1a}$$

$$c \in \mathbb{R}[\underline{X}], m \in M \implies c^2 m \in M, \tag{3.1b}$$

$$1 \in M. \tag{3.1c}$$

We observe that any preorder is a quadratic module since (3.1b) and (3.1c) implies that M contains $\mathbb{R}[\underline{X}]^{(2)}$.

[1] This definition makes sense in any commutative ring.

As for the preorder we denote by $M[B]$ the *quadratic module generated by* B where $B \subseteq \mathbb{R}[\underline{X}]$, i.e. $M[B]$ the smallest quadratic module containing all elements of B. As usual we are mainly interested in finitely generated quadratic modules.

Let $B = \{g_1, \dots, g_m\} \subseteq \mathbb{R}[\underline{X}]$ be a finite subset. Then we have

$$M[B] = \sum_{i=0}^{m} \big(\sum \mathbb{R}[\underline{X}]^{(2)}\big) g_i \subseteq \mathbb{R}[\underline{X}] \tag{3.2}$$

where $g_0 = 1 \in \mathbb{R}[\underline{X}]$.

The preorder

$$T[B] = \sum_{\nu \in \{0,1\}^m} \big(\sum \mathbb{R}[\underline{X}]^{(2)}\big) g_1^{\nu_1} \dots g_m^{\nu_m} \tag{3.3}$$

considered in the previous chapter contains $M[B]$ as a subset. But $T[B]$ is also itself a finitely generated quadratic module, namely the quadratic module generated by the $2^m - 1$ products $g_1^{\nu_1} \cdots g_m^{\nu_m}$, where $0 \neq \nu \in \{0,1\}^m$.

Like in $T[B]$, any polynomial $p \in M[B]$ is nonnegative on $\mathcal{W}(g_1, \dots, g_m)$ and admits a certificate of this fact, namely a representation of of the form

$$p = \sum_{i=0}^{m} \sigma_i g_i, \quad \sigma_i \in \sum \mathbb{R}[\underline{X}]^{(2)}. \tag{3.4}$$

In terms of the number of polynomials involved in the certificate, elements in $M[B]$ possess prettier (meaning smaller) certificates than those in $T[B]$.[2]

Definition 3.2 *A quadratic module M is said to b* Archimedian *if for any $f \in \mathbb{R}[\underline{X}]$ there is an $N \in \mathbb{N}$ such that $N - f \in M$.*

This seems like a quite restrictive property, and it is not clear how one should check if it holds. We will discuss these matters later. For now we just point out that S has to be compact.

Lemma 3.3 *Let $M = M[B]$ and $S = \mathcal{W}(B)$ where $B = \{g_1, \dots, g_m\} \subseteq \mathbb{R}[\underline{X}]$. Then*

$$M \text{ is Archimedian} \implies S \text{ is compact.} \tag{3.5}$$

Proof If M is Archimedian then $N - \sum_{i=1}^{n} X_i^2 \in M$ for some N and so $N - \sum_{i=1}^{n} x_i^2 = N - \|x\|^2 \geq 0$ for all $x \in S$. This means that S is contained in the ball $B_{\sqrt{N}}(0)$ of radius $\sqrt{N}$ and so S is compact being closed and bounded. □

[2]However, it might be the case that a polynomial on the form (3.4) requires a higher degree of the SOS polynomials σ_i so that allowing mixed products $g_1^{\nu_1} \cdots g_m^{\nu_m}$ would produce a preferable certificate.

We can now state one of the two main theorems in this chapter:

Theorem 3.4 (Putinar's Positivstellensatz) *Let $S = \mathcal{W}(B)$ and suppose $M = M[B]$ is Archimedian. Then*

$$f > 0 \text{ on } S \implies f \in M \tag{3.6}$$

The proof of Putinar's Theorem in this section is based on [28]. A main ingredient is a theorem of Pólya which is interesting and useful in its own as it gives an alternative way of certifying positivity on the non-negative orthant in case the polynomial is homogeneous.

Theorem 3.5 (Pólya) *Suppose $G \in \mathbb{R}[\underline{X}]$ is homogeneous and*

$$G(x) > 0, \quad \forall x \in (\mathbb{R}^+)^n \backslash \{0\} \tag{3.7}$$

Then for sufficiently large $K \in \mathbb{N}$ the polynomial $(X_1 + \ldots, X_n)^K G$ has non-negative coefficients.

The proof which is is purely analytic and not particularly interesting is included in Appendix A.1.

Positivity of the coefficients of $(X_1 + \cdots + X_n)^K G(\underline{X})$ is a simple certificate of the fact that G is positive on $(\mathbb{R}^+)^n \backslash \{0\}$: Both $(X_1 + \cdots + X_n)^K G(\underline{X})$ and $(X_1 + \cdots + X_n)^K$ (having positive coefficients) are positive on $\mathbb{R}^n_{\geq 0} \backslash \{0\}$, so G must be positive as well.

Example 3.6 Let $G_\epsilon = (X - Y)^2 + \epsilon XY$. Note that for $\epsilon > 0$, $G_\epsilon \in T[X, Y]$ is a certificate of non-negativity on $(\mathbb{R}^+)^2 \mathcal{W}(X, Y)$. It is easy to see that $G_\epsilon$is strictly positive on $(\mathbb{R}^+)^2 \backslash \{0\}$.

Let us compute a Pólya certificate of this. If $\epsilon = \frac{1}{2}$, then

$$\begin{aligned}
(X+Y)G_{\frac{1}{2}} &= X^3 - \tfrac{1}{2}X^2Y - \tfrac{1}{2}XY^2 + Y^3, \\
(X+Y)^2 G_{\frac{1}{2}} &= X^4 + \tfrac{1}{2}X^3Y - X^2Y^2 + \tfrac{1}{2}XY^3 + Y^4, \\
(X+Y)^3 G_{\frac{1}{2}} &= X^5 + \tfrac{3}{2}X^4Y - \tfrac{1}{2}X^3Y^2 - \tfrac{1}{2}X^2Y^3 + \tfrac{3}{2}XY^4 + Y^5, \\
(X+Y)^4 G_{\frac{1}{2}} &= X^6 + \tfrac{5}{2}X^5Y + X^4Y^2 - X^3Y^3 + X^2Y^4 + \tfrac{5}{2}XY^5 + Y^6, \\
(X+Y)^5 G_{\frac{1}{2}} &= X^7 + \tfrac{7}{2}X^6Y + \tfrac{7}{2}X^5Y^2 + \tfrac{7}{2}X^2Y^5 + \tfrac{7}{2}XY^6 + Y^7,
\end{aligned}$$

where the last polynomial has positive coefficients.

If instead $\epsilon = \frac{1}{10}$, then

$$(X+Y)^{37} G_{\frac{1}{10}} = X^{39} + 35.1X^{38}Y + 596.7X^{37}Y^2 + \cdots + 35.1XY^{38} + Y^{39}$$

is the first polynomial with positive coefficients.

Inspired by an exercise in [3] we have made a small experiment to examine the the size of the power needed as a function of $\epsilon = \frac{1}{n}$ (see Figure 3.1). There

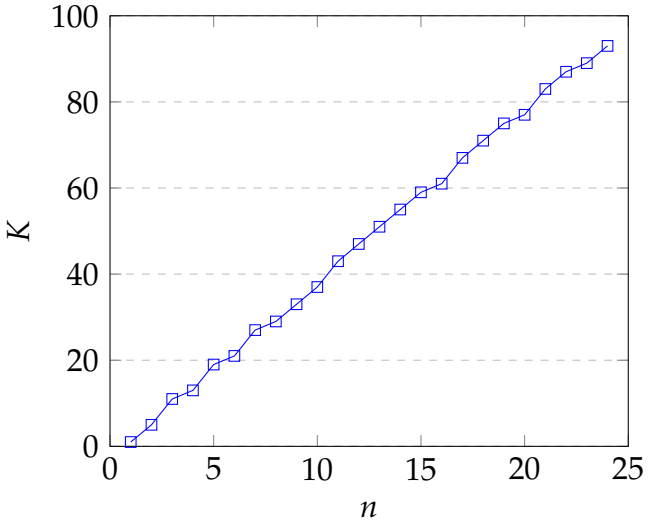

Figure 3.1: Minimal K such that $(X+Y)^K((X-Y)^2+\frac{1}{n}XY)$ has positive coefficients.

are general complexity bounds for Theorem 3.5 in terms of the coefficients of G and the minimum of G on the standard n-simplex Δ_n [21]. ○

We need a few lemmas before we continue:

Lemma 3.7 *For any $y \in (-\infty, 1]$ the sequence $(y_k)_{k\in\mathbb{N}}$ given by $y_k = (y-1)^{2k}y$ is decreasing and dominated by $\frac{1}{2k+1}$. I.e*

$$y_{k+1} \leq y_k \leq \frac{1}{2k+1} \tag{3.8}$$

for all $k \in \mathbb{N}$.

Proof The inequalities are clearly satisfied for $y < 0$ so let $y \in [0,1]$. $(y_k)_{k\in\mathbb{N}}$ is clearly decreasing.

Observe that

$$0 = \frac{d}{dy}(y-1)^{2k}y = (y-1)^{2k-1}((2k+1)y-1) \implies y \in \{\tfrac{1}{2k+1}, 1\}.$$

Among the candidates $0, 1$ or $\frac{1}{2k+1}$ the latter is clearly the where the maximum is attained. Thus

$$(y-1)^{2k}y \leq \left(\frac{1}{2k+1}-1\right)^{2k}\frac{1}{2k+1} \leq \frac{1}{2k+1}. \qquad \square$$

The next lemma is our own work and a simple application of Dini's Theorem from analysis. We use it to simplify the proof of Lemma 3.9[3] below.

Lemma 3.8 *Let $(f_k)_{k\in\mathbb{N}}$ be a monotonic increasing sequence of continuous functions $f_k : C \to \mathbb{R}$ on a compact set C, i.e. $f_1(x) \leq f_2(x) \leq \dots$ for all $x \in C$. Suppose that there is an $\epsilon > 0$ such that $f_k(x) \geq \epsilon$ for all sufficiently large k dependent on x. More precisely:*

$$\exists \epsilon > 0 \, \forall x \in C \, \exists K_x \in \mathbb{N} \, \forall k \geq K_x : \quad f_k(x) \geq \epsilon. \tag{3.9}$$

Then there is a uniform *$K \in \mathbb{N}$ such that*

$$f_k > 0 \ \text{ on } C, \quad \forall k \geq K. \tag{3.10}$$

Proof Let $(f_k)_{k\in\mathbb{N}}$ and ϵ satisfy the assumptions and let $h_k = \min(f_k, \epsilon)$ for all $k \in \mathbb{N}$. Then $(h_k)_{k\in\mathbb{N}}$ is monotonic increasing and by (3.9) converging point-wise to ϵ. As h_k is continuous for all k and C is compact Dini's Theorem implies that $(h_k)_{k\in\mathbb{N}}$ converges *uniformly* to ϵ. So choose $K > 0$ such that $|h_k(x) - \epsilon| = \epsilon - h_k(x) \leq \frac{\epsilon}{2}$ for all $x \in C$. Then

$$f_k(x) \geq h_k(x) \geq \tfrac{\epsilon}{2} > 0, \quad \forall x \in C \, \forall k \geq K, \tag{3.11}$$

as wanted. □

Lemma 3.9 *Let $S = \mathcal{W}(g_1, \dots, g_m)$ be compact and suppose $C \subseteq \mathbb{R}^n$ is compact with $g_i \leq 1$ on C for all $i = 1, \dots, m$. If $p > 0$ on S then there exist $K \in \mathbb{N}$ such that*

$$q_k := p - \sum_{i=1}^{m} (1 - g_i)^{2k} g_i > 0 \ \text{ on } C, \quad \forall k \geq K. \tag{3.12}$$

Remark 3.10 This is a stronger version of [28, Lemma 8] which states that there is an $s \in \mathbb{N}$ such that

$$p - s \sum_{i=1}^{m} (1 - g_i)^{2k} g_i > 0 \text{ on } C, \quad \forall k \geq K. \tag{3.13}$$

We prove that we can always choose $s = 1$. Using Lemma 3.8 we get this stronger result with a simpler proof – this is however at cost of transparency on the complexity. △

[3] compared to [28, Lemma 8]

Proof Choose $\epsilon > 0$ such that $p \geq 2\epsilon$ on S (this can be done since p attains its minimum on S). Note that $(q_k)_{k\in N}$ is monotonic increasing on C by the assumption $g_i \leq 1$ for all i and the first inequality in Lemma 3.7. We now consider two cases:

Assume that $x \in S \cap C$. Then since $g_i(x) \leq 1$ for all $i = 1, \ldots, m$, Lemma 3.7 implies that

$$q_k(x) \geq 2\epsilon - \frac{m}{2k+1} \geq \epsilon \tag{3.14}$$

for all sufficiently large k.

Assume now that $x \in C\backslash S$. Then $g_j(x) < 0$ for some j and so $(1 - g_j(x))^{2k} g_j(x) \to -\infty$ as $k \to \infty$. Using Lemma 3.7 on the remaining $m-1$ terms yield

$$q_k(x) \geq p(x) - \frac{m-1}{2k+1} - (1 - g_j(x))^{2k} g_j(x) \to \infty \tag{3.15}$$

as $k \to \infty$. In particular $q_k(x) \geq \epsilon$ for sufficiently large k.

The conclusion now follows from Lemma 3.8. □

We are now ready to prove Putinar's Positivstellensatz. We follow the main arguments of [28] but make an other choice of N: Instead of choosing N such that $N - \sum_{i=1}^{n} X_i^2 \in M$ (this is the way they define the Archimedian property – we show later that our definition is equivalent) we choose N such that $N \pm X_i \in M$ for $i = 1, \ldots, n$. The simplex Δ below is also a bit different as where our constraint is $y_1 + \cdots + y_{2n} = 2nN$, [28] chooses $2n(N + \frac{1}{4})$. This does not affect the proof other than we can skip an argument when we show $\phi(Y_i) \in M$.

Proof (of Theorem 3.4) As M is Archimedian we can choose N such that $N \pm X_i \in M$ for $i = 1, \ldots, n$. Consider the $2n$-simplex

$$\Delta = \left\{ y \in [0, \infty)^{2n} \,\middle|\, y_1 + \cdots + y_{2n} = 2nN \right\} \subseteq \mathbb{R}^{2n} \tag{3.16}$$

and the linear map

$$l \colon \mathbb{R}^{2n} \to \mathbb{R}^n, \quad y \mapsto \left(\tfrac{y_1 - y_{n+1}}{2}, \ldots, \tfrac{y_n - y_{2n}}{2} \right). \tag{3.17}$$

We then put $C := l(\Delta) \subseteq \mathbb{R}^n$. This is a compact set being the image of a compact set under a continuous function.

We can WLOG assume that $g_i \leq 1$ on C for all $i = 1, \ldots, m$: If $g_i(x) > 1$ for some i and $x \in C$ we can put $a := \sup\{g_i(x) \mid x \in C,\ i = 1, \ldots, m\}$ and note that $0 < a < \infty$ by compactness OF C. If we scale g_i by the positive factor a^{-1} both S and M remain invariant. To be precise let $\tilde{g}_i = a^{-1} g_i$, $i = 1 \ldots, m$. Then

$g_i(x) \geq 0$ iff $\tilde{g}_i(x) \geq 0$ so $\mathcal{W}(g_1, \ldots, g_m) = \mathcal{W}(\tilde{g}_1, \ldots, \tilde{g}_m)$. Also $a^{-1} \in \mathbb{R}[\underline{X}]^{(2)}$ so $M[g_1, \ldots, g_m] = M[\tilde{g}_1, \ldots, \tilde{g}_m]$.

By Lemma 3.9 we can choose k large enough that

$$q := q_k = p - \sum_{i=1}^{m} (1 - g_i)^{2k} g_i > 0 \text{ on } C. \tag{3.18}$$

If we can show that $q \in M$ we are done, since then

$$p = q + \sum_{i=1}^{m} (1 - g_i)^{2k} g_i \in M.$$

Write $q = \sum_{i=1}^{d} Q_i$, where $d = \deg q$ and Q_i is homogeneous of degree i. Then define

$$F := \sum_{i=1}^{m} Q_i\Big(\frac{Y_1 - Y_{n+1}}{2}, \ldots, \frac{Y_n - Y_{2n}}{2}\Big)\Big(\frac{Y_1 + \cdots + Y_{2n}}{2nN}\Big)^{d-i} \in \mathbb{R}[\underline{Y}], \tag{3.19}$$

which is homogeneous of degree d. Now observe that

$$F(y) = \sum_{i=1}^{d} Q(l(y)) = q(l(y)) > 0, \ \forall y \in \Delta. \tag{3.20}$$

Here we used that $y_1 + \cdots + y_{2n} = 2nN$ in the first equality and that $l(y) \in C$ in the inequality.

As F is homogeneous $F(\lambda y) = \lambda^d F(y)$ for all $\lambda > 0$. Hence $F > 0$ on $(\mathbb{R}^+)^{2n} \backslash \{0\}$ and by Theorem 3.5 there exists a $K > 0$ such that

$$G := \Big(\frac{Y_1 + \cdots + Y_{2n}}{2nN}\Big)^K F \tag{3.21}$$

has only non-negative coefficients.

Consider now the homomorphism

$$\phi : \mathbb{R}[\underline{Y}] \to \mathbb{R}[\underline{X}], \tag{3.22}$$

$$Y_i \mapsto N + X_i, \tag{3.23}$$

$$Y_{n+i} \mapsto N - X_i, \tag{3.24}$$

for $i = 1, \ldots, n$ and observe that

$$\phi(G) = \phi(F) = \sum_{i=1}^{d} Q_i\Big(\frac{\phi(Y_1) - \phi(Y_{n+1})}{2}, \ldots, \frac{\phi(Y_n) - \phi(Y_{2n})}{2}\Big) \tag{3.25}$$

$$= \sum_{i=1}^{d} Q_i = \ q. \tag{3.26}$$

By the choice of N we see that $\phi(Y_i) \in M$ for $i = 1, \ldots, 2n$. Since M is closed under addition, multiplication and non-negative scaling (non-negative scalars are squares) we conclude that $\phi(G) = q \in M$. □

Remark 3.11 In the above proof we used only the assumption that $N \pm X_i \in M$ for sufficiently large M and not the seemingly stronger Archimedian property. In the next section we will see that this is essentially *not* a weaker assumption. △

3.2 Characterizing Archimedian quadratic modules

We now aim at a characterization of the finitely generated Archimedian quadratic modules $M[B]$. We have seen in Lemma 3.3 that $\mathcal{W}(B)$ has to be compact. However compactness is not a sufficient condition as the following example (which is based on an exercise in [22]) illustrates.

Example 3.12 Consider $g_1 = X_1 - \frac{1}{2}$, $g_2 = X_2 - \frac{1}{2}$, and $g_3 = 1 - X_1X_2$.

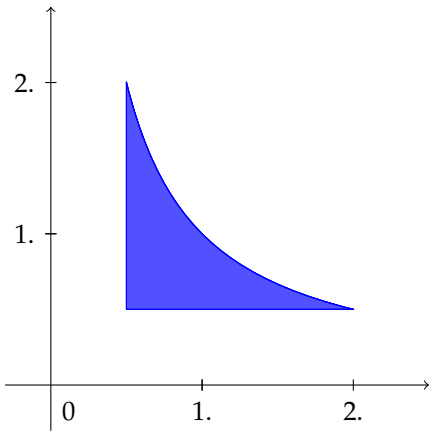

Figure 3.2: $\mathcal{W}(g_1, g_2, g_3)$

We claim that $M = M[g_1, g_2, g_3]$ is not Archimedian.

To show this we define an invariant which holds for all elements in M. Fix a term order, say lexicographic order with $X_1 \geq X_2$. We let $\mathrm{lc}(f)$ denote the leading coefficient of f with respect to this term order and $\mathrm{mdeg}(f)$ the multidegree of f. E.g. if $f = -7X_1X_2^2 + X_1^2 + 3$ then $\mathrm{lc}(f) = -7$ and $\mathrm{mdeg}(f) = (1, 2)$. By definition $\mathrm{lc}(0) = 0$.

Let $\eta : \mathbb{Z}^2 \to \{-1, 1\}$ be given by

$$\eta(\nu) = \begin{cases} -1 & \text{if } \nu \equiv (1,1) \mod 2, \\ 1 & \text{otherwise.} \end{cases} \tag{3.27}$$

and put

$$\Psi(f) = \mathrm{lc}(f)\eta(\mathrm{mdeg}(f)). \tag{3.28}$$

We now show that $\Psi(f) \geq 0$ for all $f \in M$. We do this by structural induction:

First we note that

$$\Psi(1) = 1 \cdot \nu(0,0) = 1 \geq 0, \tag{3.29}$$
$$\Psi(g_1) = 1 \cdot \nu(1,0) = 1 \geq 0, \tag{3.30}$$
$$\Psi(g_2) = 1 \cdot \nu(0,1) = 1 \geq 0, \tag{3.31}$$
$$\Psi(g_3) = -1 \cdot \nu(1,1) = 1 \geq 0. \tag{3.32}$$

Suppose now $\Psi(h_1) \geq 0$ and $\Psi(h_2) \geq 0$. We must show that $\Psi(h_1 + h_2) \geq 0$. We can safely assume $h_1, h_2 \neq 0$.

There are now two cases: If $\mathrm{mdeg}(h_1) > \mathrm{mdeg}(h_2)$ then $\Psi(h_1 + h_2) = \Psi(h_1) \geq 0$.

If $\mathrm{mdeg}(h_1) = \mathrm{mdeg}(h_2)$ then

$$\mathrm{lc}(h_1 + h_2) = \mathrm{lc}(h_1) + \mathrm{lc}(h_2) \text{ and } \mathrm{mdeg}(h_1 + h_2) = \mathrm{mdeg}(h_1) = \mathrm{mdeg}(h_2).$$

Here we used that the leading terms can not cancel since $\Psi(h_i) \geq 0$, $i = 1, 2$, implies that $\mathrm{lc}(h_1)$ and $\mathrm{lc}(h_2)$ have same sign. Hence

$$\Psi(h_1 + h_2) = \Psi(h_1) + \Psi(h_2) \geq 0.$$

Now suppose $\Psi(h) \geq 0$ and $f^2 \in \mathbb{R}[X_1, X_2]^{(2)}$. We must show that $\Psi(f^2 h) \geq 0$. Assume again $h, f \neq 0$. Then $\mathrm{lc}(f^2 h) = \mathrm{lc}(f)^2 \mathrm{lc}(h)$ and

$$\mathrm{mdeg}(f^2 h) = 2\mathrm{mdeg}(f) + \mathrm{mdeg}(h) \equiv \mathrm{mdeg}(h) \mod 2.$$

So

$$\Psi(f^2 h) = \mathrm{lc}(f)^2 \Psi(h) \geq 0. \tag{3.33}$$

This finishes the induction step.

To show that M is not Archimedian we simply note that

$$\Psi(N - X_1^2 - X_2^2) = -1 \cdot \nu(2,0) = -1 < 0 \tag{3.34}$$

for all $N \in \mathbb{N}$. ○

Putinar's Positivstellensatz gives the following characterization of finitely generated Archimedian quadratic modules:

Proposition 3.13 *Let $S = \mathcal{W}(g_1, \ldots, g_m)$ and $M = M[g_1, \ldots, g_m]$. The following conditions are equivalent:*

i) M is Archimedian.

ii)
$$\exists N > 0 : \; N - \sum_{i=1}^{n} X_i^2 \in M. \tag{3.35}$$

iii)
$$\exists N > 0 : \; N \pm X_i \in M, \; i = 1, \ldots, n. \tag{3.36}$$

iv) S is bounded and
$$f > 0 \text{ on } S \implies f \in M. \tag{3.37}$$

for all $f \in \mathbb{R}[\underline{X}]$.

Proof $i) \Rightarrow ii)$: Clear.

$ii) \Rightarrow iii)$: Choose N so that $N - \sum_{i=1}^{n} X_i^2 \in M$. Then observe that

$$N + \tfrac{1}{4} \pm X_i = \Big(N - \sum_{j=1}^{n} X_j^2\Big) + \sum_{j \neq i} X_j^2 + (X_i \pm \tfrac{1}{2})^2 \in M. \tag{3.38}$$

$iii) \Rightarrow iv)$: Choose n as in $iii)$. Then $S \subseteq [-N, N]^n$. (3.37) is exactly what we proved in Theorem 3.4. Indeed we only assumed the condition $iii)$ in the proof.

$iv) \Rightarrow i)$: Clear: For any $f \in \mathbb{R}[\underline{X}]$ we can choose N big enough that $N - f > 0$ on S. Then $N - f \in M$ by $iv)$. □

The identity (3.38) is found in the proof of Putinar's theorem appearing in [28] – this was exactly the argument we could skip by the other choice of N.

There are also some sufficient condition for M to be Archimedian. The following shows that the quadratic module associated to a compact *polytope* is always Archimedian.

Proposition 3.14 *Suppose $g_1, \ldots, g_m \in \mathbb{R}[\underline{X}]$ are linear and $\mathcal{W}(g_1, \ldots, g_m)$ is compact and non-empty. Then $M[g_1, \ldots, g_m]$ is Archimedian.*

Following an exercise in [22] we will use a result by Minkowski which we state without proof.

Theorem 3.15 *Suppose $f, g_1, \ldots, g_m \in \mathbb{R}[\underline{X}]_1$ are linear functions such that $S = \mathcal{W}(g_1, \ldots, g_m) \neq \emptyset$, i.e. S is a non-empty polytope, and $f \geq 0$ on S. Then there exist non-negative real numbers $\beta_0, \ldots, \beta_m \geq 0$ such that*

$$f = \beta_0 + \beta_1 g_1 + \cdots + \beta_m g_m. \tag{3.39}$$

In particular $f \in M[g_1, \ldots, g_m]$.

Proof (of Proposition 3.14) By compactness choose $N > 0$ such that $N \pm X_i \geq 0$ on $\mathcal{W}(g_1, \ldots, g_l)$ for all $i = 1, \ldots, m$. As $N \pm X_i$ is linear Theorem 3.15 implies

$$N \pm X_i \in M[g_1, \ldots, g_m], \quad i = 1, \ldots, m. \tag{3.40}$$

Hence by Proposition 3.13 $M[g_1, \ldots, g_m]$ is Archimedian. □

3.3 Schmüdgen's Positivstellensatz

The Archimedian assumption in Putinar's Positivstellensatz allowed for a quite low level proof using only simple analysis. We never invoked the theorems of Krivine-Stengle and neither used any of the abstract Artin-Schreir theory which they depended on.

Note also that when S is compact we can always choose $N \geq 0$ such that $g = N - \sum_i X_i^2 \geq 0$ on S. Adding the single redundant inequality $g \geq 0$ to the definition of S we obtain the quadratic module $M[g_1, \ldots, g_m, g]$ which is automatically Archimedian. Hence if $f > 0$ on S then $f \in M[g_1, \ldots, g_m, g]$.

However we might be in the situation that the bound N is not known to us. In this case Schmüdgen's Theorem [27] below guarantees Archimedianness if we add a great number of redundant inequalities: namely by going from the quadratic module $M[B]$ to the preoder $T[B]$.

When the Archimedian property is removed from the assumptions we do not avoid abstract theory behind the Krivine-Stengle Positivstellensatz.

The original proof of Schmüdgen [27] uses only functional analysis and does not mention quadratic modules at all. Using Theorem 3.4 it is possible to give a purely algebraic proof. We present the algebraic version below which is due to Wörman [33] and is also found in [17].

Theorem 3.16 (Schmüdgen's Positivstellensatz) *Suppose $S = \mathcal{W}(g_1, \ldots, g_m)$ is compact and let $T = T[g_1, \ldots, g_m]$. Then*

$$f > 0 \text{ on } S \implies f \in T \tag{3.41}$$

for all $f \in \mathbb{R}[\underline{X}]$.

Proof We recall that the preorder $T = T[g_1, \ldots, g_m]$ is a finitely generated quadratic module. By Proposition 3.13 we can equivalently show that T is Archimedian.

Choose N such that $g = N - \sum_{i=1}^{n} X_i^2 > 0$ on S. Then by Theorem 2.48 *ii*) we get a relation of the form

$$1 + a_1 = a_2 g \tag{3.42}$$

for some $a_1, a_2 \in T[g_1, \ldots, g_m]$.

This implies that

$$(1 + a_1)T' \subseteq T, \tag{3.43}$$

where $T' = M[g] = \sum \mathbb{R}[\underline{X}]^{(2)} + g \sum \mathbb{R}[\underline{X}]^{(2)}$. Consider namely $\sigma_1, \sigma_2 \in \sum \mathbb{R}[\underline{X}]^{(2)}$. Then

$$\begin{aligned}(1 + a_1)(\sigma_1 + \sigma_2 g) &= \sigma_1(1 + a_1) + \sigma_2 g(1 + a_1) \\ &= \sigma_1(1 + a_1) + \sigma_2 g^2 a_2 \in T.\end{aligned}$$

Further more

$$g + Na_1 = g + \Big(g + \sum_{i=1}^{n} X_i^2\Big)a_1 = (1 + a_1)g + \Big(\sum_{i=1}^{n} X_i^2\Big)a_1 \in T \tag{3.44}$$

by (3.43).

As T' is Archimedian by Proposition 3.13 *ii*), we can choose N' such that $N' - a_1 \in T'$. Now the identity

$$(N' + 1)(N' - a_1) = \underbrace{(1 + a_1)(N' - a_1)}_{\in (1+a_1)T' \subseteq T} + (N' - a_1)^2 \in T$$

reveals that

$$N' - a_1 \in T \tag{3.45}$$

since we can divide by $1 + N' > 0$ and stay in T.

Finally

$$N(N' + 1) - \sum_{i=1}^{n} X_i^2 = NN' + g = \underbrace{(g + Na_1)}_{\in T \text{ by } (3.44)} + \underbrace{N(N' - a_1)}_{\in T \text{ by } (3.45)} \in T, \tag{3.46}$$

and so T is Archimedian by Proposition 3.13. □

Example 3.17 Consider again $g_1 = X_1 - \frac{1}{2}$, $g_2 = X_2 - \frac{1}{2}$, and $g_3 = 1 - X_1X_2$. We saw before that $M = M[g_1, g_2, g_3]$ was not Archimedian although $S = \mathcal{W}(g_1, g_2, g_3)$ was compact.

By Theorem 3.16 any f which is positive on S can be written in the form:

$$f = \sigma_1 + \sigma_2 g_1 + \sigma_3 g_2 + \sigma_4 g_3 + \sigma_5 g_1 g_2 + \sigma_6 g_1 g_3 + \sigma_7 g_2 g_3 + \sigma_8 g_1 g_2 g_3$$

for some σ_i which are sums of squares.

In this example it is not difficult to see that $S \subseteq [0,2] \times [0,2]$ so that $g_4 = 8 - X_1^2 - X^2$ is positive on S. Thus $S = \mathcal{W}(g_1, g_2, g_3, g_4)$ and Proposition 3.13 ensures that $\tilde{M} = M[g_1, g_2, g_3, g_4]$ is Archimedian. Therefore Theorem 3.4 guarantees a certificate of the form

$$f = \sigma_0 + \sigma_1 g_1 + \sigma_2 g_2 + \sigma_3 g_3 + \sigma_4(8 - X^2 - Y^2) \tag{3.47}$$

for any f which is positive on S.

Consider now

$$f = X_1X_2 - \tfrac{1}{2}X_1 - \tfrac{1}{2}X_2 + \tfrac{5}{4} = g_1g_2 + 1 \in T = T[g_1, g_2, g_3].$$

Clearly $f > 0$ on S. However $f \notin M$ since

$$\Psi(f) = 1 \cdot \eta(1,1) = -1 < 0$$

(see Example 3.12).

The relation $f = g_1g_2 + 1$ is a very simple certificate of membership in T and hence non-negativity on S. It does not certify membership in $\tilde{M}$ however, since it is not on the form (3.47). We know that a certificate on that form exist, but it is hard to believe that it will be equally simple. ○

3.4 The Moment Problem

We round off this chapter by illustrating an interesting connection between the algebraic theory we have seen so far and the field of functional analysis. From the functional analysis point of view, characterizing positive polynomials amounts to characterize linear functionals $L : \mathbb{R}[\underline{X}] \to \mathbb{R}$ which can be expressed as integration with respect to a probability (or Borel) measure μ, meaning that $L(f) = \int f \, d\mu$ for all $f \in \mathbb{R}[\underline{X}]$. The problem of characterizing such linear functionals is also known as a Moment Problem.

This dual point of view becomes useful when we study constraint optimization in Chapter 5.

Let $S \subseteq \mathbb{R}^n$ be a closed set. The Moment Problem is the problem of finding necessary and sufficient conditions for a linear functional $L : \mathbb{R}[\underline{X}] \to \mathbb{R}$ to be expressible by a Borel measure[4] μ on S in terms of integration, i.e. such that:

$$L(f) = \int_S f \, d\mu, \qquad \forall f \in \mathbb{R}[\underline{X}]. \tag{3.48}$$

A linear functional $L : \mathbb{R}[\underline{X}] \to \mathbb{R}$ is uniquely determined by it values on the monomials in $\mathcal{T}^n$, since $L(\sum_{\alpha \in \mathbb{N}^n} b_\alpha \underline{X}^\alpha) = \sum_{\alpha \in \mathbb{N}^n} b_\alpha L(\underline{X}^\alpha)$. So L can be represented by the multi-sequence $(a_\alpha)_{\alpha \in \mathbb{N}^n}$ where $a_\alpha := L(\underline{X}^\alpha)$. In that way the Moment Problem can be reformulated in a way that explains its name: Given a multi-sequence $(a_\alpha)_{\alpha \in \mathbb{N}^n} \in \mathbb{R}^{\mathbb{N}^n}$, can this sequence be realized as the moments $\int_S X^\alpha \, d\mu$ for some Borel measure μ?

The following theorem of Haviland [17, Cha. 3] gives a solution to the general Moment Problem in terms of non-negative polynomials:

Theorem 3.18 (Haviland 1935) *Let $L : \mathbb{R}[\underline{X}] \to \mathbb{R}$ be a linear functional and $S \subseteq \mathbb{R}^n$ a closed set. Then the following is equivalent*

i) L is expressible by a Borel measure μ as in (3.48).

ii) $L(f) \geq 0$ for all $f \in \mathbb{R}[\underline{X}]$ such that $f \geq 0$ on S.

Proof The proof is based on some deep results from functional analysis which are out of scope here. See instead [17, Sec. 3.2]. In Appendix A.2 we give a proof under the stronger assumption that S is compact and $L(1) = 1$. This is what is assumed in Theorem 3.19 below. □

One could argue that this is not entirely a *solution* to the Moment Problem since in general the condition *ii*) is not particularly easy to verify. However in combination with the theorems in the earlier chapters we can get good solutions to a broad class of moment problems. Particularly Putinar's Positivstellensatz Theorem 3.4 gives the following solution in case S is a compact semi-algebraic set such that the associated quadratic module is Archimedian. The following result appears in [28]:

Theorem 3.19 *Let $L : \mathbb{R}[\underline{X}] \to \mathbb{R}$ be linear and suppose $S = \mathcal{W}(B)$ where $B = \{g_1, \ldots, g_m\} \subseteq \mathbb{R}[\underline{X}]$ and $M = M[B]$ is Archimedian. The following is equivalent:*

i) L is expressible by a probability *measure μ as in* (3.48).

ii) $L(M) \subseteq [0, \infty)$ and $L(1) = 1$.

[4] I.e. μ is a σ-additive map $\mathcal{B}(S) \to [0, \infty]$ with $\mu(\emptyset) = 0$. Here $\mathcal{B}(S)$ is the σ-algebra in S generated by the open sets in S [31, Def. 4.4.4])

Proof Clearly $i) \implies ii)$ since any $f \in M$ is non-negative on S and hence $L(f) = \int_S f\, d\mu \geq 0$. Also $L(1) = \int_S 1\, d\mu = \mu(S) = 1$ as μ was a probability measure.

For the other direction, suppose $L(M) \subseteq [0, \infty)$ and $L(1) = 1$. Then by linearity and the property $L(1) = 1$ we get for any $f \geq 0$ on S:

$$L(f) = \lim_{\epsilon \to 0^+} (L(f) + \epsilon) = \lim_{\epsilon \to 0^+} L(f + \epsilon) \geq 0, \tag{3.49}$$

where we used that $f + \epsilon > 0$ on S so $f + \epsilon \in M$ by Theorem 3.4. Then by Theorem 3.18 L is expressible by a Borel measure μ on S. Further more $\mu(S) = \int_S 1\, d\mu = L(1) = 1$, so μ is a probability measure on S. □

For example in case $n = 1$ and $S = [0, 1]$ we get:

Corollary 3.20 (Hausdorff 1923) *$L : \mathbb{R}[X] \to \mathbb{R}$ is represented by a probability measure on $[0, 1]$ if and only if $L(1) = 1$ and $L(\sigma_0 + \sigma_1 X + \sigma_2(1 - X)) \geq 0$ for any $\sigma_0, \sigma_1, \sigma_2 \in \sum \mathbb{R}[X]^2$.*

4 Computing certificates

In the previous chapters proved the existence of different types of sums of squares-based certificates. Some of the proofs were far from being constructive as they were heavily dependent on Zorn's Lemma. However it turns out that the close connection between sums of squares and semi-definite matrices which we saw in Chapter 1 allows one to efficiently compute certificates by the use of a tool from convex optimization known as semi-definite programming (SDP).

In this chapter we will briefly introduce SDP and examine some of its many applications. The main goal is to compute sos-certificates. Later this will be used to efficiently solve (or approximate) difficult and non-convex polynomial optimization problems.

4.1 Semidefinite programming

Semidefinite programs form a certain class of convex optimization problems [2] [3]. A *primal* semidefinite programming problem (SDP) takes the form

$$\begin{aligned} \text{Minimize}_G \quad & \langle C, G \rangle \\ \text{s.t.} \quad & \langle A_i, G \rangle = b_i, \quad i = 1, \ldots, m \\ & G \succeq 0 \end{aligned} \tag{4.1}$$

where C and A_i are symmetric matrices and the decision variable G is also real symmetric. So we are minimizing a linear function $\langle C, G \rangle$ on the feasible set which is the intersection of cone $\mathcal{S}_+^n$ with an affine subspace of $\mathcal{S}^n$ cut out by the relations $\langle A_i, G \rangle = b_i$. Such a set is called a *spectrahedron* – see e.g. Figure 4.1 for an example. Note that any spectrahedron is convex being the intersection of convex sets. Also, according to Proposition 1.3 (iv) spectrahedra are a basic closed semialgebraic sets.

There has been developed and implemented efficient algorithms to numerically solve SDPs. Thus the solvers will return a numerical solution within a specified accuracy $\epsilon > 0$, i.e. a $G \succeq -\epsilon I$ (G is *almost* semidefinite) with $\langle C, G\rangle \leq \text{opt} + \epsilon$ (the solution is *almost optimal*), where opt is the optimal solution to (4.1). We will mainly use SDP as a black-box tool and avoid going into the complexity analysis and the underlying algorithmic theory behind the solvers – see e.g. [2, Chapter 5, 6] for an exposition.

Like in linear programming each primal SDP problem has an associated *dual problem*:

$$\begin{aligned} \text{Maximize}_y \quad & \sum_{i=1}^{m} b_i y_i \\ \text{s.t.} \quad & C - \sum_{i=1}^{m} A_i y_i \succeq 0, \end{aligned} \tag{4.2}$$

where $y \in \mathbb{R}^m$. Feasible solutions to the dual problem gives lower bounds on the primal optimal solution and feasible solutions to the primal bound the optimal solution to the dual from abouve. This is what is known as *weak duality*:

Theorem 4.1 (Weak duality) *For any G and $y \in \mathbb{R}^m$ which are feasible for* (4.1) *and* (4.2) *respectively, we have*

$$\sum_{i=1}^{m} b_i y_i \leq \langle C, G\rangle. \tag{4.3}$$

Proof Let G and $y \in \mathbb{R}^m$ be feasible for (4.1) and (4.2) respectively. Then

$$\langle C, G\rangle - \sum_{i=1}^{m} b_i y_i = \langle C, G\rangle - \sum_{i=1}^{m} \langle A_i, G\rangle y_i = \Big\langle \underbrace{C - \sum_{i=1}^{m} A_i y_i}_{\succeq 0}, G \Big\rangle \geq 0, \tag{4.4}$$

where the first equality follows from feasibility of G in (4.1) and the inequality follows from Proposition 1.3 *vi*). □

Example 4.2 The set

$$S = \{A \in \mathcal{S}^3 \mid A \succeq 0 \text{ and } A_{ii} = 1,\ i = 1, 2, 3\}, \tag{4.5}$$

is a spectrahedron. Any matrix $A \in S$ takes the form

$$A = \begin{bmatrix} 1 & a & b \\ a & 1 & c \\ b & c & 1 \end{bmatrix}, \tag{4.6}$$

and since all principle minors are non-negative by Proposition 1.3, a, b, and c must satisfy:

$$1 - a^2 \geq 0, \quad 1 - b^2 \geq 0, \quad 1 - c^2 \geq 0$$
$$1 - a^2 - b^2 - c^2 + 2abc \geq 0.$$

○

In Figure 4.1 we have pictured S as a subset of $\mathbb{R}^3$.

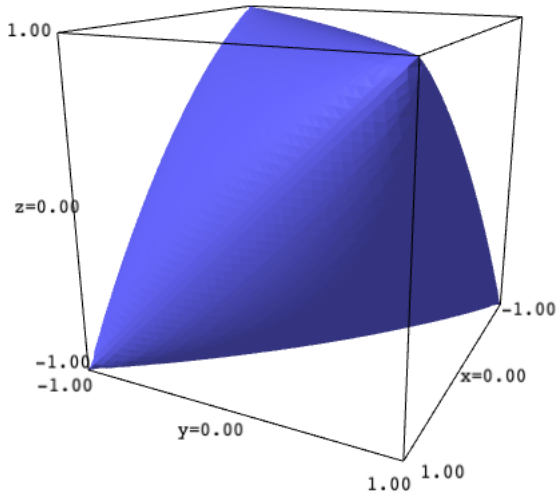

Figure 4.1: The spectrahedron S.

Multiple constraints

Given $C_j, A_{ij} \in \mathcal{S}^{n_j}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, k$, consider the optimization problem

$$\begin{aligned} \text{Minimize}_{G_j} \quad & \sum_{j=1}^{k} \langle C_j, G_j \rangle \\ \text{s.t.} \quad & \sum_{j=1}^{k} \langle A_{ij}, G_j \rangle = b_i, \quad i = 1, \ldots, m \\ & G_j \succeq 0, \; j = 1, \ldots, k. \end{aligned} \tag{4.7}$$

This can be formulated as a single SDP. We can namely form the block diagonal matrices

$$A_i = \mathrm{Diag}(A_{i1}, \ldots, A_{ik}) = \begin{bmatrix} A_{i1} & & 0 \\ & \ddots & \\ 0 & & A_{ik} \end{bmatrix} \in \mathcal{S}^n \tag{4.8}$$

and $C = \mathrm{Diag}(C_1, \ldots, C_j) \in \mathcal{S}^n$, where $n = n_1 + \cdots + n_k$. Then (4.7) is equivalent to (4.1) with $n = n_1 + \cdots + n_k$: If $G_1, \ldots, G_k$ are feasible for (4.7) then $G = \mathrm{Diag}(G_1, \ldots, G_k)$ is feasible for (4.1):

$$\mathrm{Diag}(G_1, \ldots, G_k) \succeq 0 \iff G_j \succeq 0,\ j = 1, \ldots, k \tag{4.9}$$

follows from Proposition 1.3 since any principal minor of G is either 0 or a principal minor of G_j for some j. We also have the relation

$$\sum_{j=1}^{k} \langle A_{ij}, G_j \rangle = \langle A_i, G \rangle. \tag{4.10}$$

Since the objective agree, i.e.

$$\sum_{j=1}^{k} \langle C_j, G_j \rangle = \langle C, G \rangle, \tag{4.11}$$

the optimal value of (4.1) is greater than or equal to the optimal value of (4.7). On the other hand, if

$$G = \begin{bmatrix} G_1 & & * \\ & \ddots & \\ * & & G_k \end{bmatrix} \in \mathcal{S}^n \tag{4.12}$$

is feasible for (4.1) with arbitrary values away from the blog diagonal, then $G_1, \ldots, G_k$ are feasible for (4.7) and (4.11) still holds. Thus the optimal values are equal.

The dual problem associated to (4.7) is equivalent to

$$\begin{aligned} \text{Maximize}_y \quad & \sum_{i=1}^{m} b_i y_i \\ \text{s.t.} \quad & C_j - \sum_{i=1}^{m} A_{ij} y_i \succeq 0,\ j = 1, \ldots, k, \end{aligned} \tag{4.13}$$

by (4.9) with $G_j = C - \sum_{i=1}^{m} y_i A_{ij}$.

4.2 SOS-certificates through SDP

In Chapter 1 we saw that SOS-polynomials and semidefinite matrices were closely connected. Given an $f = \sum_{\alpha \in \mathbb{N}_d^n} b_\alpha \underline{X}^\alpha \in \mathbb{R}[\underline{X}]_{2d}$ we could determine if it was SOS by deciding if

$$\mathcal{L}_f \cap \mathcal{S}_+^{\mathbb{N}_d^n} \neq \emptyset. \tag{4.14}$$

We can write the affine subspace $\mathcal{L}_f$ in terms of inner products and see that

$$\mathcal{L}_f \cap \mathcal{S}_+^{\mathbb{N}_d^n} = \Big\{ G \in \mathcal{S}_+^{\mathbb{N}_d^n} \Big| \sum_{\substack{\beta,\gamma \in \mathbb{N}_d^n \\ \beta+\gamma=\alpha}} G_{\beta\gamma} = b_\alpha, \quad \forall \alpha \in \operatorname{supp}(f) \Big\} \tag{4.15}$$

$$= \{ G \in \mathcal{S}_+^{\mathbb{N}_d^n} \mid \langle A_\alpha, G \rangle = b_\alpha, \quad \alpha \in \mathbb{N}_{2d}^n \}, \tag{4.16}$$

where $A_\alpha \in \mathcal{S}^{\mathbb{N}_d^n}$ is given by

$$A_\alpha(\beta, \gamma) = \begin{cases} 1 & \text{if } \beta + \gamma = \alpha, \\ 0 & \text{otherwise.} \end{cases} \tag{4.17}$$

In other words we can determine if f is SOS by solving the semidefinite program:

$$\begin{aligned} \text{Minimize} \quad & \langle 0, G \rangle \\ \text{s.t.} \quad & \langle A_\alpha, G \rangle = b_\alpha, \quad \alpha \in \mathbb{N}_{2d}^n \\ & G \succeq 0 \end{aligned} \tag{4.18}$$

The objective function here is simply $\langle 0, G \rangle = 0$ which turns the SDP into a feasibility problem.

SOS-programs

By parameterizing the coefficients of polynomials the SOS/SDP relation just discussed, can be pushed further to a nice optimization framework, known as SOS-programming [3].

An SOS-program can be formulated as

$$\begin{aligned} \text{Maximize}_y \quad & \sum_{i=1}^{m} b_i y_i \\ \text{s.t.} \quad & f_j(\underline{X}; y) \in \sum \mathbb{R}[\underline{X}]^{(2)}, \; j = 1, \ldots, k \end{aligned} \tag{4.19}$$

where $f_j(\underline{X}; y) = c_j + a_{j1} y_1 + \cdots + a_{jm} y_m$ and $c_j, a_{ji} \in \mathbb{R}[\underline{X}]$. I.e. f_j is a polynomial in $\mathbb{R}[\underline{X}]$ with coefficients parametrized affinely in the variables

$y_1, \ldots, y_k$. Note the similarity to the dual SDP formulation. We can think of an SOS-program as the problem maximizing a linear function over the intersection of the convex cone $\sum \mathbb{R}[\underline{X}]^{(2)} \subseteq \mathbb{R}[\underline{X}]$ and the k affine subspaces of $\mathbb{R}[\underline{X}]$ parametrized by y. Each of the constraints $f_j(\underline{X}, y) \in \sum \mathbb{R}[\underline{X}]^{(2)}$ translates into SDP constraints and SOS-problems are in that way just SDPs in disguise.[1] The parsing from an SOS-program to an SDP has been automated and is available in the Matlab packages `SOSTOOL` [6] and `YALMIP` [15] [16]. The next example is based on an exercise in [3].

Example 4.3 Consider the butterfly curve $x^6 + y^6 - x^2 = 0$. We want to find a minimal disc containing its real points. I.e. we will minimize $\lambda \in \mathbb{R}$ such that $\lambda - X^2 - Y^2 > 0$ on $\mathbb{V}(\langle f \rangle)$ where $f = X^6 + Y^6 - X^2$. This is equivalent to minimize λ such that $\lambda - X^2 - Y^2 > 0$ on $S = \mathcal{W}(-f)$ which is just the "filled" butterfly.

As S is compact Schmüdgen's Positivstellensatz Theorem 3.16 implies that $\lambda - X^2 - Y^2 > 0$ on S if and only if

$$\lambda - X^2 - Y^2 \in T[-f] = \sum \mathbb{R}[\underline{X}]^{(2)} - f \sum \mathbb{R}[\underline{X}]^{(2)} \tag{4.20}$$

So our problem can be formulated as the following

$$\begin{aligned} \text{Maximize}_{\lambda,\sigma} \quad & \lambda \\ \text{s.t.} \quad & \lambda - X^2 - Y^2 - \sigma f \in \sum \mathbb{R}[\underline{X}]^{(2)} \\ & \sigma \in \sum \mathbb{R}[\underline{X}]^{(2)} \end{aligned} \tag{4.21}$$

In order to express this as an SOS-program we need to parametrize σ. If we let σ be a general degree 2 polynomial, we can parametrize it as $\sigma_c = \sigma(X, Y; c) = c_0 + c_1 X + c_2 Y + c_3 X^2 + c_4 XY + c_5 Y^2$. Then the problem can be expressed as the SOS-program

$$\begin{aligned} \text{Maximize}_{\lambda,c} \quad & \lambda \\ \text{s.t.} \quad & \lambda - X^2 - Y^2 - \sigma_c f \in \sum \mathbb{R}[\underline{X}]^{(2)} \\ & \sigma_c \in \sum \mathbb{R}[\underline{X}]^{(2)} \end{aligned} \tag{4.22}$$

○

This can be solved with a small script in `YALMIP`:

```
sdpvar X Y lambda;
[sigma, c] = polynomial([X, Y], 2);
f = X^6 + Y^6 - X^2;
```

[1] We make these arguments more clear in Section 5.3 where we give an explicit SDP formulation of a broad class of SOS-programs

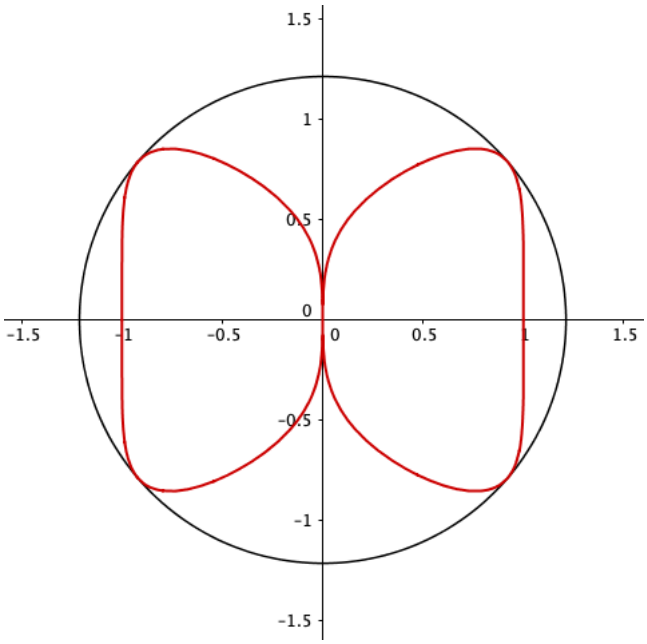

Figure 4.2: Fitting a disc to the butterfly curve.

```
F = [sos(lambda - X^2 - Y^2 - sigma*f), sos(sigma)];
solvesos(F,lambda,[],[c;lambda]);
value(lambda)
```

This first outputs a bunch of statistics on how things went (they went fine) and eventually it returns the value $\lambda = 1.4679$. According to Figure 4.2 this seems like a satisfiable result – we could have been in the situation that we got a an upper bound and had to increase the size (i.e. the degree of σ) of the problem.

4.3 Exact certificates and assisted theorem proving

As semidefinite programming is a numerical technique it will not give us exact solutions. This is often not a problem in optimization applications, but if we want SOS-certificates in pure mathematics then we can not be satisfied with floating points and error. What we want are exact rational certificates.

For this we need to improve Lemma 1.6 to polynomials with rational coefficients. This result is based on an exercise in [3].

Lemma 4.4 *$f \in \mathbb{Q}[\underline{X}]_{2d}$ is a admits a rational SOS-decomposition, i.e. $f = \sum_i r_i f_i^2$ for some $f_i \in \mathbb{Q}[\underline{X}]$ and $r_i \in \mathbb{Q}^+$, if and only if there exists a psd $G \in \mathcal{S}_+^{\mathbb{N}_d^n}$ with rational entries such that*

$$f = z^T G z. \tag{4.23}$$

Here we mention the LDL decomposition: If G is psd then one can always write $G = L^T D L$ where L is lower triangular and D is diagonal with non-negative entries. The advantage of the LDL decomposition is that it can be computed without the need of taking square roots ([32]) so if G is rational then so are L and D.

Proof If $f = z^T G z$ for a rational psd matrix G then we write $G = L^T D L$ and obtain

$$f = z^T G z = z^T L^T \sqrt{D}^T \sqrt{D} L z = \|\sqrt{D} L z\|^2 = \sum_{i=1}^{n} D_{ii} [Lz]_i^2 \tag{4.24}$$

which is a rational SOS-decomposition.

Similarly if $f = \sum_{i=1}^m r_i f_i^2$ then we can define $L \in \mathbb{R}^{m \times \mathbb{N}_d^n}$ such that $f_i = [Lz]_i$ and define $G = L^T D L$ where $D = \text{Diag}(r_1, \ldots, r_m)$. Then $f = z^T G z$ by (4.24). □

Round and project procedure

We will now describe a procedure to obtain an exact rational SOS-decomposition of an $f \in \mathbb{Q}[\underline{X}]_{2d}$ from a numerical one. The main idea is to round and project the Gram matrix to obtain something which satisfies Lemma 4.4. The material in this subsection is based on exercises in [3].

We make the general assumptions that f admits a rational SOS-decomposition and that the semidefinite program (4.18) is strictly feasible.

Let $\tau > 0$ be the accuracy of the SDP-solver. Then the numerical solution G to (4.18) will satisfy $\|G - \pi_f(G)\|_F \leq \tau$ where π_f is the orthogonal projection onto the affine subspace $\mathcal{L}_f$ and $\|\cdot\|_F$ is the Frobenius norm, i.e. the distance from G to $\mathcal{L}_f$ is less that τ. As the program is strictly feasible the solver will ensure that G is positive definite, i.e. it is in the interior of the semidefinite cone and hence if $\|G - G'\|_F \leq \epsilon$ then $G' \succeq 0$.

One can then approximate G with a rational matrix $\tilde{G}$ such that $\|G - \tilde{G}\|_F \leq \delta$ for some small $\delta > 0$.[2] Then $\pi_f(\tilde{G}) \in \mathcal{L}_f$ has rational entries because $\mathcal{L}_f$ is defined by rational datum: In general one can show that, if

[2]There are more or less naive approaches to rational approximations which we will not go into.

$\mathcal{L} = \{x \in \mathbb{R}^m \mid Ax = b\}$ where $A \in \mathbb{Q}^{n\times m}$ and $b \in \mathbb{Q}^n$ and A has full rank, i.e. the rows of A are independent, then the projection of x_0 to $\mathcal{L}$ is given by

$$\pi(x_0) = x_0 - A^T(AA^T)^{-1}(Ax_0 - b). \tag{4.25}$$

It follows from Cramer's rule that $(AA^T)^{-1} \in \mathbb{Q}^{n\times n}$, so if x_0 is rational, then so is $\pi(x_0)$. In our situation with $\mathcal{L} = \mathcal{L}_f$, we have $m = |\mathbb{N}^n_d \times \mathbb{N}^n_d|$, and $n = |\mathbb{N}^n_{2d}|$. Our x_0 is $\tilde{G}$, the vector b is the coefficients $(b_\alpha)_{\alpha\in\mathbb{N}^n_{2d}}$ of f, and the n "rows" of the matrix A are the matrices A_α defined in (4.17). These are clearly linear independent: Suppose

$$\sum_{\alpha\in\mathbb{N}^n_{2d}} a_\alpha A_\alpha = 0 \in \mathcal{S}^{\mathbb{N}^n_d}. \tag{4.26}$$

Let $\alpha' \in \mathbb{N}^n_{2d}$ and write $\alpha' = \beta + \gamma$ for some $\beta, \gamma \in \mathbb{N}^n_d$. Then

$$0 = \sum_{\alpha\in\mathbb{N}^n_{2d}} a_\alpha A_\alpha(\beta, \gamma) = a_{\alpha'}. \tag{4.27}$$

Now if we choose $\tau < \epsilon$ and $\delta \leq \sqrt{\epsilon^2 - \tau^2}$ (See Figure 4.3) then

$$\begin{aligned}
\|G - \pi_f(\tilde{G})\|^2 &= \|G - \pi_f(G) + \pi_f(G) - \pi_f(\tilde{G})\|^2 \\
&= \|G - \pi_f(G)\|^2 + \|\pi_f(G) - \pi_f(\tilde{G})\|^2 \\
&\leq \|G - \pi_f(G)\|^2 + \|G - \tilde{G}\|^2 \\
&= \tau^2 + \delta^2 \leq \epsilon^2.
\end{aligned}$$

This ensures that $\pi_f(\tilde{G})$ is psd and hence satisfies the conditions in Lemma 4.4.

Geometrical theorem proving

The procedure just presented has been implemented in the package `SumsOfSquares.m2` [20] in the CAS system Macauley 2 [9] which is freely available on the web. Searching on Wikipedia we found a theorem suited for an SOS-proof – this is known as the Hadwiger-Finsler inequality. The proof below is our own.

Example 4.5 Consider a triangle with side lengths $a, b, c > 0$ and area K. The Hadwiger-Finsler inequality states that:

$$a^2 + b^2 + c^2 - (a-b)^2 + (a-c)^2 + (b-c)^2 \geq 4\sqrt{3}K. \tag{4.28}$$

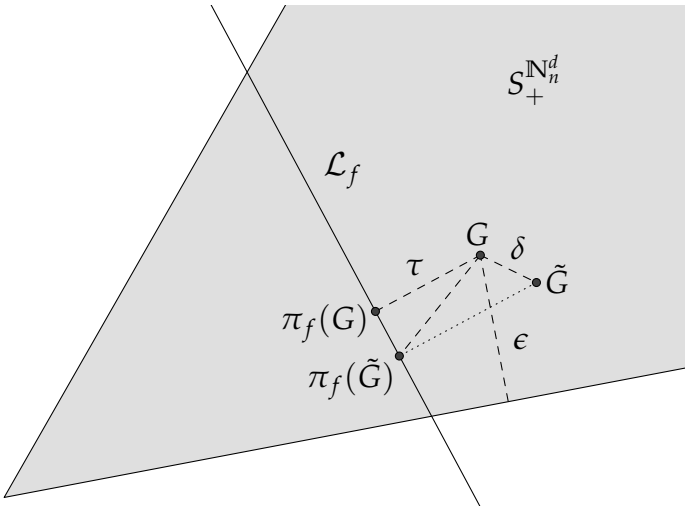

Figure 4.3: The rounded and projected matrix $\pi_f(\tilde{G})$ is psd as $\|G - \pi_f(\tilde{G})\| \leq \epsilon$.

This is a refinement of the Weitzenböck inequality:

$$a^2 + b^2 + c^2 \geq 4\sqrt{3}K. \tag{4.29}$$

We will prove (4.28) by computing an sos certificate. First we need an expression of the area in terms of a, b and c. This is given by Heron's fomula:

$$K = \tfrac{1}{4}\sqrt{(a+b+c)(-a+b+c)(a-b+c)(a+b-c)}. \tag{4.30}$$

We note that both sides of (4.28) are non-negative for valid side lengths a, b, c, since they satisfy the reverse triangle inequality $|a-b| \leq c$, $|b-c| \leq a$ and $|b-c| \leq a$. Squarring both sides and subtracting the left-hand side it then suffices to show that the polynomial

$$p(a,b,c) = \big(a^2+b^2+c^2 - ((a-b)^2+(a-c)^2+(b-c)^2)\big)^2 - (4\sqrt{3}K)^2$$

is non-negative on the basic semi-algebraic set

$$S = \{(a,b,c) \in (0,\infty]^3 \,|\, a^2-(b-c)^2 \geq 0, b^2-(a-c)^2 \geq 0, c^2-(a-b)^2 \geq 0\},$$

which defines the valid triples of side-lengths. However this constraint is not needed. In fact p is *globally* non-negative. Using the package SumsOfSquares.m2 we find the SOS-decomposition:

$$p(a,b,c) = \big(2a^2 + b(a+b-c) + c(a-b+c)\big)^2 + 3\big(b(a-b) + c(a-c)\big)^2,$$

This is a simple certificate for (4.28). Although it is difficult to make any geometrical sense of this formula other than it proves the theorem, it does provide some extra information, namely that equality holds if and only if the triangle is equilateral: If the equality holds then p vanishes and consequently $b(a-b)+c(a-c)=0$. As $a,b,c>0$ we immediately get $a-b=0$ and $a-c=0$, so $a=b=c$. ○

4.4 The Max-Cut Problem

In this section we will leave the topic of certificates and show another very nice application of semidefinite programming: approximation algorithms. The Max-Cut Problem it a combinatorial problem in graph theory, which is known to be NP-hard. We will see how SDP can be used to attack this problem efficiently. We will return to the Max-Cut Problem at the end of Chapter 5 when we discuss constrained polynomial optimization. In this section we present the original approach which was discovered by Williamson and Goemans [8]. Our exposition is mainly based on [17].

Problem formulation

We consider an undirected graph $G=(V,E)$ where $V=\{1,\dots,n\}$ is the set of *vertices* and $E\subseteq V\times V$ is the set of *edges* in G. An edge between vertex i and j is represented by the pair (i,j) if $i<j$ and (j,i) otherwise. A *cut* in G is a subset of the edges of the form $C=C_x=\{(i,j)\in E\mid x_i\neq x_j\}$ where $x\in\{-1,1\}^n$ (See e.g Figure 4.5 and 4.6). Here we can think of x as representing the partion $V=V_+\cup V_-$ given by $V_+=\{i\in V\mid x_i=1\}$ and $V_-=\{i\in V\mid x_i=-1\}$. In that sense C consist of edges connecting the two components V_+ and V_-. The cardinality of a cut C_x can be expressed

$$|C_x|=\sum_{(i,j)\in E}1-\delta_{x_i,x_j}=\sum_{(i,j)\in E}\tfrac{1}{2}(1-x_ix_j), \tag{4.31}$$

where $\delta_{x,y}$ is the Dirac delta function.

The Max-Cut Problem asks to find the maximal cardinality $|C|$ among any cut C in G, i.e. to compute

$$\text{opt}=\max\{|C_x|\mid x\in\{-1,1\}^n\}. \tag{4.32}$$

Relaxing the Max-Cut Problem

As Max-Cut is NP-hard the only realistic goal is to approximate opt. The quality of an approximation algorithm is measured by means of the *approximation*

ratio ρ which is the ratio between the output of the algorithm and opt (we choose the direction of the ratio such that $\rho \geq 1$). A really trivial approximation algorithm for Max-Cut is to simply output $|E|$ without considering any other structure of G. This algorithm has an approximation ratio of $\rho = 2$ because of the following:

Lemma 4.6 *Any graph $G = (V, E)$ admits a cut C of cardinality $|C| \geq \frac{1}{2}|E|$.*

Proof If we choose $x \in \{-1, 1\}^n$ uniformly at random, e.g. by flipping a coin for each vertex, then the expected cardinality of the associated random cut C_x is

$$\mathbb{E}[|C_x|] = \mathbb{E}\Big[\sum_{(i,j)\in E} 1 - \delta_{x_i,x_j}\Big] = \sum_{(i,j)\in E} \mathbb{E}[1 - \delta_{x_i,x_j}] = \sum_{(i,j)\in E} \tfrac{1}{2} = \tfrac{1}{2}|E|. \quad (4.33)$$

As the expectation is $\frac{1}{2}|E|$ then at least one cut has cardinality at least $\frac{1}{2}|E|$. □

In order to improve the trivial algorithm above we can cast the Max-Cut problem as a quadratic programming problem with integer constraints:

$$\begin{aligned} \text{Maximize} \quad & \sum_{(i,j)\in E} \tfrac{1}{2}(1 - x_i x_j) \\ \text{s.t.} \quad & x_i \in \{-1, 1\}, \quad i = 1, \ldots, n \end{aligned} \quad (4.34)$$

This is just a reformulation of the problem so it still needs some kind of relaxation. One approach is to replace the integer constraints $x_i \in \{-1, 1\}$ with the inequalities $-1 \leq x_i \leq 1$. One could then hope that the relaxed optimal value is close to the true optimum. However this is not the case. The relaxed problem is very easy to solve: simply put $x_i = 0$ for all $i = 1, \ldots, n$. Then the optimal value is $|E|$, which does not get us any closer than the trivial approximation.

Actually it was not until 1995 that an improvement appeared. Williamson and Goemans used a semi-definite relaxation of (4.34) to get an approximation ratio of $\rho_{WG} \approx 1.1382$ which is given as the smallest constant such that

$$\frac{1 - \cos(x)}{2} \leq \frac{\rho_{WG}}{\pi} x, \quad \forall x \in [0, \pi]. \quad (4.35)$$

Their relaxation is as follow:

$$\begin{aligned} \text{Maximize} \quad & \sum_{(i,j)\in E} \tfrac{1}{2}(1 - \langle v_i, v_j \rangle) \\ \text{s.t.} \quad & \|v_i\| = 1,\ v_i \in \mathbb{R}^n, \quad i = 1, \ldots, n \end{aligned} \quad (4.36)$$

We let opt_{WG} denote the solution to (4.36).

Theorem 4.7 *The problem* (4.36) *can be solved using semi-definite programming (and hence in polynomial time) and the solution* opt_{WG} *satisfies* $\text{opt}_{WG} \in [\text{opt}, \rho_{WG} \cdot \text{opt}]$.

Proof First we note that (4.36) can be formulated as

$$\begin{aligned} \text{Maximize} \quad & \sum_{(i,j)\in E} \tfrac{1}{2}(1 - X_{ij}) \\ \text{s.t.} \quad & X_{ii} = 1, \quad i = 1, \dots, n \\ & X \succeq 0. \end{aligned} \tag{4.37}$$

This is because X is psd if and only if $X = V^T V = (\langle v_i, v_j \rangle)_{ij}$ for some $v_i \in \mathbb{R}^n$ by Proposition 1.3 (appending zeroes to v_i if $m = \text{rk}\, X < n$) and $X_{ii} = 1$ if and only if $\|v_i\|^2 = \langle v_i, v_i \rangle = 1$. It now suffices to solve the semidefinite program

$$\begin{aligned} \text{Minimize} \quad & \sum_{(i,j)\in E} X_{ij} \\ \text{s.t.} \quad & X_{ii} = 1, \quad i = 1, \dots, n \\ & X \succeq 0, \end{aligned} \tag{4.38}$$

since if opt'_{WG} is the solution to (4.38) then we can compute

$$\text{opt}_{WG} = \tfrac{1}{2}(|E| - \text{opt}'_{WG}).$$

Next we show that $\text{opt} \leq \text{opt}_{WG}$. Let $x \in \{-1, 1\}^n$ be feasible for (4.34) and define $v_i = x_i u$, $i = 1, \dots, n$, where $u \in \mathbb{R}^n$ is any unit vector. Then $v_1, \dots, v_n$ are feasible for (4.36) and $x_i x_j = \langle v_i, v_j \rangle$ so the objectives of (4.34) and (4.36) are equal. This was what we wanted.

Now for the interesting inequality: $\text{opt}_{WG} \leq \rho_{WG}\text{opt}$. Let $v_1, \dots, v_n$ be unit vectors that solve (4.36). The idea is now to construct $x \in \{-1, 1\}^n$ from $v_1, \dots, v_n$ using so called *random hyperplane rounding*: Pick a random unit vector $u \in \mathbb{R}^n$ (i.e. uniformly in the unit sphere) and let $H = \{u\}^\perp$ be the associated random hyperplane. Then define $x \in \{-1, 1\}^n$ by

$$x_i = \begin{cases} 1 & \text{if } \langle u, v_i \rangle \geq 0 \\ -1 & \text{if } \langle u, v_i \rangle < 0, \end{cases} \tag{4.39}$$

i.e. x_i is chosen according to the side of H on which v_i falls.

Let $(i, j) \in E$. The propability that $x_i \neq x_j$ is the probability that H separates v_i and v_j. This is proportional to the angle θ_{ij} between v_i and v_j (see Figure 4.4):

$$\Pr((i, j) \in C_x) = \Pr(x_i \neq x_j) = \frac{\theta_{ij}}{\pi}. \tag{4.40}$$

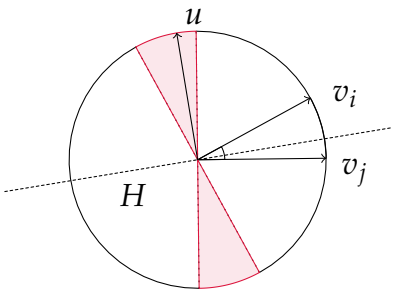

Figure 4.4: The propability that H separates v_i and v_j is $\frac{\theta_{ij}}{\pi}$.

We also have $\langle v_i, v_j \rangle = \cos(\theta_{ij})$ so

$$\sum_{(i,j)\in E} \tfrac{1}{2}(1 - \langle v_i, v_j \rangle) = \sum_{(i,j)\in E} \tfrac{1}{2}(1 - \cos(\theta_{ij})) \tag{4.41}$$

$$\leq \sum_{(i,j)\in E} \frac{\rho_{WG}}{\pi}\theta_{ij} \tag{4.42}$$

$$= \rho_{WG} \sum_{(i,j)\in E} \Pr((i,j) \in C_x) \tag{4.43}$$

$$= \rho_{WG} \cdot \mathbb{E}[|C_x|]. \tag{4.44}$$

As the above inequality holds in expectation, it also holds for at least one cut C_x. And since X was arbitrary in the feasible set, the inequality also holds for the optimal solution. □

Not only does the algorithm of Williamson and Goemans give a good estimate on the cardinality of the maximal cut, the hyperplane rounding in their proof also provide concrete cuts, namely C_x, which are good *in expectation*.

Example 4.8 Consider the graph $G = (V, E)$ which is drawn in Figure 4.6.

Using the convexpy library we solve the associated SDP as described in (4.38). This gives the optimal solution $\text{opt}_{WG} \approx 9.0258$ to (4.36). By Theorem 4.7 the maximal cut has cardinality $\text{opt} \in [\rho_{WG}^{-1}\text{opt}_{WG}, \text{opt}_{WG}] = [7.9299, 9.0258]$. As opt is an integer $\text{opt} \in \{8, 9\}$.

We have also implemented the hyperplane rounding described in the proof of Theorem 4.7. The hyperplane orthogonal to the random vector

$$u = (-0.2810, 0.3651, -0.2784, -0.3212, 0.3262, 0.6703, -0.2267),$$

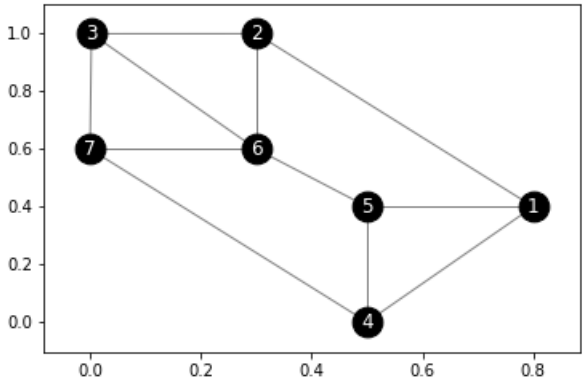

Figure 4.5: The graph G.

yielded the partition vextor $x = (-1, +1, -1, -1, +1, -1, +1)$ which gave rise to a cut C_x (see Figure 4.6) of cardinality $|C_x| = 9$ which is consequently maximal.

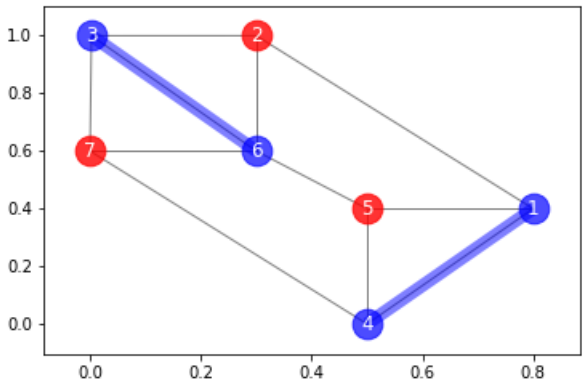

Figure 4.6: The maximal cut C_x if G obtained from hyperplane rounding.

Sampling 10.000 random unit vectors the average cut size of the induced cuts was 8.52. This is in line with the proof of Theorem 4.7 where we saw that the expected cut size is at leat $\frac{\text{opt}_{WG}}{\rho_{WG}} = 7.9299$. ○

5 Constrained polynomial optimization

In this final chapter we discuss how the presented theory and especially Theorem 3.4 and theorem 3.19 give rise to a general algorithm for solving (or approximating) constrained polynomial optimization problems, known as The Lasserre Hierarchy. As in Chapter 3 our main reference is [28].

Concretely, we consider the optimization problem of minimizing a polynomial $f \in \mathbb{R}[\underline{X}]$ on a basic closed semi-algebraic set $S = \mathcal{W}(B)$, $B = \{g_1, \dots, g_m\}$. Throughout the chapter we will have the general assumption that *the quadratic module $M = M[B]$ is Archimedian*. In particular, S is assumed to be compact c.f. Lemma 3.3.

The problem is to find (or bound) the optimal value

$$f^* = \inf\{f(x) \mid x \in S\}. \tag{5.1}$$

In Section 5.2 we further consider the problem of finding a minimizer x^*, i.e. a point satisfying $f(x^*) = f^*$.

Two equivalent problems

The optimization problem (5.1) is not assumed to be convex and S might not even be connected. This makes it very hard to solve in general, and one has to aim for an approximation.

The first step is to turn the problem into a convex one. This can be done in at least two ways.

One way is to consider the set of positive polynomials on S and realize that

$$f^* = a^* := \sup\{a \in \mathbb{R} \mid f(x) - a > 0, \forall x \in S\} \tag{5.2}$$

As $f(x) - (f^* - \epsilon) \geq \epsilon > 0$ for any $x \in S$ and $\epsilon > 0$ we see that $a^* \geq f^* - \epsilon$ for all $\epsilon > 0$ and so $a^* \geq f^*$. On the other hand $a^* = \lim_{k\to\infty} a_k$ for some sequence

(a_k) with $f(x) - a_k > 0$ for all $x \in S$ and so $f(x) - a^* = \lim_{k\to\infty} f(x) - a_k \geq 0$ for all $x \in S$. Hence $f^* - a^* = f(x^*) - a^* \geq 0$ where $x^* \in S$ is a point where the infimum is attained (using that S is compact).

The other approach is to look at the probability measures supported on S. We observe that

$$f^* = \inf\Big\{ \int_S f \, d\mu \;\Big|\; \mu \in \mathcal{M}^1(S) \Big\}. \tag{5.3}$$

where $\mathcal{M}^1(S)$ denote the set of probability measures supported on S.

To see this note that

$$f^* = f(x^*) = \int_S f \, d\delta_{x^*}$$

for some x^* where $\delta_{x_0} \in \mathcal{M}^1(S)$ is the Dirac measure

$$\delta_{x_0}(A) = \begin{cases} 1 & \text{if } x_0 \in A \\ 0 & \text{otherwise,} \end{cases} \qquad (A \in \mathcal{B}(S)). \tag{5.4}$$

So l.h.s $\geq$ r.h.s. in (5.3). On the other hand for any $\mu \in \mathcal{M}^1(S)$ we have

$$\int_S f(x)\, \mu(dx) \geq \int_S f^* \mu(dx) = f^* \mu(S) = f^*, \tag{5.5}$$

so we also have l.h.s $\leq$ r.h.s.

Both the set of positive polynomials on S and the set of probability measures on S are convex. This does not mean that the associated optimization problems are easy to solve.

Putinar's theorems help in characterizing these sets in terms of the quadratic module $M[B]$.

By Theorem 3.4, (5.2) can then be expressed as

$$f^* = \sup\{a \in \mathbb{R} \mid f - a \in M\}. \tag{5.6}$$

Likewise by Theorem 3.19, (5.3) becomes

$$f^* = \inf\{L(f) \mid L \in \chi\}. \tag{5.7}$$

where

$$\chi = \{L \colon \mathbb{R}[\underline{X}] \to \mathbb{R} \mid L \text{ is linear, } L(M) \subseteq [0, \infty), \text{ and } L(1) = 1\}. \tag{5.8}$$

5.1 The Lasserre Hierarchy

One of the obstacles with the optimization problems (5.6) and (5.7) is that M lives in $\mathbb{R}[\underline{X}]$ which is infinite dimensional. Even if f has a fixed degree d, the certificate provided by Putinar's theorem might involve polynomials of much higher degree, so we can not expect that $f \in M \cap \mathbb{R}[\underline{X}]_d$. The way around this is to make a finite dimensional approximation of M:

For $k \geq N_{\deg} := \max\{\deg(g_1), \ldots, \deg(g_m), \deg(f)\}$ we define

$$M_k = \sum_{i=0}^{m} \mathbb{R}[\underline{X}]^{(2)}_{d_i} g_i \tag{5.9}$$

$$= \left\{ \sum_{i=0}^{m} b_i g_i \;\middle|\; b_i \in \sum \mathbb{R}[\underline{X}]^{(2)} \text{ and } \deg(b_i g_i) \leq k \right\} \tag{5.10}$$

where

$$d_i = \max\{d \in \mathbb{N} \mid 2d + \deg(g_i) \leq k\}. \tag{5.11}$$

Elements in M_k are in that way the polynomials which are non-negative on S and admits a Putinar-like certificate where every involved polynomial $b_i g_i$ has degree at most k. We also define the following finite dimensional approximation of χ:

$$\chi_k = \{L \colon \mathbb{R}[\underline{X}]_k \to \mathbb{R} \mid L \text{ is linear, } L(M_k) \subseteq [0, \infty), \text{ and } L(1) = 1\}. \tag{5.12}$$

Now we get a natural pair of relaxed optimization problems:

$$\begin{aligned} \text{Minimize}_L \quad & L(f) \\ \text{s.t.} \quad & L \in \chi_k \end{aligned} \tag{Las$_k$}$$

$$\begin{aligned} \text{Maximize}_{a \in \mathbb{R}} \quad & a \\ \text{s.t.} \quad & f - a \in M_k \end{aligned} \tag{Las$'_k$}$$

We call this a k-th order relaxation of the optimization problem (5.1).

We denote by f_k^* and $\bar{f}_k^*$ the optimal solution to (Las$_k$) and (Las$'_k$) respectively. By increasing k we get finer and finer relaxations of the original optimization problem (5.1). The idea with this is that whereas (5.1) is in general NP-hard, each relaxed problem can be realized as semidefinite programs and hence solved efficiently (we show this in Section 5.3). The algorithm is called *Lasserre's method* or *The Lasserre Hierarchy*.

The next theorem shows that this method works, meaning that f_k^* and $\bar{f}_k^*$ will (in principle[1]) approximate f^* to any desired degree.

[1]We say in principle because we do not get an estimate of the complexity of the algorithm.

Theorem 5.1 *$(\bar{f}_k^*)_{k\geq N_{deg}}$ and $(f_k^*)_{k\geq N_{deg}}$ are increasing sequences which converge to f^* and satisfy*

$$\bar{f}_k^* \leq f_k^* \leq f^*, \tag{5.13}$$

for all $k \geq N_{deg}$.

Proof As $M_k \subseteq M_{k+1}$ for all $k \geq N_{\deg}$ the sequences $(\bar{f}_k^*)$ and (f_k^*) are increasing.

If L and a are in the feasible set of (Las_k) and (Las_k') respectively, then $f - a \in M_k$ so $0 \leq L(f-a) = L(f) - L(a) = L(f) - a$. Hence $\bar{f}_k^* \leq f_k^*$.

To see that $f_k^* \leq f^*$ we note that for any $x \in S$ the functional $\varepsilon_x(p) = p(x)$ wich evaluates at x is feasible for (Las_k) so $f_k^* \leq f^*$.

It remains to show that $(\bar{f}_k^*)$ converges to the optimum f^*: Let $\epsilon > 0$ be given. Then $f(x) - f^* + \epsilon > 0$ for all $x \in S$ and so $f - (f^* - \epsilon) \in M = \cup_{k\geq N_{\deg}} M_k$ by Theorem 3.4. Hence for sufficiently large K, $f - (f^* - \epsilon)$ is a candidate solution to (Las_k') so $\bar{f}_K^* \geq f^* - \epsilon$. This shows that $(\bar{f}_k^*)$ converges to f^* as claimed. □

5.2 Obtaining a minimizer

We will now see how the viewpoint of the relaxed problem (Las_k) can be used to find a minimizer x^*, i.e. a point in the set $S^* = \{x^* \in S \mid f^* = f(x^*)\}$. We note that whenever $S \neq \emptyset$ we also have $S^* \neq \emptyset$ by compactness of S and continuity of f.

Suppose for some $k \geq N_{\deg}$ that $f_k^* = L(f)$, for some feasible $L : \mathbb{R}[\underline{X}]_k \to \mathbb{R}$ for (Las_k) (in general f_k^* is an infimum over all feasible solutions and not a minimum). Assume further that L is integration with respect to a probability measure $\mu \in \mathcal{M}^1(S)$, i.e.

$$L(p) = \int_S p \, d\mu, \tag{5.14}$$

for all $p \in \mathbb{R}[\underline{X}]_k$. Again this is in general more than we can hope for since L acts on $\mathbb{R}[\underline{X}]_k$ so Theorem 3.19 does not apply.

Since f^* is the minimum of f over S we get

$$f^* \leq \int_S f \, d\mu = L(f) = f_k^* \leq f^* \tag{5.15}$$

where last inequality comes from Theorem 5.1. Hence $L_k(f) = f^*$ and μ is in fact a probability measure on S^*:

$$\begin{aligned} 0 = L(f) - f^* = L(f - f^*) = \int_S (f - f^*)\, d\mu \\ = \int_{S\setminus S^*} (f - f^*)\, d\mu + \int_{S^*} \underbrace{(f - f^*)}_{=0}\, d\mu = \int_{S\setminus S^*} \underbrace{(f - f^*)}_{>0}\, d\mu \end{aligned} \tag{5.16}$$

so $0 = \mu(S\setminus S^*) = 1 - \mu(S^*)$, i.e. $\mu \in \mathcal{M}^1(S^*)$. If f has a unique minimizer on S, i.e. $S^* = \{x^*\}$, then μ must be the Dirac measure on x^* and hence

$$(L(X_1), \ldots, L(X_n)) = x^*. \tag{5.17}$$

We will next see how one can approximate x^* under less restrictive conditions.

Definition 5.2 *Let $(L_k)_{k\geq N_{deg}}$ be a sequence of feasible solutions to* (Las_k) *such that* $\lim_{k\to\infty} L_k(f) = f^*$. *Then (L_k) is said to solve* (Las_k) nearly to optimality.

Theorem 5.3 *Suppose $S^* = \{x^*\}$ and (L_k) solves* (Las_k) *nearly to optimality. Then*

$$x^* = \lim_{k\to\infty} (L_k(X_1), \ldots, L_k(X_n)). \tag{5.18}$$

This will follow from the next general result which is a reformulation of [28, Thm. 12]. We have elaborated the compactness argument bit.

Theorem 5.4 *Let $\epsilon > 0$ and $d \in \mathbb{N}$ be given and suppose (L_k) solves* (Las_k) *nearly to optimality. Then for any $k > 0$ sufficiently large*[2]*, there exists a probability measure $\mu_k \in \mathcal{M}^1(S^*)$ such that*

$$\left\| \left(L_k(X_i^\alpha) - \int_{S^*} X_i^\alpha \, d\mu_k \right)_{|\alpha|\leq d} \right\| < \epsilon. \tag{5.19}$$

The idea in the proof is to find linear maps $\tilde{L}_k : \mathbb{R}[\underline{X}] \to \mathbb{R}$ which agree with L_k on $\mathbb{R}[\underline{X}]_t$ for some large t and then approximate $\tilde{L}_k$ by a probability measure by use of Theorem 3.19. The main ingredient is a compactness argument in which it will be helpful to represent linear maps by their infinite multi-sequences $(a_\alpha)_{\alpha\in\mathbb{N}^n}$ as introduced in Section 3.4. I.e.

$$a_\alpha = L(\underline{X}^\alpha), \qquad \alpha \in \mathbb{N}^n. \tag{5.20}$$

[2]meaning $k \geq K$ for some fixed K

Proof Let $\epsilon > 0$ and $d \in \mathbb{N}$. For each $k \geq N_{\deg}$ choose $N_k \geq N_{\deg}$ such that

$$N_k \pm \underline{X}^\alpha \in M_{N_k} \tag{5.21}$$

for all $\alpha \in \mathbb{N}^n$ with $|\alpha| = k$.

This can be done in the following way: Choose N'_k such that $N'_k \pm \underline{X}^\alpha > 0$ on S for all α with $|\alpha| = k$. Then $N'_k \pm \underline{X}^\alpha \in M = \cup_{i=1}^\infty M_i$ by Theorem 3.4 and there is an N''_k such that $N'_k \pm \underline{X}^\alpha \in M_{N''_k}$. Let $N_k = \max\{N'_k, N''_k\}$. Then $M_{N''_k} \subseteq M_{N_k}$ and $N'_k \pm \underline{X}^\alpha \leq N_k \pm \underline{X}^\alpha$ on S and (5.21) follows.

Then we define

$$Z = \prod_{\alpha \in \mathbb{N}^n} [-N_{|\alpha|}, N_{|\alpha|}], \tag{5.22}$$

which is compact by Tychonoff's Theorem. We think of elements $(a_\alpha)_{\alpha \in \mathbb{N}^n} \in Z$ as representing linear maps $L : \mathbb{R}[\underline{X}] \to \mathbb{R}$ with bounded moments.

We claim that if we choose t sufficiently large and $\delta > 0$ sufficiently small, then for any linear functional $L : \mathbb{R}[\underline{X}] \to \mathbb{R}$ which can be represented by a multi-sequence $(a_\alpha) = (L(\underline{X}^\alpha)) \in Z$ and satisfying $L(1) = 1$, $L(M_t) \subseteq [0, \infty)$, and $|L(f) - f^*| \leq \delta$, there is a $\mu \in \mathcal{M}^1(S^*)$ such that

$$\left\| \left(L(X_i^\alpha) - \int_{S^*} X_i^\alpha \, d\mu \right)_{|\alpha| \leq d} \right\| < \epsilon. \tag{5.23}$$

To show this we consider the following sets:

$$\begin{aligned}
A &= \{(a_\alpha) \in Z \mid a_0 = 1\}, \\
B &= \bigcap_{p \in M} B_p = \bigcap_{p \in M} \{(a_\alpha) \in Z \mid \sum_\alpha b_\alpha a_\alpha \geq 0 \text{ when } p = \sum_\alpha b_\alpha a_\alpha\}, \\
C &= \bigcap_{\delta > 0} C_\delta = \bigcap_{\delta > 0} \left\{ (a_\alpha) \in Z \;\middle|\; \left| \sum_\alpha c_\alpha a_\alpha - f^* \right| \leq \delta \right\} \\
&\quad = \left\{ (a_\alpha) \in Z \;\middle|\; \sum_\alpha c_\alpha a_\alpha - f^* = 0 \right\}, \\
D &= \bigcup_{\mu \in \mathcal{M}^1(S^*)} D_\mu = \bigcup_{\mu \in \mathcal{M}^1(S^*)} \left\{ (a_\alpha) \in Z \;\middle|\; \left\| \left(a_\alpha - \int_{S^*} X_i^\alpha \, d\mu \right)_{|\alpha| \leq d} \right\| < \epsilon \right\}.
\end{aligned}$$

In C we assume $f = \sum_\alpha c_\alpha \underline{X}^\alpha$. The intuition behind these sets is that they represent linear maps which satisfies the conditions in the claim above. E.g. B represent linear maps $L : \mathbb{R}[\underline{X}] \to \mathbb{R}$ satisfying $L(p) = \sum_\alpha b_\alpha a_\alpha \geq 0$ for all $p \in M$, i.e. $L(M) \subseteq [0, \infty)$.

First we observe that

$$A \cap B \cap C \subseteq D \tag{5.24}$$

Suppose namely that $(a_\alpha) \in A \cap B \cap C$. Then (a_α) represents a linear map $L : \mathbb{R}[\underline{X}] \to \mathbb{R}$ which satisfies $L(1) = 1$ and $L(M) \subseteq [0, \infty)$ so Theorem 3.19 L can be represented by a probability measure $\mu \in \mathcal{M}^1(S)$. I.e.

$$a_\alpha = L(\underline{X}^\alpha) = \int_S \underline{X}^\alpha d\mu, \tag{5.25}$$

for all $\alpha \in \mathbb{N}^n$. As we also have $(a_\alpha) \in C$ we see that $L(f) = f^*$. But then μ is in fact a probability measure on S^*, c.f. (5.16). By (5.25) we then get

$$\left\| \left(a_\alpha - \int_{S^*} X_i^\alpha \, d\mu_k \right)_{|\alpha| \leq d} \right\| = 0 < \epsilon, \tag{5.26}$$

and hence $(a_\alpha) \in D$.

Notice that the sets A, B_p, and C_δ are closed for any $p \in M$ and $\delta > 0$ and that D is open. (5.24) now yields an open cover of Z:

$$Z = A^c \cup \bigcup_{p \in M} B_p^c \cup \bigcup_{\delta > 0} C_\delta^c \cup D. \tag{5.27}$$

By compactness there is a finite subcover of Z. I.e. there are $\delta_1, \ldots, \delta_r > 0$ and $p_1, \ldots, p_s \in M$ such that

$$Z = A^c \cup \bigcup_{i=1}^{s} B_{p_i}^c \cup C_\delta^c \cup D. \tag{5.28}$$

taking $\delta = \min\{\delta_1, \ldots, \delta_r\}$. Consequently

$$E := A \cap \bigcap_{i=1}^{s} B_{p_i} \cap C_\delta \subseteq D. \tag{5.29}$$

Now take $t \geq \max\{N_{\deg}, d\}$ such that $p_i \in M_t$ for all $i = 1, \ldots, s$ (again this is possible since $M = \bigcup_k M_k$). For any $L : \mathbb{R}[\underline{X}] \to \mathbb{R}$ represented by $(a_\alpha) \in Z$ and satisfying $L(1) = 1$, $L(M_t) \subseteq [0, \infty)$, and $|L(f) - f^*| \leq \delta$, we then have $(a_\alpha) \in E \subseteq D$. Hence there is a $\mu \in \mathcal{M}^1(S^*)$ such that (5.23) holds. This proves the claim.

We now return to the sequence (L_k). This is assumed to solve (Las_k) nearly to optimality. Hence we can choose

$$K \geq \max\{t, N_0, \ldots, N_t\}, \tag{5.30}$$

so large that $|L_k(f) - f^*| \leq \delta$ for all $k \geq K$. For each $k \geq K$ we now define

$$\tilde{L}_k(\underline{X}^\alpha) = \begin{cases} L_k(\underline{X}^\alpha) & \text{if } |\alpha| \leq t, \\ 0 & \text{otherwise.} \end{cases} \tag{5.31}$$

L_k and $\tilde{L}_k$ agree on polynomials of degree $\leq t$, so we immediately get that $\tilde{L}_k(1) = 1$ and $\tilde{L}_k(M_t) \subseteq [0, \infty)$. As $\deg f \leq N_{\deg} \leq t$ and $k \geq K$ we also have $|\tilde{L}_k(f) - f^*| = |L_k(f) - f^*| \leq \delta$. Finally $(\tilde{L}_k(\underline{X}^\alpha))_\alpha \in Z$: If $|\alpha| \leq t$ then

$$N_{|\alpha|} \pm \tilde{L}_k(\underline{X}^\alpha) = N_{|\alpha|} \pm L_k(\underline{X}^\alpha) = L_k(N_{|\alpha|} \pm \underline{X}^\alpha) \geq 0, \tag{5.32}$$

by (5.21) and Theorem 3.4 where we also used that $N_{|\alpha|} \leq K \leq k$ so that $M_{N_{|\alpha|}} \subseteq M_k$. If $|\alpha| > 0$ then clearly

$$\tilde{L}_k(\underline{X}^\alpha) = 0 \in [-N_{|\alpha|}, N_{|\alpha|}]. \tag{5.33}$$

By the earlier claim there is a $\mu_k \in \mathcal{M}^1(S^*)$ such that

$$\left\| \left(L_k(X_i^\alpha) - \int_{S^*} X_i^\alpha \, d\mu_k \right)_{|\alpha| \leq d} \right\| = \left\| \left(\tilde{L}_k(X_i^\alpha) - \int_{S^*} X_i^\alpha \, d\mu_k \right)_{|\alpha| \leq d} \right\| < \epsilon, \tag{5.34}$$

here using that $d \leq t$. This finishes the proof. □

Proof (of Theorem 5.3) Take $d = 1$. As $S^* = \{x^*\}$ the Dirac measure δ_{x^*} is the only probability measure on S^*. The result then follows directly from Theorem 5.4. □

In case S is a compact polytope we can bound f^* from both below *and above*. By Proposition 3.14 we automatically have the Archimedian property, so the following theorem applies to all compact polytopes.

Theorem 5.5 *Suppose $S = \mathcal{W}(g_1, \ldots, g_m)$ is a compact polytope, i.e.* $\deg(g_i) = 1$, *and that f has a unique minimizer on S. Then the following holds:*

i) If (L_k) is feasible for (Las_k), *then*

$$f^* \in \big[L_k(f), f\big(L_k(X_1), \ldots, L_k(X_n)\big) \big]. \tag{5.35}$$

ii) If (L_k) solves (Las_k) *nearly to optimality, then the interval converges to $\{f^*\}$, i.e.*

$$\lim_{k \to \infty} L_k(f) = \lim_{k \to \infty} f(L_k(X_1), \ldots, L_k(X_n)) = f^*. \tag{5.36}$$

Proof i): Suppose L_k is feasible for (Las_k). By Theorem 5.1 we only need to show that $f^* \leq f(L_k(X_1), \ldots, L_k(X_n))$. Write $g_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$. Then

$$\begin{aligned} g_i(L_k(X_1), \ldots, L_k(X_n)) &= \beta_0 + \beta_1 L_k(X_1) + \cdots + \beta_n L_k(X_n) \\ &= L_k(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n) = L_k(g_i) \geq 0, \end{aligned}$$

so $(L_k(X_1), \dots, L_k(X_n)) \in S$. Hence $f^* \leq f(L_k(X_1), \dots, L_k(X_n))$.

ii): Suppose (L_k) solves (Las$_k$) nearly to optimality. Again by Theorem 5.1 it suffices to show that $\lim_{k\to\infty} f(L_k(X_1), \dots, L_k(X_n)) = f^*$. This follows directly from Theorem 5.3 and continuity of f:

$$\lim_{k\to\infty} f(L_k(X_1), \dots, L_k(X_n)) = f(\lim_{k\to\infty}(L_k(X_1), \dots, L_k(X_n))) = f(x^*) = f^*.$$

□

5.3 Expressing the Lasserre Hirachy as an SDP

We will show that the relaxed problems (Las$_k$) and (Las$'_k$) can be expressed as semidefinite programs and in fact form a primal dual pair of SDPs. We have used this to implement the Lasserre Hierarchy in python and in the end of this chapter we give some concrete examples.

It is not difficult to see that the problem (Las$'_k$) is an SOS-program as defined in (4.22) and hence translates into an SDP. We will give an explicit SDP-formulation of (Las$'_k$) anyways since this will illuminate the duality and since we were not too explicit when stating that SOS-programs could be formulated as SDPs in Section 4.2. This section is also based on [28].

Defining the problem data

As before, k will denote a number $\geq N_{\deg}$. We will make the assumption that $f(0) = 0$ (this is not restrictive, as we can always substitute f by $f - f(0)$) and write

$$f = \sum_{\alpha \in \mathbb{N}^n_k \setminus \{0\}} b_\alpha \underline{X}^\alpha. \tag{5.37}$$

We define the following symmetric $\mathbb{R}[\underline{X}]_k$ valued matrices[3]:

$$T_i = (\underline{X}^{\beta+\gamma} g_i)_{\beta,\gamma \in \mathbb{N}^n_{d_i}}, \quad i = 0 \dots, m,$$

Here d_i is defined as in (5.11) and ensures that each entry is in $\mathbb{R}[\underline{X}]_k$. So T_i is an $s_i \times s_i$ matrix where $s_i := |\mathcal{T}^n_{d_i}| = \binom{n+d_i}{d_i} = \binom{n+d_i}{n}$.

For each $\alpha \in \mathbb{N}^n_k$ and $i = 0, \dots, m$ we define $A_{\alpha i} \in \mathcal{S}^{s_i}$ such that $A_{\alpha i}(\beta, \gamma)$ is the coefficient of $\underline{X}^\alpha$ in $\underline{X}^{\beta+\gamma} g_i$. I.e.

$$T_i(\beta, \gamma) = \underline{X}^{\beta+\gamma} g_i = \sum_{\alpha \in \mathbb{N}^n_k} A_{\alpha i}(\beta, \gamma) \underline{X}^\alpha. \tag{5.38}$$

[3]Recall that $g_0 = 1$ is for convenience. It is not a describing polynomial of S.

and hence

$$T_i = \sum_{\alpha \in \mathbb{N}^n_k} A_{\alpha i} \underline{X}^\alpha. \tag{5.39}$$

Example 5.6 Let $k = 3$ and $g_1 = 1 - X + 2Y$. Then $d_1 = 1$, $s_1 = \binom{2+1}{1} = 3$, and

$$T_1 = \begin{bmatrix} 1 - X + 2Y & X - X^2 + 2XY & Y - XY + 2Y^2 \\ X - X^2 + 2XY & X^2 - X^3 + 2X^2Y & XY - X^2Y + 2XY^2 \\ Y - XY + 2Y^2 & XY - X^2Y + 2XY^2 & Y^2 - XY^2 + 2Y^3 \end{bmatrix}.$$

Among the $|\mathbb{N}^2_3| = 10$ matrices $A_{(0,0)1}, \ldots, A_{(0,3)1}$ we have for instance

$$A_{(1,0)1} = \begin{bmatrix} -1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_{(1,1)1} = \begin{bmatrix} 0 & 2 & -1 \\ 2 & 0 & 1 \\ -1 & 1 & 0, \end{bmatrix} \tag{5.40}$$

which represent the coefficients of X and XY in T_1. ○

SDP formulation of (Las'_k)

First we need to characterize M_k.

Lemma 5.7

$$M_k = \Big\{ \sum_{i=0}^m \langle T_i, G_i \rangle \;\Big|\; G_i \in \mathcal{S}^{s_i}_+, i = 1, \ldots, m \Big\}. \tag{5.41}$$

Proof This follows directly from the definition of M_k and Lemma 1.6. □

From Lemma 5.7, (5.39), and (5.37) it follows that

$$\begin{aligned} f - a \in M_k &\iff -a + \sum_{\alpha \in \mathbb{N}^n_k \setminus \{0\}} b_\alpha \underline{X}^\alpha = \sum_{\alpha \in \mathbb{N}^n_k} \Big(\sum_{i=0}^m \langle A_{\alpha i}, G_i \rangle \Big) \underline{X}^\alpha \\ &\iff -a = \sum_{i=0}^m \langle A_{0i}, G_i \rangle \text{ and } b_\alpha = \sum_{i=0}^m \langle A_{\alpha i}, G_i \rangle,\ \alpha \in \mathbb{N}^n_k \setminus \{0\}, \end{aligned}$$

for some $G_i \succeq 0$, $i = 0, \ldots, m$. Here the second bi-implication follows by matching terms of same degree.

Therefore (Las'_k) can be restated as

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=0}^m \langle A_{0i}, G_i \rangle \\ \text{s.t.} \quad & b_\alpha = \sum_{i=0}^m \langle A_{\alpha i}, G_i \rangle,\ \alpha \in \mathbb{N}^n_k \setminus \{0\} \\ & G_i \succeq 0,\ i = 0, \ldots, m. \end{aligned} \tag{5.42}$$

Note that we switched from maximizing a to minimizing $-a = \sum_{i=0}^{m} \langle A_{0i}, G_i \rangle$. This is indeed a primal SDP formulation with multiple constraints.

SDP formulation of (Las_k)

We will now derive an SDP formulation of (Las_k). We denote by $L(T_i)$ the matrix obtained by applying L entry-wise on T_i, i.e. $L(T_i)(\beta, \gamma) = L(\underline{X}^{\beta+\gamma} g_i)$. We can then express $L(M_k) \subseteq [0, \infty)$ as a positive semi-definite condition on these matrices:

Lemma 5.8 *Suppose $L \in \mathbb{R}[\underline{X}]_k \to \mathbb{R}$ is linear. Then*

$$L(M_k) \subseteq [0, \infty) \iff L(T_i) \succeq 0,\ i = 0, \dots, m. \tag{5.43}$$

Proof Let $c_i = (c_{\alpha i})_{\alpha \in \mathbb{N}^n_{d_i}} \in \mathbb{R}^{s_i}$ and $p_i = \sum_{\alpha \in \mathbb{N}^n_{d_i}} c_\alpha \underline{X}^\alpha \in \mathbb{R}[\underline{X}]_{d_i}$ be the corresponding polynomial. Then

$$\begin{aligned} c_i^T L(T_i) c_i &= \sum_{\beta, \gamma \in \mathbb{N}^n_{d_i}} c_{\beta i} L(\underline{X}^{\beta+\gamma} g_i) c_{\gamma i} \\ &= L\Big(\sum_{\beta, \gamma \in \mathbb{N}^n_{d_i}} c_{\beta i} c_{\gamma i} \underline{X}^{\beta+\gamma} g_i \Big) = L(p_i^2 g_i). \end{aligned}$$

So

$$\begin{aligned} L(T_i) \succeq 0,\ i = 0, \dots, m &\iff c_i^T L(T_i) c_i \geq 0,\ \forall c_i \in \mathbb{R}^{s_i},\ i = 0, \dots, m \\ &\iff L(p_i^2 g_i) \geq 0,\ \forall p_i \in \mathbb{R}[\underline{X}]_{d_i},\ i = 0, \dots, m \\ &\iff L\Big(\sum_{i=0}^{m} (\sum_j p_{ij}^2) g_i\Big) = \sum_{i=0}^{m} \sum_j L(p_{ij}^2 g_i) \\ &\qquad\qquad \geq 0,\ \forall p_{ij} \in \mathbb{R}[\underline{X}]_{d_i} \\ &\iff L(M_k) \subseteq [0, \infty). \end{aligned}$$

□

Next, note that a linear map $L : \mathbb{R}[\underline{X}]_k \to \mathbb{R}$ satisfying $L(1) = 1$ is represented by a multi-sequence $(a_\alpha)_{\alpha \in \mathbb{N}^n_k}$ with $a_0 = 1$. Hence L can be represented uniquely by the multi-sequence

$$(y_\alpha)_{\alpha \in \mathbb{N}^n_k \setminus \{0\}} = (-a_\alpha)_{\alpha \in \mathbb{N}^n_k \setminus \{0\}} \tag{5.44}$$

and we can then write

$$L(f) = - \sum_{\alpha \in \mathbb{N}^n_k \setminus \{0\}} b_\alpha y_\alpha. \tag{5.45}$$

In terms of the matrices $A_{\alpha i}$ we also have

$$L(T_i) = \sum_{\alpha \in \mathbb{N}_k^n} A_{\alpha i} L(\underline{X}^\alpha) = A_{0i} - \sum_{\alpha \in \mathbb{N}_k^n \setminus \{0\}} A_{\alpha i} y_\alpha. \tag{5.46}$$

This follows from linearity and (5.39).

Using Lemma 5.8 and (5.46) we can now restate (Las_k) as

$$\begin{aligned} \text{Maximize} \quad & \sum_{\alpha \in \mathbb{N}_k^n \setminus \{0\}} b_\alpha y_\alpha \\ \text{s.t.} \quad & A_{0i} - \sum_{\alpha \in \mathbb{N}_k^n \setminus \{0\}} A_{\alpha i} y_\alpha \succeq 0,\ i = 0, \ldots, m. \end{aligned} \tag{5.47}$$

Here we switched from minimizing $L(f)$ to maximizing

$$-L(f) = \sum_{\alpha \in \mathbb{N}_k^n \setminus \{0\}} b_\alpha y_\alpha.$$

We note that (5.42) and (5.47) indeed form a primal dual pair of SDPs with multiple constraints.

Implementation

The translation of the polynomial optimization problem into a semidefinite program as just described is completely algorithmic although very cumbersome to do at hand. We have made an implementation of this in Python. We used the `SymPy` library to manipulate polynomials and extract coefficient in order to construct the data $A_{\alpha i}$ and b_α for the SDP. We then used the library `CVXPY` to formulate and solve the SDPs by its internal solver.

Our implementation was more of an exercise than a need of a good solver. There are options freely available which most likely do a better job, e.g. `GloptiPoly 3` [10] which is also interfaced through `YALMIP`.

Examples

Example 5.9 Consider the following minimization problem taken from [14]:

$$\begin{aligned} \text{Minimize} \quad & X + 3Y, \\ \text{subject to} \quad & g_1 = 1 - X^2 + 2Y^2 \geq 0, \\ & g_2 = 1 - X + Y \geq 0, \\ & g_3 = X - Y \geq 0. \end{aligned}$$

We want to approximate the solution by a 3rd order relaxation. Put $k = 3$. Then $d_0 = d_2 = d_3 = 1$ and $d_1 = 0$. There are $4 \cdot |\mathbb{N}_3^2| = 40$ matrices $A_{\alpha i}$ in play: 30 of dimension 3×3 (indexed by $\mathbb{N}_1^2 = \{(0,0),(1,0),(0,1)\}$ or equivalently $\mathcal{T}_1^2 = \{1, X, Y\}$) and 10 of dimension 1×1.

The relaxed problem Las_3 can then be cast as the dual SDP problem (5.47) and we are left with maximizing $y_{10} + 3y_{01}$ subject to the four linear matrix inequalities:

$$\begin{bmatrix} 1 & -y_{10} & -y_{01} \\ -y_{10} & -y_{20} & -y_{11} \\ -y_{01} & -y_{11} & -y_{02} \end{bmatrix} \succeq 0,$$

$$\begin{bmatrix} 2y_{02} + y_{20} + 1 \end{bmatrix} \succeq 0,$$

$$\begin{bmatrix} y_{01} + y_{10} + 1 & -y_{10} + y_{11} + y_{20} & -y_{01} + y_{02} + y_{11} \\ -y_{10} + y_{11} + y_{20} & -y_{20} + y_{21} + y_{30} & -y_{11} + y_{12} + y_{21} \\ -y_{01} + y_{02} + y_{11} & -y_{11} + y_{12} + y_{21} & -y_{02} + y_{03} + y_{12} \end{bmatrix} \succeq 0,$$

$$\begin{bmatrix} y_{01} - y_{10} & y_{11} - y_{20} & y_{02} - y_{11} \\ y_{11} - y_{20} & y_{21} - y_{30} & y_{12} - y_{21} \\ y_{02} - y_{11} & y_{12} - y_{21} & y_{03} - y_{12} \end{bmatrix} \succeq 0.$$

Here the second matrix inequality simplifies to $2y_{02} + y_{20} + 1 \geq 0$. This SDP has an optimal value 2.3452078 and since we switch sign in (5.47) the optimal value to (Las_k) becomes -2.3452078. This in fact coincide with the true optimum of (5.9). The SDP-solver also returns the solution vector

$$(y_{10}, \ldots, y_{03}) = (0.426, 0.640, -0.18, -0.27, -0.41, -0.01, 0.21, 0.14, 1.05)$$

which represent L as

$L(X)$	$L(Y)$	$L(X^2)$	$L(XY)$	$L(Y^2)$	$L(X^3)$	$L(X^2Y)$	$L(XY^2)$	$L(Y^3)$
-0.426	-0.640	0.18	0.27	0.41	0.01	-0.21	-0.14	-1.05

In fact $x^* = (-0.426, -0.640)$ is the (unique) minimizer. The table suggests that L acts as *evaluation at* $(-0.426, -0.640)$, e.g. $L(Y^2) \approx 0.41 \approx (-0.640)^2$. In fine terms L solves the moment problem as it can be represented as integration with respect to the Dirac measure at x^*. So for this specific problem the situation turned out to be exactly as described (but not predicted) in the beginning of Section 5.2. ○

Example 5.10 Consider the six-hump camel back function

$$f = 4X^2 + XY - 4Y^2 - 2.1X^4 + 4Y^4 + 1/3X^6 \tag{5.48}$$

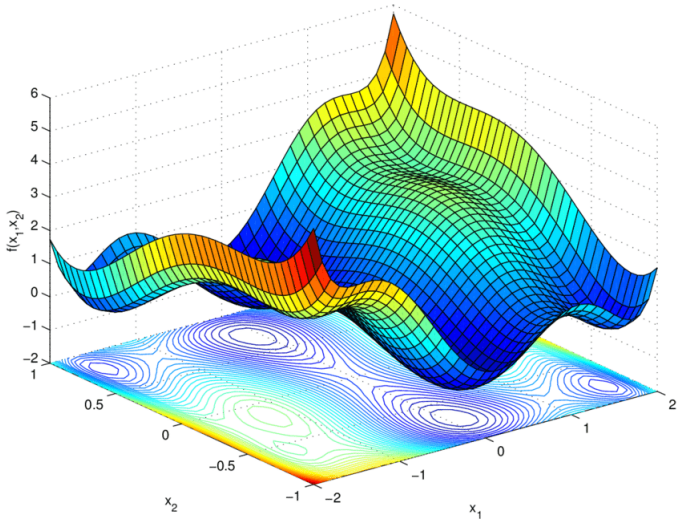

Figure 5.1: Six-hump camel back function [7].

which has six local minima, two of which are global. Using our implementation we minimize f over the three compact polytopes

$$S_1 = \mathcal{W}(2 - X, Y + \tfrac{1}{2}X + \tfrac{1}{2}, -Y + \tfrac{1}{2}X - 1) \tag{5.49}$$
$$S_2 = \mathcal{W}(2 - X, Y + \tfrac{1}{2}X + \tfrac{1}{2}, -Y + \tfrac{1}{2}X - 1.12) \tag{5.50}$$
$$S_3 = \mathcal{W}(2 - X, Y + \tfrac{1}{2}X + \tfrac{1}{2}, -Y + \tfrac{1}{2}X - 1.1) \tag{5.51}$$

by solving (Las_6).

We have chosen the polytopes such that the minimum over S_1 is at the boundary while on S_2 it is in the interior (one of the six local minima). It is very hard to tell whether the minimum is attained in the interior or at the boundary for S_3.

For S_1 we get the solution $f_6^* = -0.4854$ which is a lower bound on f^* by Theorem 5.1. We also obtain $(L_6(X), L_6(Y)) = (0.5000, -0.7499)$. As $f(0.5000, -0.7499) = -0.4854 = f_6^*$ is both an upper and a lower bound on f^*, we conclude the that $f^* = -0.4854$ and that $x^* = (0.5000, -0.7499)$.

Similarly for S_2 we find that $f^* = f(1.7034, -0.7963) = -0.2154$.

For S_3 however we get $f_6^* = -0.2180$ and

$$(L_6(X), L_6(Y)) = (0.6449, -0.7998).$$

We compute $f(0.6449, -0.7998) = -0.1134$ and conclude by Theorem 5.5 that $f^* \in [-0.2180, -0.1134]$. ○

5.4 Returning to the Max-Cut Problem

In Section 4.4 we saw how semidefinite programming could tackle the Max-Cut Problem. As mentioned the solution by Williamson and Goemanns was the first non-trivial approximation and hence a big breakthrough. Their method build on a clever relaxation of the quadratic integer problem:

$$\begin{aligned} \text{Maximize} \quad & \sum_{(i,j)\in E} \tfrac{1}{2}(1 - x_i x_j) \\ \text{s.t.} \quad & x_i \in \{-1, 1\}, \quad i = 1, \dots, n. \end{aligned} \tag{5.52}$$

The Lasserre Hiearchy gives a different and straight forward approach to this problem. We just have to realize that the feasible set $S = \{-1, 1\}^n$ is a compact semi-algebraic set: it is the 2^n corners of the hypercube in $\mathbb{R}^n$ and can be expressed as

$$S = \mathbb{V}(\langle X_1^2 - 1, \dots, X_n^2 - 1\rangle) = \mathcal{W}(B), \tag{5.53}$$

with $B = \{\pm(X_1^2 - 1), \dots, \pm(X_n^2 - 1)\}$. In fact the associated quadratic module $M = M[B]$ is Archimedian since $n - \sum_{i=1}^n X_i^2 = \sum_{i=1}^n (1 - X_i^2) \in M$.

By Theorem 5.1 the primal and dual relaxed solutions will converge to the true solution as the relaxation order goes to infinity. Of course this does not yield a polynomial time algorithm for solving the problem (since Max-Cut is NP-hard this would otherwise imply $P = NP$) and it might not even terminate. However for a fixed number n, the k-th order relaxation can be solved in polynomial time in k.[4]

The question is then how well the k-th order relaxations approximate the solution. Answering this is out of scope of this thesis, but it seems to be the case that the Lasserre Hierarchy performs well – see e.g. [13] and [26].

We round off by returning to Example 4.8 and apply our implementation of the Lasserre Hierarchy to a specific instance of the Max-Cut Problem.

[4]For a fixed n, the size (i.e. number of entries) of the variable matrix $G = \text{Diag}(G_1, \dots, G_m)$ in the semidefinite program (5.42) (when translated to a single constraint SDP) is at most

$$\Big(\sum_{i=1}^m |\mathcal{T}_{d_i}^n|\Big)^2 \le \Big(\sum_{i=1}^m |\mathcal{T}_k^n|\Big)^2 = m^2 \binom{n+k}{n}^2 = m^2 \Big(\frac{(k+n)\cdots(k+1)}{n!}\Big)^2 = O(k^{2n}).$$

Hence it is bounded by a polynomial expression in k when n and m (the number of polynomials defining S, i.e. $2n$ in this situation) is fixed. This is our own very rough analysis and there might be much better upper bounds. As semidefinite programs can be solved in polynomial time in the problem size the overall complexity is polynomial in k.

Example 5.11 Consider the graph $G = (V, E)$ from Example 4.8. Here $n = |V| = 7$ so we must maximize the polynomial

$$\begin{aligned}\tfrac{1}{2}(11 &- X_1X_2 - X_1X_4 - X_1X_5 - X_2X_3 \\ &- X_2X_6 - X_3X_6 - X_4X_5 - X_5X_6 - X_6X_7)\end{aligned}$$

subject to the 14 polynomial constraints $\pm(X_i^2 - 1) \geq 0, i = 1, \ldots, 7$.

The relaxed solution of order $k = 2$ gives the upper bound[5] $f_2^* = 9.3231$ which is not as good as the approximation 9.0258 obtained by the William and Goemanns relaxation but still fairly close to the true optimum, 9. Although the graph is rather small and the complexity grows polynomially in k our proof of concept implementation faced its limits when we tried to improve the upper bound by a higher order relaxation. ○

[5] As we are maximizing instead of minimizing this is an upper bound on the solution c.f. Theorem 5.1.

6 Conclusion

It is always fascinating how the most simple and concrete questions in mathematics can turn out to be just the opposite, and how their solutions develop into whole new fields. Hilbert's 17th problem is a good example. We have seen how the abstract theory on ordered and real fields developed by Artin and Schreier, together with the powerful Transfer Principle gave rise to a solution. In fact, the same theory applied to the much more general Krivine-Stengle theorems, which we proved (obtaining some simplifications) based on the original article of Stengle [29].

We have seen several different results concerning the problem of certifying properties of polynomial systems – the recurring property being positivity (or non-negativity) of a polynomial on some semi-algebraic set. Instead of focusing on *solving* equations, the concern is rather to *decide* whether a solution *exists* and even more at the core, to provide a good *reason* for this answer. When a solution exists it is conceptually not difficult to give a good reason for its existence – one can simply provide the point satisfying the equations. The problem is much more complex when no solution exists. The main result, being the Semi-algebraic Nullstellensatz Theorem 2.35, tells us that if a real polynomial system in its most general form (i.e. involving both equations, inequations, and inequalities) is infeasible, then there is a reason: this "reason" is provided by a purely syntactic and easily verifiable *certificate*.

In the real world (and on many different levels) existence is not very interesting if one cannot find what is "promised to be out there". We have seen how the power of modern convex optimization methods based on semidefinite programming makes the search for certificates a possible task. Concretely, we have used these methods to give a very short proof of a result in simple geometry, knowns as the Hadwiger-Finsler inequality (see Example 4.5). The

proof being the following relation:

$$\begin{aligned}&\big(a^2+b^2+c^2-((a-b)^2+(a-c)^2+(b-c)^2)\big)^2\\&\qquad -3(a+b+c)(-a+b+c)(a-b+c)(a+b-c)\\&\qquad\qquad =\big(2a^2+b(a+b-c)+c(a-b+c)\big)^2+3\big(b(a-b)+c(a-c)\big)^2.\end{aligned}$$

Finally we have seen how these techniques and results, together with theory in functional analysis, provided the basis for a powerful and general approach to constrained polynomial optimization. We studied propperties of the Lasserre Hierarchy, implemented the algorithm, and applied it to different optimization problems, including an approximation to the Max-Cut Problem.

A Proofs

A.1 Proof of Theorem 3.5

The proof is left out in [28] and we have instead followed an exercise in [22].

Proof (of (A.2)*)* Let d be the degree of G and write

$$G = \sum_{\nu \in I_d} \alpha_\nu X_1^{\nu_1} \cdots X_n^{\nu_n} \tag{A.1}$$

with $I_r = \{(\nu_1, \ldots, \nu_n) \in \mathbb{N}_0^n \mid \nu_1 + \cdots + \nu_n = r\}$ for $r \in \mathbb{N}$. Define $H \in \mathbb{R}[X_1, \ldots, X_n, Y]$ as

$$H = \sum_{\nu \in I_d} \alpha_\nu \prod_{i=1}^{n} \prod_{j=0}^{\nu_i - 1} (X_i - jY)$$

We will show that

$$(X_1 + \cdots + X_n)^k G = \sum_{\gamma \in I_{d+k}} \alpha_\gamma^{(k)} X_1^{\gamma_1} \cdots X_n^{\gamma_n}, \tag{A.2}$$

with the coefficients being

$$\alpha_\gamma^{(k)} = \frac{k!(d+k)^d}{\gamma_1! \ldots \gamma_n!} H\left(\frac{\gamma_1}{d+k}, \ldots, \frac{\gamma_n}{d+k}, \frac{1}{d+k}\right) \tag{A.3}$$

for all $k \in \mathbb{N}$.

Let $k \geq 0$ be given. We calculate the lefthand side using the multinomial formula:

$$\begin{aligned}(X_1 + \cdots + X_n)^k G &= \Big(\sum_{\mu \in I_k} \binom{k}{\mu_1, \ldots, \mu_n} X_1^{\mu_1} \cdots X_n^{\mu_n} \Big) \Big(\sum_{\nu \in I_d} \alpha_\nu X_1^{\nu_1} \cdots X_n^{\nu_n} \Big) \\ &= \sum_{\mu \in I_k} \sum_{\nu \in I_d} \frac{k!}{\mu_1! \cdots \mu_n!} \alpha_\nu X_1^{\nu_1 + \mu_1} \cdots X_n^{\nu_n + \mu_n} \\ &= \sum_{\gamma \in I_{d+k}} \sum_{(\nu,\mu) \in A_\gamma} \frac{k!}{\mu_1! \cdots \mu_n!} \alpha_\nu X_1^{\gamma_1} \cdots X_n^{\gamma_n} \qquad \text{(A.4)}\end{aligned}$$

Here $A_\gamma = \{(\nu, \mu) \in I_d \times I_k \mid \nu + \mu = \gamma\}$.

Now for the righthand side we use the fact that H is homogeneous of degree d to move denomiantors out:

$$\begin{aligned}&\sum_{\gamma \in I_{d+k}} \frac{k!(d+k)^d}{\gamma_1! \ldots \gamma_n!} H\big(\frac{\gamma_1}{d+k}, \ldots, \frac{\gamma_n}{d+k}, \frac{1}{d+k}\big) X_1^{\gamma_1} \cdots X_n^{\gamma_n} \\ &\qquad = \sum_{\gamma \in I_{d+k}} \frac{k!}{\gamma_1! \ldots \gamma_n!} H(\gamma_1, \ldots, \gamma_n, 1) X_1^{\gamma_1} \cdots X_n^{\gamma_n} \qquad \text{(A.5)}\end{aligned}$$

Comparing (A.4) and (A.5) it remains to show that

$$H(\gamma_1, \ldots, \gamma_n, 1) = \sum_{(\nu,\mu) \in A_\gamma} \frac{(\nu_1 + \mu_1)! \cdots (\nu_n + \mu_n)!}{\mu_1! \cdots \mu_n!} \alpha_\nu. \qquad \text{(A.6)}$$

It follows from the definition og H that for natural numbers $a_1, \ldots, a_n \in \mathbb{N}_0$ we have the identity

$$H(a_1, \ldots, a_n, 1) = \sum_{\substack{\nu \in I_d \\ \nu_i \leq a_i}} \frac{a_1! \cdots a_n!}{(a_1 - \nu_1)! \cdots (a_n - \nu_n)!} \alpha_\nu. \qquad \text{(A.7)}$$

Now note that for any $\gamma \in I_{d+k}$ bijection between $\{\nu \in I_d \mid \nu_i \leq \gamma_i, i = 1, \ldots, n\}$ and A_γ. If $(\mu, \nu) \in A_\gamma$ then $\nu_i = \gamma_i - \mu_i \leq \gamma_i$ for all i. On the other hand if $\nu \in I_d$ with $\nu_i \leq \gamma_i$ for all i, then

$$\begin{aligned}0 \leq (\gamma_i - \nu_i) &\leq (\gamma_1 - \nu_1) + \cdots + (\gamma_n - \nu_n) \\ &= (\gamma_1 + \cdots + \gamma_n) - (\nu_1 + \cdots + \nu_n) = k\end{aligned}$$

for all i and so we find $\mu = \gamma - \nu \in I_k$ such that $(\nu, \mu) \in A_\gamma$. Rewriting (A.7) we finally get:

$$\begin{aligned} H(\gamma_1, \dots, \gamma_n, 1) &= \sum_{\substack{\nu \in I_d \\ \nu_i \leq \gamma_i}} \frac{\gamma_1! \cdots \gamma_n!}{(\gamma_1 - \nu_1)! \cdots (\gamma_n - \nu_n)!} \alpha_\nu \\ &= \sum_{(\nu,\mu) \in A_\gamma} \frac{(\nu_1 + \mu_1)! \cdots (\nu_n + \mu_n)!}{\mu_1! \cdots \mu_n!} \alpha_\nu, \end{aligned}$$

establishing (A.6) as wanted.

We will now find a $K \in \mathbb{N}$ such that $\alpha_\gamma^{(K)} > 0$ for all $\gamma \in I_{d+K}$. Let

$$\gamma^{(k)} = \arg\min_{\gamma \in I_{d+k}} H(\frac{\gamma_1}{d+k}, \dots, \frac{\gamma_n}{d+k}, \frac{1}{d+k}), \quad k \in \mathbb{N} \tag{A.8}$$

and define the sequence $y_k = (\frac{\gamma_1^{(k)}}{d+k}, \dots, \frac{\gamma_n^{(k)}}{d+k})$. Since y_k is contained in the compact simplex

$$\Delta = \{(x_1, \dots, x_n) \in \mathbb{R}^n \mid x_1, \dots, x_n \geq 0,\ x_1 + \cdots + x_n = 1\} \tag{A.9}$$

for all k, we get a convergent subsequence $\{y_{k_i}\}_i$ with limit $y = \lim_{i \to \infty} y_{k_i} \in \Delta$.

Note that $G = H(X_1, \dots, X_n, 0)$ and $G > 0$ on Δ by assumption, so we get

$$\epsilon := \lim_{i \to \infty} H(y_{k_i}, \tfrac{1}{d+k_i}) = H(y, 0) = G(y) > 0. \tag{A.10}$$

There is then a $K \in \mathbb{N}$ such that

$$H(y_K, \frac{1}{d+K}) = H(\frac{\gamma_1^{(K)}}{d+K}, \dots, \frac{\gamma_n^{(K)}}{d+K}), \frac{1}{d+K}) \geq \tfrac{\epsilon}{2} \tag{A.11}$$

and so

$$\alpha_\gamma^{(K)} = \frac{K!(d+K)^d}{\gamma_1! \dots \gamma_n!} H(\frac{\gamma_1}{d+K}, \dots, \frac{\gamma_n}{d+K}, \frac{1}{d+K}) \tag{A.12}$$

$$\geq \frac{K!(d+K)^d}{\gamma_1! \dots \gamma_n!} H(\frac{\gamma_1^{(K)}}{d+K}, \dots, \frac{\gamma_n^{(K)}}{d+K}, \frac{1}{d+K}) > 0 \tag{A.13}$$

for all γ in I_{d+K} as wanted. □

A.2 Proof of special version of Haviland's Theorem

The following proof of Havilands theorem in case S is compact and $L(1) = 1$, is based on the proof of Theorem 3.19 as it appears in [28, Thm. 2].

Proof $ii) \implies i)$. Clear.

$i) \implies ii)$. We want to extend L to a positive linear functional $\bar{L}$ on $\mathcal{C}(S, \mathbb{R})$, the vector space of continuous functions supported on S.

For this consider the embedding

$$\phi : \mathbb{R}[\underline{X}] \hookrightarrow \mathcal{C}(S, \mathbb{R}),\ p \mapsto p\big|_S. \tag{A.14}$$

It follows from our assumption that $\ker\phi \subseteq \ker L$: If $p \in \ker\phi$ then also $-p \in \ker\phi$ so $L(p) \geq 0$ and $-L(p) = L(-p) \geq 0$, i.e. $p \in \ker L$.

We then define

$$\bar{L} : \phi(\mathbb{R}[\underline{X}]) \to \mathbb{R}, \phi(p) \mapsto L(p). \tag{A.15}$$

This is well defined: Suppose $\phi(p) = \phi(q)$. Then $\phi(p-q) = \phi(p) - \phi(q) = 0$ so $p - q \in \ker\phi \subseteq \ker L$, and hence $L(p) - L(q) = L(p-q) = 0$.

$\bar{L}$ is continuous in the supremum norm $\|p\|_\infty = \sup\{p(x) \mid x \in S\}$: For any $p \in \mathbb{R}[\underline{X}]$ we have $L(p) \leq \|p\|_\infty$ since $\|p\|_\infty - p \geq 0$ implies $L(p) - \|p\|_\infty = L(p - \|p\|_\infty) \geq 0$.

As $\phi(\mathbb{R}[\underline{X}])$ is a dense subspace of $\mathcal{C}(S, \mathbb{R})$ (equipped with the supremum norm) c.f. the Stone-Weirstrass Theorem, $\bar{L}$ extends to a continuous linear functional on $\mathcal{C}(S, \mathbb{R})$.

We must show that $\bar{L}(f) \geq 0$ for all $f \in \mathcal{C}(S, \mathbb{R})$ such that $f(x) \geq 0$ for all $x \in S$ (i.e. $\bar{L}$ is a *positive* linear functional).

Let $\epsilon > 0$ and suppose $f(x) \geq 0$ for all $x \in S$. As $\bar{L}$ is continuous choose $\delta \in (0, \epsilon]$ such that $\|f - g\|_\infty \leq \delta \implies |\bar{L}(f) - \bar{L}(g)| \leq \epsilon$.

As $\phi(\mathbb{R}[\underline{X}])$ is dense in $\mathcal{C}(S, \mathbb{R})$ there is $p \in \mathbb{R}[X]$ such that $\|f - \phi(p)\|_\infty \leq \delta$ and hence $\bar{L}(\phi(p)) \leq \bar{L}(f) + \epsilon$. Furthermore $p(x) + \epsilon \geq p(x) + \delta \geq f(x) \geq 0$ for any $x \in S$. This implies

$$0 \leq L(p + \epsilon) = \bar{L}(\phi(p)) + \epsilon \leq \bar{L}(f) + 2\epsilon \tag{A.16}$$

As ϵ was arbitrary $\bar{L}(f) \geq 0$ as wanted.

Now Riesz Representation Theorem [17, Thm. 3.2.1] yields a Borel measure μ such that

$$\bar{L}(f) = \int_S f\, d\mu, \tag{A.17}$$

for all $f \in \mathcal{C}(S, \mathbb{R})$. In particular for any $p \in \mathbb{R}[\underline{X}]$,

$$L(p) = \bar{L}(\phi(p)) = \int_S p \, d\mu. \tag{A.18}$$

□

Bibliography

[1] Saugata Basu, Richard Pollack, and Marie-Fran¸coise Roy. *Algorithms in Real Algebraic Geometry*. Springer, New York, 2016. URL: https://perso.univ-rennes1.fr/marie-francoise.roy/bpr-ed2-posted3.pdf.

[2] Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.

[3] Grigoriy Blekherman, Pablo A Parrilo, and Rekha R Thomas. *Semidefinite optimization and convex algebraic geometry*. SIAM, 2012.

[4] Stephen Boyd and Lieven Vandenberghe. "Convex sets". In: *Convex Optimization*. Cambridge University Press, 2004. DOI: 10.1017/CBO9780511804441.003.

[5] David Cox, John Little, and Donal OShea. *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer Science & Business Media, 2013.

[6] A. Papachristodoulou et.al. "SOSTOOLS: Sum of squares optimization toolbox for MATLAB". In: Available from http://www.eng.ox.ac.uk/control/sostools, http://www.cds.caltech.edu/sostools and http://www.mit.edu/~parrilo/sostools. http://arxiv.org/abs/1310.4716, 2013.

[7] *GloptiPoly: Global Optimization over Polynomials with Matlab and SeDuMi - Scientific Figure on ResearchGate.* Available from https://www.researchgate.net/figure/Six-hump-camel-back-function-fig1-4006246.

[8] Michel X Goemans and David P Williamson. "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming". In: *Journal of the ACM (JACM)* 42.6 (1995), pp. 1115–1145.

[9] Daniel R. Grayson and Michael E. Stillman. *Macaulay2, a software system for research in algebraic geometry*. Available at https://faculty.math.illinois.edu/Macaulay2/.

[10] Didier Henrion, Jean-Bernard Lasserre, and Johan Löfberg. "GloptiPoly 3: moments, optimization and semidefinite programming". In: *Optimization Methods & Software* 24.4-5 (2009), pp. 761–779.

[11] Jens Carsten Jantzen. *Advanced Algebra – The note formerly know as Algebra 2*. 2005.

[12] Jean-Louis Krivine. "Anneaux préordonnés". In: *Journal d'analyse mathématique* 12.1 (1964), pp. 307–326.

[13] Monique Laurent. "Semidefinite relaxations for max-cut". In: *The sharpest cut: The Impact of Manfred Padberg and his work*. SIAM, 2004, pp. 257–290.

[14] Niels Lauritzen. *Undergraduate convexity: from Fourier and Motzkin to Kuhn and Tucker*. World Scientific, 2013.

[15] Johan Lofberg. "YALMIP: A toolbox for modeling and optimization in MATLAB". In: *2004 IEEE international conference on robotics and automation (IEEE Cat. No. 04CH37508)*. IEEE. 2004, pp. 284–289.

[16] Johan Löfberg. "Pre- and post-processing sum-of-squares programs in practice". In: *IEEE Transactions on Automatic Control* 54.5 (2009), pp. 1007–1011.

[17] Murray Marshall. *Positive polynomials and sums of squares*. 146. American Mathematical Soc., 2008.

[18] Murray A Marshall. *Spaces of orderings and abstract real spectra*. Springer, 2006.

[19] Pablo A Parrilo. "Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization". PhD thesis. California Institute of Technology, 2000.

[20] Helfried Peyrl and Pablo A Parrilo. "A Macaulay 2 package for computing sum of squares decompositions of polynomials with rational coefficients". In: *Proceedings of the 2007 international workshop on Symbolic-numeric computation*. 2007, pp. 207–208.

[21] Victoria Powers and Bruce Reznick. "A new bound for Pólya's theorem with applications to polynomials positive on polyhedra". In: *Journal of pure and applied algebra* 164.1-2 (2001), pp. 221–229.

[22] Alexander Prestel and Charles Delzell. *Positive polynomials: from Hilbert's 17th problem to real algebra*. Springer, 2004.

[23] A. R. Rajwade. *Squares*. London Mathematical Society Lecture Note Series. Cambridge University Press, 1993. DOI: 10.1017/CBO9780511566028.

[24] Bruce Reznick et al. "Extremal PSD forms with few terms". In: *Duke mathematical journal* 45.2 (1978), pp. 363–374.

[25] Bruce Reznick. "Some concrete aspects of Hilbert's 17th problem". In: *Contemporary mathematics* 253 (2000), pp. 251–272.

[26] Thomas Rothvoß. "The Lasserre hierarchy in approximation algorithms". In: *Lecture Notes for the MAPSP* (2013), pp. 1–25.

[27] Konrad Schmüdgen. "The k-moment problem for compact semi-algebraic sets". In: *Mathematische Annalen* 289.1 (1991), pp. 203–206.

[28] Markus Schweighofer. "Optimization of polynomials on compact semi-algebraic sets". In: *SIAM Journal on Optimization* 15.3 (2005), pp. 805–825.

[29] Gilbert Stengle. "A Nullstellensatz and a Positivstellensatz in semialgebraic geometry". In: *Mathematische Annalen* 207.2 (1974), pp. 87–97.

[30] Bernd Sturmfels. *Solving systems of polynomial equations*. American Mathematical Soc., 2002.

[31] Steen Thorbjornsen. *Grundlæggende mål-og integralteori*. ISD LLC, 2014.

[32] David S Watkins. *Fundamentals of matrix computations*. Vol. 64. John Wiley & Sons, 2004.

[33] T Wörmann. "Short algebraic proofs of theorems of Schmüdgen and Pólya". In: *preprint* (1996).