Université catholique de Louvain
Faculté des sciences appliquées
Département d'ingénierie mathématique

Center for Operations Research and Econometrics
Centre for Systems Engineering and Applied Mechanics

Convex Optimization over Non-Negative Polynomials: Structured Algorithms and Applications

Yvan Hachez

Thesis submitted in partial fulfillment
of the requirements for the degree of
Docteur en sciences appliquées

Dissertation committee:
Yves Genin
Michel Gevers (chair)
Yurii Nesterov (advisor)
Marc Van Barel
Paul Van Dooren (advisor)
Jean-Philippe Vial

May 2003

Contents

List of Figures

List of Tables

List of Algorithms

List of Notations

Basics

$\mathbb{N}$	set of non-negative integer numbers
$\mathbb{Z}$	set of integer numbers
$\mathbb{K}$	generic field
$\mathbb{R}$	field of real numbers
$\mathbb{C}$	field of complex numbers
$\mathbb{F}$	$\mathbb{R}$ or $\mathbb{C}$ (depending on the context)
$\mathbb{T}$	complex unit circle, $\mathbb{T} = \{z \in \mathbb{C} : \lvert z\rvert = 1\}$
$\mathbb{D}$	complex open unit disk, $\mathbb{D} = \{z \in \mathbb{C} : \lvert z\rvert < 1\}$
$\jmath$	complex unit, $\jmath = \sqrt{-1}$
$\operatorname{Re} z$	real part of z
$\operatorname{Im} z$	complex part of z
$\overline{z}$	complex conjugate of z
$\lvert z\rvert$	modulus of z
$\iff$	if and only if
$\doteq$	equal by definition to
$\binom{k}{n}$	binomial coefficient, $\binom{k}{n} \doteq \frac{n!}{(n-k)!k!}$
δ_{ij}	Kronecker delta
$\dim X$	dimension of X
$\operatorname{span}\{\cdot\}$	linear space generated by
$\mathcal{O}(\cdot)$	"big-oh" asymptotic complexity bound

Inner products and norms

$\langle \cdot, \cdot \rangle$	generic inner product
$\langle\langle \cdot, \cdot \rangle\rangle$	generic matrix-valued inner product
$\langle \cdot, \cdot \rangle_{\mathbb{R}}$	real-valued inner product
$\langle\langle \cdot, \cdot \rangle\rangle_{\mathbb{R}}$	matrix-valued inner product corresponding to $\langle \cdot, \cdot \rangle_{\mathbb{R}}$
$\langle \cdot, \cdot \rangle_{\mathbb{C}}$	complex-valued inner product
$\langle\langle \cdot, \cdot \rangle\rangle_{\mathbb{C}}$	matrix-valued inner product corresponding to $\langle \cdot, \cdot \rangle_{\mathbb{C}}$
$\lVert \cdot \rVert_p$	p-norm ($1 \leq p \leq +\infty$)
$\lVert \cdot \rVert_2$	Euclidean norm (vectors) / spectral norm (matrices)
$\lVert \cdot \rVert_F$	Frobenius norm
$\lVert \cdot \rVert_A$	norm induced by positive definite matrix A

Matrix theory

X^T	transpose of matrix X
X^*	conjugate transpose of matrix X
$\mathcal{S}^n$	set of real symmetric matrices (of order n)
$\mathcal{H}^n$	set of complex Hermitian matrices (of order n)
$\operatorname{rank} X$	rank of matrix X
$\det X$	determinant of square matrix X
$\operatorname{Trace} X$	trace of square matrix X
$\lambda_k(X)$	kth largest eigenvalue of matrix X
$\lambda_{\min}(X)$	minimal eigenvalue of matrix X
$\lambda_{\max}(X)$	maximal eigenvalue of matrix X
$\mathcal{F}(A)$	field of value of matrix A
$\mathcal{W}(\mathcal{A})$	joint numerical range of the matrix set $\mathcal{A}$
$\preceq$, $\succeq$	Löwner partial ordering
$\otimes$	Kronecker product
Z	(block) shift matrix
J	(block) permutation matrix

Sets and cones

$Q_1 + Q_2$	sum of Q_1 and Q_2
$Q_1 \times Q_2$	direct product of Q_1 and Q_2
$\operatorname{int} Q$	interior of Q
$\operatorname{rint} Q$	relative interior of Q
$\operatorname{cl} Q$	closure of Q
$\operatorname{co} Q$	convex hull of Q
$\mathcal{K}$	generic cone
$\mathcal{K}^\star$	generic dual cone
$\mathcal{L}_n$	Lorentz cone
$\mathcal{S}^n_+$	cone of real symmetric positive semidefinite matrices (of order n)
$\mathcal{H}^n_+$	cone of complex Hermitian positive semidefinite matrices (of order n)
$\mathcal{K}_\mathbb{R}$	cone of non-negative polynomials on the real line
$\mathcal{K}_\mathbb{C}$	cone of non-negative trigonometric polynomials

Linear operators

$\Lambda(\cdot)$	generic linear operator, generic structured matrix
$\Lambda^*(\cdot)$	generic dual linear operator
$\operatorname{diag}(\cdot)$	(block) diagonal linear operator, (block) diagonal matrix
$\operatorname{Diag}(\cdot)$	(block) vector defined by the (block) diagonal elements
$H(\cdot)$	(block) Hankel linear operator, (block) Hankel matrix
$T(\cdot)$	(block) Toeplitz linear operator, (block) Toeplitz matrix

Functional systems

$\psi(\cdot), \upsilon(\cdot)$	generic lower-level and upper-level functional systems
$\pi_n(\cdot)$	column vector of monomials, from degree 0 to n
$T_n(\cdot)$	first-order Chebyshev polynomial of degree n
$\tau_n(\cdot)$	column vector of first-order Chebyshev polynomials, from degree 0 to n
$\Psi(\cdot), \Upsilon(\cdot)$	matrix equivalents of $\psi(\cdot)$ and $\upsilon(\cdot)$, respectively
$\Pi_n(\cdot)$	matrix equivalent of $\pi_n(\cdot)$

Acronyms

CPU	central processing unit
DARE	discrete-time algebraic Riccati equation
flops	floating point operations
KYP	Kalman-Yakubovich-Popov
LMI	linear matrix inequality (-ies)
LP	linear programming
QEP	quadratic eigenvalue problem(s)
QQP	quadratically constrained quadratic problem(s)
SDP	semidefinite programming (problems)

Preface

Mathematical engineering is concerned with the development of theoretical models, the formulation of real-life problems using these models and the computational aspects of solving these formulations. Thus, this field has a wide range of practical applications. In this thesis, we focus on a particular subfield of research in this area, namely *optimization*. Optimization problems occur frequently in practice because their solutions provide us with an answer to the following question :

> Given a set of constraints (budget, design...), what is the best feasible solution with respect to a given criterion?

In order to answer this question, quantitative tools must be developed that clearly provide decision support at multiple levels (strategic, operational...) within organizations.

For instance, optimization problems were intensively investigated in the military context during the Cold War to assist in planning military actions and in designing robust supply networks. The denomination *operations research* originates from that period. Optimization problems are important in other contexts, as well. *Systems engineering* is probably the best illustration of this fact. In this case, optimization problems are auxiliary problems that need to be solved so as to yield a relevant solution to the original problem.

There exists a wide class of optimization problems for which efficient numerical methods exist. This class consists of *convex optimization problems* and covers a huge number of applications. In this thesis, we are interested in convex problems related to *non-negative matrix polynomials* and to *moment spaces*. Although these mathematical objects have been extensively studied for more than a century, their use in a convex optimization framework has not been thoroughly investigated, probably owing to the recent and rapid achievements in this field.

Our motivation for this research can be explained as follows. Non-negative polynomials play a fundamental role in systems and control theory; they represent, among other things, spectral density functions and therefore show up in spectral factorization. Moment spaces are clearly associated with probability theory and statistics; moments usually need to be identified so as to obtain the parameters of a mathematical model. Therefore, convex optimization problems involving these mathematical objects are frequent in practice. However, the underlying problem structure is usually disregarded when solving them; straightforward approximations or generic algorithms are used. Obviously, this is a source of inefficiency.

This work proposes a novel and self-contained treatment of cones of non-negative polynomials and of moment spaces, with a convex optimization perspective. Theoret-

ical issues, as well as numerical aspects, are considered. In particular, the structural properties of several important optimization problems involving non-negative polynomials or moment spaces are exploited in order to obtain specific algorithms; our approach leads to a dramatic improvement in complexity. Of course, convex reformulations of initial problems are sometimes required in order to reveal the appropriate properties (convexity, structure...). This key step is often overlooked in the literature. Several applications (in linear algebra, systems and control, signal processing, probability and statistics...) illustrate the use of structured algorithms and the relevance of our results in practice.

This thesis is organized as follows.

In Chapter 1, we remind the reader of several theoretical concepts used throughout this thesis. We focus on our three different research areas (linear algebra, convex analysis, convex optimization). Note that only the most important definitions and properties are stated.

In Chapter 2, we prove our main characterization results, which are related to sum-of-squares representations. These original results are at the core of this thesis because they can be applied to various cones of non-negative matrix polynomials. Interestingly enough, the corresponding dual cones are related to moment spaces. Although this fact is well known in the literature, we provide a self-contained convex optimization approach. Finally, we highlight the difficulties of representing non-negative multivariate polynomials via sums of squares.

In Chapter 3, we focus on conic optimization problems involving non-negative polynomials (in the primal space) and moment spaces (in the dual space), which are frequently encountered in practice. Indeed, the compact parametrizations obtained in the previous chapter allow us to optimize over these convex cones. Several polynomial-time solution methods are investigated. In particular, we point out that the dual problems can be solved very efficiently by making use of the underlying problem structure. We propose a new dual algorithm that has the best-known worst-case complexity estimate. The proximal analytic center cutting plane method is also evaluated using this dual approach.

In Chapter 4, we use the spectral factorization representation of non-negative polynomials to investigate quadratic optimization problems. Although these problems are hard to solve in general, we identify an important convexity condition that makes these problems easier to solve. This condition is related to a well-known concept in linear algebra, namely the joint numerical range. Moreover, we point out several new classes of easy quadratic optimization problems, which are related to non-negative polynomials and interpolation constraints.

In Chapter 5, we consider quadratic matrix polynomials, which have many applications in mechanical engineering. We focus on two important problems. The first relates to determining whether a pair of matrices is definite and the second to distance problems. The proposed solutions, which take advantage of the problem structures, improve those given in the literature.

In Chapter 6, we describe several applications of non-negative polynomials and moment spaces in mathematical engineering (systems and control, signal processing, probability and statistics, coding theory). The results obtained previously provide us with new approaches and new tools to solve the associated problems efficiently.

We conclude each chapter with a section entitled "Research summary", which clearly emphasizes our contributions. Although "Notes and references" are also pro-

vided, they undoubtedly reflect a personal and intentionally limited point of view.

Acknowledgements

I would like to thank my advisors, Yurii Nesterov and Paul Van Dooren, for their encouragement and helpfulness during the last three and a half years. I have enjoyed our collaboration, which I found very instructive. Of course, they have communicated (part of) their knowledge of convex optimization and linear algebra to me. Just as important, however, were our numerous discussions, which were an invaluable experience from which I have learnt that suitably used primary concepts often yield extremely elegant and important results.

I am also grateful to Yves Genin and Marc Van Barel for their willingness to share their expertise on orthogonal polynomials and structured matrices. I wish to thank Jean-Philippe Vial for giving me the opportunity to spend three and a half months in Geneva and for sharing his expertise on cutting plane methods. I would also like to thank Michel Gevers for his key role in the interuniversity attraction pole on systems and control (IAP Phases IV & V), which provides essential financial support to the department, and for accepting the role of chair of the dissertation committee.

I am also indebted to Philippe Delsarte for providing me with several useful references on coding theory and association schemes, to Alphonse Magnus for granting me access to the former Philips Research library, and to Hugo Woerdeman for discussing several spectral factorization problems and topics on multivariate polynomials.

This research was financially supported by the Belgian National Fund for Scientific Research by means of a research fellowship. Travel grants from the Belgian National Fund for Scientific Research and the Communauté Française de Belgique are also gratefully acknowledged.

The scientific surroundings at the Center for Operations Research and Econometrics (CORE) and at the Centre for Systems Engineering and Applied Mechanics (CESAME) were particularly motivating. I especially enjoyed the friendship of Radu Ştefan and Antoine Vandendorpe, who both shared my office.

Finally, I would like to thank my family and all of my friends for their encouragement and support. I dedicate this thesis to my son Antoine.

Louvain-la-Neuve, February 2003

Et il revint vers le renard :
–Adieu, dit-il...
–Adieu, dit le renard. Voici mon secret. Il est très simple : on ne voit bien qu'avec le coeur. L'essentiel est invisible pour les yeux.
–L'essentiel est invisible pour les yeux, répéta le petit prince, afin de se souvenir.
–C'est le temps que tu as perdu pour ta rose qui fait ta rose si importante.
–C'est le temps que j'ai perdu pour ma rose... fit le petit prince, afin de se souvenir.
–Les hommes ont oublié cette vérité, dit le renard. Mais tu ne dois pas l'oublier. Tu deviens responsable pour toujours de ce que tu as apprivoisé. Tu es responsable de ta rose...
–Je suis responsable de ma rose... répéta le petit prince, afin de se souvenir.

Antoine de Saint Exupéry, Le petit Prince

Chapter 1

Preliminaries

This opening chapter introduces several important concepts used throughout this thesis. Since we mainly focus on relevant examples which define our notation, only the most important definitions and properties are mentioned. Relevant references are given for each statement and in the last section, "Notes and references".

Section 1.1 is devoted to linear algebra, which lies at the core of modern mathematical engineering, both from a theoretical and a practical point of view. Section 1.2 deals with convex analysis, which is the mathematical theory concerned with convex sets and functions, our main area of interest. Section 1.3 introduces our primary field of research, i.e., convex optimization, which allows us to formulate various optimization problems, such as minimizing a production cost subject to satisfaction of demand and limited availability of resources.

1.1 Linear algebra and matrix theory

Linear algebra is very important in applied mathematics. On the one hand, the theory underlying this field of research is very rich. It allows us to formulate problems at a higher level of abstraction by using its building blocks, which are called matrices. On the other hand, efficient algorithms usually exist for computing the theoretical properties of interest. Keeping in mind our final goal, i.e., numerical convex optimization, this section defines the concepts of vector space, of inner product and of linear operator.

1.1.1 Vector space

A *vector space* over a field $\mathbb{K}$ is composed of

1. a set X of vectors: $X = \{(x_0, \ldots, x_{n-1})\}$;
2. an additive law $+ : X \times X \to X$ such that $(X, +)$ is an Abelian group:
   (a) $\forall u, v, w \in X, (u + v) + w = u + (v + w)$,
   (b) $\exists 0 \in X$ such that $\forall u \in X, u + 0 = u$,
   (c) $\forall u, v \in X, u + v = v + u$,
   (d) $\forall u \in X$, $\exists v \in X$ such that $u + v = v + u = 0$.

3. an external law $\circ : \mathbb{K} \times X \to X$ such that, for all $\alpha, \beta \in \mathbb{K}$ and for all $v, w \in X$, the following identities hold:

   (a) $\alpha \circ (v + w) = \alpha \circ v + \alpha \circ w$,

   (b) $(\alpha + \beta) \circ v = \alpha \circ v + \beta \circ v$,

   (c) $\alpha \circ (\beta \circ v) = (\alpha\beta) \circ v$,

   (d) $1_{\mathbb{K}} \circ x = x$.

Let us comment this definition of vector space. First, the symbol $\circ$, which denotes the external law, is usually dropped as it can be inferred from the context. Then, from a computational point of view, it is clear that only finite-dimensional vector spaces can be used, i.e., $n < +\infty$. In this thesis, we only consider such vector spaces and their elements are stacked into column vectors

$$X \ni x = \begin{bmatrix} x_0 \\ \vdots \\ x_{n-1} \end{bmatrix}.$$

Finally, in applied mathematics, the field $\mathbb{K}$ of interest is either the field of real numbers $\mathbb{R}$ or the field of complex numbers $\mathbb{C}$, with the usual set of operations $(+, \cdot)$ and properties. In order to easily treat both cases in this section, the symbol $\mathbb{F}$ is used to denote either $\mathbb{R}$ or $\mathbb{C}$. The complex unit is $\jmath \doteq \sqrt{-1}$. For a complex number $z \in \mathbb{C}$, $\operatorname{Re} z$ ($\operatorname{Im} z$) is the real (imaginary) part of z. For a vector $x \in X$, $\operatorname{Re} x$ and $\operatorname{Im} x$ are defined componentwise. The complex modulus of z is denoted by $|z|$.

Example 1.1. For any field $\mathbb{K}$, the set $\mathbb{K}^n$ associated with the componentwise addition and multiplication is a finite-dimensional vector space over $\mathbb{K}$. For $0 \leq i \leq n-1$, the canonical element e_i is the element $(x_0, \ldots, x_{n-1})$ with $x_i = 1_{\mathbb{K}}$ and $x_j = 0$ for $j \neq i$. The element e is obtained by adding all canonical elements together.

Example 1.2. The set of polynomials with real coefficients $\mathbb{R}[x]$ is a vector space over $\mathbb{R}$. It is finite-dimensional if an upper bound on the degree of these polynomials is given.

1.1.2 Norm and inner product

Let E be a finite-dimensional vector space over $\mathbb{F}$.

Definition 1.1. A *norm* $\|\cdot\|$ on E is a real-valued function

$$\|\cdot\| : E \to \mathbb{R}$$

that satisfies three fundamental properties:

1. $\|x\| \geq 0, \quad \forall x \in E$ and equality holds if and only if $x = 0$;
2. $\|\lambda x\| = |\lambda| \, \|x\|, \quad \forall x \in E, \ \forall \lambda \in \mathbb{F}$;
3. $\|x + y\| \leq \|x\| + \|y\|, \quad \forall x, y \in E$.

Example 1.3. The standard norms on $\mathbb{R}^n$ are the so-called l_p-norm. Given a real p greater or equal to 1, we define

$$\|x\|_p \doteq (\sum_{k=0}^{n-1} |x_k|^p)^{1/p}, \quad p \geq 1.$$

Two interesting p-norms are given by $p = 2$ and $p = \infty$. The l_2-norm of x is also called the *Euclidean norm*. The l_∞-norm of x is also known as *Chebyshev norm* and an alternative definition is

$$\|x\|_\infty \doteq \lim_{p \to +\infty} \|x\|_p = \max_k |x_k|.$$

Definition 1.2. An *inner product* $\langle \cdot, \cdot \rangle$ on E is a function

$$\langle \cdot, \cdot \rangle : E \times E \to \mathbb{C}$$

that satisfies three basic properties:

1. $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle, \quad \forall x, y, z \in E, \ \forall \alpha, \beta \in \mathbb{F}$;
2. $\langle y, x \rangle = \overline{\langle x, y \rangle}, \quad \forall x, y \in E$;
3. $\langle x, x \rangle \geq 0, \quad \forall x \in E$ and equality holds if and only if $x = 0$.

It is well known that any inner product $\langle \cdot, \cdot \rangle$ on E induces a norm $\|\cdot\|$ on E, which is defined by $\|x\| \doteq \langle x, x \rangle^{\frac{1}{2}}, \forall x \in E$. The celebrated *Cauchy-Schwarz* inequality relates both concepts as follows.

Proposition 1.1 (Cauchy-Schwarz inequality). *Let E be a vector space over $\mathbb{F}$ and $\langle \cdot, \cdot \rangle$ be an inner product associated with it. For all x and $y \in E$, we have*

$$|\langle x, y \rangle| \leq \|x\| \, \|y\| \tag{1.1}$$

where $\|\cdot\|$ is the norm induced by $\langle \cdot, \cdot \rangle$.

Proof. The proof is so nice that we cannot resist to reproduce it here. Let $x, y \in E$ and $\lambda \in \mathbb{F}$. Without loss of generality, assume that $y \neq 0$. Then

$$0 \leq \langle x + \lambda y, x + \lambda y \rangle = \langle x, x \rangle + \lambda \langle y, x \rangle + \overline{\lambda} \langle x, y \rangle + |\lambda|^2 \langle y, y \rangle.$$

Setting $\lambda = -\frac{\langle x, y \rangle}{\langle y, y \rangle}$ and multiplying the above inequality by $\langle y, y \rangle$, we get

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle.$$

This last inequality is equivalent to (1.1). □

Let us introduce an important convention. A real-valued inner product is denoted by $\langle \cdot, \cdot \rangle_{\mathbb{R}}$ whereas $\langle \cdot, \cdot \rangle_{\mathbb{C}}$ is used for complex-valued inner products. As we are mainly interested in real-valued inner products in the context of convex optimization, the subscript $\mathbb{R}$ is often dropped in this thesis. The precise motivation for our interest in these inner products is given in Section 1.3.

Example 1.4. The standard inner product on $\mathbb{R}^n$ is defined by

$$\langle x, y\rangle_{\mathbb{R}} \doteq \sum_{i=0}^{n-1} x_i y_i. \tag{1.2}$$

The induced norm is exactly the Euclidean norm.

Example 1.5. The standard inner product on $\mathbb{C}^n$ is defined by

$$\langle x, y\rangle_{\mathbb{C}} \doteq \sum_{i=0}^{n-1} x_i \overline{y_i}. \tag{1.3}$$

A real-valued inner product on $\mathbb{C}^n$ is given by

$$\langle x, y\rangle_{\mathbb{R}} \doteq \operatorname{Re}\langle x, y\rangle_{\mathbb{C}} = \langle \operatorname{Re} x, \operatorname{Re} y\rangle + \langle \operatorname{Im} x, \operatorname{Im} y\rangle. \tag{1.4}$$

It corresponds to the identification of $\mathbb{C}^n$ with $\mathbb{R}^{2n}$.

1.1.3 Linear operator

Let E, F be two finite-dimensional vector spaces over $\mathbb{F}$.

Definition 1.3. A *linear operator* $\mathcal{A} : E \to F$ is a function that satisfies

$$\mathcal{A}(\alpha x + \beta y) = \alpha \mathcal{A}x + \beta \mathcal{A}y, \quad \forall x, y \in E, \forall \alpha, \beta \in \mathbb{F}.$$

Definition 1.4. Given a linear operator $\mathcal{A} : E \to F$ and two inner products $\langle\cdot,\cdot\rangle_E$ and $\langle\cdot,\cdot\rangle_F$ defined on E and F, respectively, the operator $\mathcal{A}^*$ adjoint to $\mathcal{A}$ is a linear operator $\mathcal{A}^* : F \to E$ that satisfies

$$\langle \mathcal{A}x, y\rangle_F = \langle x, \mathcal{A}^* y\rangle_E, \quad \forall x \in E, \forall y \in F.$$

Using linearity, a linear operator can be characterized by the values it takes on any generating subspace of E. For instance, a standard way of representing the linear operator $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^m$ is to use the *matrix* $A \in \mathbb{R}^{m\times n}$ defined by $Ae_i \doteq \mathcal{A}e_i, \forall i$. Note that vectors of size n can be treated as $n \times 1$ matrices.

Let us now focus on the matrix representation of linear operators. Given a matrix $A \in \mathbb{C}^{m\times n}$, its transpose A^T is an element of $\mathbb{C}^{n\times m}$ that satisfies $[A^T]_{ij} = [A]_{ji}$. Its conjugate transpose A^* is an element of $\mathbb{C}^{n\times m}$ that satisfies $[A^*]_{ij} = \overline{[A]_{ji}}$. A square matrix $A \in \mathbb{C}^{n\times n}$ is said to be symmetric if $A = A^T$, anti-symmetric if $A = -A^T$, Hermitian if $A = A^*$ and skew-Hermitian if $A = -A^*$.

The sets of real symmetric matrices and of Hermitian matrices (of size n) are denoted by $\mathcal{S}^n \doteq \{X \in \mathbb{R}^{n\times n} : X = X^T\}$ and $\mathcal{H}^n \doteq \{X \in \mathbb{C}^{n\times n} : X = X^*\}$, respectively.

Let the *trace* of $A \in \mathbb{C}^{n\times n}$ be defined as the sum of its diagonal elements. Two additional inner products, defined on matrix spaces and extensively used throughout this thesis, are provided in the next examples.

Example 1.6. The standard inner product on $\mathbb{R}^{n\times n}$ is defined by

$$\langle X, Y\rangle_{\mathbb{R}} \doteq \operatorname{Trace} Y^T X.$$

The induced norm is known as the *Frobenius norm* of X, $\|X\|_F^2 = \operatorname{Trace} X^T X = \sum_{i,j=0}^{n} X_{ij}^2$.

Example 1.7. The standard real-valued inner product on $\mathbb{C}^{n\times n}$ is defined by

$$\langle X, Y\rangle_{\mathbb{R}} \doteq \operatorname{Re}\operatorname{Trace} Y^* X.$$

The induced norm is also called the *Frobenius norm* of X, $\|X\|_F^2 = \operatorname{Trace} X^*X = \sum_{i,j=0}^{n}|X_{ij}|^2$.

A very important class of matrix norms consists of *subordinate norms*

$$\|A\|_{p,q} \doteq \max\{\|Ax\|_p : \|x\|_q = 1\}.$$

Example 1.8. The subordinate matrix norm defined by Euclidean norms,

$$\|A\|_2 \doteq \max\{\|Ax\|_2 : \|x\|_2 = 1\},$$

is called the *spectral norm* of A.

Given a square matrix $A \in \mathbb{C}^{n\times n}$, an eigenpair $(x, \lambda) \in \mathbb{C}^n \times \mathbb{C}$ $(x \neq 0)$ satisfies the equality $Ax = \lambda x$. The component x is called an *eigenvector* of A and the component λ is an associated *eigenvalue*. The eigenvalues of A are the n roots of the characteristic polynomial $\det(\lambda I - A)$, where $\det X$ denotes the *determinant* of X.

It is well known that Hermitian matrices necessarily have real eigenvalues. In this case, we define $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to be the smallest and largest eigenvalues of A, respectively.

Definition 1.5. A Hermitian matrix $A \in \mathbb{C}^{n\times n}$ is said to be *positive semidefinite* if

$$x^*Ax = \langle Ax, x\rangle \geq 0, \quad \forall x \in \mathbb{C}^n.$$

It is said to be *positive definite* if strict inequality holds for all non-zero vectors $x \in \mathbb{C}^n$. These properties are denoted by $A \succeq 0$ and $A \succ 0$, respectively.

Clearly, the eigenvalues of a positive semidefinite (positive definite) matrix are non-negative (strictly positive). Positive definite matrices induce norms related to the Euclidean norm.

Example 1.9. Let A be a positive definite matrix. The *norm induced by* A is defined by

$$\|x\|_A \doteq \sqrt{\langle Ax, x\rangle}.$$

Since E is a finite-dimensional vector space, all induced norms $\|\cdot\|_A$ on E are topologically equivalent.

On matrix spaces, the standard concept of inner product can be extended so that it becomes matrix-valued. This will prove very useful in some contexts.

Definition 1.6. A *matrix-valued inner product* $\langle\langle\cdot,\cdot\rangle\rangle$ on E is a function

$$\langle\langle\cdot,\cdot\rangle\rangle : E \times E \to \mathbb{C}^{m\times m}$$

that satisfies three basic properties:

1. $\langle\langle \alpha x + \beta y, z\rangle\rangle = \alpha\langle\langle x, z\rangle\rangle + \beta\langle\langle y, z\rangle\rangle, \quad \forall x, y, z \in E,\ \forall \alpha, \beta \in \mathbb{F}$;

2. $\langle\langle y, x\rangle\rangle = \langle\langle x, y\rangle\rangle^*, \quad \forall x, y \in E;$

3. $\langle\langle x, x\rangle\rangle \succeq 0, \quad \forall x \in E$ and equality holds if and only if $x = 0$.

In this thesis, we shall encounter very frequently block matrices. By convention, the block size is always denoted by m. Given a block matrix Y, the block (i, j) is Y_{ij}. Of course, for a standard matrix W, W_{ij} is the element (i, j). Therefore $Y_{ij,kl}$ is the element (k, l) of the block (i, j) of the block matrix Y.

Given two matrices $A \in \mathbb{F}^{m_a \times n_a}$ and $B \in \mathbb{F}^{m_b \times n_b}$, the *Kronecker product* $A \otimes B$ is a block matrix of dimension $m_a m_b \times n_a n_b$ defined by

$$A \otimes B \doteq \begin{bmatrix} A_{0,0}B & \dots & A_{0,n_a-1}B \\ \vdots & \ddots & \vdots \\ A_{m_a-1,0}B & \dots & A_{m_a-1,n_a-1}B \end{bmatrix}. \tag{1.5}$$

1.2 Convex analysis

In convex analysis, our main mathematical concept is convexity. Although this property is frequently encountered in practice, it is too often hidden by the formulation of the problem. An important task is thus to reformulate the problem so as to reveal convexity. Starting from the definition of convex sets, mathematicians have started to define convex functions, to study their properties and to combine them in order to obtain more complex objects. In this section, we focus on several elements pertinent to the matter we shall discuss in the next section, namely convex optimization.

From now on, E is a finite-dimensional vector space over $\mathbb{R}$.

1.2.1 Convex sets and cones

A set $S \subset E$ is said to be *convex* if $\forall x_1, x_2 \in S, \forall \alpha \in [0, 1]$, $\alpha x_1 + (1 - \alpha)x_2 \in S$. The following propositions state three elementary properties of convex sets.

Proposition 1.2 ([126, page 16]). *Let S be a convex set. Then $\alpha S = \{y : y = \alpha x, x \in S\}$ is convex for any scalar α.*

Proposition 1.3 ([126, Theorem 3.1]). *Let S and T be two convex sets. Then $S + T = \{z : z = x + y, x \in S, y \in T\}$ is convex.*

Proposition 1.4 ([126, page 18]). *Let $\{S_i\}_{i \in I}$ be an arbitrary collection of convex sets. Then $\cap_{i \in I} S_i$ is convex.*

Convex sets have a lot of important theoretical properties and are frequently encountered in practice. For these reasons, a non-convex set is often approximated by the "closest" convex set: the *convex hull* (denoted by $\operatorname{co} S$) of an arbitrary set S is the smallest convex set containing S.

A set $\mathcal{K} \subset E$ is a *cone* if it is closed under non-negative scalar multiplication, i.e., $x \in \mathcal{K}$ implies that $\alpha x \in \mathcal{K}, \forall \alpha \geq 0$. Moreover, if given $x, y \in \mathcal{K}$, $x + y \in \mathcal{K}$, the cone is in fact convex.

A convex cone $\mathcal{K}$ is *pointed* if it contains no straight line. For closed cones, this geometric property is equivalent to $\mathcal{K} \cap -\mathcal{K} = \{0\}$. The cone $\mathcal{K}$ is said to be *solid* if its interior (denoted by $\operatorname{int} \mathcal{K}$) is not empty.

Definition 1.7. A cone $\mathcal{K}$ is said to be *proper* if it is closed, convex, pointed and solid.

Given a pointed closed convex cone $\mathcal{K} \subset E$, it induces a *partial ordering* on E:

$$x \succeq y \text{ if and only if } x - y \in \mathcal{K}.$$

Remember that partial orderings have the following three properties:

1. Reflexivity: $x \succeq x$
2. Anti-symmetry: if $x \succeq y$ and $y \succeq x$, then $x = y$.
3. Transitivity: if $x \succeq y$ and $y \succeq z$, then $x \succeq z$.

If $\mathcal{K}$ is a proper cone, the closedness of $\mathcal{K}$ makes it possible to pass to limits in $\succeq$-inequalities; if $x_i \succeq y_i$, $\lim_{i\to+\infty} x_i = x$ and $\lim_{i\to+\infty} y_i = y$, then $x \succeq y$. The non-emptiness of $\operatorname{int}\mathcal{K}$ allows us to define the strict inequality:

$$x \succ y \text{ if and only if } x - y \in \operatorname{int}\mathcal{K}.$$

As observed by Beckenback and Bellman [12, Chapter 3, §1], an element in a linear space can often be characterized most readily and revealing in terms of its interaction with a suitably chosen set of elements in a dual space. This remark paves the way for defining the dual cone.

Given a cone $\mathcal{K} \subset E$ and an inner product $\langle \cdot, \cdot \rangle$ on E, the cone

$$\mathcal{K}^\star \doteq \{s \in E : \langle s, x\rangle \geq 0,\ \forall x \in \mathcal{K}\} \tag{1.6}$$

is called the *dual* of $\mathcal{K}$. By definition, this cone is closed and convex.

Let us point out three results on the interaction between the algebra of cones and duality.

Theorem 1.5 ([16, Theorem 2.1]). *Let $\mathcal{K}_1$ and $\mathcal{K}_2$ be two non-empty cones. Then*

$$\begin{aligned}(\mathcal{K}_1 \times \mathcal{K}_2)^\star &= \mathcal{K}_1^\star \times \mathcal{K}_2^\star,\\ (\mathcal{K}_1 + \mathcal{K}_2)^\star &= \mathcal{K}_1^\star \cap \mathcal{K}_2^\star,\\ \mathcal{K}_1^\star + \mathcal{K}_2^\star &\subseteq (\mathcal{K}_1 \cap \mathcal{K}_2)^\star.\end{aligned}$$

Theorem 1.6 ([16, Corollary 2.1]). *If $\mathcal{K} \subset E$ is a convex cone, then* $\operatorname{cl}\mathcal{K} = (\mathcal{K}^\star)^\star$.

Theorem 1.7 ([16, Corollary 2.2]). *Let $\mathcal{K}_1$ and $\mathcal{K}_2$ be two closed convex cones. Then* $\operatorname{cl}(\mathcal{K}_1^\star + \mathcal{K}_2^\star) = (\mathcal{K}_1 \cap \mathcal{K}_2)^\star$.

For closed convex cones, there exists an important relationship between primal pointedness and dual solidness.

Theorem 1.8 ([16, Theorem 2.3]). *Let $\mathcal{K}$ be a closed convex cone. Then $\mathcal{K}$ is pointed if and only if $\mathcal{K}^\star$ is solid.*

Because the dual cone is closed and convex, we get the following corollary.

Corollary 1.9 ([14, Corollary 2.3.1]). *Let $\mathcal{K}$ be a closed convex cone. Then $\mathcal{K}$ is proper if and only if $\mathcal{K}^\star$ is proper.*

If $\mathcal{K}$ is a pointed closed convex cone, then $\operatorname{int}\mathcal{K}^\star$ is given algebraically by

$$\operatorname{int}\mathcal{K}^\star = \{s \in E : \langle s, x\rangle > 0,\ \forall x \neq 0 \in \mathcal{K}\}. \tag{1.7}$$

Of course, the interior of a solid closed convex cone $\mathcal{K}$ is thus equal to

$$\operatorname{int}\mathcal{K} = \{x \in E : \langle s, x\rangle > 0,\ \forall s \neq 0 \in \mathcal{K}^\star\}. \tag{1.8}$$

As a consequence, we get the following result.

Proposition 1.10 ([16, Corollary 2.3]). *Let $\mathcal{K}_1$ and $\mathcal{K}_2$ be closed convex cones with non-empty interiors. If $\mathcal{K} = \mathcal{K}_1 \times \mathcal{K}_2$, then* $\operatorname{int}\mathcal{K} = \operatorname{int}\mathcal{K}_1 \times \operatorname{int}\mathcal{K}_2$.

Example 1.10. Three important proper cones are:

1. Non-negative n-orthant: $\mathbb{R}^n_+ \doteq \{x \in \mathbb{R}^n : x_i \geq 0, 1 \leq i \leq n\}$;
2. Lorentz cone: $\mathcal{L}_n \doteq \{z = (x, \tau) \in \mathbb{R}^{n+1} : \tau \geq \|x\|_2\}$;
3. Positive semidefinite $n \times n$ real matrices:

$$\mathcal{S}^n_+ \doteq \{X \in \mathbb{R}^{n\times n} : X = X^T,\ X \succeq 0\}.$$

The Lorentz cone is also known as the ice-cream or second-order cone. The partial order on $\mathcal{S}^n_+$ is called the *Löwner partial order*. These three cones are all self-dual:

$$(\mathbb{R}^n_+)^\star = \mathbb{R}^n_+, \quad (\mathcal{L}_n)^\star = \mathcal{L}_n, \quad (\mathcal{S}^n_+)^\star = \mathcal{S}^n_+.$$

Example 1.11. The cone of positive semidefinite $n \times n$ Hermitian matrices

$$\mathcal{H}^n_+ \doteq \{X \in \mathbb{C}^{n\times n} : X = X^*,\ X \succeq 0\}$$

is proper and $(\mathcal{H}^n_+)^\star = \mathcal{H}^n_+$. It can be identified with a linear subspace of $\mathcal{S}^{2n}_+$. To see this, consider the linear transformation

$$L : \mathcal{H}_n \to \mathcal{S}_{2n} : X \to L(X) = \begin{bmatrix} \operatorname{Re} X & -\operatorname{Im} X \\ \operatorname{Im} X & \operatorname{Re} X \end{bmatrix}$$

and note that $X \succeq 0$ if and only if $L(X) \succeq 0$. Indeed, the equivalence

$$X(x + \jmath y) = \lambda(x + \jmath y) \iff \begin{bmatrix} \operatorname{Re} X & -\operatorname{Im} X \\ \operatorname{Im} X & \operatorname{Re} X \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix}$$

holds for all $\lambda \in \mathbb{R}$ and $x, y \in \mathbb{R}^n$.

1.2.2 Convex functions

Let us define the domain of the function $f : E \to \mathbb{R}$ as

$$\operatorname{dom} f \doteq \{x \in E : |f(x)| < +\infty\}. \tag{1.9}$$

The function $f : E \to \mathbb{R}$ is called *convex* if its domain $\operatorname{dom} f$ is convex and, for all $x, y \in \operatorname{dom} f$ and for all $\alpha \in [0, 1]$, the following inequality holds:

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y). \tag{1.10}$$

A function f is *concave* if $-f$ is convex.

Proposition 1.11 ([126, Theorem 4.1]). *A function $f : E \to \mathbb{R}$ is convex if and only if its* epigraph

$$\operatorname{epi} f \doteq \{(x,t) \in \operatorname{dom} f \times \mathbb{R} : t \geq f(x)\} \tag{1.11}$$

is a convex set.

A convex function $f : E \to \mathbb{R}$ is called *closed* if its epigraph $\operatorname{epi} f$ is a closed convex set.

Example 1.12. 1. Any linear function is closed and convex.

2. Any norm on E is closed and convex. Convexity is a direct consequence of Definition 1.1.

3. The function $f(x) = -\log x$ is convex on its domain. A straightforward consequence of this fact is the celebrated *arithmetic-mean – geometric-mean inequality* [12, Chapter 1, §4 – §16]. Given non-negative reals $\{x_i\}_{i=0}^{n-1}$, it reads

$$(\prod_{i=0}^{n-1} x_i)^{\frac{1}{n}} \leq \frac{\sum_{i=0}^{n-1} x_i}{n}. \tag{1.12}$$

Let us now focus on a few operations that preserve convexity. Using these rules and several elementary convex functions, we can build more complex convex functions or prove convexity of a given function.

Theorem 1.12 (Multiplication,[126, page 33]). *Let $\lambda \geq 0$ and f_1 be a closed convex function. Then the function $f(x) = \lambda f_1(x)$ is closed and convex with* $\operatorname{dom} f = \operatorname{dom} f_1$.

Theorem 1.13 (Addition,[126, Theorem 5.2]). *Let the functions f_1 and f_2 be closed and convex. Then the function $f(x) = f_1(x) + f_2(x)$ is closed and convex with* $\operatorname{dom} f = \operatorname{dom} f_1 \cap \operatorname{dom} f_2$.

Theorem 1.14 (Pointwise supremum,[126, Theorem 5.5]). *Let the functions $\{f_i(x)\}_{i\in I}$ be closed and convex. Then the function f defined by*

$$f(x) = \sup\{f_i(x), i \in I\}, \quad \forall x \in \operatorname{dom} f,$$

is closed and convex with $\operatorname{dom} f = \cap_{i\in I} \operatorname{dom} f_i$.

Theorem 1.15 (Affine-invariance, [126, Theorem 5.7] and [108, Theorem 3.1.6]). *Let $\phi : E \to \mathbb{R}$ be closed and convex. Consider the affine operator $\mathcal{L} : \mathbb{R}^n \to E : x \to Ax + b$. Then $f(x) = \phi(\mathcal{L}(x))$ is closed and convex with the following domain* $\operatorname{dom} f = \{x \in \mathbb{R}^n : \mathcal{L}(x) \in \operatorname{dom} \phi\}$.

Let $x \in \operatorname{dom} f$. We say that the function $f : E \to \mathbb{R}$ is *differentiable in the direction h at point x* if the limit

$$Df(x)[h] \doteq \lim_{\alpha \downarrow 0} \frac{f(x+\alpha h) - f(x)}{\alpha} \tag{1.13}$$

exists. Of course, this definition can be applied recursively to define the kth directional derivative of f in the direction h at point x. Convex functions are not necessarily smooth, but are differentiable in any direction, as stated in the next theorem.

Theorem 1.16 ([126, Theorem 23.1]). *A convex function f is differentiable in any direction at any point of the interior of its domain.*

For twice differentiable functions f, the first two derivatives

$$[f'(x)]_i \doteq \frac{\partial f(x)}{\partial x_i}, \quad [f''(x)]_{ij} \doteq \frac{\partial^2 f(x)}{\partial x_i \partial x_j},$$

are called the *gradient* and the *Hessian*, respectively. In the context of convex optimization, we need to compute the directional derivatives of $\mathcal{C}^3$ closed convex functions with open domain. At any $x \in \operatorname{dom} f$ and for any direction h, they can be obtained by computing the derivatives of the function

$$\phi(x;t) = f(x+th)$$

with respect to $t \in \operatorname{dom} \phi(x;\cdot)$. We have

$$\begin{aligned} Df(x)[h] &= \phi'(x;t) = \langle f'(x), h\rangle, \\ D^2 f(x)[h,h] &= \phi''(x;t) = \langle f''(x)h, h\rangle, \\ D^3 f(x)[h,h,h] &= \phi'''(x;t). \end{aligned}$$

In general, the Hessian provides some information on the convexity of f.

Proposition 1.17 ([126, Theorem 4.5]). *Let $f : E \to \mathbb{R}$ be a twice differentiable function on an open convex set Q. Then f is convex on Q if and only if its Hessian is positive semidefinite for every $x \in Q$.*

For non-smooth functions, we need some replacement for the gradient. A vector $g(x_0)$ is called the *subgradient* of a convex function f at $x_0 \in \operatorname{dom} f$ if, for any $x \in \operatorname{dom} f$,

$$f(x) \geq f(x_0) + \langle g(x_0), x - x_0\rangle.$$

The set of all subgradients of f at x_0, denoted by $\partial f(x_0)$, is called the *subdifferential* of function f at the point x_0. By definition, this set is closed and convex.

In the remaining part of this section, we focus on a particular class of convex barrier functions, which is extremely important in the context of modern convex optimization. The main feature of these functions is that they satisfy some Lipschitz conditions, on the function itself and on the Hessian, with respect to the local metric defined by the Hessian. It has been shown that both characteristics are essential for designing interior-point methods running in polynomial time [112].

Let $Q \subset E$ be an open non-empty convex domain. A $\mathcal{C}^3$ closed convex function $F : Q \to \mathbb{R}$ is a *self-concordant function* if $\lim_{x \to \partial Q} F(x) = +\infty$ and, for all $x \in Q$ and $h \in E$,

$$|D^3 F(x)[h,h,h]| \leq M_f (D^2 F(x)[h,h])^{\frac{3}{2}}, \tag{1.14}$$

where $M_f \geq 0$. If $M_f = 2$, the self-concordant function is called standard.

An important consequence of self-concordancy is as follows.

Theorem 1.18 ([112, Corollary 2.1.1]). *Let the function $F : Q \to \mathbb{R}$ be self-concordant. If $\operatorname{dom} F = Q$ contains no straight line, then the Hessian $F''(x)$ is non-degenerate at any $x \in \operatorname{dom} F$.*

Furthermore, it has been shown that self-concordant functions can be efficiently minimized using Newton method, see [112, Section 2.2].

A standard self-concordant function $F : Q \to \mathbb{R}$ is called a *ν-self-concordant barrier* for the closed convex domain $\operatorname{cl} Q$ if, in addition,

$$\max_{h \in E} \{2\langle F'(x), h\rangle - \langle F''(x)h, h\rangle\} \leq \nu, \quad \forall x \in \operatorname{dom} f. \tag{1.15}$$

The value ν is called the parameter of the barrier. If the Hessian $F''(x)$ is non-degenerate, inequality (1.15) is equivalent to

$$\langle [F''(x)]^{-1} F'(x), F'(x)\rangle \leq \nu, \quad \forall x \in \operatorname{dom} f. \tag{1.16}$$

Example 1.13. The function $F(x) = -\log x$ is a 1-self-concordant barrier for the ray $\{x \in \mathbb{R} : x \geq 0\}$. Indeed, $\operatorname{dom} F = \{x \in \mathbb{R} : x > 0\}$ is an open non-empty convex set, $\lim_{x \to 0} F(x) = +\infty$ and the derivatives

$$F'(x) = -\frac{1}{x}, \quad F''(x) = \frac{1}{x^2}, \quad F'''(x) = -\frac{2}{x^3},$$

satisfy (1.14) and (1.16) with $M_f = 2$ and $\nu = 1$.

Let us now state a major existence theorem in the theory of self-concordant functions, the consequences of which are crucial for numerical convex optimization.

Theorem 1.19 (Universal barrier, [112, Theorem 2.5.1]). *Let $Q \in \mathbb{R}^n$ be a closed convex set with non-empty interior. If Q contains no straight line, then there exists a self-concordant universal barrier for Q with a parameter in $\mathcal{O}(n)$.*

The paramount importance of this result will become evident in view of Section 1.3. For the moment, it proves that the concept of self-concordant barriers is viable as it covers a wide range of convex sets. However, this result is mainly of theoretical interest. Indeed, it is crucial for convex optimization schemes that the barrier function and its differential characteristics are computable in an explicit form and, in any case, in polynomial time.

For arbitrary convex sets, the universal barrier is not very often known explicitly because of its implicit definition. Therefore, we must rely on barrier calculus in order to find self-concordant barrier functions for most practical problems. Let us mention a few basic combination rules.

Theorem 1.20 (Direct product, [112, Proposition 5.1.2]). *Let F_i be ν_i-self-concordant barriers for Q_i, $(i = 1, 2)$. Then the function $F(x) = F_1(x) + F_2(x)$ is a $\nu_1 + \nu_2$-self-concordant barrier for the convex set $Q = Q_1 \times Q_2$.*

Theorem 1.21 (Intersection, [112, Proposition 5.1.3]). *Let F_i be ν_i-self-concordant barriers for Q_i, $(i = 1, 2)$. Then the function $F(x) = F_1(x) + F_2(x)$ is a $\nu_1 + \nu_2$-self-concordant barrier for the convex set $Q = Q_1 \cap Q_2$.*

Theorem 1.22 (Affine invariance, [112, Proposition 5.1.1]). *Let $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^m : x \to Ax + b$ be an affine operator. Assume that the function $F(y)$ is a ν-self-concordant barrier for the convex set $\operatorname{cl} Q \subset \mathbb{R}^m$. Then $\phi(x) = F(\mathcal{L}(x))$ is a ν-self-concordant barrier for the set $\{x \in \mathbb{R}^n : \mathcal{L}(x) \in \operatorname{cl} Q\}$.*

Let $\mathcal{K} \subset E$ be a closed convex cone with non-empty interior. A self-concordant function for $\mathcal{K}$, say $F(x)$, is said to be a *ν-normal barrier* for $\mathcal{K}$ if it is logarithmically homogeneous:

$$F(\tau x) = F(x) - \nu \ln \tau, \quad x \in \operatorname{int} \mathcal{K}, \tau > 0. \tag{1.17}$$

The next three results explain the practical significance of logarithmically homogeneous self-concordant functions. First, they are natural self-concordant barriers. Then, they also provide us with a self-concordant barrier for the dual cone. Finally, they define a one-to-one transformation between a cone and its dual. Notice that these three properties are all valuable and that they cannot be expected from arbitrary barrier functions.

Proposition 1.23 ([112, Corollary 2.3.2]). *A ν-normal barrier for $\mathcal{K}$ is a ν-self-concordant barrier for $\mathcal{K}$.*

Proposition 1.24 ([112, Theorem 2.4.4]). *If $F(x)$ is a ν-normal barrier for $\mathcal{K}$, then the* conjugate function

$$F_*(s) = \max\{\langle -s, x\rangle - F(x) : x \in \mathcal{K}\}$$

is a ν-normal barrier for $\mathcal{K}^\star$.

Proposition 1.25 (Gradient mapping, [126, Theorem 26.5]). *Let $F(x)$ be a ν-normal barrier for $\mathcal{K}$ and $F_*(s)$ its conjugate. For any $x \in \operatorname{int} \mathcal{K}$ and $s \in \operatorname{int} \mathcal{K}^\star$,*

$$-F'(x) \in \operatorname{int} \mathcal{K}^\star, \quad -F'_*(s) \in \operatorname{int} \mathcal{K}.$$

Moreover, $x = -F'_*(-F'(x))$ and $s = -F'(-F'_*(s))$.

In this thesis, our cones of interest are linear images of semidefinite cones. For these cones, there exist explicit formulae for the corresponding normal barriers.

Proposition 1.26. *The function $F(X) = -\log\det X$ is a n-normal barrier for the cones $\mathcal{S}^n_+$ and $\mathcal{H}^n_+$.*

Proof. Although a proof of this important result can be found in the literature, see e.g. [112, Proposition 5.4.5], we derive it completely below. The function F is clearly $\mathcal{C}^\infty$ on $\operatorname{int} \mathcal{S}^n_+$ and tends to infinity as the argument approach the boundary of $\mathcal{S}^n_+$.

The first three directional derivatives of F can be calculated as follows. Let us fix $X \in \operatorname{int} \mathcal{S}^n_+$ and $H \in \mathcal{S}^n$. Then the following equalities hold

$$\begin{aligned} -\ln\det(X+tH) + \ln\det X &= -\ln\det X^{-\frac{1}{2}}(X+tH)X^{-\frac{1}{2}} \\ &= -\ln\det(I + tX^{-\frac{1}{2}}HX^{-\frac{1}{2}}) \\ &= -\sum_{i=1}^n \ln(1 + t\lambda_i(X^{-\frac{1}{2}}HX^{-\frac{1}{2}})) \end{aligned}$$

and we get

$$\begin{aligned} DF(X)[H] &= \lim_{t \downarrow 0} \frac{-\ln\det(X+tH) + \ln\det X}{t} \\ &= -\sum_{i=1}^n \lambda_i(X^{-\frac{1}{2}}HX^{-\frac{1}{2}}) = -\langle I, X^{-\frac{1}{2}}HX^{-\frac{1}{2}}\rangle = -\langle X^{-1}, H\rangle. \end{aligned}$$

The last equality proves that $F'(X) = -X^{-1}$. In order to compute the second differential derivative, we note that

$$\begin{aligned} -(X+tH)^{-1} + X^{-1} &= -X^{-1}(X-(X+tH))(X-tH)^{-1} \\ &= tX^{-1}H(X+tH)^{-1}. \end{aligned}$$

Hence, we have

$$\lim_{t \downarrow 0} \frac{-(X+tH)^{-1} + X^{-1}}{t} = X^{-1}HX^{-1}$$

and $D^2F(X)[H,H] = \langle X^{-1}HX^{-1}, H\rangle$. The third directional derivative can be obtained similarly by making use of the identity

$$\begin{aligned} &(X+tH)^{-1}H(X+tH)^{-1} - X^{-1}HX^{-1} \\ &\qquad = (X+tH)^{-1}(-2tHX^{-1}H - t^2HX^{-1}HX^{-1}H)(X+tH)^{-1}. \end{aligned}$$

We get that $D^3F(X)[H,H,H] = -2\langle X^{-1}HX^{-1}HX^{-1}, H\rangle$.

Because the Hessian is positive semidefinite, i.e.,

$$D^2F(X)[H,H] = \langle X^{-1}HX^{-1}, H\rangle \geq 0, \forall H \in \mathcal{S}^n,$$

convexity of F follows from Proposition 1.17. In order to prove self-concordancy, we note that

$$D^2F(X)[H,H] = \langle F''(X)H, H\rangle = \sum_{i=0}^{n-1} \lambda_i^2(X^{-\frac{1}{2}}HX^{-\frac{1}{2}}),$$

$$D^3F(X)[H,H,H] = -2\sum_{i=0}^{n-1} \lambda_i^3(X^{-\frac{1}{2}}HX^{-\frac{1}{2}}).$$

From the inequality $|\sum_{i=0}^{n-1} \lambda_i^3| \leq (\sum_{i=0}^{n-1} \lambda_i^2)^{3/2}$ [12, Chapter 1, §16], we obtain that inequality (1.14) with $M_f = 2$ holds for all $X \in \operatorname{int} \mathcal{S}_+^n$ and $H \in \mathcal{S}^n$.

We complete the proof by making use of Proposition 1.23 and of the identity

$$-\log\det(\tau X) = -\log(\tau^n \det X) = -\log\det X - n\log\tau, \quad \forall \tau > 0, X \in \operatorname{int} \mathcal{S}_+^n.$$

The proof for the cone $\mathcal{H}_+^n$ is identical, provided that some obvious changes are made. □

1.3 Convex optimization

Given an arbitrary optimization problem, it is usually difficult to certify that the local minimum found by an algorithm is the global minimum. Convex optimization problems do have this nice additional property that any local minimum is also global. For this reason, convex optimization should preferably be used to model and to solve real-life problems. In fact, convex programming is so powerful that non-convex problems are often approximated by convex problems (or by sequences of convex problems). In this section, the class of convex problems that can be solved in polynomial time, as well as some relevant optimization schemes, are described.

Remember that E denotes a finite-dimensional vector space over $\mathbb{R}$.

1.3.1 Structural programming

A convex programming problem is an optimization problem where the objective function f_0, the functional constraints $\{f_i\}_{i=1}^k$ and the basic feasible set Q are convex:

$$\begin{array}{ll} \inf & f_0(y) \\ \text{s.t.} & f_i(y) \leq u_i, \qquad i = 1, \ldots, k, \\ & y \in Q \subseteq \mathbb{R}^n. \end{array}$$

Given an arbitrary optimization problem, checking convexity of its functional components clearly requires some knowledge about its structure. Therefore the functions $\{f_i\}_{i=0}^k$ cannot be part of a black-box. In order to elude this important conceptual contradiction, structural optimization must be considered.

Let us pause with convex optimization and remember how the linear system $Ax = b$, with A positive definite, is solved in numerical linear algebra. First the Cholesky factor of A is computed in $\mathcal{O}(n^3)$ operations. Then two triangular systems are solved in $\mathcal{O}(n^2)$ operations. Note that solving the system with a different right-hand side is cheap as it only costs $\mathcal{O}(n^2)$. The essence of this procedure is the clever identification and use of the problem structure. This kind of two-step procedure can also be devised in convex optimization.

In structural optimization, we need to describe a class of convex problems which are efficiently solvable. An instance from this class is called a *standard minimization problem* and its formulation is

$$\min\{\langle c, x\rangle : x \in Q \subseteq E\}, \tag{1.18}$$

where $c \in \mathbb{R}^n$ and Q is a closed convex set with non-empty interior. The key point is the fact that a *computable* self-concordant barrier for the set Q must be available. Note that the inner product on E, which appears in (1.18), must be real-valued. Indeed, we cannot minimize complex-valued functions.

In order to solve a given convex programming problem, with structural programming in mind, we apply a two-step procedure. First, the problem of interest is rewritten or reformulated so as to get a standard minimization problem. Then an optimal solution is found using a standard minimization scheme, which uses our knowledge of a self-concordant function for Q. This can be done efficiently via Newton method because of self-concordancy.

In conclusion, structural programming is a standard way of solving a general convex programming problem, which emphasizes the structure of the objective function, of the basic convex set and of the functional constraints, using the framework of self-concordant barriers. By keeping some relevant information about the structure, we obtain an efficient way of solving the problem.

1.3.2 Conic formulation

We can even be more specific about the structure of the convex set Q showing up in a standard minimization problem. Indeed, the convex set Q in (1.18) can be lifted to the cone $\mathcal{K} = \{(tx, t) : x \in Q, t \geq 0\}$. Clearly, the set Q is recovered by considering the cross-section of $\mathcal{K}$ with the hyperplane $t = 1$. That is to say, any closed convex set could be obtained as the intersection of a closed convex cone with an appropriate subspace. Figure 1.1 illustrates the whole procedure. This paves the way for the conic

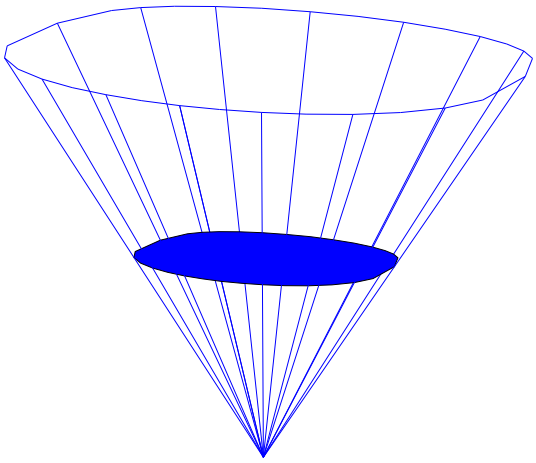

Figure 1.1. *Lifting a convex set into a cone*

formulation of a convex minimization problem.

Let $\mathcal{K}$ be a proper cone in E. Given an objective $c \in E$, a linear mapping $\mathcal{A} : E \to \mathbb{R}^k$ and a right hand side $b \in \mathbb{R}^k$, consider the primal minimization problem

$$\begin{array}{ll} \min & \langle c, x \rangle \\ \text{s.t.} & \mathcal{A}x = b, \\ & x \in \mathcal{K}. \end{array} \tag{1.19}$$

Without loss of generality, we assume that $\mathcal{A}$ is a surjective operator. Formulation (1.19) is known as the *conic formulation* of a convex optimization problem and any convex optimization problem can be reformulated like this.

Let $\mathcal{K}^\star$ be the cone dual to $\mathcal{K}$ and $\mathcal{A}^* : \mathbb{R}^k \to E$ be the operator adjoint to $\mathcal{A}$. Then the problem dual to (1.19) is defined by

$$\begin{array}{ll} \max & \langle b, y \rangle \\ \text{s.t.} & s + \mathcal{A}^* y = c, \\ & s \in \mathcal{K}^\star. \end{array} \tag{1.20}$$

Primal and dual problems have the same structure; a linear function is optimized over the intersection of an affine plane with a cone in both settings. This observation is strengthened in the next proposition.

Proposition 1.27 ([14, Theorem 2.4.1]). *Duality is symmetric; the dual problem* (1.20) *is conic and the problem dual to the dual is equivalent to the primal problem* (1.19).

In order to avoid a misleading terminology, the problems in which the feasible set is formed as an intersection of a convex cone with an affine subspace, given by the equality constraints, will always be called the *primal problems*. The convex cone in this setting will be seen as the primal cone. The problems in which the feasible set is formed by an intersection of an image of a linear operator with some convex cone will be called the *dual problems*. The convex cone in this setting will be seen as the dual cone.

Conic formulation of optimization problems covers a wide spectrum of applications. Let us consider two famous examples.

Example 1.14 (Linear programming, [33]). Let $E = \mathbb{R}^n, \mathcal{K} = \mathbb{R}^n_+, \mathcal{A} = A \in \mathbb{R}^{k\times n}$. Then the primal problem

$$\begin{array}{ll} \min & \langle c, x\rangle \\ \text{s.t.} & Ax = b, \\ & x \geq 0 \end{array}$$

is a canonical linear programming problem and the dual

$$\begin{array}{ll} \max & \langle b, y\rangle \\ \text{s.t.} & s = c - A^T y, \\ & s \geq 0 \end{array}$$

is the standard dual problem. This example shows that the assumption on $\mathcal{A}$ is a natural one; if the row rank of A is equal to $\tilde{k}$, $\tilde{k} < k$, then either one can eliminate $(k - \tilde{k})$ redundant constraints or the system $Ax = b$ is not consistent.

Example 1.15 (Semidefinite programming, [144, 150]). Let $E = \mathcal{S}^n, \mathcal{K} = \mathcal{S}^n_+$. Given a set of symmetric matrices $A_0, \cdots, A_{k-1} \in \mathcal{S}^n$, we can define a linear operator $\mathcal{A}$ by

$$\mathcal{A} : \mathcal{S}^n \to \mathbb{R}^k : X \to \big[\langle A_0, X\rangle, \quad \ldots, \quad \langle A_{k-1}, X\rangle)\big]^T$$

The adjoint operator $\mathcal{A}^* : \mathbb{R}^k \to \mathcal{S}^n$ is $\mathcal{A}^* y = \sum_{i=0}^{k-1} y_i A_i$. Then the primal problem is a standard semidefinite programming problem

$$\begin{array}{ll} \min & \langle C, X\rangle \\ \text{s.t.} & \langle A_i, X\rangle = b_i, \quad i = 0, \ldots, k-1, \\ & X \succeq 0 \end{array}$$

and the dual is

$$\begin{array}{ll} \max & \langle b, y\rangle \\ \text{s.t.} & S = C - \sum_{i=0}^{k-1} y_i A_i, \\ & S \succeq 0. \end{array}$$

By convention, the optimal objective value of the primal (dual) problem is set to $+\infty$ ($-\infty$) if this problem is infeasible. Of course, the primal and dual optimal values are strongly related, as shown by the following proposition.

Proposition 1.28 (Duality gap, [14, Theorem 2.4.1]). *The value of the dual objective at every dual feasible solution y is less or equal to the value of the primal objective at every primal feasible solution x, so that the duality gap*

$$\langle c, x\rangle - \langle b, y\rangle \tag{1.21}$$

is non-negative at every primal-dual feasible pair (x, y).

In the context of conic programming, there might exist a non-zero duality gap at a pair (x, y) of optimal solutions. An example of this non-trivial situation is as follows.

Example 1.16 (Non-zero duality gap). Consider the semidefinite programming problem defined by the matrices

$$C = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, A_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, A_1 = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and the vector $b = \begin{bmatrix} 0 & 1 \end{bmatrix}^T$. On the one hand, the primal problem is equivalent to

$$\begin{array}{ll} \min & 2X_{01} \\ \text{s.t.} & \begin{bmatrix} 0 & X_{01} & X_{02} \\ X_{01} & X_{11} & X_{12} \\ X_{02} & X_{12} & 1 + 2X_{01} \end{bmatrix} \succeq 0 \end{array} .$$

In order to satisfy the constraint, we must have $X_{01} = X_{02} = 0$ because of the zero in the upper left corner. The optimal (primal) value is thus equal to 0. On the other hand, the dual problem is

$$\begin{array}{ll} \max & y_1 \\ \text{s.t.} & \begin{bmatrix} -y_0 & 1 + y_1 & 0 \\ 1 + y_1 & 0 & 0 \\ 0 & 0 & -y_1 \end{bmatrix} \succeq 0 \end{array} .$$

The zero entry at the center of the constraint implies that $y_1 = -1$ so that the optimal (dual) value is equal to -1. Our semidefinite programming problem exhibits a duality gap equal to 1.

Fortunately, non-zero duality gaps at optimal solutions are not generic as they are related to pathological instances. The next results illustrate this claim.

Proposition 1.29 (Primal (dual) strict feasibility, [14, Theorem 2.4.1]). *If the primal (dual) is bounded below (above) and strictly feasible, then the dual (primal) is solvable and the optimal values of both problems are equal to each other.*

Proposition 1.30 (Optimality, [14, Theorem 2.4.1]). *If the primal problem and/or the dual problem is bounded and strictly feasible, a primal-dual feasible pair (x, y) is optimal if and only if the duality gap (1.21) at this pair is equal to zero and complementary slackness holds, i.e. $\langle x, c - \mathcal{A}^* y \rangle = 0$.*

Corollary 1.31 ([14, Corollary 2.4.1]). *Assume that both primal and dual problems are strictly feasible. Then (i) both problems are solvable, (ii) the optimal values of both problems are equal to each other and (iii) either zero duality gap or complementary slackness are necessary and sufficient conditions for optimality of a primal-dual feasible pair.*

1.3.3 Interior-point methods

In order to solve convex optimization problems, efficient interior-point methods have been developed and implemented during the last decade. These *iterative methods* minimize (or maximize) the objective function by moving *inside the feasible set*, up to an *approximate solution*. An important feature of these interior-point methods is

that they run in polynomial time (provided that a *computable* self-concordant barrier is available).

Although there exist many different interior-point methods, we focus in this thesis on short-step path-following methods. These methods start from an initial point and follow some central trajectory. Of course, any interior-point method usually requires our feasible sets to be strictly feasible. The regularity assumption is as follows.

Assumption 1.1 (Regularity). There exist points $\tilde{x} \in \operatorname{int} \mathcal{K}$, $\tilde{s} \in \operatorname{int} \mathcal{K}^\star$ and $\tilde{y} \in \mathbb{R}^k$ that satisfy the following linear system

$$A\tilde{x} = b, \quad \tilde{s} + A^*\tilde{y} = c. \tag{1.22}$$

Note that Assumption 1.1 enforces the conditions of Corollary 1.31, which ensures solvability and zero duality gap. If a problem fulfils this regularity condition, it is thus necessarily tamed. Otherwise, the problem is extremely inappropriate for interior-point methods. Indeed, the regularity assumption is associated with the existence of central trajectories.

Proposition 1.32 ([109, Lemmata 1 and 2]).

1. *The* primal central path

$$x(t) = \arg\min\{t\langle c, x\rangle + F(x) | Ax = b, x \in \mathcal{K}\}, t > 0$$

is well defined if and only if problem (1.19) *satisfies Assumption 1.1.*

2. *The* dual central path

$$y(t) = \arg\min\{-t\langle b, y\rangle + F_*(c - A^*y) | c - A^*y \in \mathcal{K}^\star\}, t > 0$$

is well defined if and only if problem (1.20) *satisfies Assumption 1.1.*

Because of this result, Assumption 1.1 is of paramount importance from a theoretical, as well as a practical, point of view. If an optimization problem does not satisfy it, one should necessarily think about a reformulation.

In Chapter 3, we focus on an efficient scheme for solving a dual problem of the form (1.20). Therefore, we are mainly interested in the dual central path, which can be tracked using a path-following algorithm [108, Chapter 4]. Indeed, this iterative process allows us to solve the dual problem up to an accuracy ϵ. Reasonable values for the parameters of the schemes given below are

$$\beta = \frac{1}{9}, \quad \gamma = \frac{\sqrt{\beta}}{1 + \sqrt{\beta}} - \beta = \frac{5}{36}.$$

Algorithm 1.1 Main path-following scheme (dual problem)

1. Set $t_0 = 0$. Choose an accuracy $\epsilon > 0$ and $y_0 \in \operatorname{dom} F_*$ satisfying the centering condition

$$\langle F_*''(y_0)^{-1} F_*'(y_0), F_*'(y_0)\rangle^{1/2} \leq \beta. \tag{1.23}$$

2. kth iteration ($k \geq 0$). Set

$$t_{k+1} = t_k + \frac{\gamma}{\langle F''_*(y_k)^{-1}b, b\rangle^{1/2}},$$
$$y_{k+1} = y_k - F''_*(y_k)^{-1}(-t_{k+1}b + F'_*(y_k)).$$

3. Stop the process if $\nu + \frac{(\beta+\sqrt{\nu})\beta}{1-\beta} \leq \epsilon t_k$.

In order to find a starting point y_0 in the neighbourhood of the central path, i.e., satisfying (1.23), an auxiliary scheme can be performed. For instance, we may use the next algorithm, which follows the auxiliary central path

$$x^*(t) = \arg\min\{-t\langle F'_*(x_0), x\rangle + F_*(x) : x \in \operatorname{dom} F_*\} \tag{1.24}$$

with $x_0 \in \operatorname{dom} F_*$. Note that $x^*(1) = x_0$ and $x^*(0)$ is the analytic center of $\operatorname{dom} F_*$.

Algorithm 1.2 Auxiliary path-following scheme

1. Set $t_0 = 1$. Choose $x_0 \in \operatorname{dom} F_*$.
2. kth iteration ($k \geq 0$). Set

$$t_{k+1} = t_k - \frac{\gamma}{\langle F''_*(x_k)^{-1}F'_*(x_0), F'_*(x_0)\rangle^{1/2}},$$
$$x_{k+1} = x_k - F''_*(x_k)^{-1}(t_{k+1}F'_*(x_0) + F'_*(x_k)).$$

3. Stop the process if $\langle F''_*(x_k)^{-1}F'_*(x_k), F'_*(x_k)\rangle^{1/2} \leq \frac{\sqrt{\beta}}{1+\sqrt{\beta}}$.
4. Set $y_0 = x_k - F''_*(x_k)^{-1}F'_*(x_k)$ so that

$$\langle F''_*(y_0)^{-1}F'_*(y_0), F'_*(y_0)\rangle^{1/2} \leq \beta.$$

It can be proved that the total complexity of the path-following scheme described above is $\mathcal{O}(\sqrt{\nu}\log\frac{\nu}{\epsilon})$, where ϵ is the required accuracy and ν is the barrier parameter [108].

If the self-concordant barrier functions $F(x)$ and $F_*(s)$ are both computable, problems (1.19) and (1.20) can be unified in the *primal-dual conic formulation*

$$\begin{array}{ll} \min & \langle c, x\rangle - \langle b, y\rangle \\ \text{s.t.} & Ax = b, \\ & s + A^*y = c, \\ & x \in \mathcal{K}, s \in \mathcal{K}^\star. \end{array} \tag{1.25}$$

In this context, we define the primal-dual central path, which can be followed using an appropriate scheme. The theoretical result is stated in the next theorem.

Theorem 1.33 ([109, Theorem 1]). *Under Assumption 1.1,*

1. *The optimal value of problem* (1.25) *is zero.*

2. *For any* $\mu > 0$ *the sets*

$$Q(\mu) = \{z = (x, s, y) \ : \ \mathcal{A}x = b, \ s + \mathcal{A}^* y = c, \\ \langle c, x\rangle - \langle b, y\rangle = \mu, x \in \mathcal{K}, s \in \mathcal{K}^\star\}$$

   are non-empty and bounded. The optimal set of problem (1.25) *is also non-empty and bounded.*

3. *The points of the primal-dual central path*

$$z(t) = (x(t), s(t), y(t)), \ t > 0, \ s(t) = c - \mathcal{A}^* y(t)$$

   are well defined and for any $t > 0$*, the following relations hold*

$$\begin{aligned} \langle c, x(t)\rangle - \langle b, y(t)\rangle &= \langle s(t), x(t)\rangle = \nu/t, \\ F(x(t)) + F_*(s(t)) &= -\nu + \nu \ln t, \\ s(t) &= -(1/t)F'(x(t)), \\ x(t) &= -(1/t)F'_*(s(t)). \end{aligned}$$

Although the complexity estimate of an adaptative primal-dual path-following method based on this theorem is $\mathcal{O}(\sqrt{\nu} \log \frac{\nu}{\epsilon})$, the numerical behaviour is much better in practice (between 20 and 80 Newton steps). Indeed, using both primal and dual information, the step size is adapted in function of the curvature of the central path. In non-degenerate situation, we get a quadratic convergence at the end.

Notes and references

Many books on linear algebra and matrix analysis exist. To our point of view, four of them deserve a special attention. The two volumes by Horn and Johnson [81, 82] contain all of the important research results and abundant references. For numerical aspects, the best book is most certainly "Matrix computations" [66]. For a quick reference guide, the book of Marcus and Minc [101] is probably the most complete.

The standard reference book on convex analysis was written by Rockafellar [126]. Our short introduction on cones is based on the first chapter of Berman's lecture notes [16]. The theory of self-concordant functions can be found in [108, 112].

For modern convex optimization, two books are recommended. Nesterov's lecture notes [108] contain a complete overview of convex optimization methods, starting from the basics. The recent book of Ben Tal and Nemirovskii [14] is oriented towards engineering applications and therefore it deserves a special attention. Both books are much easier to read than the original monograph of Nesterov and Nemirovskii [112], which contains the first self-contained theory of polynomial-time interior-point methods for convex optimization. Section 1.3 is based upon these references. Interior-point methods for linear programming have proven to be competitive with the simplex method and a recent book on these methods is [127].

Chapter 2

Non-negative polynomials

In this chapter, our mathematical objects of interest are non-negative matrix polynomials. Given a non-negative real, say $x \in \mathbb{R}_+$, there are several ways of representing it as a sum of squares. For instance, we could write

$$x = \frac{a^2 + b^2}{2} = \left(\frac{a+b}{2}\right)^2 + \left(\frac{a-b}{2}\right)^2,$$

for some real numbers a and b. This example illustrates that a sum-of-squares representation of x is possible, but generically not unique. For non-negative matrix polynomials, sum-of-squares representations are also available. Our goal is to describe several cones of non-negative matrix polynomials and to characterize them using computationally tractable representations based on sum-of-squares decompositions.

Section 2.1 focuses on cones generated by matrix sums of squares of certain functional systems, for which efficient convex parametrizations have been obtained. These results are then applied to several cones of non-negative matrix polynomials in Section 2.2 and of non-negative trigonometric matrix polynomials in Section 2.3. Our results are noteworthy as they rely on the semidefinite representability of our cones. Although our characterizations are original, they are probably hidden in the literature on moment spaces, as shown in Section 2.4. Indeed, non-negative polynomials and moment spaces are dual objects. In the systems theory literature, part of our characterizations can also be obtained via the Kalman-Yakubovich-Popov Lemma, as described in Section 2.5. Unfortunately, our results cannot be extended to multivariate polynomials in a straightforward manner. Section 2.6 illustrates this fact with several examples and with references to the literature.

2.1 Sum-of-squares representations

This first section contains the most important results of this chapter. We focus on cones related to sums of squares of matrix functions. These cones are shown to permit semidefinite representations; they are linear images of cones of semidefinite matrices. Real-valued and complex-valued functional systems are treated separately. Of course, the consequences of our convex characterizations are stressed. Finally, functional weight factors allow us to consider weighted sum-of-squares decompositions.

2.1.1 Functional systems with real value

Let $\Gamma \subseteq \mathbb{C} \times \cdots \times \mathbb{C}$ be an arbitrary continuous set. Given an arbitrary system of linearly independent real-valued functions $\mathcal{S} = \{\psi_0(x), \ldots, \psi_{r-1}(x)\}$ and a real-valued weight function $\phi(x)$ that is non-negative on Γ, we are interested in the characterization of the finite-dimensional cone

$$\mathcal{K} = \{P(x) : P(x) = \phi(x) \sum_{i=0}^{N-1} Q_i^*(x) Q_i(x);\ Q_i(x) \in \mathcal{F}(\mathcal{S}),\ i = 0, \ldots, N-1\} \quad (2.1)$$

where

$$\mathcal{F}(\mathcal{S}) \doteq \{Q(x) : Q(x) = \sum_{k=0}^{r-1} Q_k \psi_k(x);\ Q_k \in \mathbb{C}^{m\times m},\ k = 0, \ldots, r-1\}.$$

In this definition, $N \in \mathbb{N}$ is fixed and large enough, say $N \geq r$.

Our description requires the definition of the squared functional system

$$\mathcal{S}^2 \doteq \{\upsilon_{ij}(x) = \phi(x)\psi_i(x)\psi_j(x); i, j = 0, \ldots, r-1\}$$

and of two vector functions $\psi(x)$ and $\upsilon(x)$. The components of $\psi(x)$ are the elements of $\mathcal{S}$ and those of $\upsilon(x) = \begin{bmatrix}\upsilon_0(x), & \ldots, & \upsilon_{s-1}(x)\end{bmatrix}^T$ are obtained from any basis of the finite-dimensional functional space that covers $\mathcal{S}^2$. This basis is said to be *minimal* if all its elements are needed to span $\mathcal{S}^2$. Let us exemplify the objects involved.

Example 2.1. Let $\Gamma = \mathbb{R}$. If $\mathcal{S} = \{1, x, x^2\}$ and $\phi(x) = 1$, then $\psi(x) = \begin{bmatrix}1, & x, & x^2\end{bmatrix}^T$. A minimal basis for $\mathcal{S}^2$ corresponds to the vector function $\upsilon(x) = \begin{bmatrix}1, & x, & x^2, & x^3, & x^4\end{bmatrix}^T$. Note that $r = 3$ and $s = 5$.

Consider the linear operator

$$\Lambda : \mathbb{R}^s \to \mathbb{R}^{r\times r} : v \to \Lambda(v) = \sum_{k=0}^{s-1} \Lambda_k v_k$$

for which the identity

$$\Lambda(\upsilon(x)) = \phi(x)\psi(x)\psi(x)^T \quad (2.2)$$

holds for all $x \in \Gamma$. This operator Λ can be extended to a block-linear operator from the set

$$E = \{V : V_k = V_k^* \in \mathbb{C}^{m\times m}, k = 0, \ldots, s-1\}$$

to the set

$$F = \{W : W_{ij} = W_{ji}^* \in \mathbb{C}^{m\times m}, i, j = 0, \ldots, r-1\}$$

as follows

$$\Lambda : E \to F : V \to \Lambda(V) \doteq \sum_{k=0}^{s-1} \Lambda_k \otimes V_k. \quad (2.3)$$

If the spaces E and F are endowed with the block inner products

$$\langle\langle \cdot, \cdot \rangle\rangle_E : E \times E \to \mathbb{C}^{m\times m} : X, Y \to \langle\langle X, Y \rangle\rangle_E = \sum_{i=0}^{s-1} Y_i^* X_i,$$

$$\langle\langle \cdot, \cdot \rangle\rangle_F : F \times F \to \mathbb{C}^{m\times m} : X, Y \to \langle\langle X, Y \rangle\rangle_F = \sum_{i,j=0}^{r-1} Y_{ij}^* X_{ij},$$

then the dual block-linear operator $\Lambda^* : F \to E$, which is defined by

$$\langle\langle X, \Lambda^*(Y)\rangle\rangle_E = \langle\langle \Lambda(X), Y\rangle\rangle_F, \quad \forall X \in E, Y \in F,$$

also satisfies

$$\langle X, \Lambda^*(Y)\rangle_E = \langle \Lambda(X), Y\rangle_F, \quad \forall X \in E, Y \in F.$$

The linear operator $\Lambda^* : F \to E$ is thus dual to Λ for both real-valued and matrix-valued inner products. Therefore, the dual operator $\Lambda^* : E \to F$ could also be defined using the inner products

$$\langle \cdot, \cdot \rangle_E : E \times E \to \mathbb{R} : X, Y \to \langle X, Y\rangle_E \doteq \operatorname{Re}\operatorname{Trace}\langle\langle X, Y\rangle\rangle_E,$$
$$\langle \cdot, \cdot \rangle_F : F \times F \to \mathbb{R} : X, Y \to \langle X, Y\rangle_F \doteq \operatorname{Re}\operatorname{Trace}\langle\langle X, Y\rangle\rangle_F,$$

and this definition would be equivalent to the previous one. Note that

$$\Lambda(\Upsilon(x)) = \phi(x)\Psi(x)\Psi(x)^T,$$

where $\Psi(x) \doteq \psi(x) \otimes I_m$ and $\Upsilon(x) \doteq \upsilon(x) \otimes I_m$. Let us continue Example 2.1.

Example 2.2 (Example 2.1, continued). Identity (2.2), rewritten with the functional objects $\phi(x)$, $\psi(x)$ and $\upsilon(x)$ defined in Example 2.1, is given by

$$\begin{bmatrix} 1 & x & x^2 \\ x & x^2 & x^3 \\ x^2 & x^3 & x^4 \end{bmatrix} = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix} \begin{bmatrix} 1 & x & x^2 \end{bmatrix}.$$

Therefore, the matrices $\{\Lambda_k\}_{k=0}^4$ which define the linear operator $\Lambda(v)$ are equal to

$$\Lambda_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \Lambda_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \Lambda_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$
$$\Lambda_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad \Lambda_4 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The block linear operator $\Lambda(V)$ is then defined as follows

$$\Lambda : E \to F : V = \begin{bmatrix} V_0 \\ V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} \to \Lambda(V) = \begin{bmatrix} V_0 & V_1 & V_2 \\ V_1 & V_2 & V_3 \\ V_2 & V_3 & V_4 \end{bmatrix}.$$

We are now able to state and prove our main characterization result.

Theorem 2.1. *1. A function $P(x) = \langle\langle P, \Upsilon(x)\rangle\rangle_E$ belongs to the cone $\mathcal{K}$ defined in (2.1) if and only if there exists a Hermitian positive semidefinite $mr \times mr$ matrix Y such that $P = \Lambda^*(Y)$:*

$$\mathcal{K} = \{P \in E : \exists Y \in \mathcal{H}_+^{mr}, P = \Lambda^*(Y)\}. \tag{2.4}$$

This cone is closed, convex and pointed.

2. *The dual cone $\mathcal{K}^\star$ is equal to $\{S \in E : \Lambda(S) \succeq 0\}$.*

3. *If $\upsilon(x)$ is a minimal system for $\mathcal{S}^2$, then $\mathcal{K}^\star$ is a pointed cone and $\mathcal{K}$ has a non-empty interior.*

Proof. 1. If $P(x) = \langle\langle P, \Upsilon(x)\rangle\rangle_E$ with $P = \Lambda^*(Y)$ and $Y = Y^* \succeq 0$, then we have

$$\begin{aligned} P(x) &= \langle\langle \Lambda^*(Y), \Upsilon(x)\rangle\rangle_E = \langle\langle Y, \Lambda(\Upsilon(x))\rangle\rangle_F \\ &= \langle\langle Y, \phi(x)\Psi(x)\Psi(x)^T\rangle\rangle_F = \phi(x)\Psi(x)^T Y \Psi(x) \succeq 0 \end{aligned}$$

for any $x \in \Gamma$. As Y is positive semidefinite, it can be factorized as $Y = \sum_{i=0}^{N-1} Q_i Q_i^*$ with $Q_i \in \mathbb{C}^{mr\times m}$. We thus obtain a factorization of $P(x)$, i.e., $P(x) = \phi(x)\sum_{i=0}^{N-1}[\Psi(x)^T Q_i][Q_i^*\Psi(x)]$, which proves that $P(x) \in \mathcal{K}$.

Conversely, if $P(x) \in \mathcal{K}$, then there exists a set of functions

$$Q_i(x) = Q_i^*\Psi(x),\ Q_i \in \mathbb{C}^{mr\times m}, \quad i = 0, \ldots, N-1,$$

such that

$$\begin{aligned} P(x) &= \phi(x)\sum_{i=0}^{N-1} \Psi(x)^T Q_i Q_i^* \Psi(x) = \langle\langle \sum_{i=0}^{N-1} Q_i Q_i^*, \phi(x)\Psi(x)\Psi(x)^T\rangle\rangle_F \\ &= \langle\langle \sum_{i=0}^{N-1} Q_i Q_i^*, \Lambda(\Upsilon(x))\rangle\rangle_F = \langle\langle \Lambda^*(\sum_{i=0}^{N-1} Q_i Q_i^*), \Upsilon(x)\rangle\rangle_E. \end{aligned}$$

Thus, we can set $Y = \sum_{i=0}^{N-1} Q_i Q_i^*$ and $P = \Lambda^*(Y)$.

In view of (2.4), $\mathcal{K}$ is a linear image of the convex cone $\mathcal{H}_+^{mr}$ so that it is convex [126, Theorem 3.4]. A sufficient condition which guarantees that $\mathcal{K}$ is closed is that $Y = 0$ is the only positive semidefinite matrix such that $\Lambda^*(Y) = 0$ [126, Theorem 9.1]. Let us prove that the dual operator Λ^* satisfies this condition. For any $Y \in \mathcal{H}_+^{mr}$ such that $\Lambda^*(Y) = 0$, we have

$$0 = \langle\langle \Lambda^*(Y), \Upsilon(x)\rangle\rangle_E, \quad \forall x \in \Gamma.$$

As the matrix Y can be factorized as $Y = \sum_{i=0}^{N-1} Q_i Q_i^*$ with $Q_i \in \mathbb{C}^{mr\times m}$ for all i, the previous identity can be recast as

$$0 = \phi(x)\sum_{i=0}^{N-1} \Psi(x)^T Q_i Q_i^* \Psi(x), \quad \forall x \in \Gamma.$$

Because $\Psi(x)^T Q_i Q_i^* \Psi(x)$ is positive semidefinite for all i, we necessarily have that

$$0 = \langle Q_i Q_i^*, \Psi(x)\Psi(x)^T\rangle_E, \quad \forall x \in \Gamma \setminus \{x : \phi(x) = 0\}.$$

The semi-infinite nature of this equality constraint and the linear independence of the components of $\psi(x)$ then yield that $Q_i = 0$ for all i. Therefore, $Y = 0$ and $\mathcal{K}$ is closed. By a completely similar argument, we prove that $\mathcal{K}$ is pointed. Indeed, if $P \in \mathcal{K} \cap -\mathcal{K}$, then $P(x) \succeq 0$ and $P(x) \preceq 0$ for all $x \in \Gamma$. As a consequence $P(x) = 0$ for all $x \in \Gamma$. By making use of the parametrization $P = \Lambda^*(Y)$, we get

$$0 = P(x) = \langle\langle P, \Upsilon(x)\rangle\rangle_E = \langle\langle \Lambda^*(Y), \Upsilon(x)\rangle\rangle_E, \quad \forall x \in \Gamma.$$

The above analysis can then carried out as before to show that $P = 0$.

2. The expression of the dual cone $\mathcal{K}^\star$ can be derived using the equivalences

$$\begin{aligned} S \in \mathcal{K}^\star &\iff \langle P, S\rangle_E \geq 0, \quad \forall P \in \mathcal{K} \\ &\iff \operatorname{Re}\operatorname{Trace}\langle\langle P, S\rangle\rangle_E \geq 0, \quad \forall P \in \mathcal{K} \\ &\iff \operatorname{Re}\operatorname{Trace}\langle\langle \Lambda^*(Y), S\rangle\rangle_E \geq 0, \quad \forall Y \succeq 0 \\ &\iff \operatorname{Re}\operatorname{Trace}\langle\langle Y, \Lambda(S)\rangle\rangle_F \geq 0, \quad \forall Y \succeq 0 \\ &\iff \langle Y, \Lambda(S)\rangle \equiv \langle Y, \Lambda(S)\rangle_F \geq 0, \quad \forall Y \succeq 0 \\ &\iff \Lambda(S) \succeq 0 \end{aligned}$$

   Since the primal cone is pointed, the interior of $\mathcal{K}^\star$ is non-empty.

3. First, consider the case $m = 1$. In view of Theorems 1.6 and 1.8, it is sufficient to guarantee that the cone $\mathcal{K}^\star$ is pointed. This is equivalent to non-degeneracy of the linear operator Λ because of item 2. By contradiction, assume that there exists $c \neq 0$ such that $\Lambda(c) = 0$. Without loss of generality, we can think that $c_0 = 1$. Then in view of (2.3), we have

$$\Lambda(\upsilon(x)) = \phi(x)\psi(x)\psi(x)^T = \Lambda(\upsilon(x) - \upsilon_0(x)c).$$

   Thus the system $\mathcal{S}^2$ is covered by the functional system $\tilde{\upsilon}(x) = \upsilon(x) - \upsilon_0(x)c$. However, $\tilde{\upsilon}_0(x)$ is equal to zero for all x, which contradicts our minimality assumption on $\mathcal{S}^2$. Therefore, $\mathcal{K}^\star$ is pointed and $\operatorname{int}\mathcal{K}$ is not empty. Finally, let us prove that $\operatorname{int}\mathcal{K}$ is not empty for an arbitrary m. Indeed, an interior point $P \in \operatorname{int}\mathcal{K}$ is obtained via any interior point p from the initial case $m = 1$ by setting $P = p \otimes I_m$. As a consequence, $\mathcal{K}^\star$ is pointed.

□

2.1.2 Functional systems with complex value

Let $\Gamma \subseteq \mathbb{C} \times \cdots \times \mathbb{C}$ be an arbitrary continuous set. Given an arbitrary system of linearly independent complex-valued functions $\mathcal{S} = \{\psi_0(z), \ldots, \psi_{r-1}(z)\}$ and a real-valued weight function $\phi(z)$ that is non-negative on Γ, we would like to characterize the finite-dimensional cone

$$\mathcal{K} = \{P(z) : P(z) = \phi(z) \sum_{i=0}^{N-1} Q_i^*(z)Q_i(z);\ Q_i \in F(\mathcal{S}),\ i = 0, \ldots, N-1\} \tag{2.5}$$

where

$$F(\mathcal{S}) \doteq \{Q(z) : Q(z) = \sum_{k=0}^{r-1} Q_k\psi_k(z);\ Q_k \in \mathbb{C}^{m\times m},\ k = 0, \cdots, r-1\}.$$

In this definition, $N \in \mathbb{N}$ is fixed and large enough, say $N \geq r$.

Our characterization is based on objects similar to that of real-valued functional systems. We define the squared functional system

$$\mathcal{S}^2 = \{\upsilon_{ij}(z) = \phi(z)\psi_i(z)\overline{\psi_j(z)}; i, j = 0, \ldots, r-1\}$$

and two vector functions $\psi(z)$ and $\upsilon(z)$. The components of $\psi(z)$ are the elements of $\mathcal{S}$ and those of $\upsilon(z) = \begin{bmatrix}\upsilon_0(z), & \ldots, & \upsilon_{s-1}(z)\end{bmatrix}^T$ span the finite-dimensional functional space that covers $\mathcal{S}^2$. Let us exemplify the objects involved.

Example 2.3. Let $\Gamma = \mathbb{T}$, the complex unit circle. If $\Gamma = \mathbb{T}$, $\mathcal{S} = \{1, z, z^2\}$ and $\phi(z) = 1$, then $\psi(z) = \begin{bmatrix}1 & z & z^2\end{bmatrix}^T$. An acceptable basis for $\mathcal{S}^2$ corresponds to the vector function $\upsilon(z) = \begin{bmatrix}1 & z & z^2 & z^{-1} & z^{-2}\end{bmatrix}^T$. In what follows, we shall argue that a minimal basis for $\mathcal{S}^2$ corresponds to the vector function $\upsilon(z) = \begin{bmatrix}1 & z & z^2\end{bmatrix}^T$. Therefore, we have $r = 3$ and $s = 3$.

As it was done in the previous subsection, we can define the linear operator

$$\Lambda : \mathbb{C}^s \to \mathbb{C}^{r\times r} : v \to \Lambda(v) = \frac{1}{2}\sum_{k=0}^{s-1}(\Lambda_k v_k + \Lambda_k^* \overline{v_k})$$

which satisfies the identities

$$\Lambda(\upsilon(z)) = \phi(z)\psi(z)\psi(z)^*, \quad \forall z \in \Gamma. \tag{2.6}$$

It can be extended to a block linear operator from the set

$$E = \{V : V_k \in \mathbb{C}^{m\times m}, k = 0, \ldots, s-1\}$$

to the set

$$F = \{W : W_{ij} = W_{ji}^* \in \mathbb{C}^{m\times m}, i, j = 0, \ldots, r-1\}$$

as follows

$$\Lambda : E \to F : V \to \Lambda(V) \doteq \frac{1}{2}\sum_{k=0}^{s-1}(\Lambda_k \otimes V_k + \Lambda_k^* \otimes V_k^*). \tag{2.7}$$

If the spaces E and F are endowed with the block inner products

$$\langle\langle\cdot,\cdot\rangle\rangle_E : E \times E \to \mathbb{C}^{m\times m} : X, Y \to \langle\langle X, Y\rangle\rangle_E = \frac{1}{2}\sum_{i=0}^{s-1}(Y_i^* X_i + Y_i X_i^*),$$

$$\langle\langle\cdot,\cdot\rangle\rangle_F : F \times F \to \mathbb{C}^{m\times m} : X, Y \to \langle\langle X, Y\rangle\rangle_F = \sum_{i,j=0}^{r-1} Y_{ij}^* X_{ij},$$

then the dual block-linear operator $\Lambda^* : F \to E$, which is defined by

$$\langle\langle X, \Lambda^*(Y)\rangle\rangle_E = \langle\langle \Lambda(X), Y\rangle\rangle_F, \quad \forall X \in E, Y \in F,$$

also satisfies

$$\langle X, \Lambda^*(Y)\rangle_E = \langle \Lambda(X), Y\rangle_F, \quad \forall X \in E, Y \in F.$$

As before, the linear operator $\Lambda^* : F \to E$ is thus dual to Λ for both real-valued and matrix-valued inner products. Note that

$$\Lambda(\Upsilon(z)) = \phi(z)\Psi(z)\Psi(z)^*,$$

where $\Psi(z) \doteq \psi(z) \otimes I_m$ and $\Upsilon(z) \doteq \upsilon(z) \otimes I_m$. Let us continue Example 2.3.

Example 2.4 (Example 2.3, continued). Identity (2.6), rewritten with the functional objects $\phi(z)$, $\psi(z)$ and $\upsilon(z)$ defined in Example 2.3, is given by

$$\begin{bmatrix}1 & z^{-1} & z^{-2}\\ z & 1 & z^{-1}\\ z^2 & z & 1\end{bmatrix} = \begin{bmatrix}1\\ z\\ z^2\end{bmatrix}\begin{bmatrix}1 & z^{-1} & z^{-2}\end{bmatrix}.$$

Therefore, the matrices $\{\Lambda_k\}_{k=0}^2$ which define the linear operator $\Lambda(v)$ are equal to

$$\Lambda_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \Lambda_1 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \Lambda_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

Because $\Lambda(v)$ is Hermitian, we only need the components of $\mathcal{S}^2$ that are not conjugate on the set Γ. This motivates the change from $v(z) = \begin{bmatrix} 1, & z, & z^2, & z^{-1}, & z^{-2} \end{bmatrix}^T$ to $v(z) = \begin{bmatrix} 1, & z, & z^2 \end{bmatrix}^T$, see Example 2.3. The block linear operator $\Lambda(V)$ is then defined as follows

$$\Lambda(V) : E \to F : V = \begin{bmatrix} V_0 \\ V_1 \\ V_2 \end{bmatrix} \to \Lambda(V) = \begin{bmatrix} V_0 & V_1^* & V_2^* \\ V_1 & V_0 & V_1^* \\ V_2 & V_1 & V_0 \end{bmatrix}.$$

The following proposition provides us with a closed formula for $P = \Lambda^*(Y)$.

Proposition 2.2. *If* $\Lambda(V) = \frac{1}{2}\sum_{k=0}^{s-1}(\Lambda_k \otimes V_k + \Lambda_k^* \otimes V_k^*)$ *and* $Y = Y^* \in \mathbb{C}^{mr \times mr}$,

$$P = \Lambda^*(Y) \iff P_k = \langle\langle Y, \Lambda_k \otimes I_m \rangle\rangle, \quad k = 0, \ldots, s-1.$$

Proof. First, note that

$$\begin{aligned} \langle\langle Y, \Lambda(V) \rangle\rangle_F &= \langle\langle Y, \frac{1}{2}\sum_{k=0}^{s-1}(\Lambda_k \otimes V_k + \Lambda_k^* \otimes V_k^*) \rangle\rangle_F \\ &= \frac{1}{2}\sum_{k=0}^{s-1}\langle\langle Y, \Lambda_k \otimes V_k + \Lambda_k^* \otimes V_k^* \rangle\rangle_F \\ &= \frac{1}{2}\sum_{k=0}^{s-1}(V_k^*\langle\langle Y, \Lambda_k \otimes I_m \rangle\rangle_F + V_k\langle\langle Y, \Lambda_k^* \otimes I_m \rangle\rangle_F). \end{aligned}$$

Since $\langle\langle Y, \Lambda_k \otimes I_m \rangle\rangle_F = (\langle\langle Y, \Lambda_k^* \otimes I_m \rangle\rangle_F)^*$ for any Hermitian matrix Y, it is clear that $P = \Lambda^*(Y)$ if and only if its components are obtained from

$$P_k = \langle\langle Y, \Lambda_k \otimes I_m \rangle\rangle, \quad k = 0, \ldots, s-1.$$

Indeed, in this case,

$$\begin{aligned} \langle\langle Y, \Lambda_H(V) \rangle\rangle_F &= \frac{1}{2}\sum_{k=0}^{s-1}(V_k^*\langle\langle Y, \Lambda_k \otimes I_m \rangle\rangle_F + V_k\langle\langle Y, \Lambda_k^* \otimes I_m \rangle\rangle_F) \\ &= \frac{1}{2}\sum_{k=0}^{s-1}(V_k^* P_k + V_k P_k^*) = \langle\langle P, V \rangle\rangle_E. \end{aligned}$$

□

We are now able to state and prove our main characterization result.

Theorem 2.3. *1. A function* $P(z) = \langle\langle P, \Upsilon(z) \rangle\rangle_E$ *belongs to the cone* $\mathcal{K}$ *defined in* (2.5) *if and only if there exists a Hermitian positive semidefinite* $(mr \times mr)$*-matrix* Y *such that* $P = \Lambda^*(Y)$*:*

$$\mathcal{K} = \{P \in E : \exists Y \in \mathcal{H}_+^{mr}, P = \Lambda^*(Y)\}. \tag{2.8}$$

This cone is closed, convex and pointed.

2. *The dual cone $\mathcal{K}^\star$ is equal to $\{S \in E : \Lambda(S) \succeq 0\}$.*

3. *If $\upsilon(z)$ is a minimal system for $\mathcal{S}^2$, then $\mathcal{K}^\star$ is a pointed cone and $\mathcal{K}$ has a non-empty interior.*

Proof. 1. If $P(z) = \langle\langle P, \Upsilon(z)\rangle\rangle_E$ with $P = \Lambda^*(Y)$ and $Y = Y^* \succeq 0$, then for any $z \in \Gamma$, we have

$$\begin{aligned} P(z) &= \langle\langle \Lambda^*(Y), \Upsilon(z)\rangle\rangle_E = \langle\langle Y, \Lambda(\Upsilon(z))\rangle\rangle_F \\ &= \langle\langle Y, \phi(z)\Psi(z)\Psi(z)^*\rangle\rangle_F = \phi(z)\Psi(z)^* Y \Psi(z). \end{aligned}$$

The positive semidefinite matrix Y is then factorized as $Y = \sum_{i=0}^{N-1} Q_i Q_i^*$ with $Q_i \in \mathbb{C}^{m\times mr}$. We thus obtain an appropriate factorization of $P(z)$, i.e., $P(z) = \phi(z)\sum_{i=0}^{N-1}[\Psi(z)^* Q_i][Q_i^*\Psi(z)]$.

Conversely, if $P(z) \in \mathcal{K}$, there exists a set of functions

$$Q_i(z) = Q_i^* \Upsilon(z),\ Q_i \in \mathbb{C}^{m\times mr}, i = 0, \ldots, N-1,$$

such that

$$\begin{aligned} P(z) &= \phi(z)\sum_{i=0}^{N-1} \Psi(z)^* Q_i Q_i^* \Psi(z) = \langle\langle \sum_{i=0}^{N-1} Q_i Q_i^*, \phi(z)\Psi(z)\Psi(z)^*\rangle\rangle_F \\ &= \langle\langle \sum_{i=0}^{N-1} Q_i Q_i^*, \Lambda(\Upsilon(z))\rangle\rangle_F = \langle\langle \Lambda^*(\sum_{i=0}^{N-1} Q_i Q_i^*), \Upsilon(z)\rangle\rangle_E. \end{aligned}$$

Thus, we can set $Y = \sum_{i=0}^{N-1} Q_i Q_i^*$ and $P = \Lambda^*(Y)$.

In view of (2.8), $\mathcal{K}$ is a linear image of the convex cone $\mathcal{H}_+^{mr}$ so that it is convex [126, Theorem 3.4]. A sufficient condition which guarantees that $\mathcal{K}$ is closed is that $Y = 0$ is the only positive semidefinite matrix such that $\Lambda^*(Y) = 0$ [126, Theorem 9.1]. Let us prove that the dual operator Λ^* satisfies this condition. For any $Y \in \mathcal{H}_+^{mr}$ such that $\Lambda^*(Y) = 0$, we have

$$0 = \langle\langle \Lambda^*(Y), \Upsilon(z)\rangle\rangle_E, \quad \forall z \in \Gamma.$$

As the matrix Y can be factorized as $Y = \sum_{i=0}^{N-1} Q_i Q_i^*$ with $Q_i \in \mathbb{C}^{m\times mr}$ for all i, the previous identity can be recast as

$$0 = \phi(z)\sum_{i=0}^{N-1} \Psi(z)^* Q_i Q_i^* \Psi(z), \quad \forall z \in \Gamma.$$

Because $\Psi(z)^* Q_i Q_i^* \Psi(z)$ is positive semidefinite for all i, we must have that

$$0 = \langle Q_i Q_i^*, \Psi(z)\Psi(z)^*\rangle_E, \quad \forall z \in \Gamma \setminus \{z : \phi(z) = 0\}.$$

The semi-infinite nature of this equality constraint and the linear independence of the components of $\psi(z)$ then yield that $Q_i = 0$ for all i. Therefore, $Y = 0$ and $\mathcal{K}$ is closed. By a completely similar argument we prove that $\mathcal{K}$ is pointed. Indeed, if $P \in \mathcal{K} \cap -\mathcal{K}$, then $P(z) \succeq 0$ and $P(z) \preceq 0$ for all $z \in \Gamma$. As a consequence $P(z) = 0$ for all $z \in \Gamma$. By making use of the parametrization $P = \Lambda^*(Y)$, we get

$$0 = P(z) = \langle\langle P, \Upsilon(z)\rangle\rangle_E = \langle\langle \Lambda^*(Y), \Upsilon(z)\rangle\rangle_E, \quad \forall z \in \Gamma.$$

The above analysis can then carried out as before to show that $P = 0$.

2. The expression of the dual cone $\mathcal{K}^\star$ can be derived as follows:

$$\begin{aligned} S \in \mathcal{K}^\star &\iff \langle P, S\rangle \geq 0, \quad \forall P \in \mathcal{K} \\ &\iff \operatorname{Re}\operatorname{Trace}\langle\langle P, S\rangle\rangle_E \geq 0, \quad \forall P \in \mathcal{K} \\ &\iff \operatorname{Re}\operatorname{Trace}\langle\langle \Lambda^*(Y), S\rangle\rangle_E \geq 0, \quad \forall Y \succeq 0 \\ &\iff \operatorname{Re}\operatorname{Trace}\langle\langle Y, \Lambda(S)\rangle\rangle_F \geq 0, \quad \forall Y \succeq 0 \\ &\iff \langle Y, \Lambda(S)\rangle \equiv \langle Y, \Lambda(S)\rangle_F \geq 0, \forall Y \succeq 0 \\ &\iff \Lambda(S) \succeq 0. \end{aligned}$$

Since the primal cone is pointed, the interior of $\mathcal{K}^\star$ is non-empty.

3. First, consider the case $m = 1$. In view of Theorems 1.6 and 1.8 it is sufficient to guarantee that the cone $\mathcal{K}^\star$ is pointed. This is equivalent to non-degeneracy of the linear operator Λ because of item 2. By contradiction, assume that there exists $c \neq 0$ such that $\Lambda(c) = 0$. Without loss of generality, we can think that $c_0 = 1$. Then in view of (2.7), we have

$$\Lambda(v(z)) = \phi(z)\psi(z)\psi(z)^* = \Lambda(v(z) - v_0(z)c).$$

Thus the system $\mathcal{S}^2$ is covered by the functional system $\tilde{v}(z) = v(z) - v_0(z)c$. However, $\tilde{v}_0(z) \equiv 0$, which contradicts our minimality assumption on $\mathcal{S}^2$. Therefore, $\mathcal{K}^\star$ is pointed and $\operatorname{int}\mathcal{K}$ is not empty. Finally, let us prove that $\operatorname{int}\mathcal{K}$ is not empty for an arbitrary m. Indeed, an interior point $P \in \operatorname{int}\mathcal{K}$ is obtained via any interior point p from the initial case $m = 1$ by setting $P = p \otimes I_m$. As a consequence, $\mathcal{K}^\star$ is pointed.

□

2.1.3 Consequences of semidefinite representability

Let us elaborate on the results obtained previously in this section. We have proved that cones admitting a sum-of-squares representation are semidefinite representable. In other words, they can be parametrized using a linear operator and a cone of positive semidefinite matrices.

Several consequences of these representations must be highlighted in the context of convex analysis and optimization. We focus on the generic case

$$\begin{aligned} \mathcal{K} &= \{P \in E : P = \Lambda^*(Y), Y = Y^* \succeq 0\}, \\ \mathcal{K}^\star &= \{S \in E : \Lambda(S) \succeq 0\}, \end{aligned}$$

with the assumptions that both cones are proper, see Theorems 2.1 and 2.3.

Our cone $\mathcal{K}$ can be equipped with the implicit self-concordant barrier function

$$F(P) = \min_Y \{-\log\det Y : P = \Lambda^*(Y), Y = Y^* \succeq 0\}$$

The barrier parameter ν is equal to the dimension of the square matrix Y. We can avoid this implicit definition by considering the matrix Y as a decision variable. In this case, the extended cone is

$$\mathcal{K} = \{(P, Y) : P = \Lambda^*(Y), Y = Y^* \succeq 0\}.$$

This cone is closed and the function $F(P,Y) = -\log\det Y$ is an appropriate self-concordant barrier.

The structural treatment of the dual cone $\mathcal{K}^\star$ is even simpler as a self-concordant barrier for $\mathcal{K}^\star$ is readily given by

$$F_*(S) = -\log\det\Lambda(S).$$

This function is well-defined on $\operatorname{int}\mathcal{K}^\star$, which is not empty. Its Hessian is non-degenerate at any feasible point, provided that $\upsilon(\cdot)$ is a minimal system, see Theorem 1.18. The barrier parameter ν is equal to the dimension of the square matrix $\Lambda(S)$.

Since $F_*(S)$ is a normal barrier, its gradient induces a one-to-one correspondence between the interiors of $\mathcal{K}$ and $\mathcal{K}^\star$. As a consequence, we get the following result.

Theorem 2.4. *Let the system $\upsilon(\cdot)$ be minimal for $\mathcal{K}$. Then a polynomial P belongs to the interior of $\mathcal{K}$ if and only if there exists $S \in \operatorname{int}\mathcal{K}^\star$ such that*

$$P = \Lambda^*(\Lambda(S)^{-1}).$$

This point S is uniquely defined.

Proof. Since $F_*(S) = -\log\det\Lambda(S)$, the directional derivative at S in the direction H is equal to

$$DF_*(S)[H] = -\langle\Lambda(S)^{-1}, \Lambda(H)\rangle$$

so that $F'_*(S) = -\Lambda^*(\Lambda(S)^{-1})$. The statement then follows from the gradient mapping property of $F_*(S)$, see Proposition 1.25. □

2.1.4 Weighted sums of squares

The above results can be extended so as to treat more complicated convex cones. Indeed, several primal cones can be added together with different weights and the resultant cone is composed of weighted sums of squares. Of course, real-valued and complex-valued functional systems, which are typically defined on different curves of the complex plane, cannot be mixed. Let us formalize this description at a high level of abstraction so that all situations are treated at the same time.

First we need to fix which class of systems (real-valued or complex-valued) is to be represented. Then given a set $\{\mathcal{S}_i\}_{i=1}^p$ of systems of linearly independent functions (belonging to the chosen class) and a set $\{\phi_i\}_{i=1}^p$ of weight functions, each pair $(\mathcal{S}_i, \phi_i)$ defines a cone $\mathcal{K}_i$ according to (2.1) or (2.5).

Let the components of the vector function $\upsilon(\cdot)$ be some basis of the finite dimensional functional subspace that covers $\mathcal{S}^2 = \cup_{i=1}^p \mathcal{S}_i^2$. For each pair $(\mathcal{S}_i, \phi_i)$, the linear operator Λ_i is defined using the triplet $(\phi_i, \psi_i, \upsilon)$ according to (2.3) or (2.7). As before, we obtain the semidefinite characterization of the sum of these cones.

Theorem 2.5. *With the notation introduced above, we have that:*

1. *Primal cone:*

$$\mathcal{K} = \mathcal{K}_1 + \cdots + \mathcal{K}_p = \{P \in E : P = \sum_{i=1}^p \Lambda_i^*(Y_i); Y_i \succeq 0, i = 1, \ldots, p\}. \qquad (2.9)$$

This cone is closed, convex and pointed.

2. *Dual cone:*

$$\mathcal{K}^\star = \mathcal{K}_1^\star \cap \ldots \cap \mathcal{K}_p^\star = \{S \in E : \Lambda_i(S) \succeq 0, i = 1, \ldots, p\}.$$

3. *If the identities* $\{\Lambda_i(C) = 0\}_{i=1}^p$ *imply that* $C = 0$*, then* $\mathcal{K}^\star$ *is a pointed cone and* $\mathcal{K}$ *has a non-empty interior.*

Proof. Let us focus on real-valued functional systems because the proof for complex-valued functional systems is almost identical (up to some obvious changes).

1. The cone $\mathcal{K}$ is a sum of convex cones so that it is convex. In order to prove that it is closed, we proceed as before. A sufficient condition which guarantee that $\mathcal{K}$ is closed is that $Y \doteq \operatorname{diag}(Y_1, \ldots, Y_p) = 0$ is the only positive semidefinite matrix such that $\Lambda^*(Y) \doteq \sum_{i=1}^p \Lambda_i^*(Y_i) = 0$ [126, Theorem 9.1]. Let us prove that the dual operator Λ^* satisfies this condition. For any $Y_1, \ldots, Y_p$ such that $Y \succeq 0$ and $\Lambda^*(Y) = 0$, we have

$$0 = \langle\langle \Lambda^*(Y), \Upsilon(x) \rangle\rangle_E = \sum_{i=1}^p \langle\langle \Lambda_i^*(Y_i), \Upsilon(x) \rangle\rangle_E, \quad \forall x \in \Gamma.$$

   As the matrices Y_i can be factorized as $Y_i = \sum_{j=0}^{N-1} Q_{i,j} Q_{i,j}^*$ with $Q_{i,j} \in \mathbb{C}^{mr_i \times m}$ for all i, j, the previous identity can be recast as

$$0 = \sum_{i=1}^p \phi_i(x) \sum_{j=0}^{N-1} \Psi_i(x)^T Q_{i,j} Q_{i,j}^T \Psi_i(x), \quad \forall x \in \Gamma.$$

   Because $\Psi_i(x)^T Q_{i,j} Q_{i,j}^T \Psi_i(x)$ is positive semidefinite for all i, j, we necessarily have that

$$0 = \langle Q_{i,j} Q_{i,j}^T, \Psi_i(x) \Psi_i(x)^T \rangle_E, \quad \forall x \in \Gamma \setminus \{x : \phi_i(z) = 0\}, i = 1, \ldots, p.$$

   The semi-infinite nature of this equality constraint and the linear independence of the components of $\psi_i(x)$ then yield that $Q_{i,j} = 0$ for all i, j. Therefore, $Y = 0$ and $\mathcal{K}$ is closed. By a completely similar argument, we prove that $\mathcal{K}$ is pointed. Indeed, if $P \in \mathcal{K} \cap -\mathcal{K}$, then $P(x) \succeq 0$ and $P(x) \preceq 0$ for all $x \in \Gamma$. As a consequence, $P(x) = 0$ for all $x \in \Gamma$. By making use of the parametrization $P = \Lambda^*(Y)$, we get

$$0 = P(x) = \langle\langle P, \Upsilon(z) \rangle\rangle_E = \langle\langle \Lambda^*(Y), \Upsilon(z) \rangle\rangle_E, \quad \forall x \in \Gamma.$$

   The above analysis can then carried out as before to show that $P = 0$.

2. Item 2 directly follows from Theorem 1.5.

3. Because of item 2, the dual cone contains a line if and only if there exists a non-zero vector C such that $\Lambda_i(C) = 0$ for all i. By assumption, this is not the case so that $\mathcal{K}^\star$ is a pointed cone and $\mathcal{K}$ has a non-empty interior.

□

If the cone $\mathcal{K}$ defined by (2.9) is proper, then the dual cone can be equipped with a non-degenerate self-concordant barrier

$$F_*(S) = -\sum_{i=1}^{p} \log\det \Lambda_i(S).$$

Using this barrier function, we directly obtain the counterpart of Theorem 2.4 by making use of the gradient mapping, see Proposition 1.25.

Theorem 2.6. *If the cone $\mathcal{K}$ defined by (2.9) is proper. Then a polynomial P belongs to the interior of $\mathcal{K}$ if and only if there exists $S \in \operatorname{int}\mathcal{K}^\star$ such that*

$$P = \sum_{i=1}^{p} \Lambda_i^*(\Lambda_i(S)^{-1}).$$

This point S is uniquely defined.

In this section, we have shown that cones generated by sums of squares can be parametrized by semidefinite matrices and linear operators. These original results are extremely valuable in the context of univariate non-negative polynomials. Indeed, sum-of-squares representations, which clearly guarantee that the non-negativity constraint is satisfied, always exist in this context. The next sections contain more details on this last statement. In particular, we obtain several convex parametrizations of cones of non-negative polynomials by merely applying the previous results to specific functional systems.

2.2 Self-adjoint polynomials

Self-adjoint matrix polynomials (of degree d) are matrix functions parametrized by $d+1$ Hermitian matrices $\{P_0, \ldots, P_d\}$. Although the exact parametrization depends on the functional basis, they are commonly defined using the standard monomial basis:

$$P(x) = \sum_{k=0}^{d} P_k x^k. \tag{2.10}$$

The space of coefficients is

$$E = \{P : P = \begin{bmatrix} P_0, & P_1, & \ldots, & P_d \end{bmatrix}; P_k \in \mathcal{H}^m\} \tag{2.11}$$

and the associated inner-products are

$$\langle \cdot, \cdot \rangle_{\mathbb{R}} : E \times E \to \mathbb{R} : X, Y \to \langle X, Y \rangle_{\mathbb{R}} = \operatorname{Re}\operatorname{Trace} \sum_{i=0}^{d} Y_i^* X_i,$$

$$\langle\langle \cdot, \cdot \rangle\rangle_{\mathbb{R}} : E \times E \to \mathbb{C}^{m\times m} : X, Y \to \langle\langle X, Y \rangle\rangle_{\mathbb{R}} = \sum_{i=0}^{d} Y_i^* X_i.$$

The vector of monomials $\pi_n(x)$ is defined by

$$\pi_n(x) \doteq \begin{bmatrix} 1, & x, & \ldots, & x^n \end{bmatrix}^T, \quad \Pi_n(x) \doteq \pi_n(x) \otimes I_m.$$

With these definitions, (2.10) can be recast as $P(x) = \langle\langle P, \Pi_d(x) \rangle\rangle_{\mathbb{R}}$ with $P \in E$.

2.2.1 Real line $\mathbb{R}$

As a direct consequence of the Fundamental Theorem of Algebra, a polynomial non-negative on the real axis is necessarily a sum of squares of polynomials. First, it must be of even degree, say $d = 2n$. Then the real roots $\{\lambda_k\}$ must be of even multiplicities and the complex roots must appear in conjugate pairs $\{\mu_k, \overline{\mu_k}\}$. Consequently, any polynomial

$$p(x) = \sum_{k=0}^{2n} p_k x^k$$

non-negative on the real line is a sum of two squares. More specifically, we have

$$p(x) = |q(x)|^2 = [q_1(x)]^2 + [q_2(x)]^2, \quad \forall x \in \mathbb{R} \tag{2.12}$$

where $q = q_1 + \jmath q_2$ and $q_1, q_2 \in \mathbb{R}^{n+1}$, see e.g. [121, Part 6, Problem 44]. Given a non-negative polynomial $p(x)$, there exist up to 2^n different polynomials $q(x)$ that satisfy (2.12).

Consider the cone of (self-adjoint) matrix polynomials (of degree $d = 2n$) that are non-negative on the real line

$$\mathcal{K}_{\mathbb{R}} \doteq \{P \in E : P(x) = \sum_{k=0}^{2n} P_k x^k \succeq 0, \forall x \in \mathbb{R}\},$$

with E as defined in (2.11). It is well known that any non-negative matrix polynomial on the real line also possesses a *spectral factorization*

$$P(x) = Q(x)^* Q(x), \quad \forall x \in \mathbb{R}, \tag{2.13}$$

with $Q(x) = \sum_{k=0}^{n} Q_k x^k, Q_k \in \mathbb{C}^{m\times m}$, see e.g. [128, Section 6.7].

Because all elements of $\mathcal{K}_{\mathbb{R}}$ can be rewritten as squares, see (2.13), the following characterization of $\mathcal{K}_{\mathbb{R}}$ is obtained by making use of Theorem 2.1.

Theorem 2.7. *Given $S \in E$, define the block linear operator H by*

$$[H(S)]_{ij} = S_{i+j}, \quad 0 \le i, j \le n.$$

Then

1. *Primal cone:*

$$\begin{aligned}\mathcal{K}_{\mathbb{R}} &= \{P \in E : P(x) = \sum_{k=0}^{2n} P_k x^k \succeq 0, \forall x \in \mathbb{R}; P_k = P_k^* \in \mathbb{C}^{m\times m}\} \\ &= \{P \in E : P = H^*(Y), Y \in \mathcal{H}_+^{m(n+1)}\}.\end{aligned}$$

2. *Dual cone:*

$$\mathcal{K}_{\mathbb{R}}^\star = \{S \in E : H(S) \succeq 0\}.$$

3. *Both cones are closed, convex and pointed. Consequently, they also have non-empty interiors.*

Proof. Because of spectral factorization, the cone $\mathcal{K}_{\mathbb{R}}$ admits a sum-of-squares representation, see (2.13). Therefore, we get a description of $\mathcal{K}_{\mathbb{R}}$ from Theorem 2.1. To see this, we set

$$\phi(x) = 1, \quad \psi(x) = \pi_n(x), \quad \upsilon(x) = \pi_{2n}(x), \quad \Psi(x) = \Pi_n(x), \quad \Upsilon(x) = \Pi_{2n}(x).$$

Note that υ is minimal. The structure of the linear operator Λ is obtained from the identity (2.2), which now reads

$$[\pi_n(x)\pi_n(x)^T]_{ij} = x^{i+j} = [\upsilon(x)]_{i+j}.$$

Consequently, the block linear operator Λ is equal to $H : E \to \mathcal{H}^{m(n+1)}$, with

$$[H(S)]_{ij} = S_{i+j}, \quad 0 \leq i, j \leq n.$$

We complete the proof by making use of Theorem 2.1. □

The linear operator $H : E \to \mathcal{H}^{m(n+1)}$ is associated with a Hermitian block Hankel matrix. Indeed, by definition,

$$H(S) = \begin{bmatrix} S_0 & S_1 & \dots & S_n \\ S_1 & \iddots & \iddots & \vdots \\ \vdots & \iddots & \iddots & S_{2n-1} \\ S_n & \dots & S_{2n-1} & S_{2n} \end{bmatrix}. \tag{2.14}$$

The dual operator $H^* : \mathbb{C}^{m(n+1)\times m(n+1)} \to E$ is defined as

$$P = H^*(Y) \iff P_k = \sum_{i+j=k} Y_{ij}, \quad k = 0, \dots, 2n. \tag{2.15}$$

Remember that Y_{ij} is the block (i, j) of dimension m of the matrix Y. By convention, the blocks Y_{ij} are equal to 0 for i and j outside their definition range.

In view of Theorem 2.4, we have the following dual representation of $P \in \operatorname{int} \mathcal{K}_{\mathbb{R}}$:

$$P(x) = \langle\langle H(S)^{-1}\Pi_n(x), \Pi_n(x)\rangle\rangle, \tag{2.16}$$

where the point $S \in \operatorname{int} \mathcal{K}_{\mathbb{R}}^{\star}$ is uniquely defined.

Example 2.5. Consider a quadratic polynomial $p(x) = ax^2 + bx + c$. It is non-negative on the real line if and only if $a \geq 0$ and $b^2 - 4ac \leq 0$. The matrix

$$Y = \begin{bmatrix} c & b/2 \\ b/2 & a \end{bmatrix}$$

is such that $p(x) = \langle Y\pi_1(x), \pi_1(x)\rangle$, i.e., $p = H^*(Y)$. Note that positive semidefiniteness of Y is equivalent to the standard non-negativity conditions for $p(x)$. For $p \in \operatorname{int} \mathcal{K}_{\mathbb{R}}$, we have that

$$Y^{-1} = \frac{\begin{bmatrix} 4a & -2b \\ -2b & 4c \end{bmatrix}}{4ac - b^2}$$

is a positive definite Hankel matrix. Our matrix Y thus provides us with the dual representation of $p(x)$!

Note that the identity

$$P(x) = \langle\langle Y\Pi_n(x), \Pi_n(x)\rangle\rangle = \Pi_n(x)^T Y \Pi_n(x), \forall x \in \mathbb{R}$$

is algebraically equivalent to $P_k = \sum_{i+j=k} Y_{ij}, k = 0, \ldots, 2n$. A simple choice for Y so as to obtain this identity is found to be

$$Y_0 = \begin{bmatrix} P_0 & \frac{1}{2}P_1 & & \\ \frac{1}{2}P_1 & P_2 & \ddots & \\ & \ddots & \ddots & \frac{1}{2}P_{2n-1} \\ & & \frac{1}{2}P_{2n-1} & P_{2n} \end{bmatrix}. \tag{2.17}$$

However, this matrix is not unique; any matrix in the kernel of H^* can be added to Y_0. The following theorem formally describes the whole class of equivalence.

Theorem 2.8. *Let*

$$Z \doteq \begin{bmatrix} 0 & & & \\ I_m & 0 & & \\ & \ddots & \ddots & \\ & & I_m & 0 \end{bmatrix}$$

be the block shift matrix. A Hermitian matrix Y satisfies (2.15) *if and only if it can be expressed as*

$$Y = Y_0 + ZX - XZ^T, \tag{2.18}$$

where the skew-Hermitian matrix X has the form

$$X \doteq \left[\begin{array}{c|c} X_0 & 0 \\ \hline 0 & 0 \end{array}\right], \quad X_0 = -X_0^* \in \mathbb{C}^{mn\times mn}. \tag{2.19}$$

Proof. The *if* part is obvious since one has

$$\Pi_n(x)^T(ZX - XZ^T)\Pi_n(x) = (x - x)\Pi_{n-1}(x)^T X_0 \Pi_{n-1}(x) = 0$$

for any matrix X of the form (2.19). Conversely, let Y be a solution of $P(x) = \Pi_n(x)^T Y \Pi_n(x)$ and let us set X as

$$X = \sum_{k=0}^{n} (Z^{k+1})^T (Y - Y_0)(Z^k)^T. \tag{2.20}$$

It turns out that X has the structure (2.19) with $X = -X^*$ and satisfies (2.18). To see this, observe first that X has the structure (2.19) as an immediate consequence of relations (2.15). Next, inserting (2.20) in (2.18), one obtains successively

$$\begin{aligned} & Y_0 + ZX - XZ^T \\ &= Y_0 + ZZ^T \sum_{k=0}^{n} (Z^k)^T (Y - Y_0)(Z^k)^T - \sum_{k=0}^{n} (Z^{k+1})^T (Y - Y_0)(Z^{k+1})^T \\ &= Y_0 + ZZ^T (Y - Y_0) + (ZZ^T - I_{(n+1)m}) \sum_{k=0}^{n-1} (Z^{k+1})^T (Y - Y_0)(Z^{k+1})^T \\ &= Y_0 + (Y - Y_0) = Y \end{aligned}$$

again in view of relations (2.15). Finally, one establishes the skew-Hermitian property of X from the fact that $ZX - XZ^T = X^*Z^T - ZX^*$ necessarily implies $X = -X^*$ for any matrix X of algebraic structure (2.19). □

2.2.2 Interval $[a, b]$

Polynomials that are non-negative on an interval $[a, b]$ also possess sum-of-squares representations. This important result is usually derived from the characterization of non-negative trigonometric polynomials, see e.g. [121, Part 6, Problem 47].

Theorem 2.9 (Markov-Lukács, [98]). *Let $p(x) \in \mathbb{R}[x]$ be a real polynomial of degree d and $[a, b]$ be a segment. Then $p(x)$ is non-negative on $[a, b]$ if and only if*

even degree, $d = 2n$

$$p(x) = (q_1(x))^2 + (x - a)(b - x)(q_2(x))^2, \quad \forall x \in \mathbb{R},$$

where $q_1(x)$ and $q_2(x)$ are real polynomials of degree n and $n - 1$, respectively.

odd degree, $d = 2n + 1$

$$p(x) = (x - a)(q_3(x))^2 + (b - x)(q_4(x))^2, \quad \forall x \in \mathbb{R},$$

where $q_3(x)$ and $q_4(x)$ are real polynomials of degree n.

This result has also been extended to the cone of (self-adjoint) matrix polynomials that are non-negative on an interval:

$$\mathcal{K}_{[a,b]} \doteq \{P \in E : P(x) = \sum_{k=0}^{d} P_k x^k \succeq 0, \forall x \in [a, b]\},$$

with E as defined in (2.11).

Theorem 2.10 ([46, Theorem 2.5]). *A self-adjoint matrix polynomial P (of degree d) belongs to $\mathcal{K}_{[a,b]}$ if and only if, for all $x \in \mathbb{R}$,*

$$P(x) = \begin{cases} Q_1(x)^*Q_1(x) + (x - a)(b - x)Q_2(x)^*Q_2(x), & \text{even degree, } d=2n \\ (x - a)Q_3(x)^*Q_3(x) + (b - x)Q_4(x)^*Q_4(x), & \text{odd degree, } d=2n+1 \end{cases}$$

where Q_1, Q_2, Q_3 and Q_4 are matrix polynomials of degree n, $n - 1$, n and n, respectively.

Using these representations, Theorems 2.1 and 2.5 lead to the following characterizations of the primal and dual cones.

Even degree, $d = 2n$

Theorem 2.11. *Given $S \in E$, define the block linear operators H_1 and H_2 by*

$$[H_1(S)]_{ij} = S_{i+j}, \quad 0 \le i, j \le n,$$
$$[H_2(S)]_{ij} = -S_{i+j+2} + (b + a)S_{i+j+1} - abS_{i+j}, \quad 0 \le i, j \le n - 1.$$

Then

1. *Primal cone:*

$$\begin{aligned}\mathcal{K}_{[a,b]} &= \{P \in E : P(x) = \sum_{k=0}^{2n} P_k x^k \succeq 0, \forall x \in [a,b]\} \\ &= \{P \in E : P = H_1^*(Y_1) + H_2^*(Y_2), Y_1 \in \mathcal{H}_+^{m(n+1)}, Y_2 \in \mathcal{H}_+^{mn}\}.\end{aligned}$$

2. *Dual cone:*

$$\mathcal{K}_{[a,b]}^\star = \{S \in E : H_1(S) \succeq 0, H_2(S) \succeq 0\}.$$

3. *Both cones are closed, convex and pointed. Consequently, they also have non-empty interiors.*

Proof. Theorem 2.10 implies that $\mathcal{K}_{[a,b]}$ is a sum of two sub-cones, which admit weighted sums of squares representations. Therefore, the statement follows from applying Theorem 2.5 to the appropriate sub-cones. Let us be more specific. If the notation of Section 2.1 is used, these two sub-cones are characterized by the functional elements

$$\phi(x) = 1, \quad \psi(x) = \pi_n(x), \quad \Psi(x) = \Pi_n(x),$$

and

$$\phi(x) = (x-a)(b-x), \quad \psi(x) = \pi_{n-1}(x), \quad \Psi(x) = \Pi_{n-1}(x),$$

respectively. We use the upper-level functional system

$$\upsilon(x) = \pi_{2n}(x), \quad \Upsilon(x) = \Pi_{2n}(x),$$

which is minimal. Then the structure of the associated linear operators H_1 and H_2 are obtained from the identities

$$\begin{aligned}[\pi_n(x)\pi_n(x)^T]_{ij} &= x^{i+j}, \\ [(x-a)(b-x)\pi_{n-1}(x)\pi_{n-1}(x)^T]_{ij} &= -x^{i+j+2} + (b+a)x^{i+j+1} - abx^{i+j},\end{aligned}$$

which are easy to check. Note that the relations $H_1(S) = 0$ and $H_2(S) = 0$ imply that $S = 0$. To see this, consider the first block column and the last block row of $H_1(S)$. The statement then follows from Theorems 2.1 and 2.5: the first one is used to characterize both sub-cones of interest and the second one leads to the characterization of their sum, which is equal to $\mathcal{K}_{[a,b]}$. □

Note that both linear operators $H_1 : E \to \mathcal{H}^{m(n+1)}$ and $H_2 : E \to \mathcal{H}^{mn}$ are associated with Hermitian block Hankel matrices. Moreover, we have $H_1(S) = H(S)$, see (2.14).

Theorem 2.4 yields the dual representation of $P \in \operatorname{int} \mathcal{K}_{[a,b]}$

$$P(x) = \langle\langle H_1(S)^{-1}\Pi_n(x), \Pi_n(x)\rangle\rangle + (x-a)(b-x)\langle\langle H_2(S)^{-1}\Pi_{n-1}(x), \Pi_{n-1}(x)\rangle\rangle,$$

where the point $S \in \operatorname{int} \mathcal{K}_{[a,b]}^\star$ is uniquely defined.

Odd degree, $d = 2n+1$

Theorem 2.12. *Given $S \in E$, define the block linear operators H_3 and H_4 by*

$$[H_3(S)]_{ij} = S_{i+j+1} - aS_{i+j}, \quad 0 \leq i, j \leq n,$$
$$[H_4(S)]_{ij} = bS_{i+j} - S_{i+j+1}, \quad 0 \leq i, j \leq n.$$

Then

1. *Primal cone:*

$$\begin{aligned}\mathcal{K}_{[a,b]} &= \{P \in E : P(x) = \sum_{k=0}^{2n+1} P_k x^k \succeq 0, \forall x \in [a,b]\} \\ &= \{P \in E : P = H_3^*(Y_3) + H_4^*(Y_4), Y_3 \in \mathcal{H}_+^{m(n+1)}, Y_4 \in \mathcal{H}_+^{m(n+1)}\}.\end{aligned}$$

2. *Dual cone:*

$$\mathcal{K}_{[a,b]}^\star = \{S \in E : H_3(S) \succeq 0, H_4(S) \succeq 0\}.$$

3. *Both cones are closed, convex and pointed. Consequently, they also have non-empty interiors.*

Proof. Theorem 2.10 implies that $\mathcal{K}_{[a,b]}$ is a sum of two sub-cones, which admit sum-of-squares representations. Therefore, the statement follows from applying Theorem 2.5 to the appropriate sub-cones. The proof is almost identical to that of Theorem 2.11. The main differences are the functional systems which describe our two sub-cones. If the notation of Section 2.1 is used, these two sub-cones are characterized by the functional elements

$$\phi(x) = (x-a), \quad \psi(x) = \pi_n(x), \quad \Psi(x) = \Pi_n(x),$$

and

$$\phi(x) = (b-x), \quad \psi(x) = \pi_n(x), \quad \Psi(x) = \Pi_n(x),$$

respectively. We use the upper-level functional system

$$\upsilon(x) = \pi_{2n+1}(x), \quad \Upsilon(x) = \Pi_{2n+1}(x),$$

which is minimal. As before, the structure of the associated linear operators H_3 and H_4 are obtained from the identities

$$[(x-a)\pi_n(x)\pi_n(x)^T]_{ij} = x^{i+j+1} - ax^{i+j},$$
$$[(b-x)\pi_n(x)\pi_n(x)^T]_{ij} = bx^{i+j} - x^{i+j+1},$$

which are easy to check. The relations $H_3(S) = 0$ and $H_4(S) = 0$ are equivalent to

$$S_{i+j+1} - aS_{i+j} = 0, \quad bS_{i+j} - S_{i+j+1} = 0, \quad \forall i, j,$$

so that $(b-a)S_{i+j} = 0$ for all i and j. Because $a < b$, we obtain that $S = 0$. The statement then follows from Theorems 2.1 and 2.5: the first one is used to characterize both sub-cones of interest and the second one leads to the characterization of their sum, which is equal to $\mathcal{K}_{[a,b]}$. □

Note that both linear operators $H_3 : E \to \mathcal{H}^{m(n+1)}$ and $H_4 : E \to \mathcal{H}^{m(n+1)}$ are associated with Hermitian block Hankel matrices.

Theorem 2.4 yields the dual representation of $P \in \operatorname{int} \mathcal{K}_{[a,b]}$

$$P(x) = (x-a)\langle\langle H_3(S)^{-1}\Pi_n(x), \Pi_n(x)\rangle\rangle + (b-x)\langle\langle H_4(S)^{-1}\Pi_n(x), \Pi_n(x)\rangle\rangle,$$

where the point $S \in \operatorname{int} \mathcal{K}^{\star}_{[a,b]}$ is uniquely defined.

2.2.3 Semi-infinite interval $[0, +\infty)$

An interesting representation of real polynomials non-negative on the semi-infinite interval $[0, +\infty)$ is somehow similar to the Markov-Lukács Theorem. Define the following convex cone

$$\mathcal{K}_{[0,+\infty)} = \{p \in \mathbb{R}^{n+1} : p(x) = \sum_{k=0}^{n} p_k x^k \geq 0, \forall x \in [0, +\infty)\}$$

and set $n_1 = \lfloor n/2 \rfloor$ and $n_2 = \lfloor (n-1)/2 \rfloor$. Then $p(x)$ belongs to $\mathcal{K}_{[0,+\infty)}$ if and only if

$$p(x) = p_1(x) + x p_2(x), \quad \forall x \in \mathbb{R},$$

where $p_1(x)$ and $p_2(x)$ are non-negative polynomials of degree $2n_1$ and $2n_2$, respectively. Note that this characterization can be obtained without relying on trigonometric polynomials, see [121, Part 6, Problem 45].

This result can be extended to the cone of (self-adjoint) matrix polynomials (of degree n) that are non-negative on the semi-infinite interval $[0, +\infty)$:

$$\mathcal{K}_{[0,+\infty)} \doteq \{P \in E : P(x) = \sum_{k=0}^{n} P_k x^k \succeq 0, \forall x \in [0, +\infty)\},$$

with E as defined in (2.11).

Theorem 2.13 ([46, Theorem 5.1]). *A self-adjoint matrix polynomial $P(x)$ (of degree n) belongs to $\mathcal{K}_{[0,+\infty)}$ if and only if*

$$P(x) = P_1(x) + x P_2(x), \quad \forall x \in \mathbb{R}, \tag{2.21}$$

where $P_1(x)$ and $P_2(x)$ are non-negative matrix polynomials of degree $2n_1$ and $2n_2$, respectively.

Of course, Theorems 2.1 and 2.5 then lead to the appropriate characterization of the primal and dual cones.

Theorem 2.14. *Given $S \in E$, define the block linear operators H_5 and H_6 by*

$$[H_5(S)]_{ij} = S_{i+j}, \quad 0 \leq i, j \leq n_1,$$
$$[H_6(S)]_{ij} = S_{i+j+1}, \quad 0 \leq i, j \leq n_2.$$

Then

1. *Primal cone:*

$$\begin{aligned}\mathcal{K}_{[0,+\infty)} &= \{P \in E : P(x) = \sum_{k=0}^{n} P_k x^k \succeq 0, \forall x \in [0,+\infty)\} \\ &= \{P \in E : P = H_5^*(Y_5) + H_6^*(Y_6), Y_5 \in \mathcal{H}_+^{mn_1}, Y_6 \in \mathcal{H}_+^{mn_2}\}.\end{aligned}$$

2. *Dual cone:*

$$\mathcal{K}_{[0,+\infty)}^\star = \{S \in E : H_5(S) \succeq 0, H_6(S) \succeq 0\}.$$

3. *Both cones are closed, convex and pointed. Consequently, they also have non-empty interiors.*

Proof. Theorem 2.13 implies that $\mathcal{K}_{[0,+\infty)}$ is a sum of two sub-cones, which admit sum-of-squares representations. Consequently the statement follows from applying Theorem 2.5 to the appropriate sub-cones. If the notation of Section 2.1 is used, these two sub-cones are characterized by the functional elements

$$\phi(x) = 1, \quad \psi(x) = \pi_{n_1}(x), \quad \Psi(x) = \Pi_{n_1}(x),$$

and

$$\phi(x) = x, \quad \psi(x) = \pi_{n_2}(x), \quad \Psi(x) = \Pi_{n_2}(x),$$

respectively. We use the upper-level functional system

$$\upsilon(x) = \pi_n(x), \quad \Upsilon(x) = \Pi_n(x),$$

which is necessarily minimal for one of our sub-cones. As before, the structure of the associated linear operators H_5 and H_6 are obtained from the identities

$$\begin{aligned}[\pi_{n_1}(x)\pi_{n_1}(x)^T]_{ij} &= x^{i+j}, \quad 0 \le i,j \le n_1, \\ [x\pi_{n_2}(x)\pi_{n_2}(x)^T]_{ij} &= x^{i+j+1}, \quad 0 \le i,j \le n_2,\end{aligned}$$

which are easy to check. Let us check that $H_5(S) = 0$ and $H_6(S) = 0$ imply $S = 0$. The first conditions are equivalent to $S_i = 0$ for $i = 0, \ldots, 2n_1$. The second conditions are equivalent to $S_i = 0$ for $i = 1, \ldots, 2n_2+1$. By definition, we have $n_1+n_2 = n-1$ so that we conclude that $S_i = 0$ for $i = 0, \ldots, n$. The statement then follows from Theorems 2.1 and 2.5: the first one is used to characterize both sub-cones of interest and the second one leads to the characterization of their sum, which is equal to $\mathcal{K}_{[0,+\infty)}$. □

As before, Theorem 2.4 yields the dual representation of $P \in \operatorname{int} \mathcal{K}_{[0,+\infty)}$:

$$P(x) = \langle\langle H_5(S)^{-1}\Pi_{n_1}(x), \Pi_{n_1}(x)\rangle\rangle + x\langle\langle H_6(S)^{-1}\Pi_{n_2}(x), \Pi_{n_2}(x)\rangle\rangle,$$

where the point $S \in \operatorname{int} \mathcal{K}_{[0,+\infty)}^\star$ is uniquely defined.

2.3 Self-adjoint trigonometric polynomials

Self-adjoint trigonometric matrix polynomials (of degree n) are matrix functions parametrized by $2n+1$ Hermitian matrices $\{A_0, \ldots, A_n, B_1, \ldots, B_n\}$. They are usually defined by

$$P(\theta) = \sum_{k=0}^{n} A_k \cos(k\theta) + \sum_{k=1}^{n} B_k \sin(k\theta). \tag{2.22}$$

By convention, we set $B_0 = 0$. The above definition does not emphasize the polynomial nature of $P(\theta)$. Setting $z = \cos\theta + \jmath \sin\theta$ in (2.22), the equivalent mathematical object is

$$P(z) = \frac{1}{2}(\sum_{k=0}^{n}(A_k + \jmath B_k)z^{-k} + \sum_{k=0}^{n}(A_k - \jmath B_k)z^{k}). \tag{2.23}$$

Our trigonometric polynomial $P(\theta)$ can thus be treated as a Laurent polynomial $P(z)$, which is specified on the unit circle. Consequently, $P(z)$ is defined on the whole complex plane.

The space of coefficients is

$$E = \{P = \begin{bmatrix} P_0, & \dots, & P_n \end{bmatrix} : P_0 = P_0^*, P_k \in \mathbb{C}^{m\times m}, \forall k\} \tag{2.24}$$

and the associated inner products are

$$\langle \cdot, \cdot \rangle_{\mathbb{R}} : E \times E \to \mathbb{R} : X, Y \to \langle X, Y \rangle_{\mathbb{R}} = \operatorname{Re}\operatorname{Trace}(\frac{1}{2}\sum_{i=0}^{n}(Y_i^* X_i + Y_i X_i^*)),$$

$$\langle\langle \cdot, \cdot \rangle\rangle_{\mathbb{R}} : E \times E \to \mathbb{C}^{m\times m} : X, Y \to \langle\langle X, Y \rangle\rangle_{\mathbb{R}} = \frac{1}{2}\sum_{i=0}^{n}(Y_i^* X_i + Y_i X_i^*).$$

As before, the vector of monomials $\pi_n(z)$ is defined as

$$\pi_n(z) \doteq \begin{bmatrix} 1, & z, & \dots & z^n \end{bmatrix}^T, \quad \Pi_n(z) \doteq \pi_n(z) \otimes I_m.$$

With these definitions, (2.23) can be rewritten as $P(z) = \langle\langle P, \Pi_n(z) \rangle\rangle_{\mathbb{R}}$ where $P_k = A_k + \jmath B_k, k = 0, \dots, n$.

Remark. The matrix coefficients $\{P_k\}_{k=0}^n$ are all Hermitian if and only if the matrices $\{B_k\}_{k=1}^n$ are all equal to the zero matrix. In this particular case, the original polynomial $P(\theta)$ is exactly a (trigonometric) cosine polynomial. Indeed, the original coefficients of $P(\theta)$ can be recovered by setting

$$A_k = \frac{P_k + P_k^*}{2}, \quad k = 0, \dots, n, \tag{2.25}$$

$$B_k = \frac{P_k - P_k^*}{2\jmath}, \quad k = 1, \dots, n. \tag{2.26}$$

Remark. In the scalar case($m = 1$), we have $E = \mathbb{R} \times \mathbb{C}^{n+1}$. Moreover, the inner-products $\langle \cdot, \cdot \rangle_{\mathbb{R}}$ and $\langle\langle \cdot, \cdot \rangle\rangle_{\mathbb{R}}$ are equivalent to the one of Example 1.4. This motivates our common notation.

2.3.1 Unit circle $\mathbb{T}$

In the scalar case ($m = 1$), the Fundamental Theorem of Algebra provides us with a sum-of-squares decomposition of any non-negative trigonometric polynomial. The associated factorization result dates back to the beginning of the *20*th century and is known as the Fejér-Riesz Theorem, see e.g. [121, Part 6, Problems 40 and 41].

Theorem 2.15 (Fejér-Riesz, [51, §1]). *A trigonometric polynomial $p(z) = \langle p, \pi_n(z) \rangle_{\mathbb{R}}$ (of degree n) is non-negative on the unit circle , i.e., $p(z) \geq 0, \forall z \in \mathbb{T}$, if and only if there exists a complex polynomial $q(z) = \sum_{k=0}^{n} q_k z^k$ such that $p(z) = |q(z)|^2, \forall z \in \mathbb{T}$. Moreover, $p \in \mathbb{R}^{n+1}$ if and only if $q \in \mathbb{R}^{n+1}$.*

Consider the cone of trigonometric matrix polynomials (of degree n) that are non-negative on the unit circle

$$\mathcal{K}_{\mathbb{C}} \doteq \{P \in E : P(z) = \frac{1}{2}\sum_{k=0}^{n}(P_k z^{-k} + P_k^* z^k) \succeq 0, \forall z \in \mathbb{T}\}.$$

It is well known that the Fejér-Riesz Theorem can be extended to non-negative matrix trigonometric polynomials. The *spectral factorization* of $P \in \mathcal{K}_{\mathbb{C}}$ is given by

$$P(z) = Q(z)^* Q(z), \quad \forall z \in \mathbb{T}, \tag{2.27}$$

with $Q(z) = \sum_{k=0}^{n} Q_k z^k$, see e.g. [128, Section 6.6]

From this result, we see that any element of $\mathcal{K}_{\mathbb{C}}$ can be written as a square. Theorem 2.3 then produces the following characterization result.

Theorem 2.16. *Given $S \in E$, define the block linear operator T by*

$$[T(S)]_{ij} = S_{i-j}, \quad 0 \leq i, j \leq n,$$

with $S_{-k} = S_k^*$. Then

1. *Primal cone:*

$$\begin{aligned}\mathcal{K}_{\mathbb{C}} &= \{P \in E : P(x) = \frac{1}{2}\sum_{k=0}^{n}(P_k z^{-k} + P_k^* z^k) \succeq 0, \forall z \in \mathbb{T}\} \\ &= \{P \in E : P = T^*(Y), Y \in \mathcal{H}_+^{m(n+1)}\}.\end{aligned}$$

2. *Dual cone:*

$$\mathcal{K}_{\mathbb{C}}^\star = \{S \in E : T(S) \succeq 0\}.$$

3. *Both cones are closed, convex and pointed. Consequently, they also have non-empty interiors.*

Proof. Due to spectral factorization, the cone $\mathcal{K}_{\mathbb{C}}$ admits a sum-of-squares representation, see (2.27). Therefore, we can get a description of $\mathcal{K}_{\mathbb{C}}$ from Theorem 2.3. To see this, we set

$$\phi(z) = 1, \quad \psi(z) = \pi_n(z), \quad \upsilon(x) = \pi_n(z), \quad \Psi(z) = \Pi_n(z), \quad \Upsilon(x) = \Pi_n(z).$$

The structure of the linear operator Λ is obtained from the identity (2.6), which now reads

$$[\pi_n(z)\pi_n(z)^*]_{ij} = z^{i-j} = [\upsilon(z)]_{i-j}.$$

with $[\upsilon(z)]_{-k} = [\upsilon(z)]_k^*, |z| = 1$. Consequently, the block linear operator Λ is equal to $T : E \to \mathcal{H}^{m(n+1)}$ with

$$[T(S)]_{ij} = S_{i-j}, \quad 0 \leq i, j \leq n,$$

and $S_{-k} = S_k^*$. We complete the proof by making use of Theorem 2.3. □

The linear operator $T : E \to \mathbb{C}^{m(n+1)\times m(n+1)}$ is associated with a Hermitian block Toeplitz matrix. Indeed, by definition,

$$T(S) = \begin{bmatrix} S_0 & S_1^* & \dots & S_n^* \\ S_1 & S_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & S_1^* \\ S_n & \dots & S_1 & S_0 \end{bmatrix}. \tag{2.28}$$

The dual operator $T^* : \mathbb{C}^{m(n+1)\times m(n+1)} \to E$ is defined as

$$P = T^*(Y) \iff P_k = \begin{cases} \sum_{i-j=0} Y_{ij}, & k = 0 \\ 2\sum_{i-j=k} Y_{ij}, & k = 1, \dots, n \end{cases} \tag{2.29}$$

where Y_{ij} is the block (i, j) of dimension m of the matrix Y. Note that $Y_{ij} = 0$ for i and j outside their definition range.

In view of Theorem 2.4, we have the following dual representation for $P \in \operatorname{int} \mathcal{K}_{\mathbb{C}}$:

$$P(z) = \langle\langle T(S)^{-1}\Pi_n(z), \Pi_n(z)\rangle\rangle, \tag{2.30}$$

where the point $S \in \operatorname{int} \mathcal{K}_{\mathbb{C}}^\star$ is uniquely defined.

Example 2.6. Consider a trigonometric polynomial $p(\theta) = a + b\cos\theta + c\sin\theta$. As its minima satisfy $\tan\hat{\theta} = c/b$, it is non-negative if and only if $a \geq \sqrt{b^2 + c^2}$. Indeed, the minimal value $p(\hat{\theta})$ is equal to $a - \sqrt{b^2 + c^2}$. Since

$$p(z) = \operatorname{Re}(a + (b + \jmath c)z^{-1}) = \frac{b - \jmath c}{2} z + a + \frac{b + \jmath c}{2} z^{-1}$$

with $|z| = 1$, the matrix

$$Y = \begin{bmatrix} a/2 & (b - \jmath c)/2 \\ (b + \jmath c)/2 & a/2 \end{bmatrix}$$

is such that $p(z) = \langle Y\pi_1(z), \pi_1(z)\rangle$. Note that non-negativity of $p(z)$ is equivalent to positive semidefiniteness of Y. For $p \in \operatorname{int} \mathcal{K}_{\mathbb{C}}$, one easily checks that

$$Y^{-1} = \frac{\begin{bmatrix} 2a & -2(b - \jmath c) \\ -2(b + \jmath c) & 2a \end{bmatrix}}{a^2 - b^2 - c^2}$$

is Toeplitz. Our matrix Y thus provides us with the dual representation of $p(z)$!

The relation $P(z) = \langle\langle Y\Pi_n(z), \Pi_n(z)\rangle\rangle = \Pi_n(z)^* Y \Pi_n(x), |z| = 1$ is algebraically equivalent to (2.29). Clearly, the choice

$$Y_0 = \begin{bmatrix} P_0 & P_1^*/2 & \dots & P_n^*/2 \\ P_1/2 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ P_n/2 & 0 & \dots & 0 \end{bmatrix} \tag{2.31}$$

is an admissible matrix Y. However, this matrix is not unique: any matrix in the kernel of T^* can be added to Y_0. The following theorem formally describes the whole class of equivalence.

Theorem 2.17. *Let*

$$Z \doteq \begin{bmatrix} 0 & & & \\ I_m & 0 & & \\ & \ddots & \ddots & \\ & & I_m & 0 \end{bmatrix}$$

be the block shift matrix. A Hermitian matrix Y satisfies equation (2.29) if and only if it can be expressed as

$$Y = Y_0 + X - ZXZ^T, \tag{2.32}$$

where the Hermitian matrix X has the form

$$X \doteq \left[\begin{array}{c|c} X_0 & 0 \\ \hline 0 & 0 \end{array}\right], \quad X_0 = X_0^* \in \mathbb{C}^{mn \times mn}. \tag{2.33}$$

Proof. The *if* part is obvious since one has

$$\Pi_n(z)^*(X - ZXZ^T)\Pi_n(z) = (1 - |z|^2)\Pi_{n-1}(z)^* X_0 \Pi_{n-1}(z) = 0$$

for any matrix X of the form (2.33) and $|z| = 1$. Conversely, let Y be a solution of $P(z) = \Pi_n(z)^* Y \Pi_n(z)$ and let us set X as

$$X = \sum_{k=0}^{n} Z^k (Y - Y_0)(Z^k)^T. \tag{2.34}$$

It turns out that X has the structure (2.33) with $X = X^*$ and satisfies (2.32). To see this, observe first that X has the structure (2.33) as an immediate consequence of relations (2.29). Next, inserting (2.34) in (2.32), one obtains successively

$$\begin{aligned} Y_0 + X - ZXZ^T &= Y_0 + \sum_{k=0}^{n} Z^k (Y - Y_0)(Z^k)^T - \sum_{k=1}^{n+1} Z^k (Y - Y_0)(Z^k)^T \\ &= Y_0 + (Y - Y_0) = Y. \end{aligned}$$

Finally, one establishes the Hermitian property of X from the fact that $X - ZXZ^T = X^* - ZX^*Z^T$ necessarily implies $X = X^*$ for any matrix X of algebraic structure (2.33). □

2.3.2 Arc $[-\omega, +\omega]$ of the unit circle

Let $[-\omega, +\omega] \subset [-\pi, +\pi]$ be an symmetric interval. Trigonometric polynomials that are non-negative on $[-\omega, +\omega]$ also possess representations as sums of squares. This result is similar to the Markov-Lukács Theorem (concerning polynomials of even degree).

Consider the cone of trigonometric matrix polynomials (of degree n) that are non-negative on the interval $[-\omega, +\omega]$

$$\mathcal{K}_{[-\omega,+\omega]} \doteq \{P \in E : P(z) = \frac{1}{2}\sum_{k=0}^{n}(P_k z^{-k} + P_k^* z^k) \succeq 0, z = e^{\jmath\theta}, \forall \theta \in [-\omega, +\omega]\}.$$

Theorem 2.18. *A self-adjoint trigonometric matrix polynomial $P(z)$ (of degree n) belongs to $\mathcal{K}_{[-\omega,+\omega]}$ if and only if*

$$P(z) = P_1(z) + (z + z^{-1} - 2\cos\omega)P_2(z), \quad \forall z \in \mathbb{T},$$

where $P_1(z)$ and $P_2(z)$ are non-negative trigonometric polynomials of degree n and $n-1$, respectively.

Proof. To prove this well-known result, we use the conformable transformations

$$z = \frac{1+\jmath x}{1-\jmath x} \in \mathbb{T} \quad \Longleftrightarrow \quad x = \jmath\frac{1-z}{1+z} \in \mathbb{R}$$

and Theorem 2.10. Indeed, these transformations are one-to-one correspondence between the arc $e^{\jmath[-\omega,+\omega]} \subset \mathbb{T}$ and the segment $[-a,a] \doteq [-\jmath\frac{1-e^{\jmath\omega}}{1+e^{\jmath\omega}}, \jmath\frac{1-e^{\jmath\omega}}{1+e^{\jmath\omega}}] \subset \mathbb{R}$. Therefore, $P(e^{\jmath\theta})$ is non-negative on $[-\omega,+\omega]$ if and only if the polynomial $P(\frac{1+\jmath x}{1-\jmath x})$ is non-negative on $[-a,a]$. Elementary algebraic manipulations show that $P(\frac{1+\jmath x}{1-\jmath x}) = \frac{\tilde{P}(x)}{(1+x^2)^n}$ for some self-adjoint polynomial $\tilde{P}$ of degree $2n$. Consequently, Theorem 2.10 allows us to write the equivalences

$$\begin{aligned}
& P(e^{\jmath\theta}) \geq 0, \forall\theta \in [-\omega,\omega] \\
\Longleftrightarrow\ & P(\frac{1+\jmath x}{1-\jmath x}) = \frac{\tilde{P}(x)}{(1+x^2)^n} \geq 0, \forall x \in [-a,a] \\
\Longleftrightarrow\ & P(\frac{1+\jmath x}{1-\jmath x}) = \frac{\tilde{P}_1(x) + (a-x)(x+a)\tilde{P}_2(x)}{(1+x^2)^n} \geq 0, \forall x \in [-a,a] \\
\Longleftrightarrow\ & P(\frac{1+\jmath x}{1-\jmath x}) = \frac{\tilde{P}_1(x)}{(1+x^2)^n} + \frac{(a-x)(x+a)}{1+x^2}\frac{\tilde{P}_2(x)}{(1+x^2)^{n-1}} \geq 0, \forall x \in [-a,a]
\end{aligned}$$

Going back to the original variable z, we finally get that

$$P(z) = P_1(z) + (z + z^{-1} - 2\cos\omega)P_2(z), \quad \forall z \in \mathbb{T}$$

for some non-negative trigonometric matrix polynomials $P_1(z)$ and $P_2(z)$ of degree n and $n-1$, respectively. □

Because any element of $\mathcal{K}_{[-\omega,+\omega]}$ can be written as a weighted sum of squares, Theorems 2.3 and 2.5 lead to the following result.

Theorem 2.19. *Given $S \in E$, define the block linear operators T_1 and T_2 by*

$$\begin{aligned}
[T_1(S)]_{ij} &= S_{i-j}, \quad 0 \leq i,j \leq n, \\
[T_2(S)]_{ij} &= S_{i-j+1} + S_{i-j-1} - 2S_{i-j}\cos\omega, \quad 0 \leq i,j \leq n-1.
\end{aligned}$$

with $S_{-i} = S_i^*$. Then

1. *Primal cone:*

$$\mathcal{K}_{[-\omega,+\omega]} = \{P : P = T_1^*(Y_1) + T_2^*(Y_2), Y_1 \in \mathcal{H}_+^{m(n+1)}, Y_2 \in \mathcal{H}_+^{mn}\}.$$

2. *Dual cone:*
$$\mathcal{K}^{\star}_{[-\omega,+\omega]} = \{S : T_1(S) \succeq 0, T_2(S) \succeq 0\}.$$

3. *Both cones are closed, convex and pointed. Consequently, they also have non-empty interiors.*

Proof. Theorem 2.18 implies that $\mathcal{K}_{[-\omega,+\omega]}$ is a sum of two sub-cones, which admit sum-of-squares representations. Therefore the statement follows from applying Theorem 2.5 to the appropriate sub-cones. Let us be more specific. If the notation of Section 2.1 is used, these two sub-cones are characterized by the functional elements

$$\phi(z) = 1, \quad \psi(z) = \pi_n(z), \quad \Psi(z) = \Pi_n(z),$$

and

$$\phi(z) = (z + z^{-1} - 2\cos\omega), \quad \psi(z) = \pi_{n-1}(z), \quad \Psi(z) = \Pi_{n-1}(z),$$

respectively. We use the upper-level functional system

$$\upsilon(z) = \pi_n(z), \quad \Upsilon(z) = \Pi_n(z),$$

which is minimal. Then the structure of the associated linear operators T_1 and T_2 are obtained from the identities

$$[\pi_n(z)\pi_n(z)^*]_{ij} = z^{i-j},$$
$$[(z + z^{-1} - 2\cos\omega)\pi_{n-1}(z)\pi_{n-1}(z)^*]_{ij} = z^{i-j+1} + z^{i-j-1} - 2z^{i-j}\cos\omega,$$

which are easy to check. Note that the relations $T_1(S) = 0$ and $T_2(S) = 0$ imply that $S = 0$. To see this, consider the first block column of $T_1(S)$. The statement then follows from Theorems 2.3 and 2.5: the first one is used to characterize both sub-cones of interest and the second one leads to the characterization of their sum, which is equal to $\mathcal{K}_{[-\omega,+\omega]}$. □

Of course, we get the usual dual representation of $P \in \mathcal{K}_{[-\omega,+\omega]}$:

$$P(z) = \langle\langle T_1(S)^{-1}\Pi_n(z), \Pi_n(z)\rangle\rangle + (z + z^{-1} - 2\cos\omega)\langle\langle T_2(S)^{-1}\Pi_{n-1}(z), \Pi_{n-1}(z)\rangle\rangle,$$

where the point $S \in \operatorname{int}\mathcal{K}^{\star}_{[-\omega,+\omega]}$ is uniquely defined.

2.3.3 Interval $[-1, 1] \subset \mathbb{R}$

Our characterizations of cones of non-negative polynomials obviously depend on our choice of the functional basis; different functional basis yield different parametrizations. Let us illustrate this fact using the cone of self-adjoint polynomials (of degree $d = n$) that are non-negative on the interval $[-1, 1]$

$$\mathcal{K}_{[-1,1]} = \{P \in E : P(x) \succeq 0, \forall x \in [-1, 1]\}$$

with E as defined in (2.11).

Theorems 2.11 and 2.12 with $a = -b = 1$ give us an interesting parametrization of $\mathcal{K}_{[-1,1]}$, provided that the standard monomial basis is used. Let us modify the

functional basis. For instance, we consider a Chebyshev basis. Remember that first-order Chebyshev polynomials are orthogonal polynomials that satisfy a three-term recurrence

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$$

initialized with $T_0(x) = 1$ and $T_1(x) = x$. Now,

$$P(x) = \sum_{k=0}^{n} P_k T_k(x) \succeq 0, \quad \forall x \in [-1, 1]$$

if and only if

$$P(\theta) = \sum_{k=0}^{n} P_k T_k(\cos\theta) \succeq 0, \quad \forall \theta \in [0, 2\pi].$$

By definition, $T_k(\cos\theta) = \cos k\theta$ for all k. From this observation, Theorem 2.16 directly yields the following corollary.

Corollary 2.20. *Given $S \in E$, define the block linear operator T_c by*

$$[T_c(S)]_{ij} = S_{|i-j|}, \quad 0 \le i, j \le n.$$

Then

1. *Primal cone:*

$$\begin{aligned}\mathcal{K}_{[-1,1],c} &= \{P \in E : \sum_{k=0}^{n} P_k T_k(x) \succeq 0, \forall x \in [-1, 1]\} \\ &= \{P \in E : P = T_c^*(Y), Y \in \mathcal{H}_+^{m(n+1)}\}.\end{aligned}$$

2. *Dual cone:*

$$\mathcal{K}^\star_{[-1,1],c} = \{S \in E : T_c(S) \succeq 0\}.$$

3. *Both cones are closed, convex and pointed. Consequently, they also have non-empty interiors.*

Proof. Indeed, $\mathcal{K}_{[-1,1],c}$ is a linear image of $\mathcal{K}_{\mathbb{C}}$. To see this, note that $P \in \mathcal{K}_{[-1,1],c}$ if and only if there exists $\tilde{P} \in \mathcal{K}_{\mathbb{C}}$ such that

$$P_k = \frac{\tilde{P}_k + \tilde{P}_k^*}{2}, \quad k = 0, \ldots, n.$$

Moreover, the linear transformation from $\tilde{P}$ to P is nonsingular. This observation concludes the proof. □

The dual operator $T_c^* : \mathbb{C}^{m(n+1)\times m(n+1)} \to E$ is defined as

$$P = T_c^*(Y) \iff P_k = \sum_{|i-j|=k} Y_{ij}, k = 0, \ldots, n$$

where Y_{ij} is the block (i, j) of dimension m of the matrix Y.

In the next section, the dual parametrizations obtained in the two previous sections are shown to be related with moment problems.

2.4 Moment problems

Moment problems have been studied for a long time in the literature, see e.g. [2, 3]. The mathematical foundations of this class of problems were laid by several famous mathematicians, e.g. Chebyshev, Hausdorff, Markov, Stieltjes. They consider several variations of a simple question:

> Given a sequence of moments, what is the associated measure (if any)?

Two aspects of this question must be taken into account. First, one need to characterize solvability of the problem. Then, appropriate solutions must be provided for solvable instances. The aim of this section is to put into light the strong connection between moment problems and non-negative polynomials.

In the case of $m \times m$ Hermitian moments, the standard definition of measure [32] is extended as follows. A *Hermitian (non-negative) matrix measure* μ is an $m \times m$ matrix of complex measures μ_{ij} on the Borel field of $\mathbb{R}$ such that, for each Borel set A, the matrix $\mu(A)$ is Hermitian and non-negative definite.

2.4.1 Classical moments

Hamburger moment problem

Let us discuss the *Hamburger moment problem*, which can be stated as follows:

> Given an infinite set of Hermitian matrices $\{S_k\}_{k\geq 0}$, find a measure μ such that the identities
> $$S_k = \int_{-\infty}^{+\infty} t^k \mu(dt), \quad k \geq 0$$
> hold.

From this statement, it is clear that sets of Hermitian matrices for which this problem is solvable are elements of a convex cone. In applied mathematics, this problem has few applications as infinite sets of matrices are difficult to treat. For this reason, mathematicians have started to look at the *truncated Hamburger moment problem*, which only fix the first $2n+1$ moments

$$S_k = \int_{-\infty}^{+\infty} t^k \mu(dt), \quad k = 0, \ldots, 2n.$$

This truncation procedure is similar to our implicit assumption that only polynomials of bounded degree are considered in Section 2.2. This truncated problem is solvable if and only if there exist S_{2n+1} and S_{2n+2} such that the block Hankel matrix $H(S)$ defined by $\{S_0, \ldots, S_{2n+2}\}$ is positive semidefinite [28, Theorem 3.1]. In particular, it is solvable if the block Hankel matrix $H(S)$ is positive definite.

A variation of the truncated moment problem relax the last equality constraint into an inequality

$$S_{2n} \succeq \int_{-\infty}^{+\infty} t^{2n} \mu(dt).$$

Now, this modified problem is solvable if and only if the block Hankel matrix $H(S)$ defined by $\{S_0, \ldots, S_{2n}\}$ is positive semidefinite [28, Theorem 3.2]. This last condition is equivalent to $S \in \mathcal{K}_{\mathbb{R}}^{\star}$.

For any $S \in \mathcal{K}_{\mathbb{R}}^{\star}$, the truncated moment problem is solvable. Therefore it sounds reasonable to look for a solution. For instance, one could decompose the block Hankel matrix $H(S)$ into factors. The next result provides an explicit answer for $S \in \operatorname{int} \mathcal{K}_{\mathbb{R}}^{\star}$.

Theorem 2.21 ([132, Theorem 1.4]). *If $S \in \operatorname{int} \mathcal{K}_{\mathbb{R}}^{\star}$, the block Hankel matrix $H(S)$ can be factorized as*

$$H(S) = \sum_{i=1}^{m(n+1)} w_i (\pi_n(x_i) \otimes v_i)(\pi_n(x_i) \otimes v_i)^*$$

with $w_i > 0, x_i \in \mathbb{R}$ and $v_i \in \mathbb{C}^m$, for all i. Moreover the Vandermonde-like matrix

$$V \doteq \begin{bmatrix} \pi_n(x_1) \otimes v_1, & \ldots, & \pi_n(x_i) \otimes v_i, & \ldots, & \pi_n(x_{m(n+1)}) \otimes v_{m(n+1)} \end{bmatrix}$$

is nonsingular.

Theorem 2.21 has an interesting consequence in the context of scalar polynomials ($m = 1$). Let $p \in \operatorname{int} \mathcal{K}_{\mathbb{R}}$ be a scalar polynomial of degree $2n$. Then the dual representation (2.16) yields the existence of a unique $s \in \operatorname{int} \mathcal{K}_{\mathbb{R}}^{\star}$ such that

$$p = H^*(Y), \quad Y = H(s)^{-1} \succ 0.$$

Using Theorem 2.21, the Hankel matrix $H(s)$ admits a Vandermonde factorization

$$H(s) = V D V^T, \tag{2.35}$$

where D is a diagonal matrix and V is a nonsingular Vandermonde matrix. Let $\{w_i\}_{i=0}^n$ be the elements of D and $\{x_i\}_{i=0}^n$ be the nodes of V, i.e.,

$$V = \begin{bmatrix} 1 & \ldots & 1 \\ x_0 & \ldots & x_n \\ \vdots & & \vdots \\ x_0^n & \ldots & x_n^n \end{bmatrix}.$$

If L^T denotes the inverse of V, the rows of L^T are the coefficients of $n+1$ Lagrange polynomials $\{l_i(x)\}_{i=0}^n$. These polynomials are defined by the distinct nodes $\{x_i\}_{i=0}^n$:

$$l_i(x_j) = \langle L e_i, \pi_n(x_j) \rangle = \delta_{ij}, \quad 0 \leq i, j \leq n,$$

where δ_{ij} is the Kronecker delta. An explicit expression of Y in terms of the Lagrange polynomials is

$$Y = L D^{-1} L^T. \tag{2.36}$$

This representation of Y is equivalent to the decomposition of $p(x)$ as a sum of $n+1$ squared Lagrange polynomials

$$p(x) = \sum_{i=0}^{n} w_i^{-1} (l_i(x))^2.$$

Indeed, let us plug the expression (2.36) into the identity $p(x) = \langle Y\pi_n(x), \pi_n(x)\rangle$. We have

$$\begin{aligned} p(x) &= \langle LD^{-1}L^T\pi_n(x), \pi_n(x)\rangle = \sum_{i=0}^{n} w_i^{-1}\langle Le_ie_i^TL^T, \pi_n(x)\pi_n(x)^T\rangle \\ &= \sum_{i=0}^{n} w_i^{-1}(\langle Le_i, \pi_n(x)\rangle)^2 = \sum_{i=0}^{n} w_i^{-1}(l_i(x))^2. \end{aligned}$$

As $\{l_i(x)\}_{i=0}^n$ are the Lagrange polynomials associated with the points $\{x_i\}_{i=0}^n$, it is straightforward to check that $w_i^{-1} = p(x_i), \forall i$.

Hausdorff moment problem

Let $[a, b] \subset \mathbb{R}$ be a finite interval. Given a set of Hermitian matrices $\{S_k\}_{k=0}^{2n}$, the *Hausdorff moment problem* consists in finding a measure μ such that the identities

$$S_k = \int_a^b t^k \mu(dt), \quad k = 0, \ldots, 2n$$

hold. This problem is solvable if and only if the block Hankel matrices $H_1(S)$ and $H_2(S)$ defined by $\{S_0, \ldots, S_{2n}\}$ are positive semidefinite [7, Theorem 3]. This last condition is equivalent to $S \in \mathcal{K}^\star_{[a,b]}$.

The main difference between this moment problem and the truncated Hamburger moment problem is the support of the measure μ. Indeed, the Hausdorff moment problem imposes that the support of μ is contained in $[a, b]$. This restriction is enforced by adding the constraint $H_2(S) \succeq 0$.

Stieltjes moment problem

Given a set of Hermitian matrices $\{S_k\}_{k=0}^{2n}$, the *Stieltjes moment problem* consists in finding a measure μ such that the identities

$$\begin{aligned} S_k &= \int_0^{+\infty} t^k \mu(dt), \quad k = 0, \ldots, n-1, \\ S_n &\succeq \int_0^{+\infty} t^n \mu(dt), \end{aligned}$$

hold. This problem is solvable if and only if the block Hankel matrices $H_5(S)$ and $H_6(S)$ defined by $\{S_0, \ldots, S_n\}$ are positive semidefinite [29, Lemma 1.2]. This last condition is equivalent to $S \in \mathcal{K}^\star_{[0,+\infty)}$.

The support of μ should be within the interval $[0, +\infty)$ so that we get the additional localization constraint $H_6(S) \succeq$. The inequality on S_n is related to points at infinity. These points require a special attention and their specificity was sometimes overlooked in the literature.

2.4.2 Trigonometric moments

Trigonometric moment problem

Given a set of Hermitian matrices $\{S_k\}_{k=0}^n$, the *partial trigonometric moment problem* consists in finding a measure μ such that the identities

$$S_k = \int_{-\pi}^{+\pi} e^{\jmath k\theta} \mu(d\theta), \quad k = 0, \dots, n$$

hold. This problem is solvable if and only if the block Toeplitz matrix $T(S)$ defined by $\{S_0, \dots, S_n\}$ is positive semidefinite [42]. This last condition is equivalent to $S \in \mathcal{K}_{\mathbb{C}}^{\star}$. For the sake of completeness, we mention that the *trigonometric moment problem* considers the case with $n = +\infty$.

Positive definite block Toeplitz matrices also admit a Vandermonde factorization, as shown in the following theorem.

Theorem 2.22 ([133, Theorem 3.2]). *If $S \in \operatorname{int} \mathcal{K}_{\mathbb{C}}^{\star}$, the block Toeplitz matrix $T(S)$ can be factorized as*

$$T(S) = \sum_{i=1}^{m(n+1)} w_i (\pi_n(z_i) \otimes v_i)(\pi_n(z_i) \otimes v_i)^*$$

with $w_i > 0, z_i \in \mathbb{T}$ and $v_i \in \mathbb{C}^m$, for all i. Moreover the Vandermonde-like matrix

$$V \doteq \begin{bmatrix} \pi_n(z_1) \otimes v_1, & \dots, & \pi_n(z_i) \otimes v_i, & \dots, & \pi_n(z_{m(n+1)}) \otimes v_{m(n+1)} \end{bmatrix}$$

is nonsingular.

In the scalar case ($m = 1$), this theorem also gives a decomposition of $p(z)$, which is obtained from the dual representation. The analysis carried out for polynomials on the real line can be reproduced with minor changes. Given a non-negative trigonometric polynomial $p \in \operatorname{int} \mathcal{K}_{\mathbb{C}}$ and its dual representation, say

$$p = T^*(Y), \quad Y = T(s)^{-1}, \quad s \in \operatorname{int} \mathcal{K}_{\mathbb{C}}^{\star},$$

one can factorize $T(s)$ as $T(s) = VDV^*$ by making use of Theorem 2.22.

Remember that D is a diagonal matrix and V is a nonsingular Vandermonde matrix. The nodes $\{z_i\}_{i=0}^n$ of V are distinct and belongs to the unit circle, so that the associated Lagrange polynomials $\{l_i(z)\}_{i=0}^n$ are well defined. Their coefficients are obtained via the rows of the matrix L^* such that $L^*V = I_{n+1}$. Therefore, Y can be factorized as $Y = LD^{-1}L^*$. As a consequence, the dual representation (2.30) is linked to the decomposition of $p(z) \in \operatorname{int} \mathcal{K}_{\mathbb{C}}$ as a weighted sum of squared Lagrange polynomials. Indeed, we have

$$p(z) = \sum_{i=0}^{n} p(z_i) |l_i(z)|^2, \quad |z| = 1.$$

Trigonometric moment problem on an interval

Given a set of Hermitian matrices $\{S_k\}_{k=0}^n$, the *partial trigonometric moment problem on an interval* $[-\omega, +\omega] \subset [-\pi, \pi]$ consists in finding a measure μ with support $[-\omega, +\omega]$

for which

$$S_k = \int_{-\pi}^{+\pi} e^{\jmath k\theta} \mu(d\theta), \quad k = 0, \ldots, n.$$

This problem is solvable if and only if the block Toeplitz matrices $T_1(S)$ and $T_2(S)$ defined by $\{S_0, \ldots, S_n\}$ are positive semidefinite [5, Section 2]. This last condition is equivalent to $S \in \mathcal{K}^{\star}_{[-\omega,+\omega]}$.

The second constraint $T_2(s) \succeq 0$ has the same purpose than in the classical moment problems; it locates the support of μ. Note that, for trigonometric moments, there is no need to treat infinite points in a specific manner.

2.4.3 Link with non-negative polynomials

From these results, the strong link between non-negative polynomials of degree $2n$ and non-negative measures is apparent. These objects are dual in the sense that they basically belong to dual spaces. Non-negative polynomials are best represented using their coefficients, whereas measures are identified by means of their moments.

To formalize our previous statement, let $P(x) = \sum_{k=0}^{d} P_k x^k$ be a matrix polynomial and μ be a measure. Then, for any degree d and any interval $I \subseteq \mathbb{R}$, we get the equivalences

$$P(x) \succeq 0, \forall x \in I \iff \operatorname{Re}\operatorname{Trace} \int_I P(x)\mu(dx) \geq 0, \forall \mu \,\text{s. t.}\, \mu \succeq 0, \operatorname{supp}\mu \subseteq I,$$

$$\mu \succeq 0 \text{ on } I \iff \operatorname{Re}\operatorname{Trace} \int_I P(x)\mu(dx) \geq 0, \forall P \,\text{s. t.}\, P(x) \succeq 0, \forall x \in I.$$

Of course, similar relations hold for trigonometric polynomials. Although the focus of this thesis is the polynomial setting, most results and formulations could be written in their dual form, using the moment setting.

2.5 Non-negativity and linear matrix inequalities in systems and control

In the systems theory literature, there exists an important result, known as the Kalman-Yakubovich-Popov Lemma, which links a non-negativity constraint in the frequency domain with a linear matrix inequality. It is not surprising that our characterizations of the primal cones $\mathcal{K}_{\mathbb{R}}$ and $\mathcal{K}_{\mathbb{C}}$ can also be obtained from this lemma. Before illustrating this important link, we provide the reader with an overview of certain fundamental concepts in this research area.

A matrix function $\Phi : \mathbb{C} \to \mathbb{C}^{n\times n}$ maps the complex variable λ to a complex matrix $\Phi(\lambda)$. In this section we restrict ourselves to *rational* matrix functions; the elements of $\Phi(\lambda)$ are rational functions of λ. We introduce the *paraconjugate transpose* of this function with respect to a particular curve Γ in the complex plane. For convenience, we use a different symbol for the variable λ for each choice of Γ. The three cases are (a) the real axis (including the point at infinity), (b) the imaginary axis (relevant to continuous-time dynamical systems) and (c) the unit circle (relevant to

discrete-time dynamical systems):

$$\Phi_{\mathbb{R}}(x) \doteq [\Phi(\bar{x})]^* \qquad \text{for } \Gamma = \mathbb{R}, \tag{2.37a}$$

$$\Phi_{\mathbb{C}}(s) \doteq [\Phi(-\bar{s})]^* \qquad \text{for } \Gamma = \jmath\mathbb{R}, \tag{2.37b}$$

$$\Phi_{\mathbb{D}}(z) \doteq [\Phi(1/\bar{z})]^* \qquad \text{for } \Gamma = e^{\jmath\mathbb{R}}. \tag{2.37c}$$

Because these three particular cases can be treated in much the same way, we use the generic notation $\Phi_*(.)$ for $\Phi_{\mathbb{R}}(.)$, $\Phi_{\mathbb{C}}(.)$ and $\Phi_{\mathbb{D}}(.)$ and λ for the variables x, s and z.

If $\Phi(\lambda)$ is rational and analytic in λ, so is its paraconjugate transpose as it can be easily verified. We now say that $\Phi(\lambda)$ is a *para-Hermitian matrix function*, if it is its own paraconjugate transpose, i.e., if

$$\Phi_*(\lambda) = \Phi(\lambda), \quad \forall \lambda \in \Gamma,$$

for the three cases (2.37a),(2.37b),(2.37c). It is a well-known fact that there exist first-order conformable transformations of the complex plane that link the above cases with each other. Since substituting the real axis for the imaginary axis is a trivial operation ($\jmath x = s$), only the cases (2.37b) and (2.37c) will be emphasized in what follows, in view of the importance of their role in continuous-time and discrete-time control systems, respectively. Furthermore, results specific to the above two situations will be identified by using the *Laplace operator* ($\lambda = s$) for continuous-time systems and the *delay operator* ($\lambda = z$) for discrete-time systems.

2.5.1 Generalized state-space realizations

Rational transfer matrices can always be represented as simple expressions involving first order polynomial matrix functions, i.e., pencils. Indeed, every rational transfer matrix $G(\lambda)$ of dimension $p \times m$ is known from realization theory to admit a *generalized state-space model* [142] of the form

$$G(\lambda) = (C - \lambda F)(\lambda E - A)^{-1} B + D, \tag{2.38}$$

which is the Schur complement of the *system matrix* $S(\lambda)$ of dimension $(n+p)\times(n+m)$

$$S(\lambda) = \begin{bmatrix} A - \lambda E & B \\ C - \lambda F & D \end{bmatrix} \tag{2.39}$$

with respect to its top left block entry. The minimum dimension n of the invertible pencil $(A - \lambda E)$ is the McMillan degree of $G(\lambda)$, and the generalized eigenvalues of $A - \lambda E$ are then the poles of $G(\lambda)$ [142, 84]. A test for the minimality of the realization $S(\lambda)$ is the following set of conditions [115]:

$$\text{(i)} \ \operatorname{rank} \begin{bmatrix} A - \lambda E & B \end{bmatrix} = n, \quad \forall \lambda \in \mathbb{C}, |\lambda| < \infty; \tag{2.40a}$$

$$\text{(ii)} \ \operatorname{rank} \begin{bmatrix} E & B \end{bmatrix} = n; \tag{2.40b}$$

$$\text{(iii)} \ \operatorname{rank} \begin{bmatrix} A - \lambda E \\ C - \lambda F \end{bmatrix} = n, \quad \forall \lambda \in \mathbb{C}, |\lambda| < \infty; \tag{2.40c}$$

$$\text{(iv)} \ \operatorname{rank} \begin{bmatrix} E \\ F \end{bmatrix} = n. \tag{2.40d}$$

If all these conditions are not satisfied, then the system matrix (2.39) is not minimal and the state-space dimension is reducible to the minimal dimension [142]. Minimal realizations are not unique, even though their dimension n is. A simple class of transformations acting on (2.39) and preserving minimality is defined by

$$\hat{S}(\lambda) = \begin{bmatrix} \hat{A} - \lambda\hat{E} & \hat{B} \\ \hat{C} - \lambda\hat{F} & \hat{D} \end{bmatrix} = \begin{bmatrix} P & 0 \\ Y & I_p \end{bmatrix} \begin{bmatrix} A - \lambda E & B \\ C - \lambda F & D \end{bmatrix} \begin{bmatrix} Q & X \\ 0 & I_m \end{bmatrix}, \tag{2.41}$$

with P, Q invertible matrices and where X is chosen so as to have $EX = FX = 0$. Indeed, the Schur complements of $S(\lambda)$ and $\hat{S}(\lambda)$ are easily checked to be identical so that they are both well defined realizations of $G(\lambda)$. The minimality conditions (2.40) for $\hat{S}(\lambda)$ are easily seen to hold as well since left and right invertible transformation of the relevant pencils do not change their rank. Let us finally point out that there is no unique definition of minimality for generalized state-space systems.

The zeros of $G(\lambda)$ can also be computed as generalized eigenvalues of a smaller pencil, derived from $S(\lambda)$. Let M be any invertible row transformation such that

$$\begin{bmatrix} \tilde{E} \\ 0 \end{bmatrix} \doteq M \begin{bmatrix} E \\ F \end{bmatrix}, \tag{2.42}$$

where $\tilde{E}$ is $n \times n$ (one can choose M to be unitary to achieve this). Applying this transformation to $S(\lambda)$ defines the matrices $\tilde{A}, \tilde{B}, \tilde{C}$ and $\tilde{D}$ as follows

$$\tilde{S}(\lambda) \doteq MS(\lambda) = \begin{bmatrix} \tilde{A} - \lambda\tilde{E} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{bmatrix}.$$

If $\tilde{D}$ is invertible, it follows from [84] that the Schur complement

$$(\tilde{A} - \lambda\tilde{E}) - \tilde{B}\tilde{D}^{-1}\tilde{C}$$

is a *zero pencil* of the system: its generalized eigenvalues are the zeros of $G(\lambda)$. If $\tilde{D}$ is not invertible, it is shown in [84] that one can still derive a zero pencil but we will not further elaborate on this here.

2.5.2 Para-Hermitian realizations

It is natural to expect that para-Hermitian transfer matrices admit realizations and system matrices, which reflect this property in some structural way. In fact, we shall consider generalized state-space realizations of Popov type

$$S(s) = \begin{bmatrix} 0 & A^* + sE^* & C^* + sF^* \\ A - sE & Y_{11} & Y_{12} \\ C - sF & Y_{21} & Y_{22} \end{bmatrix} \qquad \text{for } \Gamma = \jmath\mathbb{R}, \tag{2.43a}$$

$$S(z) = \begin{bmatrix} 0 & zA^* - E^* & zC^* - F^* \\ A - zE & Y_{11} & Y_{12} \\ C - zF & Y_{21} & Y_{22} \end{bmatrix} \qquad \text{for } \Gamma = e^{\jmath\mathbb{R}}, \tag{2.43b}$$

where

$$Y = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} \in \mathcal{H}^{(n+p)}.$$

Setting the matrix $T(\lambda) = (C - \lambda F)(\lambda E - A)^{-1}$, one finds after some algebraic manipulations that the transfer function corresponding to these system matrices is given by

$$\Phi(\lambda) = \begin{bmatrix} T(\lambda) & I \end{bmatrix} \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} \begin{bmatrix} T_*(\lambda) \\ I \end{bmatrix}, \tag{2.44}$$

which is clearly para-Hermitian. Note that the above realizations are not necessarily minimal. If it is the case, then the poles of the transfer function are the eigenvalues of, respectively,

$$\begin{bmatrix} 0 & A^* + sE^* \\ A - sE & Y_{11} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & zA^* - E^* \\ A - zE & Y_{11} \end{bmatrix}$$

which are symmetric with respect to Γ (i.e., $s_i, -\bar{s}_i$ or $z_i, 1/\bar{z}_i$). The zeros of the transfer function are those of the respective system matrices, which clearly exhibit the same form of symmetry. If we define a transformation matrix M as in (2.42) and then define accordingly:

$$\begin{bmatrix} \tilde{A} \\ \tilde{C} \end{bmatrix} \doteq M \begin{bmatrix} A \\ C \end{bmatrix}, \quad \begin{bmatrix} \tilde{Y}_{11} & \tilde{Y}_{12} \\ \tilde{Y}_{21} & \tilde{Y}_{22} \end{bmatrix} \doteq M \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} M^*, \tag{2.45}$$

this yields a new system matrix $\tilde{S}(\lambda)$ which is similar to $S(\lambda)$ but with $\tilde{F} = 0$, and which has the same zeros. If now $\tilde{Y}_{22}$ is invertible, then the zeros are also the generalized eigenvalues of the respective Schur complements of the system matrices:

$$\begin{bmatrix} 0 & \tilde{A}^* + s\tilde{E}^* \\ \tilde{A} - s\tilde{E} & \tilde{Y}_{11} \end{bmatrix} - \begin{bmatrix} \tilde{C}^* \\ \tilde{Y}_{12} \end{bmatrix} \tilde{Y}_{22}^{-1} \begin{bmatrix} \tilde{C} & \tilde{Y}_{21} \end{bmatrix}, \tag{2.46a}$$

$$\begin{bmatrix} 0 & z\tilde{A}^* - \tilde{E}^* \\ \tilde{A} - z\tilde{E} & \tilde{Y}_{11} \end{bmatrix} - \begin{bmatrix} z\tilde{C}^* \\ \tilde{Y}_{12} \end{bmatrix} \tilde{Y}_{22}^{-1} \begin{bmatrix} \tilde{C} & \tilde{Y}_{21} \end{bmatrix} \tag{2.46b}$$

which are known in the literature as the Hamiltonian and the symplectic pencils describing the zeros of the respective para-Hermitian transfer functions. This could suggest that para-Hermitian transfer functions would be always of even degree. This is actually not the case, as exemplified by the first degree scalar para-Hermitian transfer function $G(s) = 2\jmath/s$, which has a pole at $s = 0$ and a zero at $s = \infty$; note that odd degree para-Hermitian transfer functions must have at least one pole and one zero on the curve Γ. However, if the transfer function, as considered in the sequel, is not only para-Hermitian but also non-negative definite on the contour Γ, then it has always a minimal realization of type (2.43); this is easily established from the well-known spectral factorization property $\Phi(\lambda) = G(\lambda)G_*(\lambda)$ of any such transfer function [154].

2.5.3 Linear and Riccati matrix inequalities

Realizations of the type (2.43) are obviously not unique, since they are invariant under transformations of the form (2.41). We now introduce the following transformations which leave (E, A, C, F) invariant:

$$\begin{bmatrix} I & 0 & 0 \\ EX & I & 0 \\ FX & 0 & I \end{bmatrix} S(s) \begin{bmatrix} I & XE^* & XF^* \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} = \begin{bmatrix} 0 & A^* + sE^* & C^* + sF^* \\ A - sE & Y_{11}(X) & Y_{12}(X) \\ C - sF & Y_{21}(X) & Y_{22}(X) \end{bmatrix}, \tag{2.47a}$$

$$\begin{bmatrix} I & 0 & 0 \\ EX & I & 0 \\ FX & 0 & I \end{bmatrix} S(z) \begin{bmatrix} I & XA^* & XC^* \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} = \begin{bmatrix} 0 & zA^* - E^* & zC^* - F^* \\ A - zE & Y_{11}(X) & Y_{12}(X) \\ C - zF & Y_{21}(X) & Y_{22}(X) \end{bmatrix}, \tag{2.47b}$$

where the matrices $Y(X)$ are given by

$$Y(X) \doteq \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} + \begin{bmatrix} E \\ F \end{bmatrix} X \begin{bmatrix} A^* & C^* \end{bmatrix} + \begin{bmatrix} A \\ C \end{bmatrix} X \begin{bmatrix} E^* & F^* \end{bmatrix}, \tag{2.48a}$$

$$Y(X) \doteq \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} + \begin{bmatrix} A \\ C \end{bmatrix} X \begin{bmatrix} A^* & C^* \end{bmatrix} - \begin{bmatrix} E \\ F \end{bmatrix} X \begin{bmatrix} E^* & F^* \end{bmatrix}, \tag{2.48b}$$

respectively. Let us emphasize that the Schur complement (i.e., the transfer function) of these realizations does not change under these transformations.

Since the entries of the matrix inequality

$$Y(X) \succeq 0$$

are linear in the elements of the unknown matrix X, it is called a *linear matrix inequality*. Such inequalities are related to transfer functions $\Phi(\lambda)$ that are non-negative definite on Γ. Let us now further assume that $Y_{22}(X)$ is positive definite. Then, the Schur complement of $Y(X)$ with respect to $Y_{22}(X)$ must be non-negative definite. It is easy to check that this amounts to the constraints

$$\begin{aligned} &Y_{11} + EXA^* + AXE^* - (Y_{12} + EXC^* + AXF^*) \\ &\quad (Y_{22} + FXC^* + CXF^*)^{-1}(Y_{21} + FXA^* + CXE^*) \succeq 0, \end{aligned} \tag{2.49a}$$

$$\begin{aligned} &Y_{11} + AXA^* - EXE^* - (Y_{12} + AXC^* - EXF^*) \\ &\quad (Y_{22} + CXC^* - FXF^*)^{-1}(Y_{21} + CXA^* - FXE^*) \succeq 0, \end{aligned} \tag{2.49b}$$

in the continuous-time and discrete-time cases, respectively. These inequalities are known as the *Riccati matrix inequalities* and were introduced in [149].

For appropriate choices of X, one has that $\operatorname{rank} Y_{22}(X) = \operatorname{rank} Y(X)$ so that its Schur complement in $Y(X)$ must be zero. The above inequalities then become equalities, which have the form of the celebrated *Riccati equations* for which X appears therefore to be a solution. These equations are rewritten below for the case that $F = 0$ since this can always be obtained under a transformation (2.42)–(2.45) of the system:

$$Y_{11} + EXA^* + AXE^* - (Y_{12} + EXC^*)(Y_{22})^{-1}(Y_{21} + CXE^*) = 0, \tag{2.50a}$$

$$Y_{11} + AXA^* - EXE^* - (Y_{12} + AXC^*)(Y_{22} + CXC^*)^{-1}(Y_{21} + CXA^*) = 0. \tag{2.50b}$$

The solution X of these equations is obtained from the calculation of eigenspaces of the zero pencils (2.46) [92]. If there are no repeated eigenvalues in these pencils (this is the generic case) then the number of solutions X to these equations are finite, whereas the inequalities have typically an infinite solution set.

Remark. Alternatively, one can consider realizations of the form

$$\tilde{G}(\lambda) = \tilde{C}(\lambda\tilde{E} - \tilde{A})^{-1}(\tilde{B} - \lambda\tilde{K}) + \tilde{D},$$

which are "dual" to the generalized state-space representation in (2.38), in the sense that $\tilde{G}_*(\lambda)$ has precisely a realization of the form (2.38). Then, the state-space representation (2.44) or the expressions of the matrices $Y(X)$ showing up in (2.48) can be rewritten accordingly.

2.5.4 Kalman-Yakubovich-Popov Lemma

It turns out that the celebrated Kalman-Yakubovich-Popov Lemma [87, 152, 123] allows one to express the non-negative definiteness of $\Phi(\lambda)$ on the curve Γ in terms of conditions on $Y(X)$. This theorem was initially formulated for standard state-space transfer functions, i.e., $E = I$ and $F = 0$.

Theorem 2.23 (KYP Lemma, [123]). *Given $A \in \mathbb{C}^{n\times n}$, $C \in \mathbb{C}^{p\times n}$ and $Y \in \mathcal{H}^{(n+p)}$ with $\det(\lambda I - A) \neq 0, \lambda \in \Gamma$ and $\operatorname{rank}[A^* - \lambda I, C^*] = n, \forall \lambda \in \mathbb{C}$, define $T(\lambda) = C(\lambda I - A)^{-1}$. Then the transfer function*

$$\Phi(\lambda) = \begin{bmatrix} T(\lambda) & I \end{bmatrix} \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} \begin{bmatrix} T_*(\lambda) \\ I \end{bmatrix}$$

is non-negative definite on Γ if and only if there exists a Hermitian matrix X such that

$$Y(X) \doteq Y + \begin{bmatrix} XA^* + AX & XC^* \\ CX & 0 \end{bmatrix} \succeq 0, \tag{2.51a}$$

$$Y(X) \doteq Y + \begin{bmatrix} AXA^* - X & AXC^* \\ CXA^* & CXC^* \end{bmatrix} \succeq 0. \tag{2.51b}$$

This result can be extended to generalized state-space transfer functions.

Theorem 2.24 (Extended KYP Lemma, [56, Theorem 2.1]). *Every para-Hermitian transfer function $\Phi(\lambda)$ has a realization as in* (2.43)*, with condition* (2.40) *satisfied. Moreover, it is non-negative definite on Γ,*

$$\Phi(\lambda) \succeq 0, \qquad \text{for almost all } \lambda \in \Gamma, \tag{2.52}$$

if and only if there exists a Hermitian matrix X such that

$$Y(X) \succeq 0 \tag{2.53}$$

where $Y(X)$ is defined in (2.48)*.*

The above result is linked to that of *spectral factorization*. If $Y(X)$ is positive semidefinite, it can be factorized as

$$Y(X) = \begin{bmatrix} L \\ W \end{bmatrix} \begin{bmatrix} L^* & W^* \end{bmatrix},$$

so that $\hat{S}(\lambda)$ is easily seen to be the system matrix of

$$\Phi(\lambda) = \Xi(\lambda)\Xi_*(\lambda)$$

with the *spectral factor* $\Xi(\lambda)$ defined by

$$\Xi(\lambda) = (C - \lambda F)(\lambda E - A)^{-1}L + W.$$

For more details on spectral factorizations of rational matrices, see [115, 154].

2.5.5 Non-negative matrix polynomials

Our parametrization of non-negative matrix polynomials, obtained in Theorems 2.7 and 2.16, can alternatively be obtained from the theory of positive para-Hermitian transfer functions. More precisely, it follows from a straightforward application of the Kalman-Yakubovich-Popov Lemma to the subclass of positive para-Hermitian transfer functions that have a pseudo-polynomial form.

Real line $\mathbb{R}$

From Theorem 2.24, a necessary and sufficient condition for $\Phi(s)$ to be a para-Hermitian transfer function non-negative on the whole of the imaginary axis is the existence of a Hermitian matrix X such that $Y(X)$ is non-negative definite. Let us apply this result to the transfer function

$$\Phi(s) = \begin{bmatrix} -\jmath F(-sI_n + \jmath Z)^{-1} & I_m \end{bmatrix} Y_0 \begin{bmatrix} (sI_n - \jmath Z^T)^{-1}\jmath F^T \\ I_m \end{bmatrix}, \tag{2.54}$$

where $F = [0, \ldots, 0, I_m]$, Z is the block shift matrix

$$Z \doteq \begin{bmatrix} 0 & & & \\ I_m & 0 & & \\ & \ddots & \ddots & \\ & & I_m & 0 \end{bmatrix},$$

and

$$Y_0 = \begin{bmatrix} P_0 & \frac{1}{2}P_1 & & \\ \frac{1}{2}P_1 & P_2 & \ddots & \\ & \ddots & \ddots & \frac{1}{2}P_{2n-1} \\ & & \frac{1}{2}P_{2n-1} & P_{2n} \end{bmatrix}.$$

Assume that $\{P_k\}_{k=0}^{2n}$ are Hermitian matrices. Then $\Phi(s)$ is a well-defined para-conjugate transfer function. As

$$\left.\begin{bmatrix} (sI_n - \jmath Z^T)^{-1}\jmath F^T \\ I_m \end{bmatrix}\right|_{s=\jmath x} = x^{-n}\Pi_n(x),$$

one has by construction the relation

$$\Phi(\jmath x) = x^{-2n}\sum_{k=0}^{2n} P_k x^k = x^{-2n}P(x).$$

Therefore, $\Phi(s)$ is a non-negative para-conjugate transfer function if and only if $P(x)$ is a non-negative polynomial matrix. In view of the Kalman-Yakubovich-Popov Lemma, it finally appears that $P(x)$ is non-negative if and only if there exists a Hermitian matrix X such that the Hermitian matrix

$$Y(X) = Y_0 + \begin{bmatrix} \jmath XZ^T - \jmath ZX & \jmath XF^T \\ -\jmath FX & 0 \end{bmatrix} \tag{2.55}$$

is non-negative definite. If one sets $X_0 \doteq -\jmath X$, this is precisely the characterization provided by Theorems 2.7 and 2.8.

Unit circle $\mathbb{T}$

An alternative proof of Theorems 2.16 and 2.17 can be obtained on the basis of a similar argument. Consider a state-space realization of a para-conjugate transfer function of the form

$$\Phi(z) = \begin{bmatrix} zF(I - zE)^{-1} & I_m \end{bmatrix} Y_0 \begin{bmatrix} (zI - E^*)^{-1}F^* \\ I_m \end{bmatrix}$$

with Y_0 some Hermitian matrix. Incidentally, this realization can also be deduced from (2.54) by means of the variable transformation $s = (z-1)/(z+1)$, which maps the unit circle onto the imaginary axis.

The transfer function $\Phi(z)$ is non-negative on the unit circle if the matrix $\Phi(e^{\jmath\theta})$ is non-negative definite definite for all θ in the interval $[0, 2\pi]$. In this setting, Theorem 2.24 states that $\Phi(z)$ is a non-negative para-conjugate transfer function if and only there exits a Hermitian matrix X such that

$$Y(X) = Y_0 + \begin{bmatrix} I \\ 0 \end{bmatrix} X \begin{bmatrix} I & 0 \end{bmatrix} - \begin{bmatrix} E \\ F \end{bmatrix} X \begin{bmatrix} E^* & F^* \end{bmatrix}$$

is non-negative definite. With $E = Z$, $F = [0, \ldots, 0, I_m]$ and

$$Y_0 = \begin{bmatrix} P_0 & P_1^*/2 & \ldots & P_n^*/2 \\ P_1/2 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ P_n/2 & 0 & \ldots & 0 \end{bmatrix},$$

the following equality holds

$$\Phi(z) = \frac{1}{2} \sum_{k=0}^{n} (P_k z^{-k} + P_k^* z^k).$$

Therefore, the pseudo-polynomial matrix $P(z)$ is found to be non-negative definite on the unit circle if and only if there exists a Hermitian matrix X such that the matrix

$$Y(X) = Y_0 + \begin{bmatrix} X - ZXZ^* & -ZXF^* \\ -FXZ^* & -FXF^* \end{bmatrix} \tag{2.56}$$

is non-negative definite. Here again, this is exactly the characterization proposed in Section 2.3 provided one substitutes X for X_0.

2.5.6 Analytic center

Given a closed convex set Q with non-empty interior, there exist many different ways of defining its center. In the context of structural programming, the center of Q is naturally defined as

$$\hat{y} = \arg\min F(y), \tag{2.57}$$

where $F(\cdot)$ is a self-concordant barrier for Q. This central point is called the *analytic center* of Q. Note that it depends on the choice of the barrier function F.

Consider the convex set Q characterized by the linear matrix inequality

$$C - \sum_{i=0}^{k-1} A_i y_i \succeq 0$$

with Hermitian matrices C and $\{A_i\}_{i=0}^{k-1}$ and assume that Q has a non-empty interior. We describe the structure of Q by making use of the self-concordant barrier

$$F(y) = -\log\det(C - \sum_{i=0}^{k-1} A_i y_i).$$

Then the analytic center of Q, say $\hat{y}$, satisfies the first-order optimality conditions

$$\langle (C - \sum_{i=0}^{k-1} A_i \hat{y}_i)^{-1}, A_i \rangle = 0, \quad i = 1, \ldots, k.$$

Clearly, the matrix $C - \sum_{i=0}^{k-1} A_i \hat{y}_i$ is nonsingular as a consequence of (2.57).

In our context, the analytic centers of particular linear matrix inequalities provide us with the dual representations of non-negative interior polynomials.

Proposition 2.25. *Let $P \in \operatorname{int} \mathcal{K}_{\mathbb{R}}$. Then the analytic center $\hat{X}$ of linear matrix inequality* (2.55) *satisfies*

$$\langle Y(\hat{X})^{-1}, \begin{bmatrix} \jmath HZ^T - \jmath ZH & \jmath HF^T \\ -\jmath FH & 0 \end{bmatrix} \rangle = 0, \quad \forall H \in \mathcal{H}^{mn}.$$

Consequently, $Y(\hat{X})^{-1}$ is a block Hankel matrix.

Proof. Indeed, the directional derivatives of $F(X) = -\log\det Y(X)$ with $Y(X)$ defined in (2.55) is equal to

$$DF(X)[H] = -\langle Y(X)^{-1}, \begin{bmatrix} \jmath HZ^T - \jmath ZH & \jmath HF^T \\ -\jmath FH & 0 \end{bmatrix} \rangle.$$

As $\hat{X}$ is an extremal point of $F(X)$ if and only if $DF(\hat{X})[H] = 0$ for all $H \in \mathcal{H}^{mn}$, the desired characterization of the analytic center is obtained. From the identities

$$\langle Y(\hat{X})^{-1}, \begin{bmatrix} \jmath HZ^T & \jmath HF^T \\ 0 & 0 \end{bmatrix} \rangle = \langle Y(\hat{X})^{-1}, \begin{bmatrix} \jmath ZH & 0 \\ \jmath FH & 0 \end{bmatrix} \rangle, \quad \forall H \in \mathcal{H}^{mn}, \tag{2.58}$$

we obtain that $[Y(\hat{X})^{-1}]_{ij} = S_{i+j}$ for some $m \times m$ Hermitian matrices $S_0, \ldots, S_{2n}$. To see this, note that (2.58) is equivalent to

$$\langle Y(\hat{X})^{-1}, \begin{bmatrix} 0 & \jmath H \\ 0 & 0 \end{bmatrix} \rangle = \langle Y(\hat{X})^{-1}, \begin{bmatrix} 0 & 0 \\ \jmath H & 0 \end{bmatrix} \rangle, \quad \forall H \in \mathcal{H}^{mn},$$

and let H run through the canonical basis of $\mathcal{H}^{mn}$. □

Proposition 2.26. *Let $P \in \operatorname{int} \mathcal{K}_{\mathbb{C}}$. Then the analytic center $\hat{X}$ of linear matrix inequality* (2.56) *satisfies*

$$\langle Y(\hat{X})^{-1}, \begin{bmatrix} H - ZHZ^* & -ZHF^* \\ -FHZ^* & -FHF^* \end{bmatrix} \rangle = 0, \quad \forall H \in \mathcal{H}^{mn}.$$

Consequently, $Y(\hat{X})^{-1}$ is a block Toeplitz matrix.

Proof. Indeed, the directional derivatives of $F(X) = -\log\det Y(X)$ with $Y(X)$ defined in (2.56) is equal to

$$DF(X)[H] = -\langle Y(X)^{-1}, \begin{bmatrix} H - ZHZ^* & -ZHF^* \\ -FHZ^* & -FHF^* \end{bmatrix}\rangle.$$

As $\hat{X}$ is an extremal point of $F(X)$ if and only if $DF(\hat{X})[H] = 0$ for all $H \in \mathcal{H}^{mn}$, the desired characterization of the analytic center is obtained. From the identities

$$\langle Y(\hat{X})^{-1}, \begin{bmatrix} H & 0 \\ 0 & 0 \end{bmatrix}\rangle = \langle Y(\hat{X})^{-1}, \begin{bmatrix} ZHZ^* & ZHF^* \\ FHZ^* & FHF^* \end{bmatrix}\rangle, \quad \forall H \in \mathcal{H}^{mn}, \tag{2.59}$$

we obtain that $[Y(\hat{X})^{-1}]_{ij} = S_{i-j}$ for some $m \times m$ Hermitian matrices $S_0, \ldots, S_n$ and $S_{-k} = S_k^*$. To see this, note that (2.59) is equivalent to

$$\langle Y(\hat{X})^{-1}, \begin{bmatrix} H & 0 \\ 0 & 0 \end{bmatrix}\rangle = \langle Y(\hat{X})^{-1}, \begin{bmatrix} 0 & 0 \\ 0 & H \end{bmatrix}\rangle, \quad \forall H \in \mathcal{H}^{mn},$$

and let H run through the canonical basis of $\mathcal{H}^{mn}$. □

From both propositions, we conclude that the analytic centers of (2.55) and (2.56) provide us with the dual representations of $P(x) = \langle\langle Y_0\Pi_n(x), \Pi_n(x)\rangle\rangle$ and $P(z) = \langle\langle Y_0\Pi_n(z), \Pi_n(z)\rangle\rangle$, respectively.

Given a self-adjoint matrix polynomial P, the feasible domain of (2.55) (if the polynomial $P(x) = \langle\langle Y(X)\Pi_n(x), \Pi_n(x)\rangle\rangle$ is defined on the real line) or (2.56) (if the polynomial $P(z) = \langle\langle Y(X)\Pi_n(z), \Pi_n(z)\rangle\rangle$ is defined on the unit circle) can be used to classify P according to its non-negativity:

Empty feasible set $P(x)$ (or $P(z)$) is not non-negative; $P \notin \mathcal{K}_{\mathbb{R}}$ ($P \notin \mathcal{K}_{\mathbb{C}}$).

Feasible set with empty interior $P(x)$ (or $P(z)$) is non-negative, but on the boundary of $\mathcal{K}_{\mathbb{R}}$ (or $\mathcal{K}_{\mathbb{C}}$).

Feasible set with non-empty interior $P(x)$ (or $P(z)$) is positive and belongs to the interior of $\mathcal{K}_{\mathbb{R}}$ (or $\mathcal{K}_{\mathbb{C}}$). Moreover, the analytic center is related to the dual representation of P.

Example 2.7. Consider a quadratic polynomial $p(x) = ax^2 + bx + c$, with $a > 0$. It is non-negative on the real line if and only if $a \geq 0$, and $b^2 - 4ac \leq 0$. In this case, the linear matrix inequality (2.55) is equivalent to

$$Y(X) = \begin{bmatrix} c & b/2 + \jmath X \\ b/2 - \jmath X & a \end{bmatrix} \succeq 0, \quad X \in \mathbb{R}.$$

Note that $\det Y(X) = ac - \frac{b^2}{4} - X^2$. If $p(x)$ has two distinct real roots, then $Y(X)$ is indefinite for all X as $\det Y(X)$ is strictly negative (empty feasible set). If $p(x)$ has one root with multiplicity 2, then $Y(X)$ is positive semidefinite if and only if $X = 0$ (feasible set with empty interior). If $p(x)$ has two distinct complex roots, then $Y(X)$ is positive semidefinite on a symmetric interval. Its analytic center is thus $X = 0$ and corresponds to our standard matrix Y for a quadratic polynomial, see Example 2.5.

2.6 Multivariate polynomials and Hilbert's 17th problem

In Section 2.2, we could have relaxed our implicit assumption that $\Gamma \subseteq \mathbb{R}$ and used any set $\Gamma \subseteq \mathbb{R}^p$. Consequently, multivariate sum-of-squares functions could also be represented using a word-by-word extension of our previous results. Unfortunately, the situation is not as simple as in the univariate case. Indeed, non-negativity is usually not equivalent to representability as sums of squares. Let us start this section with an illustration of this fact.

Example 2.8 (Motzkin [104]). Consider *Motzkin's polynomial*

$$p(x,y) = 1 + x^4y^2 + x^2y^4 - 3x^2y^2.$$

Although this polynomial is non-negative on $\mathbb{R}^2$, it cannot be written as a sum of squares. Note that non-negativity of $p(x,y)$ is a direct consequence of the arithmetic-mean – geometric-mean inequality

$$\frac{1 + x^4y^2 + x^2y^4}{3} \geq (x^6y^6)^{1/3}.$$

In order to check that $p(x,y)$ is not semidefinite representable, we consider the multivariate extension of our previous results, i.e., $\Gamma = \mathbb{R}^2$. Note that the minimal functional basis $\mathcal{S}$ such that p can be generated by the elements of $\mathcal{S}^2$, corresponds to

$$\psi(x,y) = \begin{bmatrix} 1 & x^2y & xy^2 & xy \end{bmatrix}^T.$$

To see this, one proceeds via term-inspection [125]. In this basis, the unique representation of the form $p(x,y) = \psi(x,y)^T Y \psi(x,y)$ is obtained with

$$Y = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -3 \end{bmatrix}.$$

Obviously, this matrix is not positive semidefinite so that $p(x,y)$ is not semidefinite representable.

This celebrated example illustrates the fact that non-negative multivariate polynomials are not necessarily sums of squares. Although we cannot expect to represent an arbitrary cone of non-negative multivariate polynomials using the cone of positive semidefinite matrices, several specific cones are semidefinite representable. Let us give a non-exhaustive list.

Quadratic polynomials The cone of non-negative quadratic polynomials is equivalent to

$$\mathcal{K}_{n+1,2} = \{p : p(x) = x^TAx + b^Tx + c \geq 0, \forall x \in \mathbb{R}^n\}.$$

Given $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}_+$, the polynomial $p(x) = x^TAx + 2b^Tx + c$ is non-negative if and only if the matrix

$$Y = \begin{bmatrix} c & b^T/2 \\ b/2 & A \end{bmatrix},$$

is positive semidefinite. Therefore, we conclude that $\mathcal{K}_{n+1,2}$ is parametrized by $\mathcal{S}_+^{n+1}$. The representation of $p(x)$ as a sum of squares follows from the eigenvalue/eigenvector decomposition of Y. The dual operator is obtained from the decomposition of Y as shown in the above equation. Note that this result has been known for a long time, see e.g. [21].

Bivariate polynomials of degree 4 From Hilbert's results on the representability of non-negative ternary quartic forms as sums of squares [78], we directly obtain that the cone

$$\mathcal{K}_{3,4} = \{p \in \mathbb{R}^{15} : p(x,y) = \sum_{0 \leq i+j \leq 4} p_{ij} x^i y^j \geq 0, \forall x, y \in \mathbb{R}\}$$

is semidefinite representable.

Specific cone of bivariate real polynomials Consider the cone

$$\mathcal{K}_{3,\{2n,2\}} = \{p \in \mathbb{R}^{6n+3} : p(x,y) = \sum_{k=0}^{2n} \sum_{l=0}^{2} p_{k+l(2n+1)} x^k y^l \geq 0; x, y \in \mathbb{R}\}.$$

We claim that this cone is semidefinite representable. Indeed, any polynomial $p \in \mathcal{K}_{3,\{2n,2\}}$ is of the form

$$a(x)y^2 + b(x)y + c(x), \tag{2.60}$$

where $a(x)$, $b(x)$ and $c(x)$ are real polynomials of degree $2n$. From Example 2.5, we conclude that non-negativity of $p(x,y)$ is equivalent to non-negativity of the symmetric matrix polynomial

$$P(x) = \begin{bmatrix} c(x) & b(x)/2 \\ b(x)/2 & a(x) \end{bmatrix} \in \mathcal{S}^2.$$

Consequently, semidefinite representability of $\mathcal{K}_{3,\{2n,2\}}$ follows from Theorem 2.7. Indeed, there exists a one-to-one correspondence between the coefficients of (2.60) and the elements of $P(x)$. This observation directly yields the following corollary.

Corollary 2.27. *Given $s \in \mathbb{R}^{6n+3}$, define the Hankel-block-Hankel matrix $H_{\{2n,2\}}(s)$ by*

$$H_{\{2n,2\}}(s) = \begin{bmatrix} s_0 & \cdots & s_n & s_{2n+1} & \cdots & s_{3n+1} \\ \vdots & \iddots & \vdots & \vdots & \iddots & \vdots \\ s_n & \cdots & s_{2n} & s_{3n+1} & \cdots & s_{4n+1} \\ s_{2n+1} & \cdots & s_{3n+1} & s_{4n+2} & \cdots & s_{5n+2} \\ \vdots & \iddots & \vdots & \vdots & \iddots & \vdots \\ s_{3n+1} & \cdots & s_{4n+1} & s_{5n+2} & \cdots & s_{6n+2} \end{bmatrix}.$$

Then, we obtain

1. *Primal cone:*

$$\begin{aligned} \mathcal{K}_{3,\{2n,2\}} &= \{p \in \mathbb{R}^{6n+3} : p(x,y) = \sum_{k=0}^{2n} \sum_{l=0}^{2} p_{k+l(2n+1)} x^k y^l \geq 0\} \\ &= \{p \in \mathbb{R}^{6n+3} : p = H^*_{\{2n,2\}}(Y), Y \in \mathcal{S}_+^{2(n+1)}\}. \end{aligned}$$

2. *Dual cone:*
$$\mathcal{K}^{\star}_{3,\{2n,2\}} = \{s \in \mathbb{R}^{6n+3} : H_{\{2n,2\}}(s) \succeq 0\}.$$

3. *Both cones are closed, convex and pointed. Consequently, they also have non-empty interiors.*

Note that the functional systems $\psi(x,y)$ and $\upsilon(x,y)$ are given by $\pi_1(y) \otimes \pi_n(x)$ and $\pi_2(y) \otimes \pi_{2n}(x)$, respectively.

Using a conformable transformation, our last example is equivalent to a particular cone of bivariate trigonometric polynomials. Consequently, that cone is also semidefinite representable.

Specific cone of bivariate trigonometric polynomials Consider the cone

$$\mathcal{K}_{3,\{2n,2\}} = \{p \in \mathbb{R} \times \mathbb{C}^{3n+1} : p(z,w) = \operatorname{Re} \sum_{k=0}^{n} \sum_{l=0}^{1} p_{k+l(n+1)} z^k w^l \geq 0; z, w \in \mathbb{T}\}.$$

The cone is semidefinite representable [10, 52]. Let us derive its representation using our previous results. Indeed, any polynomial $p \in \mathcal{K}_{3,\{2n,2\}}$ is of the form

$$a(z) + b(z)\cos\omega + c(z)\sin\omega \tag{2.61}$$

where $w = e^{\jmath\omega}$, $a(z)$, $b(z)$ and $c(z)$ are trigonometric polynomials of degree n. From Example 2.6, we know that non-negativity of $p(z,w)$ is equivalent to non-negativity of the matrix Laurent polynomial

$$P(z) = \begin{bmatrix} a(z)/2 & (b(z) - \jmath c(z))/2 \\ (b(z) + \jmath c(z))/2 & a(z)/2 \end{bmatrix} \in \mathcal{H}^2.$$

Although the semidefinite representability follows from Theorem 2.16, the correspondence between trigonometric polynomials of the form (2.61) and non-negative Laurent polynomials $P(z) \in \mathcal{H}^2$ is not one-to-one. However, $\mathcal{K}_{3,\{2n,2\}}$ is a linear image of $\mathcal{K}_{\mathbb{C}}$ (with $m = 2$) and the associated linear operator is non-singular on the recession cone of $\mathcal{K}_{\mathbb{C}}$. This proves the next corollary.

Corollary 2.28. *Given $s \in \mathbb{R} \times \mathbb{C}^{3n+1}$, define the Toeplitz-block-Toeplitz matrix $T_{\{2n,2\}}(s)$ by*

$$T_{\{2n,2\}}(s) = \begin{bmatrix} s_0 & \cdots & \overline{s}_n & \overline{s}_{n+1} & \cdots & \overline{s}_{2n+1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ s_n & \cdots & s_0 & \overline{s}_{3n+1} & \cdots & \overline{s}_{n+1} \\ s_{n+1} & \cdots & s_{3n+1} & s_0 & \cdots & \overline{s}_n \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ s_{2n+1} & \cdots & s_{n+1} & s_n & \cdots & s_0 \end{bmatrix}$$

Then, we obtain

1. *Primal cone:*

$$\begin{aligned}\mathcal{K}_{3,\{2n,2\}} &= \{p \in \mathbb{R} \times \mathbb{C}^{3n+1} : p(z,w) = \mathrm{Re} \sum_{k=0}^{n} \sum_{l=0}^{1} p_{k+l(n+1)} z^k w^l \geq 0\} \\ &= \{p \in \mathbb{R} \times \mathbb{C}^{3n+1} : p = T^*_{\{2n,2\}}(Y), Y \in \mathcal{H}_+^{2(n+1)}\}.\end{aligned}$$

2. *Dual cone:*

$$\mathcal{K}^\star_{3,\{2n,2\}} = \{s \in \mathbb{R} \times \mathbb{C}^{3n+1} : T_{\{2n,2\}}(s) \succeq 0\}.$$

3. *Both cones are closed, convex and pointed. Consequently, they also have non-empty interiors.*

Note that the functional systems are given by

$$\begin{aligned}\psi(z,w) &= \pi_1(w) \otimes \pi_n(z), \\ \nu(z,w) &= [1, z, \cdots, z^n, w, wz, \cdots, wz^n, w^{-1}z, \cdots, w^{-1}z^n]^T.\end{aligned}$$

As shown in our previous examples, non-negative multivariate polynomials are related to cones of positive semidefinite matrices, to non-negative matrix polynomials or to non-negative forms. Indeed, multivariate polynomials are usually homogenized to get forms

$$\sum_{i_0+\cdots+i_n=d} p_{i_0,\ldots,i_n} x_0^{i_0} \ldots x_n^{i_n}.$$

In our context, the equivalence between polynomials and forms follows from inheritance of non-negativity and sum-of-squares representation under dehomogenization and under homogenization into forms of even degree.

The problem of representing non-negative (or definite) forms dates back to the end of the *19*th century. During his talk at the Second International Congress of Mathematicians in Paris (1900), David Hilbert outlined 23 mathematical problems, which have influenced mathematical research all over the world. As every non-negative forms (or polynomials) cannot be represented as a sum of squares, Hilbert asked whether such forms (or polynomials) can be expressed as a quotient of sums of squares. The original problem quoted below is known as Hilbert's *17*th problem.

> *17. Darstellung definiter Formen durch Quadrate.*
> Definit heißt eine solche ganze rationale Funktion oder Form beliebig vieler Veränderlichen mit reellen Coefficienten, die für keine reellen Werte dieser Veränderlichen negativ ausfällt. Das System aller definiten Funktionen verhält sich invariant gegenüber den Operationen der Addition und der Multiplikation; aber auch der Quotient zweier definiten Funktionen ist - sofern er eine ganze Funktion der Veränderlichen wird - eine definite Form. Das Quadrat einer jeden beliebigen Form ist offenbar stets eine definite Form; da aber, wie ich gezeigt habe [Mathematische Annalen Bd. 32] [77], nicht jede definite Form durch Addition aus Formenquadraten zusammengesetzt werden kann, so entsteht die Frage - die ich für den Fall ternärer Formen in bejahendem Sinne entschieden habe [Acta mathematica Bd. 17] [78] -, ob nicht jede definite Form als Quotient von Summen

> von Formenquadraten dargestellt werden kann. Zugleich ist es für gewisse Fragen hinsichtlich der Möglichkeit gewisser geometrischer Konstruktionen wünschenswert, zu wissen, ob die Coefficienten der bei der Darstellung zu verwendenden Formen stets in demjenigen Rationalitätsbereiche angenommen werden dürfen, der durch die Coefficienten der dargestellten Form gegeben ist [Hilbert, Grundlagen der Geometrie, Leipzig 1899, Kap. VII, insbesondere § 38] [79].

As mentioned above, the problem for ternary forms (bivariate polynomials) was solved by Hilbert himself. Hilbert's *17*th problem was answered in the affirmative by Artin in 1927 [8]. Let us check this result on Motzkin's polynomial.

Example 2.9. The identity

$$1 + x^4y^2 + x^4y^2 - 3x^2y^2 = \frac{x^2y^2(x^2+y^2+1)(x^2+y^2-2)^2 + (x^2-y^2)^2}{(x^2+y^2)^2}$$

holds for all $x, y \in \mathbb{R}$. This representation of Motzkin's polynomial is exactly the one predicted by Hilbert's 1893 Theorem [78] as it is represented as a sum of $2^{3-1} = 4$ squares of rational functions.

Pólya gave an explicit solution to Hilbert's *17*th Problem [120] [71, Sections 2.23 and 2.24] in the context of strictly positive forms f of degree $2n$ (even degree multivariate polynomials). Indeed, he proved that, for sufficiently large $r \in \mathbb{N}$,

$$(\sum_{k=1}^{p} x_k^2)^r f(x_1, \ldots, x_p)$$

is a sum of squares. Although Motzkin's polynomial is not strictly positive, it admits such a representation.

Example 2.10. Let $p(x, y) = 1 + x^4y^2 + x^4y^2 - 3x^2y^2$. Then the identity

$$\begin{aligned}(1 + x^2 + y^2)p(x, y) = (x^2y - y)^2 + (xy^2 - x)^2 + (x^2y^2 - 1)^2 \\ + 0.25(xy^3 - x^3y)^2 + 0.75(xy^3 + x^3y - 2xy)^2\end{aligned}$$

holds for all $x, y \in \mathbb{R}$. This representation was obtained using a semidefinite formulation of the problem, see [119].

The problem of representing non-negative forms and of determining those which are sums of squares is still challenging today. A huge specific literature is available, see e.g. [125]. On the one hand, classes of non-negative forms that are representable as sums of squares can be obtained as linear images of semidefinite cones. Note that our results of Section 2.1 are still relevant in this context. Indeed, we can consider a set Γ belonging to $\mathbb{R}^p, p \geq 1$. On the other hand, Pólya's theorem allows us to deal with positive forms that are not sums of squares. For any given $r \in \mathbb{N}$, we multiply the functional systems by $(1 + \sum_{k=1}^{n} x_k^2)^r$. Then there exists some finite integer r such that a sum-of-squares representation is obtained. Although this kind of procedure is studied by several authors in different contexts, see e.g. [39, 119], this method is not efficient for two reasons: no polynomial bound is known on r and it requires iterating

on r. Therefore, the number of tentative semidefinite representations that we need to compute is not known in advance.

In conclusion, our results can be extended to cones of non-negative multivariate polynomials provided that these cones admit representations as sums of squares. Several examples were given to illustrate that property. Otherwise, we need to rely on an iterative scheme to obtain a rational sum-of-squares representation of an arbitrary positive multivariate polynomial. This procedure is not robust in the sense that different elements of the same cone may require different multiplicative factors and that no efficient bound is available on the degree of these factors.

2.7 Research summary

The parametrization of functional cones representable as sums of squared functions, initially described in [111], has been extended to matrix-valued problems, see Section 2.1. The most relevant examples of these cones, which are linked to non-negative matrix polynomials, have been developed so as to illustrate the theory. In particular, our results allow us to parametrize matrix polynomials that are non-negative on segments. This feature is important in practice because such polynomials are often encountered in practical problems, see Chapter 6.

Note that the primal objects, non-negative polynomials, are often used to model important properties whereas the dual objects, moments, are omnipresent in probability theory and statistics. This dual interaction between polynomials and moments is well known in the literature. Our framework proposes a unified treatment of these objects.

Some of our characterizations can be derived from simple algebraic facts or from a well-known result in Systems and Control, the Kalman-Yakubovich-Popov Lemma. We have pointed out these important facts and we have shown how they can be related with previous concepts.

Finally we have treated cones of sums of squares, which involve multivariate functionals. Although these cones are usually not equivalent to cones of non-negative multivariate polynomials, some particular cones do enjoy this remarkable property. Nevertheless, our characterizations are still required in iterative procedures for generic multivariate positive polynomials.

Notes and references

Non-negative polynomials were studied very actively during the period 1890–1920. The strong link with moment problems, see Section 2.4, was recognized early. This probably explains why many well-known mathematicians (e.g. Hermite, Goursat, Bernstein, Fejér, Markov, Lukács, Fekete, ...) have contributed to the underlying theory. The strong interest for these mathematical objects has continued ever since. Of course, the associated problems have been generalized in different directions: from scalar polynomials to matrix polynomials, from univariate to multivariate polynomials, from existence results to computationally tractable formulas,... Classic references on moment problems include the books of Akhiezer [3], Akhiezer and Kreĭn [2], Kreĭn and Nudel'man [91] and Karlin and Studden [89].

Section 2.5 is based on the paper [56]. An interesting survey paper on Hilbert's *17*th problem is [125].

Chapter 3

Conic optimization problems

Given a convex cone and a linear subspace intersecting this cone, a legitimate goal is to find a point that belongs to the intersection and that minimizes some linear functional. This is exactly the conic formulation of a convex problem, see Section 1.3 and references therein. In this chapter, we provide a complete overview of this problem class in the context of non-negative polynomial cones.

After formulating the optimization problems of interest in Section 3.1, several generic methods are investigated in Section 3.2. However, efficient algorithms must rely on the structure of the dual problem, as emphasized in Section 3.3. An alternative approach based on separating hyperplanes is considered in Section 3.4. Numerical results are provided in Section 3.5.

3.1 Conic formulation

Remember that conic optimization problems on a closed convex cone $\mathcal{K}$ have the generic structure

$$\begin{array}{ll} \min & \langle C, P \rangle \\ \text{s.t.} & \langle A_i, P \rangle = b_i, \quad i = 0, \ldots, k-1, \ . \\ & P \in \mathcal{K}. \end{array} \tag{3.1}$$

In this chapter, we focus on problems where $\mathcal{K}$ is a cone of non-negative matrix polynomials. Note that the corresponding vector space and inner products have been described in Chapter 2. As the constraint $P \in \mathcal{K}$ is a set of inequality constraints on the eigenvalues of $P(\cdot)$ to be enforced at an infinite number of points, Problem (3.1) is a *semi-infinite optimization problem*. In the sequel, only the cones $\mathcal{K}_{\mathbb{R}}$ and $\mathcal{K}_{\mathbb{C}}$ are considered. However, our methodology can be applied to other polynomial cones (e.g. polynomials non-negative on intervals). As usual, the subscript $\mathbb{R}$ or $\mathbb{C}$ is dropped when the corresponding statement is valid for both $\mathcal{K}_{\mathbb{R}}$ and $\mathcal{K}_{\mathbb{C}}$.

Remark. For reasons explained in Section 1.3, primal-dual strict feasibility is assumed throughout this chapter.

3.1.1 Real line $\mathbb{R}$

Consider the cone of (self-adjoint) matrix polynomials (of degree $2n$) that are non-negative on the real line

$$\mathcal{K}_{\mathbb{R}} = \{P \in E : P(x) = \sum_{k=0}^{2n} P_k x^k \succeq 0, \forall x \in \mathbb{R}\},$$

where the space of coefficients is

$$E = \{P : P = \begin{bmatrix} P_0, & P_1, & \dots, & P_{2n} \end{bmatrix}; P_k = P_k^* \in \mathbb{C}^{m \times m}\}.$$

The inner-product on E is defined by

$$\langle \cdot, \cdot \rangle_{\mathbb{R}} : E \times E \to \mathbb{R} : X, Y \to \langle X, Y \rangle_{\mathbb{R}} = \operatorname{Re} \operatorname{Trace} \sum_{i=0}^{2n} Y_i^* X_i.$$

Problem (3.1) can be rewritten in the primal form

$$\begin{array}{ll} \min & \langle C, P \rangle \\ \text{s.t.} & \langle A_i, P \rangle = b_i, \quad i = 0, \dots, k-1, \\ & P(x) \succeq 0, \qquad \forall x \in \mathbb{R} \end{array} \tag{3.2}$$

or, by making use of Theorem 2.7, in the dual form

$$\begin{array}{ll} \max & \langle b, y \rangle \\ \text{s.t.} & S = C - \sum_{i=0}^{k-1} y_i A_i, \\ & H(S) \succeq 0. \end{array} \tag{3.3}$$

Remember that $H(S)$ is a block Hankel matrix, see (2.14).

As $P \in \mathcal{K}_{\mathbb{R}}$ necessarily implies that the matrix coefficients P_k are Hermitian for all k, we are not restricted to assuming that all the $m \times m$ blocks C_k of C and blocks $A_{i,j}$ of A_i are Hermitian as well. Clearly, the anti-Hermitian part of these matrices would disappear anyway in the inner products.

From a numerical point of view, the dual formulation (3.3) has a considerable advantage over the primal form (3.2). Indeed, it involves an optimization scheme in a space of dimension k rather than $(2n+1)m^2$.

3.1.2 Unit circle

Consider the cone of trigonometric matrix polynomials (of degree n) that are non-negative on the unit circle

$$\mathcal{K}_{\mathbb{C}} = \{P \in E : P(z) = \frac{1}{2} \sum_{k=0}^{n} (P_k z^{-k} + P_k^* z^k) \succeq 0, \forall z \in \mathbb{T}\},$$

where the space of coefficients is

$$E = \{P = \begin{bmatrix} P_0, & \dots, & P_n \end{bmatrix} : P_0 = P_0^*, P_k \in \mathbb{C}^{m \times m}, \forall k\}.$$

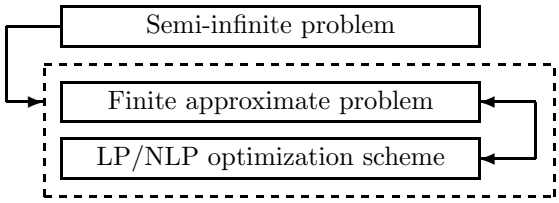

Figure 3.1. *Generic semi-infinite optimization scheme*

The inner product on E is defined by

$$\langle \cdot, \cdot \rangle_{\mathbb{R}} : E \times E \to \mathbb{R} : X, Y \to \langle X, Y \rangle_{\mathbb{R}} = \operatorname{Re} \operatorname{Trace}(\frac{1}{2} \sum_{i=0}^{n} (Y_i^* X_i + Y_i X_i^*)).$$

As in the real line setting, problem (3.1) can be rewritten in the primal form

$$\begin{array}{ll} \min & \langle C, P \rangle \\ \text{s.t.} & \langle A_i, P \rangle = b_i, \quad i = 0, \ldots, k-1, \\ & P(z) \succeq 0, \qquad \forall z \in \mathbb{T}, \end{array} \tag{3.4}$$

or in the dual form

$$\begin{array}{ll} \max & \langle b, y \rangle \\ \text{s.t.} & S = C - \sum_{i=0}^{k-1} y_i A_i, \\ & T(S) \succeq 0. \end{array} \tag{3.5}$$

Remember that $T(S)$ is a block Toeplitz matrix, see (2.28).

As before, we are not restricted to assuming that the $m \times m$ blocks C_k of C and $m \times m$ blocks $A_{i,j}$ of A_i have the same type of symmetry as the blocks of P, since this do not affect the scalar products. The dual formulation (3.5) has also a considerable advantage over the primal form (3.4) because of the space dimensions.

3.2 Generic algorithms

In order to solve problems (3.2) and (3.4), several numerical schemes are available. In this section, we review two approaches which do not rely on the problem structure. The first technique is not restricted to convex optimization problems whereas the second one relies on the semidefinite representability of non-negative polynomial cones.

3.2.1 Semi-infinite programming

Since the constraint $P \in \mathcal{K}$ is semi-infinite, the first idea would be to run algorithms developed for semi-infinite programming. In this context, the semi-infinite problem is approximated by a finite problem, which is solved by a standard linear or nonlinear algorithm. If the approximate solution is not sufficiently accurate, then the finite problem is tuned using some auxiliary process. This procedure is iterated until a good solution is obtained, see Figure 3.1.

Three classes of numerical methods are often used to get finite approximate problems, see e.g. [74, Section 7]. Let us briefly describe them.

Exchange methods The constraint $P \in \mathcal{K}$ is enforced at a finite number of points. At each iteration, a number of new constraints are added and some of the old constraints are deleted. From a conceptual point of view, an exchange of constraints is performed. Probably the most famous exchange algorithm in approximation theory is *Remez algorithm* [124], which solves the linear Chebyshev approximation problem.

Discretization methods The constraint $P \in \mathcal{K}$ is enforced on a grid, which is refined at each iteration using the information collected so far.

Local reduction methods Given a candidate P, all local minima which violate the constraint $P \in \mathcal{K}$ are computed. The reduced optimization problem, which is obtained from these points, allows us to obtain the next candidate.

The survey paper [74] contains a complete overview of semi-infinite programming, as well as numerous references to the literature.

Because the semi-infinite approach is general, it does not take advantage of the sum-of-squares representation of $P \in \mathcal{K}$; the non-negativity constraint is only enforced at a finite number of points and the optimal solution may slightly violate this non-negativity requirement. Moreover, the auxiliary process, which solves approximate problems, might be computationally demanding. For these reasons, we favour solving the semidefinite formulation of problems (3.2) and (3.4).

3.2.2 Semidefinite programming

Because the cone $\mathcal{K}$ is semidefinite representable, see our representation theorems in Chapter 2, the constraint $P \in \mathcal{K}$ can be treated exactly. Indeed, it is equivalent to $P = \Lambda^*(Y)$ for an appropriate linear operator Λ and some positive semidefinite matrix Y. Consequently, we get the semidefinite formulations of (3.2) and (3.4).

Real line The operator Λ is block Hankel so that we get

$$\begin{array}{ll} \min & \langle H(C), Y \rangle \\ \text{s.t.} & \langle H(A_i), Y \rangle = b_i, \quad i = 0, \ldots, k-1, \\ & Y \in \mathcal{H}_+^{m(n+1)}. \end{array} \tag{3.6}$$

Unit circle The operator Λ is block Toeplitz so that we get

$$\begin{array}{ll} \min & \langle T(C), Y \rangle \\ \text{s.t.} & \langle T(A_i), Y \rangle = b_i, \quad i = 0, \ldots, k-1, \\ & Y \in \mathcal{H}_+^{m(n+1)}. \end{array} \tag{3.7}$$

Since the matrix Y is Hermitian, the number of variables in these primal formulations is equal to $m^2(n+1)^2$. The price to pay for treating the non-negativity constraint $P \in \mathcal{K}$ exactly is a huge increase in the number of (primal) variables. Therefore, generic primal or primal-dual semidefinite solvers are very slow for high dimensions. Indeed, each iteration involves solving an oversized linear system.

It can be shown [136, Table 1] that the total complexity of solving the SDP (3.6) and (3.7), up to an accuracy ϵ, by making use of a path-following primal-dual solver is

$$\mathcal{O}((km^{3.5}n^{3.5} + k^2m^{2.5}n^{2.5} + k^3m^{0.5}n^{0.5}) \log \frac{1}{\epsilon}). \tag{3.8}$$

The semidefinite approach is the best approach because the non-negativity constraint is enforced from the beginning. Moreover, interior-point methods find an ϵ-solution in polynomial time. However, tailor-made algorithms which exploit the underlying problem structure must be devised in order to avoid computationally expensive schemes. Such a design is the goal of the next section.

3.3 Fast dual solver

In general, any dual optimization problem can be solved efficiently with the help of interior point methods, provided that a *computable* self-concordant barrier for the dual cone is known. Indeed, their numerical implementation requires the calculation of the first and second derivatives of the barrier function. For our dual problems (3.3) and (3.5), these objects are defined as follows.

Real line The self-concordant barrier function is

$$f(u) = -\ln\det H(C - \sum_{\ell=0}^{k-1} A_\ell u_\ell). \tag{3.9}$$

The first two derivatives of f can be expressed as follows:

$$\frac{\partial f(u)}{\partial u_\ell} = \langle H(S)^{-1}, H(A_\ell)\rangle, \tag{3.10a}$$

$$\frac{\partial^2 f(u)}{\partial u_\ell \partial u_s} = \langle H(S)^{-1} H(A_\ell) H(S)^{-1}, H(A_s)\rangle, \tag{3.10b}$$

where $S = C - \sum_{\ell=0}^{k-1} A_\ell u_\ell$.

Unit circle The self-concordant barrier function is

$$f(u) = -\ln\det T(C - \sum_{\ell=0}^{k-1} A_\ell u_\ell). \tag{3.11}$$

As in the block Hankel case, its derivatives can be expressed as follows:

$$\frac{\partial f(u)}{\partial u_\ell} = \langle T(S)^{-1}, T(A_\ell)\rangle, \tag{3.12a}$$

$$\frac{\partial^2 f(u)}{\partial u_\ell \partial u_s} = \langle T(S)^{-1} T(A_\ell) T(S)^{-1}, T(A_s)\rangle, \tag{3.12b}$$

where $S = C - \sum_{\ell=0}^{k-1} A_\ell u_\ell$.

Efficient numerical schemes to solve conic optimization problems require repeated calculations of the differential characteristics of the barrier function, i.e., the gradient $\partial f(u)/\partial u_\ell$ and the Hessian $\partial^2 f(u)/\partial u_\ell \partial u_s$. The block Toeplitz or block Hankel structure underlying the optimization space allows us to carry out these computations in a fast, and even superfast, manner. The aim of this section is to explain this procedure in detail.

3.3.1 Fast algorithms for matrices with structure

Displacement structure theory

Let us first consider Hermitian $(n+1)\times(n+1)$ block Toeplitz matrices with arbitrary $m\times m$ matrix blocks T_i:

$$T \doteq \begin{bmatrix} T_0 & T_1^* & \cdots & T_n^* \\ T_1 & T_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & T_1^* \\ T_n & \cdots & T_1 & T_0 \end{bmatrix},$$

and $(n+1)\times(n+1)$ block Hankel matrices with Hermitian $m\times m$ matrix blocks H_i:

$$H \doteq \begin{bmatrix} H_0 & H_1 & \cdots & H_n \\ H_1 & \iddots & \iddots & \vdots \\ \vdots & \iddots & \iddots & H_{2n-1} \\ H_n & \cdots & H_{2n-1} & H_{2n} \end{bmatrix}.$$

Note that T and H are defined by $(2n+1)m^2$ parameters.

Remember that

$$Z \doteq \begin{bmatrix} 0 & & & \\ I_m & 0 & & \\ & \ddots & \ddots & \\ & & I_m & 0 \end{bmatrix} \tag{3.13}$$

is the block *shift matrix*. Let us also set the block *permutation matrix* J

$$J \doteq \begin{bmatrix} 0 & \cdots & 0 & I_m \\ \vdots & \iddots & \iddots & 0 \\ 0 & \iddots & \iddots & \vdots \\ I_m & 0 & \cdots & 0 \end{bmatrix}, \tag{3.14}$$

that will play a special role in the subsequent developments.

The displacement theory of Toeplitz and Hankel matrices is well established [85, 86] and is the basis underlying most fast algorithms for decomposing such matrices. Using the block shift matrix, one defines a *Toeplitz displacement operator* ∇_t and a *Hankel displacement operator* ∇_h as follows:

$$\nabla_t T \doteq T - ZTZ^T, \quad \nabla_h H \doteq H - Z^T H Z^T. \tag{3.15}$$

The reader may easily check that the following equalities hold

$$\nabla_t T = \begin{bmatrix} T_0 & T_1^* & \cdots & T_n^* \\ T_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ T_n & 0 & \cdots & 0 \end{bmatrix}, \tag{3.16}$$

$$\nabla_h H = \begin{bmatrix} H_0 & 0 & \cdots & 0 \\ H_1 & \vdots & \ddots & \vdots \\ \vdots & 0 & \cdots & 0 \\ H_n & \cdots & H_{2n-1} & H_{2n} \end{bmatrix}. \tag{3.17}$$

From the above expressions, we notice that the original matrices T and H can be recovered from their respective displacement. The inverse operators are obtained by merely applying the displacement operator again and again to both sides of (3.15) to produce:

$$T = \nabla_t T + Z \cdot \nabla_t T \cdot Z^T + \cdots + Z^n \cdot \nabla_t T \cdot Z^{nT} \tag{3.18}$$

and

$$H = \nabla_h H + Z^T \cdot \nabla_h H \cdot Z^T + \cdots + Z^{nT} \cdot \nabla_h H \cdot Z^{nT}. \tag{3.19}$$

It is also useful to point out that both displacements are closely related to each other. Permuting the block rows of a block Hankel matrix H indeed yields a block Toeplitz matrix JH, which can be defined as T by setting $T_i = H_{n-i},\ i = -n, \ldots, n$. Since $Z^T = JZJ$ the displacement operators are related in a similar fashion as follows:

$$T = JH \iff \nabla_t T = J \nabla_h H.$$

From the sparsity structure of matrices (3.16) and (3.17), it is obvious that the ranks of $\nabla_t T$ and $\nabla_h H$ cannot be larger than $2m$. This rank is called the *displacement rank* of the corresponding matrix.

Generalized Schur algorithm

Since the displacement rank of a block Toeplitz or block Hankel matrix is typically much lower than the dimensions of the corresponding matrix, and since the displacement operator can be inverted, it is economical to represent such a matrix by a rank factorization of its displacement.

From the expressions (3.16) and (3.17), it is simple to construct low rank factorizations of $\nabla_t T$ or $\nabla_h H$ as follows :

$$\nabla_t T = F_t \cdot G_t^*, \quad \nabla_h H = F_h \cdot G_h^*, \tag{3.20}$$

where the number of columns of F_t and G_t equals $r_t \doteq \operatorname{rank} \nabla_t T$ and the number of columns of F_h and G_h equals $r_h \doteq \operatorname{rank} \nabla_h H$. The pairs of matrices $\{F_t, G_t\}$ and $\{F_h, G_h\}$ are called the *generators* of T and H, respectively.

The theory of displacement ranks [85, 86] tells us that the inverse of T or H (when it exists) has the same displacement as that of the matrix itself as follows:

$$\operatorname{rank} \nabla_t^\star T^{-1} = \operatorname{rank} \nabla_t T, \quad \operatorname{rank} \nabla_h H^{-1} = \operatorname{rank} \nabla_h H,$$

where $\nabla_t^\star$ stands for the transposed Toeplitz displacement operator, i.e., $\nabla_t^\star T^{-1} = T^{-1} - Z^T T^{-1} Z$.

Given such factorizations, fast generalized Schur based algorithms can be used to derive from them the corresponding factorizations of the displacement of the inverses:

$$\nabla_t^\star T^{-1} = A_t \cdot B_t^*, \quad \nabla_h H^{-1} = A_h \cdot B_h^*$$

and these precise decompositions are used in what follows. This feature originates from an important result of displacement structure theory: Schur complements of a structured matrix inherit its displacement structure and generalized Schur based algorithms yield generator matrices for these Schur complements. These algorithms can be implemented in a fast manner with $\mathcal{O}(rm^2n^2)$ flops, see e.g. [86, Algorithm 3.3.1]. This complexity can be reduced to $\mathcal{O}(rm^2 n \log^2 n)$ flops by means of a divide-and-conquer strategy, which produces superfast algorithms, see e.g. [55, 141].

Incidentally, let us note that these factorizations (3.20) are not unique and that, for positive definite matrices T and H, there exist particular choices of factorizations that can benefit from these properties. For instance, one can choose in the Toeplitz case

$$F_t = \begin{bmatrix} T_0 & 0 \\ T_1 & -T_1 \\ T_2 & -T_2 \\ \vdots & \vdots \\ T_n & -T_n \end{bmatrix}, \quad G_t = F_t \begin{bmatrix} T_0 & 0 \\ 0 & -T_0 \end{bmatrix}^{-1}.$$

In what follows, these aspects will be disregarded since they only marginally affect the complexity results.

Displacement-based factorization

Let us focus first on the case of Toeplitz displacement of an $m(n+1) \times m(n+1)$ matrix X and suppose that a rank r_t factorization of its Toeplitz displacement $\nabla_t X$ has been computed

$$\nabla_t X = F \cdot G^*,$$

where F and G have dimensions $m(n+1) \times r_t$.

Let us also define a lower block triangular Toeplitz matrix $L(F)$ as a function of the partitioned matrix F, where each sub-block has dimensions $m \times r_t$

$$F \doteq \begin{bmatrix} F_0 \\ F_1 \\ \vdots \\ F_n \end{bmatrix}, \quad L(F) \doteq \begin{bmatrix} F_0 & 0 & \cdots & 0 \\ F_1 & F_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ F_n & \cdots & F_1 & F_0 \end{bmatrix}.$$

Doing the same for the matrix G, one obtains

$$G \doteq \begin{bmatrix} G_0 \\ G_1 \\ \vdots \\ G_n \end{bmatrix}, \quad L(G)^* \doteq \begin{bmatrix} G_0^* & G_1^* & \cdots & G_n^* \\ 0 & G_0^* & \ddots & \vdots \\ \vdots & \ddots & \ddots & G_1^* \\ 0 & \cdots & 0 & G_0^* \end{bmatrix}.$$

It follows from the displacement equation $\nabla_t X = F \cdot G^*$ that

$$X = \sum_{j=0}^{n} (Z^j F)(Z^j G)^* = L(F)L(G)^* \tag{3.21a}$$

$$= \begin{bmatrix} F_0 & 0 & \cdots & 0 \\ F_1 & F_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ F_n & \cdots & F_1 & F_0 \end{bmatrix} \cdot \begin{bmatrix} G_0^* & G_1^* & \cdots & G_n^* \\ 0 & G_0^* & \ddots & \vdots \\ \vdots & \ddots & \ddots & G_1^* \\ 0 & \cdots & 0 & G_0^* \end{bmatrix}. \tag{3.21b}$$

This formula, when applied to a particular choice of displacement factors F and G for the inverse of a Toeplitz matrix T, is also known as the *Gohberg-Semencul formula* for $X = T^{-1}$.

For the Hankel displacement $\nabla_h X$ of an $m(n+1) \times m(n+1)$ matrix X, there exists a similar representation starting based upon the rank r_h factorization of $\nabla_h X$

$$\nabla_h X = F \cdot G^*,$$

where F and G have dimension $m(n+1) \times r_h$. If the matrix F is partitioned in the reversed order

$$F \doteq \begin{bmatrix} F_0 \\ F_1 \\ \vdots \\ F_n \end{bmatrix} \iff JF \doteq \begin{bmatrix} F_n \\ F_{n-1} \\ \vdots \\ F_0 \end{bmatrix},$$

then it follows from the relation $J\nabla_h X = \nabla_t(JX)$ that

$$X = J \sum_{j=0}^{n} (Z^j JF)(Z^j G)^* = JL(JF)L(G)^* \tag{3.22a}$$

$$= \begin{bmatrix} F_0 & \cdots & F_{n-1} & F_n \\ \vdots & \iddots & \iddots & 0 \\ F_{n-1} & F_n & \iddots & \vdots \\ F_n & 0 & \cdots & 0 \end{bmatrix} \cdot \begin{bmatrix} G_0^* & G_1^* & \cdots & G_n^* \\ 0 & G_0^* & \ddots & \vdots \\ \vdots & \ddots & \ddots & G_1^* \\ 0 & \cdots & 0 & G_0^* \end{bmatrix}. \tag{3.22b}$$

When applied to a particular choice of displacement factors F and G for the inverse of a Hankel matrix, this formula is also known as the *Christoffel-Darboux formula* for $X = H^{-1}$.

3.3.2 Gradient and Hessian computation

The numerical solution of the optimization problem considered in Section 3.1 requires evaluations of the gradient $\partial f(u)/\partial u_\ell$ and the Hessian $\partial^2 f(u)/\partial u_\ell \partial u_s$ as given by (3.10) or (3.12). Let us now focus on the fast computation of these elements using the displacement techniques mentioned above.

Structure of inner products

Consider the inner product $\langle X, T(A_s)\rangle$ that appears in (3.12) with

$$X = T(S)^{-1} \quad \text{or} \quad X = T(S)^{-1}T(A_l)T(S)^{-1}.$$

Since

$$T(A_s) = I_{n+1} \otimes A_{s,0} + \sum_{k=0}^{n}[Z^k(I_{n+1} \otimes A_{s,k}) + (Z^k)^T(I_{n+1} \otimes A_{s,k}^*)],$$

the computation can be broken down into a summation of scalar products of the type

$$\langle X, Z^k(I_{n+1} \otimes W)\rangle, \quad \langle X, (Z^k)^T(I_{n+1} \otimes W^*)\rangle. \tag{3.23}$$

For Hermitian matrices X, it turns out that

$$\langle X, Z^k(I_{n+1} \otimes W)\rangle = \langle X, (Z^k)^T(I_{n+1} \otimes W^*)\rangle$$

so that only one expression has to be evaluated. Indeed, the following identities hold

$$\begin{aligned}\langle X, Z^k(I_{n+1} \otimes W)\rangle &= \sum_{i-j=k}\langle X_{ij}, W\rangle = \sum_{i-j=k}\langle X_{ij}^*, W^*\rangle \\ &= \sum_{i-j=k}\langle X_{ji}, W^*\rangle = \langle X, (Z^k)^T(I_{n+1} \otimes W^*)\rangle.\end{aligned}$$

Similarly, the inner product $\langle X, H(A_s)\rangle$ that appears in (3.10) with $X = H(S)^{-1}$ or $X = H(S)^{-1}H(A_l)H(S)^{-1}$ requires the evaluation of scalar products of the type

$$\langle X, JZ^k(I_{n+1} \otimes W_1)\rangle, \quad \langle X, J(Z^k)^T(I_{n+1} \otimes W_2)\rangle, \tag{3.24}$$

where W_1 and W_2 are Hermitian matrices of order m.

Gradient

In addition, since the matrices X can be described by their Hankel or Toeplitz displacement, one can speed up the computation of (3.23) and (3.24). Let us first consider matrices X given by their Toeplitz displacement $\nabla_t X = F \cdot G^*$. Since

$$L(F) = \sum_{k=0}^{n} Z^k(I_{n+1} \otimes F_k), \quad L(G) = \sum_{k=0}^{n} Z^k(I_{n+1} \otimes G_k)$$

and, as

$$\langle Z^j(I_{n+1} \otimes X), Z^i(I_{n+1} \otimes Y)\rangle = \delta_{i,j}(n+1-i)\langle X, Y\rangle,$$

one obtains the expression

$$\begin{aligned}&\langle L(F)L(G)^*, Z^j(I_{n+1} \otimes W)\rangle = \\ &\qquad \langle (n+1-j)F_jG_0^* + \cdots + 2F_{n-1}G_{n-j-1}^* + F_nG_{n-j}^*, W\rangle \doteq \langle M_j, W\rangle.\end{aligned}$$

Since the matrix $X = L(F)L(G)^*$ is Hermitian, the roles of the F_i and G_i can be interchanged in the above formula. Moreover, the quantities $\{M_j\}_{j=0}^n$ can be evaluated as the convolution of the block vectors

$$\begin{bmatrix}(n+1)F_0, & nF_1, & \ldots, & 2F_{n-1}, & F_n\end{bmatrix} \quad , \quad \begin{bmatrix} G_0^* \\ G_1^* \\ \vdots \\ G_{n-1}^* \\ G_n^* \end{bmatrix}.$$

Using Discrete Fourier Transform, this convolution has a complexity of $\mathcal{O}(r_t m^2 n \log_2 n)$ flops, see [86, Algorithm 8.3.1].

As the computation of the inner product $\langle M_j, W\rangle$ requires $\mathcal{O}(m^2)$ operations, the overall complexity of computing $\langle X, T(A_s)\rangle$ is found to be equal to $\mathcal{O}(r_t m^2 n \log_2 n + m^2 n)$ flops for a matrix of displacement rank r_t, *provided that* the matrices F and G are given.

If the matrix X is given by its transposed displacement $\nabla_t^\star X = A^* \cdot B$, one can easily adapt the above formula and check that the overall complexity is also equal to $\mathcal{O}(r_t m^2 n \log_2 n + m^2 n)$ flops, *provided that* the matrices A and B are given.

Hessian

The calculations involving the Hessian, i.e., when $X = T(S)^{-1}T(A_\ell)T(S)^{-1}$, require some elaboration. With the matrix $\tilde{T}$ defined by

$$\tilde{T} = \begin{bmatrix} -T(A_\ell) & T(S) \\ T(S) & 0 \end{bmatrix},$$

note first that the following relation holds:

$$\tilde{T}^{-1} = \begin{bmatrix} 0 & T(S)^{-1} \\ T(S)^{-1} & X \end{bmatrix}.$$

Furthermore, as $T(S)$ and $T(A_\ell)$ are block Toeplitz matrices, the rank of the matrix factors F and G in the block displacement equation

$$\nabla_t \tilde{T} = \tilde{T} - \begin{bmatrix} Z & 0 \\ 0 & Z \end{bmatrix} \tilde{T} \begin{bmatrix} Z^T & 0 \\ 0 & Z^T \end{bmatrix} = F \cdot G^*$$

is equal to $4m$, as is easily verified. The corresponding factorization of the block displacement of the inverse can be achieved at low computational cost in the form

$$\nabla_t^\star \tilde{T}^{-1} = \tilde{T}^{-1} - \begin{bmatrix} Z^T & 0 \\ 0 & Z^T \end{bmatrix} \tilde{T}^{-1} \begin{bmatrix} Z & 0 \\ 0 & Z \end{bmatrix} = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \cdot \begin{bmatrix} B_1^* & B_2^* \end{bmatrix}.$$

Therefore, the expression of the transposed Toeplitz displacement of X is given by:

$$\nabla_t^\star X = A_2 \cdot B_2^*.$$

The formalism described above for the fast computation of the relevant inner products can thus be applied to construct the entries of the Hessian (3.12). If the displacement factors are computed using a superfast algorithm, then the overall complexity of constructing the Hessian is equal to $\mathcal{O}(k r_t m^2 n \log^2 n + k^2 m^2 n)$ flops.

Let us now consider matrices X given by their Hankel displacement $\nabla_h X = F \cdot G^*$. The inner products of interest can be rewritten in terms of JX as follows:

$$\langle X, JZ^i(I_{n+1} \otimes W_1)\rangle = \langle JX, Z^i(I_{n+1} \otimes W_1)\rangle,$$
$$\langle X, J(Z^i)^T(I_{n+1} \otimes W_2)\rangle = \langle (JX)^*, Z^i(I_{n+1} \otimes W_2)\rangle.$$

where W_1 and W_2 are Hermitian matrices of order m. Since JX is block Toeplitz, the above formulas could, in theory, be applied *mutatis mutandis*. From a practical point of view, however, this does not make much sense. As explained later in this section, the Hankel setting of the optimization problem considered is numerically ill-conditioned. Hence, the problem formulation itself needs to be redesigned so as to circumvent this inherent difficulty. This issue is addressed in Subsection 3.3.4.

3.3.3 Complexity of the optimization scheme

The actual solution of the dual optimization problem of Section 3.1 is often achieved with the help of a path-following scheme, see Section 1.3. In particular, this iterative process requires frequent evaluations of the so-called Newton directions, which involve the product of the inverse of the current Hessian by an appropriate given vector.

As the barrier parameter of (3.9) or (3.11) is equal to mn, interior-points methods require $\mathcal{O}(\sqrt{nm}\log\frac{1}{\epsilon})$ Newton steps to solve the optimization problems (3.3) or (3.5) up to an accuracy ϵ, see Section 1.3. Furthermore, the cost of a Newton step is dominated by the cost of evaluating the Hessian. The overall complexity of solving these problems is thus found to be

$$\mathcal{O}((kr_t m^{2.5} n^{1.5} \log^2 n + k^2 m^{2.5} n^{1.5} + k^3 m^{0.5} n^{0.5}) \log \frac{1}{\epsilon})$$

flops, provided that the superfast procedure outlined above is applied.

Remark. In particular, for non-negative *scalar* polynomials, i.e., $m = 1$, each Newton iteration requires $\mathcal{O}(kn(\log^2 n + k) + k^3)$ flops.

By solving the dual problem and using the matrix structure, we get a remarkable result for solving an optimization problem in a $(2n+1)m^2$-dimensional vector space, subject to k linear constraints and m semi-infinite inequality constraints, see (3.2) and (3.4). Indeed, a straightforward evaluation of the Hessian, which does not take advantage of the Hankel or Toeplitz structure, would cost $\mathcal{O}(km^3n^3 + k^2m^2n^2)$ flops. Clearly, the fast procedure described above reduces this complexity by a factor n. We point out that efficient algorithms must necessarily rely on the structure of the inner products of interest, which allow us to assemble the Hessian at a lower computational cost.

The Newton step $\Delta y = y_+ - y$ performed by interior-point schemes updates the vector S of coefficients accordingly, $\Delta S = S_+ - S$. If the matrix $T(S)^{-1}$ is a good approximation of the matrix $T(S_+)^{-1}$, e.g. the Newton step is sufficiently small, the generators of $T(S_+)^{-1}$ can be obtained from the generators of $T(S)^{-1}$ via a low-cost Newton-based iterative scheme. However, as the matrix $T(S)$ often converges to a singular matrix, the convergence condition, which involves the condition number of $T(S)$, is increasingly difficult to satisfy [86, 117].

The computation of the product of the Hessian by a vector can also be done at low cost with the help of the inner product formalism explained in the present section.

Let us briefly clarify this issue. Assume that the optimization problem is defined on the unit circle and consider the product of the Hessian by a vector x to yield a vector y. By definition , one has in view of (3.12) that the sth component of y is given by

$$\begin{aligned} y_s &= \sum_{\ell=0}^{k-1} \frac{\partial^2 f(u)}{\partial u_\ell \partial u_s} x_\ell \\ &= \sum_{\ell=0}^{k-1} \langle T(S)^{-1} T(A_\ell) T(S)^{-1}, T(A_s) \rangle x_\ell \\ &= \langle T(S)^{-1} T(A_s) T(S)^{-1}, T(D) \rangle, \end{aligned}$$

where $T(D)$ stands for the block Toeplitz matrix defined by $D = \sum_{\ell=0}^{k-1} A_\ell x_\ell$. The vector y can thus be evaluated in $\mathcal{O}(kr_t m^2 n \log^2 n + km^2 n)$ flops. For instance, this kind of matrix-vector product would appear in conjugate gradient steps.

3.3.4 Functional basis and numerical behaviour

The formulation of the real line optimization problem exhibits a serious drawback: it involves positive definite Hankel matrices, which are numerically ill conditioned [13, 139]. The celebrated Hilbert matrix is a good illustration of this fact. More generally, the Euclidean condition number $\kappa(H) = \|H\|_2 \left\|H^{-1}\right\|_2$ of any positive definite Hankel matrix H of order $n+1$ was shown recently [13, Theorem 3.6] to be bounded from below by

$$\kappa(H) \geq \frac{(3.210)^n}{16(n+1)}, \quad n \geq 2.$$

Therefore, solving the real line optimization problem as considered previously is inherently hazardous, and all the more so if the problem dimension is large.

To get around this, let us first observe that the occurrence of the block Hankel structure originates from the choice of the natural powers

$$1,\ x,\ x^2, \ldots$$

as a functional basis for describing the optimization space of the polynomial matrices $P(x) = \sum_{k=0}^{2n} P_k x^k$, positive semi-definite on the real line. Obviously, other choices are possible. In this section, the alternative use of a basis of Chebyshev polynomials to describe the primal optimization space is specifically investigated together with the consequences of this choice.

The first-order Chebyshev polynomials $T_k(x)$ are well known to satisfy, for $k \geq 1$, the recurrence formula

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$$

initialized with $T_0(x) = 1$ and $T_1(x) = x$. In particular, one has the relation

$$T_i(x)T_j(x) = \frac{1}{2}[T_{i+j}(x) + T_{|i-j|}(x)], \qquad \forall\, i, j \geq 0. \tag{3.25}$$

Let us denote by $\tilde{P}_k = \tilde{P}_k^*$ the matrix coordinates of any polynomial matrix $P(x)$ in our Chebyshev basis, i.e.,

$$P(x) = \sum_{k=0}^{2n} \tilde{P}_k T_k(x).$$

Using the notation introduced in Chapter 2, let us consider the set of Hermitian matrices Y such that one has the identity

$$P(x) = \tilde{\Pi}_n(x)^* Y \tilde{\Pi}_n(x) \tag{3.26}$$

with $\tilde{\Pi}_n(x) = \begin{bmatrix} T_0(x)I_m, & T_1(x)I_m, & \ldots, & T_n(x)I_m \end{bmatrix}^T$.

The algebraic constraints on Y implied by (3.26) can be expressed in terms of the Chebyshev basis in a simple manner. Using identities (3.25), one can easily check that the following relations hold:

$$\tilde{P}_k = \frac{1}{2}\left(\sum_{i+j=k} Y_{ij} + \sum_{|i-j|=k} Y_{ij} \right), \qquad \text{for } k = 0, \ldots, 2n. \tag{3.27}$$

If L stands for the block lower triangular matrix transforming $\Pi_n(x)$ into $\tilde{\Pi}_n(x)$, i.e., $\tilde{\Pi}_n(x) = L\,\Pi_n(x)$, a simple consequence of Theorem 2.8 is that the set of all solutions Y to (3.26) is parametrized by the relation

$$Y = Y_0 + L^{-T}(Z\,X - X\,Z^T)L^{-1},$$

where Y_0 is a particular solution of (3.26) and X is any skew-Hermitian matrix of form (2.19).

In view of Theorem 2.7, the existence of a positive semidefinite solution Y to (3.26) is found to be the necessary and sufficient condition such that $P(x)$ is a well-defined non-negative polynomial matrix on the real line. Note incidentally that a particular matrix Y_0 satisfying (3.26) is provided by

$$Y_0 = \begin{bmatrix} \frac{1}{2}\tilde{P}_0 - \sum_{k=1}^{n} \tilde{P}_{2k} & \frac{1}{2}\tilde{P}_1 - \frac{1}{2}\sum_{k=2}^{n} \tilde{P}_{2k-1} & & & \\ \frac{1}{2}\tilde{P}_1 - \frac{1}{2}\sum_{k=2}^{n} \tilde{P}_{2k-1} & 2\tilde{P}_2 & \tilde{P}_3 & & \\ & \tilde{P}_3 & 2\tilde{P}_4 & \ddots & \\ & & \ddots & \ddots & \tilde{P}_{2n-1} \\ & & & \tilde{P}_{2n-1} & 2\tilde{P}_{2n} \end{bmatrix}.$$

When such a Chebyshev basis is chosen, the optimization space is transformed into the convex cone $\mathcal{K}_{\mathbb{R},c}$ of (Chebyshev) matrix coefficients of polynomial matrices non-negative definite on the real axis, i.e.

$$P(x) = \sum_{k=0}^{2n} \tilde{P}_k T_k(x) \succeq 0, \quad x \in \mathbb{R}.$$

Furthermore and as shown above, $\tilde{P}$ belongs to the cone $\mathcal{K}_{\mathbb{R},c}$ if and only if there exists a non-negative block matrix Y with blocks Y_{ij} satisfying (3.27).

By definition, the dual cone $\mathcal{K}^{\star}_{\mathbb{R},c}$ consists of the coefficients

$$\tilde{Q} \doteq \begin{bmatrix} \tilde{Q}_0, & \tilde{Q}_1, & \ldots, & \tilde{Q}_{2n} \end{bmatrix}$$

satisfying the constraints

$$\langle \tilde{Q}, \tilde{P} \rangle_{\mathbb{R}} \geq 0, \quad \forall \tilde{P} \in \mathcal{K}_{\mathbb{R},c}.$$

Recall that we are not restricted to assuming that the matrix coefficients $\tilde{P}_k$ and $\tilde{Q}_k$ are Hermitian for all k. For any $\tilde{Q}$, let us set the block Toeplitz-plus-Hankel matrix

$$T_H(\tilde{Q}) \doteq \begin{bmatrix} \tilde{Q}_0 & \tilde{Q}_1 & \cdots & \tilde{Q}_n \\ \tilde{Q}_1 & \iddots & \iddots & \vdots \\ \vdots & \iddots & \iddots & \tilde{Q}_{2n-1} \\ \tilde{Q}_n & \cdots & \tilde{Q}_{2n-1} & \tilde{Q}_{2n} \end{bmatrix} + \begin{bmatrix} \tilde{Q}_0 & \tilde{Q}_1 & \cdots & \tilde{Q}_n \\ \tilde{Q}_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \tilde{Q}_1 \\ \tilde{Q}_n & \cdots & \tilde{Q}_1 & \tilde{Q}_0 \end{bmatrix}. \tag{3.28}$$

In view of (3.27) and the scalar product definition, one derives the relation

$$\langle \tilde{P}, \tilde{Q} \rangle_{\mathbb{R}} = \sum_{k=0}^{2n} \langle \tilde{Q}_k, \tilde{P}_k \rangle = \frac{1}{2} \sum_{k=0}^{2n} \left[\sum_{i+j=k} \langle \tilde{Q}_k, Y_{ij} \rangle + \sum_{|i-j|=k} \langle \tilde{Q}_k, Y_{ij} \rangle \right] = \frac{1}{2} \langle T_H(\tilde{Q}), Y \rangle,$$

which shows that the dual cone $\mathcal{K}^{\star}_{\mathbb{R},c}$ is characterized by $T_H(\tilde{Q}) \succeq 0$. Note that we could have applied the methodology of Chapter 2 to derive these results.

Therefore, the dual form of the conic optimization problem can be expressed in the present case as

$$\max_{u_0,\ldots,u_{k-1}} \left\{ \sum_{\ell=0}^{k-1} b_\ell u_\ell : T_H \left(C - \sum_{\ell=0}^{k-1} u_\ell A_\ell \right) \succeq 0 \right\}.$$

The corresponding barrier function is $f(u) = -\ln\det T_H(C - \sum_{\ell=0}^{k-1} A_\ell u_\ell)$ and the differential characteristics of interest now read

$$\frac{\partial f(u)}{\partial u_\ell} = \langle T_H(S)^{-1}, T_H(A_\ell) \rangle, \tag{3.29a}$$

$$\frac{\partial^2 f(u)}{\partial u_\ell \partial u_s} = \langle T_H(S)^{-1} T_H(A_\ell)\, T_H(S)^{-1}, T_H(A_s) \rangle, \tag{3.29b}$$

where $S = C - \sum_{\ell=0}^{k-1} A_\ell u_\ell$.

From a numerical point of view, this reformulation of the optimization problem on the real line exhibits a considerable advantage over its initial formulation in the sense that is not intrinsically ill-conditioned. Indeed, for all degree n, there exist non-negative matrices $T_H(\tilde{Q})$ with a condition number equal to 2, as illustrated by the trivial example $\tilde{Q} = \big[I_m, \quad 0, \quad \ldots, \quad 0\big]$. As a result, the numerical behaviour of the computational optimization scheme is expected to be substantially improved.

Finally, let us point out that the differential characteristics of the Chebyshev basis reformulated barrier function (3.29) can also be computed in a fast way with the help of structured algorithms. However, a divide-and-conquer strategy for computing efficiently the inner products (3.29) as presented above is not likely to exist. Indeed, our strategy for block-Toeplitz or block-Hankel matrices is based on the existence of a displacement shift matrix Z, compatible with these structures, and such that finite expansions are computable from the generators using this shift, see (3.18) and (3.19).

3.4 Proximal analytic center cutting plane method

Consider the general formulation of a convex optimization problem:

$$\min\{f(y) : y \in Q \subseteq \mathbb{R}^n\}, \tag{3.30}$$

where $f(y)$ is a convex function and Q is a closed convex set. The class of optimization schemes studied in this section is based on the property that any closed convex set can be described as an intersection of half-spaces, see [126, Theorem 11.5]. For a given convex set Q, any finite subset of such half-spaces is a *polyhedral relaxation*, which is defined by a set of separating hyperplanes. The terminology *cutting planes* denotes these separating hyperplanes, which cut the space into two parts. It is also associated, by extension, with several methods based on successive polyhedral relaxations.

3.4.1 Method description

Standard cutting plane methods belong to the set of optimization methods based on *first-order oracles*. The basic assumption is the existence of a "black-box" oracle, which produces separation hyperplanes. From a theoretical point of view, it is well known that, for any $y \notin Q$, there always exists such an hyperplane [126, Chapter 11]. Moreover, it is proved in the literature that optimization and separation are polynomially equivalent [67]. These results explain the practical importance of cutting plane methods for solving optimization problems. Since the convex function $f(y)$ could be non-smooth, the solution methods belong to the realm of *nondifferentiable optimization*.

Given a convex set $Q \subset \mathbb{R}^n$ and a point $y \in \mathbb{R}^n$, a generic first-order oracle returns the following information:

- If $y \in Q$ (*optimality cut*), it computes the function value $f(y)$ and an associated subgradient $\gamma(y)$: $f(x) \geq f(y) + \langle \gamma(y), x - y \rangle, \ \forall x \in \operatorname{dom} f$.
- If $y \notin Q$ (*feasibility cut*), it returns at least one separating hyperplane: $\langle \gamma, x - y \rangle + \delta \leq 0, \ \forall x \in Q$.

An optimality cut reduces the objective function value by making use of the subgradient $\gamma(y)$. Indeed, the optimal point $\hat{y}$ satisfies the inequalities

$$f(y) + \langle \gamma(y), \hat{y} - y \rangle \leq f(\hat{y}) \leq f(y),$$

for any $y \in \operatorname{dom} f$. Therefore, an optimality cut, which is of the form $\langle \gamma(y), x - y \rangle \leq 0, \ \forall x \in Q$, selects the appropriate part of the polyhedral relaxation. A feasibility cut separates the current query point from the feasible set. Depending on δ, it is said to be shallow ($\delta < 0$), central ($\delta = 0$) or deep($\delta > 0$). Figures 3.2 and 3.3 illustrate these different cuts with respect to the convex set Q (previous feasibility cuts are also drawn).

Using an embedding of the original problem in an $\mathbb{R}^n \times \mathbb{R}$ dimensional space, our generic problem (3.30) can be rewritten in the canonical form

$$\begin{array}{ll} \min & z \\ \text{s. t.} & f(y) - z \leq 0, \\ & y \in Q. \end{array} \tag{3.31}$$

Polyhedral relaxation schemes are based on approximating the non-smooth components of (3.31) by increasingly refined polyhedral sets. An additional constraint is given by the *epigraph cut*: the objective function achieves at most the best recorded value. If the best value is reached at point $\hat{y}$ this constraint takes the form

$$z \leq \hat{\theta} \doteq f(\hat{y}),$$

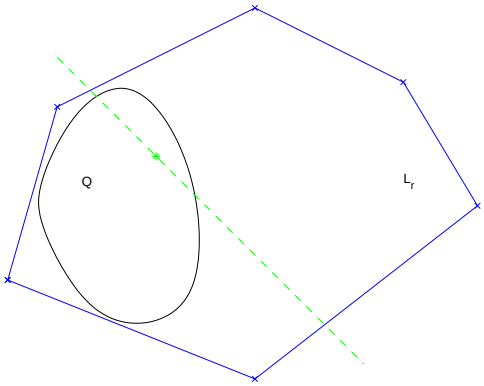

Figure 3.2. *Optimality cut*

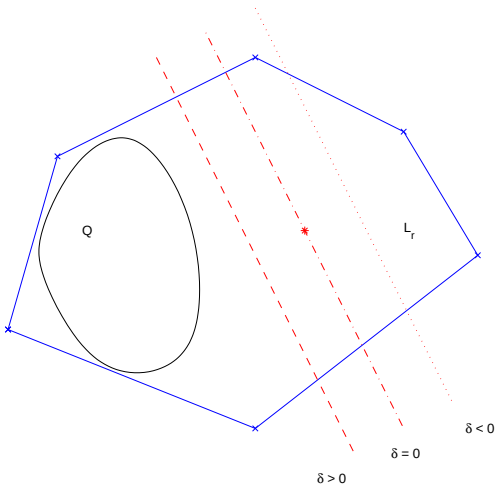

Figure 3.3. *Feasibility cuts*

where z is a surrogate for the non-smooth objective. The set we described is named the *localization set* and is formally written

$$\mathcal{L}(A, E, c, \hat{\theta}) = \{(y, z) \in \mathbb{R}^n \times \mathbb{R} : A^T y - E^T z \leq c,\ z \leq \hat{\theta}\}. \tag{3.32}$$

Each column of A (and of E) is associated with a cutting plane generated previously. Note that E is a simple$\{0, 1\}$-row vector: its ith component is equal to 1 if the ith cut is an optimality cut associated with f and equal to 0 if that cut is a feasibility cut associated with Q.

A generic cutting plane algorithm works as follows. At each iteration r, it refines the current localization set $\mathcal{L}_r$, which contains the optimal solution x^*, using the information returned by the oracle. More precisely, the following operations are performed.

Algorithm 3.1 Generic cutting plane method

1. Choose an accuracy $\epsilon > 0$ and a starting point y_0.
2. rth iteration ($r \geq 0$).
   (a) Generate a new query point y_r as a "center" of the localization set $\mathcal{L}_r$;
   (b) Call the oracle at this point, so as to generate an associated cutting plane;
   (c) Update the localization set: $\mathcal{L}_{r+1} \leftarrow \mathcal{L}_r \cap \{(y, z) : A_r^T y - E_r^T z \leq c_r\}$;
   (d) Test the termination criterion.

There exist several cutting plane methods (Kelley-Cheney-Goldstein [90, 30], center of gravity [96, 114], ellipsoid [106], inscribed ellipsoid [131], volumetric center [140], analytic center [63]), which differ in the selection of the next query point, in the availability of a lower bound and in the termination criterion. As theoretical properties (convergence, complexity estimates,...) of these schemes are extensively studied in the optimization literature, it gives a strong background to any problem-specific algorithm based on them.

From a practical point of view, all cutting planes generated by the oracle must be stored somehow in the computer memory. Clearly, the memory requirement is an important drawback of these solution methods. However, several cutting planes can often be pruned during the optimization process (because they are either redundant or inactive). If the generated cuts are sufficiently deep and an efficient pruning procedure is used, then the memory issue can be avoided. Note that both aspects are strongly problem-specific and can be based on heuristics.

The main advantages of cutting plane methods are the practical efficiency estimate and the possible avoidance of extra variables introduced by a problem reformulation. This last fact explains the growing interest for cutting plane methods in Systems and Control[88, 118]. Remember that it is extremely important to provide these schemes with oracles adapted to the problem structure; a well-devised oracle is the key ingredient for an efficient scheme.

Let us now describe the *proximal generalized analytic center cutting plane* method. Given a standard weighted logarithmic barrier for the localization set $\mathcal{L}(A, E, c, \hat{\theta})$

$$F(s) = -\sum_{i=0}^{r} w_i \log s_i,$$

with $s \in \operatorname{int} \mathbb{R}_+^{r+1}$ defined by

$$s_0 = \hat{\theta} - z, \quad s_i = c_i - (A^T y - E^T z)_i, \quad i = 1, \ldots, r,$$

it is augmented with a proximal term to yield the *augmented barrier*

$$\frac{1}{2}\langle Q(y - \bar{y}), y - \bar{y}\rangle + F(s).$$

In the above expression, the weight matrix Q is positive definite and completely defines the proximal term. The point $\bar{y}$ is known as the *proximal center*. In view of the original

minimization problem, we compound the augmented barrier with the approximation z of the non-smooth objective $f(y)$ to get the *generalized augmented barrier*

$$tz + \frac{1}{2}\langle Q(y-\bar{y}), y-\bar{y}\rangle + F(s). \tag{3.33}$$

For an increased flexibility, an additional parameter $t \geq 0$ has been introduced to control the weight associated with the objective.

The proximal generalized analytic center cutting plane method then defines the next query point as the y-component of

$$\begin{array}{rl} (\hat{y},\hat{z}) = \arg\min & tz + \frac{1}{2}\langle Q(y-\bar{y}), y-\bar{y}\rangle - \sum_{i=0}^{r} w_i \log s_i \\ \text{s.t.} & s_0 = \hat{\theta} - z, \\ & s_i = c_i - (A^T y - E^T z)_i, \quad i = 1,\ldots,r. \end{array} \tag{3.34}$$

To make sure that the minimum exists, we introduce the following assumption.

Assumption 3.1. The localization set has a non-empty interior.

Remark. The variant $t = 0$ is known as the standard proximal analytic center cutting plane method [107].

Any convex minimization scheme (primal, dual, primal/dual, ...) can be used to solve (3.34), see e.g. [48]. The number of iterations needed to obtain a new query point is the number of *inner iterations*. This terminology is necessary to make a clear difference with the number of *outer iterations*, i.e., the number of calls to the oracle. When adding linear constraints to the localization set, problem (3.34) is only slightly modified. A good starting point can therefore be obtained from the solution of the previous outer iteration by using an appropriate *re-entering direction*.

In general, complexity estimates of analytic center cutting plane methods depend on several assumptions, which are made in order to simplify the theoretical analysis [9, 64, 107]. Unfortunately, these complexity estimates are often too large to be valuable in practice, especially for high accuracy. However, the practical behaviour of these schemes is usually much better.

3.4.2 Separating hyperplanes

Let us now apply the general framework to the optimization problems described in Section 3.1. We focus on the scalar case $m = 1$ as the extension to matrix polynomials is evident. In practice, we need to specify how to generate the separating hyperplanes for our convex sets.

Let us start with our dual problems (3.3) and (3.5).

Dual problems

General setting Remember that a conic dual problem related to some cone $\mathcal{K}$ of non-negative polynomials has the following structure

$$\begin{array}{ll} \max & \langle b, y\rangle \\ \text{s.t.} & \Lambda(c - \mathcal{A}^* y) \succeq 0 \end{array} \tag{3.35}$$

where Λ is a structured Hermitian linear operator. This linear operator is directly related to the functional basis chosen to represent $\mathcal{K}$, see Chapter 2 or Table 3.1.

Non-negative polynomials	Functional basis	Structure
real line	$\pi(\cdot)$ - "powers" $\tau(\cdot)$ - "Chebyshev"	Hankel Toeplitz-plus-Hankel
unit circle	$\pi(\cdot)$ - "powers"	Toeplitz

Table 3.1. *Standard linear operators* Λ

As the closed convex cone of interest is

$$Q = \{y : \Lambda(c - \mathcal{A}^* y) \succeq 0\},$$

any standard oracle checks whether the matrix $\Lambda(c-\mathcal{A}^*y)$ is positive semidefinite at the current query point $y = y_r$. It then generates either a feasibility cut or an optimality cut:

Feasibility cut $(y_r \notin Q)$

As y_r is not a feasible point, there exists at least one non-zero vector $v \in \mathbb{C}^n$ such that

$$\langle \Lambda(c - \mathcal{A}^* y_r)v, v\rangle < 0.$$

Without loss of generality, we assume that $\|v\|_2 = 1$. If Λ^* denotes the linear operator dual to Λ with respect to the appropriate inner products, the feasibility cut associated with v is defined as

$$\langle \gamma, y - y_r\rangle + \delta \leq 0, \quad \forall y \in Q \tag{3.36}$$

with $\gamma = \mathcal{A}\Lambda^*(vv^*)$ and $\delta = -\langle \Lambda(c - \mathcal{A}^* y_r)v, v\rangle$. As $\delta > 0$, the associated cut is deep. It is an open question to determine how deep this cut is, as the answer depends on the ratio $\frac{-\langle \Lambda(c-\mathcal{A}^* y_r)v,v\rangle}{\|\mathcal{A}\Lambda^*(vv^*)\|}$ and thus on the data.

Remark. The same procedure could be applied *mutatis mutandis* to a set of linearly independent directions of negative curvature. In this case, each vector of the set generates a deep cut.

Remark. If our problem constraints are based on interpolation conditions, there is a nice interpretation of our cut vector γ. We have chosen to exemplify this fact using the real line setting. In this case, the linear operator $\mathcal{A}$ is representable as an interpolation matrix

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{k-1} & x_{k-1}^2 & \dots & x_{k-1}^{2n} \end{bmatrix}, \quad x_i \neq x_j. \tag{3.37}$$

Then the vector $\gamma = \begin{bmatrix}\gamma_0, & \dots, & \gamma_{k-1}\end{bmatrix}^T$ can be computed as

$$\gamma_i = (v(x_i))^2 \geq 0, \qquad i = 0, \dots, k-1,$$

where $v(\cdot)$ is the polynomial of degree n defined by the vector v of negative curvature. To sum up, the vector γ is obtained by squaring the values of the polynomial $v(\cdot)$ at the interpolation points $x_0, \dots, x_{k-1}$.

Optimality cut $(y_r \in Q)$

The oracle returns the function value $f(y_r) = -\langle b, y_r \rangle$ and the subgradient $\gamma_r = -b$. Note that the minus sign comes from the fact that (3.35) is a maximization problem while our canonical problem (3.31) is a minimization problem. Moreover the optimality cut has the structure (3.36) because of the linear objective function. Eventually, the straightforward full disaggregation

$$f(y_r) = -\sum_{\ell=0}^{k-1} b_\ell y_{r,\ell}, \quad \gamma_r = -\sum_{\ell=0}^{k-1} b_\ell e_\ell$$

yields a different oracle. As this straightforward disaggregation is artificial, it will not be considered in the sequel.

Strategies for checking feasibility Cutting plane methods heavily rely on the existence of an efficient procedure to check feasibility of the current query point. This is equivalent to the existence of an efficient separation procedure. For our class of problems, we need to check whether a given Hermitian matrix $\Lambda(c - \mathcal{A}^* y)$ is positive semidefinite. Although there exist several linear algebra algorithms that are appropriate for this purpose, the selected algorithm should provide us with sufficiently deep cuts as a by-product. Let us review a selection of strategies for generating cutting planes at the current query point y.

Minimum eigenvalue Given the Hermitian matrix $\Lambda(c - \mathcal{A}^* y)$, we compute its minimal eigenvalue with any eigenvalue solver (for instance, the symmetric QR algorithm [66, Section 8.3] or Lanczos method [66, Section 9.1] can be used). If it is negative, our current point y is not feasible for Q and the corresponding eigenvector v, which can be computed via inverse iterations [66, Section 8.2], generates an appropriate cutting plane. Because

$$\delta = -\langle \Lambda(c - \mathcal{A}^* y)v, v \rangle = -\lambda_{\min}(\Lambda(c - \mathcal{A}^* y)),$$

this strategy clearly gives us a deep separating hyperplane.

Negative eigenvalues Of course the previous strategy could be extended to a selection of negative eigenvalues of $\Lambda(c - \mathcal{A}^* y)$. For each infeasible point y, we then insert the associated cutting planes. As this strategy can possibly add a large number of cuts at each iteration, we could add only p cuts $(p \geq 1)$, corresponding to the p smallest negative eigenvalues.

Direct factorization Another strategy would be to perform a direct factorization (an LDL^*-factorization or the Cholesky factorization) of the matrix $\Lambda(c - \mathcal{A}^* y)$ [66, Sections 4.1 and 4.2]. A vector v of negative curvature is then obtained as soon as a negative pivot appears in the factorization scheme. This strategy is not appropriate for generating efficient cutting planes; the associated cuts are almost central and the actual convergence rate of our cutting plane scheme is thus negatively affected by this strategy. However, iterative refinement of this direction v could be made to improve its quality, see [22].

Specific issues

Structured matrices As the matrix $\Lambda(c - \mathcal{A}^* y)$ is structured, see Table 3.1, one should preferably use structured algorithms [86] to compute its eigenvalues. In the case of symmetric Toeplitz matrices, such computations can be efficiently done via specific algorithms, which take into account the matrix structure and symmetry [36, 99, 100, 102, 147, 148].

If the structure of Λ is not used when checking feasibility, the complexity estimate for the above strategies is $\mathcal{O}(n^3)$ flops. If its structure is suitably used, we may obtain a complexity as low as $\mathcal{O}(n^2)$ flops.

Ill-conditioned matrices As positive definite Hankel matrices are exponentially ill-conditioned [13], the dual problem in the real line setting is challenging. Indeed, checking positive definiteness of a matrix in finite arithmetics implies the existence of numerical errors, which are usually bounded by a function of its condition number or of its norm [75, 66].

Therefore, we have chosen to shift all Hankel matrices before checking positive semidefiniteness

$$H(c - A^T \bar{y}) \quad \rightarrow \quad H(c - A^T \bar{y}) + \sigma I_{n+1},\ \sigma > 0.$$

Note that this shift destroys the Hankel structure. The regularization parameter σ must be fixed *before* starting the optimization process as it modifies the feasible region. Let us clarify the meaning of such a shift. Consider the "shifted dual problem"

$$\begin{array}{ll} \max & \langle b, y \rangle \\ \text{s.t.} & s + A^T y = c, \\ & H(s) + \sigma I_{n+1} \succeq 0. \end{array} \tag{3.38}$$

The associated primal problem reads as follows

$$\begin{array}{ll} \min & \langle c, p \rangle + \sigma \langle I_{n+1}, Y \rangle \equiv \langle c, p \rangle + \sigma \operatorname{Trace} Y \\ \text{s.t.} & Ap = b, \\ & p = H^*(Y), \quad Y \in \mathcal{S}_+^{n+1}. \end{array} \tag{3.39}$$

Recall that last two constraints of the "shifted primal problem" are equivalent to $p \in \mathcal{K}$. By introducing a small shift in the dual constraint $H(s) \succeq 0$, we have modified our primal objective function and added a *regularization term*. There is a trade-off between not altering the primal problem too much (small σ) and regularizing the dual oracle (σ big enough).

Remark. Other regularization terms (for instance, $\frac{\sigma}{2} \|p\|_2^2$) added in the primal objective function would lead to modified dual problems, which could also be written explicitly.

Primal problems

Another course of action is to directly tackle the primal problem

$$\begin{array}{ll} \min & \langle c, p \rangle \\ \text{s.t.} & \mathcal{A}p = b, \\ & p \in \mathcal{K}, \end{array} \tag{3.40}$$

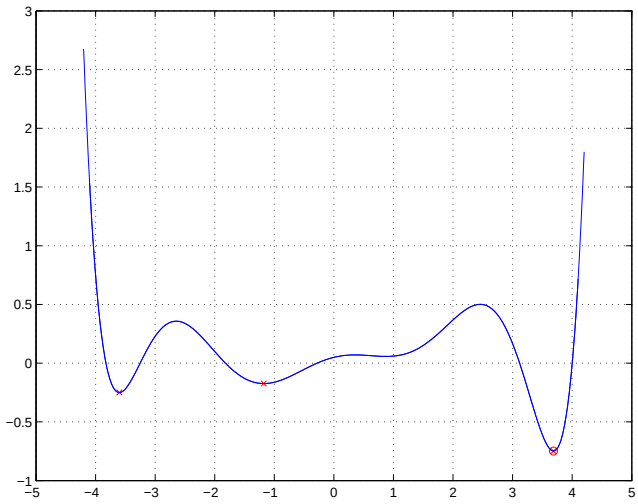

Figure 3.4. *Illustration of primal oracle*

where $\mathcal{K}$ is some cone of polynomials non-negative on a curve $\Gamma \subset \mathbb{C}$.

As the intersection of the linear subspace $Ap = b$ with the cone $\mathcal{K}$ is a convex set, an oracle can be written for the convex set

$$Q = \{p : \mathcal{A}p = b, p \in \mathcal{K}\}.$$

Indeed, given a query point p, checking that the associated polynomial $p(\cdot)$ is non-negative on Γ can be done by checking the values of $p(\cdot)$ on an appropriate grid or by checking the roots of $p'(\cdot)$. The former solution is related to the standard discretization approach used in semi-infinite programming; it is not really appropriate in our polynomial context. The latter is apparently the most appropriate to our problem class; the polynomial $p(\cdot)$ is non-negative on Γ if and only if its local extrema are all non-negative on Γ and the values at $\pm\infty$ are positive. The local extrema with (minimum) negative value can then be used to generate cutting planes, see Figure 3.4. This strategy is very similar to the local reduction method in semi-infinite programming.

Example 3.1. Let $\mathcal{K} = \mathcal{K}_{\mathbb{R}}$ and p_r be an infeasible query point. If x_r is a local extremum of the polynomial p_r with negative value, then a valid cutting plane is given by

$$\langle -\pi_{2n}(x_r), p - p_r \rangle - \langle \pi_{2n}(x_r), p_r \rangle \leq 0.$$

From a theoretical point of view, such an oracle based on computing roots is extremely appropriate and leads to deep cutting planes. Note that any implementation is potentially affected by numerical errors in computing the zero of $p'(\cdot)$, as illustrated by Wilkinson's famous example, see Figure 3.5. An accurate computation of the extrema might be numerically difficult. This is also illustrated on Figure 3.4: the polynomial is almost constant on the interval $[0, 1]$, so that the precise identification of an extremum might be delicate. Even if the computational cost of this primal oracle is attractive, numerical errors might prevent the whole optimization scheme from converging.

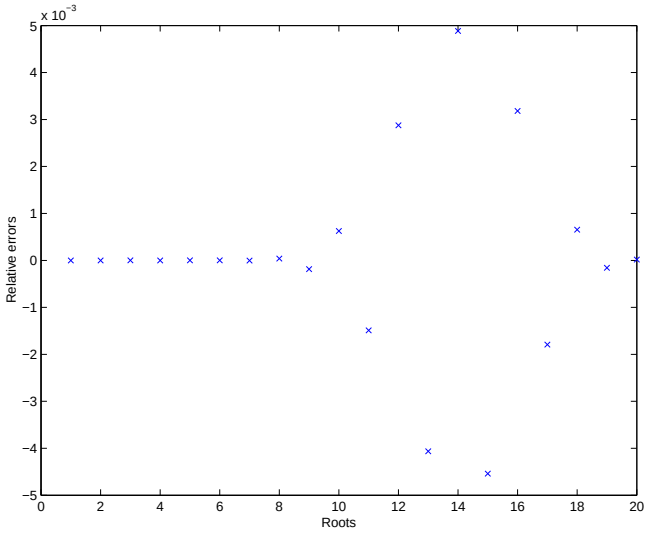

Figure 3.5. *Relative errors in computing the roots of* $p(x) = \prod_{\ell=1}^{20}(x-\ell)$

3.5 Numerical results

In this section, several test instances are solved using three different approaches: a primal-dual semidefinite scheme, a dual path-following scheme and a cutting plane method. More theoretical details on these approaches are given in the preceding sections. Our aim is to asses the practical efficiency of these optimization schemes and to compare them.

All computations were performed on a dual-processor Sun Ultra 80 workstation with 2GB of random access memory (RAM), using MATLAB 6.1.0.450 (R12.1). Note that the computer performance, as well as the programming language, influence our execution times. Therefore, central processing unit (CPU) execution times are not reported hereafter. In order to compare execution times of two strategies based on the same method, the ratio

$$\text{CPU Ratio} = \frac{\text{CPU Time of Strategy A}}{\text{CPU Time of Strategy B}}$$

is used. Except otherwise stated, MATLAB toolboxes were used with their default settings. All optimization problems are solved up to an accuracy $\epsilon = 10^{-6}$.

Since non-negative polynomials on the real line and on the unit circle are related by a one-to-one mapping, they are very similar. The main difference is the existence of infinite points in the real line setting, which is responsible for numerical difficulties. Moreover the geometry of the corresponding dual spaces is essentially the same; there exists a congruence, known as Fischer-Frobenius transformation [83, Chapter 4, Section 19], which establishes a one-to-one mapping between Hankel and Toeplitz matrices. For these reasons, we have chosen to focus on conic problems of the form (3.1) with $\mathcal{K} = \mathcal{K}_{\mathbb{C}}$, the cone of scalar trigonometric polynomials.

Our test instances were generated as follows:

- The linear operator $\mathcal{A}$ and the vector b represent interpolation constraints on the

trigonometric polynomial $p(\cdot)$. The interpolation points are uniformly distributed on the unit circle and the components of b on the interval $[0, 10]$.

- The objective vector $c = e_0$ corresponds to minimizing the integral of $p(\cdot)$ on the whole unit circle, i.e.,

$$\langle c, p\rangle_{\mathbb{R}} = \frac{1}{2\pi}\int_{-\pi}^{\pi} p(e^{\jmath\theta})d\theta = \langle e_0, p\rangle_{\mathbb{R}}.$$

They have three main advantages over completely random instances: the dual problems are always strictly feasible, there is no duality gap and the associated optimization problems are bounded.

3.5.1 Primal-dual semidefinite solver

Table 3.2 contains a description of our test instances as well as their optimal value. These problems were solved via the straightforward semidefinite embedding

$$\begin{array}{ll} \min & \langle T(c), Y\rangle \\ \text{s.t.} & \langle T(a_i), Y\rangle = b_i, \quad i = 0, \ldots, k-1 \\ & Y \in \mathcal{H}_+^{n+1} \end{array} .$$

with MATLAB SDPT3 toolbox (version 2.3) [137, 138]. Although the worst-case analytical complexity is $\mathcal{O}(\sqrt{n}\log\frac{1}{\epsilon})$, it is well known that, in practice, the number of Newton iterations is usually between 20 and 80. In our case, we see that this number belongs to the interval $[10, 40]$.

We have also mentioned in Table 3.2 the results obtained after reformulating our initial problems. Indeed, our instances are generated using interpolation constraints so that a specific reformulation, which is derived in Section 4.3, can be used. Consequently, the problem dimension is reduced to the number k of constraints and the optimization problems involve only unstructured matrices of size k. Clearly, the values of CPU ratio indicate that the reformulation should always be used (if possible). This observation is backed by theoretical results; the associated worst-case complexity bound is reduced from $\mathcal{O}((kn^{3.5} + k^2n^{2.5} + k^3n^{0.5})\log\frac{1}{\epsilon})$ flops to $\mathcal{O}((kn^2 + k^2n + k^3) + k^{4.5}\log\frac{1}{\epsilon})$ flops.

3.5.2 Dual short-step path-following method

Let us now focus on a fast dual solver. Table 3.3 contains the numerical results of a short-step path-following method applied to the dual formulation (3.5). Our method consists of two phases: an auxiliary process (Algorithm 1.2) and the main scheme (Algorithm 1.1). Remember that the worst-case analytical complexity of this scheme is $\mathcal{O}(\sqrt{n}\log\frac{1}{\epsilon})$. In practice, the number of Newton iterations approximately satisfies this bound.

3.5.3 Proximal analytic center cutting plane method

MATLAB P-ACCPM is an implementation of the (generalized) proximal analytic center cutting plane method; the next query point is found by minimizing the generalized augmented barrier (3.33). This toolbox is currently developed at Logilab (Université de Genève, HEC). In particular, a future public release might (slightly) modify our

Table 3.2. *Test instances solved via a primal-dual semidefinite solver*

		std embedding	interp. reform.		
n	k	iterations	iterations	CPU ratio	obj.
25	5	11	9	2.34e+00	4.32231e-01
25	10	14	13	1.35e+00	8.48939e+00
25	15	13	13	1.19e+00	2.06493e+00
25	20	16	15	1.25e+00	4.63573e+01
25	25	12	19	7.27e-01	9.22868e+00
50	10	15	12	3.63e+00	8.16722e-01
50	20	13	13	2.89e+00	1.81945e+00
50	30	15	20	2.11e+00	2.13887e+00
50	40	19	17	2.68e+00	5.00080e+01
75	15	12	12	5.28e+00	9.96334e-01
75	30	16	15	5.51e+00	8.83442e+00
75	45	20	21	4.54e+00	1.33466e+01
75	60	22	22	4.10e+00	4.38644e+02
100	20	13	12	1.08e+01	9.80036e-01
100	40	21	18	1.13e+01	4.18237e+01
100	60	32	21	1.21e+01	4.34991e+02
100	80	23	25	6.63e+00	4.24888e+01
125	25	13	14	1.57e+01	7.84959e-01
125	50	13	16	1.16e+01	1.53064e+00
125	75	28	19	1.82e+01	1.32573e+03
125	100	40	30	1.45e+01	3.85243e+05
150	25	13	13	2.52e+01	7.21620e-01
150	50	19	17	2.38e+01	3.50461e+00
150	75	18	22	1.63e+01	5.51209e+00
150	100	24	29	1.43e+01	6.32946e+01
175	25	16	13	4.53e+01	1.37099e+00
175	50	13	14	2.84e+01	1.31951e+00
175	75	27	20	3.75e+01	7.99402e+02
175	100	16	18	2.15e+01	4.54349e+00
175	125	31	24	2.53e+01	2.41865e+03
200	25	13	12	5.60e+01	5.61280e-01
200	50	16	17	4.09e+01	1.64215e+00
200	75	21	18	3.95e+01	3.55920e+01
200	100	27	21	4.38e+01	2.91542e+02
200	125	27	22	3.35e+01	1.99622e+03
200	150	35	30	2.72e+01	1.93318e+03
500	200	25	26	1.78e+02	6.26670e+00

results or conclusions. Let us discuss how to provide this MATLAB toolbox with the appropriate dual oracles, as well as the primal ones. Note that we also focus on various method parameters, which could influence the practical rate of convergence.

Table 3.3. *Test instances solved via a short-step path-following method*

		iterations		
n	k	auxiliary	main	total
25	5	102	154	256
25	10	120	186	306
25	15	89	258	347
25	20	188	204	392
25	25	135	360	495
50	10	94	162	256
50	20	80	276	356
50	30	93	381	474
50	40	220	449	669
75	15	92	276	368
75	30	144	393	537
75	45	199	468	667
75	60	357	627	984
100	20	79	284	363
100	40	242	367	609
100	60	406	523	929
100	80	297	574	871
125	25	95	296	391
125	50	92	503	595
125	75	480	656	1136
150	25	97	295	392
150	50	176	456	632
150	75	190	606	796
150	100	362	695	1057
175	25	123	279	402
175	50	109	464	573

Dual problems

Oracles Remember that direct factorization methods of $\Lambda(c - \mathcal{A}^* y)$ are usually not appropriate to generate efficient cutting planes. This general fact also holds for our problem class. If feasibility is checked with such a direct strategy, then the whole optimization scheme breaks down after a few iterations. This behaviour originates from numerical errors associated to almost central feasibility cuts (no iterative refinement was performed to improve the directions of negative curvature). Therefore, we focus on two oracles based on the negative eigenvalues of $\Lambda(c - \mathcal{A}^* y)$. Oracle MIN is based on the minimum eigenvalue of this matrix and oracle NEG on all negative eigenvalues. The numerical results are given in Table 3.4. Query points were generated using the standard proximal analytic center of the localization set and a dual method. Our starting point is the origin.

The following observations, which seem to be valid for problems with completely random linear constraints too, can be made:

- The ratio between the number of outer iterations and the problem dimension

Table 3.4. *Comparison of oracles MIN and NEG*

		oracle MIN			oracle NEG			
n	k	outer	inner	cuts	outer	inner	cuts	CPU ratio
25	5	46	134	46	45	140	48	9.60e-01
25	10	95	227	95	94	233	108	8.73e-01
25	15	125	325	125	110	304	139	1.02e+00
50	10	91	239	91	88	244	100	8.85e-01
50	20	165	422	165	140	373	176	9.44e-01
50	30	229	593	229	148	450	286	1.07e+00
75	15	122	324	122	97	288	132	9.17e-01
75	30	242	589	242	195	489	255	1.01e+00
75	45	379	888	379	319	828	434	9.10e-01
100	20	156	405	156	133	380	177	8.04e-01
100	40	324	781	324	290	741	353	8.36e-01
100	60	1142	2775	1142	1055	2733	1191	8.14e-01
125	25	194	524	194	140	426	231	8.63e-01
125	50	346	930	346	208	629	470	9.20e-01
125	75	1072	2553	1072	966	2537	1135	7.50e-01
150	25	190	499	190	157	436	212	7.04e-01
150	50	402	1024	402	294	808	449	8.96e-01
150	75	581	1465	581	396	1135	706	9.03e-01
175	25	206	531	206	177	473	212	7.54e-01
175	50	370	991	370	245	714	466	8.17e-01
175	75	1241	2970	1241	1068	2908	1212	7.59e-01
200	25	184	514	184	119	371	216	9.07e-01
200	50	378	1033	378	281	793	428	8.40e-01
200	75	591	1493	591	422	1186	696	8.31e-01

usually belongs to the interval $[4.1, 9.5]$, which indicates that the underlying geometry of our convex set Q is strongly non-polyhedral.

- Three instances, namely $(100, 60)$, $(125, 75)$ and $(175, 75)$, require more outer iterations than the others. This behaviour is related to the proximal term, which prevents the next query point from being too far from the current proximal center. This is known as the *spring effect* of the proximal term. In these three cases, the optimal points are far away from the origin; many iterations are therefore needed before arriving in their neighbourhood. This is a known drawback of the proximal term, which could be avoided by dynamically adjusting the weight on the proximal term or by using initial box constraints.

- Oracle NEG reduces the number of outer iterations. Note that one of the feasibility cuts generated by oracle NEG at a given query point is the one generated by oracle MIN and that both oracles coincide if there is less than one negative eigenvalue. Our observation is therefore natural: we get a better convergence speed because the localization set is reduced more efficiently ! However, the total number of cuts is slightly increased.

Table 3.5. *Influence of box constraints*

		without proximal term			with proximal term			
n	k	outer	inner	cuts	outer	inner	cuts	CPU ratio
25	5	46	152	46	47	140	47	1.20e+00
25	10	97	252	97	98	238	98	9.61e-01
25	15	131	364	131	131	344	131	1.01e+00
50	10	98	277	98	91	237	91	1.13e+00
50	20	161	438	161	168	438	168	9.47e-01
50	30	235	652	235	227	589	227	1.05e+00
75	15	118	347	118	122	328	122	1.00e+00
75	30	247	673	247	236	569	236	1.09e+00
75	45	411	1150	411	377	881	377	1.16e+00
100	20	164	467	164	161	419	161	1.05e+00
100	40	363	1046	363	340	847	340	1.11e+00
100	60	580	1798	580	1160	2830	1 160	3.87e-01
125	25	200	592	200	201	558	201	9.94e-01
125	50	357	1060	357	347	932	347	1.06e+00
125	75	603	1901	603	1120	3029	1 120	3.88e-01
150	25	202	600	202	193	502	193	1.06e+00
150	50	411	1177	411	394	995	394	1.07e+00
150	75	611	1780	611	595	1503	595	1.07e+00
175	25	204	572	204	216	568	216	9.33e-01
175	50	380	1154	380	374	1028	374	1.04e+00
175	75	594	1896	594	1270	3205	1270	3.64e-01
200	25	184	555	184	180	501	180	1.04e+00
200	50	381	1114	381	392	1090	392	9.82e-01
200	75	598	1708	598	606	1545	606	1.02e+00

- As the re-entering direction was not implemented in the tested version of MATLAB P-ACCPM, the ratio between the number of inner and of outer iterations belongs to the interval $[2, 3]$.

Outer approximation Initial box constraints can be introduced at the beginning of the optimization process. On the one hand, they speed up the convergence of a number of instances in the absence of proximal term, see Table 3.5. In this case, there is no spring effect of the proximal term. On the other hand, initial box constraints which are combined with a proximal term very often lead to numerical errors at the end of the optimization scheme. Therefore, they must be used with care and probably discarded during the optimization scheme. This numerical issue needs to be investigated further. Instead of specifying arbitrary box constraints, one could also provide a (partial) outer approximation of the convex set

$$Q = \{y : \Lambda(c - \mathcal{A}^* y) \succeq 0\}.$$

For instance, any set of random complex unit vectors $\{v_r\}_{r \in R}$ defines such an approximation

$$Q \subset \mathcal{L}_0 = \{(y, z) : \langle \Lambda(c - \mathcal{A}^* y) v_r, v_r \rangle \succeq 0, r \in R\}.$$

Table 3.6. *Influence of dynamic epigraph cut weights*

		static weight			dynamic weight			
n	k	outer	inner	cuts	outer	inner	cuts	CPU ratio
25	5	46	134	46	46	145	46	1.04e+00
25	10	95	227	95	108	373	108	6.91e-01
25	15	126	329	126	131	364	131	8.91e-01
50	10	91	239	91	92	280	92	9.19e-01
50	20	165	423	165	175	490	175	8.97e-01
50	30	232	606	232	215	566	215	1.08e+00
75	15	122	323	122	120	352	120	9.62e-01
75	30	237	579	237	291	816	291	7.55e-01
75	45	376	867	376	452	1251	452	7.41e-01
100	20	156	401	156	156	415	156	9.86e-01
100	40	327	801	327	423	1258	423	6.94e-01
100	60	1146	2991	1146	1575	8630	1575	2.86e-01
125	25	195	523	195	202	550	202	9.59e-01
125	50	348	934	348	328	864	328	1.06e+00
125	75	1095	2844	1095	1513	7507	1513	3.42e-01
150	25	190	496	190	219	609	219	8.06e-01
150	50	402	1023	402	461	1322	461	8.44e-01
150	75	587	1481	587	656	1860	656	8.57e-01
175	25	209	541	209	223	621	223	9.20e-01
175	50	377	1015	377	411	1159	411	8.87e-01
175	75	1227	2979	1227	1743	9943	1743	3.06e-01
200	25	179	497	179	186	549	186	9.56e-01
200	50	387	1059	387	409	1133	409	9.39e-01
200	75	583	1473	583	741	2165	741	7.35e-01

As we cannot guarantee that this initial polyhedra $\mathcal{L}_0$ is closed, it cannot be used without a proximal term or initial box constraints. Several unreported numerical experiments show that this outer approximation does not reduce the number of outer iterations.

Method parameters For several problem classes [48, 146], a standard rule is to increase the weight of the epigraph cut by 1 at each iteration. The efficiency of this rule is clearly problem-specific as it depends on the problem geometry. Our numerical results given in Table 3.6 confirm that this strategy is not efficient for our problem class. All previous results were obtained using the standard proximal analytic center of the localization set, i.e., $t = 0$ in (3.34). We have also studied the use of the generalized proximal analytic center as query point generator. Of course, the parameter t in (3.34) must be dynamically adjusted during the optimization scheme. Although several rules based on the duality gap have been tested, we were not able to find any successful rule, which generically reduces the number of outer iterations.

Problem reformulation As the linear operator $\mathcal{A}$ of our test instances are related to interpolation constraints, the dual optimization problem can be reformulated to be

Table 3.7. *Impact of problem reformulation*

		without reformulation			with reformulation			
n	k	outer	inner	cuts	outer	inner	cuts	CPU ratio
25	20	204	468	204	205	516	205	1.10e+00
25	25	207	475	207	199	504	199	1.14e+00
50	40	364	844	364	336	822	336	1.42e+00
75	60	535	1333	535	503	1141	503	1.49e+00
100	80	704	1645	704	665	1655	665	1.32e+00
150	100	782	1876	782	721	1833	721	1.54e+00
175	100	745	1947	745	762	2087	762	1.50e+00
175	125	2465	7484	2465	2232	4863	2232	2.13e+00
200	100	952	2368	952	857	2007	857	2.26e+00
200	125	1631	4307	1631	1483	3328	1483	1.95e+00
200	150	2811	7916	2811	2507	5427	2507	1.88e+00
500	200	1431	3885	1431	1406	4068	1406	2.56e+00

independent of the degree n, see Section 4.3. This has already been done for the SDP solver in Table 3.2. Several large problems (with $k \geq min\{\frac{2}{3}n, 100\}$) have been solved by applying MATLAB P-ACCPM to the reformulated dual, see Table 3.7. Let us make two observations.

- Reformulating the dual problem does not improve the behaviour of our cutting plane as much as it does for the SDP solver.
- Without reformulation, the cutting plane approach is competitive with the SDP embedding for large problems (k and/or n are big). For instance, the problem $(n = 500, k = 200)$ is solved in $7.26{\cdot}10^3$ seconds and $2.90{\cdot}10^4$ seconds, respectively.

These facts can be explained by the inherent drawbacks of a primal-dual SDP solver, i.e., memory requirement and oversized linear systems, which dramatically slow down the whole optimization scheme. For these reasons, we advise using structure-specific dual solvers, see Section 3.3. As cutting plane methods avoid using extra variables, but are intrinsically slower than the semi-definite programming approach, the reformulation has obviously less impact on that method.

Specific issues related to the real line problem Table 3.8 describes seven test instances based on interpolation constraints and related to the real line setting. We have solved the shifted dual problems, with $\sigma = 10^{-10}$, and the dual problems corresponding to the use of Chebyshev functional basis. Our previous remarks are still valid, but the reader might notice the odd result for instance F in Table 3.9. This result illustrates the discrepancy between the theoretical oracle and any practical implementation with a shift parameter σ. Indeed, the huge number of outer iterations is related to an inappropriate shift value (it can be solved by increasing this shift) and to numerical errors (they are introduced during the reformulation in Chebyshev basis). In order to avoid these numerical troubles, the rule of thumb for solving real line problems is to reformulate the problem on the unit circle, to solve that reformulation and then to translate the solution back to the original setting. Of course, we could also solve

Table 3.8. *Selected test instances*

	$2n$	k	x_i	b	c
A	8	3	$[2,0,-2]^T$	$[2,3,1]^T$	$c_{0:2:2n}=1; c_{2n}=-1$
B	8	3	$[2,0,-2]^T$	$[2,3,1]^T$	$c_{0:2:2n}=1; c_{2n}=-1; c_{n+1}=0.1$
C	8	3	$[2,0,-2]^T$	$[2,3,1]^T$	$c_{0:2:2n}=1; c_{2n}=-1; c_{n+1}=0.4$
D	8	3	$[1,0,-1]^T$	$[2,1,2]^T$	$c_{0:2:2n}=1$
E	16	3	$[1,0,-1]^T$	$[2,1,2]^T$	$c_{0:2:2n}=1$
F	50	3	$[1,0,-1]^T$	$[2,1,2]^T$	$c_{0:2:2n}=1$
G	8	3	$[1,0,-0.5]^T$	$[2,1,0.5]^T$	$c_{0:2:2n}=1$

Table 3.9. *Box term*

	power basis			Chebyshev basis			
	outer	inner	cuts	outer	inner	cuts	CPU ratio
A	42	147	42	43	157	43	8.95e-01
B	45	172	45	43	149	43	1.06e+00
C	42	157	42	45	163	45	8.12e-01
D	19	64	19	20	81	20	8.33e-01
E	19	63	19	20	81	20	8.57e-01
F	19	64	19	409	3945	409	2.07e-02
G	24	72	24	25	85	25	8.37e-01

Table 3.10. *Proximal term*

	power basis			Chebyshev basis			
	outer	inner	cuts	outer	inner	cuts	CPU ratio
A	36	114	36	38	127	38	8.73e-01
B	37	125	37	39	145	39	8.48e-01
C	35	109	35	35	114	35	9.08e-01
D	13	39	13	19	63	19	6.57e-01
E	13	41	13	14	44	14	9.23e-01
F	13	35	13	22	106	22	4.07e-01
G	23	63	23	23	69	23	8.75e-01

the optimization problem with a primal-dual path-following scheme, but this strategy is not viable for huge problems.

Complexity estimate We have used the dual method of MATLAB P-ACCPM to generate the next query point, see problem (3.34). Its complexity depends on the number r of generated cuts and on the problem dimension k. It requires $\mathcal{O}(rk^2+k^3)$ flops. If the oracle returns only one cut at each iteration and is not adapted to the problem structure, e.g. oracle MIN is used, then the whole complexity is $\mathcal{O}(k^2N^2+Nk^3+Nn^3)$ flops, where N is the number of outer iterations. If N is assumed to be proportional to k (for a fixed ϵ, this condition is empirically satisfied), the total complexity is $\mathcal{O}(k^4+kn^3)$ flops.

Primal problems

As mentioned before, the primal formulation of our problems is intrinsically difficult to solve. The bad numerical behaviour of our primal oracle is also worsen by the nature of our objective functions. Let us be more specific. We have seen that minimizing an integral of a non-negative polynomial is a legitimate objective. For that particular class of objective functions, the cutting plane scheme minimizes the associated integral. Therefore, it very often returns an infeasible query point after each optimality cut. Indeed, non-negativity is only enforced on a grid and the integral is reduced by making the polynomial negative outside this grid... As recovering feasibility could take several iterations, the whole scheme might be completely inefficient.

Oracle There exist several ways of enforcing the linear constraints $Ap = b$:

- They are implicitly taken into account by reformulating the optimization problem in a vector space of smaller dimension. Checking non-negativity of a given query point then implies to "inflate" this point in order to get the corresponding polynomial. After generating the appropriate cutting planes, they are "deflated" to the initial subspace.
- They are explicitly taken into account in the query point generator.
- They are explicitly taken into account by adding a non-smooth convex penalty function, which is related to the constraint violation, in the objective function. The weight associated to this penalty must be dynamically adapted during the optimization scheme.
- They are explicitly taken into account by using any convex approximation. For instance, the constraints $\mathcal{A}p = b$ could be approximated by $\|\mathcal{A}p - b\| \leq \sigma$.

We have used the last procedure. Indeed, the first one leads to an infinite number of iterations (MATLAB P-ACCPM was not able to compute a lower bound), the second one needs to re-engineer the query point generator implemented in the tested version of MATLAB P-ACCPM and the third one is far too sensitive to the weight dynamic. We have chosen to use the infinity norm, which is the most natural distance measure in our linear context, and a maximal constraint violation σ equal to 10^{-5}. Our results are given in Table 3.11. At each iteration, we have generated all cuts related to local extrema with negative value. Our results show that the primal oracle is far more expensive (from a computational point of view) than the dual one. To see this, compare the numbers of outer iterations with those of Table 3.4. A possible improvement would involve the use of a pruning procedure, which eliminates unnecessary cuts. As it reduces the dimensions of the matrices used by the query point generator, this strategy could reduce the CPU time accordingly.

Remark. If only the best feasibility cut is generated at each iteration, the total number of iterations is approximately equal to the number of cuts given in Table 3.11.

Complexity estimate We have used the dual method of MATLAB P-ACCPM to generate the next query point, see Problem (3.34). Its complexity depends on the number r of generated cuts and on the problem dimension $\mathcal{O}(n)$. It requires approximately $\mathcal{O}(rn^2 + n^3)$ flops. If the oracle (root computation) is based on an eigen-

Table 3.11. *Primal oracle*

n	k	outer	inner	cuts
25	5	177	618	579
25	10	163	577	501
25	15	159	545	434
50	10	264	1006	990
50	20	254	918	874
50	30	221	818	797
75	15	338	1334	1419
75	30	311	1175	1279
75	45	278	1055	1146
100	20	381	1549	1811
100	40	551	2251	1835
100	60	3025	14473	6046
125	25	456	1881	2222
125	50	428	1731	2031
125	75	*5000	10201	5645

*: maximal number of iterations reached.

value solver and generates only one cut at each iteration, then the whole complexity is $\mathcal{O}(n^2N^2 + Nn^3 + Nkn)$ flops, where N is the total number of outer iterations.

3.5.4 Comparison of experimental results

For our class of problems, solving the dual problem with the proximal analytic center cutting plane method is far more efficient than solving the primal problem. This conclusion can be drawn from our complexity estimates and is clearly backed by our numerical tests.

The proximal analytic center cutting plane method applied to the dual problem is competitive with primal-dual semidefinite programming solvers for solving conic optimization problems on cones of non-negative polynomials, depending on the problem data. For large problems (see e.g. our instance $(n = 500, k = 200)$), the cutting plane scheme clearly outperforms the primal-dual solver for reasons detailed above. However, for small problems, the primal-dual solver has a better convergence rate as it uses both dual and primal information. Note that only linear cuts have been used to generate the localization sets. We therefore expect that other types of cut (e.g. quadratic cuts, semidefinite cuts) could improve the convergence rate of the proximal analytic center cutting plane method. This approach should be investigated further and is left as a future research topic.

Our structure-specific tailor-made dual solver is probably the most efficient numerical method. Indeed, it enjoys a remarkable complexity analysis, which enables it to outperform all other schemes. However, our experimental results show that the performance of our three methods are somehow similar for small problems.

3.6 Research summary

Our contribution to convex optimization over cones of non-negative matrix polynomials is composed of two parts.

On the one hand, we have shown how the matrix structure (Hankel/Toeplitz) can be used to improve the complexity of solving the associated dual problems. We have obtained a fast dual solver with a remarkable complexity estimate; it provides us with the best known complexity result for this problem class.

On the other hand, we have investigated the use of cutting plane methods in our "polynomial" context. This study was motivated by recent claims in the systems and control literature, see e.g. [118]. We have shown that the cutting plane approach favourably compares with the SDP approach, especially for large problems. Nevertheless, this is no longer true if the problem structure (interpolation constraints, dual matrix structure) is suitably used.

Notes and references

Section 3.3 is based on the forthcoming paper [58]. Alkire and Vandenberghe [4] independently obtained an algorithm to solve optimization problems involving autocorrelation sequences. The associated cone consists of non-negative cosine polynomials, which are particular matrix polynomials. In their case, the barrier function $f(u)$ is thus equal to the logarithmic barrier of a Toeplitz matrix $T(u)$,

$$f(u) = -\log\det T(u). \tag{3.41}$$

They apply the Levinson-Durbin algorithm to factor the inverse Toeplitz matrix and they use a DFT to assemble the gradient and the Hessian. The overall complexity of their scheme is equal to $\mathcal{O}(n^3)$. If our method is applied directly to the function (3.41) and fast algorithms are used, we obtain the same complexity estimate. Although their method is similar to the one proposed in Section 3.3 (if applied to this particular setting), the techniques presented above are more general. On the one hand, they can be applied to structured matrices with low displacement rank, in particular, block Hankel or block Toeplitz matrices. On the other hand, we consider the more general setting of conic optimization problems, for which the barrier function is clearly more general.

Section 3.4 is based on the technical reports [48, 146]. The numerical results of Section 3.5 were obtained during a research stay at Logilab (Université de Genève, HEC), which was supported by a travel grant from the Communauté Française de Belgique.

Chapter 4

Quadratic optimization problems

Optimization problems on cones of non-negative polynomials naturally lead to quadratic optimization problems due to spectral factorization. This chapter investigates the underlying structural property which allows us to solve quadratic problems in polynomial time.

In Section 4.1, quadratically constrained quadratic problems are formulated, and the general difficulty of solving these problems is presented. Moreover, the link with non-negative polynomials is clearly emphasized. In Section 4.2, the joint numerical range of a set of matrices is defined. Convexity of the cone derived from this mathematical object is shown to be of paramount importance for efficiently solving quadratic problems. In Section 4.3, we focus on conic optimization problems with interpolation constraints, which yield several new classes of easy quadratic problems.

4.1 Quadratically constrained quadratic programs

4.1.1 Problem formulation

Let $\{A_0, \ldots, A_k\}$ be a set of real square matrices. Consider the *quadratic optimization problem with quadratic equality constraints*:

$$\begin{array}{ll} \min & \langle A_0 x, x\rangle \\ \text{s.t.} & \langle A_i x, x\rangle = b_i, \quad i = 1, \ldots, k, \\ & x \in \mathbb{R}^n. \end{array} \tag{4.1}$$

Without loss of generality, the matrices $\{A_0, \ldots, A_k\}$ are assumed to be linearly independent and symmetric. Hereafter we only focus on the case $k \leq n$. Of course, a similar optimization problem can be formulated using complex Hermitian matrices and $x \in \mathbb{C}^n$. For this complex case, we assume that $k \leq 2n - 1$.

Quadratically constrained quadratic programs (QQP) are very often used as theoretical models, e.g. for hard combinatorial optimization problems. Therefore these problems are too hard to solve in general. Let us prove this well-known statement. Remember that the *subset sum problem*

Given non-negative integers $c, d_0, d_1, \ldots, d_{n-1}$, is there a solution to

$$\sum_{i=0}^{n-1} d_i y_i = c, \quad y_i \in \{0,1\},\ i = 0, \ldots, n-1 \quad ? \tag{4.2}$$

is a discrete optimization problem, which is known to be NP-complete [54, SP13]. Now, consider the following quadratic problem

$$\begin{array}{ll} \min & (\sum_{i=0}^{n-1} d_i y_i - c)^2 \\ \text{s.t.} & (y_i - \frac{1}{2})^2 = \frac{1}{4}, \qquad i = 0, \ldots, n-1. \end{array} \tag{4.3}$$

Clearly, (4.2) has a feasible solution if and only if the global minimum objective value in (4.3) is zero. Therefore, computing the global minimum of a QQP is a potentially NP-hard problem.

However, there exist many non-trivial QQP instances that are solvable in polynomial-time, as shown in the following example.

Example 4.1 (Trigonometric polynomials). Consider the cone $\mathcal{K}_{\mathbb{C}}$ of scalar trigonometric polynomials (see Section 2.3) and the conic optimization problem (see Section 3.1)

$$\begin{array}{ll} \min & \langle c, p \rangle_{\mathbb{R}} \\ \text{s.t.} & \langle a_i, p \rangle_{\mathbb{R}} = b_i, \quad i = 0, \ldots, k-1, \\ & p \in \mathcal{K}_{\mathbb{C}}. \end{array}$$

Using Fejér-Riesz Theorem, this conic problem can be recast as

$$\begin{array}{ll} \min & \langle T(c)q, q \rangle \\ \text{s.t.} & \langle T(a_i)q, q \rangle = b_i, \quad i = 0, \ldots, k-1, \\ & q \in \mathbb{R} \times \mathbb{C}^n, \end{array}$$

which is exactly a quadratic programming problem of the form (4.1). However, this problem can be solved in polynomial-time as shown in Chapter 3, provided that the conic formulation is used.

In this chapter we focus on several problem classes for which the QQP formulation is easy to solve. Let us describe how a quadratic problem can be approximated or even solved.

4.1.2 Approximate or exact solution?

Even if the original QQP is difficult to solve, we would like to get a good approximate solution. The standard rule of thumb is to solve the *semidefinite relaxation*

$$\begin{array}{ll} \min & \langle A_0, X \rangle \\ \text{s.t.} & \langle A_i, X \rangle = b_i, \quad i = 1, \ldots, k \\ & X \in \mathcal{S}^n_+. \end{array} \tag{4.4}$$

The feasible sets of (4.1) and (4.4) are strongly related as the latter is the convex hull of the former. Therefore, the relaxation (4.4) provides us a lower bound on the optimal value of (4.1). For particular problems, this relaxation can sometimes be strengthened but we shall not elaborate this aspect in the sequel.

The semidefinite relaxation has two important features. First its tightness can be theoretically assessed for several important problem classes [62, 105, 110, 153]. Then semidefinite relaxations of easy QQP are expected to yield the same optimal value as the original formulation. Let us illustrate our claims on two well-known examples.

Example 4.2 (maxcut). Given an undirected graph $G = (V, E)$ with non-negative weights $\{w_{ij} = w_{ji}\}_{(i,j)\in E}$, the *maximum cut problem* MAXCUT is to find the set of vertices S that maximizes the weight of the cut (S, S^c), i.e., the weights of the edges with one endpoint in S and the other in S^c. Even if all weights are equal, MAXCUT is known to be NP-hard. Assume that the set of vertices V is labelled from 0 to n. Then the quadratic formulation is

$$\begin{array}{ll} \max & \frac{1}{2}\sum_{i<j} w_{ij}(1 - y_i y_j) \\ \text{s.t.} & y_i \in \{-1, +1\}, \qquad \forall i \in V. \end{array}$$

The set S is given by $S = \{i | y_i = +1\}$. It is known that the semidefinite relaxation is optimal within a factor 0.87865 and that it is NP-hard to approximate MAXCUT within a factor $16/17 + \epsilon(\epsilon > 0)$ [60, 61].

Example 4.3 (Constrained least-squares). Let $A \in \mathbb{R}^{n\times n}$ be symmetric and $b \in \mathbb{R}^n$. The constrained least squares problem is

$$\min\{x^T A x - 2b^T x : x^T x = \alpha^2\}$$

and its homogenized formulation is

$$\begin{array}{ll} \min & \langle \begin{bmatrix} 1 & -b^T \\ -b & A \end{bmatrix} y, y\rangle \\ \text{s.t.} & \langle e_0 e_0^T y, y\rangle = 1, \\ & \langle I_{n+1} y, y\rangle = 1 + \alpha^2, \\ & y \in \mathbb{R}^{n+1}. \end{array}$$

Although this linear algebra problem is a quadratic programming problem, its SDP relaxation provides us the optimal value. Moreover, it can also be solved using standard linear algebra algorithms [66, Section 12.1].

Clearly, there must exist some intrinsic property of our set $\{A_0, \ldots, A_k\}$ that *guarantees* an exact semidefinite relaxation. In the next section, this property is shown to be related to a well-known concept of linear algebra.

4.2 Joint numerical range

4.2.1 Definition and convexity

Let $\mathcal{A} = (A_0, \ldots, A_{k-1})$ be a set of $n \times n$ real symmetric matrices. The *real joint numerical range* of $\mathcal{A}$ is defined as

$$\mathcal{W}_{\mathbb{R}}(\mathcal{A}) = \{(\langle A_0 x, x\rangle, \ldots, \langle A_{k-1} x, x\rangle) : \langle x, x\rangle = 1, x \in \mathbb{R}^n\} \subseteq \mathbb{R}^k. \tag{4.5}$$

Similarly, if the matrices $(A_0, \ldots, A_{k-1})$ are complex Hermitian matrices, the *complex joint numerical range* of $\mathcal{A}$ is

$$\mathcal{W}_{\mathbb{C}}(\mathcal{A}) = \{(\langle A_0 x, x\rangle, \ldots, \langle A_{k-1} x, x\rangle) : \langle x, x\rangle = 1, x \in \mathbb{C}^n\} \subseteq \mathbb{R}^k. \tag{4.6}$$

From now on the subscript is dropped whenever the corresponding statement is valid for both joint numerical ranges. Note that joint numerical ranges are linear image of the n-dimensional unit sphere. Consequently, they are compact and invariant with respect to unitary transformations.

Proposition 4.1. *Let $\mathcal{A} = (A_0, \ldots, A_{k-1})$ be a k-tuple of real symmetric (complex Hermitian) matrices. Then*

$$\mathcal{W}(\mathcal{A}) = \mathcal{W}(U^*A_0U, \ldots, U^*A_{k-1}U)$$

for any unitary matrix U.

From an historical point of view, the *field of values* of a matrix A

$$\mathcal{F}(A) = \{\langle Ax, x\rangle_{\mathbb{C}} : x \in \mathbb{C}^n, x^*x = 1\} \tag{4.7}$$

was introduced by Toeplitz [135]. Obviously, joint numerical ranges are straightforward extensions of this concept to several matrices. To see this, consider the Hermitian decomposition

$$A = \frac{A + A^*}{2} + \jmath\frac{A - A^*}{2\jmath} = A_0 + \jmath A_1$$

of the matrix A. Because A_0 and A_1 are Hermitian, we obtain the identity

$$\langle Ax, x\rangle_{\mathbb{C}} = \langle A_0x, x\rangle_{\mathbb{C}} + \jmath\langle A_1x, x\rangle_{\mathbb{C}}, \quad \forall x \in \mathbb{C}^n.$$

By considering the real and complex parts of $\langle Ax, x\rangle_{\mathbb{C}}$ separately, we get that $\mathcal{F}(A)$ is equivalent to $\mathcal{W}(A_0, A_1)$, which is the joint numerical range of two matrices. Toeplitz proved that the field of values has a convex outer boundary, and a short time later Hausdorff showed that it is convex [72]. It is therefore not surprising that the main geometrical issue about joint numerical ranges is related to convexity. Let us summarize a few facts about convexity of $\mathcal{W}_{\mathbb{R}}$ and $\mathcal{W}_{\mathbb{C}}$.

First, the constraint $\langle x, x\rangle = 1$ can be rewritten as $\langle I_nx, x\rangle = 1$ so that we get the next proposition.

Proposition 4.2. *Let $\mathcal{A} = (A_0, \ldots, A_{k-1})$ be a k-tuple of real symmetric (complex Hermitian) matrices. Then $\mathcal{W}(\mathcal{A})$ is convex if and only if $\mathcal{W}(I_n, A_0, \ldots, A_{k-1})$ is convex.*

Moreover convexity is preserved under linear transformation so that the following proposition is merely an observation.

Proposition 4.3. *Let $\mathcal{A} = (A_0, \ldots, A_{k-1})$ be a k-tuple of real symmetric (complex Hermitian) matrices. Then*

1. *Let $\{B_0, \ldots, B_{s-1}\}$ be a basis for $\operatorname{span}\{A_0, \ldots, A_{k-1}\}$. Then $\mathcal{W}(\mathcal{A})$ is convex if and only if $\mathcal{W}(B_0, \ldots, B_{s-1})$ is convex.*

2. *Suppose $\mathcal{W}(\mathcal{A})$ is convex. If $B_j \in \operatorname{span}\{I_n, A_0, \ldots, A_{k-1}\}$ for $0 \le j \le s-1$, then $\mathcal{W}(B_0, \ldots, B_{s-1})$ is convex.*

Table 4.1. *Convexity (C) or non-convexity (NC) of* $\mathcal{W}_{\mathbb{R}}(\mathcal{A})$

	$k=1$	$k=2$	$k\geq 3$
$n=1$	C	C	C
$n=2$	C	NC	NC
$n\geq 3$	C	C	NC

Table 4.2. *Convexity (C) or non-convexity (NC) of* $\mathcal{W}_{\mathbb{C}}(\mathcal{A})$

	$k=1$	$k=2$	$k=3$	$k\geq 4$
$n=1$	C	C	C	C
$n=2$	C	C	NC	NC
$n\geq 3$	C	C	C	NC

This proposition implies that only linearly independent matrices have to be considered for studying convexity of $\mathcal{W}(\mathcal{A})$. Hereafter this assumption is always implicitly made.

In addition, convexity of $\mathcal{W}(\mathcal{A})$ with *arbitrary* matrices $\{A_i\}_{i=0}^{k-1}$ depends on the dimension n and on the number of matrices k. Generic convexity statements [82, page 86], which do not rely on the matrix structure, are summarized in Tables 4.1 and 4.2. Note that real and complex joint numerical ranges have different convexity results.

Although these general facts appear to be quite restrictive, there exist families of matrices for which convexity holds. However, the number of linearly independent matrices in these families is bounded by a function of the problem dimension. This feature originates from a geometric fact. Given a set $\{Q_0,\ldots,Q_{q-1}\}$ of linearly independent symmetric matrices, the mapping

$$\mathcal{Q}:\mathbb{R}^p\to\mathbb{R}^q : x\to\langle Q_0x,x\rangle,\ldots,\langle Q_{q-1}x,x\rangle$$

is a nonlinear description of a p-dimensional surface immersed in a q-dimensional space. Consequently, this surface is not convex for $q>p$ in general. This result can be applied to joint numerical ranges and the resulting bounds are confirmed by more theoretical results on orbits and tangent spaces [97, Theorems 2.3 and 5.2]. For real joint numerical ranges, we must have

$$\dim \operatorname{span}\{I,A_0,\ldots,A_{k-1}\}\leq n,$$

whereas, for complex joint numerical ranges, this bound is

$$\dim \operatorname{span}\{I,A_0,\ldots,A_{k-1}\}\leq 2n-1.$$

Finally convexity of $\mathcal{W}(\mathcal{A})$ depends on some structural property of the set $\mathcal{A}=(A_0,\ldots,A_{k-1})$, as stated in the following theorem.

Theorem 4.4. *Let* $\mathcal{A}=(A_0,\ldots,A_{k-1})$ *be a* k*-tuple of real symmetric (complex Hermitian) matrices. Then* $\mathcal{W}(\mathcal{A})$ *is convex if and only if, for any* $X\in\mathcal{S}_+^n$ *(*$X\in\mathcal{H}_+^n$*) with* $\operatorname{Trace} X=1$*, there exists a unit vector* $x\in\mathbb{R}^n$ *(*$x\in\mathbb{C}^n$*) so that* $\langle A_ix,x\rangle=\langle A_i,X\rangle$ *for* $i=0,\ldots,k-1$*.*

Proof. If part. Let x_v and x_w be two vectors of $\mathbb{R}^n$ ($\mathbb{C}^n$) such that $\langle x_v, x_v\rangle = \langle x_w, x_w\rangle = 1$. Then the vectors v and w defined by

$$v_i = \langle A_i x_v, x_v\rangle, \quad w_i = \langle A_i x_w, x_w\rangle, \quad i = 0, \ldots, k-1$$

belongs to $\mathcal{A}$. For any $\alpha \in [0,1]$ the matrix $X = \alpha x_v x_v^* + (1-\alpha) x_w x_w^*$ is positive semidefinite with unit trace. Consequently, we obtain that $y = \alpha v + (1-\alpha) w \in \mathcal{A}$. Indeed, our hypothesis implies the existence of a vector $x \in \mathbb{R}^n$ ($\mathbb{C}^n$) such that

$$y_i = \langle A_i, X\rangle = \langle A_i x, x\rangle, \quad i = 0, \ldots, k-1.$$

Only if part. Assume that $\mathcal{W}(\mathcal{A})$ is convex and let $X \in \mathcal{S}_+^n$ ($X \in \mathcal{H}_+^n$) be an arbitrary matrix such that $\operatorname{Trace} X = 1$. The eigenvectors and eigenvalues of X are given by $\{x_j\}_{j=0}^{n-1} \subset \mathbb{R}^n$ ($\{x_j\}_{j=0}^{n-1} \subset \mathbb{C}^n$) and $\{\lambda_j\}_{j=0}^{n-1} \subset \mathbb{R}$, respectively. Note that $\sum_{j=0}^{n-1} \lambda_j = 1$. By convexity, we obtain the existence of $x \in \mathbb{R}^n$ ($\mathbb{C}^n$) such that the identities

$$\langle A_i, X\rangle = \langle A_i, \sum_{j=0}^{n-1} \lambda_j x_j x_j^*\rangle = \sum_{j=0}^{n-1} \lambda_j \langle A_i, x_j x_j^*\rangle = \langle A_i, xx^*\rangle$$

hold for all i. □

The first interesting family of linearly independent matrices with convex joint numerical ranges consists of (pairwise) commuting Hermitian matrices. This is a direct consequence of Proposition 4.1 and Theorem 4.4. Indeed, sets of commuting Hermitian matrices are characterized by the existence of a unitary matrix that simultaneously diagonalize all matrices, see [81, Theorem 2.3.3]. For diagonal matrices, the joint numerical range is obviously convex because of Theorem 4.4. Therefore, we have proved the next statement.

Proposition 4.5. *If $\mathcal{A} = (A_0, \ldots, A_{k-1})$ is a k-tuple of (pairwise) commuting Hermitian matrices, then $\mathcal{W}(\mathcal{A})$ is a convex set for all k and n.*

Moreover, the maximal number of linearly independent commuting matrices is equal to n. Since the unit matrix is diagonal, this family is maximal for $\mathcal{W}_{\mathbb{R}}(\mathcal{A})$. This family has been extended so as to obtain a maximal linearly independent family for $\mathcal{W}_{\mathbb{C}}(\mathcal{A})$.

Theorem 4.6 ([97, Theorem 3.1]). *Let $\mathcal{A} = (A_0, \ldots, A_{k-1})$ be a k-tuple of Hermitian matrices. If there exists a unitary U such that $U^* A_j U$ is a symmetric tridiagonal matrix for all j, then $\mathcal{W}_{\mathbb{C}}(\mathcal{A})$ is convex.*

By combining Proposition 4.1 and Proposition 4.3, other maximal families with convex joint numerical ranges have been discovered.

Theorem 4.7 ([97, Theorem 5.1]). *Let $\mathcal{A} = (A_0, \ldots, A_{k-1})$ be a k-tuple of real symmetric matrices. Suppose that $n \geq 2$ and*

$$S = \{\alpha I_n + \begin{pmatrix} 0 & x^T \\ x & 0 \end{pmatrix} : \alpha \in \mathbb{R}, x \in \mathbb{R}^{n-1}\}, \quad \text{or} \quad S = \{\alpha \oplus \beta I_{n-1} : \alpha, \beta \in \mathbb{R}\}.$$

If there exists a real orthogonal Q such that $Q^T A_j Q \in S$ for all j, then $\mathcal{W}_{\mathbb{R}}(\mathcal{A})$ is convex.

Theorem 4.8 ([97, Theorem 2.2]). *Let $\mathcal{A} = (A_0, \ldots, A_{k-1})$ be a k-tuple of Hermitian matrices. Suppose that $n \geq 2$ and*

$$S = \{\alpha I_n + \begin{pmatrix} 0 & x^* \\ x & 0 \end{pmatrix} : \alpha \in \mathbb{R}, x \in \mathbb{C}^{n-1}\}.$$

If there exists a unitary U such that $U^* A_j U \in S$ for all j, then $\mathcal{W}_{\mathbb{C}}(\mathcal{A})$ is convex.

In general, checking whether $(A_0, \ldots, A_{k-1})$ belongs to one of the previous families requires some extra computational work, except for diagonal matrices. In this case, the matrix structure is a direct certificate of convex joint numerical ranges. Another matrix structure for which convexity of the joint numerical range is directly visible is the Toeplitz structure. Both real symmetric and complex Hermitian Toeplitz matrices are covered by the following theorems, the proofs of which are related to non-negative polynomials.

Theorem 4.9. *Let $\mathcal{A} = (T_0, \ldots, T_{k-1})$ be a k-tuple of real symmetric Toeplitz matrices. Then $\mathcal{W}_{\mathbb{R}}(\mathcal{A})$ is convex.*

Proof. Because of Proposition 4.3 the most general case corresponds to the canonical basis of symmetric Toeplitz matrices, i.e., $k = n$ and $T_i = T(e_i)$ for all i. Let $v, w \in \mathcal{W}_{\mathbb{R}}(\mathcal{A})$. Then we have

$$\begin{aligned} v_i &= \langle T(e_i)x_v, x_v\rangle = \langle T(e_i), x_v x_v^T\rangle, \quad i = 0, \ldots, n-1 \\ w_i &= \langle T(e_i)x_w, x_w\rangle = \langle T(e_i), x_w x_w^T\rangle, \quad i = 0, \ldots, n-1 \end{aligned}$$

for some x_v and $x_w \in \mathbb{R}^n$. For any $\alpha \in [0, 1]$, let us consider $y = \alpha v + (1-\alpha)w$. Of course, we get

$$\begin{aligned} y_i &= \langle T(e_i), \alpha x_v x_v^T + (1-\alpha) x_w x_w^T\rangle \\ &= \langle e_i, T^*(\alpha x_v x_v^T + (1-\alpha) x_w x_w^T)\rangle \end{aligned}$$

for all i. From the Fejér-Riesz Theorem, there exist $x \in \mathbb{R}^n$ such that

$$T^*(\alpha x_v x_v^T + (1-\alpha) x_w x_w^T) = T^*(xx^T).$$

Indeed, the vector x is a spectral factor of the non-negative cosine polynomial $p(z) = \pi_n(z)^*(\alpha x_v x_v^T + (1-\alpha) x_w x_w^T)\pi_n(z)$. Because the identity matrix belongs to the basis $\{T_i\}_i$, we conclude that $y \in \mathcal{W}_{\mathbb{R}}(\mathcal{A})$. □

Theorem 4.10. *Let $\mathcal{A} = (T_0, \ldots, T_{k-1})$ be a k-tuple of complex Hermitian Toeplitz matrices. Then $\mathcal{W}_{\mathbb{C}}(\mathcal{A})$ is convex.*

Proof. Because of Proposition 4.3 we focus on the canonical basis of Hermitian Toeplitz matrices, say $\{T_i = T(s_i)\}_{i=0}^{2n-2}$. Given $v, w \in \mathcal{W}_{\mathbb{C}}(\mathcal{A})$ we have that

$$\begin{aligned} v_i &= \langle T(e_i)x_v, x_v\rangle_{\mathbb{R}} = \langle T(e_i), x_v x_v^*\rangle_{\mathbb{R}}, \quad i = 0, \ldots, 2n-2 \\ w_i &= \langle T(e_i)x_w, x_w\rangle_{\mathbb{R}} = \langle T(e_i), x_w x_w^*\rangle_{\mathbb{R}}, \quad i = 0, \ldots, 2n-2 \end{aligned}$$

for some x_v and $x_w \in \mathbb{C}^n$. For any $\alpha \in [0, 1]$, set $y = \alpha v + (1-\alpha)w$ and note that

$$\begin{aligned} y_i &= \langle T(e_i), \alpha x_v x_v^* + (1-\alpha) x_w x_w^*\rangle_{\mathbb{R}} \\ &= \langle e_i, T^*(\alpha x_v x_v^* + (1-\alpha) x_w x_w^*)\rangle_{\mathbb{C}} \end{aligned}$$

for all i. From the Fejér-Riesz Theorem, there exist $x \in \mathbb{C}^n$ such that

$$T^*(\alpha x_v x_v^* + (1-\alpha) x_w x_w^*) = T^*(xx^*).$$

Indeed, the vector x is a spectral factor of the non-negative trigonometric polynomial $p(z) = \pi_n(z)^*(\alpha x_v x_v^* + (1-\alpha) x_w x_w^*)\pi_n(z)$. Because the identity matrix belongs to the basis $\{T_i\}_{i=0}^{2n-2}$, we conclude that $y \in \mathcal{W}_{\mathbb{R}}(\mathcal{A})$. □

These well-known convexity statements are direct consequences of the Fejér-Riesz Theorem, see e.g.,[1, Theorem 7]. We elaborate the strong relationship between non-negative polynomials and convex joint numerical ranges later in this section.

4.2.2 Associated cone and closedness

Given a set of real symmetric (complex Hermitian) matrices $\{A_0, \ldots, A_{k-1}\}$, both joint numerical ranges can be lifted into cones

$$\mathcal{K}_{\mathbb{R}}(\mathcal{A}) = \{y \in \mathbb{R}^k : y_i = \langle A_i x, x\rangle, i = 0, \ldots, k-1; x \in \mathbb{R}^n\},$$
$$\mathcal{K}_{\mathbb{C}}(\mathcal{A}) = \{y \in \mathbb{R}^k : y_i = \langle A_i x, x\rangle, i = 0, \ldots, k-1; x \in \mathbb{C}^n\}.$$

As before, the subscript is dropped whenever the corresponding statement is valid for both cones. Of course, we get

$$\mathcal{W}(\mathcal{A}) \equiv \mathcal{K}(\mathcal{A} \cup I_n) \cap \{(y, \lambda) : \lambda = 1; y \in \mathbb{R}^k\}.$$

Therefore, convexity of $\mathcal{K}(\mathcal{A})$ is inherited from convexity of $\mathcal{W}(\mathcal{A})$.

Proposition 4.11. *Let $\mathcal{A} = (A_0, \ldots, A_{k-1})$ be a k-tuple of real symmetric (complex Hermitian) matrices. Then $\mathcal{W}(\mathcal{A})$ is convex if and only if $\mathcal{K}(\mathcal{A} \cup I_n)$ is convex. Moreover, if $\mathcal{K}(\mathcal{A} \cup I_n)$ is convex, then $\mathcal{K}(\mathcal{A})$ is convex.*

Proof. From Proposition 4.2, convexity of $\mathcal{W}(\mathcal{A})$ is equivalent to convexity of $\mathcal{W}(\mathcal{A} \cup I_n)$. Clearly, the set $\mathcal{W}(\mathcal{A} \cup I_n)$ is convex if and only if the lifted cone $\mathcal{K}(\mathcal{A} \cup I_n)$ is convex. Finally, our last statement follows from the fact that projections preserve convexity. □

Note that convexity of $\mathcal{K}(\mathcal{A})$ is also preserved under unitary transformation and linear mappings.

Proposition 4.12. *Let $\mathcal{A} = (A_0, \ldots, A_{k-1})$ be a k-tuple of real symmetric (complex Hermitian) matrices. Then*

$$\mathcal{K}(\mathcal{A}) = \mathcal{K}(U^* A_0 U, \ldots, U^* A_{k-1} U)$$

for any unitary matrix U.

Proposition 4.13. *Let $\mathcal{A} = (A_0, \ldots, A_{k-1})$ be a k-tuple of real symmetric (complex Hermitian) matrices. Then*

1. *Let $\{B_0, \ldots, B_{s-1}\}$ be a basis for $\operatorname{span}\{A_0, \ldots, A_{k-1}\}$. Then $\mathcal{K}(\mathcal{A})$ is convex if and only if $\mathcal{K}(B_0, \ldots, B_{s-1})$ is convex.*
2. *Suppose $\mathcal{K}(\mathcal{A})$ is convex. If $B_j \in \operatorname{span}\{I_n, A_0, \ldots, A_{k-1}\}$ for $0 \le j \le s-1$, then $\mathcal{K}(B_0, \ldots, B_{s-1})$ is convex.*

However, there exist convex cones $\mathcal{K}(\mathcal{A})$ for which the associated joint numerical range $\mathcal{W}(\mathcal{A})$ is *not* convex. In general, convexity of $\mathcal{K}(\mathcal{A})$ is characterized as follows.

Theorem 4.14. *Let $\mathcal{A} = (A_0, \ldots, A_{k-1})$ be a k-tuple of real symmetric (complex Hermitian) matrices. Then $\mathcal{K}(\mathcal{A})$ is convex if and only if, for any $X \in \mathcal{S}_+^n$ ($X \in \mathcal{H}_+^n$), there exists a vector $x \in \mathbb{R}^n$ ($x \in \mathbb{C}^n$) such that $\langle A_i x, x\rangle = \langle A_i, X\rangle$ for $i = 0, \ldots, k-1$.*

Since the convex hulls of $\mathcal{K}_{\mathbb{R}}(\mathcal{A})$ and $\mathcal{K}_{\mathbb{C}}(\mathcal{A})$ are given by

$$\operatorname{co}\mathcal{K}_{\mathbb{R}}(\mathcal{A}) = \{y \in \mathbb{R}^k : y_i = \langle A_k, X\rangle, i = 0, \ldots, k-1; X \in \mathcal{S}_+^n\},$$
$$\operatorname{co}\mathcal{K}_{\mathbb{C}}(\mathcal{A}) = \{y \in \mathbb{R}^k : y_i = \langle A_k, X\rangle, i = 0, \ldots, k-1; X \in \mathcal{H}_+^n\},$$

this theorem can be restated as follows.

Theorem 4.15. *Let $\mathcal{A} = (A_0, \ldots, A_{k-1})$ be a k-tuple of real symmetric (complex Hermitian) matrices. Then $\mathcal{K}(\mathcal{A}) = \operatorname{co}\mathcal{K}(\mathcal{A})$ if and only if the set $\mathcal{K}(\mathcal{A})$ is convex.*

Proof. If part. In any case, the inclusion $\mathcal{K}(\mathcal{A}) \subseteq \operatorname{co}\mathcal{K}(\mathcal{A})$ is easily checked. In order to prove that $\operatorname{co}\mathcal{K}(\mathcal{A}) \subseteq \mathcal{K}(\mathcal{A})$, let y be an arbitrary element of $\operatorname{co}\mathcal{K}(\mathcal{A})$. Then there exists a positive semidefinite matrix X such that:

$$y_i = \langle A_i, X\rangle, \quad i = 0, \ldots, k-1, \quad X = \sum_{j=0}^{n-1} x_j x_j^*.$$

Define the vectors $\{u^{(j)}\}_{j=0}^{n-1}$ as follows

$$u_i^{(j)} = \langle A_i x_j, x_j\rangle, \quad j = 0, \ldots, n-1;\ i = 0, \ldots, k-1.$$

Note that $u^{(j)} \in \mathcal{K}(\mathcal{A})$ for all j. Since the cone $\mathcal{K}(\mathcal{A})$ is convex by assumption, we conclude that $y = \sum_{j=0}^{n-1} u^{(j)} \in \mathcal{K}(\mathcal{A})$. Thus $\operatorname{co}\mathcal{K}(\mathcal{A}) \subseteq \mathcal{K}(\mathcal{A})$

Only if part. This correspond to the definition of convexity. □

Consequently, the next convexity result is straightforward to prove.

Theorem 4.16. *Let $\mathcal{A} = (H_0, \ldots, H_{k-1})$ be a k-tuple of real symmetric upper-left triangular Hankel matrices. Then $\mathcal{K}_{\mathbb{R}}(\mathcal{A})$ is convex.*

Proof. For any $X \in \mathcal{S}_+^n$, the nonlinear system of equations

$$\langle A_i x, x\rangle = \langle A_i, X\rangle, \quad i = 0, \ldots, k-1. \tag{4.8}$$

has a triangular structure so that it can be solved recursively. Theorem 4.15 is then invoked to complete the proof. □

Example 4.4 (Upper Hankel matrices). If the matrices $\{H_i\}_{i=0}^{n-1}$ span the set of $n \times n$ upper-left triangular Hankel matrices, then the set $\mathcal{K}(\{H_i\}_{i=0}^{n-1})$ is convex. For instance, choose $n = 3$ and

$$H_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad H_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad H_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

The triangular system (4.8) can be solved sequentially as follows

$$\begin{aligned} x_0^2 = X_{00} &\to x_0, \\ 2x_0x_1 = X_{01} + X_{10} = 2X_{01} &\to x_1, \\ 2x_0x_2 + x_1^2 = X_{02} + X_{20} + X_{11} = 2X_{02} + X_{11} &\to x_2. \end{aligned}$$

Note that this set is maximal for $\mathcal{K}_{\mathbb{R}}(\mathcal{A})$.

The last theorem can be generalized to arbitrary Hankel matrices [1, Theorem 7], provided that complex vectors are considered.

Theorem 4.17. *Let $\mathcal{A} = (H_0, \ldots, H_{k-1})$ be a k-tuple of real symmetric Hankel matrices. Then $\mathcal{K}_{\mathbb{C}}(\mathcal{A})$ is convex.*

Proof. Because of Proposition 4.3, the most general case corresponds to the canonical basis of symmetric Hankel matrices, i.e., $k = 2n-1$ and $H_i = H(e_i)$ for all i. Let $v, w \in \mathcal{W}_{\mathbb{C}}(\mathcal{A})$. Then we have

$$\begin{aligned} v_i = \langle H(e_i)x_v, x_v\rangle = \langle H(e_i), x_v x_v^*\rangle, \quad i = 0, \ldots, n-1 \\ w_i = \langle H(e_i)x_w, x_w\rangle = \langle H(e_i), x_w x_w^*\rangle, \quad i = 0, \ldots, n-1 \end{aligned}$$

for some x_v and $x_w \in \mathbb{C}^n$. For any $\alpha \in [0,1]$, let us consider $y = \alpha v + (1-\alpha)w$. Of course, we get

$$\begin{aligned} y_i &= \langle H(e_i), \alpha x_v x_v^* + (1-\alpha)x_w x_w^*\rangle \\ &= \langle e_i, H^*(\alpha x_v x_v^* + (1-\alpha)x_w x_w^*)\rangle \end{aligned}$$

for all i. From spectral factorization, there exist $x \in \mathbb{C}^n$ such that

$$H^*(\alpha x_v x_v^* + (1-\alpha)x_w x_w^*) = H^*(xx^*).$$

Indeed, the vector x is a spectral factor of the non-negative polynomial $p(t) = \pi_n(t)^T(\alpha x_v x_v^* + (1-\alpha)x_w x_w^*)\pi_n(t)$. Therefore $y \in \mathcal{W}_{\mathbb{C}}(\mathcal{A})$. □

Although $\mathcal{K}_{\mathbb{R}}(H_0, \ldots, H_{k-1})$ is not convex, its convex hull has a Caratheodory number [126] equal to 2. Thus we obtain the following result.

Theorem 4.18. *Let $(H_0, \ldots, H_{k-1})$ be a k-tuple of real symmetric Hankel matrices. Then $\mathcal{K}_{\mathbb{R}}(I_2 \otimes H_0, \ldots, I_2 \otimes H_{k-1})$ is convex.*

Proof. Because of Proposition 4.3, the most general case corresponds to the canonical basis of symmetric Hankel matrices, i.e., $k = 2n-1$ and $H_i = H(e_i)$ for all i. This result follows from Theorem 4.17 and from the identification of $\mathbb{C}^n$ with $\mathbb{R}^{2n}$. To see this, write $x \in \mathbb{C}^n$ as $x = \operatorname{Re} x + \jmath \operatorname{Im} x$. Then the identities

$$\begin{aligned} \langle H(e_i)x, x\rangle &= \langle H(e_i)(\operatorname{Re} x + \jmath \operatorname{Im} x), \operatorname{Re} x + \jmath \operatorname{Im} x\rangle \\ &= \langle H(e_i)\operatorname{Re} x, \operatorname{Re} x\rangle + \langle H(e_i)\operatorname{Im} x, \operatorname{Im} x\rangle \\ &= \langle I_2 \otimes H(e_i) \begin{bmatrix} \operatorname{Re} x \\ \operatorname{Im} x \end{bmatrix}, \begin{bmatrix} \operatorname{Re} x \\ \operatorname{Im} x \end{bmatrix}\rangle, \end{aligned}$$

hold for any $x \in \mathbb{C}$ and $0 \le i \le 2n-2$. Consequently $\mathcal{K}_{\mathbb{R}}(I_2 \otimes H_0, \ldots, I_2 \otimes H_{k-1})$ is convex. □

Given a conic optimization problem formulated on a cone $\mathcal{K}$, a desirable property of the associated convex set is closedness. If the problem is feasible and bounded, this property guarantees that the minimum is well defined. Therefore, given a set $\{A_0, \ldots, A_k\}$ of real symmetric (complex Hermitian) matrices, we would like to guarantee that the corresponding cone $\mathcal{K}_{\mathbb{C}}(\mathcal{A})$ is closed.

From now on, the cone $\mathcal{K}(\mathcal{A})$ is assumed to be convex. By Theorem 4.15, $\mathcal{K}(\mathcal{A})$ is thus the image of the set of positive semi-definite matrices under the linear mapping

$$\mathcal{L} : X \to \langle A_0, X\rangle, \ldots, \langle A_{k-1}, X\rangle. \tag{4.9}$$

It is well known that the linear mapping of convex sets preserves relative interiors, but may not preserve closures. In general, given a convex set C and a linear mapping $\mathcal{L}$, one has [126, Theorem 6.6]

$$\operatorname{rint} \mathcal{L}C = \mathcal{L} \operatorname{rint} C, \tag{4.10}$$

$$\operatorname{cl} \mathcal{L}C \supseteq \mathcal{L} \operatorname{cl} C. \tag{4.11}$$

Given a convex set $Q \subset \mathbb{R}^n$, the *recession cone* of Q is the set $\{v \in \mathbb{R}^n : Q + v \subseteq \operatorname{cl} Q\}$. The next result is a sufficient condition to get equality in (4.11).

Theorem 4.19 ([126, Theorem 9.1]). *Let C be a non-empty closed convex set in $\mathbb{R}^n$, and let $\mathcal{L}$ be a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. If $z = 0$ is the only z in the recession cone of C such that $\mathcal{L}z = 0$, then $\mathcal{L}C$ is closed.*

Cones of semidefinite matrices are not only closed and convex, but also their own recession cones. Therefore we obtain the following corollary.

Corollary 4.20. *Let $A_0, \ldots, A_{k-1}$ be k matrices defining the cone $\mathcal{K}(\mathcal{A})$ and assume that this cone $\mathcal{K}(\mathcal{A})$ is convex. If there exists a vector $\alpha \in \mathbb{R}^k$ such that*

$$\sum_{i=0}^{k-1} \alpha_i A_i \succ 0, \tag{4.12}$$

then the convex cone $\mathcal{K}(\mathcal{A})$ is pointed and closed.

Proof. Let $\mathcal{L}$ be the linear operator defined in (4.9) and α be a non-zero vector such that $\sum_{i=0}^{k-1} \alpha_i A_i$ is positive definite. Assume that the matrix Z is positive semidefinite and satisfies the identity $\mathcal{L}Z = 0$. Then, we have that

$$0 = \sum_{i=0}^{k-1} \alpha_i \langle A_i, Z\rangle = \langle \sum_{i=0}^{k-1} \alpha_i A_i, Z\rangle.$$

Consequently, the matrix Z is necessarily equal to 0 and closedness follows from Theorem 4.19.

From inequality (4.12), the matrix A_0 can be assumed to be positive definite (otherwise a nonsingular linear operator transforms the original data to data having this property). If $y \in \mathcal{K}(\mathcal{A}) \cap -\mathcal{K}(\mathcal{A})$, we have $\langle A_0 x, x\rangle = 0$ for some vector x. This is possible if and only if $x = 0$. Thus, $y = 0$ and $\mathcal{K}(\mathcal{A})$ is pointed. □

Remark. Inequality (4.12) states that the dual cone $\mathcal{K}^\star(\mathcal{A})$ has a non-empty interior.

4.2.3 Consequences for quadratic optimization

Consider the quadratically constrained quadratic program (4.1) and denote by $\mathcal{A} = \{A_0, A_1, \ldots, A_k\}$ the corresponding set of matrices. This set also defines a joint numerical range $\mathcal{W}(A)$ and a cone $\mathcal{K}(\mathcal{A})$. Our main result of this section characterizes a wide class of QQP for which the semidefinite relaxation yields a tight solution.

Theorem 4.21. *If $\mathcal{K}(\mathcal{A})$ is convex, then the semidefinite relaxation* (4.4) *is completely equivalent to the quadratically constrained quadratic program* (4.1)*.*

Proof. Our convexity assumption on $\mathcal{K}(\mathcal{A})$ implies that $\mathcal{K}(\mathcal{A})$ is equivalent to its convex hull $\operatorname{co}\mathcal{K}(\mathcal{A})$ by Proposition 4.3. Consider the minimization problem (4.4) and let t be the objective value at an arbitrary point $X \in \mathcal{S}^n_+$. Since $\mathcal{K}(\mathcal{A}) = \operatorname{co}\mathcal{K}(\mathcal{A})$, there exists a vector x such that

$$\begin{aligned} t = \langle A_0, X\rangle &= \langle A_0, xx^T\rangle, \\ \langle A_i, X\rangle &= \langle A_i, xx^T\rangle, \quad i = 1, \ldots, k. \end{aligned}$$

This means that the value t can also be reached by making use of a rank-one positive semidefinite matrix. In particular, this reasoning is valid at the optimal point, provided that it exists. □

By combining several previous results together, we obtain a sufficient characterization of QQP families with exact semidefinite relaxation.

Theorem 4.22. *Let $\mathcal{B} = \{B_0, \ldots, B_{p-1}\}, p > 0$, be a set of linearly independent real symmetric matrices. If $\mathcal{K}_{\mathbb{R}}(\mathcal{B})$ is convex, then the quadratic optimization problem* (4.1) *defined by the matrices*

$$A_i \in U^T \mathit{span}(\mathcal{B})U, \quad i = 0, \ldots, k, \quad U^T U = I_n, \tag{4.13}$$

$$\sum_{i=0}^{k} \alpha_i A_i \succ 0, \quad \alpha_0 > 0 \tag{4.14}$$

can be solved in polynomial time (up to an accuracy $\epsilon > 0$). Moreover, there is no duality gap.

Proof. As convexity is not affected by linear substitution or congruence, our statement follows from Theorem 4.21. Assume that the cone $\mathcal{K}_{\mathbb{R}}(\mathcal{B})$ is convex. By Proposition 4.13 and (4.13), $\mathcal{K}_{\mathbb{R}}(A_0, A_1, \ldots, A_m)$ is convex. Condition (4.14) guarantees that it is pointed and closed. It remains to use Theorem 4.21. The absence of a duality gap follows from (4.14), which states that the dual feasible set has a non-empty interior. □

Theorem 4.23. *Let $\mathcal{B} = \{B_0, \ldots, B_{p-1}\}, p > 0$, be a set of linearly independent complex Hermitian matrices. If $\mathcal{K}_{\mathbb{C}}(\mathcal{B})$ is convex, then the quadratic optimization problem* (4.1) *defined by the matrices*

$$A_i \in U^* \mathit{span}(\mathcal{B})U, \quad i = 0, \ldots, k, \quad U^* U = I_n, \tag{4.15}$$

$$\sum_{i=0}^{k} \alpha_i A_i \succ 0, \quad \alpha_0 > 0 \tag{4.16}$$

can be solved in polynomial time (up to an accuracy $\epsilon > 0$). Moreover, there is no duality gap.

Remark. Our result is only a sufficient characterization. Indeed, some QQP may have an exact semidefinite relaxation because of their objective functions. This situation is completely different from the one considered above as it follows from an appropriate choice of the objective.

Example 4.5 (Minimum eigenvalue). Let $A = A^T$ be a symmetric matrix. It is well known that the smallest eigenvalue of A is obtained from the Rayleigh quotient

$$\lambda_{\min}(A) = \min\{\langle Ax, x\rangle : \|x\|_2 = 1\}.$$

Since A is not necessarily positive semidefinite, we minimize a non-convex function on a non-convex set. However, the underlying numerical range $\mathcal{W}_{\mathbb{R}}(A)$ is convex. This problem can thus be solved exactly using the semidefinite relaxation

$$\lambda_{\min}(A) = \min\{\langle A, X\rangle : \|X\|_2 = 1, X \succeq 0\}.$$

This is the well-known variational characterization of the minimal eigenvalue.

Example 4.6 (Constrained least-squares). The constrained least-squares problem of Example 4.3 can be relaxed as follows

$$\begin{array}{ll} \min & \langle \begin{bmatrix} 1 & -b^T \\ -b & A \end{bmatrix}, Y\rangle \\ \text{s.t.} & \langle e_0 e_0^T, Y\rangle = 1, \\ & \langle I_{n+1}, Y\rangle = 1 + \alpha^2, \\ & Y \in \mathcal{S}_+^{n+1}. \end{array}$$

The underlying joint numerical range

$$\mathcal{W}_{\mathbb{R}}(\begin{bmatrix} 1 & -b^T \\ -b & A \end{bmatrix}, e_0 e_0)$$

is convex for $n \geq 2$ as predicted in Table 4.1. Therefore, the above relaxation is tight as stated in Theorem 4.21. Note that the dual problem is given by

$$\begin{array}{ll} \max & y_1 + (1+\alpha^2) y_2 \\ \text{s.t.} & \begin{bmatrix} 1 & -b^T \\ -b & A \end{bmatrix} - y_1 e_0 e_0^T - y_2 I_{n+1} \succeq 0. \end{array}$$

Since both primal and dual problems are strictly feasible, there is no duality gap. A similar reasoning explains why the *trust region subproblem* can be solved efficiently [66, 113].

4.2.4 Link with non-negative polynomials

Theorems 4.10 and 4.21 clearly explain why the relaxation of Example 4.1, which involves non-negative trigonometric polynomials, is exact. In order to clarify the link between our previous results and non-negative polynomials, let us change our notation and assume that the matrices $\{A_i\}_{i=0}^k$ are generated from a linear operator Λ and a set of vectors, i.e.,

$$A_i = \Lambda(s_i), \quad i = 0, \ldots, k.$$

As before, the linear operator dual to Λ (with respect to the appropriate real-valued inner products) is denoted by Λ^*.

Our sets $\mathcal{K}(\mathcal{A})$ and $\mathcal{K}_R(\mathcal{A})$ can be rewritten using this linear operator and this set of vectors:

$$\begin{aligned}
\mathcal{K}_{\mathbb{R}}(\mathcal{A}) \equiv \mathcal{K}_{\mathbb{R}}(\Lambda) &= \{y \in \mathbb{R}^m : y_i = \langle s_i, \Lambda^*(xx^T)\rangle, x \in \mathbb{R}^n\},\\
\mathcal{K}_{\mathbb{C}}(\mathcal{A}) \equiv \mathcal{K}_{\mathbb{C}}(\Lambda) &= \{y \in \mathbb{R}^m : y_i = \langle s_i, \Lambda^*(xx^*)\rangle, x \in \mathbb{C}^n\},\\
\operatorname{co}\mathcal{K}_{\mathbb{R}}(\mathcal{A}) \equiv \operatorname{co}\mathcal{K}_{\mathbb{R}}(\Lambda) &= \{y \in \mathbb{R}^m : y_i = \langle s_i, \Lambda^*(X)\rangle, X \in \mathcal{S}^n_+\},\\
\operatorname{co}\mathcal{K}_{\mathbb{C}}(\mathcal{A}) \equiv \operatorname{co}\mathcal{K}_{\mathbb{C}}(\Lambda) &= \{y \in \mathbb{R}^m : y_i = \langle s_i, \Lambda^*(X)\rangle, X \in \mathcal{H}^n_+\}.
\end{aligned}$$

A sufficient condition for convexity of $\mathcal{K}(\Lambda)$ is that, for any $X \in \mathcal{S}^n_+$ ($X \in \mathcal{H}^n_+$), there exists a vector $x \in \mathbb{R}^n$ ($x \in \mathbb{C}^n$) such that $\Lambda^*(xx^T) = \Lambda^*(X)$ ($\Lambda^*(xx^*) = \Lambda^*(X)$).

Recall that this observation was very important to prove Theorems 4.9, 4.10, 4.17 and 4.18. Indeed, we have shown that convexity follows from the existence of spectral factors. Remember that the Hankel and Toeplitz structures are inherited from the functional basis υ and ψ, see Section 2.2 and 2.3. By modifying these basis, the matrix structure is altered but spectral factorization still holds. Consequently, a wide range of matrix families $\mathcal{A}$ with convex cone $\mathcal{K}(\mathcal{A})$ can be generated. Let us exemplify this procedure.

Example 4.7 (First-kind Chebyshev functional basis). Consider the functional basis generated by *first-kind Chebyshev polynomials*

$$T_0(t) = 1, \quad T_1(t) = t, \quad T_k(t) = 2tT_{k-1}(t) - T_{k-2}(t),\ k \geq 2$$

to describe the cone $\mathcal{K}_{\mathbb{R}}$ (scalar non-negative polynomials of degree $2n$). The vector functions $\psi(t)$ and $\upsilon(t)$ are defined by

$$\begin{aligned}
\psi(t) &= \begin{bmatrix} T_0(t), & T_1(t), & T_2(t), & \cdots, & T_n(t)\end{bmatrix}^T,\\
\upsilon(t) &= \begin{bmatrix} T_0(t), & T_1(t), & T_2(t), & \cdots, & T_{2n}(t)\end{bmatrix}^T.
\end{aligned}$$

The linear operator Λ such that $\Lambda(\upsilon(t)) = \psi(t)\psi(t)^T$, for all $t \in \mathbb{R}$, is defined by

$$\Lambda(v) = \sum_{i=0}^{2n} H(e_i)v_i + \sum_{i=0}^{n} T(e_i)v_i$$

and $\Lambda(v)$ is thus a Hankel-plus-Toeplitz matrix. By construction the cone

$$\{y \in \mathbb{R}^{2n+1} : y_i = \langle e_i, \Lambda^*(xx^*)\rangle, i = 0, \ldots, 2n; x \in \mathbb{C}^n\} \subseteq \mathbb{R}^{2n+1}$$

is convex. Note that this family of linearly independent matrices is maximal with respect to convexity.

4.3 Interpolation constraints

The problem of optimizing over cone of non-negative polynomials, subject to linear constraints on the polynomial coefficients, has already been studied in Chapter 3. In this section, we focus more specifically on scalar polynomials and interpolation constraints.

First, we recall the optimization problems of interest. Then we show that they can be reformulated as QQP that can be solved in polynomial time. Finally, we discuss how the previous results can be extended to matrix polynomials and interpolation constraints on the derivatives.

4.3.1 Problem formulation

Of course, the classes of problems to be considered in this section are particular instances of those presented in Chapter 3. Nevertheless, the next description is easier to read as it only emphasizes our formulations for scalar polynomials.

Real line

Several important optimization problems on the real line can be formulated as the following *primal* problem

$$\begin{array}{ll} \min & \langle c, p\rangle \\ \text{s.t.} & \langle a_i, p\rangle = b_i, \quad i = 0, \ldots, k-1, \\ & p \in \mathcal{K}_{\mathbb{R}}, \end{array} \tag{4.17}$$

where the matrix of constraints $A = \{a_i\}_{i=0}^{k-1} \in \mathbb{R}^{k\times(2n+1)}$ is a full row rank matrix. Clearly, the constraints $Ap = b$ are linear constraints on the coefficients of the polynomial $p(x) = \sum_{i=0}^{2n} p_i x^i$ whereas the constraint $p \in \mathcal{K}_{\mathbb{R}}$ is semi-infinite. Note that the number k of linear constraints must satisfy $1 \leq k \leq 2n+1$. Moreover, if $k = 2n+1$, (4.17) is clearly not an optimization problem.

From a computational point of view, the problem dual to (4.17) has a considerable advantage over its primal counterpart. It reads as follows

$$\begin{array}{ll} \max & \langle b, y\rangle \\ \text{s.t.} & s + \sum_{i=0}^{k-1} a_i y_i = c, \\ & s \in \mathcal{K}_{\mathbb{R}}^{\star}. \end{array} \tag{4.18}$$

Since its constraints are equivalent to $H(c - A^T y) \succeq 0$, the Hankel structure allows us to solve this dual problem efficiently, see Chapter 3.

Using Theorem 2.7, the primal optimization problem (4.17) can also be recast as a semidefinite programming problem:

$$\begin{array}{ll} \min & \langle H(c), Y\rangle \\ \text{s.t.} & \langle H(a_i), Y\rangle = b_i, \quad i = 0, \ldots, k-1, \\ & Y \in \mathcal{S}_+^{n+1}. \end{array}$$

Let us now consider interpolation constraints. Note that an interpolation constraint on a polynomial p is a linear constraint:

$$p(x_i) = \langle p, \pi_{2n}(x_i)\rangle = b_i.$$

Assume that all linear constraints of (4.17) are interpolation constraints, i.e.,

$$\langle a_i, p\rangle \doteq \langle \pi_{2n}(x_i), p\rangle = b_i, \quad i = 0, \ldots, k-1. \tag{4.19}$$

Then the dual problem (4.18) is equivalent to

$$\begin{array}{ll} \max & \langle b, y\rangle \\ \text{s.t.} & H(c) - \sum_{i=0}^{k-1} y_i H(\pi_{2n}(x_i)) \succeq 0. \end{array}$$

As the Hankel structure satisfies

$$H(\pi_{2n}(x)) = \pi_n(x)\pi_n(x)^T, \quad \forall x \in \mathbb{R},$$

we finally obtain the following formulation

$$\begin{array}{ll} \max & \langle b, y\rangle \\ \text{s.t.} & H(c) - V \operatorname{diag}(y) V^T \succeq 0, \end{array} \tag{4.20}$$

where the *Vandermonde matrix* V is defined by the nodes $\{x_0, \ldots, x_{k-1}\}$, i.e.,

$$V = \begin{bmatrix} 1 & \ldots & 1 \\ x_0 & \ldots & x_{k-1} \\ \vdots & & \vdots \\ x_0^n & \ldots & x_{k-1}^n \end{bmatrix}.$$

Assumption 4.1. The components of the vector b are strictly positive.

Remark. Since we work with non-negative polynomials, this assumption is not restrictive. If there exists an integer i such that $b_i = 0$, one can factorize $p(x)$ as $p(x) = \tilde{p}(x)(x - x_i)^2$ and rewrite the optimization problem using the polynomial $\tilde{p}(x)$.

Unit circle

Several important optimization problems on the unit circle can be formulated as the following *primal* problem

$$\begin{array}{ll} \min & \langle c, p\rangle \\ \text{s.t.} & \langle a_i, p\rangle = b_i, \quad i = 0, \ldots, k-1, \\ & p \in \mathcal{K}_{\mathbb{C}}. \end{array} \tag{4.21}$$

with linearly independent constraints. In this context, the inner product $\langle \cdot, \cdot \rangle$ must be the real-valued inner product defined in (1.4). From a computational point of view, the problem dual to (4.21) has again a considerable advantage over its primal counterpart. This dual problem reads as follows

$$\begin{array}{ll} \max & \langle b, y\rangle \\ \text{s.t.} & s + \sum_{i=0}^{k-1} y_i a_i = c, \\ & s \in \mathcal{K}_{\mathbb{C}}^{\star}. \end{array} \tag{4.22}$$

As in the real line setting, one can use the Toeplitz structure of its constraints to get fast algorithms. Using Theorem 2.16, the primal optimization problem (4.21) can be reformulated as the semidefinite programming problem

$$\begin{array}{ll} \min & \langle T(c), Y\rangle \\ \text{s.t.} & \langle T(a_i), Y\rangle = b_i, \quad i = 0, \ldots, k-1, \\ & Y \in \mathcal{H}_+^{n+1}. \end{array}$$

An interpolation constraint on the trigonometric polynomial p corresponds to

$$p(\theta_i) = \sum_{k=0}^{n} [a_k \cos(k\theta_i) + b_k \sin(k\theta_i)] = b_i \geq 0, \quad \theta_i \in [0, 2\pi],$$

and is equivalent to the linear constraint

$$\langle a_i, p \rangle \doteq p(z_i) = \langle p, \pi_n(z_i) \rangle = b_i, \quad z_i = e^{\jmath\theta_i}. \tag{4.23}$$

Note that the identity

$$T(\pi_{2n}(z)) = \pi_n(z)\pi_n(z)^*, \quad \forall z \in \mathbb{T},$$

holds for the Toeplitz structure. If all linear constraints of (4.21) are interpolation constraints, the dual can therefore be written as

$$\begin{array}{ll} \max & \langle b, y \rangle \\ \text{s.t.} & T(c) - V \operatorname{diag}(y) V^* \succeq 0, \end{array} \tag{4.24}$$

where the *Vandermonde matrix* V is defined by the points $\{z_0, \ldots, z_{k-1}\}$, i.e.,

$$V = \begin{bmatrix} 1 & \ldots & 1 \\ z_0 & \ldots & z_{k-1} \\ \vdots & & \vdots \\ z_0^n & \ldots & z_{k-1}^n \end{bmatrix}.$$

As before, we make the next non-restrictive assumption.

Assumption 4.2. The components of the vector b are strictly positive.

4.3.2 Solving the optimization problem

The specific problem structure, which originates from the interpolation constraints (4.19) or (4.23), allows us to discuss whether Assumption 1.1, the regularity assumption, is satisfied. As this assumption holds under mild hypotheses, we provide explicit solutions to three optimization problems (one interpolation constraint, two interpolation constraints, property of the objective function). In the general setting, we focus on QQP reformulations and show that the semidefinite relaxations are exact. This observation yields efficient algorithms for solving this class of problems.

Strict feasibility

Remember that a standard assumption on the primal and dual problems is the regularity assumption, also known as "strict feasibility" assumption. This theoretical hypothesis is essential for defining the primal and dual central paths appropriately, see Section 1.3, and it ensures that the optimal values of both problems coincide. This important property is highly desirable for solving our class of problems efficiently and reads as follows.

Table 4.3. *Interiors of primal and dual cones*

	$\mathcal{K} = \mathcal{K}_{\mathbb{R}}$	$\mathcal{K} = \mathcal{K}_{\mathbb{C}}$
$p \in \operatorname{int} \mathcal{K}$	$p(t) > 0, \forall t \in \mathbb{R}$ and $p_{2n} > 0$	$p(z) > 0, \forall \lvert z\rvert = 1$
$s \in \operatorname{int} \mathcal{K}^\star$	$H(s)$ is positive definite	$T(s)$ is positive definite

Assumption 4.3 (Strict feasibility). There exist points $\tilde{p} \in \operatorname{int} \mathcal{K}$, $\tilde{s} \in \operatorname{int} \mathcal{K}^\star$ and $\tilde{y} \in \mathbb{R}^k$ that satisfy the linear systems

$$\begin{aligned} \langle a_i, \tilde{p} \rangle &= b_i, \quad 0 \leq i \leq k-1, \\ \hat{s} + \sum_{i=0}^{k-1} a_i \tilde{y}_i &= c. \end{aligned}$$

As mentioned in Table 4.3, the interiors of the primal and dual cones are characterized in terms of polynomials and structured matrices, respectively. However, our particular problem classes allow us to further discuss the interpretation of the previous assumption. More specifically, we shall see that some information about strict feasibility of our problems is known in advance.

Real line First, we analyse the strict feasibility of the primal constraints. If the number of interpolation points is less or equal to $n+1$, i.e., $k \leq n+1$, it is clear that there exists a strictly positive polynomial $\tilde{p}$ such that $A\tilde{p} = b$. Assume that $k = n+1$ and let $\{l_i(x)\}_{i=0}^n$ be the set of Lagrange polynomials of degree n associated with the interpolation points. By definition, these polynomials satisfy the identities

$$l_i(x_j) = \delta_{ij}, \quad 0 \leq i, j \leq n,$$

where δ_{ij} is the well-known Kronecker delta. The polynomial $\tilde{p}(x) = \sum_{i=0}^n b_i (l_i(x))^2$ clearly satisfies all our interpolation constraints and belongs to $\operatorname{int} \mathcal{K}_{\mathbb{R}}$. For the case $k < n+1$, we can add $n+1-k$ "extra" interpolation constraints and check that the (original) primal problem is always strictly feasible. If the number of interpolation points is strictly greater that $n+1$, we cannot say anything in advance about primal strict feasibility.

Let us now analyse the strict feasibility of the dual constraints. Because of the structure of our interpolation constraints, the interior of the dual space is the set of vectors $s = c - A^T y$ such that

$$H(s) = H(c - A^T y) = H(c) - \sum_{i=0}^{k-1} y_i \pi_n(x_i) \pi_n(x_i)^T \succ 0.$$

If $k \geq n+1$, we conclude from this inequality that there always exists $s = c - A^T y \in \operatorname{int} \mathcal{K}_{\mathbb{R}}^\star$. Another simple situation arises when $c \in \operatorname{int} \mathcal{K}_{\mathbb{R}}^\star$, i.e., $H(c) \succ 0$. Then the dual problem is always strictly feasible. For instance, this situation occurs when minimizing the integral of the polynomial $p(x)$ on a finite interval $I \subset (-\infty, +\infty)$:

$$\langle c, p \rangle = \int_I p(x) dx = \sum_{i=0}^{2n} p_i \Big(\int_I x^i dx \Big),$$

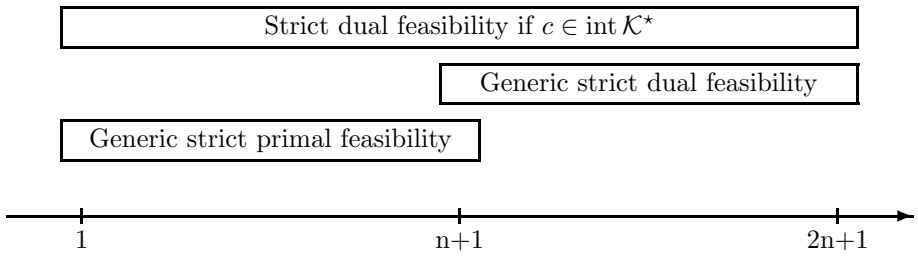

Figure 4.1. *Generic strict feasibility as a function of the number of interpolation constraints*

subject to interpolation constraints. This situation is frequent in practice and one easily checks that $c \in \operatorname{int} \mathcal{K}_{\mathbb{R}}^\star$ in this case. Indeed, the inner product $\langle c, p \rangle$ is positive for all non-zero $p \in \mathcal{K}_{\mathbb{R}}$.

Remark. If the dual problem is strictly feasible, one can always reformulate the problem in order to ensure that $c \in \operatorname{int} \mathcal{K}_{\mathbb{R}}^\star$.

We have summarized our discussion on Figure 4.1. Let us point out a remarkable property of our class of problems, which is clearly exhibited on this figure. If the number of constraints is equal to $n + 1$, both primal and dual problems are strictly feasible and this property is *independent* of the data. Except for this particular case, there usually exists a trade-off between strict primal and dual feasibility.

Unit circle Using exactly the same argument, one can show that the primal problem is always strictly feasible if the number of interpolation constraints is less or equal to $n+1$. As in the real line, there exists a trade-off between strict feasibility of the primal and dual constraints unless $k = n + 1$. If $c \in \operatorname{int} \mathcal{K}_{\mathbb{C}}^\star$, i.e., $T(c) \succ 0$, the dual problem is always strictly feasible.

Guaranteed strict feasibility Therefore, the largest class of interpolation problems on non-negative polynomials (degree $2n$ or n, in the real line or unit circle setting, respectively) for which strict feasibility holds and does not depend on the interpolation points, satisfies the following assumption.

Assumption 4.4. The number k of interpolation constraints is less or equal to $n + 1$ and the objective vector c satisfies $H(c) \succ 0$ (real line setting) or $T(c) \succ 0$ (unit circle setting), i.e., $c \in \operatorname{int} \mathcal{K}^\star$.

From now on, we focus on problems that fulfil this assumption. First, we consider several problems for which explicit solutions are easily computed from the data.

One interpolation constraint

Real line Suppose that one wants to solve the primal problem

$$\min\{\langle c, p \rangle : p(\bar{x}) = b, p \in \mathcal{K}_{\mathbb{R}}\}.$$

The dual problem reads as follows

$$\begin{array}{ll} \max & by \\ \text{s.t.} & H(c) \succeq y\pi_n(\bar{x})[\pi_n(\bar{x})]^T. \end{array}$$

Without loss of generality, the scalar b is assumed to be equal to 1. The optimal value of this problem is

$$\frac{1}{\langle H(c)^{-1}\pi_n(\bar{x}), \pi_n(\bar{x})\rangle},$$

which is equal to the optimal value of y. Using Assumption 4.4, the optimal value of p is thus given by

$$p = H^*(qq^T), \quad q = \frac{H(c)^{-1}\pi_n(\bar{x})}{\langle H(c)^{-1}\pi_n(\bar{x}), \pi_n(\bar{x})\rangle}.$$

One can check that

$$\begin{aligned} p(\bar{x}) = \langle \pi_{2n}(\bar{x}), p\rangle = \langle \pi_n(\bar{x})\pi_n(\bar{x}), qq^T\rangle = (\langle \pi_n(\bar{x}), q\rangle)^2 = 1, \\ \langle c, p\rangle = \langle H(c), qq^T\rangle = \frac{1}{\langle H(c)^{-1}\pi_n(\bar{x}), \pi_n(\bar{x})\rangle}. \end{aligned}$$

As p is feasible and the corresponding objective value $\langle c, p\rangle$ is equal to the dual optimal one, the polynomial $p(x) = \langle p, \pi_{2n}(x)\rangle$ is optimal.

Unit circle Let us now solve the primal problem

$$\min\{\langle c, p\rangle : p(\bar{z}) = \langle p, \pi_n(\bar{z})\rangle_{\mathbb{R}} = b, p \in \mathcal{K}_{\mathbb{C}}\}. \tag{4.25}$$

As in the real line setting, both primal and dual optimal solutions are computed explicitly by making use of Assumption 4.4. They are equal to:

$$\begin{aligned} y = \frac{1}{\langle T(c)^{-1}\pi_n(\bar{z}), \pi_n(\bar{z})\rangle}, \\ p = T^*(qq^*), \quad q = \frac{T(c)^{-1}\pi_n(\bar{z})}{\langle T(c)^{-1}\pi_n(\bar{z}), \pi_n(\bar{z})\rangle}. \end{aligned}$$

Example 4.8 (Moving average system, [116]). Let $h[n]$ be a discrete-time signal and $\mathcal{H}(e^{\jmath\omega})$ be its Fourier transform. The function $|\mathcal{H}(e^{\jmath\omega})|^2$ is known as the *energy density spectrum* because it determines how the energy is distributed in frequency. Let us compute the signal that has the minimum energy

$$2\pi E = \int_{-\pi}^{\pi} |\mathcal{H}(e^{\jmath\omega})|^2 d\omega$$

and satisfies $|\mathcal{H}(e^{\jmath 0})| = 1$.

This is exactly an example of the problem class (4.25). Since $p(e^{\jmath\omega}) = |\mathcal{H}(e^{\jmath\omega})|^2$ is a trigonometric polynomial, $\int_{-\pi}^{\pi} p(e^{\jmath\omega})d\omega = p_0$. The vector c that defines the objective function is thus equal to $c = [1, 0, \ldots, 0]^T = e_0$. The interpolation constraint is obviously defined by $\bar{z} = \pi_n(e^{\jmath 0}) = e$ and $b = 1$.

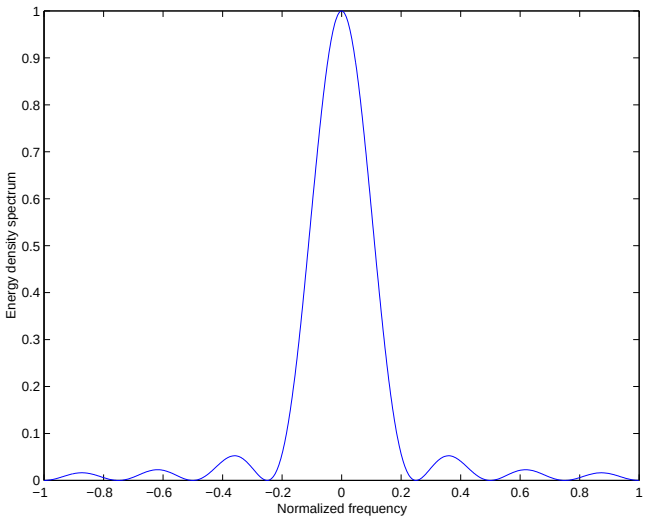

Figure 4.2. *Energy density spectrum ($|\mathcal{H}(e^{\jmath\omega})|^2$ – n=7)*

Therefore, the optimal primal solution is given by

$$p = T^*(qq^*), \quad q = \frac{[1, \ldots, 1]^T}{n+1}.$$

and the corresponding Fourier transform $\mathcal{H}(e^{\jmath\omega})$ can be set to

$$\mathcal{H}(e^{\jmath\omega}) = \sum_{i=0}^{n} \frac{1}{n+1} e^{-\jmath\omega}.$$

Note that $|\mathcal{H}(e^{\jmath\omega})|^2$ is an approximation of a low-pass filter, see Figure 4.2. The corresponding signal is exactly the impulse response of the *moving average system*:

$$h[k] = \begin{cases} \frac{1}{n+1}, & 0 \leq k \leq n+1, \\ 0, & \text{otherwise.} \end{cases}$$

Since convolution of a discrete signal $x[n]$ with $h[n]$ returns a signal $y[n]$ such that

$$y[k] = \frac{1}{n+1} \sum_{l=0}^{n} x[k-l],$$

$y[n]$ is the "moving average" of $x[n]$.

Two interpolation constraints

Before investigating problems with two interpolation constraints, we need to solve explicitly a 2-dimensional semidefinite programming problem.

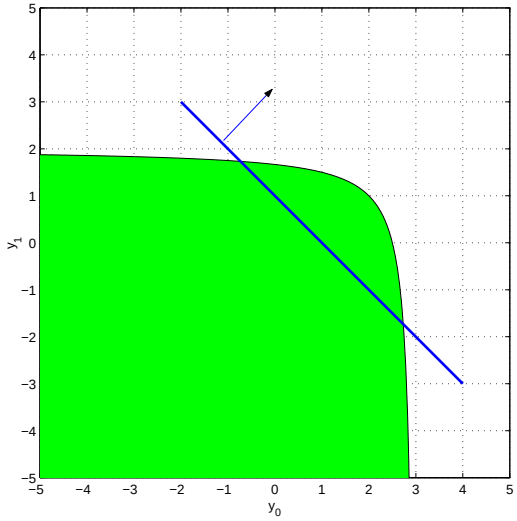

Figure 4.3. *Feasibility region of* (4.26) *with* $\alpha = 3, \beta = (1+\jmath)/\sqrt{2}$ *and* $\gamma = 2$

Proposition 4.24. *Let* $b_0, b_1 \in \operatorname{int} \mathbb{R}_+$ *and* $\alpha, \gamma \in \mathbb{R}$ *and* $\beta \in \mathbb{C}$. *The optimal value of the optimization problem*

$$\begin{array}{ll} \max & b_0 y_0 + b_1 y_1 \\ \text{s.t.} & \begin{bmatrix} \alpha & \overline{\beta} \\ \beta & \gamma \end{bmatrix} \succeq \begin{bmatrix} y_0 & 0 \\ 0 & y_1 \end{bmatrix} \end{array} \tag{4.26}$$

is reached at the optimal point

$$y_0 = \alpha - |\beta| \sqrt{\frac{b_1}{b_0}}, \qquad y_1 = \gamma - |\beta| \sqrt{\frac{b_0}{b_1}}$$

and is equal to $b_0\alpha + b_1\gamma - 2|\beta|\sqrt{b_0 b_1}$.

Proof. The constraints are equivalent to

$$\alpha - y_0 \geq 0, \qquad \gamma - y_1 - \frac{|\beta|^2}{\alpha - y_0} \geq 0.$$

Maximizing the linear function $b_0 y_0 + b_1 y_1$ on this 2-dimensional convex region is straightforward (see Figure 4.3). Clearly, the system of equations

$$\frac{|\beta|^2}{(\alpha - y_0)^2} = \frac{b_0}{b_1}, \quad y_1 = \gamma - \frac{|\beta|^2}{\alpha - y_0},$$

provides us with the optimal point (y_0, y_1). □

Real line If the number of interpolation constraints is equal to 2, the dual problem (4.18) is given by

$$\begin{array}{ll} \max & \langle b, y \rangle \\ \text{s.t.} & H(c) \succeq y_0 \pi_n(x_0)[\pi_n(x_0)]^T + y_1 \pi_n(x_1)[\pi_n(x_1)]^T. \end{array}$$

Equivalently, the dual constraint can be recast as

$$H(c) - \begin{bmatrix}\pi_n(x_0) & \pi_n(x_1)\end{bmatrix} \begin{bmatrix} y_0 & 0 \\ 0 & y_1 \end{bmatrix} \begin{bmatrix}\pi_n(x_0) & \pi_n(x_1)\end{bmatrix}^T \succeq 0.$$

Let us define the matrix $M_H(c; x_0, x_1)$ by

$$M_H(c; x_0, x_1) = \begin{bmatrix} \langle H(c)^{-1}\pi_n(x_0), \pi_n(x_0)\rangle & \langle H(c)^{-1}\pi_n(x_1), \pi_n(x_0)\rangle \\ \langle H(c)^{-1}\pi_n(x_0), \pi_n(x_1)\rangle & \langle H(c)^{-1}\pi_n(x_1), \pi_n(x_1)\rangle \end{bmatrix}.$$

If $\operatorname{diag}(y)$ is positive definite at the optimum, then the previous linear matrix inequalities can be recast as

$$M_H(c; x_0, x_1)^{-1} \succeq \operatorname{diag}(y).$$

Indeed, this reformulation follows from the Schur complement formula. Otherwise, our hypothesis on the objective function, $c \in \operatorname{int} \mathcal{K}_{\mathbb{R}}^\star$, can be used so as to obtain the same reformulation. We delay the proof of this fact to the general setting, see Proposition 4.26.

Consequently, Proposition 4.24 allows us to solve our dual problem explicitly:

$$y_0 = \frac{1}{\det(M_H)} \left[\langle H(c)^{-1}\pi_n(x_1), \pi_n(x_1)\rangle - |\langle H(c)^{-1}\pi_n(x_0), \pi_n(x_1)\rangle| \sqrt{\frac{b_1}{b_0}} \right],$$
$$y_1 = \frac{1}{\det(M_H)} \left[\langle H(c)^{-1}\pi_n(x_0), \pi_n(x_0)\rangle - |\langle H(c)^{-1}\pi_n(x_0), \pi_n(x_1)\rangle| \sqrt{\frac{b_0}{b_1}} \right],$$

with $\det(M_H) = \det(M_H(c; x_0, x_1))$.

Our primal optimization problem can also be solved explicitly. To see this, define the vector $v = \begin{bmatrix} v_1 & v_2 \end{bmatrix}^T$ as the solution of the linear system

$$\begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix} M_H(c; x_0, x_1) \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \end{bmatrix} = \begin{bmatrix} \sqrt{b_0} \\ \sqrt{b_1} \end{bmatrix},$$

where $\sigma \in \{-1, +1\}$ is the sign of $\langle H(c)^{-1}\pi_n(x_0), \pi_n(x_1)\rangle$. Then the vector

$$p = H^*(qq^*), \quad q = H(c)^{-1} \begin{bmatrix}\pi_n(x_0) & \pi_n(x_1)\end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \end{bmatrix}$$

defines a non-negative polynomial $p(x) = (\langle q, \pi_n(x)\rangle)^2$ that satisfies $p(x_0) = b_0$ and $p(x_1) = b_1$. Indeed, we have

$$\begin{bmatrix} q(x_0) \\ q(x_1) \end{bmatrix} = \begin{bmatrix}\pi_n(x_0) & \pi_n(x_1)\end{bmatrix}^T q = \begin{bmatrix} \sqrt{b_0} \\ \sigma\sqrt{b_1} \end{bmatrix}.$$

Moreover, the inner product $\langle c, p\rangle$ is equal to the optimal dual value: the vector p is thus optimal.

Unit circle As in the real line setting, the dual problem can be rewritten as

$$\begin{array}{ll} \max & \langle b, y\rangle \\ \text{s.t.} & M_T(c; z_0, z_1)^{-1} \succeq \operatorname{diag}(y), \end{array}$$

where

$$M_T(c; z_0, z_1) = \begin{bmatrix} \langle T(c)^{-1}\pi_n(z_0), \pi_n(z_0)\rangle & \langle T(c)^{-1}\pi_n(z_1), \pi_n(z_0)\rangle \\ \langle T(c)^{-1}\pi_n(z_0), \pi_n(z_1)\rangle & \langle T(c)^{-1}\pi_n(z_1), \pi_n(z_1)\rangle \end{bmatrix}.$$

The optimal dual solution is now equal to

$$y_0 = \frac{1}{\det(M_T)} \left[\langle T(c)^{-1}\pi_n(z_1), \pi_n(z_1)\rangle - |\langle T(c)^{-1}\pi_n(z_0), \pi_n(z_1)\rangle| \sqrt{\frac{b_1}{b_0}} \right],$$
$$y_1 = \frac{1}{\det(M_T)} \left[\langle T(c)^{-1}\pi_n(z_0), \pi_n(z_0)\rangle - |\langle T(c)^{-1}\pi_n(z_0), \pi_n(z_1)\rangle| \sqrt{\frac{b_0}{b_1}} \right],$$

with $\det(M_T) = \det(M_T(c; z_0, z_1))$. Let us define the vector $\begin{bmatrix} v_0 & v_1 \end{bmatrix}^T$ as the solution of the linear system

$$\begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix}^* M_T(c; z_0, z_1) \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \end{bmatrix} = \begin{bmatrix} \sqrt{b_0} \\ \sqrt{b_1} \end{bmatrix},$$

where σ is equal to $e^{-\jmath \arg\langle T(c)^{-1}\pi_n(z_1), \pi_n(z_0)\rangle}$. The vector

$$p = T^*(qq^*), \quad q = T(c)^{-1} \begin{bmatrix} \pi_n(z_0) & \pi_n(z_1) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \end{bmatrix}$$

corresponds to a trigonometric polynomial $p(z) = |q(z)|^2$ that satisfies our interpolation constraints and such that $\langle c, p\rangle = \langle b, y\rangle$. This vector p is thus the (primal) optimal one.

More interpolation constraints ($k \leq n+1$)

If Assumption 4.4 holds and $k \leq n+1$, the previous analysis can always be carried out. We first focus on the unit circle setting and show the connection with spectral factorization of trigonometric polynomials. The real line problem is then solved using a similar methodology. Let us start with two preliminary results.

Preliminary results

Proposition 4.25. *Let $C \in \operatorname{int} \mathcal{H}_+^n$ be a positive definite matrix and $V = \begin{bmatrix} V_0 & V_1 \end{bmatrix} \in \mathbb{C}^{n\times n}$ be a nonsingular matrix. If the matrix $W = \begin{bmatrix} W_0 \\ W_1 \end{bmatrix}$ is the (left) inverse of V with compatible partitions, i.e., $\begin{bmatrix} W_0V_0 & W_0V_1 \\ W_1V_0 & W_1V_1 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$, then we have*

$$(V_1^*C^{-1}V_1)^{-1} = W_1CW_1^* - W_1CW_0^*(W_0CW_0^*)^{-1}W_0CW_1.$$

Proof. Let us apply the well-known *Schur complement identity*

$$\begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} I & 0 \\ GE^{-1} & I \end{bmatrix} \begin{bmatrix} E & 0 \\ 0 & H - GE^{-1}F \end{bmatrix} \begin{bmatrix} I & E^{-1}F \\ 0 & I \end{bmatrix}$$

to the matrix product

$$WCW^* = \begin{bmatrix} W_0CW_0^* & W_0CW_1^* \\ W_1CW_0^* & W_1CW_1^* \end{bmatrix}.$$

Clearly, we obtain that

$$WCW^* = \begin{bmatrix} I & 0 \\ (W_1CW_0^*)(W_0CW_0^*)^{-1} & I \end{bmatrix} \begin{bmatrix} W_0CW_0^* & 0 \\ 0 & W_1/W_0 \end{bmatrix} \begin{bmatrix} I & (W_0CW_0^*)^{-1}(W_0CW_1^*) \\ 0 & I \end{bmatrix}$$

with $W_1/W_0 = W_1CW_1^* - (W_1CW_0^*)(W_0CW_0^*)^{-1}(W_0CW_1^*)$. Because the matrix WCW^* is nonsingular (by assumption), we have

$$(WCW^*)^{-1} = \begin{bmatrix} I & -(W_0CW_0^*)^{-1}(W_0CW_1^*) \\ 0 & I \end{bmatrix} \begin{bmatrix} W_0CW_0^* & 0 \\ 0 & W_1/W_0 \end{bmatrix}^{-1} \begin{bmatrix} I & 0 \\ -(W_1CW_0^*)(W_0CW_0^*)^{-1} & I \end{bmatrix}.$$

Hence, the lower right block of the identity

$$(WCW^*)^{-1} = V^*C^{-1}V = \begin{bmatrix} V_0^*C^{-1}V_0 & V_0^*C^{-1}V_1 \\ V_1^*C^{-1}V_0 & V_1^*C^{-1}V_1 \end{bmatrix}$$

is exactly equivalent to

$$V_1^*C^{-1}V_1^* = (W_1CW_1^* - (W_1CW_0^*)(W_0CW_0^*)^{-1}(W_0CW_1^*))^{-1}.$$

□

Proposition 4.26. *Let $C \in \operatorname{int} \mathcal{H}_+^n$ be a positive definite matrix and $V_1 \in \mathbb{C}^{n\times k}$ be a matrix with full column rank ($k \leq n$). Then the linear matrix inequality*

$$C \succeq V_1 \operatorname{diag}(y) V_1^* \tag{4.27}$$

is equivalent to

$$(V_1^*C^{-1}V_1)^{-1} \succeq \operatorname{diag}(y). \tag{4.28}$$

Proof. If $k = n$, the proof is trivial. Indeed, both inequalities (4.27) and (4.28) are congruent. This congruence is defined by the nonsingular matrix V_1^{-1}. If $k < n$, Proposition 4.25 must be used. Let $V_0 \in \mathbb{C}^{n\times(n-k)}$ be a matrix such that $V = \begin{bmatrix} V_0 & V_1 \end{bmatrix} \in \mathbb{C}^{n\times n}$ is nonsingular. The (left) inverse of V is denoted by $W = \begin{bmatrix} W_0 \\ W_1 \end{bmatrix}$. If the rows of W are partitioned according to the partition of V, we have

$$WV = \begin{bmatrix} W_0V_0 & W_0V_1 \\ W_1V_0 & W_1V_1 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

The linear matrix inequality (4.27), which can be rewritten as

$$C - \begin{bmatrix} V_0 & V_1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \operatorname{diag}(y) \end{bmatrix} \begin{bmatrix} V_0^* \\ V_1^* \end{bmatrix} \succeq 0,$$

is thus equivalent to

$$\begin{bmatrix} W_0 \\ W_1 \end{bmatrix} C \begin{bmatrix} W_0^* & W_1^* \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & \operatorname{diag}(y) \end{bmatrix} \succeq 0 \tag{4.29}$$

by congruence. Because $W_0 C W_0^*$ is positive definite by assumption, the previous inequality is equivalent to positive semidefiniteness of its Schur complement in (4.29),

$$W_1 C W_1^* - (W_1 C W_0^*)(W_0 C W_0^*)^{-1}(W_0 C W_1^*) \succeq \operatorname{diag}(y).$$

We complete the proof by making use of Proposition 4.25. □

Unit circle Remember that the optimization problem of interest is

$$\begin{array}{ll} \min & \langle c, p \rangle_{\mathbb{R}} \\ \text{s.t.} & \langle p, \pi_n(z_i) \rangle_{\mathbb{R}} = b_i, \quad i = 0, \ldots, k-1, \\ & p \in \mathcal{K}_{\mathbb{C}}. \end{array} \tag{4.30}$$

If the non-negative trigonometric polynomial $p(z)$ is written as a square by making use of an arbitrary spectral factor $q(z)$, i.e., $p(z) = |q(z)|^2$ or $p = T^*(qq^*)$, the primal optimization problem can be rewritten as

$$\begin{array}{ll} \min & \langle T(c)q, q \rangle \\ \text{s.t.} & \langle q, \pi_n(z_i) \rangle = \sqrt{b_i} e^{\jmath\theta_i}, \quad i = 0, \ldots, k-1, \end{array} \tag{4.31}$$

where $\{\theta_i\}_{i=0}^{k-1}$ is a set of phases.

Define the vector σ by $\sigma_i = \sqrt{b_i} e^{\jmath\theta_i}, i = 0, \ldots, k-1$ and the matrix M_T by

$$M_T(c; z_0, \ldots, z_{k-1}) = \begin{bmatrix} \pi_n(z_0) & \cdots & \pi_n(z_{k-1}) \end{bmatrix}^* T(c)^{-1} \begin{bmatrix} \pi_n(z_0) & \cdots & \pi_n(z_{k-1}) \end{bmatrix}.$$

As a function of σ, the optimal solution of (4.31) is equal to

$$q = T(c)^{-1} \begin{bmatrix} \pi_n(z_0) & \cdots & \pi_n(z_{k-1}) \end{bmatrix} M_T(c; z_0, \ldots, z_{k-1})^{-1} \sigma \tag{4.32}$$

and the corresponding optimal value is

$$\langle T(c)q, q \rangle = \langle M_T(c; z_0, \ldots, z_{k-1})^{-1} \sigma, \sigma \rangle.$$

Remark. A direct consequence of (4.32) is that the spectral factor $q(z)$ is decomposed as a sum of "Lagrange-like" polynomials:

$$q(z) = \langle q, \pi_n(z) \rangle_{\mathbb{C}} = \sum_{i=0}^{k-1} e^{\jmath\theta_i} \sigma_i l_i(z),$$

where $l_i(z_j) = \delta_{ij}, \forall i, j$.

Finally, the optimal solution of problem (4.31) is obtained by minimizing over the vector σ,

$$\begin{array}{ll} \min & \langle M_T(c; z_0, \ldots, z_{k-1})^{-1} \sigma, \sigma \rangle \\ \text{s.t.} & |\sigma_i|^2 = b_i, \quad i = 0, \ldots, k-1. \end{array} \tag{4.33}$$

If $m > 2$, an explicit solution is difficult to obtain easily from this new formulation. However, we can solve the semidefinite relaxation of problem (4.33):

$$\begin{array}{ll} \min & \langle M_T^{-1}(z_0, \ldots, z_{k-1}), X \rangle, \\ \text{s.t.} & \operatorname{Diag}(X) = b, \\ & X \in \mathcal{H}_+^k, \end{array} \tag{4.34}$$

where $\operatorname{Diag}(X)$ is the vector defined by the diagonal elements of X. In general, a QQP formulation of the form (4.33) is NP-hard to solve, see the Appendix. Nevertheless, the particular structure of the quadratic objective function yields an extremely interesting result.

Theorem 4.27. *If Assumption 4.4 holds, relaxation* (4.34) *of the quadratically constrained quadratic problem* (4.33) *is exact.*

Proof. Using standard convex duality theory, the dual of problem (4.34) is

$$\begin{array}{ll} \max & \langle b, y \rangle \\ \text{s.t.} & M_T^{-1}(z_0, \ldots, z_{k-1}) \succeq \operatorname{diag}(y), \end{array} \tag{4.35}$$

which is exactly the dual of the original problem (4.30):

$$\begin{array}{ll} \max & \langle b, y \rangle \\ \text{s.t.} & T(c) \succeq \begin{bmatrix} \pi_n(z_0) & \ldots & \pi_n(z_{k-1}) \end{bmatrix} \operatorname{diag}(y) \begin{bmatrix} \pi_n(z_0) & \ldots & \pi_n(z_{k-1}) \end{bmatrix}^*. \end{array} \tag{4.36}$$

To see this, we define the matrix V_1 as $V_1 = \begin{bmatrix} \pi_n(z_0) & \ldots & \pi_n(z_{k-1}) \end{bmatrix}$ and we apply Proposition 4.26 with $C = T(c)$. Because the (dual) constraints of (4.35) and (4.36) are equivalent, both problems are identical.

By assumption the original problem (4.30) has no duality gap. Since both problems (4.30) and (4.34) have the same dual, the relaxation has also a zero duality gap. This last observation completes our proof. □

The optimal coefficients p can be obtained from the solution X of (4.34) via the identity

$$p = T^*(T(c)^{-1} V_1 M_T^{-1} X M_T^{-1} V_1^* T(c)^{-1})$$

where $V_1 = \begin{bmatrix} \pi_n(z_0) & \ldots & \pi_n(z_{k-1}) \end{bmatrix}$ and $M_T = M_T(c; z_0, \ldots, z_{k-1})$. To see this, note that

$$\begin{aligned} \langle c, p \rangle &= \langle T(c), T(c)^{-1} V_1 M_T^{-1} X M_T^{-1} V_1^* T(c)^{-1} \rangle \\ &= \langle T(c)^{-1} V_1 M_T^{-1} X M_T^{-1} V_1^*, I \rangle \\ &= \langle V_1^* T(c)^{-1} V_1 M_T^{-1} X M_T^{-1}, I \rangle \\ &= \langle X, M_T^{-1} \rangle \end{aligned}$$

and that, for all i,

$$\begin{aligned} \langle p, \pi_n(z_i) \rangle =& \langle T(\pi_n(z_i)), T(c)^{-1} V_1 M_T^{-1} X M_T^{-1} V_1^* T(c)^{-1} \rangle \\ =& \langle \pi_n(z_i) \pi_n(z_i)^*, T(c)^{-1} V_1 M_T^{-1} X M_T^{-1} V_1^* T(c)^{-1} \rangle \\ =& \langle (\pi_n(z_i)^* T(c)^{-1} V_1 M_T^{-1}) X (M_T^{-1} V_1^* T(c)^{-1} \pi_n(z_i)), I \rangle \\ =& \langle e_i e_i^*, X \rangle = b_i. \end{aligned}$$

Real line Remember that the optimization problem of interest is

$$\begin{array}{ll} \min & \langle c, p\rangle \\ \text{s.t.} & \langle p, \pi_{2n}(x_i)\rangle = b_i, \quad i = 0, \ldots, k-1, \ . \\ & p \in \mathcal{K}_{\mathbb{R}}. \end{array} \tag{4.37}$$

If we use any complex spectral factor $q(x)$ of our unknown polynomial $p(x) = |q(x)|^2$ as a variable, the previous analysis can be carried out in the real line setting. It leads exactly to the same formulae *provided that* the following substitutions are performed:

1. $T(c)$ is replaced by $H(c)$;
2. the interpolation points $\{z_i\}_{i=0}^{k-1}$ are replaced by $\{x_i\}_{i=0}^{k-1}$;
3. the matrix $M_T(c; z_0, \ldots, z_{k-1})$ is replaced by its "Hankel counterpart"

$$[M_H(c; x_0, \ldots, x_{k-1})]_{ij} = \pi_n(x_i)^* H(c)^{-1} \pi_n(x_j).$$

Let us summarize the most important steps. First, the primal optimization problem (4.37) is reformulated as

$$\begin{array}{ll} \min & \langle H(c)q, q\rangle \\ \text{s.t.} & \langle q, \pi_n(x_i)\rangle = \sqrt{b_i} e^{\jmath\theta_i}, \quad i = 0, \ldots, k-1, \end{array} \tag{4.38}$$

which is equivalent to

$$\begin{array}{ll} \min & \langle M_H(c; x_0, \ldots, x_{k-1})^{-1}\sigma, \sigma\rangle \\ \text{s.t.} & |\sigma_i|^2 = b_i, \quad i = 0, \ldots, k-1. \end{array} \tag{4.39}$$

In practice, this last optimization problem is solved using the following relaxation

$$\begin{array}{ll} \min & \langle M_H^{-1}(x_0, \ldots, x_{k-1}), X\rangle, \\ \text{s.t.} & \mathrm{Diag}(X) = b, \\ & X \in \mathcal{H}_+^k. \end{array} \tag{4.40}$$

As before, the structure of QQP (4.39) leads to an exact semidefinite relaxation.

Theorem 4.28. *If Assumption 4.4 holds, relaxation* (4.40) *of the quadratically constrained quadratic problem* (4.39) *is exact.*

Proof. Using standard convex duality theory, the dual of problem (4.34) is

$$\begin{array}{ll} \max & \langle b, y\rangle \\ \text{s.t.} & M_H^{-1}(x_0, \ldots, x_{k-1}) \succeq \mathrm{diag}(y), \end{array} \tag{4.41}$$

which is exactly the dual of the original problem (4.37):

$$\begin{array}{ll} \max & \langle b, y\rangle \\ \text{s.t.} & H(c) \succeq \begin{bmatrix} \pi_n(x_0) & \ldots & \pi_n(x_{k-1}) \end{bmatrix} \mathrm{diag}(y) \begin{bmatrix} \pi_n(x_0) & \ldots & \pi_n(x_{k-1}) \end{bmatrix}^*. \end{array} \tag{4.42}$$

To see this, we define the matrix V_1 as $V_1 = \begin{bmatrix} \pi_n(z_0) & \ldots & \pi_n(z_{k-1}) \end{bmatrix}$ and we apply Proposition 4.26 with $C = T(c)$ and V_1. Because the (dual) constraints of (4.41) and (4.42) are equivalent, both problems are identical.

By assumption the original problem (4.37) has no duality gap. Since both problems (4.37) and (4.40) have the same dual, the relaxation has also a zero duality gap. This last observation completes our proof. □

Complexity The complexity of solving relaxation (4.34) or (4.40) is only a function of the desired accuracy ϵ and the number of interpolation constraints k. If Assumption 4.4 holds and if the original problem has been preprocessed, it can be solved in a number of iterations that does *not* depend on the degree n. Indeed, solving the dual problem (4.35) or (4.41) using a *standard* path-following scheme requires $\mathcal{O}(\sqrt{k}\log\frac{1}{\epsilon})$ Newton steps. At each iteration, computing the gradient and the Hessian of a barrier function of the type

$$f(y) = -\log\det(M^{-1} - \operatorname{diag}(y))$$

requires $\mathcal{O}(k^3)$ flops. Note that the preprocessing (computing M_H or M_T) can be done via fast Hankel or Toeplitz solvers, see [86].

Still more interpolation constraints ($k > n+1$)

If the number of interpolation constraints is strictly greater than $n+1$, strict feasibility of the primal problem depends on the data. Therefore, a *general* procedure that solves efficiently the primal problem and uses the structure of the interpolation constraints is not likely to exist. Indeed, the primal problem might be infeasible ! Let us illustrate this fact by a simple example.

Example 4.9. Consider the set of polynomials of degree $2n = 4$, non-negative on the real line, and four interpolation points $x = [-2, -1, 1, 2]$. The vector $b = [1, 1, 1, 1]$ gives a strictly feasible primal problem. Indeed, the polynomial $p(x) = \frac{1}{3}(x^4 - 5x^2 + 7)$ satisfies our interpolation constraints and belongs to $\operatorname{int}\mathcal{K}_{\mathbb{R}}$. If the vector b is equal to $[1, 10, 1, 1]$, the polynomial family that satisfies our interpolation constraints is $p(x;\lambda) = \frac{1}{4}((\lambda-7)x^4+6x^3+(29-5\lambda)x^2-24x+4\lambda)$, $\lambda \in \mathbb{R}$. If $p(x;\lambda)$ belonged to $\operatorname{int}\mathcal{K}_{\mathbb{R}}$, λ would be greater than 7. As $p(\frac{5}{4};\lambda) = \frac{1}{1024}(-371\lambda - 3255), \forall\lambda > 0$, these data correspond to an infeasible primal set...

Of course, the dual structure can still be exploited to try reducing the computational cost. For instance, consider a problem on the unit circle with $m > n+1$ interpolation constraints. Clearly, the corresponding Vandermonde matrix V can be divided into a nonsingular square Vandermonde matrix V_0 and a rectangular one V_1

$$V = \begin{bmatrix} V_0 & V_1 \end{bmatrix}, \quad \det V_0 \neq 0.$$

If the dual vector is divided accordingly, the dual constraint can be recast as $T(c) \succeq V_0 \operatorname{diag}(y_0) V_0^* + V_1 \operatorname{diag}(y_1) V_1^*$. Since V_0 is nonsingular, it is equivalent to

$$V_0^{-1} T(c) V_0^{-*} - V_0^{-1} V_1 \operatorname{diag}(y_1) V_1^*) V_0^{-*} \succeq \operatorname{diag}(y_0).$$

Therefore, an appropriate preprocessing leads to the following dual constraint

$$\hat{C} - \hat{V} \operatorname{diag}(y_1) \hat{V}^* \succeq \operatorname{diag}(y_0).$$

Since the Toeplitz structure of the dual is lost, the resulting algorithm cannot use the underlying displacement operator nor a divide-and-conquer strategy to evaluate the gradient and the Hessian of the self-concordant barrier function. This strategy will thus be slower than the one designed in Chapter 3.

Property of the objective function

If $H(c)$ or $T(c)$ is not positive definite, the corresponding dual problem can sometimes be solved explicitly.

Real line If the vector c is such that $H(c)$ can be factorized as

$$H(c) = \begin{bmatrix} V & W \end{bmatrix} \begin{bmatrix} \operatorname{diag}(\lambda_v) & 0 \\ 0 & \operatorname{diag}(\lambda_w) \end{bmatrix} \begin{bmatrix} V^T \\ W^T \end{bmatrix}, \tag{4.43}$$

where $V \in \mathbb{R}^{k\times(n+1)}$ is the Vandermonde matrix defined by the interpolation constraints and $W \in \mathbb{R}^{(n+1-k)\times(n+1)}$ is such that $\begin{bmatrix} V & W \end{bmatrix}$ is full rank, one can easily compute an explicit solution of the optimization problem. From a theoretical point of view, there exist vectors c such that the proposed factorization does not exist. From a computational point of view, it may also be difficult to compute accurately.

The dual constraint now reads as follows

$$\begin{bmatrix} \operatorname{diag}(\lambda_v - y) & 0 \\ 0 & \operatorname{diag}(\lambda_w) \end{bmatrix} \succeq 0.$$

If $\operatorname{diag}(\lambda_w)$ is not positive semidefinite, the dual optimization problem is infeasible and the primal problem is unbounded. Otherwise, the solution is obtained by setting the dual variables y_i to their upper bounds, i.e., $y = \lambda_v$. This provides us with either a lower bound or the exact value of the optimization problem, depending on whether the problem has a duality gap.

Unit circle The same factorization technique can be applied to $T(c)$, i.e.,

$$T(c) = \begin{bmatrix} V & W \end{bmatrix} \begin{bmatrix} \operatorname{diag}(\lambda_v) & 0 \\ 0 & \operatorname{diag}(\lambda_w) \end{bmatrix} \begin{bmatrix} V^* \\ W^* \end{bmatrix}, \tag{4.44}$$

and leads to the same results and drawbacks.

4.3.3 Extensions

Matrix polynomials

Most of the previous results still hold in the context of non-negative matrix polynomials. To avoid unnecessary redundancies, we only consider the cone $\mathcal{K}_{\mathbb{R}}$ of matrix polynomials non-negative on the real line, see Section 2.2. Remember that the dual cone is the set of matrix coefficients $S = \begin{bmatrix} S_0, & S_1, & \ldots, & S_{2n} \end{bmatrix}$ such that the block Hankel matrix

$$H(S) = \begin{bmatrix} S_0 & S_1 & \cdots & S_n \\ S_1 & \iddots & \iddots & \vdots \\ \vdots & \iddots & \iddots & S_{2n-1} \\ S_n & \cdots & S_{2n-1} & S_{2n} \end{bmatrix},$$

is positive semidefinite. That is, $\mathcal{K}_{\mathbb{R}}^{\star} = \{S : H(S) \succeq 0\}$.

The optimization problem Using matrix interpolation constraints, our optimization problem (4.17) could be extended to

$$\begin{array}{ll} \min & \langle C, P\rangle \equiv \sum_{\ell=0}^{2n} \langle C_\ell, P_\ell\rangle \\ \text{s.t.} & P(x_i) = \sum_{\ell=0}^{2n} P_\ell x_i^\ell = B_i, \quad i = 0, \ldots, k-1, \\ & P(x) \succeq 0, \ \forall x \in \mathbb{R}, \end{array} \tag{4.45}$$

where $\{B_i\}_{i=0}^{k-1}$ is a set of positive definite matrices. Its dual is readily seen to be equal to

$$\begin{array}{ll} \max & \langle B, Y\rangle \equiv \sum_{i=0}^{k-1}\langle B_i, Y_i\rangle \\ \text{s.t.} & S_\ell + \sum_{i=0}^{k-1} x_i^\ell Y_i = C_\ell, \quad \ell = 0,\dots,2n, \\ & H(S) \succeq 0. \end{array} \tag{4.46}$$

Remark. Unfortunately, we cannot treat tangential interpolation constraints, which are frequent in systems and control.

Strict feasibility As before, primal strict feasibility holds if the number k of matrix interpolation constraints is less or equal to $n+1$. To see this, consider $n+1$ distinct interpolation points $\{x_i\}_{i=0}^n$ and the associated Lagrange polynomials $\{L_i(x)\}_{i=0}^n$ of degree n. These polynomials are defined by the identities

$$L_j(x_i) = \delta_{ij} I_m, \qquad 0 \le i, j \le n.$$

Then the polynomial

$$P(x) = \sum_{i=0}^{n} L_i(x) P(x_i) L_i^T(x) = \sum_{i=0}^{n} L_i(x) B_i L_i^T(x)$$

can be rewritten as

$$P(x) = \langle L\,\mathrm{diag}(\{P(x_i)\}_{i=0}^{k-1}) L^T \Pi_n(x), \Pi_n(x)\rangle$$

where L is nonsingular and $\mathrm{diag}(\{P(x_i)\}_{i=0}^{k-1})$ is positive definite. By construction, we see that $P(x) \in \operatorname{int}\mathcal{K}_\mathbb{R}$ and $P(x_i) = B_i, \forall i$.

Since the dual constraints (4.46) are equivalent to

$$H(C) \succeq \sum_{i=0}^{k-1} \Pi_n(x_i) Y_i \Pi_n(x_i)^T,$$

the dual is strictly feasible if $k \ge n+1$.

Let us state the matrix counterpart of Assumption 4.4 for future use.

Assumption 4.5. The number k of interpolation constraints is less or equal to $n+1$ and the objective block vector C satisfies $H(C) \succ 0$.

Hereafter, we focus on problems satisfying this assumption.

One interpolation constraint Let us consider a matrix interpolation problem with one constraint:

$$\begin{array}{ll} \min & \langle C, P\rangle \\ \text{s.t.} & P(\bar{x}) = \sum_{\ell=0}^{2n} P_\ell \bar{x}^\ell = B \succ 0, \\ & P(x) \succeq 0, \ \forall x \in \mathbb{R}. \end{array}$$

Without loss of generality, B is assumed to be the identity matrix, i.e., $B = I_m$. Using the dual matrix variable Y, the dual problem reads

$$\begin{array}{ll} \max & \langle I, Y\rangle \\ \text{s.t.} & H(C) \succeq \Pi_n(\bar{x}) Y \Pi_n(\bar{x})^T. \end{array}$$

Because $H(C) \succ 0$, a standard Schur complement approach shows that the optimal dual variable is

$$Y = [\Pi_n(\bar{x})^T H(C)^{-1} \Pi_n(\bar{x})]^{-1}.$$

The spectral factor

$$Q = H(C)^{-1}\Pi_n(\bar{x})[\Pi_n(\bar{x})^T H(C)^{-1}\Pi_n(\bar{x})]^{-1}$$

allows us to compute the optimal primal variable P

$$P(x) = Q(x)Q(x)^* \quad \Longleftrightarrow \quad P = H^*(QQ^*).$$

It is easy to check that this value of P is optimal,

$$\begin{aligned}\langle C, P\rangle &= \sum_{\ell=0}^{2n} \langle C_\ell, P_\ell\rangle = \langle H(C)Q, Q\rangle \\ &= \langle I, [\Pi_n(\bar{x})^T H(C)^{-1}\Pi_n(\bar{x})]^{-1}\rangle = \langle I, Y\rangle\end{aligned}$$

and

$$P(\bar{x}) = \Pi_n(\bar{x})^T QQ^*\Pi_n(\bar{x}) = I_m = B.$$

More interpolation constraints If the number of matrix interpolation constraints is less or equal to $n+1$, we can again use an arbitrary spectral factor to get an efficient algorithm, the complexity of which mainly depends on k and m.

Indeed, let $Q(x)$ be an arbitrary spectral factor $Q(x)$ of our unknown polynomial $P(x)$, i.e $P(x) = Q(x)Q(x)^*$. Then the optimization problem can be rewritten as

$$\begin{array}{ll}\min & \langle H(C)Q, Q\rangle \\ \text{s.t.} & Q(x_i) = \sum_{\ell=0}^{2n} Q_\ell x_i^\ell = B_i^{1/2} U_i, \quad i = 0, \ldots, k-1,\end{array} \tag{4.47}$$

where $\{U_i\}_{i=0}^{k-1}$ is a set of unitary matrices ($U_i^* U_i = I_m, \forall i$).

If the definition of M_H is adapted to the matrix case,

$$[M_H(C; x_0, \ldots, x_{k-1})]_{ij} = \Pi_n(x_i)^* H(C)^{-1}\Pi_n(x_j),$$

then the optimal solution of (4.47), written as a function of

$$U = \begin{bmatrix} U_0, \\ \vdots, \\ U_{k-1}\end{bmatrix},$$

is equal to

$$Q = H(C)^{-1}\begin{bmatrix}\Pi_n(x_0) & \cdots & \Pi_n(x_{k-1})\end{bmatrix} M_H(C; x_0, \ldots, x_{k-1})^{-1} \operatorname{diag}(\{B_i^{1/2}\}_{i=0}^{k-1})U.$$

As in the scalar case, the optimal solution of the original problem is obtained via the quadratic optimization problem

$$\begin{array}{ll}\min & \langle \operatorname{diag}(\{B_i^{1/2}\}_{i=0}^{k-1}) M_H(C; x_0, \ldots, x_{k-1})^{-1} \operatorname{diag}(\{B_i^{1/2}\}_{i=0}^{k-1})U, U\rangle \\ \text{s.t.} & U_i^* U_i = I_m, \quad i = 0, \ldots, k-1.\end{array} \tag{4.48}$$

The semidefinite relaxation of this QQP is

$$\begin{array}{lll} \min & \langle M_H(C; x_0, \ldots, x_{k-1})^{-1}, X \rangle & \\ \text{s.t.} & X_{ii} = B_i, & i = 0, \ldots, k-1, \\ & X \in \mathcal{H}_+^{mk}, & \end{array} \tag{4.49}$$

where X_{ii} is the ith $m \times m$ diagonal block of X. Its dual is given by

$$\begin{array}{ll} \max & \langle B, Y \rangle \\ \text{s.t.} & M_H(C; x_0, \ldots, x_{k-1})^{-1} \succeq \operatorname{diag}(\{Y_i\}_{i=0}^{k-1}) \end{array} \tag{4.50}$$

and is equal to the dual of the original problem. Therefore, we could proceed as before to obtain the following theorem:

Theorem 4.29. *If Assumption 4.5 holds, relaxation* (4.49) *of the quadratically constrained quadratic problem* (4.48) *is exact.*

Provided that the original problem has been preprocessed, solving the dual problem (4.50) does *not* depend on the degree $2n$ of $P(x)$. This result is similar to the scalar case. As Assumption 4.5 guarantees that strict feasibility holds, we obtain an efficient algorithm to solve our problem class.

Derivatives

Hereafter, we present the straightforward extension of our previous results to interpolation conditions on the derivatives. We only consider the scalar case to keep our equations as small as possible.

Real line In the real line setting, interpolation constraints on the derivatives are formulated as

$$p^{(\ell)}(x_i) = \langle p, \pi_{2n}^{(\ell)}(x_i) \rangle = b_i,$$

where $\pi_{2n}^{(\ell)}(\cdot)$ is the componentwise ℓth derivative of $\pi_{2n}(\cdot)$. Such constraints will be called "interpolation-like" constraints.

If all the linear constraints of (4.17) are interpolation-like constraints, i.e.,

$$\langle a_i, p \rangle \doteq \langle \pi_{2n}^{(\ell_i)}(x_i), p \rangle = b_i, \quad i = 0, \ldots, k-1,$$

the dual problem (4.18) reads as follows

$$\begin{array}{ll} \max & \langle b, y \rangle \\ \text{s.t.} & H(c) - \sum_{i=0}^{k-1} y_i H(\pi_{2n}^{(\ell_i)}(x_i)) \succeq 0. \end{array} \tag{4.51}$$

Let us now prove that $H(\pi_{2n}^{(\ell_i)}(x_i))$ has a special structure.

Proposition 4.30. *Let $\ell \geq 0$. Then*

$$H(\pi_{2n}^{(\ell)}(x)) = \sum_{r=0}^{\ell} \binom{r}{\ell} \pi_n^{(r)}(x) (\pi_n^{(\ell-r)}(x))^T, \quad \forall x \in \mathbb{R} \tag{4.52}$$

and the rank of this matrix is $\min\{\ell, 2n-\ell\} + 1$.

Proof. Since $H(\pi_{2n}(x)) = \pi_n(x)\pi_n(x)^T$ and $H(\cdot)$ is a linear operator, equation (4.52) is a direct consequence of the chain rule. The rank condition originates from the fact that $\pi_n^{(n+1)}(x) = 0$. □

This proposition allows us to improve the formulation (4.51) of the dual problem. First of all, assume that the interpolation points are distinct and that $\ell_i \leq n, \forall i$. Let us define a block diagonal matrix

$$\Delta(y) = \operatorname{diag}(\{\Delta_0(y), \ldots, \Delta_{k-1}(y)\})$$

where $\Delta_i(y)$ is an $(\ell_i + 1) \times (\ell_i + 1)$ matrix defined by

$$\Delta_i(y) = \begin{bmatrix} 0 & & \binom{\ell_i}{\ell_i} y_i \\ & \iddots & \\ \binom{0}{\ell_i} y_i & & 0 \end{bmatrix}, \quad i = 0, \ldots, k-1.$$

Using the above proposition, the dual problem can be written as

$$\begin{array}{ll} \max & \langle b, y \rangle \\ \text{s.t.} & H(c) - V\Delta(y)V^T \succeq 0, \end{array} \tag{4.53}$$

where V is the non-square *confluent Vandermonde matrix*

$$V = \left[\pi_n^{(0)}(x_0) \quad \ldots \quad \pi_n^{(\ell_1)}(x_0) \quad | \quad \ldots \quad | \quad \pi_n^{(0)}(x_m) \quad \ldots \quad \pi_n^{(\ell_m)}(x_m) \right].$$

If the interpolation points are not distinct or if there exists at least one index i such that $\ell_i > n$, the matrix V and the block-diagonal matrix $\Delta(y)$ must be redefined in order to get a dual problem similar to (4.53). Because the appropriate reformulation is evident, but cumbersome, it has been omitted.

If $H(c) \succ 0$ and the numbers of rows of V is greater than its number of columns, the dual constraint (4.53) is easily recast using Proposition 4.26:

$$(V^T H(c)^{-1} V)^{-1} \succeq \Delta(y).$$

The complexity of solving the dual problem (4.53) depends mostly on the dimension of $\Delta(y)$. That is, an appropriate preprocessing tends to eliminate the dependence on the degree $2n$. Because primal strict feasibility cannot be guaranteed from the knowledge of k, we cannot guarantee that the semidefinite relaxation is exact.

Unit circle In the unit circle setting, interpolation constraints on the derivatives, $p^{(\ell_i)}(\theta_i) = b_i$, are equivalent to the linear constraints

$$p^{(\ell_i)}(z_i) = \langle (-\jmath N)^{\ell_i} p, \pi_n(z_i) \rangle = \langle p, (\jmath N)^{\ell_i} \pi_n(z_i) \rangle = b_i, \quad z_i = e^{\jmath\theta_i}, \tag{4.54}$$

where $N = \operatorname{diag}(0, 1, \ldots, n)$.

If all linear constraints of (4.21) are interpolation-like constraints, i.e.,

$$\langle a_i, p \rangle \doteq \langle p, (\jmath N)^{\ell_i} \pi_n(z_i) \rangle = b_i, \quad z_i = e^{\jmath\theta_i}, \quad i = 0, \ldots, k-1,$$

the dual problem (4.22) reads now as follows

$$\begin{array}{ll} \max & \langle b, y\rangle \\ \text{s.t.} & T(c) - \sum_{i=0}^{k-1} y_i T((\jmath N)^{\ell_i}\pi_n(z_i)) \succeq 0 \end{array} . \tag{4.55}$$

Note that $T((\jmath N)^m \pi_n(z))$ has a special structure.

Proposition 4.31. *Let $\ell \geq 0$. Then*

$$T((\jmath N)^\ell \pi_n(z)) = \sum_{r=0}^{\ell} \binom{r}{\ell} (\jmath N)^r \pi_n(z)[(\jmath N)^{\ell-r}\pi_n(z)]^* \tag{4.56}$$

and the rank of this matrix is $\min\{\ell, n\} + 1$.

Proof. Since $T(\pi_n(z)) = \pi_n(z)\pi_n(z)^*$, $\frac{\partial}{\partial\theta}(\pi_n(z)|_{z=e^{\jmath\theta}}) = \jmath N(\pi_n(z)|_{z=e^{\jmath\theta}})$ and $T(\cdot)$ is a linear operator, it is straightforward to check equation (4.56). □

Assume that the interpolation points are distinct and define the block diagonal matrix

$$\Delta(y) = \operatorname{diag}(\{\Delta_0(y), \ldots, \Delta_{k-1}(y)\})$$

as before. Using the above proposition, the dual problem can be written as

$$\begin{array}{ll} \max & \langle b, y\rangle \\ \text{s.t.} & T(c) - W\Delta(y)W^* \succeq 0, \end{array}$$

where W is the non-square matrix

$$W = \Big[(\jmath N)^0\pi_n^{(0)}(z_0), \ldots, (\jmath N)^{\ell_1}\pi_n^{(\ell_1)}(z_0)| \ldots | (\jmath N)^0\pi_n^{(0)}(z_{k-1}), \ldots, (\jmath N)^{\ell_{k-1}}\pi_n^{(\ell_{k-1})}(x_{k-1})\Big] .$$

If $\ell_i \leq 1, \forall i$, the matrix W is the product of a confluent Vandermonde matrix V and a diagonal scaling D, i.e., $W = VD$. If $T(c) \succ 0$ and the numbers of rows of V is greater than its number of columns, the complexity of solving the reformulated dual problem

$$\begin{array}{ll} \max & \langle b, y\rangle \\ \text{s.t.} & (W^*T(c)^{-1}W)^{-1} \succeq \Delta(y) \end{array}$$

depends mostly on the dimension of $\Delta(y)$. That is, an appropriate preprocessing tends to eliminate the dependence on the degree n. However, primal strict feasibility cannot be guaranteed from the knowledge of k so that the exact semidefinite relaxation cannot be certified in general.

4.4 Research summary

Our contribution to quadratic optimization developed in this chapter is twofold.

On the one hand, convexity of some cone related to joint numerical ranges is shown to be a sufficient condition for exact relaxation of QQP, see Theorems 4.22 and 4.23. Although this result unifies several well-known statements about easy quadratic problems, it is undervalued in the literature.

On the other hand, new classes of QQP with exact semidefinite relaxation are described. They are related to non-negative polynomials by means of interpolation constraints, see Theorems 4.27 and 4.28.

Notes and references

To our knowledge, only a few papers emphasize the importance of convex joint numerical range in the context of quadratic optimization. B. Polyak [122] considers non-convex quadratic problems with this hidden convexity property. More specifically, he focuses on problems generated by three real matrices. Various optimization and control problems (ellipsoidal approximation, S-procedure) can be unified using his approach. In systems and control, an application of convex joint numerical ranges with three Hermitian matrices appears in μ-analysis, see [50]. Hidden convexity in non-convex quadratic programming was partially investigated in [15]. Our more general approach encompasses part of these results as well as some of their conclusions.

Section 4.3 is based on the conference paper [69]. Standard references on numerical ranges are [82, Chapter 1] and [68]. A recent review of semidefinite relaxation of QQP problems is [113].

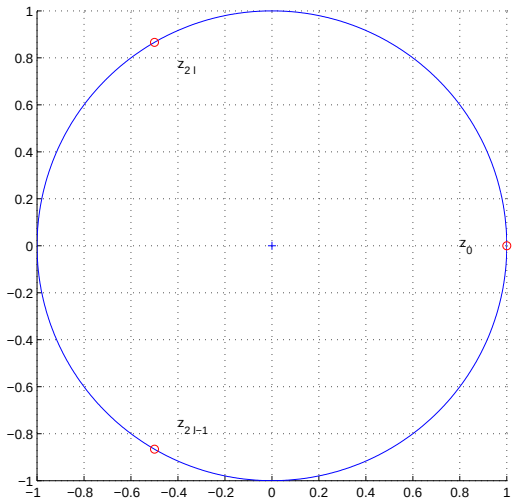

Figure 4.4. *Solution of* $\max\{P_\ell(z) : |z_i| = 1, \forall i\}$

4.5 Appendix

Proposition 4.32. *Let* $A = A^*$ *be a Hermitian matrix of order* $2n+1$. *Then the quadratic optimization problem*

$$\begin{array}{ll} \min & \langle Az, z\rangle \\ \text{s.t.} & |z_i| = 1, \quad i = 0, \ldots, 2n \end{array} \tag{4.57}$$

is NP*-hard.*

Proof. This proof is based on ideas of A. Nemirovskii. Let $\{a_i\}_{i=0}^n \subseteq \mathbb{Z}$ be a finite set of integers. Checking whether there exist $\{x_i\}_{i=0}^n \subseteq \{-1, +1\}$ such that the equality

$$\sum_{i=0}^{2n} a_i x_i = 0 \tag{4.58}$$

holds is related to the subset sum problem and is thus NP-complete.

Let $\{z_\ell\}_{\ell=0}^{2n} \subseteq \mathbb{C}$ be a finite set of complex numbers of modulus one and define the quadratic functions

$$P_\ell(z) = |z_0 - z_{2\ell-1}|^2 + |z_{2\ell-1} - z_{2\ell}|^2 + |z_0 - z_{2\ell}|^2, \quad \ell = 1, \ldots, n.$$

Assume that z_0 is equal to 1 without loss of generality. Then the optimization problem

$$\max\{\sum_{\ell=1}^{n} P_\ell(z) : |z_i| = 1, \forall i\}$$

can be solved explicitly, see Figure 4.4. Note that the inequality

$$\max\{\sum_{\ell=1}^{n} P_\ell(z) - |\sum_{\ell=0}^{n} a_\ell (z_{2\ell+1} - z_{2\ell+2})|^2 : |z_i| = 1, \forall i\} \leq \max\{\sum_{\ell=1}^{n} P_\ell(z) : |z_i| = 1, \forall i\}$$

is tight if and only if Problem (4.58) is solvable. Since its left hand side is an instance of (4.57), this quadratic problem is hard to solve. □

Chapter 5

Quadratic matrix polynomials

Quadratic polynomials, of the form

$$p(x) = ax^2 + bx + c, \quad a, b, c \in \mathbb{R}$$

are well-known mathematical objects and related problems are often solved explicitly. Although quadratic matrix polynomials are generalizations of the above, they do not enjoy their simplicity. For this reason, they deserve special mention.

Section 5.1 describes eigenvalue problems associated with quadratic matrix polynomials. These problems can be classified using the eigenvalue location. Section 5.2 describes several algorithms for computing the Crawford number of a pair (A, B). This problem is considered because some structured matrix pairs are related to some quadratic matrix polynomials. Section 5.3 solves several distance problems, which originate from the classification. Several of our solutions are based on our characterization of non-negative matrix polynomials.

5.1 Quadratic eigenvalue problem

5.1.1 Definition

Let $A, B, C \in \mathbb{C}^{n\times n}$ be three complex matrices. The *quadratic eigenvalue problem* (QEP) is to find scalars λ and vectors x, y such that

$$Q(\lambda)x = (\lambda^2 A + \lambda B + C)x = 0, \tag{5.1}$$

$$y^* Q(\lambda) = y^*(\lambda^2 A + \lambda B + C) = 0. \tag{5.2}$$

Since the roots of the quadratic polynomial

$$Q(\lambda) = \lambda^2 A + \lambda B + C \tag{5.3}$$

are defined as the roots of $\det Q(\lambda)$, the eigenvalues of (5.1) are exactly the roots of $Q(\lambda)$. This eigenvalue problem is an extension of the generalized eigenvalue problem, for which the matrix polynomial $Q(\lambda)$ is of degree 1. The main algebraic difference between both problems is that the QEP has $2n$ eigenvalues (finite or infinite) with up to $2n$ right and $2n$ left eigenvectors.

The quadratic eigenvalue problem has a wide range of applications, from vibration analysis to fluid dynamics. The recent survey [134] contains a list of its many applications, its mathematical properties and several numerical methods for this problem class. The next examples have been selected because of their relevance.

Example 5.1 (Second-order differential equation). Consider a linear homogeneous second-order differential equation

$$Ax''(t) + Bx'(t) + Cx(t) = 0, \tag{5.4}$$

where $A, B, C \in \mathbb{C}^{n\times n}$ are three Hermitian matrices. In many engineering applications, this equation models the physical behaviour of a system from a mathematical point of view. For instance, second-order differential equations are very common in electrical and mechanical engineering. In this context, the matrix A is often related to the kinetic energy so that its positive definiteness is a natural assumption. A solution to (5.4) can be expressed in terms of the eigenvalues and eigenvectors of the corresponding QEP.

Example 5.2 (Constrained least-squares problem). Let $A \in \mathbb{R}^{n\times n}$ be symmetric and $b \in \mathbb{R}^n$. Consider the *constrained least-squares problem*

$$\min\{x^T Ax - 2b^T x : x^T x = \alpha^2\}. \tag{5.5}$$

This problem can be reduced to a QEP by using the optimality conditions:

$$(A - \lambda I)x = b, \quad x^T x = \alpha^2, \tag{5.6}$$

where $\lambda \in \mathbb{R}$ is the Lagrange multiplier associated with the equality constraint $x^T x = \alpha^2$. The smallest solution λ of these equations is needed to solve our optimization problem (5.5) [53]. Assuming that λ is not an eigenvalue of A and setting

$$y = (A - \lambda I)^{-2} b = (A - \lambda I)^{-1} x,$$

conditions (5.6) are equivalent to

$$(A - \lambda I)^2 y = b, \quad b^T y = \alpha^2.$$

By expanding these equations, we get the symmetric QEP

$$(\lambda^2 I_n - 2\lambda A + (A^2 - \alpha^{-2} bb^T))y = 0. \tag{5.7}$$

The solution of (5.5) is $x = (A-\lambda I)^{-1}b$, where λ is the smallest *real* eigenvalue of (5.7). Therefore, it can be obtained in polynomial time using standard linear algebra tools.

In this chapter, we focus on self-adjoint quadratic eigenvalue problems: A, B and C are Hermitian matrices. Moreover, we assume that A is positive definite.

5.1.2 Classification

Quadratic eigenvalue problems have been classified in different sets according to the eigenvalues of A, B and C and to the location of their eigenvalues on the complex plane, see [134, Table 1.1]. Let us describe two important classes of self-adjoint QEP.

A self-adjoint quadratic eigenvalue problem is said to be *elliptic* if

$$(x^*Bx)^2 < 4(x^*Ax)(x^*Cx) \quad \text{for all non-zero } x \in \mathbb{C}^n.$$

Because any eigenpair (x, λ) satisfies

$$\lambda^2(x^*Ax) + \lambda(x^*Bx) + x^*Cx = 0,$$

it follows that elliptic QEP have non-real eigenvalues. Moreover, the matrix C must be positive definite. A self-adjoint QEP is said to be *hyperbolic* if

$$(x^*Bx)^2 > 4(x^*Ax)(x^*Cx) \quad \text{for all non-zero } x \in \mathbb{C}^n.$$

Note that hyperbolic QEP have real eigenvalues. Of course, there exist self-adjoint QEP which are neither elliptic nor hyperbolic.

Remark. Ellipticity or hyperbolicity of a given QEP can sometimes be asserted from its origin, i.e., from the application it models.

5.1.3 Eigenvalue computation

Different methods can be used to compute the eigenvalues of a matrix polynomial. Given a self-adjoint quadratic polynomial $Q(\lambda) = \lambda^2 A + \lambda B + C$, the most straightforward method is based on the *Hermitian linearization*

$$\begin{bmatrix} -C & 0 \\ 0 & A \end{bmatrix} - \lambda \begin{bmatrix} B & A \\ A & 0 \end{bmatrix}.$$

Indeed, the eigenvalues of this pencil are exactly the eigenvalues of $Q(\lambda)$. However, this linearization is not unique. For a given generic QEP, one can also consider the first companion form

$$\begin{bmatrix} 0 & X \\ -C & -B \end{bmatrix} - \lambda \begin{bmatrix} X & 0 \\ 0 & A \end{bmatrix}$$

or the second companion form

$$\begin{bmatrix} -C & 0 \\ 0 & X \end{bmatrix} - \lambda \begin{bmatrix} B & A \\ X & 0 \end{bmatrix},$$

where X is any nonsingular $n \times n$ matrix. The eigensolutions of (5.1) and (5.2) can be obtained from the eigensolutions of the matrix pencil

$$\tilde{A} - \lambda \tilde{B} \doteq \begin{bmatrix} 0 & X \\ -C & -B \end{bmatrix} - \lambda \begin{bmatrix} X & 0 \\ 0 & A \end{bmatrix}.$$

Indeed, the equations $Q(\lambda_i)x_i = 0$ and $y_i^*Q(\lambda_i) = 0$ are equivalent to $(\tilde{A} - \lambda_i \tilde{B})v_i = 0$ and $w_i^*(\tilde{A} - \lambda_i \tilde{B}) = 0$ with

$$v_i = \begin{bmatrix} x_i \\ \lambda_i x_i \end{bmatrix}, \quad \text{and} \quad w_i = \begin{bmatrix} y_i \\ \overline{\lambda_i} y_i \end{bmatrix}.$$

Using standard linear algebra software, a numerical solution to this eigenvalue problem can be calculated in $\mathcal{O}(n^3)$ flops. We refer the reader to the recent survey [134] for a complete overview of all available methods.

5.2 Crawford number

Given two Hermitian matrices A and B, we would like to compute the generalized eigenvalues of the pencil $A - \lambda B$. Clearly, the problem is easy to solve for pairs of upper triangular matrices. In this section, we focus on definite pairs (A, B), which permit a simultaneous diagonalization of both matrices [66, Section 8.7].

5.2.1 Definite pair

Two Hermitian matrices $A, B \in \mathbb{C}^{n\times n}$ form a *definite pair* if the Crawford number

$$\gamma(A, B) \doteq \min\{\sqrt{(z^*Az)^2 + (z^*Bz)^2} : z \in \mathbb{C}^n, \|z\|_2 = 1\} \tag{5.8}$$

is strictly greater than 0. Equivalently, the value $\gamma(A, B)$ can be obtained from the following optimization problem

$$\begin{array}{ll} \min & |w| \\ \text{s.t.} & w \in \mathcal{F}(A + \jmath B), \end{array} \tag{5.9}$$

where the *field of values* of $G \in \mathbb{C}^{n\times n}$ is defined by

$$\mathcal{F}(G) = \{z^*Gz : \|z\|_2 = 1, z \in \mathbb{C}^n\}.$$

Thus $\gamma(A, B)$ is positive if and only if $\mathcal{F}(A + \jmath B)$ and $\gamma(A, B)$ is the distance from the origin to the nearest point in $\mathcal{F}(A + \jmath B)$.

Finding the distance between a convex set Q and a given point x_0 is related to its projection onto the convex set, so it is a convex optimization problem. As the field of values is convex, computing the Crawford number can thus be recast using a convex formulation.

Proposition 5.1.

$$\begin{array}{rll} \gamma(A, B) = \max & \lambda_{\min}(Au_0 + Bu_1) \\ \text{s.t.} & u_0^2 + u_1^2 \leq 1 \end{array} \tag{5.10}$$

Proof. First of all, convexity of the joint numerical range

$$W = \{(\langle Az, z\rangle, \langle Bz, z\rangle) : z \in \mathbb{C}^n, \|z\|_2 = 1\}$$

implies that the original problem (5.8) can be rewritten as a convex optimization problem in conic form

$$\begin{array}{ll} \min & \tau \\ \text{s.t.} & \langle A, Z\rangle - x = 0, \\ & \langle B, Z\rangle - y = 0, \\ & \langle I_n, Z\rangle = 1, \\ & (x, y, \tau) \in \mathcal{L}_2, \\ & Z \in \mathcal{H}_+^n, \end{array} \tag{5.11}$$

where $\mathcal{L}_2$ is a Lorentz cone and $\mathcal{H}_+^n$ is a cone of Hermitian positive semidefinite matrices. Although the number of constraints is low, the number of variables in this primal formulation is rather high (approximately $\mathcal{O}(n^2)$). Therefore the problem dual to (5.11)

$$\begin{array}{ll} \max & \sigma \\ \text{s.t.} & Au_0 + Bu_1 \succeq \sigma I_n, \\ & u_0^2 + u_1^2 \leq 1 \end{array} \tag{5.12}$$

is computationally more attractive. Moreover, strict feasibility of both problems implies a zero duality gap. Using the standard characterization of $\lambda_{\min}(Au_0 + Bu_1)$ completes our proof. □

Since the objective function $\lambda_{\min}(Au_0 + Bu_1)$ is homogeneous, the constraint $u_0^2 + u_1^2 \leq 1$ is active at the optimum, provided that $\gamma(A, B) > 0$. In this case, we may assume that $u_0 = \cos\theta$ and $u_1 = \sin\theta$ without loss of generality.

Although a sufficient condition for definiteness of (A, B) is that one of A or B is definite, Proposition 5.1 shows that definiteness of an appropriate linear combination is the relevant characterization.

Since the generalized eigenvalues of $Ax = \lambda Bx$ are real for definite pairs, there exists a strong link between definite pairs and hyperbolic systems.

Theorem 5.2 ([76]). *The self-adjoint QEP* (5.1) *with A positive definite is hyperbolic if and only if the pair $(\tilde{A}, \tilde{B})$ is definite, where*

$$\tilde{A} = \begin{bmatrix} -C & 0 \\ 0 & A \end{bmatrix}, \quad \tilde{B} = -\begin{bmatrix} B & A \\ A & 0 \end{bmatrix}.$$

This theorem is a nice consequence of the linearization. Remember that the eigenvalues of $Q(\lambda)$ are the eigenvalues of a linearization defined by $\tilde{A}$ and $\tilde{B}$; all eigenvalues of this linearization are real if the pair $(\tilde{A}, \tilde{B})$ is definite.

5.2.2 Standard numerical methods

In order to check whether a pair (A, B) of matrices is definite, Crawford and Moon [35] present a bisection-like algorithm. At each iteration, a Cholesky factorization is used to test definiteness of $-A\sin\theta_k + B\cos\theta_k$ for the current estimate θ_k. This algorithm requires at most $\mathcal{O}(n)$ steps, so that its total complexity is $\mathcal{O}(n^4)$ flops. As it does not compute the Crawford number explicitly, it will not be considered in the sequel.

Bisection algorithm

Note that computing $\gamma(A, B)$ is a one-dimensional optimization problem. Indeed, we obtain from Proposition 5.1 that

$$\gamma(A, B) = \max\{0, \max_{0 \leq \theta \leq 2\pi} \lambda_{\min}(A\cos\theta + B\sin\theta)\}.$$

Let $C = A + \jmath B$ and rewrite the above expression using the associated trigonometric polynomial $M(z) = z^{-1}C + zC^*$ to get the optimization problem

$$\theta(A, B) = \max\{f(z) \doteq \lambda_{\min}(M(z)) : |z| = 1\}.$$

Note that $\gamma(A, B) = \max\{0, \theta(A, B)\}$. Higham, Tisseur and Van Dooren have adapted an algorithm of Byers [24] so as to iteratively refine an interval $[a, b]$ containing $\theta(A, B)$ [76]. The idea underlying their scheme is very simple. Since z belongs to the unit circle, the function $f(z)$ is a continuous function of z so it achieves every value between its minimum and its maximum. Given a level ξ, their algorithm checks whether this level is attained by $f(z)$. Depending on the answer, it updates either the lower bound a or the upper bound b. The formal description is as follows.

Algorithm 5.1 Crawford number – bisection method [76, Algorithm 2.3]

1. Choose an accuracy $\epsilon > 0$.
   Set $a = \max\{\lambda_{\min}(A), -\lambda_{\max}(A), \lambda_{\min}(B), -\lambda_{\max}(B)\}$ and $b = \sigma_n\left(\begin{bmatrix} A \\ B \end{bmatrix}\right)$. If (A, B) is known to be definite, set $a = \max(a, 0)$.
2. kth iteration ($k \geq 0$).
   (a) Set the level ξ to $\frac{a+b}{2}$.
   (b) Compute the eigenvalues $\{z_i\}_i$ of $Q(z) \doteq C - 2\xi z I + z^2 C^*$
   (c) If $\lambda_{\min}(M(z_i)) = \xi$ for some eigenvalue z_i with unit modulus, set $a = \xi$. Otherwise, set $b = \xi$.
3. Stop the process if $b - a \leq \epsilon$ or $b \leq 0$. If $b \leq 0$, set $a = b = 0$.

Note that the eigenvalues of $Q(z)$ with unit modulus are exactly the crossing of the eigenvalues of $M(z)$ with the level ξ. Indeed, we have

$$\begin{aligned} \det(A\cos\theta + B\sin\theta - \xi I) = 0 &\iff \det(M(z) - \xi I) = 0, z = e^{\jmath\theta}, \\ &\iff \det(C - 2\xi z I + z^2 C^*) = 0, z = e^{\jmath\theta}. \end{aligned}$$

Consequently, these crossing can be computed at each step from the generalized eigenvalue problem

$$\lambda \begin{bmatrix} I_n & 0 \\ 0 & C^* \end{bmatrix} - \begin{bmatrix} 0 & I_n \\ -C & 2\xi I_n \end{bmatrix} \tag{5.13}$$

at a cost of $\mathcal{O}(n^3)$ flops. Up to $2n$ eigenvalue problems must be solved at each iteration. Due to its bisection rule, the whole scheme has a linear rate of convergence so the total complexity is equal to $\mathcal{O}(n^4 \log \frac{1}{\epsilon})$ flops.

Level set algorithm

A faster scheme is obtained by applying a variant of the level set algorithm presented in [130, 56]. The idea is to consider all eigenvalues of $A\cos\theta + B\sin\theta - \xi I$ as a function of θ. At the "zero crossings", the derivatives

$$\frac{\partial}{\partial\theta}\lambda_i(A\cos\theta + B\sin\theta - \xi I)$$

can be evaluated using the eigensolution of the linearization (5.13). Analysing these derivatives then yields the pair(s) of zero crossing corresponding to the smallest eigenvalue and so the interval(s) containing the optimal solution, see Figure 5.1 ($\xi = 0$). The signs of the derivatives are indicated with $\triangle$ (positive crossing) and with ∇ (negative crossing). Clearly, the pattern of relevant intervals is $[\theta(\triangle), \theta(\nabla)]$. Finally, the minimal eigenvalue of $A\cos\theta + B\sin\theta$ is computed at the midpoint of the largest interval to obtain the next level.

Algorithm 5.2 Crawford number – level set method [76, Algorithm 2.4]

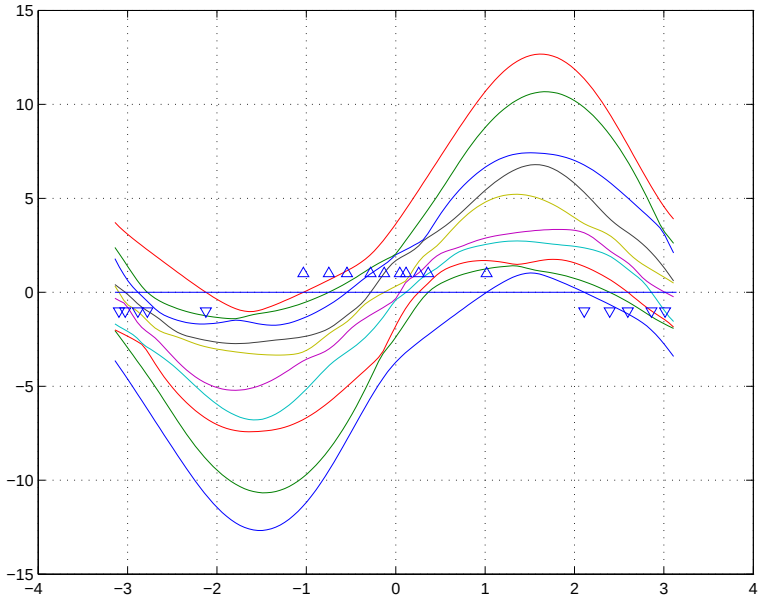

Figure 5.1. *Eigenvalues of a* 10×10 *matrix* $A\cos\theta + B\sin\theta$ *as a function of* θ

1. Choose an accuracy $\epsilon > 0$. Set $\xi = 0$.
2. kth iteration ($k \geq 0$).
   (a) Compute the eigenvalues $\{z_i\}_i$ of $Q(z) \doteq C - 2\xi zI + z^2C^*$ via the linearization (5.13).
   (b) From the values of the derivatives at each crossing, find some candidate interval(s) containing the optimal solution.
   (c) Set $\xi = \lambda_{\min}(A\cos\theta_k + B\sin\theta_k)$, where θ_k is the midpoint of the largest interval.
3. Stop the process if ξ cannot be improved, i.e., the lowest eigenvalue does not cross the level $\xi + \epsilon$.

Each iteration of this scheme costs $\mathcal{O}(n^3)$ flops. It has a guaranteed linear rate of convergence, so the total complexity is equal to $\mathcal{O}(n^3 \log \frac{1}{\epsilon})$. If $\lambda_{\min}(A\cos\theta + B\sin\theta)$ is not multiple at its maximum, then this scheme converges quadratically [59, 76]. In order to get a higher order of convergence (under mild smoothness assumptions), variants were proposed that exploit the value of the derivatives at the zero crossing [59].

5.2.3 Structure specific method

Although they do not rely on convexity nor on problem structure, the previous algorithms are valuable because they computed the Crawford number in polynomial time. Let us consider both aspects, convexity and structure, in order to prove that the above complexity results may be improved.

Problem structure

Let us start from the structure of the problem. The next proposition establishes partial convexity of our objective function (5.10) in the non-degenerate case.

Proposition 5.3. *Let $A, B \in \mathbb{R}^{n\times n}$ be Hermitian matrices. The function $\lambda_{\min}(A\cos\theta + B\sin\theta)$ is concave in θ on the intervals*

$$\{\theta \in \mathbb{R} : \lambda_{\min}(A\cos\theta + B\sin\theta) \geq 0\}.$$

In addition,

$$\lambda_{\min}(A\cos\theta + B\sin\theta) = \min\{x\cos\theta + y\sin\theta : x + \jmath y \in \mathcal{F}(A + \jmath B)\}.$$

Proof. The value $\gamma(A, B)$ can also be obtained from rotating the field of values $\mathcal{F}(A + \jmath B)$ around the origin. To see this, write $C = A + \jmath B$ and note that $\mathcal{F}(e^{-\jmath\theta}C) = e^{-\jmath\theta}\mathcal{F}(C)$. Using this observation, it can be proved that $\gamma(A, B)$ is the maximal distance to the imaginary axis of these rotated convex sets [31, Theorem 2.1].

Let $x + \jmath y$ be an arbitrary point of $\mathcal{F}(A + \jmath B)$ and apply a rotation θ. The distance to the imaginary axis of the rotated point is equal to $x\cos\theta + y\sin\theta$. Using elementary trigonometric formula, we get that

$$x\cos\theta + y\sin\theta = \|[x\ y]\|_2 \cos(\theta - \theta_{xy})$$

where $\theta_{xy} = \arctan(y/x)$. Note that this distance has the same period 2π for any $x + \jmath y \in \mathcal{F}(A + \jmath B)$ and any rotation θ.

Consider an interval where the curves $x\cos\theta + y\sin\theta$ are non-negative for all pair $x + \jmath y \in \mathcal{F}(A + \jmath B)$. Concavity of the cosine function on its domain of non-negativity yields that the minimum of all these curves is concave, see Figure 5.2. To conclude, note that this minimum function is exactly our function $\lambda_{\min}(A\cos\theta + B\sin\theta)$, i.e., the distance to the imaginary axis of these rotated fields of values. □

Cutting plane method

As shown above, the Crawford number can be computed via the convex optimization problem

$$\begin{array}{ll} \max & \sigma \\ \text{s.t.} & Au_0 + Bu_1 \succeq \sigma I_n, \\ & u_0^2 + u_1^2 \leq 1. \end{array}$$

At the optimal solution, we have either $u = 0$ or $\|u\|_2 = 1$. This formulation is well adapted to the problem structure and it can be solved very efficiently using a one-dimensional tailor-made cutting plane method. Remember that the objective function is $f(\theta) = \lambda_{\min}(A\cos\theta + B\sin\theta)$.

Assume that $u \neq 0$ at the optimum. The optimal solution is then obtained at a point u on the unit circle. Our cutting plane algorithm locates the optimal u within an interval of the unit circle using an iterative scheme. Any cutting plane, which contradicts the current localization set, is a certificate that $\sigma = u = 0$. Indeed, in this case, the next localization set would be empty so that our starting assumption cannot be true. The complete algorithm is as follows.

Algorithm 5.3 Crawford number – cutting plane method

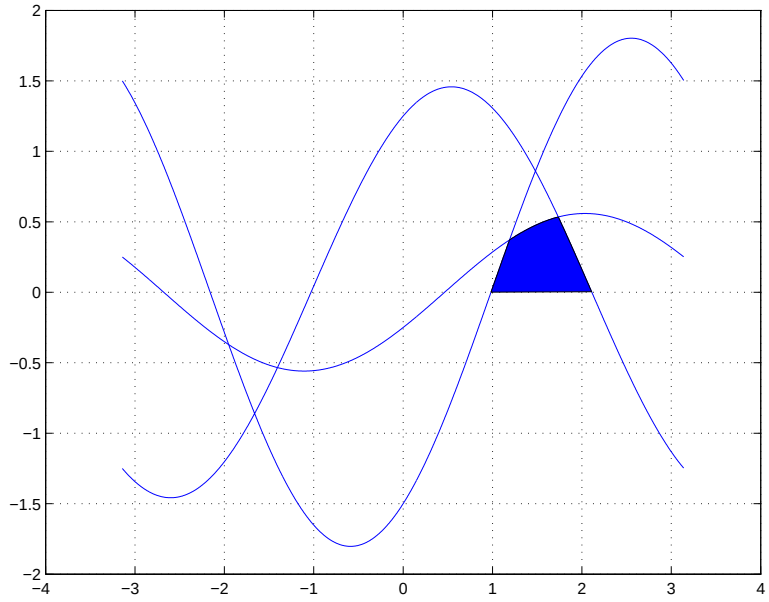

Figure 5.2. *Concavity of* $\lambda_{\min}(A\cos\theta + B\sin\theta)$ *on its non-negativity domain*

1. Choose an accuracy ϵ. Let $\sigma = 0$. Pick a vector $x \in \mathbb{C}^n$ at random and define the initial localization set as

$$\mathcal{L}_0 = \{u_0 + \jmath u_1 : (x^*Ax)u_0 + (x^*Bx)u_1 \geq 0, \|u\|_2 = 1\}.$$

   The starting point u_0 is chosen as the center of $\mathcal{L}_0$.

2. kth iteration ($k \geq 0$).

   (a) Given the current query point u_k, compute the minimal eigenvalue λ_k of $Au_{k,1} + Bu_{k,2}$ and an associated eigenvector v_k.

   (b) If $\lambda_k < \sigma$, update the localization set using the feasibility cut generated by v_k

$$\mathcal{L}_{k+1} \leftarrow \mathcal{L}_k \cap \{u_0 + \jmath u_1 : (v_k^*Av_k)u_0 + (v_k^*Bv_k)u_1 \geq \sigma\}.$$

   (c) If $\lambda_k \geq \sigma$, update the objective value and the localization set using the optimality cut generated by v_k

$$\begin{aligned} \sigma &\leftarrow \lambda_k, \\ \mathcal{L}_{k+1} &\leftarrow \mathcal{L}_k \cap \{u_0 + \jmath u_1 : (v_k^*Av_k)(u_0 - u_{k,0}) + (v_k^*Bv_k)(u_1 - u_{k,1}) \geq 0\}. \end{aligned}$$

   (d) Improve the new localization using structure specific information, if possible. Then select the next query point inside the new localization set $\mathcal{L}_{k+1}$.

3. Stop the process if an ϵ-solution is obtained or if the localization set is detected to be empty. In the latter case, return $u = 0$.

This algorithm is essentially a one-dimensional cutting plane method. The only structure specific step is the one performed at the end of each iteration. In the sequel, we detail three different strategies for improving the localization set and for choosing the next query point.

Query point generator

The first strategy for choosing the next query point is to use a standard bisection rule. Let $\mathcal{L}_k$ be the current localisation set. This set is an arc of the unit circle and the next query point is chosen as the mid-point of this arc. This strategy ensures that the length of the localization arc is divided by a factor greater or equal to two at each iteration.

Algorithm 5.4 Query point selection – bisection rule
Given the current localisation set $\mathcal{L}_k$ return the mid-point of the corresponding arc.

The next strategies take into account the structure of the objective function, see Proposition 5.3.

Let u_k be the current query point, which is on the unit circle, and let θ_k be its phase. The values of the objective function f and of its derivative at the query point are obtained via an eigenvalue problem in $\mathcal{O}(n^3)$ flops at the first step of each iteration. From the structure of the objective function, we know that $(\theta_k, f(\theta_k))$ belongs to the graph of a trigonometric function

$$F_k(\theta) = \alpha_k \cos(\theta - \beta_k).$$

The parameters (α_k, β_k) are obtained from the system of equations

$$\begin{aligned} \alpha_k \cos(\theta_k - \beta_k) &= f(\theta_k), \\ -\alpha_k \sin(\theta_k - \beta_k) &= v_k^*(-A \sin\theta_k + B\cos\theta_k)v_k. \end{aligned}$$

The above procedure allows us to identify a cosine function, which is responsible for the shape of f at the current query point.

Our strategies to be described next use either the full information (function value, derivative and cosine shape) or a subset (function value and derivative) for improving the localization set and for generating the next query point.

In the case of a feasibility cut, we have $f(\theta_k) < \sigma$. If $f'(\theta_k) < 0$, the function $F_k(\theta)$ crosses the current level σ at a point $\tilde{\theta}_k$ so that any point between $\tilde{\theta}_k$ and θ_k cannot belong to the current localization set, see Figure 5.3. Therefore the localization set must be modified to reflect this property. If $f'(\theta_k) > 0$, we clearly get the symmetric procedure. The next query point is then obtained by a bisection rule.

In the case of an optimality cut, the localization set cannot be improved using the problem structure while an interesting guess of the next query point is possible. On the one hand, the cosine structure of the problem can be fully used, see the generic situation on Figure 5.4. All intersection points of the trigonometric functions $\{F_k\}_{k\in O}$, which were obtained during optimality cuts, are computed. At each intersection abscissa ($\circ$), the minimum of our functions $\{F_k\}_{k\in O}$ is then calculated. The next query point is selected as the one that maximizes these minimum values. This strategy is clearly a discrete version of our initial optimization problem. To avoid slow downs,

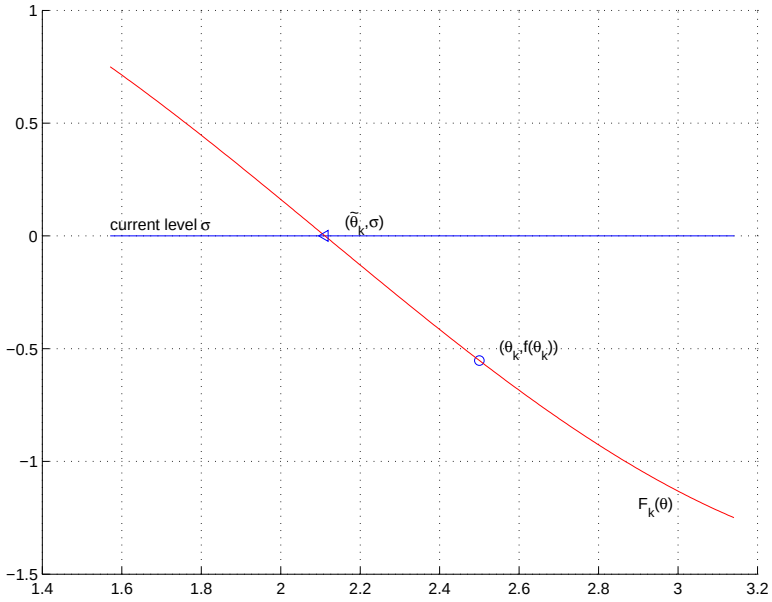

Figure 5.3. *Improvement of localisation set for feasibility cuts.*

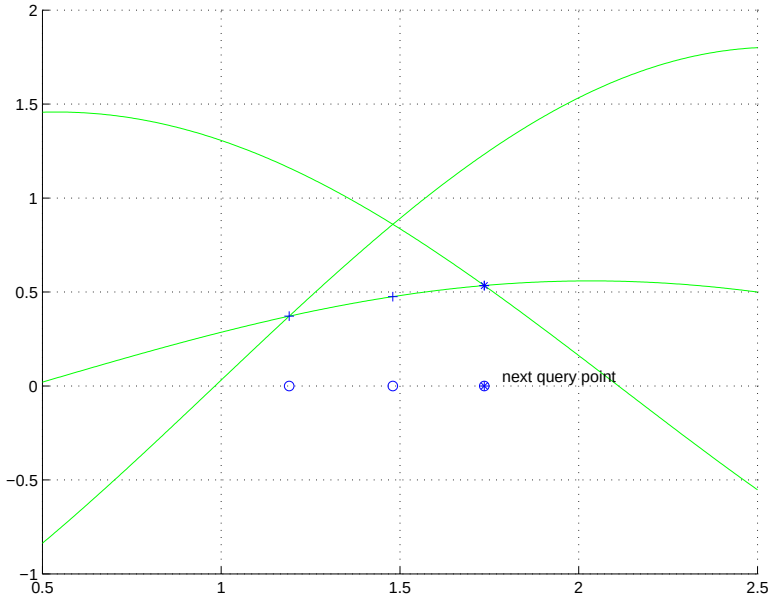

Figure 5.4. *Next query point using cosine structure.*

old trigonometric functions can be discarded as the algorithm proceeds. On the other hand, after a few steps, the values of f and f' (or good approximations) are known at both endpoints $[\triangleright, \triangleleft]$ of the localization set, see Figure 5.5. We can thus fit a cubic

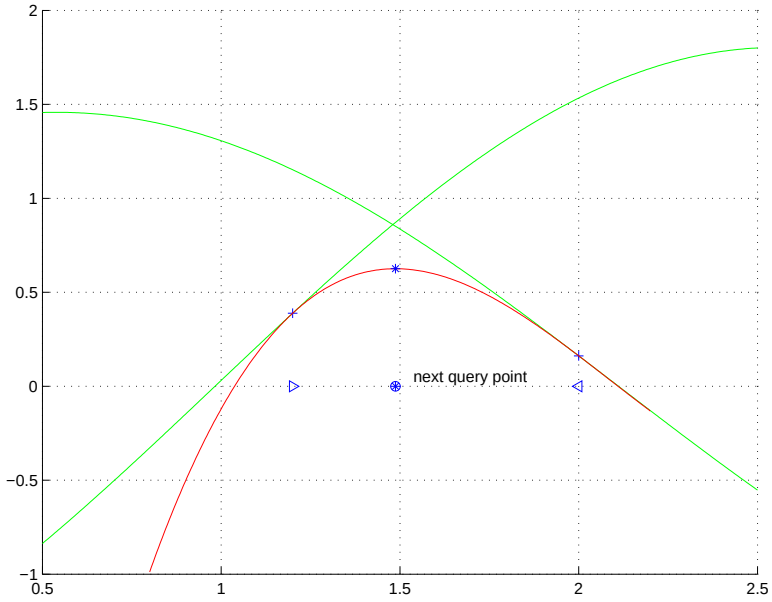

Figure 5.5. *Next query point using cubic fit.*

function through these points. The next query point is selected as the extremum of this cubic function that lies within the current localization set. Finally, it may happen for numerical reasons that the next query point does not belong to the current localization set. In this case, the next query point is obtained by a bisection rule.

Algorithm 5.5 Query point selection – cosine structure

1. Feasibility cut
   (a) Compute the intersection of $F_k(\theta)$ with the level σ. Let $\hat{\theta}_k \in \mathcal{L}_{k+1}$ be the corresponding point.
   (b) Reduce the localization set from either the left side or the right side:
   $$\mathcal{L}_{k+1} \leftarrow \mathcal{L}_{k+1} \cap \{u = e^{\jmath\theta} : \theta \in [\hat{\theta}_k - \pi, \hat{\theta}_k]\} \text{ if } F_k'(\theta_k) < 0$$
   $$\mathcal{L}_{k+1} \leftarrow \mathcal{L}_{k+1} \cap \{u = e^{\jmath\theta} : \theta \in [\hat{\theta}_k, \hat{\theta}_k + \pi]\} \text{ if } F_k'(\theta_k) \geq 0$$
   (c) Select the next query point as the midpoint of L.
2. Optimality cut
   (a) Compute all intersections of $\{F_k\}_{k\in O}$ and select those belonging to the current localisation set.
   (b) For each intersection abscissa, compute the minimum of our functions $\{F_k\}_{k\in O}$.

(c) Select the next query point as the one that maximizes these minimum values. If this point does not belong to the current localization set, select the next query point as the midpoint of $\mathcal{L}_{k+1}$.

Algorithm 5.6 Query point selection – cubic fit

1. Feasibility cut: see Algorithm 5.5
2. Optimality cut
   (a) If estimates of f and f' are not available at the endpoints of the localization set, select the next query point as the midpoint of $\mathcal{L}_{k+1}$. This case is only use at the very beginning of the cutting plane scheme
   (b) If estimates of f and f' are available at the endpoints of the localization set, fit a cubic polynomial and select the next query point as the local extremum within the current localization set. If there is no such point, select the next query point as the midpoint of $\mathcal{L}_{k+1}$.

Upper bound and complexity

The convergence of our cutting plane scheme is clearly influenced by our ability to derive appropriate upper bounds. At the end of each iteration, the trigonometric function F_k and our new query point, say t, yield such an upper bound. Indeed, we have

$$\theta(A, B) \leq F_k(t),$$

provided that $f'(\theta_k) \equiv F'_k(\theta_k)$ and $F'_k(t)$ have the same sign.

Although our cutting plane schemes have a guaranteed linear rate of convergence, they are competitive with more efficient schemes and in particular with Algorithm 5.2. Indeed, each iteration of Algorithm 5.3 involves the computation of the smallest eigenvalue (and an associated eigenvector) of a Hermitian matrix, which can be done in $\mathcal{O}(n^3)$ flops. Our strategies for getting the next query point do not have a prohibitive cost. Note that, in Algorithm 5.2, we need to compute all eigenvalues and a subset of eigenvectors of the $2n \times 2n$ matrix pencil (5.13). Although this can be done in $\mathcal{O}(n^3)$ flops too, the constant factor of n^3 is much greater than that for standard eigenvalue problems [66].

5.2.4 Numerical results

All algorithms described in this section (see Table 5.1) have been applied to three different problems. Our first two examples are those considered in [76, Section 2].

Example 5.3. The matrices A and B are random real symmetric matrices of order 5. This pair is definite and its Crawford number is equal to $1.3541 \cdot 10^{-1}$. The convergence of our schemes is illustrated on Figure 5.6.

Table 5.1. *Algorithms for computing the Crawford number of* (A, B)

SDP	Semidefinite formulation (5.12) solved with SDPT3
Byers-like	Bisection method – Algorithm 5.1
Level set	Level set method – Algorithm 5.2
CP bisection	Cutting plane scheme 5.3 with generator 5.4
CP cubic fit	Cutting plane scheme 5.3 with generator 5.6
CP cosine structure	Cutting plane scheme 5.3 with generator 5.5

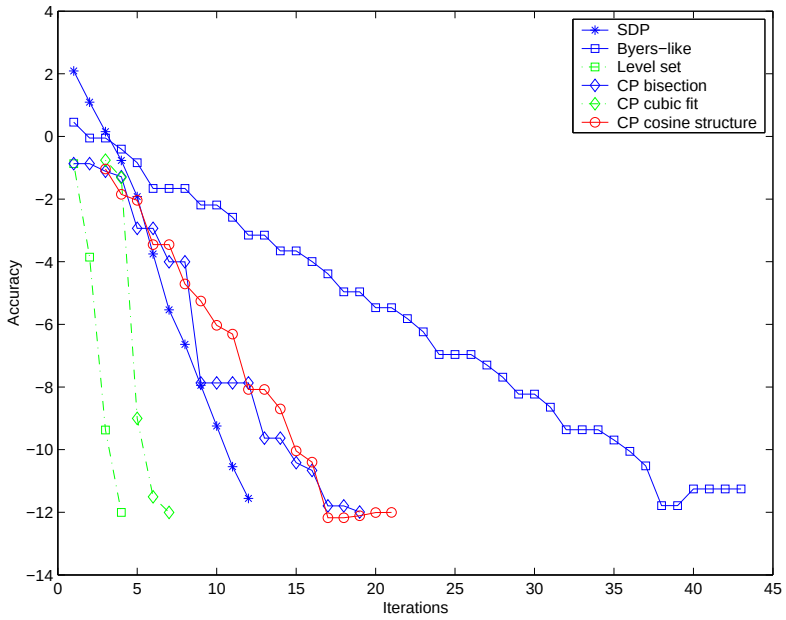

Figure 5.6. *Numerical convergence for Example 5.3*

Example 5.4. The matrices A and B are 10×10 Fiedler and Moler matrices, respectively. This pair is definite and its Crawford number is equal to $1.8678 \cdot 10^{-1}$. The convergence of our schemes is illustrated on Figure 5.7.

Let us comment these numerical examples. Note that Algorithm 5.1 is the worst scheme in both cases. It thus makes sense to look for more efficient schemes. The semidefinite solver has linear convergence and is very stable. Two cutting plane schemes (Algorithms 5.4 and 5.5) display a chaotic linear convergence, as predicted by the theory. The most interesting fact is that Algorithms 5.2 and 5.6 exhibit a quadratic convergence. This is rather surprising for the cutting plane scheme and this could be explained by the efficiency of our query point generator. Indeed, for both examples the objective function is smooth so that a cubic fit is a good local approximation. As Algorithm 5.6 is less expensive from a computational point of view than Algorithm 5.2, we would recommend the former. However, this recommendation is no longer true for pairs (A, B) with non-smooth field of values $A + \jmath B$, as shown by our last example.

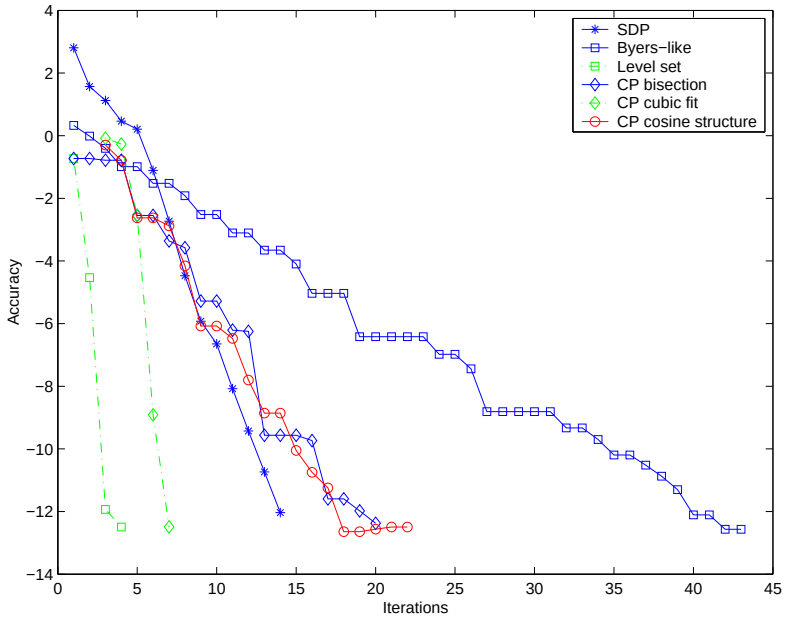

Figure 5.7. *Numerical convergence for Example 5.4*

Example 5.5. The matrices A and B are diagonal matrices

$$A = \begin{bmatrix} -1/4 & 0 & 0 \\ 0 & 5/4 & 0 \\ 0 & 0 & -3/2 \end{bmatrix}, \quad B = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 3/4 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The field of values $\mathcal{F}(A+\jmath B)$ is the convex hull of $\{-0.25+0.5\jmath, 1.25+0.75\jmath, -1.5+\jmath\}$ so the exact Crawford number is $\frac{13\cdot\sqrt{37}}{4\cdot 37}$. The convergence of our schemes is illustrated on Figure 5.8.

This last test instance is related to a pair (A, B) with a polyhedral field of values. Although the Crawford number can be computed explicitly, this example is very interesting. As $\lambda_{\min}(A\cos\theta + B\sin\theta)$ is the minimum of three cosine functions (see Figure 5.2), the objective function is not smooth. For this reason, Algorithm 5.2 does not converge quadratically any more and Algorithm 5.6 does not perform very well too. These observations contrast with our previous examples. Because Algorithm 5.5 takes advantage of the cosine structure for getting the next query point, the optimal solution is found with $4 = 3+1$ iterations. Note that the last iteration is used to assert optimality.

5.3 QEP distance problems

Given an elliptic or hyperbolic self-adjoint quadratic eigenvalue problem, our aim is to compute the smallest perturbation such that the corresponding property is lost by the perturbed problem. This is clearly a distance problem (how much the problem can be altered without losing its defining property?). The converse problem is also of interest

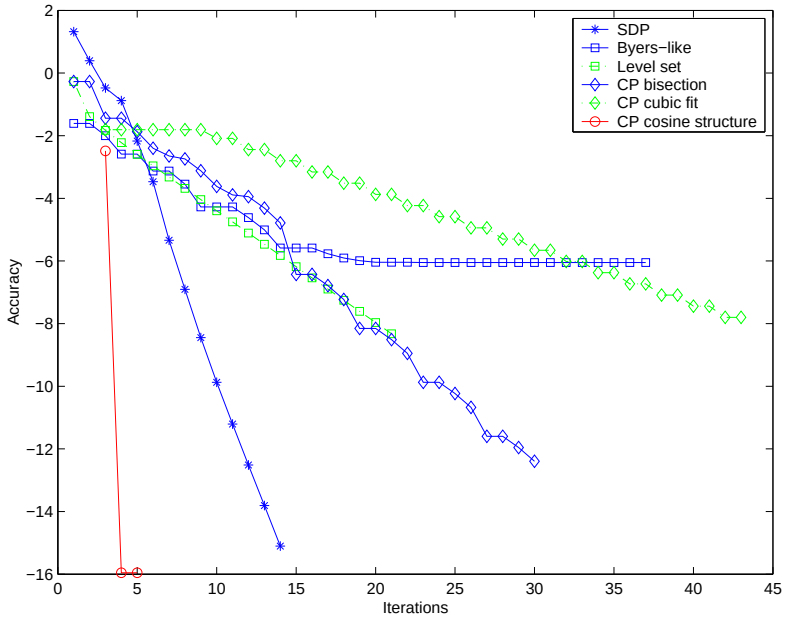

Figure 5.8. *Numerical convergence for Example 5.5*

(what is the nearest QEP having the desired property?). This section proposes explicit formulas to compute these distances and the optimal perturbations. Let us first review the method proposed in the literature.

In [76] the distance to the nearest non-hyperbolic or non-elliptic quadratic eigenvalue problem is solved by making use of the Hermitian matrix

$$W(x, A, B, C) = \begin{bmatrix} 2x^*Ax & x^*Bx \\ x^*Bx & 2x^*Cx \end{bmatrix}.$$

Indeed, $\det W(x, A, B, C)$ is strictly positive for all non-zero $x \in \mathbb{C}^n$ if the QEP is elliptic and $\det W(x, A, B, C)$ is strictly negative for all non-zero $x \in \mathbb{C}^n$ if the QEP is hyperbolic. The minimal distance is computed by solving a non-convex global optimization problem, from which the optimal perturbations can then be recovered. As there is no guarantee to obtain a global optimum, this optimization problem can be considered to be difficult to solve efficiently. Moreover, the perturbations have no easy interpretation in the original polynomial setting.

In the sequel, we propose a simpler approach based on the trigonometric matrix polynomial

$$\begin{aligned} P(\omega) &= \sin^2\omega A + \cos\omega \sin\omega B + \cos^2\omega C \\ &= \begin{bmatrix} \sin(\omega)I & \cos(\omega)I \end{bmatrix} \begin{bmatrix} A & B/2 \\ B/2 & C \end{bmatrix} \begin{bmatrix} \sin(\omega)I \\ \cos(\omega)I \end{bmatrix}, \end{aligned}$$

from which optimal perturbations can be efficiently obtained. Note that $P(\omega) \equiv \frac{Q(\lambda)}{\lambda^2+1}$ with $\lambda = \tan\omega$.

Our solutions are based on the minimal and maximal eigenvalues of $P(\omega)$, regarded as a matrix function of ω. Once the critical frequency $\hat{\omega}$ is identified, an appropriate eigenvector $\hat{x}$ of $P(\hat{\omega})$ allows us to construct the optimal perturbation $\Delta Q(\lambda)$. Subsections 5.3.1 and 5.3.2 deal with elliptic-related and hyperbolic-related distance problems, respectively.

5.3.1 Distance problems related to elliptic QEP

The original definition of ellipticity is not convenient when dealing with distance problems. However, an important characterization of elliptic systems is obtained by considering the matrix polynomial $Q(\lambda)$.

Theorem 5.4. *The self-adjoint QEP* (5.1) *with A positive definite is elliptic if and only if $Q(\lambda)$ is positive definite for all $\lambda \in \mathbb{R}$*

Proof. This standard proof is based on the quadratic polynomial

$$x^*Q(\lambda)x = (x^*Ax)\lambda^2 + (x^*Bx)\lambda + (x^*Cx)$$

where x is any non-zero vector. Since A is positive definite, this polynomial is positive for all non-zero vector x if and only if $(x^*Bx)^2 - 4(x^*Ax)(x^*Cx)$ is strictly negative for all non-zero vector x. Thus, the matrix polynomial $Q(\lambda)$ is positive definite on the real line if and only if the QEP (5.1) is elliptic. □

Corollary 5.5. *The self-adjoint QEP* (5.1) *with A positive definite is elliptic if and only if $P(\omega)$ is positive definite for all frequencies $\omega \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.*

A Hermitian quadratic polynomial $Q(\lambda)$ (or the associated trigonometric matrix polynomial $P(\omega)$) is therefore said to be elliptic if the corresponding QEP (5.1) is elliptic.

Two distance problems related to elliptic QEP are:

- If $Q(\lambda)$ is elliptic, find

  $$\Delta Q(\lambda) = \lambda^2 \Delta A + \lambda \Delta B + \Delta C$$

  of smallest norm

  $$\left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|$$

  such that $Q(\lambda) + \Delta Q(\lambda)$ is not elliptic.

- If $Q(\lambda)$ is not elliptic, find

  $$\Delta Q(\lambda) = \lambda^2 \Delta A + \lambda \Delta B + \Delta C$$

  of smallest norm

  $$\left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|$$

  such that $Q(\lambda) + \Delta Q(\lambda)$ is elliptic.

The first distance problem is solved in both spectral and Frobenius norms by the following theorem.

Theorem 5.6. *Let $Q(\lambda)$ be elliptic. Any perturbation $\Delta Q(\lambda)$ such that $Q(\lambda)+\Delta Q(\lambda)$ is not elliptic satisfies the inequality*

$$r_E \leq \left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|_2 \leq \left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|_F,$$

where $r_E = \min_\omega \lambda_{\min} P(\omega) > 0$. Moreover, equality holds for the rank-one perturbations

$$\begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} = -r_E(\begin{bmatrix} \sin(\hat{\omega}) \\ \cos(\hat{\omega}) \end{bmatrix} \begin{bmatrix} \sin(\hat{\omega}) & \cos(\hat{\omega}) \end{bmatrix}) \otimes (\hat{x}\hat{x}^*),$$

with $\hat{\omega} = \arg\min_\omega \lambda_{\min} P(\omega)$ and $P(\hat{\omega})\hat{x} = r_E\hat{x}$ ($\|\hat{x}\|_2 = 1$).

Proof. In order to get compact expressions within this proof, we define the matrix function

$$f(\omega) = \begin{bmatrix} \sin(\omega) I \\ \cos(\omega) I \end{bmatrix}.$$

Any perturbation of $Q(\lambda)$ that makes it non-elliptic must also perturb the appropriate eigenvalues of $P(\omega)$ so that it is not a strictly positive polynomial any more. For a given frequency ω, standard perturbation theory can be applied to $P(\omega)$. Because of the inequality

$$\lambda_{\min}(P(\lambda) + \Delta P(\lambda)) \geq \lambda_{\min}(P(\lambda)) - \lambda_{\max}(\Delta P(\lambda)),$$

any perturbation such that $P(\omega)$ loses its definiteness satisfies the following inequality

$$\lambda_{\min}(f(\omega)^* \begin{bmatrix} A & B/2 \\ B/2 & C \end{bmatrix} f(\omega)) \leq \left\| f(\omega)^* \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} f(\omega) \right\|_2.$$

From norm consistency,

$$\left\| f(\omega)^* \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} f(\omega) \right\|_2 \leq \left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|_2 \|f(\omega)\|_2^2.$$

As $f(\omega)^* f(\omega) = I$, we have that $\|f(\omega)\|_2 = 1$. Therefore, a minimization with respect to ω yields

$$r_E \leq \left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|_2.$$

The inequality $\|\cdot\|_2 \leq \|\cdot\|_F$ is well known.

Let $\hat{\omega} = \arg\min_\omega \lambda_{\min} P(\omega)$ and $P(\hat{\omega})\hat{x} = r_E\hat{x}$ ($\|\hat{x}\|_2 = 1$). The perturbations defined by

$$\begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} = -r_E(\begin{bmatrix} \sin(\hat{\omega}) \\ \cos(\hat{\omega}) \end{bmatrix} \begin{bmatrix} \sin(\hat{\omega}) & \cos(\hat{\omega}) \end{bmatrix}) \otimes (\hat{x}\hat{x}^*)$$

satisfy $\left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|_{2,F} = r_E$ and produce a non-elliptic polynomial. Indeed, by construction, we have

$$\hat{x}^*(P(\hat{\omega}) + \Delta P(\hat{\omega}))\hat{x} = 0,$$

from which we conclude that the polynomial $P(\omega) + \Delta P(\omega)$ is not strictly positive definite. As the quadratic polynomial $Q(\lambda) + \Delta Q(\lambda)$ is non-elliptic, our perturbation is optimal. □

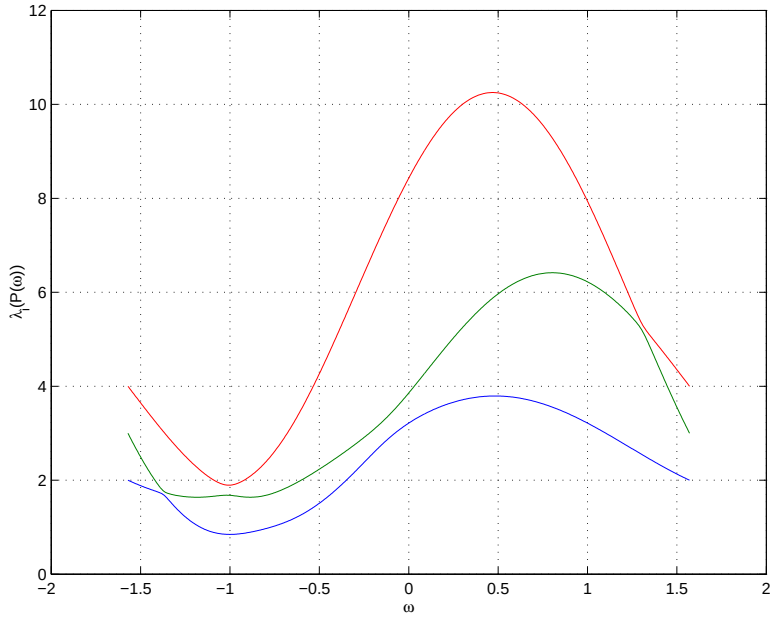

Figure 5.9. *Eigenvalues of $P(\omega)$ for Example 5.6*

Remark. The above theorem also treats the case where only the matrix A is perturbed. In this case, the non-elliptic QEP is obtained by modifying the matrix A so that it loses positive definiteness.

Example 5.6. The QEP defined by

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}, \quad B = \begin{bmatrix} 7/4 & 0 & 0 \\ 0 & 15/2 & 0 \\ 0 & 0 & 5 \end{bmatrix}, \quad C = \begin{bmatrix} 7/2 & 1 & 0 \\ 1 & 8 & 1 \\ 0 & 1 & 4 \end{bmatrix}$$

is elliptic, see Figure 5.9. Its minimal distance to a non-elliptic QEP is $r_E = 0.8460$ and the critical frequency is $\hat{\omega} = -1.0011$. The optimal perturbations are obtained via the eigenvector $\hat{x} = \begin{bmatrix} 0.3281 & -0.8972 & 0.2956 \end{bmatrix}^T$.

A straightforward consequence of Theorem 5.4 is that the set of elliptic QEP is an *open* convex set. Computing the distance between a non-elliptic QEP and this set, which is our second distance problem, is therefore a badly defined problem. However, the distance to the closure of this set and the associated boundary point can be easily obtained. This problem can be solved using our parametrization of non-negative matrix polynomials by positive semidefinite matrices, see Chapter 2.

Theorem 5.7. *The Hermitian quadratic polynomial $\lambda^2 A + \lambda B + C$ is non-negative on the real line if and only if there exists a matrix X such that*

$$\begin{bmatrix} A & B/2 - X \\ B/2 + X & C \end{bmatrix} \succeq 0$$

and $X = -X^*$.

Indeed, the closure of the set of elliptic QEP is exactly the set of Hermitian quadratic matrix polynomials non-negative on the real line, see Theorem 5.4. Depending on the measure, the following convex problems provide us with the asymptotically optimal perturbations:

- Spectral norm

$$\begin{aligned} \min \quad & \tau \\ \text{s.t.} \quad & \tau^2 I \succeq \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix}^* \\ & \begin{bmatrix} A + \Delta A & B/2 + \Delta B/2 - X \\ B/2 + \Delta B/2 + X & C + \Delta C \end{bmatrix} \succeq 0, \\ & \Delta A = \Delta A^*, \Delta B = \Delta B^*, \Delta C = \Delta C^*, X = -X^*. \end{aligned}$$

- Frobenius norm

$$\begin{aligned} \min \quad & \left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|_F^2 \\ \text{s.t.} \quad & \begin{bmatrix} A + \Delta A & B/2 + \Delta B/2 - X \\ B/2 + \Delta B/2 + X & C + \Delta C \end{bmatrix} \succeq 0, \\ & \Delta A = \Delta A^*, \Delta B = \Delta B^*, \Delta C = \Delta C^*, X = -X^*. \end{aligned}$$

Both optimization problems can be recast as semidefinite programming problems in a straightforward manner. Therefore, they are efficiently solvable in polynomial time (up to any given accuracy) using modern interior-point methods [14, 112, 144].

An arbitrarily close elliptic QEP is then obtained from their solutions using an appropriate shift. For instance, the polynomial $Q(\lambda) + \Delta Q(\lambda) + \epsilon I$ is elliptic, for all $\epsilon > 0$. Note that both optimization problems allow us to treat *structured* perturbations.

If we are only interested in the spectral norm, an optimal unstructured perturbation is easy to obtain.

Theorem 5.8. *Let $Q(\lambda)$ be non-elliptic. Any perturbation $\Delta Q(\lambda)$ such that $Q(\lambda) + \Delta Q(\lambda)$ is elliptic satisfies the strict inequality*

$$-r_E < \left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|_2,$$

where $r_E = \min_\omega \lambda_{\min} P(\omega) \leq 0$. For $\epsilon > 0$, an arbitrarily close perturbation $\Delta Q(\lambda)$ corresponds to $\Delta A = \Delta C = (-r_E + \epsilon)I$ and $\Delta B = 0$.

Proof. This proof is completely similar to the first part of the proof of Theorem 5.6 and is therefore omitted. □

Unfortunately, we were not able to obtain an explicit expression of the optimal perturbations for the Frobenius norm.

Example 5.7. As shown on Figure 5.10, the QEP defined by

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad B = \begin{bmatrix} 7/4 & 0 & 0 \\ 0 & 15/2 & 0 \\ 0 & 0 & 5 \end{bmatrix}, \quad C = \begin{bmatrix} 3/2 & 1 & 0 \\ 1 & 6 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

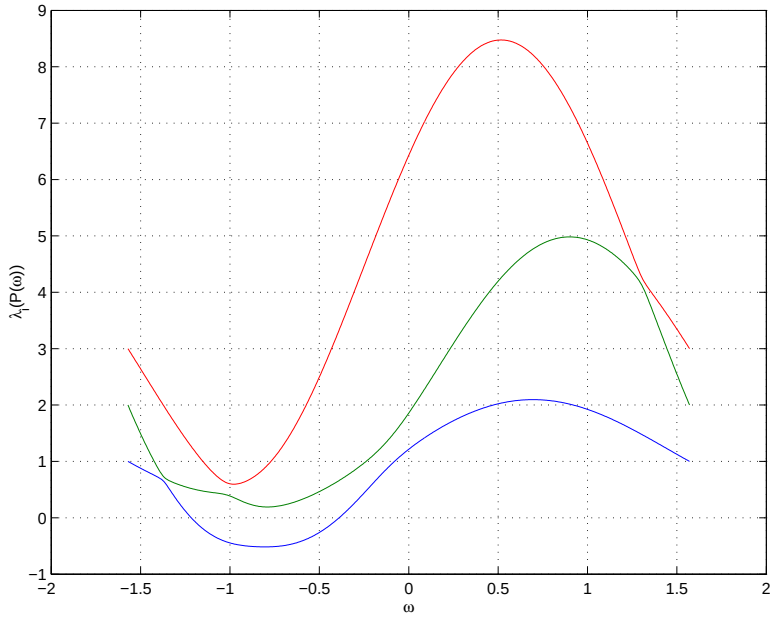

Figure 5.10. *Eigenvalues of $P(\omega)$ for Example 5.7*

is neither elliptic nor hyperbolic. Its minimal distance to a boundary point of the set of elliptic QEP is equal to $-r_E = 0.5163$. The semidefinite formulation for the spectral norm produces the optimal perturbations

$$\Delta A = \begin{bmatrix} 0.1639 & -0.0698 & 0.0455 \\ -0.0698 & 0.3047 & -0.0126 \\ 0.0455 & -0.0126 & 0.2644 \end{bmatrix}, \quad \Delta B = \begin{bmatrix} -0.2008 & 0.0190 & -0.0463 \\ 0.0190 & -0.3208 & 0.0077 \\ -0.0463 & 0.0077 & -0.4003 \end{bmatrix},$$

$$\Delta C = \begin{bmatrix} 0.1062 & -0.1081 & 0.0300 \\ -0.1081 & 0.2476 & -0.0379 \\ 0.0300 & -0.0379 & 0.2278 \end{bmatrix}$$

and $\left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|_2 = 0.5163 = -r_E$. If the Frobenius norm is used, we obtain

$$\Delta A = \begin{bmatrix} 0.0247 & -0.1497 & 0.0605 \\ -0.1497 & 0.2314 & -0.1454 \\ 0.0605 & -0.1454 & 0.1092 \end{bmatrix}, \quad \Delta B = \begin{bmatrix} -0.0194 & 0.1123 & -0.0658 \\ 0.1123 & -0.1680 & 0.1525 \\ -0.0658 & 0.1525 & -0.1298 \end{bmatrix},$$

$$\Delta C = \begin{bmatrix} 0.0163 & -0.0898 & 0.0745 \\ -0.0898 & 0.1283 & -0.1683 \\ 0.0745 & -0.1683 & 0.1548 \end{bmatrix}$$

and $\left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|_F = 0.5899.$

In Theorems 5.6 and 5.8, our solutions are based on the quantity $r_E = \min_\omega \lambda_{\min} P(\omega)$, which can be computed in polynomial time (up to any given accu-

racy). Indeed, this problem can be recast as a convex optimization problem, for which a global minimum can be easily obtained. Of course, other standard methods in linear algebra can also be adapted. In particular, the bisection and level set methods described in Algorithms 5.1 and 5.2 can be modified in a straightforward way to obtain a global minimum.

5.3.2 Distance problems related to hyperbolic QEP

There also exists a characterization of hyperbolic systems in terms of the matrix polynomial $Q(\lambda)$.

Theorem 5.9. *The self-adjoint QEP* (5.1) *with A positive definite is hyperbolic if and only if $Q(\lambda)$ is negative definite for some $\lambda \in \mathbb{R}$.*

Proof. See the proof of Theorem 1 in [11]. □

Corollary 5.10. *The self-adjoint QEP* (5.1) *with A positive definite is hyperbolic if and only if $P(\omega)$ is negative definite for some $\omega \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.*

As before, two distance problems are related to hyperbolic QEP:

- If $Q(\lambda)$ is hyperbolic, find

$$\Delta Q(\lambda) = \lambda^2 \Delta A + \lambda \Delta B + \Delta C$$

of smallest norm

$$\left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|$$

such that $Q(\lambda) + \Delta Q(\lambda)$ is not hyperbolic.

- If $Q(\lambda)$ is not hyperbolic, find

$$\Delta Q(\lambda) = \lambda^2 \Delta A + \lambda \Delta B + \Delta C$$

of smallest norm

$$\left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|$$

such that $Q(\lambda) + \Delta Q(\lambda)$ is hyperbolic.

Let us focus on the first distance problem. First, note that hyperbolicity of the QEP (5.1) is lost by adding to A a perturbation of spectral norm equal to $\lambda_{\min}(A)$, which makes this matrix lose its definiteness. Hereafter, these perturbations $\Delta Q(\lambda)$ with $\Delta B \equiv 0$ and , $\Delta C \equiv 0$ are said to be trivial. Of course, there also exist non-trivial perturbations.

Theorem 5.11. *Let $Q(\lambda)$ be hyperbolic. Any non-trivial perturbation $\Delta Q(\lambda)$ such that $Q(\lambda) + \Delta Q(\lambda)$ is not hyperbolic satisfies the inequality*

$$-r_H \leq \left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|_2 \leq \left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|_F,$$

where $r_H = \min_\omega \lambda_{\max} P(\omega) < 0$. *Moreover, equality holds for the rank-one perturbations*

$$\begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} = -r_H \left(\begin{bmatrix} \sin(\hat{\omega}) \\ \cos(\hat{\omega}) \end{bmatrix} \begin{bmatrix} \sin(\hat{\omega}) & \cos(\hat{\omega}) \end{bmatrix} \right) \otimes (\hat{x}\hat{x}^*),$$

with $\hat{\omega} = \arg\min_\omega \lambda_{\max} P(\omega)$ *and* $P(\hat{\omega})\hat{x} = r_H \hat{x}$ *(*$\|\hat{x}\|_2 = 1$*).*

Proof. As $Q(\lambda)$ is hyperbolic, the matrix $P(\omega)$ is negative definite for at least one frequency ω. In order to get at least one non-negative eigenvalue at all frequencies, the eigenvalues of $P(\omega)$ must be shifted by a quantity greater than $-r_H = -\min_\omega \lambda_{\max} P(\omega)$. This is a necessary condition for $P(\omega)$ to have a non-negative eigenvalue at the frequency $\hat{\omega} = \arg\min_\omega \lambda_{\max} P(\omega)$. Therefore, the lower bound on the norm of the perturbation is

$$-r_H \leq \left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|_2 .$$

The inequality $\|\cdot\|_2 \leq \|\cdot\|_F$ is well known.

The perturbations defined by

$$\begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} = -r_H \left(\begin{bmatrix} \sin(\hat{\omega}) \\ \cos(\hat{\omega}) \end{bmatrix} \begin{bmatrix} \sin(\hat{\omega}) & \cos(\hat{\omega}) \end{bmatrix} \right) \otimes (\hat{x}\hat{x}^*)$$

satisfy $\left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|_2 = -r_H$ and yield a non-hyperbolic polynomial. Indeed, the perturbation $\Delta Q(\lambda)$ is non-trivial so that the leading coefficient $A + \Delta A$ of $Q(\lambda) + \Delta Q(\lambda)$ is still positive definite. Since we have

$$\hat{x}^*(P(\omega) + \Delta P(\omega))\hat{x} = \begin{bmatrix} \sin(\hat{\omega}) & \cos(\hat{\omega}) \end{bmatrix} \begin{bmatrix} \hat{x}^*(A + \Delta A)\hat{x} & \hat{x}^*(B + \Delta B)\hat{x}/2 \\ \hat{x}^*(B + \Delta B)\hat{x}/2 & \hat{x}^*(C + \Delta C)\hat{x} \end{bmatrix} \begin{bmatrix} \sin(\hat{\omega}) \\ \cos(\hat{\omega}) \end{bmatrix} = 0,$$

the Schur complement of the two-by-two matrix

$$\begin{bmatrix} \hat{x}^*(A + \Delta A)\hat{x} & \hat{x}^*(B + \Delta B)\hat{x}/2 \\ \hat{x}^*(B + \Delta B)\hat{x}/2 & \hat{x}^*(C + \Delta C)\hat{x} \end{bmatrix}$$

with respect to its $(1,1)$-entry is necessarily equal to 0. Up to the positive factor $\hat{x}^*(A + \Delta A)\hat{x}$, this is equivalent to

$$(\hat{x}^*(B + \Delta B)\hat{x})^2 - 4(\hat{x}^*(A + \Delta A)\hat{x})(\hat{x}^*(C + \Delta C)\hat{x}) = 0.$$

Since $\hat{x}$ is a non-zero vector, $Q(\lambda) + \Delta Q(\lambda)$ is by definition non-hyperbolic. □

Remark. In general, the above theorem does not consider any trivial perturbations. It is therefore of paramount importance to compare the distance $-r_H$ with $\lambda_{\min}(A)$ in order to select the optimal perturbation. The complete procedure is illustrated in Examples 5.8 and 5.9.

Example 5.8. The QEP defined by

$$A = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 3/2 & 0 \\ 0 & 0 & 5/2 \end{bmatrix}, \quad B = \begin{bmatrix} 7/4 & 0 & 0 \\ 0 & 15/2 & 0 \\ 0 & 0 & 5 \end{bmatrix}, \quad C = \begin{bmatrix} -1/2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

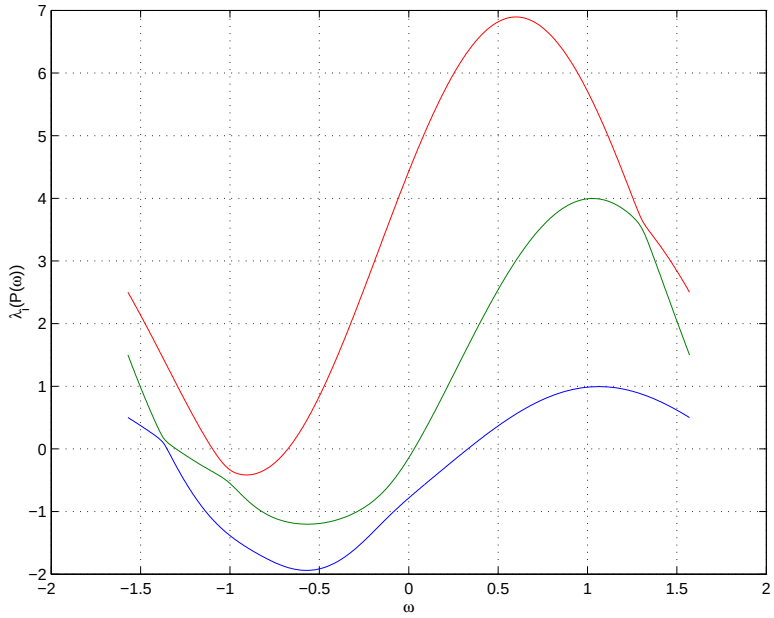

Figure 5.11. *Eigenvalues of $P(\omega)$ for Example 5.8*

is hyperbolic, see Figure 5.11. Its minimal distance to a non-hyperbolic QEP is $-r_H = 0.4161$ and the critical frequency is $\hat{\omega} = -0.9080$. The optimal perturbations are obtained via the associated eigenvector $\hat{x} = \begin{bmatrix}0.6831 & 0.5617 & 0.4667\end{bmatrix}^T$.

Example 5.9. The QEP defined by

$$A = \begin{bmatrix}1/2 & 0 & 0\\ 0 & 3/2 & 0\\ 0 & 0 & 5/2\end{bmatrix}, \quad B = \begin{bmatrix}7/4 & 0 & 0\\ 0 & 15/2 & 0\\ 0 & 0 & 5\end{bmatrix}, \quad C = \begin{bmatrix}-3/2 & 1 & 0\\ 1 & 3 & 1\\ 0 & 1 & -1\end{bmatrix}$$

is hyperbolic, see Figure 5.12. As the distance $-r_H = 0.8263$ is greater than the minimal eigenvalue of A, the optimal perturbation is the trivial one. We get that $\Delta A = -0.5 e_0 e_0^T$, $\Delta B = 0$ and $\Delta C = 0$ where e_0 is the first canonical vector.

Let us now consider the second distance problem. Since the set of hyperbolic QEP is not closed, we can only expect to compute a boundary point. As before, an arbitrarily close hyperbolic QEP can then be obtained using an appropriate shift.

Theorem 5.12. *Let $Q(\lambda)$ be non-hyperbolic with positive definite A. Any perturbation $\Delta Q(\lambda)$ such that $Q(\lambda) + \Delta Q(\lambda)$ is hyperbolic satisfies the strict inequality*

$$r_H < \left\| \begin{bmatrix}\Delta A & \Delta B/2\\ \Delta B/2 & \Delta C\end{bmatrix} \right\|_2,$$

where $r_H = \min_\omega \lambda_{\max} P(\omega) \geq 0$. For any $\epsilon > 0$, an arbitrarily close perturbation is

$$\begin{bmatrix}\Delta A & \Delta B/2\\ \Delta B/2 & \Delta C\end{bmatrix} = -(r_H + \epsilon)\left(\begin{bmatrix}\sin(\hat{\omega})\\ \cos(\hat{\omega})\end{bmatrix}\begin{bmatrix}\sin(\hat{\omega}) & \cos(\hat{\omega})\end{bmatrix}\right) \otimes I,$$

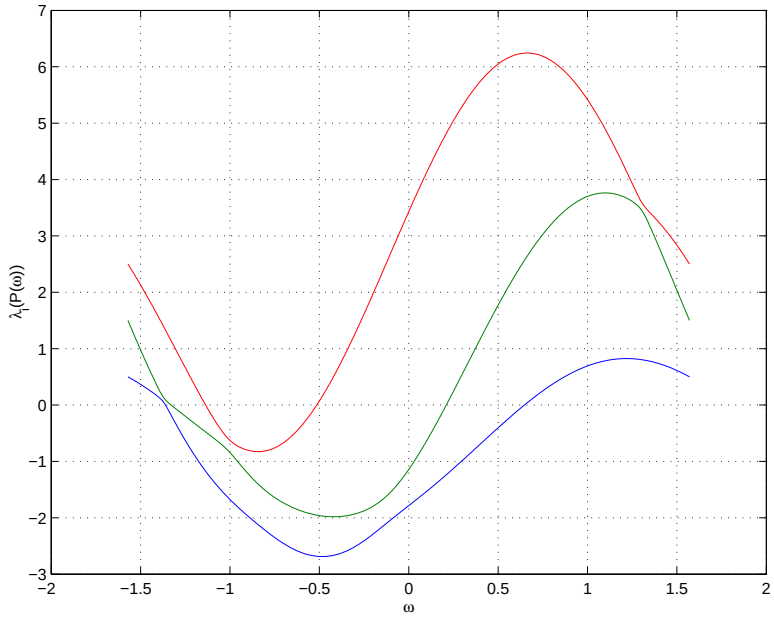

Figure 5.12. *Eigenvalues of $P(\omega)$ for Example 5.9*

with $\hat{\omega} = \arg\min_\omega \lambda_{\max} P(\omega)$.

Proof. As $Q(\lambda)$ is non-hyperbolic, the matrix $P(\omega)$ has at least one non-negative eigenvalue at all frequencies ω. In order to make $P(\omega)$ hyperbolic, we need to shift the eigenvalues of $P(\omega)$ by a quantity strictly greater than $r_H = \min_\omega \lambda_{\max} P(\omega)$. At the corresponding frequency $\hat{\omega} = \arg\min_\omega \lambda_{\max} P(\omega)$, $P(\omega)$ could then become negative definite. Therefore, the strict lower bound on the norm of the perturbation is

$$r_H < \left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|_2 .$$

For any $\epsilon > 0$, the perturbations defined by

$$\begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} = -r_H(1+\epsilon)\left(\begin{bmatrix} \sin(\hat{\omega}) \\ \cos(\hat{\omega}) \end{bmatrix} \begin{bmatrix} \sin(\hat{\omega}) & \cos(\hat{\omega}) \end{bmatrix}\right) \otimes I$$

satisfy $\left\| \begin{bmatrix} \Delta A & \Delta B/2 \\ \Delta B/2 & \Delta C \end{bmatrix} \right\|_2 = r_H(1+\epsilon)$ and yield an hyperbolic polynomial.

Indeed, at the frequency $\omega = \hat{\omega}$ and for all unit vectors x, $x^*(P(\omega) + \Delta P(\omega))x \leq -r_H\epsilon$. At the frequency $\omega = \hat{\omega}$, $P(\omega) + \Delta P(\omega)$ is thus negative definite and the polynomial $Q(\lambda) + \Delta Q(\lambda)$ is hyperbolic. □

Example 5.10. The QEP defined in Example 5.7 is neither elliptic nor hyperbolic. Its minimal distance to a boundary point of the set of hyperbolic QEP is equal to $r_H = 0.5957$. The critical frequency is $\hat{\omega} = -0.9785$ and the associated perturbations are

$$\Delta A = -0.4101(1+\epsilon)I_3, \quad \Delta B = 0.5519(1+\epsilon)I_3, \quad \Delta C = -0.1857(1+\epsilon)I_3,$$

with $\epsilon > 0$.

In Theorems 5.11 and 5.12, our solutions are based on the quantity $r_H = \min_\omega \lambda_{\max} P(\omega)$. Although this problem cannot be recast as a convex optimization problem, other standard methods in linear algebra can be applied to obtained the global minimum. In particular, the bisection and level set methods described in Algorithms 5.1 and 5.2 can be modified in a straightforward way.

5.4 Research summary

Several important problems have been solved in this chapter by making use of original techniques.

First, we have described two new cutting plane algorithms for computing the Crawford number. Their efficiency follows from a (partial) use of the problem structure to get the next query point, see Algorithms 5.5 and 5.6. The subproblem to be solved at each iteration is easier to solve than the subproblem in standard methods. Their performance have been illustrated by three different test instances; their computational advantages over standard algorithms have also been discussed.

Then, we have focused on distance problems related to quadratic polynomials. These problems have been solved so as to highlight the structure of the optimal perturbations. The exact distance can be obtained using either convex optimization (elliptic systems) or standard linear algebra algorithms (elliptic and hyperbolic systems).

Notes and references

Section 5.1 is partially based on the remarkable survey paper [134]. The possible use of convex optimization for computing the Crawford number is mentioned in [76], but is not investigated throughout. Section 5.2 fills in this gap. Distance problems related to QEP were also considered in this paper. Our approach is definitely superior as optimal perturbations are explicitly obtained in Section 5.3, which is based on the forthcoming paper [70].

Chapter 6
Applications

In this chapter, we review several selected applications of non-negative polynomials, which belong to four different research areas:

- Systems and control;
- Signal processing;
- Probability theory and statistics;
- Coding theory.

Our aim is to illustrate the impact of our results in these disciplines. For this reason, we restrict ourselves to well-known academic examples.

In Section 6.1, we focus on two methods for computing a spectral factor of a given non-negative trigonometric polynomial, which is a well-studied problem in the systems theory literature. One of the most important problems in signal processing is to design digital filters. In Section 6.2, we show that non-negative polynomials allow us to formulate such problems. Practical problems involving moments are very often encountered in probability theory and statistics, which is not surprising at all. We solve several related estimation problems in Section 6.3. Finally, we concentrate on a very particular problem of coding theory, the kissing number problem, in Section 6.4. Although this last problem is difficult to solve exactly, upper bounds can be obtained from an optimization problem with a polynomial non-negativity constraint.

6.1 Spectral factorization

A standard problem of systems theory is that of spectral factorization. In this section, we focus on non-negative trigonometric polynomials, for which the Fejér-Riesz Theorem yields the existence of a spectral decomposition.

Theorem 6.1 (Fejér-Riesz). *A trigonometric polynomial $p(z) = \langle p, \pi_n(z)\rangle_{\mathbb{R}}$ (of degree n) is non-negative on the unit circle , i.e., $p(z) \geq 0, \forall z \in \mathbb{T}$, if and only if there exists a complex polynomial $q(z) = \sum_{k=0}^{n} q_k z^k$ such that $p(z) = |q(z)|^2, \forall z \in \mathbb{T}$. Moreover, $p \in \mathbb{R}^{n+1}$ if and only if $q \in \mathbb{R}^{n+1}$.*

The problem of computing $q(z)$ from $p(z)$ is known as *spectral factorization* in the literature. The aim of this section is to illustrate the relevance of our semidefinite parametrization to the spectral factorization problem. In particular, it gives us some insight on this problem in a convex optimization framework.

Given a non-negative trigonometric polynomial $p(z)$, there exist (up to) 2^n distinct factors $q(z)$ such that $p(z) = |q(z)|^2,\ \forall z \in \mathbb{T}$. To see this, recall that the roots of $p(z)$ are necessarily paired and that any spectral factor of $p(z)$ has only one root of a given pair. Moreover, spectral factors are unique up to a phase $e^{\jmath\theta}$. For these reasons, spectral factorization methods are concerned with the computation of a *canonical* factor. The next lemma summarizes our observations.

Lemma 6.2 ([129]). *Consider a trigonometric polynomial $p(z) = \langle p, \pi_n(z)\rangle_{\mathbb{R}}$ (of degree n) non-negative on the unit circle. Then the following facts hold:*

1. *If a_k is a zero of $p(z)$, then $a_k \neq 0$ and $\overline{a_k}^{-1}$ is also a zero. It follows that if $p(z)$ has n zeros $\{a_k\}_{k=0}^{n-1}$ on and inside the unit circle, then it also has n additional zeros $\{b_k = \overline{a_k}^{-1}\}_{k=0}^{n-1}$ on and outside the unit circle.*
2. *The canonical factorization of $p(z)$ is given by $p(z) = q(z)^* q(z)$ where $q(z) = \sum_{k=0}^{n} q_k z^k$ has all its zeros outside the open unit disk.*
3. *If $p(z)$ is strictly positive on the unit circle, then $q(z)$ has all its zeros strictly outside the closed unit disk.*

Although the factorization is easy to perform for low-order polynomials by computing their roots, more systematic procedures are needed for higher-order polynomials.

6.1.1 Spectral factorization via semidefinite programming

Using our characterization of non-negative trigonometric polynomials and an appropriate conic optimization problem, we can compute the stable spectral factor. In order to prove this result, we need a few preliminary lemmata.

Lemma 6.3. *If $z_0 \in \mathbb{C}$ belongs to the open unit disk, i.e., $|z_0| < 1$, then any Hermitian Toeplitz matrix $T(s)$ of order $n+1$ satisfying $T(s) \succeq \pi_n(z_0)\pi_n(z_0)^*$ is positive definite.*

Proof. Let z_0 be an element of the open unit disk in the complex plane. Any positive semidefinite Toeplitz matrix $T(s)$ can be factorized as

$$T(s) = VDV^*,$$

where $D = \operatorname{diag}(\{d_k\}_{k=0}^n)$ is a positive semidefinite diagonal matrix and V is a non-singular Vandermonde matrix, the nodes $\{z_k\}_{k=0}^n$ of which belong to the complex unit circle [83].

Let L be the matrix of Lagrange polynomials associated with V, i.e., $V^*L = I_{n+1}$. Then the following equivalence holds

$$T(s) \succeq \pi_n(z_0)\pi_n(z_0)^* \iff D \succeq L^*\pi_n(z_0)\pi_n(z_0)^* L.$$

Therefore, we get that, for all $k \in \{0, \ldots, n\}$,

$$d_k \geq (e_k^* L^* \pi_n(z_0))(\pi_n(z_0)^* L e_k) = \left| \frac{\prod_{l\neq k}(z_0 - \overline{z_l})}{\prod_{l\neq k}(\overline{z_k} - \overline{z_l})} \right|^2 \geq \frac{|\prod_{l\neq k}(z_0 - \overline{z_l})|^2}{4^n} > 0. \quad (6.1)$$

This implies that $T(s)$ is necessarily strictly positive definite. Moreover, it cannot be too close to the boundary of the set of positive definite matrices. □

Remark. If $z_0 = 0$, then (6.1) can be rewritten as

$$d_k \geq (e_k^* L^* e_0)(e_0^* L e_k) = \frac{1}{\prod_{l \neq k} |z_k - z_l|^2} \geq \frac{1}{4^n} > 0.$$

Remark. The above lemma cannot be improved. Indeed, let $z_0 = 1$. Then the inequality

$$T(e) \equiv \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix} \succeq \pi_n(1)\pi_n(1)^*,$$

holds, but the left-hand side matrix is singular. However, we could have chosen $z_0 \in \mathbb{C}$ outside the closed unit disk, i.e., $|z_0| > 1$. In this case, the proof is exactly similar.

Lemma 6.4. *Let $q(z) = \sum_{i=0}^n q_i z^i$ be a spectral factor of the non-negative trigonometric polynomial $p(z) = \langle p, \pi_n(z) \rangle_{\mathbb{R}}$. Then*

$$q(z) = e^{\jmath\theta} \frac{\sqrt{|p_n|} \prod_{k=1}^n (z - z_k)}{\sqrt{\prod_{k=1}^n |z_k|}}$$

where $\{z_k\}_{k=1}^n$ are the roots of $q(z)$ and θ is an arbitrary phase.

Proof. The polynomial $q(z)$ can be written as $q(z) = q_n \prod_{k=1}^n (z - z_k)$, so that the dependence on the roots $\{z_k\}_{k=1}^n$ is now explicit.

As the identity $p(z) = |q(z)|^2$ holds for any $z \in \mathbb{T}$, we get that $p = T^*(\overline{q} \cdot \overline{q}^*)$ by making use our semidefinite parametrization. This identity leads to the relation

$$p_n = 2\overline{q_0} q_n = 2\overline{q_n} q_n \prod_{k=1}^n (-\overline{z_k}).$$

Consequently, the modulus of q_n is equal to

$$|q_n| = \sqrt{\left| \frac{p_n}{\prod_{k=1}^n z_k} \right|}.$$

The result then follows from the fact that q_n is determined up to an arbitrary phase. □

Lemma 6.5. *Let $z, w \in \mathbb{C}$. If $|z| < 1$ and $0 < |w| \leq 1$, then*

$$\frac{|z - w|^2}{|w|} \leq \frac{|z - \frac{1}{\overline{w}}|^2}{|\frac{1}{\overline{w}}|}.$$

Moreover, strict inequality holds if $|w| < 1$.

Proof. Without loss of generality, we can assume that

$$z = r_1, \quad 0 \leq r_1 < 1, \qquad w = r_2[\cos\theta + \jmath \sin\theta], \quad 0 < r_2 \leq 1.$$

Since r_1 and r_2 are such that $1 - r_1^2 > 0$, $1 - r_2^2 \geq 0$ and $r_2 \neq 0$, we successively get that

$$\begin{aligned} 0 &\leq (1 - r_1^2)(1 - r_2^2) = r_1^2 r_2^2 + 1 - r_1^2 - r_2^2 \\ &= [(r_1 r_2 - \cos\theta)^2 + \sin\theta^2] - [(r_1 - r_2\cos\theta)^2 + (r_2 \sin\theta)^2] \end{aligned}$$

Therefore,

$$\frac{(r_1 - r_2\cos\theta)^2 + (r_2\sin\theta)^2}{r_2} \leq \frac{(r_1 - \frac{1}{r_2}\cos\theta)^2 + (\frac{1}{r_2}\sin\theta)^2}{\frac{1}{r_2}}$$

This exactly proves the inequality

$$\frac{|z - w|^2}{|w|} \leq \frac{|z - \frac{1}{\overline{w}}|^2}{|\frac{1}{\overline{w}}|}.$$

If $|w| < 1$, i.e., $r_2 < 1$, then $0 < (1 - r_1^2)(1 - r_2^2)$ and the above inequality is strict. □

We can now state our main result.

Theorem 6.6. *Let $\tilde{p} \in \operatorname{int} \mathcal{K}_{\mathbb{C}}$ and $z_0 \in \mathbb{C}$ such that $|z_0| < 1$. The primal problem*

$$\begin{array}{ll} \max & \langle \pi_n(z_0)\pi_n(z_0)^*, Y \rangle \\ \text{s. t.} & \tilde{p} = T^*(Y) \\ & Y = Y^* \succeq 0 \end{array} \tag{6.2}$$

has an optimal rank-one solution $\hat{Y}$, which yields the stable spectral factor of $\tilde{p}$. If $\hat{Y} = qq^*$, then $q(z) = q^*\pi_n(z)$ has all its roots outside the closed unit disk. The optimal solution $\hat{s}$ of the dual problem

$$\begin{array}{ll} \min & \langle s, \tilde{p} \rangle \\ \text{s. t.} & T(s) \succeq \pi_n(z_0)\pi_n(z_0)^* \end{array}$$

is such that $T(\hat{s})$ is positive definite. Moreover, the solutions of both problems are strictly complementary.

Proof. The dual problem is readily obtained from standard duality theory.

By Lemma 6.3, $T(\hat{s})$ is necessarily positive definite. Since both problems are strictly feasible, there is no duality gap and the complementary slackness condition holds [144], that is

$$\hat{Y}(T(\hat{s}) - \pi_n(z_0)\pi_n(z_0)^*) = 0. \tag{6.3}$$

As the matrix

$$T(\hat{s}) = (T(\hat{s}) - \pi_n(z_0)\pi_n(z_0)^*) + \pi_n(z_0)\pi_n(z_0)^*$$

is of rank $n + 1$, the rank of $(T(\hat{s}) - \pi_n(z_0)\pi_n(z_0)^*)$ is at least equal to n. Since $\hat{Y}$ is of rank at least one, $\hat{Y} = qq^*$ by complementary slackness. Moreover, q is necessarily proportional to $T(\hat{s})^{-1}\pi_n(z_0)$. To see this, note that (6.3) is equivalent to

$$q^*(T(\hat{s}) - \pi_n(z_0)\pi_n(z_0)^*)q = 0$$

and that $q = \alpha T(\hat{s})^{-1}\pi_n(z_0)$ satisfies this equation for some non-zero $\alpha \in \mathbb{C}$.

Since the rank of the optimal solution of (6.2) is equal to 1, the original problem can be rewritten as

$$\begin{array}{ll} \max & \langle \pi_n(z_0)\pi_n(z_0)^*, qq^* \rangle \\ \text{s.t.} & \tilde{p} = T^*(qq^*). \end{array}$$

Therefore, the optimal vector q corresponds to a polynomial $q(z) = q^*\pi_n(z)$ such that

- $q(z)$ is a spectral factor of $\tilde{p}(z)$, i.e., $\tilde{p} = T^*(qq^*)$;
- the value of $q(z_0)$ has a maximal modulus.

By Lemmata 6.4 and 6.5, this clearly corresponds to the spectral factor having all its roots outside the closed unit disk. Note that, if $z_0 = 0$, then the second condition states that the product of the roots of $q(z)$ is of maximal modulus. □

Proposition 6.7. *For any vector $\tilde{s} \in \mathbb{R} \times \mathbb{C}^n$ such that $T(\tilde{s}) \succ 0$ and any point $z_0 \in \mathbb{D}$, the polynomial $q(z) = \pi_n(z_0)^* T(\tilde{s})^{-1} \pi_n(z)$ has all its roots outside the closed unit disk.*

Proof. This polynomial can be obtained as the solution of an optimization problem similar to (6.2). Let $\tilde{Y} = T(\tilde{s})^{-1}\pi_n(z_0)\pi_n(z_0)^* T(\tilde{s})^{-1}$, $\tilde{p} = T^*(\tilde{Y})$ and $C = \pi_n(z_0)[\pi_n(z_0)^* T(\tilde{s})^{-1}\pi_n(z_0)]^{-1}\pi_n(z_0)^*$.

The dual of the optimization problem

$$\begin{array}{ll} \max & \langle C, Y \rangle \\ \text{s.t.} & \tilde{p} = T^*(Y), \\ & Y = Y^* \succeq 0 \end{array}$$

is exactly

$$\begin{array}{ll} \min & \langle s, \tilde{p} \rangle \\ \text{s.t.} & T(s) \succeq C. \end{array}$$

The equality $\langle C, \tilde{Y} \rangle = \langle \tilde{s}, \tilde{p} \rangle = \langle T(\tilde{s})^{-1}\pi_n(z_0), \pi_n(z_0)^* \rangle$ proves that the points $\tilde{Y}$ and $\tilde{s}$ are optimal for the corresponding problems. As $[\pi_n(z_0)^* T(\tilde{s})^{-1}\pi_n(z_0)]^{-1}$ is a scalar, applying Theorem 6.6 to $\tilde{p}$ ends the proof. □

Corollary 6.8. *For any $s \in \mathbb{R} \times \mathbb{C}^n$ such that $T(s) \succ 0$, the function*

$$\phi(z_0, z_1) : \mathbb{C} \times \mathbb{C} \to \mathbb{C} : (z_0, z_1) \to \phi(z_0, z_1) = \pi_n(z_0)^* T(s)^{-1} \pi_n(z_1)$$

has no roots in the region $\mathbb{D}^2 = \{(z_1, z_2) \in \mathbb{C}^2 : |z_1| < 1, |z_2| < 1\}$.

Proof. This result is a consequence of Proposition 6.7. Indeed, note that $q(z) = \phi(z_0, z) = \pi_n(z_0)^* T(s)^{-1} \pi_n(z)$ has no zero inside the unit disk. The result then follow by moving z_0 inside the unit disk. □

Theorem 6.6 can be extended to non-negative matrix trigonometric polynomials. The proof is technically more difficult because it relies on the theory of orthogonal polynomials, see e.g. [42, 43].

Let us highlight several important concepts that are required to prove the matrix counterpart of Lemma 6.3. It is well known that any positive definite block Toeplitz

matrix $T(S)$ defines a set of (left) *orthogonal matrix polynomials* with respect to the (left) matrix inner product

$$\langle\langle P, Q\rangle\rangle = \begin{bmatrix} P_0 & \cdots & P_n \end{bmatrix} T(S) \begin{bmatrix} Q_0^* \\ \vdots \\ Q_n^* \end{bmatrix}.$$

Given $S \in \operatorname{int} \mathcal{K}_{\mathbb{C}}^{\star}$, the block Levinson algorithm applied to the matrix $JT(S)J^*$ yields the set of (left) orthogonal matrix polynomials associated with the measure with moments

$$S_0^*, S_1^*, \ldots, S_n^*.$$

Remember that J is the block permutation matrix, see (3.14). From these orthogonal polynomials, say

$$P_k(z) = \sum_{l=0}^{k} P_{k,l} z^l, \quad k = 0, \ldots, n,$$

we obtain a factorization of $T(S)$,

$$\left(\begin{bmatrix} P_{0,0} & & & \\ P_{1,0} & P_{1,1} & & \\ \vdots & & \ddots & \\ P_{n,0} & P_{n,1} & \cdots & P_{n,n} \end{bmatrix} J \right) T(S) \left(J^* \begin{bmatrix} P_{0,0}^* & P_{1,0}^* & \cdots & P_{n,0}^* \\ & P_{1,1}^* & \cdots & P_{n,1}^* \\ & & \ddots & \vdots \\ & & & P_{n,n}^* \end{bmatrix} \right) = I_{m(n+1)}.$$

Clearly, the leading coefficient of P_k is nonsingular, i.e., $\det P_{k,k} \neq 0, \forall k$. The *reversed predictor polynomials* defined by

$$A_k(z) = P_{k,k}^{-1} P_k(z^{-1}) z^n = \sum_{l=0}^{k-1} P_{k,k}^{-1} P_{k,l} z^{n-l} + z^{n-k} I_m \doteq \sum_{l=0}^{k} A_{k,l} z^{n-l}$$

produces a factorization of $T(S)^{-1}$, which proves to be convenient in the sequel,

$$T(S)^{-1} = \begin{bmatrix} & & & A_{n,n}^* \\ & & \iddots & \vdots \\ & A_{1,1}^* & \cdots & A_{n,1}^* \\ A_{0,0}^* & A_{1,0}^* & \cdots & A_{n,0}^* \end{bmatrix} \begin{bmatrix} D_0^{-1} & & & \\ & D_1^{-1} & & \\ & & \ddots & \\ & & & D_n^{-1} \end{bmatrix} \begin{bmatrix} & & & A_{0,0} \\ & & A_{1,1} & A_{1,0} \\ & \iddots & & \vdots \\ A_{n,n} & \cdots & A_{n,1} & A_{n,0} \end{bmatrix}. \tag{6.4}$$

In this equation, the elements of the block diagonal matrix, which are equal to

$$D_k = P_{k,k}^{-1} P_{k,k}^{-*},$$

can alternatively be obtained from the *reflection coefficients* $\{R_k\}_{k=0}^n$. These coefficients are computed by the block Levinson algorithm in a recursive way and they summarize all information on positive definiteness of $T(S)$ [42, 43]. Indeed, it is well known that $\|R_k\|_2 < 1, \forall k$ if and only if $T(S)$ is positive definite. In particular, the last diagonal block D_n is of paramount importance because of the recurrence relations

$$\det D_0 = \det S_0, \quad \det D_k = \det(I_m - R_k R_k^*) \det D_{k-1}, \ k = 1, \ldots, n.$$

For any $S \in \operatorname{int} \mathcal{K}_{\mathbb{C}}$, the diagonal block D_n associated with $T(S)$ has clearly a positive determinant. Moreover, its determinant tends to zero as S approaches the boundary.

These theoretical results allow us to generalize Lemma 6.3.

Lemma 6.9. *Let R be an $m \times m$ positive definite matrix. If $z_0 \in \mathbb{C}$ belongs to the open unit disk, i.e., $|z_0| < 1$, then any Hermitian block Toeplitz matrix $T(S)$ of order $m(n+1)$ satisfying*

$$T(S) \succeq \Pi_n(z_0) R^{-1} \Pi_n(z_0)^* \tag{6.5}$$

is necessarily positive definite.

Proof. First we consider the case $z_0 = 0$. By contradiction, assume that a singular matrix $T(S)$ satisfies (6.5). Then there exists a sequence $\{S_k\}_{k=1}^{\infty}$ of positive definite matrices such that

$$\lim_{k \to +\infty} S_k = S \quad \text{and} \quad T(S_k) \succeq \Pi_n(z_0) R^{-1} \Pi_n(z_0)^*, \forall k.$$

For instance, one can use the sequence $T(S_k) = T(S) + \frac{1}{k} I_{m(n+1)}$.

On the one hand, the value

$$F(S_k) = \log \det(\Pi_n(z_0)^* T(S_k)^{-1} \Pi_n(z_0))$$

is unbounded as k goes to infinity. Indeed, from a factorization of $T(S_k)^{-1}$ of the type (6.4), we conclude that $F(S_k) = -\log \det D_{k,n}$.

On the other hand, any $T(S_k)$ is positive definite. A Schur complement approach thus reveals the following inequalities

$$R \succeq \Pi_n(z_0)^* T(S_k)^{-1} \Pi_n(z_0), \quad \forall k.$$

Therefore, $\log \det(\Pi_n(z_0)^* T(S_k)^{-1} \Pi_n(z_0))$ is bounded from above by a constant depending on R for all matrices S_k in the sequence $\{S_k\}_{k=1}^{\infty}$. This last fact contradicts our previous observation so $S \in \operatorname{int} \mathcal{K}_{\mathbb{C}}$.

Finally, the case $z_0 \in \mathbb{D}$ is obtained by considering the transformation

$$z = \frac{\hat{z} + z_0}{1 + \overline{z_0} \hat{z}} \quad \Longleftrightarrow \quad \hat{z} = \frac{z - z_0}{1 - \overline{z_0} z},$$

which maps the open unit disk into itself and which preserves the unit circle. Clearly, $z = z_0$ is mapped on $\hat{z} = 0$. Moreover, the block Toeplitz structure is not altered. Indeed, this follows from the Vandermonde decomposition of $T(S)$, see Theorem 2.22. □

Proposition 6.10. *Let $z_0 \in \mathbb{C}$ be a point of the open unit disk and $S \in \operatorname{int} K_{\mathbb{C}}^*$. Then the matrix polynomial $\Pi_n(z_0)^* T(S)^{-1} \Pi_n(z)$ has all its zeros outside the closed unit disk. Equivalently, the matrix pseudo-polynomial $\Pi_n(z^{-1})^T T(S)^{-1} \Pi_n(z_0)$ has all its zeros in the open unit disk.*

Proof. By [42, Theorem 6], the matrix polynomial

$$\Pi_n(z)^T T(S)^{-1} \Pi_n(0)$$

is devoid of zeros in the closed unit disc. Consequently, the matrix pseudo-polynomial

$$\Pi_n(z^{-1})^T T(S)^{-1} \Pi_n(0)$$

has all its zeros in the open unit disc. This proves the Proposition for $z_0 = 0$.

If z_0 is different from the origin, we may use the transformation

$$z = \frac{\hat{z} + z_0}{1 + \overline{z_0}\hat{z}} \quad \Longleftrightarrow \quad \hat{z} = \frac{z - z_0}{1 - \overline{z_0}z},$$

which maps the open unit disk into itself and which preserves the unit circle. Clearly, $z = z_0$ is mapped on $\hat{z} = 0$. Moreover, the block Toeplitz structure is not altered. Indeed, this follows from the Vandermonde decomposition of $T(S)$, see Theorem 2.22. □

Theorem 6.11. *Let $\tilde{P} \in \operatorname{int} \mathcal{K}_{\mathbb{C}}$, $z_0 \in \mathbb{C}$ such that $|z_0| < 1$ and R be an $m \times m$ positive definite matrix. The primal problem*

$$\begin{array}{ll} \max & \langle \Pi_n(z_0) R^{-1} \Pi_n(z_0)^*, Y \rangle \\ \text{s.t.} & \tilde{P} = T^*(Y) \\ & Y = Y^* \succeq 0 \end{array} \tag{6.6}$$

has an optimal solution $\hat{Y}$ of rank m, which gives the stable spectral factor of $\hat{P}$. If $\hat{Y} = QQ^*$, then $Q(z) = Q^*\Pi_n(z)$ has all its roots outside the unit disk. The optimal solution $\hat{S}$ of the dual problem

$$\begin{array}{ll} \min & \langle S, \tilde{P} \rangle \\ \text{s.t.} & T(S) \succeq \Pi_n(z_0) R^{-1} \Pi_n(z_0)^* \end{array}$$

is such that $T(\hat{S})$ is positive definite. Moreover, the solutions of both problems are strictly complementary.

Proof. The proof is similar to the scalar case. By Lemma 6.9, $T(\hat{S})$ is necessarily positive definite. Since both problems are strictly feasible, there is no duality gap and the complementary slackness condition holds [144], that is

$$\hat{Y}(T(\hat{S}) - \Pi_n(z_0) R^{-1} \Pi_n(z_0)^*) = 0.$$

Therefore,

$$\begin{aligned} \hat{Y} &= \hat{Y}(\Pi_n(z_0) R^{-1} \Pi_n(z_0)^*) T(\hat{S})^{-1} \\ &= T(\hat{S})^{-1} (\Pi_n(z_0) R^{-1} \Pi_n(z_0)^*) \hat{Y} (\Pi_n(z_0) R^{-1} \Pi_n(z_0)^*) T(\hat{S})^{-1} \\ &= (T(\hat{S})^{-1} \Pi_n(z_0) R^{-1/2})(R^{-1/2} \Pi_n(z_0)^* \hat{Y} \Pi_n(z_0) R^{-1/2})(R^{-1/2} \Pi_n(z_0)^* T(\hat{S})^{-1}). \end{aligned}$$

As a direct consequence of these equalities, we get that

$$\hat{Y} = T(\hat{S})^{-1} \Pi_n(z_0) R^{-1/2} X R^{-1/2} \Pi_n(z_0)^* T(\hat{S})^{-1}$$

for some positive definite matrix $X \in \mathbb{C}^{m \times m}$. Note that X must be positive definite because $\hat{Y}$ corresponds to a polynomial $\tilde{P}$ belonging to the interior of $\mathcal{K}_{\mathbb{C}}$. Taking the trace of the complementary slackness condition, we obtain that

$$\operatorname{Trace} X[M - MM] = 0,$$

where $M = R^{-1/2} \Pi_n(z_0)^* T(\hat{S})^{-1} \Pi_n(z_0) R^{-1/2}$. As $X \succ 0$, $M = M^* \succ 0$ and $M - MM \succeq 0$, we conclude that $\Pi_n(z_0)^* T(\hat{S})^{-1} \Pi_n(z_0) = R$.

Finally, $Q(z) = X^{1/2} R^{-1/2} \Pi_n(z_0)^* T(\hat{S})^{-1} \Pi_n(z)$ has all its roots outside the unit disk as a consequence of Proposition 6.10. □

As problem (6.6) has $\mathcal{O}(nm^2)$ linear constraints, solving the semidefinite formulation is rather expensive. Indeed, the generic complexity of a path-following primal-dual solver (such as SDPT3) is $\mathcal{O}(m^6n^4)$ flops per iteration. Our fast dual solver allows us to speed up the computations by a factor n so that the complexity is $\mathcal{O}(m^6n^3)$ flops per iteration. Remember that the number of iterations is $\mathcal{O}(\sqrt{mn}\log\frac{1}{\epsilon})$ in the worst case.

Remark. From a numerical point of view, the choice $z_0 = 0$ is clearly the most appropriate for solving (6.2) or (6.6).

Remark. The result presented above can be extended to other curves of the complex plane, namely the real line and the imaginary axis, by means of a one-to-one transformation of the objective function, e.g. Fischer-Frobenius transformation [73, 83].

6.1.2 Spectral factorization via linear algebra

Consider the linear operator

$$X \in \mathcal{H}_n \to Y(X) = \begin{bmatrix} R & S^* \\ S & Q \end{bmatrix} + \begin{bmatrix} F^*XF & F^*XZ \\ Z^*XF & Z^*XZ - X \end{bmatrix} \in \mathcal{H}_{n+1},$$

where $F = e_0$, Z is the shift matrix and

$$\left[\begin{array}{c|c} R & S^* \\ \hline S & Q \end{array}\right] \doteq \left[\begin{array}{c|ccc} p_0 & p_1^*/2 & \dots & p_n^*/2 \\ \hline p_1/2 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ p_n/2 & 0 & \dots & 0 \end{array}\right].$$

By construction, we have that

$$\pi_n(z)^*Y(X)\pi_n(z) = \langle p, \pi_n(z)\rangle_{\mathbb{R}} = p(z), \quad z \in \mathbb{T}.$$

If the trigonometric polynomial $p(z)$ is non-negative, then there exist Hermitian matrices X such that $Y(X)$ is positive semidefinite, see Theorems 2.16 and 2.17. Note that the linear operator $Y(X)$ is exactly equivalent to (2.56).

The *discrete-time algebraic Riccati equation* (DARE)

$$Q + Z^*XZ - X - (S + Z^*XF)(R + F^*XF)^{-1}(S^* + F^*XZ) = 0 \tag{6.7}$$

plays a crucial role since it yields a spectral factorization of $p(z)$. Indeed, $\operatorname{rank} Y(X) = 1$ if and only if the Schur complement of $R + F^*XF$ in $Y(X)$ is equal to 0. This last condition is equivalent to the Riccati equation.

In order to compute a solution to the DARE, several different algorithms can be used. Let us focus on the scheme based on the *Riccati difference equation*

$$X_0 = 0, \tag{6.8a}$$

$$X_{k+1} = Q + Z^*X_kZ - (S + Z^*X_kF)(R + F^*X_kF)^{-1}(S^* + F^*X_kZ). \tag{6.8b}$$

Remember that we are interested in a solution X for which the spectral decomposition of $Y(X)$ yields the stable spectral factor of $p(z)$. It turns out that the above recursion computes a stabilizing solution $\hat{X} = \lim_{k\to+\infty} X_k$ of (6.7). Indeed, the properties of this iterative scheme are as follows [6]:

1. the iterates $\{X_k\}_k$ satisfy $0 \succeq X_k \succeq X_{k+1}$;
2. the matrix $\hat{X}$ satisfies the DARE and $Y(\hat{X})$ is positive semidefinite;
3. for any negative semidefinite X such that $Y(X) \succeq 0$, one has $\hat{X} \succeq X$;
4. the feedback matrix

$$Z - F(R + F^*\hat{X}F)^{-1}(S^* + F^*\hat{X}Z)$$

has all its eigenvalues in the closed unit disk.

Remark. Property 3 is equivalent to the fact that $\hat{X}$ is an extremal solution of the LMI $Y(X) \succeq 0$. Therefore, the convex optimization problem

$$\begin{array}{ll} \max & \langle I, X \rangle \\ \text{s.t.} & Y(X) \succeq 0, \end{array}$$

yields the stabilizing solution $\hat{X}$.

After computing the stabilizing solution $\hat{X}$, the spectral factor is obtained via the following factorization

$$\begin{bmatrix} 1 \\ (S + Z^*\hat{X}F)(R + F^*\hat{X}F)^{-1} \end{bmatrix} (R + F^*\hat{X}F) \begin{bmatrix} 1 & (R + F^*\hat{X}F)^{-1}(S^* + F^*\hat{X}Z) \end{bmatrix}$$

of $Y(\hat{X})$. From Property 4, we conclude that the spectral factor

$$q(z) = \begin{bmatrix} (R + F^*\hat{X}F)^{1/2} & (R + F^*\hat{X}F)^{-1/2}(S^* + F^*\hat{X}Z) \end{bmatrix} \pi_n(z)$$

has all its roots outside the open unit disk.

The Riccati difference equation (6.8) is a Schur-complement based algorithm for computing the spectral factor. To see this, rewrite the iterative step as

$$X_{k+1} = X_k + [Q + Z^*X_kZ - X_k - (S + Z^*X_kF)(R + F^*X_kF)^{-1}(S^* + F^*X_kZ)]$$

and note that the matrix

$$Q + Z^*X_kZ - X_k - (S + Z^*X_kF)(R + F^*X_kF)^{-1}(S^* + F^*X_kZ)$$

is the Schur complement of $R + F^*X_kF$ in the matrix $Y(X_k)$. At each iteration this Schur complement is added to X_k, i.e it is added to the upper left block of $Y(X_k)$ and substracted from the lower right block.

The extension to matrix-valued trigonometric polynomials is straightforward. Indeed, the matrices R and S are now defined by the block coefficients, i.e.,

$$\left[\begin{array}{c|c} R & S^* \\ \hline S & Q \end{array}\right] \doteq \left[\begin{array}{c|ccc} P_0 & P_1^*/2 & \dots & P_n^*/2 \\ \hline P_1/2 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ P_n/2 & 0 & \dots & 0 \end{array}\right],$$

whereas $F = e_0 \otimes I_m$ and Z is the block shift matrix.

The complexity of iterating the Riccati difference equation is $\mathcal{O}(m^3n^2)$ flops. In the non-degenerate case $P \in \operatorname{int}\mathcal{K}_{\mathbb{C}}$, the convergence rate of the Riccati difference equation is linear. The rate of convergence is determined by the maximal magnitude of the roots of $P(z)$ belonging to the closed unit disk [6]. Indeed, we have that

$$\left\|X_k - \hat{X}\right\|_2 \leq \rho^{2k} \left\|\hat{X}\right\|_2, \tag{6.9}$$

where ρ is the spectral radius of the feedback matrix. In the degenerate case, the convergence of $\left\|X_k - \hat{X}\right\|_2$ to zero occurs at least as fast as k^{-1}.

Anderson, Hitz and Diem [6] have made the link between the iterative scheme based on Riccati difference equation and Bauer's algorithm [129] for computing the canonical spectral factor.

6.2 Filter design

In this section, we describe an important problem in signal processing, i.e., digital filter design with lowpass constraints.

6.2.1 Problem statement

Consider a class of discrete-time systems whose input u and output y satisfy a *linear constant coefficient difference equation* of the form

$$\sum_{k=0}^{n} a_k y[t-k] = \sum_{k=0}^{n} b_k u[t-k]$$

with real coefficients $\{a_k\}_{k=0}^n$ and $\{b_k\}_{k=0}^n$. From a mathematical point of view, a convenient way of handling this difference equation is to make use of the complex exponential. That is, we use the frequency-domain representation of corresponding discrete-time system. The *frequency response*, i.e., the system transfer function evaluated on the unit circle, has the form

$$H(e^{\jmath\omega}) = \frac{\sum_{k=0}^{n} a_k e^{-\jmath\omega k}}{\sum_{k=0}^{n} b_k e^{-\jmath\omega k}}.$$

In practice, linear time-invariant discrete time systems are often used to implement *frequency-selective filters*. In this section we focus on *discrete-time infinite impulse response* (IIR) lowpass filters, for which the passband is centered around zero. Figure 6.1 illustrates the corresponding specifications, the mathematical formulation of which is

$$1 - \delta_1 \leq |H(e^{\jmath\omega})| \leq 1 + \delta_1, \quad 0 \leq \omega \leq \omega_s, \tag{6.10a}$$

$$0 \leq |H(e^{\jmath\omega})| \leq \delta_2, \quad \omega_p \leq \omega \leq \pi. \tag{6.10b}$$

Remark. Since the function $H(e^{\jmath\omega})$ is 2π-periodic and satisfies $H(e^{-\jmath\omega}) = \overline{H(e^{\jmath\omega})}$, it is sufficient to specify it over the interval $[0, \pi]$.

Although several classical methods exist for designing lowpass filters, see [116, Chapter 7], the aim of this section is to propose an original method based on our

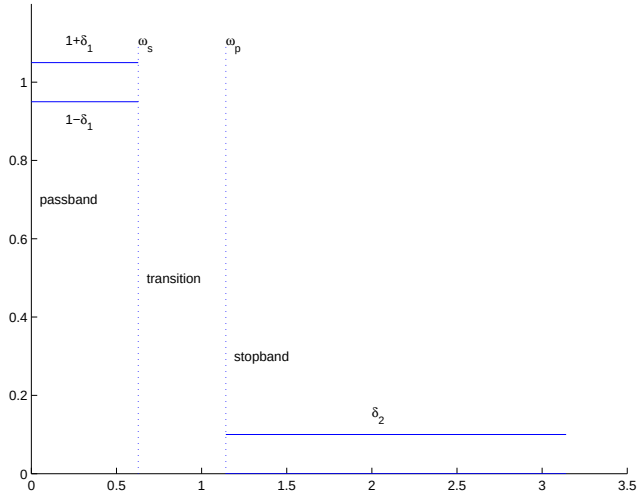

Figure 6.1. *Specifications for discrete-time IIR lowpass filter*

parametrization of non-negative polynomials. A similar methodology has been applied recently by a number of independent researchers to other filter design problems and promising results have been obtained [4, 37, 38, 49].

In order to completely formulate the optimal design problem, an objective function must be added to the constraints (6.10). Three standard characteristics can be taken into account:

1. the passband ripple;
2. the stopband attenuation;
3. the degree n of the filter.

and the objective functions are $\min \delta_1$, $\min \delta_2$ and $\min n$, respectively.

Because magnitude filter design optimization problems are not convex in general, numerical methods are not guaranteed to find the globally optimal design [151]. To get around this convexity issue, the constraints (6.10) are best written using the squared magnitude of the filter frequency response,

$$R(\omega) = |H(e^{\jmath\omega})|^2.$$

The semi-infinite inequality constraints are then rewritten as

$$(1-\delta_1)^2 \leq R(\omega) \leq (1+\delta_1)^2, \quad 0 \leq \omega \leq \omega_s, \tag{6.11a}$$

$$0 \leq R(\omega) \leq \delta_2^2, \quad \omega_p \leq \omega \leq \pi, \tag{6.11b}$$

$$R(\omega) \geq 0, \quad 0 \leq \omega \leq \pi, \tag{6.11c}$$

where $0 < \omega_s < \omega_p < \pi$.

A straightforward *approximation* of the semi-infinite inequality constraints uses N sampling frequencies

$$0 \leq \omega_1 \leq \cdots \leq \omega_N \leq \pi$$

and replaces the semi-infinite inequality constraints with the corresponding ordinary inequalities. A standard rule of thumb is to choose $N \simeq 15n$ linearly spaced sampling frequencies [151].

Although this discretization strategy yields good approximate solutions, the semi-infinite constraints can be enforced exactly! The key point of our development is the existence of two (trigonometric) cosine polynomials $a(z)$ and $b(z)$ such that

$$R(\omega) = |H(e^{\jmath\omega})|^2 = \frac{|\sum_{k=0}^{n} a_k e^{-\jmath\omega k}|^2}{|\sum_{k=0}^{n} b_k e^{-\jmath\omega k}|^2} = \frac{a(e^{-\jmath\omega})}{b(e^{-\jmath\omega})}.$$

Clearly, the real coefficients $\{a_k\}_{k=0}^n$ and $\{b_k\}_{k=0}^n$ can be recovered by taking the stable spectral factors of $a(z)$ and $b(z)$, respectively. By making use of our parametrization of non-negative trigonometric polynomials, we propose an original scheme for finding optimal designs of lowpass IIR filters. The proposed algorithm combines a bisection rule on the design parameter and an algorithm for checking feasibility of some auxiliary problems.

6.2.2 Non-negative trigonometric polynomials on an arbitrary segment

Our previous characterizations of non-negative trigonometric polynomials, Theorems 2.16 and 2.19, are not sufficient in our filter design context. Indeed, we need to characterize trigonometric polynomials that are non-negative on an arbitrary arc of the unit circle. Given ω_a and ω_b with $\omega_a < \omega_b$ and $|\omega_b - \omega_a| < 2\pi$, let us set

$$\omega = \frac{\omega_b - \omega_a}{2}, \tag{6.12}$$

$$\theta_s = \frac{\omega_a + \omega_b}{2}. \tag{6.13}$$

Note that ω is the angular width and θ_s is the bisector angle.

Consider the cone of trigonometric matrix polynomials (of degree n) that are non-negative on the arc $[\omega_a, \omega_b]$

$$\mathcal{K}_{[\omega_a,\omega_b]} = \{P \in E : P(z) = \frac{1}{2}\sum_{k=0}^{n}(P_k z^{-k} + P_k^* z^k) \succeq 0, z = e^{\jmath\theta}, \forall\theta \in [\omega_a, \omega_b]\},$$

where E is defined in (2.24).

Theorem 6.12. *The self-adjoint trigonometric matrix polynomial P (of degree n) belongs to $\mathcal{K}_{[\omega_a,\omega_b]}$ if and only if*

$$P(z) = P_1(z) + (e^{-\jmath\theta_s} z + e^{\jmath\theta_s} z^{-1} - 2\cos\omega)P_2(z), \quad \forall z \in \mathbb{T},$$

where $P_1(z)$ and $P_2(z)$ are non-negative trigonometric polynomials of degree n and $n-1$, respectively.

Proof. Note that the trigonometric polynomial $P(e^{\jmath\theta})$ is non-negative on $[\omega_a, \omega_b]$ if and only if the trigonometric polynomial $\tilde{P}(e^{\jmath\theta}) = P(e^{\jmath(\theta+\theta_s)})$ is non-negative on $[-\omega, +\omega]$. Hence Theorem 2.18 yields the representation

$$\tilde{P}(z) = \tilde{P}_1(z) + (z + z^{-1} - 2\cos\omega)\tilde{P}_2(z), \quad z \in \mathbb{T}.$$

Therefore, the result follows from the identity $P(z) = \tilde{P}(ze^{-\jmath\theta_s}), \forall z \in \mathbb{T}$. □

Theorem 6.13. *Given $S \in E$, define the block linear operators T_3 and T_4 by*

$$[T_3(S)]_{ij} = S_{i-j}, \quad 0 \leq i, j \leq n,$$
$$[T_4(S)]_{ij} = e^{-\jmath\theta_s} S_{i-j+1} + e^{\jmath\theta_s} S_{i-j-1} - 2S_{i-j}\cos\omega, \quad 0 \leq i, j \leq n-1,$$

with $S_{-i} = S_i^*$. Then

1. *Primal cone:*

$$\mathcal{K}_{[\omega_a,\omega_b]} = \{P : P = T_3^*(Y_3) + T_4^*(Y_4), Y_3 \in \mathcal{H}_+^{m(n+1)}, Y_4 \in \mathcal{H}_+^{mn}\}.$$

2. *Dual cone:*

$$\mathcal{K}^\star_{[\omega_a,\omega_b]} = \{S : T_3(S) \succeq 0, T_4(S) \succeq 0\}.$$

3. *Both cones are closed, convex and pointed. Consequently, they also have non-empty interiors.*

Proof. This proof has exactly the same pattern as the one described in Theorem 2.19, provided that the functional systems of the sub-cones are adapted. In this proof, theses functional elements are

$$\phi(z) = 1, \quad \psi(z) = \pi_n(z), \quad \Psi(z) = \Pi_n(z),$$

and

$$\phi(z) = (e^{-\jmath\theta_s} z + e^{\jmath\theta_s} z^{-1} - 2\cos\omega), \quad \psi(z) = \pi_{n-1}(z), \quad \Psi(z) = \Pi_{n-1}(z),$$

respectively. The structure of the associated linear operators are obtained from the identities

$$[\pi_n(z)\pi_n(z)^*]_{ij} = z^{i-j}$$

and

$$[(e^{-\jmath\theta_s} z + e^{\jmath\theta_s} z^{-1} - 2\cos\omega)\pi_{n-1}(z)\pi_{n-1}(z)^*]_{ij} = e^{-\jmath\theta_s} z^{i-j+1} + e^{\jmath\theta_s} z^{i-j-1} - 2z^{i-j}\cos\omega.$$

The rest of the proof is unchanged, except for some minor and obvious changes. □

In the context of filter design problems, this result is applied to symmetric intervals of the form $[\omega_p, 2\pi - \omega_p]$. Consequently we have that $\omega = \pi - \omega_p$ and $\theta_s = \pi$. For these values, the linear operators T_3 and T_4 are given by

$$[T_3(S)]_{ij} = S_{i-j}, \quad 0 \leq i, j \leq n,$$
$$[T_4(S)]_{ij} = -S_{i-j+1} - S_{i-j-1} - 2S_{i-j}\cos(\pi - \omega_p), \quad 0 \leq i, j \leq n-1.$$

6.2.3 Filter design algorithm

Feasibility problem

Given δ_1, δ_2 and n, we define

$$l_1 = 1 - \delta_1, \quad l_2 = 0,$$
$$u_1 = 1 + \delta_1, \quad u_2 = \delta_2,$$
$$\Omega_1 = [0, \omega_s], \quad \Omega_2 = [\omega_p, \pi].$$

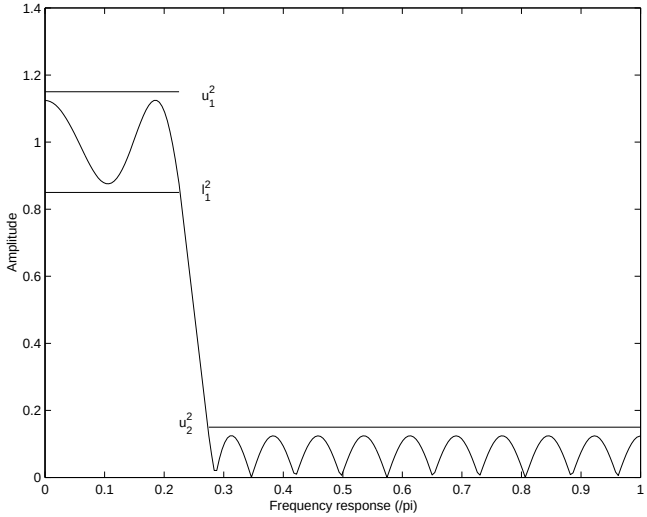

Figure 6.2. *Bandpass filter design ($\omega_s = 0.225$, $\omega_p = 0.275$, $n = 26$)*

so as to get more degrees of freedom. The feasibility problem corresponding to our lowpass rational filter design problem can then be formulated as follows:

> Find the (trigonometric) cosine polynomials $a(z)$ and $b(z)$ such that the design constraints
>
> $$l_i^2 \leq \frac{a(z)}{b(z)} \leq u_i^2, \quad z = e^{\jmath\omega}, \forall \omega \in \Omega_i, \quad i = 1, 2$$
>
> are satisfied (see Figure 6.2).

We have slightly modified these design constraints in the proposed solution so as to control the well-known Gibbs phenomenon, which is associated with a non-uniform convergence. The altered constraints are given by

$$\begin{aligned}
l_1^2 &\leq \frac{a(z)}{b(z)}, && z = e^{\jmath\omega}, \forall \omega \in [0, \omega_s], \\
\frac{a(z)}{b(z)} &\leq u_1^2, && z = e^{\jmath\omega}, \forall \omega \in [0, \omega_p], \\
0 = l_2^2 &\leq \frac{a(z)}{b(z)}, && z = e^{\jmath\omega}, \forall \omega \in [\omega_s, \pi], \\
\frac{a(z)}{b(z)} &\leq u_2^2, && z = e^{\jmath\omega}, \forall \omega \in [\omega_p, \pi].
\end{aligned}$$

Let us rewrite these inequalities in order to get non-negativity constraints on (trigonometric) cosine polynomials:

$$\begin{aligned}
p_0(z) &\doteq a(z) \geq 0, && z = e^{\jmath\omega}, \forall \omega \in [0, \pi], \\
p_1(z) &\doteq a(z) - l_1^2 b(z) \geq 0, && z = e^{\jmath\omega}, \forall \omega \in [0, \omega_s], \\
p_2(z) &\doteq u_1^2 b(z) - a(z) \geq 0, && z = e^{\jmath\omega}, \forall \omega \in [0, \omega_p], \\
p_3(z) &\doteq u_2^2 b(z) - a(z) \geq 0, && z = e^{\jmath\omega}, \forall \omega \in [\omega_p, \pi].
\end{aligned}$$

By setting $\mathcal{K} = \mathcal{K}_{\mathbb{C}} \times \mathcal{K}_{[0,\omega_s]} \times \mathcal{K}_{[0,\omega_p]} \times \mathcal{K}_{[\omega_p, 2\pi-\omega_p]}$, the auxiliary feasibility problem becomes

> Find two (trigonometric) cosine polynomials $a(z) = \langle a, \pi_n(z) \rangle_{\mathbb{R}}$ and $b(z) = \langle b, \pi_n(z) \rangle_{\mathbb{R}}$ such that

$$\mathcal{A} \begin{bmatrix} a \\ b \end{bmatrix} \doteq \begin{bmatrix} I_{n+1} & 0 \\ I_{n+1} & -l_1^2 I_{n+1} \\ -I_{n+1} & u_1^2 I_{n+1} \\ -I_{n+1} & u_2^2 I_{n+1} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \in \mathcal{K}. \tag{6.14}$$

> Let $F_*(s)$ be the normal barrier function of the dual cone

$$\mathcal{K}^\star = \mathcal{K}_{\mathbb{C}}^\star \times \mathcal{K}_{[0,\omega_s]}^\star \times \mathcal{K}_{[0,\omega_p]}^\star \times \mathcal{K}_{[\omega_p, 2\pi-\omega_p]}^\star.$$

Note that the complexity parameter of this barrier is $7n + 1$. If $\tilde{s} \in \operatorname{int} \mathcal{K}^\star$, then the constraint (6.14) is feasible if and only if the optimization problem

$$\begin{aligned}
\min \quad & F_*(s) \\
\text{s.t.} \quad & \mathcal{A}^* s = \mathcal{A}^* \tilde{s} \\
& s \in \operatorname{int} \mathcal{K}^\star \subset \mathbb{R}^{4(n+1)}
\end{aligned} \tag{6.15}$$

is bounded. From the associated analytic center $\hat{s} \in \operatorname{int} \mathcal{K}^\star$, the gradient mapping property of $F_*(s)$ yields an interior point of $\mathcal{K}$. Furthermore, the structure of the linear operator $\mathcal{A}$ is such that both feasible polynomials $a(z)$ and $b(z)$ can be easily obtained from the first $2(n+1)$ components of this mapping, i.e.,

$$a = \begin{bmatrix} \hat{s}_0 \\ \vdots \\ \hat{s}_n \end{bmatrix}, \quad b = \frac{1}{l_1^2} \left(\begin{bmatrix} \hat{s}_0 \\ \vdots \\ \hat{s}_n \end{bmatrix} - \begin{bmatrix} \hat{s}_{n+1} \\ \vdots \\ \hat{s}_{2n+1} \end{bmatrix} \right).$$

Algorithm

Because the whole optimization problem (objective and constraints) is not convex but quasi-convex, our iterative scheme combines a bisection rule on the design parameter and solving auxiliary feasibility subproblems.

For instance, assume that δ_1 and n are fixed and that we minimize the stopband attenuation. The corresponding algorithm can be described by the following steps :

Algorithm 6.1 Filter design algorithm

1. Choose an accuracy $\epsilon > 0$. Set δ_2 at some starting value.

2. Solve the analytic center problem (6.15). If the problem is unbounded, then the constraint (6.14) is not feasible and δ_2 is increased. Otherwise, δ_2 is decreased, e.g. $\delta_2 \leftarrow \delta_2/2$. Before iterating this step, check whether δ_2 is less than ϵ.

3. Using the gradient mapping associated with F_*, recover $a(z)$ and $b(z)$. Compute the spectral factorization of these non-negative trigonometric polynomials to get the real coefficients $\{a_k\}_{k=0}^n$ and $\{b_k\}_{k=0}^n$.

Of course, the bisection rule on δ_2 can be replaced by one on δ_1 or n.

Even if our formulation is correct, its numerical behaviour is not always adequate. In fact, this algorithm has been implemented in MATLAB and it breaks down as soon as $n \geq 7$. As a matter of fact, it turns out that the problem formulation is intrinsically ill-conditioned as put into light by numerical experiments. Similar numerical problems were observed in the finite-impulse response (FIR) magnitude filter design formulation of [4, Section 4.3].

Therefore, an appropriate reformulation of our problem approach should be looked for that keeps the property of our original treatment: it should not be based on approximations of the semi-infinite inequality constraints. This is still an open problem and we think that optimizing over polynomials non-negative on intervals is a challenging and interesting problem (especially if several different intervals are combined).

6.3 Probability theory and statistics

Optimization problems on cones of non-negative polynomials and moment spaces are common in probability theory and in statistics. They also appear in related fields such as signal processing, system identification and telecommunications. In the previous chapters, we have proposed original characterizations and efficient methods for solving these problems. In this section, we have selected three important classes of problems, which are frequently encountered in practice. Their wide range of applications (mathematical finance, systems engineering,...) motivates our brief description of these problems. Because they were initially solved by other researchers, we refer to the literature for a complete description, see the section "Notes and references".

Apart from the usual mathematical objects, we also need several well-known concepts of probability theory [32]. Let $(\Omega, \mathcal{F}, P)$ be a probability space. A measurable function X on this probability space is called a *random variable*. The *expectation* of X is defined as the integral

$$EX \doteq \int_\Omega X dP.$$

The *indicator function* $\mathbb{1}_{\{A\}}(x)$ is equal to 1 if $x \in A$ and 0 otherwise.

6.3.1 Bounds on probability

Given a random variable X on $(\Omega, \mathcal{F}, P)$, we would like to bound the probability

$$P\{X \in A\} \doteq \int_A dP = E(\mathbb{1}_{\{X \in A\}})$$

from above using the first n moments of X, i.e., EX^k, $k = 1, \ldots, n$. This bound can be derived because the expectation operator, which is linear, also preserves order. To be more specific, we have the following well-known result.

Theorem 6.14 (Order preservation, [32, Theorem 4.1(iii)]). *If X and Y are measurable functions on $(\Omega, \mathcal{F}, P)$ and $X \geq Y \geq 0$ almost certainly, then*

$$EX \geq EY \geq 0.$$

For any non-negative random variable X, a direct consequence of this Theorem is the *Markov inequality* [32]

$$P\{X \geq a\} \leq \frac{1}{a} EX, \quad a \in \operatorname{int} \mathbb{R}_+. \tag{6.16}$$

Indeed, the first-order polynomial $p(x) = \frac{x}{a}$ yields an upper bound on $\mathbb{1}_{\{X \geq a\}}(x)$ and the inequality follows by taking the expectation on both sides

$$P\{X \geq a\} = E\,\mathbb{1}_{\{X \geq a\}}(x) \leq Ep(X) = E\left(\frac{x}{a}\right).$$

This idea can be generalized to a polynomial

$$p(x) = \sum_{k=0}^{n} p_k x^k$$

of arbitrary degree n, for which $p(x) \geq \mathbb{1}_{\{X \in A\}}(x)$. Of course, it requires the knowledge of the first n moments of X, say $x_k = EX^k$. These moments can be obtained directly or after some regularization process. A tight polynomial upper bound on $P\{X \in A\}$ is then obtained via the convex optimization problem

$$\begin{array}{lll} \min & Ep(X) = \sum_{k=0}^{n} p_k EX^k & \\ \text{s.t.} & p(x) \geq 0, & \forall x \in \mathbb{R}, \\ & p(x) \geq 1, & \forall x \in A. \end{array}$$

Clearly, the polynomial constraints of this problem are non-negativity constraints, which imply the inequality $p(x) \geq \mathbb{1}_{\{X \in A\}}(x)$ for all $x \in \mathbb{R}$. This optimization problem can be solved without any approximation of the constraints by making use of our characterization results, see Section 2.2.

Example 6.1 (Chebyshev inequality). Let X be a random variable X on $(\Omega, \mathcal{F}, P)$ and define $X_c = X - EX$. We are interested in a polynomial upper bound on the probability $P\{|X_c| \geq a\}$, which is obtained via some quadratic polynomial $p(x) = p_0 + p_1 x + p_2 x^2$. By symmetry, we must have $p_1 = 0$. Moreover, we have $p_0 = 0$ and $p_2 = \frac{1}{a^2}$, see Figure 6.3. Consequently, we obtain the *Chebyshev inequality*

$$P\{|X - EX| \geq a\} = P\{|X_c| \geq a\} \leq \frac{E(X_c)^2}{a^2} = \frac{E(X - EX)^2}{a^2}.$$

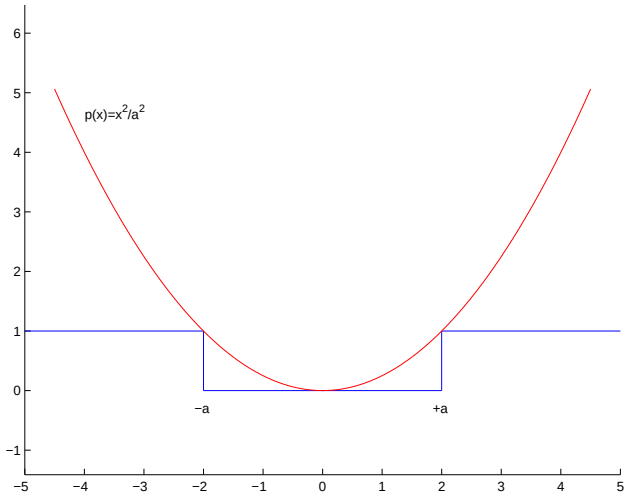

Figure 6.3. *Optimal quadratic polynomial (Chebyshev inequality)*

6.3.2 Estimation of moments

Let X be a random variable on $(\Omega, \mathcal{F}, P)$. In practice, the moments of X are estimated from a set of observations $\{X_i\}_{i=0}^{N-1}$. For instance, it is well known that an estimator of the expectation of X is

$$\overline{X} = \frac{\sum_{i=0}^{N-1} X_i}{N}.$$

If several moments are estimated, inconsistent results can be obtained. Indeed, they are potentially affected by measurement and numerical errors, i.e., estimation errors. Two different methods can be applied in order to correct these estimates, which are denoted by

$$\tilde{s}_k \approx EX^k, \quad k = 0, \ldots, n.$$

The first strategy is based on a *least-squares approximation*. Given a positive definite weight matrix W, the corrected estimates $\hat{s}$ are obtained via the convex optimization problem

$$\begin{array}{rl} \hat{s} = \arg\min & \frac{1}{2}\langle W(s - \tilde{s}), s - \tilde{s}\rangle \\ \text{s. t.} & s \in \mathcal{K}_{\mathbb{R}}^{\star}. \end{array}$$

The constraint $s \in \mathcal{K}_{\mathbb{R}}^{\star}$ is the moment constraint, which imposes that s is a sequence of moments, see Section 2.4. This optimization problem can be efficiently solved using interior-point methods.

The second strategy is based on a *determinant maximization problem*. Indeed, statistical methods often provide $(1 - \alpha)$ confidence intervals for estimates. Practical values of the significance level α are 0.01 and 0.05. Assume that the confidence limits are equivalent to the inequalities

$$\underline{s}_k \leq s_k \leq \overline{s}_k, \quad k = 1, \ldots, n.$$

Then the corrected estimates $\hat{s}$ are obtained via the convex optimization problem

$$\begin{array}{rll} \hat{s} = \arg\min & -\log\det H(s) & \\ \text{s.t.} & s_0 = 1, & \\ & \underline{s}_k \leq s_k \leq \overline{s}_k, & k \in V \subseteq \{1,\ldots,n\} \\ & s \in \mathcal{K}_{\mathbb{R}}^{\star} & \end{array}$$

where $H(s)$ is the Hankel matrix defined by s. In general, the optimal point $\hat{s}$ is an interior point of $\mathcal{K}_{\mathbb{R}}^{\star}$ so that $\hat{s}$ is a robust estimate of the moments. This property is highly desirable in a probabilistic context.

Of course, the minimal expectation of any polynomial function of the moments can be estimated via a similar methodology. In this case the problem formulation is as follows

$$\begin{array}{ll} \min & Ep(X) = \sum_{k=0}^{n} p_k s_k \\ \text{s.t.} & s_0 = 1, \\ & \underline{s}_k \leq s_k \leq \overline{s}_k, \qquad\qquad k \in V \subseteq \{1,\ldots,n\} \\ & s \in \mathcal{K}_{\mathbb{R}}^{\star}. \end{array}$$

6.3.3 Estimation of spectral densities

A vector $r \in \mathbb{R}^{n+1}$ is a *finite autocorrelation sequence* if there exists a vector $h \in \mathbb{R}^{n+1}$ such that

$$r_k = \sum_{i=0}^{n-k} h_i h_{i+k}, \quad k = 0,\ldots,n. \tag{6.17}$$

Given $r \in \mathbb{R}^{n+1}$, a necessary and sufficient condition which guarantees the identities (6.17) for some vector $h \in \mathbb{R}^{n+1}$ is

$$r_0 + 2\sum_{k=1}^{n} r_k \cos k\theta \geq 0, \quad \forall \theta \in [0, 2\pi].$$

Therefore, r represents a finite autocorrelation sequence if and only if

$$\begin{bmatrix} r_0 \\ 2r_1 \\ \vdots \\ 2r_n \end{bmatrix} \in \mathcal{K}_{\mathbb{C}} \cap \mathbb{R}^{n+1}.$$

Suppose that the vector $\tilde{r} \in \mathbb{R}^{n+1}$ contains the estimates of the first $n+1$ autocorrelation coefficients $Ew(t)w(t+k)$ of a stationary time serie $w(t)$. Because of estimation errors, these estimates might be inaccurate and the (trigonometric) cosine polynomial

$$r_0 + 2\sum_{k=1}^{n} r_k \cos k\theta$$

might be negative for some $\theta \in [0, 2\pi]$. We can correct them via a projection on the appropriate set.

Let W be a positive definite weight matrix. Given $\tilde{x} \in \mathbb{R}^{n+1}$, the least-squares problem

$$\begin{array}{rll} \hat{x} = \arg\min & \frac{1}{2}\langle W(x-\tilde{x}), x-\tilde{x}\rangle & \\ \text{s.t.} & x \in \mathcal{K}_{\mathbb{C}} \cap \mathbb{R}^{n+1} & \end{array}$$

yields the nearest element in the cone $\mathcal{K}_{\mathbb{C}} \cap \mathbb{R}^{n+1}$. From a computational point of view, the dual problem

$$\begin{array}{ll} \max & -\frac{1}{2}\langle W^{-1}s, s\rangle - \langle \tilde{x}, s\rangle \\ \text{s. t.} & s \in \mathcal{K}_{\mathbb{C}}^{\star} \cap \mathbb{R}^{n+1} \end{array}$$

can be solved more efficiently. Indeed, this dual formulation avoids extra variables and the dual Toeplitz structure can be efficiently exploited in the solution method, see Section 3.3.

Let us consider another interesting optimization problem over cones of non-negative polynomials, which solves an extension problem related to covariance sequences. Given a positive partial covariance sequence, say $\tilde{c} \in \operatorname{int} \mathcal{K}_{\mathbb{C}}^{\star} \cap \mathbb{R}^{n+1}$, and a polynomial

$$\sigma(z) = \sigma_0 + \cdots + \sigma_{n-1} z^{n-1} + z^n,$$

with all roots in the open unit disc, the *rational covariance extension problem* is formulated as follows [25].

Find the unique real polynomial

$$a(z) = a_0 + \cdots + a_{n-1} z^{n-1} + a_n z^n, \quad a_n > 0,$$

with all roots in the open unit disc, such that the rational function $w(z) = \frac{\sigma(z)}{a(z)}$ is a minimum-phase spectral factor of a spectral density

$$\Phi(z) = \tilde{c}_0 + \sum_{k=1}^{n} \tilde{c}_k (z^k + z^{-k}) + \sum_{k=n+1}^{+\infty} c_k (z^k + z^{-k}), \quad c_k \in \mathbb{R}.$$

In order to compute the solution of this extension problem, an algorithm involving non-negative polynomials is completely described in [25]. This numerical scheme was motivated by the effectiveness of interior-point methods for solving convex optimization problems; the rational covariance extension problem can be solved via the convex optimization problem

$$\begin{array}{ll} \min & \langle \tilde{c}, p\rangle - \frac{1}{2\pi} \int_{-\pi}^{\pi} |\sigma(e^{\jmath\theta})|^2 \log \langle p, \pi_n(e^{\jmath\theta})\rangle_{\mathbb{R}} d\theta \\ \text{s. t.} & p \in \operatorname{int} \mathcal{K}_{\mathbb{C}} \cap \mathbb{R}^{n+1}. \end{array} \tag{6.18}$$

Indeed, the optimal positive polynomial $\hat{p}(z)$ yields the solution $\hat{a}(z)$ via spectral factorization, $\hat{p}(z) = \hat{a}(z)\hat{a}(z^{-1})$.

Remark. The case $\sigma(z) = z^n$ leads to the maximum entropy solution, which is usually computed via the Levinson algorithm.

Remark. Although the function

$$F(p) = -\frac{1}{2\pi} \int_{-\pi}^{\pi} |\sigma(e^{\jmath\theta})|^2 \log \langle p, \pi_n(e^{\jmath\theta})\rangle_{\mathbb{R}} d\theta$$

is not a barrier function for the cone $\mathcal{K}_{\mathbb{C}}$, it can be proved that the optimal value of (6.18) is never attained on the boundary of $\mathcal{K}_{\mathbb{C}} \cap \mathbb{R}^{n+1}$, see [25, Lemma 5.4].

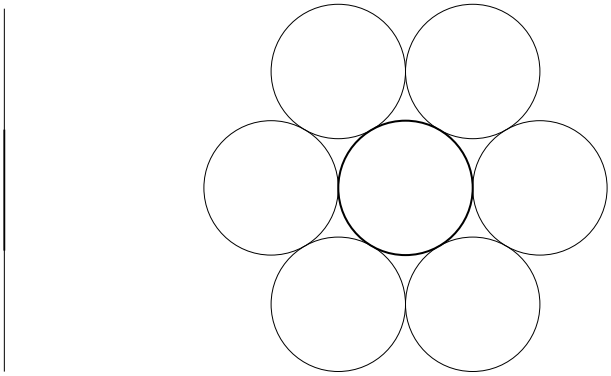

Figure 6.4. *Kissing numbers,* $n = 1$ *and* $n = 2$

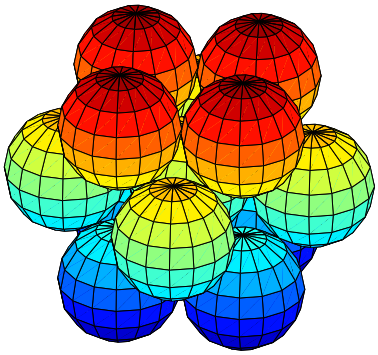

Figure 6.5. *Kissing number,* $n = 3$

6.4 Sphere packing and coding theory

The maximal number of unit spheres in $\mathbb{R}^n$ that can touch ("kiss") a central unit sphere without any intersections is known as the *kissing number* in dimension n (notation: τ_n). Other names for τ_n that have been used in the literature are *Newton number*, *contact number*, *coordination number* or *ligancy*.

Obviously, for $n = 1$ or 2, the exact value can be obtained via an explicit construction, see Figure 6.4. For $n = 3$, Isaac Newton believed that the answer was $\tau_3 = 12$, which was not proved until the *19*th century, see Figure 6.5.

Alternatively, the kissing number problem can be stated in the following terms:

> How many points can be placed on the surface of a unit sphere in $\mathbb{R}^n$ so that the angular separation between any two points is at least $\frac{\pi}{3}$?

The kissing number problem can therefore be regarded as a packing problem on the unit sphere.

This last formulation paves the way for a generalization of the kissing number

problem. Let $S_n = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ be the surface of the n-dimensional unit ball. A *spherical code* is a finite subset T of S_n. Its *minimal angle* ϕ is defined as the largest angle for which

$$\langle x, y \rangle \leq \cos \phi, \quad \forall x, y \in T, x \neq y.$$

For given n and ϕ, the generalized question is:

> What is the maximal number of points, say $A(n, \theta)$, in a spherical code of $\mathbb{R}^n$ with minimal angle ϕ ?

Although this very specific question is difficult to answer in general, upper bounds on $A(n, \theta)$ can be devised. The next theorem provides us with an interesting result in this direction.

Theorem 6.15 (Delsarte's LP bound, [44, Theorem 4.3]). *Let $N \in \mathbb{N}$ be a positive integer and denote by $P_k^{(\alpha,\beta)}$ the Jacobi polynomial of degree k with parameters (α, β). If*

$$f(t) = \sum_{k=0}^{N} f_k P_k^{(\alpha,\alpha)}(t), \quad \alpha = \frac{n-3}{2}$$

is a polynomial satisfying $f_0 > 0, f_k \geq 0 (k = 1, \ldots, N)$ and $f(t) \leq 0$ for $t \in [-1, \cos\theta]$, then

$$A(n, \theta) \leq \frac{f(1)}{f_0}.$$

From a computational point of view, an upper bound on $A(n, \theta)$ can be obtained via the convex optimization problem

$$\begin{array}{ll} \min & f(1) \\ \text{s.t.} & f_0 = 1, \\ & f(t) = \sum_{k=0}^{N} f_k P_k^{(\alpha,\alpha)}(t) \leq 0, \quad \forall t \in [-1, \cos\theta]. \end{array} \tag{6.19}$$

Of course the non-negativity constraint on $-f(t)$ can be treated efficiently and exactly by using the characterization obtained in Section 2.2. From a theoretical point of view, the situation is even better as an explicit formula is known for the optimal polynomial, provided that N is fixed, see [45, Theorem 9]. This last example shows that complex optimization problems involving non-negativity constraints can sometimes be solved explicitly. In this particular case, the theoretical solution is obtained via adjacent systems of orthogonal polynomials. We have not reproduced the closed formula of the solution. Indeed, this would entail the statement of many preliminary results, which are out of the scope of this section.

6.5 Research summary

The applications presented in this chapter can be classified in three different groups.

The first category is composed of applications that give us some insight into the geometry of our semidefinite characterizations of cones of non-negative polynomials. This is clearly the case for spectral factorization. Note that convex optimization problems, with non-negativity constraints on polynomials, might have a particular, but hidden, algebraic structure. This is exemplified by the kissing number problem.

The second category is comprised of signal processing applications. For these problems, efficient and reliable algorithms already exist, but they can be improved by exact treatment of non-negativity constraints on polynomials. In particular, we have shown that convex optimization might have an important role to play in the future.

The last category consists of problems appearing in probability theory and statistics. In this context, convex optimization formulations can definitely improve the quality of the numerical estimates. Because many mathematical models require several parameters to be estimated, the general methods presented in this chapter have potentially a huge number of practical applications. Of course, the fast dual solver outlined in Section 3.3 should preferably be used to efficiently solve these problems.

Notes and references

Theoretical results on spectral factorization of matrix polynomials are extensively mentioned in the literature, see e.g. [65, 128]. From a computational point of view, several algorithms have been developed to solve the spectral factorization problem. The recent survey of spectral factorization methods [129] emphasizes the relations that exist among them; numerical aspects of these schemes can be found in another survey paper [143]. Theorems 6.6 and 6.11 with $z_0 = 0$ were also observed in [47], where it was derived from results in [128]. In these references, the coefficients were allowed to be operators on a Hilbert space. The general result with $|z_0| < 1$ can be derived in that way as well. In addition, in [103] a simple proof is given for the rank property of the optimum. Finding the stabilizing solution of a Riccati difference equation is a well-studied problem in the systems and control literature, see e.g. [6, 19, 40]. To our point of view, two books on algebraic Riccati equations also deserve a special attention [20, 92].

Digital filter design is of the utmost importance in signal processing. Although hundreds of papers have been published on this subject, the semidefinite parametrization of non-negative polynomials was only mentioned recently, see e.g. [4, 37, 38, 49, 57]. However, there are still numerical troubles that need to be investigated further.

Tight upper bounds on probabilities using non-negative polynomials were obtained by Bertsimas and his co-authors. They have investigated their use in probability theory and finance applications, see e.g. [17, 18]. Estimation problems are well known in system identification. In particular, robust identification is a very active field of research in systems and control. Our applications can be found in [145], [4, Section 4.1] and references therein. Another recent result in this area, which involves optimization problems over moment spaces, can be found in [80]. The rational covariance extension problem as formulated in this chapter was recently solved by Byrnes et al., see [25, 26, 27] and references therein.

Delsarte's linear programming bound was first mentioned in Delsarte's thesis [41]. Sphere packing and coding theory are still very active fields of research. In particular, mathematicians are still trying to improve Delsarte's LP bound for particular codes. Our presentation of the kissing number problem is based on [34, Chapters 1, 9 and 13].

Conclusion

In this thesis, optimization problems over cones of non-negative polynomials have been addressed by making use of compact convex characterizations of these cones. Special attention has been devoted to the problem structures in order to obtain efficient algorithms. Several applications have been considered in order to highlight the importance of our results in the context of applied mathematics.

More specifically, the original contributions of this thesis can be divided into three groups.

Theory On the one hand, we have proposed a unified treatment of cones generated by sums of squares. Because of their semidefinite representability, these cones can be efficiently included in convex optimization problems. Our results can then be applied to various cones of non-negative polynomials, the duals of which are moment spaces. Some of them are related to the Kalman-Yakubovich-Popov Lemma. On the other hand, we have identified a property that is responsible for exact relaxation of quadratic optimization problems. We have shown that this property holds in the context of problems generated by non-negative polynomials. Furthermore, we have identified new classes of easy quadratic optimization problems, which are related to interpolation constraints.

Algorithms On the one hand, we have obtained a fast dual solver for conic optimization problems involving non-negative polynomials. The proposed algorithm dramatically reduces the computational cost by cleverly using the dual matrix structure; it has the best complexity known in the literature. On the other hand, we have shown how convexity can be exploited to efficiently compute the Crawford number. Our numerical experiments show that our cutting plane schemes are possible challengers to well-established algorithms. Their efficiency is related to an appropriate use of the problem structure.

Applications Several interesting problems can be solved by applying the results presented in this thesis. First, we have proposed new solutions to several distance problems associated with quadratic eigenvalue problems. The main advantage of our solutions is that they explicitly provide us with the structure of the optimal perturbations. In addition, our characterizations can be applied to solve well-known problems in systems and control (spectral factorization) and in signal processing (design of lowpass infinite-impulse response filters). Although our results on spectral factorization are probably hidden in the literature, they provide some insight into the underlying geometry of our characterizations. However, the use of semidefinite programming in signal processing is clearly original. Because efficient dual algorithms have been obtained, our results can also be applied

to identification problems (estimation of moments). Finally, we point out that several problems involving non-negative polynomials can at times be solved explicitly, see e.g. our discussion about the kissing number.

The methodology developed in this thesis is based on two fundamental ideas. We must first seek a convex parametrization or formulation of a given problem and, second, the associated structure has to be taken into account to obtain fast interior-point algorithms. Although this powerful methodology can be applied to a wide range of problems, it consists of two non-trivial tasks. In particular, further research topics related to our work can be outlined.

Non-negative multivariate polynomials There is currently a great interest in non-negative multivariate polynomials, a subject dating back to the early *20*th century (see Hilbert's address in Section 2.6). Indeed, sets of multivariate polynomials are theoretically investigated in optimization because they provide relaxation of hard combinatorial problems [39, 93, 94, 95, 119]. Another interesting topic of theoretical interest is related to the computation of spectral factors for multivariate polynomials. Although these spectral factors are not guaranteed to exist, approximate solutions can be sought.

Structure-specific optimization schemes Many optimization problems in systems and control enjoy structural properties [23]. Because they are usually not exploited in the solution methods, more efficient algorithms can be obtained by making use of the problem structure. Recall that our fast dual solver is based on such an observation. In the future, problem-specific algorithms for optimization problems in systems and control should become widespread. Moreover, other important issues (loss of numerical stability, loss of accuracy, reduction of worst-case complexity...) must also be addressed.

Applications We have seen that non-negative polynomials are very common objects and might be used as modelling tools. For this reason, the possible applications of our characterizations should be investigated further. Indeed, we have pointed out only several applications although the number is probably unlimited. Moreover, each application would appear to require a specific treatment so as to exploit its underlying structure.

In conclusion, this thesis provides an overview of convex optimization over cones of non-negative polynomials (primal spaces) and over moment spaces (dual spaces). We have obtained several structured algorithms for solving these optimization problems, as well as other related problems. We have shown the relevance of our results with several applications. As highlighted above, scientific activity in these research areas is expected to increase in the coming years. The self-contained treatment of several fundamental problems, as presented in this thesis, should thus prove extremely valuable in this evolving context.

Bibliography

[1] Y. S. Abramov, *Numerical ranges, zones and spectra of families of selfadjoint operators*, Dokl. Akad. Nauk SSSR, 257 (1981), pp. 1033–1037. English translation: Soviet Math. Dokl. 23 (1981), no. 2, 384–388.

[2] N. I. Ahiezer and M. Kreĭn, *Some questions in the theory of moments*, vol. 2 of Translation of Mathematical Monographs, American Mathematical Society, 1962.

[3] N. I. Akhiezer, *The classical moment problem and some related questions in analysis*, Oliver and Boyd, Edinburgh, 1965.

[4] B. Alkire and L. Vandenberghe, *Convex optimization problems involving finite autocorrelation sequences*, Math. Programming, 93 (2002), pp. 331–359.

[5] D. Alpay and P. Loubaton, *The partial trigonometric moment problem on an interval: the matrix case*, Linear Algebra Appl., 225 (1995), pp. 141–161.

[6] B. D. O. Anderson, K. L. Hitz, and N. D. Diem, *Recursive algorithm for spectral factorization*, IEEE Trans. Circuits and Systems, 21 (1974), pp. 742–750.

[7] T. Andô, *Truncated moment problems for operators*, Acta Sci. Math. (Szeged), 31 (1970), pp. 319–334.

[8] E. Artin, *Über die Zerlegung definiter Funktionen in Quadrate*, Abh. Math. Semin. Univ. Hamb., 5 (1927), pp. 100–115.

[9] D. S. Atkinson and P. M. Vaidya, *A cutting plane algorithm for convex programming that uses analytic centers*, Math. Programming, 69 (1995), pp. 1–43.

[10] M. Bakonyi and G. Nævdal, *The finite subsets of $\mathbb{Z}^2$ having the extension property*, J. London Math. Soc. (2), 62 (2000), pp. 904–916.

[11] L. Barkwell and P. Lancaster, *Overdamped and gyroscopic vibrating systems*, Trans. ASME J. Appl. Mech., 59 (1992), pp. 176–181.

[12] E. F. Beckenbach and R. Bellman, *Inequalities*, vol. 30 of Ergebnisse der Mathematik und ihrer Grenzgebiete, Springer-Verlag, New York, 1971. Third printing.

[13] B. Beckermann, *The condition number of real Vandermonde, Krylov and positive definite Hankel matrices*, Numer. Math., 85 (2000), pp. 553–577.

[14] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, SIAM, Philadelphia, PA, 2001.

[15] A. Ben-Tal and M. Teboulle, *Hidden convexity in some nonconvex quadratically constrained quadratic programming*, Math. Programming, 72 (1996), pp. 51–63.

[16] A. Berman, *Cones, matrices and mathematical programming*, vol. 79 of Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1973.

[17] D. Bertsimas and I. Popescu, *On the relation between option and stock prices: a convex optimization approach*, Oper. Res., 50 (2002), pp. 358–374.

[18] D. Bertsimas and J. Sethuraman, *Moment problems and semidefinite optimization*, in Handbook of semidefinite programming, vol. 27 of Internat. Ser. Oper. Res. Management Sci., Kluwer Acad. Publ., Boston, MA, 2000, ch. 16, pp. 469–509.

[19] R. R. Bitmead, M. R. Gevers, I. R. Petersen, and R. J. Kaye, *Monotonicity and stabilizability properties of solutions of the Riccati difference equation: propositions, lemmas, theorems, fallacious conjectures and counterexamples*, Systems Control Lett., 5 (1985), pp. 309–315.

[20] S. Bittanti, A. J. Laub, and J. C. Willems, eds., *The Riccati equation*, Communications and Control Engineering Series, Springer-Verlag, Berlin, 1991.

[21] M. Bqcher, *Introduction to higher algebra*, Dover Publications Inc., New York, 1964.

[22] E. G. Boman, *Infeasibility and negative curvature in optimization*, PhD thesis, Stanford University, Feb. 1999.

[23] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory*, vol. 15 of Studies in Applied Mathematics, SIAM, Philadelphia, PA, 1994.

[24] R. Byers, *A bisection method for measuring the distance of a stable matrix to the unstable matrices*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 875–881.

[25] C. I. Byrnes, S. V. Gusev, and A. Lindquist, *A convex optimization approach to the rational covariance extension problem*, SIAM J. Control Optim., 37 (1998), pp. 211–229.

[26] ——, *From finite covariance windows to modeling filters: a convex optimization approach*, SIAM Rev., 43 (2001), pp. 645–675.

[27] C. I. Byrnes, A. Lindquist, S. V. Gusev, and A. S. Matveev, *A complete parameterization of all positive rational extensions of a covariance sequence*, IEEE Trans. Automat. Control, 40 (1995), pp. 1841–1857.

[28] G.-n. Chen and Y.-j. Hu, *The truncated Hamburger matrix moment problems in the nondegenerate and degenerate cases, and matrix continued fractions*, Linear Algebra Appl., 277 (1998), pp. 199–236.

[29] ——, *A unified treatment for the matrix Stieltjes moment problem in both non-degenerate and degenerate cases*, J. Math. Anal. Appl., 254 (2001), pp. 23–34.

[30] E. W. Cheney and A. A. Goldstein, *Newton's method for convex programming and Tchebycheff approximation.*, Numer. Math., 1 (1959), pp. 253–268.

[31] S. H. Cheng and N. J. Higham, *The nearest definite pair for the Hermitian generalized eigenvalue problem*, Linear Algebra Appl., 302/303 (1999), pp. 63–76.

[32] Y. S. Chow and H. Teicher, *Probability theory: Independence, interchangeability, martingales*, Springer Texts in Statistics, Springer-Verlag, New York, third ed., 1997.

[33] V. Chvátal, *Linear programming*, A Series of Books in the Mathematical Sciences, W. H. Freeman, New York, 1983.

[34] J. H. Conway and N. J. A. Sloane, *Sphere packings, lattices and groups*, vol. 290 of Grundlehren der Mathematischen Wissenschaften, Springer-Verlag, New York, 1988.

[35] C. R. Crawford and Y. S. Moon, *Finding a positive definite linear combination of two Hermitian matrices*, Linear Algebra Appl., 51 (1983), pp. 37–48.

[36] G. Cybenko and C. Van Loan, *Computing the minimum eigenvalue of a symmetric positive definite Toeplitz matrix*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 123–131.

[37] T. N. Davidson, Z.-Q. Luo, and J. F. Sturm, *Linear matrix inequality formulation of spectral mask constraints with applications to FIR filter design*, IEEE Trans. Signal Process., 50 (2002), pp. 2702–2715.

[38] T. N. Davidson, Z.-Q. Luo, and K. M. Wong, *Design of orthogonal pulse shapes for communications via semidefinite programming*, IEEE Trans. Signal Process., 48 (2000), pp. 1433–1445.

[39] E. de Klerk and D. V. Pasechnik, *Approximation of the stability number of a graph via copositive programming*, SIAM J. Optim., 12 (2002), pp. 875–892.

[40] C. E. de Souza, M. R. Gevers, and G. C. Goodwin, *Riccati equations in optimal filtering of nonstabilizable systems having singular state transition matrices*, IEEE Trans. Automat. Control, 31 (1986), pp. 831–838.

[41] P. Delsarte, *An algebraic approach to the association schemes of coding theory*, Philips Res. Rep. Suppl., (1973), pp. vi+97.

[42] P. Delsarte, Y. Genin, and Y. Kamp, *Orthogonal polynomial matrices on the unit circle*, IEEE Trans. Circuits and Systems, 25 (1978), pp. 149–160.

[43] ——, *Schur parametrization of positive definite block-Toeplitz systems*, SIAM Journal on Applied Mathematics, 36 (1979), pp. 34–46.

[44] P. Delsarte, J. M. Goethals, and J. J. Seidel, *Spherical codes and designs*, Geometriae Dedicata, 6 (1977), pp. 363–388.

[45] P. Delsarte and V. I. Levenshtein, *Association schemes and coding theory*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2477–2504.

[46] H. Dette and W. J. Studden, *Matrix measures, moment spaces and Favard's theorem for the interval* $[0,1]$ *and* $[0,\infty)$, Linear Algebra Appl., 345 (2002), pp. 169–193.

[47] M. A. Dritschel, S. McCullough, and H. J. Woerdeman, *Model theory for ρ-contractions, $\rho \leq 2$*, J. Operator Theory, 41 (1999), pp. 321–350.

[48] O. du Merle and J.-P. Vial, *Proximal ACCPM, a cutting plane method for column generation and lagrangian relaxation: Application to the p-median problem*, tech. report, Université de Genève, 2002. Cahiers de recherche HEC, 2002.23.

[49] B. Dumitrescu, I. Tăbuş, and P. Stoica, *On the parameterization of positive real sequences and MA parameter estimation*, IEEE Trans. Signal Process., 49 (2001), pp. 2630–2639.

[50] M. K. H. Fan and A. L. Tits, *m-form numerical range and the computation of the structured singular value*, IEEE Trans. Automat. Control, 33 (1988), pp. 284–289.

[51] L. Fejér, *Über trigonometrische Polynome*, J. Reine Angew. Math., (1915), pp. 53–82.

[52] J.-P. Gabardo, *Trigonometric moment problems for arbitrary finite subsets of $\mathbb{Z}^n$*, Trans. Amer. Math. Soc., 350 (1998), pp. 4473–4498.

[53] W. Gander, G. H. Golub, and U. von Matt, *A constrained eigenvalue problem*, Linear Algebra Appl., 114–115 (1989), pp. 815–839.

[54] M. R. Garey and D. S. Johnson, *Computers and intractability: a guide to the theory of NP-completeness*, W. H. Freeman, San Francisco, Calif., 1979.

[55] L. Gemignani, *Schur complements of Bezoutians and the inversion of block Hankel and block Toeplitz matrices*, Linear Algebra Appl., 253 (1997), pp. 39–59.

[56] Y. Genin, Y. Hachez, Y. Nesterov, R. Stefan, P. Van Dooren, and S. Xu, *Positivity and linear matrix inequalities*, Eur. J. Control, 8 (2002), pp. 275–298.

[57] Y. Genin, Y. Hachez, Y. Nesterov, and P. Van Dooren, *Convex optimization over positive polynomials and filter design*, in Proceedings UKACC International Conference on Control 2000, 2000. CD-ROM Paper SS-41, University of Cambridge, UK, 4-7 September 2000.

[58] ——, *Optimization problems over positive pseudopolynomial matrices*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 57–79.

[59] Y. Genin, P. Van Dooren, and V. Vermaut, *Convergence of the calculation of H_∞ norms and related questions*, in Proceedings of the International Symposium Mathematical Theory of Networks and Systems, Padova, Italy, July 1998, pp. 429–432.

[60] M. X. Goemans, *Semidefinite programming in combinatorial optimization*, Math. Programming, 79 (1997), pp. 143–161.

[61] ——, *Semidefinite programming and combinatorial optimization*, Doc. Math., Extra Vol. III (1998), pp. 657–666.

[62] M. X. Goemans and D. P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. Assoc. Comput. Mach., 42 (1995), pp. 1115–1145.

[63] J. Goffin, A. Haurie, and J. Vial, *Decomposition and nondifferentiable optimization with the projective algorithm*, Manage. Sci., 38 (1992), pp. 284–302.

[64] J.-L. Goffin, Z.-Q. Luo, and Y. Ye, *Complexity analysis of an interior cutting plane method for convex feasibility problems*, SIAM J. Optim., 6 (1996), pp. 638–652.

[65] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix polynomials*, Computer Science and Applied Mathematics, Academic Press Inc., New York, 1982.

[66] G. H. Golub and C. F. Van Loan, *Matrix computations*, Johns Hopkins Studies in the Mathematical Sciences, The Johns Hopkins University Press, Baltimore, MD, third ed., 1996.

[67] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric algorithms and combinatorial optimization*, vol. 2 of Algorithms and Combinatorics, Springer-Verlag, Berlin, second ed., 1993.

[68] K. E. Gustafson and D. K. M. Rao, *Numerical range: The field of values of linear operators and matrices*, Universitext, Springer-Verlag, New York, 1997.

[69] Y. Hachez and Y. Nesterov, *Optimization problems over nonnegative polynomials with interpolation constraints*, in Proceedings 15th IFAC World Congress, July 2002. CD-ROM Paper 1650, Barcelona, ES, 21-26 July 2002.

[70] Y. Hachez and P. Van Dooren, *Elliptic and hyperbolic quadratic eigenvalue problems and associated distance problems*, Linear Algebra Appl., (2003). DOI 10.1016/S0024-3795(03)00489-0 (in press).

[71] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, second ed., 1952.

[72] F. Hausdorff, *Der Wertvorrat einer Bilinearform*, Math. Z., 3 (1919), pp. 314–316.

[73] G. Heinig and K. Rost, *Algebraic methods for Toeplitz-like matrices and operators*, vol. 19 of Mathematical Research, Akademie-Verlag, Berlin, 1984.

[74] R. Hettich and K. O. Kortanek, *Semi-infinite programming: theory, methods, and applications*, SIAM Rev., 35 (1993), pp. 380–429.

[75] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, second ed., 2002.

[76] N. J. Higham, F. Tisseur, and P. M. Van Dooren, *Detecting a definite Hermitian pair and a hyperbolic or elliptic quadratic eigenvalue problem, and associated nearness problems*, Linear Algebra Appl., 351–352 (2002), pp. 455–474.

[77] D. Hilbert, *Über die Darstellung definiter Formen als Summe von Formenquadraten*, Math. Ann., 32 (1888), pp. 342–350.

[78] ——, *Über ternäre definite Formen*, Acta Math., 17 (1893), pp. 169–197.

[79] ——, *Grundlagen der Geometrie*, B.G. Teubner, 1899.

[80] R. Hildebrand and M. Gevers, *Identification for control: optimal input design with respect to a worst-case ν-gap cost function*, SIAM J. Control Optim., 41 (2003), pp. 1586–1608.

[81] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, 1985.

[82] ——, *Topics in matrix analysis*, Cambridge University Press, Cambridge, 1991.

[83] I. S. Iohvidov, *Hankel and Toeplitz matrices and forms: Algebraic theory*, Birkhäuser Boston, Mass., 1982.

[84] T. Kailath, *Linear systems*, Prentice-Hall Information and System Sciences Series, Prentice-Hall Inc., Englewood Cliffs, NJ, 1980.

[85] T. Kailath and A. H. Sayed, *Displacement structure: theory and applications*, SIAM Rev., 37 (1995), pp. 297–386.

[86] ——, eds., *Fast reliable algorithms for matrices with structure*, SIAM, Philadelphia, PA, 1999.

[87] R. Kalman, *Lyapunov functions for the problem of Lur'e in automatic control*, Proc. Nat. Acad. Sci. USA, 49 (1963), pp. 201–205.

[88] C. Kao, U. Jonsson, and A. Megretski, *A cutting plane algorithm for robustness analysis of periodically time-varying systems*, IEEE Transactions on Automatic Control, 46 (2001), pp. 579–592.

[89] S. Karlin and W. J. Studden, *Tchebycheff systems: With applications in analysis and statistics*, vol. 15 of Pure and Applied Mathematics, Interscience Publishers John Wiley & Sons, New York-London-Sydney, 1966.

[90] J. E. Kelley, Jr., *The cutting-plane method for solving convex programs*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 703–712.

[91] M. G. Kreĭn and A. A. Nudel'man, *The Markov moment problem and extremal problems*, vol. 50 of Translations of Mathematical Monographs, American Mathematical Society, Providence, R.I., 1977.

[92] P. Lancaster and L. Rodman, *Algebraic Riccati equations*, Oxford Science Publications, The Clarendon Press Oxford University Press, New York, 1995.

[93] J. B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.

[94] ——, *An explicit equivalent positive semidefinite program for nonlinear* 0-1 *programs*, SIAM J. Optim., 12 (2002), pp. 756–769.

[95] ——, *Semidefinite programming vs. LP relaxations for polynomial programming*, Math. Oper. Res., 27 (2002), pp. 347–360.

[96] A. J. Levin, *An algorithm for minimizing convex functions*, Dokl. Akad. Nauk SSSR, 160 (1965), pp. 1244–1247.

[97] C.-K. Li and Y.-T. Poon, *Convexity of the joint numerical range*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 668–678.

[98] F. Lukács, *Verschärfung des ersten Mittelwertsatzes der Integralrechnung für rationale Polynome*, Math. Z., 2 (1918), pp. 295–305.

[99] W. Mackens and H. Voss, *A projection method for computing the minimum eigenvalue of a symmetric positive definite Toeplitz matrix*, Linear Algebra Appl., 275–276 (1998), pp. 401–415.

[100] ——, *Computing the minimum eigenvalue of a symmetric positive definite Toeplitz matrix by Newton-type methods*, SIAM J. Sci. Comput., 21 (2000), pp. 1650–1656.

[101] M. Marcus and H. Minc, *A survey of matrix theory and matrix inequalities*, Allyn and Bacon Inc., Boston, Mass., 1964.

[102] N. Mastronardi and D. Boley, *Computing the smallest eigenpair of a symmetric positive definite Toeplitz matrix*, SIAM J. Sci. Comput., 20 (1999), pp. 1921–1927.

[103] J. W. McLean and H. J. Woerdeman, *Spectral factorizations and sums of squares representations via semidefinite programming*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 646–655.

[104] T. S. Motzkin, *Selected papers*, Contemporary Mathematicians, Birkhäuser Boston, Mass., 1983.

[105] A. S. Nemirovskii, C. Roos, and T. Terlaky, *On maximization of quadratic form over intersection of ellipsoids with common center*, Math. Programming, 86 (1999), pp. 463–476.

[106] A. S. Nemirovsky and D. B. Yudin, *Problem complexity and method efficiency in optimization*, Wiley-Interscience Series in Discrete Mathematics, John Wiley & Sons Inc., New York, 1983.

[107] Y. Nesterov, *Complexity estimates of some cutting plane methods based on the analytic barrier*, Math. Programming, 69 (1995), pp. 149–176.

[108] ——, *Introductory lectures on convex programming.* Unpublished lecture notes, 1996.

[109] ——, *Long-step strategies in interior-point primal-dual methods*, Math. Programming, 76 (1997), pp. 47–94.

[110] ——, *Semidefinite relaxation and nonconvex quadratic optimization*, Optim. Methods Softw., 9 (1998), pp. 141–160.

[111] ——, *Squared functional systems and optimization problems*, in High performance optimization, H.Frenk et al., eds., vol. 33 of Appl. Optim., Kluwer Acad. Publ., Dordrecht, 2000, ch. 17, pp. 405–440.

[112] Y. Nesterov and A. Nemirovskii, *Interior-point polynomial algorithms in convex programming*, vol. 13 of Studies in Applied Mathematics, SIAM, Philadelphia, PA, 1994.

[113] Y. Nesterov, H. Wolkowicz, and Y. Ye, *Semidefinite programming relaxations of nonconvex quadratic optimization*, in Handbook of semidefinite programming, vol. 27 of Internat. Ser. Oper. Res. Management Sci., Kluwer Acad. Publ., Boston, MA, 2000, ch. 13, pp. 361–419.

[114] D. J. Newman, *Location of the maximum on unimodal surfaces*, J. Assoc. Comput. Mach., 12 (1965), pp. 395–398.

[115] C. Oara and A. Varga, *Computation of general inner-outer and spectral factorization*, IEEE Trans. Automat. Control, 45 (2000), pp. 2307–2325.

[116] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*, Prentice Hall Signal Processing Series, Prentice Hall, Englewood Cliffs, NJ, 1989.

[117] V. Y. Pan, Y. Rami, and X. Wang, *Structured matrices and Newton's iteration: unified approach*, Linear Algebra Appl., 343/344 (2002), pp. 233–265.

[118] P. A. Parrilo, *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*, PhD thesis, California Institute of Technology, May 2000.

[119] ——, *Semidefinite programming relaxations for semialgebraic problems*, Math. Programming, 96 (2003), pp. 293–320.

[120] G. Pólya, *Über positive Darstellung von Polynomen*, Vierteljahresschr. Naturforsch. Ges. Zürich, 73 (1928), pp. 141–145.

[121] G. Pólya and G. Szegő, *Problems and theorems in analysis. Vol. II: Theory of functions, zeros, polynomials, determinants, number theory, geometry*, vol. 216 of Die Grundlehren der Mathematischen Wissenschaften, Springer-Verlag, New York, 1976.

[122] B. T. Polyak, *Convexity of quadratic transformations and its use in control and optimization*, J. Optim. Theory Appl., 99 (1998), pp. 553–583.

[123] V.-M. Popov, *Hyperstability of control systems*, vol. 204 of Die Grundlehren der mathematischen Wissenschaften, Springer Verlag, Berlin, 1973.

[124] E. Y. Remez, *General computation methods for Čebyšev approximation. Problems with real parameters entering linearly*, Izdat. Akad. Nauk Ukrainsk. SSR. Kiev, 1957. (Russian).

[125] B. Reznick, *Some concrete aspects of Hilbert's 17th Problem*, in Real algebraic geometry and ordered structures (Baton Rouge, LA, 1996), vol. 253 of Contemp. Math., Amer. Math. Soc., Providence, RI, 2000, pp. 251–272.

[126] R. T. Rockafellar, *Convex analysis*, vol. 28 of Princeton Mathematical Series, Princeton University Press, Princeton, NJ, 1970.

[127] C. Roos, T. Terlaky, and J.-P. Vial, *Theory and algorithms for linear optimization: an interior point approach*, Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons Ltd., Chichester, 1997.

[128] M. Rosenblum and J. Rovnyak, *Hardy classes and operator theory*, Oxford Mathematical Monographs, Oxford University Press, New York, 1985.

[129] A. H. Sayed and T. Kailath, *A survey of spectral factorization methods*, Numer. Linear Algebra Appl., 8 (2001), pp. 467–496.

[130] J. Sreedhar, P. Van Dooren, and A. L. Tits, *A fast algorithm to compute the real structured stability radius*, in Stability theory (Ascona, 1995), vol. 121 of Internat. Ser. Numer. Math., Birkhäuser, Basel, 1996, pp. 219–230.

[131] S. P. Tarasov, L. G. Khachiyan, and I. I. Èrlikh, *The method of inscribed ellipsoids*, Dokl. Akad. Nauk SSSR, 298 (1988), pp. 1081–1085.

[132] M. Tismenetsky, *Factorizations of Hermitian block Hankel matrices*, Linear Algebra Appl., 166 (1992), pp. 45–63.

[133] ———, *Matrix generalizations of a moment problem theorem. I. the Hermitian case*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 92–112.

[134] F. Tisseur and K. Meerbergen, *The quadratic eigenvalue problem*, SIAM Review, 43 (2001), pp. 235–286.

[135] O. Toeplitz, *Das algebraische Analogon zu einem Satze von Féjer*, Math. Z., 2 (1918), pp. 187–197.

[136] K.-C. Toh, *Some new search directions for primal-dual interior point methods in semidefinite programming*, SIAM J. Optim., 11 (2000), pp. 223–242.

[137] K. C. Toh, M. J. Todd, and R. H. Tütüncü, *SDPT3—a MATLAB software package for semidefinite programming, version 1.3*, Optim. Methods Softw., 11/12 (1999), pp. 545–581.

[138] R. H. Tütüncü, K. C. Toh, and M. J. Todd, *Solving semidefinite-quadratic-linear programs using SDPT3*, Math. Programming, 95 (2003), pp. 189–217.

[139] E. E. Tyrtyshnikov, *How bad are Hankel matrices?*, Numer. Math., 67 (1994), pp. 261–269.

[140] P. M. Vaidya, *A new algorithm for minimizing convex functions over convex sets*, Math. Programming, 73 (1996), pp. 291–341.

[141] M. Van Barel, *private communication*.

[142] P. Van Dooren, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Control, 26 (1981), pp. 111–129.

[143] ——, *Numerical aspects of system and control algorithms*, J. A, 30 (1989), pp. 25–32.

[144] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.

[145] ——, *Connections between semi-infinite and semidefinite programming*, in Semi-infinite programming, vol. 25 of Nonconvex Optim. Appl., Kluwer Acad. Publ., Boston, MA, 1998, pp. 277–294.

[146] J.-P. Vial and The NewAccpm Team, *Generalized analytic center cutting plane method*, tech. report, Université de Genève, 2002. (manuscript).

[147] H. Voss, *Symmetric schemes for computing the minimum eigenvalue of a symmetric Toeplitz matrix*, Linear Algebra Appl., 287 (1999), pp. 359–371.

[148] ——, *A symmetry exploiting Lanczos method for symmetric Toeplitz matrices*, Numer. Algorithms, 25 (2000), pp. 377–385.

[149] J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automatic Control, 16 (1971), pp. 621–634.

[150] H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., *Handbook of semidefinite programming*, vol. 27 of Internat. Ser. Oper. Res. Management Sci., Kluwer Academic Publishers, Boston, MA, 2000.

[151] S.-P. Wu, S. Boyd, and L. Vandenberghe, *FIR filter design via spectral factorization and convex optimization*, vol. 1 of Appl. Comput. Control Signals Circuits, Birkhäuser Boston, Boston, MA, 1999, pp. 215–245.

[152] V. Yakubovich, *Solution of certain matrix inequalities in the stability theory of nonlinear control systems*, Dokl. Akad. Nauk. SSSR, 143 (1962), pp. 1304–1307.

[153] Y. Ye, *Approximating quadratic programming with bound and quadratic constraints*, Math. Programming, 84 (1999), pp. 219–226.

[154] D. C. Youla, *On the factorization of rational matrices*, IRE Trans. Information Theory, IT-7 (1961), pp. 172–189.

[155] K. Zhou, J. C. Doyle, and K. Glover, *Robust and optimal control*, Prentice Hall, Upper Saddle River, NJ, 1996.