

NONSTANDARD DUALITY IN OPTIMIZATION

NONSTANDARD DUALITY CONCEPTS IN CONIC
AND QUADRATIC OPTIMIZATION

by
IMRE PÓLIK, M.Sc.

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Doctor of Philosophy

McMaster University
© Copyright by Imre Pólik, 2007

DOCTOR OF PHILOSOPHY (2007)
(Mathematics)

McMaster University
Hamilton, Ontario

TITLE: Nonstandard duality concepts in conic and quadratic optimization

AUTHOR: Imre Pólik, M.Sc.

SUPERVISOR: Dr. Tamás Terlaky

NUMBER OF PAGES: xvii, 138

Abstract

This thesis is centred around the topic of duality. It presents the classical duality theories in optimization and identifies their key ingredients as convexity and constraint qualification. The thesis answers questions on what we can salvage from the theories if these conditions fail to hold.

First we present a special duality theory for systems of quadratic functions without assuming convexity. The results stem from the S-lemma, a fundamental result in control theory. We show several proofs for this basic results, then investigate the theory behind each of them. New theorems about nonconvex quadratic systems are also proved.

Then we look at relaxing the other key factor in duality theories: the constraint qualification or regularity condition. We present a strong duality theory for optimization over symmetric cones without assuming any constraint qualification. We show that the dual problem is defined over a homogeneous cone and we analyze its complexity. Specializing the dual to semidefinite optimization the result improves the best currently known complexity bound.

Finally, we show how the classical theory changes if the computations are carried out in finite precision arithmetic. This is an important issue in practice since duality theorems provide the theoretical foundation for stopping criteria and optimality conditions in optimization software. Along with a brief overview of the existing results we present a new set of stopping criteria for convex conic optimization. We prove that the criteria do not change the complexity of the algorithm they are applied with.

A long list of open problems and topics for future research conclude the thesis.

Acknowledgements

Although this thesis is the final result of my five-year study at McMaster University, its roots go back a long time before that. I was very fortunate throughout my life to have the right people around me, people who provided me what I needed at the moment when I needed it.

My father, an engineer himself, planted the seed of mathematics in me and raised me hoping that one day I would understand all his mathematical books (including the ones he never quite mastered.) My mother provided all the care and understanding. They, together with my two brothers and my sister gave me all the support that was humanly possible, which I am greatly indebted for.

I was very fortunate to have great teachers at all levels of my studies. At the end of elementary school Sándor Kiss and his special math circle assured me about my goal to become a mathematician. During high school Ákos Binzberger OSB opened up the horizon towards the university, where András Lőrincz and his Neural Information Processing Group introduced me to the world of research. It was Tibor Illés, my MSc supervisor, who introduced me to optimization and interior point methods, and was very influential in arranging my “trip” to McMaster. My PhD supervisor, Tamás Terlaky kept my mind focused on my projects from the very first day I met him. He provided me all the guidance that I could ask for. His optimization seminars brought the greatest people to our group, who on many occasions moved me forward: the right speaker with the right talk at the right time. I am also grateful to my friends at the Advanced Optimization Lab. Their company made my five years at McMaster very enjoyable.

It was also partly due to mathematics that I met my wife, Andi. She has always kept an interest in my work and encouraged me throughout the past 5 years. Her love and support helped me to be where I am now.

The comments and suggestions of Chek Beng Chua, Etienne de Klerk, Jean-Baptiste Hiriart-Urruty, Michael Overton and Dima Pasechnik are greatly appreciated, but most importantly, discussions with Gábor Pataki had the most significant impact on my work. Special thanks also to the members of my PhD supervisory committee: Tom Hurd, Nicholas Kevlahan and Jiming Peng.

Finally, I would like to thank NSERC (Discovery Grant #5-48923) and MITACS for their financial support during the course of my studies.

Contents

List of Figures	ix
List of Definitions	xi
List of Lemmas, Propositions and Theorems	xiii
Notation and symbols	xv
Abbreviations	xvii
1 Introduction	1
1.1 The structure of the thesis	2
2 Background and motivation	3
2.1 Basic concepts in convex analysis	3
2.2 Lagrange duality	6
2.2.1 The classical theory	6
2.2.2 Lagrange-Wolfe duality	10
2.3 Fenchel duality	10
2.3.1 The Fenchel dual as a Lagrange dual	10
2.4 Special cases	13
2.4.1 Linear optimization	13
2.4.2 Quadratic optimization	14
2.4.3 Conic optimization	15
2.5 Theorems of the alternative, optimality conditions and duality	18
2.6 Historical remarks	18
2.6.1 The history of optimization, duality and convexity . . .	19
2.6.2 Literature	20
2.7 Motivation: convexity+CQ=duality?	20
2.7.1 A nonconvex system	20
2.7.2 A nonregular system	21

2.7.3	Nonexact duality	22
3	Duality for nonconvex quadratic systems	25
3.1	Introduction	26
3.1.1	Motivation	26
3.1.2	Historical background	28
3.1.3	The structure of this chapter	29
3.2	Proofs for the basic S-lemma	30
3.2.1	The two faces of the S-lemma	30
3.2.2	The traditional approach	31
3.2.3	A modern approach	34
3.2.4	A new elementary proof	38
3.3	Special results and counterexamples	40
3.3.1	Other variants	40
3.3.2	General results	41
3.3.3	Counterexamples	45
3.4	Applications	49
3.4.1	Stability analysis	49
3.4.2	Sum of two ellipsoids	51
3.5	Convexity of the joint numerical range	52
3.5.1	Motivation	52
3.5.2	Theoretical results	54
3.5.3	Implications	60
3.6	Rank-constrained LMI	64
3.6.1	Motivation	64
3.6.2	Theoretical results	65
3.6.3	Implications	67
3.6.4	Higher rank solutions	69
3.6.5	Rank constraints and convexity	71
3.7	Generalized convexities	73
3.7.1	Motivation	73
3.7.2	Theoretical results	73
3.7.3	Implications	76
3.8	Miscellaneous topics	78
3.8.1	Trust region problems	78
3.8.2	SDP relaxation of quadratically constrained quadratic programs (QCQP)	79
3.8.3	Algebraic geometry	80
3.8.4	Computational complexity	82
3.9	New duality theorems	82

3.9.1	Polyhedral case	83
3.9.2	Conic case	84
3.10	Summary	85
4	Nonregular duality for symmetric cones	87
4.1	Introduction	87
4.1.1	Historical background	87
4.1.2	The structure of this chapter	88
4.1.3	Preliminaries	88
4.2	The minimal cone and the facial reduction algorithm	89
4.3	The geometry of homogeneous cones	93
4.3.1	Definition, basic properties	93
4.3.2	Constructing homogeneous cones from T-algebras	93
4.3.3	Recursive construction of homogeneous cones	97
4.3.4	Another representation for homogeneous cones	98
4.3.5	Self dual homogeneous (symmetric) cones	99
4.4	An exact duality theory for symmetric cones	99
4.4.1	The facial reduction algorithm for symmetric cones	100
4.4.2	Complexity of the dual problem	105
4.5	Special cases	106
4.5.1	Semidefinite optimization	106
4.5.2	Second order conic optimization	107
4.6	Summary	108
5	Nonexact duality for conic optimization	109
5.1	Introduction	109
5.2	An approximate Farkas theorem	110
5.3	Stopping criteria for self-dual models	114
5.3.1	Homogeneous self-dual model for conic optimization	114
5.3.2	New stopping criteria	115
5.3.3	Complexity of the criteria	117
5.4	Practical considerations	119
5.4.1	How big is big enough?	119
5.4.2	Handling weak infeasibility	119
5.5	Summary	120
6	Conclusions and further directions	121
6.1	Contributions	121
6.1.1	S-lemma	121
6.1.2	Nonconvex duality	122

6.1.3	Nonregular duality	122
6.1.4	Approximate duality	122
6.2	Problems for future research	122
6.2.1	Nonconvex systems	122
6.2.2	Nonregular systems	124
6.2.3	Approximate duality	125
Bibliography		127

List of Figures

2.1	A nonconvex objective function	21
2.2	Numerically sensitive linear problems	23
3.1	The quadratic regions defined by system (3.3.16) with $x_3 = 1$.	46
3.2	Separating a convex cone from a nonconvex set	76

List of Definitions

Convex combination	3
Convex set	3
Convex hull	4
Extreme point	4
Cone	4
Pointed/solid cone	4
Ordering defined by a cone	4
Extreme direction	4
Dual cone	5
Self-dual cone	5
Irreducible cone	5
Convex and concave functions	6
Slater point	6
Singular/regular constraints	7
Saddle point	8
Convex conjugate functions	10
k^{th} joint numerical range	58
General convex functions	73
General convex-linear functions	74
Nice cone	90
Homogeneous cones	93
T-algebra	95

Homogeneous symmetric form on a cone	97
Siegel cone	97

List of Theorems

Existence of extreme points	4
Existence of extreme directions	4
Separation of convex sets	5
Properties of the dual cone	5
Self-dual cones	5
Farkas theorem	7
Farkas lemma	8
Karush-Kuhn-Tucker theorem	8
Weak duality for NLP	9
Strong duality for NLP	9
Fenchel duality	12
Fenchel duality (dual form)	13
Duality for LP	14
Duality for QP	15
Duality for conic optimization	17
S-lemma	31
Quadratic systems – nonnegative eigenvalues	42
Circle criterion	50
Sum of two ellipsoids	52
Convexity of the 2D quadratic image of \mathbb{R}^n	54
Convexity of the 3D quadratic image of \mathbb{R}^n	54
ICON maps	54

LICON maps	55
Convexity of the 2D quadratic image of the ball	56
Convexity of the n D quadratic image of a ball	57
Nonconvexity of the quadratic image	58
Convexity of the joint numerical range	59
Nonconvexity, complex case	59
S-lemma with a norm constraint	60
S-lemma, local version	61
S-lemma with a LICON map	63
S-lemma, complex case	64
Low rank semidefinite matrices	66
Lower rank semidefinite matrices	67
Quadratic systems with multiple terms	69
Duality for König convex functions	74
Duality for K -convexlike functions	75
Duality for Ky Fan convex-linear functions	75
Nullstellensatz, complex case	80
Nullstellensatz, real case	81
Effective Nullstellensatz	81
New duality theorem (polyhedral case)	83
New duality theorem (conic case)	84
Facial reduction algorithm for nice cones	91
Representation of homogeneous cones	95
Properties of the Siegel cone	97
The description of $((\mathcal{F}(\mathcal{C}))^c)^\perp$	102
Approximate Farkas theorem for conic problems	111
Properties of the HSD model	114
Complexity of the IPM with (R1) and (R2)	117
Complexity of the IPM with (R1) and (R3)	118

Notation and symbols

Basic objects

A, B, \dots	matrices
A_{ij}	the j^{th} element of the i^{th} row of A
u, v, \dots	column vectors
u_i	the i^{th} component of v
u^i	the i^{th} vector in a list
i, j, \dots	indices
\mathbf{i}	the imaginary unit, $\mathbf{i}^2 = -1$
$u_{i:j}$	the vector (u_i, \dots, u_j)
$A_{:,i}$	the i^{th} column of A
$A_{i,:}$	the i^{th} row of A
I	the identity matrix
e	the unit element of a T-algebra
f, g	functions
\mathcal{K}	a (usually convex) cone
$\mathcal{C}, \mathcal{G}, \mathcal{H}, \dots$	sets

Sets

\mathbb{R}^n	the real n -dimensional vector space
\mathbb{C}^n	the complex n -dimensional vector space
\mathbb{R}_+^n	the set of n -dimensional vectors with nonnegative components
\mathbb{R}_-^n	the set of n -dimensional vectors with nonpositive components
\mathbb{S}^n	the space of $n \times n$ real symmetric matrices
\mathbb{PS}^n	the cone of real symmetric positive semidefinite matrices
\mathbb{L}	the Lorentz cone, $\mathbb{L} = \{x \in \mathbb{R}^n : x_1 \geq \ x_{2:n}\ _2\}$
\mathbb{L}_r	the rotated Lorentz cone, $\mathbb{L}_r = \{x \in \mathbb{R}^n : x_1 x_2 \geq \ x_{3:n}\ _2\}$
\mathcal{F}	a face of a convex cone

\mathcal{A}	a T-algebra
\mathcal{T}	upper triangular elements in a T-algebra
\mathcal{I}	upper triangular elements with positive diagonal in a T-algebra
\mathcal{H}	Hermitian ($u = u^*$) elements in a T-algebra
\mathcal{K}_{\min}	the minimal cone of an optimization problem defined over \mathcal{K} $\mathcal{K}_{\min} = \mathcal{F}(\{c - A^T y : y \in \mathbb{R}^m\} \cap \mathcal{K})$

Relations

$x \geq 0$	the components of x are all nonnegative
$X \succeq 0$	X is symmetric positive semidefinite
$X \succ 0$	X is symmetric positive definite
$x \succeq_{\mathcal{K}} 0$	$x \in \mathcal{K}$
$x \succ_{\mathcal{K}} 0$	$x \in \text{int}(\mathcal{K})$
$\mathcal{F} \trianglelefteq \mathcal{K}$	\mathcal{F} is a face of \mathcal{K}

Operators, functions

$\langle u, v \rangle$	dot product of u and v , in special cases $u^T v$ is also used
$A \bullet B$	the scalar product of two matrices, $\text{Tr}(A^T B)$
\mathcal{K}^*	the dual cone of \mathcal{K}
$\text{int}(\mathcal{C})$	the interior of \mathcal{C}
$\text{rel int}(\mathcal{C})$	the relative interior of \mathcal{C}
$\text{cl}(\mathcal{C})$	the closure of \mathcal{C}
$\partial(\mathcal{C})$	the boundary of \mathcal{C}
$\text{span}(\mathcal{C})$	the subspace spanned by \mathcal{C}
$\text{conv}(\mathcal{C})$	the convex hull of \mathcal{C}
$\text{Ker}(A)$	the null space or kernel of A
V^\perp	the orthogonal complement of V $V^\perp = \{u \in \mathbb{R}^n : \langle u, v \rangle = 0, \forall v \in V\}$
$\mathcal{F}(\mathcal{C})$	the face generated by \mathcal{C}
\mathcal{F}^c	the complementary or conjugate face of \mathcal{F}
$\text{Aut}(\mathcal{K})$	the group of automorphisms of a cone
$\text{Tr}(X)$	the trace of a matrix
$\text{tr}(x)$	the trace of an element in a T-algebra
$\text{SC}(\mathcal{K}, B)$	the Siegel cone of \mathcal{K} with the bilinear form B
$(\cdot)^*$	an involution of a T-algebra
$\vartheta(\mathcal{K})$	the complexity parameter of \mathcal{K}

Abbreviations

CQ:	constraint qualification
HSD:	homogeneous self-dual
ICON:	image convex map
IPM:	interior point method
KKT:	Karush–Kuhn–Tucker
LICON:	linear image convex map
LMI:	linear matrix inequality
LP:	linear optimization
NLP:	nonlinear optimization
QCQP:	quadratically constrained quadratic optimization
QP:	quadratic optimization
SDP:	semidefinite optimization
SOCP:	second-order cone optimization
SOS:	sum of squares

Chapter 1

Introduction

Just as we have two eyes and two feet, duality is part of life.

CARLOS SANTANA

In everyday life duality refers to the fact that there are two of something. Two eyes, two feet, government and opposition, two genders, two sides in chess. They coexist, they are equivalent, but they are more than the same thing twice. (Imagine a team playing itself.) We need both parts for proper functionality even if they are only mirrored images or complete opposites of each other.

In physics one particular example of duality is the dual nature of light: wave and/or particle. Neither describes the behaviour of light in full detail, but together they form a powerful theory.

In mathematics the term duality refers to a certain symmetry of looking at the same problem in different ways. Beyond aesthetic beauty this symmetry has theoretical advantages: looking at one formulation we can draw useful conclusions about the other one, or combining the two we can solve the problem more efficiently. Both forms carry the full information about the problem, but together they become more powerful.

In optimization, duality has a special meaning. Vaguely speaking, two optimization problems – the primal and the dual problem – are referred to as being duals of each other if the solution of the primal problem has some implication on the solution of the dual problem, and there is some relation between the optimal values.

One simple example is the isoperimetric problem: find the body with the largest volume among the bodies with a given surface area. The dual problem is to find the body with the smallest surface area among the bodies

with a given volume.

In this thesis we present three nonstandard duality concepts in optimization. They relax the conditions of the classical duality theorems and establish the results under weaker conditions in special cases. The three concepts are convexity, regularity and accuracy.

1.1 The structure of the thesis

First, in Chapter 2 we present the classical duality theories in optimization, most importantly the Lagrange, the Lagrange-Wolfe and the Fenchel-Moreau dualities. The special cases of linear, quadratic and conic optimization are discussed in detail. As it turns out, the key ingredients for these results are convexity and constraint qualification. We show examples to illustrate how the theorems fail to hold in the absence of those conditions. These examples will motivate the rest of the thesis.

In Chapter 3 we present a special duality theory for quadratic functions, without assuming convexity. This chapter stems from the S-lemma, a fundamental result in control theory. We show three different proofs for this basic results, then investigate the theory behind each of them. Though the primary contribution of this chapter is the analysis of the different proof techniques and their interplay, some new theorems about nonconvex quadratic systems are also proved. Results in this chapter were first published in [102].

Chapter 4 looks at relaxing the other key factor in duality theories: the constraint qualification or regularity condition. We present a strong duality theory for optimization over symmetric cones without assuming any sort of constraint qualification. We show that the dual problem is defined over a homogeneous cone and we analyze its complexity. Specializing the dual to semidefinite optimization the result improves the best currently known complexity bound. This chapter is based on [103].

In Chapter 5 we show how the classical theory changes if the computations are carried out in finite precision arithmetic. This is an important issue in practice since duality theorems provide the theoretical foundation for stopping criteria and optimality conditions in optimization software. Along with a brief overview of the existing results we present a new set of stopping criteria for convex conic optimization. We prove that the new criteria do not change the complexity of the algorithm they are applied with. Most of the material in the chapter is from [104].

Finally, Chapter 6 summarizes the results of the thesis and presents some open questions for future research.

Chapter 2

Background and motivation

It is a capital mistake to theorize
before one has data.

SHERLOCK HOLMES

In this chapter we review the different duality concepts for optimization and identify their key components. The goal here is twofold. On one hand, to familiarize the reader with these concepts, on the other hand, to motivate the subsequent chapters.

2.1 Basic concepts in convex analysis

The following concepts from convex analysis play a crucial role in the thesis. This is not a concise review of convex analysis, we only discuss the topic at the level necessary to follow the rest of the thesis. Readers interested in the proofs can find the references at the end of §2.6.1. Familiarity with basic terminology of classical analysis is assumed.

The most common ingredient of duality theories is convexity:

Definition 2.1.1 (Convex combination). *Let V be a linear space, and consider $x^1, \dots, x^k \in V$. If $\lambda_1, \dots, \lambda_k \in [0, 1]$ with $\sum_{i=1}^k \lambda_i = 1$ then the sum $\sum_{i=1}^k \lambda_i x^i$ is called a convex combination of the points x^1, \dots, x^k .*

Definition 2.1.2 (Convex set). *Let V be a linear space, then a set $\mathcal{C} \subseteq V$ is convex if for every $x, y \in \mathcal{C}$ and $\lambda \in [0, 1]$ we have $\lambda x + (1 - \lambda)y \in \mathcal{C}$, i.e., the set contains all the line segments between its points. Equivalently, a convex set contains all the convex combinations of its points.*

Definition 2.1.3 (Convex hull). If $\mathcal{H} \subseteq V$ is any set then its convex hull (denoted by $\text{conv}(\mathcal{H})$) is the set containing all the possible convex combinations of points from \mathcal{H} , i.e.,

$$\text{conv}(\mathcal{H}) = \left\{ \sum_{i=1}^k \lambda_i x^i : x^1, \dots, x^k \in \mathcal{H}, \lambda_1, \dots, \lambda_k \in [0, 1], \sum_{i=1}^k \lambda_i = 1 \right\}. \quad (2.1.1)$$

Moreover, $\text{conv}(\mathcal{H})$ is the smallest convex set containing \mathcal{H} .

In analyzing the boundary structure of a convex set the following notion is very important.

Definition 2.1.4 (Extreme point). If \mathcal{C} is a convex set then $x \in \partial(\mathcal{C})$ is an extreme point if it is not an interior point of any line segment from \mathcal{C} .

Theorem 2.1.5 (Existence of extreme points). Every closed convex set not containing a line has at least one extreme point. Moreover, a convex compact set is the convex hull of its extreme points.

Cones play a special part in duality theories:

Definition 2.1.6 (Cone). A set $\mathcal{K} \subseteq V$ is a cone if for every $x \in \mathcal{K}$ and $\lambda \geq 0$ we have $\lambda x \in \mathcal{K}$.

Definition 2.1.7 (Pointed/solid cone). A cone \mathcal{K} is pointed if it does not contain a line, or in other words, the origin is an extreme point of \mathcal{K} . A cone is solid if its interior is nonempty.

Definition 2.1.8 (Ordering defined by a cone). Let \mathcal{K} be a closed, convex, pointed, solid cone. Given a vector $x \in \mathbb{R}^n$ we say that $x \succeq_{\mathcal{K}} 0$ if $x \in \mathcal{K}$. Further, if we have two vectors then $x \succeq_{\mathcal{K}} y \Leftrightarrow x - y \succeq_{\mathcal{K}} 0$. Similarly we write $x \succ_{\mathcal{K}} 0$ if $x \in \text{relint}(\mathcal{K})$. If \mathcal{K} is missing from the notation then it is assumed to be the cone of real symmetric positive semidefinite matrices.

For cones, we can define a notion similar to the extreme points of convex sets:

Definition 2.1.9 (Extreme direction). If \mathcal{K} is a convex cone then $x \in \mathcal{K}$ is an extreme direction if it is not a positive convex combination of other directions from \mathcal{K} .

Theorem 2.1.10 (Existence of extreme directions). Every pointed convex cone has at least one extreme direction.

The reason that convex sets are so important is that they can be separated with a hyperplane.

Theorem 2.1.11 (Separation of convex sets). *If $\mathcal{C} \subseteq \mathbb{R}^n$ is a convex set and $w \in \mathbb{R}^n \setminus \mathcal{C}$ then there exists a hyperplane separating w and \mathcal{C} , such that the hyperplane does not contain \mathcal{C} . More precisely there is an $a \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ such that $a^T u \geq \alpha$ for all $u \in \mathcal{C}$, $a^T w \leq \alpha$ and $\mathcal{C} \not\subseteq \{x : a^T x = \alpha\}$.*

For a convex cone we can define a dual cone. This will be an important point in the construction of the dual problem.

Definition 2.1.12 (Dual cone). *If $\mathcal{K} \subseteq \mathbb{R}^n$ is a cone then its dual cone is defined as*

$$\mathcal{K}^* = \{y \in \mathbb{R}^n : \langle x, y \rangle \geq 0\}. \quad (2.1.2)$$

Theorem 2.1.13 (Properties of the dual cone). *The dual cone is always closed and convex. If \mathcal{K} is pointed then its dual is solid, and if \mathcal{K} is solid then its dual is pointed. If \mathcal{K} is a closed, convex cone then $\mathcal{K}^{**} = \mathcal{K}$. If $\mathcal{K}_1, \dots, \mathcal{K}_m$ are closed convex cones then*

- $\mathcal{K}_1 \subseteq \mathcal{K}_2$ implies $\mathcal{K}_2^* \subseteq \mathcal{K}_1^*$,
- $\mathcal{K}_1^* \cap \mathcal{K}_2^* = (\mathcal{K}_1 + \mathcal{K}_2)^*$,
- $(\mathcal{K}_1 \cap \mathcal{K}_2)^* = \text{cl}(\mathcal{K}_1^* + \mathcal{K}_2^*)$
- $(\mathcal{K}_1 \times \dots \times \mathcal{K}_m)^* = \mathcal{K}_1^* \times \dots \times \mathcal{K}_m^*$

Definition 2.1.14 (Self-dual cone). *A cone \mathcal{K} is self-dual if $\mathcal{K} = \mathcal{K}^*$.*

Definition 2.1.15 (Irreducible cone). *A cone is irreducible if it is not a product of other cones.*

Theorem 2.1.16 (Self-dual cones). *The product of self-dual cones is self-dual. The following three examples are real, irreducible, self-dual cones.*

Nonnegative orthant: *The set \mathbb{R}_+^n is a self dual cone with the usual scalar product.*

Lorentz cone: *The set*

$$\mathbb{L} = \{(x_0, x) \in \mathbb{R}^{n+1} : x_0 \geq \|x\|\} \quad (2.1.3)$$

is a closed convex cone. It is sometimes referred to as second order or quadratic cone. With the usual scalar product it is self-dual.

Positive semidefinite cone: *The cone of symmetric positive semidefinite matrices is self-dual. The scalar product is defined as $\langle X, Y \rangle = \text{Tr}(X^T Y)$. Considering the $n \times n$ matrices as vectors in $\mathbb{R}^{n \times n}$ we can use the scalar product for vectors, i.e., the sum of the products of the corresponding elements. Denoting this scalar product of two matrices A and B by $A \bullet B$ we have the following properties, see, e.g., [66].*

1. $A \bullet B = \text{Tr}(AB) = \text{Tr}(BA)$.
2. If $B = bb^T$ is a rank-1 matrix then $A \bullet B = b^T Ab$.
3. If A and B are positive semidefinite matrices then $A \bullet B \geq 0$.

Finally, we extend the concept of convexity to functions.

Definition 2.1.17 (Convex and concave functions). *If $\mathcal{C} \subseteq V$ is a convex set, then a function $f : \mathcal{C} \rightarrow \mathbb{R}$ is convex if for every $x, y \in \mathcal{C}$ and $\lambda \in [0, 1]$ we have $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$. In other words the epigraph $\{(x, y) : y \geq f(x)\}$ is convex. A function f is concave if $-f$ is convex.*

Background material specific to only one chapter will be included in the corresponding chapter.

2.2 Lagrange duality

The classical approach to optimization problems is Lagrange duality. First we review the classical theory followed by some special cases. The proofs can be found in almost any textbook, see e.g. [115, 120]. Here we follow [35].

2.2.1 The classical theory

In this section we will investigate the optimization problem

$$\begin{aligned} \min f(x) \\ g_j(x) \leq 0, \quad j = 1, \dots, m \\ x \in \mathcal{C}, \end{aligned} \tag{NLP}$$

where $\mathcal{C} \subseteq \mathbb{R}^n$ is a convex set, f and $g_j, j = 1, \dots, m$ are convex functions on \mathcal{C} . We will require the following constraint qualification or regularity condition.

Definition 2.2.1 (Slater point). *A point $\bar{x} \in \text{relint}(\mathcal{C})$ is a Slater point if*

$$\begin{aligned} g_j(\bar{x}) < 0, & \quad \text{if } g_j \text{ is nonlinear,} \\ g_j(\bar{x}) \leq 0, & \quad \text{if } g_j \text{ is linear.} \end{aligned}$$

If there exists a Slater point then we call the problem Slater regular.

Instead of talking about linear and nonlinear constraints we can introduce the following definition:

Definition 2.2.2 (Singular/regular constraints). *A constraint is called singular¹ if it is satisfied with equality for all feasible points. If $J = \{1, \dots, m\}$ is the index set for the constraints then J_s will denote the index set of singular constraints. The remaining constraints are called regular, their index set is denoted by J_r .*

If the problem is Slater regular then the singular constraints must be linear, but there can be linear regular constraints as well. If we have a Slater point then it is easy to construct a point that satisfies not only the nonlinear but all the regular constraints with strict inequality. Such a point is called an *ideal Slater point*.

The theory of the Lagrange dual is based on the following theorem of the alternatives, which can be proved using the separation theorem of convex sets (Theorem 2.1.11). For a detailed proof see [115, Theorem 21.1], [120, §6.10] or [35].

Theorem 2.2.3 (Farkas theorem). *Consider the inequality system*

$$\begin{aligned} f(x) &< 0 \\ g_j(x) &\leq 0, \quad j = 1, \dots, m \\ x &\in \mathcal{C} \end{aligned} \tag{ConvP}$$

and assume that there is a Slater point. Then the system (ConvP) has no solutions if and only if the following system is solvable:

$$\begin{aligned} f(x) + \sum_{j=1}^m y_j g_j(x) &\geq 0, \quad \forall x \in \mathcal{C}, \\ y &\geq 0. \end{aligned} \tag{ConvD}$$

Moreover, y can be chosen in a way that $y_j > 0$ whenever $j \in J_s$ corresponds to a singular constraint.

Specializing this result for linear systems provides the Farkas lemma, first proved by Julius (Gyula) Farkas [43]:

¹Singular constraints are often called implicit equality constraints.

Theorem 2.2.4 (Farkas lemma). *Exactly one of the following two systems has a solution:*

$$\begin{aligned} c^T x &< 0 \\ a_j^T x &\leq 0, \quad j = 1, \dots, m \\ x &\geq 0 \end{aligned} \tag{2.2.1}$$

$$\begin{aligned} \sum_{j=1}^m y_j a_j + c &\geq 0, \\ y &\geq 0. \end{aligned} \tag{2.2.2}$$

After this introduction let us define the Lagrange function

$$L(x, y) = f(x) + \sum_{j=1}^m y_j g_j(x), \tag{2.2.3}$$

where $x \in \mathcal{C}$, $y_j \geq 0$, $j = 1, \dots, m$. This function is convex in x and linear in y , let us investigate its saddle points.

Definition 2.2.5 (Saddle point). *A point $(\tilde{x}, \tilde{y}) \in \mathcal{C} \times \mathbb{R}_{\oplus}^m \subseteq \mathbb{R}^{n+m}$ is a saddle point of $L(x, y)$ if*

$$L(\tilde{x}, y) \leq L(\tilde{x}, \tilde{y}) \leq L(x, \tilde{y}) \tag{2.2.4}$$

for all $x \in \mathcal{C}$ and $y \geq 0$, or equivalently if

$$L(\tilde{x}, y) \leq L(x, \tilde{y}) \tag{2.2.5}$$

for all $x \in \mathcal{C}$ and $y \geq 0$.

Using the following technical lemma we can get the fundamental Karush–Kuhn–Tucker theorem.

Lemma 2.2.6. *A point $(\tilde{x}, \tilde{y}) \in \mathcal{C} \times \mathbb{R}_{\oplus}^m \subseteq \mathbb{R}^{n+m}$ is a saddle point of $L(x, y)$ if and only if*

$$\inf_{x \in \mathcal{C}} \sup_{y \geq 0} L(x, y) = L(\tilde{x}, \tilde{y}) = \sup_{y \geq 0} \inf_{x \in \mathcal{C}} L(x, y). \tag{2.2.6}$$

Theorem 2.2.7 (Karush-Kuhn-Tucker theorem).

If the convex optimization problem (NLP) is Slater regular then a point \tilde{x} is an optimal solution if and only if there is a \tilde{y} such that (\tilde{x}, \tilde{y}) is a saddle point of $L(x, y)$.

Several useful optimality criteria can be derived from this theorem. Since our main interest is duality we skip these statements.

Based on the Lagrange function we can define the dual functional:

$$\psi(y) = \inf_{x \in \mathcal{C}} L(x, y) = \inf_{x \in \mathcal{C}} \left\{ f(x) + \sum_{j=1}^m y_j g_j(x) \right\} \quad (2.2.7)$$

and the dual problem

$$\begin{aligned} \sup \psi(y) \\ y \geq 0. \end{aligned} \quad (\text{NLP-Dual})$$

It is easy to prove that the dual problem is convex, i.e., $\psi(y)$ is a concave function. An immediate result is the weak duality theorem:

Theorem 2.2.8 (Weak duality for NLP). *If \tilde{x} is a feasible solution of the convex optimization problem and $\tilde{y} \geq 0$ then*

$$\psi(\tilde{y}) \leq f(\tilde{x}). \quad (2.2.8)$$

The necessary and sufficient condition for equality is

$$\inf_{x \in \mathcal{C}} \left\{ f(x) + \sum_{j=1}^m \tilde{y}_j g_j(x) \right\} = f(\tilde{x}). \quad (2.2.9)$$

As we mentioned in the introduction this type of duality is usually easy to prove even without further assumptions. The difficult question is whether there are optimal points \tilde{x} and \tilde{y} for which $\psi(\tilde{y}) = f(\tilde{x})$. We need some regularity assumption to prove such a result.

Theorem 2.2.9 (Strong duality for NLP). *Assume the convex optimization problem (ConvP) satisfies the Slater condition. Then a feasible vector \tilde{x} can be an optimal solution if and only if there exists an optimal solution $\tilde{y} \geq 0$ to the dual problem and the duality gap is zero, ie.*

$$\psi(\tilde{y}) = f(\tilde{x}). \quad (2.2.10)$$

This nice theory is difficult to apply in full generality, since without derivative information simply calculating $\psi(y)$ can be as difficult as the original problem. In the next subsection we will discuss how to simplify this theory for smooth convex functions.

2.2.2 Lagrange-Wolfe duality

Let us take another look at the optimality conditions (2.2.9). If all the functions are convex, continuously differentiable and $\mathcal{C} = \mathbb{R}^n$ (or \mathcal{C} is a full dimensional open set) then we can write the first order optimality condition for unconstrained optimization: The infimum is taken at a point where the gradient is 0 and under the convexity assumption this condition is sufficient as well, assuming this infimum is attained. This observation gives us the Wolfe dual:

$$\begin{aligned} \sup f(x) + \sum_{j=1}^m y_j g_j(x) \\ \nabla f(x) + \sum_{j=1}^m y_j \nabla g_j(x) = 0 \\ y \geq 0. \end{aligned} \tag{LW}$$

Since this system is equivalent to the Lagrangian dual system (assuming convexity and smoothness) we have the same weak and strong duality results.

2.3 Fenchel duality

In this section we present a newer and more abstract way of dualizing the convex optimization problem. In the last section we will use this technique to derive the dual for the conic optimization problem.

Given any function we can define its conjugate function.

Definition 2.3.1 (Convex conjugate functions). *Let $f : X \rightarrow \mathbb{R}$ be a function where $X \subseteq \mathbb{R}^n$. The convex conjugate of f is a function $g : \Lambda \rightarrow \mathbb{R}$ defined by*

$$g(\lambda) = \sup_{x \in X} \{x^T \lambda - f(x)\}, \tag{2.3.1}$$

where $\Lambda = \{\lambda : g(\lambda) < \infty\}$.

The theory of convex conjugate functions was developed by Fenchel [46, 47]. Later it was generalized by Rockafellar [115].

2.3.1 The Fenchel dual as a Lagrange dual

Instead of the tedious and lengthy discussion of Fenchel duality we will now make use of what we derived in the previous section and use Lagrange duality

to introduce the Fenchel dual.² The material presented here is from [20].

Let us consider the optimization problem in the form

$$\begin{aligned} \min \quad & f_1(x) - f_2(x) \\ & x \in X_1 \cap X_2, \end{aligned} \tag{FP}$$

where $f_{1,2} : \mathbb{R}^n \rightarrow \mathbb{R}$ are convex functions and $X_{1,2} \subseteq \mathbb{R}^n$ are convex sets. Rewriting the problem in the following equivalent form

$$\begin{aligned} \min \quad & f_1(x^1) - f_2(x^2) \\ & x^1 = x^2 \\ & x^1 \in X_1, \\ & x^2 \in X_2 \end{aligned} \tag{2.3.2}$$

enables us to use Lagrange duality discussed in the previous section.³ The objective function of the Lagrange dual is

$$\begin{aligned} q(\lambda) &= \inf_{x^1 \in X_1, x^2 \in X_2} \left\{ f_1(x^1) - f_2(x^2) + (x^2 - x^1)^T \lambda \right\} \\ &= \inf_{x^2 \in X_2} \left\{ x^{2T} \lambda - f_2(x^2) \right\} + \inf_{x^1 \in X_1} \left\{ f_1(x^1) - x^{1T} \lambda \right\}, \end{aligned} \tag{2.3.3}$$

where $\lambda \in \mathbb{R}^n$. Introducing

$$g_i(\lambda) = \sup_{x \in X_i} \{x^T \lambda - f_i(x)\}, \quad i = 1, 2 \tag{2.3.4}$$

we can write our dual problem in the form

$$\begin{aligned} \max \quad & g_2(\lambda) - g_1(\lambda) \\ & \lambda \in \Lambda_1 \cap \Lambda_2, \end{aligned} \tag{FD}$$

where

$$\Lambda_1 = \{\lambda : g_1(\lambda) < \infty\}, \quad \Lambda_2 = \{\lambda : g_2(\lambda) > -\infty\}. \tag{2.3.5}$$

Note that g_1 and g_2 are the corresponding convex conjugate functions to f_1 and f_2 , introduced in (2.3). Applying the Lagrange duality theory we have the following necessary and sufficient optimality conditions:

²The two duals are in fact equivalent, see, e.g., [84].

³Before applying the dual we convert the equality constraint into two inequalities.

Theorem 2.3.2. *The pair x^* and λ^* are optimal primal and dual solutions with zero duality gap if and only if*

$$x^* \in X_1 \cap X_2 \quad (\text{primal feasibility}) \quad (2.3.6a)$$

$$\lambda^* \in \Lambda_1 \cap \Lambda_2 \quad (\text{dual feasibility}) \quad (2.3.6b)$$

$$\begin{aligned} x^* &= \arg \min_{x \in X_1} \{x^T \lambda^* - f_1(x)\} \\ &= \arg \min_{x \in X_2} \{x^T \lambda^* - f_2(x)\} \quad (\text{Lagrangian optimality}). \end{aligned} \quad (2.3.6c)$$

Again, using the conjugate functions g_1 and g_2 we can write the last two optimality conditions as

$$f_i(x^*) + g_i(\lambda^*) = x^{*T} \lambda^*, \quad i = 1, 2. \quad (2.3.7)$$

To guarantee this we need convexity assumption on the sets and functions. We have the following general result:

Theorem 2.3.3 (Fenchel duality). *Let us consider the primal and dual problems (FP) and (FD). Assume that*

1. X_1 is the intersection of a polyhedron and a convex set C_1 and f_1 is convex over C_1 ,
2. X_2 is the intersection of a polyhedron and a convex set C_2 and f_2 is concave over C_2
3. and the sets $X_1 \cap \text{rel int}(C)_1$ and $X_2 \cap \text{rel int}(C)_2$ have nonempty intersection.

With this condition the dual problem is solvable and for the optimal solution λ^ we have*

$$\inf_{x \in X_1 \cap X_2} \{f_1(x) - f_2(x)\} = \sup_{\lambda \in \Lambda_1 \cap \Lambda_2} \{g_2(\lambda) - g_1(\lambda)\} = g_2(\lambda^*) - g_1(\lambda^*) \quad (2.3.8)$$

Remark 2.3.4. Note that we cannot state anything about primal solvability.

If we look closer then we can find well known elements from the Lagrange theory. The convexity assumption is always a must, the intersection conditions on X_i corresponds to the singular and regular constraints and finally the relative interior condition is a form of the Slater condition.

Finally it remains to show the symmetry of this dual: Under what conditions is it true that dualizing the dual problem we get the primal one? Or

in other words, when can dual feasibility guarantee primal-dual solvability with zero duality gap? Recall from the theory of convex conjugate functions that if f is a closed function over a convex set, i.e., the epigraph $\{(x, y) : y \geq f(x)\}$ is closed, then the double conjugate of f is itself. Based on this result we can state the dual form of the previous theorem.

Theorem 2.3.5 (Fenchel duality (dual form)). *Consider the above primal and dual problems. Assume that*

1. Λ_1 is the intersection of a polyhedron and a convex set C_1 and g_1 can be extended to a real valued convex function over C_1 ,
2. Λ_2 is the intersection of a polyhedron and a convex set C_2 and g_2 can be extended to a real valued concave function over C_2 ,
3. the sets $\Lambda_1 \cap \text{rel int}(C)_1$ and $\Lambda_2 \cap \text{rel int}(C)_2$ have nonempty intersection,
4. and the primal functions f_1 and f_2 are closed.

With these conditions the primal problem is solvable and for the optimal solution x^* we have

$$f_1(x^*) - f_2(x^*) = \inf_{x \in X_1 \cap X_2} \{f_1(x) - f_2(x)\} = \sup_{\lambda \in \Lambda_1 \cap \Lambda_2} \{g_2(\lambda) - g_1(\lambda)\}. \quad (2.3.9)$$

Finally we can combine this result with the previous one. What makes this approach very attractive is that the dual problem shares the same structure as the primal problem.

2.4 Special cases

In this section we derive the duals for some common problems.

2.4.1 Linear optimization

The standard linear optimization problem is the following:

$$\begin{aligned} \min \quad & c^T x \\ & Ax = b \\ & x \geq 0 \end{aligned} \quad (\text{LP-P})$$

where $A \in \mathbb{R}^{m \times n}$, $x, c \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$. First of all we write $Ax = b$ equivalently as $Ax - b \leq 0$ and $-Ax + b \leq 0$, and $x \geq 0$ as $-x \leq 0$. The

corresponding Lagrange multipliers are $y^+, y^- \in \mathbb{R}^m$ and $s \in \mathbb{R}^n$. Since all the functions are convex and smooth and we are working on \mathbb{R}^n the easiest way is to apply Wolfe duality. The dual has the following form:

$$\begin{aligned} \max \quad & c^T x + (y^-)^T (Ax - b) + (y^+)^T (-Ax + b) + s^T (-x) \\ & c + A^T y^- - A^T y^+ - s = 0 \\ & y^- \geq 0, \\ & y^+ \geq 0, \\ & s \geq 0. \end{aligned} \tag{2.4.1}$$

Introducing $y = y^+ - y^-$ and substituting the equality constraint into the objective we get the usual dual form:

$$\begin{aligned} \max \quad & b^T y \\ & A^T y + s = c \\ & s \geq 0 \end{aligned} \tag{LP-D}$$

The optimality conditions can be derived from the KKT theorem. We obtain the following theorem:

Theorem 2.4.1 (Duality for LP). *Consider the problems (LP-P)–(LP-D). If both problems are feasible then the optimal values are equal, and there are x, s and y such that the optimum is realized, i.e., $c^T x = b^T y$, the duality gap is 0. In other words, a feasible primal-dual solution (x, s, y) is optimal if and only if $x^T s = 0$.*

Since x and s are nonnegative this is the well-known complementary slackness condition.

2.4.2 Quadratic optimization

Let us consider the quadratic optimization problem in the form

$$\begin{aligned} \min \quad & c^T x + \frac{1}{2} x^T Q x \\ & Ax = b, \\ & x \geq 0, \end{aligned} \tag{QP-P}$$

where $A \in \mathbb{R}^{m \times n}$, $Q \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^m$ and $x, c \in \mathbb{R}^n$. Assume further that Q is positive semidefinite, i.e., the objective function is convex. Using Lagrange multipliers $y \in \mathbb{R}^m$ and $s \in \mathbb{R}^n$ we can write the dual form (after

simplification):

$$\begin{aligned} \max \quad & b^T y - \frac{1}{2} x^T Q x \\ -Qx + A^T y + s &= c \\ s &\geq 0. \end{aligned} \tag{2.4.2}$$

It is quite odd that the primal variable x is present in the dual. Moreover, it is important to note that variable x in this form is not necessarily a feasible primal solution. To circumvent this little drawback we have to eliminate x . This is done by factorizing Q as $P^T P$ and introducing $z = Px$. This gives us the dual form:

$$\begin{aligned} \max \quad & b^T y - \frac{1}{2} z^T z \\ -P^T z + A^T y + s &= c \\ s &\geq 0. \end{aligned} \tag{QP-D}$$

The optimality conditions are obtained from the general theory.

Theorem 2.4.2 (Duality for QP). *The vectors x , s , z and y are optimal solutions for the primal-dual quadratic problems (QP-P) and (QP-D) if and only if*

$$\begin{aligned} x^T s &= 0, \\ y^T (Ax - b) &= 0, \\ z &= Px. \end{aligned} \tag{2.4.3}$$

2.4.3 Conic optimization

This problem is much more general than the previous ones. As such, duality results are weaker, and practical solvability is limited only to a subset of problems discussed later.

Let $\mathcal{K} \subseteq \mathbb{R}^n$ be a closed, convex, pointed, solid cone, and let us use the ordering defined in Definition 2.1.8. The standard form of the conic linear program is the following:

$$\begin{aligned} \min \quad & c^T x \\ Ax &= b, \\ x &\succeq_{\mathcal{K}} 0, \end{aligned} \tag{ConicP}$$

where $A \in \mathbb{R}^{m \times n}$, $x, c \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$. Taking the nonnegative orthant \mathbb{R}_{\oplus}^n as \mathcal{K} we have linear optimization as a special case. As an illustration we will now use Fenchel duality to derive the dual of the conic linear optimization problem. Recall from §2.3.1 the notations of the Fenchel duality. Here we have the following setting:

$$X_1 = \mathcal{K}, \quad f_1(x) = c^T x \quad (2.4.4a)$$

$$X_2 = \{x \in \mathbb{R}^n : Ax = b\}, \quad f_2(x) = 0. \quad (2.4.4b)$$

The next step is to find the conjugate functions $g_i(\lambda)$.

$$g_1(\lambda) = \sup_x \{x^T(\lambda - c) : x \in \mathcal{K}\} \quad (2.4.5)$$

If there is an $x \in \mathcal{K}$ such that $x^T(\lambda - c) > 0$ then the supremum is unbounded since \mathcal{K} is a cone. On the other hand if $x^T(\lambda - c) \leq 0$ for all $x \in \mathcal{K}$ then the supremum is 0 with $x = 0$. This condition is equivalent to $c - \lambda \in \mathcal{K}^*$. The final form of the conjugate function is

$$g_1(\lambda) = \begin{cases} 0, & \text{if } c - \lambda \in \mathcal{K}^* \\ +\infty, & \text{if } c - \lambda \notin \mathcal{K}^* \end{cases}. \quad (2.4.6)$$

The other function

$$g_2(\lambda) = \inf_x \{x^T \lambda : Ax = b\} \quad (2.4.7)$$

requires more work. Using the usual assumption that A has full row rank we can write it as $[A_B | A_N]$, where A_B is an invertible matrix. Dividing x the same way we have $x_B = A_B^{-1}b - A_B^{-1}A_N x_N$. Plugging this into the objective gives

$$x^T \lambda = x_B^T \lambda_B + x_N^T \lambda_N = (A_B^{-1}b)^T + x_N^T (\lambda_N - A_N^T A_B^{-1T} \lambda_B). \quad (2.4.8)$$

Now we have the conjugate function:

$$g_2(\lambda) = \begin{cases} b^T A_B^{-1T} \lambda_B, & \text{if } \lambda_N = A_N^T A_B^{-1T} \lambda_B \\ -\infty & \text{else.} \end{cases}. \quad (2.4.9)$$

The corresponding sets of properness are

$$\Lambda_1 = \{\lambda : c - \lambda \in \mathcal{K}^*\} \quad (2.4.10a)$$

$$\Lambda_2 = \left\{ \lambda : \lambda_N = A_N^T A_B^{-1T} \lambda_B \right\}, \quad (2.4.10b)$$

so the Fenchel dual of (ConicP) is

$$\begin{aligned} \max \quad & (A_B^{-1}b)^T \lambda_B \\ & \lambda_N = A_N^T A_B^{-1T} \lambda_B \\ & c - \lambda \in \mathcal{K}^*. \end{aligned} \tag{2.4.11}$$

Let us introduce $y = A_B^{-1T} \lambda_B$. The constraint gives $\lambda_N = A_N^T y$, so λ can be expressed as $A^T y$. Rewriting the system gives the usual dual:

$$\begin{aligned} \max \quad & b^T y \\ & A^T y \preceq_{\mathcal{K}^*} c. \end{aligned} \tag{ConicD}$$

Now we can turn our attention to duality. Translating the result of the Fenchel duality theorem we get the following:

Theorem 2.4.3 (Duality for conic optimization). *Consider the primal and dual conic optimization problem.*

1. *If the primal problem (ConicP) is strictly feasible, i.e., there exists an x for which $Ax = b$ and $x \succ_{\mathcal{K}} 0$, then the dual problem is solvable and the optimal values are the same.*
2. *If the dual problem (ConicD) is strictly feasible, i.e., there exists an y for which $A^T y \prec_{\mathcal{K}^*} c$, then the primal problem is solvable and the optimal values are the same.*

Note that in the linear case it was enough to assume feasibility instead of strict feasibility.

These types of duality theorems give us certificates for optimality, feasibility and solvability. Based on this information we can design algorithms to solve these problems.

2.5 Theorems of the alternative, optimality conditions and duality

All the results in the previous section can be expressed in two equivalent ways represented in the following diagram:

$$\begin{array}{ccc}
 x^* \text{ is optimal in} & & \\
 \min f(x) & \stackrel{\text{CQ}}{\iff} & \max_{y \geq 0} \min_{x \in \mathcal{C}} f(x) + y^T g(x) \geq f(x^*) \\
 g(x) \leq 0 & & \\
 x \in \mathcal{C} & & \\
 \Downarrow & & \Downarrow \\
 \nexists x \in \mathcal{C} & \stackrel{\text{CQ}}{\iff} & \exists y \geq 0 \text{ such that } \forall x \in \mathcal{C} \\
 f(x) - f(x^*) < 0 & & f(x) + y^T g(x) \geq f(x^*) \\
 g(x) \leq 0 & &
 \end{array}$$

The top row is a duality theorem, it characterizes the optimality of a solution to the primal problem. The bottom row is a theorem of the alternative as exactly one of the two systems can have a solution. The equivalence between the top and bottom rows is evident. From right to left the implications correspond to the weak duality theorem, i.e., a sufficient condition. From left to right we need a CQ, and we get the strong duality theorem.

In this thesis we will use both formulations depending on which one is more natural. It is left to the reader to convert the results into the other form if needed. Usually the theorem of the alternative is easier to understand, but it is the duality theorem that is used in practice in the form of optimality conditions.

2.6 Historical remarks

This section is based on [52, 115, 130]. References to historical works are omitted as they are usually inaccessible and newer books offer a better treatment of the subjects. A brief survey of the modern literature is presented in the next section.

2.6.1 The history of optimization, duality and convexity

The solvability of linear and nonlinear systems and the existence and identification of optimal solutions has long been an active research area. The first noteworthy result dates back to the mid 1600s and is from Fermat who proved the necessary condition $f'(x) = 0$ for the unconstrained (local) minimum/maximum of a nonlinear function f . The first result about constrained problems came in 1788 by Lagrange, who minimized the potential function of a mechanical system to find its equilibrium. He considered nonlinear problems with equality constraints and used what is today known as the Lagrange-Wolfe theorem (see §2.2.2). The multipliers he introduced are the Lagrange multipliers. The theorem was proved by Euler, Lagrange's supervisor. The optimality conditions for general (convex) nonlinear optimization were first obtained (and published in his MSc thesis) by Karush in 1939 and later, independently, by Kuhn and Tucker. Fenchel's duality theorem for nonlinear problems dates back to 1949. Wolfe proved his theorem about differentiable constrained nonlinear optimization in 1961. Minimax theorems and the saddle points were first investigated by von Neumann in the context of game theory.

Inequality constrained linear problems were first solved in 1826 by Fourier, and later linear problems were also studied by Cournot (1827) and Ostrogradski (1834). The first complete proof of the theory was provided only in 1898 by Julius (Gyula) Farkas. Theorems of the alternative with strict positivity constraints were proved by Gordan (1873) and Stiemke (1951). Most of these results can also be derived from Motzkin's thesis written in 1936.

The linear optimization problem was stated formally by Kantorovich in 1939. The first general algorithm to solve such problems (the simplex method) was given by Dantzig in 1947. The strong duality theorem of linear optimization was proved in 1951 by Gale, Kuhn and Tucker, using Farkas's result about systems of homogeneous linear inequalities.

Convex sets, in particular polyhedra were first studied by Minkowski in 1896. He was the first to prove separation results for convex sets, which were later generalized to infinite dimensional spaces by Hahn and Banach independently around 1920. Minkowski's results were extended by Carathéodory (1907) and Weyl (1937). The theory of convex conjugate functions was developed by Fenchel in 1951 with later extensions by Moreau (1966). The extremal structure of convex sets was described by Klee in the late 1950s.

Unquestionably, the most influential work on convex optimization is Rockafellar's classic book [115], first published in 1970. He summarized all the classical theory of convexity and duality, proved many new results, introduced new techniques and opened whole new directions for further results. His book

closed the classical era of convex optimization.

The interior point revolution started in 1984 with Karmarkar’s algorithm for linear optimization problems. The idea was first introduced by Fiacco and McCormick in the early 1960s for nonlinear optimization. They even had a working computer code (called the *sledgehammer*) to solve nonlinear optimization problems. Karmarkar’s method had polynomial complexity and it initiated a new line of research that resulted in efficient algorithms. The modern theory of interior point methods for nonlinear optimization was developed by Nesterov and Nemirovski in 1994.

2.6.2 Literature

The modern literature of the classical theory of duality and optimization is quite extensive, see [21, 26, 35, 64, 115, 120] to mention just a few of the books. By today, convex analysis has become a well-established, almost classical field of research. For an introduction to the geometry of cones we suggest the short monograph by Bergman [19]. Polytopes, their facial structure and the properties of the faces are the central topic of Brønsted’s book [30].

In the theory of interior point methods for convex optimization the classical reference is [89], but [114] and [131] may be easier to read and satisfy most readers.

2.7 Motivation: convexity+CQ=duality?

For all the theorems in this section we had to assume that the sets and functions are convex, and to prove strong duality results we also needed a constraint qualification, typically the Slater condition. As illustrated by the following examples, these ingredients are crucial.

2.7.1 A nonconvex system

Convexity seems to be a crucial element of the classical duality theories. Consider, e.g., the following one-dimensional nonconvex problem:

$$\min x^4 - 10x^2 \tag{2.7.1a}$$

$$x^2 \geq 4 \tag{2.7.1b}$$

$$x \in \mathbb{R} \tag{2.7.1c}$$

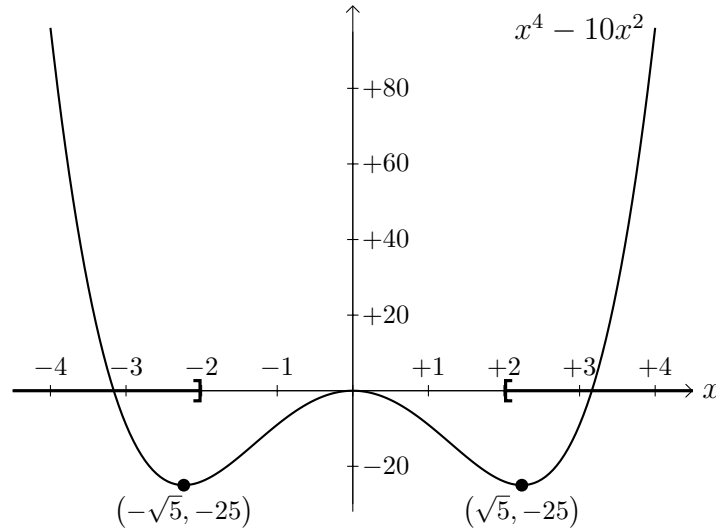


Figure 2.1: The objective function, the feasible set and the global minima of the nonconvex optimization problem (2.7.1).

The primal system (2.7.1) is nonconvex, it has a local (but infeasible) maximum at $x = 0$ and two global minima at $\pm\sqrt{5}$, the optimal value is -25 and the problem clearly satisfies the Slater condition. The objective function is depicted in Figure 2.1. The Lagrange-Wolfe dual of the problem is:

$$\max x^4 - 10x^2 + y(4 - x^2) \quad (2.7.2a)$$

$$4x^3 - 20x - y(-2x) = 0 \quad (2.7.2b)$$

$$y \geq 0 \quad (2.7.2c)$$

Unfortunately, this dual problem is unbounded, if $x = 0$ then y can be any nonnegative number, thus the dual optimum value is $+\infty$. Even the weak duality theorem fails for this example.

In Chapter 3 we show how the classical theory can be changed for nonconvex quadratic systems.

2.7.2 A nonregular system

The second crucial point in the duality theories is the constraint qualification or regularity condition. The natural question is: how much does the theory change if we do not assume the Slater condition? The following example (taken from [89]) shows that the strong duality theorem will no longer hold. Consider

the semidefinite optimization problem:

$$\begin{aligned} & \max u_2 \\ & \begin{pmatrix} u_2 & 0 & 0 \\ 0 & u_1 & u_2 \\ 0 & u_2 & 0 \end{pmatrix} \preceq \begin{pmatrix} \alpha & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \end{aligned} \quad (2.7.3)$$

For any feasible solution of this problem we have $u_2 = 0$ so the optimal objective value is 0. However, the dual problem

$$\begin{aligned} & \min \alpha v_{11} \\ & v_{22} = 0 \\ & v_{11} + 2v_{23} = 1 \\ & \begin{pmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{pmatrix} \succeq 0 \end{aligned} \quad (2.7.4)$$

has $v_{22} = v_{23} = 0$ and $v_{11} = 1$, thus the optimal value is α . This shows that the duality gap can be arbitrarily large. The set of feasible solutions for the primal problem is $u_1 \leq 0, u_2 = 0$, so there is no strictly feasible primal solution. The Slater condition is not satisfied and the strong duality theorem of conic optimization (Theorem 2.4.3) does not hold.

A special case of the theory, when duality holds without the Slater condition, is the topic of Chapter 4.

2.7.3 Nonexact duality

So far we assumed that all the computations can be carried out in exact arithmetic. In practice this is not the case. Numbers are represented in finite precision using truncation. Similarly, the result of different operations will only be approximate. The result of an operation that in theory should yield zero can be a small nonzero number, or we might get a small negative number where a nonnegative number is expected. Similarly, the duality theorems of this chapter have to be adapted to this new situation. We have to live with the fact that optimality conditions are barely satisfied exactly, there is always a small error. In practice, small deviations up to the order of 10^{-9} are routinely ignored.

Consider the following linear example, illustrated in the left of Fig-

Figure 2.2.

$$\begin{aligned} \min \quad & 10^6 x_1 + 10^{-6} x_2 \\ & x_1, x_2 \leq 1 \\ & x_1, x_2 \geq 0 \end{aligned} \tag{2.7.5}$$

The optimal solution is clearly $x_1 = x_2 = 0$ with an optimal value of 0. The

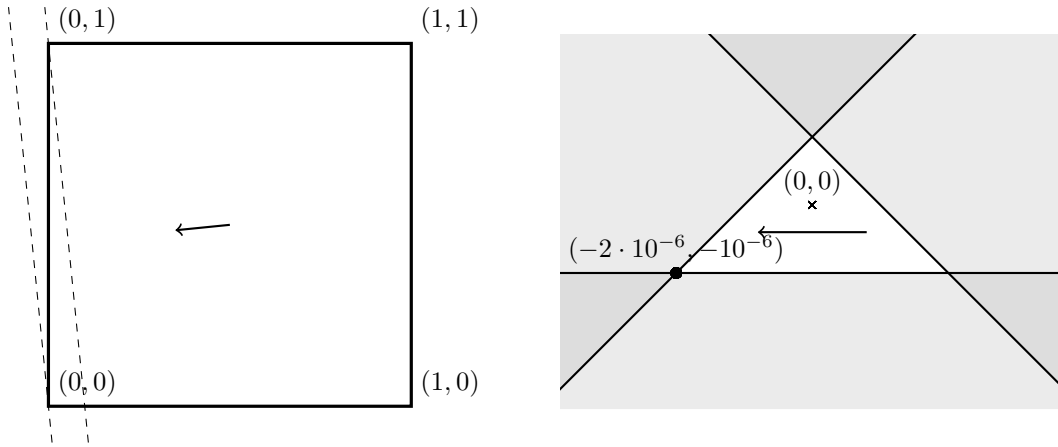


Figure 2.2: Two numerically sensitive linear problems. **Left:** An almost optimal solution far from the real one. The feasible set is the unit square, the objective function is decreasing in the direction of the arrow. The dashed lines are level sets of the objective function. The optimal solution is the vertex $(0,0)$, but $(0,1)$ is almost as good as the optimum. (The picture is not drawn to scale.) **Right:** An almost feasible and optimal solution for an infeasible problem. The lines are the boundaries of the constraints, the shaded regions are the feasible sets of the constraints. The arrow in the middle shows the direction where the objective function is decreasing. The highlighted point is almost optimal, but not feasible.

dual problem is:

$$\begin{aligned} \max \quad & -y_1 - y_2 \\ & y_1 \geq -10^6 \\ & y_2 \geq -10^{-6} \\ & y_1, y_2 \geq 0 \end{aligned} \tag{2.7.6}$$

The optimal solution of the dual problem is $y_1 = -10^6$, $y_2 = -10^{-6}$ and the optimal value is $10^6 + 10^{-6}$. However, setting $x_1 = 0$, $x_2 = 1$, $y_1 = -10^{-6}$ and $y_2 = 0$ the duality gap is $2 \cdot 10^{-6}$, a small number. Given that these are feasible solutions, one might treat this solution as almost optimal, and it is correct in

the sense that the optimal value is almost the best possible. Unfortunately, although the value of the objective function is close to the optimal value, the optimal solution is relatively far from this point.

The following example is even more evil:

$$\begin{aligned} \min \quad & x_1 \\ & x_2 \leq -10^{-6} \\ & x_1 + x_2 \geq 10^{-6} \\ & x_1 - x_2 \leq -10^{-6}, \end{aligned} \tag{2.7.7}$$

see the right hand side of Figure 2.2 for a graphical representation. Its dual problem is:

$$\begin{aligned} \max \quad & 10^{-6}(y_1 + y_2 + y_3) \\ & -y_2 + y_3 = -1 \\ & y_1 - y_2 - y_3 = 0 \\ & y_1, y_2, y_3 \geq 0 \end{aligned} \tag{2.7.8}$$

The point $x_1 = -2 \cdot 10^{-6}$, $x_2 = -10^{-6}$ is almost feasible for the primal problem and $y_1 = y_2 = 1$, $y_3 = 0$ is feasible for dual problem. The duality gap is 0, so we might accept it as an optimal solution. But not only this point is not optimal, it is not feasible, and in fact the problem does not even have a feasible solution.

These two examples show that we have to be very careful with finite precision arithmetic. We might not be able to tell apart an optimal solution from a suboptimal one, or we might declare a solution optimal when no feasible solution exists. These problems may look very artificially crafted but these are the typical situations in practice. This is the motivation behind Chapter 5.

In the next two chapters we will present two special cases where strong duality holds either without convexity or without regularity condition. Following that, we will show what we can do in nonexact (i.e., machine) arithmetic.

Chapter 3

Duality for nonconvex quadratic systems

Con.vex *a.* Rising or swelling into a spherical or rounded form; regularly protuberant or bulging; – said of a spherical surface or curved line when viewed from without, in opposition to *concave*.

THE 1913 WEBSTER
UNABRIDGED DICTIONARY

In this chapter we review the many faces of the S-lemma, a result about the correctness of the S-procedure. The basic idea of this widely used method came from control theory but it has important consequences in quadratic and semidefinite optimization, convex geometry and linear algebra as well. These were active research areas, but as there was little interaction between researchers in these different areas, their results remained mainly isolated. Here we give a unified analysis of the theory by providing three different proofs (one of which is new) for the S-lemma and revealing hidden connections with various areas of mathematics. We prove some new duality results and present applications from control theory, error estimation and computational geometry.

This chapter can be divided into two parts. The first part gives a general overview of the S-lemma. It starts from the basics, provides three different proofs of the fundamental result, discusses possible extensions and presents some counterexamples. Some illustrative applications from control theory and error estimation are also discussed.

The second part, starting with §3.5, shows how the basic theory is related to various fields of mathematics: functional analysis, rank-constrained optimization and generalized convexities. This part goes beyond the proofs and demonstrates how the same result was discovered several times throughout the 60-year history of the S-lemma. New duality results and further directions are also presented.

3.1 Introduction

In this section we expose the basic question of the S-lemma and provide some historical background.

3.1.1 Motivation

The fundamental question of the theory of the S-lemma is the following:

*When is a quadratic (in)equality a consequence
of other quadratic (in)equalities?*

If we ask the same question for linear or general convex (in)equalities then the Farkas lemma (Theorem 2.2.4 or [16, Theorem 1.3.1.]) and the Farkas Theorem (Theorem 2.2.3 in this thesis, or [120, §6.10]) give the answer, respectively. These results essentially state that a concave inequality is a (logical) consequence of some convex inequalities if and only if it is a nonnegative linear combination of those convex inequalities and an identically true inequality. This is important, since it is relatively easy to check if an inequality is a linear combination of some other inequalities.

However, this notion is not guaranteed to work in a general setting, as is demonstrated by the following example, taken from [16]. Consider the inequalities

$$\begin{aligned} u^2 &\geq 1, \\ v^2 &\geq 1, \\ u &\geq 0, \\ v &\geq 0, \end{aligned} \tag{3.1.1}$$

then the inequality

$$uv \geq 1 \tag{3.1.2}$$

is clearly a logical consequence of these four inequalities. However, taking those four and the trivially true inequalities (such as $0 > -1$, $u^2 \pm 2uv + v^2 \geq 0$, etc.) we cannot combine them in a linear way to obtain (3.1.2). Thus only a fraction of the logical consequences can be reached using linear combinations. In this chapter we discuss when we can apply this approach to quadratic systems. As shown in the above simple example, it does not work in general, but there are some special cases when we can answer our question using various methods.

Quadratic inequalities arise naturally in many areas of theoretical and applied mathematics. Consider the following examples.

Quadratic intersections: This problem arises in computer graphic; see [76] for a standard reference. Let us take two quadratic surfaces represented by the equations $x^T A_i x + b_i^T x + c_i = 0$, $i = 1, 2$. How can we decide whether the two surfaces intersect without actually computing the intersections? Can we compute the intersection efficiently? How can we do the same for the solid bodies bounded by the surfaces?

Noise estimation: Level sets for the density function of measurements in \mathbb{R}^n with a Gaussian noise are ellipsoids. Given two such measurements each with a fixed noise level, we need to find a bound on the noise of their sum. In other words, we are looking for the smallest ellipsoid that contains the sum of the two ellipsoids.

Trust region problem: A broad range of functions can be efficiently approximated locally with quadratic functions. The trust region problem is a quadratic optimization problem with a single quadratic constraint, i.e.,

$$\begin{aligned} \min \quad & x^T A x + b^T x + c, \\ & \|x - \hat{x}\|_2 \leq \alpha, \end{aligned} \tag{3.1.3}$$

where $\alpha, c \in \mathbb{R}$, $b, x, \hat{x} \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$. This problem is easily solvable, and methods based on these approximations are widely applied in nonlinear optimization; for more details, see [33].

Quadratically constrained quadratic minimization: Finally, we can replace the norm constraint (3.1.3) with a set of general quadratic constraints:

$$\min \quad x^T A x + b^T x + c, \tag{3.1.4a}$$

$$x^T A^i x + b^{iT} x + c^i \leq 0, \quad i = 1, \dots, m. \tag{3.1.4b}$$

Besides the solvability of this problem it is often necessary to decide whether the problem is feasible (i.e., whether system (3.1.4b) is solvable). If the problem is not feasible then one might require a certificate that proves infeasibility.

Since this last problem includes integer programming we cannot hope for very general results.

The S-procedure is a frequently used technique originally arising from the stability analysis of nonlinear systems. Despite being used in practice quite frequently, its theoretical background is not widely understood. On the other hand, the concept of the S-procedure is well known in the optimization community, although not under this name.

In short, the S-procedure is a relaxation method: it tries to solve a system of quadratic inequalities via a linear matrix inequality (LMI) relaxation. Yakubovich [133] was the first to give sufficient conditions for which this relaxation is exact, i.e., when it is possible to obtain a solution for the original system using the solution of the LMI; this result is the S-lemma. The advantage we gain is the computational cost: while solving a general quadratic system can take an exponential amount of work, LMIs can be solved more efficiently.

3.1.2 Historical background

The earliest result of this kind is due to Finsler [48], and was later generalized by Hestenes and McShane [61]. In 1937 Finsler proved that if A and B are two symmetric matrices and $x^T B x = 0$ ($x \neq 0$) implies $x^T A x > 0$ then there exists an $y \in \mathbb{R}$ such that $A + yB$ is positive definite.

On the practical side, this idea was first used probably by Lur'e and Postnikov [81, 82, 83] in the 1940s, but at that time there was no well-founded theory behind the method. The theoretical background was developed some 30 years later by Yakubovich: in the early 1970s he proved a theorem known as the S-lemma [133, 135] using an old theoretical result of Dines [38] on the convexity of homogeneous quadratic mappings. The simplicity of the method allowed rapid advances in control theory. Applications grew but it was not until the 1990s that new theoretical results appeared in the area. Megretsky and Treil [87] extended the results to infinite-dimensional spaces, giving rise to more general applications. Recently, there is active research in the area, see [15, 14].

Yakubovich himself presented some applications [134], which were followed by many others [27], including contemporary ones [54, 78], spanning

a broad range of engineering, financial mathematics and abstract dynamical systems. We will discuss various applications in §3.4.

Although the result emerged mainly from practice, Yakubovich himself was aware of the theoretical implications [49] of the S-lemma. The theoretical line was then continued by others (see, e.g., [27], or recently, [36, 37, 80, 122] but apart from a few exceptions such as [16, 27, 78] or [70] the results did not reach the control community. Moreover, to the best of my knowledge no thorough study presenting all these approaches has been written so far. The collection of lecture notes by Ben-Tal and Nemirovski [16] contains some of the ideas explained here, and several aspects of this theory have been presented in the context of the LMI relaxation of systems of quadratic inequalities.

The term *S-method* was coined by Aizerman and Gantmacher in their book [1], but later it changed to *S-procedure*. The S-method tries to decide the stability of a system of linear differential equations by constructing a Lyapunov matrix. During the process an auxiliary matrix S (for stability) is introduced. This construction leads to a system of quadratic equations (the Lure’s resolving equations, [82, 83]). If that quadratic system can be solved then a suitable Lyapunov function can be constructed. The term *S-lemma* refers to results stating that such a system can be solved under some conditions; the first such result is due to Yakubovich [133]. In this chapter my main interest is the S-lemma, but we will present an example from control theory in §3.4.

3.1.3 The structure of this chapter

In this chapter we show how the S-lemma relates to well known concepts in optimization, relaxation methods and functional analysis. This chapter is structured as follows. In §3.2 we give three independent proofs for the S-lemma, illustrating how the original result is connected to more general theories. In §3.3 we show some examples and counterexamples, and present other variants of the S-lemma. Applications from control theory and computational geometry are discussed in §3.4.

In the second part we go deeper and investigate three major topics. First, in §3.5 we discuss a classical topic, the convexity of the numerical range of a linear operator. Following that, in §3.6 we present a seemingly different field, rank-constrained optimization. In §3.6.5 we merge the results of these two fields and show the equivalence of the theories. Finally, in §3.7 we put the problem in a more general context and show that the S-lemma is a special case of a duality theorem due to Illés and Kassay [67, 68, 69].

Some miscellaneous topics (trust region problems, algebraic geometric connections and complexity issues) are then discussed in §3.8. We briefly

summarize their connection to the S-lemma and indicate directions for further research. To illustrate the power of the methods discussed in the thesis, in §3.9 we present a few surprising duality theorems. Possible future directions and some open questions are discussed in §3.10.

We have made every effort to make the major sections (§3.5-3.7) self-contained, i.e., any one of them can be skipped depending on the reader's area of interest. Each of them starts with a motivation part where we describe how the S-lemma is related to the selected area, then we summarize the major theoretical results of the field, and finally, we apply the theory to the problem and conclude with the results.

3.2 Proofs for the basic S-lemma

In this section we present three proofs for the basic S-lemma. We start with the original proof of Yakubovich, then we present a modern proof based on LMIs, and we conclude with an elementary, analytic proof. The key concepts of these proofs will be further investigated and generalized in the remaining sections.

3.2.1 The two faces of the S-lemma

Now we present the central result of the theory starting from the very basics and showing the main directions of the rest of the survey. The theorems we are going to discuss can be viewed in two ways. As illustrated by the example in §3.4, the original application of the S-lemma is to decide whether a quadratic (in)equality is satisfied over a domain. As these domains are usually defined by quadratic inequalities, the question we are investigating is when a quadratic inequality is a consequence of other quadratic inequalities. This idea can be formalized as follows:

$$g_j(x) \leq 0, \quad j = 1, \dots, m \stackrel{?}{\Rightarrow} f(x) \geq 0, \quad (3.2.1)$$

where $x \in \mathbb{R}^n$ and $f, g_j : \mathbb{R}^n \rightarrow \mathbb{R}, j = 1, \dots, m$ are quadratic functions.

The alternative approach views the question as a feasibility problem: Is there any point in the domain where the inequality in question does not hold? This problem is of the same form as the Farkas Theorem, a fundamental theorem of alternatives in convex analysis.

When writing this thesis we had to decide which form to use. Since in our opinion the latter one is easier to write and more popular in the optimization community we will present all the results in the form of the Farkas

theorem, i.e., a theorem of the alternative. All the results in this chapter can be converted into duality theorems or optimality conditions using the chart in §2.5.

The theorem we present here was first proved by Yakubovich [133, 135] in 1971.

Theorem 3.2.1 (S-lemma). *Let $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ be real quadratic functions and suppose that there is an $\bar{x} \in \mathbb{R}^n$ such that $g(\bar{x}) < 0$. Then the following two statements are equivalent.*

(i) *There is no $x \in \mathbb{R}^n$ such that*

$$\begin{aligned} f(x) &< 0 \\ g(x) &\leq 0. \end{aligned} \tag{3.2.2a}$$

(ii) *There is a non-negative number $y \geq 0$ such that*

$$f(x) + yg(x) \geq 0, \forall x \in \mathbb{R}^n. \tag{3.2.2b}$$

3.2.2 The traditional approach

Yakubovich used the following convexity result to prove the S-lemma. It was first published by Dines in 1941, see [38].

Proposition 3.2.2 (Convexity of the 2D quadratic image of \mathbb{R}^n).

If $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ are homogeneous⁴ quadratic functions then the set $\mathcal{M} = \{(f(x), g(x)) : x \in \mathbb{R}^n\} \subset \mathbb{R}^2$ is convex.

Proof. We will verify the definition of convexity directly. Let us take two points, $u = (u_f, u_g)$ and $v = (v_f, v_g)$. If these two points and the origin are collinear then obviously the line segment between u and v belongs to \mathcal{M} , since the functions are homogeneous. From now on we will assume that these points are not collinear with the origin. Since they belong to \mathcal{M} there are points $x_u, x_v \in \mathbb{R}^n$ such that

$$u_f = f(x_u), \quad u_g = g(x_u) \tag{3.2.3a}$$

$$v_f = f(x_v), \quad v_g = g(x_v). \tag{3.2.3b}$$

We can further assume without loss of generality that

$$v_f u_g - u_f v_g = d^2 > 0, \tag{3.2.4}$$

⁴By a homogeneous quadratic function we mean a function without linear and constant terms, i.e., a function of the form $x^T M x$ with a square symmetric matrix M .

otherwise we change the sign of f . Let $\lambda \in (0, 1)$ be a constant. We try to show that there exists an $x_\lambda \in \mathbb{R}^n$ such that

$$(f(x_\lambda), g(x_\lambda)) = (1 - \lambda)u + \lambda v. \quad (3.2.5)$$

Let us look for x_λ in the form⁵

$$x_\lambda = \rho(x_u \cos \theta + x_v \sin \theta), \quad (3.2.6)$$

where ρ and θ are real variables. Substituting these into the defining equation of x_λ we get:

$$\rho^2 f(x_u \cos \theta + x_v \sin \theta) = (1 - \lambda)u_f + \lambda v_f \quad (3.2.7a)$$

$$\rho^2 g(x_u \cos \theta + x_v \sin \theta) = (1 - \lambda)u_g + \lambda v_g. \quad (3.2.7b)$$

Eliminating ρ^2 from these equations and expressing λ as a function of θ we get

$$\lambda(\theta) = \frac{u_g f(x_u \cos \theta + x_v \sin \theta) - u_f g(x_u \cos \theta + x_v \sin \theta)}{(u_g - v_g) f(x_u \cos \theta + x_v \sin \theta) - (u_f - v_f) g(x_u \cos \theta + x_v \sin \theta)}. \quad (3.2.8)$$

Here the denominator of $\lambda(\theta)$ is a quadratic function of $\cos \theta$ and $\sin \theta$, let us denote it by $T(\cos \theta, \sin \theta) = \alpha \cos^2 \theta + \beta \sin^2 \theta + 2\gamma \cos \theta \sin \theta$. Computing $T(0)$, $T(\pm\pi/2)$, and using (3.2.3) we get that $\alpha = \beta = d^2 > 0$, thus $T(\cos \theta, \sin \theta) = d^2 + \gamma \sin(2\theta)$. If $\gamma \geq 0$ then $T(\cos \theta, \sin \theta) > 0$ for $\theta \in [0, \pi/2]$, and similarly, if $\gamma \leq 0$ then $T(\cos \theta, \sin \theta) > 0$ for $\theta \in [-\pi/2, 0]$. We can assume without loss of generality that the former holds, then $\lambda(\theta)$ is defined on the whole interval $[0, \frac{\pi}{2}]$ and is also continuous. Since $\lambda(0) = 0$ and $\lambda(\frac{\pi}{2}) = 1$, we can find a value $\theta_\lambda \in [0, \frac{\pi}{2}]$ such that $\lambda(\theta_\lambda) = \lambda$. Using this θ_λ we get ρ from (3.2.7) and the desired vector x_λ from (3.2.6). This completes the proof. \square

Yakubovich used this result to prove Theorem 3.2.1.

Proof. (Yakubovich, 1971, [133]) It is obvious that (ii) implies (i). On the other hand let us assume (i) and try to prove (ii). First let f and g be homogeneous functions, then by Proposition 3.2.2 the two-dimensional image of \mathbb{R}^n under the mapping (f, g) is convex, and by (i) this image does not intersect the

⁵This is simply a clever parameterization of the plane spanned by x_u and x_v . If we used the form $x_\lambda = px_u + qx_v$, where $p, q \in \mathbb{R}$, then we could reduce the problem to a 2×2 convexity problem, which can be solved by an elementary but tedious analysis. This is the idea of the proof of the S-lemma in [16].

convex cone $\mathcal{C} = \{(u_1, u_2) : u_1 < 0, u_2 \leq 0\} \subset \mathbb{R}^2$, thus they can be separated by a line. This means that there are real numbers y_1 and y_2 such that

$$y_1 u_1 + y_2 u_2 \leq 0, \forall (u_1, u_2) \in \mathcal{C} \quad (3.2.9a)$$

$$y_1 f(x) + y_2 g(x) \geq 0, \forall x \in \mathbb{R}^n. \quad (3.2.9b)$$

Taking $(-1, 0) \in \mathcal{C}$ we have $y_1 \geq 0$ and setting $(-\varepsilon, -1) \in \mathcal{C}$ where ε is arbitrarily small gives $y_2 \geq 0$. The case $y_1 = 0$ can be ruled out by substituting \bar{x} in the second equation, so we have $y_1 > 0$. Letting $y = y_2/y_1 \geq 0$ then satisfies (ii).

Now let f and g be general, not necessarily homogeneous quadratic functions satisfying (i). First notice that we can assume $\bar{x} = 0$, if this is not the case then let $\bar{g}(x) = g(x + \bar{x})$ be our new function. Let the functions be defined as

$$f(x) = x^T A_f x + b_f^T x + c_f, \quad (3.2.10a)$$

$$g(x) = x^T A_g x + b_g^T x + c_g, \quad (3.2.10b)$$

then the Slater condition is equivalent to $g(0) = c_g < 0$. Let us introduce the homogeneous version of our functions

$$\tilde{f} : \mathbb{R}^{n+1} \rightarrow \mathbb{R}, \quad \tilde{f}(x, \tau) = x^T A_f x + \tau b_f^T x + \tau^2 c_f \quad (3.2.11a)$$

$$\tilde{g} : \mathbb{R}^{n+1} \rightarrow \mathbb{R}, \quad \tilde{g}(x, \tau) = x^T A_g x + \tau b_g^T x + \tau^2 c_g. \quad (3.2.11b)$$

Now we prove that the new functions satisfy (i), i.e., there is no $(x, \tau) \in \mathbb{R}^{n+1}$ such that

$$\tilde{f}(x, \tau) < 0 \quad (3.2.12a)$$

$$\tilde{g}(x, \tau) \leq 0. \quad (3.2.12b)$$

Let us assume on the contrary that there is an $(x, \tau) \in \mathbb{R}^{n+1}$ with these properties. If $\tau \neq 0$ then

$$f(x/\tau) = \tilde{f}(x, \tau)/\tau^2 < 0, \quad (3.2.13a)$$

$$g(x/\tau) = \tilde{g}(x, \tau)/\tau^2 \leq 0, \quad (3.2.13b)$$

which contradicts (i). If $\tau = 0$ then $x^T A_f x < 0$ and $x^T A_g x \leq 0$, therefore

$$\underbrace{(\lambda x)^T A_f (\lambda x)}_{<0} + \lambda b_f^T x + c_f < 0, \text{ if } |\lambda| \text{ is large enough, and} \quad (3.2.14a)$$

$$\underbrace{(\lambda x)^T A_g (\lambda x)}_{\leq 0} + \lambda b_g^T x + \underbrace{c_g}_{<0} < 0, \text{ if } \lambda \text{ has the proper sign,} \quad (3.2.14b)$$

contradicting (i). This implies that the new system (3.2.12) is not solvable. Further, taking $(0, 1)$ gives

$$\tilde{g}(0, 1) = g(0) < 0, \quad (3.2.15)$$

therefore the Slater condition is satisfied so we can apply the already proved homogeneous version of the theorem. We find that there exists a $y \geq 0$ such that

$$\tilde{f}(x, \tau) + y\tilde{g}(x, \tau) \geq 0, \quad \forall (x, \tau) \in \mathbb{R}^{n+1}, \quad (3.2.16)$$

and with $\tau = 1$ we get (ii). \square

Problems similar to Proposition 3.2.2 were first investigated by Hausdorff [60] and Toeplitz [124] in the late 1910s in a more general context: the joint numerical range of Hermitian operators. The importance of this simple fact becomes more obvious if we recall that the S-lemma is actually a non-convex theorem of alternatives, an extended version of the Farkas theorem.

3.2.3 A modern approach

This proof is similar to the one found in [16], but extends it for the nonhomogeneous case. The following lemma from [122] plays a crucial role in this theory:

Lemma 3.2.3 (Semidefinite rank-1 decomposition). *Let $G, X \in \mathbb{R}^{n \times n}$ be symmetric matrices X being positive semidefinite and rank r . Then $G \bullet X \leq 0$ ($G \bullet X = 0$) if and only if there are $p^1, \dots, p^r \in \mathbb{R}^n$ such that*

$$X = \sum_{i=1}^r p^i p^{iT} \quad \text{and} \quad G \bullet p^i p^{iT} = p^{iT} G p^i \leq 0 \quad (p^{iT} G p^i = 0), \quad \forall i = 1, \dots, r. \quad (3.2.17)$$

Proof. We prove only the first version of the lemma, the second one is very similar. The proof is based on [122] and is constructive. Consider Algorithm 3.1.

If the iterations take the first branch of the **if** construct then $w^{iT} G w^i$ has the same sign for all $i = k, \dots, r$. Since the sum of these terms is negative we have $p^{kT} G p^k \leq 0$ and $W - p^k p^{kT} = \sum_{i=k+1}^r w^i w^{iT}$ implying that the remaining matrix is positive semidefinite and has rank $r - 1$.

If it takes the **else** branch then the quadratic equation for α must have two distinct roots, and by definition $0 = p^{kT} G p^k \geq G \bullet W$. Finally, we can see that

$$W - p^k p^{kT} = uu^T + \sum_{i \in \{k, k+1, \dots, r\} \setminus j} w^i w^{iT}, \quad (3.2.19a)$$

Algorithm 3.1 The semidefinite decomposition algorithm

Input: X and $G \in \mathbb{R}^{n \times n}$ such that $X \succeq 0$ and $G \bullet X \leq 0$, $\text{rank}(X) = r$.

$W = X$

for $k = 1$ to r

Compute the rank-1 decomposition of W , i.e., $W = \sum_{i=k}^r w^i w^{iT}$

if $(w^{kT} G w^k)(w^{iT} G w^i) \geq 0$ for all $i = k + 1, \dots, r$ **then**

return $p^k = w^k$.

else

We have a j such that $(w^{kT} G w^k)(w^{jT} G w^j) < 0$. Since $w^{kT} G w^k$ and $w^{jT} G w^j$ have opposite sign we must have an $\alpha \in \mathbb{R}$ such that

$$(w^k + \alpha w^j)^T G (w^k + \alpha w^j) = 0. \quad (3.2.18)$$

In this case set $p^k = \frac{w^k + \alpha w^j}{\sqrt{1 + \alpha^2}}$

end if

$W = W - p^k p^{kT}$

end for

Output: p^1, \dots, p^r such that $X = \sum_{i=1}^r p^i p^{iT}$ and $p^{iT} G p^i \leq 0 \forall i = 1, \dots, r$.

where

$$u = \frac{w^j - \alpha w^k}{\sqrt{1 + \alpha^2}}. \quad (3.2.19b)$$

This decomposition ensures that $W - p^k p^{kT}$ has rank $r - 1$ and the procedure is correct. Applying the procedure r times we get the statement of the lemma.

□

Remark 3.2.4. The lemma can also be proved using the rank-1 approach presented in the proof of Theorem 3.6.1.

Now we can finish the proof of Theorem 3.2.1.

Proof. (Theorem 3.2.1, S-lemma) It is obvious that if either of the two systems has a solution then the other cannot have one, so what we have to prove is that at least one system has a solution. Let the functions be given as

$$f(x) = x^T A_f x + b_f^T x + c_f, \quad (3.2.20a)$$

$$g(x) = x^T A_g x + b_g^T x + c_g, \quad (3.2.20b)$$

and let us consider the following notation:

$$H_f = \begin{bmatrix} c_f & \frac{1}{2}b_f^T \\ \frac{1}{2}b_f & A_f \end{bmatrix}, \quad (3.2.21a)$$

$$H_g = \begin{bmatrix} c_g & \frac{1}{2}b_g^T \\ \frac{1}{2}b_g & A_g \end{bmatrix}. \quad (3.2.21b)$$

Using this notation we can rewrite the first system as:

$$\begin{aligned} H_f \bullet \begin{bmatrix} 1 & x^T \\ x & xx^T \end{bmatrix} &< 0, \\ H_g \bullet \begin{bmatrix} 1 & x^T \\ x & xx^T \end{bmatrix} &\leq 0, \end{aligned} \quad (3.2.22)$$

$x \in \mathbb{R}^n.$

Here the rank-1 matrices

$$\begin{bmatrix} 1 & x^T \\ x & xx^T \end{bmatrix} \quad (3.2.23)$$

are positive semidefinite and symmetric. This can inspire us to look at the following relaxation of (3.2.22):

$$\begin{aligned} H_f \bullet Z &< 0, \\ H_g \bullet Z &\leq 0, \\ Z &\succeq 0. \end{aligned} \quad (3.2.24)$$

The key idea of the proof is to show that this relaxation is exact in the sense that problem (3.2.22) is solvable if and only if the relaxed problem (3.2.24) is solvable. More specifically, we will prove the following lemma.

Lemma 3.2.5. *Using the notation introduced in (3.2.21)-(3.2.24), if the relaxed system (3.2.24) has a solution then it has a rank-1 solution of the form $Z = zz^T$, where the first coordinate of z is 1. This gives a solution for (3.2.22). Moreover, (3.2.24) has a solution that strictly satisfies all the inequalities including the semidefinite constraint.*

Proof. Let Z be a solution of (3.2.24). Then, since Z is positive semidefinite it can be written as

$$Z = \sum_{j=1}^r q^j q^{jT}, \quad (3.2.25)$$

where $q^j \in \mathbb{R}^{n+1}$ and $r = \text{rank}(Z)$. Applying Lemma 3.2.3 we see that it is possible to choose these vectors such that

$$H_g \bullet q^j q^{jT} = q^{jT} H_g q^j \leq 0, \quad j = 1, \dots, r. \quad (3.2.26)$$

Now, from the strict inequality of (3.2.24) we can conclude that there is a vector $q = q^j$ for some $1 \leq j \leq r$ such that

$$H_f \bullet q q^T < 0, \quad (3.2.27)$$

otherwise the sum of these terms could not be negative. This means that $Z = q q^T$ is a rank-1 solution of (3.2.24). Observe that this result was obtained without using the Slater condition.

If the first coordinate of q is nonzero then $x = q_{2:n+1}/q_1$ gives a solution for (3.2.22). If this is not the case then let us introduce

$$\tilde{q} = q + \alpha \begin{bmatrix} 1 \\ \bar{x} \end{bmatrix}, \quad (3.2.28)$$

where \bar{x} is the point satisfying the Slater condition. Notice that

$$H_f \bullet \tilde{q} \tilde{q}^T = \underbrace{H_f \bullet q q^T}_{<0} + 2\alpha H_f \bullet \begin{bmatrix} 1 \\ \bar{x} \end{bmatrix} q^T + \alpha^2 H_f \bullet \begin{bmatrix} 1 & \bar{x}^T \\ \bar{x} & \bar{x} \bar{x}^T \end{bmatrix} < 0 \quad (3.2.29a)$$

if $|\alpha|$ is small and

$$H_g \bullet \tilde{q} \tilde{q}^T = \underbrace{H_g \bullet q q^T}_{\leq 0} + 2\alpha H_g \bullet \begin{bmatrix} 1 \\ \bar{x} \end{bmatrix} q^T + \alpha^2 \underbrace{H_g \bullet \begin{bmatrix} 1 & \bar{x}^T \\ \bar{x} & \bar{x} \bar{x}^T \end{bmatrix}}_{=g(\bar{x}) < 0} < 0 \quad (3.2.29b)$$

if the sign of α is chosen to make the middle term negative. Further, if $\alpha \neq -q_1$ then $\tilde{q}_1 \neq 0$. It is obvious that these conditions can be satisfied simultaneously, i.e., we have $\tilde{q} \tilde{q}^T$ that solves (3.2.24) and $x = \tilde{q}_{2:n+1}/\tilde{q}_1$ gives a solution for (3.2.22). Finally, letting

$$\tilde{Z} = \tilde{q} \tilde{q}^T + \beta I, \quad (3.2.30)$$

where $\beta \in \mathbb{R}$ and $I \in \mathbb{R}^{(n+1) \times (n+1)}$ is the identity matrix, provides $H_f \bullet \tilde{Z}$ and $H_g \bullet \tilde{Z} < 0$ if $|\beta|$ is small enough and $\tilde{Z} \succ 0$. In other words, \tilde{Z} satisfies the strict version of all the inequalities. This completes the proof of the lemma. \square

Now we can easily finish the proof of Theorem 3.2.1. It follows directly from the Farkas Theorem (see Theorem 2.2.3) that system (3.2.24) is solvable if and only if the dual system

$$H_f + yH_g \succeq 0 \quad (3.2.31a)$$

$$y \geq 0. \quad (3.2.31b)$$

is not solvable. Now, by Lemma 3.2.5, the solvability of the original quadratic system (3.2.22) is equivalent to the solvability of its LMI relaxation (3.2.24), which—by duality—is equivalent to the non-solvability of the dual system (3.2.31). This means that there is a $y \geq 0$ such that $f(x) + yg(x) \geq 0$ for all $x \in \mathbb{R}^n$. This completes the proof of the S-lemma. \square

Quadratic systems can always be relaxed using linear matrix inequalities, so the key question asks when this relaxation is exact. This topic is further discussed in §3.6.

3.2.4 A new elementary proof

Unlike the previous two proofs, this proof is new, it is based on [137, Lemma 2.3], and it is the most elementary of the proofs presented in this section. We prove only the homogeneous version, the nonhomogeneous case can be handled similarly. We will use the following lemma first proved by Yuan in [137]:

Lemma 3.2.6 (Yuan’s lemma). *Let $A, B \in \mathbb{R}^{n \times n}$ be two symmetric matrices and let $\mathcal{F}, \mathcal{G} \subseteq \mathbb{R}^n$ be closed sets such that $\mathcal{F} \cup \mathcal{G} = \mathbb{R}^n$. If*

$$x^T A x \geq 0, \quad \forall x \in \mathcal{F} \quad (3.2.32a)$$

$$x^T B x \geq 0, \quad \forall x \in \mathcal{G} \quad (3.2.32b)$$

then there is a $t \in [0, 1]$ such that $tA + (1 - t)B$ is positive semidefinite.

Proof. The lemma is trivially true if either of the two sets is empty, so we can assume that both are non-empty. Further, we can assume that both sets are symmetric about the origin, i.e., $\mathcal{F} = -\mathcal{F}$ and $\mathcal{G} = -\mathcal{G}$. Let $\lambda(t)$ be the smallest eigenvalue of $tA + (1 - t)B$. If $\lambda(t) \geq 0$ for some $t \in [0, 1]$ then the lemma is true. Let us assume now that $\lambda(t) < 0$ for all $t \in [0, 1]$. We define the following set:

$$S(t) = \{x : (tA + (1 - t)B)x = \lambda(t)x, \|x\| = 1\}. \quad (3.2.33)$$

Since $\lambda(t)$ is an eigenvalue, $S(t)$ is not empty and it is closed by continuity, so

$$S(t) \supseteq \left\{ x : x = \lim_{k \rightarrow \infty} x_k, x_k \in S(t_k), t = \lim_{k \rightarrow \infty} t_k \right\}. \quad (3.2.34)$$

If $x \in S(0)$, then $x^T Bx = \lambda(0) < 0$, thus $x \notin \mathcal{G}$, so $x \in \mathcal{F}$. This shows that $S(0) \subseteq \mathcal{F}$, thus $S(0) \cap \mathcal{F} = S(0) \neq \emptyset$. Let t_{\max} be the largest number in $[0, 1]$ such that $S(t_{\max}) \cap \mathcal{F} \neq \emptyset$, this number exists due to (3.2.34). If $t_{\max} = 1$ then by the assumptions on \mathcal{G} , we have $S(1) \cap \mathcal{G} = S(1) \neq \emptyset$. If $t_{\max} < 1$ then for every $t \in (t_{\max}, 1]$ we have

$$S(t) \cap \mathcal{G} = (S(t) \cap \mathcal{G}) \cup \underbrace{(S(t) \cap \mathcal{F})}_{=\emptyset} = S(t) \neq \emptyset, \quad (3.2.35)$$

where we use the assumption that $\mathcal{F} \cup \mathcal{G} = \mathbb{R}^n$. Again, using (3.2.34) we get that

$$S(t_{\max}) \cap \mathcal{G} \neq \emptyset. \quad (3.2.36)$$

Since $S(t)$ is the intersection of a subspace (the eigenspace of $\lambda(t)$) and the unit ball, it is a unit ball in some dimension, therefore $S(t)$ is either connected, or it is the union of two points, symmetric about the origin.

If $S(t_{\max})$ is a connected ball, then any path connecting a point in $S(t_{\max}) \cap \mathcal{F}$ with a point in $S(t_{\max}) \cap \mathcal{G}$ contains a point in $S(t_{\max}) \cap \mathcal{F} \cap \mathcal{G}$, since both \mathcal{F} and \mathcal{G} are closed. This shows that $S(t_{\max}) \cap \mathcal{F} \cap \mathcal{G} \neq \emptyset$, thus there exists an $x \in \mathcal{F} \cap \mathcal{G}$ such that $x^T (tA + (1-t)B)x = \lambda(t) < 0$, but then either $x^T Ax < 0$ or $x^T Bx < 0$, contradicting (3.2.32).

If on the other hand $S(t_{\max})$ consists of two points then since $\mathcal{F} = -\mathcal{F}$ and $\mathcal{G} = -\mathcal{G}$, we have $S(t_{\max}) \subseteq \mathcal{F}$ and $S(t_{\max}) \subseteq \mathcal{G}$, thus we reach the same conclusion. This completes the proof of Lemma 3.2.6. \square

Now the S-lemma (Theorem 3.2.1) can be proved easily. Let A and B be symmetric matrices and assume that the system

$$x^T Ax < 0 \quad (3.2.37a)$$

$$x^T Bx \leq 0 \quad (3.2.37b)$$

is not solvable, but the Slater condition is satisfied, i.e., $\exists \bar{x} \in \mathbb{R}^n : \bar{x}^T B\bar{x} < 0$. Defining the closed sets

$$\mathcal{F} = \{x : x^T Bx \leq 0\}$$

$$\mathcal{G} = \{x : x^T Bx \geq 0\}$$

one has $\mathcal{F} \cup \mathcal{G} = \mathbb{R}^n$. By the assumption of nonsolvability we have that $x^T A x \geq 0 : \forall x \in \mathcal{F}$ and $x^T B x \geq 0 : \forall x \in \mathcal{G}$, thus all the conditions of Lemma 3.2.6 are satisfied and we can conclude that there is a $t \in [0, 1]$ such that $tA + (1 - t)B$ is positive semidefinite. Now t cannot be 0, otherwise B would be positive semidefinite and the Slater condition could not be satisfied. Dividing by t we get that $A + \frac{1-t}{t}B$ is positive semidefinite.

Remark 3.2.7. In the proof of Lemma 3.2.6 we used little about the quadratic nature of the functions, this gives an incentive to try to extend this lemma for more general functions.

3.3 Special results and counterexamples

In this section we present some related results and counterexamples.

3.3.1 Other variants

For the sake of completeness we enumerate other useful forms of the basic S-lemma. One can get these results by modifying the original proof slightly. For references see [70, 80].

Proposition 3.3.1 (S-lemma with equality).

Let $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ be quadratic functions where g is assumed to be strictly concave (or strictly convex) and let us assume a stronger form of the Slater condition, namely g takes both positive and negative values. Then the following two statements are equivalent:

(i) *The system*

$$f(x) < 0 \tag{3.3.1a}$$

$$g(x) = 0 \tag{3.3.1b}$$

is not solvable.

(ii) *There exists a multiplier $y \in \mathbb{R}$ such that*

$$f(x) + yg(x) \geq 0, \quad \forall x \in \mathbb{R}^n. \tag{3.3.2}$$

In the presence of two non-strict inequalities we have the following result.

Proposition 3.3.2 (S-lemma with non-strict inequalities).

Let $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ be homogeneous quadratic functions. The following two statements are equivalent.

(i) The system

$$f(x) \leq 0 \tag{3.3.3a}$$

$$g(x) \leq 0 \tag{3.3.3b}$$

is not solvable.

(ii) There exist nonnegative multipliers $y_1, y_2 \geq 0$ such that

$$y_1 f(x) + y_2 g(x) > 0, \quad \forall x \in \mathbb{R}^n \setminus \{0\}. \tag{3.3.4}$$

If we assume the Slater condition for one of the functions then we can make the corresponding multiplier positive.

3.3.2 General results

In this section we present some known results on how the solvability of the system

$$f(x) < 0 \tag{3.3.5a}$$

$$g_i(x) \leq 0, \quad i = 1, \dots, m \tag{3.3.5b}$$

$$x \in \mathbb{R}^n \tag{3.3.5c}$$

and the existence of a dual vector $y = (y_1, \dots, y_m) \geq 0$ such that

$$f(x) + \sum_{i=1}^m y_i g_i(x) \geq 0 \quad \forall x \in \mathbb{R}^n \tag{3.3.6}$$

are related to each other. We will assume that

$$f(x) = x^T A x + b^T x + c \tag{3.3.7a}$$

$$g_i(x) = x^T B_i x + p_i^T x + q_i, \quad i = 1, \dots, m \tag{3.3.7b}$$

where A and B_i are symmetric but not necessarily positive semidefinite matrices. Unfortunately, the S-lemma is not true in this general setting. It fails to hold even if we restrict ourselves to $m = 2$ and assume the Slater condition. This does not prevent us from studying the idea of the S-procedure even when

the procedure is theoretically not exact. It is trivial that the two systems in the S-lemma cannot be solved simultaneously, so if we are lucky enough to find a solution for the second system then we can be sure that the first system is not solvable. However, the non-solvability of the second system does not always guarantee the solvability of the first one.

First let me present what can be found in the literature about this general setting. My main sources are [16] and [122]. We will outline the proofs as necessary.

Let us start with a general result that contains the S-lemma as a special case.

Theorem 3.3.3 (Quadratic systems – nonnegative eigenvalues).

Consider the systems (3.3.5)-(3.3.6) and let us assume that the functions f and $g_i, i = 1, \dots, m$ are all homogeneous and $m \leq n$. If system (3.3.5) is not solvable then there exist a nonnegative vector $y = (y_1, \dots, y_m) \geq 0$ and an $(n - m + 1)$ -dimensional subspace $V^{n-m+1} \subseteq \mathbb{R}^n$ such that

$$f(x) + \sum_{i=1}^m y_i g_i(x) \geq 0 \quad \forall x \in V^{n-m+1}. \quad (3.3.8)$$

In other words, the matrix

$$A + \sum_{i=1}^m y_i B_i \quad (3.3.9)$$

has at least $n - m + 1$ nonnegative eigenvalues (counted with multiplicities) and the above subspace is spanned by the corresponding eigenvectors.

Remark 3.3.4. If $m = 1$ this gives the usual S-lemma.

The proof of this theorem is based on differential geometric arguments, see [16, §4.10]. Despite the relative strength of the theorem, it is not straightforward to apply it in practice. One such way is to exploit the possible structure of the linear combination and rule out a certain number of negative eigenvalues.

Besides this general result we have only very special ones.

Proposition 3.3.5. *If A and $B_i, i = 1, \dots, m$ are all*

- *diagonal matrices, i.e., $f(x)$ and $g_i(x)$ are all weighted sums of squares, or*
- *linear combinations of two fixed matrices, i.e., $\text{rank}(A, B_1, \dots, B_m) \leq 2$,*

then the system

$$f(x) < 0 \quad (3.3.10a)$$

$$g_i(x) \leq 0, \quad i = 1, \dots, m \quad (3.3.10b)$$

$$x \in \mathbb{R}^n \quad (3.3.10c)$$

is not solvable if and only if there exists a nonnegative vector $y = (y_1, \dots, y_m)$ such that

$$f(x) + \sum_{i=1}^m y_i g_i(x) \geq 0 \quad \forall x \in \mathbb{R}^n. \quad (3.3.10d)$$

In other words, the matrix

$$A + \sum_{i=1}^m y_i B_i \quad (3.3.10e)$$

is positive semidefinite.

The first part of this proposition can be proved easily using the substitution $z_i = x_i^2$ and applying the Farkas Lemma. The second part is a new result, which we will prove in §3.9, see Theorem 3.9.1.

If we want to incorporate more quadratic constraints, we need extra conditions.

Proposition 3.3.6 (S-lemma, 3 inequalities). *Let f , g_1 and g_2 be homogeneous quadratic functions and assume that the Slater condition holds, i.e., there is an $\bar{x} \in \mathbb{R}^n$ such that $g_1(\bar{x}), g_2(\bar{x}) < 0$. If either*

- $m = 2$, $n \geq 3$ and there is a positive definite linear combination of A , B_1 and B_2 , or
- $m = n = 2$ and there is a positive definite linear combination of B_1 and B_2 ,

then the following two statements are equivalent:

(i) The system

$$f(x) < 0 \quad (3.3.11a)$$

$$g_i(x) \leq 0, \quad i = 1, 2 \quad (3.3.11b)$$

$$x \in \mathbb{R}^n \quad (3.3.11c)$$

is not solvable.

(ii) *There are nonnegative numbers y_1 and y_2 such that*

$$f(x) + y_1 g_1(x) + y_2 g_2(x) \geq 0 \quad \forall x \in \mathbb{R}^n. \quad (3.3.12)$$

Remark 3.3.7. The condition $n \geq 3$ in the first part is necessary, see the counterexample for $n = 2$ in §3.3.3.

Remark 3.3.8. An equivalent condition on when some matrices have a positive definite linear combination is given in [39]. If $n \geq 3$ then the property that two symmetric matrices have a positive definite linear combination is equivalent to the nonexistence of a common root of the quadratic forms, see [48]. Further, in this case the matrices are simultaneously diagonalizable by a real congruence, see [6, 100, 113, 129].⁶ In the general case, symmetric matrices A_1, \dots, A_m have a positive definite linear combination if and only if $A_i \bullet S = 0, i = 1, \dots, m$ implies that S is indefinite. This result is a trivial corollary of the duality theory of convex optimization, but was only rediscovered around the middle of the 20th century, see [39].

One additional step is to include linear constraints. First, we include some equalities:

Proposition 3.3.9 (S-lemma with linear equalities). *Let $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ be quadratic functions, $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Assume that there exists an $\bar{x} \in \mathbb{R}^n$ such that $A\bar{x} = b$ and $g(\bar{x}) < 0$. The following two statements are equivalent:*

(i) *The system*

$$f(x) < 0 \quad (3.3.13a)$$

$$g(x) \leq 0 \quad (3.3.13b)$$

$$Ax = b \quad (3.3.13c)$$

$$x \in \mathbb{R}^n \quad (3.3.13d)$$

is not solvable.

(ii) *There is a nonnegative number y such that*

$$f(x) + yg(x) \geq 0 \quad \forall x \in \mathbb{R}^n, Ax = b. \quad (3.3.13e)$$

⁶For a complete review of simultaneous diagonalizability see [66, Theorem 7.6.4.] or [126]. For computational methods see [31, 51].

With some convexity assumption we can include a linear inequality:

Proposition 3.3.10 (S-lemma with a linear inequality).

Let $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ be quadratic functions, $c \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Assume that $g(x)$ is convex and there exists an $\bar{x} \in \mathbb{R}^n$ such that $c^T \bar{x} < b$ and $g(\bar{x}) < 0$. The following two statements are equivalent:

(i) The system

$$f(x) < 0 \quad (3.3.14a)$$

$$g(x) \leq 0 \quad (3.3.14b)$$

$$c^T x \leq b \quad (3.3.14c)$$

$$x \in \mathbb{R}^n \quad (3.3.14d)$$

is not solvable.

(ii) There is a nonnegative multiplier $y \geq 0$ and a vector $(u_0, u) \in \mathbb{R}^{n+1}$ such that

$$f(x) + yg(x) + (u^T x - u_0)(c^T x - b) \geq 0, \quad \forall x \in \mathbb{R}^n, \quad (3.3.15)$$

$$u^T x \geq 0, \quad \forall x \in \mathbb{R}^n : x^T A_g x \leq 0, b_g^T x \leq 0$$

$$u^T x - u_0 \geq 0, \quad \forall x \in \mathbb{R}^n : g(x) \leq 0, c_g + b_g^T x \leq 0,$$

where $g(x) = x^T A_g x + b_g^T x + c_g$.

For a proof see [122]. We can see that including even one linear inequality is difficult, and the dual problem (3.3.15) is not any easier than the primal problem (3.3.14).

3.3.3 Counterexamples

In this section we present some counterexamples.

More inequalities

The generalization to the case $m \geq 3$ is practically hopeless as is illustrated by the following example taken from [16]. Consider the matrices

$$A = \begin{pmatrix} 1 & 1.1 & 1.1 \\ 1.1 & 1 & 1.1 \\ 1.1 & 1.1 & 1 \end{pmatrix}, \quad B_1 = \begin{pmatrix} -2.1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad (3.3.16a)$$

$$B_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -2.1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad B_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2.1 \end{pmatrix}. \quad (3.3.16b)$$

Lemma 3.3.11. *There is an \bar{x} such that $\bar{x}^T B_i \bar{x} < 0$, $i = 1, 2, 3$ (the Slater condition holds). Further, the system*

$$x^T A x < 0 \tag{3.3.17a}$$

$$x^T B_i x \leq 0, \quad i = 1, 2, 3 \tag{3.3.17b}$$

is not solvable in \mathbb{R}^3 .

Proof. The Slater condition is satisfied with $\bar{x} = (1, 1, 1)$. Let us now look for a solution in the form $x = (x_1, x_2, x_3)$. If $x_3 = 0$ then by the last three inequalities we can conclude that $x_1 = x_2 = 0$, which does not satisfy the first inequality. Since all the functions are homogeneous, and since x and $-x$ are essentially the same solution we can assume that $x_3 = 1$, thus we can reduce the dimension of the problem. Now all four of the inequalities define some quadratic areas in \mathbb{R}^2 . Instead of a formal and tedious proof we simply plot

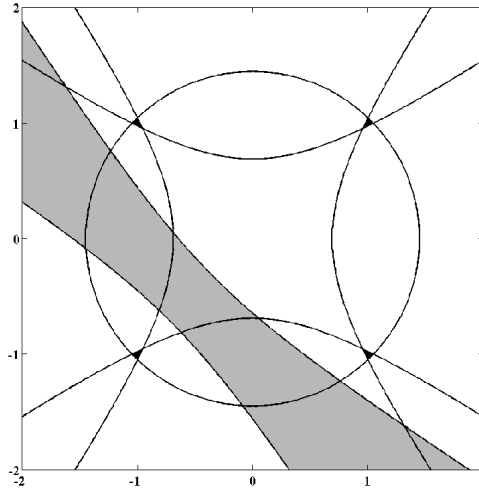


Figure 3.1: The quadratic regions defined by system (3.3.16) with $x_3 = 1$.

these areas in Fig. 3.1. The grey area represents the set of points (x_1, x_2) where $x = (x_1, x_2, 1)$ satisfies $x^T A x < 0$, while the four black corners are the feasible set for the remaining three inequalities. Intuitively, the last three inequalities are satisfied for values close to ± 1 . However, it is easy to see that such values cannot satisfy the first inequality. \square

Lemma 3.3.12. *There are no nonnegative multipliers y_1, y_2, y_3 for which $A + y_1B_1 + y_2B_2 + y_3B_3 \succeq 0$.*

Proof. Consider the matrix

$$X = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}, \quad (3.3.18)$$

which is positive semidefinite with eigenvalues 0 and 3. Moreover,

$$A \bullet X = -0.6 < 0 \quad (3.3.19a)$$

$$B_i \bullet X = -0.2 \leq 0, \quad i = 1, 2, 3. \quad (3.3.19b)$$

Now for any nonnegative linear combination of the matrices, we have

$$(A + y_1B_1 + y_2B_2 + y_3B_3) \bullet X = -0.6 - 0.2(y_1 + y_2 + y_3) < 0, \quad (3.3.20)$$

therefore $A + y_1B_1 + y_2B_2 + y_3B_3$ can never be positive semidefinite, since the scalar product of positive semidefinite matrices is nonnegative. \square

These two lemmas show that neither of the alternatives is true in the general theorem.

The $n = 2$ case

Discussing the $m = 2$ case we noted that the result fails to hold if $n = 2$. Here is a counterexample taken from [16] to demonstrate this.

Let us consider the following three matrices:

$$A = \begin{pmatrix} \lambda\mu & 0.5(\mu - \lambda) \\ 0.5(\mu - \lambda) & -1 \end{pmatrix} \quad (3.3.21a)$$

$$B = \begin{pmatrix} -\mu\nu & -0.5(\mu - \nu) \\ -0.5(\mu - \nu) & 1 \end{pmatrix} \quad (3.3.21b)$$

$$C = \begin{pmatrix} -\lambda^2 & 0 \\ 0 & 1 \end{pmatrix}. \quad (3.3.21c)$$

We will verify the following claims:

Proposition 3.3.13. *With the proper choice of the parameters we have the following results.*

- (i) *There is a positive definite linear combination of A, B and C .*

(ii) *There is a vector \bar{x} such that $\bar{x}^T B \bar{x} < 0$ and $\bar{x}^T C \bar{x} < 0$.*

(iii) *The quadratic system*

$$x^T A x < 0 \quad (3.3.22a)$$

$$x^T B x \leq 0 \quad (3.3.22b)$$

$$x^T C x \leq 0 \quad (3.3.22c)$$

is not solvable.

(iv) *There are no $y_1, y_2 \geq 0$ such that*

$$A + y_1 B + y_2 C \succeq 0. \quad (3.3.23)$$

Proof. Let $\lambda = 1.1$, $\mu = 0.818$ and $\nu = 1.344$.

(i) *The linear combination*

$$-1.15A - 0.005B - C = \begin{pmatrix} 0.18072696 & 0.160835 \\ 0.160835 & 0.1450 \end{pmatrix} \quad (3.3.24)$$

is positive definite as it is seen by the diagonal elements and the determinant.

(ii) Let $\bar{x} = (1, 0)^T$ then $\bar{x}^T B \bar{x} < 0$ and $\bar{x}^T C \bar{x} < 0$.

(iii) Let us exploit the special structure of the matrices. If we are looking for a solution $x = (x_1, x_2) \in \mathbb{R}^2$ then we get

$$x^T A x = \lambda \mu x_1^2 - x_2^2 + (\mu - \lambda)x_1 x_2 = (\lambda x_1 + x_2)(\mu x_1 - x_2) \quad (3.3.25a)$$

$$x^T B x = -\mu \nu x_1^2 + x_2^2 - (\mu - \nu)x_1 x_2 = (\nu x_1 + x_2)(-\mu x_1 + x_2) \quad (3.3.25b)$$

$$x^T C x = -\lambda^2 x_1^2 + x_2^2 = (-\lambda x_1 + x_2)(\lambda x_1 + x_2). \quad (3.3.25c)$$

Now, in order to satisfy (3.3.22) we need to solve one of the following two systems corresponding to which terms are negative and positive:

$$\begin{array}{ll} \lambda x_1 + x_2 > 0, & \lambda x_1 + x_2 < 0, \\ \mu x_1 - x_2 < 0, & \mu x_1 - x_2 > 0, \\ \nu x_1 + x_2 \leq 0, & \nu x_1 + x_2 \geq 0, \\ -\lambda x_1 + x_2 \leq 0, & -\lambda x_1 + x_2 \geq 0. \end{array} \quad (3.3.26)$$

It is easy to check that with the values specified in the statement both of these systems are inconsistent, and therefore (3.3.22) is not solvable.

(iv) The proof of this part is similar to the proof of Lemma 3.3.12. The matrix

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (3.3.27)$$

satisfies

$$A \bullet X < 0 \quad (3.3.28a)$$

$$B \bullet X \leq 0 \quad (3.3.28b)$$

$$C \bullet X \leq 0, \quad (3.3.28c)$$

thus no nonnegative linear combination $A + y_1 B + y_2 C$ of A , B and C is positive semidefinite.

This completes the proof of the lemma. \square

Let us examine briefly why the S-lemma fails to hold for this example. We have already shown in Proposition 3.3.6 that if $n = m = 2$ then in order for the result to hold we need to assume that a certain linear combination of B and C is positive definite. However, taking the positive definite matrix

$$X = \begin{pmatrix} 1 & \frac{\lambda^2 - \mu\nu}{\mu - \nu} \\ \frac{\lambda^2 - \mu\nu}{\mu - \nu} & \lambda^2 \end{pmatrix} \quad (3.3.29)$$

yields

$$B \bullet X = 0 \quad (3.3.30a)$$

$$C \bullet X = 0, \quad (3.3.30b)$$

therefore no linear combination of B and C can be positive definite.

3.4 Applications

3.4.1 Stability analysis

The first example is based on the one presented in [70].

Let us consider the following dynamical system

$$\dot{x} = Ax + Bw, \quad x(0) = x_0 \quad (3.4.1a)$$

$$v = Cx \quad (3.4.1b)$$

with a so-called sector constraint

$$\sigma(v, w) = (\beta v - w)^T(w - \alpha v) \geq 0, \quad (3.4.1c)$$

where $\alpha < \beta$ are real numbers. We would like to use the basic tool of Lyapunov functions [27]: for the quadratic stability of the system it is necessary and sufficient to have a symmetric matrix P such that $V(x) = x^T P x$ is a Lyapunov function, i.e.,

$$\dot{V}(x) = 2x^T P(Ax + Bw) < 0, \quad \forall (x, w) \neq 0 \text{ s.t. } \sigma(Cx, w) \geq 0. \quad (3.4.2)$$

Introducing the quadratic forms

$$\begin{aligned} \sigma_0(x, w) &= \begin{bmatrix} x \\ w \end{bmatrix}^T \begin{bmatrix} A^T P + P A & P B \\ B^T P & 0 \end{bmatrix} \begin{bmatrix} x \\ w \end{bmatrix} \\ \sigma_1(x, w) &= 2\sigma(Cx, w) = \begin{bmatrix} x \\ w \end{bmatrix}^T \begin{bmatrix} -2\beta\alpha C^T C & (\beta + \alpha)C^T \\ (\beta + \alpha) & -2 \end{bmatrix} \begin{bmatrix} x \\ w \end{bmatrix} \end{aligned} \quad (3.4.3)$$

the Lyapunov condition can be written as

$$\sigma_0(x, w) < 0, \quad \forall (x, w) \neq 0 \text{ s.t. } \sigma_1(x, w) \geq 0, \quad (3.4.4)$$

or in other words, we have to decide the solvability of the quadratic system

$$\sigma_0(x, w) \geq 0 \quad (3.4.5a)$$

$$\sigma_1(x, w) \geq 0. \quad (3.4.5b)$$

Using $\alpha < \beta$ we can see that the strict version of the second inequality can be satisfied. Based on a suitable form of the S-lemma (Proposition 3.3.2) we find that the non-solvability of this system is equivalent to the existence of $y \geq 0$ such that

$$\sigma_0(x, w) + y\sigma_1(x, w) < 0 \quad \forall (x, w) \neq (0, 0). \quad (3.4.6)$$

Now $y = 0$ would imply that σ_0 is negative definite and would contradict the non-solvability of (3.4.5). Thus we can divide by y and use P to denote P/y . Finally, we can state the criterion using the LMI formulation. We get the following theorem (cf. [27]):

Theorem 3.4.1 (Circle criterion). *A necessary and sufficient condition for the quadratic stability of the system*

$$\dot{x} = Ax + Bw, \quad x(0) = x_0 \quad (3.4.7a)$$

$$v = Cx \quad (3.4.7b)$$

$$\sigma(v, w) = (\beta v - w)^T(w - \alpha v) \geq 0, \quad (3.4.7c)$$

where $\alpha < \beta$, is the existence of a symmetric matrix P such that

$$\begin{bmatrix} A^T P + PA - 2\beta\alpha C^T C & PB + (\beta + \alpha)C^T \\ B^T P + (\beta + \alpha)C & -2 \end{bmatrix} \quad (3.4.7d)$$

is negative definite.

3.4.2 Sum of two ellipsoids

This example is taken from [105].

Let $E_i = E(a^i, A_i) \subseteq \mathbb{R}^n$, $n \geq 2$ be an ellipsoid with centre a^i and shape A_i , i.e.,

$$E(a^i, A_i) = \{x \in \mathbb{R}^n : (x - a^i)^T A_i (x - a^i) \leq 1\}. \quad (3.4.8)$$

The sum of two such ellipsoids is the usual Minkowski sum:

$$Q = E_1 + E_2 = \{x^1 + x^2 : x^1 \in E_1, x^2 \in E_2\}. \quad (3.4.9)$$

We are looking for a general description of all the ellipsoids that contain this sum. This is an important question in error estimation. Errors on measurements are usually modelled by Gaussian distributions. If we are looking for the error of the sum of two measurements then we need to solve this problem.

First notice that since

$$Q = a^1 + a^2 + E(0, A_1) + E(0, A_2) \quad (3.4.10)$$

we can assume that all the ellipsoids are centred at the origin. Our target object is an ellipsoid $E(0, A_0)$ such that

$$E(0, A_0) \supset E(0, A_1) + E(0, A_2). \quad (3.4.11)$$

This condition can be stated equivalently using the notation $x = (x^1, x^2) \in \mathbb{R}^{2n}$. The ellipsoid E_0 contains the sum $E_1 + E_2$ if and only if the following system is not solvable:

$$x^T \begin{pmatrix} -A_0 & -A_0 \\ -A_0 & -A_0 \end{pmatrix} x < 1 \quad (3.4.12a)$$

$$x^T \begin{pmatrix} A_1 & 0 \\ 0 & 0 \end{pmatrix} x \leq 1 \quad (3.4.12b)$$

$$x^T \begin{pmatrix} 0 & 0 \\ 0 & A_2 \end{pmatrix} x \leq 1. \quad (3.4.12c)$$

Since the matrix

$$\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} \quad (3.4.13)$$

is positive definite and $n \geq 2$, we can apply a slightly modified version of Proposition 3.3.6 to obtain the following characterization:

Theorem 3.4.2 (Sum of two ellipsoids). *An ellipsoid $E(0, A_0)$ contains the Minkowski sum of the ellipsoids $E(0, A_1)$ and $E(0, A_2)$ if and only if there exist nonnegative numbers y_1 and y_2 such that $y_1 + y_2 \leq 1$ and the matrix*

$$\begin{pmatrix} -A_0 + y_1 A_1 & -A_0 \\ -A_0 & -A_0 + y_2 A_2 \end{pmatrix} \quad (3.4.14)$$

is positive semidefinite.

This condition can be validated in polynomial time. Further, we can use this result to build an algorithm to minimize the maximum eigenvalue of A_0 . A similar argument can be repeated for the intersection of two ellipsoids, see [105].

Several other applications can be found in the literature, such as distance geometry [12], portfolio management [4, 5, 54], statistics [65], signal processing [78], or control and stability problems [16, 27, 86, 90], to mention just a few.

The results and examples presented in the first part show how diverse areas are contributing to the theory of the S-lemma. Now we give a summary of these connections. In each section we first present how the S-lemma is related to the specific theory then we summarize the relevant results of the field and finally we discuss the consequences of the results and draw the conclusions.

3.5 Convexity of the joint numerical range

In this section we investigate the theory behind the first proof for the S-lemma, see §3.2.2.

3.5.1 Motivation

Recall that the key step in Yakubovich's proof was to use Dines's result about the convexity of the set

$$\{(x^T A x, x^T B x) : x \in \mathbb{R}^n\} \subseteq \mathbb{R}^2. \quad (3.5.1)$$

The separation idea we used can be extended to more inequalities. Let us assume that the system

$$x^T A x < 0 \quad (3.5.2a)$$

$$x^T B_i x \leq 0, \quad i = 1, \dots, m \quad (3.5.2b)$$

is not solvable and assume that the Slater condition is satisfied, i.e., there exists an $\bar{x} \in \mathbb{R}^n$ such that $\bar{x}^T B_i \bar{x} < 0$ for all $i = 1, \dots, m$. If the set

$$H_{\mathbb{R}}(A, B_1, \dots, B_m) = \{(x^T A x, x^T B_1 x, \dots, x^T B_m x) : x \in \mathbb{R}^n\} \subseteq \mathbb{R}^{m+1} \quad (3.5.3)$$

is convex then an equivalent characterization of the non-solvability of (3.5.2) is that

$$H_{\mathbb{R}}(A, B_1, \dots, B_m) \cap \mathcal{C} \quad (3.5.4)$$

is empty, where

$$\mathcal{C} = \{(u_0, u) : u_0 < 0, u \leq 0\} \subset \mathbb{R}^{m+1} \quad (3.5.5)$$

is a convex cone. A well known basic fact in convex analysis (see [115, 120]) is that disjoint convex sets can be separated by a hyperplane, i.e., there exist $(y_0, y) \in \mathbb{R}^{m+1} \setminus (0, 0)$ such that

$$y_0 u_0 + \sum_{i=1}^m y_i u_i \geq 0, \quad \forall (u_0, u) \in H_{\mathbb{R}}, \quad (3.5.6a)$$

$$y_0 u_0 + \sum_{i=1}^m y_i u_i \leq 0, \quad \forall (u_0, u) \in \mathcal{C}. \quad (3.5.6b)$$

Since $(-1, 0) \in \mathcal{C}$ we get $y_0 \geq 0$, and using $(-\varepsilon, -e^i) \in \mathcal{C}$ where $e^i \in \mathbb{R}^m$ is the i^{th} unit vector we get $y \geq 0$.

Using the Slater condition we have a $(\bar{u}_0, \bar{u}) \in H_{\mathbb{R}}$ where $\bar{u} < 0$. If $y = 0$ then, since all the coefficients cannot be zero, we have $y_0 > 0$. On the other hand, if $y \neq 0$ then using the Slater point we have

$$y_0 \bar{u}_0 + \underbrace{\sum_{i=1}^m y_i \bar{u}_i}_{< 0} \geq 0, \quad (3.5.7)$$

implying that $y_0 > 0$. After dividing by y_0 we get the desired coefficients. We have thus proved that the convexity of $H_{\mathbb{R}}(A, B_1, \dots, B_m)$ implies the validity of the S-lemma.

3.5.2 Theoretical results

First we present results over the field of real numbers, then we generalize the concept to complex numbers.

Results over real numbers

The key question is the following: How can we guarantee the convexity of $H_{\mathbb{R}}(A, B_1, \dots, B_m)$?

Most of the results (see [8, 58, 77]) on the convexity of the numerical range investigate the question over the complex field. However, we need convexity results for $H_{\mathbb{R}}(A, B_1, \dots, B_m)$. The first such result has already been mentioned. It is credited to Dines [38] and dates back to 1941.

Theorem 3.5.1 (Convexity of the 2D quadratic image of \mathbb{R}^n).

If $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ are homogeneous quadratic functions then the set

$$\mathcal{M} = \{(f(x), g(x)) : x \in \mathbb{R}^n\} \subset \mathbb{R}^2 \quad (3.5.8)$$

is convex.

An analogous result for three quadratic forms was proved by Polyak [105] in 1998, using a theorem by Brickman [29].

Theorem 3.5.2 (Convexity of the 3D quadratic image of \mathbb{R}^n).

If $n \geq 3$ and $f, g, h : \mathbb{R}^n \rightarrow \mathbb{R}$ are homogeneous quadratic functions such that there exists a positive definite linear combination of them, then the set

$$\mathcal{M} = \{(f(x), g(x), h(x)) : x \in \mathbb{R}^n\} \subset \mathbb{R}^3 \quad (3.5.9)$$

is convex.

Another interesting source is Ramana and Goldman's article [109] from 1995. They characterized the quadratic transformations $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ for which the set $F(\mathbb{R}^n)$ is convex. They called such maps image convex (ICON) and established the following theorem:

Theorem 3.5.3 (ICON maps). Let $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a constant-free ($F(0) = 0$) quadratic map. Then F is ICON if and only if

$$F(\mathbb{R}^n) = F^Q(\mathbb{R}^n) + F(\mathbb{R}^n) \quad (3.5.10)$$

where

$$F^Q(x) = \frac{F(x) + F(-x)}{2} \quad (3.5.11)$$

is the quadratic part of F , and the sum is the Minkowski-sum.

If F is homogeneous, as it is so in our case, then the equivalent condition reduces to $F(\mathbb{R}^n) = F(\mathbb{R}^n) + F(\mathbb{R}^n)$, which is trivial.⁷ Further, they proved that the identification of ICON maps is NP-hard.

Similarly, they investigated quadratic maps under which the image of every linear subspace is convex. They called these maps linear image convex (LICON). Obviously, all LICON maps are ICON maps: thus equivalent conditions for the LICON property provide sufficient conditions for the ICON property. The following equivalent condition is presented in their paper:

Theorem 3.5.4 (LICON maps). *Let $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a constant-free ($F(0) = 0$) quadratic map, then F is LICON if and only if at least one of the following conditions holds:*

- (i) F is of the form $(u^T x + a)Ax$, or
- (ii) $\chi(F) \leq 1$, or
- (iii) F is homogeneous and

$$\text{rank}(F(x), F(y), F(x+y)) \leq 2, \forall x, y \in \mathbb{R}^n, \quad (3.5.12)$$

where $\chi(F)$ is the polynomial rank⁸ of F .

Remark 3.5.5. This theorem can be exploited to check the LICON property in polynomial time. This shows that the recognition of LICON maps is much simpler than that of the ICON maps.

A similar problem is the convexity of the image of the unit sphere. Since we are dealing with homogeneous quadratic functions, the image of the whole space is the cone spanned by the image of the unit sphere, thus if the image of the unit sphere is convex, then so is the image of the whole space. The corresponding theorem for the real case was proved by Brickman in 1961:

⁷Using the terms of §3.7 this means that $F(\mathbb{R}^n)$ is König linear.

⁸Given a polynomial map of the form

$$F(x) = \sum_{\alpha \in \mathcal{A}} x^\alpha v_\alpha, \quad (3.5.13)$$

where $\{x^\alpha : \alpha \in \mathcal{A}\}$ is the set of monomials appearing in F and $v_\alpha \in \mathbb{R}^m$, the polynomial rank of F is defined as

$$\chi(F) = \text{rank}(\{v_\alpha : \alpha \in \mathcal{A}\}). \quad (3.5.14)$$

The condition $\chi(F) \leq 1$ requires that all the vector coefficients of the terms x_i^2 , $x_i x_j$ and x_i are constant multiples of each other.

Theorem 3.5.6 (Convexity of the 2D quadratic image of the ball).

Let $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ be homogeneous quadratic functions. If $n \geq 3$ then the set

$$W_{f,g} = \{(f(x), g(x)) : x \in \mathbb{R}^n, \|x\| = 1\} \subset \mathbb{R}^2 \quad (3.5.15)$$

is convex.

Remark 3.5.7. If $n = 2$ then this set is not necessarily convex. Let us take $x = (x_1, x_2)$ and define $f(x) = x_1^2 - x_2^2$ and $g(x) = 2x_1x_2$. These functions satisfy $f(x)^2 + g(x)^2 = 1$, thus the image is the unit circle line, which is not convex.

Polyak used this result to prove his previously mentioned theorem (see Theorem 3.5.2). The original proof of this theorem uses advanced differential geometric arguments. The following elementary proof is by Pépin from [99].⁹ It uses only the characterization of quadratic surfaces in \mathbb{R}^3 .

Proof. The proof is based on the following two simple lemmas.

Lemma 3.5.8. *Let $P : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be an affine map, $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ homogeneous quadratic functions, then there are homogeneous quadratic functions $\tilde{f}, \tilde{g} : \mathbb{R}^n \rightarrow \mathbb{R}$ such that $P(W_{f,g}) = W_{\tilde{f},\tilde{g}}$.*

Proof. Let P be of the form $P(u, v) = (a_1u + b_1v + c_1, a_2u + b_2v + c_2)$, then $\tilde{f}(x) = a_1f(x) + b_1g(x) + c_1\|x\|^2$ and $\tilde{g}(x) = a_2f(x) + b_2g(x) + c_2\|x\|^2$ satisfy the requirements. \square

Lemma 3.5.9. *Let $V \subseteq \mathbb{R}^n$ be a subspace with $\dim(V) \geq 3$. If there are two points $x, y \in V$ such that $f(x) = f(y) = 0$ and $g(x)g(y) < 0$ then there is a third point $z \in V$ for which $\|z\| = 1$ and $f(z) = g(z) = 0$.*

Proof. We can assume without loss of generality that $\dim(V) = 3$. Let us define the following cone:

$$\mathcal{C} = \{x \in V : f(x) = 0\}, \quad (3.5.16)$$

then $x, y \in \mathcal{C}$, and they must be linearly independent since $g(x)g(y) < 0$. As \mathcal{C} is a three dimensional homogeneous quadratic surface, and it is not a trivial cone of one point or one direction, it can either be a second-order cone, a plane, a union of two planes or the whole subspace V . In any case, the set $\mathcal{C} \setminus \{0\}$ is either connected or it consists of two centrally symmetric connected components. Now taking $-y$ instead of y we can assume that x and y belong to the same connected component of \mathcal{C} and by the continuity of $g(x)$ we have a point u in the component satisfying $g(u) = 0$. Finally, $z = u/\|u\|$ satisfies all the requirements of the lemma. \square

⁹This reference was provided by Jean-Baptiste Hiriart-Urruty.

Using these two lemmas we can finish the proof easily. Let $V \subseteq \mathbb{R}^n$ be a subspace with $\dim(V) \geq 3$, let $f, g : V \rightarrow \mathbb{R}$ be homogeneous quadratic functions, and assume that $W_{f,g}$ is not only the origin. Let a and b be two distinct points in $W_{f,g}$ and let c be a point on the open line segment between a and b . Let x and y be the pre-images of a and b , respectively. Let us choose an affine bijection $P : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ for which $P(c) = (0, 0)$, $P(a) = (0, 1)$; then there is a $\beta < 0$ such that $P(b) = (0, \beta)$. Applying Lemma 3.5.8 we get functions \tilde{f} and \tilde{g} such that $P(W_{f,g}) = W_{\tilde{f},\tilde{g}}$. Now we have $\tilde{f}(x) = 0$, $\tilde{g}(x) = 1$, $\tilde{f}(y) = 0$ and $\tilde{g}(y) = \beta < 0$. Applying Lemma 3.5.9 we get a point $z \in V$ such that $\|z\| = 1$ and $\tilde{f}(z) = \tilde{g}(z) = 0$. This means that $(0, 0) \in W_{\tilde{f},\tilde{g}} = P(W_{f,g})$ and therefore $c = P^{-1}(0, 0) \in W_{f,g}$. This completes the proof of Theorem 3.5.6. \square

More recently, Polyak [106] proved a local version of these theorems:

Theorem 3.5.10 (Convexity of the n D quadratic image of a ball).

Consider the function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $f(x) = (f_1(x), \dots, f_m(x))$ where

$$f_i(x) = \frac{1}{2}x^T A_i x + a^{iT} x, \quad i = 1, \dots, m \quad (3.5.17)$$

are quadratic functions. Let A be an $n \times m$ matrix with columns a^i and let us define

$$L = \sqrt{\sum_{i=1}^m \|A_i\|^2} \quad (3.5.18a)$$

$$\nu = \sigma_{\min}(A) = \sqrt{\lambda_{\min}(A^T A)}, \quad (3.5.18b)$$

where $\|A_i\|$ is the operator norm of A_i , and $\sigma_{\min}(A)$ denotes the smallest singular value of A . If $\varepsilon < \nu/(2L)$ then the image $\{f(x) : x \in \mathbb{R}^n, \|x\| \leq \varepsilon\}$ is a convex set in \mathbb{R}^m .

The statement remains true if we take a small ellipsoid instead of the ball. In that case the norm constraint becomes $x^T H x < \varepsilon$, where H is some positive definite matrix. Notice that the theorem says nothing if all the functions are homogeneous, since in that case $A = 0$.

Polyak proved his result for much more general nonlinear functions in Hilbert spaces, but the general idea is the same: if f is “close” to its linear approximation then it will preserve convexity (as its linear approximation would also do).

Finally, let me mention a negative result. Under some structural assumptions it is impossible that the image of the real unit surface is convex.

Definition 3.5.11 (k^{th} joint numerical range). *Let $A = (A_1, \dots, A_m)$ be an m -tuple of $n \times n$ real symmetric matrices, and let $1 \leq k \leq n$. The set*

$$W_k^{\mathbb{R}}(A) = \{(\text{Tr}(X^T A_1 X), \dots, \text{Tr}(X^T A_m X)) : X \in \mathbb{R}^{n \times k}, X^T X = I_k\} \quad (3.5.19)$$

is called the real k^{th} joint numerical range of A (see [77]). This object is closely connected to the quadratic system presented in §3.6.4.

Theorem 3.5.12 (Nonconvexity of the quadratic image). *Assume that the matrices A_1, \dots, A_m are linearly independent. If $m > k(n - k) + 1$ then $W_k^{\mathbb{R}}(A)$ is not convex. Further, if the identity matrix is not a linear combination of A_1, \dots, A_m and $m > k(n - k)$ then $W_k^{\mathbb{R}}(A)$ is not convex.*

It is interesting to contrast this theorem with Brickman's result (Theorem 3.5.6). Brickman proved that if $n \geq 3$ then $W_1(A)$ is convex for $m = 2$. Further generalizations are blocked by Li's negative result: to prove a similar convexity theorem for $m = 3$ it is necessary (but not sufficient!) to assume $n \geq 4$, or the existence of a positive definite linear combination of the matrices and $n \geq 3$. It is surprising, though, that the convexity of the joint numerical range is a structural property, i.e., for certain values of m and n the image is convex for all possible linearly independent families of matrices, while for other values convexity is not possible at all.

If the image of the unit sphere is not convex then one might wonder how much nonconvex it is. One way to tell this is through the description of the convex hull. The Carathéodory Theorem [115] states that every point in the convex hull of an m -dimensional set can be written as the convex combination of $m + 1$ points from the set. In our case we have much stronger results due to Poon, [107].

Theorem 3.5.13. *Let $A = (A_1, \dots, A_m)$ be an m -tuple of $n \times n$ real symmetric matrices, and let $W_1^{\mathbb{R}}(A)$ denote the joint numerical range defined earlier, i.e., $W_1^{\mathbb{R}}(A)$ is the quadratic image of the n -dimensional real unit sphere. Then every point in the convex hull of $W_1^{\mathbb{R}}(A)$ can be expressed as the convex combination of at most $k^{\mathbb{R}}(m, n)$ points from $W_1^{\mathbb{R}}(A)$, where*

$$k^{\mathbb{R}}(m, n) = \min \left\{ n, \left\lfloor \frac{\sqrt{8m+1}-1}{2} \right\rfloor + \delta_{\frac{n(n+1)}{2}, m} \right\} \quad (3.5.20)$$

and $\delta_{a,b}$ is the Kronecker symbol, i.e., $\delta_{a,b} = 1$ or 0 depending on whether $a = b$ or $a \neq b$, respectively.

Results over complex numbers

Similar questions were first investigated by Hausdorff [60] and Toeplitz [124] who proved that if A, B_1, \dots, B_m are complex Hermitian matrices then the set

$$H_{\mathbb{C}}(A, B_1, \dots, B_m) = \{(z^*Az, z^*B_1z, \dots, z^*B_mz) : z \in \mathbb{C}^n, \|z\| = 1\} \subseteq \mathbb{R}^{m+1}, \quad (3.5.21)$$

where z^* is the complex conjugate-transpose, is convex if $m = 1$. This object is called the *joint numerical range* of the matrices. Later, Au-Yeung and Poon [7] proved that if $n \geq 3$ then the joint numerical range of three Hermitian matrices is also convex.

A general differential geometric characterization of the cases when the joint numerical range is convex can be found in [58]. This is the strongest known theorem for the general case.

Theorem 3.5.14 (Convexity of the joint numerical range).

Let A_1, \dots, A_m be $n \times n$ complex Hermitian matrices. If the multiplicity of the largest eigenvalue of $\sum_{i=1}^m \mu_i A_i$ is the same for all μ_1, \dots, μ_m , where $\sum_{i=1}^m \mu_i^2 = 1$, and the union of the eigenspaces corresponding to the largest eigenvalue is not the whole \mathbb{C}^n then the set

$$\{(z^*A_1z, \dots, z^*A_mz) : z \in \mathbb{C}^n, z^*z = 1\} \subseteq \mathbb{R}^m \quad (3.5.22)$$

is convex.

Remark 3.5.15. The second condition is redundant unless $m = n + 1$ and the multiplicity of the largest eigenvalue is $n/2$. Moreover, if $m \geq 4$ and the conditions of the theorem fail for some matrices but the image is still convex, then there is an arbitrarily small perturbation of A_1, \dots, A_m that destroys convexity. Thus, the above condition is almost necessary. For more details and proofs see [58].

Now let me state the complex analogue of Theorem 3.5.12:

Theorem 3.5.16 (Nonconvexity, complex case). Let $A = (A_1, \dots, A_m)$ be an m -tuple of $n \times n$ complex Hermitian matrices, and let $1 \leq k \leq n$. The set

$$W_k^{\mathbb{C}}(A) = \{(\mathrm{Tr}(X^*A_1X), \dots, \mathrm{Tr}(X^*A_mX)) : X \in \mathbb{C}^{n \times k}, X^*X = I_k\} \quad (3.5.23)$$

is the complex k^{th} joint numerical range of A (see [77]). Assume that the matrices A_1, \dots, A_m are linearly independent. If $m > 2k(n-k)+1$ then $W_k^{\mathbb{C}}(A)$ is not convex. Further, if the identity matrix is not a linear combination of A_1, \dots, A_m and $m > 2k(n-k)$ then $W_k^{\mathbb{C}}(A)$ is not convex.

Finally, we can state the complex counterpart of Theorem 3.5.13, also due to Poon, [107].

Theorem 3.5.17. *Let $A = (A_1, \dots, A_m)$ be an m -tuple of $n \times n$ complex Hermitian matrices, and let $W_1^{\mathbb{C}}(A)$ denote the joint numerical range defined earlier, i.e., $W_1^{\mathbb{C}}(A)$ is the quadratic image of the n -dimensional complex unit sphere. Every point in the convex hull of $W_1^{\mathbb{C}}(A)$ can be expressed as the convex combination of at most $k^{\mathbb{C}}(m, n)$ points from $W_1^{\mathbb{C}}(A)$, where*

$$k^{\mathbb{C}}(m, n) = \min \left\{ n, \lfloor \sqrt{m} \rfloor + \delta_{n^2, m+1} \right\}. \quad (3.5.24)$$

One might wonder why the complex case is more deeply developed than the real one. The reason for this is twofold. First, from a purely differential geometric point of view the complex field has a much nicer structure, which allows for more advanced proof techniques. Second, as we will argue later in §3.6.5, the problem is structurally simpler over the complex field.

Recently, Faybusovich [45] put all these results in a unified context and provided general proofs using Jordan algebras.

3.5.3 Implications

It is straightforward to apply these results to our case. Whenever a collection of matrices satisfies the convexity property, we can characterize the solvability of the corresponding quadratic system with a simple LMI. Moreover, we can extend the results to more general cases.

Results over real numbers

Let me start with the results over real numbers. As we have already seen in the first proof, Dines's result (Theorem 3.5.1) gives rise to the basic homogeneous S-lemma. The generalization for three inequalities (Proposition 3.3.6) comes from Polyak's convexity result (Theorem 3.5.2). The norm constrained results will give us something new. Using Theorem 3.5.6 with a simple separation idea we get the following theorem:

Theorem 3.5.18 (S-lemma with a norm constraint). *Let $n \geq 3$, and let $A, B \in \mathbb{R}^{n \times n}$ be real symmetric matrices. Assume further that there exists a Slater point $\bar{x} \in \mathbb{R}^n$ such that $\|\bar{x}\| = 1$ and $\bar{x}^T B \bar{x} < \beta$. The following two statements are equivalent:*

(i) *The system*

$$x^T Ax < \alpha \quad (3.5.25a)$$

$$x^T Bx \leq \beta \quad (3.5.25b)$$

$$\|x\| = 1 \quad (3.5.25c)$$

is not solvable.

(ii) *There is a nonnegative multiplier y such that*

$$x^T Ax - \alpha + y(x^T Bx - \beta) \geq 0, \forall x \in \mathbb{R}^n, \|x\| = 1, \quad (3.5.26a)$$

or equivalently

$$A - \alpha I + y(B - \beta I) \succeq 0. \quad (3.5.26b)$$

The latter condition is an LMI, thus it can be verified in essentially polynomial time.¹⁰ This result, however, cannot be extended to general non-homogeneous functions.

Polyak's local convexity result (Theorem 3.5.10) can be used to prove the following local duality result:

Theorem 3.5.19 (S-lemma, local version).

Let $x \in \mathbb{R}^n$ and $f(x) = (f_1(x), \dots, f_m(x))$, where

$$f_i(x) = \frac{x^T A_i x}{2} + a_i^T x \quad (3.5.27)$$

are quadratic functions. If the vectors a^i , $i = 1, \dots, m$ are linearly independent then there exists an $\bar{\varepsilon} > 0$ such that for all $\varepsilon < \bar{\varepsilon}$ the following two statements are equivalent:

(i) *The system*

$$f_i(x) \leq \alpha_i, \quad i = 1, \dots, m \quad (3.5.28a)$$

$$\|x\| \leq \varepsilon \quad (3.5.28b)$$

is not solvable.

(ii) *There exists a vector of nonnegative multipliers y_1, \dots, y_m (not all of them are zero) such that*

$$\sum_{i=1}^m y_i (f_i(x) - \alpha_i) \geq 0, \quad \forall x \in \mathbb{R}^n, \|x\| \leq \varepsilon. \quad (3.5.29)$$

¹⁰For a complete discussion on the complexity issues see [16, §6.6].

Proof. If the vectors a^i , $i = 1, \dots, m$ are linearly independent then the smallest singular value of $A = (a^1 | \dots | a^m)$ is positive, therefore from Theorem 3.5.10 the set

$$\{(f_1(x), \dots, f_m(x)) : x \in \mathbb{R}^n, \|x\| < \varepsilon\} \subset \mathbb{R}^m \quad (3.5.30)$$

is convex for any $\varepsilon < \bar{\varepsilon} = \nu/(2L)$, where ν and L are defined in Theorem 3.5.10. Knowing this we can apply the usual separation idea to finish the proof, see e.g., §3.2.2. \square

These kinds of results usually require the convexity of the functions (see, [115, 120]). Here, however, we do not need to impose convexity on the functions. It is important to note that unlike other results presented so far, Theorem 3.5.19 uses quantitative information about the problem data.

Now we can put together the pieces. If the image is convex then we can use the separation argument presented in §3.2.2 to get the dual statement. If the image is nonconvex then we can use Theorem 3.5.13 to characterize how much the image is not convex. This idea yields the following result.

Theorem 3.5.20. *If the system*

$$x^T A_i x \leq \alpha_i, \quad i = 1, \dots, m \quad (3.5.31a)$$

$$\|x\| = 1 \quad (3.5.31b)$$

is not solvable then one of the following two statements is true:

(i) *There are nonnegative multipliers y_1, \dots, y_m (not all of them zero) such that the matrix*

$$\sum_{i=1}^m y_i (A_i - \alpha_i I) \quad (3.5.32)$$

is positive semidefinite.

(ii) *Using the notations of Theorem 3.5.13, there are at most $k^{\mathbb{R}}(m, n)$ vectors $x_1, \dots, x_k \in \mathbb{R}^n$ such that*

$$\sum_{j=1}^{k^{\mathbb{R}}(m, n)} x_j^T A_i x_j \leq \alpha_i, \quad \forall i = 1, \dots, m. \quad (3.5.33)$$

Remark 3.5.21. System (3.5.33) will be discussed later in §3.6.4 in connection with rank-constrained optimization.

Finally, let us see what we can derive from Ramana's result about ICON maps (Theorem 3.5.3). His characterization of ICON maps is different in nature from the previous equivalent conditions. Ramana's conditions are more dependent on the actual data in the matrices. However, the result about LICON maps serves two goals. First, it gives a polynomial time verifiable sufficient condition for the ICON property. Second, if $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a constant-free LICON map then we can include a set of linear constraints in the system. The following theorem is obtained:

Theorem 3.5.22 (S-lemma with a LICON map). *Let $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a constant-free ($F(0) = 0$) quadratic map and $M \in \mathbb{R}^{n \times k}$ any matrix. If F is a LICON map (see Theorem 3.5.4 for an equivalent condition of this) then the following two statements are equivalent:*

(i) *The system*

$$F_i(x) \leq \alpha_i, \quad i = 1, \dots, m \quad (3.5.34a)$$

$$Mx = 0 \quad (3.5.34b)$$

is not solvable.

(ii) *There is a nonzero vector $y = (y_1, \dots, y_m) \in \mathbb{R}_+^m \setminus \{0\}$ of nonnegative multipliers such that*

$$y^T(F(x) - \alpha) \geq 0, \quad \forall x : Mx = 0. \quad (3.5.35)$$

Remark 3.5.23. A special case of this result was also presented in Proposition 3.3.9.

Results over the complex field

The theorems in §3.5.2 can be applied in two ways. First we can easily prove the complex counterparts of all the theorems in the previous subsection. This way we can characterize the solvability of the system

$$z^* A_i z \leq \alpha_i, \quad i = 1, \dots, m \quad (3.5.36a)$$

$$z^* z = 1, \quad (3.5.36b)$$

where $A_i, i = 1, \dots, m$ are $n \times n$ complex Hermitian matrices. However, as my main topic is the solvability of real quadratic systems, the enumeration of all these results is beyond the scope of this thesis. We just briefly mention the strongest duality result here because we will refer to it later. It is an easy exercise to derive all the other results.

Theorem 3.5.24 (S-lemma, complex case). *Let $n \geq 3$, $A_i \in \mathbb{C}^n$, $i = 1, 2, 3$. The following two statements are equivalent:*

(i) *The system*

$$z^* A_i z = 0, \quad i = 1, 2, 3 \quad (3.5.37a)$$

$$z \neq 0 \quad (3.5.37b)$$

is not solvable.

(ii) *There are multipliers y_1, y_2, y_3 such that*

$$\sum_{i=1}^3 y_i A_i \succ 0. \quad (3.5.38)$$

On the other hand the convex results have important consequences for real systems, too. Let A be a real symmetric matrix. If $z = x + \mathbf{i}y \in \mathbb{C}^n$ is a complex vector, then

$$z^* A z = (x - \mathbf{i}y)^T A (x + \mathbf{i}y) = x^T A x + y^T A y. \quad (3.5.39)$$

What we can get this way is a real quadratic system

$$x^T A_i x + y^T A_i y \leq \alpha_i, \quad i = 1, \dots, m \quad (3.5.40a)$$

$$\|x\|^2 + \|y\|^2 = 1. \quad (3.5.40b)$$

These equations are strongly related to rank-constrained optimization and will be discussed in more detail in §3.6.5.

3.6 Rank-constrained LMI

Looking at the second proof (see §3.2.3) of the S-lemma we can see that the crucial step is to show that the LMI relaxation of the system of quadratic inequalities is exact. This idea leads to the concept of rank-constrained optimization.

3.6.1 Motivation

Consider the homogeneous case

$$x^T A x < 0 \quad (3.6.1a)$$

$$x^T B_i x \leq 0, \quad i = 1, \dots, m \quad (3.6.1b)$$

and assume that the Slater condition is satisfied, i.e., there exists an $\bar{x} \in \mathbb{R}^n$ such that $\bar{x}^T B_i \bar{x} < 0$ for all $i = 1, \dots, m$. Using the standard notation introduced in §3.2.3, this system is equivalent to the following LMI:

$$A \bullet Z < 0 \tag{3.6.2a}$$

$$B_i \bullet Z \leq 0, \quad i = 1, \dots, m \tag{3.6.2b}$$

$$Z \succeq 0, \tag{3.6.2c}$$

$$\text{rank}(Z) = 1. \tag{3.6.2d}$$

After relaxing the condition on the rank

$$A \bullet Z < 0 \tag{3.6.3a}$$

$$B_i \bullet Z \leq 0, \quad i = 1, \dots, m \tag{3.6.3b}$$

$$Z \succeq 0 \tag{3.6.3c}$$

we have to establish the following:

1. Prove the equivalence of the solvability of (3.6.3) and (3.6.2).
2. Prove that the Slater condition holds for (3.6.3).

The latter is simple, let us take $Z = \bar{x}\bar{x}^T + \alpha I$ where \bar{x} is the Slater point of (3.6.1) and I is the identity matrix. If $\alpha > 0$ is small enough then all the linear constraints are satisfied while Z is positive definite. The former is a more difficult problem and there are only few general results in that area. This is the subject of the next section.

After proving these two statements, we can apply the Farkas Theorem (see Theorem 2.2.3) to derive the dual equivalent of system (3.6.3):

$$A + \sum_{i=1}^m y_i B_i \succeq 0 \tag{3.6.4a}$$

$$y \geq 0. \tag{3.6.4b}$$

This is exactly the second part of the S-lemma. Putting the parts together we get that if the solvability of (3.6.3) implies the existence of a rank-1 solution then the S-procedure is exact. Let us examine when this can happen.

3.6.2 Theoretical results

Finding low rank solutions of linear matrix inequalities is a relatively new research field. The earliest general result, due to Pataki [92, 93], was later discovered in other contexts, too.

Theorem 3.6.1 (Low rank semidefinite matrices).

If $\mathcal{A} \subseteq \mathbb{S}^n$ is an affine subspace such that the intersection $\mathbb{P}\mathbb{S}^n \cap \mathcal{A}$ is non-empty and $\dim(\mathcal{A}) \geq \binom{n+1}{2} - \binom{r+2}{2} + 1$ then there is a matrix $X \in \mathbb{P}\mathbb{S}^n \cap \mathcal{A}$ such that $\text{rank}(X) \leq r$.

Proof. There are many possible ways to prove the theorem but the key observation is that any extreme point of the intersection will have a sufficiently low rank. This also helps one to find such a matrix. The theorem is intuitively plausible: in order to have low rank matrices we have to intersect $\mathbb{P}\mathbb{S}^n$ with a high dimensional subspace \mathcal{A} .

Let $X \in \mathbb{P}\mathbb{S}^n \cap \mathcal{A}$ be an extreme point of the intersection, then we can assume without loss of generality that

$$X = \begin{pmatrix} X_{11} & 0 \\ 0 & 0 \end{pmatrix}, \quad (3.6.5)$$

where X_{11} is positive definite. If $\text{rank}(X) \leq r$ then we have the requested low rank matrix, so let us assume that $\text{rank}(X) \geq r + 1$. As the dimension of $(r + 1) \times (r + 1)$ symmetric matrices is $\binom{r+2}{2}$ and X_{11} is constrained by at most $\binom{r+2}{2} - 1$ linear equalities we have a nonzero matrix Y such that

$$Y \bullet A = 0 \quad \forall A \in \mathcal{A} \quad (3.6.6a)$$

$$Y = \begin{pmatrix} Y_{11} & 0 \\ 0 & 0 \end{pmatrix} \quad (3.6.6b)$$

$$Y_{11} \neq 0. \quad (3.6.6c)$$

Now $X \pm \varepsilon Y \in \mathbb{P}\mathbb{S}^n \cap \mathcal{A}$ for small values of ε , thus X is not an extreme point of the intersection, contradicting our assumption. This proves that X is of sufficiently low rank. \square

Remark 3.6.2. The bound is sharp in the sense that if $r < n$ then one can find a subspace $\mathcal{A} \subseteq \mathbb{S}^n$ such that $\mathbb{P}\mathbb{S}^n \cap \mathcal{A} \neq \emptyset$, $\dim(\mathcal{A}) = \binom{n+1}{2} - \binom{r+2}{2}$ and for every matrix $X \in \mathbb{P}\mathbb{S}^n \cap \mathcal{A}$ we have $\text{rank}(X) > r$.

If \mathcal{A} is nontrivial then the conditions of the theorem are obviously satisfied for $r = n - 1$, implying the following simple corollary:

Corollary 3.6.3. *For any nontrivial (i.e., at least one dimensional) affine subspace \mathcal{A} , the intersection $\mathbb{P}\mathbb{S}^n \cap \mathcal{A}$ (provided that it is not empty) contains a matrix X such that $\text{rank}(X) \leq n - 1$.*

The result in Theorem 3.6.1 was generalized by Barvinok [12]:

Theorem 3.6.4 (Lower rank semidefinite matrices).

Let $r > 0$, $n \geq r+2$ and $\mathcal{A} \subseteq \mathbb{S}^n$ be an affine subspace such that the intersection $\mathbb{P}\mathbb{S}^n \cap \mathcal{A}$ is non-empty, bounded and $\dim(\mathcal{A}) = \binom{n+1}{2} - (r+2)$. Then there is a matrix $X \in \mathbb{P}\mathbb{S}^n \cap \mathcal{A}$ such that $\text{rank}(X) \leq r$.

Barvinok’s proof uses a differential geometric argument based on the structure of the cone of positive semidefinite matrices. There is a crucial difference between the two theorems. In the Pataki Theorem any extremal point of the intersection always has a sufficiently low rank, while in the Barvinok Theorem we can guarantee only that there is a low rank extremal point. This has important algorithmic consequences: while it is easy to find a suitable low rank matrix for the Pataki Theorem, there is no known algorithm to find such a matrix for the Barvinok Theorem.

Barvinok was aware that this result was not new, but he was the first to state it and to provide a direct proof. He related his result to a theorem of Au-Yeung and Poon [7] on the image of the sphere under a quadratic map. This result is discussed in §3.5.

Finally, both of the theorems in this section are extended for general symmetric matrices in [45] using Jordan algebraic techniques.

3.6.3 Implications

Now let us see what we can obtain using the above theorems. In view of the relaxation argument in §3.6.1, we are interested in rank-1 solutions, i.e., we use the theorems with $r = 1$.

Let us assume that system (3.6.3) is solvable, then we have a matrix Z such that

$$A \bullet Z = w_0 < 0 \tag{3.6.7a}$$

$$B_i \bullet Z = w_i \leq 0, \quad i = 1, \dots, m \tag{3.6.7b}$$

$$Z \succeq 0. \tag{3.6.7c}$$

Here each equality corresponds to a hyperplane. Let \mathcal{A} be the intersection of these hyperplanes, then

$$\dim(\mathcal{A}) \geq \binom{n+1}{2} - m - 1. \tag{3.6.8}$$

In order to apply Theorem 3.6.1 with $r = 1$ we need to have

$$\dim(\mathcal{A}) \geq \binom{n+1}{2} - 2, \tag{3.6.9}$$

so we must have $m \leq 1$, i.e., we can have at most one non-strict inequality. Then Theorem 3.6.1 guarantees the existence of a rank-1 solution to (3.6.3), thus the Pataki Theorem implies the basic S-lemma.

Notice that this way we proved slightly more than the existence of a rank-1 solution. Namely, we proved that the rank-1 matrix xx^T can be chosen such that

$$A \bullet xx^T = w_0 < 0 \quad (3.6.10a)$$

$$B_i \bullet xx^T = w_i \leq 0, \quad i = 1, \dots, m \quad (3.6.10b)$$

$$xx^T \succeq 0, \quad (3.6.10c)$$

where $w_i, i = 0, \dots, m$ are the same numbers as in system (3.6.7).

Now let us try to apply the Barvinok theorem. Using the same setup and an argument similar to that used before, and setting $r = 1$, we have $m = 2$, so we can have two non-strict inequalities. However, there is an extra condition to satisfy, namely the solution set of

$$A \bullet Z = w_0 \quad (3.6.11a)$$

$$B_i \bullet Z = w_i, \quad i = 1, 2 \quad (3.6.11b)$$

$$Z \succeq 0 \quad (3.6.11c)$$

must be bounded and since $n \geq r + 2$ we must have $n \geq 3$.

Unfortunately the boundedness does not always hold, we need some extra condition to force it. It is well known from convex analysis (see [115, 120]) that a convex set is unbounded if and only if it has a nontrivial recession direction, or, in other words, if it contains a half-line. Applying this to our case we get that the solution set of system (3.6.11) is bounded if and only if the following system is not solvable:

$$A \bullet Z = 0 \quad (3.6.12a)$$

$$B_i \bullet Z = 0, \quad i = 1, 2 \quad (3.6.12b)$$

$$Z \succeq 0 \quad (3.6.12c)$$

$$Z \neq 0. \quad (3.6.12d)$$

Let us note that this system is independent of w , therefore the boundedness of the solution set of (3.6.11) does not depend on the actual choice of w_0 and w . Either all of them are bounded or all are unbounded (provided they are not empty).

Since this is an LMI it is easy to characterize its solvability. Namely, by the duality theory of LMIs (see [16, Proposition 2.4.2.]), system (3.6.12)

is not solvable if and only if there are real numbers λ_0, λ_1 and λ_2 such that $\lambda_0 A + \lambda_1 B_1 + \lambda_2 B_2$ is positive definite. This way we proved Proposition 3.3.6, i.e., the S-lemma with three inequalities.

3.6.4 Higher rank solutions

So far we have applied the theorems to the $r = 1$ case. However, the higher rank results also have some consequences for quadratic systems. Consider the following system:

$$\sum_{j=1}^r x^{jT} A x^j < 0 \quad (3.6.13a)$$

$$\sum_{j=1}^r x^{jT} B_i x^j \leq 0, \quad i = 1, \dots, m, \quad (3.6.13b)$$

where $x^j \in \mathbb{R}^n$, $j = 1, \dots, r$. Introducing

$$X = \sum_{j=1}^r x^{jT} x^j \quad (3.6.14)$$

and using the \bullet notation this system can be written as:

$$A \bullet X < 0 \quad (3.6.15a)$$

$$B_i \bullet X \leq 0, \quad i = 1, \dots, m \quad (3.6.15b)$$

$$X \succeq 0 \quad (3.6.15c)$$

$$\text{rank}(X) \leq r, \quad (3.6.15d)$$

since a positive semidefinite matrix X can always be decomposed as the sum of $\text{rank}(X)$ positive semidefinite rank-1 matrices. Now relaxing the rank constraint we get an LMI identical to (3.6.3):

$$A \bullet X < 0 \quad (3.6.16a)$$

$$B_i \bullet X \leq 0, \quad i = 1, \dots, m \quad (3.6.16b)$$

$$X \succeq 0. \quad (3.6.16c)$$

Applying Theorems 3.6.1 and 3.6.4, and using a similar argument we obtain the following theorem:

Theorem 3.6.5 (Quadratic systems with multiple terms).

Let $A, B_1, \dots, B_m \in \mathbb{R}^{n \times n}$ be symmetric matrices such that there are vectors $\bar{x}^1, \dots, \bar{x}^r \in \mathbb{R}^n$ for which the Slater condition is satisfied, i.e.,

$$\sum_{j=1}^r \bar{x}^{jT} B_i \bar{x}^j < 0, \quad i = 1, \dots, m. \quad (3.6.17)$$

If

1. $m \leq \binom{r+2}{2} - 2$, or
2. $m = \binom{r+2}{2} - 1$, $n \geq r+2$ and there is a positive definite linear combination of A, B_1, \dots, B_m ,

then the following two statements are equivalent:

(i) The quadratic system

$$\sum_{j=1}^r x^{jT} A x^j < 0 \quad (3.6.18a)$$

$$\sum_{j=1}^r x^{jT} B_i x^j \leq 0, \quad i = 1, \dots, m \quad (3.6.18b)$$

is not solvable.

(ii) The LMI

$$A + \sum_{i=1}^m y_i B_i \succeq 0 \quad (3.6.19a)$$

$$y_1, \dots, y_m \geq 0 \quad (3.6.19b)$$

is solvable.

Now let $z = x^1 + \mathbf{i}x^2 \in \mathbb{C}^n$ be a complex vector, then for any real symmetric matrix A we have $z^* A z = x^{1T} A x^1 + x^{2T} A x^2$, therefore the above theorems can be used to decide the solvability of the following complex quadratic system, where A and B_i , ($i = 1, \dots, m$) are real symmetric matrices:

$$z^* A z < 0 \quad (3.6.20a)$$

$$z^* B_i z \leq 0, \quad i = 1, \dots, m. \quad (3.6.20b)$$

The value of m can be at most 3 without any further assumptions, or 4, if there is a positive definite linear combination of the matrices.

3.6.5 Rank constraints and convexity

We have seen that one can have more inequalities in the S-lemma in the complex case than in the real case. The reason for this is obvious from the previous section. Real solutions come from rank-1 solutions of system (3.6.3), while for complex solutions it is enough to have a rank-2 solution. This observation sheds some light on the convexity issues discussed in the previous section, particularly, it answers the question of why the joint numerical range over the complex field possesses a much nicer structure and has a more straightforward characterization.

There is an interesting connection between the rank constraints and the convexity of the joint numerical range.

Recall Dines's result (Proposition 3.2.2): $\{(x^T Ax, x^T Bx) : x \in \mathbb{R}^n\}$ is convex. Let us take two points in the image, (u_1, u_2) and (v_1, v_2) , and let $0 < \lambda < 1$. Now we have to find a point $x \in \mathbb{R}^n$ such that $x^T Ax = \lambda u_1 + (1 - \lambda)v_1$ and $x^T Bx = \lambda u_2 + (1 - \lambda)v_2$. Using the notation of this section we are looking for a rank-1 solution of the following system:

$$A \bullet X = \lambda u_1 + (1 - \lambda)v_1 \quad (3.6.21a)$$

$$B \bullet X = \lambda u_2 + (1 - \lambda)v_2. \quad (3.6.21b)$$

Using Theorem 3.6.1 we can see that this system always has a rank-1 solution, $X = xx^T$, and thus we have proved the convexity of the set $\{(x^T Ax, x^T Bx) : x \in \mathbb{R}^n\}$. The same argument can be applied to the $m = 2, n \geq 3$ case, then the Barvinok theorem (Theorem 3.6.4) will yield Polyak's convexity result (Theorem 3.5.2). The connection works both ways, from the convexity of the joint numerical range we can deduce rank constraints for real or complex LMIs.¹¹ This is particularly useful in the complex case, since the theory of the complex numerical ranges has been investigated thoroughly, while there are no results for low rank solutions of complex LMIs. On the other hand, there are very few results about the convexity of the image of the real space under higher rank quadratic maps.

Finally, we can contrast the Pataki Theorem for low rank matrices (Theorem 3.6.1) and Poon's Theorem on the number of terms in the convex combinations (Theorem 3.5.13). Although they are stated in different contexts they have some applications in common. Consider the system

$$A_i \bullet X = b_i, \quad i = 1, \dots, m \quad (3.6.22a)$$

$$X \succeq 0, \quad (3.6.22b)$$

¹¹The idea of this argument was sparked during a discussion with Gábor Pataki.

where $A_i, X \in \mathbb{R}^{n \times n}$ and assume that the system is solvable. Using Theorem 3.6.1 we find that there is a solution X such that

$$\text{rank}(X) \leq \left\lfloor \frac{\sqrt{8m+1}-1}{2} \right\rfloor = R_1(m, n), \quad (3.6.23)$$

and of course $\text{rank}(X) \leq n$ also holds.

Now we try to get a similar bound from Theorem 3.5.13. Let X be a solution for system (3.6.22) and let us consider its rank-1 decomposition:

$$X = \sum_{j=1}^r \lambda^j x^j x^{jT}, \quad (3.6.24)$$

where $\|x^j\| = 1$, and since X is positive semidefinite, we have $\lambda^j \geq 0$ for $j = 1, \dots, m$. Let us define the following scaled quantities:

$$\lambda = \sum_{j=1}^r \lambda^j \quad (3.6.25a)$$

$$\bar{x}^j = \frac{x^j}{\lambda} \quad (3.6.25b)$$

$$\bar{X} = \frac{X}{\lambda} = \sum_{j=1}^r \frac{\lambda^j}{\lambda} x^j x^{jT}, \quad (3.6.25c)$$

where the last line shows that \bar{X} is a convex combination of the rank-1 matrices $x^j x^{jT}$. Now let \mathcal{A} denote the m -tuple (A_1, \dots, A_m) and let us use the notations of Theorem 3.5.13. By the definition of the image set we have

$$(A_1 \bullet \bar{x}^j \bar{x}^{jT}, \dots, A_m \bullet \bar{x}^j \bar{x}^{jT}) \in W_1^{\mathbb{R}}(\mathcal{A}), \quad \forall j = 1, \dots, r, \quad (3.6.26)$$

and using the decomposition result on \bar{X} we find that

$$(A_1 \bullet \bar{X}, \dots, A_m \bullet \bar{X}) \in \text{conv}(W_1^{\mathbb{R}}(\mathcal{A})), \quad \forall j = 1, \dots, r. \quad (3.6.27)$$

We can use Theorem 3.5.13 to deduce that there is a matrix \tilde{X} such that

$$A_i \bullet \tilde{X} = A_i \bullet \bar{X}, \quad \forall j = 1, \dots, m \quad (3.6.28)$$

$$\tilde{X} \succeq 0$$

$$\text{rank}(\tilde{X}) \leq \min \left\{ n, \left\lfloor \frac{\sqrt{8m+1}-1}{2} \right\rfloor + \delta_{\frac{n(n+1)}{2}, m} \right\} = R_2(m, n),$$

and the matrix $\lambda \tilde{X}$ solves system (3.6.22). Finally, if $m = \frac{n(n+1)}{2}$ then $R_2(m, n) = n$. This shows that the two bounds are identical.

3.7 Generalized convexities

“Hidden convexity” (see [17]) seems to play an important role in the S-lemma. Although we made no convexity assumptions, the image of the quadratic map in §3.2.2 turned out to be convex. In this section we shed some more light on this interesting phenomenon.

3.7.1 Motivation

The fact that the S-lemma can be viewed as a non-convex generalization of the Farkas Theorem inspires us to look for a more general sense of convexity, which includes both the classical convex and the quadratic case. The convexity results in §3.5 show that even though the functions are not convex they describe a convex object, thus the problems admit some hidden convexity.

3.7.2 Theoretical results

There are many different convexity notions in the literature. Let us assume that X is a nonempty set, Y is a topological vector space over the real numbers with dual space Y^* , i.e., Y^* contains all the linear functions that map from Y to \mathbb{R} .

Let us assume that K is a convex, closed, pointed, solid cone, and consider the orderings \succeq_K and \succ_K defined by K . Let us note that the functions in this section are not necessarily quadratic. Using these notations we can define the following convexity notions:

Definition 3.7.1 (General convex functions). *A function $f : X \rightarrow Y$ is called*

Ky Fan convex, *if for every $x_1, x_2 \in X$ and $\lambda \in [0, 1]$ there exists an $x_3 \in X$ such that*

$$f(x_3) \preceq_K (1 - \lambda)f(x_1) + \lambda f(x_2). \quad (3.7.1)$$

König convex, *if for every $x_1, x_2 \in X$ there exists an $x_3 \in X$ such that*

$$f(x_3) \preceq_K \frac{1}{2}f(x_1) + \frac{1}{2}f(x_2). \quad (3.7.2)$$

K -convexlike, *if there exists a $\lambda \in (0, 1)$ such that for every $x_1, x_2 \in X$ there exists an $x_3 \in X$ such that*

$$f(x_3) \preceq_K (1 - \lambda)f(x_1) + \lambda f(x_2). \quad (3.7.3)$$

Further, in order to deal with both equalities and inequalities we can introduce the mixed versions of the above definitions.

Definition 3.7.2 (General convex-linear functions). *Let Y and Z be locally convex¹² topological vector spaces and let X be any set. The function pair $(f, g) : X \rightarrow Y \times Z$ is called*

Ky Fan convex-linear, *if for each $x_1, x_2 \in X$ and $\lambda \in [0, 1]$ there exists an $x_3 \in X$ such that*

$$f(x_3) \preceq_K (1 - \lambda)f(x_1) + \lambda f(x_2) \quad (3.7.4a)$$

$$g(x_3) = (1 - \lambda)g(x_1) + \lambda g(x_2), \quad (3.7.4b)$$

König convex-linear, *if for each $x_1, x_2 \in X$ there exists $x_3 \in X$ such that*

$$2f(x_3) \preceq_K f(x_1) + f(x_2) \quad (3.7.5a)$$

$$2g(x_3) = g(x_1) + g(x_2). \quad (3.7.5b)$$

The generality of these definitions is obvious if we notice that X can be any set, without any topology. Notice, e.g., that any continuous function mapping a compact set into \mathbb{R} is König convex, since it has a minimum.

Special cases of these definitions were first introduced by Ky Fan [41] and König [75]. Obviously, all Ky Fan convex functions are König convex, and all König convex functions are K -convexlike. However, there seems to be a large gap between the two extremes in these definitions. This gap is not that large, as it is shown in the next proposition (see [28]).

Proposition 3.7.3. *If f is K -convexlike then the set of λ 's satisfying the definition of the Ky Fan convexity is dense in $[0, 1]$.*

Corollary 3.7.4. *If f is continuous (or at least lower semicontinuous) then the convexity notions in Definition 3.7.1 coincide.*

Illés, Joó and Kassay proved several versions of the Farkas Theorem for these generalized convexities [67, 68, 69].

Theorem 3.7.5 (Duality for König convex functions). *Let $f : X \rightarrow Y$ be König convex, where Y is a locally convex space. If there is no $x \in X$ such that $f(x) \prec_K 0$ then there exists a $y^* \in Y^* \setminus \{0\}$ such that*

$$y^*(f(x)) \geq 0, \forall x \in X. \quad (3.7.6)$$

¹²A topological vector space is locally convex if every point has a neighbourhood basis consisting of open convex sets. Every normed space is locally convex.

For the following theorem let Y_1 and Y_2 be two topological vector spaces over the reals, and let $K_1 \subseteq Y_1$ and $K_2 \subseteq Y_2$ be convex cones with vertices O_{Y_1} and O_{Y_2} such that $\text{int}(K)_1 \neq \emptyset$. Let $f : X \rightarrow Y_1$ and $g : X \rightarrow Y_2$ be two functions and define $Y = Y_1 \times Y_2$, $K = K_1 \times K_2$ and $F = (f, g) : X \rightarrow Y$.

Theorem 3.7.6 (Duality for K -convexlike functions).

Suppose $F = (f, g) : X \rightarrow Y$ is K -convexlike and the set $F(X) + K$ has nonempty interior. The following assertions hold:

(i) If there is no $x \in X$ such that

$$f(x) \prec_K 0 \quad (3.7.7a)$$

$$g(x) \preceq_K 0 \quad (3.7.7b)$$

then there exist $y_1^* \in K_1^*$ and $y_2^* \in K_2^*$ (not both being the origin) such that

$$y_1^*(f(x)) + y_2^*(g(x)) \geq 0, \forall x \in X. \quad (3.7.8)$$

(ii) If there exist $y_1^* \in K_1^* \setminus \{O_{Y_1^*}\}$ and $y_2^* \in K_2^*$ such that (3.7.8) holds then system (3.7.7) is not solvable.

The following theorem deals with systems containing equality constraints.

Theorem 3.7.7 (Duality for Ky Fan convex-linear functions).

Let Y and Z be locally convex topological vector spaces and let X be any set. Let $(f, g) : X \rightarrow Y \times Z$ be Ky Fan convex-linear with K and define

$$M = \{(f(x) + v, g(x)) : x \in X, v \in K\} \subset Y \times Z. \quad (3.7.9)$$

If $\text{int}(M) \neq \emptyset$ and there is no $x \in X$ such that

$$f(x) \prec_K 0$$

$$g(x) = 0$$

then there exist $y^* \in K^*$ and $z^* \in Z^*$ (not both being the origin) such that

$$y^*(f(x)) + z^*(g(x)) \geq 0, \forall x \in X. \quad (3.7.10)$$

Remark 3.7.8. The authors in [68] conjecture that this result generalizes to König convex functions.

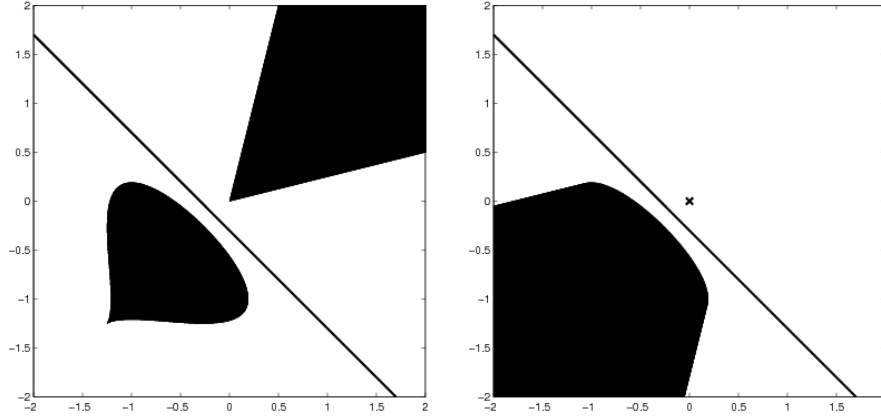


Figure 3.2: **Left:** Separating a convex cone from a nonconvex set. **Right:** The Minkowski sum of the set and the negative cone is separable from the origin.

The proof of the theorems above relies on the following simple observation (see Fig. 3.2). In order to separate a convex cone from a set, the set does not have to be convex. It is enough to be convex on the side “facing the cone.” More precisely, if K is a convex cone disjoint from set \mathcal{C} , and $\mathcal{C} + (-K)$ is convex then K and \mathcal{C} can be separated by a hyperplane. Another view of this idea is that $K \cap \mathcal{C} = \emptyset$ if and only if $0 \notin \mathcal{C} + (-K)$. Separating this latter set from the origin is equivalent to separating K and \mathcal{C} .

For a summary of different convexity concepts see [2, 28] and the references therein.

3.7.3 Implications

It is straightforward to apply these theorems to our problems, we only need to verify the convexity assumptions. Notice however, that if we restrict ourselves to homogeneous quadratic functions then the above three notions coincide. Thus, e.g., in order to apply the results we need to prove that a pair of quadratic functions is König convex, i.e., for any $x_1, x_2 \in \mathbb{R}^n$ there exists an $x_3 \in \mathbb{R}^n$ such that

$$x_3^T A x_3 \leq \frac{1}{2} x_1^T A x_1 + \frac{1}{2} x_2^T A x_2 \quad (3.7.11a)$$

$$x_3^T B x_3 \leq \frac{1}{2} x_1^T B x_1 + \frac{1}{2} x_2^T B x_2. \quad (3.7.11b)$$

In fact, in Proposition 3.2.2 we proved that these two equations can be satisfied with equality, therefore a pair of quadratic functions is both König convex and

König concave, also called König linear. This result seems to be new in the context of generalized convexities.

The three theorems presented in the previous section yield the following results:

Theorem 3.7.9. *Let $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ be homogeneous quadratic functions. The following two statements are equivalent:*

(i) *The system*

$$f(x) < 0 \tag{3.7.12a}$$

$$g(x) < 0 \tag{3.7.12b}$$

is not solvable.

(ii) *There exist nonnegative multipliers y_1 and y_2 (not both of them being zero) such that*

$$y_1 f(x) + y_2 g(x) \geq 0, \forall x \in \mathbb{R}^n. \tag{3.7.13}$$

This result is a variant of the S-lemma with two strict inequalities. Notice that the second statement immediately implies the first one, without any assumption on the functions.

Theorem 3.7.10. *Suppose $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ are homogeneous quadratic functions. The following assertions hold:*

(i) *If there is no $x \in \mathbb{R}^n$ such that*

$$f(x) < 0 \tag{3.7.14a}$$

$$g(x) \leq 0, \tag{3.7.14b}$$

then there exist nonnegative multipliers y_1 and y_2 (not both being the origin) such that

$$y_1 f(x) + y_2 g(x) \geq 0, \forall x \in \mathbb{R}^n. \tag{3.7.15}$$

(ii) *If there exists $y_1 > 0$ and $y_2 \geq 0$ such that (3.7.15) holds then there is no solution for (3.7.14).*

The gap in this theorem comes from that fact that we did not assume the Slater condition. If we further assume that there exists an $\bar{x} \in \mathbb{R}^n$ with $g(\bar{x}) < 0$ then the nonsolvability of (3.7.14) automatically implies the existence of multipliers in (ii).

Finally, let us see what we can get if we allow equality constraints.

Theorem 3.7.11. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \mathbb{R}^n \rightarrow Y$ be homogeneous quadratic functions, where Y is one of $\{0\}$, \mathbb{R} , \mathbb{R}_+ or \mathbb{R}_- . If there is no $x \in \mathbb{R}^n$ such that*

$$f(x) < 0 \tag{3.7.16a}$$

$$g(x) = 0 \tag{3.7.16b}$$

then there exist $y_1 \geq 0$ and $y_2 \in Y^$ (not both being the origin) such that*

$$y_1 f(x) + y_2 g(x) \geq 0, \forall x \in \mathbb{R}^n. \tag{3.7.17}$$

In fact, depending on the set Y , we have obtained four theorems.

If $Y = \{0\}$ (i.e., $g(x) \equiv 0$) then the solvability of (3.7.16) is equivalent to the solvability of $f(x) < 0$.

If $Y = \mathbb{R}$, i.e., $g(x)$ takes both positive and negative values, then $Y^* = \{0\}$, thus $y_2 = 0$ and consequently $y_1 > 0$. This means that $f(x) \geq 0$ for all $x \in \mathbb{R}^n$. This result (at least for the special homogeneous case) is stronger than the one found in [122], because we showed that the multiplier of $g(x)$ is actually 0. In other words, if $g(x)$ takes both positive and negative values then system (3.7.16) is not solvable if and only if $f(x) < 0$ is not solvable, i.e., $f(x)$ is positive semidefinite.

If $Y = \mathbb{R}_+$ (or $Y = \mathbb{R}_-$) then $Y^* = \mathbb{R}_+$ ($Y^* = \mathbb{R}_-$) and $g(x) = 0$ is equivalent to $g(x) \leq 0$ ($g(x) \geq 0$), therefore the problem is reduced to Theorem 3.7.10.

3.8 Miscellaneous topics

In the last section before the summary we discuss some miscellaneous topics. These are all related to the S-lemma or, more generally, to systems of quadratic equations. The goal of this section is not to give a full-detailed review of all these topics, but rather to show the connections and raise some further questions.

3.8.1 Trust region problems

Trust region algorithms [33] are among the most successful methods of solving unconstrained nonlinear optimization problems. At each iteration of the algorithm we minimize a quadratic approximation of the objective function over a ball, called the trust region. Thus, given the current iterate \hat{x} , the trust region

subproblem is

$$\min x^T Hx + b^T x \quad (3.8.1a)$$

$$\|x - \hat{x}\|_2 \leq \alpha. \quad (3.8.1b)$$

In a slightly more general setting we can replace the ball constraint (3.8.1b) with an ellipsoidal constraint:

$$\min x^T Hx + b^T x \quad (3.8.2a)$$

$$(x - \hat{x})^T A(x - \hat{x}) \leq \alpha, \quad (3.8.2b)$$

where A is a symmetric positive semidefinite matrix, or, more generally, we can allow any matrix A . This way we get indefinite trust region subproblems [62, 63, 98, 119, 137]. These problems can still be solved in polynomial time, and that is what makes them suitable for building an algorithm. The reason behind the polynomial solvability is basically the S-lemma: a solution \tilde{x} is optimal, if and only if the following system is not solvable:

$$x^T Hx + b^T x < \tilde{x}^T H\tilde{x} + b^T \tilde{x} \quad (3.8.3a)$$

$$(x - \hat{x})^T A(x - \hat{x}) \leq \alpha. \quad (3.8.3b)$$

Using the S-lemma (Theorem 3.2.1) we can conclude that the optimality of \tilde{x} is further equivalent to the solvability of the following system:

$$\begin{aligned} x^T Hx + b^T x - \tilde{x}^T H\tilde{x} - b^T \tilde{x} + y((x - \hat{x})^T A(x - \hat{x}) - \alpha) &\geq 0, \quad \forall x \in \mathbb{R}^n, \\ y &\geq 0, \end{aligned} \quad (3.8.4)$$

which can be homogenized and written as an LMI. Thus, using a simple bisection scheme we can solve the indefinite trust region subproblem to any given precision in polynomial time.

Moreover, we can build more complex trust regions, i.e., minimize a quadratic function over the intersection of two ellipsoids, or use one ellipsoidal and one general (possibly indefinite) constraint. Exact details and properties of these algorithms are to be developed.

3.8.2 SDP relaxation of quadratically constrained quadratic problems (QCQP)

A broad range of optimization problems can be written as QCQPs. Recently, quite a number of articles considered the solvability of these problems. The research has two main branches. On one hand, one might consider the cases

when the SDP relaxation of general quadratic problems is exact [50, 72, 136], or, if the relaxation is not exact, then we can ask for bounds on the quality of the approximation [36, 88, 101, 125, 132]. Further, Kojima and Tunçel [73] proposed a successive SDP relaxation scheme to solve these problems.

The exact relaxation results of this area are quite different in nature from the results discussed so far. The sufficient conditions for the validity of the S-lemma are usually structural conditions, i.e., they do not involve explicit data in the matrices. On the other hand, the sufficient conditions in the SDP relaxation theory are more dependent on the problem data. The reason for this might be purely practical: as we have shown, structural conditions allow for only a relatively small number of equations and inequalities.

The quantity of results makes it impossible to include any of them here, so the reader is thus referred to the references mentioned in the first paragraph.

3.8.3 Algebraic geometry

So far little has been said about the algebraic nature of the problem, since the discussion has been more geometric and analytic. In this section we briefly review the algebraic connections.

The S-lemma is a surprising result from the point of view of algebraic geometry. Recall Hilbert's classical theorem, which can be found in any advanced algebra textbook, see, e.g., [13, 22, 55, 85]:

Theorem 3.8.1 (Nullstellensatz, complex case). *Let $p_1, \dots, p_m : \mathbb{C}^n \rightarrow \mathbb{C}$ be complex polynomials. The following two statements are equivalent:*

(i) *The system*

$$p_i(z) = 0, \quad i = 1, \dots, m \quad (3.8.5a)$$

$$z \in \mathbb{C}^n \quad (3.8.5b)$$

is not solvable.

(ii) *There exist complex polynomials $y_1(z), \dots, y_m(z)$ such that*

$$\sum_{i=1}^m y_i(z)p_i(z) \equiv -1. \quad (3.8.6)$$

This theorem holds only for the complex case. For real polynomials we have the following theorem:

Theorem 3.8.2 (Nullstellensatz, real case). *Let $p_1, \dots, p_m : \mathbb{R}^n \rightarrow \mathbb{R}$ be real polynomials. The following two statements are equivalent:*

(i) *The system*

$$p_i(x) = 0, \quad i = 1, \dots, m \quad (3.8.7a)$$

$$x \in \mathbb{R}^n \quad (3.8.7b)$$

is not solvable.

(ii) *There exist real polynomials $y_1(x), \dots, y_m(x)$ and $s_1(x), \dots, s_k(x)$ such that*

$$\sum_{j=1}^k s_j^2(x) + \sum_{i=1}^m y_i(x)p_i(x) \equiv -1. \quad (3.8.8)$$

If we know an *a priori* bound on the degree of the multipliers $y_i(x)$ (and the degree and number of $s_j(x)$ in the real case) then we can find them easily by equating the coefficients and solving a large linear system. Since any combinatorial optimization problem can be written as a system of nonlinear equations, such bounds must be exponential in the number of variables. The best available degree bounds are summarized in the following theorem (for simplicity we are dealing with the complex case only).

Theorem 3.8.3 (Effective Nullstellensatz). *Let $p_1, \dots, p_m : \mathbb{C}^n \rightarrow \mathbb{C}$ be complex polynomials with $\max_i \deg p_i = d$. If there exist complex polynomials $y_1(z), \dots, y_m(z)$ such that*

$$\sum_{i=1}^m y_i(z)p_i(z) \equiv -1, \quad (3.8.9)$$

then these polynomials can be chosen to satisfy

$$\max_i \deg y_i \leq \begin{cases} d^n & \text{if } d \geq 3 \text{ (see [74])} \\ 2^{\min\{m,n\}} & \text{if } d = 2 \text{ (see [118]).} \end{cases} \quad (3.8.10)$$

These bounds are essentially sharp, better estimates use some additional information about the polynomials, such as the geometric degree [117], the sparsity [118] or the height [59], to mention just a few. It would be interesting to specialize those bounds for quadratic systems and also to derive similar results for the real case. More details can also be found in [18].

In the simple case of two complex quadratic polynomials Theorem 3.8.3 gives a rather weak corollary (compare with Theorem 3.5.24). In this regard it is surprising that under the conditions of the S-lemma, the multiplier polynomials can be chosen to be constants. This also suggests further directions for the research on degree bounds in the Nullstellensatz.

3.8.4 Computational complexity

Let $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, \dots, m$ be quadratic functions. Using more advanced techniques (see [11, 56]) one can determine the solvability of a system

$$f_i(x) = 0, \quad i = 1, \dots, m \quad (3.8.11a)$$

$$\|x\|_2 = 1 \quad (3.8.11b)$$

$$x \in \mathbb{R}^n, \quad (3.8.11c)$$

and an exact solution can also be obtained. These problems include most combinatorial optimization problems thus in general we cannot hope for a polynomial time algorithm. However, there are some special cases. If $m = 1$ then solving system (3.8.11) is equivalent to determining the definiteness of $f_1(x)$, which can be done in polynomial time. Further, the S-lemma gives rise to a polynomial time algorithm if $m = 2$.

In [56] Grigoriev and Pasechnik proposed an algorithm to solve system (3.8.11). The algorithm is polynomial in n and exponential in m . Whence, if m is fixed, or at least bounded, then (3.8.11) can be solved in polynomial time.

3.9 New duality theorems

The particular strength of the techniques described in §3.5 is that they can be applied to more general constraints. We can characterize the solvability of the system

$$(x^T A_1 x, \dots, x^T A_m x) \in \mathcal{C} \quad (3.9.1a)$$

$$x \in \mathbb{R}^n \quad (3.9.1b)$$

$$(\|x\| = 1), \quad (3.9.1c)$$

where $\mathcal{C} \subset \mathbb{R}^m$ is a convex and possibly closed set. We can include the norm constraint, if we want to use the results about the joint numerical range. If this system is not solvable and the set of possible left-hand side vectors is convex then that set can be separated from \mathcal{C} by a hyperplane. The actual form of these duality results depends on the form of \mathcal{C} . In the results derived so far \mathcal{C} was the cone of nonnegative vectors. In what follows we present some examples when \mathcal{C} is a polyhedron or a Lorentz cone. Similar theorems can be proved using other special forms of set \mathcal{C} . To the best of my knowledge these results have not yet been stated explicitly.

The strength of these theorems is that the primal problems are non-convex while the dual problems are convex. In other words, we can decide the solvability of a nonconvex system by solving a convex one.

3.9.1 Polyhedral case

Let $\mathcal{C} \subset \mathbb{R}^m$ be a nonempty polyhedron defined by the inequalities $Mu \leq h$, where $M \in \mathbb{R}^{l \times m}$, $u \in \mathbb{R}^m$ and $h \in \mathbb{R}^l$. Consider the following system:

$$(x^T A_1 x, \dots, x^T A_m x) = u \quad (3.9.2a)$$

$$Mu \leq h. \quad (3.9.2b)$$

We can prove the following theorem:

Theorem 3.9.1 (New duality theorem (polyhedral case)).

Let $A_1, A_2 \in \mathbb{R}^{n \times n}$ be real symmetric matrices, $x \in \mathbb{R}^n$, $M \in \mathbb{R}^{l \times 2}$, $h \in \mathbb{R}^l$. Let us assume that the polyhedron $\{u \in \mathbb{R}^2 : Mu \leq h\}$ is not empty. The following two statements are equivalent:

(i) The quadratic system

$$M_{i1}x^T A_1 x + M_{i2}x^T A_2 x \leq h_i, \quad i = 1, \dots, l \quad (3.9.3a)$$

$$x \in \mathbb{R}^n \quad (3.9.3b)$$

is not solvable.

(ii) There exists a vector of nonnegative multipliers $y = (y_1, \dots, y_l)$ such that

$$\sum_{i=1}^l y_i (M_{i1}A_1 + M_{i2}A_2) \succeq 0 \quad (3.9.4a)$$

$$y^T h < 0 \quad (3.9.4b)$$

$$y \geq 0. \quad (3.9.4c)$$

Proof. First let us assume that we have a vector y satisfying (3.9.4), then multiplying the inequalities in (3.9.3) by the corresponding multiplier and taking the sum we get that for any solution of (3.9.3)

$$0 \leq x^T \underbrace{\left(\sum_{i=1}^l y_i (M_{i1}A_1 + M_{i2}A_2) \right)}_{\succeq 0} x \leq y^T h < 0, \quad (3.9.5)$$

which is a contradiction.

Let us assume now that (3.9.3) is not solvable. In this case the image

$$\{(x^T A_1 x, x^T A_2 x) : x \in \mathbb{R}^n\} \quad (3.9.6)$$

and the polyhedron $\{u \in \mathbb{R}^2 : Mu \leq h\}$ are nonempty, disjoint, convex sets, therefore they can be separated by a hyperplane, i.e., there exist multipliers z_1, z_2 such that

$$z_1 x^T A_1 x + z_2 x^T A_2 x \geq 0, \forall x \in \mathbb{R}^n \quad (3.9.7a)$$

$$z^T u < 0, \forall u : Mu \leq h. \quad (3.9.7b)$$

The first inequality states that a linear combination of the matrices is positive semidefinite, while the second one can be written in equivalent form using the well-known Farkas lemma (see, e.g., [35, 115, 120]). After these substitutions we get the statement of the theorem. \square

Remark 3.9.2. There is another way to look at this theorem. All the inequalities in system (3.9.3) are of the form $x^T B_i x \leq h_i$ where $B_i = M_{i1} A_1 + M_{i2} A_2$. In these terms what we proved is that the S-lemma remains true for the multi-inequality case, provided that all the matrices in the system are linear combinations of two matrices. This explains and generalizes an observation of Sturm and Zhang in [122, §6].

3.9.2 Conic case

For a second example let us use Theorem 3.5.2 and assume that \mathcal{C} is the three dimensional Lorentz cone, i.e.,

$$\mathcal{C} = \{(u, v, w) \in \mathbb{R}^3 : u^2 \geq v^2 + w^2, u \geq 0\}. \quad (3.9.8)$$

This set is a closed, convex cone. In what follows we will use the fact that the Lorentz cone is self-dual. We obtain the following theorem:

Theorem 3.9.3 (New duality theorem (conic case)). *Let $n \geq 3$, let $A, B, C \in \mathbb{R}^{n \times n}$ be symmetric matrices and assume that they have a positive definite linear combination. The following two statements are equivalent:*

(i) *The system*

$$(x^T A x)^2 \geq (x^T B x)^2 + (x^T C x)^2 \quad (3.9.9a)$$

$$x^T A x \geq 0 \quad (3.9.9b)$$

$$x \neq 0 \quad (3.9.9c)$$

is not solvable.

(ii) *There exist multipliers y_1, y_2, y_3 such that*

$$y_1A + y_2B + y_3C \prec 0 \quad (3.9.10a)$$

$$y_1^2 \geq y_2^2 + y_3^2 \quad (3.9.10b)$$

$$y_1 \geq 0. \quad (3.9.10c)$$

Proof. First let us assume that (3.9.9) is solvable and the multipliers in system (3.9.10) exist. Then for these solutions we have

$$y_1x^T Ax + y_2x^T Bx + y_3x^T Cx < 0. \quad (3.9.11)$$

On the other hand both (y_1, y_2, y_3) and $(x^T Ax, x^T Bx, x^T Cx)$ are in a three dimensional Lorentz cone, and since the Lorentz cone is self-dual, we have

$$y_1x^T Ax + y_2x^T Bx + y_3x^T Cx \geq 0 \quad (3.9.12)$$

contradicting (3.9.11).

Now assume that system (3.9.9) is not solvable. This means that

$$\{(x^T Ax, x^T Bx, x^T Cx) : x \in \mathbb{R}^n\} \cap \mathcal{C} = \{0\} \quad (3.9.13)$$

and since both sets are convex (the first one by Theorem 3.5.2, the other one by definition), they can be separated by a hyperplane going through the origin, i.e., there exist y_1, y_2, y_3 such that

$$y_1u + y_2v + y_3w \geq 0, \quad \forall u, v, w : u^2 \geq v^2 + w^2, u \geq 0 \quad (3.9.14a)$$

$$y_1x^T Ax + y_2x^T Bx + y_3x^T Cx < 0, \quad \forall x \neq 0. \quad (3.9.14b)$$

The first equation requires that (y_1, y_2, y_3) be in the dual cone of the Lorentz cone, but since the Lorentz cone is self-dual this is equivalent to (y_1, y_2, y_3) being in a Lorentz cone. This shows that we have a solution for system (3.9.10). \square

3.10 Summary

We have gone through several different approaches and examples showing that Yakubovich's S-lemma is only the tip of the iceberg. The theories that lead to this result are much more general and the S-lemma has a wealth of applications in various areas of applied mathematics. We demonstrated that the generalization for more inequalities is not practical, as the conditions needed become more and more complex. In fact, the minimal conditions under which the S-lemma holds with more inequalities can be obtained easily from the characterization theorems for convex images in §3.5. This answers a question posed as open in [37].

Chapter 4

Nonregular duality for symmetric cones

Music before all else,
and for that choose the irregular,
which is vaguer and melts better
into the air...

PAUL VERLAINE

In this chapter we present a strong duality theory for optimization problems over symmetric cones without assuming any constraint qualification. We show important complexity implications of the result to semidefinite and second order conic optimization. The result is an application of Borwein and Wolkowicz's facial reduction procedure to express the minimal cone. We use Pataki's simplified analysis and provide an explicit formulation for the minimal cone of a symmetric cone optimization problem. In the special case of semidefinite optimization the dual has better complexity than Ramana's strong semidefinite dual. After specializing the dual for second order cone optimization we argue that new software for homogeneous cone optimization problems should be developed. This chapter is based on [103].

4.1 Introduction

4.1.1 Historical background

In his seminal paper [110], Ramana presented an exact duality theory for semidefinite optimization without any constraint qualification. Later, in a joint paper [112] with Tunçel and Wolkowicz, they showed that the result can

be derived from a more general theorem for convex problems [25]. Our original intention was to develop a similar theory for second order cone optimization problems. We had hoped that a purely second order strong dual could be constructed. The continued failure to do so motivated us to look for more general classes of cones on the dual side, this is how we found the class of homogeneous cones. They are quite general, yet possess all the properties we need for the construction of the dual. The primal problem is defined over a symmetric (i.e., homogeneous and self dual) cone, but the strong dual problem will involve non self dual homogeneous cones. We use the general theory of homogeneous cones during the construction.

The discussion uses Pataki’s simplified analysis of the facial reduction algorithm for general convex conic optimization [96].

4.1.2 The structure of this chapter

We will use the terminology of Lagrange duality theory, see §2 for details. First, we present the facial reduction algorithm of Borwein and Wolkowicz with Pataki’s simplified analysis. In §4.3 we introduce homogeneous cones, discuss their constructions and basic properties. Section 4.4 contains the new results, the application of the facial reduction algorithm to the symmetric cone optimization problem. We prove that symmetric cones satisfy Pataki’s sufficient conditions and give an explicit formula for the cones in the dual problem. In §4.5 we specialize the result to semidefinite and second order conic optimization. We show that in the semidefinite case our dual has better complexity than Ramana’s strong semidefinite dual.

4.1.3 Preliminaries

Let $\mathcal{K} \subset \mathbb{R}^n$ be a closed, convex, pointed, solid cone. The optimization problem in the focus of this chapter is defined as follows:

$$\begin{aligned} \max \quad & b^T y \\ & A^T y + s = c \\ & s \in \mathcal{K}, \end{aligned} \tag{P}$$

where $A \in \mathbb{R}^{m \times n}$, $c, s \in \mathbb{R}^n$, $b, y \in \mathbb{R}^m$. The importance of this conic form comes from the result that under mild assumptions any convex optimization problem can be written this way, see [89] for details. The primal problem (P)

has a corresponding Lagrange-Slater dual (see §2):

$$\begin{aligned} \min \quad & \langle c, x \rangle \\ & Ax = b \\ & x \in \mathcal{K}^*, \end{aligned} \tag{D}$$

where $x \in \mathbb{R}^n$. We do not specify the exact form of the scalar product $\langle \cdot, \cdot \rangle$ yet. Recall that for strong duality we usually have to assume some constraint qualification, otherwise strong duality fails to hold, see §2.7.2. In this chapter no such condition is assumed.

4.2 The minimal cone and the facial reduction algorithm

Let \mathcal{K} be a closed, convex cone, then a closed set $\mathcal{F} \subseteq \mathcal{K}$ is by definition a *face* of \mathcal{K} if for every $x, y \in \mathcal{K}$, $x + y \in \mathcal{F}$ implies $x, y \in \mathcal{F}$. This relation is denoted by $\mathcal{F} \trianglelefteq \mathcal{K}$. If $\mathcal{C} \subseteq \mathcal{K}$ is a subset of \mathcal{K} then $\mathcal{F}(\mathcal{C})$ denotes the face generated by \mathcal{C} in \mathcal{K} , i.e., the smallest face of \mathcal{K} containing \mathcal{C} . This is given as

$$\mathcal{F}(\mathcal{C}) = \bigcap \{ \mathcal{F} : \mathcal{C} \subseteq \mathcal{F} \trianglelefteq \mathcal{K} \}. \tag{4.2.1}$$

Since the intersection of faces is a face, $\mathcal{F}(\mathcal{C})$ is indeed a face and its minimality is due to the construction. Let \mathcal{F} be a face of \mathcal{K} , then its complementary (or conjugate) face is defined as $\mathcal{F}^c = \mathcal{F}^\perp \cap \mathcal{K}^* \trianglelefteq \mathcal{K}^*$. A face and its complementary face belong to the primal and dual cones and are orthogonal. If $\mathcal{F} \trianglelefteq \mathcal{K}^*$ then we use the same notation for $\mathcal{F}^c = \mathcal{F}^\perp \cap \mathcal{K}$. This ambiguity will not cause a problem as \mathcal{F} always determines the appropriate cone.

For the optimization problem (P) we can define a corresponding *minimal cone* (see [25, 110, 112]), which is the smallest face of \mathcal{K} containing all the feasible solutions s of (P).

$$\mathcal{K}_{\min} = \mathcal{F}(\{c - A^T y : y \in \mathbb{R}^m\} \cap \mathcal{K}). \tag{4.2.2}$$

Replacing cone \mathcal{K} with \mathcal{K}_{\min} in (P) we get an equivalent problem:

$$\begin{aligned} \max \quad & b^T y \\ & A^T y + s = c \\ & s \in \mathcal{K}_{\min}. \end{aligned} \tag{P}_{\min}$$

Moreover, due to the minimality of \mathcal{K}_{\min} this new problem satisfies the Slater condition,¹³ thus it is in strong duality with its Lagrange dual:

$$\begin{aligned} \min \quad & \langle c, x \rangle & (\text{D}_{\min}) \\ \text{Ax} = & b \\ x \in & \mathcal{K}_{\min}^*. \end{aligned}$$

This abstract dual is quite useless unless we can express either the primal minimal cone or its dual explicitly. One algorithmic description is given by Borwein and Wolkowicz in [25]. The following simplified form is due to Pataki [96]. First let us define an interesting set of cones:

Definition 4.2.1 (Nice cone). *A convex cone \mathcal{K} is nice if $\mathcal{F}^* = \mathcal{K}^* \oplus \mathcal{F}^\perp$ for all faces $\mathcal{F} \trianglelefteq \mathcal{K}$. Equivalently, since $\mathcal{F}^* \subseteq \text{cl}(\mathcal{K}^* \oplus \mathcal{F}^\perp)$, \mathcal{K} is nice if $\mathcal{K}^* \oplus \mathcal{F}^\perp$ is closed.*

Most of the cones arising in practical applications (polyhedral, semidefinite and Lorentz cones) are nice. Now consider the following system:

$$\begin{aligned} A(u + v) &= 0 \\ \langle c, u + v \rangle &= 0 & (\text{FR}) \\ (u, v) &\in \mathcal{K}^* \times \mathcal{F}^\perp. \end{aligned}$$

The next lemma was proved by Borwein and Wolkowicz in [25].

Lemma 4.2.2 (Facial reduction lemma). *If $\mathcal{K}_{\min} \trianglelefteq \mathcal{F} \trianglelefteq \mathcal{K}$ then for any solution u, v of system (FR) we have $\mathcal{K}_{\min} \subseteq \{u\}^\perp \cap \mathcal{F} \trianglelefteq \mathcal{F}$. If $\mathcal{K}_{\min} \subsetneq \mathcal{F}$ then there is a solution (u, v) such that $\mathcal{F} \cap \{u\}^\perp \subsetneq \mathcal{F}$. In this case we say that u is a reducing certificate for the system $c - A^T y \in \mathcal{F}$.*

Proof. As $0 = \langle u + v, (A^T y - c) \rangle = \langle u, (A^T y - c) \rangle$ for all values of y , we get that $\{u\}^\perp \supseteq \mathcal{K}_{\min}$. Moreover, if $x, y \in \mathcal{F}$ and $x + y \in \{u\}^\perp \cap \mathcal{F}$ then $\langle u, (x + y) \rangle = 0$, but since $u \in \mathcal{F}^*$, both $\langle u, x \rangle$ and $\langle u, y \rangle$ are nonnegative, thus they are both 0. This shows that $x, y \in \{u\}^\perp \cap \mathcal{F}$, thus $\{u\}^\perp \cap \mathcal{F}$ is a face of \mathcal{F} .

To prove the second statement let us choose an $f \in \text{relint}(\mathcal{F})$, then $\mathcal{K}_{\min} \neq \mathcal{F}$ if and only if $c - A^T y - \alpha f \in \mathcal{F}$ implies $\alpha \leq 0$. This system is strictly feasible, thus there is a certificate for this implication, i.e., $\mathcal{K}_{\min} \neq \mathcal{F}$

¹³This is true because we use a relaxed version of the Slater condition, see Definition 2.2.1.

if and only if

$$\begin{aligned}
& \exists x \in \mathcal{F}^* \\
& Ax = 0 \\
& \langle c, x \rangle \leq 0 \\
& \langle f, x \rangle = 1,
\end{aligned} \tag{4.2.3}$$

which (since \mathcal{K} is a nice cone) is further equivalent to

$$\begin{aligned}
& \exists (u, v) \in \mathcal{K}^* \times \mathcal{F}^\perp \\
& A(u + v) = 0 \\
& \langle c, u + v \rangle \leq 0 \\
& \langle f, u + v \rangle = 1.
\end{aligned} \tag{4.2.4}$$

Notice that we must have $\langle c, x \rangle = 0$ otherwise (P) would be infeasible by the convex Farkas theorem ([120], §6.10). Further, $\langle f, v \rangle = 0$, since $f \in \mathcal{F}$ and $v \in \mathcal{F}^\perp$. This implies that $\langle f, u \rangle = 1$, thus $u \notin \mathcal{F}^\perp$, which translates to $\{u\}^\perp \not\subseteq \mathcal{F}$, proving that $\mathcal{F} \cap \{u\}^\perp \subsetneq \mathcal{F}$. \square

We can turn the result into an algorithm to construct \mathcal{K}_{\min} by repeatedly intersecting \mathcal{K} with subspaces until we arrive at \mathcal{K}_{\min} . This is the idea behind the dual construction of Borwein and Wolkowicz in [25]. We use a simplified form, see Algorithm 4.1.

The correctness of the general algorithm was established in [25], for our simplified form see Pataki's papers [96] and [94, 95].

Theorem 4.2.3 (Facial reduction algorithm for nice cones).

If \mathcal{K} is a nice cone then

1. *During the algorithm ($i = 0, \dots, \ell$):*

(a) $\mathcal{K}_{\min} \subseteq \mathcal{F}_i$ and

(b) $\mathcal{F}_i = \mathcal{F}_{i-1} \cap \{u^i\}^\perp = \mathcal{K} \cap \{u^0 + \dots + u^i\}^\perp = (\mathcal{F}(u^0 + \dots + u^i))^c$.

2. *The facial reduction algorithm is finite and it returns*

$$\mathcal{K}_{\min} = \mathcal{F}_\ell = \mathcal{K} \cap \{u^0 + \dots + u^\ell\}^\perp = (\mathcal{F}(u^0 + \dots + u^\ell))^c. \tag{4.2.6}$$

3. *The number of iterations is $\ell \leq L$, where*

$$L = \min \{ \dim(\text{Ker}(A) \cap c^\perp), \text{length of the longest chain of faces in } \mathcal{K} \}. \tag{4.2.7}$$

Algorithm 4.1 The facial reduction algorithm**Input:** A , c and \mathcal{K} Set $(u^0, v^0) = (0, 0)$, $\mathcal{F}_0 = \mathcal{K}$, $i = 0$ **while** $\mathcal{K}_{\min} \neq \mathcal{F}_i$ Find (u^{i+1}, v^{i+1}) satisfying

$$\begin{aligned} A(u + v) &= 0 \\ \langle c, u + v \rangle &= 0 \\ (u, v) &\in \mathcal{K}^* \times (\mathcal{F}_i)^\perp. \end{aligned} \tag{FR}_i$$

Set $\mathcal{F}_{i+1} = \mathcal{F}_i \cap \{u^{i+1}\}^\perp$, $i = i + 1$ **end while****Output:** $\ell \geq 0$, $u^0, u^1, \dots, u^\ell \in \mathcal{K}^*$ such that

$$\mathcal{K}_{\min} = \mathcal{K} \cap \{u^0 + u^1 + \dots + u^\ell\}^\perp \tag{4.2.5}$$

In general, this algorithm is not a viable method to express the minimal cone, since the solution of the auxiliary system is comparable to solving the original problem. The significance of the algorithm is its finiteness: this implies that by combining all the auxiliary systems (FR_i) from the iterations we can construct an extended strong dual problem for (P), assuming \mathcal{K} is a nice cone.

$$\begin{aligned} \min \quad & \langle c, (u^{L+1} + v^{L+1}) \rangle \\ A(u^{L+1} + v^{L+1}) &= b \\ A(u^i + v^i) &= 0, \quad i = 1, \dots, L \\ \langle c, (u^i + v^i) \rangle &= 0, \quad i = 1, \dots, L \\ u^0, v^0 &= 0 \\ u^i &\in \mathcal{K}^*, \quad i = 1, \dots, L + 1 \\ v^i &\in \mathcal{F}_{i-1}^\perp, \quad i = 1, \dots, L + 1. \end{aligned} \tag{ED}_{\text{nice}}$$

Our original dual variable would only be $x = u^{L+1} + v^{L+1}$, the additional variables are needed to describe the dual of the minimal cone. Unfortunately, this dual is useful only if we can give an explicit expression for \mathcal{F}_i^\perp . For the case when \mathcal{K} is the cone of positive semidefinite matrices then this description is given in [112]. In [96], Pataki poses the open problem to find a broader class of cones for which such a description is possible. In this chapter we show that the theory extends to symmetric cones and we also show a possible way to extend the algorithm to general homogeneous cones.

4.3 The geometry of homogeneous cones

4.3.1 Definition, basic properties

The central objects of this chapter are the homogeneous cones:

Definition 4.3.1 (Homogeneous cones). *A closed, convex cone $\mathcal{K} \subseteq \mathbb{R}^n$ with nonempty interior is homogeneous if for any $u, v \in \text{int}(\mathcal{K})$ there exists an invertible linear map $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that*

1. $\varphi(\mathcal{K}) = \mathcal{K}$, i.e., φ is an automorphism of \mathcal{K} , and
2. $\varphi(u) = v$.

In other words, the group of automorphisms of \mathcal{K} acts transitively on the interior of \mathcal{K} .

Typical examples of homogeneous cones are polyhedral cones in \mathbb{R}^n with exactly n extreme rays (e.g., the nonnegative orthant), the set of symmetric positive semidefinite matrices and the second order or Lorentz cone.

It is important to note here that if \mathcal{K} is homogeneous then so is its dual, \mathcal{K}^* . A cone is self-dual if $\mathcal{K} = \mathcal{K}^*$; a typical homogeneous cone is not self-dual. Self-dual homogeneous cones are called symmetric, and they are very special, see [42] for more details. We have to note here that the usual approach to symmetric cones is to use Jordan algebras. In our case it will be very important that the cone is homogeneous and therefore T-algebras provide a more powerful analysis. For a classical text on symmetric cones and T-algebras see [128].

4.3.2 Constructing homogeneous cones from T-algebras

Definition 4.3.1 does not give too much insight into the structure of homogeneous cones. Fortunately, there is a constructive way to build homogeneous cones using generalized matrices. We briefly summarize the necessary properties and techniques based on [32].

A generalized matrix is a matrix-like structure whose elements are vectors. Formally, let r be a positive integer, $1 \leq i, j \leq r$, and let $\mathcal{A}_{ij} \subseteq \mathbb{R}^{n_{ij}}$ be a vector space of dimension n_{ij} . Let us assume that for every $1 \leq i, j, k, \ell \leq r$ we have a bilinear product such that if $a_{ij} \in \mathcal{A}_{ij}$ and $a_{k\ell} \in \mathcal{A}_{k\ell}$ then

$$a_{ij}a_{k\ell} \in \begin{cases} \mathcal{A}_{i\ell}, & \text{if } j = k, \\ \{0\}, & \text{if } j \neq k. \end{cases} \quad (4.3.1)$$

Now consider $\mathcal{A} = \bigoplus_{i,j=1}^r \mathcal{A}_{ij}$, an algebra of rank r . Every element $a \in \mathcal{A}$ is a generalized matrix, while a_{ij} is an n_{ij} -dimensional generalized element of the matrix, $a_{ij} \in \mathcal{A}_{ij}$. In other words, a_{ij} is the projection of a onto \mathcal{A}_{ij} . The multiplication of two elements $a, b \in \mathcal{A}$ is analogous to the multiplication of matrices, i.e.,

$$(ab)_{ij} = \sum_{k=1}^r a_{ik}b_{kj}. \quad (4.3.2)$$

An involution $*$ of the matrix algebra \mathcal{A} is a linear mapping such that for every $a, b \in \mathcal{A}$ and $1 \leq i, j \leq r$

1. $*$: $\mathcal{A} \rightarrow \mathcal{A}$,
2. $a^{**} = a$,
3. $(ab)^* = b^*a^*$,
4. $\mathcal{A}_{ij}^* = \mathcal{A}_{ji}$,
5. $(a^*)_{ij} = (a_{ij})^*$.

This generalizes the classical notion of the (conjugate) transpose. From now on we assume that our matrix algebra is equipped with an involution. Let

$$\mathcal{T} = \bigoplus_{i \leq j} \mathcal{A}_{ij} \quad (4.3.3)$$

be the set of upper triangular elements, and let

$$\mathcal{H} = \{a \in \mathcal{A} : a = a^*\} \quad (4.3.4)$$

be the set of Hermitian elements. Assume that for every i , \mathcal{A}_{ii} is isomorphic to \mathbb{R} , let ρ_i be the isomorphism and let e_i denote the unit element of \mathcal{A}_{ii} . We define the trace of an element¹⁴ as

$$\text{tr}(a) = \sum_{i=1}^r \rho_i(a_{ii}). \quad (4.3.5)$$

We will need some technical conditions about \mathcal{A} , these are summarized in the following definition, originally introduced by Vinberg in [127].

¹⁴Actually, we will not make use of the explicit form of this definition. Any formula satisfying Definition 4.3.2 would work.

Definition 4.3.2 (T-algebra). *The set of generalized matrices \mathcal{A} is a T-algebra if for all $a, b, c \in \mathcal{A}$, $t, u, v \in \mathcal{T}$ and $1 \leq i, j \leq r$*

1. \mathcal{A}_{ii} is isomorphic to \mathbb{R} ,
2. $e_i a_{ij} = a_{ij} e_j = a_{ij}$ for all $a_{ij} \in \mathcal{A}_{ij}$,
3. $\text{tr}(ab) = \text{tr}(ba)$,
4. $\text{tr}(a(bc)) = \text{tr}((ab)c)$,
5. $\text{tr}(a^*a) \geq 0$, and $\text{tr}(a^*a) = 0$ implies $a = 0$,
6. $t(uv) = (tu)v$,
7. $t(uv^*) = (tu)v^*$.

It is important to note that multiplication in a T-algebra is neither commutative nor associative.

Based on the properties of the trace we can define a natural inner product on \mathcal{A} , namely $\langle a, b \rangle = \text{tr}(a^*b)$, which will provide us a Hilbert-space structure. Let now

$$\mathcal{I} = \{t \in \mathcal{T} : \rho_i(t_{ii}) > 0, 1 \leq i \leq r\} \quad (4.3.6)$$

be the set of upper triangular matrices with positive diagonal elements, and define

$$\mathcal{K}(\mathcal{A}) = \{tt^* : t \in \mathcal{I}\} \subseteq \mathcal{H}. \quad (4.3.7)$$

The fundamental representation theorem of homogeneous cones was proved by Vinberg [127]:

Theorem 4.3.3 (Representation of homogeneous cones). *A cone \mathcal{K} is homogeneous if and only if there is a T-algebra \mathcal{A} such that $\text{int}(\mathcal{K})$ is isomorphic to $\mathcal{K}(\mathcal{A})$. Moreover, given $\mathcal{K}(\mathcal{A})$, the representation of an element from $\text{int}(\mathcal{K})$ in the form tt^* is unique. Finally, the interior of the dual cone \mathcal{K}^* can be represented as $\{t^*t : t \in \mathcal{I}\}$.*

The rank of the homogeneous cone \mathcal{K} is the rank of the generalized matrix algebra \mathcal{A} . It is denoted by $\text{rank}(\mathcal{K})$ and it is the size of the generalized matrices used in the construction.

Remark 4.3.4. This theorem is analogous to the representation of symmetric cones as the set of squares over a Jordan algebra. However, for T-algebras and homogeneous cones it is not true that for a given $a \in \mathcal{A}$ we have $aa^* \in \mathcal{K}$.

Remark 4.3.5. The positivity of the diagonal elements ($t \in \mathcal{I}$) is only required to ensure uniqueness. In general every $x \in \mathcal{K}$ can be expressed as $x = tt^*$ for some $t \in \mathcal{T}$.

Example 4.3.6. The cone of positive semidefinite matrices is homogeneous, since every positive definite matrix admits a unique Cholesky factorization of the form UU^T with positive diagonal elements in U .

Example 4.3.7 (Rotated Lorentz cone). By the classical definition, a rotated Lorentz cone is the following set:

$$\mathbb{L}_r = \left\{ (x_0, x_1, x) \in \mathbb{R}^{n+2} : x_0 \geq 0, x_0x_1 - \|x\|^2 \geq 0 \right\}. \quad (4.3.8)$$

Rotated Lorentz cones can be represented as generalized matrices in the following way. Let the T-algebra be defined as

$$\mathcal{A}_{\mathbb{L}_r} = \left\{ \begin{pmatrix} v_0 & v^T \\ u & u_0 \end{pmatrix} : u_0, v_0 \in \mathbb{R}, u, v \in \mathbb{R}^n \right\}, \quad (4.3.9)$$

with the product

$$\begin{pmatrix} v_0 & v^T \\ u & u_0 \end{pmatrix} \begin{pmatrix} q_0 & q^T \\ p & p_0 \end{pmatrix} = \begin{pmatrix} v_0q_0 + v^Tq & v_0q^T + p_0v^T \\ q_0u + u_0p & u_0p_0 + q^Tu \end{pmatrix}. \quad (4.3.10)$$

The involution is the transpose. It is straightforward to verify that this algebra does satisfy all the axioms of Definition 4.3.2, therefore it is a T-algebra. This implies that the set

$$\mathcal{K}(\mathcal{A}_{\mathbb{L}_r}) = \left\{ \begin{pmatrix} t_1 & t^T \\ 0 & t_0 \end{pmatrix} \begin{pmatrix} t_1 & 0 \\ t & t_0 \end{pmatrix} : t_0, t_1 > 0 \right\} \quad (4.3.11)$$

is a homogeneous cone. Now considering the element

$$\begin{pmatrix} x_1 & x^T \\ x & x_0 \end{pmatrix} \in \mathcal{A}_{\mathbb{L}_r} \quad (4.3.12)$$

with $x_0, x_1 > 0$ and $(x_0, x_1, x) \in \mathbb{L}_r$. We have the factorization

$$\begin{pmatrix} x_1 & x^T \\ x & x_0 \end{pmatrix} = \begin{pmatrix} \sqrt{x_1 - \frac{\|x\|^2}{x_0}} & \frac{x^T}{\sqrt{x_0}} \\ 0 & \sqrt{x_0} \end{pmatrix} \begin{pmatrix} \sqrt{x_1 - \frac{\|x\|^2}{x_0}} & 0 \\ \frac{x}{\sqrt{x_0}} & \sqrt{x_0} \end{pmatrix}, \quad (4.3.13)$$

establishing an isomorphism between \mathbb{L}_r and $\mathcal{K}(\mathcal{A}_{\mathbb{L}_r})$. It is interesting to note that the rotated Lorentz cone arises more naturally in this framework, while standard Lorentz cones are easier to construct in Jordan algebras. Naturally, they are isomorphic.

4.3.3 Recursive construction of homogeneous cones

Homogeneous cones can also be constructed in a recursive way, due to Vinberg [127]. Here we follow [57].

Definition 4.3.8 (Homogeneous symmetric form on a cone).

Let $\mathcal{K} \subseteq \mathbb{R}^n$ be a homogeneous cone, and consider a symmetric bilinear mapping $B : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}^n$ such that for every $u, v \in \mathbb{R}^p$ and $\lambda_1, \lambda_2 \in \mathbb{R}$

1. $B(\lambda_1 u_1 + \lambda_2 u_2, v) = \lambda_1 B(u_1, v) + \lambda_2 B(u_2, v)$,
2. $B(u, v) = B(v, u)$,
3. $B(u, u) \in \mathcal{K}$,
4. $B(u, u) = 0$ implies $u = 0$.

A symmetric bilinear form B is called homogeneous if \mathcal{K} is a homogeneous cone and there is a transitive subgroup $G \subseteq \text{Aut}(\mathcal{K})$ such that for every $g \in G$ there is a linear transformation \tilde{g} on \mathbb{R}^p such that

$$g(B(u, v)) = B(\tilde{g}(u), \tilde{g}(v)), \quad (4.3.14)$$

in other words, the diagram

$$\begin{array}{ccc} \mathbb{R}^p \times \mathbb{R}^p & \xrightarrow{\tilde{g} \times \tilde{g}} & \mathbb{R}^p \times \mathbb{R}^p \\ B \downarrow & & \downarrow B \\ \mathbb{R}^n & \xrightarrow{g} & \mathbb{R}^n \end{array} \quad (4.3.15)$$

is commutative.

Having such a bilinear function we can define a new set:

Definition 4.3.9 (Siegel cone). The Siegel cone $\text{SC}(\mathcal{K}, B)$ of \mathcal{K} and B is defined as

$$\text{SC}(\mathcal{K}, B) = \text{cl}(\{(x, u, t) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R} : t > 0, tx - B(u, u) \in \mathcal{K}\}). \quad (4.3.16)$$

Theorem 4.3.10 (Properties of the Siegel cone).

The Siegel cone has the following properties:

1. If \mathcal{K} is a homogeneous cone and B is a homogeneous symmetric bilinear form then $\text{SC}(\mathcal{K}, B)$ is a homogeneous cone.

2. Every homogeneous cone can be obtained as the Siegel cone of another homogeneous cone using an appropriate bilinear function B .
3. For the rank of the Siegel cone we have

$$\text{rank}(\text{SC}(\mathcal{K}, B)) = \text{rank}(\mathcal{K}) + 1, \quad (4.3.17)$$

i.e., the rank of a homogeneous cone is exactly the number of Siegel extension steps needed to construct the cone starting from $\{0\}$, the cone of rank 0.

This result is analogous to the concept of the Schur complement for semidefinite matrices. It expresses a quadratic relation with a linear one of higher rank and dimension.

Example 4.3.11 (The rotated Lorentz cone as a Siegel cone). Let us see how the rotated Lorentz cone \mathbb{L}_r can be constructed this way. Starting with $\mathcal{K} = \mathbb{R}_+$ as a homogeneous cone of rank 1, we choose the bilinear function $B : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ to be the usual scalar product, i.e., $B(u, v) = u^T v$. All the automorphisms of \mathbb{R}_+ are multiplications with a positive number, thus if $g(u) = \alpha u$ then with $\tilde{g}(u) = \sqrt{\alpha} u$ we have

$$g(B(u, v)) = \alpha u^T v = (\sqrt{\alpha} u)^T (\sqrt{\alpha} v) = B(\tilde{g}(u), \tilde{g}(v)). \quad (4.3.18)$$

This shows that B is a homogeneous symmetric form. Now the resulting Siegel cone is

$$\text{SC}(\mathcal{K}, B) = \{(x, u, t) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} : t \geq 0, tx \geq \|u\|^2\}, \quad (4.3.19)$$

which is exactly the rotated Lorentz cone. This shows that the rank of this cone is indeed 2.

4.3.4 Another representation for homogeneous cones

Despite the quite abstract definition, it turns out that homogeneous cones are slices of the positive semidefinite cone. More precisely (see [32]):

Proposition 4.3.12 (Homogeneous cones as semidefinite slices).

If $\mathcal{K} \subset \mathbb{R}^n$ is a homogeneous cone then there exist an $m \leq n$ and an injective linear map $M : \mathbb{R}^n \rightarrow \mathbb{S}^{m \times m}$ such that

$$M(\mathcal{K}) = \mathbb{S}_+^{m \times m} \cap M(\mathbb{R}^n) \quad (4.3.20)$$

A similar result was obtained independently by Faybusovich, [44]. We have to note here that not all slices of the positive semidefinite cone are homogeneous, see [32] for a counterexample.

4.3.5 Self dual homogeneous (symmetric) cones

Throughout the rest of this chapter we will assume that \mathcal{K} is a symmetric cone, i.e., it is homogeneous and self dual. In this case the theory can be simplified a bit, see [128]. In particular, we will use the following proposition:

Proposition 4.3.13 (Special properties of symmetric cones). *If \mathcal{A} is a T-algebra and \mathcal{K} is the corresponding self dual homogeneous cone, then for every $a, b \in \mathcal{A}$ we have*

$$\operatorname{tr}((aa^*)(bb^*)) = \operatorname{tr}((ab)(ab)^*). \quad (4.3.21)$$

Consequently, $aa^* \in \mathcal{K}$ for every $a \in \mathcal{A}$. Moreover, the mapping

$$g_w : u \mapsto wuw^* \quad (4.3.22)$$

is well-defined for every $u \in \mathcal{H}$.

Remark 4.3.14. Remember that in the case of a general T-algebra the identity (4.3.21) holds typically only if $a, b^* \in \mathcal{T}$, and in that case $aa^* \in \mathcal{K}$ and $bb^* \in \mathcal{K}^*$.

4.4 An exact duality theory for symmetric cones

Now we are ready to present our results about the strong dual for the symmetric cone optimization problems. First let us repeat the primal problem:

$$\begin{aligned} \max \quad & b^T y \\ A^T y + s &= c \\ s &\in \mathcal{K}, \end{aligned} \quad (\text{P})$$

where $A \in \mathbb{R}^{m \times n}$, $c, s \in \mathbb{R}^n$, $b, y \in \mathbb{R}^m$. We assume that $\mathcal{K} = \mathcal{K}^*$ is a homogeneous cone given in the form of $\mathcal{K}(\mathcal{A})$, i.e., every element in $\operatorname{int}(\mathcal{K})$ is represented as tt^* with $t \in \mathcal{I}$. Our goal is to derive a dual for (P) satisfying the following requirements (cf. [110, §1.4]):

1. The dual problem is a homogeneous cone optimization problem that can be generated easily from the primal input data.
2. If the primal problem is feasible and bounded then the duality gap (the difference of the optimal primal and dual objective values) is 0, and the optimum is attained on the dual side.

3. It yields a theorem of the alternative for symmetric conic feasibility systems, i.e., the infeasibility of a symmetric conic system can be characterized by the feasibility of a homogeneous conic system.

4.4.1 The facial reduction algorithm for symmetric cones

We will use the facial reduction algorithm presented in §4.2. In light of Theorem 4.2.3 we have to establish two things:

1. prove that symmetric cones are nice, therefore the facial reduction algorithm can be applied;
2. find an explicit expression for the space $(\mathcal{F}_i)^\perp$ and show that it can be expressed with homogeneous cones.

For the proof of the first part we need some lemmas.

Lemma 4.4.1 (Faces of intersections of cones). *If $\mathcal{K} = \mathcal{K}_1 \cap \mathcal{K}_2$ then $\mathcal{F} \trianglelefteq \mathcal{K}$ if and only if $\mathcal{F} = \mathcal{F}_1 \cap \mathcal{F}_2$ where $\mathcal{F}_1 \trianglelefteq \mathcal{K}_1$ and $\mathcal{F}_2 \trianglelefteq \mathcal{K}_2$. In other words, a face of the intersection of cones is the intersection of faces of the cones.*

Remark 4.4.2. This classical result, commonly attributed to Dubins [40], first appeared in [23]. An easily accessible source is [120, Theorem 3.6.19].

Pataki [91] proves the following lemma about the intersection of nice cones. We include a short proof for completeness.

Lemma 4.4.3 (Intersections of nice cones).

The intersection of nice cones is nice.

Proof. Let \mathcal{K}_1 and \mathcal{K}_2 be nice cones, $\mathcal{K} = \mathcal{K}_1 \cap \mathcal{K}_2$, $\mathcal{F} \trianglelefteq \mathcal{K}$. From the previous lemma we know that \mathcal{F} can be expressed as $\mathcal{F} = \mathcal{F}_1 \cap \mathcal{F}_2$, with $\mathcal{F}_i \trianglelefteq \mathcal{K}_i$, $i = 1, 2$.

$$\begin{aligned} \mathcal{F}^* &= (\mathcal{F}_1 \cap \mathcal{F}_2)^* = \text{cl}(\mathcal{F}_1^* \oplus \mathcal{F}_2^*) = \text{cl}((\mathcal{K}_1^* \oplus \mathcal{F}_1^\perp) \oplus (\mathcal{K}_2^* \oplus \mathcal{F}_2^\perp)) \\ &= \text{cl}((\mathcal{K}_1^* \oplus \mathcal{K}_2^*) \oplus (\mathcal{F}_1^\perp \oplus \mathcal{F}_2^\perp)) = (\mathcal{K}_1 \cap \mathcal{K}_2)^* \oplus (\mathcal{F}_1 \cap \mathcal{F}_2)^\perp \\ &= \mathcal{K}^* \oplus \mathcal{F}^\perp, \end{aligned} \tag{4.4.1}$$

where we used the basic properties of the dual cones. □

These two lemmas lead us to the following theorem:

Theorem 4.4.4. *Homogeneous cones (and thus symmetric cones) are nice.*

Proof. By Proposition 4.3.12 homogeneous cones are slices of the positive semidefinite cone. The semidefinite cone and all the subspaces are nice, thus their intersection is nice. \square

Remark 4.4.5. So far we have not used the fact the \mathcal{K} is self dual.

For the second part we will consider the following bilinear function:

$$B : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n \quad (4.4.2a)$$

$$B(u, v) = \frac{1}{2}(uv^* + vu^*) \quad (4.4.2b)$$

Our first result ensures that B is an appropriate bilinear form for our construction:

Lemma 4.4.6. *The bilinear function B satisfies all the requirements of Definition 4.3.8 thus it is a homogeneous symmetric form for \mathcal{K} and consequently $\text{SC}(\mathcal{K}, B)$ is a valid Siegel cone.*

Proof. Bilinearity and symmetry are obvious. Further,

$$B(u, u) = uu^* \in \mathcal{K} \quad (4.4.3)$$

by Proposition 4.3.13. If $B(u, u) = 0$ then necessarily $u = 0$.

To prove the homogeneity consider the group (see [127, 128])

$$G = \{g_w \in \text{Aut}(\mathcal{K}) : g_w(u) = wuw^* \text{ for an invertible element } w \in \mathcal{T}\}. \quad (4.4.4)$$

The mapping g_w is well defined by Proposition 4.3.13 and G is indeed a group. Its transitivity is due to the fact that $g_{w^{-1}}(ww^*) = ee^* = e$. For $g_w \in G$ let us define $\tilde{g}_w(u) = wu$. We then have

$$\begin{aligned} g_w(B(u, v)) &= \frac{1}{2}w(uv^* + vu^*)w^* \\ &= \frac{1}{2}((wu)(wv)^* + (wv)(wu)^*) \\ &= B(\tilde{g}_w(u), \tilde{g}_w(v)). \end{aligned} \quad (4.4.5)$$

This completes the proof of the homogeneity of B . \square

Remember that if $\mathcal{C} \subseteq \mathcal{K}$ is a convex subset of \mathcal{K} , then $\mathcal{F}(\mathcal{C})$ denotes the smallest face of \mathcal{K} containing \mathcal{C} . If $\mathcal{F} \trianglelefteq \mathcal{K}$ is a face of \mathcal{K} then its conjugate face is defined as $\mathcal{F}^c = \mathcal{F}^\perp \cap \mathcal{K}^*$, where the orthogonal complement is now defined relative to the set \mathcal{H} of self-adjoint elements. The following result is a generalization of Lemma 2.1 in [112] and it plays a key role in the construction of the dual problem:

Theorem 4.4.7 (The description of $((\mathcal{F}(\mathcal{C}))^c)^\perp$). *If $\mathcal{C} \subseteq \mathcal{K}$ is a convex subset of the symmetric cone \mathcal{K} then*

$$((\mathcal{F}(\mathcal{C}))^c)^\perp = \{w + w^* : \exists u \in \mathcal{C}, u - B(w, w) \in \mathcal{K}\}. \quad (4.4.6)$$

Proof. First, let $u \in \mathcal{C}$ and $w \in \mathcal{A}$ be such that $u - B(w, w) \in \mathcal{K}$ and let us choose an arbitrary element $v \in (\mathcal{F}(\mathcal{C}))^c \subseteq \mathcal{K}^*$. By the properties of the dual cone we have

$$\langle v, u - B(w, w) \rangle \geq 0. \quad (4.4.7)$$

Since $u \in \mathcal{F}(\mathcal{C})$ and $v \in \mathcal{F}(\mathcal{C})^\perp$, we have $\langle v, u \rangle = 0$. Now $B(w, w) \in \mathcal{K}$ implies that

$$0 = \langle v, B(w, w) \rangle = \langle v, ww^* \rangle. \quad (4.4.8)$$

Now as $v \in \mathcal{K}^*$ it can be written as $v = t^*t$ for some upper triangular element $t \in \mathcal{T}$. This gives

$$0 = \langle v, ww^* \rangle = \text{tr}((tw)(tw)^*), \quad (4.4.9)$$

which implies that $tw = 0$ thus

$$\langle v, w + w^* \rangle = \text{tr}(v(w + w^*)) = \text{tr}(t^*t(w + w^*)) = \text{tr}(t^*(tw) + (w^*t^*)t) = 0. \quad (4.4.10)$$

This shows that $w + w^* \in ((\mathcal{F}(\mathcal{C}))^c)^\perp$.

For the converse direction let us take $w + w^* \in ((\mathcal{F}(\mathcal{C}))^c)^\perp$ and $u \in \mathcal{C} \cap \text{relint}(\mathcal{F}(\mathcal{C}))$. At this point we do not specify w , we only fix the sum $w + w^*$. The exact form of w will be specified later. We will show that there exists an $\alpha \geq 0$ such that $\alpha u - B(w, w) \in \mathcal{K}$, for this we need to show that $B(w, w) \in \mathcal{F}(\mathcal{C})$.

Due to the homogeneity of \mathcal{K} there is an automorphism $\varphi \in \text{Aut}(\mathcal{K})$ such that $\varphi(u)$ is a diagonal element,¹⁵ or in other words,

$$\varphi(u) = \begin{pmatrix} \text{Diag}(d_{11}, \dots, d_{kk}) & 0 \\ 0 & 0 \end{pmatrix}, \quad (4.4.11)$$

where the diagonal elements are positive, i.e., $\rho(d_{11}), \dots, \rho(d_{kk}) > 0$. The generated face is

$$\mathcal{F}(\mathcal{C}) = \mathcal{F}(u) = \left\{ \varphi^{-1} \left[\begin{pmatrix} s & 0 \\ 0 & 0 \end{pmatrix} \right] : \begin{pmatrix} s & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{K} \right\}, \quad (4.4.12)$$

¹⁵Remember that we consider the elements of the homogeneous cone in matrix form.

and the complementary face is

$$\mathcal{F}(\mathcal{C})^c = \mathcal{F}(u)^c = \left\{ \varphi^{-1} \left[\begin{pmatrix} 0 & 0 \\ 0 & z \end{pmatrix} \right] : \begin{pmatrix} 0 & 0 \\ 0 & z \end{pmatrix} \in \mathcal{K}^* \right\}. \quad (4.4.13)$$

Now if $w + w^* \in (\mathcal{F}(\mathcal{C})^c)^\perp$ then it can be expressed as

$$w + w^* = \varphi^{-1} \left[\begin{pmatrix} p_1 & p_2 \\ p_2^* & 0 \end{pmatrix} \right] \quad (4.4.14)$$

with a symmetric p_1 . Defining

$$w = \varphi^{-1} \left[\begin{pmatrix} \frac{1}{2}p_1 & p_2 \\ 0 & 0 \end{pmatrix} \right] \quad (4.4.15)$$

we have

$$\begin{aligned} B(w, w) &= ww^* = \varphi^{-1} \left[\begin{pmatrix} \frac{1}{2}p_1 & p_2 \\ 0 & 0 \end{pmatrix} \right] \varphi^{-1} \left[\begin{pmatrix} \frac{1}{2}p_1 & 0 \\ p_2 & 0 \end{pmatrix} \right] \\ &= \varphi^{-1} \left[\begin{pmatrix} \frac{1}{2}p_1 & p_2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{2}p_1 & 0 \\ p_2^* & 0 \end{pmatrix} \right] = \varphi^{-1} \left[\begin{pmatrix} \frac{1}{4}p_1^2 + p_2p_2^* & 0 \\ 0 & 0 \end{pmatrix} \right] \in \mathcal{F}(\mathcal{C}). \end{aligned} \quad (4.4.16)$$

Finally, as $B(w, w) \in \mathcal{F}(\mathcal{C})$ and $u \in \text{rel int}(\mathcal{F}(\mathcal{C}))$ we have an $\alpha \geq 0$ such that $\alpha u - B(w, w) \in \mathcal{F}(\mathcal{C}) \subseteq \mathcal{K}$, see [9, 10] for the proof of this classical but simple result. \square

Now we can give an explicit expression for $(\mathcal{F}_i)^\perp$:

Theorem 4.4.8. *The space $(\mathcal{F}_i)^\perp$, $(i = 1, \dots, L + 1)$ can be expressed as*

$$(\mathcal{F}_i)^\perp = \{w + w^* : \exists \alpha \geq 0, (u^1 + \dots + u^i, w, \alpha) \in \text{SC}(\mathcal{K}^*, B)\}, \quad (4.4.17)$$

where $\text{SC}(\mathcal{K}^*, B)$ is the Siegel cone built on \mathcal{K}^* with the bilinear form B .

Proof. Theorem 4.2.3 about the facial reduction algorithm establishes that

$$(\mathcal{F}_i)^\perp = ((\mathcal{F}(u^1 + \dots + u^i))^c)^\perp, \quad (4.4.18)$$

where $u^1 + \dots + u^i \in \mathcal{K}^*$, and the complementary face is taken in \mathcal{K} . We can now apply the previous proposition with $\mathcal{C} = u^1 + \dots + u^i$ and \mathcal{K}^* instead of \mathcal{K} to get

$$(\mathcal{F}_i)^\perp = \{w + w^* : \exists \alpha \geq 0, \alpha (u^1 + \dots + u^i) - B(w, w) \in \mathcal{K}^*\}. \quad (4.4.19)$$

Lemma 4.4.6 enables us to rewrite the condition $\alpha(u^1 + \dots + u^i) - B(w, w) \in \mathcal{K}^*$ using the Siegel cone, this yields the final form for $(\mathcal{F}_i)^\perp$:

$$(\mathcal{F}_i)^\perp = \{w + w^* : \exists \alpha \geq 0, (u^1 + \dots + u^i, w, \alpha) \in \text{SC}(\mathcal{K}^*, B)\}. \quad (4.4.20)$$

This formulation suits our needs as it is expressed with homogeneous cones using dual information only. \square

Corollary 4.4.9. *Now we can apply this theorem along with the facial reduction algorithm described in §4.2 to obtain an explicit strong dual for (P). This provides the following extended dual problem:*

$$\begin{aligned} & \min \langle c, (x^{L+1} + z^{L+1}) \rangle \\ & A(x^{L+1} + z^{L+1}) = b \\ & A(x^i + z^i) = 0, \quad i = 1, \dots, L \\ & \langle c, (x^i + z^i) \rangle = 0, \quad i = 1, \dots, L \\ & x^0, z^1 = 0 \\ & (x^1 + \dots + x^{i-1}, z^i, \alpha^i) \in \text{SC}(\mathcal{K}^*, B), \quad i = 2, \dots, L+1 \\ & x^i \in \mathcal{K}^*, \quad i = 1, \dots, L+1, \end{aligned} \quad (\text{ED})$$

where L is the lesser of $\dim(\text{Ker}(A) \cap \{c\}^\perp)$ and the length of the longest chain of faces in \mathcal{K}^* . We used the fact that $\langle g, w + w^* \rangle = 2\langle g, w \rangle$ if $g = g^*$ to simplify the expressions.

The dual problem is a homogeneous cone optimization problem as the dual cone $\mathcal{K}^* = \mathcal{K}$ and its Siegel cone $\text{SC}(\mathcal{K}^*, B)$ are both homogeneous cones.

Although this extended dual problem satisfies all the requirements we set up at the beginning it can still be improved using a reduction technique due to Pataki [96]. The idea is to disaggregate the LHS of the Siegel cone constraint. This provides the following dual problem:

$$\begin{aligned} & \min \langle c, (x^{L+1} + z^{L+1}) \rangle \\ & A(x^{L+1} + z^{L+1}) = b \\ & A(x^i + z^i) = 0, \quad i = 1, \dots, L \\ & \langle c, (x^i + z^i) \rangle = 0, \quad i = 1, \dots, L \\ & x^0, z^1 = 0 \\ & (x^{i-1}, z^i, 1) \in \text{SC}(\mathcal{K}^*, B), \quad i = 2, \dots, L+1 \\ & x^{L+1} \in \mathcal{K}^*. \end{aligned} \quad (\text{ED}_{\text{disagg}})$$

4.4.2 Complexity of the dual problem

Let us examine system (ED) from an algorithmic point of view.

Optimization problems over convex cones are generally solved with interior point methods. From the viewpoint of this thesis the internal workings of these algorithms are not so important, the interested reader is directed to the literature, most importantly [89, 114]. These methods solve problems (P) and (D) to precision ε in at most $\mathcal{O}\left(\sqrt{\vartheta} \log(1/\varepsilon)\right)$ iterations, where ϑ is a complexity parameter depending only on the cone \mathcal{K} . It is important to note that the iteration complexity does not depend on the dimension of \mathcal{K} or the number of linear equalities, but of course the cost of one iteration is determined by these quantities. There are several estimates for ϑ depending on the geometric and algebraic properties of the cone, for the Lorentz cone in any dimension $\vartheta = 2$, for the cone of $n \times n$ positive semidefinite matrices it is n . For homogeneous cones the two most important results are (see [53, 57]):

1. $\vartheta(\mathcal{K}) = \vartheta(\mathcal{K}^*)$, and
2. $\vartheta(\mathcal{K}) = \text{rank}(\mathcal{K})$.

Moreover, we know that a Siegel extension increases the rank of the cone and thus the complexity by 1. We can also give an explicit barrier function for the Siegel cones, so we can solve the extended dual problems with interior point methods.

The complexity of the dual problem can be computed easily from the rank of the cones. Let $r = \vartheta(\mathcal{K}) = \text{rank}(\mathcal{K})$ be the complexity parameter (and also the rank) of \mathcal{K} , then the complexity of the dual problem is $r + L(r + 1)$. This seems to contradict the fact that $\text{rank}(\mathcal{K}) \geq \text{rank}(\mathcal{K}_{\min}) = \text{rank}(\mathcal{K}_{\min}^*)$, thus we might expect to have a lower complexity. The contradiction is easily removed by noticing that the dual problem (ED) uses only one possible representation of \mathcal{K}_{\min} , with several other cones linked to each other through linear and conic constraints. In this construction \mathcal{K}_{\min} is represented as an intersection of L cones therefore its dual will be a sum of L cones. Better representations may also exist. The following is an interesting open question: what is the best possible complexity for an explicitly formed, easily computable strong dual problem to (P)?

These complexity results are constructive in the sense that the appropriate barrier function of $\text{SC}(\mathcal{K}, B)$ can be constructed from the barrier function of \mathcal{K} . The only practical difficulty can be the actual computability of these functions, but we can always use an approximation [116] to drive the algorithm.

4.5 Special cases

In this section we specialize the new dual for semidefinite and second-order conic optimization.

4.5.1 Semidefinite optimization

For the case when $\mathcal{K} = \mathbb{S}_+^{n \times n}$ is the set of $n \times n$ positive semidefinite matrices, Ramana's original paper [110] proposes the following extended dual problem:

$$\begin{aligned}
 & \min \operatorname{Tr} (c^T (x^{L+1} + z^{L+1})) \\
 & A (x^{L+1} + z^{L+1}) = b \quad (\text{ED}_{\text{Ramana}}) \\
 & \operatorname{Tr} (c^T (x^i + z^i)) = 0, \quad i = 1, \dots, L \\
 & A (x^i + z^i) = 0, \quad i = 1, \dots, L \\
 & x^0, z^1 = 0 \\
 & \begin{pmatrix} x^{i-1} & z^{iT} \\ z^i & I \end{pmatrix} \in \mathbb{S}_+^{2n \times 2n}, \quad i = 2, \dots, L + 1 \\
 & x^{L+1} \in \mathbb{S}_+^{n \times n},
 \end{aligned}$$

where c and all the variables are $n \times n$ matrices, $b \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times (n \times n)}$.

Our dual construction specialized for semidefinite optimization is only slightly different:

$$\begin{aligned}
 & \min \operatorname{Tr} (c^T (x^{L+1} + z^{L+1})) \\
 & A (x^{L+1} + z^{L+1}) = b \quad (\text{ED}_{\text{HomPSD}}) \\
 & \operatorname{Tr} (c^T (x^i + z^i)) = 0, \quad i = 1, \dots, L \\
 & A (x^i + z^i) = 0, \quad i = 1, \dots, L \\
 & x^0, z^1 = 0 \\
 & (x^{i-1}, z^i, 1) \in \text{SC} (\mathbb{S}_+^{n \times n}, B), \quad i = 2, \dots, L + 1 \\
 & x^{L+1} \in \mathbb{S}_+^{n \times n},
 \end{aligned}$$

with the bilinear form $B(u, v) = (uv^T + vu^T)/2$. The difference between the two duals is that the Ramana dual uses the Schur complement and thus the dual is a semidefinite problem, while our dual uses the Siegel cone, and arrives at a general homogeneous conic problem. The Schur complement gives the semidefinite representation of the resulting Siegel cone.

This shows that we actually derived a new strong dual for the semidefinite optimization problem, which is not a semidefinite problem. The benefit of our dual is its improved complexity.

The complexity parameter (see §4.4.2) of $(\text{ED}_{\text{Ramana}})$ is $2nL + n$, since the Schur complement results in $2n \times 2n$ matrices. The complexity parameter of $(\text{ED}_{\text{HomPSD}})$ is only $(n + 1)L + n$, as the Siegel cone construction increases the complexity only by 1. The difference is explained by the fact that our dual is not a semidefinite problem.

The difference in complexity answers another question. Since homogeneous cones are slices of an appropriate dimensional positive semidefinite cone (see Proposition 4.3.12, we can rewrite the original problem (P) as a semidefinite problem, apply the classical Ramana dual and then transform the problem back to homogeneous cones. But as our dual for the semidefinite problem has better complexity than the Ramana dual, we cannot hope to recover our dual from this process.

4.5.2 Second order conic optimization

Originally, our research was sparked by the idea to construct a purely second order conic strong dual to the second order conic optimization problem. Our continued failure to construct such a problem led us to believe that the strong dual cannot be a second order conic optimization problem. After developing the theory for symmetric cones we can partially answer this question.

For notational simplicity let us consider the problem with only one cone:¹⁶

$$\begin{aligned} \max \quad & b^T y \\ & A^T y + s = c \\ & s \in \mathbb{L}, \end{aligned} \tag{P}_{\text{Lor}}$$

where $\mathbb{L} = \{u \in \mathbb{R}^n : u_1 \geq \|u_{2:n}\|_2\}$. The extended dual for problem (P_{Lor}) is ($L \leq 2$, the length of the longest chain of faces in \mathbb{L}):

$$\begin{aligned} \min \quad & \langle c, (x^3 + z^3) \rangle \\ & A(x^3 + z^3) = b \\ & A(x^i + z^i) = 0, \quad i = 1, 2 \\ & \langle c, (x^i + z^i) \rangle = 0, \quad i = 1, 2 \\ & x^0, z^1 = 0 \\ & (x^{i-1}, z^i, 1) \in \text{SC}(\mathbb{L}, B), \quad i = 2, 3 \\ & x^3 \in \mathbb{L}. \end{aligned} \tag{ED}_{\text{Lor}}$$

¹⁶There is a closed form solution to this problem due to Alizadeh and Goldfarb [3], but even this simplest form is enough to illustrate our point.

The rank of the Siegel cone built over the Lorentz cone is 3 and thus it cannot be a Lorentz cone itself, see the classification of lower dimensional homogeneous cones in [71]. It is still an open question whether a purely Lorentz cone strong dual exists to (P_{Lor}) . For duals based on the facial reduction algorithm the answer is negative.

Actually, for second order cones the bilinear form simplifies to $B(u, v) = (u^T v, u_1 v_{2:n} + v_1 u_{2:n})$, i.e., the usual product in the Jordan algebra of the Lorentz cone. This product is commutative, which simplifies the proofs. The Siegel cone corresponding to this form will be

$$\begin{aligned} \text{SC}(\mathbb{L}, B) &= \{(x, z, t) \in \mathbb{R}^{n-1} \times \mathbb{R}^{n-1} \times \mathbb{R}_+ : tx - B(z, z) \in \mathbb{L}\} \\ &= \{(x, z, t) \in \mathbb{R}^{n-1} \times \mathbb{R}^{n-1} \times \mathbb{R}_+ : tx_1 - \|z\|^2 \geq \|tx_{2:n} - 2z_1 z_{2:n}\|\}. \end{aligned} \quad (4.5.1)$$

One might wonder how to express the Siegel cone $\text{SC}(\mathbb{L}, B)$ with a semidefinite constraint. An n dimensional Lorentz cone can be represented by $n \times n$ semidefinite matrices, then the direct application of the Schur complement for the Siegel extension yields a representation of $\text{SC}(\mathbb{L}, B)$ with $2n \times 2n$ matrices. Moreover, Proposition 4.3.12 guarantees the existence of a representation with $(2n - 1) \times (2n - 1)$ matrices. Both are much worse than the actual complexity of the cone, which is 3. It is unknown if there is a better construction, the theoretical lower bound (by a simple dimension argument) is in the order of $2\sqrt{n}$.

4.6 Summary

The central result of this chapter was to provide an explicit strong dual for optimization problems over symmetric cones without assuming any constraint qualification. The construction is based on the facial reduction algorithm and it uses the machinery of homogeneous cones quite heavily. Using the dual we showed a new strong dual for semidefinite problems, which has better complexity than the classical Ramana dual. We have also specialized the dual for second order cone optimization and found that the dual problem cannot be expressed with second order cones, this answers an open question.

Chapter 5

Nonexact duality for conic optimization

Sometimes it is useful to know how large your zero is.

UNKNOWN AUTHOR

Detecting infeasibility in conic optimization and providing certificates pose a bigger challenge than in the linear case due to the lack of strong duality in general. For example, weakly infeasible conic problems can be treated as either feasible or infeasible problems from a numerical point of view. In this chapter we generalize the approximate Farkas lemma of Todd and Ye [123] from the linear to the general conic setting, and use it to propose stopping criteria for interior point algorithms using self-dual embedding. We will prove that – no matter what the duality properties of the problem are – one of the stopping criteria is satisfied after $\mathcal{O}\left(\sqrt{\vartheta} \log(1/\varepsilon)\right)$ iterations, and after that we either have an ε -optimal ε -feasible solution, or we can guarantee that every feasible (or optimal) solution has a norm larger than $1/\varepsilon$. Some practical issues are also discussed. This chapter is based on [104].

5.1 Introduction

Recall the conic feasibility problem in primal and dual form:

$$\begin{aligned} Ax &= b \\ x &\succeq_{\kappa} 0 \end{aligned} \tag{ConFeas_P}$$

and

$$\begin{aligned} A^T y &\preceq_{\mathcal{K}^*} 0 \\ b^T y &= 1. \end{aligned} \tag{ConFeas_D}$$

The Farkas theorem gives us a tool to provide a certificate for infeasibility: a solution for (ConFeas_P) certifies that (ConFeas_D) is not solvable. The difficulty is that in practice we never get an exact solution for these systems. Whence a more relevant question is: what can we deduce, if anything, from an almost certificate?

5.2 An approximate Farkas theorem

Let us consider the primal-dual conic optimization problem in standard form,

$$\begin{aligned} \inf c^T x & & (P) \\ Ax &= b \\ x &\succeq_{\mathcal{K}} 0, \end{aligned}$$

and

$$\begin{aligned} \sup b^T y & & (D) \\ A^T y + s &= c \\ s &\succeq_{\mathcal{K}^*} 0, \end{aligned}$$

where $x, s, c \in \mathbb{R}^n$, $y, b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$, $\mathcal{K} \subseteq \mathbb{R}^n$ is a closed, pointed, solid cone and $\mathcal{K}^* \subseteq \mathbb{R}^n$ is the dual cone associated to \mathcal{K} . Now define some useful quantities.

Definition 5.2.1 (α_x and β_y). *Let $\|\cdot\|$ be any norm on \mathbb{R}^n , and let $\|\cdot\|^*$ be its dual norm, i.e., $\|v\|^* = \max \{v^T y : \|y\| = 1\}$. Now define*

$$\begin{aligned} \alpha_x &= \inf \|x\| & (P_{\alpha_x}) \\ Ax &= b \\ x &\succeq_{\mathcal{K}} 0, \end{aligned}$$

and

$$\begin{aligned} \beta_u &= \inf \|u\|^* & (D_{\beta_u}) \\ A^T y &\preceq_{\mathcal{K}^*} u \\ b^T y &= 1. \end{aligned}$$

In other words, α_x is the norm of the smallest primal solution and β_y is the norm of the smallest dual improving direction.

The results of this section are based on the following theorem:

Theorem 5.2.2 (Approximate Farkas theorem for conic problems).

Assuming the convention $0 \cdot \infty = 1$ we have $\alpha_x \beta_u = 1$.

Proof. This proof is based on an idea of Jos Sturm [121].

If $\alpha_x = 0$ then necessarily $b = 0$, therefore (D_{β_u}) is infeasible and $\beta_u = +\infty$.

If $\alpha_x = +\infty$ then there is no $x \succeq_{\mathcal{K}} 0$ such that $Ax = b$ (i.e., (P_{α_x}) is infeasible) and in this case we can apply the Farkas theorem (Thm. 2.2.3) to deduce that (D_{β_u}) is almost feasible, i.e., for every $\varepsilon > 0$ there is y_ε and u_ε such that $A^T y_\varepsilon \preceq_{\mathcal{K}^*} u_\varepsilon$, $b^T y_\varepsilon = 1$ and $\|u_\varepsilon\|^* < \varepsilon$, hence $\beta_u = 0$.

Now we can assume that $0 < \alpha_x < +\infty$. Let us introduce the so-called norm-cone:

$$\mathcal{K}_{\text{norm}} := \{(x_0, x) \in \mathbb{R}^{n+1} : x_0 \geq \|x\|\}. \quad (5.2.1)$$

It is straightforward to check that this set is indeed a closed, convex, pointed, solid cone. Now α_x can be written as

$$\begin{aligned} \alpha_x = \quad & \inf x_0 \\ & Ax = b \\ & x \succeq_{\mathcal{K}} 0 \\ & (x_0, x) \succeq_{\mathcal{K}_{\text{norm}}} 0. \end{aligned} \quad (\text{P}')$$

Let \bar{x} be a feasible solution for (P_{α_x}) . Now $(\|\bar{x}\| + 1, \bar{x})$ provides a strictly feasible solution for (P') and applying the strong duality theorem we get that

$$\begin{aligned} \alpha_x = \quad & \sup b^T y \\ & \|u\|^* \leq 1 \\ & A^T y \preceq_{\mathcal{K}^*} u. \end{aligned} \quad (\text{D}')$$

As both problem (P') and (D') are strictly feasible we know by strong duality that (D') is solvable, let \hat{y}, \hat{u} be an optimal solution pair. We can assume that for this solution $\|u\|^* = 1$, since changing u does not change the objective function. Defining $\bar{y} = \hat{y}/\alpha_x$ and $\bar{u} = \hat{u}\alpha_x$ we get

$$\begin{aligned} \|\bar{u}\|^* &= 1/\alpha_x \\ A^T \bar{y} &\preceq_{\mathcal{K}^*} \bar{u} \\ b^T \bar{y} &= 1, \end{aligned}$$

implying $\beta_u \leq 1/\alpha_x$, thus $\alpha_x\beta_u \leq 1$.

Finally, if either one of (P_{α_x}) or (D_{β_u}) is not feasible, then the corresponding value is $+\infty$, thus $\alpha_x\beta_u \geq 1$. If both are feasible, then let x and u, y be feasible solutions. In this case:

$$1 = b^T y = \underbrace{x^T}_{\in \mathcal{K}} \underbrace{A^T y}_{\preceq_{\mathcal{K}^*} u} \leq x^T u \leq \|x\| \|u\|^*,$$

giving $\alpha_x\beta_u \geq 1$. This completes the proof. \square

First let us try to understand the geometry of these problems, especially weak infeasibility. For this we introduce the following quantity:

$$\begin{aligned} \alpha_x^\varepsilon = \quad & \inf \|x\| & (P_\varepsilon) \\ & Ax = b^\varepsilon \\ & x \succeq_{\mathcal{K}} c^\varepsilon \\ \|b - b^\varepsilon\|^* & \leq \varepsilon \\ \|c^\varepsilon\|^* & \leq \varepsilon \end{aligned}$$

In this notation our previous α_x is α_x^0 , i.e., the norm of the smallest solution for the unperturbed system. It is obvious that if $\varepsilon_1 \leq \varepsilon_2$ then $\alpha_x^{\varepsilon_1} \geq \alpha_x^{\varepsilon_2}$, since the infimum is taken over a smaller set. This means that α_x^ε converges as $\varepsilon \rightarrow 0$, allowing the limit to be infinity. Let the limit be denoted by $\bar{\alpha}_x$. The first interesting question is how it relates to $\alpha_x = \alpha_x^0$:

Proposition 5.2.3 (Properties of α_x^ε).

$\bar{\alpha}_x = \alpha_x^0$, thus the function $\varepsilon \mapsto \alpha_x^\varepsilon$ is continuous at $\varepsilon = 0$.

Proof. First, let us notice that since α_x^ε is increasing in ε , $\alpha_x^0 \geq \alpha_x^\varepsilon$ for all $\varepsilon \geq 0$, and taking the limit as $\varepsilon \rightarrow 0$ we have $\bar{\alpha}_x \leq \alpha_x^0$. So it remains to prove the opposite inequality, i.e., $\bar{\alpha}_x \geq \alpha_x^0$.

In the case when $\bar{\alpha}_x = +\infty$ the inequality is trivial, so we can restrict ourselves to $\bar{\alpha}_x$ being finite. Since α_x^ε is defined as an infimum the following system of (conic) inequalities must be solvable for all positive ε and arbitrary n .

$$\begin{aligned} Ax_n^\varepsilon &= b_n^\varepsilon \\ x_n^\varepsilon &\succeq_{\mathcal{K}} c_n^\varepsilon \\ \|b - b_n^\varepsilon\|^* &\leq \varepsilon \\ \|c_n^\varepsilon\|^* &\leq \varepsilon \\ \|x_n^\varepsilon\| &\leq \alpha_x^\varepsilon + \frac{1}{n} \end{aligned}$$

Let us fix $\varepsilon > 0$. It is clear, that

$$\forall n : \|x_n^\varepsilon\| \leq \alpha_x^\varepsilon + 1, \quad \|b - b_n^\varepsilon\|^* \leq \varepsilon, \quad \|c_n^\varepsilon\|^* \leq \varepsilon.$$

Using the fact that $\bar{\alpha}_x$ is finite we have that α_x^ε is finite for all $\varepsilon > 0$. This means that the solutions $(x_n^\varepsilon, b_n^\varepsilon, c_n^\varepsilon)$ are in a compact set for all n . Thus, there is a convergent subsequence such that

$$x_{n_k}^\varepsilon \rightarrow x^\varepsilon, \quad b_{n_k}^\varepsilon \rightarrow b^\varepsilon, \quad c_{n_k}^\varepsilon \rightarrow c^\varepsilon,$$

as $k \rightarrow \infty$. The limit points satisfy the system

$$\begin{aligned} Ax^\varepsilon &= b^\varepsilon \\ x^\varepsilon &\succeq_{\mathcal{K}} c^\varepsilon \\ \|b - b^\varepsilon\|^* &\leq \varepsilon \\ \|c^\varepsilon\|^* &\leq \varepsilon \\ \|x^\varepsilon\| &\leq \alpha_x^\varepsilon \end{aligned}$$

which means that they form a feasible solution of (P_ε) with the additional property that $\|x^\varepsilon\| \leq \alpha_x^\varepsilon$. These two facts imply that $\|x^\varepsilon\| = \alpha_x^\varepsilon$, i.e., the infimum in (P_ε) is attained, so it is in fact a minimum.

Now we have $\|x^\varepsilon\| = \alpha_x^\varepsilon \leq \bar{\alpha}_x < +\infty$, so there is a sequence $\varepsilon_k \rightarrow 0$ that $x^{\varepsilon_k} \rightarrow x$, $b^{\varepsilon_k} \rightarrow b$, and $c^{\varepsilon_k} \rightarrow 0$ as $k \rightarrow \infty$. Taking the limit of the inequalities we have that

$$\begin{aligned} \|x\| &\leq \bar{\alpha}_x \\ Ax &= b \\ x &\succeq_{\mathcal{K}} 0 \end{aligned}$$

Here the last two equations imply that x is a feasible solution of (P_ε) with $\varepsilon = 0$, so

$$\alpha_x^0 \leq \|x\| \leq \bar{\alpha}_x.$$

This completes the proof. □

From the last part of the proof we can draw a useful corollary:

Corollary 5.2.4. *If $\alpha_x < \infty$ then there exists $x \in \mathbb{R}^n$ for which $\alpha_x = \alpha_x^0 = \|x\|$, in other words, the infimum in (P_{α_x}) is attained.*

In the general conic case α_x alone does not tell everything about the infeasibility of the primal system (P) , the further distinction is made by α_x^ε . If $\alpha_x = \infty$ then (P) is clearly infeasible, but if now α_x^ε is finite for every $\varepsilon > 0$ then (P) is only weakly infeasible. Similarly, if there is an $\varepsilon > 0$ such that $\alpha_x^\varepsilon = \infty$ then the problem is strongly infeasible, meaning that slightly perturbing the problem does not make it feasible.

Similar results can be proved for the dual case.

5.3 Stopping criteria for self-dual models

In this section we derive two sets of stopping criteria for the homogeneous self-dual model for conic optimization.

5.3.1 Homogeneous self-dual model for conic optimization

This model has already been described and analyzed in many works, see [34, 79]. Given the primal-dual pair (P) and (D) consider the following system:

$$\begin{array}{rccccccc}
 & & & & \min & (\bar{x}^T \bar{s} + 1)\theta & & \\
 & Ax & -b\tau & +\bar{b}\theta & & & = & 0 \\
 -A^T y & & +c\tau & -\bar{c}\theta & -s & & = & 0 \\
 b^T y & -c^T x & & +\bar{z}\theta & & -\kappa & = & 0 \\
 -\bar{b}^T y & -\bar{c}^T x & -\bar{z}^T \tau & & & & = & -\bar{x}^T \bar{s} - 1 \\
 x \succeq_{\mathcal{K}} 0 & \tau \geq 0 & & s \succeq_{\mathcal{K}^*} 0 & \kappa \geq 0, & & &
 \end{array} \tag{HSD}$$

where $\bar{x}, \bar{s} \in \mathbb{R}^n$, $\bar{y} \in \mathbb{R}^m$ are arbitrary starting points, τ, θ are scalars, $\bar{b} = b - A\bar{x}$, $\bar{c} = c - A^T \bar{y} - \bar{s}$ and $\bar{z} = c^T \bar{x} - b^T \bar{y} + 1$. This model has the following properties.

Theorem 5.3.1 (Properties of the HSD model). *System (HSD) is self-dual and it has a strictly feasible starting point, namely $(x, s, y, \tau, \theta, \kappa) = (\bar{x}, \bar{s}, \bar{y}, 1, 1, 0)$.¹⁷ The optimal value of these problems is $\theta = 0$, and if $\tau > 0$ at optimality then $(x/\tau, y/\tau, s/\tau)$ is an optimal solution for the original primal-dual problem. If $\tau = 0$, then the problem is either unbounded, infeasible, or the duality gap at optimality is nonzero.*

¹⁷In practice this makes it possible to use a feasible-start interior point method to solve these problems.

Let us solve the homogeneous problem with a path-following interior point method that generates a sequence of iterates $(x^k, s^k, y^k, \tau^k, \theta^k, \kappa^k)$, such that

$$\tau^k \kappa^k \geq (1 - \beta)\theta^k, \quad (5.3.1)$$

where β is a fixed constant. This is a standard assumption for interior point methods, see [89] for details.

5.3.2 New stopping criteria

In practice we have to live with the fact that we never have exact optimal solutions. Similarly, we can rarely guarantee infeasibility, but we can certify that every feasible solution should have a large norm. If the threshold is set high enough then this is sufficient in most cases. The results of these sections are generalizations of [123, §4].

Let ρ be a large enough number (about 10^9 or more in general) and ε a small number (typically 10^{-9}). The first stopping criterion deals with optimality, it is activated if the current iterate provides an ε -optimal and ε -feasible solution:

$$\begin{aligned} \|Ax - b\tau\|^* &\leq \varepsilon\tau & (R1) \\ \|A^T y + s - c\tau\|^* &\leq \varepsilon\tau \\ c^T x - b^T y &\leq \varepsilon\tau. \end{aligned}$$

As we mentioned earlier, if $\tau = 0$ is an optimal solution for the self-dual model then the original problems do not have an optimal solution with zero duality gap. A small value of τ gives us slightly less, that is our second criterion:

$$\tau \leq \frac{1 - \beta}{1 + \rho}. \quad (R2)$$

We have the following result:

Lemma 5.3.2 (Identification of large optimal solutions).

If stopping rule (R2) is satisfied then for every optimal solution x^ , (y^*, s^*) of the original primal-dual problem we have $x^{*T}\bar{s} + s^{*T}\bar{x} \geq \rho$, where \bar{x} and \bar{s} are the starting points of the HSD model.*

Proof. Assume that the stopping rule is activated, let x^* , (y^*, s^*) be an optimal solution of the original primal-dual problem, and assume that they have small norm, i.e., $x^{*T}\bar{s} + s^{*T}\bar{x} < \rho$. Let

$$\alpha = \frac{\bar{x}^T \bar{s} + 1}{x^{*T} \bar{s} + s^{*T} \bar{x} + 1} > 0, \quad (5.3.2)$$

and define the following quantities:

$$\begin{aligned}\bar{x} &= \alpha x^* \\ \bar{y} &= \alpha y^* \\ \bar{s} &= \alpha s^* \\ \bar{\tau} &= \alpha \\ \bar{\theta} &= 0 \\ \bar{\kappa} &= 0.\end{aligned}$$

These quantities form an optimal solution for the self-dual system. Now let us subtract the two self dual systems:

$$\begin{aligned}A(x - \bar{x}) & -b(\tau - \bar{\tau}) & +\bar{b}(\theta - \bar{\theta}) & = 0 \\ -A^T(y - \bar{y}) & +c(\tau - \bar{\tau}) & -\bar{c}(\theta - \bar{\theta}) & = s - \bar{s} \\ b^T(y - \bar{y}) & -c^T(x - \bar{x}) & +\bar{z}(\theta - \bar{\theta}) & = \kappa - \bar{\kappa} \\ -\bar{b}^T(y - \bar{y}) & -\bar{c}^T(x - \bar{x}) & -\bar{z}^T(\tau - \bar{\tau}) & = 0.\end{aligned}\tag{5.3.3}$$

Premultiplying this system by $(y - \bar{y}, x - \bar{x}, \tau - \bar{\tau}, \theta - \bar{\theta})$ gives

$$0 = (x - \bar{x})^T(s - \bar{s}) + (\tau - \bar{\tau})(\kappa - \bar{\kappa}),\tag{5.3.4}$$

since the coefficient matrix on the LHS is skew symmetric. Rearranging the terms and using that $x^T s + \tau \kappa = \theta(1 - \bar{x}^T \bar{s})$ we get

$$x^T \bar{s} + s^T \bar{x} + \tau \bar{\kappa} + \kappa \bar{\tau} = (\bar{x}^T s^* + \bar{s}^T x^* + 1) \theta.\tag{5.3.5}$$

Now

$$\begin{aligned}\tau &= \tau \frac{x^T \bar{s} + s^T \bar{x} + \tau \bar{\kappa} + \kappa \bar{\tau}}{(\bar{x}^T s^* + \bar{s}^T x^* + 1) \theta} \geq \\ &\geq \frac{\tau \kappa \bar{\tau}}{(\bar{x}^T s^* + \bar{s}^T x^* + 1) \theta} \geq \frac{(1 - \beta) \bar{\tau}}{\bar{x}^T s^* + \bar{s}^T x^* + 1} > \frac{1 - \beta}{1 + \rho},\end{aligned}\tag{5.3.6}$$

contradicting (R2). This proves the lemma. \square

Stopping rule (R2) guarantees that the optimal solutions have large norm. There might be some moderate-sized feasible solutions, but they are far from optimal. This usually suggests some modelling error or unboundedness.

The third set of criteria identifies large feasible solutions.

$$b^T y \geq (\tau \|c\|^* + \theta \|\bar{c}\|^*) \bar{\rho}\tag{R3a}$$

$$c^T x \leq -(\tau \|b\|^* + \theta \|\bar{b}\|^*) \bar{\rho}\tag{R3b}$$

Lemma 5.3.3 (Identification of large feasible solutions).

If stopping criterion (R3a) is activated then for every feasible solution x of the primal problem we have $\|x\| \geq \bar{\rho}$. Similarly, if stopping criterion (R3b) is activated then for every feasible solution (y, s) of the dual problem we have $\|s\|^* \geq \bar{\rho}$.

Proof. We only prove the first statement as the second one is very similar. Let $\bar{y} = y/b^T y$ and $\bar{u} = (\tau c - \theta \bar{c})/b^T y$, then \bar{y} and \bar{u} are feasible in (D_{β_u}) , thus $\beta_u \leq 1/\bar{\rho}$ and by Theorem 5.2.2 we have $\alpha_x \geq \bar{\rho}$. \square

This technique is used in most conic optimization software packages.

5.3.3 Complexity of the criteria

In this subsection we show that these conditions are indeed practical in the sense that one of them is activated after a polynomial number of iterations. This is trivial if we know in advance that our original (not the self-dual) problem is feasible. Now let us assume that we use a feasible interior point method to solve the self-dual problem. Such algorithms can produce a solution satisfying $\theta < \varepsilon$ (if such a solution exists) in $\mathcal{O}(\sqrt{\vartheta} \log(1/\varepsilon))$ iterations, where ϑ depends only on \mathcal{K} , see, e.g., [89, 114]. The question is what happens if this is not the case, i.e., if either the original problem is infeasible or there are no solutions with duality gap 0. First let us see what we get from rules (R1) and (R2).

Theorem 5.3.4 (Complexity of the IPM with (R1) and (R2)).

Either rule (R1) or (R2) is activated in

$$\mathcal{O}\left(\sqrt{\vartheta} \log\left(\frac{\max\{\|\bar{b}\|^*, \|\bar{c}\|^*, \bar{z}\}(1+\rho)}{\varepsilon}\right)\right) \quad (5.3.7)$$

iterations.

Proof. Using that $Ax - b\tau = -\bar{b}\theta$, $A^T y + s - c\tau = -\bar{c}\theta$ and $c^T x - b^T y = \bar{z}\theta - \kappa$ we get that criterion (R1) is satisfied if

$$\frac{\theta}{\tau} \leq \frac{\varepsilon}{\max\{\|\bar{b}\|^*, \|\bar{c}\|^*, \bar{z}\}}. \quad (5.3.8)$$

Assume that rule (R2) is not activated during the iterations, this means that $\tau > (1 - \beta)/(1 + \rho)$ for every fixed β . This implies that rule (R1) is satisfied if

$$\theta < \frac{(1 - \beta)\varepsilon}{(1 + \rho) \max\{\|\bar{b}\|^*, \|\bar{c}\|^*, \bar{z}\}}, \quad (5.3.9)$$

and using assumption (5.3.1) about the algorithm we get the statement of the theorem. \square

Corollary 5.3.5. *Setting $\rho = 1/\varepsilon$ the complexity of the algorithm with stopping criteria (R1) and (R2) is $\mathcal{O}(\sqrt{\vartheta} \log(1/\varepsilon))$, the same order as the original algorithm running on a feasible problem. In this many iterations either the algorithm finds an ε -optimal solution or it proves that the norm of any optimal solution is larger than $1/\varepsilon$.*

The second theorem deals with rules (R1), (R3a) and (R3b). Let ε and ρ be the parameters for accuracy and feasible solution size. Let us define the following quantities:

$$\bar{\rho} = \max \{ \bar{z}, \rho \max \{ \|c\|^*, \|\bar{c}\|^*, \|b\|^*, \|\bar{b}\|^* \} \} \quad (5.3.10)$$

$$\bar{\varepsilon} = \min \left\{ \frac{2}{3}, \frac{\varepsilon}{\max \{ \|\bar{b}\|^*, \|\bar{c}\|^*, \bar{z} \}} \right\} \quad (5.3.11)$$

Theorem 5.3.6 (Complexity of the IPM with (R1) and (R3)).

Either rule (R1) or (R3) is activated in not more than

$$\mathcal{O} \left(\sqrt{\vartheta} \log \frac{\bar{\rho}}{\bar{\varepsilon}} \right) \quad (5.3.12)$$

iterations.

Proof. We prove that if

$$\theta \leq \frac{(1 - \beta)\bar{\varepsilon}^2}{4\bar{\rho}} \quad (5.3.13)$$

then one of the stopping criteria (R1), (R3a), (R3b) is satisfied. Assume to the contrary that none of the three criteria is satisfied. Since (R1) is not active we get

$$\frac{\theta}{\tau} \geq \bar{\varepsilon} \quad (5.3.14)$$

and combining this with the assumption on the IPM we have

$$\kappa \geq (1 - \beta) \frac{\theta}{\tau} > (1 - \beta) \bar{\varepsilon}. \quad (5.3.15)$$

Using (5.3.13) we can continue the estimate:

$$\begin{aligned} \frac{\kappa}{\theta} &> \frac{4\bar{\rho}}{\bar{\varepsilon}} \geq \left(3 + \frac{2}{\bar{\varepsilon}} \right) \bar{\rho} \geq \bar{z} + 2 \left(1 + \frac{\tau}{\theta} \right) \bar{\rho} \\ &\geq \bar{z} + \frac{\rho}{\theta} \left(\tau (\|c\|^* + \|b\|^*) + \theta (\|\bar{c}\|^* + \|\bar{b}\|^*) \right), \end{aligned}$$

where we used that $\bar{\varepsilon} \leq 2/3$, thus $4/\bar{\varepsilon} \geq 3 + 2/\bar{\varepsilon}$. Since (R3a) and (R3b) are not satisfied we can continue with

$$\frac{\kappa}{\theta} > \bar{z} + \frac{1}{\theta} (b^T y - c^T x) = \bar{z} + \frac{1}{\theta} (\kappa - \bar{z}\theta) = \frac{\kappa}{\theta}, \quad (5.3.16)$$

which is a contradiction. The complexity result follows easily, using assumption (5.3.1) about the algorithm. \square

Corollary 5.3.7. *Again, setting $\rho = 1/\varepsilon$ the complexity of the algorithm with stopping criteria (R1), (R3a) and (R3b) is $\mathcal{O}(\sqrt{\vartheta} \log(1/\varepsilon))$, the same order as the original algorithm running on a feasible problem. In this many iterations either the algorithm finds an ε -optimal and ε -feasible solution or it proves that the norm of any feasible solution is larger than $1/\varepsilon$.*

5.4 Practical considerations

Here we present a couple of topics on how this theory relates to typical practical situations.

5.4.1 How big is big enough?

In all the criteria discussed in this chapter we have a parameter ρ , which is supposed to be big. The role of ρ is the same in all the conditions: if some computed value exceeds ρ , then we declare infeasibility. The natural question is: how big should ρ be? In practice $\rho = 1/\varepsilon$ is a popular choice, with a reasonable justification. Either we find an ε -optimal solution, or we find an ε -certificate for infeasibility. This guarantees a certain primal-dual symmetry.

Another option is to apply a preprocessing phase to obtain bounds on the variables in the problem, and compute ρ accordingly. For very large scale problems with 5000×5000 semidefinite matrices if the elements are on the order of some hundreds then the norm of the matrix is more than 10^9 . This method is yet to be implemented in major conic optimization solvers.

5.4.2 Handling weak infeasibility

Weakly infeasible problems are infeasible but there is no certificate that proves that. From another point of view, the problem is infeasible, but perturbing it slightly we can get either a feasible or an infeasible problem. Since practical solvers usually deal with approximate solutions and certificates, weakly infeasible problems are hard (if not practically impossible) to identify. An almost

solution and an almost certificate for infeasibility can coexist, and there is no easy way to decide if the problem is in fact feasible or infeasible.

This is exactly how this problem manifests in practice: in the set of stopping criteria more conditions are satisfied simultaneously. Of course, this does not mean weak infeasibility automatically, it only indicates that the problem is close to the borderline between feasible and infeasible problems. We do not have a way to decide which side the problem lies unless we carry out the operations more accurately.

The conclusions we get from the criteria are also twofold. We end up with an approximate solution (usually of large norm) and we also deduce that all feasible or optimal solutions must have large norm.

5.5 Summary

In this chapter we presented two new stopping criteria for IPMs for general conic optimization. They can detect if the norm of feasible or optimal solutions is large, thus give an indication of infeasibility. We fully analyzed the complexity of the resulting algorithms and showed that the new criteria do not change the order of the number of iterations.

Chapter 6

Conclusions and further directions

The science of today is the
technology of tomorrow.

EDWARD TELLER

Science never solves a problem
without creating ten more.

GEORGE BERNARD SHAW

In this chapter we summarize my results and present some open problems for future research.

6.1 Contributions

The most important results of this thesis are summarized below.

6.1.1 S-lemma

We surveyed the literature of the S-lemma extensively. We presented three frameworks (convexity of quadratic images, rank constrained optimization, generalized convexities) to analyze systems of quadratic (in)equalities and showed the various interplays between them. We also showed how similar or identical results had been discovered independently within these frameworks. We answered some open questions about quadratic systems.

Based on an auxiliary lemma, we presented a new, elementary proof for the S-lemma. The technique of the proof extends to more general functions.

6.1.2 Nonconvex duality

We proved several new nonlinear, nonconvex duality theorems based on various convexity results. What makes these theorems remarkable is that while the primal problem is nonconvex, thus it is hard to solve, the dual problem is convex, whence the solvability of the original system can be decided efficiently.

6.1.3 Nonregular duality

We developed a complete strong duality theory for optimization over symmetric cones. The dual problem can be formulated easily from the problem data, and is defined over a homogeneous cone. We analyzed the complexity of the dual problem and tightened the complexity bound for the strong dual of a semidefinite optimization problem. We answered an open questions about the second-order conic case in the negative.

6.1.4 Approximate duality

We provided a new, elementary proof for the approximate duality theorems for conic optimization and we used the results to derive new stopping criteria for interior point methods. We proved that the new criteria do not change the complexity of the interior point method. The criteria can be used to identify that the optimal solutions are large, this usually suggests an error in the model.

6.2 Problems for future research

To finish this thesis let me present some problems and projects worthy of further research.

6.2.1 Nonconvex systems

SOCO relaxations

Remaining in the quadratic world, second order conic (SOCO) relaxations offer an alternative way to treat quadratic systems. This procedure is similar to semidefinite relaxation, which was discussed in §3.2.3 and §3.6. Currently,

there are not many results on when an SOCO relaxation of a quadratic system is exact.

Real numerical range

A characterization theorem for the convexity of the joint numerical range over the real numbers (similar to Theorem 3.5.14) could be developed.

Polynomial systems

The question arises of whether it is possible to extend the concept of the S-lemma to polynomial or even more general functions. Since any polynomial system can be written equivalently as a quadratic system with more equations and variables, this case essentially reduces to the quadratic case, but finding these results and providing suitable applications remains a challenging task.

Convex images

The convexity of the image of a set under some transformation is crucial to this theory. However, there are hardly any results about the convexity of the image of sets under a nonlinear transformation, Polyak's local convexity result (Theorem 3.5.10) is one of them. In [109] Ramana and Goldman posed the same question for general maps, but then only dealt with the quadratic case. This is an open and mainly untouched research field. Generalized convexities (see §3.7) offer another framework for this research.

Algebraic methods

Leaving the concept of the S-lemma and concentrating on more general related results we find that polynomial and sum of squares (SOS) optimization (see [108]) have recently drawn a lot of interest in both the optimization and control communities. This research directed the attention of optimizers towards algebraic geometric results, e.g., the Nullstellensatz (see §3.8.3). Since polynomials are well studied in advanced algebra, more research should be carried out in this direction. On the other hand, the example of the S-lemma shows that under some conditions stronger degree bounds are possible in the effective Nullstellensatz.

Applications

As of today, there are few practical applications of the S-lemma with more than two inequalities. This might be due to the fact that nonconvex quadratic systems are usually believed to be difficult and people often use alternative formulations instead. However, the fact that the S-lemma allows for polynomial time solvability (see §3.8.4) of certain quadratic systems should be promoted. Similarly, one can look for applications of Theorem 3.5.19 or Theorem 3.9.3.

6.2.2 Nonregular systems

Better complexity

The best possible complexity of an easily constructible strong dual to (P) is still an open question. Moreover, the extended dual is deficient in the sense that it only guarantees zero duality gap and dual (but not primal) solvability. This follows from the properties of the construction in [25]. Another idea, stemming from [111] is to investigate the Lagrange dual of the extended dual (called the corrected primal).

More general cones

A natural question is whether these results extend to more general cones. We proved that homogeneous cones are nice, thus the facial reduction algorithm can be applied. To obtain a description of $(\mathcal{F}(\mathcal{C})^c)^\perp$ all we need is a bilinear function B that satisfies the requirements of Definition 4.3.8 and works in the proof of Theorem 4.4.7. The existence and possible construction of such a function is an open question.¹⁸

Generalizations to more general cones are probably not possible. Of course, if a cone is a slice of a higher dimensional semidefinite cone then the problem can be rewritten as an SDP and the strong dual can be applied. However, the complexity of the dual problem can be much worse than the complexity of the original problem, see §4.5.2 about the complexity of the strong dual of an SOCP. The bottleneck seems to be some Schur complement-like result, which can convert a quadratic conic constraint into a linear one of good complexity. For homogeneous cones the Siegel cone construction of §4.3.3 gives the answer, but for more general cones there are no such results. Homogeneous cones possess some remarkable properties, which are essential for this theory.

¹⁸We strongly believe such a function exists.

Modelling and software

Homogeneous cones also provide an efficient modelling tool for optimization problems currently modelled directly as semidefinite problems. Siegel cones give rise to new modelling formulations, especially since the Schur complement is a fundamental tool wherever semidefinite optimization is applied, see [16, 27, 138] for some examples. Using Siegel cones instead of the Schur complement would greatly improve the complexity of the models. This should serve as an incentive to develop software for homogeneous conic optimization problems. Another approach to solve these problems is to make use of specialized linear algebra to solve system (ED_{Lor}) effectively.

6.2.3 Approximate duality

Most of the future questions in this area are about the implementation of the stopping criteria.

Large optimal solutions

The set of stopping criteria that can certify that all the optimal solutions of a problem are large has not been implemented in any conic optimization solver yet. By notifying the user about the problem early on, this method would save computing time.

Weak infeasibility

One particular challenge is weakly infeasible problems. The only clue we get about this situation is the coexistence of an almost optimal solution and an almost exact certificate for infeasibility. In inexact arithmetic we cannot really detect if the problem is really weakly infeasible or it is just close to the boundary between feasibility and infeasibility. We cannot even decide which side the problem is on. Currently, as far as we know none of the conic optimization solvers indicate to the user that the feasibility of the problem is in question, they simply return an approximate solution or declare feasibility depending on their stopping criteria.

Usually, if a problem is close to being infeasible that indicates that the model is not perfect. Either the data is wrong or the problem should be stated differently.

A library of infeasible problems

Problem libraries for linear, quadratic and general nonlinear optimization usually contain a good selection of *practical* infeasible problems. Unfortunately, the infeasible problems in the SDPLIB library (see [24]) or among the DIMACS Challenge problems [97] are rather small and trivial. A comprehensive library of infeasible problems would make testing and development of new stopping criteria easier.

Bibliography

- [1] M. A. AIZERMAN AND F. R. GANTMACHER, *Absolute Stability of Regulator Systems*, Holden-Day Series in Information Systems, Holden-Day, Inc, San Francisco, 1964. Originally published as *Absolutnaya Ustoyichivost' Reguliruyemykh Sistem*, by The Academy of Sciences of the USSR, Moscow, 1963.
- [2] A. ALEMAN, *On some generalizations of convex sets and convex functions*, *L'Analyse Numérique et la Théorie de l'Approximation*, 14 (1985), pp. 1–6.
- [3] F. ALIZADEH AND D. GOLDFARB, *Second-order cone programming*, *Mathematical Programming*, 95 (2003), pp. 3–51.
- [4] M. ANITESCU, *Degenerate nonlinear programming with a quadratic growth condition*, *SIAM Journal on Optimization*, 10 (2000), pp. 1116–1135.
- [5] —, *A superlinearly convergent sequential quadratically constrained quadratic programming algorithm for degenerate nonlinear programming*, *SIAM Journal on Optimization*, 12 (2002), pp. 949–978.
- [6] Y.-H. AU-YEUNG, *A theorem on a mapping from a sphere to the circle and the simultaneous diagonalization of two Hermitian matrices*, *Proceedings of the AMS*, 20 (1969), pp. 545–548.
- [7] Y.-H. AU-YEUNG AND Y. T. POON, *A remark on the convexity and positive definiteness concerning Hermitian matrices*, *Southeast Asian Bulletin of Mathematics*, 3 (1979), pp. 85–92.
- [8] Y.-H. AU-YEUNG AND N.-K. TSING, *An extension of the Hausdorff-Toeplitz theorem on the numerical range*, *Proceedings of the AMS*, 89 (1983), pp. 215–218.

-
- [9] G. P. BARKER, *The lattice of faces of a finite dimensional cone*, Linear Algebra and Its Applications, 7 (1973), pp. 71–82.
- [10] G. P. BARKER AND D. CARLSON, *Cones of diagonally dominant matrices*, Pacific Journal of Mathematics, 57 (1975), pp. 15–32.
- [11] A. I. BARVINOK, *Feasibility testing for systems of real quadratic equations*, Discrete & Computational Geometry, 10 (1993), pp. 1–13.
- [12] ———, *A remark on the rank of positive semidefinite matrices subject to affine constraints*, Discrete & Computational Geometry, 25 (2001), pp. 23–31.
- [13] S. BASU, R. POLLACK, AND M.-F. ROY, *Algorithms in Real Algebraic Geometry*, Springer-Verlag, 2003.
- [14] A. BECK, *Quadratic matrix programming*, SIAM Journal on Optimization, 17 (2006), pp. 1224–1238.
- [15] A. BECK AND M. TEBoulLE, *Global optimality conditions for quadratic optimization problems with binary constraints*, SIAM Journal on Optimization, 11 (2000), pp. 179–188.
- [16] A. BEN-TAL AND A. NEMIROVSKI, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPS-SIAM Series on Optimization, SIAM, Philadelphia, PA, 2001.
- [17] A. BEN-TAL AND M. TEBoulLE, *Hidden convexity in some nonconvex quadratically constrained quadratic programming*, Math. Programming, 72 (1996), pp. 51–63.
- [18] C. BERENSTEIN AND D. STRUPPA, *Recent improvements in the complexity of the effective Nullstellensatz*, Linear Algebra and Its Applications, 157 (1991), pp. 203–215.
- [19] A. BERMAN, *Cones, Matrices and Mathematical Programming*, no. 79 in Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1973.
- [20] D. P. BERTSEKAS, *Convexity, duality, and Lagrange multipliers*. Lecture Notes, MIT, Spring 2001.

- [21] D. P. BERTSEKAS, A. NEDIĆ, AND A. E. OZDAGLAR, *Convex Analysis and Optimization*, no. 1 in Athena Scientific Optimization and Computation Series, Athena Scientific, Belmont, Massachusetts, 2003.
- [22] J. BOCHNAK, M. COSTE, AND M.-F. ROY, *Real Algebraic Geometry*, Springer-Verlag, Berlin, 1998.
- [23] T. BONNESEN AND W. FENCHEL, *Theorie der konvexen Körper*, Chelsea Publishing Company, New York, 1948.
- [24] B. BORCHERS, *SDPLIB 1.2, a library of semidefinite programming test problems*, Optimization Methods and Software, 11 (1999), pp. 683–690.
- [25] J. BORWEIN AND H. WOLKOWICZ, *Regularizing the abstract convex program*, Journal of Mathematical Analysis and Applications, 83 (1981), pp. 495–530.
- [26] J. M. BORWEIN AND A. S. LEWIS, *Convex Analysis and Nonlinear Optimization: Theory and Examples*, no. 3 in Canadian Mathematical Society Books in Mathematics, Springer, New York, 2000.
- [27] S. BOYD, L. EL GHAOU, E. FERON, AND V. BALAKRISHNAN, *Linear Matrix Inequalities in System and Control Theory*, vol. 15 of SIAM Studies in Applied Mathematics, SIAM, Philadelphia, PA, 1994.
- [28] W. W. BRECKNER AND G. KASSAY, *A systemization of convexity concepts for sets and functions*, Journal of Convex Analysis, 4 (1997), pp. 109–127.
- [29] L. BRICKMAN, *On the field of values of a matrix*, Proceedings of the AMS, 12 (1961), pp. 61–66.
- [30] A. BRØNDSTED, *An Introduction to Convex Polytopes*, no. 90 in Graduate texts in mathematics, Springer-Verlag, New York, 1983.
- [31] A. BUNSE-GERSTNER, R. BYERS, AND V. MEHRMANN, *Numerical methods for simultaneous diagonalization*, SIAM Journal of Matrix Analysis and Applications, 14 (1993), pp. 927–949.
- [32] C. B. CHUA, *Relating homogeneous cones and positive cones via T -algebras*, SIAM Journal on Optimization, 14 (2003), pp. 500–506.

- [33] A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *Trust-Region Methods*, MPS-SIAM Series on Optimization, SIAM, Philadelphia, PA, 2000.
- [34] E. DE KLERK, C. ROOS, AND T. TERLAKY, *Infeasible-start semidefinite programming algorithms via self-dual embeddings*, in Topics in Semidefinite and Interior Point Methods, P. Pardalos and H. Wolkowicz, eds., vol. 18 of Fields Institute Communications, AMS, Providence, RI, 1998, pp. 215–236.
- [35] —, *Nonlinear Optimization*. Lecture notes, Delft University of Technology, Delft, The Netherlands, 2003.
- [36] K. DERINKUYU, M. Ç. PINAR, AND A. CAMCI, *An improved probability bound for the approximate S-lemma*, Oper. Res. Lett., (2007). To appear.
- [37] K. DERINKUYU AND M. Ç. PINAR, *On the S-procedure and some variants*, Mathematical Methods of OR, 64 (2006), pp. 55–77.
- [38] L. L. DINES, *On the mapping of quadratic forms*, Bulletin of the AMS, 47 (1941), pp. 494–498.
- [39] —, *On linear combinations of quadratic forms*, Bulletin of the AMS, 49 (1943), pp. 388–393.
- [40] L. E. DUBINS, *On extreme points of convex sets*, J. Math. Anal. Appl., 5 (1962), pp. 237–244.
- [41] K. FAN, *Minimax theorems*, Proceedings of the National Academy of Sciences, (1953), pp. 42–47.
- [42] J. FARAUT AND A. KORÁNYI, *Analysis on Symmetric Cones*, Oxford Mathematical Monographs, Oxford University Press, 1994.
- [43] GY. FARKAS, *Theorie der einfachten Ungleichungen*, J. Reine Angew. Math, 124 (1902), pp. 1–27.
- [44] L. FAYBUSOVICH, *On Nesterov’s approach to semi-infinite programming*, Acta Applicandae Mathematicae, 74 (2002), pp. 195–215.
- [45] —, *Jordan-algebraic approach to convexity theorems for quadratic mappings*, preprint, Department of Mathematics, University of Notre Dame, Notre Dame, IN, 2005.

- [46] W. FENCHEL, *On conjugate convex functions*, Canadian Journal of Mathematics, 1 (1949), pp. 73–77.
- [47] —, *Convex cones, sets, and functions*. Mimeographed notes, Princeton University, 1951.
- [48] P. FINSLER, *Über das Vorkommen definitiver und semidefiniter Formen in Scharen quadratischer Formen*, Commentaria Mathematicae Helvetiae, 9 (1937), pp. 188–192.
- [49] A. L. FRADKOV AND V. A. YAKUBOVICH, *The S-procedure and a duality relation in nonconvex problems of quadratic programming*, Vestnik Leningrad University, 5 (1979), pp. 101–109. Originally in Russian in 1973.
- [50] T. FUJIE AND M. KOJIMA, *Semidefinite programming relaxation for nonconvex quadratic programs*, Journal of Global Optimization, 10 (1997), pp. 367–380.
- [51] S. D. GARVEY, F. TISSEUR, M. I. FRISWELL, J. E. T. PENNY, AND U. PRELLS, *Simultaneous tridiagonalization of two symmetric matrices*, International Journal for Numerical Methods in Engineering, 57 (2003), pp. 1643–1660.
- [52] S. I. GASS AND A. A. ASSAD, *An Annotated Timeline of Operations Research: An Informal History*, vol. 75 of International Series in Operations Research & Management Science, Springer, 2004.
- [53] S. G. GINDIKIN, *Tube Domains and the Cauchy Problem*, vol. 111 of Translation of Mathematical Monographs, AMS, Providence, RI, 1992.
- [54] D. GOLDFARB AND G. IYENGAR, *Robust portfolio selection problems*, Mathematics of OR, 28 (2003), pp. 1–38.
- [55] P. GRIFFITHS AND J. HARRIS, *Principles of Algebraic Geometry*, Wiley Classics Library, John Wiley & Sons, 1994.
- [56] D. GRIGORIEV AND D. V. PASECHNIK, *Polynomial-time computing over quadratic maps I: Sampling in real algebraic sets*, Computational Complexity, 14 (2005), pp. 20–52.
- [57] O. GÜLER AND L. TUNÇEL, *Characterization of the barrier parameter of homogeneous convex cones*, Math. Programming, 81 (1998), pp. 55–76.

- [58] E. GUTKIN, E. A. JONCKHEERE, AND M. KAROW, *Convexity of the joint numerical range: Topological and differential geometric viewpoints*, Linear Algebra and Its Applications, 376 (2004), pp. 143–171.
- [59] K. HAGELE, J. E. MORAIS, L. M. PARDO, AND M. SOMBRA, *On the intrinsic complexity of the arithmetic Nullstellensatz*, Journal of Pure and Applied Algebra, 146 (2000), pp. 103–183.
- [60] F. HAUSDORFF, *Der Wertvorrat einer Bilinearform*, Math. Zentralblatt, 3 (1919), pp. 314–316.
- [61] M. R. HESTENES AND E. J. MCSHANE, *A theorem on quadratic forms and its application in the calculus of variations*, Transactions of the AMS, 47 (1940), pp. 501–512.
- [62] J.-B. HIRIART-URRUTY, *Conditions for global optimality 2*, Journal of Global Optimization, 13 (1998), pp. 349–367.
- [63] —, *Global optimality conditions in maximizing a convex quadratic function under convex quadratic constraints*, Journal of Global Optimization, 21 (2001), pp. 445–455.
- [64] J. B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms*, no. 305 in A Series of Comprehensive Studies in Mathematics, Springer-Verlag, 1993.
- [65] A. E. HOERL AND R. W. KENNARD, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 12 (1970), pp. 55–67.
- [66] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, 1985.
- [67] T. ILLÉS, I. JOÓ, AND G. KASSAY, *On a nonconvex Farkas theorem and its application in optimization theory*, Report 1992-03, Eötvös University, Budapest, Hungary, 1992.
- [68] T. ILLÉS AND G. KASSAY, *Farkas type theorems for generalized convexities*, Pure Mathematics and Applications, 5 (1994), pp. 225–239.
- [69] —, *Theorems of the alternative and optimality conditions for convex-like and general convexlike programming*, Journal of Optimization Theory and Applications, 101 (1999), pp. 243–257.

- [70] U. T. JÖNSSON, *A lecture on the S-procedure*. Lecture notes, Division of Optimization and Systems Theory, Royal Institute of Technology, Stockholm, Sweden, May 2001.
- [71] S. KANEYUKI AND T. TSUYI, *Classification of homogeneous bounded domains of lower dimension*, Nagoya Math. Journal, 53 (1974), pp. 1–46.
- [72] S. KIM AND M. KOJIMA, *Exact solutions of some nonconvex quadratic optimization problems via SDP and SOCP relaxations*, Computational Optimization and Applications, 26 (2003), pp. 143–154.
- [73] M. KOJIMA AND L. TUNÇEL, *On the finite convergence of successive SDP relaxation methods*, European Journal of Operations Research, 143 (2002), pp. 325–341.
- [74] J. KOLLÁR, *Sharp effective Nullstellensatz*, Journal of the AMS, 1 (1988), pp. 963–975.
- [75] H. KÖNIG, *Über das von Neumannsche Minimax-Theorem*, Archiv der Mathematik, 19 (1968), pp. 482–487.
- [76] J. LEVIN, *Mathematical models for determining the intersections of quadric surfaces*, Computer Graphics and Image Processing, 11 (1979), pp. 73–87.
- [77] C.-K. LI AND Y.-T. POON, *Convexity of the joint numerical range*, SIAM Journal on Matrix Analysis and Applications, 21 (2000), pp. 668–678.
- [78] Z.-Q. LUO, *Applications of convex optimization in signal processing and digital communication*, Math. Programming (Series B), 97 (2003), pp. 177–207.
- [79] Z.-Q. LUO, J. F. STURM, AND S. ZHANG, *Conic linear programming and self-dual embedding*, Optimization Methods and Software, 14 (2000), pp. 169–218.
- [80] —, *Multivariate nonnegative quadratic mappings*, SIAM Journal on Optimization, 14 (2004), pp. 1140–1162.
- [81] A. I. LUR’E, *On the problem of stability of control systems*, Prikl. Mat. i Mekh., 15 (1951).

-
- [82] —, *Some Nonlinear Problems in the Theory of Automatic Control*, Gostekhizdat, Moscow, 1951.
- [83] A. I. LUR'E AND V. N. POSTNIKOV, *On the theory of stability of control systems*, Prikl. Mat. i Mekh., 8 (1944), pp. 3–13.
- [84] T. L. MAGNANTI, *Fenchel and Lagrange duality are equivalent*, Mathematical Programming, (1974), pp. 253–258.
- [85] H. MATSUMURA, *Commutative Ring Theory*, Cambridge University Press, Cambridge, UK, 1986.
- [86] A. MEGRETSKY, *S-procedure in optimal non-stochastic filtering*, Technical Report TRITA/MAT-92-0015, Department of Mathematics, Royal Institute of Technology, S-100 44 Stockholm, Sweden, 1992.
- [87] A. MEGRETSKY AND S. TREIL, *Power distribution in optimization and robustness of uncertain systems*, Journal of Mathematical Systems, Estimation, and Control, 3 (1993), pp. 301–319.
- [88] A. NEMIROVSKI, C. ROOS, AND T. TERLAKY, *On maximization of quadratic forms over intersection of ellipsoids with common center*, Math. Programming, 86 (1999), pp. 463–473.
- [89] Y. NESTEROV AND A. NEMIROVSKI, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Publications, Philadelphia, PA, 1994.
- [90] R. ORTEGA, *An energy amplification condition for decentralized adaptive stabilization*, IEEE Transactions on Automatic Control, 41 (1996), pp. 285–288.
- [91] G. PATAKI, *A partial characterization of nice cones*. Forthcoming.
- [92] —, *On the facial structure of cone-LP's and semidefinite programs*, Management Science Research Report MSRR-#595, Graduate School of Industrial Administration, Carnegie Mellon University, 1994.
- [93] —, *On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues*, Mathematics of Operations Research, 23 (1998), pp. 339–358.

-
- [94] —, *On the geometry of cone LPs*, in Handbook of Semidefinite Programming: Theory, Algorithms, and Applications, H. Wolkowicz, L. Vandenberghe, and R. Saigal, eds., Kluwer, 2000.
- [95] —, *On the closedness of the linear image of a closed convex cone*, Mathematics of Operations Research, (2007). Accepted for publication.
- [96] —, *A simple derivation of a facial reduction algorithm and extended dual systems*, technical report, Department of Statistics and OR, University of North Carolina, Chapel Hill, 2007. In preparation.
- [97] G. PATAKI AND S. SCHMIETA, *The DIMACS library of semidefinite-quadratic-linear programs*, preliminary draft, Columbia University, Computational Optimization Research Center, 2002.
- [98] J.-M. PENG AND Y.-X. YUAN, *Optimality conditions for the minimization of a quadratic with two quadratic constraints*, SIAM Journal on Optimization, 7 (1997), pp. 579–594.
- [99] H. PÉPIN, *Answer to a problem posed by J.-B. Hiriart-Urruty*, Revue de la filière Mathématiques (RMS), 3 (2004), pp. 171–172. In French.
- [100] E. PESONEN, *Über die Spektraldarstellung quadratischer Formen in linearen Räumen mit indefiniter Metrik*, vol. 227 of Annales Academiae Scientiarum Fennicae Series A. I. Mathematica, Academia Scientiarum Fennica, 1956.
- [101] M. C. PINAR AND M. TEBoulLE, *On semidefinite bounds for maximization of a non-convex quadratic objective over the ℓ_1 unit ball*, RAIRO Oper. Res., 40 (2006), pp. 253–265.
- [102] I. PÓLIK AND T. TERLAKY, *A survey of the S-lemma*, SIAM Review, 49 (2007), pp. 371–418.
- [103] —, *Strong duality for optimization over symmetric cones*, AdvOL Report 2007/10, McMaster University, Advanced Optimization Lab, Hamilton, Canada, 2007.
- [104] —, *New stopping criteria for detecting infeasibility in conic optimization*, AdvOL Report 2007/9, McMaster University, Advanced Optimization Lab, Hamilton, Canada, 2007.

-
- [105] B. T. POLYAK, *Convexity of quadratic transformations and its use in control and optimization*, Journal of Optimization Theory and Applications, 99 (1998), pp. 553–583.
- [106] —, *Convexity of nonlinear image of a small ball with applications to optimization*, Set-Valued Analysis, 9 (2001), pp. 159–168.
- [107] Y. T. POON, *On the convex hull of the multiform numerical range*, Linear and Multilinear Algebra, 37 (1994), pp. 221–223.
- [108] A. PRESTEL AND C. N. DELZELL, *Positive Polynomials*, Springer-Verlag, Berlin, 2001.
- [109] M. RAMANA AND A. J. GOLDMAN, *Quadratic maps with convex images*, Report 36-94, Rutgers Center for Operations Research, Rutgers, The State University of New Jersey, 1994.
- [110] M. V. RAMANA, *An exact duality theory for semidefinite programming and its complexity implications*, Math. Programming, (1997), pp. 129–162.
- [111] M. V. RAMANA AND R. M. FREUND, *On the ELSD duality theory for SDP*. Unpublished manuscript., 1996.
- [112] M. V. RAMANA, L. TUNÇEL, AND H. WOLKOWICZ, *Strong duality for semidefinite programming*, SIAM Journal on Optimization, 7 (1997), pp. 641–662.
- [113] C. R. RAO AND S. K. MITRA, *Generalized Inverse of Matrices and Its Applications*, John Wiley & Sons, Inc, New York, 1971.
- [114] J. RENEGAR, *A Mathematical View of Interior-Point Methods in Convex Optimization*, MPS/SIAM series on optimization, SIAM, 2001.
- [115] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N. J., 1970.
- [116] S. P. SCHURR, *An interior-point method for convex optimization using inexact barrier function evaluations*. McMaster Optimization Seminar, May 29, 2006.
- [117] M. SOMBRA, *Bounds for the Hilbert function of polynomial ideals and for the degrees in the Nullstellensatz*, Journal of Pure and Applied Algebra, 117&118 (1991), pp. 565–559.

-
- [118] —, *A sparse effective Nullstellensatz*, *Advances in Applied Mathematics*, 22 (1999), pp. 271–295.
- [119] R. J. STERN AND H. WOLKOWICZ, *Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations*, *SIAM Journal on Optimization*, 5 (1995), pp. 286–313.
- [120] J. STOER AND C. WITZGALL, *Convexity and Optimization in Finite Dimensions*, vol. I, Springer-Verlag, Heidelberg, 1970.
- [121] J. F. STURM, *Primal-Dual Interior Point Approach to Semidefinite Programming*, PhD thesis, Tinbergen Institute Research Series vol. 156, Tilburg University, 1997.
- [122] J. F. STURM AND S. ZHANG, *On cones of nonnegative quadratic functions*, *Mathematics of Operations Research*, 28 (2003), pp. 246–267.
- [123] M. J. TODD AND Y. YE, *Approximate Farkas lemmas and stopping rules for iterative infeasible-point algorithms for linear programming*, *Mathematical Programming*, 81 (1998), pp. 1–22.
- [124] O. TOEPLITZ, *Das algebraische Analogon zu einem Satze von Fejér*, *Math. Zentralblatt*, 2 (1918), pp. 187–197.
- [125] P. TSENG, *Further results on approximating nonconvex quadratic optimization by semidefinite programming relaxation*, *SIAM Journal on Optimization*, 14 (2003), pp. 268–283.
- [126] F. UHLIG, *A recurring theorem about pairs of quadratic forms and extensions: A survey*, *Linear Algebra and Its Applications*, 25 (1979), pp. 219–237.
- [127] E. B. VINBERG, *The theory of convex homogeneous cones*, *Transactions of the Moscow Mathematical Society*, 12 (1963), pp. 340–403.
- [128] —, *The structure of the group of automorphisms of a homogeneous convex cone*, *Transactions of the Moscow Mathematical Society*, 13 (1965), pp. 63–93.
- [129] K. WEIERSTRASS, *Zur Theorie der bilinearen und quadratischen Formen*, *Monatsberichte der Königl. Preuss. Akademie der Wissenschaften zu Berlin*, May (1868), pp. 310–338.
- [130] H. WOLKOWICZ. Personal communication., 2007.

-
- [131] S. WRIGHT, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, PA, 1997.
- [132] D. XU AND S. ZHANG, *Approximation bounds for quadratic maximization with semidefinite programming relaxation*, Technical Report SEEM 2003-01, Department of Systems Engineering & Engineering Management, The Chinese University of Hong Kong, Hong Kong, 2003.
- [133] V. A. YAKUBOVICH, *S-procedure in nonlinear control theory*, Vestnik Leningrad University, 1 (1971), pp. 62–77. In Russian.
- [134] —, *Minimization of quadratic functionals under quadratic constraints and the necessity of a frequency condition in the quadratic criterion for absolute stability of nonlinear control systems*, Soviet Math. Doklady, 14 (1973), pp. 593–597.
- [135] —, *S-procedure in nonlinear control theory*, Vestnik Leningrad University, 4 (1977), pp. 73–93. (English translation).
- [136] Y. YE AND S. ZHANG, *New results on quadratic minimization*, SIAM Journal on Optimization, 14 (2003), pp. 245–267.
- [137] Y.-X. YUAN, *On a subproblem of trust region algorithms for constrained optimization*, Math. Programming, 47 (1990), pp. 53–63.
- [138] F. ZHANG, ed., *The Schur Complement and its Applications*, vol. 4 of Numerical Methods and Algorithms, Springer, 2005.