Lectures on Modern Convex Optimization

MPS/SIAM Series on Optimization

This series is published jointly by the Mathematical Programming Society and the Society for Industrial and Applied Mathematics. It includes research monographs, textbooks at all levels, books on applications, and tutorials. Besides being of high scientific quality, books in the series must advance the understanding and practice of optimization and be written clearly, in a manner appropriate to their level.

Series Volumes

Ben-Tal, Aharon and Nemirovski, Arkadi, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*

Conn, Andrew R., Gould, Nicholas I. M., and Toint, Phillippe L., *Trust-Region Methods*

Lectures on Modern Convex Optimization

Analysis, Algorithms, and Engineering Applications

Aharon Ben-Tal
Arkadi Nemirovski
Technion-Israel Institute of Technology
Haifa, Israel

siam
Society for Industrial and Applied Mathematics
Philadelphia

MPS
Mathematical Programming Society
Philadelphia

10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data
Ben-Tal, A.
Lectures on modern convex optimization : analysis, algorithms, and engineering applications / Aharon Ben-Tal, Arkadi Nemirovski.
p. cm. — (MPS-SIAM series on optimization)
Includes bibliographical references and index.
ISBN 0-89871-491-5
1. Convex programming. 2. Mathematical optimization. I. Nemirovski, Arkadi Semenovich. II. Title. III. Series.

T57.815 .B46 2001
519.7'6—dc21

2001020818

This book is dedicated
to our friend and colleague, Jochem Zowe

Contents

Preface

The goals. To make decisions optimally is a basic human desire. Whenever the situation and the objectives can be described quantitatively, this desire can be satisfied, to some extent, by using mathematical tools, specifically those provided by optimization theory and algorithms. For our purposes, a sufficiently general mathematical setting of an optimization problem is offered by *mathematical programming*:

$$\begin{array}{lrcl} \text{Minimize} & f_0(x) & & \\ \text{subject to (s.t.)} & & & \\ & f_i(x) & \leq & b_i,\ i = 1, \ldots, m, \\ & x & \in & X \subset \mathbf{R}^n. \end{array} \tag{P}$$

In this problem, we are given an *objective* $f_0(x)$ and finitely many *functional constraints* $f_i(x)$, $i = 1, \ldots, m$, which are real-valued functions of n-dimensional *design vector* x varying in a given domain X. Our goal is to minimize the objective over the *feasible set* of the problem—the set that is cut off the domain X by the system of inequalities $f_i(x) \leq b_i$, $i = 1, \ldots, m$.

In typical engineering applications, the design vector specifies a decision to be made (e.g., the physical sizes of the bars in a trusslike structure), the domain X is the set of "meaningful" values of the design vector, and the functional constraints represent design specifications—restrictions (physical, technical, financial) on certain characteristics of the decision.

The last decade has witnessed major progress in optimization, especially in the area of convex programming. Powerful modeling languages and database technology extended our abilities to model large-scale, complex real-life problems; progress in complexity theory improved our understanding of the advantages of certain algorithms, and the limitations of others, and resulted in development of efficient interior point algorithms for a large family of convex programs. Combined with the dramatic improvement of computers, this progress enables us today to solve problems that were considered out of reach for optimization just 10 years ago.

Regrettably, this promising state of affairs is yet unrealized, and consequently not utilized, by potential end users (engineers, managers, etc.). This book presents modern optimization, combining it with applications (mainly from engineering) and thus helping to bridge the gap between researchers and practitioners. This ultimate goal determines our approach and dictates specific targets on which we should focus our attention.

Theoretically, what modern optimization can solve well are *convex optimization problems*. In essence, the two-decade-long investigation of complexity issues in optimization can be summarized as follows:

> *From the viewpoint of the numerical processing of problem* (P)*, there exists a "solvable case"—the one of* convex *optimization problems, those where the domain X is a closed convex subset of $\mathbf{R}^n$, and the objective $f_0(x)$ and the functional constraints $f_i(x)$, $i = 1, \ldots, m$, are convex functions on X.*
>
> *Under minimal additional computability assumptions* (which are satisfied in basically all applications)*, a convex optimization problem is computationally tractable—the computational effort required to solve the problem to a given accuracy grows moderately with the dimensions of the problem and the required number of accuracy digits.*
>
> *In contrast to this, general-type nonconvex problems are too difficult for numerical solution; the computational effort required to solve such a problem, by the best numerical methods known, grows prohibitively fast with the dimensions of the problem and the number of accuracy digits. Moreover, there are serious theoretical reasons to conjecture that this is an intrinsic feature of nonconvex problems rather than a drawback of the existing optimization techniques.*

As an example, consider the pair of optimization problems (A) and (B). The first is the nonconvex problem

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{i=1}^{n} x_i \\ \text{subject to} & \\ & x_i^2 - x_i = 0, \ i = 1, \ldots, n; \\ & x_i x_j = 0 \quad \forall (i, j) \in \Gamma, \end{array} \tag{A}$$

where Γ is a given set of pairs (i, j) of indices i, j. This is a fundamental combinatorial problem of computing the *stability number* of a graph. It arises, e.g., in the following *channel communication problem*:

There is an alphabet of n letters a_i, $i = 1, 2, \ldots, n$, say, the 256 usual bytes. A letter a_i can be sent through a communication channel; when passing through it, it either remains unchanged or can be converted to another letter a_j due to transmission errors. The errors are assumed to be "symmetric" (if a_i can be converted into a_j, then a_j can be converted into a_i as well), and Γ is the set of (indices of) those pairs of letters that can be converted from one into another. Assume that we are interested in a "nonconfusing" communication, where the addressee either gets a correct letter or is able to conclude that a transmission error has occurred but never misreads the input letter. In this case we should restrict the subalphabet S to be *independent*, meaning that no two distinct letters from S can be converted from one to another by the channel. To get the most from the channel, we would like to use an independent subalphabet of maximal cardinality. It turns out that the optimal value in (A) is exactly the cardinality of such a maximal independent subalphabet.

The second problem is the convex program

$$
\begin{array}{l}
\text{minimize } x_0 \\
\text{subject to} \\
\lambda_{\min}\begin{pmatrix} x_1 & & & x_1^\ell \\ & \ddots & & \vdots \\ & & x_m & x_m^\ell \\ x_1^\ell & \cdots & x_m^\ell & x_0 \end{pmatrix} \;\geq\; 0,\ \ell = 1,\ldots,k, \\
\qquad\qquad \displaystyle\sum_{j=1}^m a_j x_j^\ell \;=\; b^\ell,\ \ell = 1,\ldots,k, \\
\qquad\qquad \displaystyle\sum_{j=1}^m x_j \;=\; 1,
\end{array}
\tag{B}
$$

where $\lambda_{\min}(A)$ denotes the minimum eigenvalue of a symmetric matrix A. This problem originates from design of a *truss* (a mechanical construction built from thin elastic bars, like an electric mast, a bridge, or the Eiffel Tower) able to withstand k nonsimultaneous loads.

Looking at the analytical forms of problems (A) and (B), it seems that the first problem is easier than the second: the constraints in (A) are simple explicit quadratic equations, while the constraints in (B) involve much more complicated functions of the design variables—the eigenvalues of certain matrices depending on the design vector. The truth, however, is that the first problem is in a sense as difficult as an optimization problem can be, and the worst-case computational effort to solve it within absolute inaccuracy 0.5 is about 2^n operations for all known optimization methods. For $n = 256$ ("alphabet of bytes"), the complexity $2^n \approx 10^{77}$ is, for all practical purposes, the same as $+\infty$. In contrast to this, the second problem is quite computationally tractable. For example, for $k = 6$ (six loading scenarios) and $m = 100$ (a 100-bar construction), the problem has 701 variables (2.7 times the number of variables in the byte version of (A)); however, it can be reliably solved within six accuracy digits in a couple of minutes. The dramatic difference in the computational effort required to solve (A) and (B) is due to the fact that (A) is a nonconvex optimization problem, while (B) is convex.

The above discussion explains the words *Convex Programming* in the title of our book. We now explain the word *modern*. The emphasis in the book is on *well-structured convex problems* such as linear programming (LP), conic quadratic programming (CQP), and semidefinite programming (SDP). These are the areas most affected by the recent progress in optimization, areas where we possess well-developed techniques for building large-scale models, analyzing them theoretically ("on paper") and processing them numerically. Except for LP, these areas did not exist 10 years ago; thus most users who could potentially benefit from recent developments are not aware of their existence, let alone their usefulness. In enlarging the scope of classical optimization (LP, general nonlinear programming) by introduction of CQP and SDP, new application areas were created. Examples include semidefinite relaxations of difficult combinatorial problems, linear matrix inequality–based techniques in control, and mathematical programming with uncertain data. These new applications create synergies between optimization, computer science, and engineering, with

potentially far-reaching consequences for our ability to solve complex real-life decision problems.

At this point, we want to address experts in optimization rather than general readers. The history of convex programming, as far as applied aspects are concerned, started with the invention of LP (Dantzig, circa 1948). LPs possess the simplest possible and transparent analytical structure, which from the beginning was heavily exploited both theoretically (the completely algorithmic LP duality theory) and computationally (the simplex method). With subsequent nonlinear extensions, the focus was shifted, in one giant step, from the simplest possible linear structure to the most general one, where all we know about the entities occurring in (P) is that they are called f_i, $i = 0, 1, \ldots, m$, and X, that they are convex, and that f_i are $0/1/2/\ldots$ times continuously differentiable. At the theoretical level, the focus on this general setting yielded very deep results in convex analysis (Lagrange duality, Karush–Kuhn–Tucker (KKT) optimality conditions, etc.); however, the price paid for these fundamental achievements was a lack of an *algorithmic* content of the subject. For example, the Lagrange dual of a general-type convex program (P) is something that exists and possesses very nice properties; this "something," however, normally cannot be written explicitly (in sharp contrast with the "fully algorithmic" LP duality). At the computational level, the emphasis on generality resulted in general-purpose "near-sighted" (and thus slow, as compared to LP algorithms) optimization methods, those utilizing purely local information on the problem.

To some extent, recent trends (the last decade or so) in convex optimization stem from the realization that there is something between the relatively narrow LP and the completely unstructured universe of convex programming; what is between are well-structured generic convex optimization problems like CQP and SDP. Needless to say, interest in special cases is not a complete novelty for our area (recall linearly constrained convex quadratic and geometric programming); what is a novelty is the recent emphasis on well-structured generic problems, along with outstanding achievements resulting from this emphasis. The most remarkable of these achievements is, of course, the interior point revolution, which has extended dramatically our abilities to process convex programs numerically, while creating a completely new, challenging, and attractive area of theoretical research. The emphasis on well-structured special cases has, however, another, less-evident (perhaps not less-important) consequence, which can be summarized as follows:

• *When restricted to "nice" generic convex programs, like LP, CQP, and SDP, convex analysis becomes an algorithmic calculus—as algorithmic as in the LP case*. For example, the SDP duality is as explicit and symmetric as the LP one. In fact, the same can be said about all other basic operations of convex analysis, from the simplest (like taking intersections and affine images of convex sets, or sums of convex functions) to the more sophisticated ones (like passing from a convex set to its polar or from a convex function to its Legendre transformation). Whenever the operands of such a construction can be represented, in a properly defined precise sense, via, say, SDP, the same is true for the resulting entity, and the SDP representation of the result is readily given by those of the operands.

As a result,

• *An instance of a nice generic convex problem can be processed, up to a certain point, on paper (and in a routine fashion). In many cases this allows one to obtain important qualitative information on the problem or to convert it into an equivalent one, better suited for subsequent numerical processing.* For example, applying SDP duality, one can reduce

dramatically the design dimension of the truss topology design (TTD) problem and, as a result, increase by orders of magnitude the sizes of the TTD problems that can be processed numerically in practice.

Moreover,

• *nice generic convex problems, like CQP and especially SDP, possess vast expressive abilities*, which allow one to utilize the above advantages in a very wide spectrum of applications, much wider than the one covered by LP.

When writing this book, our major concern was to emphasize the issues just raised, and this emphasis is perhaps the most characteristic (and, hopefully, to some extent novel) feature of the book.

Restricting ourselves to well-structured convex optimization problems, it becomes logical to skip a significant part of what is traditionally thought of as the theory and algorithms of mathematical programming. Readers interested in the gradient descent, quasi-Newton methods, and even sequential quadratic programming, are kindly advised to use the excellent existing textbooks on these important subjects; our book should be thought of as a self-contained complement to, and not as an extended version of, these textbooks. We even have dared to omit the KKT optimality conditions in their standard form, since they are too general to be algorithmic; the role of the KKT conditions in our exposition is played by their particular (and truly algorithmic) case, expressed by the so-called conic duality theorem.

The book is addressed primarily to potential users (mainly engineers). Consequently, our emphasis is on building and processing instructive engineering models rather than on describing the details of optimization algorithms. The underlying motivation is twofold. First, we wish to convince engineers that optimization indeed has something to do with their professional interests. Second, we believe that a crucial necessary condition for successful practical applications of optimization is understanding what is desirable and what should be avoided at the *modeling* stage. Thus, important questions to be addressed are (a) *What optimization models can be efficiently processed* (to a certain point on paper and then on a computer)? and (b) *How one can convert* (provided that it is possible) *a seemingly bad initial description of a problem into a tractable and well-structured optimization model*?

We believe that the best way to provide relevant insight for potential users of optimization is to present, along with general concepts and techniques, many applications of these concepts and techniques. We believe that the presentation of *optimization algorithms* in a user-oriented book should be as nontechnical as possible (to drive a car, no knowledge of engines is necessary). The section devoted to algorithms presents the ellipsoid method (due to its simplicity, combined with its capability to answer affirmatively the fundamental question of whether convex programming is computationally tractable) and an overview of polynomial-time interior-point methods for LP, CQP, and SDP.

Although the book is user oriented, it is a mathematical book. Our goal is to demonstrate that when processing "meaningful" *mathematical* models by *rigorous* mathematical methods (not by their engineering surrogates), one can obtain results that have meaningful and instructive engineering interpretation. Whether we have reached this goal, is another story; this judgment rests upon the reader.

Last, but not least, a reader should keep in mind that what follows are *lecture notes*; our intention is to highlight those issues that we find most interesting and instructive, rather than to present a complete overview of convex optimization. Consequently, we are ready to take the blame for being boring or unclear or for focusing on issues of minor importance,

in any material in the book. However, we do not accept "money back" requests based on claims that something (however important) is *not* included. Along with immunity with regard to what is absent, a lecture notes–type book provides us with some other privileges, like a style which is a bit more vivid compared to the academic standards, and a rather short list of bibliography references (embedded as footnotes in the body of the text). In this latter respect, a reader should be aware that if a statement appears in the text without a reference, this does not mean that we are claiming authorship; it merely reflects that our focus is on the state of convex optimization rather than on its history.

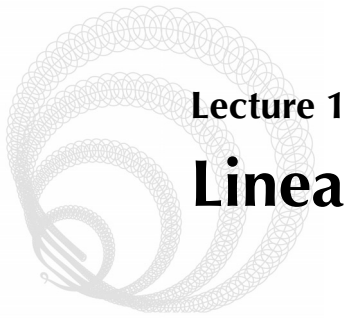
Audience and prerequisites. Formally, readers should know basic facts from linear algebra and analysis—those presented in standard undergraduate mathematical courses for engineers. For optimization-related areas, readers are assumed to know not much more than the definitions of a convex set and a convex function and to have heard (no more than that!) about mathematical programming problems. Informally, it is highly desirable that a reader is in possession of the basic elements of mathematical culture.

The exercises. Exercises accompanying each lecture form a significant part of the book. They are organized in small groups, each devoted to a specific topic related to the corresponding lecture. Typically, the task in an exercise is to prove something. Most of the exercises are not easy. The results stated by the exercises are used in the subsequent parts of the book in the same way as the statements presented and proved in the main body of the book; consequently, a reader is kindly asked to at least read all exercises carefully. The order of exercises is of primary importance: in many cases preceding exercises contain information and hints necessary to succeed in subsequent ones.

Acknowledgments. A significant part of the applications discussed in our book is taken from the papers of Prof. Stephen Boyd of Stanford University and his colleagues. We are greatly indebted to Prof. Boyd for providing us with access to this material and for stimulating discussions. Applications related to structural design were developed in tight collaboration with Prof. Jochem Zowe and Dr. Michael Kočvara of Erlangen University. Parts of the book were written when the authors were visiting the Statistics and Operations Research Department of the Technical University of Delft, and we are thankful to our hosts, Prof. Kees Roos and Prof. Tamas Terlaky.

Aharon Ben-Tal and Arkadi Nemirovski

August 2000, Haifa, Israel



Lecture 1
Linear Programming

In this chapter our primary goal is to present the basic results on the linear programming (LP) duality in a form that makes it easy to extend these results to the nonlinear case.

1.1 Linear programming: Basic notions

An LP program is an optimization program of the form

$$\min\left\{c^T x \;\middle|\; Ax \geq b\right\}, \tag{LP}$$

where

- $x \in \mathbf{R}^n$ is the *design vector*,
- $c \in \mathbf{R}^n$ is a given vector of coefficients of the *objective function* $c^T x$,
- A is a given $m \times n$ *constraint matrix*, and
- $b \in \mathbf{R}^m$ is a given *right-hand side of the constraints*.

(LP) is called

—*feasible* if its *feasible set*

$$\mathcal{F} = \{x \mid Ax - b \geq 0\}$$

is nonempty; a point $x \in \mathcal{F}$ is called a *feasible solution* to (LP);

—*bounded below* if it is infeasible or if its objective $c^T x$ is bounded below on $\mathcal{F}$.

For a feasible bounded-below problem (LP), the quantity

$$c^* \equiv \inf_{x: Ax-b\geq 0} c^T x$$

is called the *optimal value* of the problem. For an infeasible problem, we set $c_* = +\infty$, while for a feasible *un*bounded-below problem we set $c_* = -\infty$.

Linear programming is called *solvable* if it is feasible and bounded below and the optimal value is attained, i.e., there exists $x \in \mathcal{F}$ with $c^T x = c^*$. An x of this type is called an *optimal solution* to LP.

A priori it is unclear whether a feasible and bounded-below LP program is solvable: why should the infimum be achieved? It turns out, however, that a feasible and bounded-below program (LP) *always* is solvable. This nice fact (we shall establish it later) is specific for LP. Indeed, a very simple *nonlinear* optimization program

$$\min\left\{\frac{1}{x}\,\middle|\, x \geq 1\right\}$$

is feasible and bounded below, but it is not solvable.

1.2 Example: Tschebyshev approximation and its applications

In most textbooks known to us, examples of LP programs have to do with economics, production planning, etc., and indeed the major applications of LP are in these areas. In this book, however, we prefer to use, as a basic example, a problem related to engineering. Let us start with the mathematical formulation.

1.2.1 Best uniform approximation

Problem 1.2.1. Tschebyshev approximation. *Given an $M \times N$ matrix*

$$A = \begin{bmatrix} a_1^T \\ a_2^T \\ \cdots \\ a_M^T \end{bmatrix}$$

and a vector $b \in \mathbf{R}^M$, solve the problem

$$\min_{x \in \mathbf{R}^N} \|Ax - b\|_\infty, \ \textit{where } \|Ax - b\|_\infty \equiv \max_{i=1,\ldots,M} |a_i^T x - b_i|. \tag{1.2.1}$$

As stated, problem (1.2.1) is *not* an LP program—its objective is nonlinear. We can, however, immediately convert (1.2.1) to an equivalent LP program

$$\min_{x \in \mathbf{R}^n, t \in \mathbf{R}} \left\{ t \,\middle|\, -t \leq a_i^T x - b_i \leq t, \ i = 1, \ldots, M \right\}, \tag{1.2.2}$$

where t is an additional design variable. Thus, (1.2.1) is equivalent to an LP program.

A typical situation giving rise to the Tschebyshev problem is as follows. We want to approximate a given target function $\beta(t)$ on, say, the unit interval $[0, 1]$ by a linear

combination $\sum_{j=1}^{N} x_j\alpha_i(t)$ of N given functions $\alpha_j(t)$. The quality of approximation is measured by its uniform distance from β, i.e., by the quantity

$$\|\beta - \sum_{j=1}^{N} x_j\alpha_j\|_\infty \equiv \sup_{0\le t\le 1} |\beta(t) - \sum_{j=1}^{N} x_j(t)\alpha_j(t)|, \tag{1.2.3}$$

so that the best approximation is solution of the problem

$$\min_{x\in\mathbf{R}^N} \|\beta - \sum_{j=1}^{N} x_j\alpha_j\|_\infty. \tag{1.2.4}$$

As we shall see, problem (1.2.4) is important for several engineering applications. From the computational viewpoint, the drawback of the problem is that its objective is implicit—it involves maximization with respect to a continuously varying variable. As a result, already the related analysis problem (to evaluate the quality of the approximation corresponding to a given x) can be quite difficult numerically. The simplest way to overcome this drawback is to approximate in (1.2.3) the interval $[0, 1]$ by a "fine finite grid," e.g., the grid

$$T_M = \left\{ t_i = \frac{i}{M} \mid i = 1, \dots, M \right\}.$$

With this approximation, the objective in the problem (1.2.4) becomes

$$\max_{i=1,\dots,M} |\beta(t_i) - \sum_{j=1}^{N} x_j\alpha_j(t_i)| \equiv \|Ax - b\|_\infty,$$

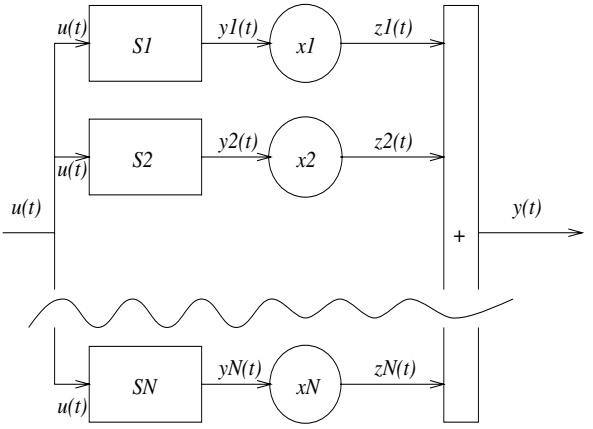
where the columns of A are the restrictions of the functions $\alpha_j(\cdot)$ on the grid T_M, and b is the restriction of $\beta(\cdot)$ on this grid. Consequently, the optimization problem (1.2.1) can be viewed as a discrete version of problem (1.2.4).

1.2.2 Application: Synthesis of filters

An interesting engineering problem corresponding to (1.2.4) is the problem of synthesizing a linear time-invariant (LTI) dynamic system (a "filter") with a given impulse response.[1]

> A (continuous-time) time-invariant linear dynamic system S is, mathematically, a transformation from the space of signals—functions on the axis—to the same space, given by the convolution with certain fixed function:
>
> $$u(t) \mapsto y(t) = \int_{-\infty}^{\infty} u(s)h(t-s)ds,$$
>
> where $u(\cdot)$ is an *input* and $y(\cdot)$ is the corresponding *output*. The convolution kernel $h(\cdot)$ is a characteristic function of the system S called the *impulse response*.

[1] The filter synthesis and the subsequent antenna array examples are taken from M.S. Lobo, L. Vanderbeghe, S. Boyd, and H. Lebret, *Second-order cone programming*, Linear Algebra Appl., 284 (1998), pp. 193–228.



$$\begin{aligned} y_j(t) &= \int_{-\infty}^{\infty} u(s)h_j(t-s)ds, \\ z_j(t) &= x_j y_j(t), \\ y(t) &= z_1(t) + z_2(t) + \cdots + z_N(t). \end{aligned}$$

Figure 1.1. *Parallel structure with amplifiers.*

Consider the simplest synthesis problem, as follows.

Problem 1.2.2. Filter synthesis, I. *Given a desired impulse response $h_*(t)$ along with N building blocks—standard systems S_j with impulse responses $h_j(\cdot)$, $j = 1, \ldots, N$—assemble these building blocks as shown in Fig. 1.1 into a system S in such a way that the impulse response of the assembled system will be as close as possible to the desired impulse response $h_*(\cdot)$.*

Note that the structure of S is given, and all we can play with are the amplification coefficients x_j, $j = 1, \ldots, N$.

The impulse response of the structure in Fig. 1.1 is clearly

$$h(t) = \sum_{j=1}^{N} x_j h_j(t).$$

Assuming further that h_* and all h_j vanish outside $[0, 1]$[2] and that we are interested in the best possible uniform approximation of the desired impulse response h_* on $[0, 1]$, we can pose our synthesis problem as (1.2.4) and further approximate it by (1.2.1). As we remember, the latter problem is equivalent to the LP program (1.2.2) and can therefore be solved by LP tools.

[2]Assumptions of this type have a natural interpretation. That impulse response vanishes to the left of the origin means that the corresponding system is *casual*—its output until any time t depends solely on the input until t and is independent of what happens with the input after time t. The fact that impulse response vanishes after certain $T > 0$ means that the *memory* of the corresponding system is at most T: output at a time t depends on what is the input starting with time $t - T$ and is independent of what the input was before time $t - T$.

1.2.3 Filter synthesis revisited

In the filter synthesis problem we wanted to combine given building blocks S_i to get a system with an impulse response as close as possible to the target one. A somewhat more typical design requirement is to get a system with a desired *transfer function*: the Fourier transform of the impulse response. The role of the transfer function becomes clear when we represent the action of an LTI system S in the frequency domain—the space of the Fourier transforms of inputs and outputs. In the frequency domain the transformation carried out by the system becomes

$$U(\omega) \mapsto Y(\omega) = U(\omega)H(\omega), -\infty < \omega < \infty, \tag{1.2.5}$$

where uppercase letters denote the Fourier transforms of their lowercase counterparts (e.g., $H(\omega)$ stands for the Fourier transform of the impulse response $h(t)$). Relation (1.2.5) demonstrates that the action of an LTI system in the frequency domain is very simple—it is just multiplication by the transfer function; this is why the analysis of an LTI system is carried out usually in the frequency domain, and thus typical design requirements on LTI systems are formulated in terms of their transfer functions.

The frequency domain version of the filter synthesis problem is as follows.

Problem 1.2.3. Filter synthesis, II. *Given a target transfer function $H_*(t)$ along with N building blocks—standard systems S_j with transfer function $H_j(\cdot)$, $j = 1, \ldots, N$—assemble these building blocks (as shown in Fig. 1.1) into a system S in such a way that the transfer function of the latter system will be as close as possible to the target function $H_*(\cdot)$.*

Again, we measure the quality of approximating the target transfer function on a given segment $[\omega_{\min}, \omega_{\max}]$ in the frequency domain by the uniform norm of the approximation error. Thus, the problem of interest is

$$\min_{x \in \mathbf{R}^n} \sup_{\omega_{\min} \le \omega \le \omega_{\max}} |H_*(\omega) - \sum_{j=1}^{N} x_j H_j(\omega)|,$$

and its computationally tractable approximation is

$$\min_{x \in \mathbf{R}^n} \max_{i=1,\ldots,M} |H_*(\omega_i) - \sum_{j=1}^{N} x_j H_j(\omega_i)|, \tag{1.2.6}$$

where $\{\omega_1, \omega_2, \ldots, \omega_M\}$ is a grid in $[\omega_{\min}, \omega_{\max}]$. Mathematically, the latter problem looks exactly as (1.2.1), and one could think that it can be immediately converted to an LP program. We should, however, take into account an important consideration:

> In contrast to impulse response, a transfer function is, generally, *complex-valued*. Consequently, the absolute values in (1.2.6) are absolute values of complex numbers. As a result, the conversion of (1.2.1) to an LP program now fails—the possibility to represent the nonlinear inequality $|a| \le t$ by two linear inequalities $a \le t$ and $a \ge -t$ exists in the case of real data only!

The difficulty we have met can be overcome in two ways:

1. The inequality $|a| \le t$ with *complex-valued* a can be represented as the inequality $\sqrt{\Re^2(a) + \Im^2(a)} \le t$ with real data ($\Re(a)$ is the real and $\Im(a)$ the imaginary part of a). As a result, problem (1.2.6) can be posed as a *conic quadratic* problem. Such problems will be our subject in Lecture 3.

2. The inequality $|a| \le t$ with complex-valued a can be easily *approximated* by a number of linear inequalities on $\Re(a)$ and $\Im(a)$. Indeed, let us inscribe into the unit circle D on the complex plane $\mathbf{C} = \mathbf{R}^2$ a $(2k)$-vertex perfect polygon P_k:

$$P_k = \{(u, v) \in \mathbf{R}^2 : |u\cos(\ell\phi) + v\sin(\ell\phi)| \le \cos(\phi/2), \ell = 1, \dots, k\}, \quad \left[\phi = \frac{\pi}{k}\right].$$

For $z = u + iv$, denote

$$p_k(z) = \max_{l=1,\dots,k} |u\cos(l\phi) + v\sin(l\phi)|.$$

We claim that for every $z = (u, v) \in \mathbf{R}^2$ one has

$$|z| \ge p_k(z) \ge \cos(\phi/2)|z|. \tag{1.2.7}$$

The left inequality in (1.2.7) follows from the Cauchy inequality: for $z = u + iv$ one has

$$|u\cos(\ell\phi) + v\sin(\ell\phi)| \le |z|\sqrt{\cos^2(\ell\phi) + \sin^2(\ell\phi)} = |z|,$$

hence $p_k(z) \le |z|$. The right inequality follows from the fact that P_k is inside D:

$$|z| = 1 \Rightarrow z \notin \mathrm{int} P_k \Rightarrow p_k(z) \ge \cos(\phi/2).$$

Since both $|z|$ and $p_k(z)$ are positive homogeneous of degree 1, i.e.,

$$p_k(\lambda z) = \lambda p_k(z), |\lambda z| = \lambda|z| \quad \forall \lambda \ge 0,$$

the validity of the inequality $p_k(z) \ge \cos(\phi/2)|z|$ for $|z| = 1$ implies the validity of the inequality for all z.

We see that the absolute value $|z|$ of a complex number can be approximated, within relative accuracy $1 - \cos(\phi/2) = 1 - \cos(\pi/(2k))$, by the polyhedral norm $p_k(z)$—the maximum of absolute values of k linear forms of $\Re(z)$ and $\Im(z)$. As an illustration, for $k = 4$ we approximate $|z|$ within the 7.7% margin; see Fig. 1.2. For most engineering applications, it is basically the same: approximate H_* in the uniform norm on a grid $\Omega = \{\omega_1, \dots, \omega_M\}$ or in the polyhedral norm

$$\max_{i=1,\dots,M} p_4\left(H_*(\omega_i) - \sum_{j=1}^N x_j H_j(\omega_i)\right);$$

the corresponding measures of the quality of an approximation differ by less than 8%. Consequently, one can pass from problem (1.2.6) to its approximation,

$$\min_x \max_{i=1,\dots,M} p_4\left(H_*(\omega_i) - \sum_{j=1}^N x_j H_j(\omega_i)\right). \tag{1.2.8}$$

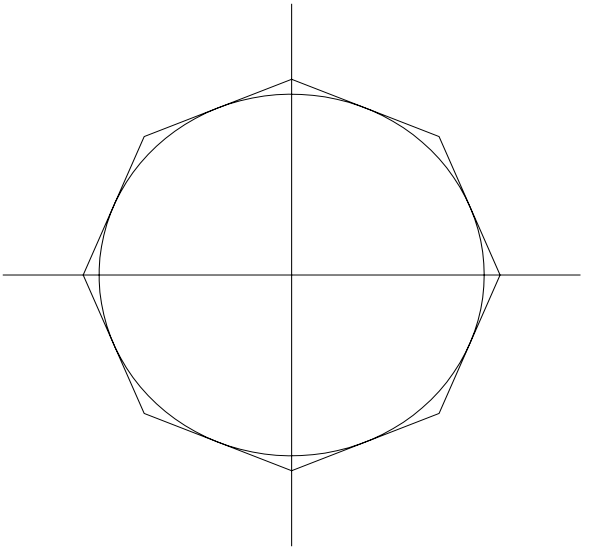


Figure 1.2. *The contours* $|z| = 1$ *(circle) and* $p_4(z) = 1$ *(polygon).*

Problem (1.2.8) is equivalent to the program

$$\min_{x,t} \left\{ t \;\middle|\; p_4\left(H_*(\omega_i) - \sum_{j=1}^{N} x_j H_j(\omega_i) \right) \le t,\ i = 1, \ldots, M \right\},$$

which (due to the structure of $p_4(\cdot)$) is equivalent to the LP program

$$\min_{x,t} \left\{ t \;\middle|\; -t \le \cos(\ell\phi)\Re\left(H_*(\omega_i) - \sum_{j=1}^{N} x_j H_j(\omega_i) \right) + \sin(\ell\phi)\Im\left(H_*(\omega_i) - \sum_{j=1}^{N} x_j H_j(\omega_i) \right) \le t, \quad \begin{matrix} i = 1, \ldots, M \\ \ell = 1, \ldots, 4 \end{matrix} \right\}. \quad (1.2.9)$$

1.2.4 Synthesis of arrays of antennae

An important engineering application of the Tschebyshev approximation problem is the synthesis of arrays of antennae. An antenna is an electromagnetic device that can generate (or receive) electromagnetic waves. The main characteristic of a monochromatic antenna is its *diagram* $Z(\delta)$, which is a complex-valued function of a three-dimensional (3D) direction δ. The absolute value $|Z(\delta)|$ of the diagram is responsible for the directional density of the energy sent by the antenna in a direction δ; this density is proportional to $|Z(\delta)|^2$. The argument $\arg Z(\delta)$ of the diagram corresponds to the initial phase of the wave propagating

in the direction δ, so that the electric field generated by the antenna at a point $P = r\delta$ (we assume that the antenna is placed at the origin) is proportional to

$$E(r\delta, t) = |Z(\delta)|r^{-1}\cos(\arg Z(\delta) + t\omega - 2\pi r/\lambda), \tag{1.2.10}$$

where t stands for time, ω is the frequency of the wave, and λ is the wavelength.[3] For a complex antenna comprising N antenna elements with diagrams $Z_1(\delta), \ldots, Z_N(\delta)$, the diagram $Z(\cdot)$ is

$$Z(\delta) = \sum_{j=1}^{N} Z_j(\delta).$$

When designing an array of antennae comprising several antenna elements, an engineer starts with N given building blocks with diagrams $Z_1, \ldots, Z_N$. For each block, the engineer can *amplify* the signal sent by a block by a factor ρ_j and *shift* the initial phase $\phi_j(\cdot)$ by a constant ψ_j. In other words, the engineer can modify the original diagrams of the blocks according to

$$\begin{aligned} Z_j(\delta) &\equiv a_j(\delta)[\cos(\phi_j(\delta)) + i\sin(\phi_j(\delta))] \\ &\mapsto Z_j^+(\delta) = \rho_j a_j(\delta)[\cos(\phi_j(\delta) + \psi_j) + i\sin(\phi_j(\delta) + \psi_j)]. \end{aligned}$$

Thus, it is possible to multiply the initial diagram of every block by an arbitrary complex constant ("weight")

$$z_j = \rho_j(\cos\psi_j + i\sin\psi_j) \equiv u_j + iv_j.$$

The diagram of the resulting complex antenna will be

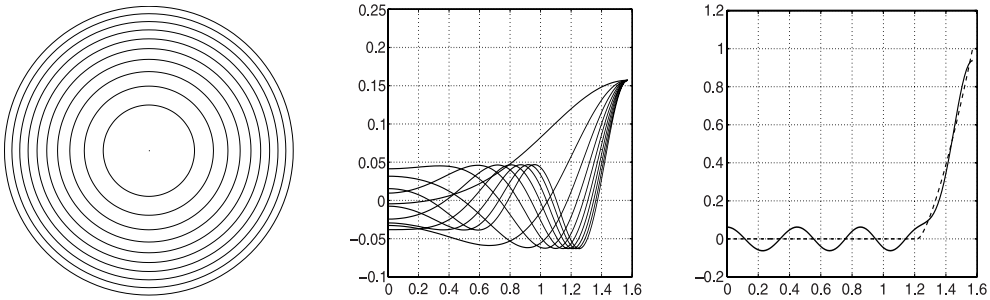
$$Z(\delta) = \sum_{j=1}^{N} z_j Z_j(\delta). \tag{1.2.11}$$

A typical design problem associated with the above description is to choose the design parameters z_j, $j = 1, \ldots, N$, in order to get a diagram (1.2.11) as close as possible to a given target diagram $Z_*(\delta)$. In many cases a relevant measure of closeness is in terms of the uniform norm, and the corresponding synthesis problem becomes

$$\min_{z_1,\ldots,z_N \in \mathbf{C}^n} \max_{\delta:\|\delta\|_2=1} |Z_*(\delta) - \sum_{j=1}^{N} z_j Z_j(\delta)|,$$

where the design variables are N complex numbers $z_1, \ldots, z_N$, or, which is the same, $2N$ real numbers $\Re(z_j)$, $\Im(z_j)$. Mathematically, the resulting problem is completely similar to the one discussed in the previous section, up to the fact that now the inner maximum in the objective is taken over the unit sphere in $\mathbf{R}^3$ rather than an interval on the axis. Even this difference disappears when we approximate the maximum over continuously varying direction by the maximum over a finite grid on the sphere (this in any case is necessary to get an efficiently solvable optimization program). Thus, the problem of synthesis of an

[3] Relation (1.2.10) works when the distance r between P and the antenna is much larger than the linear sizes of the antenna. Mathematically, the difference between the left and the right sides in (1.2.10) is $o(r^{-1})$ as $r \to \infty$.



(a) (b) (c)

Figure 1.3. *Synthesis of antennae array. (a): 10 array elements of equal areas in the XY-plane the outer radius of the largest ring is 1m, the wavelength is 50cm. (b): "building blocks" – the diagrams of the rings as functions of the altitude angle θ. (c): the target diagram (dashed) and the synthesied diagram (solid).*

array of antennae is, mathematically, identical to the problem of synthesis of an LTI system with a desired transfer function; in particular, we can approximate the problem by an LP program.

Example. Let a planar antenna array comprise a central circle and nine concentric rings of the same area as the circle (Fig. 1.3(a)). The array is placed in the XY-plane (Earth's surface), and the outer radius of the outer ring is 1 m.

One can easily see that the diagram of a ring $\{a \le r \le b\}$ in the plane XY (r is the distance from a point to the origin) as a function of a 3D direction δ depends on the altitude (the angle θ between the direction and the plane) only. The resulting function of θ turns out to be *real-valued*, and its analytic expression is

$$Z_{a,b}(\theta) = \frac{1}{2}\int_a^b \left[\int_0^{2\pi} r\cos\left(2\pi r\lambda^{-1}\cos(\theta)\cos(\phi)\right) d\phi\right] dr,$$

where λ is the wavelength. Figure 1.3(b) represents the diagrams of our 10 rings for $\lambda = 50$ cm.

Assume that our goal is to design an array with a real-valued diagram that should be axial symmetric with respect to the Z-axis and should be concentrated in the cone $\pi/2 \ge \theta \ge \pi/2 - \pi/12$. In other words, our target diagram is a real-valued function Z_* of the altitude θ with $Z_*(\theta) = 0$ for $0 \le \theta \le \pi/2 - \pi/12$ and $Z_*(\theta)$ somehow approaching 1 as θ approaches $\pi/2$. The target diagram $Z_*(\theta)$ used in this example is given in Fig. 1.3(c) (the dashed curve).

Our design problem is simplified considerably by the fact that the diagrams of our building blocks and the target diagram are real valued; thus, we need no complex numbers, and the problem we should finally solve is

$$\min_{x\in\mathbf{R}^{10}}\left\{\max_{\theta\in T}|Z_*(\theta) - \sum_{j=1}^{10} x_j Z_{r_{j-1},r_j}(\theta)|\right\},$$

Table 1.1. *Optimal weights (rounded to five significant digits).*

element	1	2	3	4	5	6	7	8	9	10
coefficient	1624.4	−14700	55383	−107247	95468	19221	−138620	144870	−69303	13311

where T is a finite grid on the segment $[0, \pi/2]$. In the design represented in Fig. 1.3(c), the 120-point equidistant grid is used. Both the data and the design variables in the problem are real, so that we can immediately convert the problem into an equivalent LP program.

The solid line in Fig. 1.3(c) represents the optimal diagram of the array of antennae given by the synthesis. The uniform distance between the actual and the target diagram is ≈ 0.0621 (recall that the target diagram varies from 0 to 1). Table 1.1 displays the optimal weights (i.e., the coordinates x_j of the optimal solution).

Why the uniform approximation? The antenna array example raises a natural question: Why is the distance between the target diagram $Z_*(\cdot)$ and the synthesized one $Z(\cdot)$ measured by the uniform norm of the residual $\|Z_* - Z\|_\infty = \max_\theta |Z_*(\theta) - Z(\theta)|$ and not by, say, the 2-norm $\|Z_* - Z\|_2 = \sqrt{\sum_\theta |Z_*(\theta) - Z(\theta)|^2}$? With this latter norm—i.e., with the standard least squares approximation—the (squared) objective to be minimized would be a sum of squares of affine forms of the design variables, i.e., a convex quadratic form $\frac{1}{2}x^T Ax + b^T x + c$, and we could immediately write the optimal solution $x^* = -A^{-1}b$, thus avoiding any need for numerical optimization.

Note, however, that the *only* advantage of the $\|\cdot\|_2$-accuracy measure is that it leads to a computationally cheap approximation routine. From the modeling viewpoint, least squares are not attractive in many cases. Indeed, consider the case when the target function is nearly singular—it is close to one constant, say, to 0, in the major part A_0 of its domain and is close to another constant, say, to 1, in another relatively small part A_1 of the domain. This is the case in our antenna synthesis example: we are trying to concentrate the signal in a small cone, and what is of interest is exactly the nearly singular behavior of the signal. Now, with an integral-type norm of the residual, like $\|\cdot\|_2$, the typical squared deviation between an approximation and the target in A_0 is taken with relatively large weight (proportional to the cardinality of A_0), while the typical squared deviation between the functions in A_1 is taken with relatively small weight. It follows that in order to get a good $\|\cdot\|_2$-approximation, it pays to concentrate on a good approximation of the background behavior of the target (the one in A_0), even at the price of poor reproduction of the singular behavior of the target (the one in A_1). As a result, least squares designs usually result in oversmoothed approximations that badly capture the near singularity of the target—a feature of the target we are most interested in. In contrast, $\|\cdot\|_\infty$ design pays the same attention to how well we reproduce the background behavior of the target and to how well we reproduce its singularities; this feature of the uniform norm makes it a better candidate than the $\|\cdot\|_2$-norm to be used in approximation problems with singularities. To illustrate this point, let us look at what the least squares yield in our example (Fig. 1.4). We see that the least squares approximation indeed pays more attention to the background than to the singularity. The uniform distance

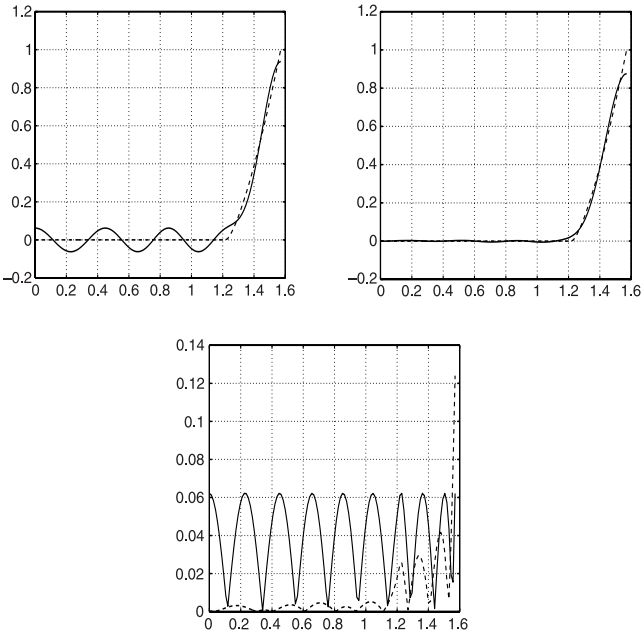


Figure 1.4. *Top: Best uniform approximation (left) versus the least squares approximation (right). Bottom: Errors of the least squares (dashed) and the best uniform (solid) approximations.*

between the target and the least squares approximation is 0.1240—more than twice the distance corresponding to the best uniform approximation!

1.3 Duality in linear programming

The most important and interesting feature of LP as a mathematical entity (other than computations and applications) is the wonderful *LP duality theory* we are about to consider. We motivate this topic by first addressing the following question:

Given an LP program

$$c^* = \min_x \left\{ c^T x \;\middle|\; Ax - b \geq 0 \right\}, \tag{LP}$$

how do we find a systematic way to bound from below its optimal value c^*?

Why this is an important question, and how the answer helps us deal with LP, will be seen later. For the time being, let us accept that the question is worthy of the effort.

A trivial answer to the posed question is to solve (LP) and see what is the optimal value. There is, however, a smarter and much more instructive way to answer our question. Let us look at the following example:

$$\min\left\{x_1+x_2+\cdots+x_{1999}\;\middle|\;\begin{array}{rcl} x_1+2x_2+\cdots+1998x_{1998}+1999x_{1999}-1 & \geq & 0,\\ 1999x_1+1998x_2+\cdots+2x_{1998}+x_{1999}-100 & \geq & 0,\\ \cdots & \cdots & \cdots \end{array}\right\}.$$

We claim that the optimal value in the problem is $\geq \frac{101}{2000}$. How could one certify this bound? This is immediate: add the first two constraints to get the inequality

$$2000(x_1+x_2+\cdots+x_{1998}+x_{1999})-101\geq 0$$

and divide the resulting inequality by 2000. LP duality is nothing but a straightforward generalization of this simple trick.

1.3.1 Certificates for solvability and insolvability

Consider a (finite) system of scalar inequalities with n unknowns. To be as general as possible, we do not assume for now that the inequalities are linear, and we allow for both nonstrict and strict inequalities in the system, as well as for equalities. Since an equality can be represented by a pair of nonstrict inequalities, our system can always be written as

$$f_i(x)\;\Omega_i\;0,\;\; i=1,\ldots,m, \qquad (\mathcal{S})$$

where every Ω_i is either the relation $>$ or the relation $\geq$.

The basic question about $(\mathcal{S})$ is whether $(\mathcal{S})$ has a solution. When we can answer this question, we can answer many other questions. For example, verifying whether a given real a is a lower bound on the optimal value c^* of (LP) is the same as verifying whether the system

$$\left\{\begin{array}{rcl} -c^Tx+a & > & 0,\\ Ax-b & \geq & 0 \end{array}\right.$$

has no solutions.

The general question above is too difficult, and it makes sense to pass from it to a seemingly simpler one: How do we certify that $(\mathcal{S})$ has, or does not have, a solution? Imagine that you are very smart and know the correct answer to the first question; how could you convince somebody that your answer is correct? What would certify the validity of your answer for everybody?

If your claim is that $(\mathcal{S})$ is solvable, certification could come from pointing out a solution x^* to $(\mathcal{S})$. Then one can substitute x^* into the system and check whether x^* indeed is a solution.

Assume now that your claim is that $(\mathcal{S})$ has no solutions. What could be a simple certificate of this claim? How one could certify a *negative* statement? This is a highly nontrivial problem and not just for mathematics; for example, in criminal law, how should someone accused in a murder prove his innocence? The real-life answer to how to certify a negative statement is discouraging: such a statement normally *cannot* be certified. In

mathematics, however, the situation is different: in some cases there exist simple certificates of negative statements. For example, to certify that $(\mathcal{S})$ has no solutions, it suffices to demonstrate that a consequence of $(\mathcal{S})$ is a contradictory inequality, such as

$$-1 \geq 0.$$

For example, assume that λ_i, $i = 1, \ldots, m$, are nonnegative weights. Combining inequalities from $(\mathcal{S})$ with these weights, we come to the inequality

$$\sum_{i=1}^{m} \lambda_i f_i(x) \ \Omega \ 0, \qquad \text{(Cons}(\lambda)\text{)}$$

where Ω is either $>$ (this is the case when the weight of at least one strict inequality from $(\mathcal{S})$ is positive) or $\geq$ (otherwise). Since the resulting inequality, due to its origin, is a consequence of the system $(\mathcal{S})$ (i.e., it is satisfied by every solution to $(\mathcal{S})$), it follows that if (Cons(λ)) has no solutions at all, we can be sure that $(\mathcal{S})$ has no solution. Whenever this is the case, we may treat the corresponding vector λ as a simple certificate of the fact that $(\mathcal{S})$ is infeasible.

Let us look at what the outlined approach means when $(\mathcal{S})$ is made up of *linear* inequalities:

$$(\mathcal{S}): \quad \{\alpha_i^T x \ \Omega_i \ b_i, \ i = 1, \ldots, m\}, \quad \left[\Omega_i = \begin{cases} > \\ \geq \end{cases}\right].$$

Here the combined inequality is linear as well:

$$\text{(Cons}(\lambda)\text{)}: \qquad \left(\sum_{i=1}^{m} \lambda a_i\right)^T x \ \Omega \ \sum_{i=1}^{m} \lambda b_i$$

(Ω is $>$ whenever $\lambda_i > 0$ for at least one i with $\Omega_i =>$, and Ω is $\geq$ otherwise). Now, when can a linear inequality

$$d^T x \ \Omega \ e$$

be contradictory? Of course, it can happen only when the left-hand side is identically zero, i.e., only when $d = 0$. Whether in this latter case the inequality is contradictory depends on the relation Ω: if Ω is $>$, then the inequality is contradictory if and only if $e \geq 0$, and if Ω is $\geq$, the inequality is contradictory if and only if $e > 0$. We have established the following simple result.

Proposition 1.3.1. *Consider a system of linear inequalities*

$$(\mathcal{S}): \qquad \begin{cases} a_i^T x & > & b_i, \ i = 1, \ldots, m_{\mathrm{s}}, \\ a_i^T x & \geq & b_i, \ i = m_{\mathrm{s}} + 1, \ldots, m, \end{cases}$$

with n-dimensional vector of unknown x. Let us associate with $(\mathcal{S})$ two systems of linear inequalities and equations with m-dimensional vector of unknown λ:

$$\mathcal{T}_{\mathrm{I}}: \quad \begin{cases} (a) & \lambda & \geq & 0, \\ (b) & \sum_{i=1}^{m} \lambda_i a_i & = & 0, \\ (c_{\mathrm{I}}) & \sum_{i=1}^{m} \lambda_i b_i & \geq & 0, \\ (d_{\mathrm{I}}) & \sum_{i=1}^{m_{\mathrm{s}}} \lambda_i & > & 0, \end{cases}$$

$$\mathcal{T}_{\mathrm{II}}: \quad \begin{cases} (a) & \lambda & \geq & 0, \\ (b) & \sum_{i=1}^{m} \lambda_i a_i & = & 0, \\ (c_{\mathrm{II}}) & \sum_{i=1}^{m} \lambda_i b_i & > & 0. \end{cases}$$

Assume that at least one of the systems $\mathcal{T}_{\mathrm{I}}$, $\mathcal{T}_{\mathrm{II}}$ is solvable. Then the system $(\mathcal{S})$ is infeasible.

Proposition 1.3.1 says that in some cases it is easy to certify infeasibility of a linear system of inequalities: a simple certificate is a solution to another system of linear inequalities. Note, however, that the existence of a certificate of this latter type is to the moment only a *sufficient*, but not a *necessary*, condition for the infeasibility of $(\mathcal{S})$. A fundamental result in the theory of linear inequalities is that the sufficient condition in question is in fact also necessary.

Theorem 1.3.1. General theorem on the alternative. *In the notation from Proposition* 1.3.1, *system $(\mathcal{S})$ has no solution if and only if either $\mathcal{T}_{\mathrm{I}}$ or $\mathcal{T}_{\mathrm{II}}$, or both systems, is solvable.*

The proof of the theorem on the alternative, as well as a number of useful particular cases of it, is an exercise topic of Lecture 1. We explicitly formulate here two very useful principles following from the theorem:

1. A system of linear inequalities

$$a_i^T x \ \Omega_i \ b_i, \ \ i = 1, \ldots, m,$$

has no solutions if and only if one can combine the inequalities of the system in a linear fashion (i.e., multiplying the inequalities by nonnegative weights, adding the results, and passing, if necessary, from an inequality $a^T x > b$ to the inequality $a^T x \geq b$) to get a contradictory inequality, namely, either the inequality $0^T x \geq 1$ or the inequality $0^T x > 0$.

2. A linear inequality

$$a_0^T x \ \Omega_0 \ b_0$$

is a consequence of a solvable system of linear inequalities

$$a_i^T x \ \Omega_i \ b_i, \ \ i = 1, \ldots, m,$$

if and only if it can be obtained by combining, in a linear fashion, the inequalities of the system and the trivial inequality $0 > -1$.

It should be stressed that the above principles are highly nontrivial and very deep. Consider, e.g., the following system of four linear inequalities with two variables u, v:

$$\begin{array}{c} -1 \le u \le 1, \\ -1 \le v \le 1. \end{array}$$

From these inequalities it follows that

$$u^2 + v^2 \le 2, \tag{i}$$

which in turn implies, by the Cauchy inequality, the linear inequality $u + v \le 2$:

$$u + v = 1 \times u + 1 \times v \le \sqrt{1^2 + 1^2}\sqrt{u^2 + v^2} \le (\sqrt{2})^2 = 2. \tag{ii}$$

The concluding inequality is linear and is a consequence of the original system, but in the demonstration of this fact both steps (i) and (ii) are highly nonlinear. It is absolutely unclear a priori why the same consequence can, as stated by principle 1, be derived from the system in a linear manner as well. (Of course it can—just add two inequalities $u \le 1$ and $v \le 1$.)

Note that the theorem on the alternative and its corollaries 1 and 2 heavily exploit the fact that we are speaking about *linear* inequalities. For example, consider two quadratic and two linear inequalities with two variables,

$$\begin{array}{lrcl} \text{(a)} & u^2 & \ge & 1, \\ \text{(b)} & v^2 & \ge & 1, \\ \text{(c)} & u & \ge & 0, \\ \text{(d)} & v & \ge & 0, \end{array}$$

along with the quadratic inequality

$$\text{(e)} \quad uv \ \ge \ 1.$$

The inequality (e) is clearly a consequence of (a)–(d). However, if we extend the system of inequalities (a)–(b) by all "trivial" (i.e., identically true) linear and quadratic inequalities with two variables, like $0 > -1$, $u^2 + v^2 \ge 0$, $u^2 + 2uv + v^2 \ge 0$, $u^2 - uv + v^2 \ge 0$, etc., and ask whether (e) can be derived in a *linear* fashion from the inequalities of the extended system, the answer will be negative. Thus, principle 1 fails to be true already for quadratic inequalities (which is a great sorrow—otherwise there were no difficult problems at all!).

We are about to use the theorem on the alternative to obtain the basic results of the LP duality theory.

1.3.2 Dual to a linear programming program: Origin

As mentioned, the motivation for constructing the problem dual to an LP program

$$c^* = \min_x \left\{ c^T x \;\middle|\; Ax - b \geq 0 \right\}, \quad A = \begin{bmatrix} a_1^T \\ a_2^T \\ \dots \\ a_m^T \end{bmatrix} \in \mathbf{R}^{m \times n} \tag{LP}$$

is the desire to generate, in a systematic way, lower bounds on the optimal value c^* of (LP). As previously explained, $a \in \mathbf{R}$ is such a lower bound if and only if $c^T x \geq a$ whenever $Ax \geq b$ or, which is the same, if and only if the system of linear inequalities

$$(\mathcal{S}_a): \qquad -c^T x > -a,\ Ax \geq b$$

has no solution. We know by the theorem on the alternative that the latter fact means that some other system of linear equalities (specifically, at least one of a certain pair of systems) does have a solution. More precisely,

(*) *($\mathcal{S}_a$) has no solutions if and only if at least one of the following two systems with $m+1$ unknowns has a solution:*

$$\mathcal{T}_{\mathrm{I}}: \quad \begin{cases} \text{(a)} & \lambda = (\lambda_0, \lambda_1, \dots, \lambda_m) & \geq & 0, \\ \text{(b)} & -\lambda_0 c + \sum_{i=1}^m \lambda_i a_i & = & 0, \\ (\mathrm{c_I}) & -\lambda_0 a + \sum_{i=1}^m \lambda_i b_i & \geq & 0, \\ (\mathrm{d_I}) & \lambda_0 & > & 0, \end{cases}$$

or

$$\mathcal{T}_{\mathrm{II}}: \quad \begin{cases} \text{(a)} & \lambda = (\lambda_0, \lambda_1, \dots, \lambda_m) & \geq & 0, \\ \text{(b)} & -\lambda_0 c - \sum_{i=1}^m \lambda_i a_i & = & 0, \\ (\mathrm{c_{II}}) & -\lambda_0 a - \sum_{i=1}^m \lambda_i b_i & > & 0. \end{cases}$$

Now assume that (LP) is feasible. We claim that under this assumption ($\mathcal{S}_a$) has no solutions if and only if $\mathcal{T}_{\mathrm{I}}$ has a solution.

The implication "$\mathcal{T}_{\mathrm{I}}$ has a solution $\Rightarrow$ $\mathcal{S}_a$ has no solution" is readily given by the above remarks. To verify the inverse implication, assume that ($\mathcal{S}_a$) has no solutions and the system $Ax \geq b$ has a solution, and let us prove then that $\mathcal{T}_{\mathrm{I}}$ has a solution. If $\mathcal{T}_{\mathrm{I}}$ has no solution, then by (*) $\mathcal{T}_{\mathrm{II}}$ has a solution and, moreover, $\lambda_0 = 0$ for (every) solution to $\mathcal{T}_{\mathrm{II}}$ (since a solution to the latter system with $\lambda_0 > 0$ solves $\mathcal{T}_{\mathrm{I}}$ as well). But the fact that $\mathcal{T}_{\mathrm{II}}$ has a solution λ with $\lambda_0 = 0$ is independent of the values of a and c; if this fact would take place, it would

mean, by the same theorem on the alternative, that, e.g., the following instance of $(\mathcal{S}_a)$ has no solution:

$$0^T x \geq -1,\ Ax \geq b.$$

The latter means that the system $Ax \geq b$ has no solutions—a contradiction with the assumption that (LP) is feasible.

Now, if $\mathcal{T}_I$ has a solution, this system has a solution with $\lambda_0 = 1$ as well. (To see this, pass from a solution λ to the one λ/λ_0; this construction is well defined, since $\lambda_0 > 0$ for every solution to $\mathcal{T}_I$.) Now, an $(m+1)$-dimensional vector $\lambda = (1, y)$ is a solution to $\mathcal{T}_I$ if and only if the m-dimensional vector y solves the system of linear inequalities and equations

$$\begin{aligned} y &\geq 0, \\ A^T y \equiv \sum_{i=1}^{m} y_i a_i &= c, \\ b^T y &\geq a. \end{aligned} \tag{D}$$

Summarizing our observations, we come to the following result.

Proposition 1.3.2. *Assume that system* (D) *associated with the LP program* (LP) *has a solution* (y, a). *Then* a *is a lower bound on the optimal value in* (LP). *Likewise, if* (LP) *is feasible and* a *is a lower bound on the optimal value of* (LP), *then* a *can be extended by a properly chosen* m*-dimensional vector* y *to a solution to* (D).

We see that the entity responsible for lower bounds on the optimal value of (LP) is the system (D): every solution to the latter system induces a bound of this type, and when (LP) is feasible, all lower bounds can be obtained from solutions to (D). Now note that if (y, a) is a solution to (D), then the pair $(y, b^T y)$ also is a solution to the same system, and the lower bound $b^T y$ on c^* is not worse than the lower bound a yielded by the former one. Thus, as far as lower bounds on c^* are concerned, we lose nothing by restricting ourselves to the solutions (y, a) of (D) with $a = b^T y$; the best lower bound on c^* given by (D) is therefore the optimal value of the problem

$$\max_y \left\{ b^T y \,\middle|\, A^T y = c,\ y \geq 0 \right\}, \tag{LP*}$$

which we call the problem *dual* to the *primal* problem (LP). Note that (LP*) is also an LP program.

All we know about the dual problem at the moment is the following.

Proposition 1.3.3. *Whenever* y *is a feasible solution to* (LP*), *the corresponding value of the dual objective* $b^T y$ *is a lower bound on the optimal value* c^* *in* (LP). *If* (LP) *is feasible, then for every* $a \leq c^*$ *there exists a feasible solution* y *of* (LP*) *with* $b^T y \geq a$.

1.3.3 Linear programming duality theorem

Proposition 1.3.3 is in fact equivalent to the following theorem.

THEOREM 1.3.2. Duality theorem in linear programming. *Consider an LP program*

$$\min_x \left\{ c^T x \;\middle|\; Ax \geq b \right\} \tag{LP}$$

along with its dual

$$\max_y \left\{ b^T y \;\middle|\; A^T y = c,\ y \geq 0 \right\}. \tag{LP*}$$

Then

1. *the duality is symmetric: the problem dual to dual is equivalent to the primal;*
2. *the value of the dual objective at every dual feasible solution is $\leq$ the value of the primal objective at every primal feasible solution;*
3. *the following properties are equivalent to each other:*
   (i) the primal is feasible and bounded below,
   (ii) the dual is feasible and bounded above,
   (iii) the primal is solvable,
   (iv) the dual is solvable,
   (v) both primal and dual are feasible.

Whenever (i) $\equiv$ (ii) $\equiv$ (iii) $\equiv$ (iv) $\equiv$ (v) *is the case, the optimal values of the primal and the dual problems are equal to each other.*

Proof. Item 1 is quite straightforward: writing the dual problem (LP*) in our standard form, we get

$$\min_y \left\{ -b^T y \;\middle|\; \begin{bmatrix} I_m \\ A^T \\ -A^T \end{bmatrix} y - \begin{pmatrix} 0 \\ c \\ -c \end{pmatrix} \geq 0 \right\},$$

where I_m is the m-dimensional unit matrix. Applying the duality transformation to the latter problem, we come to the problem

$$\max_{\xi,\eta,\zeta} \left\{ 0^T\xi + c^T\eta + (-c)^T\zeta \;\middle|\; \begin{array}{rcl} \xi & \geq & 0 \\ \eta & \geq & 0 \\ \zeta & \geq & 0 \\ \xi - A\eta + A\zeta & = & -b \end{array} \right\},$$

which is clearly equivalent to (LP) (set $x = \eta - \zeta$).

Point 2 is readily given by Proposition 1.3.3. The proof of point 3 is as follows:

(i)$\Rightarrow$(iv): If the primal is feasible and bounded below, its optimal value c^* (which of course is a lower bound on itself) can, by Proposition 1.3.3, be (non-strictly) majorized by a quantity $b^T y^*$, where y^* is a feasible solution to (LP*). In the situation in question, of course, $b^T y^* = c^*$ (by already proved item 2)); on the other hand, in view of the same Proposition 1.3.3, the optimal value in the dual is $\leq c^*$. We conclude that the optimal value in the dual is attained and is equal to the optimal value in the primal.

(iv)$\Rightarrow$(ii): This is evident.

(ii)$\Rightarrow$(iii): This implication, in view of the primal-dual symmetry, follows from the implication (i)$\Rightarrow$(iv).

(iii)$\Rightarrow$(i): This is evident.

We have seen that (i) ≡ (ii) ≡ (iii) ≡ (iv) and that the first (and consequently each) of these four equivalent properties implies that the optimal value in the primal problem is equal to the optimal value in the dual one. All that remains is to prove the equivalence between (i)–(iv), on one hand, and (v), on the other hand. This is immediate: (i)–(iv), of course, imply (v); conversely, in the case of (v) the primal not only is feasible but also is bounded below (this is an immediate consequence of the feasibility of the dual problem; see point 2), and (i) follows. □

An immediate corollary of the LP duality theorem is the following *necessary and sufficient* optimality condition in LP.

Theorem 1.3.3. Necessary and sufficient optimality conditions in linear programming. *Consider an LP program* (LP) *along with its dual* (LP*). *A pair* (x, y) *of primal and dual feasible solutions is made up of optimal solutions to the respective problems if and only if*

$$y_i[Ax - b]_i = 0, \ \ i = 1, \ldots, m, \qquad \text{[complementary slackness]}$$

likewise as if and only if

$$c^T x - b^T y = 0 \qquad \text{[zero duality gap].}$$

Indeed, the zero duality gap optimality condition is an immediate consequence of the fact that the value of primal objective at every primal feasible solution is greater than or equal to the value of the dual objective at every dual feasible solution, while the optimal values in the primal and the dual are equal to each other; see Theorem 1.3.2. The equivalence between the zero duality gap and the complementary slackness optimality conditions is given by the following computation: whenever x is primal feasible and y is dual feasible, the products $y_i[Ax - b]_i$, $i = 1, \ldots, m$, are nonnegative, while the sum of these products is precisely the duality gap:

$$y^T[Ax - b] = (A^T y)^T x - b^T y = c^T x - b^T y.$$

Thus, the duality gap can vanish at a primal-dual feasible pair (x, y) if and only if all products $y_i[Ax - b]_i$ for this pair are zeros.

1.3.4 Illustration: Problem dual to the Tschebyshev approximation problem

Let us look at the program dual to the (LP form of) the Tschebyshev approximation problem. Our primal LP program is

$$\min_{t,x} \left\{ t \,\middle|\, t - [b_i - a_i^T x] \geq 0, t - [-b_i + a_i^T x] \geq 0, \ \ i = 1, \ldots, M \right\}. \qquad \text{(P)}$$

Consequently, the dual problem is the LP program

$$\max_{\eta,\zeta}\left\{\sum_{i=1}^{M} b_i[\eta_i-\zeta_i]\;\middle|\;\begin{array}{rcl}\eta_i,\zeta_i & \geq & 0,\ i=1,\dots,M,\\ \sum_{i=1}^{M}[\eta_i+\zeta_i] & = & 1,\\ \sum_{i=1}^{N}[\eta_i-\zeta_i]a_i & = & 0.\end{array}\right\}.$$

To simplify the dual problem, let us pass from the variables η_i, ζ_i to the variables $p_i=\eta_i+\zeta_i$, $q_i=\eta_i-\zeta_i$. With respect to the new variables the problem becomes

$$\max_{p,q}\left\{\sum_{i=1}^{M} b_i q_i\;\middle|\;\begin{array}{rcl}p_i\pm q_i & \geq & 0,\ i=1,\dots,M,\\ \sum_{i=1}^{M} p_i & = & 1,\\ \sum_{i=1}^{M} q_i a_i & = & 0.\end{array}\right\}.$$

In the resulting problem one can easily eliminate the p-variables, thus coming to the problem

$$\max_{q}\left\{\sum_{i=1}^{M} b_i q_i\;\middle|\;\begin{array}{rcl}\sum_{i=1}^{M} q_i a_i & = & 0,\\ \sum_{i=1}^{M}|q_i| & \leq & 1.\end{array}\right\}. \tag{D}$$

The primal-dual pair (P)–(D) admits a nice geometric interpretation. Geometrically, the primal problem (P) is as follows:

> *Given a vector $b\in\mathbf{R}^M$ and a linear subspace L in $\mathbf{R}^M$ spanned by the columns of the matrix*
>
> $$\begin{bmatrix}a_1^T\\ \cdots\\ a_M^T\end{bmatrix},$$
>
> *find an element of L closest to b in the uniform norm*
>
> $$\|z\|_\infty=\max_{i=1,\dots,M}|z|_i$$
>
> *on $\mathbf{R}^M$.*

Observing that the equality constraints $\sum_{i=1}^{M} q_i a_i = 0$ in (D) say *exactly* that the M-dimensional vector q must be orthogonal to the columns of the matrix

$$\begin{bmatrix}a_1^T\\ \cdots\\ a_M^T\end{bmatrix}$$

or, which is the same, that the linear functional $z \mapsto q^T z$ vanishes on L, we see that the dual problem (D) is as follows:

> *Given the same data as in* (P)*, find a linear functional* $z \mapsto q^T z$ *on* $\mathbf{R}^M$ *of the* $\|\cdot\|_1$*-norm*
>
> $$\|q\|_1 = \sum_{i=1}^{M} |q_i|$$
>
> *not exceeding* 1*, which separates best of all the point* b *and the linear subspace* L*, i.e., which is identically* 0 *on* L *and is as large as possible at* b*.*

The duality theorem says, in particular, that the optimal values in (P) and in (D) are equal to each other; in other words,

> The $\|\cdot\|_\infty$-distance from a point $b \in \mathbf{R}^M$ to a linear subspace $L \subset \mathbf{R}^M$ is equal to the maximum quantity by which b can be separated from L by a linear functional of $\|\cdot\|_1$-norm 1.

This is the simplest case of a very general and useful statement (a version of the Hahn–Banach theorem):

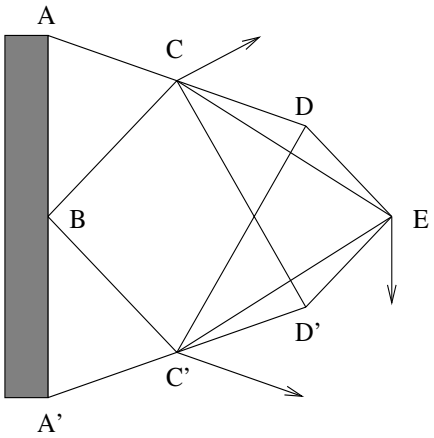
> The distance from a point b in a linear normed space $(E, \|\cdot\|)$ to a linear subspace $L \subset E$ is equal to the supremum of quantities by which b can be separated from L by a linear functional $q(b)$ of the conjugate to $\|\cdot\|$ norm at most 1:
>
> $$\inf\{\|b - x\| \mid x \in L\} = \sup\{q(b) \mid q(\cdot) : E \to \mathbf{R} \text{ is linear}, \|q\|_* \le 1\},$$
>
> $$\left[\|q\|_* = \sup\{q(x) \mid x \in E,\ \|x\| \le 1\}\right].$$

1.3.5 Application: Truss topology design

Surprisingly, LP in general, and the Tschebyshev approximation problem in particular, may serve to solve seemingly highly nonlinear optimization problems. One of the most interesting examples of this type is the truss topology design (TTD) problem.

Truss topology design

A truss is a mechanical construction comprising thin elastic bars linked to each other, such as an electric mast, a railroad bridge, or the Eiffel Tower. The points at which the bars are linked are called nodes. A truss can be subjected to an external load—a collection of simultaneous forces acting at the nodes, as shown on Fig. 1.5. Under a load, the construction deforms a bit, until the tensions caused by the deformation compensate the external forces. When deformed, the truss stores certain potential energy; this energy is called the compliance of the truss with respect to the load. The less the compliance, the more rigid the truss with respect to the load in question.



nodes: A,A′,B,C,C′,D,D′,E
bars: AC,A′C′,BC,BC′,CD,CD′,C′D′,C′D,CE,C′E,DE,D′E
forces: arrows

Figure 1.5. *A simple planar truss and a load.*

In the simplest TTD problem, we are given

- a nodal set, which is a finite set of points on the plane or in the space where the bars of the truss to be designed can be linked,
- boundary conditions specifying the nodes that are supported and cannot move (like nodes A,B,A′ on the wall AA′ in Fig. 1.5),
- a load, which is a collection of external forces acting at the nodes.

The goal is to design a truss of a given total weight best able to withstand the given load, i.e., to link some pairs of the nodes by bars of appropriate sizes, not exceeding a given total weight, in such a way that the compliance of the resulting truss with respect to the load of interest will be as small as possible.

An attractive feature of the TTD problem is that although it seems to deal with the size (weights) of the bars only, it finds the geometric shape (layout) of the truss as well. Indeed, we may start with a dense nodal grid and allow all pairs of nodes to be connected by bars. In the optimal truss, yielded by the optimization process, some of the bars (typically the majority of them) will get zero weights. In other words, the optimization problem will by itself decide which nodes to use and how to link them, i.e., it will find both the optimal pattern (topology) of the construction and the optimal sizing.

Derivation of the model

To pose the TTD problem as an optimization program, let us look in more detail at what happens with a truss under a load. Consider a particular bar AB in the unloaded truss (Fig. 1.6); after the load is applied, the nodes A and B move a little bit, as shown on Fig. 1.6.

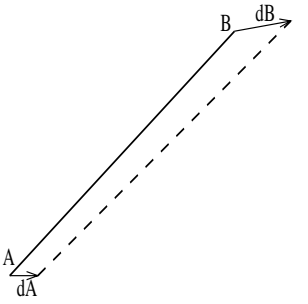


Figure 1.6. *A bar before (solid) and after (dashed) load is applied.*

Assuming the nodal displacements dA and dB to be small and neglecting the second order terms, the elongation dl of the bar under the load is the projection of the vector $dB - dA$ on the direction of the bar:

$$dl = (dB - dA)^T(B - A)/\|B - A\|.$$

The tension (the magnitude of the reaction force) caused by this elongation, by Hooke's law, is

$$\kappa \frac{dl \times S_{AB}}{\|B - A\|} = \kappa \frac{dl \times t_{AB}}{\|B - A\|^2},$$

where κ is a characteristic of the material (Young's modulus), S_{AB} is the cross-sectional area of the bar AB, and t_{AB} is the volume of the bar. Thus, the tension is

$$\tau = \kappa t_{AB}(dB - dA)^T(B - A)\|B - A\|^{-3}.$$

The reaction force at point B associated with the tension is the vector

$$\begin{aligned} -\tau(B - A)\|B - A\|^{-1} &= \kappa t_{AB}[(dB - dA)^T(B - A)](B - A)\|B - A\|^{-4} \\ &= -t_{AB}[(dB - dA)^T\beta_{AB}]\beta_{AB}, \\ &\quad \beta_{AB} = \sqrt{\kappa}(B - A)\|B - A\|^{-2}. \end{aligned} \tag{1.3.12}$$

Note that the vector β_{AB} depends on the positions of the nodes linked by the bar and is independent of the load and of the design.

Now let us look at the potential energy stored by our bar as a result of its elongation. Mechanics says that this energy is the half-product of the tension and the elongation, i.e., it is

$$\begin{aligned} \frac{\text{tension} \times \text{elongation}}{2} = \frac{\tau dl}{2} &= \frac{[\kappa t_{AB}(dB - dA)^T(B - A)\|B - A\|^{-3}][(dB - dA)^T(B - A)\|B - A\|^{-1}]}{2} \\ &= \frac{1}{2}t_{AB}\left[(dB - dA)^T\beta_{AB}\right]^2. \end{aligned} \tag{1.3.13}$$

Now it is easy to build the relevant model. Let M be the number of nodes in the nodal grid and M_{f} be the number of the "free" nodes—those that are not fixed by the boundary

conditions.[4] We define the space $\mathbf{R}^m$ of *virtual displacements* of the construction as the direct sum of the spaces of displacements of the free nodes; thus, m is either $2M_{\mathrm{f}}$ or $3M_{\mathrm{f}}$, depending on whether we are speaking about planar or spatial trusses. A vector v from $\mathbf{R}^m$ represents a displacement of the nodal grid: a free node ν corresponds to a pair (planar case) or a triple (spatial case) of (indices of) the coordinates of v, and the corresponding subvector $v[\nu]$ of v represents the "physical" two-dimensional (2D) or 3D displacement of the node ν. It is convenient to define the subvectors $v[\nu]$ for fixed nodes ν as well; by definition, these subvectors are zeros.

A load—a collection of external forces acting at the free nodes[5]—can be represented by a vector $f \in \mathbf{R}^m$; for every free node ν, the corresponding subvector $f[\nu]$ of f is the external force acting at ν.

Let n be the number of tentative bars (i.e., pair connections between distinct nodes from the grid, at least one node in the pair being free). Let us somehow order all our n tentative bars and consider the ith of them. This bar links two nodes $\nu'(i)$, $\nu''(i)$, i.e., two points A_i and B_i from our physical space (which is the 2D plane in the planar case and the 3D space in the spatial case). Let us associate with tentative bar i a vector $b_i \in \mathbf{R}^m$ as follows (cf. (1.3.12)):

$$b_i[\nu] = \begin{cases} \beta_{A_iB_i}, & \nu = \nu''(i) \text{ and } \nu \text{ is free,} \\ -\beta_{A_iB_i}, & \nu = \nu'(i) \text{ and } \nu \text{ is free,} \\ 0 & \text{in all remaining cases.} \end{cases} \tag{1.3.14}$$

A particular truss can be identified with a nonnegative vector $t = (t_1, \ldots, t_n)$, where t_i is the volume of bar i in the truss. Consider a truss t, and let us look at the reaction forces caused by a displacement v of the nodes of the truss. From (1.3.12) and (1.3.14) it follows that for every free node ν, the component of the reaction force caused, under the displacement v, by the ith bar at the node ν is $-t_i(b_i^T v)b_i[\nu]$. Consequently, the total reaction force at the node ν is

$$-\sum_{i=1}^{n} t_i(b_i^T v)b_i[\nu],$$

and the collection of the reaction forces at the nodes is

$$-\sum_{i=1}^{n} t_i(b_i^T v)b_i = -\left[\sum_{i=1}^{n} t_i b_i b_i^T\right] v.$$

We see that the m-dimensional vector representing the reaction forces caused by a displacement v depends on v linearly:

$$f_{\mathrm{r}} = -A(t)v,$$

[4] We assume for simplicity that a node is either completely fixed or completely free. In some TTD problems it makes sense to speak also of *partially fixed* nodes, which can move along a given line (or along a given 2D plane in the 3D space in the case of spatial trusses). It turns out that the presence of partially fixed nodes does not change the mathematical structure of the resulting optimization problem.

[5] It makes no sense to speak about external force acting at a fixed node. Such a force will be compensated by the physical support that makes the node fixed.

where

$$A(t) = \sum_{i=1}^{n} t_i b_i b_i^T \tag{1.3.15}$$

is the so called *bar-stiffness matrix* of the truss. This is an $m \times m$ symmetric matrix, which depends linearly on the design variables—the volumes of tentative bars.

Now, at equilibrium the reaction forces must compensate the external ones, which gives us a system of linear equations determining the displacement of the truss under an external load f:

$$A(t)v = f. \tag{1.3.16}$$

To complete the model, we should also write an expression for the compliance—the potential energy stored by the truss at equilibrium. According to (1.3.13)–(1.3.14), this energy is

$$\begin{aligned} \tfrac{1}{2}\sum_{i=1}^{n} t_i \left[(v[\nu''(i)] - v[\nu'(i)])^T \beta_{A_i B_i}\right]^2 &= \tfrac{1}{2}\sum_{i=1}^{n} t_i (v^T b_i)^2 \\ &= \tfrac{1}{2} v^T \left[\sum_{i=1}^{n} t_i b_i b_i^T\right] v \\ &= \tfrac{1}{2} v^T A(t) v \\ &= \tfrac{1}{2} f^T v, \end{aligned}$$

the concluding equality being given by (1.3.16). Thus, the compliance of a truss t under a load f is

$$\mathrm{Compl}_f(t) = \frac{1}{2} f^T v, \tag{1.3.17}$$

where v is the corresponding displacement; see (1.3.16).

The expression for the compliance possesses a transparent mechanical meaning:

> *The compliance is just one half of the mechanical work performed by the external load on the displacement of the truss until equilibrium.*

Remark 1.3.1. Mathematically speaking, there is a gap in the above considerations: the linear system (1.3.16) can have more than one solution v (or no solution at all). Why do we know that in the former case the value of the right-hand side of (1.3.17) is independent of the particular choice of the solution to the equilibrium equation? And what do we do if (1.3.16) has no solution?

The answers to these questions are as follows. If (1.3.16) has no solution, that means that the truss t cannot carry the load f: it is crushed by this load. In this case it makes sense to define $\mathrm{Compl}_f(t)$ as $+\infty$. If (1.3.16) is solvable, then the quantity $f^T v$ does not depend on a particular choice of a solution to the equation. Indeed, if v solves (1.3.16), then

$$f = \sum_{i=1}^{n} t_i b_i (b_i^T v) \Rightarrow f^T v = \sum_{i=1}^{n} t_i (b_i^T v)^2.$$

The resulting quantity is independent of a particular choice of v due to the following observation:

> *If $t \geq 0$, then for every i for which $t_i > 0$, the quantity $b_i^T v$ does not depend on a particular choice of a solution to* (1.3.16).

Indeed, if v, v' are solutions to (1.3.16), then

$$\begin{aligned} \sum_{i=1}^{n} t_i b_i (b_i^T [v - v']) &= 0 \Rightarrow \\ [v - v']^T \sum_{i=1}^{n} t_i b_i (b_i^T [v - v']) &= 0 \Rightarrow \\ \sum_{i:t_i>0} t_i \left(b_i^T [v - v']\right)^2 &= 0. \end{aligned}$$

Now we can formulate the problem of designing the stiffest truss (with respect to a given load) of a given weight as the following optimization problem.

Problem 1.3.1. The simplest TTD problem. *Given a ground structure*[6]

$$m; n; \{b_i \in \mathbf{R}^m\}_{i=1}^n,$$

a load $f \in \mathbf{R}^m$, and a total bar volume $w > 0$, find a truss $t = (t_1, \ldots, t_n)$ with nonnegative t_i satisfying the resource constraint

$$\sum_{i=1}^{n} t_i \leq w \tag{1.3.18}$$

with the minimum possible compliance $\mathrm{Compl}_f(t)$ *with respect to the load f.*

When speaking about the TTD problem, we always make the following assumption.

Assumption 1.3.1. *The vectors $\{b_i\}_{i=1}^n$ span the entire $\mathbf{R}^m$.*

This assumption has a very transparent mechanical interpretation. Let us look at a full truss—one where all tentative bars are of positive weights. Assumption 1.3.1 says that there should be no nonzero displacement v orthogonal to all b_i, or, which is the same, an arbitrary nonzero displacement should cause nonzero tensions in some bars of our full truss. In other

[6] From the engineering viewpoint, a ground structure is the data of a particular TTD problem, i.e., a particular nodal set along with its partition into fixed and free nodes, the Young modulus of the material, etc. The engineering data define, as explained above, the mathematical data of the TTD problem.

words, the assumption says that our boundary conditions forbid rigid body motions of the nodal set.

Linear programming form of the truss topology design problem

As stated above, the TTD problem, Problem 1.3.1, does not resemble an LP program at all: although the constraints

$$t \geq 0, \ \sum_{i=1}^{n} t_i \leq w$$

are linear, the objective $\mathrm{Compl}_f(t)$ given by (1.3.15)–(1.3.17) is highly nonlinear.

Surprisingly, the TTD problem can be converted into an LP program. The corresponding transformation is as follows.

For a loaded truss, a *stress* in a bar is the absolute value of the corresponding tension (i.e., the magnitude of the reaction force caused by bar's deformation) divided by the cross-sectional area of the bar; the larger this quantity, the worse the conditions the material is working in. According to (1.3.12), (1.3.14), the stress in bar i is (up to the constant factor $\sqrt{\kappa}$) a simple function of the displacement vector v:

$$s_i = |b_i^T v|. \tag{1.3.19}$$

Now let us formulate the following intermediate problem.

Problem 1.3.2. *Given a ground structure $m, n, \{b_i\}_{i=1}^n$ and a load f, find a displacement v that maximizes the work $f^T v$ of the load under the constraint that all stresses are ≤ 1:*

$$\min_v \left\{ f^T v \mid |b_i^T v| \leq 1, \ i = 1, \ldots, n \right\}. \tag{1.3.20}$$

A derivation completely similar to the one in section 1.3.4 demonstrates that the problem dual to (1.3.20) is (equivalent to) the program

$$\min_{q_1, \ldots, q_n} \left\{ \sum_{i=1}^{n} |q_i| \mid \sum_{i=1}^{n} q_i b_i = f \right\}. \tag{1.3.21}$$

Note that both the primal and the dual are feasible (for the primal it is evident; the feasibility of the dual follows from Assumption 1.3.1). By the LP duality theorem, both problems are solvable with common optimal value w_*. Let v^* be an optimal solution to (1.3.20) and let q^* be an optimal solution to (1.3.21). It is easy to see that the complementary slackness optimality condition results in

$$|q_i^*| = q_i^* (b_i^T v^*), \ i = 1, \ldots, n. \tag{1.3.22}$$

Assuming $f \neq 0$ (this is the only case of interest in the TTD problem), we ensure that $w_* = \sum_{i=1}^n |q_i^*| > 0$, so that the vector

$$t^* : \quad t_i^* = \frac{w}{w_*} |q_i^*|, \ i = 1, \ldots, n, \tag{1.3.23}$$

(w is the material resource in the TTD problem) is well defined and is a feasible truss (i.e., t^* is nonnegative and satisfies the resource constraint). We claim that

> ($*$) *The vector t^* is an optimal solution to the TTD problem, and $v^+ = \frac{w_*}{w} v^*$ is the corresponding displacement.*

Indeed, we have

$$\begin{aligned}
\sum_{i=1}^n t_i^* b_i b_i^T v^+ &= \sum_{i=1}^n |q_i^*| b_i (b_i^T v^*) && \text{[by construction of } t^*, v^+], \\
&= \sum_{i=1}^n q_i^* b_i && \text{[by (1.3.22)]}, \\
&= f && \text{[see (1.3.21)]}
\end{aligned}$$

so that v^+ is the displacement of the truss t^* under the load f. The corresponding compliance is

$$\begin{aligned}
\mathrm{Compl}_f(t^*) &= \frac{1}{2} f^T v^+ \\
&= \frac{1}{2} \sum_{i=1}^n q_i^* b_i^T v^+ \qquad \left[\text{since } f = \sum_i q_i^* b_i \right] \\
&= \frac{1}{2} \frac{w_*}{w} \sum_{i=1}^n q_i^* b_i^T v^* \qquad (1.3.24) \\
&= \frac{w_*}{2w} \sum_{i=1}^n |q_i^*| \qquad \text{[see (1.3.22)]} \\
&= \frac{w_*^2}{2w}.
\end{aligned}$$

Thus, t^* is a feasible solution to the TTD problem with the value of the objective $\frac{w_*^2}{2w}$. To prove that the solution is optimal, it suffices to demonstrate that the latter quantity is a lower bound on the optimal value in the TTD problem. To see this, let t be a feasible solution to the TTD problem and let v be the corresponding displacement. Let also

$$q_i = t_i (b_i^T v).$$

We have

$$\sum_{i=1}^n q_i b_i = \sum_{i=1}^n t_i (b_i^T v) b_i = f \qquad (1.3.25)$$

(the equilibrium equation; see (1.3.15)–(1.3.16)). Thus, q is a feasible solution to (1.3.21), and

$$\begin{aligned}
\mathrm{Compl}_f(t) &= \tfrac{1}{2} f^T v \\
&= \tfrac{1}{2}\sum_{i=1}^{n} t_i (b_i^T v)^2 \qquad \text{[see (1.3.25)]} \\
&= \tfrac{1}{2}\sum_{i:t_i\neq 0} \tfrac{q_i^2}{t_i} \\
&\geq \tfrac{1}{2}\left[\sum_{i:t_i\neq 0} |q_i|\right]^2 \left[\sum_{i:t_i\neq 0} t_i\right]^{-1}
\end{aligned}$$

since by the Cauchy inequality

$$\left(\sum_{i:t_i\neq 0} |q_i|\right)^2 = \left(\sum_{i:t_i\neq 0} [t_i^{-1/2}|q_i|]t_i^{1/2}\right)^2 \leq \left(\sum_{i:t_i\neq 0} q_i^2/t_i\right)\left(\sum_i t_i\right)$$
$$\geq \tfrac{1}{2}\tfrac{w_*^2}{w},$$

Note that (*) not only provides us with a possibility to compute an optimal solution to the TTD problem via LP techniques but also establishes a very important fact:

(**) *In an optimal truss t^* the stresses, caused by load f, in all bars of nonzero weight are equal to each other, so that the material in all bars is under the same working conditions.*

As stated by (*), the displacement of t^* under the load f is v^+, i.e., it is proportional to v^*, and it remains to look at (1.3.22) and (1.3.23).

Strictly speaking, the above reasoning is incomplete. First, v^+ is *a* solution to the equilibrium equation associated with t^*; how do we know that (**) holds true for other solutions to this equation? The answer: the stresses in those bars that are actually present in a truss are uniquely defined by the truss and the load; see Remark 1.3.1. Second, (**) is established for *an* optimal solution t^* to the TTD problem, one which can be obtained, in the aforementioned fashion, from an optimal solution to (1.3.25). A priori it may happen that the TTD problem has other optimal solutions. However, it can be shown that every optimal solution to the TTD problem can be obtained from an optimal solution to (1.3.25).

Remark 1.3.2. Note that problem (1.3.21) is, basically, the Tschebyshev approximation problem. Indeed, instead of asking what is the largest possible value of $f^T v$ under the constraints $|b_i^T v| \leq 1, i = 1, \ldots, n$, we might ask what is the minimum value of $\max_i |b_i^T v|$ under the constraint that $f^T v = 1$. The optimal solutions to these two problems can be easily obtained from each other. The second problem is equivalent to a Tschebyshev approximation problem: we can use the equation $f^T v = 1$ to eliminate one of the design variables, thus coming to a Tschebyshev problem in the remaining variables.

1.4 Exercises to Lecture 1

1.4.1 Uniform approximation

We have seen that the Tschebyshev approximation problem normally arises as a discrete version of the best uniform approximation problem:

Given a segment $\Delta = [a, b]$, N basic functions $f_1, \ldots, f_N$, and a target function f_0 on Δ, find the best uniform approximation of f_0 by a linear combination of $f_1, \ldots, f_N$:

$$\min_{x\in\mathbf{R}^n}\left\{\left\|f_0-\sum_{j=1}^N x_j f_j\right\|_\infty \equiv \sup_{t\in\Delta}\left|f_0(t)-\sum_{j=1}^N x_j f_j(t)\right|\right\}. \qquad \text{(Appr}(\Delta)\text{)}$$

The discrete version of this problem is obtained by replacing Δ with a finite set $T \subset \Delta$:

$$\min_{x\in\mathbf{R}^n}\left\{\left\|f_0-\sum_{j=1}^N x_j f_j\right\|_{T,\infty} = \sup_{t\in T}\left|f_0(t)-\sum_{j=1}^N x_j f_j(t)\right|\right\}. \qquad \text{(Appr}(T)\text{)}$$

The following questions related to the above problems are of primary interest:

1. What is the quality of approximation of (Appr(Δ)) by (Appr(T))? Specifically, can we write down an inequality

$$\left\|f_0-\sum_{j=1}^N x_i f_i\right\|_\infty \le \kappa\left\|f_0-\sum_{j=1}^N x_i f_i\right\|_{T,\infty} \qquad \forall x \qquad (1.4.26)$$

with appropriately chosen κ? If it is the case, then κ can be viewed as a natural measure of the quality of approximating the original problem by its discrete version—the closer κ is to 1, the better the quality.

2. Given the total number M of points in the finite set T, how should we choose these points to get the best possible quality of approximation?

The goal of the subsequent series of problems is to provide some answers to these two questions. The answers will be given in terms of properties of the functions from the *linear space L spanned by* $f_0, f_1, \ldots, f_N$:

$$L = \left\{f = \sum_{j=0}^N \xi_j f_j\right\}_{\xi\in\mathbf{R}^{N+1}}.$$

Given a finite set $T \subset \Delta$, let us say that T is *L-dense*, if there exists $\kappa < \infty$ such that

$$\|f\|_\infty \le \kappa\|f\|_{T,\infty} \quad \forall f \in L;$$

the smallest κ with the latter property is denoted by $\kappa_L(T)$. If T is not L-dense, we set $\kappa_L(T) = \infty$. Note that $\kappa_L(T)$ majorates the quality of approximating the problem (Appr(Δ)) by (Appr(T)), and this is the quantity we shall focus on.

Exercise 1.1. *Let L be a finite-dimensional space comprising continuous functions on a segment Δ, and let T be a finite subset in Δ. Prove that T is L-dense if and only if the only function from L that vanishes on T is $\equiv 0$.*

Exercise 1.2. *Let $\alpha < \infty$, and assume L is α-regular, i.e., the functions from L are continuously differentiable and*

$$\|f'\|_\infty \le \alpha \|f\|_\infty \quad \forall f \in L.$$

Assume that $T \subset \Delta$ is such that the distance from every point in Δ to the closest point of T does not exceed $\beta < \alpha^{-1}$. Prove that under these assumptions

$$\kappa_L(T) \le \frac{1}{1-\alpha\beta}.$$

Exercise 1.3. *Let L be a k-dimensional linear space comprising continuously differentiable functions on a segment Δ. Prove that L is α-regular for some α; consequently, by choosing a fine-enough finite grid $T \subset \Delta$, we can ensure a given quality of approximating* (Appr(Δ)) *by* (Appr(T)).

To use the simple result stated in Exercise 1.2, we should know something about regular linear spaces L of functions. The most useful result of this type is the following fundamental fact.

Theorem 1.4.1. Bernshtein's theorem on trigonometric polynomials. *Let $\Delta = [0, 2\pi]$ and let f be a trigonometric polynomial of degree $\le k$ on Δ:*

$$f(t) = a_0 + \sum_{l=1}^{k} [a_0 \cos(lt) + b_0 \sin(lt)]$$

with real or complex coefficients. Then

$$\|f'\|_\infty \le k \|f\|_\infty.$$

Note that the inequality stated in the Bernshtein theorem is tight. Indeed, for the trigonometric polynomial $f(t) = \cos(kt)$ the inequality becomes equality.

We see that the space of trigonometric polynomials of degree $\le k$ on $[0, 2\pi]$ is k-regular. What about the space of *algebraic* polynomials of degree $\le k$ on the segment, say, $[-1, 1]$? Specifically, let us look at the *Tschebyshev polynomial* of degree k given on $\Delta = [-1, 1]$ by the relation

$$T_k(t) = \cos(k \operatorname{acos}(t)), \quad -1 \le t \le 1.$$

(Check that this indeed is a polynomial in t of degree k.) This polynomial possesses the following property:

(A) $\|T_k\|_\infty = 1$, *and there are $k+1$ points of alternance $t_\ell = \cos(\frac{\pi(k-\ell)}{k}) \in \Delta$, $\ell = 0, \dots, k$, where $|T_k(t_\ell)| = 1$ and the signs of the values $T_k(t_\ell)$ alternate* (*see Fig.* 1.7).

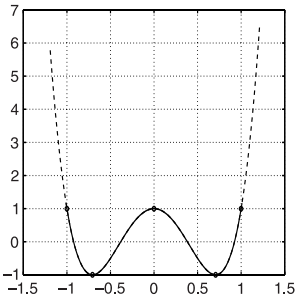


Figure 1.7. *The Tschebyshev polynomial T_4 and its five points of alternance.*

The derivative of T_k at the point $t = 1$ is k^2; thus, the factor α in an inequality

$$\|T_k'\|_\infty \leq \alpha \|T_k\|_\infty$$

is at least k^2. We conclude that the space L_k of real algebraic polynomials of degree $\leq k$ on the segment $[-1, 1]$ is *not* α-regular for $\alpha < k^2$. Is our space k^2-regular? We guess that the answer is positive, but we were too lazy to find out if this is true. What we intend to demonstrate here is that L_k is $2k^2$-regular.

Exercise 1.4. *Prove that if $f \in L_k$ and $\|f\|_\infty = 1$, then*

$$|f'(1)| \leq k^2 = T_k'(1). \qquad (*)$$

Hint. Assuming that $f'(1) > T_k'(1)$, look at the polynomial

$$p(t) = T_k(t) - \frac{T_k'(1)}{f'(1)} f(t).$$

Verify that the values of this polynomial at the points of alternance of T_k are of the same alternating signs as those of the values of T_k, so that p has at least k distinct zeros on $(-1, 1)$. Taking into account the latter fact and the equality $p'(1) = 0$, count the number of zeros of $p'(t)$.

Derive from (*) *that*

$$|f'(t)| \leq 2k^2$$

for all $t \in [-1, 1]$ and conclude that L_k is $2k^2$-regular.

Now let us apply the information collected so far to investigating questions 1 and 2 in the two simplest cases, where L comprises the trigonometric and the algebraic polynomials, respectively.

Exercise 1.5. *Assume that $\Delta = [0, 2\pi]$, and let L be a linear space of functions on Δ comprising all trigonometric polynomials of degree $\le k$. Let T be an equidistant M-point grid on Δ:*

$$T = \left\{ \frac{(2\ell+1)\pi}{M} \right\}_{\ell=0}^{M-1}.$$

1. *Prove that if $M > k\pi$, then T is L-dense, with*

$$\kappa_L(T) \le \frac{M}{M - k\pi}.$$

2. *Prove that the above inequality remains valid if we replace T with its arbitrary shift modulo 2π, i.e., treat Δ as the unit circumference and rotate T by an angle.*

3. *Prove that if T is an arbitrary M-point subset of Δ with $M \le k$, then $\kappa_L(T) = \infty$.*

Exercise 1.6. *Let $\Delta = [-1, 1]$ and let L be the space of all algebraic polynomials of degree $\le k$.*

1. *Assume that $2M > \pi k$ and T is the M-point set on Δ as follows:*

$$T = \left\{ t_l = \cos\left(\frac{(2\ell+1)\pi}{2M} \right) \right\}_{\ell=0}^{M-1}.$$

Then T is L-dense, with

$$\kappa_L(T) \le \frac{2M}{2M - \pi k}.$$

2. *Let T be an M-point set on Δ with $M \le k$. Then $\kappa_L(T) = \infty$.*

Exercise 1.7. The result stated in Exercise 1.6 says that when L_k is made up of all real algebraic polynomials of degree $\le k$ on $[-1, 1]$ and we want to ensure $\kappa_L(T) = O(1)$, then it suffices to take $M \equiv \mathrm{card}(T) = O(k)$ point grid. Note that the corresponding grid is *nonuniform*. Is it possible to achieve similar results with a *uniform* grid? The answer is no:

Prove that for the equidistant M-point grid $T = \left\{-1 + \frac{2\ell}{M}\right\}_{\ell=0}^{M}$ on $\Delta = [-1, 1]$ one has

$$\kappa_{L_k}(T) \ge c_1 M^{-1} \exp\{-c_2 k/\sqrt{M}\}$$

for some positive absolute constants c_1, c_2. Thus, in order to get $\kappa_{L_k}(T) = O(1)$ for an equidistant grid T, the cardinality of the grid should be nearly quadratic in k.

Hint. Let $t_1 = -1, t_2 = -1 + \frac{2}{M}, \ldots, t_M = +1$ be the points of T. Reduce the question to the following:

> Given a polynomial $f(t)$ of degree k which is ≤ 1 in absolute value on $[-1, t_{M-1}]$ and is equal to 0 at the point 1, how large could the polynomial be at the point $0.5(t_{M-1} + 1)$?

To answer this, look how the Tschebyshev polynomial T_k grows outside the segment $[-1, 1]$. (Note that for $t \ge 1$ the polynomial is given by $T_k(t) = \cosh(k \operatorname{acosh}(t))$.)

1.4.2 Theorem on the alternative

The goal of the subsequent exercises is to prove the general theorem on the alternative (Theorem 1.3.1).

From the homogeneous Farkas lemma to the theorem on the alternative

Consider the very particular case of the theorem on the alternative. We want to verify whether the system of homogeneous linear inequalities in $\mathbf{R}^n$ of the form

$$\begin{array}{rcl} a^T x & < & 0, \\ a_i^T x & \geq & 0, \ i = 1, \ldots, m, \end{array} \tag{F}$$

has no solutions. The answer is given by the following.

Lemma 1.4.1. The homogeneous Farkas lemma.

$$\{\text{(F) } \textit{is infeasible}\} \Leftrightarrow \left\{ \exists \lambda \geq 0 : a = \sum_{i=1}^{m} \lambda_i a_i \right\}.$$

Exercise 1.8. *Prove that Lemma* 1.4.1 *is exactly what is said by the theorem on the alternative as applied to the particular system* (F).

Our plan of attack is as follows. We shall demonstrate that the general theorem on the alternative can be easily obtained from the homogeneous Farkas lemma, and in section 1.4.3 we shall present a direct proof of the lemma. Thus, for the time being you may take Lemma 1.4.1 for granted.

Exercise 1.9. *Consider the same system of linear inequalities as in the theorem on the alternative:*

$$(\mathcal{S}): \qquad \left\{ \begin{array}{rcl} a_i^T x & > & b_i, \ i = 1, \ldots, m_{\mathrm{s}}, \\ a_i^T x & \geq & b_i, \ i = m_{\mathrm{s}} + 1, \ldots, m. \end{array} \right.$$

Prove that this system has no solution if and only if this is the case for the following homogeneous system:

$$(\mathcal{S}^*): \qquad \left\{ \begin{array}{rcl} -s & < & 0, \\ t - s & \geq & 0, \\ a_i^T x - b_i t - s & \geq & 0, \ i = 1, \ldots, m_{\mathrm{s}}, \\ a_i^T x - b_i t & \geq & 0, \ i = m_{\mathrm{s}} + 1, \ldots, m, \end{array} \right.$$

where the unknowns are x and two additional real variables s, t.

Derive from the above observation and the homogeneous Farkas lemma the general theorem on the alternative.

The next exercise presents several useful consequences of the general theorem on the alternative.

Exercise 1.10. *Prove the following statements.*

1. (Gordan's theorem on the alternative.) *One of the inequality systems*

$$Ax < 0,\ x \in \mathbf{R}^n, \qquad [A : m \times n]$$

$$A^T y = 0,\ 0 \neq y \geq 0,\ y \in \mathbf{R}^m,$$

has a solution if and only if the other one has no solution.

2. (Inhomogeneous Farkas lemma.) *A linear inequality*

$$a^T x \leq p \tag{1.4.27}$$

is a consequence of a solvable system of inequalities

$$Ax \leq b \qquad [A : m \times n]$$

if and only if

$$a = A^T v$$

for some nonnegative vector v such that

$$v^T b \leq p.$$

3. (Motzkin's theorem on the alternative.) *The system*

$$Sx < 0,\ \ Nx \leq 0 \qquad [S : p \times n,\ N : q \times n]$$

has no solution if and only if the system

$$S^T \sigma + N^T v = 0,\ \sigma \geq 0,\ \ v \geq 0,\ \ \sigma \neq 0,$$

has a solution.

1.4.3 Proof of the homogeneous Farkas lemma

Here we present a series of exercises aimed at proving the homogeneous Farkas lemma. In fact we present two proofs: a quick and dirty one based on separation arguments, and a more intelligent proof based on the Helley theorem. Note that the homogeneous Farkas lemma states that the system (F) has no solution if and only if a is a linear combination, with nonnegative coefficients, of $a_1, \ldots, a_m$. The only nontrivial here is the "only if" part: if the system has no solutions, then $a = \sum_i \lambda_i a_i$, $\lambda_i \geq 0$, and later we focus on this "only if" part.

From the separation theorem to the Farkas lemma

Exercise 1.11. *Let K be a nonempty closed convex set in $\mathbf{R}^n$ and let $x \in \mathbf{R}^n$ be a point not belonging to K.*

1. *Prove that the distance from x to K is achieved: there exists $x^* \in K$ such that*

$$\|x - x^*\|_2 = \min_{y \in K} \|x - y\|_2.$$

Moreover, x^* is unique.

2. *Prove that $e = x - x^*$ strictly separates x and K, namely,*

$$e^T(x - y) \geq \|e\|_2^2 > 0 \quad \forall y \in K.$$

3. *Assume in addition that K is a cone, i.e.,*

$$y \in K, \lambda \geq 0 \Rightarrow \lambda y \in K.$$

Prove that in this case $e = x - x^*$ satisfies the relations

$$e^T x > 0 \quad \& \quad e^T y \leq 0 \quad \forall y \in K.$$

Exercise 1.12. *Let $a_1, \ldots, a_m \in \mathbf{R}^n$. Consider the conic hull of these vectors, i.e., the set K of all their linear combinations with nonnegative coefficients:*

$$K = \left\{ p = \sum_{i=1}^{m} \lambda_i a_i \mid \lambda \geq 0 \right\}.$$

1. *Prove that the set K is a convex cone.*
2. *Prove that the set K is closed.*

Hint. Let $p_j = \sum_{i=1}^{m} \lambda_{ij} a_i$ with $\lambda_{ij} \geq 0$, and let the sequence $\{p_j\}$ converge to a point p; we should prove that p can also be represented as a linear combination, with nonnegative coefficients, of $a_1, \ldots, a_m$.

A. Assume that $p \neq 0$ (this is the only nontrivial case). For every j with $p_j \neq 0$ consider a *minimal* representation of p_j as a nonnegative linear combination of $a_1, \ldots, a_m$, i.e., a representation $p_j = \sum_i \lambda_{ij} a_i$, $\lambda_{ij} \geq 0$, with the least number of positive weights λ_{ij}'s. Prove that the a_i's participating in (any) minimal representation of p_j with positive weights are linearly independent.

B. Derive from A that the weight vectors $(\lambda_{1j}, \lambda_{2j}, \ldots, \lambda_{mj})$ associated with minimal representations of p_j's form a bounded sequence.

C. Derive from B that $p \in K$.

3. *Given* 1, 2, *and the results of the previous exercise, demonstrate that for any vector $a \notin K$ there exists a vector x such that*

$$a^T x < 0, a_i^T x \geq 0, \ i = 1, \ldots, m,$$

and derive from this fact the homogeneous Farkas lemma.

Hint. Use as x the negation of the vector that separates a from K.

An intelligent proof

We start with the following basic fact.

THEOREM 1.4.2. Helley theorem. *Let $A_1, \dots, A_M$ be a collection of convex sets in $\mathbf{R}^n$. Assume that the intersection of every $k \le n+1$ sets from the collection is nonempty. Then the intersection of all M sets $A_1, \dots, A_M$ is nonempty.*

Let us derive the homogeneous Farkas lemma from the Helley theorem. Let $a, a_1, \dots, a_m$ be vectors in $\mathbf{R}^n$ such that the system of inequalities

$$\begin{array}{rcl} a^T x & < & 0, \\ a_i^T x & \ge & 0, \ i = 1, \dots, m, \end{array} \tag{F}$$

has no solution. We should prove that under this assumption a can be represented as a linear combination, with nonnegative coefficients, of $a_1, \dots, a_m$; this is exactly what is said by the homogeneous Farkas lemma. The statement is evident when $a = 0$, so that from now on we assume $a \neq 0$.

Set

$$\begin{array}{rcl} \Pi & = & \{x \mid a^T x = -1\}, \\ A_i & = & \{x \in \Pi \mid a_i^T x \ge 0\} \\ & = & \{x \mid a^T x = -1, a_i^T x \ge 0\}. \end{array}$$

Let us call a nonempty subcollection of the collection $\{a_1, \dots, a_m\}$ a contradiction if the sets A_i corresponding to the vectors from the subcollection have no common point.

EXERCISE 1.13. 1. *Prove that the entire collection $\{a_1, \dots, a_m\}$ is a contradiction.*

According to 1, contradictions exist. Consequently, there exists a minimal contradiction (one with the smallest number of vectors). By reordering the vectors a_i, we may assume that $\{a_1, \dots, a_k\}$ is a minimal contradiction.

2. *Prove that the vector a belongs to the linear span of the vectors $a_1, \dots, a_k$.*

Hint. Assuming that a does not belong to the linear span of $a_1, \dots, a_k$, prove that there exists a vector x that is orthogonal to $a_1, \dots, a_k$ and is not orthogonal to a, and conclude that $\{a_1, \dots, a_k\}$ is not a contradiction.

3. *Prove that the vectors $a_1, \dots, a_k$ are linearly independent.*

Hint. Assuming that $a_1, \dots, a_k$ are linearly dependent, consider the linear space L spanned by $a_1, \dots, a_k$ along with its subsets

$$\begin{array}{rcl} \Pi' & = & \{x \in L : a^T x = -1\}, \\ A_i' & = & \{x \in \Pi' : a_i^T x \ge 0\}, \ i = 1, \dots, k. \end{array}$$

3.1. Consider a subcollection of the collection $A_1', \dots, A_k'$. Prove that the sets of this subcollection have a point in common if and only if the corresponding sets of the collection $A_1, \dots A_k$ have a point in common.

3.2. Taking into account that $\{a_1, \ldots, a_k\}$ is a minimal contradiction, verify that every $k-1$ sets from the collection $A'_1, \ldots, A'_k$ have a point in common.

Applying the Helley theorem to the sets $A'_1, \ldots, A'_k$ (they are convex subsets of Π', i.e., essentially, of a $(\dim(L)-1)$-dimensional linear space), prove that under the assumption $\dim(L) < k$ the sets $A'_1, \ldots, A'_k$ have a point in common, which is impossible (since $\{a_1, \ldots, a_k\}$ is a contradiction).

4. *Derive from* 2, 3, *and the fact that* $\{a_1, \ldots, a_k\}$ *is a contradiction that* a *is a linear combination of* $a_1, \ldots, a_k$*, and all coefficients in this combination are nonnegative, thus concluding the proof of the homogeneous Farkas lemma.*

It is time to explain why the proof of the homogeneous Farkas lemma sketched in Exercise 1.13 is more intelligent than the proof coming from the separation scheme (Exercises 1.11, 1.12). The reason is that the Helley theorem itself, as well as the reasoning outlined in Exercise 1.13, are *purely algebraic* facts: they do not use compactness or other topological arguments, as does the reasoning in Exercise 1.11. As a result, the proof sketched in Exercise 1.13 remains valid also in the case when we replace our universe $\mathbf{R}^n$, with, say, the linear space $\mathbf{Q}^n$ of all n-dimensional vectors with rational coefficients.[7] From this observation we conclude that the theorem on the alternative remains valid when we speak about rational solutions to systems of linear inequalities with rational coefficients. This is a nontrivial and useful observation. (It implies, e.g., that a solvable LP program with rational data has a rational solution.)

Note that the proof via the separation heavily exploits the compactness and does not work at all in the case of a percolated space such as $\mathbf{Q}^n$. Consider, e.g., the rational plane $\mathbf{Q}^2$ along with the convex cone

$$K = \{(u, v) \in \mathbf{Q}^2 \mid u + \sqrt{2}v \le 0\}.$$

A point x from $\mathbf{Q}^2$ not belonging to K cannot be separated from K by a legitimate linear functional on $\mathbf{Q}^2$—there is no rational vector e such that $e^T x > e^T y \; \forall y \in K$. Consequently, in the case of the rational universe an attempt to prove the Farkas lemma via a separation-type reasoning fails at the very first step.

1.4.4 Helley theorem

The goal of the subsequent exercises is to establish the Helley theorem and to illustrate some of its applications.

Exercise 1.14. *Prove the following.*

Theorem 1.4.3. Radon. *Let* $a_1, \ldots, a_m$ *be a collection of* $m \ge n+2$ *vectors in* $\mathbf{R}^n$*. There exists a partitioning of the index set* $\{1, \ldots, m\}$ *into two nonempty disjoint subsets* I *and* J *such that the convex hull of the points* $\{a_i\}_{i \in I}$ *intersects the convex hull of the points* $\{a_i\}_{i \in J}$*.*

[7] $\mathbf{Q}^n$ should be treated as a linear space over the field $\mathbf{Q}$ of rationals, i.e., we allow multiplying the vectors from $\mathbf{Q}^n$ by rational scalars only, not by arbitrary reals.

Hint. Note that the system of $n+1 < m$ homogeneous linear equations

$$\begin{aligned} \sum_{i=1}^{m} \lambda_i a_i &= 0, \\ \sum_{i=1}^{m} \lambda_i &= 0 \end{aligned}$$

has a nontrivial solution λ^* and set $I = \{i : \lambda_i^* \geq 0\}$, $J = \{i : \lambda_i^* < 0\}$.

Exercise 1.15. *Derive the Helley theorem from the Radon theorem.*

Hint. Apply induction on the number M of the sets $A_1, \ldots, A_M$. To justify the inductive step, it suffices to demonstrate that if the Helley theorem is valid for every collection of $M \geq n+1$ sets, then it is valid for a collection of $M+1$ sets $A_1, \ldots, A_{M+1}$. To verify this implication, consider the $M+1$ nonempty (by the inductive hypothesis) sets

$$\begin{aligned} B_1 = A_2 \cap A_3 \cap \cdots \cap A_{M+1};\ B_2 = A_1 \cap A_3 \cap A_4 \cap \cdots \cap A_{M+1}; \\ \cdots;\ B_{M+1} = A_1 \cap A_2 \cap \cdots \cap A_M. \end{aligned}$$

For every i, choose a point $a_i \in B_i$, apply to the resulting collection of $M+1 \geq n+2$ points the Radon theorem, and demonstrate that every point from the intersection of the corresponding convex hulls is a common point of the sets $A_1, \ldots, A_{M+1}$.

Exercise 1.16. *Consider the Tschebyshev approximation problem,*

$$\sigma^* = \min_x \left\{ \max_{i=1,\ldots,M} |a_i^T x - b_i| \right\}, \tag{T}$$

and let k be the rank of the system $a_1, \ldots, a_M$. Prove that one can choose a subset $J \subset \{1, \ldots, M\}$, containing no more than $k+1$ indices, in such a way that the optimal value in the relaxed problem

$$\min_x \left\{ \max_{i \in J} |a_i^T x - b_i| \right\} \tag{T_J}$$

is equal to σ^*.

Hint. Look at the convex sets $X_i = \{x \mid |a_i^T x - b_i| < \sigma^*\}$.

Prove that if every k of the vectors $a_1, \ldots, a_M$ are linearly independent and $\sigma^* > 0$, then for every optimal solution x^* to (T) there exist $k+1$ indices of alternance—there exists a subset $J \subset \{1, \ldots, M\}$ of the cardinality $k+1$ such that

$$|a_i^T x^* - b_i| = \sigma^* \quad \forall i \in J.$$

Integration formulas and Gauss points

An integration formula is a formula of the type

$$\int_\Delta f(t)dt \approx \sum_{i=1}^N \alpha_i f(t_i)$$

with *nonnegative* weights α_i. Given an integration formula (i.e., a finite set of grid points $t_1, \ldots, t_N$ and nonnegative weights $\alpha_1, \ldots, \alpha_N$), one may ask how rich is the set of functions for which the formula is exact. For example, the equidistant two-point formula

$$\int_{-1}^1 f(t)dt \approx f(-1/2) + f(1/2)$$

is exact for linear functions but is not exact for quadratic polynomials. In contrast to this, the Gauss formula

$$\int_{-1}^1 f(t)dt \approx f(-1/\sqrt{3}) + f(1/\sqrt{3})$$

is exact on all polynomials of degree ≤ 3.

It turns out that the Helley theorem and the Farkas lemma allow one to get the following very general result:

> $(*)$ *Let Δ be a subset of $\mathbf{R}^k$, let L be an n-dimensional linear space comprising continuous real-valued functions on Δ, and let $I(f) : L \to \mathbf{R}$ be an integral—a linear functional on L that is nonnegative at every $f \in L$ such that $f(t) \geq 0$ everywhere on Δ. Assume also that if a function $f \in L$ is nonnegative on Δ and is not identically 0, then $I(f) > 0$. Then there exists an exact n-point cubature formula for I, i.e., there are n points $t_1, \ldots, t_n \in \Delta$ and n nonnegative weights $\alpha_1, \ldots, \alpha_n$ such that*
>
> $$I(f) = \sum_{i=1}^n \alpha_i f(t_i) \quad \forall f \in L.$$

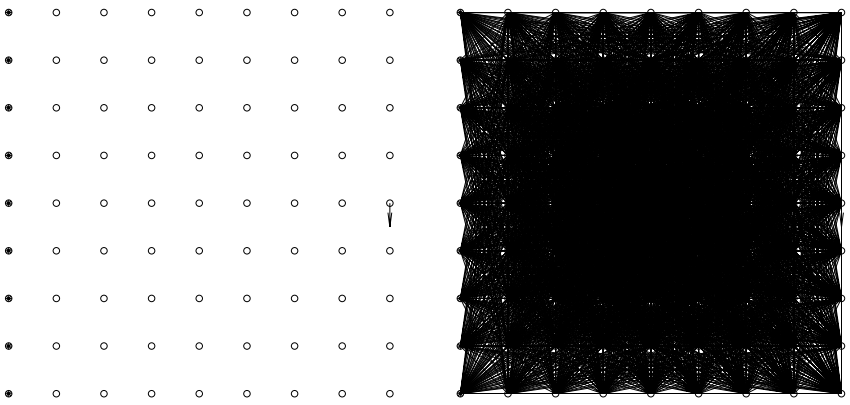
Exercise 1.17. 1. *Prove $(*)$ for the case of finite Δ.*

> *Hint.* Assuming that I is not identically zero, associate with points $t \in \Delta$ the convex sets $A_t = \{f \in L \mid f(t) \leq 0, I(f) = 1\}$ and prove that there exist n sets $A_{t_1}, \ldots, A_{t_n}$ of this type with empty intersection. Apply the homogeneous Farkas lemma to the linear forms $f(t_1), \ldots, f(t_n), I(f)$ of $f \in L$.

2. *Prove $(*)$ for the general case.*

1.4.5 How many bars are needed in an optimal truss?

Let us look at the ground structure shown on Fig. 1.8. We see that the optimal bar uses just 24 of the 3204 tentative bars. Is this phenomenon typical or not? As you shall see in a while, the answer is positive: there exists an optimal truss with no more than $m + 1$ bars, where



9×9 planar nodal grid and the load (left); 3204 tentative bars (right).
(The most left nodes are fixed; the dimension of the space of
displacements is $m = 2 \times 8 \times 9 = 144$.)

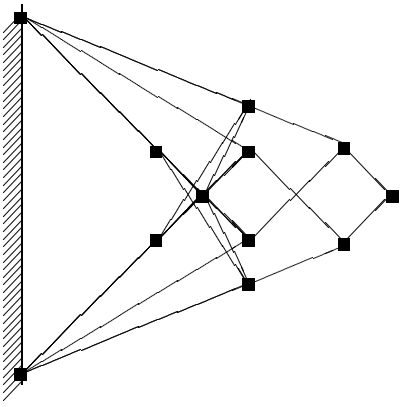


Figure 1.8. *Ground structure and optimal truss (*24 *bars).*

m is the dimension of the space displacements. Thus, with the above ground structure we know in advance that there exists an optimal truss with at most 145 bars; this is more than the 24 bars in the optimal truss we have found but still is by an order of magnitude less than the number of tentative bars.

Exercise 1.18. *Consider an LP problem*

$$\min \left\{ c^T x \;\middle|\; Ax = b, x \geq 0 \right\}$$

with $k \times n$ matrix A of rank r. Assuming that the program is solvable, prove that there exists an optimal solution x^* to the problem with at most r nonzero coordinates.

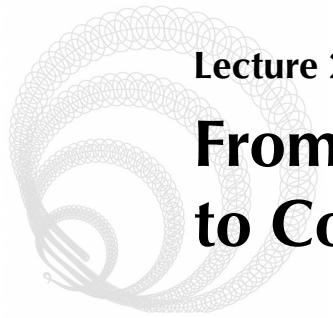
Hint. Look at an optimal solution with the minimum possible number of positive coordinates.

Exercise 1.19. *Consider a TTD problem with M-dimensional space of virtual displacements. Prove that there exists an optimal solution to this problem with no more than $m + 1$ bars of positive weight.*

Hint. Given an optimal truss t^* along with the associated displacement v^*, demonstrate that every solution t to the system

$$\begin{aligned} \sum_{i=1}^{n} t_i (b_i^T v^*) b_i &= f, \\ \sum_{i=1}^{n} t_i &= w, \\ t &\geq 0, \end{aligned}$$

is also an optimal truss.



Lecture 2

From Linear Programming to Conic Programming

Linear programming models cover numerous applications. Whenever applicable, LP allows one to obtain useful quantitative and qualitative information on the problem at hand. The specific analytic structure of LP programs gives rise to a number of general results (e.g., those of the LP duality theory) that provide us in many cases with valuable insight and understanding (see, e.g., Exercise 1.19). At the same time, this analytic structure underlies some specific computational techniques for LP; these techniques, which by now are perfectly well developed, allow one to solve routinely quite large (tens or hundreds of thousands of variables and constraints) LP programs. Nevertheless, there are real-life situations that cannot be covered by LP models. To handle these essentially nonlinear cases, one needs to extend the basic theoretical results and computational techniques known for LP beyond the bounds of LP.

For the time being, the widest class of optimization problems to which the basic results of LP were extended is the class of *convex* optimization programs. There are several equivalent ways to define a general convex optimization problem; the one we are about to use is not the traditional one, but it is well suited to encompass the range of applications covered in this book.

When passing from a generic LP problem

$$\min_x \left\{ c^T x \;\middle|\; Ax \geq b \right\}, \quad [A : m \times n], \tag{LP}$$

to its nonlinear extensions, we should expect to encounter some nonlinear components in the problem. The traditional way here is to say, "Well, in (LP) there are a linear objective function $f_0(x) = c^T x$ and inequality constraints $f_i(x) \geq b_i$ with linear functions $f_i(x) = a_i^T x$, $i = 1, \ldots, m$. Let us allow some or all of these functions $f_0, f_1, \ldots, m$ to be nonlinear." In contrast to this traditional way, we intend to keep the objective and the constraints linear, but introduce nonlinearity in the inequality sign $\geq$.

2.1 Orderings of $\mathbf{R}^m$ and convex cones

The constraint inequality $Ax \geq b$ in (LP) is an inequality between *vectors*; as such, it requires a definition, and the definition is well known: given two vectors $a, b \in \mathbf{R}^m$, we write $a \geq b$ if the coordinates of a majorate the corresponding coordinates of b:

$$a \geq b \Leftrightarrow \{a_i \geq b_i,\ i = 1, \ldots, m\}. \qquad (\geq)$$

In the latter relation, we again meet with the inequality sign $\geq$, but now it stands for the "arithmetic $\geq$"—a well known relation between real numbers. The above coordinatewise partial ordering of vectors in $\mathbf{R}^m$ satisfies a number of basic properties of the standard ordering of reals; namely, for all vectors $a, b, c, d, \ldots \in \mathbf{R}^m$ one has

1. *reflexivity:* $a \geq a$;
2. *antisymmetry:* if both $a \geq b$ and $b \geq a$, then $a = b$;
3. *transitivity:* if both $a \geq b$ and $b \geq c$, then $a \geq c$;
4. *compatibility with linear operations:*
   (a) *homogeneity:* if $a \geq b$ and λ is a nonnegative real, then $\lambda a \geq \lambda b$ ("one can multiply both sides of an inequality by a nonnegative real");
   (b) *additivity:* if both $a \geq b$ and $c \geq d$, then $a + c \geq b + d$ ("one can add two inequalities of the same sign").

It turns out that

- *A significant part of the nice features of LP programs comes from the fact that the vector inequality $\geq$ in the constraint of* (LP) *satisfies the properties* 1–4.
- *The definition* $(\geq)$ *is neither the only possible nor the only interesting way to define the notion of a vector inequality fitting the axioms* 1–4.

As a result,

> *a generic optimization problem that looks exactly the same as* (LP)*, up to the fact that the inequality $\geq$ in* (LP) *is now replaced by a vector inequality different from the componentwise ordering, inherits a significant part of the properties of LP problems. Specifying properly the ordering of vectors, one can obtain from* (LP) *generic optimization problems covering many important applications that cannot be treated by the standard LP.*

To the moment what is said is just a declaration. Let us see how this declaration comes to life.

We start by clarifying the geometry of a vector inequality satisfying the axioms 1–4. Thus, we consider vectors from $\mathbf{R}^m$ and assume that $\mathbf{R}^m$ is equipped with a partial ordering, denoted by $\succeq$; in other words, we say what are the pairs of vectors a, b from $\mathbf{R}^m$ linked by the inequality $a \succeq b$. We call the ordering good if it obeys the axioms 1–4, and we want to understand what these good orderings are.

Our first observation follows:

A good inequality $\succeq$ is completely identified by the set $\mathbf{K}$ of $\succeq$-nonnegative vectors:

$$\mathbf{K} = \{a \in \mathbf{R}^m \mid a \succeq 0\}.$$

Namely,

$$a \succeq b \Leftrightarrow a - b \succeq 0, \quad [\Leftrightarrow a - b \in \mathbf{K}].$$

Indeed, let $a \succeq b$. By 1 we have $-b \succeq -b$, and by 4(b) we may add the latter inequality to the former one to get $a - b \succeq 0$. Conversely, if $a - b \succeq 0$, then, adding to this inequality the one $b \succeq b$, we get $a \succeq b$.

The set $\mathbf{K}$ in the observation cannot be arbitrary. It is easy to verify (do it!) that it must be a *pointed convex cone*, i.e., it must satisfy the following conditions:

1. $\mathbf{K}$ is nonempty and closed under addition:

$$a, a' \in \mathbf{K} \Rightarrow a + a' \in \mathbf{K}.$$

2. $\mathbf{K}$ is a conic set:

$$a \in \mathbf{K}, \lambda \geq 0 \Rightarrow \lambda a \in \mathbf{K}.$$

3. $\mathbf{K}$ is pointed:

$$a \in \mathbf{K} \text{ and } -a \in \mathbf{K} \Rightarrow a = 0.$$

Geometrically, $\mathbf{K}$ does not contain straight lines passing through the origin.

Thus, every nonempty pointed convex cone $\mathbf{K}$ in $\mathbf{R}^m$ induces a partial ordering on $\mathbf{R}^m$, which satisfies the axioms 1–4. We denote this ordering by $\geq_{\mathbf{K}}$:

$$a \geq_{\mathbf{K}} b \Leftrightarrow a - b \geq_{\mathbf{K}} 0 \Leftrightarrow a - b \in \mathbf{K}.$$

Which cone is responsible for the standard coordinatewise ordering $\geq$ we have started with? The answer is clear: this is the cone made up of vectors with nonnegative entries—the *nonnegative orthant*

$$\mathbf{R}^m_+ = \{x = (x_1, \ldots, x_m)^T \in \mathbf{R}^m : x_i \geq 0, \ i = 1, \ldots, m\}.$$

(Thus, to express that a vector a is greater than or equal to, in the componentwise sense, a vector b, we were supposed to write $a \geq_{\mathbf{R}^m_+} b$. However, we will not be that formal and shall use the standard shorthand notation $a \geq b$.)

The nonnegative orthant $\mathbf{R}^m_+$ is not just a pointed convex cone; it possesses two useful additional properties:

1. The cone is closed: if a sequence of vectors a^i from the cone has a limit, the latter also belongs to the cone.

2. The cone possesses a nonempty interior: there exists a vector such that a ball of positive radius centered at the vector is contained in the cone.

These additional properties are very important. For example, property 1 is responsible for the possibility to pass to the termwise limit in an inequality:

$$a^i \geq b^i \quad \forall i, a^i \to a, b^i \to b \text{ as } i \to \infty \Rightarrow a \geq b.$$

It makes sense to restrict ourselves to good partial orderings coming from cones $\mathbf{K}$ sharing properties 1 and 2. Thus,

> *From now on, speaking about good partial orderings $\geq_{\mathbf{K}}$, we always assume that the underlying set $\mathbf{K}$ is a pointed and closed convex cone with a nonempty interior.*

Note that the closedness of $\mathbf{K}$ makes it possible to pass to limits in $\geq_{\mathbf{K}}$-inequalities:

$$a^i \geq_{\mathbf{K}} b^i,\ a^i \to a, b^i \to b \text{ as } i \to \infty \Rightarrow a \geq_{\mathbf{K}} b.$$

The nonemptiness of the interior of $\mathbf{K}$ allows us to define, along with the nonstrict inequality $a \geq_{\mathbf{K}} b$, the strict inequality according to the rule

$$a >_{\mathbf{K}} b \Leftrightarrow a - b \in \text{int}\mathbf{K},$$

where int $\mathbf{K}$ is the interior of the cone $\mathbf{K}$. For example, the strict coordinatewise inequality $a >_{\mathbf{R}^m_+} b$ (shorthand: $a > b$) simply says that the coordinates of a are strictly greater, in the usual arithmetic sense, than the corresponding coordinates of b.

Examples. The partial orderings we are especially interested in are given by the following cones:

- the nonnegative orthant $\mathbf{R}^m_+$;
- the *Lorentz* (or the second order, or the ice cream) cone

$$\mathbf{L}^m = \left\{ x = (x_1, \ldots, x_{m-1}, x_m)^T \in \mathbf{R}^m : x_m \geq \sqrt{\sum_{i=1}^{m-1} x_i^2} \right\};$$

- The positive semidefinite cone $\mathbf{S}^m_+$. This cone lives in the space $\mathbf{S}^m$ of $m \times m$ symmetric matrices and consists of all $m \times m$ matrices A which are positive semidefinite, i.e.,

$$A = A^T, \quad x^T A x \geq 0 \quad \forall x \in \mathbf{R}^m.$$

2.2 What is conic programming?

Let $\mathbf{K}$ be a cone in $\mathbf{R}^m$ (convex, pointed, closed, and with a nonempty interior). Given an objective $c \in \mathbf{R}^n$, an $m \times n$ constraint matrix A, and a right-hand side $b \in \mathbf{R}^m$, consider the optimization problem

$$\min_x \left\{ c^T x \;\middle|\; Ax - b \geq_{\mathbf{K}} 0 \right\}. \tag{CP}$$

We shall refer to conic programming (CP) as a *conic* problem associated with the cone $\mathbf{K}$. Note that the only difference between this program and an LP problem is that the latter deals with the particular choice $\mathbf{K} = \mathbf{R}^m_+$. With the formulation (CP), we can cover a much wider spectrum of applications that cannot be captured by LP. To get an idea, let us look at the following two examples:

Example 2.2.1. Synthesis of arrays of antennae (see section 1.2.4). Given N building blocks—antennae $S_1, \ldots, S_N$ with diagrams $Z_1(\delta), \ldots, Z_N(\delta)$, a target diagram $Z_*(\delta)$, and a finite grid T in the space of directions—find (complex-valued) weights $z_\ell = u_\ell + i v_\ell$ minimizing the quantity

$$\|Z_* - \sum_{\ell=1}^N z_\ell Z_\ell\|_{T,\infty} = \max_{\delta \in T} |Z_*(\delta) - \sum_{\ell=1}^N z_\ell Z_\ell(\delta)|.$$

In Lecture 1 we dealt with a particular case of this problem—the one where Z_* and all Z_i were real valued. There it was sufficient to consider real design variables z_ℓ, and consequently the problem could be posed as an LP problem. In the general case, some or all of the functions Z_*, Z_ℓ are complex valued; as a result, the conversion of the problem to an LP problem fails. However, we can pose the problem as (CP). Indeed, let $w = (u_1, v_1, \ldots, u_N, v_N)^T \in \mathbf{R}^{2N}$ be a collection of our design variables—the real and the imaginary parts of the complex-valued weights $z_1, \ldots, z_N$. For a particular direction δ, the complex number

$$Z_*(\delta) - \sum_{\ell=1}^N z_\ell Z_\ell(\delta)$$

treated as a 2D real vector, is an *affine* function of w:

$$Z_*(\delta) - \sum_{\ell=1}^N z_\ell Z_\ell(\delta) = \alpha_\delta w + \beta_\delta \qquad [\alpha_\delta \text{ is } 2 \times 2N \text{ matrix}, \beta_\delta \in \mathbf{R}^2].$$

Consequently, the problem of interest can be posed as

$$\min_{w,t} \left\{ t \;\middle|\; \|\alpha_\delta w + \beta_\delta\|_2 \le t \quad \forall \delta \in T \right\}. \tag{An}$$

Now, a constraint

$$\|\alpha_\delta w + \beta_\delta\|_2 \le t$$

means that the 3D vector

$$\begin{pmatrix} \alpha_\delta w + \beta_\delta \\ t \end{pmatrix},$$

affinely depending on the design vector $x = (w, t)$ of (An),

$$\begin{pmatrix} \alpha_\delta w + \beta_\delta \\ t \end{pmatrix} \equiv A_\delta x - b_\delta,$$

belongs to the 3D Lorentz cone $\mathbf{L}^3$. Consequently, (An) can be posed as

$$\min_{x=(w,t)} \left\{ c^T x \equiv t \;\middle|\; A_\delta x + b_\delta \in \mathbf{L}^3 \quad \forall \delta \in T \right\}.$$

Introducing the cone

$$\mathbf{K} = \prod_{\delta \in T} \mathbf{L}^3$$

along with the affine mapping

$$Ax - b = \{A_\delta x + b_\delta\}_{\delta \in T},$$

we finally express our problem as the conic problem

$$\min_x \left\{ c^T x \;\middle|\; Ax - b \geq_{\mathbf{K}} 0 \right\}.$$

The problem we end up with is a *conic quadratic* program—a conic program associated with a cone $\mathbf{K}$ which is a direct product of (finitely many) ice cream cones.

We remark that the same reduction to a conic quadratic problem can be obtained for the problem of synthesis of filters in the frequency domain (see Lecture 1).

Example 2.2.2. Stability analysis for an uncertain linear time-varying dynamic system. Consider a linear time-varying system

$$\frac{d}{dt} v(t) = Q(t) v(t) \tag{S}$$

with $m \times m$ time-varying matrix $Q(t)$. Assume that all we know about the matrix $Q(t)$ is that the matrix, at every time instant t, belongs to a given polytope:

$$Q(t) \in \mathrm{Conv}(Q_1, \ldots, Q_k).$$

Imagine, e.g., that all entries of Q but one are constant, while, say, the entry Q_{11} varies within known bounds $Q_{11}^{\min}$ and $Q_{11}^{\max}$. This is exactly the same as saying that $Q(t)$ belongs to the polytope spanned by two matrices Q_1 and Q_2, the (1,1)-entry of the matrices being $Q_{11}^{\min}$ and $Q_{11}^{\max}$, respectively, and the remaining entries being the same as in $Q(\cdot)$.

For the system (S), the question of primary interest is whether the system is stable, i.e., whether all trajectories of the system tend to 0 as $t \to \infty$. A simple sufficient condition for stability is the existence of a *quadratic Lyapunov function*—a function

$$L(v) = v^T X v,$$

where X is a symmetric positive definite matrix, such that

$$\frac{d}{dt} L(v(t)) \leq -\alpha L(v(t)) \tag{Ly}$$

for every trajectory of the system; here $\alpha > 0$ is the decay rate. Condition (Ly) clearly implies that

$$L(v(t)) \equiv v^T(t)Xv(t) \leq \exp\{-\alpha t\}L(v(0)),$$

and since X is positive definite, the latter inequality, in turn, implies that $v(t) \to 0, t \to \infty$. Thus, whenever (Ly) can be satisfied by a pair (X, α) with positive definite X and positive α, the pair can be treated as a stability certificate for (S).

Now, the left-hand side in (Ly) can be easily computed: it is simply

$$v^T(t)[Q^T(t)X + XQ(t)]v(t).$$

Consequently, (Ly) requires that

$$v^T(t)[Q^T(t)X+XQ(t)]v(t) \leq -\alpha v^T(t)Xv(t) \Leftrightarrow -v^T(t)[Q^T(t)X+XQ(t)+\alpha X]v(t) \geq 0.$$

For t given, the matrix $Q(t)$ can be an arbitrary matrix from the polytope $\mathrm{Conv}(Q_1, \ldots, Q_k)$, and $v(t)$ can be an arbitrary vector. Thus, (Ly) requires the matrix $[-Q^T X - XQ - \alpha X]$ to be positive semidefinite whenever $Q \in \mathrm{Conv}(Q_1, \ldots, Q_k)$ or, which is the same (why?), requires the validity of the inequalities

$$-Q_i^T X - XQ_i - \alpha X \geq_{\mathbf{S}^m_+} 0, \ i = 1, \ldots, k.$$

Now, a positive definite matrix X can be extended, by a positive α, to a pair (X, α) satisfying the indicated inequalities, if and only if the matrices

$$-Q_i^T X - XQ_i, \ i = 1, \ldots, k,$$

are positive definite (why?). We conclude that

> *In order to certify the stability of* (S) *by a quadratic Lyapunov function, it suffices to find a symmetric matrix X satisfying the following system of strict $\mathbf{S}^m_+$-inequalities:*
>
> $$X >_{\mathbf{S}^m_+} 0; \quad -Q_i^T X - XQ_i >_{\mathbf{S}^m_+} 0. \tag{2.2.1}$$

Now, a symmetric matrix A is positive definite if and only if the matrix $A - \tau I$, where I is the unit matrix of the same size as A, is positive semidefinite for some positive τ (see the exercises in Lecture 2). Consequently, to verify whether (2.2.1) is solvable is the same as verifying whether the optimal value of the program

$$\min_{t\in\mathbf{R}, X\in\mathbf{S}^m} \left\{ t \;\middle|\; \begin{pmatrix} X + tI & & & \\ & -Q_1^T X - XQ_1 + tI & & \\ & & \ddots & \\ & & & -Q_k^T X - XQ_k + tI \end{pmatrix} \geq_{\mathbf{S}^{m(k+1)}_+} 0 \right\} \tag{2.2.2}$$

with the design variables t and the $\frac{m(m+1)}{2}$ free entries of the symmetric matrix X is or is not negative. If the optimal value in the problem is negative, then (2.2.1) is solvable, and

one can use as a solution to (2.2.1) the X-component of an arbitrary feasible solution (X, t) to (2.2.2) with negative t. Whenever this is the case, (S) is stable, and the stability can be certified by a quadratic Lyapunov function. On the other hand, if the optimal value in (2.2.2) is nonnegative, then (2.2.1) is infeasible. Whether, in the latter case, (S) is or is not stable remains unclear; all that can be said is that the stability cannot be certified by a *quadratic* Lyapunov function.

Note that (2.2.2) is a conic problem associated with the positive semidefinite cone $\mathbf{S}_+^{m(k+1)}$. Indeed, the left-hand side in the constraint inequality in (2.2.2) depends affinely on the design variables, as required in the definition of a conic program.

2.3 Conic duality

Aside from algorithmic issues, the most important theoretical result in LP is the LP duality theorem. Can this theorem be extended to conic problems? What is the extension?

The source of the LP duality theorem was the desire to get in a systematic way a lower bound on the optimal value c^* in an LP program

$$c^* = \min_x \left\{ c^T x \;\middle|\; Ax \geq b \right\}. \tag{LP}$$

The bound was obtained by looking at the inequalities of the type

$$\lambda^T Ax \geq \lambda^T b \tag{Cons(λ)}$$

with weight vectors $\lambda \geq 0$. By its origin, an inequality of this type is a consequence of the system of constraints $Ax \geq b$ of (LP), i.e., it is satisfied at every solution to the system. Consequently, whenever we are lucky to get, as the left-hand side of (Cons(λ)), the expression $c^T x$, i.e., whenever a nonnegative weight vector λ satisfies the relation

$$A^T \lambda = c,$$

the inequality (Cons(λ)) yields a lower bound $b^T \lambda$ on the optimal value in (LP). And the dual problem

$$\max \left\{ b^T \lambda \mid \lambda \geq 0, A^T \lambda = c \right\}$$

was nothing but the problem of finding the best lower bound one can get in this fashion.

The same scheme can be used to develop the dual to a conic problem

$$\min \left\{ c^T x \mid Ax \geq_{\mathbf{K}} b \right\}. \tag{CP}$$

Here the only step that needs clarification is the following one:

> *What are the admissible weight vectors λ, that is, the vectors such that the scalar inequality*
>
> $$\lambda^T Ax \geq \lambda^T b$$
>
> *is a consequence of the vector inequality $A^T x \geq_{\mathbf{K}} b$?*

In the particular case of the coordinatewise partial ordering, i.e., in the case of $\mathbf{K} = \mathbf{R}^m_+$, the admissible vectors were those with nonnegative coordinates. These vectors, however, are not necessarily admissible for an ordering $\geq_{\mathbf{K}}$ when $\mathbf{K}$ is different from the nonnegative orthant.

Example 2.3.1. Consider the ordering $\geq_{\mathbf{L}^3}$ on $\mathbf{R}^3$ given by the 3D ice cream cone:

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \geq_{\mathbf{L}^3} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Leftrightarrow a_3 \geq \sqrt{a_1^2 + a_2^2}.$$

The inequality

$$\begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix} \geq_{\mathbf{L}^3} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

is valid; however, aggregating this inequality with the aid of a positive weight vector

$$\lambda = \begin{pmatrix} 1 \\ 1 \\ 0.1 \end{pmatrix},$$

we get the false inequality

$$-1.8 \geq 0.$$

Thus, not every nonnegative weight vector is admissible for the partial ordering $\geq_{\mathbf{L}^3}$.

Answering the question is the same as naming the weight vectors λ such that

$$\forall a \geq_{\mathbf{K}} 0: \quad \lambda^T a \geq 0. \tag{2.3.3}$$

Whenever λ possesses the property (2.3.3), the scalar inequality

$$\lambda^T a \geq \lambda^T b$$

is a consequence of the vector inequality $a \geq_{\mathbf{K}} b$:

$$\begin{array}{rrcll} & a & \geq_{\mathbf{K}} & b, & \\ \Leftrightarrow & a - b & \geq_{\mathbf{K}} & 0 & \text{(additivity of } \geq_{\mathbf{K}}\text{)}, \\ \Rightarrow & \lambda^T(a-b) & \geq & 0 & \text{(by (2.3.3))}, \\ \Leftrightarrow & \lambda^T a & \geq & \lambda^T b. & \end{array}$$

Conversely, if λ is an admissible weight vector for the partial ordering $\geq_{\mathbf{K}}$,

$$\forall (a, b : a \geq_{\mathbf{K}} b): \quad \lambda^T a \geq \lambda^T b,$$

then, of course, λ satisfies (2.3.3).

Thus the weight vectors λ that are admissible for a partial ordering $\geq_{\mathbf{K}}$ are exactly the vectors satisfying (2.3.3) or, which is the same, the vectors from the set

$$\mathbf{K}_* = \{\lambda \in \mathbf{R}^m : \lambda^T a \geq 0 \quad \forall a \in \mathbf{K}\}.$$

The set $\mathbf{K}_*$ comprises vectors whose inner products with *all* vectors from $\mathbf{K}$ are nonnegative. $\mathbf{K}_*$ is called the *cone dual to* $\mathbf{K}$. The name is correct because of the following.

Theorem 2.3.1. Properties of the dual cone. *Let $K \subset \mathbf{R}^m$ be a nonempty set. Then*

(i) *The set*

$$K_* = \{\lambda \in \mathbf{R}^m : \lambda^T a \geq 0 \quad \forall a \in K\}$$

is a closed convex cone.

(ii) *If* $\mathrm{int} K \neq \emptyset$, *then K_* is pointed.*

(iii) *If K is a closed convex pointed cone, then* $\mathrm{int} K_* \neq \emptyset$.

(iv) *If K is a closed convex cone, then so is K_*, and the cone dual to K_* is K itself:*

$$(K_*)_* = K.$$

The proof of the theorem is the subject of Exercise 2.1. An immediate corollary of the Theorem is as follows.

Corollary 2.3.1. *A set $K \subset \mathbf{R}^m$ is a closed convex pointed cone with a nonempty interior if and only if the set K_* is so.*

From the dual cone to the problem dual to (CP). Now we are ready to derive the dual problem of a conic problem (CP). As in the case of LP, we start from the observation that whenever x is a feasible solution to (CP) and λ is an admissible weight vector, i.e., $\lambda \in \mathbf{K}_*$, then x satisfies the scalar inequality

$$\lambda^T A x \geq \lambda^T b.$$

This observation is an immediate consequence of the definition of $\mathbf{K}_*$. It follows that whenever λ_* is an admissible weight vector satisfying the relation

$$A^T \lambda = c,$$

one has

$$c^T x = (A^T \lambda)^T x = \lambda^T A x \geq \lambda^T b = b^T \lambda$$

for all x feasible for (CP), so that the quantity $b^T \lambda$ is a lower bound on the optimal value of (CP). The best bound one can get in this fashion is the optimal value in the problem

$$\max \left\{ b^T \lambda \mid A^T \lambda = c, \lambda \geq_{\mathbf{K}_*} 0 \right\} \tag{D}$$

and this program is called the program *dual* to (CP).

So far, what we know about the duality just introduced is the following.

Proposition 2.3.1. Weak duality theorem. *The optimal value of* (D) *is a lower bound on the optimal value of* (CP).

2.3.1 Geometry of the primal and dual problems

The structure of problem (D) looks quite different from the one of (CP). However, a more careful analysis demonstrates that the difference in structures comes just from the way we represent the data: geometrically, the problems are completely similar. Indeed, in (D) we are asked to maximize a linear objective $b^T\lambda$ over the intersection of an affine plane $L_* = \{\lambda \mid A^T\lambda = c\}$ with the cone $\mathbf{K}_*$. And what about (CP)? Let us pass in this problem from the true design variables x to their images $y = Ax - b$. When x runs through $\mathbf{R}^n$, y runs through the affine plane $L = \{y = Ax - b \mid x \in \mathbf{R}^n\}$. x is feasible if the corresponding $y = Ax - b$ belongs to the cone $\mathbf{K}$. Thus, in (CP) we also deal with the intersection of an affine plane, namely, L, and a cone, namely, $\mathbf{K}$. Now assume that our objective $c^T x$ can be expressed in terms of $y = Ax - b$:

$$c^T x = d^T(Ax - b) + \text{const}.$$

This assumption is clearly equivalent to the inclusion

$$c \in \text{Im}A^T. \tag{2.3.4}$$

Indeed, in the latter case we have $c = A^T d$ for some d, whence

$$c^T x = d^T Ax = d^T(Ax - b) + d^T b \quad \forall x. \tag{2.3.5}$$

In the case of (2.3.4) the primal problem (CP) can be posed equivalently as the following problem:

$$\min_y \left\{d^T y \mid y \in L,\ y \geq_{\mathbf{K}} 0\right\},$$

where $L = \text{Im}A - b$ and d is (any) vector satisfying the relation $A^T d = c$. Thus,

> *in the case of* (2.3.4) *the primal problem is, geometrically, the problem to minimize a linear form over the intersection of the affine plane L with the cone $\mathbf{K}$, and the dual problem, similarly, is to maximize another linear form over the intersection of the affine plane L_* with the dual cone $\mathbf{K}_*$.*

Now, what happens if the condition (2.3.4) is *not* satisfied? The answer is simple: in this case (CP) makes no sense—it is either unbounded below or infeasible.

Indeed, assume that (2.3.4) is not satisfied. Then, by linear algebra, the vector c is not orthogonal to the null space of A, so that there exists e such that $Ae = 0$ and $c^T e > 0$. Now let x be a feasible solution of (CP); note that all points $x - \mu$, $\mu \geq 0$, are feasible, and $c^T(x - \mu e) \to \infty$ as $\mu \to \infty$. Thus, when (2.3.4) is not satisfied, problem (CP), whenever feasible, is unbounded below.

From the above observation we see that if (2.3.4) is not satisfied, then we may reject (CP) from the very beginning. Thus, from now on we assume that (2.3.4) is satisfied. In fact in what follows (until the end of the book!) we make a stronger assumption:

> **Assumption A.** *When speaking about a CP*
>
> $$\min_x \left\{c^T x \mid Ax - b \geq_{\mathbf{K}} 0\right\}, \qquad \textit{(CP)}$$
>
> *we always assume (if the opposite is not explicitly stated) that the matrix A is of full column rank (i.e., its columns are linearly independent).*

In other words, we assume that the mapping $x \mapsto Ax$ has the trivial null space. (We have eliminated from the very beginning the redundant degrees of freedom—those not affecting the value of Ax.) Under this assumption, the equation

$$A^T d = q$$

is solvable for every right-hand side vector q.

Note that an *arbitrary* conic program (CP) can be easily converted to a program satisfying Assumption A. Indeed, this statement is evident in the case when the columns of A are linearly independent, same as in the case when A is the zero matrix. Now assume that our $m \times n$ matrix A has rank k, $1 \leq k < n$. Without loss of generality, we can assume that the first k columns of A are linearly independent, and let $\bar{A}$ be the $m \times k$ submatrix of A comprising these columns. Note that there exists (and can be easily built via the standard linear algebra) a $k \times n$ matrix B of rank k such that

$$\bar{A}Bx = Ax \quad \forall x \in \mathbf{R}^n. \qquad [A = \bar{A}B]$$

Since B is of rank k, the $k \times k$ matrix BB^T is nonsingular, so that the vector

$$f = (BB^T)^{-1}Bc \in \mathbf{R}^k$$

is well defined. Now consider the CP

$$\min_{y \in \mathbf{R}^k} \left\{ f^T y \mid \bar{A}y - b \geq_{\mathbf{K}} 0 \right\}. \qquad (\mathrm{CP}')$$

This program by construction satisfies Assumption A and represents equivalently the original problem (CP) in the following sense:

1. A candidate solution $x \in \mathbf{R}^n$ is feasible for (CP) if and only if $y[x] = Bx$ is feasible for (CP′).

2. If $c \in \mathrm{Im}(A^T)$ (so that $c = A^T d$ for certain d), then the values $c^T x$ and $f^T y[x]$ of the objectives of (CP) and (CP′) on the candidate solutions x, $y[x]$ of the respective problems are equal to each other:

$$\begin{aligned} f^T y[x] &= f^T Bx = [(BB^T)^{-1}Bc]^T Bx = [(BB^T)^{-1}BA^T d]^T Bx \\ &= [(BB^T)^{-1}BA^T d]^T Bx = [(BB^T)^{-1}BB^T\bar{A}^T d]^T Bx \\ &= [\bar{A}^T d]^T Bx = d^T \bar{A}Bx = d^T Ax \\ &= c^T x. \end{aligned}$$

3. If $c \notin \mathrm{Im}(A^T)$, then (CP) is either infeasible (if (CP′) is so), or is unbounded below (if (CP′) is feasible).

We see that the feasibility–solvability status of (CP) is explicitly given by that of (CP′), and the feasible–optimal solutions of (CP), if any, can be easily obtained from the feasible–optimal solutions of (CP′).

Another way to ensure Assumption A is to represent x as the difference of two nonnegative vectors and to rewrite (CP) equivalently as

$$\min_{u,v} \left\{ c^T(u - v) \mid A(u - v) - b \geq_{\mathbf{K}} 0, u \geq 0, v \geq 0 \right\};$$

the constraints of the resulting problem form the vector inequality

$$\underbrace{\begin{bmatrix} A & -A \\ I & \\ & I \end{bmatrix}}_{\widetilde{A}} \begin{pmatrix} u \\ v \end{pmatrix} - \begin{pmatrix} b \\ 0 \\ 0 \end{pmatrix} \geq_{\widetilde{\mathbf{K}}} 0, \quad \widetilde{\mathbf{K}} = \mathbf{K} \times \mathbf{R}_+^{\dim x} \times \mathbf{R}_+^{\dim x},$$

and the matrix $\widetilde{A}$ is of full column rank.

As we have seen, in the case of $c \in \mathrm{Im}A^T$, problem (CP) can be reformulated as a problem (P) of minimizing a linear objective $d^T y$ over the intersection of an affine plane L and a cone $\mathbf{K}$. Conversely, a problem (P) of this latter type can be posed in the form of (CP). To this end it suffices to represent the plane L as the image of an affine mapping $x \mapsto Ax - b$ (i.e., to parameterize somehow the feasible plane) and to "translate" the objective $d^T y$ to the space of x-variables—to set $c = A^T d$, which yields

$$y = Ax - b \Rightarrow d^T y = c^T x + \text{const}.$$

Thus, when dealing with a conic problem, we may pass from its analytic form (CP) to the geometric form (P) and vice versa.

What are the relations between the geometric data of the primal and the dual problems? We already know that the cone $\mathbf{K}_+$ associated with the dual problem is dual of the cone $\mathbf{K}$ associated with the primal one. What about the feasible planes L and L_*? The answer is simple: they are orthogonal to each other! More exactly, the affine plane L is the translation, by vector $-b$, of the linear subspace

$$\mathcal{L} = \mathrm{Im}A \equiv \{y = Ax \mid x \in \mathbf{R}^n\}.$$

And L_* is the translation, by any solution λ_0 of the system $A^T\lambda = c$, e.g., by the solution d to the system, of the linear subspace

$$\mathcal{L}_* = \mathrm{Null}(A^T) \equiv \{\lambda \mid A^T\lambda = 0\}.$$

A well known fact of linear algebra is that the linear subspaces $\mathcal{L}$ and $\mathcal{L}_*$ are orthogonal complements of each other:

$$\mathcal{L} = \{y \mid y^T\lambda = 0 \quad \forall \lambda \in \mathcal{L}_*\}; \quad \mathcal{L}_* = \{\lambda \mid y^T\lambda = 0 \quad \forall y \in \mathcal{L}\}.$$

Thus, we come to a nice geometrical conclusion:

> *A conic problem* (CP) *with* $c \in \mathrm{Im}A$ *(in particular, a problem satisfying Assumption A) is the problem*
>
> $$\min_y \left\{d^T y \mid y \in \mathcal{L} - b, \ y \geq_{\mathbf{K}} 0\right\} \tag{P}$$
>
> *of minimizing a linear objective* $d^T y$ *over the intersection of a cone* $\mathbf{K}$ *with an affine plane* $L = \mathcal{L} - b$ *given as a translation, by vector* $-b$*, of a linear subspace* $\mathcal{L}$*.*

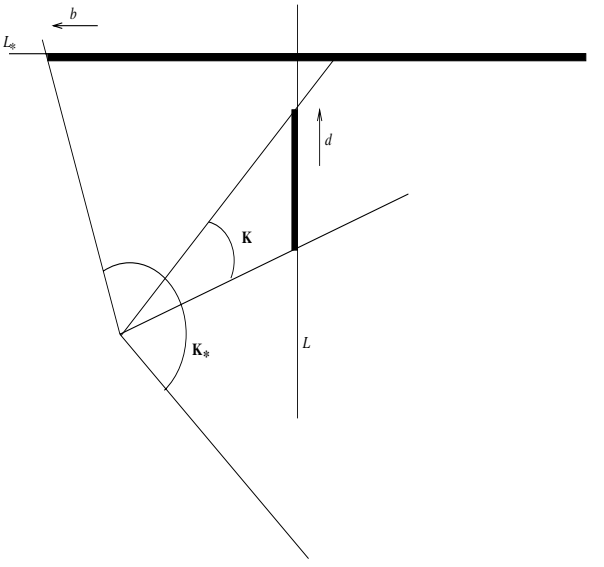


Figure 2.1. *Primal-dual pair of conic problems. Bold lines: primal (vertical segment) and dual (horizontal ray) feasible sets.*

The dual problem is the problem

$$\max_{\lambda} \left\{ b^T \lambda \mid \lambda \in \mathcal{L}^\perp + d,\ \lambda \geq_{\mathbf{K}_*} 0 \right\} \tag{D}$$

of maximizing the linear objective $b^T\lambda$ over the intersection of the dual cone $\mathbf{K}_*$ with an affine plane $L_* = \mathcal{L}^\perp + d$ given as a translation, by the vector d, of the orthogonal complement $\mathcal{L}^\perp$ of $\mathcal{L}$.

What we get is an extremely transparent geometric description of the primal-dual pair of conic problems (P), (D). Note that the duality is completely symmetric: the problem dual to (D) is (P)! Indeed, we know from Theorem 2.3.1 that $(\mathbf{K}_*)_* = \mathbf{K}$, and of course $(\mathcal{L}^\perp)^\perp = \mathcal{L}$. Switch from maximization to minimization corresponds to the fact that the shifting vector in (P) is $(-b)$, while the shifting vector in (D) is d. The geometry of the primal-dual pair (P), (D) is illustrated in Fig. 2.1. Finally, note that in the case when (CP) is an LP program (i.e., in the case when $\mathbf{K}$ is the nonnegative orthant), the conic dual problem (D) is exactly the usual LP dual; this fact immediately follows from the observation that the cone dual to $\mathbf{R}^m_+$ is $\mathbf{R}^m_+$ itself.

We have explored the geometry of a primal-dual pair of conic problems: the geometric data of such a pair are given by a pair of dual-to-each-other cones $\mathbf{K}$, $\mathbf{K}_*$ in $\mathbf{R}^m$ and a pair of affine planes $L = \mathcal{L} - b$, $L_* = \mathcal{L}^\perp + d$, where $\mathcal{L}$ is a linear subspace in $\mathbf{R}^m$ and $\mathcal{L}^\perp$ is its orthogonal complement. The first problem from the pair—let it be called (P)—is to minimize $b^T y$ over $y \in \mathbf{K} \cap L$, and the second (D) is to maximize $d^T\lambda$ over $\lambda \in \mathbf{K}_* \cap L_*$. Note that the geometric data $(\mathbf{K}, \mathbf{K}_*, L, L_*)$ of the pair do not specify completely the problems of the pair: given L, L_*, we can uniquely define $\mathcal{L}$ but not the shift vectors $(-b)$ and d: b is known

up to shift by a vector from $\mathcal{L}$, and d is known up to shift by a vector from $\mathcal{L}^\perp$. However, this nonuniqueness is of absolutely no importance; replacing a chosen vector $d \in L_*$ by another vector $d' \in L_*$, we pass from (P) to a new problem (P′), which is completely equivalent to (P). Indeed, both (P) and (P′) have the same feasible set, and on the (common) feasible plane L of the problems their objectives $d^T y$ and $(d')^T y$ differ from each other by a constant:

$$y \in L = \mathcal{L} - b, d - d' \in \mathcal{L}^\perp \Rightarrow (d-d')^T(y+b) = 0 \Rightarrow (d-d')^T y = -(d-d')^T b \quad \forall y \in L.$$

Similarly, shifting b along $\mathcal{L}$, we modify the objective in (D), but in a trivial way—on the feasible plane L_* of the problem the new objective differs from the old one by a constant.

2.4 Conic duality theorem

The weak duality (Proposition 2.3.1) we have established so far for conic problems is much weaker than the LP duality theorem. Is it possible to get results similar to those of the LP duality theorem in the general conic case as well? The answer is affirmative, *provided that the primal problem* (CP) *is strictly feasible*, i.e., that there exists x such that $Ax - b >_{\mathbf{K}} 0$, or, geometrically, $L \cap \mathrm{int}\mathbf{K} \neq \emptyset$.

The advantage of the geometrical definition of strict feasibility is that it is independent of the particular way in which the feasible plane is defined. Hence, with this definition it is clear what it means when the dual problem (D) is strictly feasible.

Our main result is the following.

Theorem 2.4.1. Conic duality theorem. *Consider a conic problem*

$$c^* = \min_x \left\{ c^T x \mid Ax \geq_{\mathbf{K}} b \right\} \tag{CP}$$

along with its conic dual

$$b^* = \max \left\{ b^T \lambda \mid A^T \lambda = c, \lambda \geq_{\mathbf{K}_*} 0 \right\}. \tag{D}$$

1. *The duality is symmetric: the dual problem is conic, and the problem dual to dual is (equivalent to) the primal.*

2. *The value of the dual objective at every dual feasible solution λ is $\leq$ the value of the primal objective at every primal feasible solution x, so that the duality gap*

$$c^T x - b^T \lambda$$

is nonnegative at every primal-dual feasible pair (x, λ).

3.a. *If the primal* (CP) *is bounded below and strictly feasible (i.e., $Ax >_{\mathbf{K}} b$ for some x), then the dual* (D) *is solvable and the optimal values in the problems are equal to each other: $c_+ = b^*$.*

3.b. *If the dual* (D) *is bounded above and strictly feasible (i.e., exists $\lambda >_{\mathbf{K}_*} 0$ such that $A^T \lambda = c$), then the primal* (CP) *is solvable and $c^* = b^*$.*

4. *Assume that at least one of the problems* (CP), (D) *is bounded and strictly feasible. Then a primal-dual feasible pair (x, λ) is a pair of optimal solutions to the respective problems*

4.a. *if and only if*

$$b^T\lambda = c^T x \qquad \text{[zero duality gap]}$$

and

4.b. *if and only if*

$$\lambda^T[Ax - b] = 0 \qquad \text{[complementary slackness]}.$$

Proof. 1. The result was obtained in the discussion of the geometry of the primal and the dual problems.

2. This is the weak duality theorem.

3. Assume that (CP) is strictly feasible and bounded below, and let c^* be the optimal value of the problem. We should prove that the dual is solvable with the same optimal value. Since we already know that the optimal value of the dual is $\leq c^*$ (see 2), all we need is to point out a dual feasible solution λ_* with $b^T\lambda_* \geq c^*$.

Consider the convex set

$$M = \{y = Ax - b \mid x \in \mathbf{R}^n, c^T x \leq c^*\}.$$

Let us start with the case of $c \neq 0$. We claim that in this case

(i) the set M is nonempty;

(ii) the plane M does not intersect the interior K of the cone $\mathbf{K}$: $M \cap \mathrm{int}\mathbf{K} = \emptyset$;

Claim (i) is evident (why?). To verify claim (ii), assume, on the contrary, that there exists a point $\bar{x}$, $c^T\bar{x} \leq c^*$, such that $\bar{y} \equiv A\bar{x} - b >_{\mathbf{K}} 0$. Then, of course, $Ax - b >_{\mathbf{K}} 0\ \forall x$ close enough to $\bar{x}$, i.e., all points x in a small-enough neighborhood of $\bar{x}$ are also feasible for (CP). Since $c \neq 0$, there are points x in this neighborhood with $c^T x < c^T\bar{x} \leq c^*$, which is impossible, since c^* is the optimal value of (CP).

Now let us make use of the following basic fact.

Theorem 2.4.2. Separation theorem for convex sets. *Let S, T be nonempty, nonintersecting convex subsets of $\mathbf{R}^m$. Then S and T can be separated by a linear functional: there exists a nonzero vector $\lambda \in \mathbf{R}^m$ such that*

$$\sup_{u\in S} \lambda^T u \leq \inf_{u\in T} \lambda^T u.$$

Applying the separation theorem to $S = M$ and $T = K$, we conclude that there exists $\lambda \in \mathbf{R}^m$ such that

$$\sup_{y\in M} \lambda^T y \leq \inf_{y\in \mathrm{int}\mathbf{K}} \lambda^T y. \tag{2.4.6}$$

From the inequality it follows that the linear form $\lambda^T y$ is bounded below on $K = \mathrm{int}\mathbf{K}$. Since this interior is a conic set,

$$y \in K, \mu > 0 \Rightarrow \mu y \in K$$

(why?), this boundedness implies that $\lambda^T y \geq 0\ \forall y \in K$. Consequently, $\lambda^T y \geq 0\ \forall y$ from the closure of K, i.e., $\forall y \in \mathbf{K}$. We conclude that $\lambda \geq_{\mathbf{K}_*} 0$, so that the inf in (2.4.6) is

nonnegative. On the other hand, the infimum of a linear form over a conic set clearly cannot be positive; we conclude that the inf in (2.4.6) is 0, so that the inequality reads

$$\sup_{u\in M} \lambda^T u \le 0.$$

Recalling the definition of M, we get

$$[A^T\lambda]^T x \le \lambda^T b \tag{2.4.7}$$

for all x from the half-space $c^T x \le c^*$. But the linear form $[A^T\lambda]^T x$ can be bounded above on the half-space if and only if the vector $A^T\lambda$ is proportional, with a nonnegative coefficient, to the vector c,

$$A^T\lambda = \mu c$$

for some $\mu \ge 0$. We claim that $\mu > 0$. Indeed, assuming $\mu = 0$, we get $A^T\lambda = 0$, whence $\lambda^T b \ge 0$ in view of (2.4.7). It is time now to recall that (CP) is strictly feasible, i.e., $A\bar{x} - b >_{\mathbf{K}} 0$ for some $\bar{x}$. Since $\lambda \ge_{\mathbf{K}_*} 0$ and $\lambda \ne 0$, the product $\lambda^T[A\bar{x} - b]$ should be strictly positive (why?), while in fact we know that the product is $-\lambda^T b \le 0$ (since $A^T\lambda = 0$ and, as we have seen, $\lambda^T b \ge 0$).

Thus, $\mu > 0$. Setting $\lambda_* = \mu^{-1}\lambda$, we get

$$\begin{array}{rcll} \lambda_* & \ge_{\mathbf{K}_*} & 0 & [\text{since } \lambda \ge_{\mathbf{K}_*} 0 \text{ and } \mu > 0], \\ A^T\lambda_* & = & c & [\text{since } A^T\lambda = \mu c], \\ c^T x & \le & \lambda_*^T b \quad \forall x : c^T x \le c^* & [\text{see (2.4.7)}]. \end{array}$$

We see that λ_* is feasible for (D), the value of the dual objective at λ_* being at least c^*, as required.

It remains to consider the case $c = 0$. Here, of course, $c^* = 0$, and the existence of the dual feasible solution with the value of the objective $\ge c^* = 0$ is evident: the required solution is $\lambda = 0$. Thus 3.a is proved.

3.b: the result follows from 3.a in view of the primal-dual symmetry.

4: Let x be primal feasible and λ be dual feasible. Then

$$c^T x - b^T\lambda = (A^T\lambda)^T x - b^T\lambda = [Ax - b]^T\lambda.$$

We get a useful identity, as follows.

PROPOSITION 2.4.1. *For every primal-dual feasible pair (x, λ) of solutions to* (CP), (D)*, the duality gap $c^T x - b^T\lambda$ is equal to the inner product of the primal slack vector $y = Ax - b$ and the dual vector λ.*

Note that the conclusion in Proposition 2.4.1 in fact does not require full primal-dual feasibility: x may be arbitrary (i.e., y should belong to the primal feasible plane $\mathrm{Im}A - b$), and λ should belong to the dual feasible plane $A^T\lambda = c$, but y and λ should not necessarily belong to the respective cones.

In view of Proposition 2.4.1, the complementary slackness holds if and only if the duality gap is zero; thus, all we need is to prove 4.a.

The primal residual $c^T x - c^*$ and the dual residual $b^* - b^T \lambda$ are nonnegative, provided that x is primal feasible and λ is dual feasible. It follows that the duality gap

$$c^T x - b^T \lambda = [c^T x - c^*] + [b^* - b^T \lambda] + [c^* - b^*]$$

is nonnegative (recall that $c^* \geq b^*$ by 2), and it is zero if and only if $c^* = b^*$ and both primal and dual residuals are zero (i.e., x is primal optimal, and λ is dual optimal). All these arguments hold without any assumptions of strict feasibility. We see that the condition "the duality gap at a primal-dual feasible pair is zero" is always sufficient for primal-dual optimality of the pair. If $c^* = b^*$, this sufficient condition is also necessary. Since in the case of 4 we indeed have $c^* = b^*$ (this is stated by 3), 4.a follows. □

A useful consequence of the conic duality theorem is the following.

Corollary 2.4.1. *Assume that both* (CP) *and* (D) *are strictly feasible. Then both problems are solvable, the optimal values are equal to each other, and each one of the conditions* 4.a, 4.b *is necessary and sufficient for optimality of a primal-dual feasible pair.*

Indeed, by the weak duality theorem, if one of the problems is feasible, the other is bounded, and it remains to use the items 3 and 4 of the conic duality theorem.

2.4.1 Is something wrong with conic duality?

The statement of the conic duality theorem is weaker than that of the LP duality theorem. In the LP case, feasibility (even nonstrict) and boundedness of either primal or dual problem implies solvability of both the primal and the dual and equality between their optimal values. In the general conic case, something nontrivial is stated only in the case of strict feasibility (and boundedness) of one of the problems. It can be demonstrated by examples that this phenomenon reflects the nature of things and is not due to our ability to analyze it. The case of nonpolyhedral cone $\mathbf{K}$ is truly more complicated than the one of the nonnegative orthant $\mathbf{K}$; as a result, a word-by-word extension of the LP duality theorem to the conic case is false.

Example 2.4.1. Consider the following conic problem with two variables $x = (x_1, x_2)^T$ and the 3D ice cream cone $\mathbf{K}$:

$$\min \left\{ x_1 \mid Ax - b \equiv \begin{bmatrix} x_1 - x_2 \\ 1 \\ x_1 + x_2 \end{bmatrix} \geq_{\mathbf{L}^3} 0 \right\}.$$

Recalling the definition of $\mathbf{L}^3$, we can write the problem equivalently as

$$\min \left\{ x_1 \mid \sqrt{(x_1 - x_2)^2 + 1} \leq x_1 + x_2 \right\},$$

i.e., as the problem

$$\min \{ x_1 \mid 4x_1 x_2 \geq 1, x_1 + x_2 > 0 \}.$$

Geometrically the problem is to minimize x_1 over the intersection of the 3D ice cream cone with a 2D plane; the inverse image of this intersection in the design plane of variables x_1, x_2

is part of the 2D nonnegative orthant bounded by the hyperbola $x_1 x_2 \geq 1/4$. The problem is clearly strictly feasible (a strictly feasible solution is, e.g., $x = (1, 1)^T$) and bounded below, with the optimal value 0. This optimal value, however, is not achieved—the problem is unsolvable!

Example 2.4.2. Consider the following conic problem with two variables $x = (x_1, x_2)^T$ and the 3D ice cream cone $\mathbf{K}$:

$$\min \left\{ x_2 \mid Ax - b = \begin{bmatrix} x_1 \\ x_2 \\ x_1 \end{bmatrix} \geq_{\mathbf{L}^3} 0 \right\}.$$

The problem is equivalent to the problem

$$\left\{ x_2 \mid \sqrt{x_1^2 + x_2^2} \leq x_1 \right\},$$

i.e., to the problem

$$\min \{ x_2 \mid x_2 = 0, x_1 \geq 0 \}.$$

The problem is clearly solvable, and its optimal set is the ray $\{x_1 \geq 0, x_2 = 0\}$.

Now let us build the conic dual to our (solvable!) primal. The cone dual to an ice cream cone is this ice cream cone itself (see Exercise 2.7). Thus, the dual problem is

$$\max_{\lambda} \left\{ 0 \mid \begin{bmatrix} \lambda_1 + \lambda_3 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \lambda \geq_{\mathbf{L}^3} 0 \right\}.$$

Although the primal is solvable, the dual is infeasible. Indeed, assuming that λ is dual feasible, we have $\lambda \geq_{\mathbf{L}^3} 0$, which means that $\lambda_3 \geq \sqrt{\lambda_1^2 + \lambda_2^2}$; since also $\lambda_1 + \lambda_3 = 0$, we come to $\lambda_2 = 0$, which contradicts the equality $\lambda_2 = 1$.

We see that the weakness of the conic duality theorem as compared to the LP duality reflects pathologies that indeed may happen in the general conic case.

2.4.2 Consequences of the conic duality theorem

Sufficient condition for infeasibility. Recall that a necessary and sufficient condition for infeasibility of a (finite) system of scalar linear inequalities (i.e., for a vector inequality with respect to the partial ordering $\geq$) is the possibility to combine these inequalities in a linear fashion so that the resulting scalar linear inequality is contradictory. In the case of cone-generated vector inequalities, a slightly weaker result can be obtained.

Proposition 2.4.2. *Consider a linear vector inequality*

$$Ax - b \geq_{\mathbf{K}} 0. \tag{I}$$

(i) *If there exists* λ *satisfying*

$$\lambda \geq_{\mathbf{K}_*} 0, A^T \lambda = 0, \lambda^T b > 0, \tag{II}$$

then (I) *has no solutions.*

(ii) *If* (II) *has no solutions, then* (I) *is almost solvable—for every positive* ϵ *there exists* b' *such that* $\|b'-b\|_2<\epsilon$ *and the perturbed system*

$$Ax-b' \geq_{\mathbf{K}} 0$$

is solvable.

Moreover,

(iii) (II) *is solvable if and only if* (I) *is not almost solvable.*

Note the difference between the simple case when $\geq_{\mathbf{K}}$ is the usual partial ordering $\geq$ and the general case. In the former, one can replace "nearly solvable" in (ii) with "solvable"; however, in the general conic case "almost" is unavoidable.

Example 2.4.3. Let system (I) be given by

$$Ax-b \equiv \begin{bmatrix} x+1 \\ x-1 \\ \sqrt{2}x \end{bmatrix} \geq_{\mathbf{L}^3} 0.$$

Recalling the definition of the ice cream cone $\mathbf{L}^3$, we can write the inequality equivalently as

$$\sqrt{2}x \geq \sqrt{(x+1)^2+(x-1)^2} \equiv \sqrt{2x^2+2}, \tag{i}$$

which of course is unsolvable. The corresponding system (II) is

$$\begin{array}{cc} \lambda_3 \geq \sqrt{\lambda_1^2+\lambda_2^2} & \left[\Leftrightarrow \lambda \geq_{\mathbf{L}^3_*} 0\right], \\ \lambda_1+\lambda_2+\sqrt{2}\lambda_3=0 & \left[\Leftrightarrow A^T\lambda=0\right], \\ \lambda_2-\lambda_1>0 & \left[\Leftrightarrow b^T\lambda>0\right]. \end{array} \tag{ii}$$

From the second of these relations, $\lambda_3=-\frac{1}{\sqrt{2}}(\lambda_1+\lambda_2)$, so that from the first inequality we get $0\leq(\lambda_1-\lambda_2)^2$, whence $\lambda_1=\lambda_2$. But then the third inequality in (ii) is impossible! We see that here both (i) and (ii) have no solutions.

The geometry of the example is as follows. Point (i) asks us to find a point in the intersection of the 3D ice cream cone and a line. This line is an asymptote of the cone (it belongs to a 2D plane that crosses the cone in such way that the boundary of the cross section is a branch of a hyperbola, and the line is one of two asymptotes of the hyperbola). Although the intersection is empty ((i) is unsolvable), small shifts of the line make the intersection nonempty (i.e., (i) is unsolvable and almost solvable at the same time). And it turns out that one cannot certify that (i) itself is unsolvable by providing a solution to (ii).

Proof of the Proposition. Point (i) is evident (why?).

Let us prove (ii). To this end it suffices to verify that if (I) is not almost solvable, then (II) is solvable. Let us fix a vector $\sigma>_{\mathbf{K}}0$ and look at the conic problem

$$\min_{x,t}\{t \mid Ax+t\sigma-b\geq_{\mathbf{K}}0\} \tag{CP}$$

in variables (x, t). Clearly, the problem is strictly feasible (why?). Now, if (I) is not almost solvable, then, first, the matrix of the problem $[A; \sigma]$ satisfies the full column rank condition A. (Otherwise the image of the mapping $(x, t) \mapsto Ax + t\sigma - b$ would coincide with the image of the mapping $x \mapsto Ax - b$, which is not the case—the first of these images intersects $\mathbf{K}$, while the second does not.) Second, the optimal value in (CP) is strictly positive (otherwise the problem would admit feasible solutions with t close to 0, and this would mean that (I) is almost solvable). From the conic duality theorem it follows that the dual problem of (CP)

$$\max_{\lambda} \left\{ b^T\lambda \mid A^T\lambda = 0, \sigma^T\lambda = 1, \lambda \geq_{\mathbf{K}_*} 0 \right\}$$

has a feasible solution with positive $b^T\lambda$, i.e., (II) is solvable.

It remains to prove (iii). Assume first that (I) is not almost solvable; then (II) must be solvable by (ii). Conversely, assume that (II) is solvable, and let λ be a solution to (II). Then λ also solves all systems of the type (II) associated with small-enough perturbations of b instead of b itself; by (i), it implies that all inequalities obtained from (I) by small-enough perturbation of b are unsolvable. □

When is a scalar linear inequality a consequence of a given linear vector inequality? The question we are interested in is as follows. Given a linear vector inequality

$$Ax \geq_{\mathbf{K}} b \tag{V}$$

and a scalar inequality

$$c^T x \geq d, \tag{S}$$

we want to check whether (S) is a consequence of (V). If $\mathbf{K}$ is the nonnegative orthant, the answer is given by the Farkas lemma:

> *Inequality* (S) *is a consequence of a feasible system of linear inequalities* $Ax \geq b$ *if and only if* (S) *can be obtained from* (V) *and the trivial inequality* $1 \geq 0$ *in a linear fashion (by taking weighted sum with nonnegative weights).*

In the general conic case we can get a slightly weaker result, as follows.

Proposition 2.4.3. (i) *If* (S) *can be obtained from* (V) *and from the trivial inequality* $1 \geq 0$ *by admissible aggregation, i.e., there exist weight vector* $\lambda \geq_{\mathbf{K}_*} 0$ *such that*

$$A^T\lambda = c, \lambda^T b \geq d,$$

then (S) *is a consequence of* (V).

(ii) *If* (S) *is a consequence of a strictly feasible linear vector inequality* (V)*, then* (S) *can be obtained from* (V) *by an admissible aggregation.*

The difference between the case of the partial ordering $\geq$ and a general partial ordering $\geq_{\mathbf{K}}$ is in the word "strictly" in (ii).

Proof of the Proposition. Point (i) is evident (why?). To prove point (ii), assume that (V) is strictly feasible and (S) is a consequence of (V), and consider the conic problem

$$\min_{x,t}\left\{t \mid \bar{A}\begin{pmatrix}x\\t\end{pmatrix}-\bar{b}\equiv\begin{bmatrix}Ax-b\\ d-c^Tx+t\end{bmatrix}\succeq_{\bar{\mathbf{K}}}0\right\},$$
$$\bar{\mathbf{K}}=\{(x,t)\mid x\in\mathbf{K},t\geq 0\}.$$

The problem is clearly strictly feasible (choose x to be a strictly feasible solution to (V) and then choose t to be large enough). The fact that (S) is a consequence of (V) says exactly that the optimal value in the problem is nonnegative. By the conic duality theorem, the dual problem

$$\max_{\lambda,\mu}\left\{b^T\lambda-d\mu \mid A^T\lambda-c=0,\mu=1,\begin{pmatrix}\lambda\\\mu\end{pmatrix}\succeq_{\bar{\mathbf{K}}_*}0\right\}$$

has a feasible solution with the value of the objective ≥ 0. Since, as it is easily seen, $\bar{\mathbf{K}}_*=\{(\lambda,\mu)\mid\lambda\in\mathbf{K}_*,\mu\geq 0\}$, the indicated solution satisfies the requirements

$$\lambda\succeq_{\mathbf{K}_*}0,\ A^T\lambda=c,\ b^T\lambda\geq d;$$

i.e., (S) can be obtained from (V) by an admissible aggregation. □

Remark 2.4.1. Although the conic duality theorem and its consequences listed in section 2.4.2 were obtained under Assumption A (by our convention, this assumption acts by default, unless the opposite is explicitly stated), it is seen from the proofs that the results in question are valid under a weaker assumption, namely, that $c\in\mathrm{Im}A$ in (CP).

2.4.3 Robust solvability status

Examples 2.4.2–2.4.3 make clear that in the general conic case we may meet pathologies that do not occur in LP. For example, a feasible and bounded problem may be unsolvable, and the dual to a solvable conic problem may be infeasible. Where do the pathologies come from? Looking at our pathological examples, we arrive at the following guess. The source of the pathologies is that in these examples, the solvability status of the primal problem is nonrobust—it can be changed by small perturbations of the data. This issue of robustness is very important in modeling, and it deserves a careful investigation.

Data of a conic problem. When asked, "What are the data of an LP program $\min\{c^Tx\mid Ax-b\geq 0\}$?" everybody will give the same answer: "The objective c, the constraint matrix A, and the right-hand side vector b." Similarly, for a conic problem

$$\min\left\{c^Tx \mid Ax-b\succeq_{\mathbf{K}}0\right\}, \tag{CP}$$

its data, *by definition*, are the triple (c,A,b), while the sizes of the problem—the dimension n of x and the dimension m of $\mathbf{K}$, same as the underlying cone $\mathbf{K}$ itself—are considered the structure of (CP).

Robustness. A question of primary importance is whether the properties of the program (CP) (feasibility, solvability, etc.) are stable with respect to perturbations of the data. This question is important for these reasons:

• In actual applications, especially those arising in engineering, the data are normally inexact: their true values, even when they exist in the nature, are not known exactly when the problem is processed. Consequently, the results of the processing say something definite about the true problem only if these results are robust with respect to small data perturbations, i.e., the properties of (CP) we have discovered are shared not only by the particular (nominal) problem we were processing but also by all problems with nearby data.

• Even when the exact data are available, we should take into account that in processing them computationally we unavoidably add noise like rounding errors (you simply cannot load something like $\frac{1}{7}$ to the standard computer). As a result, a real-life computational routine can recognize only those properties of the input problem that are stable with respect to small perturbations of the data.

Due to the above reasons, we should study not only whether a given problem (CP) is feasible, bounded, solvable, etc., but also whether these properties are robust—they remain unchanged under small data perturbations. As it turns out, the conic duality theorem allows us to recognize robust feasibility, boundedness, solvability. . . .

Let us start with introducing the relevant concepts. We say that (CP) is

- *robust feasible* if all sufficiently close problems (i.e., those of the same structure $(n, m, \mathbf{K})$ and with data close enough to those of (CP)) are feasible;
- *robust infeasible* if all sufficiently close problems are infeasible;
- *robust bounded below* if all sufficiently close problems are bounded below (i.e., their objectives are bounded below on their feasible sets);
- *robust unbounded* if all sufficiently close problems are not bounded;
- *robust solvable* if all sufficiently close problems are solvable.

Note that a problem that is not robust feasible is *not necessarily* robust infeasible, since among close problems there may be both feasible and infeasible problems. (Look at Example 2.4.2; slightly shifting and rotating the plane Im $A - b$, we may get whatever we want—a feasible bounded problem, a feasible unbounded problem, an infeasible problem…). This is why we need two kinds of definitions, one of robust presence of a property and one of robust absence of the same property.

Now let us look at necessary and sufficient conditions for the most important robust forms of the solvability status.

Proposition 2.4.4. Robust feasibility. (CP) *is robust feasible if and only if it is strictly feasible, in which case the dual problem* (D) *is robust bounded above.*

Proof. The statement is nearly tautological. Let us fix $\delta >_{\mathbf{K}} 0$. If (CP) is robust feasible, then for small enough $t > 0$ the perturbed problem $\min\{c^T x \mid Ax - b - t\delta \geq_{\mathbf{K}} 0\}$ should be feasible; a feasible solution to the perturbed problem clearly is a strictly feasible solution to (CP). The inverse implication is evident (a strictly feasible solution to (CP) remains feasible

for all problems with close enough data). It remains to note that if all problems sufficiently close to (CP) are feasible, then their duals, by the weak duality theorem, are bounded above, so that (D) is robust bounded above. □

Proposition 2.4.5. Robust infeasibility. (CP) *is robust infeasible if and only if the system*

$$b^T\lambda = 1,\ A^T\lambda = 0,\ \lambda \geq_{\mathbf{K}_*} 0$$

is robust feasible or, which is the same (by Proposition 2.4.4*), if and only if the system*

$$b^T\lambda = 1,\ A^T\lambda = 0,\ \lambda >_{\mathbf{K}_*} 0 \tag{2.4.8}$$

has a solution.

Proof. First assume that (2.4.8) is solvable, and let us prove that all problems sufficiently close to (CP) are infeasible. Let us fix a solution $\bar{\lambda}$ to (2.4.8). Since A is of full column rank, simple linear algebra says that the systems $[A']^T\lambda = 0$ are solvable for all matrices A' from a small-enough neighborhood U of A; moreover, the corresponding solution $\lambda(A')$ can be chosen to satisfy $\lambda(A) = \bar{\lambda}$ and to be continuous in $A' \in U$. Since $\lambda(A')$ is continuous and $\lambda(A) >_{\mathbf{K}_*} 0$, we have $\lambda(A') >_{\mathbf{K}_*} 0$ in a neighborhood of A; shrinking U appropriately, we may assume that $\lambda(A') >_{\mathbf{K}_*} 0\ \forall A' \in U$. Now, $b^T\bar{\lambda} = 1$. By continuity reasons, there exists a neighborhood V of b and a neighborhood U' of A such that $b' \in V$ and all $A' \in U'$ one has $(b')^T\lambda(A') > 0$.

Thus, we have seen that there exist a neighborhood U' of A and a neighborhood V of b, along with a function $\lambda(A'),\ A' \in U'$, such that

$$(b')^T\lambda(A') > 0,\ [A']^T\lambda(A') = 0,\ \lambda(A') \geq_{\mathbf{K}_*} 0$$

$\forall b' \in V$ and $A' \in U$. By Proposition 2.4.2.(i), all the problems

$$\min\left\{[c']^T x \mid A'x - b' \geq_{\mathbf{K}} 0\right\}$$

with $b' \in V$ and $A' \in U'$ are infeasible, so that (CP) is robust infeasible.

Now let us assume that (CP) is robust infeasible, and let us prove that then (2.4.8) is solvable. Indeed, by the definition of robust infeasibility, there exist neighborhoods U of A and V of b such that all vector inequalities

$$A'x - b' \geq_{\mathbf{K}} 0$$

with $A' \in U$ and $b' \in V$ are unsolvable. It follows that whenever $A' \in U$ and $b' \in V$, the vector inequality

$$A'x - b' \geq_{\mathbf{K}} 0$$

is not almost solvable (see Proposition 2.4.2). We conclude from Proposition 2.4.2.(ii) that for every $A' \in U$ and $b' \in V$ there exists $\lambda = \lambda(A', b')$ such that

$$[b']^T\lambda(A', b') > 0,\ [A']^T\lambda(A', b') = 0,\ \lambda(A', b') \geq_{\mathbf{K}_*} 0.$$

Now let us choose $\lambda_0 >_{\mathbf{K}_*} 0$. For all small-enough positive ϵ we have $A_\epsilon = A + \epsilon b[A^T\lambda_0]^T \in U$. Let us choose an ϵ with the latter property so small that $\epsilon b^T\lambda_0 > -1$ and set $A' = A_\epsilon$, $b' = b$. According to the previous observation, there exists $\lambda = \lambda(A', b)$ such that

$$b^T\lambda > 0, [A']^T\lambda \equiv A^T[\lambda + \epsilon\lambda_0(b^T\lambda)] = 0, \lambda \geq_{\mathbf{K}_*} 0.$$

Setting $\bar{\lambda} = \lambda + \epsilon\lambda_0(b^T\lambda)$, we get $\bar{\lambda} >_{\mathbf{K}_*} 0$ (since $\lambda \geq_{\mathbf{K}_*} 0$, $\lambda_0 >_{\mathbf{K}_*} 0$ and $b^T\lambda > 0$), while $A\bar{\lambda} = 0$ and $b^T\bar{\lambda} = (b^T\lambda)(1+\epsilon b^T\lambda_0) > 0$. Multiplying $\bar{\lambda}$ by an appropriate positive factor (namely, by $1/(b^T\bar{\lambda})$), we get a solution to (2.4.8). □

Now we are able to formulate our main result on robust solvability.

Proposition 2.4.6. *For a conic problem* (CP) *the following conditions are equivalent to each other:*

(i) (CP) *is robust feasible and robust bounded (below).*

(ii) (CP) *is robust solvable.*

(iii) (D) *is robust solvable.*

(iv) (D) *is robust feasible and robust bounded (above).*

(v) *Both* (CP) *and* (D) *are strictly feasible.*

In particular, under every one of these equivalent assumptions, both (CP) *and* (D) *are solvable with equal optimal values.*

Proof. (i) $\Rightarrow$ (v): If (CP) is robust feasible, it also is strictly feasible (Proposition 2.4.4). If, in addition, (CP) is robust bounded below, then (D) is robust solvable (by the conic duality theorem); in particular, (D) is robust feasible and therefore strictly feasible (again, Proposition 2.4.4).

(v) $\Rightarrow$ (ii): The implication is given by the conic duality theorem.

(ii) $\Rightarrow$ (i): The proof is trivial.

We have proved that (i) $\equiv$ (ii) $\equiv$ (v). Due to the primal-dual symmetry, we also have proved that (iii) $\equiv$ (iv) $\equiv$ (v). □

2.5 Conic duality revisited

To understand our concern now, consider a simple example, an optimization program with just two variables and four constraints:

$$\begin{array}{lrcl} \max & x_1 + 2x_2, & & \\ & x_1 + x_2 & = & 4, \\ & x_1 - x_2 & \leq & 3, \\ & x_1 & \geq & 0, \\ & x_2 & \geq & 0. \end{array}$$

We immediately recognize it as an LP program, although it is *not* a problem of minimizing a linear form over the intersection of an affine plane and the nonnegative orthant, as an LP program should formally be. What is meant when we say that our toy problem is a linear

program is that it can be routinely *converted* to a "true" LP program—the one that is to minimize a linear form on the intersection of a plane and the orthant. In principle, there are two conversion policies.

First, we can use the equality constraint(s) to express in an affine fashion part of the variables via the remaining free variables. What we end up with will be a pure inequality constrained problem of optimizing a linear objective of the free variables. Of course, the resulting problem is a true LP program—a conic program associated with the nonnegative orthant $\mathbf{R}^n_+$. (In LP the latter form is called *canonical*.)

Second, we can add to our original design variables a number of artificial variables, "slacks," and convert all nontrivial inequality constraints—those saying more than that a particular variable should be nonnegative—into equality constraints. In our example this manipulation results in the following:

$$\begin{array}{rrcl} \max & x_1 + 2x_2, & & \\ & x_1 + x_2 & = & 4, \\ x_1 - x_2 + s & & = & 3, \\ & x_1 & \geq & 0, \\ & x_2 & \geq & 0, \\ & s & \geq & 0, \end{array}$$

which is again a true LP program, now in the dual form (D). (In LP, this form is called *standard*.)

To process the problem analytically (e.g., to build its dual), the second option is incomparably better—it does not require messy computations.

What is said about LP is valid in the general conic case as well. That we can convert an optimization program to a conic form does not mean normally that the original form of the problem reads "minimize a linear form over the intersection of an affine plane and a cone." A typical original form is something like

$$\begin{array}{rrcl} \min & c^T x, & & \\ & Px & = & p, \\ & A_i x - b_i & \geq_{\mathbf{K}^i} & 0, \ i = 1, \ldots, m, \end{array} \tag{Ini}$$

where $\mathbf{K}^i$ are different cones.

Let us show how to convert (Ini) to a formal conic program, like (CP) or (D), and how to build the dual problem

If (Ini) has no equality constraints, it already is a conic problem in the form (CP). Indeed, it suffices to define a new cone $\mathbf{K}$ as the direct product of the cones $\mathbf{K}^i, i = 1, \ldots, m$:

$$\mathbf{K} = \{(y_1, \ldots, y_m) \mid y_i \in \mathbf{K}^i, \ i = 1, \ldots, m\}$$

and to write the problem

$$\min \left\{ c^T x \mid Ax - b \equiv \begin{bmatrix} A_1 x - b_1 \\ A_2 x - b_2 \\ \cdots \\ A_m x - b_m \end{bmatrix} \geq_{\mathbf{K}} 0 \right\}.$$

Exercise 2.4 states that the direct product of cones $\mathbf{K}_i$ is a cone (and its dual is the direct product of the dual cones $\mathbf{K}_i^*$), so that what we get is a conic problem.

Now, what to do if (Ini) does have equality constraints? Then we may act as in our LP example. By the same reasons as above, we prefer adding slack variables rather than eliminating variables; thus, our final target is a conic program of the form (D).

It is clear that our target is achievable. A trivial way to reach it is as follows.

1. Pass from (Ini) to an equivalent problem where the design vector is restricted to belong to some cone $\mathbf{K}_0$. This is exactly what is done in LP when representing free variables—those without restrictions on their signs—as differences of two nonnegative variables. Let us do the same, and let us also switch from minimization to maximization:

$$\text{(Ini)} \mapsto \left\{ \begin{array}{rcll} \max & c^T(v-u), & & \\ & Pu - Pv & = & p, \\ & A_i(u-v) - b_i & \geq_{\mathbf{K}^i} & 0,\ i = 1, \ldots, m, \\ & u & \geq & 0, \\ & v & \geq & 0. \end{array} \right. \qquad \text{(Med)}$$

(The optimal value in (Med) is the negation of the one in (Ini).)

2. It remains to add to (Med) slack variables. These variables should correspond to the vector inequality constraints $A_i(u-v) - b_i \geq_{\mathbf{K}_i} 0$ and should therefore be vectors of the same dimensions as b_i. Denoting these slack vectors by s_i, we transform (Med) to

$$\text{(Med)} \mapsto \left\{ \begin{array}{rcll} \max & c^T(v-u), & & \\ & Pu - Pv & = & p, \\ & A_i(u-v) - s_i & = & b_i,\ i = 1, \ldots, m, \\ & u & \geq & 0, \\ & v & \geq & 0, \\ & s_i & \geq_{\mathbf{K}^i} & 0,\ i = 1, \ldots, m. \end{array} \right. \qquad \text{(Fin)}$$

We end up with problem (Fin), which is equivalent to (Ini) and clearly is a conic problem in the form of (D).

Of course, in many cases we can act smarter than in the above formal scheme. Sometimes one can extract from the original inequality constraints $A_i x - b_i \geq_{\mathbf{K}_i} 0$ a subset I of constraints saying that x belongs to some cone $\mathbf{K}_0$. (Look at the inequality constraints $x_1, x_2 \geq 0$ in our LP example.) If this is the case, there is no need to update (Ini) $\Rightarrow$ (Med), just as there is no need to introduce slacks for the constraints from I. Sometimes there is no subset of constraints saying that x belongs to a cone, but there is a subset I saying that certain subvector x' of x belongs to a certain cone; whenever this is the case, we can modify the first step of the above scheme—to represent as $u - v$ the complement of x' in x, not the entire x—so that there is no need to introduce slacks for the constraints from I at the second step, etc.

Now, what is the conic dual to (Fin)? The cone associated with (Fin) is

$$\mathbf{K}_* = \mathbf{R}^n_+ \times \mathbf{R}^n_+ \times \mathbf{K}^1 \times \mathbf{K}^2 \times \cdots \times \mathbf{K}^m,$$

the objective (to be maximized!) in (Fin) is

$$\begin{pmatrix} -c \\ c \\ 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix}^T \begin{pmatrix} u \\ v \\ s_1 \\ s_2 \\ \cdots \\ s_m \end{pmatrix},$$

and the equality constraints are

$$\begin{pmatrix} P & -P & & & & \\ A_1 & -A_1 & -I_1 & & & \\ A_2 & -A_2 & & -I_2 & & \\ \vdots & \vdots & & & \ddots & \\ A_m & -A_m & & & & -I_m \end{pmatrix} \begin{pmatrix} u \\ v \\ s_1 \\ s_2 \\ \cdots \\ s_m \end{pmatrix} = \begin{pmatrix} p \\ b_1 \\ b_2 \\ \cdots \\ b_m \end{pmatrix},$$

where I_i are unit matrices of appropriate sizes. Taking into account that the cone $\mathbf{K}_*$ is dual to the cone

$$\mathbf{K} = \mathbf{R}^n_+ \times \mathbf{R}^n_+ \times \mathbf{K}^1_* \times \mathbf{K}^2_* \times \cdots \times \mathbf{K}^m_*,$$

we conclude that (Fin) is dual to the following conic problem of the form (CP):

$$\begin{array}{rrcl} \min & p^T\mu + \sum_{i=1}^m b_i^T\xi_i, & & \\ & P^T\mu + \sum_{i=1}^m A_i^T\xi_i + c & \geq & 0, \\ & -P^T\mu - \sum_{i=1}^m A_i^T\xi_i - c & \geq & 0, \\ & -\xi_i & \geq_{\mathbf{K}^i_*} & 0,\ i = 1, \ldots, m, \end{array}$$

in the variables μ (a vector of the same dimension as p) and ξ_i, $i = 1, \ldots, m$, of the same dimensions as $b_1, \ldots, b_m$. The first two $\geq$-constraints in the resulting problem are equivalent to the vector equation

$$P\mu + \sum_{i=1}^m A_i^T\xi_i = -c.$$

It also makes sense to pass from variables ξ_i to their negatives $\eta_i = -\xi_i$, to pass from μ to its negative $\nu = -\mu$, and to switch from minimization to maximization, thus coming to the problem

$$\begin{array}{rrcl} \max & p^T\nu + \sum_{i=1}^m b_i^T\eta_i, & & \\ & P^T\nu + \sum_{i=1}^m A_i^T\eta_i & = & c, \\ & \eta_i & \geq_{\mathbf{K}^i_*} & 0,\ i = 1, \ldots, m, \end{array} \qquad \text{(Dl)}$$

in design variables $\nu, \eta_1, \ldots, \eta_m$. The resulting problem will be called the problem dual to the primal problem (Ini). Note that we have extended somehow our duality scheme—previously it required the primal problem to be purely conic—not to have linear equality constraints. From now on this restriction is eliminated.

Summary on building the dual

Following the traditions of LP textbooks, we summarize here the recipes for building the dual.

Consider a primal problem—an optimization problem with linear objective and linear vector equality and inequality constraints

$$\begin{array}{lrcll} \textit{minimize} & c^T x & & & \\ \textit{s.t.} & & & & \\ & Px & = & p & [\dim p \textit{ scalar linear equations}], \\ & A_1 x - b_1 & \geq_{\mathbf{K}^1} & 0 & [\textit{linear vector inequality no. } 1], \\ & \cdots & & & \\ & A_m x - b_m & \geq_{\mathbf{K}^m} & 0 & [\textit{linear vector inequality no. } m]. \end{array} \tag{Pr}$$

The problem dual to (Pr) *is*

$$\begin{array}{lrcll} \textit{maximize} & p^T \nu + \sum_{i=1}^m b_i^T \eta_i & & & \\ \textit{s.t.} & & & & \\ & P^T \nu + \sum_{i=1}^m A_i^T \eta_i & = & c & [\dim x \textit{ scalar equations}], \\ & \eta_1 & \geq_{\mathbf{K}_*^1} & 0 & [\textit{linear vector inequality no. } 1], \\ & \cdots & & & \\ & \eta_m & \geq_{\mathbf{K}_*^m} & 0 & [\textit{linear vector inequality no. } m]. \end{array} \tag{Dl}$$

Note the following:

1. Dual variables correspond to the constraints of the primal problem. The dual design vector comprises a vector variable ν of the same dimension as the right-hand side p of the system of primal equalities and of m vector variables η_i, $i = 1, \ldots, m$, each of the same dimensions as those of the primal vector inequalities

2. There is a natural one-to-one correspondence between the vector inequalities of the primal problem and those of the dual problem, and the cones underlying corresponding vector inequalities of (Pr) and (Dl) are dual to each other.

3. The problem dual to (Dl) is (equivalent to) the primal problem (Pr).

Indeed, (Dl) is of the same structure as (Pr), so that we can apply the outlined construction to (Dl). The (vector) variables of the resulting problem are as follows:

- The first of them, let it be called x', is responsible for the system of equations in (Dl). The dimension of this variable is $\dim c$, i.e., it is equal to the design dimension of (Pr).
- The remaining m vector variables, let them be called w_i, $i = 1, \ldots, m$, are each responsible for its own linear vector inequality of (Dl).

Applying the outlined scheme to (Dl) (please do it and do not forget to pass first from the maximization problem (Dl) to an equivalent minimization problem; this is preassumed by our scheme), we come to the problem

$$\begin{array}{lll} \text{maximize} & c^T x' & \\ \text{s.t.} & & \\ & Px' & = -p, \\ & A_i x' + w_i & = -b_i, i = 1, \ldots, m, \\ & w_i & \geq_{\mathbf{K}^i} 0, \ i = 1, \ldots, m. \end{array}$$

The resulting problem is equivalent to (Pr). (To see it, set $x = -x'$.)

Summary on conic duality

The results on conic duality we have developed so far deal with a formal primal-dual pair (Fin)–(Dl) of problems, not with the pair of problems (Pr)–(Dl). These results, however, can be easily expressed directly in terms of (Pr) and (Dl). Here is the translation:

1. The role of Assumption **A** from page 53 is now played by the pair of requirements as follows:

A.1. *the rows of the matrix P in* (Pr) *are linearly independent;*

A.2. *there is no nonzero vector x such that $Px = 0$, $A_i x = 0$, $i = 1, \ldots, m$.*

From now on, speaking about problem (Pr), we *always* assume **A.1** and **A.2**.

Note that **A.1** and **A.2** imply that both (Fin) and (Dl) satisfy Assumption **A** (why?).

2. Strict feasibility. A problem of the form (Pr) is called strictly feasible if there exist a feasible solution $\bar{x}$ such that $A_i \bar{x} - b_i >_{\mathbf{K}_i} 0$, $i = 1, \ldots, m$. Note that (Pr) is strictly feasible if and only if (Fin) is.

3. Weak duality. The optimal value in (Dl) is less than or equal to the optimal value in (Pr).

4. Strong duality. If one of the problems (Pr), (Dl) is strictly feasible and bounded, then the other problem is solvable, and the optimal values in the problems are equal to each other. If both problems are strictly feasible, then both are solvable with equal optimal values.

5. Optimality conditions. Let x be a feasible solution to (Pr) and $\lambda = (\nu, \{\eta_i\}_{i=1}^m)$ be a feasible solution to (Dl). The duality gap at the pair (x, λ)—the quantity

$$\Delta(x, \lambda) = c^T x - \left[p^T \nu + \sum_{i=1}^m b_i^T \eta_i \right]$$

—is nonnegative and is equal to

$$\sum_{i=1}^m \eta_i^T [A_i x - b_i].$$

The duality gap is zero if and only if the complementary slackness holds:

$$\eta_i^T [A_i x - b_i] = 0, \ i = 1, \ldots, m.$$

If the duality gap $\Delta(x, \lambda)$ is zero, then x is an optimal solution to (Pr) and λ is an optimal solution to (Dl). If x is an optimal solution to (Pr) and λ is an optimal solution to (Dl) and the optimal values in the problems are equal, then the duality gap $\Delta(x, \lambda)$ is zero.

2.6 Exercises to Lecture 2

2.6.1 Cones

In this section, cone always means a pointed closed convex cone with a nonempty interior in $\mathbf{R}^n$.

Theorem 2.3.1

Exercise 2.1. 1. *Prove the following statement:*

Let S be a nonempty closed convex set in $\mathbf{R}^n$ and x be a point in $\mathbf{R}^n$ outside of S. Then the problem

$$\min\{(x-y)^T(x-y) \mid y \in S\}$$

has a unique solution x^*, and $e \equiv x - x^*$ strictly separates x and S, i.e.,

$$e^T x \geq e^T e + \sup_{y\in S} e^T y > \sup_{y\in S} e^T y.$$

2. *Derive Theorem* 2.3.1 *from item* 1 *above.*

3. *Derive from Theorem* 2.3.1 *that whenever $0 \neq x \geq_{\mathbf{K}} 0$, there exists $\lambda \geq_{\mathbf{K}_*} 0$ such that $\lambda^T x > 0$.*

Interior of a cone

Exercise 2.2. *Let $\mathbf{K}$ be a cone, and let $\bar{x} >_{\mathbf{K}} 0$. Prove that $x >_{\mathbf{K}} 0$ if and only if there exists positive real t such that $x \geq_{\mathbf{K}} t\bar{x}$.*

Exercise 2.3. 1. *Prove that if $0 \neq x \geq_{\mathbf{K}} 0$ and $\lambda >_{\mathbf{K}_*} 0$, then $\lambda^T x > 0$.*

Hint. Use the results of Exercises 2.1 and 2.2.

2. *Prove that if $\lambda >_{\mathbf{K}_*} 0$, then for every real a the set*

$$\{x \geq_{\mathbf{K}} 0 \mid \lambda^T x \leq a\}$$

is bounded.

Calculus of cones

Exercise 2.4. *Prove the following statements.*

1. (Stability with respect to direct multiplication.) *Let $\mathbf{K}_i \subset \mathbf{R}^{n_i}$ be cones, $i = 1, \ldots, k$. Prove that the direct product of the cones*

$$\mathbf{K} = \mathbf{K}_1 \times \cdots \times \mathbf{K}_k = \{(x_1, \ldots, x_k) \mid x_i \in \mathbf{K}_i,\ i = 1, \ldots, k\}$$

is a cone in $\mathbf{R}^{n_1+\cdots+n_k} = \mathbf{R}^{n_1} \times \cdots \times \mathbf{R}^{n_k}$.

Prove that the cone dual to $\mathbf{K}$ *is the direct product of the cones dual to* $\mathbf{K}_i$, $i = 1, \ldots, k$.

2. (Stability with respect to taking inverse image.) *Let* $\mathbf{K}$ *be a cone in* $\mathbf{R}^n$ *and* $u \mapsto Au$ *be a linear mapping from certain* $\mathbf{R}^k$ *to* $\mathbf{R}^n$ *with trivial null space (*$\mathrm{Null}(A) = \{0\}$*) and such that* $\mathrm{Im}A \cap \mathrm{int}\mathbf{K} \neq \emptyset$. *Prove that the inverse image of* $\mathbf{K}$ *under the mapping*

$$A^{-1}(\mathbf{K}) = \{u \mid Au \in \mathbf{K}\}$$

is a cone in $\mathbf{R}^k$.

Prove that the cone dual to $A^{-1}(\mathbf{K})$ *is* $A^T\mathbf{K}_*$, *i.e.,*

$$(A^{-1}(\mathbf{K}))_* = \{A^T\lambda \mid \lambda \in \mathbf{K}_*\}.$$

3. (Stability with respect to taking linear image.) *Let* $\mathbf{K}$ *be a cone in* $\mathbf{R}^n$ *and* $y = Ax$ *be a linear mapping from* $\mathbf{R}^n$ *onto* $\mathbf{R}^N$ *(i.e., the image of* A *is the entire* $\mathbf{R}^N$*). Assume* $\mathrm{Null}(A) \cap \mathbf{K} = \{0\}$.

Prove then that the set

$$A\mathbf{K} = \{Ax \mid x \in \mathbf{K}\}$$

is a cone in $\mathbf{R}^N$.

Prove that the cone dual to $A\mathbf{K}$ *is*

$$(A\mathbf{K})_* = \{\lambda \in \mathbf{R}^N \mid A^T\lambda \in \mathbf{K}_*\}.$$

Demonstrate by example that if in the above statement the assumption $\mathrm{Null}(A) \cap \mathbf{K} = \{0\}$ *is weakened to* $\mathrm{Null}(A) \cap \mathrm{int}\mathbf{K} = \emptyset$, *then the set* $A(\mathbf{K})$ *may happen to be nonclosed.*

> *Hint*. Look what happens when the 3D ice cream cone is projected onto its tangent plane.

Primal-dual pairs of cones and orthogonal pairs of subspaces

Exercise 2.5. *Let* A *be an* $m \times n$ *matrix of full column rank and* $\mathbf{K}$ *be a cone in* $\mathbf{R}^m$.

1. *Prove that at least one of the following facts always takes place:*
   (i) *There exists a nonzero* $x \in \mathrm{Im}\, A$ *which is* $\geq_{\mathbf{K}} 0$.
   (ii) *There exists a nonzero* $\lambda \in \mathrm{Null}(A^T)$ *which is* $\geq_{\mathbf{K}_*} 0$.

Geometrically: Given a primal-dual pair of cones $\mathbf{K}$, $\mathbf{K}_*$ *and a pair* L, $L^\perp$ *of linear subspaces that are orthogonal complements of each other, we can find a nontrivial ray in the intersection* $L \cap \mathbf{K}$ *or in the intersection* $L^\perp \cap \mathbf{K}_*$ *or both.*

2. *Prove that there exists* $\lambda \in \mathrm{Null}(A^T)$ *which is* $>_{\mathbf{K}_*} 0$ *(this is the strict version of* (ii)*) if and only if* (i) *is false. Prove that, similarly, there exists* $x \in \mathrm{Im}A$ *which is* $>_{\mathbf{K}} 0$ *(this is the strict version of* (i)*) if and only if* (ii) *is false.*

Geometrically: If $\mathbf{K}$, $\mathbf{K}_*$ *is a primal-dual pair of cones and* L, $L^\perp$ *are linear subspaces that are orthogonal complements of each other, then the intersection* $L \cap \mathbf{K}$ *is trivial (i.e., is the singleton* $\{0\}$*) if and only if the intersection* $L^\perp \cap \mathrm{int}\mathbf{K}_*$ *is nonempty.*

Several interesting cones

Given a cone $\mathbf{K}$ along with its dual $\mathbf{K}_*$, let us call a *complementary pair* every pair $x \in \mathbf{K}$, $\lambda \in \mathbf{K}_*$ such that

$$\lambda^T x = 0.$$

Recall that in good cases (e.g., under the premise of item 4 of the conic duality theorem) a pair of feasible solutions (x, λ) of a primal-dual pair of conic problems

$$\min \left\{ c^T x \mid Ax - b \geq_{\mathbf{K}} 0 \right\},$$

$$\max \left\{ b^T \lambda \mid A^T \lambda = c, \lambda \geq_{\mathbf{K}_*} 0 \right\}$$

is primal-dual optimal if and only if the primal slack $y = Ax - b$ and λ are complementary.

Exercise 2.6. Nonnegative orthant. *Prove that the n-dimensional nonnegative orthant $\mathbf{R}^n_+$ is a cone and that it is self-dual:*

$$(\mathbf{R}^n_+)_* = \mathbf{R}^n_+.$$

What are complementary pairs?

Exercise 2.7. Ice cream cone. *Let $\mathbf{L}^n$ be the n-dimensional ice cream cone:*

$$\mathbf{L}^n = \left\{ x \in \mathbf{R}^n \mid x_n \geq \sqrt{x_1^2 + \cdots + x_{n-1}^2} \right\}.$$

1. *Prove that $\mathbf{L}^n$ is a cone.*
2. *Prove that the ice cream cone is self-dual:*

$$(\mathbf{L}^n)_* = \mathbf{L}^n.$$

3. *Characterize the complementary pairs.*

Exercise 2.8. Positive semidefinite cone. *Let $\mathbf{S}^n_+$ be the cone of $n \times n$ positive semidefinite matrices in the space $\mathbf{S}^n$ of symmetric $n \times n$ matrices. Assume that $\mathbf{S}^n$ is equipped with the Frobenius inner product*

$$\langle X, Y \rangle = \mathrm{Tr}(XY) = \sum_{i,j=1}^{n} X_{ij} Y_{ij}.$$

1. *Prove that $\mathbf{S}^n_+$ indeed is a cone.*
2. *Prove that the semidefinite cone is self-dual:*

$$(\mathbf{S}^n_+)_* = \mathbf{S}^n_+.$$

That is, prove the Frobenius inner products of a symmetric matrix Λ with all positive semidefinite matrices X of the same size are nonnegative if and only if the matrix Λ itself is positive semidefinite.

3. *Prove the following characterization of the complementary pairs:*

Two matrices $X \in \mathbf{S}^n_+$, $\Lambda \in (\mathbf{S}^n_+)_* \equiv \mathbf{S}^n_+$ are complementary (i.e., $\langle \Lambda, X \rangle = 0$) if and only if their matrix product is zero: $\Lambda X = X\Lambda = 0$. In particular, matrices from a complementary pair commute and therefore share a common orthonormal eigenbasis.

2.6.2 Conic problems

Several primal-dual pairs

Exercise 2.9. The min-max Steiner problem. *Consider the following problem.*

Given N points $b_1, \ldots, b_N$ in $\mathbf{R}^n$, find a point $x \in \mathbf{R}^n$ that minimizes the maximum (Euclidean) distance from itself to the points $b_1, \ldots, b_N$; i.e., solve the problem

$$\min_x \max_{i=1,\ldots,N} \|x - b_i\|_2.$$

Imagine, e.g., that $n = 2$, $b_1, \ldots, b_N$ are locations of villages and you want to locate a fire station for which the worst-case distance to a possible fire is as small as possible.

1. *Pose the problem as a conic quadratic one—a conic problem associated with a direct product of ice cream cones.*
2. *Build the dual problem.*
3. *What is the geometric interpretation of the dual? Are the primal and the dual strictly feasible? solvable? with equal optimal values? What is the meaning of the complementary slackness?*

Exercise 2.10. The weighted Steiner problem. *Consider the following problem:*

Given N points $b_1, \ldots, b_N$ in $\mathbf{R}^n$ along with positive weights $\omega_i, i = 1, \ldots, N$, find a point $x \in \mathbf{R}^n$ that minimizes the weighted sum of its (Euclidean) distances to the points $b_1, \ldots, b_N$; i.e., solve the problem

$$\min_x \sum_{i=1}^N \omega_i \|x - b_i\|_2.$$

Imagine, e.g., that $n = 2$, $b_1, \ldots, b_N$ are locations of N villages and you want to place a telephone station for which the total cost of cables linking the station and the villages is as small as possible. The weights can be interpreted as the per-mile cost of the cables. (They may vary from village to village due to differences in populations and, consequently, in the required capacities of the cables.)

1. *Pose the problem as a conic quadratic one.*
2. *Build the dual problem.*
3. *What is the geometric interpretation of the dual? Are the primal and the dual strictly feasible? solvable? with equal optimal values? What is the meaning of the complementary slackness?*

2.6.3 Feasible and level sets of conic problems

Consider a feasible conic problem

$$\min\left\{c^T x \mid Ax - b \geq_{\mathbf{K}} 0\right\}. \tag{CP}$$

In many cases it is important to know whether the problem has

1. bounded feasible set $\{x \mid Ax - b \geq_{\mathbf{K}} 0\}$,
2. bounded level sets

$$\{x \mid Ax - b \geq_{\mathbf{K}} 0, c^T x \leq a\}$$

for all real a.

Exercise 2.11. *Let* (CP) *be feasible. Then the following four properties are equivalent:*

(i) *The feasible set of the problem is bounded.*

(ii) *The set of primal slacks* $Y = \{y \mid y \geq_{\mathbf{K}} 0, y = Ax - b\}$ *is bounded.*[8]

(iii) $\mathrm{Im}\, A \cap \mathbf{K} = \{0\}$.

(iv) *The system of vector (in)equalities*

$$A^T\lambda = 0, \lambda >_{\mathbf{K}_*} 0$$

is solvable.

Corollary. *The property of* (CP) *to have a bounded feasible set is independent of the particular value of b, provided that with this b* (CP) *is feasible!*

Exercise 2.12. *Let problem* (CP) *be feasible. Prove that the following two conditions are equivalent:*

(i) (CP) *has bounded level sets.*

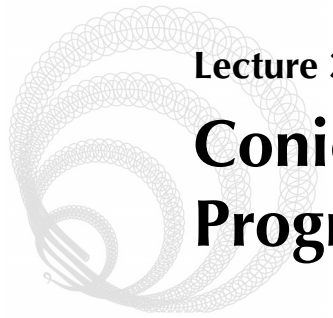
(ii) *The dual problem*

$$\max\left\{b^T\lambda \mid A^T\lambda = c, \lambda \geq_{\mathbf{K}_*} 0\right\}$$

is strictly feasible.

Corollary. *The property of* (CP) *to have bounded level sets is independent of the particular value of b, provided that with this b* (CP) *is feasible!*

[8] Recall that we always assume that **A** holds!



Lecture 3
Conic Quadratic Programming

Several generic families of conic problems are of special interest, from the viewpoint of both theory and applications. The cones underlying these problems are simple enough, so that one can describe explicitly the dual cone. As a result, the general duality machinery we have developed becomes algorithmic, as in the LP case. Moreover, in many cases this algorithmic duality machinery allows us to understand more deeply the original model, to convert it into equivalent forms better suited for numerical processing, etc. The relative simplicity of the underlying cones also enables one to develop efficient computational methods for the corresponding conic problems. The most famous example of a "nice" generic conic problem is, doubtless, LP; however, it is not the only problem of this sort. Two other nice generic conic problems of extreme importance are conic quadratic and semidefinite programs. We are about to consider the first of these two problems.

3.1 Conic quadratic problems: Preliminaries

Recall the definition of the m-dimensional ice cream ($\equiv$ second-order $\equiv$ Lorentz) cone $\mathbf{L}^m$:

$$\mathbf{L}^m = \{x = (x_1, \ldots, x_m) \in \mathbf{R}^m \mid x_m \geq \sqrt{x_1^2 + \cdots + x_{m-1}^2}\}, \quad m \geq 2.$$

A *conic quadratic problem* is a conic problem

$$\min_x \left\{c^T x \mid Ax - b \geq_{\mathbf{K}} 0\right\} \tag{CP}$$

for which the cone $\mathbf{K}$ is a direct product of several ice cream cones:

$$\begin{aligned} \mathbf{K} &= \mathbf{L}^{m_1} \times \mathbf{L}^{m_2} \times \cdots \times \mathbf{L}^{m_k} \\ &= \left\{ y = \begin{pmatrix} y[1] \\ y[2] \\ \cdots \\ y[k] \end{pmatrix} \mid y[i] \in \mathbf{L}^{m_i},\ i = 1, \ldots, k \right\}. \end{aligned} \tag{3.1.1}$$

In other words, a conic quadratic problem is an optimization problem with linear objective and finitely many *ice cream constraints*

$$A_i x - b_i \succeq_{\mathbf{L}^{m_i}} 0, \ \ i = 1, \ldots, k,$$

where

$$[A; b] = \left[\begin{array}{c} [A_1; b_1] \\ \hline [A_2; b_2] \\ \hline \cdots \\ \hline [A_k; b_k] \end{array}\right]$$

is the partition of the data matrix $[A; b]$ corresponding to the partition of y in (3.1.1). Thus, a conic quadratic program can be written as

$$\min_x \left\{ c^T x \mid A_i x - b_i \succeq_{\mathbf{L}^{m_i}} 0, \ \ i = 1, \ldots, k \right\}. \tag{3.1.2}$$

Let us recall that for a vector $z \in \mathbf{R}^m$, the inequality $z \succeq_{\mathbf{L}^m} 0$ means that the last entry in z is $\geq$ the Euclidean norm $\|\cdot\|_2$ of the subvector of z comprising the first $m-1$ entries of z. Consequently, the $\succeq_{\mathbf{L}^{m_i}} 0$-inequalities in (3.1.2) can be written as

$$\|D_i x - d_i\|_2 \leq p_i^T x - q_i,$$

where

$$[A_i; b_i] = \left[\begin{array}{cc} D_i & d_i \\ p_i^T & q_i \end{array}\right]$$

is the partitioning of the data matrix $[A_i, b_i]$ into the submatrix $[D_i; d_i]$ consisting of the first $m_i - 1$ rows and the last row $[p_i^T; q_i]$. Hence, a conic quadratic problem can be written as

$$\min_x \left\{ c^T x \mid \|D_i x - d_i\|_2 \leq p_i^T x - q_i, \ \ i = 1, \ldots, k \right\}, \tag{QP}$$

and this most explicit form is the one we prefer to use. In this form, D_i are matrices of the same row dimension as x, d_i are vectors of the same dimensions as the column dimensions of the matrices D_i, p_i are vectors of the same dimension as x and q_i are reals.

We know from Exercises 2.7 and 2.4 that (3.1.1) is indeed a cone, in fact a self-dual one: $\mathbf{K}_* = \mathbf{K}$. Consequently, the problem dual to (CP) is

$$\max_\lambda \left\{ b^T \lambda \mid A^T \lambda = c, \ \ \lambda \succeq_{\mathbf{K}} 0 \right\}.$$

Denoting

$$\lambda = \left(\begin{array}{c} \lambda_1 \\ \lambda_2 \\ \cdots \\ \lambda_k \end{array}\right)$$

with m_i-dimensional blocks λ_i (cf. (3.1.1)), we can write the dual problem as

$$\max_{\lambda_1,\dots,\lambda_m}\left\{\sum_{i=1}^k b_i^T\lambda_i \mid \sum_{i=1}^k A_i^T\lambda_i = c,\ \lambda_i \geq_{\mathbf{L}^{m_i}} 0,\ i=1,\dots,k\right\}.$$

Recalling the meaning of $\geq_{\mathbf{L}^{m_i}} 0$ and representing $\lambda_i = \binom{\mu_i}{\nu_i}$ with scalar component ν_i, we finally come to the following form of the problem dual to (QP):

$$\max_{\mu_i,\nu_i}\left\{\sum_{i=1}^k[\mu_i^T d_i + \nu_i q_i] \mid \sum_{i=1}^k[D_i^T\mu_i + \nu_i p_i] = c,\ \mid \|\mu_i\| \leq \nu_i,\ i=1,\dots,k\right\}. \quad \text{(QD)}$$

The design variables in (QD) are vectors μ_i of the same dimensions as the vectors d_i and reals ν_i, $i=1,\dots,k$.

Since from now on we will treat (QP) and (QD) as the standard forms of a conic quadratic problem and its dual, we now interpret for these two problems our basic Assumption A from Lecture 2 and notions like feasibility, strict feasibility, and boundedness. Assumption A now reads as follows (why?):

> *There is no nonzero x that is orthogonal to all rows of all matrices D_i and to all vectors p_i, $i=1,\dots,k$.*

We always make this assumption by default. Now, among notions like feasibility and solvability, the only notion that does need an interpretation is strict feasibility, which now reads as follows (why?):

> *Strict feasibility of* (QP) *means that there exists $\bar{x}$ such that $\|D_i\bar{x} - d_i\|_2 < p_i^T\bar{x} - q_i\ \forall i$.*
>
> *Strict feasibility of* (QD) *means that there exists a feasible solution $\{\bar{\mu}_i, \bar{\nu}_i\}_{i=1}^k$ to the problem such that $\|\mu_i\|_2 < \bar{\nu}_i\ \forall i = 1,\dots,k$.*

3.2 Examples of conic quadratic problems

3.2.1 Best linear approximation of complex-valued functions

Recall the Tschebyshev approximation problem from Lecture 1, which we now formulate as follows:

> *Given a finite set T, a target function f_* on T set, and n building blocks—functions $f_1,\dots,f_n$ on T—find a linear combination of the functions $f_1,\dots,f_n$ that is closest, in the uniform norm on T, to the target function f_*, i.e., solve the problem*

$$\min_x\left\{\max_{t\in T}|f_*(t) - \sum_{j=1}^n x_j f_j(t)|\right\}. \quad \text{(T)}$$

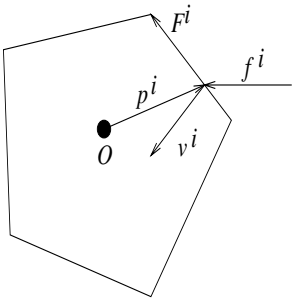


Figure 3.1. *Geometry of* ith *contact.* p^i *is the contact point;* f^i *is the contact force;* v^i *is the inward normal to the surface.*

We have seen that in the case of real-valued functions $f_*, f_1, \ldots, f_n$ the problem can be posed as an LP program. We have also seen that in some applications the functions in question are complex-valued, e.g., in the general antenna synthesis problem (section 1.2.4) and in the filter synthesis problem when the design specifications have to do with the transfer function (section 1.2.3). In these situations, our approach in Lecture 1 was to approximate the modulus of a complex number (i.e., the Euclidean norm of a real 2D vector) by a polyhedral norm—the maximum of several linear functions of the vector. With this approximation, (T) indeed becomes an LP program. If we prefer to avoid approximation, we may easily pose the complex-valued Tschebyshev problem as a conic quadratic program

$$\min_{\tau,x}\left\{\tau \mid \|f_*(t)-\sum_{j=1}^{n}x_j f_j(t)\|_2 \le \tau,\ t\in T\right\}. \tag{3.2.3}$$

In (3.2.3), we treat the complex numbers $f_*(t)$, $f_j(t)$ as real 2D vectors.

3.2.2 Contact problems with static friction

Consider a rigid body in $\mathbf{R}^3$ and a robot with N fingers.[9] When can the robot hold the body? To pose the question mathematically, let us see what happens at the point p^i of the body, which is in contact with the ith finger of the robot (Fig. 3.1). Let f^i be the contact force exerted by the ith finger, v^i be the unit inward normal to the surface of the body at the point p^i, and F^i be the friction force caused by the contact. Physics says that this force is tangential to the surface of the body,

$$(F^i)^T v^i = 0, \tag{3.2.4}$$

and its magnitude cannot exceed μ times the magnitude of the normal component of the contact force, where μ is the friction coefficient:

$$\|F^i\|_2 \le \mu (f^i)^T v^i. \tag{3.2.5}$$

[9]The examples to here, along with their analyses, are taken from M.S. Lobo, L. Vanderbeghe, S. Boyd, and H. Lebret, *Second-order cone programming*, Linear Algebra Appl., 284 (1998), pp. 193–228.

Assume that the body is subject to additional external forces (e.g., gravity). As far as their mechanical consequences are concerned, all these forces can be represented by a single force, their sum $F^{\rm ext}$, along with the *torque* $T^{\rm ext}$, the sum of vector products of the external forces and the points where they are applied.

In order for the body to be in static equilibrium, the total force acting at the body and the total torque should be zero:

$$\begin{array}{rcl} \sum_{i=1}^{N}(f^i + F^i) + F^{\rm ext} & = & 0, \\ \sum_{i=1}^{N} p^i \times (f^i + F^i) + T^{\rm ext} & = & 0, \end{array} \tag{3.2.6}$$

where $p \times q$ stands for the vector product[10] of two 3D vectors p and q.

Stable grasp analysis problem. The question of whether the robot is able to hold the body can be interpreted as follows. Assume that f^i, $F^{\rm ext}$, $T^{\rm ext}$ are given. If the friction forces F^i can adjust themselves to satisfy the friction constraints (3.2.4) and (3.2.5) and the equilibrium equations (3.2.6), i.e., if the system of constraints (3.2.4), (3.2.5), (3.2.6) with respect to unknowns F^i is solvable, then, and only then, the robot holds the body (the body is in a *stable grasp*).

Thus, the question of stable grasp is the question of solvability of the system of constraints

$$\mathrm{S} = (3.2.4)\ \&\ (3.2.5)\ \&\ (3.2.6)$$

with unknowns $F^i \in \mathbf{R}^3$. This question in fact is nothing but a conic quadratic feasibility problem—a conic quadratic problem with a trivial (identically zero) objective. We say "in fact" since the problem, as it arises, is not in our canonical form. This is typical: nice problems normally do not arise in catalogue forms, and one should know how to recognize what one is dealing with. In our case this recognition problem is easy. One way to see that S is a conic quadratic problem is to use the system of linear equations (3.2.4), (3.2.6) to express part of the unknowns via the remaining ones, letting the latter be denoted by x. With this parameterization, every F^i becomes an affine vector-valued function $D_i x - d_i$ of

[10] Here is the definition: if

$$p = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix},\ q = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix}$$

are two 3D vectors, then

$$[p,q] = \begin{pmatrix} \mathrm{Det}\begin{pmatrix} p_2 & p_3 \\ q_2 & q_3 \end{pmatrix} \\ \mathrm{Det}\begin{pmatrix} p_3 & p_1 \\ q_3 & q_1 \end{pmatrix} \\ \mathrm{Det}\begin{pmatrix} p_1 & p_2 \\ q_1 & q_2 \end{pmatrix} \end{pmatrix}.$$

The vector $[p,q]$ is orthogonal to both p and q, and $\|[p,q]\|_2 = \|p\|_2\|q\|_2 \sin(\widehat{pq})$.

the free design vector x, and the question we are interested in becomes whether the primal conic quadratic problem

$$\min_x \left\{0^T x \mid \|D_i x - d_i\|_2 \le \mu (f^i)^T v^i,\ i = 1, \ldots, N\right\}$$

is or is not feasible.

Stable grasp synthesis problems. To the moment we treated the contact forces f^i as given. Sometimes this is not the case, i.e., the robot can, to some extent, control tensions in its fingers. As a simple example, assume that the directions u^i of the contact forces—directions of fingers—are fixed, but the magnitudes of these forces can be controlled:

$$f^i = \nu_i u^i,$$

where the real ν_i are allowed to vary in a given segment $[0, F_{\max}]$. We may ask now whether the robot can choose admissible magnitudes of the contact forces to ensure a stable grasp. Mathematically, the question is whether the system

$$\begin{aligned}
\sum_{i=1}^N (\nu_i u^i + F^i) + F^{\text{ext}} &= 0, \\
\sum_{i=1}^N p_i \times (\nu_i u^i + F^i) + T^{\text{ext}} &= 0, \\
(F^i)^T v^i &= 0, \\
\|F^i\|_2 &\le [\mu (u^i)^T v^i] \nu_i,\ i = 1, \ldots, N, \\
0 \le \nu_i &\le F_{\max},\ i = 1, \ldots, N,
\end{aligned} \tag{3.2.7}$$

in variables ν_i, F^i is solvable. We again come to a conic quadratic feasibility problem. As before, we may eliminate the linear equations to end up with a system of conic quadratic and linear (i.e., also conic quadratic) constraints of the form "the Euclidean norm of something affinely depending on the design variables should be less than or equal to something else, also affinely depending on the design variables."

We could also add to our feasibility problem a meaningful objective function. For example, we may think of the quantity $\sum_{i=1}^N \nu_i$ as a measure of dissipation of power of a robot's actuators and pose the problem of minimizing this objective under the constraints (3.2.7). Another, and perhaps more adequate, measure of dissipation of power is $\sqrt{\sum_{i=1}^N \nu_i^2}$. With this objective, we again end up with a conic quadratic problem

$$\min \{t \mid (3.2.7)\ \&\ \|\nu\|_2 \le t\}, \qquad \nu = (\nu_1, \ldots, \nu_N)^T$$

in the design variables t, $\{\nu_i\}_{i=1}^N$, $\{F^i\}_{i=1}^N$.

As a concluding example of this series, consider the following situation: the robot should hold a cylinder by four fingers, all acting in the vertical direction. The external forces and torques acting at the cylinder are the gravity F_g and an externally applied torque T along the cylinder axis, as shown in Fig. 3.2. The magnitude of the contact forces may

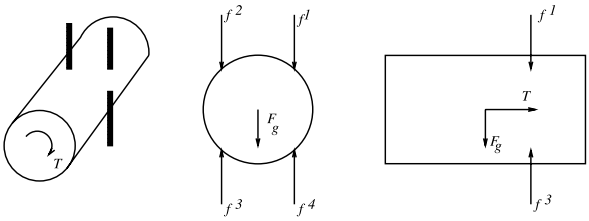


Figure 3.2. *Perspective, front and side views.*

vary in a given segment $[0, F_{\max}]$. The question is, What is the largest magnitude τ of the external torque T such that a stable grasp is still possible? Mathematically, the problem is

$$\begin{array}{lrcl}
\text{maximize} & \tau & & \\
\text{s.t.} & & & \\
& \displaystyle\sum_{i=1}^{4}(v_i u^i + F^i) + F_g & = & 0, \\
& \displaystyle\sum_{i=1}^{4} p^i \times (v_i u^i + F^i) + \tau u & = & 0 \ [u \text{ is the direction of the cylinder axis}], \\
& (v^i)^T F^i & = & 0, \ i = 1, \ldots, 4, \\
& \|F^i\|_2 & \le & [\mu [u^i]^T v^i] v_i, \ i = 1, \ldots, 4, \\
& 0 \le v_i & \le & F_{\max}, \ i = 1, \ldots, 4,
\end{array} \tag{G}$$

where the design variables are τ, v_i, F_i, $i = 1, \ldots, 4$.

3.3 What can be expressed via conic quadratic constraints?

As mentioned, optimization problems arising in applications are not normally in their catalogue forms, and thus an important skill required of anyone interested in applications of optimization is the ability to recognize the fundamental structure beneath the original formulation. The latter is frequently in the form

$$\min_x \{f(x) \mid x \in X\}, \tag{3.3.8}$$

where f is a loss function and the set X of admissible design vectors is typically given as

$$X = \bigcap_{i=1}^{m} X_i, \tag{3.3.9}$$

where every X_i is the set of vectors admissible for a particular design restriction, which in many cases is given by

$$X_i = \{x \in \mathbf{R}^n \mid g_i(x) \le 0\}, \tag{3.3.10}$$

where $g_i(x)$ is the ith *constraint function*.[11] One may interpret $g_i(x)$ as the amount of ith resource required for a design x, so that the constraint $g_i(x) \le$ const says that the resource should not exceed a given level; shifting g_i appropriately, we may make this level 0, thus coming to the representation (3.3.10).

The objective f in (3.3.8)–(3.3.9) may be nonlinear, and one might think that in these cases the problem cannot be posed in conic form. This conclusion is wrong: we can *always* pass from an optimization problem to an equivalent one with a *linear* objective. To this end it suffices to add a new design variable, say, t, and rewrite the problem equivalently as

$$\min_{t,x} \left\{ t \mid (x,t) \in \widehat{X} \equiv \{(x,t) \mid f(x) - t \le 0\} \cap \{(x,t) \mid x \in X_1\} \cap \cdots \cap \{(x,t) \mid x \in X_m\} \right\}.$$

Note that our new objective is linear in the new design variables (x,t), and the resulting problem is in the form of (3.3.8)–(3.3.9).

Let us assume that the indicated transformation was done from the very beginning, so that (3.3.8)–(3.3.9) is of the form

$$\min_x \left\{ c^T x \mid x \in X = \bigcap_{i=1}^m X_i \right\}. \tag{P}$$

In order to recognize that X is in one of our catalogue forms, one needs a kind of dictionary, where different forms of the same structure are listed. We shall build such a dictionary for the conic quadratic programs. Thus, our goal is to understand when a given set X can be represented by *conic quadratic inequalities* (CQIs), i.e., one or several constraints of the type $\|Dx - d\|_2 \le p^T x - q$. The word "represented" needs clarification, and here it is:

> *We say that a set $X \subset \mathbf{R}^n$ can be represented via CQI* (it is CQr—conic quadratic representable) *if there exists a system S of finitely many vector inequalities of the form $A_j\binom{x}{u} - b_j \ge_{\mathbf{L}^{m_j}} 0$ ($x \in \mathbf{R}^n$) in variables $x \in \mathbf{R}^n$ and additional variables u such that X is the projection of the solution set of S onto the x-space, i.e., $x \in X$ if and only if one can extend x to a solution (x,u) of the system S:*
>
> $$x \in X \Leftrightarrow \exists u : A_j \begin{pmatrix} x \\ u \end{pmatrix} - b_j \ge_{\mathbf{L}^{m_j}} 0, \quad j = 1, \ldots, N.$$
>
> *Every such system S is called a conic quadratic representation* (CQR) *of the set X.*[12]

The idea behind this definition is clarified by the following observation:

> *Consider an optimization problem*
>
> $$\min_x \left\{ c^T x \mid x \in X \right\}$$

[11] Speaking about a real-valued function on $\mathbf{R}^n$, we assume that the function is allowed to take real values and the value $+\infty$ and is defined on the entire space. The set of those x where the function is finite is called the domain of the function, denoted by $\mathrm{Dom}\, f$.

[12] Note that here we do not impose on the representing system of conic quadratic inequalities S the requirement to satisfy Assumption A; e.g., the entire space is CQr—it is a solution set of the system $|0^T x| \le 1$ comprising a single CQI.

and assume that X is CQr. Then the problem is equivalent to a conic quadratic program. The latter program can be written explicitly, provided that we are given a CQR of X.

Indeed, let S be a CQR of X, and let u be the corresponding vector of additional variables. The problem

$$\min_{x,u} \left\{ c^T x \mid (x, u) \text{ satisfy } S \right\}$$

with design variables x, u is equivalent to the original problem (P), *on one hand and is a conic quadratic program on the other hand.*

Let us call a problem of the form (P) with CQr X a *good* problem.

How do we recognize good problems, i.e., how do we recognize CQr sets? Well, how do we recognize continuity of a given function, like $f(x, y) = \exp\{\sin(x + \exp\{y\})\}$? Normally it is done not by a straightforward verification of the definition of continuity but by using two kinds of tools:

A. We know a number of simple functions—a constant, $f(x) = x$, $f(x) = \sin(x)$, $f(x) = \exp\{x\}$, etc.—that indeed are continuous: "once for the entire life" we have verified it directly, by demonstrating that the functions fit the definition of continuity.

B. We know a number of basic continuity-preserving operations, like taking products, sums, superpositions, etc.

When we see that a function is obtained from simple functions—those of type A—by operations of type B (as is the case in the above example), we immediately infer that the function is continuous.

This approach, which is common in mathematics, is the one we are about to follow. In fact, we need to answer two questions:

(i) What are CQr sets?

(ii) What are CQr functions $g(x)$, i.e., functions that possess CQr *epigraphs*

$$\mathrm{Epi}\{g\} = \{(x, t) \in \mathbf{R}^n \times \mathbf{R} \mid g(x) \le t\}?$$

Our interest in the second question is motivated by the following observation.

If a function g is CQr, then so are all its level sets $\{x \mid g(x) \le a\}$, and every CQR of (the epigraph of) g explicitly induces CQRs of the level sets.

Indeed, assume that we have a CQR of the epigraph of g:

$$g(x) \le t \Leftrightarrow \exists u : \|\alpha_j(x, t, u)\|_2 \le \beta_j(x, t, u), \quad j = 1, \ldots, N,$$

where α_j and β_j are, respectively, vector-valued and scalar affine functions of their arguments. To get from this representation a CQR of a level set $\{x \mid g(x) \le a\}$, it suffices to fix in the conic quadratic inequalities $\|\alpha_j(x, t, u)\|_2 \le \beta_j(x, t, u)$ the variable t at the value a.

We list below our raw materials—simple functions and sets admitting CQRs.

Elementary conic quadratic-representable functions and sets

1. *A constant function* $g(x) \equiv a$. Indeed, the epigraph of the function $\{(x,t) \mid a \le t\}$ is given by a linear inequality, and a linear inequality $0 \le p^T z - q$ is at the same time conic quadratic inequality $\|0\|_2 \le p^T z - q$.

2. *An affine function* $g(x) = a^T x + b$. Indeed, the epigraph of an affine function is given by a linear inequality.

3. *The Euclidean norm* $g(x) = \|x\|_2$. Indeed, the epigraph of g is given by the conic quadratic inequality $\|x\|_2 \le t$ in variables x, t.

4. *The squared Euclidean norm* $g(x) = x^T x$. Indeed, $t = \frac{(t+1)^2}{4} - \frac{(t-1)^2}{4}$, so that

$$x^T x \le t \Leftrightarrow x^T x + \frac{(t-1)^2}{4} \le \frac{(t+1)^2}{4} \Leftrightarrow \left\| \begin{pmatrix} x \\ \frac{t-1}{2} \end{pmatrix} \right\|_2 \le \frac{t+1}{2}$$

(check the second $\Leftrightarrow$!), and the last relation is a CQI.

5. *The fractional-quadratic function*

$$g(x,s) = \begin{cases} \frac{x^T x}{s}, & s > 0 \\ 0, & s = 0, x = 0 \\ +\infty & \text{otherwise} \end{cases}$$

(x vector, s scalar). Indeed, with the convention that $(x^T x)/0$ is 0 or $+\infty$, depending on whether $x = 0$, and taking into account that $ts = \frac{(t+s)^2}{4} - \frac{(t-s)^2}{4}$, we have

$$\{\tfrac{x^T x}{s} \le t, s \ge 0\} \Leftrightarrow \{x^T x \le ts, t \ge 0, s \ge 0\} \Leftrightarrow \{x^T x + \tfrac{(t-s)^2}{4} \le \tfrac{(t+s)^2}{4}, t \ge 0, s \ge 0\}$$
$$\Leftrightarrow \left\| \begin{pmatrix} x \\ \frac{t-s}{2} \end{pmatrix} \right\|_2 \le \tfrac{t+s}{2}$$

(check the third $\Leftrightarrow$!), and the last relation is a CQI.

The level sets of the CQr functions 1–5 provide us with a spectrum of elementary CQr sets. We add to this spectrum one more set:

6. *(A branch of) hyperbola* $\{(t,s) \in \mathbf{R}^2 \mid ts \ge 1, t > 0\}$. Indeed,

$$\{ts \ge 1, t > 0\} \Leftrightarrow \{\tfrac{(t+s)^2}{4} \ge 1 + \tfrac{(t-s)^2}{4} \ \&\ t > 0\} \Leftrightarrow \left\{ \left\| \begin{pmatrix} \frac{t-s}{2} \\ 1 \end{pmatrix} \right\|_2^2 \le \tfrac{(t+s)^2}{4} \right\}$$
$$\Leftrightarrow \left\{ \left\| \begin{pmatrix} \frac{t-s}{2} \\ 1 \end{pmatrix} \right\|_2 \le \tfrac{t+s}{2} \right\}$$

(check the last $\Leftrightarrow$!), and the latter relation is a CQI.

Operations preserving conic quadratic representability of sets

Next we study simple operations preserving CQ-representability of functions and sets.

A. *Intersection.* If sets $X_i \subset \mathbf{R}^n$, $i = 1, \ldots, N$, are CQr, so is their intersection $X = \bigcap_{i=1}^N X_i$.

Indeed, let S_i be a CQR of X_i and u_i be the corresponding vector of additional variables. Then the system S of constraints of the variables $(x, u_1, \ldots, u_N)$,

$$\{(x, u_i) \text{ satisfies } S_i\}, \ i = 1, \ldots, N,$$

is a system of CQIs, and this system clearly is a CQR of X.

Corollary 3.3.1. *A polyhedral set—a set in $\mathbf{R}^n$ given by finitely many linear inequalities $a_i^T x \le b_i$, $i = 1, \ldots, m$—is CQr.*

> Indeed, a polyhedral set is the intersection of finitely many level sets of affine functions, and all these functions (and thus their level sets) are CQr.

Corollary 3.3.2. *If every one of the sets X_i in problem* (P) *is CQr, then the problem is good—it can be rewritten in the form of a conic quadratic problem, and such a transformation is readily given by CQRs of the sets X_i, $i = 1, \ldots, m$.*

Corollary 3.3.3. *Adding to a good problem finitely many CQr constraints $x \in X_i$, (e.g., finitely many scalar linear inequalities), we again get a good problem.*

B. *Direct product.* If sets $X_i \subset \mathbf{R}^{n_i}$, $i = 1, \ldots, k$, are CQr, then so is their direct product $X_1 \times \cdots \times X_k$.

Indeed, if $S_i = \{\|\alpha_j^i(x_i, u_i)\|_2 \le \beta_j^i(x_i, u_i)\}_{j=1}^{N_j}$, $i = 1, \ldots, k$, are CQRs of the sets X_i, then the union over i of this system of inequalities, regarded as a system with design variables $x = (x_1, \ldots, x_k)$ and additional variables $u = (u_1, \ldots, u_k)$, is a CQR for the direct product of $X_1, \ldots, X_k$.

C. *Affine image (projection).* If a set $X \subset \mathbf{R}^n$ is CQr and $x \mapsto y = \ell(x) = Ax + b$ is an affine mapping of $\mathbf{R}^n$ to $\mathbf{R}^k$, then the image $\ell(X)$ of the set X under the mapping is CQr.

Indeed, passing to an appropriate basis in $\mathbf{R}^n$ and $\mathbf{R}^k$, we may assume that the null space of A is made up of the last $n - p$ vectors of the basis of $\mathbf{R}^n$ and that the image of A is spanned by the first p vectors of the basis in $\mathbf{R}^k$. In other words, we may assume that a vector $x \in \mathbf{R}^n$ can be partitioned as $x = \binom{x'}{x''}$ (x' is p-dimensional and x'' is $(n-p)$-dimensional) and that a vector $y \in \mathbf{R}^k$ can be partitioned as $y = \binom{y'}{y''}$ (y' is p-dimensional and y'' is $(k-p)$-dimensional) in such a way that $A\binom{x'}{x''} = \binom{Qx'}{0}$ with a nonsingular $p \times p$ matrix Q. Thus,

$$\left\{ \begin{pmatrix} y' \\ y'' \end{pmatrix} = A \begin{pmatrix} x' \\ x'' \end{pmatrix} + b \right\} \Leftrightarrow \left\{ x = \begin{pmatrix} Q^{-1}(y' - b') \\ w \end{pmatrix} \text{ for some } w \ \& \ y'' = b'' \right\}.$$

Now let $S = \{\|\alpha_j(x, u)\|_2 \le \beta_j(x, u)\}_{j=1}^N$ be a CQR of X, where u is the corresponding vector of additional variables and α_j, β_j are affine in (x, u). Then the system of CQIs in

the design variables $y = \binom{y'}{y''} \in \mathbf{R}^k$ and additional variables $w \in \mathbf{R}^{n-p}$, u,

$$S^+ = \left\{ \|\alpha_j \left(\begin{pmatrix} Q^{-1}(y' - b') \\ w \end{pmatrix}, u \right) \|_2 \le \beta_j \left(\begin{pmatrix} Q^{-1}(y' - b') \\ w \end{pmatrix}, u \right) \right\}_{j=1}^N \\ \text{and } \{\|y'' - b''\|_2 \le 0\},$$

is a CQR of $\ell(X)$. Indeed, $y = \binom{y'}{y''} \in \ell(X)$ if and only if $y'' = b''$ and there exists $w \in \mathbf{R}^{n-p}$ such that the point

$$x = \begin{pmatrix} Q^{-1}(y' - b') \\ w \end{pmatrix}$$

belongs to X, and the latter happens if and only if there exist u such that the point

$$(x, u) = \left(\begin{pmatrix} Q^{-1}(y' - b') \\ w \end{pmatrix}, u \right)$$

solves S.

Corollary 3.3.4. *A nonempty set X is CQr if and only if its characteristic function*

$$\chi(x) = \begin{cases} 0, & x \in X, \\ +\infty & \text{otherwise} \end{cases}$$

is CQr.

Indeed, Epi$\{\chi\}$ is the direct product of X and the nonnegative ray; therefore if X is CQr, so is $\chi(\cdot)$ (see B and Corollary 3.3.1). Conversely, if χ is CQr, then X is CQr by C, since X is the projection of the Epi$\{\chi\}$ on the space of x-variables.

D. *Inverse affine image*. Let $X \subset \mathbf{R}^n$ be a CQr set, and let $\ell(y) = Ay + b$ be an affine mapping from $\mathbf{R}^k$ to $\mathbf{R}^n$. Then the inverse image $\ell^{-1}(X) = \{y \in \mathbf{R}^k \mid Ay + b \in X\}$ is also CQr.

Indeed, let $S = \{\|\alpha_j(x, u)\|_2 \le \beta_j(x, u)\}_{i=1}^N$ be a CQR for X. Then the system of CQIs

$$S = \{\|\alpha_j(Ay + b, u)\|_2 \le \beta_j(Ay + b, u)\}_{i=1}^N$$

with variables y, u clearly is a CQR for $\ell^{-1}(X)$.

Corollary 3.3.5. *Consider a good problem* (P) *and assume that we restrict its design variables to be given affine functions of a new design vector y. Then the induced problem with the design vector y is also good.*

In particular, adding to a good problem arbitrarily many linear equality constraints, we end up with a good problem. (Indeed, we may use the linear equations to express affinely the original design variables via part of them; let this part be y. The problem with added linear constraints can now be posed as a problem with design vector y.)

It should be stressed that the above statements are not just existence theorems—they are algorithmic: given CQRs of the operands (say, m sets $X_1, \ldots, X_m$), we may build *completely mechanically* a CQR for the result of the operation (e.g., for the intersection $\bigcap_{i=1}^m X_i$).

Note that we have already used nearly all our corollaries in the grasp problem. To see that (G) is a conic quadratic problem, we carried out the following reasoning:

1. The problem

$$\min\left\{\tau \mid \|F^i\|_2 \le s_i,\ i = 1, \ldots, N\right\} \tag{P_0}$$

in the design variables τ, $(F^i, s_i) \in \mathbf{R}^3 \times \mathbf{R}$, $v_i \in \mathbf{R}$ is perhaps odd (part of the variables does not appear at all, the objective and the constraints are not related to each other, etc.), but clearly it is good.

2. Adding to the good problem (P_0) the linear equality constraints

$$\begin{aligned}
\sum_{i=1}^N (v_i u^i + F^i) &= -F_g,\\
\sum_{i=1}^N p^i \times (v^i u^i + F^i) + \tau u &= 0,\\
(v^i)^T F^i &= 0,\ i = 1, \ldots, N,\\
s_i - [\mu (u^i)^T v^i] v_i &= 0,\ i = 1, \ldots, N,
\end{aligned}$$

where u^i, u, F_g are given vectors, we get a good problem (P_1) (Corollary 3.3.5).

3. The original problem (G) is obtained from the good problem (P_1) by adding scalar linear inequalities

$$0 \le v_i \le F_{\max},\ i = 1, \ldots, N,$$

so that (G) itself is good (Corollary 3.3.3).

Operations preserving conic quadratic-representability of functions

Recall that a function $g(x)$ is called CQr if its epigraph $\mathrm{Epi}\{g\} = \{(x, t) \mid g(x) \le t\}$ is a CQr set; a CQR of the epigraph of g is called CQR of g. Recall also that a level set of a CQr function is CQr. Here are transformations preserving CQ-representability of functions.

E. *Taking maximum.* If functions $g_i(x)$, $i = 1, \ldots, m$, are CQr, then so is their maximum $g(x) = \max_{i=1,\ldots,m} g_i(x)$.

Indeed, $\mathrm{Epi}\{g\} = \bigcap_i \mathrm{Epi}\{g_i\}$ and the intersection of finitely many CQr sets again is CQr.

F. *Summation with nonnegative weights.* If functions $g_i(x)$, $x \in \mathbf{R}^n$, are CQr, $i = 1, \ldots, m$, and α_i are nonnegative weights, then the function $g(x) = \sum_{i=1}^m \alpha_i g_i(x)$ is also CQr.

Indeed, consider the set

$$\Pi = \left\{ (x_1, t_1; x_2, t_2; \ldots; x_m, t_m; t) \mid x_i \in \mathbf{R}^n, t_i \in \mathbf{R}, t \in \mathbf{R}, g_i(x_i) \right.$$
$$\left. \leq t_i, i = 1, \ldots, m; \sum_{i=1}^{m} \alpha_i t_i \leq t \right\}.$$

The set is CQr. Indeed, the set is the direct product of the epigraphs of g_i intersected with the half-space given by the linear inequality $\sum_{i=1}^m \alpha_i t_i \leq t$. Now, a direct product of CQr sets is also CQr, a half-space is CQr (it is a level set of an affine function, and such a function is CQr), and the intersection of CQr sets is also CQr. Since Π is CQr, so is its projection on subspace of variables $x_1, x_2, \ldots, x_m, t$, i.e., the set

$$\left\{ (x_1, \ldots, x_m, t) : \exists t_1, \ldots, t_m : g_i(x_i) \leq t_i, i = 1, \ldots, m, \sum_{i=1}^{m} \alpha_i t_i \leq t \right\}$$
$$= \left\{ (x_1, \ldots, x_m, t) : \sum_{i=1}^{m} \alpha_i g_i(x_i) \leq t \right\}.$$

Since the latter set is CQr, so is its inverse image under the mapping

$$(x, t) \mapsto (x, x, \ldots x, t),$$

and this inverse image is exactly the epigraph of g.

G. *Direct summation*. If functions $g_i(x_i), x_i \in \mathbf{R}^{n_i}, i = 1, \ldots, m$, are CQr, so is their direct sum

$$g(x_1, \ldots, x_m) = g_1(x_1) + \cdots + g_m(x_m).$$

Indeed, the functions $\hat{g}_i(x_1, \ldots, x_m) = g_i(x_i)$ are clearly CQr—their epigraphs are inverse images of the epigraphs of g_i under the affine mappings $(x_1, \ldots, x_m, t) \mapsto (x_i, t)$. It remains to note that $g = \sum_i \hat{g}_i$.

H. *Affine substitution of argument*. If a function $g(x)$, $x \in \mathbf{R}^n$, is CQr and $y \mapsto Ay + b$ is an affine mapping from $\mathbf{R}^k$ to $\mathbf{R}^n$, then the superposition $g^{\rightarrow}(y) = g(Ay + b)$ is CQr.

Indeed, the epigraph of $g^{\rightarrow}$ is the inverse image of the epigraph of g under the affine mapping $(y, t) \mapsto (Ay + b, t)$.

I. *Partial minimization*. Let $g(x)$ be CQr. Assume that x is partitioned into two subvectors: $x = (v, w)$, let $\hat{g}$ be obtained from g by partial minimization in w,

$$\hat{g}(v) = \inf_w g(v, w),$$

and assume that for every v the minimum in w is achieved. Then $\hat{g}$ is CQr.

Indeed, under the assumption that the minimum in w always is achieved, $\mathrm{Epi}\{\hat{g}\}$ is the image of $\mathrm{Epi}\{g\}$ under the projection $(v, w, t) \mapsto (v, t)$.

More operations preserving conic quadratic-representability

Let us list a number of more advanced operations with sets and functions that preserve CQ-representability.

J. *Arithmetic summation of sets.* Let X_i, $i = 1, \ldots, k$, be nonempty convex sets in $\mathbf{R}^n$, and let $X_1 + X_2 + \cdots + X_k$ be the arithmetic sum of these sets:

$$X_1 + \cdots + X_k = \{x = x^1 + \cdots + x^k \mid x^i \in X_i,\ i = 1, \ldots, k\}.$$

We claim that if all X_i are CQr, so is their sum.

Indeed, the direct product

$$X = X_1 \times X_2 \times \cdots \times X_k \subset \mathbf{R}^{nk}$$

is CQr by B; it remains to note that $X_1 + \cdots + X_k$ is the image of X under the linear mapping

$$(x^1, \ldots, x^k) \mapsto x^1 + \cdots + x^k : \mathbf{R}^{nk} \to \mathbf{R}^n,$$

and by C the image of a CQr set under an affine mapping is also CQr (see C)

J.1. inf-*convolution.* The operation with functions related to the arithmetic summation of sets is the inf-convolution defined as follows. Let $f_i : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$, $i = 1, \ldots, n$, be functions. Their inf-convolution is the function

$$f(x) = \inf\{f_1(x^1) + \cdots + f_k(x^k) \mid x^1 + \cdots + x^k = x\}. \qquad (*)$$

We claim that

> *if all f_i are CQr, their* inf-*convolution is $> -\infty$ everywhere and for every x for which the* inf *in the right-hand side of $(*)$ is finite, this infimum is achieved, then f is CQr.*

Indeed, under the assumption in question the epigraph $\mathrm{Epi}\{f\} = \mathrm{Epi}\{f_1\} + \cdots + \mathrm{Epi}\{f_k\}$.

K. *Taking conic hull of a closed convex set.* Let $X \in \mathbf{R}^n$ be a nonempty convex set. Its *conic hull* is the set

$$X^+ = \{(x, t) \in \mathbf{R}^n \times \mathbf{R} : t > 0, t^{-1}x \in X\} \cup \{0\}.$$

Geometrically, we add to the coordinates of vectors from $\mathbf{R}^n$ a new coordinate equal to 1,

$$(x_1, \ldots, x_n)^T \mapsto (x_1, \ldots, x_n, 1)^T,$$

thus getting an affine embedding of $\mathbf{R}^n$ in $\mathbf{R}^{n+1}$. We take the image of X under this mapping—lift X by one along the $(n+1)$st axis—and then form the set X^+ by taking all rays emanating from the origin and crossing the lifted X.

The conic hull of a closed convex set X is not necessarily closed; to maintain closedness, X has to be both closed and bounded. The closed convex hull of X is the closure of its conic hull:

$$\widehat{X}^+ = \mathrm{cl}X^+ = \left\{(x,t) \in \mathbf{R}^n \times \mathbf{R} : \exists\{(x_i,t_i)\}_{i=1}^\infty : t_i > 0, t_i^{-1}x_i \in X, t = \lim_i t_i, x = \lim_i x_i\right\}.$$

Note that if X is a closed convex set, then the parts of the conic hull X^+ of X and the closed convex hull $\widehat{X}^+$ belonging to the open half-space $\{t > 0\}$ are equal to each other (check!). Note also that if X is a closed convex set, you can obtain it from its (closed) convex hull by taking intersection with the hyperplane $\{t = 1\}$:

$$x \in X \Leftrightarrow (x,1) \in \widehat{X}^+ \Leftrightarrow (x,1) \in X^+.$$

Proposition 3.3.1. (i) *If X is nonempty and X is CQr,*

$$X = \{x \mid \exists u : Ax + Bu + b \geq_{\mathbf{K}} 0\}, \tag{3.3.11}$$

where $\mathbf{K}$ is a direct product of the ice cream cones, then the CQr set

$$\widetilde{X}^+ = \{(x,t) \mid \exists u : Ax + Bu + tb \geq_{\mathbf{K}} 0\} \bigcap \{(x,t) \mid t \geq 0\} \tag{3.3.12}$$

is between the conic hull X^+ and the closed conic hull $\widehat{X}^+$ of X:

$$X^+ \subset \widetilde{X}^+ \subset \widehat{X}^+.$$

In particular, if X is a closed and bounded CQr set (so that $X^+ = \widehat{X}^+$), then the conic hull of X is CQr.

(ii) *If the CQR* (3.3.11) *is such that $Bu \in \mathbf{K}$ implies that $Bu = 0$, then $\widetilde{X}^+ = \widehat{X}^+$, so that $\widehat{X}^+$ is CQr.*

Proof. (i): We should prove that the set $\widetilde{X}^+$ (which by construction is CQr) is between X^+ and $\widehat{X}^+$. First, $0 \in \widetilde{X}^+$, and if $(x,t) \in X^+\backslash\{0\}$, i.e., $t > 0$, $z = t^{-1}x \in X$, then there exists u such that

$$Az + Bu + b \geq_{\mathbf{K}} 0 \Rightarrow Ax + B(tu) + tb \geq_{\mathbf{K}} 0 \Rightarrow (x,t) \in \widetilde{X}^+.$$

Thus, $X^+ \subset \widetilde{X}^+$.

Next, let us prove that $\widetilde{X}^+ \subset \widehat{X}^+$. Let us choose a point $\bar{x} \in X$, so that for a properly chosen $\bar{u}$,

$$A\bar{x} + B\bar{u} + b \geq_{\mathbf{K}} 0$$

holds, i.e., $(\bar{x},1) \in \widetilde{X}^+$. From the description of $\widetilde{X}^+$ it is clear that whenever (x,t) belongs to $\widetilde{X}^+$, so does every pair $(x_\epsilon = x + \epsilon\bar{x}, t_\epsilon = t + \epsilon)$ with $\epsilon > 0$:

$$\exists u = u_\epsilon : Ax_\epsilon + Bu_\epsilon + t_\epsilon b \geq_{\mathbf{K}} 0.$$

It follows that $t_\epsilon^{-1}x_\epsilon \in X$, whence $(x_\epsilon, t_\epsilon) \in X^+ \subset \widehat{X}^+$. As $\epsilon \to +0$, we have $(x_\epsilon, t_\epsilon) \to (x,t)$, and since $\widehat{X}^+$ is closed, we get $(x,t) \in \widehat{X}^+$. Thus, $\widetilde{X}^+ \subset \widehat{X}^+$.

(ii): Assume that $Bu \in \mathbf{K}$ only if $Bu = 0$, and let us show that $\widetilde{X}^+ = \widehat{X}^+$. We just have to prove that $\widetilde{X}^+$ is closed, which indeed is the case due to the following lemma.

Lemma 3.3.1. *Let Y be a CQr set with CQR*

$$Y = \{y \mid \exists v : Py + Qv + r \geq_{\mathbf{K}} 0\}$$

such that $Qv \in \mathbf{K}$ only when $Qv = 0$. Then

(i) *there exists a constant $C < \infty$ such that*

$$Py + Qv + r \in \mathbf{K} \Rightarrow \|Qv\|_2 \leq C(1 + \|Py + r\|_2); \tag{3.3.13}$$

(ii) *Y is closed.*

Proof. (i): Assume, on the contrary to what should be proved, that there exists a sequence $\{y_i, v_i\}$ such that

$$Py_i + Qv_i + r \in \mathbf{K}, \ \|Qv_i\|_2 \geq \alpha_i(1 + \|Py_i + r\|_2), \ \alpha_i \to \infty \text{ as } i \to \infty. \tag{3.3.14}$$

By linear algebra, for every b such that the linear system $Qv = b$ is solvable, it admits a solution v such that $\|v\|_2 \leq C_1\|b\|_2$ with $C_1 < \infty$ depending on Q only; therefore we can assume, in addition to (3.3.14), that

$$\|v_i\|_2 \leq C_1\|Qv_i\|_2 \tag{3.3.15}$$

for all i. Now, from (3.3.14) it clearly follows that

$$\|Qv_i\|_2 \to \infty \text{ as } i \to \infty; \tag{3.3.16}$$

setting

$$\widehat{v}_i = \frac{1}{\|Qv_i\|_2} v_i,$$

we have

(a)	$\|Q\widehat{v}_i\|_2 = 1 \quad \forall i,$	
(b)	$\|\widehat{v}_i\| \leq C_1 \quad \forall i$	[by (3.3.15)],
(c)	$Q\widehat{v}_i + \|Qv_i\|_2^{-1}(Py_i + r) \in \mathbf{K} \quad \forall i,$	
(d)	$\|Qv_i\|_2^{-1}\|Py_i + r\|_2 \leq \alpha_i^{-1} \to 0 \text{ as } i \to \infty$	[by (3.3.14)].

Taking into account (b) and passing to a subsequence, we can assume that $\widehat{v}_i \to \widehat{v}$ as $i \to \infty$; by (c), (d) $Q\widehat{v} \in \mathbf{K}$, while by (a) $\|Q\widehat{v}\|_2 = 1$, i.e., $Q\widehat{v} \neq 0$, which is the desired contradiction.

(ii) To prove that Y is closed, assume that $y_i \in Y$ and $y_i \to y$ as $i \to \infty$, and let us verify that $y \in Y$. Indeed, since $y_i \in Y$, there exist v_i such that $Py_i + Qv_i + r \in \mathbf{K}$. Same as above, we can assume that (3.3.15) holds. Since $y_i \to y$, the sequence $\{Qv_i\}$ is bounded by (3.3.13), so that the sequence $\{v_i\}$ is bounded by (3.3.15). Passing to a subsequence,

we can assume that $v_i \to v$ as $i \to \infty$; passing to the limit, as $i \to \infty$, in the inclusion $Py_i + Qv_i + r \in \mathbf{K}$, we get $Py + Qv + r \in \mathbf{K}$, i.e., $y \in Y$. □

K.1. *Projective transformation of a CQr function.* The operation with functions related to taking conic hull of a convex set is the projective transformation, which converts a function $f(x) : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ into the function

$$f^+(x, s) = sf(x/s) : \{s > 0\} \times \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}.$$

The epigraph of f^+ is the conic hull of the epigraph of f with the origin excluded:

$$\begin{array}{rcl} \left\{(x, s, t) \mid s > 0, t \geq f^+(x, s)\right\} & = & \left\{(x, s, t) \mid s > 0, s^{-1}t \geq f(s^{-1}x)\right\} \\ & = & \left\{(x, s, t) \mid s > 0, s^{-1}(x, t) \in \mathrm{Epi}\{f\}\right\}. \end{array}$$

The closure $\mathrm{clEpi}\{f^+\}$ is the epigraph of a certain function, let it be denoted $\widehat{f^+}(x, s)$; this function is called the *projective transformation* of f. The fractional-quadratic function from example 5 is the projective transformation of the function $f(x) = x^T x$. Note that the function $\widehat{f^+}(x, s)$ does not necessarily coincide with $f^+(x, s)$ even in the open half-space $s > 0$. This is the case if and only if the epigraph of f is closed (or, which is the same, f is lower semicontinuous: whenever $x_i \to x$ and $f(x_i) \to a$, we have $f(x) \leq a$). We are about to demonstrate that the projective transformation nearly preserves CQ-representability.

Proposition 3.3.2. *Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ be a lower semicontinuous function that is CQr:*

$$\mathrm{Epi}\{f\} \equiv \{(x, t) \mid t \geq f(x)\} = \{(t, x) \mid \exists u : Ax + tp + Bu + b \geq_{\mathbf{K}} 0\}, \qquad (3.3.17)$$

where $\mathbf{K}$ is a direct product of ice cream cones. Assume that the CQR is such that $Bu \geq_{\mathbf{K}} 0$ implies that $Bu = 0$. Then the projective transformation $\widehat{f^+}$ of f is CQr, namely,

$$\mathrm{Epi}\{\widehat{f^+}\} = \{(x, t, s) \mid s \geq 0, \exists u : Ax + tp + Bu + sb \geq_{\mathbf{K}} 0\}.$$

Indeed, let us set

$$G = \{(x, t, s) \mid \exists u : s \geq 0, Ax + tp + Bu + sb \geq_{\mathbf{K}} 0\}.$$

As we remember from the previous combination rule, G is exactly the closed conic hull of the epigraph of f, i.e., $G = \mathrm{Epi}\{\widehat{f^+}\}$.

L. *The polar of a convex set.* Let $X \subset \mathbf{R}^n$ be a convex set containing the origin. The *polar* of X is the set

$$X_* = \left\{y \in \mathbf{R}^n \mid y^T x \leq 1 \ \forall x \in X\right\}.$$

In particular,

- the polar of the singleton $\{0\}$ is the entire space;
- the polar of the entire space is the singleton $\{0\}$;

- the polar of a linear subspace is its orthogonal complement (why?);
- the polar of a closed convex pointed cone K with a nonempty interior is $-K_*$, *minus* the dual cone (why?).

Polarity is symmetric: if X is a closed convex set containing the origin, then so is X_*, and twice taken polar is the original set—$(X_*)_* = X$.

We are about to prove that the polarity $X \mapsto X_*$ nearly preserves CQ-representability.

Proposition 3.3.3. *Let $X \subset \mathbf{R}^n$, $0 \in X$, be a CQr set,*

$$X = \{x \mid \exists u : Ax + Bu + b \geq_{\mathbf{K}} 0\}, \tag{3.3.18}$$

where $\mathbf{K}$ is a direct product of ice cream cones. Assume that the above CQR is strictly feasible, i.e., that there exists $\bar{x}, \bar{u}$ such that

$$A\bar{x} + B\bar{u} + b >_{\mathbf{K}} 0.$$

Then the polar of X is the CQr set

$$X_* = \left\{y \mid \exists \xi : A^T\xi + y = 0, B^T\xi = 0, b^T\xi \leq 1, \xi \geq_{\mathbf{K}} 0\right\}. \tag{3.3.19}$$

Indeed, consider the following conic quadratic problem:

$$\min_{x,u} \left\{-y^T x \mid Ax + Bu + b \geq_{\mathbf{K}} 0\right\}. \tag{P_y}$$

A vector y belongs to X_* if and only if (P_y) is bounded below and its optimal value is at least -1. Since (P_y) is strictly feasible, from the conic duality theorem it follows that these properties of (P_y) hold if and only if the dual problem

$$\max_{\xi} \left\{-b^T\xi \mid A^T\xi = -y, B^T\xi = 0, \xi \geq_{\mathbf{K}} 0\right\}$$

(recall that $\mathbf{K}$ is self-dual) has a feasible solution with the value of the dual objective at least -1. Thus,

$$X_* = \left\{y \mid \exists \xi : A^T\xi + y = 0, B^T\xi = 0, b^T\xi \leq 1, \xi \geq_{\mathbf{K}} 0\right\},$$

as claimed in (3.3.19). It remains to note that X_* is obtained from the CQr set $\mathbf{K}$ by operations preserving CQ-representability: intersection with the CQr set $\{\xi \mid B^T\xi = 0, b^T\xi \leq 1\}$ and subsequent affine mapping $\xi \mapsto -A^T\xi$.

L.1. *The Legendre transformation of a CQr function.* The operation with functions related to taking polar of a convex set is the *Legendre* (or *conjugate*) *transformation*. The Legendre transformation ($\equiv$ the conjugate) of a function $f(x) : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ is the function

$$f_*(y) = \sup_x \left[y^T x - f(x)\right].$$

In particular,

- the conjugate of a constant $f(x) \equiv c$ is the function
$$f_*(y) = \begin{cases} -c, & y = 0, \\ +\infty, & y \neq 0; \end{cases}$$
- the conjugate of an affine function $f(x) \equiv a^T x + b$ is the function
$$f_*(y) = \begin{cases} -b, & y = a, \\ +\infty, & y \neq a; \end{cases}$$
- the conjugate of a convex quadratic form $f(x) \equiv \frac{1}{2} x^T D^T D x + b^T x + c$ with rectangular D such that $\mathrm{Null}(D^T) = \{0\}$ is the function
$$f_*(y) = \begin{cases} \frac{1}{2}(y-b)^T D^T (DD^T)^{-2} D(y-b) - c, & y - b \in \mathrm{Im} D^T, \\ +\infty & \text{otherwise.} \end{cases}$$

It is worth mentioning that the Legendre transformation is symmetric: if f is a proper convex lower semicontinuous function (i.e., $\mathrm{Epi}\{f\}$ is nonempty, convex and closed), then so is f_*. Taken twice, the Legendre transformation recovers the original function: $(f_*)_* = f$.

We are about to prove that the Legendre transformation nearly preserves CQ-representability.

Proposition 3.3.4. *Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ be CQr:*

$$\{(x,t) \mid t \geq f(x)\} = \{(t,x) \mid \exists u : Ax + tp + Bu + b \geq_{\mathbf{K}} 0\},$$

where $\mathbf{K}$ is a direct product of ice cream cones. Assume that the above CQR is strictly feasible:

$$\exists \bar{x}, \bar{t}, \bar{u} : \quad A\bar{x} + \bar{t}p + B\bar{u} + b >_{\mathbf{K}} 0.$$

Then the Legendre transformation of f is CQr:

$$\mathrm{Epi}\{f_*\} = \left\{(y,s) \mid \exists \xi : A^T\xi = -y,\, B^T\xi = 0,\, p^T\xi = 1,\, s \geq b^T\xi,\, \xi \geq_{\mathbf{K}} 0\right\}. \quad (3.3.20)$$

Indeed, we have

$$\mathrm{Epi}\{f_*\} = \left\{(y,s) \mid y^T x - f(x) \leq s \ \forall x\right\} = \left\{(y,s) \mid y^T x - t \leq s \ \forall (x,t) \in \mathrm{Epi}\{f\}\right\}. \quad (3.3.21)$$

Consider the conic quadratic program

$$\min_{x,t,u} \left\{-y^T x + t \mid Ax + tp + Bu + b \geq \mathbf{K}_0\right\}. \quad (\mathrm{P}_y)$$

By (3.3.21), a pair (y,s) belongs to $\mathrm{Epi}\{f_*\}$ if and only if (P_y) is bounded below with optimal value $\geq -s$. Since (P_y) is strictly feasible, this is the case if and only if the dual problem

$$\max_{\xi} \left\{-b^T\xi \mid A^T\xi = -y,\, B^T\xi = 0,\, p^T\xi = 1,\, \xi \geq_{\mathbf{K}} 0\right\}$$

has a feasible solution with the value of the dual objective $\geq -s$. Thus,

$$\mathrm{Epi}\{f_*\} = \left\{(y,s) \mid \exists \xi : A^T\xi = -y,\ B^T\xi = 0,\ p^T\xi = 1,\ s \geq b^T\xi,\ \xi \geq_{\mathbf{K}} 0\right\}$$

as claimed in (3.3.20). It remains to note that the right-hand side set in (3.3.20) is CQr (as a set obtained from the CQr set $\mathbf{K} \times \mathbf{R}_s$ by operations preserving CQ-representability—intersection with the set $\{\xi \mid B^T\xi = 0,\ p^T\xi = 1, b^T\xi \leq s\}$ and subsequent affine mapping $\xi \mapsto -A^T\xi$).

Corollary 3.3.6. *Let X be a CQr set with a strictly feasible CQR:*

$$X = \{x \mid \exists u : Ax + Bu + b \geq_{\mathbf{K}} 0\}, \quad \exists \bar{x}, \bar{u} :\ A\bar{x} + B\bar{u} + b >_{\mathbf{K}} 0,$$

where $\mathbf{K}$ is a direct product of ice cream cones. Then the support function

$$\mathrm{Supp}_X(x) = \sup_{y \in X} x^T y$$

of X with a strictly feasible CQR is CQr.

Indeed, $\mathrm{Supp}_X(\cdot)$ is the conjugate of the characteristic function $\chi_X(\cdot)$ of X, and the latter, under the premise of the corollary, admits a strictly feasible CQR, namely

$$\mathrm{Epi}\{\chi_X\} = \{(x,t) \mid \exists u : Ax + Bu + b \geq_{\mathbf{K}} 0, t \geq 0\}.$$

M. *Taking convex hull of several sets.* The convex hull of a set $Y \subset \mathbf{R}^n$ is the smallest convex set that contains Y:

$$\mathrm{Conv}(Y) = \left\{x = \sum_{i=1}^{k_x} \alpha_i x_i \mid x_i \in Y, \alpha_i \geq 0, \sum_i \alpha_i = 1\right\}.$$

The closed convex hull $\overline{\mathrm{Conv}}(Y) = \mathrm{clConv}(Y)$ of Y is the smallest *closed* convex set containing Y.

Following Nesterov, let us prove that taking the convex hull nearly preserves CQ-representability.

Proposition 3.3.5. *Let $X_1, \ldots, X_k \subset \mathbf{R}^n$ be nonempty convex CQr sets:*

$$X_i = \{x \mid A_i x + B_i u_i + b_i \geq_{\mathbf{K}_i} 0,\ i = 1, \ldots, k\}, \tag{3.3.22}$$

where $\mathbf{K}_i$ is a direct product of ice cream cones.

Then the CQr set

$$\begin{array}{rcl} Y & = & \Big\{x \mid \exists \xi^1, \ldots, \xi^k, t_1, \ldots, t_k, \eta^1, \ldots, \eta^k : \\ & & \left[\begin{array}{c} A_1\xi^1 + B_1\eta^1 + t_1 b_1 \\ A_2\xi^2 + B_2\eta^2 + t_2 b_2 \\ \ldots \\ A_k\xi^k + B_k\eta^k + t_k b_k \end{array}\right] \geq_{\mathbf{K}} 0, \\ & & \begin{array}{rcl} t_1, \ldots, t_k & \geq & 0, \\ \xi^1 + \cdots + \xi^k & = & x \\ t_1 + \cdots + t_k & = & 1\Big\}, \end{array} \\ \mathbf{K} & = & \mathbf{K}_1 \times \cdots \times \mathbf{K}_k, \end{array} \tag{3.3.23}$$

is between the convex hull and the closed convex hull of the set $X_1 \cup \cdots \cup X_k$:

$$\mathrm{Conv}\left(\bigcup_{i=1}^k X_i\right) \subset Y \subset \overline{\mathrm{Conv}}\left(\bigcup_{i=1}^k X_i\right).$$

If, in addition to CQ-representability,

(i) *all X_i are closed and bounded,*

or

(ii) *$X_i = Z_i + W$, where Z_i are closed and bounded sets and W is a convex closed set,*

then

$$\mathrm{Conv}\left(\bigcup_{i=1}^k X_i\right) = Y = \overline{\mathrm{Conv}}\left(\bigcup_{i=1}^k X_i\right)$$

is CQr.

First, the set Y clearly contains $\mathrm{Conv}(\bigcup_{i=1}^k X_i)$. Indeed, since the sets X_i are convex, the convex hull of their union is

$$\left\{x = \sum_{i=1}^k t_i x^i \mid x^i \in X_i, t_i \geq 0, \sum_{i=1}^k t_i = 1\right\}$$

(why?); for a point

$$x = \sum_{i=1}^k t_i x^i, \qquad \left[x^i \in X_i, t_i \geq 0, \sum_{i=1}^k t_i = 1\right],$$

there exist u^i, $i = 1, \ldots, k$, such that

$$A_i x^i + B_i u^i + b_i \geq_{\mathbf{K}_i} 0.$$

We get

$$\begin{array}{rcll} x & = & (t_1 x^1) + \cdots + (t_k x^k) & \\ & = & \xi^1 + \cdots + \xi^k, & \\ & & [\xi^i = t_i x^i]; & \\ t_1, \ldots, t_k & \geq & 0; & \\ t_1 + \cdots + t_k & = & 1; & \\ A_i \xi^i + B_i \eta^i + t_i b_i & \geq_{\mathbf{K}_i} & 0,\ i = 1, \ldots, k, & \\ & & [\eta^i = t_i u^i], & \end{array} \tag{3.3.24}$$

so that $x \in Y$ (see the definition of Y).

To complete the proof that Y is between the convex hull and the closed convex hull of $\bigcup_{i=1}^k X_i$, it remains to verify that if $x \in Y$, then x is contained in the closed convex hull of $\bigcup_{i=1}^k X_i$. Let us somehow choose $\bar{x}^i \in X_i$; for properly chosen $\bar{u}^i$ we have

$$A_i \bar{x}^i + B_i \bar{u}^i + b_i \geq_{\mathbf{K}_i} 0,\ i = 1, \ldots, k. \tag{3.3.25}$$

Since $x \in Y$, there exist t_i, ξ^i, η^i satisfying the relations

$$\begin{array}{rcl} x & = & \xi^1 + \cdots + \xi^k, \\ t_1, \ldots, t_k & \geq & 0, \\ t_1 + \cdots + t_k & = & 1, \\ A_i\xi^i + B_i\eta^i + t_i b_i & \geq_{\mathbf{K}_i} & 0, \ i = 1, \ldots, k. \end{array} \tag{3.3.26}$$

In view of the latter relations and (3.3.25), we have for $0 < \epsilon < 1$:

$$A_i[(1-\epsilon)\xi^i + \epsilon k^{-1}\bar{x}^i] + B_i[(1-\epsilon)\eta^i + \epsilon k^{-1}\bar{u}^i] + [(1-\epsilon)t_i + \epsilon k^{-1}]b_i \geq_{\mathbf{K}_i} 0;$$

setting

$$\begin{array}{rcl} t_{i,\epsilon} & = & (1-\epsilon)t_i + \epsilon k^{-1}, \\ x^i_\epsilon & = & t_{i,\epsilon}^{-1}\left[(1-\epsilon)\xi^i + \epsilon k^{-1}\bar{x}^i\right], \\ u^i_\epsilon & = & t_{i,\epsilon}^{-1}\left[(1-\epsilon)\eta^i + \epsilon k^{-1}\bar{u}^i\right], \end{array}$$

we get

$$\begin{array}{rcl} A_i x^i_\epsilon + B_i u^i_\epsilon + b_i & \geq_{\mathbf{K}_i} & 0 \Rightarrow \\ x^i_\epsilon & \in & X_i, \\ t_{1,\epsilon}, \ldots, t_{k,\epsilon} & \geq & 0, \\ t_{1,\epsilon} + \cdots + t_{k,\epsilon} & = & 1 \\ \Rightarrow & & \\ x_\epsilon & \equiv & \displaystyle\sum_{i=1}^k t_{i,\epsilon} x^i_\epsilon \\ & \in & \mathrm{Conv}\left(\displaystyle\bigcup_{i=1}^k X_i\right). \end{array}$$

On the other hand, we have by construction

$$x_\epsilon = \sum_{i=1}^k \left[(1-\epsilon)\xi^i + \epsilon k^{-1}\bar{x}^i\right] \to x = \sum_{i=1}^k \xi^i \text{ as } \epsilon \to +0,$$

so that x belongs to the closed convex hull of $\bigcup_{i=1}^k X_i$, as claimed.

It remains to verify that in the cases (i) and (ii) the convex hull of $\bigcup_{i=1}^k X_i$ is the same as the closed convex hull of this union. Case (i) is a particular case of (ii) corresponding to $W = \{0\}$, so that it suffices to prove (ii). Assume that

$$x_t = \sum_{i=1}^k \mu_{ti}[z_{ti} + p_{ti}] \to x \text{ as } i \to \infty,$$

$$\left[z_{ti} \in Z_i,\ p_{ti} \in W,\ \mu_{ti} \geq 0,\ \sum_i \mu_{ti} = 1\right]$$

and let us prove that x belongs to the convex hull of the union of X_i. Indeed, since Z_i are closed and bounded, passing to a subsequence, we may assume that

$$z_{ti} \to z_i \in Z_i \text{ and } \mu_{ti} \to \mu_i \text{ as } t \to \infty.$$

It follows that the vectors

$$p_t = \sum_{i=1}^{m} \mu_{ti} p_{ti} = x_t - \sum_{i=1}^{k} \mu_{ti} z_{ti}$$

converge as $t \to \infty$ to some vector p, and since W is closed and convex, $p \in W$. We now have

$$x = \lim_{i\to\infty} \left[\sum_{i=1}^{k} \mu_{ti} z_{ti} + p_t \right] = \sum_{i=1}^{k} \mu_i z_i + p = \sum_{i=1}^{k} \mu_i [z_i + p]$$

so that x belongs to the convex hull of the union of X_i (as a convex combination of points $z_i + p \in X_i$).

N. *The recessive cone of a CQr set.* Let X be a closed convex set. The *recessive cone* $\mathrm{Rec}(X)$ of X is the set

$$\mathrm{Rec}(X) = \{h : x + th \in X \quad \forall (x \in X, t \geq 0)\}.$$

It can be easily verified that $\mathrm{Rec}(X)$ is a closed convex cone and that

$$\mathrm{Rec}(X) = \{h \mid \bar{x} + th \in X \quad \forall t \geq 0\} \quad \forall \bar{x} \in X,$$

i.e., that $\mathrm{Rec}(X)$ is the set of all directions h such that the ray emanating from a point of X and directed by h is contained in X.

Proposition 3.3.6. *Let X be a nonempty CQr set with CQR*

$$X = \{x \in \mathbf{R}^n \mid \exists u : Ax + Bu + b \geq_{\mathbf{K}} 0\},$$

where $\mathbf{K}$ is a direct product of ice cream cones, and let the CQR be such that $Bu \in \mathbf{K}$ only if $Bu = 0$. Then X is closed, and the recessive cone of X is CQr:

$$\mathrm{Rec}(X) = \{h \mid \exists v : Ah + Bv \geq_{\mathbf{K}} 0\}. \tag{3.3.27}$$

Proof. The fact that X is closed is given by Lemma 3.3.1. In order to prove (3.3.27), let us temporarily denote by R the set in the left-hand side of this relation; we should prove that $R = \mathrm{Rec}(X)$. The inclusion $R \subset \mathrm{Rec}(X)$ is evident. To prove the inverse inclusion, let $\bar{x} \in X$ and $h \in \mathrm{Rec}(X)$ so that for every $i = 1, 2, \ldots$ there exists u_i such that

$$A(\bar{x} + ih) + Bu_i + b \in \mathbf{K}. \tag{3.3.28}$$

By Lemma 3.3.1,

$$\|Bu_i\|_2 \leq C(1 + \|A(\bar{x} + ih) + b\|_2) \tag{3.3.29}$$

for certain $C < \infty$ and all i. Besides this, we can assume without loss of generality (w.l.o.g.) that

$$\|u_i\|_2 \leq C_1 \|Bu_i\|_2 \tag{3.3.30}$$

(cf. the proof of Lemma 3.3.1). By (3.3.29)–(3.3.30), the sequence $\{v_i = i^{-1}u_i\}$ is bounded; passing to a subsequence, we can assume that $v_i \to v$ as $i \to \infty$. By (3.3.28), we have for all i

$$i^{-1}A(\bar{x} + ih) + Bv_i + i^{-1}b \in \mathbf{K},$$

whence, passing to limit as $i \to \infty$, $Ah + Bv \in \mathbf{K}$. Thus, $h \in R$. □

O. *Theorem on superposition.* Let $f_\ell : \mathbf{R}^m \to \mathbf{R} \cup \{+\infty\}$, $\ell = 1, \ldots, m$ be CQr functions,

$$t \geq f_\ell(x) \Leftrightarrow \exists u^\ell \mid A_\ell(x, t, u^\ell) \geq_{\mathbf{K}_\ell} 0,$$

where $\mathbf{K}_\ell$ is a direct product of ice cream cones, and let

$$f : \mathbf{R}^m \to \mathbf{R} \cup \{+\infty\}$$

be CQr,

$$t \geq f(y) \Leftrightarrow \exists v : A(y, t, v) \geq_{\mathbf{K}} 0,$$

where $\mathbf{K}$ is a direct product of ice cream cones.

Assume that f is monotone with respect to the usual partial ordering of $\mathbf{R}^m$:

$$y' \geq y'' \Rightarrow f(y') \geq f(y''),$$

and consider the superposition

$$g(x) = \begin{cases} f(f_1(x), \ldots, f_m(x)), & f_\ell(x) < \infty, \ell = 1, \ldots, m, \\ +\infty & \text{otherwise.} \end{cases}$$

Theorem 3.3.1. *Under the above setting, the superposition $g(\cdot)$ is CQr with CQR*

$$t \geq g(x) \Leftrightarrow \exists t_1, \ldots, t_m, u^1, \ldots, u^m, v : \begin{cases} A_\ell(x, t_\ell, u^\ell) \geq_{\mathbf{K}_\ell} 0, \ \ell = 1, \ldots, m, \\ A(t_1, \ldots, t_m, t, v) \geq_{\mathbf{K}} 0. \end{cases} \tag{3.3.31}$$

Proof. The proof of this simple statement is left to the reader.

Remark 3.3.1. If part of the inner functions, say, $f_1, \ldots, f_k$, is affine, it suffices to require the monotonicity of the outer function f with respect to the variables $y_{k+1}, \ldots, y_m$ only. A CQR for the superposition in this case becomes

$$t \geq g(x) \Leftrightarrow \exists t_{k+1}, \ldots, t_m, u^{k+1}, \ldots, u^m, v : \begin{cases} A_\ell(x, t_\ell, u^\ell) \geq_{\mathbf{K}_\ell} 0, \ \ell = k+1, \ldots, m, \\ A(f_1(x), f_2(x), \ldots, f_k(x), t_{k+1}, t_{k+2}, \ldots, t_m, t, v) \geq_{\mathbf{K}} 0 \end{cases} \tag{3.3.32}$$

3.3.1 More examples of conic quadratic-representable functions and sets

We are sufficiently equipped to build the dictionary of CQr functions and sets. Having built already the elementary part of the dictionary, we can add now a more advanced part.

7. Convex quadratic form. $g(x) = x^T Qx + q^T x + r$ (Q is a positive semidefinite symmetric matrix) is CQr.

Indeed, Q is positive semidefinite symmetric and therefore can be decomposed as $Q = D^T D$, so that $g(x) = \|Dx\|_2^2 + q^T x + r$. We see that g is obtained from our raw materials—the squared Euclidean norm and an affine function—by affine substitution of argument and addition.

Here is an explicit CQR of g:

$$\{(x,t) \mid x^T D^T Dx + q^T x + r \le t\} = \left\{(x,t) \mid \left\| \begin{array}{c} Dx \\ \frac{t+q^T x+r}{2} \end{array} \right\|_2 \le \frac{t - q^T x - r}{2}\right\}. \quad (3.3.33)$$

8. The cone $K = \{(x, \sigma_1, \sigma_2) \in \mathbf{R}^n \times \mathbf{R} \times \mathbf{R} \mid \sigma_1, \sigma_2 \ge 0, \sigma_1\sigma_2 \ge x^T x\}$ is CQr.

Indeed, the set is just the epigraph of the fractional-quadratic function $x^T x/s$ (see Example 5); we simply write σ_1 instead of s and σ_2 instead of t.

Here is an explicit CQR for the set:

$$K = \left\{(x, \sigma_1, \sigma_2) \mid \left\| \begin{pmatrix} x \\ \frac{\sigma_1 - \sigma_2}{2} \end{pmatrix} \right\|_2 \le \frac{\sigma_1 + \sigma_2}{2}\right\}. \quad (3.3.34)$$

Surprisingly, our set is just the ice cream cone, more precisely, its inverse image under the one-to-one linear mapping

$$\begin{pmatrix} x \\ \sigma_1 \\ \sigma_2 \end{pmatrix} \mapsto \begin{pmatrix} x \\ \frac{\sigma_1 - \sigma_2}{2} \\ \frac{\sigma_1 + \sigma_2}{2} \end{pmatrix}.$$

9. The half-cone $K_+^2 = \{(x_1, x_2, t) \in \mathbf{R}^3 \mid x_1, x_2 \ge 0, 0 \le t \le \sqrt{x_1 x_2}\}$ is CQr.

Indeed, our set is the intersection of the cone $\{t^2 \le x_1 x_2, x_1, x_2 \ge 0\}$ from the previous example and the half-space $t \ge 0$.

Here is the explicit CQR of K_+:

$$K_+ = \left\{(x_1, x_2, t) \mid t \ge 0, \left\| \begin{pmatrix} t \\ \frac{x_1 - x_2}{2} \end{pmatrix} \right\|_2 \le \frac{x_1 + x_2}{2}\right\}. \quad (3.3.35)$$

10. The hypograph of the geometric mean—the set $K^2 = \{(x_1, x_2, t) \in \mathbf{R}^3 \mid x_1, x_2 \ge 0, t \le \sqrt{x_1 x_2}\}$—is CQr.

Note the difference with the previous example—here t is not required to be nonnegative!

Here is the explicit CQR for K^2 (see Example 9):

$$K^2 = \left\{(x_1, x_2, t) \mid \exists \tau : t \le \tau, \left\| \begin{pmatrix} \tau \\ \frac{x_1 - x_2}{2} \end{pmatrix} \right\|_2 \le \frac{x_1 + x_2}{2}\right\}.$$

11. The hypograph of the geometric mean of 2^l variables—the set $K^{2^l} = \{(x_1, \ldots, x_{2^l}, t) \in \mathbf{R}^{2^l+1} \mid x_i \geq 0, i = 1, \ldots, 2^l, t \leq (x_1 x_2 \ldots x_{2^l})^{1/2^l}\}$—is CQr. To see it and to get its CQR, it suffices to iterate the construction of Example 10. Indeed, let us add to our initial variables a number of additional x-variables:

1. Let us call our 2^l original x-variables the variables of level 0 and write $x_{0,i}$ instead of x_i. Let us add one new variable of level 1 per every two variables of level 0. Thus, we add 2^{l-1} variables $x_{1,i}$ of level 1.

2. Similarly, let us add one new variable of level 2 per every two variables of level 1, thus adding 2^{l-2} variables $x_{2,i}$; then we add one new variable of level 3 per every two variables of level 2, and so on, until level l with a single variable $x_{l,1}$ is built.

Now let us look at the following system S of constraints:

$$\begin{array}{ll} \text{layer 1:} & x_{1,i} \leq \sqrt{x_{0,2i-1}x_{0,2i}},\, x_{1,i}, x_{0,2i-1}, x_{0,2i} \geq 0, \quad i = 1, \ldots, 2^{l-1}, \\ \text{layer 2:} & x_{2,i} \leq \sqrt{x_{1,2i-1}x_{1,2i}},\, x_{2,i}, x_{1,2i-1}, x_{1,2i} \geq 0, \quad i = 1, \ldots, 2^{l-2}, \\ & \ldots \\ \text{layer } l\text{:} & x_{l,1} \leq \sqrt{x_{l-1,1}x_{l-1,2}},\, x_{l,1}, x_{l-1,1}, x_{l-1,2} \geq 0 \\ \hline (*) & t \leq x_{l,1} \end{array}.$$

The inequalities of the first layer say that the variables of the zero and the first level should be nonnegative and every one of the variables of the first level should be less than or equal to the geometric mean of the corresponding pair of our original x-variables. The inequalities of the second layer add the requirement that the variables of the second level should be nonnegative, and every one of them should be less than or equal to the geometric mean of the corresponding pair of the first-level variables, etc. It is clear that if all these inequalities and (*) are satisfied, then t is less than or equal to the geometric mean of $x_1, \ldots, x_{2^l}$. Conversely, given nonnegative $x_1, \ldots, x_{2^l}$ and a real t which is less than or equal to the geometric mean of $x_1, \ldots, x_{2^l}$, we always can extend these data to a solution of S. In other words, K^{2^l} is the projection of the solution set of S onto the plane of our original variables $x_1, \ldots, x_{2^l}, t$. It remains to note that the set of solutions of S is CQr (as the intersection of CQr sets $\{(v, p, q, r) \in \mathbf{R}^N \times \mathbf{R}^3_+ \mid r \leq \sqrt{qp}\}$; see example 9) so that its projection is also CQr. To get a CQR of K^{2^l}, it suffices to replace the inequalities in S with their conic quadratic equivalents, explicitly given in example 9.

What about functions that look very different from a quadratic function, e.g., what about the function $g(x) = x^{7/3}$ on the real line? Is it CQr? If the question is to be interpreted literally, the answer is a definite "no"—the function is nonconvex! An absolutely evident observation is as follows:

> *A CQr set X is always convex (as the projection of the set of solutions of a system of convex inequalities $\|\alpha_j(x, u)\|_2 - \beta_j(x, y) \leq 0$ in the space of (x, u)-variables onto the space of x-variables).*
>
> *Consequently, a CQr function is necessarily convex—since its epigraph must be a CQr and therefore a convex set.*

Our question about the function $x^{7/3}$ admits, however, a meaningful modification. Namely, the function $x^{7/3}$ (as every power function x^p with $p \geq 1$) is convex on the nonnegative ray; extending the function by the value 0 onto the negative ray, we get a convex function

x_+^p ($x_+ \equiv \max\{x, 0\}$), and we may ask whether this function is CQr. In particular, is the function $x_+^{7/3}$ CQr? The answer is affirmative, and here is the construction:

We know from example 11 that the set

$$K^8 = \{(y_1, \ldots, y_8, s) \in \mathbf{R}^9_+ \mid s \le (y_1 y_2 \ldots y_8)^{1/8}\}$$

is CQr. Now let us make all our nine variables $y_1, \ldots, y_8, s$ affine functions of just two variables ξ, t as follows:

- The variable s and the first of the variables y_i are identically equal to ξ.
- The next three of the variables y_i are identically equal to t.
- The rest of the variables (i.e., the last four variables y_i) are identically equal to 1.

We have defined certain affine mapping from $\mathbf{R}^2$ to $\mathbf{R}^9$. The inverse image of K^8 under this mapping is the set

$$\begin{array}{rcl} K & = & \{(\xi, t) \in \mathbf{R}^2_+ \mid \xi \le \xi^{1/8} t^{3/8}\} \\ & = & \{(\xi, t) \in \mathbf{R}^2_+ \mid t \ge \xi^{7/3}\}. \end{array}$$

Thus, the set K is CQr (as an inverse image of the CQr set K^8), and we can easily get an explicit CQR of K from the one of K^8 (see Example 11). On the other hand, the set K is almost the epigraph of $\xi_+^{7/3}$—it is the part of this epigraph in the first quadrant. And it is easy to get the complete epigraph from this part: it suffices to note that the epigraph E of $x_+^{7/3}$ is the projection of the 3D set

$$K' = \{(x, \xi, t) \mid \xi \ge 0, x \le \xi, \xi^{7/3} \le t\}$$

onto the (x, t)-plane and that the set K' clearly is CQr along with K. Indeed, to obtain K' from K one should first pass from K to its direct product with the x-axis—to the set

$$K^+ = \{(x, \xi, t) \mid (\xi, t) \in K\}$$

—and then to intersect K^+ with the half-space given by the inequality $x \le \xi$. Thus, to obtain the epigraph E of $x_+^{3/7}$ from the CQr set K, one should successively

- pass from K to its direct product with the real axis $\mathbf{R}$ (note that the second factor trivially is CQr!),
- intersect the result with a half-space, and
- project the result onto 2D plane of the variables (x, t).

All these operations preserve the CQ-representability and yield an explicit CQr for E:

$$\{t \ge x_+^{7/3}\} \Leftrightarrow \exists(\xi, u) : \{x \le \xi\} \ \& \ S(\xi, t, t, t, 1, 1, 1, 1, \xi; u),$$

where $S(y_1, \ldots, y_8, s; u)$ denotes the system of CQIs from the CQR of the set K[8],[13] where u is the vector of additional variables for the latter CQR.

The construction we have just described might look too sophisticated, but with a little experience the derivations of this type become much easier and much more transparent than, say, arithmetic solving a 3×3 system of linear equations.

Of course, the particular values 7 and 3 in our $x_+^{7/3}$-exercise play no significant role, and we arrive at the following.

12. The convex increasing power function $x_+^{p/q}$ of rational degree $p/q \geq 1$ is CQr.

Indeed, given positive integers p, q, $p > q$, let us choose the smallest integer l such that $p \leq 2^l$ and consider the CQr set

$$K^{2^l} = \{(y_1, \ldots, y_{2^l}, s) \in \mathbf{R}_+^{2^l+1} \mid s \leq (y_1 y_2 \ldots y_{2^l})^{1/2^l}\}. \tag{3.3.36}$$

Setting $r = 2^l - p$, consider the following affine parameterization of the variables from $\mathbf{R}^{2^l+1}$ by two variables ξ, t:

- s and r first variables y_i are all equal to ξ (note that we still have $2^l - r = p \geq q$ unused variables y_i).
- q next variables y_i are all equal to t.
- The remaining y_i's, if any, are all equal to 1.

The inverse image of K^{2^l} under this mapping is CQr and it is the set

$$K = \{(\xi, t) \in \mathbf{R}_+^2 \mid \xi^{1-r/2^l} \leq t^{q/2^l}\} = \{(\xi, t) \in \mathbf{R}_+^2 \mid t \geq \xi^{p/q}\}.$$

The epigraph of $x_+^{p/q}$ can be obtained from the CQr set K, as in the case of $p/q = 3/7$, by operations preserving CQ-representability.

13. The decreasing power function

$$g(x) = \begin{cases} x^{-p/q}, & x > 0, \\ +\infty, & x \leq 0, \end{cases}$$

(p, q are positive integers) is CQr.

As in Example 12, we choose the smallest integer l such that $2^l \geq p + q$, consider the CQr set (3.3.36), and parameterize affinely the variables y_i, s by two variables (x, t) as follows:

- s and the first $(2^l - p - q)$ y_i's are all equal to one.
- p of the remaining y_i's are all equal to x, and the q last of y_i's are all equal to t.

[13] That is, $S(y, t; u)$ is a Boolean function taking values `true` of `false` depending on whether the (y, t, u) satisfy or do not satisfy the CQIs in question

It is immediately seen that the inverse image of K^{2^l} under the indicated affine mapping is the epigraph of g.

14. The even power function $g(x) = x^{2p}$ on the axis (p positive integer) is CQr.

Indeed, we already know that the sets $P = \{(x, \xi, t) \in \mathbf{R}^3 \mid x^2 \le \xi\}$ and $K' = \{(x, \xi, t) \in \mathbf{R}^3 \mid 0 \le \xi, \xi^p \le t\}$ are CQr. (Both sets are direct products of $\mathbf{R}$ and the sets with already known to us CQRs.) It remains to note that the epigraph of g is the projection of $P \cap Q$ onto the (x, t)-plane.

Example 14 along with our combination rules allow us to build a CQR for a polynomial $p(x)$ of the form

$$p(x) = \sum_{l=1}^{L} p_l x^{2l}, \quad x \in \mathbf{R},$$

with nonnegative coefficients.

15. The hypograph of a concave monomial $x_1^{\pi_1} \dots x_n^{\pi_n}$. Let $\pi_1 = \frac{p_1}{p}, \dots, \pi_n = \frac{p_n}{p}$ be positive rational numbers with $\pi_1 + \dots + \pi_n \le 1$. The function

$$f(x) = -x_1^{\pi_1} \dots x_n^{\pi_n} : \mathbf{R}^n_+ \to \mathbf{R}$$

is CQr.

The construction is similar to the one of Example 12. Let l be such that $2^l \ge p$. We recall that the set

$$Y = \{(y_1, \dots, y_{2^l}, s) \mid y_1, \dots, y_{2^l} \ge 0, 0 \le s \le (y_1 \dots, y_{2^l})^{1/2^l}\}$$

is CQr, and therefore so is its inverse image under the affine mapping

$$(x_1, \dots, x_n, s) \mapsto (\underbrace{x_1, \dots, x_1}_{p_1}, \underbrace{x_2, \dots, x_2}_{p_2}, \dots, \underbrace{x_n, \dots, x_n}_{p_n}, \underbrace{s, \dots, s}_{2^l - p}, \underbrace{1, \dots, 1}_{p - p_1 - \dots - p_n}, s),$$

i.e., the set

$$\begin{aligned} Z &= \{(x_1, \dots, x_n, s) \mid x_1, \dots, x_n \ge 0, 0 \le s \le (x_1^{p_1} \dots x_n^{p_n} s^{2^l - p})^{1/2^l}\} \\ &= \{(x_1, \dots, x_n, s) \mid x_1, \dots, x_n \ge 0, 0 \le s \le x_1^{p_1/p} \dots x_n^{p_n/p}\}. \end{aligned}$$

Since the set Z is CQr, so is the set

$$Z' = \{(x_1, \dots, x_n, t, s) \mid x_1, \dots, x_n \ge 0, s \ge 0, 0 \le s - t \le x_1^{\pi_1} \dots x_n^{\pi_n}\},$$

which is the intersection of the half-space $\{s \ge 0\}$ and the inverse image of Z under the affine mapping $(x_1, \dots, x_n, t, s) \mapsto (x_1, \dots, x_n, s - t)$. It remains to note that the epigraph of f is the projection of Z' onto the plane of the variables $x_1, \dots, x_n, t$.

16. The convex monomial $x_1^{-\pi_1} \dots x_n^{-\pi_n}$. Let $\pi_1, \dots, \pi_n$ be positive rational numbers. The function

$$f(x) = x_1^{-\pi_1} \dots x_n^{-\pi_n} : \{x \in \mathbf{R}^n : x > 0\} \to \mathbf{R}$$

is CQr.

See Exercise 3.2.

17. The p-norm $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p} : \mathbf{R}^n \to \mathbf{R}$ ($p \geq 1$ is a rational number). We claim that the function $\|x\|_p$ is CQr.

It is immediately seen that

$$\|x\|_p \leq t \Leftrightarrow t \geq 0 \;\&\; \exists v_1, \ldots, v_n \geq 0 : |x_i| \leq t^{(p-1)/p} v_i^{1/p},\ i = 1, \ldots, n, \sum_{i=1}^n v_i \leq t. \tag{3.3.37}$$

Indeed, if the indicated v_i exist, then $\sum_{i=1}^n |x_i|^p \leq t^{p-1} \sum_{i=1}^n v_i \leq t^p$, i.e., $\|x\|_p \leq t$. Conversely, assume that $\|x\|_p \leq t$. If $t = 0$, then $x = 0$, and the right-hand side relations in (3.3.37) are satisfied for $v_i = 0$, $i = 1, \ldots, n$. If $t > 0$, we can satisfy these relations by setting $v_i = |x_i|^p t^{1-p}$.

Equation (3.3.37) says that the epigraph of $\|x\|_p$ is the projection onto the (x, t)-plane of the set of solutions to the system of inequalities

$$\begin{gathered} t \geq 0, \\ v_i \geq 0,\ i = 1, \ldots, n, \\ x_i \leq t^{(p-1)/p} v_i^{1/p},\ i = 1, \ldots, n, \\ -x_i \leq t^{(p-1)/p} v_i^{1/p},\ i = 1, \ldots, n, \\ v_1 + \cdots + v_n \leq t. \end{gathered}$$

Each of these inequalities defines a CQr set (in particular, for the nonlinear inequalities this is due to Example 15). Thus, the solution set of the system is CQr (as an intersection of finitely many CQr sets), whence its projection on the (x, t)-plane—i.e., the epigraph of $\|x\|_p$—is CQr.

17.a. The function $\|x_+\|_p = (\sum_{i=1}^n \max^p[x_i, 0])^{1/p} : \mathbf{R}^n \to \mathbf{R}$ ($p \geq 1$ a rational number) is CQr.

Indeed,

$$t \geq \|x_+\|_p \Leftrightarrow \exists y_1, \ldots, y_n : 0 \leq y_i, x_i \leq y_i,\ i = 1, \ldots, n, \|y\|_p \leq t.$$

Thus, the epigraph of $\|x_+\|_p$ is a projection of the CQr set (see Example 17) given by the system of inequalities in the right-hand side.

From the above examples it is seen that the expressive abilities of CQIs are indeed strong: they allow us to handle a wide variety of very different functions and sets.

3.4 More applications

Equipped with abilities to treat a wide variety of CQr functions and sets, we can consider now more applications of CQP.

3.4.1 Tschebyshev approximation in relative scale

In the Tschebyshev approximation problem we are looking for a linear combination of given basis functions $f_i(t)$ that is as close as possible to a certain target function $f_*(t)$ on a

given finite set T.[14] In the original version of the problem, the quality of an approximation $\sum_i x_i f_i(t)$ is measured by the maximal, over $t \in T$, *absolute* deviation of the approximation from the target function. In a number of applications where the target function is positive and so should be its approximation, a more appropriate deviation is the *relative* one. A natural way to measure the relative deviation between two positive reals a, b is to look at the smallest $\tau \equiv \tau(a, b)$ such that

$$\frac{1}{1+\tau} \le \frac{a}{b} \le 1+\tau.$$

With this approach, the relative Tschebyshev problem becomes

$$\min_x \max_{t\in T} \tau\left(f_*(t), \sum_i x_i f_i(t)\right),$$

where we should add the constraints $\sum_i x_i f_i(t) > 0$, $t \in T$, to guarantee positivity of the approximation. The resulting problem can be written equivalently as

$$\min_{x,\tau}\left\{\tau \mid \sum_i x_i f_i(t) \le (1+\tau) f_*(t), \quad f_*(t) \le (1+\tau)\sum_i x_i f_i(t)\ \forall t \in T\right\}.$$

The nonlinear constraints we get are hyperbolic constraints "the product of two nonnegative affine forms of the design variables must be greater than or equal to a positive constant," and the sets given by these constraints are CQr (see Example 6). Thus, the problem is equivalent to a conic quadratic program, specifically, to the problem

$$\min_{x,\tau}\left\{\tau \mid \forall (t \in T): \left\{\begin{array}{c} \sum_i x_i f_i(t) \le (1+\tau) f_*(t), \\ \left\|\begin{pmatrix} 2\sqrt{f_*(\tau)} \\ 1+\tau-\sum_i x_i f_i(t)\end{pmatrix}\right\|_2 \le 1+\tau+\sum_i x_i f_i(t)\end{array}\right\}\right.$$

3.4.2 Robust linear programming

Consider an LP program

$$\min_x \left\{c^T x \mid Ax - b \ge 0\right\}. \tag{LP}$$

In real-world applications, the data c, A, b of (LP) is not always known exactly; what is typically known is a domain $\mathcal{U}$ in the space of data—an uncertainty set—which for sure contains the actual (unknown) data. There are cases in reality where, in spite of this data uncertainty, our decision x must satisfy the actual constraints, whether we know them or not. Assume, e.g., that (LP) is a model of a technological process in the chemical industry, so that entries of x represent the amounts of different kinds of materials participating in the process.

[14] The material in this section originates from M.S. Lobo, L. Vanderbeghe, S. Boyd, and H. Lebret, *Second-order cone programming*, Linear Algebra Appl., 284 (1998), pp. 193–228.

Typically the process includes a number of decomposition-recombination stages. A model of this problem must take care of natural balance restrictions: the amount of every material to be used at a particular stage cannot exceed the amount of the same material yielded by the preceding stages. In a meaningful production plan, these balance inequalities must be satisfied although they involve coefficients affected by unavoidable uncertainty of the exact contents of the raw materials, of time-varying parameters of the technological devices, etc.

If indeed all we know about the data is that they belong to a given set $\mathcal{U}$, but we still have to satisfy the actual constraints, the only way to meet the requirements is to restrict ourselves to *robust feasible* candidate solutions—those satisfying all possible realizations of the uncertain constraints, i.e., vectors x such that

$$Ax - b \geq 0 \quad \forall [A; b] \text{ such that } \exists c : (c, A, b) \in \mathcal{U}. \tag{3.4.38}$$

In order to make the best possible choice from these robust feasible solutions, we should decide how to aggregate the various realizations of the objective into a single quality characteristic. To be methodologically consistent, we use the same worst-case approach and take as an objective function $f(x)$ the maximum, over all possible realizations of c, of the quantity $c^T x$:

$$f(x) = \sup\{c^T x \mid c : \exists [A; b] : (c, A, b) \in \mathcal{U}\}.$$

With this methodology, we can associate with our uncertain LP program, i.e., with the family

$$\mathcal{LP}(\mathcal{U}) = \left\{ \min_{x: Ax \geq b} c^T x \,\middle|\, (c, A, b) \in \mathcal{U} \right\}$$

of all usual (certain) LP programs with the data belonging to $\mathcal{U}$, its robust counterpart. In the latter problem we are seeking for a robust feasible solution with the smallest possible value of the guaranteed objective $f(x)$. In other words, the robust counterpart of $\mathcal{LP}(\mathcal{U})$ is the optimization problem

$$\min_{t,x} \left\{ t \mid c^T x \leq t, Ax - b \geq 0 \quad \forall (c, A, b) \in \mathcal{U} \right\}. \tag{R}$$

Note that (R) is a usual—certain—optimization problem, but typically it is not an LP program: the structure of (R) depends on the geometry of the uncertainty set $\mathcal{U}$ and can be very complicated.

In many cases it is reasonable to specify the uncertainty set $\mathcal{U}$ as an *ellipsoid*—the image of the unit Euclidean ball under an affine mapping—or, more generally, as a CQr set. As we shall see in a while, in this case the robust counterpart of an uncertain LP problem is (equivalent to) an explicit conic quadratic program. Thus, robust linear programming with CQr uncertainty sets can be viewed as a generic source of conic quadratic problems.

Let us look at the robust counterpart of an uncertain LP program

$$\left\{ \min_x \left\{ c^T x : a_i^T x - b_i \geq 0, \ i = 1, \ldots, m \right\} \,\middle|\, (c, A, b) \in \mathbf{U} \right\}$$

in the case of a simple ellipsoidal uncertainty—one where the data (a_i, b_i) of ith inequality constraint

$$a_i^T x - b_i \geq 0$$

and the objective c are allowed to run independently of each other through respective ellipsoids E_i, E. Thus, we assume that the uncertainty set is

$$\mathcal{U} = \left\{ (a_1, b_1; \ldots; a_m, b_m; c) \mid \exists (\{u_i, u_i^T u_i \le 1\}_{i=0}^m) : \quad \begin{array}{c} c = c_* + P_0 u_0, \\ \begin{pmatrix} a_i \\ b_i \end{pmatrix} = \begin{pmatrix} a_i^* \\ b_i^* \end{pmatrix} + P_i u^i, \\ i = 1, \ldots, m \end{array} \right\},$$

where c_*, a_i^*, b_i^* are the nominal data and $P_i u_i$, $i = 0, 1, \ldots, m$, represent the data perturbations; the restrictions $u_i^T u_i \le 1$ enforce these perturbations to vary in ellipsoids.

To understand that the robust counterpart of our uncertain LP problem is a conic quadratic program, note that x is robust feasible if and only if for every $i = 1, \ldots, m$ we have

$$\begin{aligned} 0 &\le \min_{u_i : u_i^T u_i \le 1} \left[a_i^T[u_i]x - b_i[u_i] \mid \begin{pmatrix} a_i[u] \\ b_i[u] \end{pmatrix} = \begin{pmatrix} a_i^* \\ b_i^* \end{pmatrix} + P_i u_i \right] \\ &= (a_i^* x)^T x - b_i^* + \min_{u_i : u_i^T u_i \le 1} u_i^T P_i^T \begin{pmatrix} x \\ -1 \end{pmatrix} \\ &= (a_i^*)^T x - b_i^* - \left\| P_i^T \begin{pmatrix} x \\ -1 \end{pmatrix} \right\|_2 . \end{aligned}$$

Thus, x is robust feasible if and only if it satisfies the system of CQIs

$$\left\| P_i^T \begin{pmatrix} x \\ -1 \end{pmatrix} \right\|_2 \le [a_i^*]^T x - b_i^*, \ i = 1, \ldots, m.$$

Similarly, a pair (x, t) satisfies all realizations of the inequality $c^T x \le t$ allowed by our ellipsoidal uncertainty set $\mathcal{U}$ if and only if

$$c_*^T x + \| P_0^T x \|_2 \le t.$$

Thus, the robust counterpart (R) becomes the conic quadratic program

$$\min_{x,t} \left\{ t \mid \| P_0^T x \|_2 \le -c_*^T x + t; \ \left\| P_i^T \begin{pmatrix} x \\ -1 \end{pmatrix} \right\|_2 \le [a_i^*]^T x - b_i^*, \ i = 1, \ldots, m \right\}. \quad \text{(RLP)}$$

Example. Robust synthesis of antenna array. Consider the antenna synthesis example of section 1.2.4. Mathematically, this was an LP program with 11 variables,

$$\min_{x,t} \left\{ t \mid -t \le Z_*(\theta_i) - \sum_{j=1}^{10} x_j Z_j(\theta_i) \le t, \ i = 1, \ldots, N \right\}, \quad \text{(Nom)}$$

with given diagrams $Z_j(\cdot)$ of 10 building blocks and a given target diagram $Z_*(\theta)$. Let x_j^* be the optimal values of the design variables. Recall that our design variables are amplification coefficients, i.e., characteristics of certain physical devices. In reality, of course, we cannot tune the devices to have precisely the optimal characteristics x_j^*; the best we may hope for

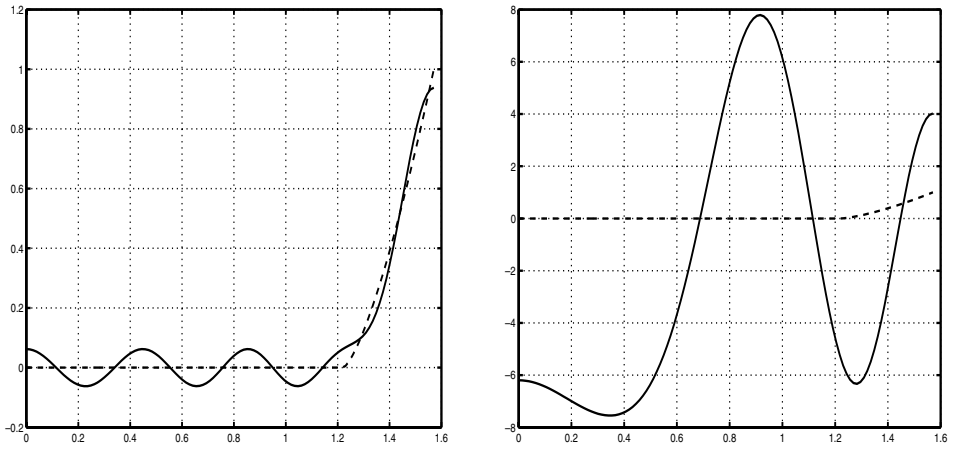


Figure 3.3. *Dream and reality: the nominal (left, solid line) and an actual (right, solid line) diagram (dashed line: the target diagram).*

is that the actual characteristics $x_j^{\rm fct}$ of the amplifiers will coincide with the desired values x_j^* within a small margin, say, 0.1% (this is a fairly high accuracy for a physical device):

$$x_j^{\rm fct} = p_j x_j^*,\ 0.999 \le p_j \le 1.001.$$

It is natural to assume that the factors p_j are random with the mean value equal to 1; it is perhaps not a great sin to assume that these factors are independent of each other.

Since the actual amplification coefficients differ from their desired values x_j^*, the actual (random) diagram of our array of antennae will differ from the nominal one we found in section 1.2.4. How large could the difference be? Look at Fig. 3.3. The right-hand diagram in Fig. 3.3 is not even the worst case: we just have taken as p_j a sample of 10 independent numbers distributed uniformly in [0.999, 1.001] and have plotted the diagram corresponding to $x_j = p_j x_j^*$. Pay attention not only to the shape (completely opposite to what we need) but also to the scale: the target diagram varies from 0 to 1, and the nominal diagram (the one corresponding to the exact optimal x_j) differs from the target by no more than by 0.0621 (this is the optimal value in the nominal problem (Nom)). The actual diagram varies from ≈ -8 to ≈ 8, and its uniform distance from the target is 7.79 (125 times the nominal optimal value!). We see that our nominal optimal design is completely meaningless. It looks as if we were trying to get the worse possible result, not the best possible one.

How could we get something better? Let us try to apply the robust counterpart approach. To this end we take into account from the very beginning that if we want the amplification coefficients to be certain x_j, then the actual amplification coefficients will be $x_j^{\rm fct} = p_j x_j$, $0.999 \le p_j \le 1.001$, and the actual discrepancies will be

$$\delta_i(x) = Z_*(\theta_i) - \sum_{j=1}^{10} p_j x_j Z_j(\theta_i).$$

Thus, we are in fact solving an uncertain LP problem where the uncertainty affects the coefficients of the constraint matrix (those corresponding to the variables x_j). These coefficients may vary within a 0.1% margin of their nominal values.

To apply the robust counterpart approach to our uncertain LP program, we should specify the uncertainty set $\mathcal{U}$. The most straightforward way is to say that our uncertainty is an interval one—every uncertain coefficient in a given inequality constraint may (independently of all other coefficients) run through its own uncertainty segment "nominal value $\pm 0.1\%$". This approach, however, is too conservative. We have completely ignored the fact that our p_j's are of stochastic nature and are independent of each other, so that it is highly improbable that all of them will simultaneously fluctuate in dangerous directions. To utilize the statistical independence of perturbations, let us look what happens with a particular inequality

$$-t \le \delta_i(x) \equiv Z_*(\theta_i) - \sum_{j=1}^{10} p_j x_j Z_j(\theta_i) \le t \tag{3.4.39}$$

when p_j's are random. For a fixed x, the quantity $\delta_i(x)$ is a random variable with mean

$$\delta_i^*(x) = Z_*(\theta_i) - \sum_{j=1}^{10} x_j Z_j(\theta_i)$$

and standard deviation

$$\sigma_i(x) = \sqrt{\mathbf{E}\{(\delta_i(x) - \delta_i^*(x))^2\}} = \sqrt{\sum_{j=1}^{10} x_j^2 Z_j^2(\theta_i)\mathbf{E}\{(p_j - 1)^2\}} \le \kappa \nu_i(x),$$

$$\nu_i(x) = \sqrt{\sum_{j=1}^{10} x_j^2 Z_j^2(\theta_i)}, \quad \kappa = 0.001.$$

Thus, a typical value of $\delta_i(x)$ differs from $\delta_i^*(x)$ by a quantity of order of $\sigma_i(x)$. Now let us act as an engineer who believes that a random variable differs from its mean by at most three times its standard deviation. Since we are not obliged to be that concrete, let us choose a safety parameter ω and ignore all events that result in $|\delta_i(x) - \delta_i^*(x)| > \omega\nu_i(x)$.[15] As for the remaining events—those with $|\delta_i(x) - \delta_i^*(x)| \le \omega\nu_i(x)$—we take upon ourselves full responsibility. With this approach, a reliable deterministic version of the uncertain constraint (3.4.39) becomes the pair of inequalities

$$\begin{aligned} -t &\le \delta_i^*(x) - \omega\nu_i(x), \\ \delta_i^*(x) + \omega\nu_i(x) &\le t. \end{aligned}$$

[15] It would be better to use σ_i here instead of ν_i. However, we did not assume that we know the distribution of p_j, and this is why we replace unknown σ_i with its known upper bound ν_i.

Replacing all uncertain inequalities in (Nom) with their reliable deterministic versions and recalling the definition of $\delta_i^*(x)$ and $\nu_i(x)$, we end up with the optimization problem

$$\begin{array}{ll} \text{minimize} & t \\ \text{s.t.} & \\ \|Q_i x\|_2 & \le \left[Z_*(\theta_i) - \sum_{j=1}^{10} x_j Z_j(\theta_i) \right] + t, \ i = 1, \ldots, N, \\ \|Q_i x\|_2 & \le -\left[Z_*(\theta_i) - \sum_{j=1}^{10} x_j Z_j(\theta_i) \right] + t, \ i = 1, \ldots, N, \\ & Q_i = \omega\kappa \mathrm{Diag}(Z_1(\theta_i), Z_2(\theta_i), \ldots, Z_{10}(\theta_i)). \end{array} \tag{Rob}$$

It is immediately seen that (Rob) is nothing but the robust counterpart of (Nom) corresponding to a simple ellipsoidal uncertainty, namely, the one as follows:

The only data of a constraint

$$\sum_{j=1}^{10} A_{ij} x_j \ \begin{matrix} \ge \\ \le \end{matrix} \ p_i t + q_i$$

(all constraints in (Nom) are of this form) affected by the uncertainty are the coefficients A_{ij} of the left-hand side. The difference $dA[i]$ between the vector of these coefficients and the nominal value $(Z_1(\theta_i), \ldots, Z_{10}(\theta_i))$ of the vector of coefficients belongs to the ellipsoid

$$\{dA[i] = \omega\kappa Q_i u \mid u \in \mathbf{R}^{10}, u^T u \le 1\}.$$

Thus, the above engineering reasoning leading to (Rob) is nothing but a reasonable way to specify uncertainty ellipsoids!

Now let us look at what diagrams are yielded by the robust counterpart approach, i.e., those given by the robust optimal solution. These diagrams are also random (neither the nominal nor the robust solution cannot be implemented exactly!). However, it turns out that they are incomparably closer to the target (and to each other) than the diagrams associated with the optimal solution to the nominal problem. (Look at a typical robust diagram shown on Fig. 3.4.) With the safety parameter $\omega = 1$, the robust optimal value is 0.0817; although it is 30% larger than the nominal optimal value 0.0621, the robust optimal value has a definite advantage in that it says something reliable about the quality of actual diagrams that we can obtain when implementing the robust optimal solution. In a sample of 40 realizations of the diagrams corresponding to the robust optimal solution, the uniform distances from the target varied from 0.0814 to 0.0830.

We have built the robust optimal solution under the assumption that the implementation errors do not exceed 0.1%. What happens if in reality the errors are larger—say, 1%? It turns out that nothing dramatic happens. Now in a sample of 40 diagrams given by the old robust optimal solution (now affected 1% implementation errors) the uniform distances from the target varied from 0.0834 to 0.116. Imagine what will happen with the nominal solution under the same circumstances.

The last issue to be addressed here is, Why is the nominal solution so unstable? And why with the robust counterpart approach were we able to get a solution that is incomparably

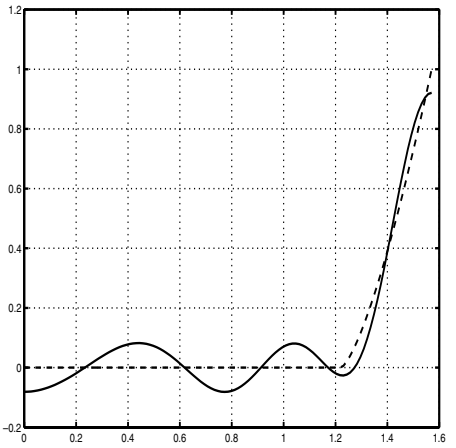


Figure 3.4. *A robust diagram. Uniform distance from the target is 0.0822. (The safety parameter for the uncertainty ellipsoids is $\omega = 1$.)*

better, as far as actual implementation is concerned? The answer becomes clear when we look at the nominal and the robust optimal amplification coefficients:

j	1	2	3	4	5	6	7	8	9	10
x_j^{nom}	1624.4	−14701	55383	−107247	95468	19221	−138622	144870	−69303	13311
x_j^{rob}	−0.3010	4.9638	−3.4252	−5.1488	6.8653	5.5140	5.3119	−7.4584	−8.9140	13.237

It turns out that the nominal problem is ill-posed. Although its optimal solution is far away from the origin, there is a massive set of nearly optimal solutions, and among the latter ones we can choose solutions of quite moderate magnitude. Indeed, here are the optimal values obtained when we add to the constraints of (Nom) the box constraints $|x_j| \leq L$, $j = 1, \ldots, 10$:

L	1	10	10^2	10^3	10^4	10^5	10^6	10^7
Opt_Val	0.09449	0.07994	0.07358	0.06955	0.06588	0.06272	0.06215	0.06215

Since the implementation inaccuracies for a solution are larger the larger it is, there is no surprise that our huge nominal solution results in a very unstable actual design. In contrast to this, the robust counterpart penalizes the (properly measured) magnitude of x (look at the terms $\|Q_i x\|_2$ in the constraints of (Rob)) and therefore yields a much more stable design. Note that this situation is typical for many applications: the nominal solution is on the boundary of the nominal feasible domain, and there are nearly optimal solutions to the nominal problem that are in the deep interior of this domain. When solving the nominal problem, we do not care if there is a reasonable tradeoff between the depth of feasibility and the optimality: *any* improvement in the objective is sufficient to make the solution just marginally feasible for the nominal problem. And a solution that is only marginally feasible in the nominal problem can easily become very infeasible when the data are perturbed. This would not be the case for a deeply interior solution. With the robust counterpart approach, we do use certain tradeoffs between the depth of feasibility and the optimality—we are trying to find something like the deepest feasible nearly optimal solution. As a result, we

normally gain a lot in stability; and if, as in our example, there are deeply interior nearly optimal solutions, we do not lose that much in optimality.

Robust counterpart to uncertain linear programming with a conic quadratic representable uncertainty set. We have seen that the robust counterpart of uncertain LP with simple constraintwise ellipsoidal uncertainty is a conic quadratic problem. This fact is a special case of the following proposition.

Proposition 3.4.1. *Consider an uncertain LP*

$$\mathcal{LP}(\mathcal{U}) = \left\{ \min_{x:Ax\geq b} c^T x \big| (c, A, b) \in \mathcal{U} \right\}$$

and assume that the uncertainty set $\mathcal{U}$ is CQr:

$$\mathcal{U} = \left\{ \zeta = (c, A, B) \in \mathbf{R}^n \times \mathbf{R}^{m\times n} \times \mathbf{R}^m \big| \exists u : \mathcal{A}(\zeta, u) \equiv P\zeta + Qu + r \geq_{\mathbf{K}} 0 \right\},$$

where $\mathcal{A}(\zeta, u)$ is an affine mapping and $\mathbf{K}$ is a direct product of ice cream cones. Assume, further, that the above CQR of $\mathcal{U}$ is strictly feasible:

$$\exists(\bar{\zeta}, \bar{u}) : \quad \mathcal{A}(\bar{\zeta}, \bar{u}) >_{\mathbf{K}} 0.$$

Then the robust counterpart of $\mathcal{LP}(\mathcal{U})$ is equivalent to an explicit conic quadratic problem.

Proof. Introducing an additional variable t and denoting by $z = (t, x)$ the extended vector of design variables, we can write down the instances of our uncertain LP in the form

$$\min_z \left\{ d^T z \mid \alpha_i^T(\zeta) z - \beta_i(\zeta) \geq 0, \ i = 1, \ldots, m+1 \right\} \qquad (\mathrm{LP}[\zeta])$$

with an appropriate vector d. Here the functions

$$\alpha_i(\zeta) = A_i\zeta + a_i, \quad \beta_i(\zeta) = b_i^T\zeta + c_i$$

are affine in the data vector ζ. The robust counterpart of our uncertain LP is the optimization program

$$\min_z \left\{ d^T z \to \min \mid \alpha_i^T(\zeta) z - \beta_i(\zeta) \geq 0 \quad \forall \zeta \in \mathcal{U} \ \forall i = 1, \ldots, m+1 \right\}. \qquad (\mathrm{RC_{ini}})$$

Let us fix i and ask ourselves what it means that a vector z satisfies the infinite system of linear inequalities

$$\alpha_i^T(\zeta) z - \beta_i(\zeta) \geq 0 \quad \forall \zeta \in \mathcal{U}. \qquad (\mathrm{C}_i)$$

Clearly, a given vector z possesses this property if and only if the optimal value in the optimization program

$$\min_{\tau,\zeta} \left\{ \tau \mid \tau \geq \alpha_i^T(\zeta) z - \beta_i(\zeta), \ \zeta \in \mathcal{U} \right\}$$

is nonnegative. Recalling the definition of $\mathcal{U}$, we see that the latter problem is equivalent to the conic quadratic program

$$\min_{\tau,\zeta,u}\left\{\tau \mid \tau \geq \alpha_i^T(\zeta)z - \beta_i(\zeta) \equiv [\underbrace{A_i\zeta + a_i}_{\alpha_i(\zeta)}]^T z - [\underbrace{b_i^T\zeta + c_i}_{\beta_i(\zeta)}],\ \mathcal{A}(\zeta,u) \equiv P\zeta + Qu + r \geq_{\mathbf{K}} 0\right\} \tag{CQ$_i[z]$}$$

in variables τ, ζ, u. Thus, z satisfies (C$_i$) if and only if the optimal value in (CQ$_i[z]$) is nonnegative.

Since by assumption the system of conic quadratic inequalities $\mathcal{A}(\zeta, u) \geq_{\mathbf{K}} 0$ is strictly feasible, the conic quadratic program (CQ$_i[z]$) is strictly feasible. By the conic duality theorem, if (a) the optimal value in (CQ$_i[z]$) is nonnegative, then (b) the dual to (CQ$_i[z]$) problem admits a feasible solution with a nonnegative value of the dual objective. By weak duality, (b) implies (a). Thus, the fact that the optimal value in (CQ$_i[z]$) is nonnegative is equivalent to the fact that the dual problem admits a feasible solution with a nonnegative value of the dual objective:

$$\begin{array}{c} z \text{ satisfies (C}_i\text{)} \\ \Updownarrow \\ \text{Opt(CQ}_i[z]) \geq 0 \\ \Updownarrow \\ \left\{\begin{array}{c} \exists \lambda \in \mathbf{R}, \xi \in \mathbf{R}^N (N \text{ is the dimension of } \mathbf{K}): \\ \lambda[a_i^T z - c_i] - \xi^T r \geq 0, \\ \lambda = 1, \\ -\lambda A_i^T z + b_i + P^T\xi = 0, \\ Q^T\xi = 0, \\ \lambda \geq 0, \\ \xi \geq_{\mathbf{K}} 0. \end{array}\right. \\ \Updownarrow \\ \left\{\begin{array}{c} \exists \xi \in \mathbf{R}^N: \\ a_i^T z - c_i - \xi^T r \geq 0 \\ -A_i^T z + b_i + P^T\xi = 0, \\ Q^T\xi = 0, \\ \xi \geq_{\mathbf{K}} 0. \end{array}\right. \end{array}$$

We see that the set of vectors z satisfying (C$_i$) is CQr:

$$\begin{array}{c} z \text{ satisfies (C}_i\text{)} \\ \Updownarrow \\ \left\{\begin{array}{c} \exists \xi \in \mathbf{R}^N: \\ a_i^T z - c_i - \xi^T r \geq 0, \\ -A_i^T z + b_i + P^T\xi = 0, \\ Q^T\xi = 0, \\ \xi \geq_{\mathbf{K}} 0. \end{array}\right. \end{array}$$

Consequently, the set of robust feasible z—those satisfying (C$_i$) $\forall i = 1, \ldots, m+1$—is CQr (as the intersection of finitely many CQr sets), whence the robust counterpart of

our uncertain LP, being the problem of minimizing a linear objective over a CQr set, is equivalent to a conic quadratic problem. Here is this problem:

$$\begin{array}{c}\text{minimize } d^T z,\\ \left\{\begin{array}{c} a_i^T z - c_i - \xi_i^T r \geq 0,\\ -A_i^T z + b_i + P^T \xi_i = 0,\\ Q^T \xi_i = 0,\\ \xi_i \geq_{\mathbf{K}} 0 \end{array}\right. \quad i = 1, \ldots, m+1,\end{array}$$

with design variables $z, \xi_1, \ldots, \xi_{m+1}$. Here A_i, a_i, c_i, b_i come from the affine functions $\alpha_i(\zeta) = A_i\zeta + a_i$ and $\beta_i(\zeta) = b_i^T\zeta + c_i$, while P, Q, r come from the description of $\mathcal{U}$:

$$\mathcal{U} = \{\zeta \mid \exists u : \quad P\zeta + Qu + r \geq_{\mathbf{K}} 0\}. \qquad \square$$

Conic quadratic representability of the optimal value in a conic quadratic program as a function of the data. Consider a conic quadratic program

$$\min_x \left\{c^T x \mid Ax - b \geq_{\mathbf{K}} 0\right\}, \tag{3.4.40}$$

where $\mathbf{K}$ is a direct product of ice cream cones. The optimal value of the problem clearly is a function of the data (c, A, b) of the problem. What can be said about CQ-representability of this function? In general, not much: the function is not even convex. There are, however, two modifications of our question that admit good answers. Namely, under mild regularity assumptions,

(a) with c, A fixed, the optimal value is a CQr function of the right-hand side vector b;

(b) with A, b fixed, the minus optimal value is a CQr function of c.

Here are the exact forms of our claims:

Proposition 3.4.2. *Let c, A be fixed and such that the system*

$$A^T y = c, \quad y >_{\mathbf{K}} 0,$$

is solvable. Then the optimal value of the problem is a CQr function of b.

The statement is quite evident. Our assumption on c, A means that the problem dual to (3.4.40) is strictly feasible. Thus, if b is such that (3.4.40) is feasible, then the optimal value $\mathrm{Opt}(b)$ in the problem is achieved (by conic duality theorem); otherwise $\mathrm{Opt}(b) = +\infty$. Thus,

$$\mathrm{Opt}(b) \leq t \Leftrightarrow \exists x : \left\{\begin{array}{l} c^T x \leq t,\\ Ax - b \geq_{\mathbf{K}} 0,\end{array}\right.$$

which is, essentially, a CQR for $\mathrm{Opt}(b)$. In this CQR, b, t are the variables of interest, and x plays the role of the additional variable.

The claim (b) is essentially a corollary of (a)—via duality, the optimal value in (3.4.40) is, up to pathological cases, the same as the optimal value in the dual problem, in which c becomes the right-hand-side vector. Here is the exact formulation.

Proposition 3.4.3. *Let A, b be such that* (3.4.40) *is strictly feasible. Then the minus optimal value* $-\mathrm{Opt}(c)$ *of the problem is a CQr function of c, with CQR induced by the equivalence*

$$\mathrm{Opt}(c) \geq t \Leftrightarrow \exists y : \left\{ \begin{array}{c} b^T y \geq t, \\ A^T y = c, \\ y \geq_{\mathbf{K}} 0. \end{array} \right.$$

A careful reader will have realized that Proposition 3.4.1 is nothing but a straightforward application of Proposition 3.4.3.

Remark 3.4.1. Looking at the proof of Proposition 3.4.1, we see that the assumption that uncertainty set $\mathcal{U}$ is CQr plays no crucial role. What is important is that $\mathcal{U}$ is the projection on the ζ-space of the solution set of a strictly feasible conic inequality associated with a certain cone $\mathbf{K}$. Whenever this is the case, the above construction demonstrates that the robust counterpart of $\mathcal{LP}(\mathcal{U})$ is a conic problem associated with the cone which is a direct product of several cones dual to $\mathbf{K}$. When the uncertainty set is polyhedral (i.e., it is given by finitely many scalar linear inequalities: $\mathbf{K} = \mathbf{R}^m_+$), the robust counterpart of $\mathcal{LP}(\mathcal{U})$ is an explicit LP program. (In this case we can eliminate the assumption that the conic inequality defining $\mathcal{U}$ is strictly feasible (why?).)

Similarly, Propositions 3.4.2 and 3.4.3 remain valid for an arbitrary conic program, up to the fact that in this general case we should speak about the representability of the epigraphs of $\mathrm{Opt}(b)$ and $-\mathrm{Opt}(c)$ via conic inequalities associated with direct products of cones $\mathbf{K}$, and their duals, rather than about CQ-representability. In particular, Propositions 3.4.1, 3.4.2, and 3.4.3 remain valid in the case of semidefinite representability, to be discussed in Lecture 4.

3.4.3 Truss topology design

We dealt with the TTD problem in Lecture 1, where we managed to pose it as an LP program. However, this was the simplest case of the problem, with only a single external load operating on the structure. What if we control the compliances with respect to several nonsimultaneous external forces? or if we have restrictions, like upper and lower bounds on bar volumes? Besides this, we do not understand the miracle that happened in section 1.3.5, where a highly nonlinear TTD problem became an LP program. Miracles are not as useful as one would think—they do not yield understanding (once understood, a miracle is not a miracle anymore) and therefore cannot be reproduced when necessary. We are about to improve our understanding of the TTD problem and, in particular, to study its more general settings.

Recalling Lecture 1, the mathematical description of the TTD problem is as follows:

> Given $m \times m$ dyadic matrices $b_i b_i^T$, $i = 1, \ldots, n$, an n-dimensional vector t with nonnegative entries (a truss), and an m-dimensional vector f (a load), we define the compliance $\mathrm{Compl}_f(t)$ of the truss t with respect to the load f as the quantity $\frac{1}{2} f^T v$, where v (the equilibrium displacement) solves the equilibrium equation
>
> $$A(t)v = f,$$

and

$$A(t) = \sum_{i=1}^{n} t_i b_i b_i^T$$

is the *bar-stiffness matrix*. If the equilibrium equation has no solutions, the compliance, by definition, is $+\infty$.

The TTD problem is as follows. Given a ground structure $(n, m, \{b_i\}_{i=1}^n)$, a load f, and a resource w, find a truss satisfying the resource restriction

$$\sum_{i=1}^{n} t_i \le w$$

(and, of course, the constraints $t \ge 0$) with the minimum possible compliance $\mathrm{Compl}_f(t)$.

There are two parts in the story just recalled:
(a) an excursion to mechanics—the definition of compliance, and
(b) formulation of a particular compliance-related optimization problem.

Instead of the particular problem (b), we could pose other quite meaningful problems (and we shall). In order to develop a unified approach to all these problems, we should better understand what, mathematically, is the compliance. Our main result will be as follows:

For a given ground structure, the compliance $\mathrm{Compl}_f(t)$*, regarded as a function of variables* (t, f)*, is CQr.*

Our goal is to justify this claim and to get an explicit CQR for the compliance. Equipped with this result, we can process routinely numerous versions of the TTD problems via the machinery of CQP.

The analysis to follow does not depend on the fact that the components $b_i b_i^T$ of the bar-stiffness matrix are dyadic; the essence of the matter is that they are positive semidefinite symmetric matrices. So, from now on, we assume that the bar-stiffness matrix $A(t)$ is of the form

$$A(t) = \sum_{i=1}^{n} t_i B_i B_i^T, \tag{3.4.41}$$

where B_i are $m \times k_i$ matrices ($k_i = 1$ for the original TTD problem). The compliance of the resulting generalized truss $\mathrm{Compl}_f(t)$ is defined exactly as before: it is the quantity $\frac{1}{2} f^T v$, where v is a solution to the equilibrium equation

$$A(t)v = f, \tag{3.4.42}$$

with the same convention as above: $\mathrm{Compl}_f(t) = +\infty$ when the equilibrium equation has no solutions.

Variational principle

Our first observation is as follows.

Proposition 3.4.4. Variational description of compliance. *Consider a ground structure $(n, m, B_1, \ldots, B_n)$ along with a load $f \in \mathbf{R}^m$, and let $t \in \mathbf{R}^n_+$ be a truss. Let us associate with these data the quadratic form (the potential energy of the loaded system)*

$$\mathcal{C}_{t,f}(v) = \frac{1}{2} v^T A(t) v - f^T v, \quad v \in \mathbf{R}^m. \tag{3.4.43}$$

Then, the compliance $\mathrm{Compl}_f(t)$ *is finite if and only if this quadratic form is bounded below on $\mathbf{R}^m$, and in this case one has*

$$-\mathrm{Compl}_f(t) = \min_v \mathcal{C}_{t,f}(v). \tag{3.4.44}$$

Proof. Since $t \geq 0$, the matrix $A(t)$ is positive semidefinite. Now let us use the following general and well-known fact.

Lemma 3.4.1. *Let*

$$\mathcal{A}(v) = \frac{1}{2} v^T A v - b^T v$$

be a quadratic form on $\mathbf{R}^m$ with symmetric positive semidefinite matrix A. Then

(i) *$\mathcal{A}(v)$ is bounded below if and only if it attains its minimum;*

(ii) *$\mathcal{A}(v)$ attains its minimum if and only if the equation*

$$Av = b \tag{$*$}$$

is solvable, in which case the set of minimizers of $\mathcal{A}(v)$ is exactly the set of solutions to the equation;

(iii) *the minimum value of the quadratic form, if it exists, is equal to $-\frac{1}{2} b^T v$, where v is (any) solution to (*).*

Proof. (i): There are two possibilities:

(a) b is orthogonal to $\mathrm{Null}(A)$,

(b) b has a nonzero projection b' onto $\mathrm{Null}(A)$.

In case of (b) the form is clearly unbounded below (look what happens when $v = tb'$ and $t \to \infty$). In case of (a) the equation $Av = b$ is solvable[16]; at every solution to this equation, the gradient of $\mathcal{A}$ vanishes, so that such a solution is a minimizer of the *convex* (A is positive semidefinite!) function $\mathcal{A}(\cdot)$. Thus, if the form is bounded below, then it attains its minimum, and, of course, vice versa.

(ii): Since the form is convex and smooth, its minimizers are exactly the same as its critical points—those where the gradient vanishes. The gradient of $\mathcal{A}(v)$ is $Av - b$, so that it vanishes exactly at the solutions to (*).

[16] Linear algebra says that a linear system $Px = q$ is solvable if and only if q is orthogonal to $\mathrm{Null}(P^T)$. We use this fact for the particular case of $P = P^T$.

(iii): Let v be a solution to (*) or, which is the same, a minimizer of $\mathcal{A}(\cdot)$. From (*) we get $v^T A v = b^T v$, so that $\mathcal{A}(v) = \frac{1}{2} b^T v - b^T v = -\frac{1}{2} b^T v$. □

In view of Lemma 3.4.1, the energy (3.4.43) is bounded below if and only if the equilibrium equation (3.4.42) is solvable, and if it is the case, the minimum of the energy is $-\frac{1}{2} f^T v$, where v is a solution to the equilibrium equation (3.4.42). Recalling the definition of the compliance, we come to the desired result. □

As a byproduct, we have obtained the following.

Variational Principle. *The equilibrium displacement of a truss t under an external load f is a minimizer of the quadratic form*

$$\frac{1}{2} v^T A(t) v - f^T v$$

of a displacement vector v; if this quadratic form is unbounded below, there is no equilibrium at all.

This is a typical variational principle in mechanics and physics. These principles state that equilibria in certain physical systems occur at critical points (in good cases—even minimizers) of certain energy functionals. Variational principles are extremely powerful, and in mechanical, electrical, and other applications an issue of primary importance is to identify a tractable variational principle governing the model.

From variational principle to conic quadratic-representation of compliance

Step 1. Let us look at the epigraph

$$\mathcal{C} = \{(t, f; \tau) \mid t \geq 0, \tau \geq \mathrm{Compl}_f(t)\}$$

of the compliance in the domain $t \geq 0$. Our goal is to find an explicit CQR of this set. To this end let us start with a slightly smaller set

$$\mathcal{C}' = \{(t, f, \tau) \mid t \geq 0, \tau > \mathrm{Compl}_f(t)\}.$$

Proposition 3.4.4 provides us with the following description of $\mathcal{C}$ and $\mathcal{C}'$:

($\mathcal{C}$): The set $\mathcal{C}$ comprises all triples $(t \geq 0, f, \tau)$ such that the quadratic form

$$Q(v) = \frac{1}{2} v^T A(t) v - f^T v + \tau$$

of $v \in \mathbf{R}^m$ is nonnegative everywhere.

($\mathcal{C}'$): The set $\mathcal{C}$ comprises all triples $(t \geq 0, f, \tau)$ such that the form $Q(v)$ is positive everywhere.

($\mathcal{C}'$) says that the convex quadratic inequality

$$Q(v) \leq 0 \tag{3.4.45}$$

has *no* solutions. Recall that a convex quadratic inequality can be represented via a conic quadratic inequality; what is the latter inequality in the case of (3.4.45)? To answer, set

$$B(t) = \sqrt{2}\begin{pmatrix} \sqrt{t_1}B_1^T \\ \sqrt{t_2}B_2^T \\ \cdots \\ \sqrt{t_n}B_n^T \end{pmatrix} \tag{3.4.46}$$

and express the bar-stiffness matrix as

$$A(t) = \sum_{i=1}^{n} t_i B_i B_i^T = \frac{1}{2}B^T(t)B(t).$$

The quadratic form $Q(v)$ now can be written as

$$Q(v) = \frac{1}{4}v^T B^T(t)B(t)v - f^T v + \tau = \frac{1}{4}\left[\|B(t)v\|_2^2 + (1 - f^T v + \tau)^2 - (1 + f^T v - \tau)^2\right]. \tag{3.4.47}$$

We obtain the following observation.

> **(O):** *Inequality* (3.4.45) *has no solutions if and only if the conic quadratic inequality in variables* v
>
> $$\left\| \begin{pmatrix} B(t)v \\ 1 - f^T v + \tau \end{pmatrix} \right\|_2 \le 1 + f^T v - \tau \tag{3.4.48}$$
>
> *has no solution.*
>
> Indeed, the relation between the inequalities (3.4.45) and (3.4.48) is as follows: The former, in view of (3.4.47), is the inequality $\frac{1}{4}P^2(v) \le \frac{1}{4}p^2(v)$, while the latter is $P(v) \le p(v)$; here $P(v)$ is the Euclidean norm of a certain vector depending on v. Taking into account that $P(\cdot)$ is always nonnegative, and that $p(v) = 1 + f^T v - \tau$ must be nonnegative at every solution of (3.4.45), we conclude that both inequalities have the same set of solutions.

Step 2. As we have seen, $\mathcal{C}'$ is exactly the set of values of the parameter (t, f, τ) for which the CQI (3.4.48) is not solvable. In fact, one can say more:

> (!) *When the triple* (t, f, τ) *is in* $\mathcal{C}'$*, the CQI* (3.4.48) *is not even almost solvable (see Proposition 2.4.2).*

Indeed, (3.4.48) is of the form

$$Av - b \equiv \begin{pmatrix} B(t)v \\ -f^T v \\ +f^T v \end{pmatrix} - \begin{pmatrix} 0 \\ -1 - \tau \\ -1 + \tau \end{pmatrix} \succeq_{\mathbf{L}^k} 0 \tag{3.4.49}$$

with certain k. Assume that $(t, f, \tau) \in \mathcal{C}'$. What we should prove is that then all CQIs of the form

$$\begin{pmatrix} B(t)v \\ -f^T v \\ +f^T v \end{pmatrix} - \begin{pmatrix} \epsilon \\ -1-\tau+\epsilon_1 \\ -1+\tau+\epsilon_2 \end{pmatrix} \geq_{\mathbf{L}^k} 0 \tag{3.4.50}$$

with small enough perturbation vector ϵ and scalars ϵ_1, ϵ_2 are not solvable. Assume, to the contrary, that (3.4.50) with some fixed perturbations $\epsilon, \epsilon_1, \epsilon_2$ has a solution. Then the quadratic inequality in variables v

$$\frac{1}{4}\|B(t)v-\epsilon\|_2^2 + \frac{1}{4}(1+\tau-f^T v-\epsilon_1)^2 \leq \frac{1}{4}(1-\tau+f^T v-\epsilon_2)^2$$

has a solution. The inequality is of the form

$$\begin{aligned} &\tfrac{1}{2}v^T A(t)v - F^T(\epsilon,\epsilon_1,\epsilon_2)v + T(\epsilon,\epsilon_1,\epsilon_2) \\ &\quad \equiv \tfrac{1}{2}v^T A(t)v - \left[\tfrac{1}{2}B^T(t)\epsilon + \left(1-\tfrac{\epsilon_1+\epsilon_2}{2}\right)f\right]^T v \\ &\qquad + \frac{(2\tau-\epsilon_1+\epsilon_2)(2-\epsilon_1-\epsilon_2)}{4} \\ &\qquad \leq 0. \end{aligned} \tag{3.4.51}$$

Now, since (3.4.45) has no solutions, we have $f = A(t)e$, where e is a minimizer of the unperturbed quadratic form $Q(v)$ (see Lemma 3.4.1). Now, since $A(t) = B^T(t)B(t)$, the image of $A(t)$ is exactly the same as the image of $B^T(t)$, and $A(t)$ is invertible on its image. In other words, there exists a matrix R such that $B^T(t)z = A(t)RB^T(t)z$ for every z, and, in particular, $B^T(t)\epsilon = A(t)RB^T(t)\epsilon$. We see that the vector

$$v(\epsilon,\epsilon_1,\epsilon_2) = \left(1-\frac{\epsilon_1+\epsilon_2}{2}\right)e + \frac{1}{2}RB^T(t)\epsilon$$

is a minimizer, over v, of the left-hand side in (3.4.51); obviously, this minimizer depends continuously on the perturbation $(\epsilon,\epsilon_1,\epsilon_2)$. The coefficients of the quadratic form of v in the left-hand side of (3.4.51) are also continuous in the perturbation; consequently, the minimum value of the form depends continuously on the perturbation. This is all we need. Assuming that the conclusion in (!) fails to be true for a fixed triple $(t, f, \tau) \in \mathcal{C}'$, we would conclude that there exist arbitrarily small perturbations such that (3.4.51) has a solution, so that the minimum over v of the left-hand side in (3.4.51) for these perturbations is nonpositive. By continuity, it follows that the minimum value of the quadratic form in the left-hand side of (3.4.51) is nonpositive when the perturbation is 0, i.e., the minimum value of the quadratic form $Q(v)$ is nonpositive. But then (3.4.45) has a solution (recall that below bounded quadratic form achieves its minimum), which is a contradiction—the parameter (t, f, τ) belongs to $\mathcal{C}'$!

Step 3. In fact, (!) can be easily replaced by the following stronger claim:

(!!) *A triple* (t, f, τ) *with* $t \geq 0$ *belongs to* $\mathcal{C}'$ *if and only if* (3.4.48) *is not almost solvable.*

Indeed, the (relatively difficult) "only if" part is given by (!). And the "if" part is immediate: we should prove that if $t \geq 0$ and (t, f, τ) does not belong to $\mathcal{C}'$ (i.e., if $\inf_v Q(v) \leq 0$; see $(\mathcal{C}')$), then (3.4.48) is almost solvable. But in the case in question, the CQI (3.4.48) is simply solvable, even without "almost." Indeed, we remember from (O) that (3.4.48) is solvable if and only if (3.4.45) is solvable. Now, if the form Q is bounded below, its minimum is attained, so that under the assumption $\inf_v Q(v) \leq 0$ (3.4.45) is solvable. Finally, in the case when Q is not bounded below, (3.4.45) is solvable by evident reasons!

Step 4. Combining (!!) with Proposition 2.4.2(iii), we come to the following result.

Lemma 3.4.2. *A triple (t, f, τ) with $t \geq 0$ belongs to the set $\mathcal{C}'$, i.e.,* $\mathrm{Compl}_f(t) < \tau$, *if and only if there exists a vector λ satisfying the relations (cf.* (3.4.49)*)*

$$A^T\lambda = 0, b^T\lambda > 0, \lambda \geq_{\mathbf{L}^k} 0. \tag{3.4.52}$$

To understand the meaning of (3.4.52), observe that by the definition of $B(t)$ and in view of (3.4.49) one has

$$\begin{array}{rcl} A^T & = & [\sqrt{2t_1}B_1; \sqrt{2t_2}B_2; \ldots; \sqrt{2t_n}B_n; -f; f], \\ b^T & = & [0; \ldots; 0; -1-\tau; -1+\tau]. \end{array}$$

Partitioning λ accordingly: $\lambda^T = [w_1; \ldots, ; w_n; p; q]$, we can rewrite (3.4.52) equivalently as

$$\begin{array}{lrcl} \text{(a)} & \displaystyle\sum_{i=1}^n (2t_i)^{1/2} B_i w_i & = & (p-q)f; \\ \text{(b)} & p(-1-\tau) + q(-1+\tau) & > & 0; \\ \text{(c)} & \sqrt{\left[\displaystyle\sum_{i=1}^n w_i^T w_i\right] + p^2} & \leq & q. \end{array} \tag{3.4.53}$$

In every solution $(\{w_i\}, p, q)$ to (3.4.53) necessarily $p \neq q$ (and, therefore, $p < q$ by (c)). Indeed, otherwise (b) would imply $-2q > 0$, which is impossible in view of (c). Consequently, we can define the vectors

$$s_i = -(q-p)^{-1}\sqrt{2t_i}w_i,$$

and with respect to these vectors (3.4.53) becomes

$$\begin{array}{rcl} \displaystyle\sum_{i=1}^n B_i s_i & = & f, \\ \displaystyle\sum_{i=1}^n \frac{s_i^T s_i}{2t_i} & \leq & \frac{q+p}{q-p}, \\ & < & \tau. \end{array} \tag{3.4.54}$$

(The concluding inequality is given by (3.4.53)(b).)

We have covered 99% of the way to our target, namely, we have basically proved the following.

LEMMA 3.4.3. *A triple $(t \geq 0, f, \tau)$ is such that $\mathrm{Compl}_f(t) < \tau$ if and only if there exist vectors s_i, $i = 1, \ldots, n$, satisfying the relations*

$$\begin{array}{rcl} \sum_{i=1}^n B_i s_i & = & f, \\ \sum_{i=1}^n \frac{s_i^T s_i}{2t_i} & < & \tau. \end{array} \tag{3.4.55}$$

(From now on, by definition, $0/0 = 0$ and $a/0 = +\infty$ when $a > 0$.)

Proof. Lemma 3.4.2 says that if $\mathrm{Compl}_f(t) < \tau$, then (3.4.52) is solvable; and as we just have seen, a solution to (3.4.52) can be converted to a solution to (3.4.55). Conversely, given a solution $\{s_i\}_{i=1}^n$ to (3.4.55), we can find $q > 1/2$ satisfying the relations

$$\sum_{i=1}^n \frac{s_i^T s_i}{2t_i} < 2q - 1 < \tau;$$

setting $p = q - 1$, $w_i = -(2t_i)^{-1/2} s_i$ ($w_i = 0$ when $t_i = 0$), it is immediately seen that we get a solution to (3.4.53) or, which is the same, to (3.4.52). Again, from Lemma 3.4.2 it follows that $\mathrm{Compl}_f(t) < \tau$. □

Step 5. We cover the remaining 1% of the way to the target in the following.

PROPOSITION 3.4.5. *A triple (t, f, τ) belongs to the epigraph of the function $\mathrm{Compl}_f(t)$ (extended by the value $+\infty$ to the set of those t's that are not nonnegative), i.e., $\mathrm{Compl}_f(t) \leq \tau$, if and only if there exist vectors s_i, $i = 1, \ldots, n$, such that the following relations are satisfied:*

$$\begin{array}{lrcl} \text{(a)} & \sum_{i=1}^n B_i s_i & = & f, \\ \text{(b)} & \sum_{i=1}^n \frac{s_i^T s_i}{2t_i} & \leq & \tau, \\ \text{(c)} & t & \geq & 0. \end{array} \tag{3.4.56}$$

In particular, the function $\mathrm{Compl}_f(t)$ is CQr.

Proof. If we can extend a given triple (t, f, τ), by properly chosen s_i's, to a solution of (3.4.56), then, by Lemma 3.4.3, $\mathrm{Compl}_f(t) < \tau'$ for every $\tau' > \tau$, whence $\mathrm{Compl}_f(t) \leq \tau$. Conversely, assume that $\mathrm{Compl}_f(t) \leq \tau$. Then for sure $\mathrm{Compl}_f(t) < \tau + 1$, and by Lemma 3.4.3 the optimization problem

$$\min_{s_1, \ldots, s_n} \left\{ \sum_{i=1}^n \frac{s_i^T s_i}{2t_i} \mid \sum_{i=1}^n B_i s_i = f \right\} \tag{P}$$

is feasible. But this is, essentially, a problem of minimizing a *convex quadratic* nonnegative objective over an *affine* plane (for those i with $t_i = 0$, s_i should be zeros, and we may simply ignore the corresponding terms). Therefore the problem is solvable. Let s_i^* form an optimal solution, and τ^* be the optimal value in (P). By Lemma 3.4.3, if $\tau > \mathrm{Compl}_f(t)$, then $\sum_{i=1}^n (2t_i)^{-1}(s_i^*)^T s_i^* = \tau^* < \tau$. Consequently, if $\tau \geq \mathrm{Compl}_f(t)$, then $\sum_{i=1}^n (2t_i)^{-1}(s_i^*)^T s_i^* \leq \tau$, so that s_i^* form the required solution to (3.4.56).

It remains to prove that the compliance, regarded as a function of t, f, is CQr. But this is clear—the function $\sum_i (2t_i)^{-1} s_i^T s_i$ is CQr (as a sum of fractional-quadratic functions), so that the second inequality in (3.4.56) defines a CQr set. All remaining relations in (3.4.56) are linear inequalities and equations, and from our calculus of CQr sets we know that constraints of this type do not spoil CQ-representability. □

Remark 3.4.2. Note that after the CQr description (3.4.56) of the epigraph of compliance is guessed, it can be justified directly by a more or less simple reasoning. (This, is, basically, how we handled the TTD problem in Lecture 1.) The goal of our exposition was not merely to justify (3.4.56), but to demonstrate that one can derive this representation quite routinely from the variational principle (which is the only methodologically correct definition of compliance) and from the rule, "When looking at a convex quadratic inequality, do not trust your eyes: what you really see is a conic quadratic inequality."

Remarks and conclusions

Mechanical interpretation of Proposition 3.4.5. Consider the case of a true truss—the one where $B_i = b_i$ are just vectors. Here s_i are reals, and from the mechanical viewpoint, the quantity $s_i b_i$ represents the reaction force in bar i (so that the constraint $\sum_{i=1}^m B_i s_i = f$ in (3.4.56) says that the sum of the reaction forces should compensate the external load). Now, recall that the magnitude of the reaction force caused by elongation Δ_i of bar i is, up to constant factor, $\frac{\Delta_i \sigma_i}{\ell_i}$, where σ_i, ℓ_i are bar's cross-sectional area and length, respectively, while $\|b_i\|_2$ is inverse proportional to ℓ_i. Comparing two expressions for the reaction force in bar i, namely, $s_i b_i$ and $\mathrm{const}\frac{\Delta_i \sigma_i}{\ell_i}\frac{b_i}{\|b_i\|_2}$, we conclude that $s_i = \mathrm{const}\,\Delta_i \sigma_i$. Thus, s_i, essentially, are the elongations of the bars multiplied by the cross-sectional areas of the bars. Note that then $\frac{s_i^2}{t_i} = \mathrm{const}\frac{\Delta_i^2 \sigma_i^2}{\sigma_i \ell_i} = \mathrm{const}\frac{\Delta_i^2 \sigma_i}{\ell_i}$, and the latter expression is, up to a constant factor, the energy stored in bar i as a result of its elongation Δ_i. Thus, the objective in (3.4.56) is proportional to the energy stored in the truss as a result of its deformation. Proposition 3.4.5 says that the compliance of a given truss under a given load is the minimum of the quantities $\sum_{i=1}^n \frac{s_i^2}{2t_i}$ over all collections $\{s_i\}$ satisfying the equation (3.4.56)(a). In other words, we get another variational principle:

> *The reaction forces in a loaded truss minimize the total energy stored by the bars under the constraint that the sum of the reaction forces compensates the external load.*

Multiload truss topology design problem. The result of Proposition 3.4.5—the fact that the compliance admits explicit CQR—allows us to process numerous versions of the TTD problem via the machinery of CQP. Consider, e.g., the multiload TTD problem.

PROBLEM 3.4.1. Multiload TTD problem. *Given a ground structure* $(n, m, b_1, \ldots, b_n)$*, a finite set of loading scenarios* $f_1, \ldots, f_k \in \mathbf{R}^m$*, and a material resource* $w > 0$*, find a truss* t *of total bar volume not exceeding* w *that is stiffest, in the worst-case sense, with respect to the loads* $f_1, \ldots, f_m$*; i.e., find* t *that minimizes the worst-case compliance* $\max_{j=1,\ldots,k} \mathrm{Compl}_{f_j}(t)$.

The origin of the problem is clear: in reality, a truss should withstand not merely a single load of interest but numerous (nonsimultaneous) loads. For example, an engineer designing a bridge should consider rush hour traffic, night traffic, earthquake, side wind, etc.

Equipped with Proposition 3.4.5, we can immediately pose the multiload TTD problem as the following conic quadratic program:

$$\begin{array}{lll} \text{minimize} & \tau & \\ \text{s.t.} & & \\ \text{(a)} & s_{ij}^2 \le 2t_i\sigma_{ij}, \ i = 1, \ldots, n, \ j = 1, \ldots, k; & \\ \text{(b)} & \sum_{i=1}^{n} \sigma_{ij} \le \tau, \ j = 1, \ldots, k; & \\ \text{(c)} & t_i, \sigma_{ij} \ge 0, \ i = 1, \ldots, n, \ j = 1, \ldots, k; & (3.4.57) \\ \text{(d)} & \sum_{i=1}^{n} t_i \le w; & \\ \text{(e)} & \sum_{i=1}^{n} s_{ij} b_i = f_j, \ j = 1, \ldots, k, & \end{array}$$

with design variables $\tau, t_i, s_{ij}, \sigma_{ij}$.

The structure of the constraints is clear. For every fixed $j = 1, \ldots, k$, the corresponding equations (e), (a), (b) and the inequalities $\sigma_{ij} \ge 0$ taken together express the fact that $\sum_{i=1}^n s_{ij} b_i = f_j$ and $\sum_{i=1}^n (2t_i)^{-1} s_{ij}^2 \le \tau$, i.e., that the compliance of truss t with respect to load f_j is $\le \tau$ (Proposition 3.4.5). The remaining inequalities $t_i \ge 0$, $\sum_{i=1}^n t_i \le w$ say that t is an admissible truss. Note that the problem is indeed conic quadratic—every one of the relations (a) says that the triple $(s_{ij}, t_i, \sigma_{ij})$ should belong to the 3D ice cream cone (more exactly, to the image of this cone under a one-to-one linear transformation of $\mathbf{R}^3$; see Example 8 in our catalogue of CQr sets). A nice feature of our approach is that it allows us to handle additional design constraints, e.g., various linear inequalities on t, like upper and lower bounds on the bar volumes t_i. Indeed, adding to (3.4.57) finitely many linear constraints, we still get a conic quadratic problem.

Another advantage is that we can—completely routinely—apply to (3.4.57) the duality machinery (see the exercises to Lecture 3), thus coming to a basically equivalent form of the problem. As we shall see later, the dual problem is incomparably better suited for numerical processing.

Where does the linear programming form of the truss topology design problem come from? We now can explain the miracle we met with in Lecture 1–the ability to pose the simplest case of the TTD problem (which is still highly nonlinear!) as an LP program.

The case in question was the *single-load* TTD problem (a problem of the form (3.4.57) with $k = 1$):

$$\min_{s,t,\tau}\left\{\tau\Big|\sum_{i=1}^n\frac{s_i^2}{2t_i}\le\tau,\ \sum_{i=1}^n s_ib_i=f,\ \sum_{i=1}^n t_i\le w,\ t\ge 0\right\}.$$

By eliminating the variable τ, the problem is reduced to

$$\min_{s,t}\left\{\sum_{i=1}^n\frac{s_i^2}{2t_i}\Big|\sum_{i=1}^n s_ib_i=f,\ \sum_{i=1}^n t_i\le w,\ t\ge 0\right\}.$$

In the latter problem, we can explicitly carry out partial optimization with respect to t by solving, for a given $\{s_i\}$, the problem

$$\min_t\left\{\sum_{i=1}^n\frac{s_i^2}{2t_i}\Big|t\ge 0,\sum_i t_i\le w\right\}.$$

It is clear that at the optimal solution the resource constraint $\sum_i t_i\le w$ is active, and the Lagrange multiplier rule yields that for some $\lambda>0$ and all i we have

$$t_i=\operatorname*{argmin}_{r>0}\left[\frac{s_i^2}{2r}-\lambda r\right]=(2\lambda)^{-1/2}|s_i|.$$

The sum of all t_i's should be w, which leads to

$$2\lambda=\left(\frac{\sum_{l=1}^n|s_l|}{w}\right)^2,$$

whence

$$t_i=\frac{w|s_i|}{\sum_{l=1}^n|s_l|},\ i=1,\dots,n.$$

Substituting the resulting t_i's into the objective, we get

$$\sum_{i=1}^n\frac{s_i^2}{2t_i}=\frac{1}{2w}\left(\sum_{i=1}^n|s_i|\right)^2.$$

The remaining problem, in the s-variables only, becomes

$$\min_s\left\{\frac{1}{2w}\left(\sum_{i=1}^n|s_i|\right)^2:\sum_{i=1}^n s_ib_i=f\right\}.$$

The latter problem is, of course, equivalent to the LP program

$$\min_s \left\{ \sum_i |s_i| \mid \sum_i s_i b_i = f \right\}.$$

The technique of eliminating the variables τ, t we just used can also be employed for the multiload case, but the resulting problem will *not* be an LP. Even the simple single-load case, but one with additional linear constraints on t (e.g., upper and lower bounds on t_i), is not reducible to LP.

3.5 Exercises to Lecture 3

3.5.1 Optimal control in discrete time linear dynamic system

Consider a discrete time linear dynamic system

$$\begin{array}{rcl} x(t) & = & A(t)x(t-1) + B(t)u(t), \ t = 1, 2, \ldots, T, \\ x(0) & = & x_0. \end{array} \tag{S}$$

Here,

- t is the (discrete) time.
- $x(t) \in \mathbf{R}^l$ is the *state* vector: its value at instant t identifies the state of the controlled plant.
- $u(t) \in \mathbf{R}^k$ is the exogeneous input at time instant t; $\{u(t)\}_{t=1}^T$ is the *control*.
- For every $t = 1, \ldots, T$, $A(t)$ and $B(t)$ are given $l \times l$ and $l \times k$ matrices, respectively.

A typical problem of optimal control associated with (S) is to minimize a given functional of the trajectory $x(\cdot)$ under given restrictions on the control. As a simple example of this type, consider the optimization model

$$\min_{u(\cdot)} \left\{ c^T x(T) \mid \frac{1}{2} \sum_{t=1}^T u^T(t) Q(t) u(t) \le w \right\}, \tag{OC}$$

where $Q(t)$ are given positive definite symmetric matrices.

(OC) can be interpreted, e.g., as follows: $x(t)$ represents the position and the velocity of a rocket, $-c^T x$ is the height of the rocket at a state x (so that our goal is to maximize the height of the rocket at the final time T), the equations in (S) represent the dynamics of the rocket, the control is responsible for the profile of the flight, and the left-hand side of the constraint in (OC) is the dissipated energy.

Exercise 3.1. 1. *Use* (S) *to express* $x(T)$ *via the control and convert* (OC) *into a quadratically constrained problem with linear objective in the* u*-variables.*

2. *Convert the resulting problem to a conic quadratic program.*
3. *Pass from the resulting problem to its dual and find the dual optimal solution.*

4. *Assuming $w > 0$, prove that both the primal and the dual are strictly feasible. What are the consequences for the solvability status of the problems? Assuming, in addition, that $x_0 = 0$, what is the optimal value?*

5. *Assume that* (S), (OC) *form a finite-difference approximation to the continuous time optimal control problem*

$$\begin{array}{ll} \textit{minimize} & c^T x(1) \\ \textit{s.t.} & \\ & \frac{d}{d\tau}x(\tau) = \alpha(\tau)x(\tau) + \beta(\tau)u(\tau), 0 \le \tau \le 1, x(0) = 0, \\ & \int_0^1 u^T(\tau)\gamma(\tau)u(\tau)d\tau \le w, \end{array}$$

where $\gamma(\tau)$, for every $\tau \in [0, 1]$, is a positive definite symmetric matrix.

Guess what should be the optimal value.

3.5.2 Conic quadratic representations

Exercise 3.2. *Let $\pi_1, \ldots, \pi_n$ be positive rational numbers. Demonstrate that the set*

$$\{(t, s_1, \ldots, s_n) \in \mathbf{R} \times \mathbf{R}^n_{++} \mid t \ge g(s) \equiv s_1^{-\pi_1} s_2^{-\pi_2} \ldots s_n^{-\pi_n}\}$$

is CQr. What is the conclusion on convexity of the function $g(s)$ with the domain $\mathrm{Dom} g = \mathbf{R}^n_{++}$*?*

Hint. Modify properly the construction from Example 15, Lecture 3.

Among important (convex) elementary functions, it appears that the only two that are *not* CQr are the exponent $\exp\{x\}$ and the minus logarithm $-\ln x$. In a sense, these are not two functions but only one: CQ-representability deals with the geometry of the epigraph of a function, and the epigraphs of $-\ln x$ and $\exp\{x\}$, geometrically, are the same—we merely are looking at the same set from two different directions. Now, why is the exponent not CQr? The answer is intuitively clear: How could we represent a set given by a transcendental inequality by algebraic inequalities? A rigorous proof, however, requires highly nontrivial tools, namely, the Zaidenberg–Tarski theorem:

> *Let B be a semialgebraic set in $\mathbf{R}^n \times \mathbf{R}^m$, i.e., a set given by a finite system of polynomial inequalities (strict as well as nonstrict). Then the projection of B onto $\mathbf{R}^n$ also is semialgebraic.*

Now, by definition, a set $Q \subset \mathbf{R}^n$ is CQr if and only if it is the projection onto $\mathbf{R}^n$ of a specific semialgebraic set Q' of larger dimension—one given by a system of inequalities of the form $\{\|A_i x - b_i\|_2^2 \le (p_i^T x - q_i)^2, p_i^T x - q_i \ge 0\}_{i=1}^N$. Therefore, assuming that the epigraph of the exponent is CQr, we would conclude that it is a semialgebraic set, which in fact is not the case.

Thus, the exponent, the minus logarithm (same as convex power functions with irrational exponentials, like x_+^π), are not captured by CQP. Let us, however, look at the funny construction as follows. As everybody knows,

$$\exp\{x\} = \lim_{k\to\infty}\left(1+\frac{1}{k}x\right)^k.$$

Let us specify here k as an integral power of 2:

$$\exp\{x\} = \lim_{l\to\infty} f_l(x), \qquad f_l(x) = (1+2^{-l}x)^{(2^l)}.$$

Note that every one of the functions $f(l)$ is CQr.

On the other hand, what is the exponent from the computational viewpoint? Something that does not exist! For a computer, the exponent is a function that is well defined on a quite moderate segment—something like $-746 < x < 710$. (SUN does not understand numbers larger than 1.0e + 309 and less than 1.0e − 344; MATLAB knows that exp{709.7} = 1.6550e + 308 and exp{−745} = 4.9407e − 324 but believes that exp{709.8} = Inf and exp{−746} = 0). And in this very limited range of values of x the computer exponent, of course, differs from the actual one—the former reproduces the latter with relative accuracy like 10^{-16}. Now, the question:

Exercise 3.3. *How large should l be in order for $f_l(\cdot)$ to be a valid substitution of the exponent in the computer, i.e., to approximate the latter in the segment $-746 < x < 710$ within relative inaccuracy 10^{-16}? What is the length of the CQR for such an f_l—how many additional variables and simple constraints of the type $s^2 \le t$ do you need to get the CQR?*

Note that we can implement our idea in a smarter way. The approximation f_l of the exponent comes from the formula

$$\exp\{x\} = \left(\exp\{2^{-l}x\}\right)^{(2^l)} \approx \left(1+2^{-l}x\right)^{(2^l)},$$

and the quality of this approximation, for given l and the range $-a \le x \le b$ of values of x, depends on how well the exponent is approximated by its linearization in the segment $-2^{-l}a < x < 2^{-l}b$. What will happen when the linearization is replaced with a polynomial approximation of a higher order, i.e., when we use approximations

$$\exp\{x\} \approx g_l(x) = \left(1+2^{-l}x+\frac{1}{2}(2^{-l}x)^2\right)^{(2^l)}$$

or

$$\exp\{x\} \approx h_l(x) = \left(1+2^{-l}x+\frac{1}{2}(2^{-l}x)^2+\frac{1}{6}(2^{-l}x)^3+\frac{1}{24}(2^{-l}x)^4\right)^{(2^l)},$$

and so on? Of course, to make these approximation useful in our context, we should be sure that the approximations are CQr.

Exercise 3.4. 1. *Assume that $s(\cdot)$ is a nonnegative CQr function and you know its CQR* (S). *Prove that for every rational $\alpha \geq 1$ the function $s^\alpha(\cdot)$ is CQr, the corresponding representation being readily given by* (S).

2. *Prove that the polynomial $1+t+t^2/2$ on the axis is nonnegative and is CQr. Find a CQR for the polynomial.*

Conclude that the approximations $g_l(\cdot)$ are CQr. How large is a sufficient l (the one reproducing the exponent with the same quality as in the previous exercise) for these approximations? How many additional variables and constraints are needed?

3. *Answer the same questions as in* 2, *but for the polynomial $1+t+t^2/2+t^3/6+t^4/24$ and the approximations $h_l(\cdot)$.*

The following table shows the best numerical results we were able to obtain with the outlined scheme.

x	$\exp\{x\}$	Rel. error of f_{40}	Rel. error of g_{27}	Rel. error of h_{18}
-512	4.337e$-$223	4.e$-$6	5.e$-$9	5.e$-$11
-256	6.616e$-$112	2.e$-$6	3.e$-$9	8.e$-$12
-128	2.572e$-$56	1.e$-$6	1.e$-$9	1.e$-$11
-64	1.603e$-$28	5.e$-$7	6.e$-$10	9.e$-$12
-32	1.266e$-$14	2.e$-$7	1.e$-$10	1.e$-$11
-16	1.125e$-$07	1.e$-$7	2.e$-$10	1.e$-$11
-1	3.678e$-$01	7.e$-$9	7.e$-$9	1.e$-$11
1	2.718e+00	7.e$-$9	7.e$-$9	1.e$-$11
16	8.886e+06	1.e$-$7	2.e$-$10	2.e$-$11
32	7.896e+13	2.e$-$7	3.e$-$10	2.e$-$11
64	6.235e+27	5.e$-$7	6.e$-$10	2.e$-$11
128	3.888e+55	1.e$-$6	1.e$-$9	2.e$-$11
256	1.511e+111	2.e$-$6	2.e$-$9	3.e$-$11
512	2.284e+222	4.e$-$6	5.e$-$9	7.e$-$11

And now, the most difficult question.

Exercise 3.5. *Why does the outlined scheme not work on a computer, or at least does not work as well as predicted by the previous analysis?*

Stable grasp

Recall the stable grasp analysis problem from section 3.2.2: to check whether the system of constraints

$$\begin{aligned}
\|F^i\|_2 &\leq \mu (f^i)^T v^i, \ i = 1, \ldots, N, \\
(v^i)^T F^i &= 0, \ i = 1, \ldots, N, \\
\sum_{i=1}^N (f^i + F^i) + F^{\text{ext}} &= 0, \\
\sum_{i=1}^N p^i \times (f^i + F^i) + T^{\text{ext}} &= 0
\end{aligned} \tag{SG}$$

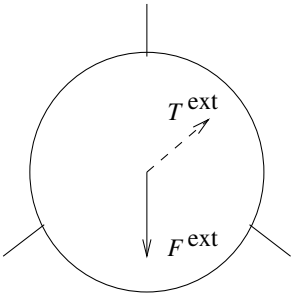


Figure 3.5. *Data of a stable grasp problem.*

in the 3D variables F^i is or is not solvable. Here the data are given by a number of 3D vectors, namely,

- vectors v^i—unit inward normals to the surface of the body at the contact points;
- contact points p^i;
- vectors f^i—contact forces;
- vectors F^{ext} and T^{ext} of the external force and torque, respectively.

Further, $\mu > 0$ is a given friction coefficient, and we assume that $(f^i)^T v^i > 0 \ \forall i$.

Exercise 3.6. 1. *Regarding* (SG) *as the system of constraints of a maximization program with trivial objective and applying the technique from section 2.5, build the dual problem.*

2. *Prove that the dual problem is strictly feasible. Derive from this observation that stable grasp is possible if and only if the dual objective is nonnegative on the dual feasible set.*

3. *Assume that* $\sum_{i=1}^N \mu[(f^i)^T v^i] < \| \sum_{i=1}^N f^i + F^{\text{ext}}\|_2$. *Is a stable grasp possible?*

4. *Let* $T = \sum_{i=1}^N p^i \times f^i + T^{\text{ext}}$, *and let* T^i *be the orthogonal projection of the vector* $p^i \times T$ *onto the plane orthogonal to* v^i. *Assume that*

$$\sum_{i=1}^N \mu[(f^i)^T v^i]\|T^i\|_2 < \|T\|_2^2.$$

Is a stable grasp possible?

5. *The data of the stable grasp problem are shown in Fig. 3.5 (the "fingers" look at the center of the circle; the contact points are the vertices of the inscribed equilateral triangle). Magnitudes of all three contact forces are equal to each other, the friction coefficient is equal to 1, magnitudes of the external force and the external torque are equal to a, and the*

torque is orthogonal to the plane of the picture. What is the smallest magnitude of contact forces that makes a stable grasp possible?

Trusses

We are about to process the multiload TTD problem 3.4.1, which we write as (see (3.4.57))

$$\begin{array}{ll} \text{minimize} & \tau \\ \text{s.t.} & s_{ij}^2 \le 4t_i r_{ij},\ i=1,\dots,n,\ j=1,\dots,k, \\ & \sum_{i=1}^n r_{ij} \le \tfrac{1}{2}\tau,\ j=1,\dots,k, \\ & \sum_{i=1}^n t_i \le w, \\ & \sum_{i=1}^n s_{ij} b_i = f_j,\ j=1,\dots,k, \\ & t_i, r_{ij} \ge 0,\ i=1,\dots,n,\ j=1,\dots,k. \end{array} \qquad \text{(Pr)}$$

Here the design variables are $s_{ij}, r_{ij}, t_i, \tau$; the variables σ_{ij} from (3.4.57) are twice the new variables r_{ij}.

Throughout this section we make the following two assumptions:

- The ground structure $(n, m, b_1, \dots, b_n)$ is such that the matrix $\sum_{i=1}^n b_i b_i^T$ is positive definite.
- The loads of interest $f_1, \dots, f_k$ are nonzero, and the material resource w is positive.

Exercise 3.7. 1. *Applying the technique from section* 2.5, *build the problem* (Dl) *dual to* (Pr).

Check that both (Pr) *and* (Dl) *are strictly feasible. What are the consequences for the solvability status of the problems and their optimal values?*

What is the design dimension of (Pr)*? of* (Dl)*?*

2. *Convert problem* (Dl) *into an equivalent problem of the design dimension* $mk + k + 1$.

Exercise 3.8. *Let us fix a ground structure* $(n, m, b_1, \dots, b_n)$ *and a material resource* w, *and let* $\mathcal{F}$ *be a finite set of loads.*

1. *Assume that* $\mathcal{F}_j \in \mathcal{F}$, $j = 1, \dots, k$, *are subsets of* $\mathcal{F}$ *with* $\bigcup_{j=1}^k \mathcal{F}_j = \mathcal{F}$. *Let* μ_j *be the optimal value in the multiload TTD problem with the set of loads* $\mathcal{F}_j$ *and* μ *be the optimal value in the multiload TTD problem with the set of loads* $\mathcal{F}$. *Is it possible that* $\mu > \sum_{j=1}^k \mu_j$*?*

2. *Assume that the ground structure includes* $n = 1998$ *tentative bars and that you are given a set* $\mathcal{F}$ *of* $N = 1998$ *loads. It is known that for every subset* $\mathcal{F}'$ *of* $\mathcal{F}$ *made up of no more than 999 loads, the optimal value in the multiload TTD problem, the set of loading scenarios being* $\mathcal{F}'$, *does not exceed 1. What can be said about the optimal value in the multiload TTD problem with the set of scenarios* $\mathcal{F}$*?*

Answer a similar question in the case when $\mathcal{F}$ *comprises* $N' = 19980$ *loads.*

3.5.3 Does conic quadratic programming exist?

Of course it does. What is meant is:

(?) *Can a conic quadratic problem be efficiently approximated by an LP one?*

To pose the question formally, let us say that a system of linear inequalities

$$Py + tp + Qu \geq 0 \tag{LP}$$

approximates the CQI

$$\|y\|_2 \leq t \tag{CQI}$$

within accuracy ϵ (or, which is the same, is an ϵ-approximation of (CQI)), if

(i) whenever (y, t) satisfies (CQI), there exists u such that (y, t, u) satisfies (LP);

(ii) whenever (y, t, u) satisfies (LP), (y, t) nearly satisfies (CQI), namely,

$$\|y\|_2 \leq (1 + \epsilon)t.$$

Note that given a conic quadratic program

$$\min_x \left\{c^T x : \|A_i x - b_i\|_2 \leq c_i^T x - d_i,\ i = 1, \ldots, m\right\} \tag{CQP}$$

with $m_i \times n$-matrices A_i and ϵ-approximations

$$P^i y_i + t_i p^i + Q^i u_i \geq 0$$

of CQIs

$$\|y_i\|_2 \leq t_i \quad [\dim y_i = m_i],$$

one can approximate (CQP) by the LP program

$$\min_{x,u} \left\{c^T x : P^i(A_i x - b_i) + (c_i^T x - d_i)p^i + Q^i u_i \geq 0,\ i = 1, \ldots, m\right\};$$

if ϵ is small enough, this program, for every practical purpose, is the same as (CQP).

Now, in principle, any closed convex cone of the form

$$\{(y, t) \mid t \geq \phi(y)\}$$

can be approximated, in the aforementioned sense, by a system of linear inequalities within any accuracy $\epsilon > 0$. The question of crucial importance, however, is how large should the approximating system be—how many linear constraints and additional variables it requires. Surprisingly, for the second-order cone these quantities are not that large:

> **Theorem** (Ben-Tal and Nemirovski, 1998). *Let n be the dimension of y in* (CQI)*, and let $0 < \epsilon < 1/2$. There exists (and can be explicitly written) a system of no more than $O(1)n \ln \frac{1}{\epsilon}$ linear inequalities of the form* (LP) *with $\dim u \leq O(1)n \ln \frac{1}{\epsilon}$ which is an ϵ-approximation of* (CQI)*. Here $O(1)$'s are appropriate absolute constants.*

To get an impression of the constant factors in the theorem, look at the numbers $I(n,\epsilon)$ of linear inequalities and $V(n,\epsilon)$ of additional variables u in an ϵ-approximation (LP) of the conic quadratic inequality (CQI) with $\dim x = n$:

n	$\epsilon = 10^{-1}$		$\epsilon = 10^{-6}$		$\epsilon = 10^{-14}$	
	$I(n,\epsilon)$	$V(N,\epsilon)$	$I(n,\epsilon)$	$V(n,\epsilon)$	$I(n,\epsilon)$	$V(n,\epsilon)$
4	6	17	31	69	70	148
16	30	83	159	345	361	745
64	133	363	677	1458	1520	3153
256	543	1486	2711	5916	6169	12710
1024	2203	6006	10899	23758	24773	51050

You can see that $I(n,\epsilon) \approx 0.7n \ln \frac{1}{\epsilon}$, $V(n,\epsilon) \approx 2n \ln \frac{1}{\epsilon}$.

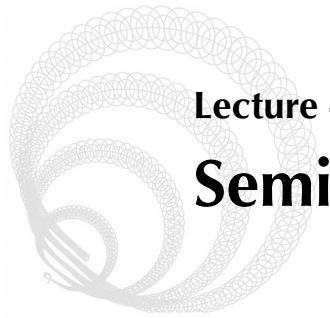
Exercise 3.9. *Prove the theorem. Specifically,*

1. *Build an ϵ-approximation of the* 3D *Lorentz cone, i.e., of the set*

$$\{(x,y,t) \mid t \geq \sqrt{x^2+y^2}\}$$

with $O(1)\ln \frac{1}{\epsilon}$ linear inequalities and additional variables.

2. *Pass from the case of $n = 2$ (i.e., from the* 3D *Lorentz cone) to the case of arbitrary n.*



Lecture 4
Semidefinite Programming

In this lecture we study semidefinite programming (SDP), a generic conic program with a vast area of applications.

4.1 Semidefinite cone and semidefinite programs

4.1.1 Preliminaries

Let $\mathbf{S}^m$ be the space of symmetric $m \times m$ matrices and $\mathbf{M}^{m,n}$ be the space of rectangular $m \times n$ matrices with real entries. From the viewpoint of their linear structure (i.e., the operations of addition and multiplication by reals), $\mathbf{S}^m$ is just the arithmetic linear space $\mathbf{R}^{m(m+1)/2}$ of dimension $\frac{m(m+1)}{2}$: by arranging the elements of a symmetric $m \times m$ matrix X in a single column, say, in row-by-row order, you get a usual m^2-dimensional column vector; multiplication of a matrix by a real and addition of matrices correspond to the same operations with the representing vector(s). When X runs through $\mathbf{S}^m$, the vector representing X runs through $m(m+1)/2$-dimensional subspace of $\mathbf{R}^{m^2}$ consisting of vectors satisfying the symmetry condition—the coordinates coming from symmetric to each other pairs of entries in X are equal to each other. Similarly, $\mathbf{M}^{m,n}$ as a linear space is just $\mathbf{R}^{mn}$, and it is natural to equip $\mathbf{M}^{m,n}$ with the inner product defined as the usual inner product of the vectors representing the matrices:

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

Here Tr stands for the trace—the sum of diagonal elements of a (square) matrix. With this inner product (called the Frobenius inner product), $\mathbf{M}^{m,n}$ becomes a legitimate Euclidean space, and we may use in connection with this space all notions based on the Euclidean structure, e.g., the (Frobenius) norm of a matrix

$$\|X\|_2 = \sqrt{\langle X, X\rangle} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2} = \sqrt{\mathrm{Tr}(X^T X)}$$

and likewise the notions of orthogonality, orthogonal complement of a linear subspace, etc. The same applies to the space $\mathbf{S}^m$ equipped with the Frobenius inner product. Of course, the Frobenius inner product of symmetric matrices can be written without the transposition sign:

$$\langle X, Y\rangle = \mathrm{Tr}(XY), \;\; X, Y \in \mathbf{S}^m.$$

Let us focus on the space $\mathbf{S}^m$. After it is equipped with the Frobenius inner product, we may speak about a cone dual to a given cone $\mathbf{K} \subset \mathbf{S}^m$:

$$\mathbf{K}_* = \{Y \in \mathbf{S}^m \mid \langle Y, X\rangle \geq 0 \quad \forall X \in \mathbf{K}\}.$$

Among the cones in $\mathbf{S}^m$, the one of special interest is the *semidefinite cone* $\mathbf{S}^m_+$, the cone of all symmetric positive semidefinite matrices.[17] It is easily seen (see Exercise 2.8) that $\mathbf{S}^m_+$ indeed is a cone, and moreover it is self-dual:

$$(\mathbf{S}^m_+)_* = \mathbf{S}^m_+.$$

Another simple fact is that the interior $\mathbf{S}^m_{++}$ of the semidefinite cone $\mathbf{S}^m_+$ is exactly the set of all positive definite symmetric $m \times m$ matrices, i.e., symmetric matrices A for which $x^T Ax > 0$ for all nonzero vectors x, or, which is the same, symmetric matrices with positive eigenvalues.

The semidefinite cone gives rise to a family of conic programs "minimize a linear objective over the intersection of the semidefinite cone and an affine plane"; these are the semidefinite programs we are about to study.

Before writing a generic semidefinite program, we should resolve a small difficulty with notation. Normally we use lowercase Latin and Greek letters to denote vectors and the uppercase letters to denote matrices; e.g., our usual notation for a conic problem is

$$\min_x \left\{c^T x : Ax - b \geq_{\mathbf{K}} 0\right\}. \tag{CP}$$

In the case of semidefinite programs, where $\mathbf{K} = \mathbf{S}^m_+$, the usual notation leads to a conflict with the notation related to the space where $\mathbf{S}^m_+$ lives. Look at (CP): without additional remarks it is unclear what A is—is it an $m \times m$ matrix from the space $\mathbf{S}^m$ or is it a linear mapping acting from the space of the design vectors—some $\mathbf{R}^n$—to the space $\mathbf{S}^m$? When speaking about a conic problem on the cone $\mathbf{S}^m_+$, we should have in mind the second interpretation of A, while the standard notation in (CP) suggests the first (wrong!) interpretation. In other words, we meet with the necessity to distinguish between linear mappings acting to or from $\mathbf{S}^m$ and elements of $\mathbf{S}^m$ (which themselves are linear mappings from $\mathbf{R}^m$ to $\mathbf{R}^m$). To resolve this difficulty, we introduce the following notational conventions.

Notational convention. To denote a linear mapping acting from a linear space to a space of matrices (or from a space of matrices to a linear space), we use uppercase script letters like $\mathcal{A}$, $\mathcal{B}$. Elements of usual vector spaces $\mathbf{R}^n$ are, as always, denoted by lowercase Latin

[17] Recall that a symmetric $n \times n$ matrix A is called positive semidefinite if $x^T Ax \geq 0 \; \forall x \in \mathbf{R}^m$. An equivalent definition is that all eigenvalues of A are nonnegative.

and Greek letters $a, b, \ldots, z, \alpha, \ldots, \zeta$, while elements of a space of matrices usually are denoted by uppercase Latin letters $A, B, \ldots, Z$. According to this convention, a semidefinite program of the form (CP) should be written as

$$\min_x \left\{ c^T x : \mathcal{A}x - B \succeq_{\mathbf{S}^m_+} 0 \right\}. \qquad (*)$$

We also simplify the sign $\succeq_{\mathbf{S}^m_+}$ to $\succeq$ and the sign $\succ_{\mathbf{S}^m_+}$ to $\succ$ (same as we write $\geq$ instead of $\geq_{\mathbf{R}^m_+}$ and $>$ instead of $>_{\mathbf{R}^m_+}$). Thus, $A \succeq B$ ($\Leftrightarrow B \preceq A$) means that A and B are symmetric matrices of the same size and $A - B$ is positive semidefinite, while $A \succ B$ ($\Leftrightarrow B \prec A$) means that A, B are symmetric matrices of the same size with positive definite $A - B$.

We further need a special notation for the conjugate (transpose) of a linear mapping $\mathcal{A}$ acting from or to a space of matrices. Recall that the conjugate of a linear mapping $\Xi : E \to F$ acting from a Euclidean space $(E, (\cdot,\cdot)_E)$ to a Euclidean space $(F, (\cdot,\cdot)_F)$ is the mapping $\Xi' : F \to E$ satisfying the identity

$$(\Xi e, f)_F = (e, \Xi' f)_E \quad \forall e \in E, f \in F.$$

When E and F are the usual coordinate spaces $\mathbf{R}^k$ and $\mathbf{R}^l$ equipped with the standard inner product $(x, y) = x^T y$, so that Ξ and Ξ' can be naturally identified with $k \times l$ and $l \times k$ matrices, respectively, these matrices are transposed to each other, and we can write Ξ^T instead of Ξ'. In the case when among the spaces E, F there is a space of matrices, the notation Ξ^T for Ξ' conflicts with the notation for the transpose of an element from E or F. This is why, when speaking about a linear mapping $\mathcal{A}$ acting to or from a space of matrices, we denote its conjugate by $\mathcal{A}^*$.

Our last convention addresses how to write expressions of the type $A\mathcal{A}xB$ ($\mathcal{A}$ is a linear mapping from some $\mathbf{R}^n$ to $\mathbf{S}^m$, $x \in \mathbf{R}^n$ $A, B \in \mathbf{S}^m$). What we are trying to denote is the result of the following operation: we first take the value $\mathcal{A}x$ of the mapping $\mathcal{A}$ at a vector x, thus getting an $m \times m$ matrix $\mathcal{A}x$, and then multiply this matrix from the left and from the right by the matrices A, B. To avoid misunderstandings, we write expressions of this type as

$$A[\mathcal{A}x]B$$

or as $A\mathcal{A}(x)B$ or as $A\mathcal{A}[x]B$.

How to specify a mapping $\mathcal{A} : \mathbf{R}^n \to \mathbf{S}^m$. Natural data specifying a linear mapping $A : \mathbf{R}^n \to \mathbf{R}^m$ consists of a collection of n elements of the destination space—n vectors $a_1, a_2, \ldots, a_n \in \mathbf{R}^m$—such that

$$Ax = \sum_{j=1}^n x_j a_j, \quad x = (x_1, \ldots, x_n)^T \in \mathbf{R}^n.$$

Similarly, a natural data specifying a linear mapping $\mathcal{A} : \mathbf{R}^n \to \mathbf{S}^m$ is a collection $A_1, \ldots, A_n$ of n matrices from $\mathbf{S}^m$ such that

$$\mathcal{A}x = \sum_{j=1}^n x_j A_j, \quad x = (x_1, \ldots, x_n)^T \in \mathbf{R}^n.$$

In terms of these data, the semidefinite program (*) can be written as

$$\min_x \left\{ c^T x : x_1 A_1 + x_2 A_2 + \cdots + x_n A_n - B \succeq 0 \right\}. \tag{SDPr}$$

Linear matrix inequality constraints and semidefinite programs. In the case of conic quadratic problems, we started with the simplest program of this type—the one with a single conic quadratic constraint $Ax - b \geq_{\mathbf{L}^m} 0$—and then defined a conic quadratic program as a program with finitely many constraints of this type, i.e., as a conic program on a direct product of the ice cream cones. In contrast to this, when defining a semidefinite program, we impose on the design vector just one linear matrix inequality (LMI) $\mathcal{A}x - B \succeq 0$. Now we should not bother about more than a single LMI, due to the following simple fact:

A system of finitely many LMIs

$$\mathcal{A}_i x - B_i \succeq 0, \ i = 1, \ldots, k,$$

is equivalent to the single LMI

$$\mathcal{A}x - B \succeq 0$$

with

$$\mathcal{A}x = \mathrm{Diag}\left(\mathcal{A}_1 x, \mathcal{A}_2 x, \ldots, \mathcal{A}_k x\right), \ B = \mathrm{Diag}(B_1, \ldots, B_k);$$

here for a collection of symmetric matrices $Q_1, \ldots, Q_k$

$$\mathrm{Diag}(Q_1, \ldots, Q_k) = \begin{pmatrix} Q_1 & & \\ & \ddots & \\ & & Q_k \end{pmatrix}$$

is the block-diagonal matrix with the diagonal blocks $Q_1, \ldots, Q_k$.

> Indeed, a block-diagonal symmetric matrix is positive (semi-)definite if and only if all its diagonal blocks are so.

Dual to a semidefinite program. As we know, the dual to conic problem (CP) is the problem

$$\max_\lambda \left\{ b^T \lambda : A^T \lambda = c, \ \lambda \geq_{\mathbf{K}^*} 0 \right\};$$

the matrix A^T defines the linear mapping conjugate to the mapping A from the primal problem. When writing the problem dual to (SDPr), we should

– replace $b^T \lambda$ by the Frobenius inner product,

– follow our notational convention and write $\mathcal{A}^*$ instead of $\mathcal{A}^T$, and

– take into account that the semidefinite cone is self-dual.

Consequently, the problem dual to (SDPr) is the semidefinite program

$$\max_\Lambda \left\{ \mathrm{Tr}(B\Lambda) : \mathcal{A}^* \Lambda = c, \ \Lambda \succeq 0 \right\}.$$

Now, let $A_1, \ldots, A_n$ be the data specifying $\mathcal{A}$; how does $\mathcal{A}^*$ act? The answer is immediate:

$$\mathcal{A}^*\Lambda = (\mathrm{Tr}(A_1\Lambda), \mathrm{Tr}(A_2\Lambda), \ldots, \mathrm{Tr}(A_n\Lambda))^T.$$

Indeed, we should verify that with the above definition we have

$$[\mathrm{Tr}([\mathcal{A}x]\Lambda) =] \quad \langle \mathcal{A}x, \Lambda\rangle = (\mathcal{A}^*\Lambda)^T x \quad \forall \Lambda \in \mathbf{S}^m, x \in \mathbf{R}^n,$$

which is immediate:

$$\begin{aligned}
\mathrm{Tr}([\mathcal{A}x]\Lambda) &= \mathrm{Tr}\left(\left(\sum_{j=1}^n x_j A_j\right)\Lambda\right) \\
&= \sum_{i=1}^n x_j \mathrm{Tr}(A_j\Lambda) \\
&= (\mathrm{Tr}(A_1\Lambda), \ldots, \mathrm{Tr}(A_n\Lambda)) \begin{pmatrix} x_1 \\ \ldots \\ x_n \end{pmatrix}.
\end{aligned}$$

We see that the explicit—given in terms of the original data—form of the problem dual to (SDPr) is the semidefinite program

$$\max_{\Lambda} \left\{\mathrm{Tr}(B\Lambda) : \mathrm{Tr}(A_j\Lambda) = c_j, \ \ j = 1, \ldots, n; \ \Lambda \succeq 0\right\}. \tag{SDDl}$$

Conic duality in the case of semidefinite programming. Let us see what we get from the conic duality theorem in the case of semidefinite programs. First note that our default Assumption A on a conic program in the form of (CP) (Lecture 2) as applied to (SDPr) says that no nontrivial linear combination of the matrices $A_1, \ldots, A_n$ is 0. Strict feasibility of (SDPr) means that there exists x such that $\mathcal{A}x - B$ is positive definite, and strict feasibility of (SDDl) means that there exists a positive definite Λ satisfying $\mathcal{A}^*\Lambda = c$. According to the conic duality theorem, if both primal and dual are strictly feasible, both are solvable, the optimal values are equal to each other, and the complementary slackness condition

$$[\mathrm{Tr}(\Lambda[\mathcal{A}x - B]) \equiv] \qquad \langle \Lambda, \mathcal{A}x - B\rangle = 0$$

is necessary and sufficient for a pair of a primal feasible solution x and a dual feasible solution Λ to be optimal for the corresponding problems.

It is easily seen (see Exercise 2.8) that for a pair X, Y of positive semidefinite symmetric matrices one has

$$\mathrm{Tr}(XY) = 0 \Leftrightarrow XY = YX = 0;$$

in particular, in the case of strictly feasible primal and dual problems, the primal slack $S_* = \mathcal{A}x^* - B$ corresponding to a primal optimal solution commutes with (any) dual optimal solution Λ_*, and the product of these two matrices is 0. Besides this, S_* and Λ_*, as a pair of commuting symmetric matrices, share a common eigenbasis, and the fact that $S_*\Lambda_* = 0$ means that the eigenvalues of the matrices in this basis are complementary: for every common eigenvector, either the eigenvalue of S_* or the one of Λ_*, or both, are equal to 0 (cf. complementary slackness in the LP case).

4.2 What can be expressed via linear matrix inequalities?

As in the previous lecture, the first thing to realize when speaking about the SDP universe is how to recognize that a convex optimization program

$$\min_x \left\{ c^T x : x \in X = \bigcap_{i=1}^m X_i \right\} \tag{P}$$

can be cast as a semidefinite program. Just as in the previous lecture, this question actually asks whether a given convex set or function is semidefinite representable (SDr). The definition of the latter notion is completely similar to the one of a CQr set or function:

We say that a convex set $X \subset \mathbf{R}^n$ is SDr if there exists an affine mapping $(x,u) \to \mathcal{A}\left(\begin{smallmatrix} x \\ u \end{smallmatrix}\right) - B : \mathbf{R}^n_x \times \mathbf{R}^k_u \to \mathbf{S}^m$ such that

$$x \in X \Leftrightarrow \exists u : \mathcal{A}\begin{pmatrix} x \\ u \end{pmatrix} - B \succeq 0;$$

in other words, X is SDr if there exists LMI

$$\mathcal{A}\begin{pmatrix} x \\ u \end{pmatrix} - B \succeq 0$$

in the original design vector x and a vector u of additional design variables such that X is a projection of the solution set of the LMI onto the x-space. An LMI with this property is called semidefinite representation (SDR) *of the set X.*

A convex function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is called SDr if its epigraph

$$\{(x,t) \mid t \geq f(x)\}$$

is an SDr set. An SDR of the epigraph of f is called an SDR of f.

By exactly the same reasons as in the case of conic quadratic problems, one has the following:

1. *If f is an SDr function, then all its level sets $\{x \mid f(x) \leq a\}$ are SDr; the SDR of the level sets are explicitly given by (any) SDR of f.*

2. *If all the sets X_i in problem* (P) *are SDr with known SDRs, then the problem can explicitly be converted to a semidefinite program.*

To understand which functions or sets are SDr, we may use the same approach as in Lecture 3. The calculus, i.e., the list of basic operations preserving SD-representability, is exactly the same as in the case of conic quadratic problems; we just may repeat word by word the relevant reasoning from Lecture 3, each time replacing CQr with SDr. Thus, the only issue to be addressed is the derivation of a catalogue of simple SDr functions or sets. Our first observation in this direction is as follows.

1–17. *We refer to Examples* 1–17 *of CQr functions and sets in section* 3.3.

If a function or set is CQr, it is also SDr, and any CQR of the function or set can be explicitly converted to its SDR.

Indeed, the notion of a CQr or an SDr function is a derivative of the notion of a CQr or an SDr set: by definition, a function is CQr or SDr if and only if its epigraph is so. Now, CQr sets are exactly those sets that can be obtained as projections of the solution sets of systems of conic quadratic inequalities, i.e., as projections of inverse images, under affine mappings, of direct products of ice cream cones. Similarly, SDr sets are projections of the inverse images, under affine mappings, of positive semidefinite cones. Consequently,

(i) To verify that a CQr set is SDr as well, it suffices to show that an inverse image, under an affine mapping, of a direct product of ice cream cones, a set of the form

$$Z = \left\{ z \mid Az - b \in \mathbf{K} = \prod_{i=1}^{l} \mathbf{L}^{k_i} \right\},$$

is the inverse image of a semidefinite cone under an affine mapping. To this end, in turn, it suffices to demonstrate that

(ii) A direct product $\mathbf{K} = \prod_{i=1}^{l} \mathbf{L}^{k_i}$ of ice cream cones is an inverse image of a semidefinite cone under an affine mapping.

Indeed, representing $\mathbf{K}$ as $\{y \mid \mathcal{A}y - b \in \mathbf{S}^m_+\}$, we get

$$Z = \{z \mid Az - b \in \mathbf{K}\} = \{z \mid \hat{\mathcal{A}}z - \hat{B} \in \mathbf{S}^m_+\},$$

where $\hat{\mathcal{A}}z - \hat{B} = \mathcal{A}(Az - b) - B$ is affine.

In turn, to prove (ii) it suffices to show that

(iii) Every ice cream cone $\mathbf{L}^k$ is an inverse image of a semidefinite cone under an affine mapping.

In fact, the implication (iii) $\Rightarrow$ (ii) is given by our calculus, since a direct product of SDr sets is again SDr.[18]

We have reached the point where no more reductions are necessary, and here is the demonstration of (iii). The case of $k = 1$ is trivial: the one-dimensional (1D) ice cream cone is exactly the same as the 1D semidefinite cone—both are the same nonnegative ray on the axis! In the case of $k > 1$ it suffices to observe that

$$\begin{pmatrix} x \\ t \end{pmatrix} \in \mathbf{L}^k \Leftrightarrow \mathcal{A}(x,t) = \begin{pmatrix} tI_{k-1} & x \\ x^T & t \end{pmatrix} \succeq 0 \tag{4.2.1}$$

(x is $(k-1)$-dimensional, t is scalar, I_{k-1} is the $(k-1)\times(k-1)$ unit matrix). Equation (4.2.1) indeed resolves the problem, since the matrix $\mathcal{A}(x,t)$ is linear in (x,t)! It remains to verify (4.2.1), which is immediate. If $(x,t) \in \mathbf{L}^k$, i.e., if $\|x\|_2 \le t$, then for every $y = \begin{pmatrix} \xi \\ \tau \end{pmatrix} \in \mathbf{R}^k$ (ξ is $(k-1)$-dimensional, τ is scalar) we have

$$\begin{aligned} y^T\mathcal{A}(x,t)y &= \tau^2 t + 2\tau x^T\xi + t\xi^T\xi \ge \tau^2 t - 2|\tau|\|x\|_2\|\xi\|_2 + t\|\xi\|_2^2 \\ &\ge t\tau^2 - 2t|\tau|\|\xi\|_2 + t\|\xi\|_2^2 \\ &\ge t(|\tau| - \|\xi\|_2)^2 \ge 0, \end{aligned}$$

[18] Just to recall where the calculus comes from, here is a direct verification. Given a direct product $\mathbf{K} = \prod_{i=1}^{l} \mathbf{L}^{k_i}$ of ice cream cones and given that every factor in the product is the inverse image of a semidefinite cone under an affine mapping,

$$\mathbf{L}^{k_i} = \{x_i \in \mathbf{R}^{k_i} \mid \mathcal{A}_i x_i - B_i \succeq 0\},$$

we can represent $\mathbf{K}$ as the inverse image of a semidefinite cone under an affine mapping, namely, as

$$\mathbf{K} = \{x = (x_1, \ldots, x_l) \in \mathbf{R}^{k_1} \times \cdots \times \mathbf{R}^{k_l} \mid \mathrm{Diag}(\mathcal{A}_1 x_i - B_1, \ldots, \mathcal{A}_l x_l - B_l) \succeq 0\}.$$

so that $\mathcal{A}(x,t) \succeq 0$. Vice versa, if $\mathcal{A}(t,x) \succeq 0$, then of course $t \geq 0$. Assuming $t = 0$, we immediately obtain $x = 0$ (since otherwise for $y = \binom{x}{0}$ we would have $0 \leq y^T \mathcal{A}(x,t)y = -2\|x\|_2^2$). Thus, $A(x,t) \succeq 0$ implies $\|x\|_2 \leq t$ in the case of $t = 0$. To see that the same implication is valid for $t > 0$, let us set $y = \binom{-x}{t}$ to get

$$0 \leq y^T \mathcal{A}(x,t)y = tx^T x - 2tx^T x + t^3 = t(t^2 - x^T x),$$

whence $\|x\|_2 \leq t$, as claimed. □

We see that the expressive abilities of SDP are even richer than those of CQP. In fact the gap is quite significant. The first new possibility is the ability to handle eigenvalues, and the importance of this possibility can hardly be overestimated.

Semidefinite-representability of functions of eigenvalues of symmetric matrices. Our first eigenvalue-related observation is as follows:

18. The largest eigenvalue $\lambda_{\max}(X)$ regarded as a function of $m \times m$ symmetric matrix X is SDr. Indeed, the epigraph of this function

$$\{(X,t) \in \mathbf{S}^m \times \mathbf{R} \mid \lambda_{\max}(X) \leq t\}$$

is given by the LMI

$$tI_m - X \succeq 0,$$

where I_m is the unit $m \times m$ matrix.

Indeed, the eigenvalues of $tI_m - X$ are t minus the eigenvalues of X, so that the matrix $tI_m - X$ is positive semidefinite—all its eigenvalues are nonnegative—if and only if t majorates all eigenvalues of X.

The latter example admits a natural generalization. Let M, A be two symmetric $m \times m$ matrices, and let M be positive definite. A real λ and a nonzero vector e are called eigenvalue and eigenvector of the *pencil* $[M, A]$ if $Ae = \lambda Me$. (In particular, the usual eigenvalues and eigenvectors of A are exactly the eigenvalues and eigenvectors of the pencil $[I_m, A]$.) Clearly, λ is an eigenvalue of $[M, A]$ if and only if the matrix $\lambda M - A$ is singular, and nonzero vectors from the kernel of the latter matrix are exactly the eigenvectors of $[M, A]$ associated with the eigenvalue λ. The eigenvalues of the pencil $[M, A]$ are the usual eigenvalues of the matrix $M^{-1/2}AM^{-1/2}$, as can be concluded from

$$\begin{aligned}\mathrm{Det}(\lambda M - A) = 0 &\Leftrightarrow \mathrm{Det}(M^{1/2}(\lambda I_m - M^{-1/2}AM^{-1/2})M^{1/2}) \\ &= 0 \Leftrightarrow \mathrm{Det}(\lambda I_m - M^{-1/2}AM^{-1/2}) = 0.\end{aligned}$$

The announced extension of example 18 is as follows.

18.a. *The maximum eigenvalue of a pencil.* Let M be a positive definite symmetric $m \times m$ matrix, and let $\lambda_{\max}(X : M)$ be the largest eigenvalue of the pencil $[M, X]$, where X is a symmetric $m \times m$ matrix. The inequality

$$\lambda_{\max}(X : M) \leq t$$

is equivalent to the matrix inequality

$$tM - X \succeq 0.$$

In particular, $\lambda_{\max}(X : M)$, regarded as a function of X, is SDr.

18.b. The spectral norm $|X|$ of a symmetric $m \times m$ matrix X, i.e., the maximum of absolute values of the eigenvalues of X, is SDr. Indeed, an SDR of the epigraph

$$\{(X,t) \mid |X| \le t\} = \{(X,t) \mid \lambda_{\max}(X) \le t, \lambda_{\max}(-X) \le t\}$$

of $|X|$ is given by the pair of LMIs

$$tI_m - X \succeq 0, \ tI_m + X \succeq 0.$$

Despite their simplicity, the indicated results are extremely useful. As a more complicated example, let us build an SDr for the sum of the *k-largest* eigenvalues of a symmetric matrix.

From now on, speaking about $m \times m$ symmetric matrix X, we denote by $\lambda_i(X)$, $i = 1, \ldots, m$, its eigenvalues counted with their multiplicities and arranged in a nonascending order:

$$\lambda_1(X) \ge \lambda_2(X) \ge \cdots \ge \lambda_m(X).$$

The vector of the eigenvalues (in the indicated order) will be denoted $\lambda(X)$:

$$\lambda(X) = (\lambda_1(X), \ldots, \lambda_m(X))^T \in \mathbf{R}^m.$$

The question we are about to address is which functions of the eigenvalues are SDr. We already know that this is the case for the largest eigenvalue $\lambda_1(X)$. Other eigenvalues cannot be SDr since they are not convex functions of X. And convexity, of course, is a necessary condition for SD-representability (cf. Lecture 3). It turns out, however, that the m functions

$$S_k(X) = \sum_{i=1}^{k} \lambda_i(X), \ k = 1, \ldots, m,$$

are convex and, moreover, are SDr:

18.c. *Sums of largest eigenvalues of a symmetric matrix.* Let X be $m \times m$ symmetric matrix, and let $k \le m$. Then the function $S_k(X)$ is SDr. Specifically, the epigraph

$$\{(X,t) \mid S_k(x) \le t\}$$

of the function admits the SDR

$$\begin{array}{lrcl} \text{(a)} & t - ks - \mathrm{Tr}(Z) & \succeq & 0, \\ \text{(b)} & Z & \succeq & 0, \\ \text{(c)} & Z - X + sI_m & \succeq & 0, \end{array} \tag{4.2.2}$$

where $Z \in \mathbf{S}^m$ and $s \in \mathbf{R}$ are additional variables.

We should prove that

(i) If a given pair X, t can be extended, by properly chosen s, Z, to a solution of the system of LMIs (4.2.2), then $S_k(X) \le t$.

(ii) Conversely, if $S_k(X) \le t$, then the pair X, t can be extended, by properly chosen s, Z, to a solution of (4.2.2).

To prove (i), we use the following basic fact (see Exercise 4.5(i)):

(A) *The vector $\lambda(X)$ is a $\succeq$-monotone function of $X \in \mathbf{S}^m$:*

$$X \succeq X' \Rightarrow \lambda(X) \ge \lambda(X').$$

Assuming that (X, t, s, Z) is a solution to (4.2.2), we get $X \preceq Z + sI_m$, so that

$$\lambda(X) \le \lambda(Z + sI_m) = \lambda(Z) + s\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix},$$

whence

$$S_k(X) \le S_k(Z) + sk.$$

Since $Z \succeq 0$ (see (4.2.2)(b)), we have $S_k(Z) \le \mathrm{Tr}(Z)$, and combining these inequalities we get

$$S_k(X) \le \mathrm{Tr}(Z) + sk.$$

The latter inequality, in view of (4.2.2)(a)), implies $S_k(X) \le t$, and (i) is proved.

To prove (ii), assume that we are given X, t with $S_k(X) \le t$, and let us set $s = \lambda_k(X)$. Then the k-largest eigenvalues of the matrix $X - sI_m$ are nonnegative, and the remaining are nonpositive. Let Z be a symmetric matrix with the same eigenbasis as X and such that the k-largest eigenvalues of Z are the same as those of $X - sI_m$ and the remaining eigenvalues are zeros. The matrices Z and $Z - X + sI_m$ are clearly positive semidefinite (the first by construction and the second since in the eigenbasis of X this matrix is diagonal with the first k diagonal entries being 0 and the remaining being the same as those of the matrix $sI_m - X$, i.e., nonnegative). Thus, the matrix Z and the real s we have built satisfy (4.2.2)(b) and (c). To see that (4.2.2)(a) is satisfied as well, note that by construction $\mathrm{Tr}(Z) = S_k(X) - sk$, whence $t - sk - \mathrm{Tr}(Z) = t - S_k(X) \ge 0$.

To proceed, we need the following highly useful technical result.

Lemma 4.2.1. Lemma on the Schur complement. *Let*

$$A = \begin{pmatrix} B & C^T \\ C & D \end{pmatrix}$$

be a symmetric matrix with $k \times k$ block B and $l \times l$ block D. Assume that B is positive definite. Then A is positive (semi-) definite if and only if the matrix

$$D - CB^{-1}C^T$$

is positive (semi-) definite. (This matrix is called the Schur complement of B in A.)

Proof. The positive semidefiniteness of A is equivalent to the fact that

$$0 \le (x^T, y^T)\begin{pmatrix} B & C^T \\ C & D \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = x^T Bx + 2x^T C^T y + y^T Dy \quad \forall x \in \mathbf{R}^k, y \in \mathbf{R}^l,$$

or, which is the same, to the fact that

$$\inf_{x \in \mathbf{R}^k} \left[x^T Bx + 2x^T C^T y + y^T Dy\right] \ge 0 \quad \forall y \in \mathbf{R}^l.$$

Since B is positive definite by assumption, the infimum in x can be computed explicitly for every fixed y: the optimal x is $-B^{-1}C^T y$, and the optimal value is

$$y^T Dy - y^T CB^{-1}C^T y = y^T[D - CB^{-1}C^T]y.$$

The positive definiteness or semidefiniteness of A is equivalent to the fact that the latter expression is, respectively, positive or nonnegative for every $y \neq 0$, i.e., to the positive definiteness or semidefiniteness of the Schur complement of B in A. □

18.d. *Determinant of a symmetric positive semidefinite matrix.* Let X be a symmetric positive semidefinite $m \times m$ matrix. Although its determinant

$$\mathrm{Det}(X) = \prod_{i=1}^{m} \lambda_i(X)$$

is neither a convex nor a concave function of X (if $m \ge 2$), it turns out that the function $\mathrm{Det}^q(X)$ is concave in X whenever $0 \le q \le \frac{1}{m}$. Functions of this type are important in many volume-related problems (see below). We are about to prove that

if q is a rational number,, $0 \le q \le \frac{1}{m}$, then the function

$$f_q(X) = \begin{cases} -\mathrm{Det}^q(X), & X \succeq 0, \\ +\infty & \textit{otherwise} \end{cases}$$

is SDr.

Consider the LMI

$$\begin{pmatrix} X & \Delta \\ \Delta^T & D(\Delta) \end{pmatrix} \succeq 0, \tag{D}$$

where Δ is $m \times m$ lower triangular matrix comprised of additional variables and $D(\Delta)$ is the diagonal matrix with the same diagonal entries as those of Δ. Let $\mathrm{Dg}(\Delta)$ denote the vector of the diagonal entries of the square matrix Δ.

As we know from Lecture 3 (see example 15), the set

$$\{(\delta, t) \in \mathbf{R}^m_+ \times \mathbf{R} \mid t \le (\delta_1 \dots \delta_m)^q\}$$

admits an explicit CQR. Consequently, this set admits an explicit SDR as well. The latter SDR is given by certain LMI $S(\delta, t; u) \succeq 0$, where u is the vector of additional variables of the SDR and $S(\delta, t, u)$ is a matrix affinely depending on the arguments. We claim that

(!) *The system of LMIs* (D) & $S(\mathrm{Dg}(\Delta), t; u) \succeq 0$ *is an SDR for the set*

$$\{(X, t) \mid X \succeq 0, t \le \mathrm{Det}^q(X)\},$$

which is basically the epigraph of the function f_q (the latter is obtained from our set by reflection with respect to the plane $t = 0$).

To support our claim, recall that by linear algebra a matrix X is positive semidefinite if and only if it can be factorized as $X = \widehat{\Delta}\widehat{\Delta}^T$ with a lower triangular $\widehat{\Delta}$, $\mathrm{Dg}(\widehat{\Delta}) \ge 0$; the resulting matrix $\widehat{\Delta}$ is called the Choleski factor of X. Now note that if $X \succeq 0$ and $t \le \mathrm{Det}^q(X)$, then

1. We can extend X by appropriately chosen lower triangular matrix Δ to a solution of (D) in such a way that if $\delta = \mathrm{Dg}(\Delta)$, then $\prod_{i=1}^m \delta_i = \mathrm{Det}(X)$.

Indeed, let $\widehat{\Delta}$ be the Choleski factor of X. Let $\widehat{D}$ be the diagonal matrix with the same diagonal entries as those of $\widehat{\Delta}$, and let $\Delta = \widehat{\Delta}\widehat{D}$, so that the diagonal entries δ_i of Δ are squares of the diagonal entries $\widehat{\delta}_i$ of the matrix $\widehat{\Delta}$: $D(\Delta) = \widehat{D}^2$. It follows that for every $\epsilon > 0$ one has $\Delta[D(\Delta) + \epsilon I]^{-1}\Delta^T = \widehat{\Delta}\widehat{D}[\widehat{D}^2 + \epsilon I]^{-1}\widehat{D}\widehat{\Delta}^T \preceq \widehat{\Delta}\widehat{\Delta}^T = X$. We see that by the Schur complement lemma all matrices of the form $\left(\begin{smallmatrix} X & \Delta \\ \Delta^T & D(\Delta)+\epsilon I \end{smallmatrix}\right)$ with $\epsilon > 0$ are positive semidefinite, whence $\left(\begin{smallmatrix} X & \Delta \\ \Delta^T & D(\Delta) \end{smallmatrix}\right) \succeq 0$. Thus, (D) is indeed satisfied by (X, Δ). And of course $X = \widehat{\Delta}\widehat{\Delta}^T \Rightarrow \mathrm{Det}(X) = \mathrm{Det}^2(\widehat{\Delta}) = \prod_{i=1}^m \widehat{\delta}_i^2 = \prod_{i=1}^m \delta_i$.

2. Since $\delta = \mathrm{Dg}(\Delta) \ge 0$ and $\prod_{i=1}^m \delta_i = \mathrm{Det}(X)$, we get $t \le \mathrm{Det}^q(X) = (\prod_{i=1}^m \delta_i)^q$, so that we can extend (t, δ) by a properly chosen u to a solution of the LMI $S(\mathrm{Dg}(\Delta), t; u) \succeq 0$.

We conclude that if $X \succeq 0$ and $t \le \mathrm{Det}^q(X)$, then one can extend the pair X, t by properly chosen Δ and u to a solution of the system of LMIs (D) & $S(\mathrm{Dg}(\Delta), t; u) \succeq 0$, which is the first part of the proof of (!).

To complete the proof of (!), it suffices to demonstrate that if for a given pair X, t there exist Δ and u such that (D) and the LMI $S(\mathrm{Dg}(\Delta), t; u) \succeq 0$ are satisfied, then X is positive semidefinite and $t \le \mathrm{Det}^q(X)$. This is immediate: denoting $\delta = \mathrm{Dg}(\Delta)$ [≥ 0] and applying the Schur complement lemma, we conclude that $X \succeq \Delta[D(\Delta) + \epsilon I]^{-1}\Delta^T$ for every $\epsilon > 0$. Applying **(A)**, we get $\lambda(X) \ge \lambda(\Delta[D(\Delta) + \epsilon I]^{-1}\Delta^T)$, whence of course $\mathrm{Det}(X) \ge \mathrm{Det}(\Delta[D(\Delta) + \epsilon I]^{-1}\Delta^T) = \prod_{i=1}^m \delta_i^2/(\delta_i + \epsilon)$. Passing to limit as $\epsilon \to 0$, we get $\prod_{i=1}^m \delta_i \le \mathrm{Det}(X)$. On the other hand, the LMI $S(\delta, t; u) \succeq 0$ takes place, which means that $t \le (\prod_{i=1}^m \delta_i)^q$. Combining the resulting inequalities, we come to $t \le \mathrm{Det}^q(X)$, as required. □

18.e. *Negative powers of the determinant*. Let q be a positive rational. Then the function

$$f(X) = \begin{cases} \mathrm{Det}^{-q}(X), & X \succ 0, \\ +\infty & \text{otherwise} \end{cases}$$

of symmetric $m \times m$ matrix X is SDr.

The construction is completely similar to the one used in example 18d. As we remember from Lecture 3, example 16, the function $g(\delta) = (\delta_1 \dots \delta_m)^{-q}$ of positive vector $\delta = (\delta_1, \dots, \delta_m)^T$ is CQr and is therefore SDr as well. Let an SDR of the function be given by LMI $\mathcal{R}(\delta, t, u) \succeq 0$. The same arguments as in example 18d demonstrate that the pair of LMIs (D) & $\mathcal{R}(\mathrm{Dg}(\Delta), t, u) \succeq 0$ is an SDR for f.

In examples 18, 18.b–18.e we discussed SD-representability of particular functions of eigenvalues of a symmetric matrix. Here is a general statement of this type.

Proposition 4.2.1. *Let $g(x_1, \dots, x_m) : \mathbf{R}^m \to \mathbf{R} \cup \{+\infty\}$ be a symmetric (i.e., invariant with respect to permutations of the coordinates $x_1, \dots, x_m$) SDr function:*

$$t \geq g(x) \Leftrightarrow \exists u : \mathcal{S}(x, t, u) \succeq 0,$$

with $\mathcal{S}$ affinely depending on x, t, u. Then the function

$$f(X) = g(\lambda(X))$$

of symmetric $m \times m$ matrix X is SDr with SDR given by the relation

$$\begin{array}{ll} \text{(a)} & t \geq f(X) \\ & \Updownarrow \\ & \exists x_1, \dots, x_m, u : \\ \text{(b)} & \left\{ \begin{array}{c} \mathcal{S}(x_1, \dots, x_m, t, u) \succeq 0, \\ x_1 \geq x_2 \geq \cdots \geq x_m, \\ S_j(X) \leq x_1 + \cdots + x_j, \ \ j = 1, \dots, m-1, \\ \mathrm{Tr}(X) = x_1 + \cdots + x_m \end{array} \right. \end{array} \tag{4.2.3}$$

(recall that the functions $S_j(X) = \sum_{i=1}^k \lambda_i(X)$ are SDr; see example 18.c*). Thus, the solution set of* (b) *is SDr (as an intersection of SDr sets), which implies SD-representability of the projection of this set onto the (X, t)-plane. By* (4.2.3) *the latter projection is exactly the epigraph of f).*

The proof of Proposition 4.2.1 is based on an extremely useful result known as Birkhoff's theorem; see the exercises to Lecture 4.

As a corollary of Proposition 4.2.1, we see that the following functions of a symmetric $m \times m$ matrix X are SDr:

• $f(X) = -\mathrm{Det}^q(X)$, $X \succeq 0$, $q \leq \frac{1}{m}$, is a positive rational; this fact was already established directly. (Here $g(x_1, \dots, x_m) = -(x_1 \dots x_m)^q : \mathbf{R}^n_+ \to \mathbf{R}$; a CQR (and thus an SDR) of g is presented in example 15 of Lecture 3.)

• $f(x) = \mathrm{Det}^{-q}(X)$, $X \succ 0$, q is a positive rational (cf. example 18.e). (Here $g(x_1, \dots, x_m) = (x_1, \dots, x_m)^{-q} : \mathbf{R}^m_{++} \to \mathbf{R}$; a CQR of g is presented in example 16 of Lecture 3.)

• $\|X\|_p = (\sum_{i=1}^m |\lambda_i(X)|^p)^{1/p}$, $p \geq 1$ is rational. ($g(x) = \|x\|_p \equiv (\sum_{i=1}^m |x_i|^p)^{1/p}$; see Lecture 3, example 17.a.)

• $\|X_+\|_p = (\sum_{i=1}^m \max^p[\lambda_i(X), 0])^{1/p}$, $p \geq 1$ is rational. (Here $g(x) = \|x_+\|_p \equiv (\sum_{i=1}^m \max^p[x_i, 0])^{1/p}$; see Lecture 3, example 17.b.)

Semidefinite representability of functions of singular values. Consider the space $\mathbf{M}^{k,l}$ of $k \times l$ rectangular matrices and assume that $k \leq l$. Given a matrix $A \in \mathbf{M}^{k,l}$, consider the symmetric positive semidefinite $k \times k$ matrix $(AA^T)^{1/2}$; its eigenvalues are called *singular values* of A and are denoted by $\sigma_1(A), \dots \sigma_k(A)$: $\sigma_i(A) = \lambda_i((AA^T)^{1/2})$. According to the

convention on how we enumerate eigenvalues of a symmetric matrix, the singular values form a nonascending sequence:

$$\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_k(A).$$

The importance of the singular values comes from the singular value decomposition theorem, which states that a $k \times l$ matrix A ($k \leq l$) can be represented as

$$A = \sum_{i=1}^{k} \sigma_i(A) e_i f_i^T,$$

where $\{e_i\}_{i=1}^k$ and $\{f_i\}_{i=1}^k$ are orthonormal sequences in $\mathbf{R}^k$ and $\mathbf{R}^l$, respectively. This is a surrogate of the eigenvalue decomposition of a symmetric $k \times k$ matrix

$$A = \sum_{i=1}^{k} \lambda_i(A) e_i e_i^T,$$

where $\{e_i\}_{i=1}^k$ form an orthonormal eigenbasis of A.

Among the singular values of a rectangular matrix, the most important is the largest $\sigma_1(A)$. This is nothing but the *operator* (or *spectral*) *norm* of A:

$$|A| = \max\{\|Ax\|_2 \mid \|x\|_2 \leq 1\}.$$

For a symmetric matrix, the singular values are exactly the modulae of the eigenvalues, and our new definition of the norm coincides with the one given in example 18.b.

It turns out that the sum of a given number of the largest singular values of A

$$\Sigma_p(A) = \sum_{i=1}^{p} \sigma_i(A)$$

is a convex and, moreover, an SDr function of A. In particular, the operator norm of A is SDr:

19. The sum $\Sigma_p(X)$ of p largest singular values of a rectangular matrix $X \in \mathbf{M}^{k,l}$ is SDr. In particular, the operator norm of a rectangular matrix is SDr:

$$|X| \leq t \Leftrightarrow \begin{pmatrix} tI_l & -X^T \\ -X & tI_k \end{pmatrix} \succeq 0.$$

Indeed, the result in question follows from the fact that the sums of p largest eigenvalues of a symmetric matrix are SDr (example 18.c) due to the following observation:

> *The singular values $\sigma_i(X)$ of a rectangular $k \times l$ matrix X ($k \leq l$) for $i \leq k$ are equal to the eigenvalues $\lambda_i(\bar{X})$ of the $(k+l) \times (k+l)$ symmetric matrix*
>
> $$\bar{X} = \begin{pmatrix} 0 & X^T \\ X & 0 \end{pmatrix}.$$

Since $\bar{X}$ linearly depends on X, SDRs of the functions $S_p(\cdot)$ induce SDRs of the functions $\Sigma_p(X) = S_p(\bar{X})$. (This is the rule on affine substitution, Lecture 3. Recall that all calculus rules established in Lecture 3 for CQRs are valid for SDRs as well.)

Let us justify our observation. Let $X = \sum_{i=1}^{k} \sigma_i(X) e_i f_i^T$ be the singular value decomposition of X. We claim that the $2k$ $(k+l)$-dimensional vectors $g_i^+ = \binom{f_i}{e_i}$ and $g_i^- = \binom{f_i}{-e_i}$ are orthogonal to each other, and they are eigenvectors of $\bar{X}$ with the eigenvalues $\sigma_i(X)$ and $-\sigma_i(X)$, respectively. Moreover, $\bar{X}$ vanishes on the orthogonal complement of the linear span of these vectors. In other words, we claim that the eigenvalues of $\bar{X}$, arranged in the nonascending order, are as follows:

$$\sigma_1(X), \sigma_2(X), \ldots, \sigma_k(X), \underbrace{0, \ldots, 0}_{2(l-k)}, -\sigma_k(X), -\sigma_{k-1}(X), \ldots, -\sigma_1(X);$$

this, of course, proves our observation.

Now, the fact that the $2k$ vectors $g_i^{\pm}$, $i = 1, \ldots, k$, are mutually orthogonal and nonzero is evident. Furthermore (we write σ_i instead of $\sigma_i(X)$),

$$\begin{aligned}
\begin{pmatrix} 0 & X^T \\ X & 0 \end{pmatrix} \begin{pmatrix} f_i \\ e_i \end{pmatrix} &= \begin{pmatrix} 0 & \sum_{j=1}^{k} \sigma_j f_j e_j^T \\ \sum_{j=1}^{k} \sigma_j e_j f_j^T & 0 \end{pmatrix} \begin{pmatrix} f_i \\ e_i \end{pmatrix} \\
&= \begin{pmatrix} \sum_{j=1}^{k} \sigma_j f_j (e_j^T e_i) \\ \sum_{j=1}^{k} \sigma_j e_j (f_j^T f_i) \end{pmatrix} \\
&= \sigma_i \begin{pmatrix} f_i \\ e_i \end{pmatrix}.
\end{aligned}$$

(We have used that both $\{f_j\}$ and $\{e_j\}$ are orthonormal systems.) Thus, g_i^+ is an eigenvector of $\bar{X}$ with the eigenvalue $\sigma_i(X)$. Similar computation shows that g_i^- is an eigenvector of $\bar{X}$ with the eigenvalue $-\sigma_i(X)$.

It remains to verify that if $h = \binom{f}{e}$ is orthogonal to all $g_i^{\pm}$ (f is l-dimensional, e is k-dimensional), then $\bar{X}h = 0$. Indeed, the orthogonality assumption means that $f^T f_i \pm e^T e_i = 0 \ \forall i$, whence $e^T e_i = 0$ and $f^T f_i = 0 \ \forall i$. Consequently,

$$\begin{pmatrix} 0 & X^T \\ X & 0 \end{pmatrix} \begin{pmatrix} f \\ e \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{k} \sigma_j f_j (e_j^T e) \\ \sum_{i=1}^{k} \sigma_j e_j (f_j^T f) \end{pmatrix} = 0.$$

Looking at Proposition 4.2.1, we see that the fact that specific functions of eigenvalues of a symmetric matrix X, namely, the sums $S_k(X)$ of k largest eigenvalues of X, are SDr underlies the possibility to build SDRs for a wide class of functions of the eigenvalues. The role of the sums of k largest singular values of a *rectangular* matrix X is equally important.

Proposition 4.2.2. *Let $g(x_1, \ldots, x_k) : \mathbf{R}^k_+ \to \mathbf{R} \cup \{+\infty\}$ be a symmetric monotone function:*

$$0 \le y \le x \in \mathrm{Dom} f \Rightarrow f(y) \le f(x).$$

Assume that g is SDr:

$$t \ge g(x) \Leftrightarrow \exists u : \mathcal{S}(x, t, u) \succeq 0$$

with $\mathcal{S}$ affinely depending on x, t, u. Then the function

$$f(X) = g(\sigma(X))$$

of $k \times l$ ($k \le l$) rectangular matrix X is SDr, with SDR given by the relation

$$\begin{array}{ll} \text{(a)} & t \ge f(X) \\ & \Updownarrow \\ & \exists x_1, \ldots, x_k, u : \\ \text{(b)} & \left\{ \begin{array}{c} \mathcal{S}(x_1, \ldots, x_k, t, u) \succeq 0, \\ x_1 \ge x_2 \ge \cdots \ge x_k, \\ \Sigma_j(X) \le x_1 + \cdots + x_j, \ \ j = 1, \ldots, m. \end{array} \right. \end{array} \tag{4.2.4}$$

Note the difference between the symmetric (Proposition 4.2.1) and the nonsymmetric (Proposition 4.2.2) situations. In the former the function $g(x)$ was assumed to be SDr and symmetric only, while in the latter the monotonicity requirement is added.

The proof of Proposition 4.2.2 is outlined in the exercises to Lecture 4.

Nonlinear matrix inequalities. There are several cases when matrix inequalities $F(x) \succeq 0$, where F is a nonlinear function of x taking values in the space of symmetric $m \times m$ matrices, can be linearized—expressed via LMIs.

20.a. *General quadratic matrix inequality.* Let X be a rectangular $k \times l$ matrix and

$$F(X) = (AXB)(AXB)^T + CXD + (CXD)^T + E$$

be a quadratic matrix-valued function of X; here $A, B, C, D, E = E^T$ are rectangular matrices of appropriate sizes. Let m be the row size of the values of F. Consider the $\succeq$-epigraph of the (matrix-valued!) function F—the set

$$\{(X, Y) \in \mathbf{M}^{k,l} \times \mathbf{S}^m \mid F(X) \preceq Y\}.$$

We claim that this set is SDr with the SDR

$$\left(\begin{array}{c|c} I_r & (AXB)^T \\ \hline AXB & Y - E - CXD - (CXD)^T \end{array} \right) \succeq 0 \qquad (B : l \times r).$$

Indeed, by the Schur complement lemma, our LMI is satisfied if and only if the Schur complement of the northwestern block is positive semidefinite, which is exactly our original quadratic matrix inequality.

20.b. *General fractional-quadratic matrix inequality*. Let X be a rectangular $k \times l$ matrix and V be a positive definite symmetric $l \times l$ matrix. Then we can define the matrix-valued function

$$F(X, V) = XV^{-1}X^T,$$

taking values in the space of $k \times k$ symmetric matrices. We claim that the closure of the $\succeq$-epigraph of this (matrix-valued!) function, i.e., the set

$$E = \mathrm{cl}\{(X, V; Y) \in \mathbf{M}^{k,l} \times \mathbf{S}^l_{++} \times \mathbf{S}^k \mid F(X, V) \equiv XV^{-1}X^T \preceq Y\}$$

is SDr, and an SDR of this set is given by the LMI

$$\begin{pmatrix} V & X^T \\ X & Y \end{pmatrix} \succeq 0. \tag{R}$$

Indeed, by the Schur complement lemma, a triple (X, V, Y) with *positive definite* V belongs to the epigraph of F—satisfies the relation $F(X, V) \preceq Y$—if and only if it satisfies (R). Now, if a triple (X, V, Y) belongs to E, i.e., it is the limit of a sequence of triples from the epigraph of F, then it satisfies (R) (as a limit of triples satisfying (R)). Conversely, if a triple (X, V, Y) satisfies (R), then V is positive semidefinite (as a diagonal block in a positive semidefinite matrix). The regularized triples $(X, V_\epsilon = V + \epsilon I_l, Y)$ associated with $\epsilon > 0$ satisfy (R) along with the triple (X, V, R); since, as we just have seen, $V \succeq 0$, we have $V_\epsilon \succ 0$, for $\epsilon > 0$. Consequently, the triples (X, V_ϵ, Y) belong to E (this was our very first observation). Since the triple (X, V, Y) is the limit of the regularized triples that, as we have seen, all belong to the epigraph of F, the triple (X, Y, V) belongs to the closure E of this epigraph.

20.c. *Matrix inequality* $Y \preceq (C^T X^{-1} C)^{-1}$. In the case of scalars x, y the inequality $y \le (cx^{-1}c)^{-1}$ in variables x, y is just an awkward way to write the linear inequality $y \le c^{-2}x$, but it leads naturally to the matrix analogy of the original inequality, namely, $Y \preceq (C^T X^{-1} C)^{-1}$, with rectangular $m \times n$ matrix C, variable symmetric $n \times n$ matrix Y, and $m \times m$ matrix X. For the matrix inequality to make sense, we should assume that the rank of C equals n (and thus $m \ge n$). Under this assumption, the matrix $(C^T X^{-1} C)^{-1}$ makes sense at least for a positive definite X. We claim that the closure of the solution set of the resulting inequality, the set

$$\mathcal{X} = \mathrm{cl}\{(X, Y) \in \mathbf{S}^m \times \mathbf{S}^n \mid X \succ 0, Y \preceq (C^T X^{-1} C)^{-1}\},$$

is SDr:

$$\mathcal{X} = \{(X, Y) \mid \exists Z : Y \preceq Z, Z \succeq 0, X \succeq CZC^T\}.$$

Indeed, let us denote by $\mathcal{X}'$ the set in the right-hand side of the latter relation; we should prove that $\mathcal{X}' = \mathcal{X}$. By definition, $\mathcal{X}$ is the closure of its intersection with the domain $X \succ 0$. It is clear that $\mathcal{X}'$ also is the closure of its intersection with the domain $X \succ 0$. Thus, all we need to prove is that a pair (Y, X) with $X \succ 0$ belongs to $\mathcal{X}$ if and only if it belongs to $\mathcal{X}'$.

"If" part. Assume that $X \succ 0$ and $(Y, X) \in \mathcal{X}'$. Then there exists Z such that $Z \succeq 0$, $Z \succeq Y$, and $X \succeq CZC^T$. Let us choose a sequence $Z_i \succ Z$ such that $Z_i \to Z$, $i \to \infty$. Since $CZ_iC^T \to CZC^T \preceq X$ as $i \to \infty$, we can find a sequence of matrices X_i such that $X_i \to X$, $i \to \infty$, and $X_i \succ CZ_iC^T$ $\forall i$. By the Schur complement lemma, the matrices $\begin{pmatrix} X_i & C \\ C^T & Z_i^{-1} \end{pmatrix}$ are positive definite; applying this lemma again, we conclude that $Z_i^{-1} \succeq C^T X_i^{-1} C$. Note that the left-hand and right-hand matrices in the latter inequality are positive definite. Now let us use the following simple fact.

Lemma 4.2.2. *Let U, V be positive definite matrices of the same size. Then*

$$U \preceq V \Leftrightarrow U^{-1} \succeq V^{-1}.$$

Proof. Note that we can multiply an inequality $A \preceq B$ by a matrix Q from the left and Q^T from the right:

$$A \preceq B \Rightarrow QAQ^T \preceq QBQ^T \quad [A, B \in \mathbf{S}^m, Q \in \mathbf{M}^{k,m}]$$

(why?). Thus, if $0 \prec U \preceq V$, then $V^{-1/2}UV^{-1/2} \preceq V^{-1/2}VV^{-1/2} = I$ (note that $V^{-1/2} = [V^{-1/2}]^T$), whence clearly $V^{1/2}U^{-1}V^{1/2} = [V^{-1/2}UV^{-1/2}]^{-1} \succeq I$. Thus, $V^{1/2}U^{-1}V^{1/2} \succeq I$. Multiplying this inequality from the left and from the right by $V^{-1/2} = [V^{-1/2}]^T$, we get $U^{-1} \succeq V^{-1}$. □

Applying Lemma 4.2.2 to the inequality $Z_i^{-1} \succeq C^T X_i^{-1} C [\succ 0]$, we get $Z_i \preceq (C^T X_i^{-1} C)^{-1}$. As $i \to \infty$, the left-hand side in this inequality converges to Z, and the right-hand side converges to $(C^T X^{-1} C)^{-1}$. Hence $Z \preceq (C^T X^{-1} C)^{-1}$, and since $Y \preceq Z$, we get $Y \preceq (C^T X^{-1} C)^{-1}$, as claimed.

"Only if" part. Let $X \succ 0$ and $Y \preceq (C^T X^{-1} C)^{-1}$; we should prove that there exists $Z \succeq 0$ such that $Z \succeq Y$ and $X \succeq CZC^T$. We claim that the required relations are satisfied by $Z = (C^T X^{-1} C)^{-1}$. The only nontrivial part of the claim is that $X \succeq CZC^T$, and here is the required justification: by its origin $Z \succ 0$ and by the Schur complement lemma the matrix $\begin{pmatrix} Z^{-1} & C^T \\ C & X \end{pmatrix}$ is positive semidefinite, whence, by the same lemma, $X \succeq C(Z^{-1})^{-1}C^T = CZC^T$.

Nonnegative polynomials. Consider the problem of the best polynomial approximation—given a function f on certain interval, we want to find its best uniform (or least squares, etc.) approximation by a polynomial of a given degree. This problem arises typically as a subproblem in all kinds of signal processing problems. In some situations the approximating polynomial is required to be nonnegative (think, e.g., of the case where the resulting polynomial is an estimate of an unknown probability density). How do we express the nonnegativity restriction? As shown by Nesterov,[19] it can be done via SDP:

> *For every k, the set of all nonnegative* (on the entire axis, or on a given ray, or on a given segment) *polynomials of degrees $\leq k$ is SDr.*

[19] Y. Nesterov, Squared functional systems and optimization problems, in *High Performance Optimization*, H. Frenk, K. Roos, T. Terlaky, S. Zhang, eds., Kluwer Academic Publishers, Dordrecht, the Netherlands, 2000, pp. 405–440.

In this statement (and everywhere below) we identify a polynomial $p(t) = \sum_{i=0}^{k} p_i t^i$ of degree (not exceeding) k with the $(k+1)$-dimensional vector $\mathrm{Coef}(p) = (p_0, p_1, \ldots, p_k)^T$ of the coefficients of p. Consequently, a set of polynomials of degrees $\le k$ becomes a set in $\mathbf{R}^{k+1}$, and we may ask whether this set is or is not SDr.

Let us look at the SDRs of different sets of nonnegative polynomials. The key here is to get an SDR for the set $P^+_{2k}(\mathbf{R})$ of polynomials of (at most) a given degree $2k$ which are nonnegative on the entire axis.[20]

21.a. *Polynomials nonnegative on the entire axis.* The set $P^+_{2k}(\mathbf{R})$ is SDr—it is the image of the semidefinite cone $\mathbf{S}^{k+1}_+$ under the affine mapping

$$X \mapsto \mathrm{Coef}(e^T(t)Xe(t)) : \mathbf{S}^{k+1} \to \mathbf{R}^{2k+1}, \quad e(t) = \begin{pmatrix} 1 \\ t \\ t^2 \\ \ldots \\ t^k \end{pmatrix}. \tag{C}$$

First note that the fact that $P^+ \equiv P^+_{2k}(\mathbf{R})$ is an affine image of the semidefinite cone indeed implies the SD-representability of P^+; see the calculus of conic representations in Lecture 3. Thus, all we need is to show that P^+ is exactly the same as the image, let it be called P, of $\mathbf{S}^{k+1}_+$ under the mapping (C).

1. The fact that P is contained in P^+ is immediate. Indeed, let X be a $(k+1)\times(k+1)$ positive semidefinite matrix. Then X is a sum of dyadic matrices:

$$X = \sum_{i=1}^{k+1} p^i (p^i)^T,\ p^i = (p_{i0}, p_{i1}, \ldots, p_{ik})^T \in \mathbf{R}^{k+1}$$

(why?). But then

$$e^T(t)Xe(t) = \sum_{i=1}^{k+1} e^T(t)p^i[p^i]^T e(t) = \sum_{i=1}^{k+1} \left(\sum_{j=0}^{k} p_{ij} t^j \right)^2$$

is the sum of squares of other polynomials and therefore is nonnegative on the axis. Thus, the image of X under the mapping (C) belongs to P^+.

Note that by reversing our reasoning, we get the following result:

(!) *If a polynomial $p(t)$ of degree $\le 2k$ can be represented as a sum of squares of other polynomials, then the vector* $\mathrm{Coef}(p)$ *of the coefficients of p belongs to the image of $\mathbf{S}^{k+1}_+$ under the mapping* (C).

With (!), the remaining part of the proof—the demonstration that the image of $\mathbf{S}^{k+1}_+$ *contains* P^+—is readily given by the following well-known algebraic fact:

(!!) *A polynomial is nonnegative on the axis if and only if it is a sum of squares of polynomials.*

[20] It is clear why we have restricted the degree to be even: a polynomial of an odd degree cannot be nonnegative on the entire axis!

The proof of (!!) is so nice that we cannot resist the temptation to present it here. The "if" part is evident. To prove the "only if" part, assume that $p(t)$ is nonnegative on the axis, and let the degree of p (it must be even) be $2k$. Now let us look at the roots of p. The real roots $\lambda_1, \dots, \lambda_r$ must be of even multiplicities $2m_1, 2m_2, \dots, 2m_r$ each (otherwise p would alter its sign in a neighborhood of a root, which contradicts the nonnegativity). The complex roots of p can be arranged in conjugate pairs $(\mu_1, \mu_1^*), (\mu_2, \mu_2^*), \dots, (\mu_s, \mu_s^*)$, and the factor of p

$$(t-\mu_i)(t-\mu_i^*) = (t-\Re\mu_i)^2 + (\Im\mu_i)^2$$

corresponding to such a pair is a sum of two squares. Finally, the leading coefficient of p is positive. Consequently, we have

$$p(t) = \omega^2[(t-\lambda_1)^2]^{m_1} \dots [(t-\lambda_r)^2]^{m_r}[(t-\mu_1)(t-\mu_1^*)] \dots [(t-\mu_s)(t-\mu_s^*)]$$

is a product of sums of squares. But such a product is itself a sum of squares (open the parentheses)!

In fact, we can say more: a nonnegative polynomial p is a sum of just two squares! To see this, note that, as we have seen, p is a product of sums of *two* squares and take into account the following fact (Louville):

The product of sums of two squares is again a sum of two squares:

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

(Compare with "The modulus of a product of two complex numbers is the product of their modulae".)

Equipped with the SDR of the set $P_{2k}^+(\mathbf{R})$ of polynomials nonnegative on the entire axis, we can immediately obtain SDRs for the polynomials nonnegative on a given ray or segment:

21.b. *Polynomials nonnegative on a ray or segment.*

1. The set $P_k^+(\mathbf{R}_+)$ of (coefficients of) polynomials of degree $\le k$ which are nonnegative on the nonnegative ray is SDr.

Indeed, this set is the inverse image of the SDr set $P_{2k}^+(\mathbf{R})$ under the linear mapping of the spaces of (coefficients of) polynomials given by the mapping

$$p(t) \mapsto p^+(t) \equiv p(t^2).$$

(Recall that the inverse image of an SDr set is SDr.)

2. The set $P_k^+([0,1])$ of (coefficients of) polynomials of degree $\le k$ which are nonnegative on the segment $[0,1]$ is SDr.

Indeed, a polynomial $p(t)$ of degree $\le k$ is nonnegative on $[0,1]$ if and only if the rational function

$$g(t) = p\left(\frac{t^2}{1+t^2}\right)$$

is nonnegative on the entire axis, or, which is the same, if and only if the polynomial

$$p^+(t) = (1+t^2)^k g(t)$$

of degree $\le 2k$ is nonnegative on the entire axis. The coefficients of p^+ depend linearly on the coefficients of p, and we conclude that $P_k^+([0,1])$ is the inverse image of the SDr set $P_{2k}^+(\mathbf{R})$ under certain linear mapping.

Our last example in this series deals with trigonometric polynomials

$$p(\phi) = a_0 + \sum_{l=1}^{k} [a_l \cos(l\phi) + b_l \sin(l\phi)].$$

Identifying such a polynomial with its vector of coefficients $\mathrm{Coef}(p) \in \mathbf{R}^{2k+1}$, we may ask how to express the set $S_k^+(\Delta)$ of those trigonometric polynomials of degree $\le k$ which are nonnegative on a segment $\Delta \subset [-\pi, \pi]$.

21.c. *Trigonometric polynomials nonnegative on a segment.* The set $S_k^+(\Delta)$ is SDr.

Indeed, $\sin(l\phi)$ and $\cos(l\phi)$ are polynomials of $\sin(\phi)$ and $\cos(\phi)$, and the latter functions, in turn, are rational functions of $\zeta = \tan(\phi/2)$:

$$\cos(\phi) = \frac{1-\zeta^2}{1+\zeta^2}, \sin(\phi) = \frac{2\zeta}{1+\zeta^2} \quad [\zeta = \tan(\phi/2)].$$

Consequently, a trigonometric polynomial $p(\phi)$ of degree $\le k$ can be represented as a rational function of $\zeta = \tan(\phi/2)$:

$$p(\phi) = \frac{p^+(\zeta)}{(1+\zeta^2)^k} \quad [\zeta = \tan(\phi/2)],$$

where the coefficients of the algebraic polynomial p^+ of degree $\le 2k$ are linear functions of the coefficients of p. Now, the requirement for p to be nonnegative on a given segment $\Delta \subset [-\pi, \pi]$ is equivalent to the requirement for p^+ to be nonnegative on a segment Δ^+ (which, depending on Δ, may be the usual finite segment or a ray or the entire axis). We see that $S_k^+(\Delta)$ is an inverse image, under certain linear mapping, of the SDr set $P_{2k}^+(\Delta^+)$, so that $S_k^+(\Delta)$ itself is SDr.

Finally, we may ask which part of the above results can be saved when we pass from nonnegative polynomials of one variable to those of two or more variables. Unfortunately, not too much. Among nonnegative polynomials of a given degree with $r > 1$ variables, exactly those that are sums of squares can be obtained as the image of a positive semidefinite cone under certain linear mapping similar to (D). The difficulty is that in the multidimensional case the nonnegativity of a polynomial is not equivalent to its representability as a sum of squares; thus, the positive semidefinite cone gives only part of the polynomials we want to describe.

4.3 Applications I: Combinatorics

Due to its tremendous expressive abilities, SDP has an extremely wide spectrum of applications. We shall overview the most important of these applications. We start with brief

presentation of what comes from inside mathematics and then continue with engineering applications. The most important mathematical applications of SDP deal with relaxations of combinatorial problems.

Combinatorial problems and their relaxations. Numerous problems of planning, scheduling, routing, etc., can be posed as *combinatorial optimization problems*, i.e., optimization programs with discrete design variables (integer or zero-one). There are several universal forms of combinatorial problems, among them LP with integer variables and LP with 0-1 variables. A problem given in one of these forms can always be converted to any other universal form, so that in principle it does not matter which form is used. Now, the majority of combinatorial problems are difficult—we do not know theoretically efficient (in a certain precise meaning of the notion) algorithms for solving these problems. What we do know is that nearly all these difficult problems are, in a sense, equivalent to each other and are *NP-complete*. The exact meaning of the latter notion will be explained in Lecture 5; for the time being it suffices to say that NP-completeness of a problem P means that the problem is as difficult as a combinatorial problem can be—if we knew an efficient algorithm for P, we would be able to convert it to an efficient algorithm for any other combinatorial problem. NP-complete problems may look extremely simple, as demonstrated by the following example:

> *Given n stones of positive integer weights (i.e., given n positive integers $a_1, \ldots, a_n$), check whether you can partition these stones into two groups of equal weight, i.e., check whether a linear equation*
>
> $$\sum_{i=1}^{n} a_i x_i = 0$$
>
> *has a solution with $x_i = \pm 1$.*

Theoretically difficult combinatorial problems happen to be difficult to solve in practice as well. An important ingredient in virtually all algorithms for combinatorial optimization is a technique for building bounds for the unknown optimal value of a given (sub)problem. A typical way to estimate the optimal value of an optimization program

$$f^* = \min_x \{f(x) : x \in X\}$$

from above is to present a feasible solution $\bar{x}$; then clearly $f^* \leq f(\bar{x})$. And a typical way to bound the optimal value from below is to pass from the problem to its relaxation,

$$f_* = \min_x \{f(x) : x \in X'\},$$

increasing the feasible set: $X \subset X'$. Clearly, $f_* \leq f^*$, so, whenever the relaxation is efficiently solvable (to ensure this, we should take care in choosing X'), it provides us with a computable lower bound on the actual optimal value.

When building a relaxation, one should take care of two issues. On one hand, we want the relaxation to be efficiently solvable. On the other hand, we want the relaxation to

be tight. Otherwise the lower bound we get may be by far too optimistic and therefore not useful. For a long time, the only practical relaxations were the LP ones, since these were the only problems one could solve efficiently. With recent progress in optimization techniques, nonlinear relaxations become more practical. As a result, we are witnessing a growing theoretical and computational activity in the area of nonlinear relaxations of combinatorial problems. In these developments, most, if not all, deal with semidefinite relaxations. Let us look how they emerge.

4.3.1 Shor's semidefinite relaxation scheme

As mentioned, there are numerous universal forms of combinatorial problems. For example, a combinatorial problem can be posed as minimizing a quadratic objective under quadratic equality constraints:

$$\begin{array}{lll} \text{minimize in} \quad x \in \mathbf{R}^n & f_0(x) & = \; x^T A_0 x + 2b_0^T x + c_0 \\ \text{s.t.} & & \\ & f_i(x) & = \; x^T A_i x + 2b_i^T x + c_i = 0, \;\; i = 1, \ldots, m. \end{array} \tag{4.3.5}$$

To see that this form is universal, note that it covers the classical universal combinatorial problem—a generic LP program with Boolean (0-1) variables:

$$\min_x \left\{ c^T x : a_i^T x - b_i \geq 0, \;\; i = 1, \ldots, m; x_j \in \{0, 1\}, \;\; j = 1, \ldots, n \right\}. \tag{B}$$

Indeed, the fact that a variable x_j must be Boolean can be expressed by the quadratic equality

$$x_j^2 - x_j = 0,$$

and a linear inequality $a_i^T x - b_i \geq 0$ can be expressed by the quadratic equality $a_i^T x - b_i - s_i^2 = 0$, where s_i is an additional variable. Thus, (B) is equivalent to the problem

$$\min_{x,s} \left\{ c^T x : a_i^T x - b_i - s_i^2 = 0, \;\; i = 1, \ldots, m; x_j^2 - x_j = 0, \;\; j = 1, \ldots, n \right\},$$

and this problem is of the form (4.3.5).

To bound from below the optimal value in (4.3.5), we may use the same technique we used for building the dual problem. We choose somehow weights λ_i, $i = 1, \ldots, m$, of arbitrary signs and add the constraints of (4.3.5) with these weights to the objective, thus coming to the function

$$\begin{array}{rcl} f_\lambda(x) & = & f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) \\ & = & x^T A(\lambda) x + 2b^T(\lambda) x + c(\lambda), \end{array} \tag{4.3.6}$$

where

$$\begin{aligned} A(\lambda) &= A_0 + \sum_{i=1}^{m} \lambda_i A_i, \\ b(\lambda) &= b_0 + \sum_{i=1}^{m} \lambda_i b_i, \\ c(\lambda) &= c_0 + \sum_{i=1}^{m} \lambda_i c_i. \end{aligned}$$

By construction, the function $f_\lambda(x)$ is equal to the actual objective $f_0(x)$ on the feasible set of the problem (4.3.5). Consequently, the unconstrained infimum of this function

$$a(\lambda) = \inf_{x \in \mathbf{R}^n} f_\lambda(x)$$

is a lower bound for the optimal value in (4.3.5). We come to the following simple result (cf. the weak duality theorem):

($*$) *Assume that* $\lambda \in \mathbf{R}^m$ *and* $\zeta \in \mathbf{R}$ *are such that*

$$f_\lambda(x) - \zeta \geq 0 \quad \forall x \in \mathbf{R}^n \tag{4.3.7}$$

(i.e., that $\zeta \leq a(\lambda)$*). Then* ζ *is a lower bound for the optimal value in* (4.3.5).

It remains to clarify what is meant that (4.3.7) holds. Recalling the structure of f_λ, we see that it means that the inhomogeneous quadratic form

$$g_\lambda(x) = x^T A(\lambda) x + 2b^T(\lambda) x + c(\lambda) - \zeta$$

is nonnegative on the entire space. Now, an inhomogeneous quadratic form

$$g(x) = x^T A x + 2b^T x + c$$

is nonnegative everywhere if and only if a certain associated homogeneous quadratic form is nonnegative everywhere. Indeed, given $t \neq 0$ and $x \in \mathbf{R}^n$, the fact that $g(t^{-1}x) \geq 0$ means exactly the nonnegativity of the homogeneous quadratic form $G(x, t)$

$$G(x, t) = x^T A x + 2tb^T x + ct^2$$

with $(n+1)$ variables x, t. We see that if g is nonnegative, then G is nonnegative whenever $t \neq 0$; by continuity, G then is nonnegative everywhere. Thus, if g is nonnegative, then G is, and of course vice versa (since $g(x) = G(x, 1)$). Now, to say that G is nonnegative everywhere is literally the same as to say that the matrix

$$\begin{pmatrix} c & b^T \\ b & A \end{pmatrix} \tag{4.3.8}$$

is positive semidefinite.

It is worthwhile to catalogue our simple observation:

Simple lemma. *A quadratic inequality with a (symmetric) $n \times n$ matrix A*

$$x^T Ax + 2b^T x + c \geq 0$$

is identically true—is valid for all $x \in \mathbf{R}^n$—if and only if the matrix (4.3.8) *is positive semidefinite.*

Applying this observation to $g_\lambda(x)$, we get the following equivalent reformulation of (*):

If $(\lambda, \zeta) \in \mathbf{R}^m \times \mathbf{R}$ satisfy the LMI

$$\begin{pmatrix} \sum_{i=1}^m \lambda_i c_i - \zeta, & b_0^T + \sum_{i=1}^m \lambda_i b_i^T \\ b_0 + \sum_{i=1}^m \lambda_i b_i, & A_0 + \sum_{i=1}^m \lambda_i A_i \end{pmatrix} \succeq 0,$$

then ζ is a lower bound for the optimal value in (4.3.5).

Now, what is the best lower bound we can get with this scheme? Of course, it is the optimal value of the semidefinite program

$$\max_{\zeta,\lambda} \left\{ \zeta : \begin{pmatrix} c_0 + \sum_{i=1}^m \lambda_i c_i - \zeta, & b_0^T + \sum_{i=1}^m \lambda_i b_i^T \\ b_0 + \sum_{i=1}^m \lambda_i b_i, & A_0 + \sum_{i=1}^m \lambda_i A_i \end{pmatrix} \succeq 0 \right\}. \tag{4.3.9}$$

We have proved the following simple proposition.

Proposition 4.3.1. *The optimal value in* (4.3.9) *is a lower bound for the optimal value in* (4.3.5).

The outlined scheme is extremely transparent, but it looks different from a relaxation scheme as explained above—where is the extension of the feasible set of the original problem? In fact the scheme is of this type. To see it, note that the value of a quadratic form at a point $x \in \mathbf{R}^n$ can be written as the Frobenius inner product of a matrix defined by the problem data and the dyadic matrix $X(x) = \binom{1}{x}\binom{1}{x}^T$:

$$x^T Ax + 2b^T x + c = \begin{pmatrix} 1 \\ x \end{pmatrix}^T \begin{pmatrix} c & b^T \\ b & A \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix} = \mathrm{Tr}\left(\begin{pmatrix} c & b^T \\ b & A \end{pmatrix} X(x) \right).$$

Consequently, (4.3.5) can be written as

$$\min_x \left\{ \mathrm{Tr}\left(\begin{pmatrix} c_0 & b_0^T \\ b_0 & A_0 \end{pmatrix} X(x) \right) : \mathrm{Tr}\left(\begin{pmatrix} c_i & b_i^T \\ b_i & A_i \end{pmatrix} X(x) \right) = 0,\ i = 1, \ldots, m \right\}. \tag{4.3.10}$$

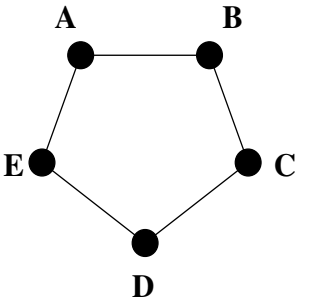


Figure 4.1. *Graph C_5.*

Thus, we may think of (4.3.6) as a problem with linear objective and linear equality constraints and with the design vector X that is a symmetric $(n+1)\times(n+1)$ matrix running through the nonlinear manifold $\mathcal{X}$ of dyadic matrices $X(x)$, $x\in\mathbf{R}^n$. Obviously, all points of $\mathcal{X}$ are positive semidefinite matrices with northwestern entry 1. Now let $\bar{\mathcal{X}}$ be the set of all such matrices. Replacing $\mathcal{X}$ by $\bar{\mathcal{X}}$, we get a relaxation of (4.3.10), which is essentially our original problem (4.3.5). This relaxation is the semidefinite program

$$\min_X \left\{\mathrm{Tr}(\bar{A}_0X) : \mathrm{Tr}(\bar{A}_iX)=0, i=1,\dots,m;\ X\succeq 0;\ X_{11}=1\right\},\quad \left[A_i=\begin{pmatrix} c_i & b_i^T\\ b_i & A_i\end{pmatrix},\ i=1,\dots,m\right]. \tag{4.3.11}$$

We get the following proposition.

Proposition 4.3.2. *The optimal value of the semidefinite program* (4.3.11) *is a lower bound for the optimal value in* (4.3.5).

One can easily verify that problem (4.3.9) is just the semidefinite dual of (4.3.11); thus, when deriving (4.3.9), we were in fact implementing the idea of relaxation. This is why in the sequel we call both (4.3.11) and (4.3.9) *semidefinite relaxations* of (4.3.5). Let us look at several interesting examples.

4.3.2 Stability number, Shannon capacity, and Lovasz capacity of a graph

Stability number of a graph. Consider a (nonoriented) graph—a finite set of nodes linked by arcs,[21] like the simple five-node graph C_5 shown in Fig. 4.1. One of the fundamental characteristics of a graph Γ is its *stability number* $\alpha(\Gamma)$, defined as the maximum cardinality of an independent subset of nodes—a subset such that no two nodes from it are linked by an arc. The stability number for the graph C_5 is, e.g., 2, and a maximal independent set is, e.g., {A; C}.

[21] One of the formal definitions of a (nonoriented) graph is as follows. An n-node graph is just an $n\times n$ symmetric matrix A with entries 0, 1 and zero diagonal. The rows (and the columns) of the matrix are identified with the nodes $1,2,\dots,n$ of the graph, and the nodes i, j are adjacent (i.e., linked by an arc) exactly for those i, j with $A_{ij}=1$.

The problem of computing the stability number of a given graph is NP-complete. This is why it is important to know how to bound this number.

Shannon capacity of a graph. An upper bound on the stability number of a graph which is interesting by its own right is the *Shannon capacity* $\vartheta(\Gamma)$, defined as follows.

Let us treat the nodes of Γ as letters of a certain alphabet and the arcs as possible errors in a certain communication channel: you can send through the channel one letter per unit time, and what arrives on the other end of the channel can be either the letter you have sent or any letter adjacent to it. Now assume that you are planning to communicate with an addressee through the channel by sending n-letter words (n is fixed). You fix in advance a dictionary D_n of words to be used and make this dictionary known to the addressee. What you are interested in when building the dictionary is to get a good one, meaning that no word from it could be transformed by the channel into another word from the dictionary. If your dictionary satisfies this requirement, you may be sure that the addressee will never misunderstand you: whatever word from the dictionary you send and whatever possible transmission errors occur, the addressee is able either to get the correct message or to realize that the message was corrupted during transmission, but there is no risk that your "yes" will be read as "no"! Now, to utilize the channel at full capacity, you want to get as large a dictionary as possible. How many words can it include? The answer is clear: this is precisely the stability number of the graph Γ^n as follows. The nodes of Γ^n are ordered n-element collections of the nodes of Γ—all possible n-letter words in your alphabet; two distinct nodes $(i_1, \dots, i_n)$, $(j_1, \dots, j_n)$ are adjacent in Γ^n if and only if for every l the lth letters i_l and j_l in the two words either coincide or are adjacent in Γ (i.e., two distinct n-letter words are adjacent, if the transmission can convert one of them into the other one). Let us denote the maximum number of words in a good dictionary D_n (i.e., the stability number of Γ^n) by $f(n)$, The function $f(n)$ possesses the following nice property:

$$f(k)f(l) \le f(k+l), \quad k, l = 1, 2, \dots. \tag{$*$}$$

Indeed, given the best (of the cardinality $f(k)$) good dictionary D_k and the best good dictionary D_l, let us build a dictionary made up of all $(k+l)$-letter words as follows: the initial k-letter fragment of a word belongs to D_k, and the remaining l-letter fragment belongs to D_l. The resulting dictionary is clearly good and contains $f(k)f(l)$ words, and (*) follows.

Now, this is a simple exercise in analysis to see that for a nonnegative function f with property (*) one has

$$\lim_{k\to\infty} (f(k))^{1/k} = \sup_{k\ge 1} (f(k))^{1/k} \in [0, +\infty].$$

In our situation $\sup_{k\ge 1}(f(k))^{1/k} < \infty$, since clearly $f(k) \le n^k$, n being the number of letters (the number of nodes in Γ). Consequently, the quantity

$$\vartheta(\Gamma) = \lim_{k\to\infty} (f(k))^{1/k}$$

is well defined. Moreover, for every k the quantity $(f(k))^{1/k}$ is a lower bound for $\vartheta(\Gamma)$. The number $\vartheta(\Gamma)$ is called the Shannon capacity of Γ. Our immediate observation is as follows.

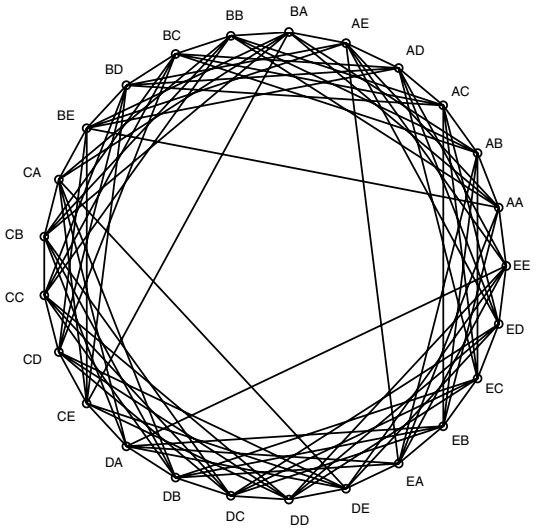


Figure 4.2. *Graph* $(C_5)^2$.

(B) *The Shannon capacity* $\vartheta(\Gamma)$ *majorates the stability number of* Γ*:*

$$\alpha(\Gamma) \le \vartheta(\Gamma).$$

Indeed, as we remember, $(f(k))^{1/k}$ is a lower bound for $\vartheta(\Gamma)$ for every $k = 1, 2, \ldots$; setting $k = 1$ and taking into account that $f(1) = \alpha(\Gamma)$, we get the desired result.

We see that the Shannon capacity number is an upper bound on the stability number, and this bound has a nice interpretation in terms of the information theory. The bad news is that we do not know how to compute the Shannon capacity. For example, what is it for the toy graph C_5?

The stability number of C_5 clearly is 2, so our first observation is that

$$\vartheta(C_5) \ge \alpha(C_5) = 2.$$

To get a better estimate, let us look the graph $(C_5)^2$ (as we remember, $\vartheta(\Gamma) \ge (f(k))^{1/k} = (\alpha(\Gamma^k))^{1/k}$ for every k). The graph $(C_5)^2$ has 25 nodes and 124 arcs and is shown in Fig. 4.2. With some effort, you can check that the stability number of $(C_5)^2$ is 5; a good five-element dictionary ($\equiv$ a five-node independent set in $(C_5)^2$) is, e.g., AA, BC, CE, DB, ED. Thus, we get

$$\vartheta(C_5) \ge \sqrt{\alpha((C_5)^2)} = \sqrt{5}.$$

Attempts to compute the subsequent lower bounds $(f(k))^{1/k}$, as long as they are implementable (think how many vertices there are in $(C_5)^8$!), do not yield any improvements, and for more than 20 years it remained unknown whether $\vartheta(C_5) = \sqrt{5}$ or is $> \sqrt{5}$. And this is for a toy graph! The breakthrough in the area of upper bounds for the stability number is due to Lovasz, who in the early 1970s found a new—computable!—bound of this type.

Lovasz capacity number. Given an n-node graph Γ, let us associate with it an affine matrix-valued function $\mathcal{L}(x)$ taking values in the space of $n \times n$ symmetric matrices, namely, as follows:

- For every pair i, j of indices ($1 \le i, j \le n$) such that the nodes i and j are not linked by an arc, the ijth entry of $\mathcal{L}$ is equal to 1.
- For a pair $i < j$ of indices such that the nodes i, j are linked by an arc, the ijth and the jith entries in $\mathcal{L}$ are equal to x_{ij}—to the variable associated with the arc (i, j).

Thus, $\mathcal{L}(x)$ is indeed an affine function of N design variables x_{ij}, where N is the number of arcs in the graph. For example, for graph C_5 the function $\mathcal{L}$ is as follows:

$$\mathcal{L} = \begin{pmatrix} 1 & x_{AB} & 1 & 1 & x_{EA} \\ x_{AB} & 1 & x_{BC} & 1 & 1 \\ 1 & x_{BC} & 1 & x_{CD} & 1 \\ 1 & 1 & x_{CD} & 1 & x_{DE} \\ x_{EA} & 1 & 1 & x_{DE} & 1 \end{pmatrix}.$$

Now, the Lovasz capacity number $\Theta(\Gamma)$ is defined as the optimal value of the optimization program

$$\min_x \{\lambda_{\max}(\mathcal{L}(x))\},$$

i.e., as the optimal value in the semidefinite program

$$\min_{\lambda, x} \{\lambda : \lambda I_n - \mathcal{L}(x) \succeq 0\}. \tag{L}$$

Proposition 4.3.3. *The Lovasz capacity number is an upper bound for the Shannon capacity,*

$$\Theta(\Gamma) \ge \vartheta(\Gamma),$$

and, consequently, for the stability number:

$$\Theta(\Gamma) \ge \vartheta(\Gamma) \ge \alpha(\Gamma).$$

For the graph C_5, the Lovasz capacity can be easily computed analytically and turns out to be exactly $\sqrt{5}$. Thus, a small byproduct of Lovasz's result is a solution of the problem that remained open for two decades.

Let us see how the Lovasz bound on the stability number can be obtained from the general relaxation scheme. To this end note that the stability number of an n-node graph Γ is the optimal value of the following optimization problem with 0-1 variables:

$$\max_x \left\{e^T x : x_i x_j = 0 \text{ whenever } i, j \text{ are adjacent nodes}, x_i \in \{0, 1\},\ i = 1, \ldots, n\right\},$$
$$e = (1, \ldots, 1)^T \in \mathbf{R}^n.$$

Indeed, 0-1 n-dimensional vectors can be identified with sets of nodes of Γ: the coordinates x_i of the vector x representing a set A of nodes are ones for $i \in A$ and zeros otherwise. The

quadratic equality constraints $x_i x_j = 0$ for such a vector express equivalently the fact that the corresponding set of nodes is independent, and the objective $e^T x$ counts the cardinality of this set.

As we remember, the 0-1 restrictions on the variables can be represented equivalently by quadratic equality constraints, so that the stability number of Γ is the optimal value of the following problem with quadratic (in fact, linear) objective and quadratic equality constraints:

$$\begin{array}{ll} \text{maximize} & e^T x \\ \text{s.t.} & \\ & \begin{array}{rcl} x_i x_j & = & 0, \ (i,j) \text{ is an arc}, \\ x_i^2 - x_i & = & 0, \ i = 1, \ldots, n. \end{array} \end{array} \tag{4.3.12}$$

The latter problem is in the form of (4.3.5); the only difference is that the objective should be maximized rather than minimized. Switching from maximization of $e^T x$ to minimization of $(-e)^T x$ and passing to (4.3.9), we get the problem

$$\max_{\zeta,\mu} \left\{ \zeta : \begin{pmatrix} -\zeta & -\frac{1}{2}(e+\mu)^T \\ -\frac{1}{2}(e+\mu) & A(\mu,\lambda) \end{pmatrix} \succeq 0 \right\},$$

where μ is n-dimensional and $A(\mu,\lambda)$ is as follows:

- The diagonal entries of $A(\mu,\lambda)$ are $\mu_1, \ldots, \mu_n$.
- The off-diagonal cells ij corresponding to nonadjacent nodes i, j (empty cells) are zeros.
- The off-diagonal cells $ij, i < j$, and the symmetric cells ji corresponding to adjacent nodes i, j (arc cells) are filled with free variables λ_{ij}.

Note that the optimal value in the resulting problem is a lower bound for minus the optimal value of (4.3.12), i.e., for minus the stability number of Γ.

Passing in the resulting problem from the variable ζ to a new variable $\xi = -\zeta$ and again switching from maximization of $\zeta = -\xi$ to minimization of ξ, we end up with the semidefinite program

$$\min_{\xi,\lambda,\mu} \left\{ \xi : \begin{pmatrix} \xi, & -\frac{1}{2}(e+\mu)^T \\ -\frac{1}{2}(e+\mu), & A(\mu,\lambda) \end{pmatrix} \succeq 0 \right\}. \tag{4.3.13}$$

The optimal value in this problem is the minus optimal value in the previous one, which, in turn, is a lower bound on the minus stability number of Γ; consequently, the optimal value in (4.3.13) is an upper bound on the stability number of Γ.

We have built a semidefinite relaxation (4.3.13) of the problem of computing the stability number of Γ; the optimal value in the relaxation is an upper bound on the stability number. To get the Lovasz relaxation, let us further fix the μ-variables at level 1. (This may only increase the optimal value in the problem, so that it still will be an upper bound for the stability number.)[22] With this modification, we come to the problem

$$\min_{\xi,\lambda} \left\{ \xi : \begin{pmatrix} \xi & -e^T \\ -e & A(e,\lambda) \end{pmatrix} \succeq 0 \right\}.$$

[22] In fact, setting $\mu_i = 1$, we do not vary the optimal value at all. See Exercise 4.32.

In every feasible solution to the problem, ξ should be ≥ 1 (it is an upper bound for $\alpha(\Gamma) \geq 1$). When $\xi \geq 1$, the LMI

$$\begin{pmatrix} \xi & -e^T \\ e & A(e,\lambda) \end{pmatrix} \succeq 0$$

by the Schur complement lemma is equivalent to the LMI

$$A(e,\lambda) \succeq (-e)\xi^{-1}(-e)^T$$

or, which is the same, to the LMI

$$\xi A(e,\lambda) - ee^T \succeq 0.$$

The left-hand-side matrix in the latter LMI is equal to $\xi I_n - B(\xi,\lambda)$, where the matrix $B(\xi,\lambda)$ is as follows:

- The diagonal cells of $B(\xi,\lambda)$ and the off-diagonal empty cells are filled with ones.
- The entries in arc cells ij, ji $(i < j)$ are equal to $1 - \xi\lambda_{ij}$. $(i < j)$ are filled with $\xi\lambda_{ij}$.

Passing from the design variables λ to the new ones $x_{ij} = 1 - \xi\lambda_{ij}$, we conclude that problem (4.3.13) with μ's set to ones is equivalent to the problem

$$\min_{\xi,x} \{\xi \to \min \mid \xi I_n - \mathcal{L}(x) \succeq 0\},$$

whose optimal value is exactly the Lovasz capacity number of Γ.

As a byproduct of our derivation, we get the easy part of the Lovasz Theorem—the inequality $\Theta(\Gamma) \geq \alpha(\Gamma)$. This inequality, however, could be easily obtained directly from the definition of $\Theta(\Gamma)$. The advantage of our derivation is that it demonstrates the origin of $\Theta(\Gamma)$.

How good is the Lovasz capacity number? The Lovasz capacity number plays a crucial role in numerous graph-related problems. There is an important subfamily of graphs—perfect graphs—for which this number coincides with the stability number. However, for a general-type graph Γ, $\Theta(\Gamma)$ may be a fairly poor bound for $\alpha(\Gamma)$. Lovasz has proved that for any graph Γ with n nodes, $\Theta(\Gamma)\Theta(\hat{\Gamma}) \geq n$, where $\hat{\Gamma}$ is the complement to Γ (i.e., two distinct nodes are adjacent in $\hat{\Gamma}$ if and only if they are not adjacent in Γ). It follows that for n-node graph Γ one always has $\max[\Theta(\Gamma), \Theta(\hat{\Gamma})] \geq \sqrt{n}$. On the other hand, it turns out that for a random n-node graph Γ (the arcs are drawn at random and independently of each other, with probability 0.5 to draw an arc linking two given distinct nodes) $\max[\alpha(\Gamma), \alpha(\hat{\Gamma})]$ is typically (with probability approaching 1 as n grows) of order of $\ln n$. It follows that for random n-node graphs a typical value of the ratio $\Theta(\Gamma)/\alpha(\Gamma)$ is at least of order of $n^{1/2}/\ln n$; as n grows, this ratio blows up to ∞.

A natural question arises: Are there difficult (NP-complete) combinatorial problems admitting good semidefinite relaxations—those with the quality of approximation not deteriorating as the sizes of instances grow? Let us look at two recent breakthrough results in this direction.

4.3.3 MAXCUT problem

The MAXCUT (maximum cut) problem is as follows.

Problem 4.3.1. *Let Γ be an n-node graph, and let the arcs (i, j) of the graph be associated with nonnegative weights a_{ij}. The problem is to find a cut of the largest possible weight, i.e., to partition the set of nodes into two parts S, S' in such a way that the total weight of all arcs linking S and S' (i.e., with one incident node in S and the other one in S') is as large as possible.*

In the MAXCUT problem, we may assume that the weights $a_{ij} = a_{ji} \geq 0$ are defined for every pair i, j of indices; it suffices to set $a_{ij} = 0$ for pairs i, j of nonadjacent nodes.

In contrast to the minimum cut problem (where we should minimize the weight of a cut instead of maximizing it), which is, basically, a nice LP program of finding the maximum flow in a net and is therefore efficiently solvable, the MAXCUT problem is as difficult as a combinatorial problem can be—it is NP-complete. However, it is easy to build a semidefinite relaxation of MAXCUT. To this end let us pose MAXCUT as a quadratic problem with quadratic equality constraints. Let Γ be an n-node graph. A cut (S, S')—a partitioning of the set of nodes in two disjoint parts S, S'—can be identified with an n-dimensional vector x with coordinates $\pm 1 - x_i = 1$ for $i \in S$, $x_i = -1$ for $i \in S'$. The quantity $\frac{1}{2}\sum_{i,j=1}^n a_{ij}x_ix_j$ is the total weight of arcs with both ends either in S or in S' minus the weight of the cut (S, S'). Consequently, the quantity

$$\frac{1}{2}\left[\frac{1}{2}\sum_{i,j=1}^n a_{ij} - \frac{1}{2}\sum_{i,j=1}^n a_{ij}x_ix_j\right] = \frac{1}{4}\sum_{i,j=1}^n a_{ij}(1 - x_ix_j)$$

is exactly the weight of the cut (S, S').

We conclude that the MAXCUT problem can be posed as the following quadratic problem with quadratic equality constraints:

$$\max_x \left\{\frac{1}{4}\sum_{i,j=1}^n a_{ij}(1 - x_ix_j) : x_i^2 = 1,\ i = 1, \ldots, n\right\}. \tag{4.3.14}$$

For this problem, the semidefinite relaxation (4.3.11) after evident simplifications becomes the semidefinite program

$$\begin{array}{ll} \text{maximize} & \frac{1}{4}\sum_{i,j=1}^n a_{ij}(1 - X_{ij}) \\ \text{s.t.} & \\ & X = [X_{ij}]_{i,j=1}^n = X^T \succeq 0, \\ & X_{ii} = 1,\ i = 1, \ldots, n. \end{array} \tag{4.3.15}$$

The optimal value in the latter problem is an upper bound for the optimal value of MAXCUT.

The fact that (4.3.15) is a relaxation of (4.3.14) can be established directly, independently of any general theory: (4.3.14) is the problem of maximizing the objective

$$\frac{1}{4}\sum_{i,j=1}^{n} a_{ij} - \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}x_i x_j \equiv \frac{1}{4}\sum_{i,j=1}^{n} a_{ij} - \frac{1}{4}\mathrm{Tr}(AX(x)), \quad X(x) = xx^T,$$

over all rank-1 matrices $X(x) = xx^T$ given by n-dimensional vectors x with entries ± 1. All these matrices are symmetric positive semidefinite with unit entries on the diagonal, i.e., they belong to the feasible set of (4.3.15). Thus, (4.3.15) indeed is a relaxation of (4.3.14).

The quality of the semidefinite relaxation (4.3.15) is given by the following brilliant result of Goemans and Williamson.[23]

THEOREM 4.3.1. *Let OPT be the optimal value of the MAXCUT problem* (4.3.14) *and SDP be the optimal value of the semidefinite relaxation* (4.3.15). *Then*

$$1 \geq \frac{OPT}{SDP} \geq 0.87856\ldots. \tag{4.3.16}$$

The reasoning used by Goemans and Williamson is so beautiful that it is impossible not to reproduce it here:

The left inequality in (4.3.16) is what we already know—it simply says that semidefinite program (4.3.15) is a relaxation of MAXCUT. To get the right inequality, Goemans and Williamson act as follows. Let $X = [X_{ij}]$ be a feasible solution to the semidefinite relaxation. Since X is positive semidefinite, it is the Gram matrix of a collection of n vectors $v_1, \ldots, v_n$:

$$X_{ij} = v_i^T v_j.$$

And since all X_{ii} are equal to 1, the vectors v_i are of the unit Euclidean norm. Given X, we can easily find v_i's (e.g., via the Choleski decomposition of X). Now let us look at the following procedure for generating random cuts of the graph. We choose at random, according to the uniform distribution on the unit sphere in $\mathbf{R}^n$, a unit vector v and build the cut

$$S = \{i \mid v^T v_i \geq 0\}.$$

What is the expected value of the weight of this random cut? The expected contribution of a particular pair i, j to the expected weight of our random cut is $\frac{1}{4}a_{ij}$ times twice the probability of the event that the nodes i and j will be separated by v, i.e., that the products $v^T v_i$ and $v^T v_j$ will have opposite signs. By elementary arguments, the probability of this event is just twice the ratio of the angle between the vectors v_i and v_j to 2π, as is seen in Fig. 4.3. Thus, the expected contribution of a pair i, j to the expected weight of our random cut is $\frac{1}{2}a_{ij}\frac{\mathrm{acos}(v_i^T v_j)}{\pi}$, and the expected weight of the random cut is

$$W[X] = \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}\frac{\mathrm{acos}(X_{ij})}{\pi}.$$

[23] M.X. Goemans and D.P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. ACM, 42 (1995), pp. 1115—1145.

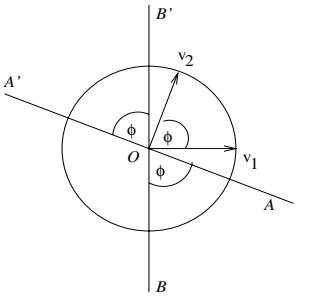


Figure 4.3. *Computing the expected weight of random cut. w is the projection of v on the 2D plane spanned by v_1 and v_2. The direction of w is uniformly distributed in $[0, 2\pi]$, and v separates v_1 and v_2 exactly when w belongs to one of the angles AOB, $A'OB'$.*

Comparing the right-hand side term by term with the value

$$f(X) = \frac{1}{4}\sum_{i,j=1}^{n} a_{ij}(1 - X_{ij})$$

of the objective in the relaxed problem (4.3.15) at X and taking into account that $a_{ij} \geq 0$ and that for $x \in [-1, 1]$ one has

$$\frac{\mathrm{acos}(x)}{\pi} \geq \alpha\frac{1}{2}(1 - x), \quad \alpha = 0.87856\ldots,$$

we come to

$$W[X] \geq \alpha f(X).$$

This inequality is valid for every feasible solution X of the semidefinite relaxation, in particular, for the optimal solution X^*. We conclude that already the expectation of the weight of random cut generated from X^* is at least $\alpha\, SDP$; the maximum possible weight OPT of a cut may be only larger than this expectation, so that $OPT/SDP \geq \alpha = 0.87856$.

Note that the above construction not only provides a proof of Theorem 4.3.1 but offers a randomized algorithm for constructing a random cut that, on average, has weight at least 0.87856 of OPT. Indeed, it suffices to solve the semidefinite relaxation (4.3.15). (This can be done efficiently, if we will be satisfied with an ϵ-solution—a feasible X such that the value of the objective of (4.3.15) at X is at least $(1 - \epsilon) \cdot SDP$—with a once forever fixed $\epsilon > 0$, say, with $\epsilon =$`1.e-6`.) After a (nearly) optimal solution X to (4.3.15) is found, we use it to generate random cuts, as explained in the above construction.

4.3.4 Extensions

In the MAXCUT problem, we are in fact maximizing the homogeneous quadratic form

$$x^T Ax \equiv \sum_{i=1}^{n}\left(\sum_{j=1}^{n} a_{ij}\right)x_i^2 - \sum_{i,j=1}^{n} a_{ij}x_i x_j$$

over the set S_n of n-dimensional vectors x with coordinates ± 1. The matrix A of this form is positive semidefinite (Exercise 4.2) and possesses a specific feature that the off-diagonal entries are nonpositive, while the sum of the entries in every row is 0. What happens when we maximize over S_n a quadratic form $x^T Ax$ with a general-type (symmetric) matrix A? An extremely nice result in this direction was recently obtained by Nesterov. The cornerstone of Nesterov's construction relates to the case when A is positive semidefinite, and this is the case we shall focus on. Note that the problem of maximizing a quadratic form $x^T Ax$ with positive semidefinite (and, say, integer) matrix A over S_n, same as MAXCUT, is NP-complete.

The semidefinite relaxation of the problem

$$\max_x \left\{x^T Ax : x \in S_n \quad [\Leftrightarrow x_i \in \{-1, 1\},\ i = 1, \ldots, n]\right\} \tag{4.3.17}$$

can be built in exactly the same way as (4.3.15) and turns out to be the semidefinite program

$$\begin{array}{lrcl} \text{maximize} & \mathrm{Tr}(AX) & & \\ \text{s.t.} & & & \\ & X = X^T = [X_{ij}]_{i,j=1}^n & \succeq & 0 \\ & X_{ii} & = & 1,\ i = 1, \ldots, n. \end{array} \tag{4.3.18}$$

The optimal value in this problem, let it again be called SDP, is $\geq$ the optimal value OPT in the original problem (4.3.17). The ratio OPT/SDP, however, cannot be too large.

Theorem 4.3.2. Nesterov's theorem. *Let A be positive semidefinite. Then*

$$SDP \geq OPT \geq \frac{2}{\pi} SDP \quad [2/\pi = 0.6366\ldots].$$

The proof utilizes the central idea of Goemans and Williamson in the following brilliant reasoning.

The inequality $SDP \geq OPT$ is valid since (4.3.18) is a relaxation of (4.3.17). Let X be a feasible solution to the relaxed problem; then $X_{ij} = v_i^T v_j$ for a system of unit vectors $v_1, \ldots, v_n$. Similar to the MAXCUT construction, we associate with this representation of X a random generator of vectors from S_n: choosing a direction v uniformly on the unit sphere, we build vector x with the ± 1-coordinates

$$x_i = \mathrm{sign}(v^T v_i),$$

where $\mathrm{sign}(a)$ is $+1$ for $a \geq 0$ and is -1 for $a < 0$. The expected value of the objective $x^T Ax$ of (4.3.17) over the generated points x is

$$V = \sum_{i,j=1}^n a_{ij} \mathbf{E}_v\{\mathrm{sign}(v^T v_i)\mathrm{sign}(v^T v_j)\},$$

where $\mathbf{E}_v$ denotes the expectation taken with respect to v uniformly distributed over the unit sphere. The expectation $\mathbf{E}_v\{\mathrm{sign}(v^T v_i)\mathrm{sign}(v^T v_j)\}$ can be easily computed: when projecting v on the 2D plane spanned by v_i, v_j, we get a vector w with the direction

uniformly distributed in $[0, 2\pi]$, and the expectation in question is the probability for w to have inner products with v_i and v_j of the same sign (i.e., to belong to the union of angles AOA′ and BOB′ in Fig. 4.3) minus probability to have inner products with v_i and v_j of opposite signs (i.e., to belong to the union of the angles AOB and A′OB′). The indicated difference is

$$\frac{1}{2\pi}\left[2\pi - 4\arccos(v_i^T v_j)\right] = \frac{2}{\pi}\arcsin(v_i^T v_j) = \frac{2}{\pi}\arcsin(X_{ij}).$$

Thus,

$$V = \frac{2}{\pi}\sum_{i,j=1}^{n} a_{ij}\arcsin(X_{ij}).$$

Recalling that V is the expected value of the objective in (4.3.17) with respect to certain probability distribution on the feasible set S_n of the problem, we get $V \le OPT$. Summarizing, we have proved the following lemma.

Lemma 4.3.1. *Let X be a feasible solution to* (4.3.18)*. Then the optimal value OPT in* (4.3.17) *satisfies the relation*

$$OPT \ge \frac{2}{\pi}\sum_{i,j=1}^{n} a_{ij}\arcsin(X_{ij}).$$

Consequently,

$$OPT \ge \frac{2}{\pi}\max\left\{\mathrm{Tr}(A\arcsin[X]) \mid X \succeq 0, X_{ii} = 1,\ i = 1, \ldots, n\right\}, \tag{4.3.19}$$

where $\arcsin[X]$ *is the matrix with the elements* $\arcsin(X_{ij})$.

Nesterov completes the proof by the following unexpected, although simple, observation.

> *For a positive semidefinite symmetric matrix X with diagonal entries ± 1 (in fact, for any positive semidefinite X with $|X_{ij}| \le 1$), one has*
>
> $$\arcsin[X] \succeq X.$$

The proof of the observation is immediate. Denoting by $[X]^k$ the matrix with the entries X_{ij}^k and making use of the Taylor series for the arcsin (this series converges uniformly on $[-1, 1]$), for a matrix X with all entries belonging to $[-1, 1]$, we get

$$\arcsin[X] - X = \sum_{k=1}^{\infty} \frac{1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2k-1)}{2^k k!(2k+1)}[X]^{2k+1},$$

and all we need is to note is that all matrices in the left-hand side are $\succeq 0$ along with X.[24]

The above observation, in view of the fact that A is positive semidefinite, implies that

$$\sum_{i,j=1}^{n} a_{ij}\mathrm{asin}(X_{ij}) = \mathrm{Tr}(A\mathrm{asin}[X]) \geq \mathrm{Tr}(AX)$$

for every feasible solution X of the semidefinite relaxation (4.3.18). Hence, the expression in the right-hand side of (4.3.19) is at least $\frac{2}{\pi}SDP$. □

Note that in fact the inequality in (4.3.19) is equality (see Exercise 4.39).

4.3.5 $\mathcal{S}$-lemma

Let us look again at the Lagrange relaxation of a quadratically constrained quadratic problem, but in the very special case when all the forms involved are homogeneous and the right-hand sides of the inequality constraints are zero:

$$\begin{array}{ll} \text{minimize} & x^T Bx \\ \text{s.t.} & x^T A_i x \geq 0,\ i = 1, \ldots, m \end{array} \tag{4.3.20}$$

($B, A_1, \ldots, A_m$ are given symmetric $m \times m$ matrices). Assume that the problem is feasible. In this case (4.3.20) is, at a first glance, a trivial problem: due to homogeneity, its optimal value is either $+\infty$ or 0, depending on whether there exists or does not exist a feasible vector x such that $x^T Bx < 0$. The challenge here is to detect which one of these two alternatives takes place, i.e., to understand whether or not a homogeneous quadratic inequality $x^T Bx \geq 0$ is a consequence of the system of homogeneous quadratic inequalities $x^T A_i x \geq 0$ or, which is the same, to understand when the implication

$$\begin{array}{ll} \text{(a)} & x^T A_i x \geq 0,\ i = 1, \ldots, m \\ & \Downarrow \\ \text{(b)} & x^T Bx \geq 0 \end{array} \tag{4.3.21}$$

holds true.

In the case of homogeneous linear inequalities it is easy to recognize when an inequality $x^T b \geq 0$ is a consequence of the system of inequalities $x^T a_i \geq 0$, $i = 1, \ldots, m$: by the Farkas lemma, it is the case if and only if the inequality is a linear consequence of the system, i.e., if b is representable as a linear combination, with nonnegative coefficients, of the vectors a_i. Now we are asking a similar question about homogeneous quadratic inequalities: When is (b) a consequence of (a)?

In general, there is no analogy of the Farkas lemma for homogeneous quadratic inequalities. Note, however, that the easy "if" part of the lemma can be extended to the

[24] That the entrywise product of two positive semidefinite matrices is positive semidefinite is a standard fact from linear algebra. The easiest way to understand it is to note that if P, Q are positive semidefinite symmetric matrices of the same size, then they are Gram matrices: $P_{ij} = p_i^T p_j$ for certain system of vectors p_i from certain (no matter from which exactly) $\mathbf{R}^N$ and $Q_{ij} = q_i^T q_j$ for a system of vectors q_i from certain $\mathbf{R}^M$. But then the entrywise product of P and Q—the matrix with the entries $P_{ij}Q_{ij} = (p_i^T p_j)(q_i^T q_j)$—also is a Gram matrix, namely, the Gram matrix of the matrices $p_i q_i^T \in \mathbf{M}^{N,M} = \mathbf{R}^{NM}$. Since every Gram matrix is positive semidefinite, the entrywise product of P and Q is positive semidefinite.

quadratic case: if the target inequality (b) can be obtained by linear aggregation of the inequalities (a) and a trivial—identically true—inequality, then the implication in question is true. Indeed, a linear aggregation of the inequalities (a) is an inequality of the type

$$x^T \left(\sum_{i=1}^m \lambda_i A_i \right) x \geq 0$$

with nonnegative weights λ_i, and a trivial—identically true—homogeneous quadratic inequality is of the form

$$x^T Q x \geq 0$$

with $Q \succeq 0$. The fact that (b) can be obtained from (a) and a trivial inequality by linear aggregation means that B can be represented as $B = \sum_{i=1}^m \lambda_i A_i + Q$ with $\lambda_i \geq 0$, $Q \succeq 0$, or, which is the same, if $B \succeq \sum_{i=1}^m \lambda_i A_i$ for certain nonnegative λ_i. If this is the case, then (4.3.21) is trivially true. We have arrived at the following simple proposition.

Proposition 4.3.4. *Assume that there exist nonnegative λ_i such that $B \succeq \sum_i \lambda_i A_i$. Then the implication* (4.3.21) *is true.*

Proposition 4.3.4 is no more than a sufficient condition for the implication (4.3.21) to be true, and in general this condition is not necessary. There is, however, an extremely fruitful particular case when the condition is both necessary and sufficient—this is the case of $m = 1$, i.e., a single quadratic inequality in the premise of (4.3.21).

Theorem 4.3.3. $\mathcal{S}$-lemma. *Let A, B be symmetric $n \times n$ matrices, and assume that the quadratic inequality*

$$x^T A x \geq 0 \tag{A}$$

is strictly feasible: there exists $\bar{x}$ such that $\bar{x}^T A \bar{x} > 0$. Then the quadratic inequality

$$x^T B x \geq 0 \tag{B}$$

is a consequence of (A) *if and only if it is a linear consequence of* (A)*, i.e., if and only if there exists a nonnegative λ such that*

$$B \succeq \lambda A.$$

We are about to present an intelligent proof of the $\mathcal{S}$-lemma based on the ideas of semidefinite relaxation. (For a straightforward proof, see the exercises to Lecture 4.)

In view of Proposition 4.3.4, all we need is to prove the "only if" part of the $\mathcal{S}$-lemma, i.e., to demonstrate that if the optimization problem

$$\min_x \left\{ x^T B x : x^T A x \geq 0 \right\}$$

is strictly feasible and its optimal value is ≥ 0, then $B \succeq \lambda A$ for certain $\lambda \geq 0$. By homogeneity reasons, it suffices to prove exactly the same statement for the optimization problem

$$\min_x \left\{ x^T B x : x^T A x \geq 0, x^T x = n \right\}. \tag{P}$$

The standard semidefinite relaxation of (P) is the problem

$$\min_X \{\mathrm{Tr}(BX) : \mathrm{Tr}(AX) \geq 0, \mathrm{Tr}(X) = n, X \succeq 0\}. \tag{P'}$$

If we could show that when passing from the original problem (P) to the relaxed problem (P′) the optimal value (which was nonnegative for (P)) remains nonnegative, we would be done. Indeed, observe that (P′) is clearly bounded below (its feasible set is compact!) and is strictly feasible (which is an immediate consequence of the strict feasibility of (A)). Thus, by the conic duality theorem the problem dual to (P′) is solvable with the same optimal value (let it be called $n\theta^*$) as the one in (P′). The dual problem is

$$\max_{\mu,\lambda} \{n\mu : \lambda A + \mu I \preceq B,\ \lambda \geq 0\},$$

and the fact that its optimal value is $n\theta^*$ means that there exists a nonnegative λ such that

$$B \succeq \lambda A + n\theta^* I.$$

If we knew that the optimal value $n\theta^*$ in (P′) is nonnegative, we would conclude that $B \succeq \lambda A$ for certain nonnegative λ, which is exactly what we are aiming at. Thus, all we need is to prove that under the premise of the $\mathcal{S}$-lemma the optimal value in (P′) is nonnegative, and here is the proof.

Proof. Observe first that problem (P′) is feasible with a compact feasible set and thus is solvable. Let X^* be an optimal solution to the problem. Since $X^* \succeq 0$, there exists a matrix D such that $X^* = DD^T$. Note that we have

$$\begin{array}{c} 0 \leq \mathrm{Tr}(AX^*) = \mathrm{Tr}(ADD^T) = \mathrm{Tr}(D^T AD), \\ n\theta^* = \mathrm{Tr}(BX^*) = \mathrm{Tr}(BDD^T) = \mathrm{Tr}(D^T BD), \\ n = \mathrm{Tr}(X^*) = \mathrm{Tr}(DD^T) = \mathrm{Tr}(D^T D). \end{array} \tag{*}$$

It remains to use the following observation:

> (!) *Let P, Q be symmetric matrices such that* $\mathrm{Tr}(P) \geq 0$ *and* $\mathrm{Tr}(Q) < 0$. *Then there exists a vector e such that $e^T Pe \geq 0$ and $e^T Qe < 0$.*

Indeed, let us believe that (!) is valid, and let us prove that $\theta^* \geq 0$. Assume, on the contrary, that $\theta^* < 0$. Setting $P = D^T BD$ and $Q = D^T AD$ and taking into account (*), we see that the matrices P, Q satisfy the premise in (!), whence, by (!), there exists a vector e such that $0 \leq e^T Pe = [De]^T A[De]$ and $0 > e^T Qe = [De]^T B[De]$; but this contradicts the premise of the $\mathcal{S}$-lemma.

It remains to prove (!). Given P and Q as in (!), note that Q, as every symmetric matrix, admits a representation

$$Q = U^T \Lambda U$$

with an orthonormal U and a diagonal Λ. Note that $\theta \equiv \mathrm{Tr}(\Lambda) = \mathrm{Tr}(Q) < 0$. Now let ξ be a random n-dimensional vector with independent entries taking values ± 1 with probabilities $1/2$. We have

$$[U^T\xi]^T Q[U^T\xi] = [U^T\xi]^T U^T \Lambda U[U^T\xi] = \xi^T \Lambda \xi = \mathrm{Tr}(\Lambda) = \theta \quad \forall \xi,$$

while

$$[U^T\xi]^T P[U^T\xi] = \xi^T[UPU^T]\xi,$$

and the expectation of the latter quantity over ξ is clearly $\mathrm{Tr}(UPU^T) = \mathrm{Tr}(P) \geq 0$. Since the expectation is nonnegative, there is at least one realization $\bar{\xi}$ of our random vector ξ such that

$$0 \leq [U^T\bar{\xi}]^T P[U^T\bar{\xi}].$$

We see that the vector $e = U^T\bar{\xi}$ is a required one: $e^T Qe = \theta < 0$ and $e^T Pe \geq 0$. □

4.4 Applications II: Stability analysis

Semidefinite programming is a natural language with which to pose and process numerous engineering problems associated with stability. Let us look at several examples.

4.4.1 Dynamic stability in mechanics

Free motions of many linearly elastic mechanical systems, i.e., their behavior in absence of external loads, are governed by systems of differential equations of the type

$$M\frac{d^2}{dt^2}x(t) = -Ax(t), \tag{N}$$

where $x(t) \in \mathbf{R}^n$ is the state vector of the system at time t, M is the (generalized) mass matrix, and A is the stiffness matrix of the system. Basically, (N) is the Newton law for a system with the potential energy $\frac{1}{2}x^T Ax$.

As a simple example, consider a system of k points of masses $\mu_1, \ldots, \mu_k$ linked by springs with given elasticity coefficients. Here x is the vector of the displacements $x_i \in \mathbf{R}^d$ of the points from their equilibrium positions e_i ($d = 1/2/3$ is the dimension of the model). The Newton equations become

$$\mu_i\frac{d^2}{dt^2}x_i(t) = -\sum_{j\neq i}\nu_{ij}(e_i - e_j)(e_i - e_j)^T(x_i - x_j), i = 1, \ldots, k,$$

where ν_{ij} are given by

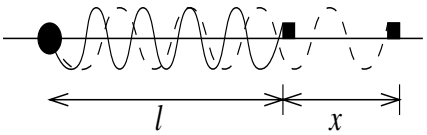
$$\nu_{ij} = \frac{\kappa_{ij}}{\|e_i - e_j\|_2^3},$$

where $\kappa_{ij} > 0$ are the elasticity coefficients of the springs. The resulting system is of the form (N) with a diagonal matrix M and a positive semidefinite symmetric matrix A. The well-known simplest system of this type is a *pendulum* (a single point able to slide along a given axis and linked by a spring to a fixed point on the axis), shown in Fig. 4.4.

Another example is given by a truss (see section 1.3.5). Here A is the bar-stiffness matrix $\sum_i t_i b_i b_i^T$, and the mass matrix is

$$M = \sum_i t_i\beta_i\beta_i^T, \quad \beta_i = \sqrt{\frac{\mu}{\kappa}}l_i b_i,$$

where μ is the material density, κ is the Young modulus, and l_i is the length of bar i.



$\frac{d^2}{dt^2}x(t) = -\nu x(t), \quad \nu = \frac{\kappa}{l}.$

Figure 4.4. *A pendulum.*

Note that in the above examples both the mass matrix M and the stiffness matrix A are symmetric positive semidefinite; in nondegenerate cases they are even positive definite, and this is what we assume from now on. Under this assumption, we can pass in (N) from the variables $x(t)$ to the variables $y(t) = M^{1/2}x(t)$; with respect to these variables the system becomes

$$\frac{d^2}{dt^2}y(t) = -\hat{A}y(t), \quad \hat{A} = M^{-1/2}AM^{-1/2}. \tag{N$'$}$$

It is well known that the space of solutions of system (N$'$) (where $\hat{A}$ is symmetric positive definite) is spanned by fundamental (perhaps complex-valued) solutions of the form $\exp\{\mu t\}f$. A nontrivial (with $f \neq 0$) function of this type is a solution to (N$'$) if and only if

$$(\mu^2 I + \hat{A})f = 0,$$

so that the allowed values of μ^2 are the minus eigenvalues of the matrix $\hat{A}$, and f's are the corresponding eigenvectors of $\hat{A}$. Since the matrix $\hat{A}$ is symmetric positive definite, the only allowed values of μ are purely imaginary, with the imaginary parts $\pm\sqrt{\lambda_j(\hat{A})}$. Recalling that the eigenvalues and eigenvectors of $\hat{A}$ are exactly the eigenvalues and eigenvectors of the pencil $[M, A]$, we come to the following result:

(C) *In the case of positive definite symmetric M, A, the solutions to* (N)*—the free motions of the corresponding mechanical system $\mathcal{S}$—are of the form*

$$x(t) = \sum_{j=1}^{n}[a_j \cos(\omega_j t) + b_j \sin(\omega_j t)]e_j,$$

where a_j, b_j are free real parameters, e_j are the eigenvectors of the pencil $[M, A]$,

$$(\lambda_j M - A)e_j = 0,$$

and $\omega_j = \sqrt{\lambda_j}$. Thus, the free motions of the system $\mathcal{S}$ are mixtures of harmonic oscillations along the eigenvectors of the pencil $[M, A]$, and the frequencies of the oscillations (the eigenfrequencies of the system) are the square roots of the corresponding eigenvalues of the pencil.

From the engineering viewpoint, the dynamic behavior of mechanical constructions such as buildings, electricity masts, and bridges is the better the larger the eigenfrequencies

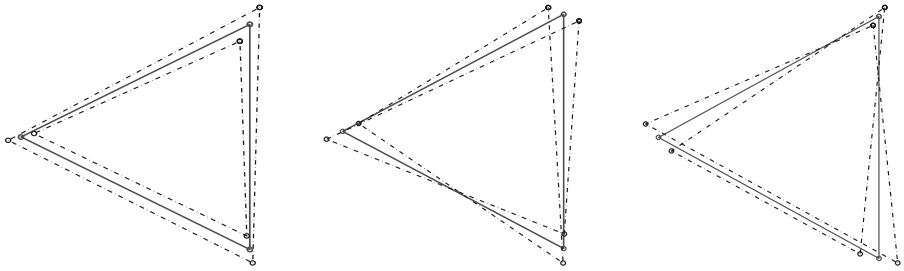


Figure 4.5. *Nontrivial modes of a spring triangle (three unit masses linked by springs). Shown are three eigenmotions (modes) of a spring triangle with nonzero frequencies. Three instant positions of the oscillating triangle are depicted. There are three more eigenmotions with zero frequency, corresponding to shifts and rotation of the triangle.*

of the system.[25] This is why a typical design requirement in mechanical engineering is a lower bound

$$\lambda_{\min}(A : M) \geq \lambda_* \quad [\lambda_* > 0] \tag{4.4.22}$$

on the smallest eigenvalue $\lambda_{\min}(A : M)$ of the pencil $[M, A]$ made up of the mass and the stiffness matrices of the would-be system. In the case of positive definite symmetric mass matrices, (4.4.22) is equivalent to the matrix inequality

$$A - \lambda_* M \succeq 0 \tag{4.4.23}$$

(why?). If M and A are affine functions of the design variables (as is the case in, e.g., truss design), the matrix inequality (4.4.23) is a linear matrix inequality in the design variables, and therefore it can be processed via the machinery of semidefinite programming. For example, when adding to the TTD problem (Lecture 3) a lower bound on the minimum eigenfrequency of the truss to be designed, we end up with a semidefinite program. Moreover, in the cases when A is affine in the design variables, and M is constant, (4.4.23) is an LMI on the design variables and λ_*, and we may play with λ_*, e.g., solve a problem of the type "given the mass matrix of the system to be designed and a number of (SDr) constraints on the design variables, build a system with the largest possible minimum eigenfrequency."

4.4.2 Lyapunov stability analysis and synthesis

The next topic, Lyapunov stability analysis, was touched on in Lecture 2, where it served as an important example of a nonpolyhedral conic problem. Consider a time-varying uncertain linear dynamic system

$$\frac{d}{dt}x(t) = A(t)x(t), \;\; x(0) = x_0. \tag{ULS}$$

[25] Think about a building and an earthquake or about sea waves and a light house: in this case the external load acting at the system is time-varying and can be represented as a sum of harmonic oscillations of different (and low) frequencies. If some of these frequencies are close to the eigenfrequencies of the system, the system can be crushed by resonance. To avoid this risk, one wants to move the eigenfrequencies of the system away from 0 as far as possible.

Here $x(t) \in \mathbf{R}^n$ represents the state of a plant at time t, the initial state is x_0, and $A(t)$ is a time-varying $n \times n$ matrix. We assume that the system is uncertain in the sense that we have no idea of what is x_0, and all we know about $A(t)$ is that this matrix, at any time t, belongs to a given uncertainty set $\mathcal{U}$. Thus, (ULS) represents a wide family of linear dynamic systems rather than a single system. It makes sense to call a trajectory of the uncertain linear system (ULS) every function $x(t)$ that is an actual trajectory of a system from the family, i.e., is such that

$$\frac{d}{dt}x(t) = A(t)x(t)$$

$\forall t \geq 0$ and certain matrix-valued function $A(t)$ taking all its values in $\mathcal{U}$.

Note that we can model a nonlinear dynamic system

$$\frac{d}{dt}x(t) = f(t, x(t)) \quad [x \in \mathbf{R}^n] \tag{NLS}$$

with a given right-hand side $f(t, x)$ and a given equilibrium $x(t) \equiv 0$ (i.e., $f(t, 0) = 0,\ t \geq 0$) as an uncertain linear system. Indeed, let us define the set $\mathcal{U}_f$ as the closed convex hull of the set of $n \times n$ matrices $\left\{\frac{\partial}{\partial x} f(t, x) \mid t \geq 0, x \in \mathbf{R}^n\right\}$. Then for every point $x \in \mathbf{R}^n$ we have

$$f(t, x) = f(t, 0) + \int_0^1 \left[\tfrac{\partial}{\partial x} f(t, sx)\right] x\, ds = A_x(t)x,$$
$$A_x(t) = \int_0^1 \tfrac{\partial}{\partial x} f(t, sx)\, ds \in \mathcal{U}.$$

We see that every trajectory of the original nonlinear system (NLS) is also a trajectory of the uncertain linear system (ULS) associated with the uncertainty set $\mathcal{U} = \mathcal{U}_f$ (this trick is called global linearization). Of course, the set of trajectories of the resulting uncertain linear system can be much wider than the set of trajectories of (NLS); however, all good news about the uncertain system (like "all trajectories of (ULS) share such and such property") are automatically valid for the trajectories of the nonlinear system of interest (NLS), and only bad news about (ULS) ("such and such property is not shared by some trajectories of (ULS)") may say nothing about the system of interest (NLS).

The basic question about a dynamic system is the one of its stability. For (ULS), this question is as follows:

(?) *Is it true that* (ULS) *is stable, i.e., that*

$$x(t) \to 0 \text{ as } t \to \infty,$$

for every trajectory of the system?

A sufficient condition for the stability of (ULS) is the existence of a *quadratic Lyapunov function*, i.e., a quadratic form $\mathcal{L}(x) = x^T Xx$ with symmetric positive definite matrix X such that

$$\frac{d}{dt}\mathcal{L}(x(t)) \leq -\alpha \mathcal{L}(x(t)) \tag{4.4.24}$$

for certain $\alpha > 0$ and all trajectories of (ULS).

Lemma 4.4.1. Quadratic stability certificate. *Assume* (ULS) *admits a quadratic Lyapunov function $\mathcal{L}$. Then* (ULS) *is stable.*

Proof. If (4.4.24) is valid with some $\alpha > 0$ for all trajectories of (ULS), then, by integrating this differential inequality, we get

$$\mathcal{L}(x(t)) \leq \exp\{-\alpha \mathcal{L}(x(0))\} \to 0 \text{ as } t \to \infty.$$

Since $\mathcal{L}(\cdot)$ is a positive definite quadratic form, $\mathcal{L}(x(t)) \to 0$ implies that $x(t) \to 0$. □

Of course, the statement of Lemma 4.4.1 also holds for nonquadratic Lyapunov functions. All we need is (4.4.24) plus the assumption that $\mathcal{L}(x)$ is smooth and nonnegative and is bounded away from 0 outside every neighborhood of the origin. The advantage of a quadratic Lyapunov function is that we more or less know how to find such a function, if it exists.

Proposition 4.4.1. Existence of quadratic stability certificate. *Let $\mathcal{U}$ be the uncertainty set associated with uncertain linear system* (ULS). *The system admits quadratic Lyapunov function if and only if the optimal value of the semi-infinite*[26] *semidefinite program*

$$\begin{array}{ll} \textit{minimize} & s \\ \textit{s.t.} & \\ & sI_n - A^T X - XA \succeq 0, \quad \forall A \in \mathcal{U}, \\ & X \succeq I_n, \end{array} \tag{Ly}$$

with the design variables $s \in \mathbf{R}$ and $X \in \mathbf{S}^n$, is negative. Moreover, every feasible solution to the problem with negative value of the objective provides a quadratic Lyapunov stability certificate for (ULS).

Proof. The derivative $\frac{d}{dt}\left[x^T(t)Xx(t)\right]$ of the quadratic function $x^T Xx$ along a trajectory of (ULS) is equal to

$$\left[\frac{d}{dt}x(t)\right]^T Xx(t) + x^T(t)X\left[\frac{d}{dt}x(t)\right] = x^T(t)[A^T(t)X + XA(t)]x(t).$$

If $x^T Xx$ is a Lyapunov function, then the resulting quantity must be at most $-\alpha x^T(t)Xx(t)$, i.e., we should have

$$x^T(t)\left[-\alpha X - A^T(t)X - XA(t)\right]x(t) \geq 0$$

for every possible value of $A(t)$ at any time t and for every possible value $x(t)$ of a trajectory of the system at this time. Since possible values of $x(t)$ fill the entire $\mathbf{R}^n$ and possible values of $A(t)$ fill the entire $\mathcal{U}$, we conclude that

$$-\alpha X - A^T X - XA \succeq 0 \quad \forall A \in \mathcal{U}.$$

[26] That is, with infinitely many LMI constraints.

By definition of a quadratic Lyapunov function, $X \succ 0$ and $\alpha > 0$; by normalization (dividing both X and α by the smallest eigenvalue of X), we get a pair $\hat{s} > 0$, $\hat{X} \geq I_n$ such that

$$-\hat{s}\hat{X} - A^T\hat{X} - \hat{X}A \succeq 0 \quad \forall A \in \mathcal{U}.$$

Since $\hat{X} \succeq I_n$, we conclude that

$$-\hat{s}I_n - A^T\hat{X} - \hat{X}A \succeq -\hat{s}\hat{X} - A^T\hat{X} - \hat{X}A \succeq 0 \quad \forall A \in \mathcal{U};$$

thus, $(s = -\hat{s}, \hat{X})$ is a feasible solution to (Ly) with negative value of the objective. We have demonstrated that if (ULS) admits a quadratic Lyapunov function, then (Ly) has a feasible solution with negative value of the objective. Reversing the reasoning, we can verify the inverse implication. □

Lyapunov stability analysis

We have just seen that to certify the stability of an uncertain linear system it suffices to provide a feasible solution to (Ly) with a negative value of the objective. It should be stressed that the existence of such a solution is only a sufficient condition for stability, and if the condition is not satisfied (i.e., if the optimal value in (Ly) is nonnegative), then all we can say is that the stability of (ULS) cannot be certified by a quadratic Lyapunov function, although (ULS) still may be stable.[27] In this sense, the stability analysis based on quadratic Lyapunov functions is conservative. This drawback, however, is in a sense compensated by the fact that this kind of stability analysis is implementable: in many cases we can efficiently solve (Ly), thus getting a quadratic stability certificate, provided that it exists, in a constructive way. Let us look at two such cases.

Polytopic uncertainty set. The first tractable case of (Ly) is when $\mathcal{U}$ is a polytope given as a convex hull of finitely many points:

$$\mathcal{U} = \mathrm{Conv}\{A_1, \ldots, A_N\}.$$

In this case (Ly) is equivalent to the semidefinite program

$$\min_{s,X} \left\{ s : sI_n - A_i^T X - XA_i \succeq 0,\ i = 1, \ldots, N;\ X \succeq I_n \right\}. \tag{4.4.25}$$

(Why?)

The assumption that $\mathcal{U}$ is a polytope given as a convex hull of a finite set is crucial for a possibility to get a computationally tractable equivalent reformulation of (Ly). If $\mathcal{U}$ is, say, a polytope given by a list of linear inequalities (e.g., all we know about the entries

[27] The only case in which the existence of a quadratic Lyapunov function is a criterion (i.e., a necessary and sufficient condition) for stability is the simplest case of a certain time-invariant linear system $\frac{d}{dt}x(t) = Ax(t)$ ($\mathcal{U} = \{A\}$). This is the case that led Lyapunov to the general concept of what is now called a Lyapunov function and what is the basic approach to establishing convergence of different time-dependent processes to their equilibria. Note also that in the case of time-invariant linear system there exists a straightforward stability criterion—all eigenvalues of A should have negative real parts. The advantage of the Lyapunov approach is that it can be extended to more general situations, which is not the case for the eigenvalue criterion.

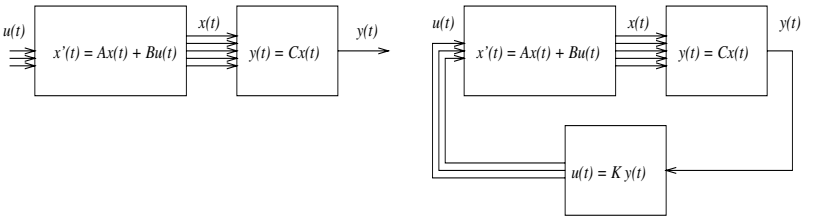


Figure 4.6. *Open-loop (left) and closed-loop (right) controlled systems.*

of $A(t)$ is that they reside in certain intervals; this case is called interval uncertainty), (Ly) may become as hard as a problem can be: it may happen that just to check whether a given pair (s, X) is feasible for (Ly) is already a computationally intractable problem. The same difficulties may occur when $\mathcal{U}$ is a general-type ellipsoid in the space of $n \times n$ matrices. There exists, however, a specific type of uncertainty ellipsoids $\mathcal{U}$ for which (Ly) is easy. Let us look at this case.

Norm-bounded perturbations. In numerous applications the $n \times n$ matrices A forming the uncertainty set $\mathcal{U}$ are obtained from a fixed nominal matrix A_* by adding perturbations of the form $B\Delta C$, where $B \in \mathbf{M}^{n,k}$ and $C \in \mathbf{M}^{l,n}$ are given rectangular matrices and $\Delta \in \mathbf{M}^{k,l}$ is the perturbation varying in a simple set $\mathcal{D}$:

$$\mathcal{U} = \{A = A_* + B\Delta C \mid \Delta \in \mathcal{D} \subset \mathbf{M}^{k,l}\} \quad \left[B \in \mathbf{M}^{n,k}, 0 \neq C \in \mathbf{M}^{l,n}\right]. \tag{4.4.26}$$

As an instructive example, consider a controlled linear time-invariant dynamic system

$$\begin{array}{rcl} \frac{d}{dt}x(t) & = & Ax(t) + Bu(t), \\ y(t) & = & Cx(t) \end{array} \tag{4.4.27}$$

(x is the state, u is the control, and y is the output we can observe) closed by a feedback

$$u(t) = Ky(t);$$

see Figure 4.6. The resulting closed-loop system is given by

$$\frac{d}{dt}x(t) = \hat{A}x(t), \quad \hat{A} = A + BKC. \tag{4.4.28}$$

Now assume that A, B, and C are constant and known, but the feedback K is drifting around certain nominal feedback K_*: $K = K_* + \Delta$. As a result, the matrix $\hat{A}$ of the closed-loop system also drifts around its nominal value $A_* = A + BK_*C$, and the perturbations in $\hat{A}$ are exactly of the form $B\Delta C$.

Note that we could get essentially the same kind of drift in $\hat{A}$ assuming, instead of perturbations in the feedback matrix K, perturbations $C = C_* + \Delta$ in the observer (or similar disturbances in the actuator B).

Now assume that the input perturbations Δ are of spectral norm $|\Delta|$ not exceeding a given ρ (norm-bounded perturbations):

$$\mathcal{D} = \{\Delta \in \mathbf{M}^{k,l} \mid |\Delta| \leq \rho\}. \tag{4.4.29}$$

Proposition 4.4.2.[28] *In the case of uncertainty set* (4.4.26), (4.4.29)*, the semi-infinite semidefinite program* (Ly) *is equivalent to the usual semidefinite program*

$$\begin{array}{lll} \textit{minimize} & s & \\ \textit{s.t.} & \left(\begin{array}{c|c} sI_n - A_*^T X - XA_* - \lambda C^T C & -\rho XB \\ \hline -\rho B^T X & \lambda I_k \end{array}\right) & \succeq 0, \\ & X & \succeq I_n \end{array} \tag{4.4.30}$$

in the design variables s, λ, X.

When shrinking the set of perturbations (4.4.29) *to the ellipsoid*[29]

$$\mathcal{E} = \{\Delta \in \mathbf{M}^{k,l} \mid \|\Delta\|_2 \equiv \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{l}\Delta_{ij}^2} \leq \rho\}, \tag{4.4.31}$$

we do not vary (Ly)*: in the case of the uncertainty set* (4.4.26), (Ly) *is still equivalent to* (4.4.30).

Proof. It suffices to verify the following general statement.

Lemma 4.4.2. *Consider the matrix inequality*

$$Y - Q^T\Delta^T P^T Z^T R - R^T Z P \Delta Q \succeq 0, \tag{4.4.32}$$

where Y *is symmetric* $n \times n$ *matrix,* Δ *is a* $k \times l$ *matrix, and* P, Q, Z, R *are rectangular matrices of appropriate sizes (i.e.,* $q \times k$, $l \times n$, $p \times q$, *and* $p \times n$, *respectively). Given* Y, P, Q, Z, R *with* $Q \neq 0$ *(this is the only nontrivial case), this matrix inequality is satisfied for all* Δ *with* $|\Delta| \leq \rho$ *if and only if it is satisfied for all* Δ *with* $\|\Delta\|_2 \leq \rho$*, and this is the case if and only if*

$$\left(\begin{array}{c|c} Y - \lambda Q^T Q & -\rho R^T Z P \\ \hline -\rho P^T Z^T R & \lambda I_k \end{array}\right) \succeq 0$$

for a properly chosen real λ.

The statement of Proposition 4.3.17 is just a particular case of Lemma 4.4.2. For example, in the case of uncertainty set (4.4.26), (4.4.29), a pair (s, X) is a feasible solution to (Ly) if and only if $X \succeq I_n$ and (4.4.32) is valid, whenever $|\Delta| \leq \rho$, for $Y = sI_n - A_*^T X - XA_*$, $P = B$, $Q = C$, $Z = X$, $R = I_n$. Lemma 4.4.2 provides us with an LMI reformulation of the latter property, and this LMI is exactly what we see in the statement of Proposition 4.3.17.

Proof of Lemma 4.4.2. The inequality (4.4.32) is valid for all Δ with $|\Delta| \leq \rho$ (let us call this property of (Y, P, Q, Z, R) "Property 1") if and only if

$$2[\xi^T R^T Z P]\Delta[Q\xi] \leq \xi^T Y \xi \quad \forall \xi \in \mathbf{R}^n \quad \forall(\Delta : |\Delta| \leq \rho)$$

[28] S. Boyd et al., *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, 1994.

[29] This indeed is a "shrinkage": $|\Delta| \leq \|\Delta\|_2$ for every matrix Δ (prove it!).

or, which is the same, if and only if

$$\max_{\Delta:|\Delta|\le\rho} 2\left[[P^TZ^TR\xi]^T\Delta[Q\xi]\right] \le \xi^TY\xi \quad \forall\xi\in\mathbf{R}^n. \tag{Property 2}$$

The maximum over Δ, $|\Delta|\le\rho$, of the quantity $\eta^T\Delta\zeta$, clearly is equal to ρ times the product of the Euclidean norms of the vectors η and ζ (why?). Thus, Property 2 is equivalent to

$$\xi^TY\xi - 2\rho\|Q\xi\|_2\|P^TZ^TR\xi\|_2 \ge 0 \quad \forall\xi\in\mathbf{R}^n. \tag{Property 3}$$

Now the trick: Property 3 is clearly equivalent to the following.

> Property 4. *Every pair $\zeta = (\xi,\eta) \in \mathbf{R}^n\times\mathbf{R}^k$ that satisfies the quadratic inequality*
>
> $$\xi^TQ^TQ\xi - \eta^T\eta \ge 0 \tag{I}$$
>
> *satisfies also the quadratic inequality*
>
> $$\xi^TY\xi - 2\rho\eta^TP^TZ^TR\xi \ge 0. \tag{II}$$

Indeed, for a fixed ξ the minimum over η satisfying (I) of the left-hand side in (II) is nothing but the left-hand side in Property 3.

It remains to use the $\mathcal{S}$-lemma. Property 4 says that the quadratic inequality (II) with variables ξ, η is a consequence of (I). By the $\mathcal{S}$-lemma (recall that $Q\ne 0$, so that (I) is strictly feasible!), this is equivalent to the existence of a nonnegative λ such that

$$\begin{pmatrix} Y & -\rho R^TZP \\ -\rho P^TZ^TR & \end{pmatrix} - \lambda\begin{pmatrix} Q^TQ & \\ & -I_k\end{pmatrix} \succeq 0,$$

which is exactly the statement of Lemma 4.4.2 for the case of $|\Delta|\le\rho$. The case of perturbations with $\|\Delta\|_2\le\rho$ is completely similar, since the equivalence between Properties 2 and 3 is valid independent of which norm of Δ – $|\cdot|$ or $\|\cdot\|_2$ – is used.

Lyapunov stability synthesis

We have seen that under reasonable assumptions on the underlying uncertainty set the question of whether a given uncertain linear system (ULS) admits a quadratic Lyapunov function can be reduced to a semidefinite program. Now let us switch from the analysis question, whether a stability of an uncertain linear system may be certified by a quadratic Lyapunov function, to the synthesis question, which is as follows. Assume that we are given an uncertain open loop controlled system

$$\begin{array}{rcl} \frac{d}{dt}x(t) & = & A(t)x(t) + B(t)u(t), \\ y(t) & = & C(t)x(t). \end{array} \tag{UOS}$$

All we know about the collection $(A(t), B(t), C(t))$ of time-varying $n\times n$ matrix $A(t)$, $n\times k$ matrix $B(t)$, and $l\times n$ matrix $C(t)$ is that this collection, at every time t, belongs to

a given uncertainty set $\mathcal{U}$. The question is whether we can equip our uncertain open-loop system (UOS) with a linear feedback

$$u(t) = Ky(t)$$

in such a way that the resulting uncertain closed-loop system

$$\frac{d}{dt}x(t) = [A(t) + B(t)KC(t)]\,x(t) \tag{UCS}$$

will be stable and, moreover, such that its stability can be certified by a quadratic Lyapunov function. In other words, now we are simultaneously looking for a stabilizing controller and a quadratic Lyapunov certificate of its stabilizing ability.

With the global linearization trick we may use the results on uncertain controlled linear systems to build stabilizing linear controllers for nonlinear controlled systems

$$\begin{array}{rcl} \frac{d}{dt}x(t) & = & f(t, x(t), u(t)), \\ y(t) & = & g(t, x(t)). \end{array}$$

Assuming $f(t, 0, 0) = 0$, $g(t, 0) = 0$ and denoting by $\mathcal{U}$ the closed convex hull of the set

$$\left\{ \left(\frac{\partial}{\partial x} f(t, x, u),\, \frac{\partial}{\partial u} f(t, x, u),\, \frac{\partial}{\partial x} g(t, x) \right) \,\middle|\, t \geq 0, x \in \mathbf{R}^n, u \in \mathbf{R}^k \right\},$$

we see that every trajectory of the original nonlinear system is a trajectory of the uncertain linear system (UOS) associated with the set $\mathcal{U}$. Consequently, if we are able to find a stabilizing controller for (UOS) and certify its stabilizing property by a quadratic Lyapunov function, then the resulting controller Lyapunov function will stabilize the nonlinear system and will certify the stability of the closed-loop system, respectively.

Exactly the same reasoning as in the previous section leads us to the following.

Proposition 4.4.3. *Let $\mathcal{U}$ be the uncertainty set associated with an uncertain open-loop controlled system* (UOS). *The system admits a stabilizing controller along with a quadratic Lyapunov stability certificate for the resulting closed-loop system if and only if the optimal value in the optimization problem*

$$\begin{array}{ll} \textit{minimize} & s \\ \textit{s.t.} & \\ & \begin{array}{rcll} [A + BKC]^T X + X[A + BCK] & \preceq & sI_n & \forall (A, B, C) \in \mathcal{U}, \\ X & \succeq & I_n, & \end{array} \end{array} \tag{LyS}$$

in design variables s, X, K, is negative. Moreover, every feasible solution to the problem with negative value of the objective provides a stabilizing controller along with a quadratic Lyapunov stability certificate for the resulting closed-loop system.

Bad news about (LyS) is that it is much more difficult to rewrite this problem as a semidefinite program than in the analysis case (i.e., the case of $K = 0$), since (LyS) is a semi-infinite system of nonlinear matrix inequalities. There is, however, an important

particular case where this difficulty can be overcome. This is the case of a feedback via the full state vector—the case when $y(t) = x(t)$ (i.e., $C(t)$ is the unit matrix). In this case, all we need to get a stabilizing controller along with a quadratic Lyapunov certificate of its stabilizing ability is to solve a system of strict matrix inequalities

$$\begin{array}{rcl} [A+BK]^T X + X[A+BK] & \preceq & Z \prec 0 \quad \forall (A,B) \in \mathcal{U}, \\ X & \succ & 0. \end{array} \qquad (*)$$

Indeed, given a solution (X, K, Z) to this system, we always can convert it by normalization of X to a solution of (LyS). Now let us make the change of variables

$$Y = X^{-1}, L = KX^{-1}, W = X^{-1}ZX^{-1} \quad \left[\Leftrightarrow X = Y^{-1}, K = LY^{-1}, Z = Y^{-1}WY^{-1}\right].$$

With respect to the new variables Y, L, K, system (*) becomes

$$\left\{ \begin{array}{rcl} [A+BLY^{-1}]^T Y^{-1} + Y^{-1}[A+BLY^{-1}] & \preceq & Y^{-1}WY^{-1} \prec 0, \\ Y^{-1} & \succ & 0 \end{array} \right.$$

$$\Updownarrow$$

$$\left\{ \begin{array}{rcl} L^T B^T + YA^T + BL + AY & \preceq & W \prec 0 \quad \forall (A,B) \in \mathcal{U}, \\ Y & \succ & 0. \end{array} \right.$$

(We have multiplied all original matrix inequalities from the left and from the right by Y.) What we end up with is a system of strict linear matrix inequalities with respect to our new design variables L, Y, W; the question of whether this system is solvable can be converted to the question of whether the optimal value in a problem of the type (LyS) is negative, and we come to the following.

Proposition 4.4.4. *Consider an uncertain controlled linear system with a full observer:*

$$\begin{array}{rcl} \frac{d}{dt}x(t) & = & A(t)x(t) + B(t)u(t), \\ y(t) & = & x(t), \end{array}$$

and let $\mathcal{U}$ be the corresponding uncertainty set (which now comprises pairs (A, B) of possible values of $(A(t), B(t))$, since $C(t) \equiv I_n$ is certain).

The system can be stabilized by a linear controller

$$u(t) = Ky(t) \quad [\equiv Kx(t)]$$

in such a way that the resulting uncertain closed-loop system

$$\frac{d}{dt}x(t) = [A(t) + B(t)K]x(t)$$

admits a quadratic Lyapunov stability certificate if and only if the optimal value in the optimization problem

$$\begin{array}{ll} \textit{minimize} & s \\ \textit{s.t.} & \\ & \begin{array}{rcl} BL + AY + L^T B^T + YA^T & \preceq & sI_n \quad \forall (A,B) \in \mathcal{U}, \\ Y & \succeq & I, \end{array} \end{array} \qquad (\text{Ly}^*)$$

in the design variables $s \in \mathbf{R}$, $Y \in \mathbf{S}^n$, $L \in \mathbf{M}^{k,n}$, is negative. Moreover, every feasible solution to (Ly*) *with negative value of the objective provides a stabilizing linear controller along with a related quadratic Lyapunov stability certificate.*

In particular, in the polytopic case,

$$\mathcal{U} = \mathrm{Conv}\{(A_1, B_1), \ldots, (A_N, B_N)\},$$

the quadratic Lyapunov stability synthesis reduces to solving the semidefinite program

$$\min_{s,Y,L} \left\{ s : B_i L + A_i Y + Y A_i^T + L^T B_i^T \preceq s I_n,\ i = 1, \ldots, N; Y \succeq I_n \right\}.$$

4.4.3 Interval stability analysis and synthesis

Consider the problem of Lyapunov stability analysis in the case of interval uncertainty:

$$\mathcal{U} = \mathcal{U}_\rho = \{A \in \mathbf{M}^{n,n} \mid |A_{ij} - A^*_{ij}| \le \rho D_{ij},\ i, j = 1, \ldots, n\}, \tag{4.4.33}$$

where A^* is the nominal matrix, $D \neq 0$ is a matrix with nonnegative entries specifying the scale for perturbations of different entries, and $\rho \ge 0$ is the level of perturbations. How can we certify the stability of the corresponding uncertain dynamic system via a quadratic Lyapunov function? Well, we are speaking about polytopic uncertainty, so that finding a quadratic Lyapunov stability certificate is the same as finding a feasible solution of the semidefinite program (4.4.25) with a negative value of the objective. The difficulty, however, is that the number N of LMI constraints in this problem is the number of vertices of the polytope (4.4.33), i.e., $N = 2^m$, where m is the number of uncertain entries in our interval matrix ($\equiv$ the number of positive entries in D). For 5×5 interval matrices with full uncertainty, $m = 25$, i.e., $N = 2^{25} = 33,554,432$, which is a bit too many; for fully uncertain 10×10 matrices, $N = 2^{100} > 1.2 \times 10^{30}$. Thus, the brute force approach fails already for small matrices affected by interval uncertainty.

In fact, the difficulty we encounter lies in the NP-hardness of the following problem:

> *Given a candidate Lyapunov stability certificate $X \succ 0$ and $\rho > 0$, check whether X indeed certifies stability of all instances of $\mathcal{U}_\rho$, i.e., whether X solves the semi-infinite system of LMIs*
>
> $$A^T X + X A \preceq -I \quad \forall A \in \mathcal{U}_\rho. \tag{4.4.34}$$
>
> *(In fact, we are interested in the system $A^T X + X A \prec 0\ \forall A \in \mathcal{U}_\rho$, but this is a minor difference—the system of interest is homogeneous in X, and therefore every feasible solution of it can be converted to a solution of* (4.4.34) *just by scaling $X \mapsto tX$.)*

The above problem, in turn, is a particular case of the following problem.

> Matrix Cube. *Given matrices $A_0, A_1, \ldots, A_m \in \mathbf{S}^n$ with $A_0 \succeq 0$, find the largest $\rho = R[A_1, \ldots, A_m : A_0]$ such that the set*
>
> $$\mathcal{A}_\rho = \left\{ A = A_0 + \sum_{i=1}^m z_i A_i \mid \|z\|_\infty \le \rho \right\} \tag{4.4.35}$$

—the image of the m-dimensional cube $\{z \in \mathbf{R}^m \mid \|z\|_\infty \le \rho\}$ *under the affine mapping* $z \mapsto A_0 + \sum_{i=1}^m z_i A_i$*—is contained in the semidefinite cone* $\mathbf{S}^n_+$.

This is the problem we will focus on.

The main result. The problem Matrix Cube (MC for short) is NP-hard; this is true also for the feasibility version MC_ρ of MC, where, given a $\rho \ge 0$, we want to verify the inclusion $\mathcal{A}_\rho \subset \mathbf{S}^n_+$. However, we can point out a simple sufficient condition for the validity of the inclusion $\mathcal{A}_\rho \subset \mathbf{S}^n_+$, as follows.

Proposition 4.4.5. *Assume that the system of LMIs*

$$\begin{array}{ll} \text{(a)} & X^i \succeq \rho A_i,\ \ X^i \succeq -\rho A_i,\ \ i = 1, \ldots, m, \\ \text{(b)} & \displaystyle\sum_{i=1}^m X^i \preceq A_0 \end{array} \qquad (\mathrm{S}_\rho)$$

in matrix variables $X^1, \ldots, X^m \in \mathbf{S}^n$ *is solvable. Then* $\mathcal{A}_\rho \subset \mathbf{S}^n_+$.

Proof. Let $X^1, \ldots, X^m$ be a solution of (S_ρ). From (a) it follows that whenever $\|z\|_\infty \le \rho$, we have $X^i \succeq z_i A_i$ for all i, whence by (b)

$$A_0 + \sum_{i=1}^m z_i A_i \succeq A_0 - \sum_i X_i \succeq 0. \qquad \square$$

Our main result is that the sufficient condition for the inclusion $\mathcal{A}_\rho \subset \mathbf{S}^n_+$ stated by Proposition 4.4.5 is not too conservative.

Theorem 4.4.1. *If the system of LMIs* (S_ρ) *is not solvable, then*

$$\mathcal{A}_{\theta(\mu)\rho} \not\subset \mathbf{S}^n_+. \qquad (4.4.36)$$

Here

$$\mu = \max_{1 \le i \le m} \mathrm{Rank}(A_i)$$

(note $i \ge 1$ *in the* max*!), and*

$$\theta(k) \le \frac{\pi\sqrt{k}}{2},\ \ k \ge 1, \quad \theta(2) = \frac{\pi}{2}. \qquad (4.4.37)$$

Proof. Below $\zeta \sim \mathcal{N}(0, I_n)$ means that ζ is a random Gaussian n-dimensional vector with zero mean and the unit covariance matrix, and $p_n(\cdot)$ stands for the density of the corresponding probability distribution:

$$p_n(u) = (2\pi)^{-n/2} \exp\left\{-\frac{u^T u}{2}\right\}, \quad u \in \mathbf{R}^n.$$

Let us set

$$\theta(k) = \frac{1}{\min\left\{\int |\alpha_i u_1^2 + \cdots + \alpha_k u_k^2| p_k(u)du \,\middle|\, \alpha \in \mathbf{R}^k, \|\alpha\|_1 = 1\right\}}. \tag{4.4.38}$$

Observe that $\theta(k)$ is nondecreasing. It suffices to verify that

(i) With the just-defined $\theta(\cdot)$, unsolvability of (S_ρ) does imply (4.4.36).

(ii) $\theta(\cdot)$ satisfies (4.4.37).

Let us prove (i).

1. Assume that (S_ρ) has no solutions. This means that the optimal value of the semidefinite problem

$$\min_{t,\{X^i\}} \left\{ t \,\middle|\, \begin{array}{c} X^i \succeq \rho A_i,\ X^i \succeq -\rho A_i,\ i = 1, \ldots, m, \\ \sum_{i=1}^m X_i \preceq A_0 + tI \end{array} \right\} \tag{4.4.39}$$

is positive. Since the problem is strictly feasible, its optimal value is positive if and only if the optimal value of the dual problem

$$\max_{W,\{U^i,V^i\}} \left\{ \rho \sum_{i=1}^m \mathrm{Tr}([U^i - V^i]A_i) - \mathrm{Tr}(WA_0) \,\middle|\, \begin{array}{c} U^i + V^i = W,\ i = 1, \ldots, m, \\ \mathrm{Tr}(W) = 1, \\ U^i, V^i, W \succeq 0 \end{array} \right\}$$

is positive. Thus, there exist matrices U^i, V^i, W such that

$$\begin{array}{ll} \text{(a)} & U^i, V^i, W \succeq 0, \\ \text{(b)} & U^i + V^i = W,\ i = 1, 2, \ldots m, \\ \text{(c)} & \rho \sum_{i=1}^m \mathrm{Tr}([U^i - V^i]A_i) > \mathrm{Tr}(WA_0). \end{array} \tag{4.4.40}$$

2. Now let us use the following simple lemma.

Lemma 4.4.3. *Let W, $A \in \mathbf{S}^n$, $W \succeq 0$. Then*

$$\max_{U,V \succeq 0, U+V=W} \mathrm{Tr}([U - V]A) = \max_{X = X^T : \|\lambda(X)\|_\infty \le 1} \mathrm{Tr}(XW^{1/2}AW^{1/2}) = \|\lambda(W^{1/2}AW^{1/2})\|_1. \tag{4.4.41}$$

Proof. We clearly have

$$U, V \succeq 0, U + V = W \Leftrightarrow U = W^{1/2}PW^{1/2}, V = W^{1/2}QW^{1/2}, P, Q \succeq 0, P + Q = I,$$

whence

$$\max_{U,V : U,V \succeq 0, U+V=W} \mathrm{Tr}([U - V]A) = \max_{P,Q : P,Q \succeq 0, P+Q=I} \mathrm{Tr}([P - Q]W^{1/2}AW^{1/2}).$$

When P, Q are linked by the relation $P+Q=I$ and vary in $\{P\succeq 0,\ Q\succeq 0\}$, the matrix $X=P-Q$ runs through the entire interval $\{-I\preceq X\preceq I\}$ (why?); we have proved the first equality in (4.4.41). When proving the second equality, we may assume w.l.o.g. that the matrix $W^{1/2}AW^{1/2}$ is diagonal, so that $\mathrm{Tr}(XW^{1/2}AW^{1/2})=\lambda^T(W^{1/2}AW^{1/2})\mathrm{Dg}(X)$, where $\mathrm{Dg}(X)$ is the diagonal of X. When X runs through the interval $\{-I\preceq X\preceq I\}$, the diagonal of X runs through the entire unit cube $\{\|x\|_\infty\le 1\}$, which immediately yields the second equality in (4.4.41). □

By Lemma 4.4.3, from (4.4.40) it follows that there exists $W\succeq 0$ such that

$$\rho\sum_{i=1}^m\|\lambda(W^{1/2}A_iW^{1/2})\|_1>\mathrm{Tr}(W^{1/2}A_0W^{1/2}). \tag{4.4.42}$$

3. Now let us use the following observation.

Lemma 4.4.4. *With $\xi\sim\mathcal{N}(0,I_n)$, for every symmetric $n\times n$ matrix A one has*

$$\begin{array}{ll}(a) & \mathbf{E}\left\{\xi^TA\xi\right\}=\mathrm{Tr}(A),\\ (b) & \mathbf{E}\left\{|\xi^TA\xi|\right\}\ge\frac{1}{\theta(\mathrm{Rank}(A))}\|\lambda(A)\|_1.\end{array} \tag{4.4.43}$$

Here $\mathbf{E}$ *stands for the expectation with respect to the distribution of* ξ.

Proof. (4.4.43)(a) is evident:

$$\mathbf{E}\left\{\xi^TA\xi\right\}=\sum_{i,j=1}^m A_{ij}\mathbf{E}\left\{\xi_i\xi_j\right\}=\mathrm{Tr}(A).$$

To prove (4.4.43)(b), by homogeneity it suffices to consider the case when $\|\lambda(A)\|_1=1$. Further, by rotational invariance of the distribution of ξ we may consider the case when A is diagonal and the first $\mathrm{Rank}(A)$ of diagonal entries of A are the nonzero eigenvalues of the matrix. With this normalization, the required relation immediately follows from the definition of $\theta(\cdot)$. □

4. Now we are ready to prove (i). Let $\xi\sim\mathcal{N}(0,I_n)$. We have

$$\begin{array}{rcll}\mathbf{E}\left\{\rho\theta(\mu)\sum_{i=1}^k|\xi^TW^{1/2}A_iW^{1/2}\xi|\right\} & = & \sum_{i=1}^m\rho\theta(\mu)\mathbf{E}\left\{|\xi^TW^{1/2}A_iW^{1/2}\xi|\right\} & \\ & \ge & \rho\sum_{i=1}^m\|\lambda(W^{1/2}A_iW^{1/2})\|_1 & \\ & & \left[\begin{array}{l}\text{by (4.4.43)(b) due to } \mathrm{Rank}(W^{1/2}A_iW^{1/2})\\ \le \mathrm{Rank}(A_i)\le\mu,\ i\ge 1, \text{ and since } \theta(\cdot)\\ \text{is nondecreasing}\end{array}\right] & \\ & > & \mathrm{Tr}(W^{1/2}A_0W^{1/2}) & \text{(by (4.4.42))}\\ & = & \mathrm{Tr}(\xi^TW^{1/2}A_0W^{1/2}\xi), & \text{(by (4.4.43)(a))],}\end{array}$$

whence

$$\mathbf{E}\left\{\rho\theta(\mu)\sum_{i=1}^{k}|\xi^T W^{1/2}A_i W^{1/2}\xi| - \xi^T W^{1/2}A_0 W^{1/2}\xi\right\} > 0.$$

It follows that there exists $r \in \mathbf{R}^n$ such that

$$\theta(\mu)\rho\sum_{i=1}^{m}|r^T W^{1/2}A_i W^{1/2}r| > r^T W^{1/2}A_0 W^{1/2}r,$$

so that setting $z_i = -\theta(\mu)\rho\,\mathrm{sign}(r^T W^{1/2}A_i W^{1/2}r)$, we get

$$r^T W^{1/2}\left(A_0 + \sum_{i=1}^{m} z_i A_i\right) W^{1/2}r < 0.$$

We see that the matrix $A_0 + \sum_{i=1}^{m} z_i A_i$ is not positive semidefinite, while by construction $\|z\|_\infty \le \theta(\mu)\rho$. Thus, (4.4.36) holds true. Point (i) is proved.

To prove (ii), let $\alpha \in \mathbf{R}^k$ be such that $\|\alpha\|_1 = 1$, and let

$$J = \int |\alpha_1 u_1^2 + \cdots + \alpha_k u_k^2| p_k(u)du.$$

Let $\beta = [\begin{smallmatrix}\alpha\\ -\alpha\end{smallmatrix}]$, and let $\xi \sim \mathcal{N}(0, I_{2k})$. We have

$$\mathbf{E}\left\{\left|\sum_{i=1}^{2k}\beta_i\xi_i^2\right|\right\} \le \mathbf{E}\left\{\left|\sum_{i=1}^{k}\beta_i\xi_i^2\right|\right\} + \mathbf{E}\left\{\left|\sum_{i=1}^{k}\beta_{i+k}\xi_{i+k}^2\right|\right\} = 2J. \tag{4.4.44}$$

On the other hand, let $\eta_i = \frac{1}{\sqrt{2}}(\xi_i - \xi_{k+i})$, $\zeta_i = \frac{1}{\sqrt{2}}(\xi_i + \xi_{k+i})$, $i = 1, \ldots, k$, and let

$$\omega = \begin{pmatrix}\alpha_1\eta_1\\ \vdots\\ \alpha_k\eta_k\end{pmatrix}, \widetilde{\omega} = \begin{pmatrix}|\alpha_1\eta_1|\\ \vdots\\ |\alpha_k\eta_k|\end{pmatrix}, \zeta = \begin{pmatrix}\zeta_1\\ \vdots\\ \zeta_k\end{pmatrix}.$$

Observe that ζ and ω are independent and $\zeta \sim \mathcal{N}(0, I_k)$. We have

$$\mathbf{E}\left\{\left|\sum_{i=1}^{2k}\beta_i\xi_i^2\right|\right\} = 2\mathbf{E}\left\{\left|\sum_{i=1}^{k}\alpha_i\eta_i\zeta_i\right|\right\} = 2\mathbf{E}\left\{|\omega^T\zeta|\right\} = 2\mathbf{E}\left\{\|\omega\|_2\right\}\mathbf{E}\left\{|\zeta_1|\right\},$$

where the concluding equality follows from the fact that $\zeta \sim \mathcal{N}(0, I_k)$ is independent of ω. We further have

$$\mathbf{E}\left\{|\zeta_1|\right\} = \int |t|p_1(t)dt = \sqrt{\frac{2}{\pi}}$$

and

$$\mathbf{E}\left\{\|\omega\|_2\right\} = \mathbf{E}\left\{\|\widetilde{\omega}\|_2\right\} \ge \|\mathbf{E}\left\{\widetilde{\omega}\right\}\|_2 = \left[\int |t|p_1(t)dt\right]\sqrt{\sum_{i=1}^{m}\alpha_i^2} = \sqrt{\frac{2}{\pi}}\sqrt{\sum_{i=1}^{m}\alpha_i^2}.$$

Combining our observations, we come to

$$\mathbf{E}\left\{\left|\sum_{i=1}^{2k}\beta_i\xi_i^2\right|\right\} \geq \frac{4}{\pi}\|\alpha\|_2 \geq \frac{4}{\pi\sqrt{k}}\|\alpha\|_1 = \frac{4}{\pi\sqrt{k}}.$$

This relation combines with (4.4.44) to yield $J \geq \frac{2}{\pi\sqrt{k}}$. Recalling the definition of $\theta(k)$, we come to $\theta(k) \leq \frac{\pi\sqrt{k}}{2}$, as required in (4.4.37).

It remains to prove that $\theta(2) = \frac{\pi}{2}$. From the definition of $\theta(\cdot)$ it follows that

$$\theta^{-1}(2) = \min_{0\leq\theta\leq 1}\int |\theta u_1^2 - (1-\theta)u_2^2| p_2(u)du \equiv \min_{0\leq\theta\leq 1} f(\theta).$$

The function $f(\theta)$ is clearly convex and satisfies the identity $f(\theta) = f(1-\theta)$, $0 \leq \theta \leq 1$, so that its minimum is attained at $\theta = \frac{1}{2}$. A direct computation says that $f(\frac{1}{2}) = \frac{2}{\pi}$. □

Corollary 4.4.1. *Let the ranks of all matrices $A_1, \ldots, A_m$ in* MC *be $\leq \mu$. Then the optimal value in the semidefinite problem*

$$\rho[A_1, \ldots, A_m : A_0] = \max_{\rho, X^i}\left\{\rho \mid \begin{array}{c} X^i \succeq \rho A_i,\ X^i \succeq -\rho A_i,\ i = 1, \ldots, m, \\ \sum_{i=1}^m X^i \preceq A_0 \end{array}\right\} \quad (4.4.45)$$

is a lower bound on $R[A_1, \ldots, A_m : A_0]$, and the true quantity is at most $\theta(\mu)$ times (see (4.4.38), (4.4.37)*) larger than the bound:*

$$\rho[A_1, \ldots, A_m : A_0] \leq R[A_1, \ldots, A_m : A_0] \leq \theta(\mu)\rho[A_1, \ldots, A_m : A_0]. \quad (4.4.46)$$

Application: Lyapunov stability analysis for an interval matrix. Now we are equipped to attack the problem of certifying the stability of uncertain linear dynamic system with interval uncertainty. The problem we are interested in is as follows:

> Interval Lyapunov. *Given a stable $n \times n$ matrix A^*[30] and an $n \times n$ matrix $D \neq 0$ with nonnegative entries, find the supremum $R[A^*, D]$ of those $\rho \geq 0$ for which all instances of the interval matrix*
>
> $$\mathcal{U}_\rho = \{A \in \mathbf{M}^{n,n} : |A_{ij} - A^*_{ij}| \leq \rho D_{ij},\ i, j = 1, \ldots, n\}$$
>
> *share a common quadratic Lyapunov function, i.e., the semi-infinite system of LMIs*
>
> $$X \succeq I; \quad A^T X + XA \preceq -I \quad \forall A \in \mathcal{U}_\rho \qquad (\mathrm{Ly}[\rho])$$
>
> *in matrix variable $X \in \mathbf{S}^n$ is solvable.*

[30] That is, with all eigenvalues from the open left half-plane or, which is the same, such that $[A^*]^T X + XA^* \prec 0$ for certain $X \succ 0$.

Observe that $X \succeq I$ solves (Ly_ρ) if and only if the matrix cube

$$\begin{aligned}\mathcal{A}_\rho[X] &= \Big\{B = \underbrace{\left[-I - [A^*]^T X - XA^*\right]}_{A_0[X]} \\ &\qquad + \sum_{(i,j)\in\mathcal{D}} z_{ij} \underbrace{\left[[D_{ij}E^{ij}]^T X + X[D_{ij}E^{ij}]\right]}_{A_{ij}[X]} \Big| \, |z_{ij}| \le \rho,\ (i,j) \in \mathcal{D}\Big\} \\ \mathcal{D} &= \{(i,j) : D_{ij} > 0\}\end{aligned}$$

is contained in $\mathbf{S}^n_+$; here E^{ij} are the basic $n \times n$ matrices. (ijth entry of E^{ij} is 1; all other entries are zero.) Note that the ranks of the matrices $A_{ij}[X]$, $(i,j) \in \mathcal{D}$, are at most 2. Therefore from Proposition 4.4.5 and Theorem 4.4.1 we get the following result.

Proposition 4.4.6. *Let $\rho \ge 0$. Then*

(i) *If the system of LMIs*

$$\begin{gathered} X \succeq I, \\ X^{ij} \succeq -\rho D_{ij}\left[[E^{ij}]^T X + XE^{ij}\right],\ \ X^{ij} \succeq \rho D_{ij}\left[[E^{ij}]^T X + XE^{ij}\right],\ \ (i,j) \in \mathcal{D}, \\ \sum_{(i,j)\in\mathcal{D}} X^{ij} \preceq -I - [A^*]^T X - XA^* \end{gathered} \tag{A[ρ]}$$

in matrix variables X, X^{ij}, $(i,j) \in \mathcal{D}$, is solvable, then so is the system (Ly[ρ])*, and the X-component of a solution of the former system solves the latter system.* (ii) *If the system of LMIs* (A[ρ]) *is not solvable, then so is the system* (Ly[$\frac{\pi\rho}{2}$]).

In particular, the supremum $\rho[A^*, D]$ of those ρ for which (A[ρ]) *is solvable is a lower bound for $R[A^*, D]$, and the true quantity is at most $\frac{\pi}{2}$ times larger than the bound:*

$$\rho[A^*, D] \le R[A^*, D] \le \frac{\pi}{2}\rho[A^*, D].$$

Computing $\rho[A^*, D]$. The quantity $\rho[A^*, D]$, in contrast to $R[A^*, D]$, is efficiently computable: applying dichotomy in ρ, we can find a high-accuracy approximation of $\rho[A^*, D]$ via solving a small series of semidefinite feasibility problems (A[ρ]). Note, however, that problem (A[ρ]), although computationally tractable, is not that simple: in the case of full uncertainty ($D_{ij} > 0\ \forall i, j$) it has $n^2 + n$ matrix variables of the size $n \times n$ each. Our local goal is to demonstrate that one can use duality to reduce dramatically the design dimension of the problem.

It makes sense to focus on the problem slightly more general than the one of finding $\rho[A^*, D]$, namely, on the problem as follows:

> (P) *We are given $m + 1$ $n \times n$ symmetric matrices $A_0[x]$, $A_1[x], \ldots, A_m[x]$ affinely depending on vector x of design variables, with $A_i[x]$, $i \ge 1$, of the form*
>
> $$A_i[x] = a_i[x]b_i^T + b_i a_i^T[x],$$

where the vectors $a_i[x]$ are affine in x. Besides this, we are given an SDr set

$$\mathcal{X} = \{x \mid Px + Qu + r \succeq 0\}$$

in the space of design variables. We want to find the supremum ρ_* of those ρ for which there exist $x \in \mathcal{X}$ and matrices $X^1, \ldots, X^m$ such that

$$\begin{array}{c} X^i \succeq \rho A_i[x], \ X^i \succeq -\rho A_i[x], \ i = 1, \ldots, m, \\ \sum_{i=1}^m X^i \prec A_0[x]. \end{array} \tag{4.4.47}$$

Note that the problem of computing $\rho[A^*, D]$ is exactly of this type, with a symmetric $n \times n$ matrix X playing the role of x.

From now on, we make the following assumptions:

I.1. The system of LMIs

$$Px + Qu + r \succeq 0, \quad A_0[x] \succ 0,$$

is feasible.

I.2. For every i we have $b_i \neq 0$ and $a_i[x] \neq 0 \ \forall x \in \mathcal{X}$.

I.3. $\rho_* < +\infty$.

Note that in the problem of computing $\rho[A^*, D]$ I.2 is trivially true, I.1 means that there exists $X \succeq I$ such that $[A^*]^T X + XA^* \prec -I$, i.e., that A^* is stable (which we have assumed from the very beginning), and I.3 requires $\rho[A^*, D] < +\infty$, which is indeed natural to assume.

Step 1. Given x such that $A_0[x] \succ 0$, consider the following semidefinite problem:

$$\rho(x) = \sup_{\rho, X^1, \ldots, X^m} \left\{ \rho \Big| \begin{array}{c} X^i \succeq \rho A_i[x], \ X^i \succeq -\rho A_i[x], \ i = 1, \ldots, m, \\ \sum_{i=1}^m X^i \preceq A_0[x] \end{array} \right\}. \tag{P[x]}$$

Note that

$$\rho_* = \sup_{x \in \mathcal{X}, A_0[x] \succ 0} \rho(x). \tag{4.4.48}$$

Since $A_0[x] \succ 0$, problem (P[x]) is strictly feasible with a positive optimal value, and since $\rho_* < +\infty$, the problem is bounded above. By the conic duality theorem, $\rho(x)$ is the optimal value in the problem

$$\min_{W, U^1, \ldots, U^m, V^1, \ldots, V^m} \left\{ \mathrm{Tr}(WA_0[x]) \Big| \begin{array}{c} \sum_{i=1}^m \mathrm{Tr}([U^i - V^i]A_i[x]) = 1, \\ U^i + V^i = W, \ i = 1, \ldots, m, \\ W \succeq 0, U^i \succeq 0, V^i \succeq 0 \end{array} \right\}$$

dual to (P[x]). Since the optimal value $\rho(x)$ in the problem is positive, by homogeneity reasons we have

$$\frac{1}{\rho(x)} = \max_{W,U^1,\ldots,U^m,V^1,\ldots,V^m} \left\{ \sum_{i=1}^m \mathrm{Tr}([U^i - V^i]A_i[x]) \,\middle|\, \begin{array}{c} \mathrm{Tr}(WA_0[x]) = 1, \\ U^i + V^i = W, \\ W \succeq 0,\, U^i \succeq 0,\, V^i \succeq 0 \end{array} \right\}.$$

Invoking Lemma 4.4.3, we can carry out partial maximization in U^i, V^i, thus coming to

$$\frac{1}{\rho(x)} = \max_W \left\{ \sum_{i=1}^m \|\lambda(W^{1/2}A_i[x]W^{1/2})\|_1 \,\middle|\, \mathrm{Tr}(WA_0[x]) = 1,\, W \succeq 0 \right\}. \tag{4.4.49}$$

Now let us make the following observation.

Lemma 4.4.5. *Let $A = fg^T + gf^T$. Then $\|\lambda(A)\|_1 = 2\|f\|_2\|g\|_2$.*

Proof. Without loss of generality, we may assume that $f = (u, 0\ldots, 0)^T$, $g = (v, w, 0, \ldots, 0)^T$. Then the eigenvalues of $fg^T + gf^T$ are, up to a number of zeros, the same as the eigenvalues of the 2×2 matrix $\left[\begin{smallmatrix} 2uv & uw \\ uw & 0 \end{smallmatrix}\right]$. The eigenvalues of the latter matrix are $-uv \pm \sqrt{u^2v^2 + u^2w^2}$, and the sum of their absolute values is $2\sqrt{u^2v^2 + u^2w^2} = 2\|f\|_2\|g\|_2$. □

Taking into account that for $i \geq 1$ we have

$$W^{1/2}A_i[x]W^{1/2} = (W^{1/2}a_i[x])(W^{1/2}b_i)^T + (W^{1/2}b_i)(W^{1/2}a_i[x])^T$$

and using Lemma 4.4.5, we can rewrite (4.4.49) as

$$\frac{1}{\rho(x)} = 2\max_W \left\{ \sum_i \sqrt{a_i^T[x]Wa_i[x]}\sqrt{b_i Wb_i} \,\middle|\, \mathrm{Tr}(WA_0[x]) = 1,\ W \succeq 0 \right\}$$

or, which is the same,

$$\frac{1}{\rho(x)} = \underbrace{\max_{W,\tau_i} \left\{ 2\sum_i \tau_i \,\middle|\, \mathrm{Tr}(WA_0[x]) = 1,\ W \succeq 0,\, \begin{pmatrix} a_i[x]^T Wa_i[x] & \tau_i \\ \tau_i & b_i^T Wb_i \end{pmatrix} \succeq 0,\ i = 1, \ldots, m \right\}}_{(\mathrm{D}[x])}. \tag{4.4.50}$$

Now we can look at the results of our effort. By definition, $\rho(x)$ is the optimal value in the semidefinite program (P[x]), which has a single scalar variable and m symmetric matrix variables (i.e., totally $\frac{mn(n+1)}{2} + 1$ scalar decision variables) and $2m + 1$ large (of the size $n \times n$ each) LMIs. Equation (4.4.49) offers an alternative description of $\rho(x)$, which requires solving a much smaller semidefinite program—one with just $m + \frac{n(n+1)}{2}$ scalar decision variables, a single large LMI $W \succeq 0$, and m small (2×2) LMIs! And this (dramatic for large n, m) simplification was achieved in a completely mechanical fashion, by a straightforward use of conic duality. Note that the problem of computing $\rho(x)$ is

important by its own right; for example, in the context of the interval Lyapunov stability analysis this problem provides a bound (accurate within the factor $\frac{\pi}{2}$) on the largest ρ such that a given quadratic function is a stability certificate for all instances of $\mathcal{U}_\rho$.

Note that our final goal is to maximize $\rho(x)$ in $x \in \mathcal{X}$, and in this respect (4.4.50) is not that useful—the problem (D[x]) is not a semidefinite program in W, x. To overcome this drawback, we intend to pass from (D[x]) (this problem is, basically, the dual of our original problem (P[x])) to its dual. At the first glance, this intention seems to be senseless: the dual of the dual is the original problem! The essence of the matter is in the word "basically." (D[x]) is not exactly the dual of (P[x]); it was obtained from this dual by eliminating part of the variables. Therefore we have no a priori understanding what will be the dual of (D[x]). Well, let us look what it is.

Step 2. From $A_0[x] \succ 0$ combined with I.2 it follows that the problem (D[x]) is strictly feasible, and of course it is bounded above (since $\rho(x) > 0$). Therefore the quantity in the right-hand side of (4.4.50) is equal to the optimal value in the problem dual to (D[x]):

$$\frac{1}{\rho(x)} = \min_{\lambda, Z, \{\xi_i, \eta_i, \zeta_i\}} \left\{ \lambda \;\middle|\; \begin{array}{c} \sum_{i=1}^m \left[\xi_i a_i[x] a_i^T[x] + \eta_i b_i b_i^T\right] + Z = \lambda A_0[x], \\ \begin{pmatrix} \xi_i & \zeta_i \\ \zeta_i & \eta_i \end{pmatrix} \succeq 0,\ i = 1, \ldots, m, \\ Z \succeq 0, \\ \zeta_i = 1,\ i = 1, \ldots, m. \end{array} \right\}$$

or, which is the same,

$$\frac{1}{\rho(x)} = \min_{\lambda, \{\eta_i\}} \left\{ \lambda \;\middle|\; \sum_{i=1}^m \left[\frac{1}{\eta_i} a_i[x] a_i^T[x] + \eta_i b_i b_i^T\right] \preceq \lambda A_0[x],\ \eta_i \geq 0,\ i = 1, \ldots, m \right\}. \tag{4.4.51}$$

Observing that when η_i are positive, the relation

$$\sum_{i=1}^m \left[\frac{1}{\eta_i} a_i[x] a_i^T[x] + \eta_i b_i b_i^T\right] \preceq Y,$$

by the Schur complement lemma, takes place if and only if

$$\begin{pmatrix} Y - \sum_{i=1}^m \eta_i b_i b_i^T & [a_1[x]; a_2[x]; \ldots; a_m[x]] \\ [a_1[x]; a_2[x]; \ldots; a_m[x]]^T & \mathrm{Diag}(\eta_1, \ldots, \eta_m) \end{pmatrix} \succeq 0,$$

we can rewrite (4.4.51) equivalently as

$$\frac{1}{\rho(x)} = \min_{\lambda, \{\eta_i\}, Y} \left\{ \lambda \;\middle|\; \begin{array}{c} \begin{pmatrix} Y - \sum_{i=1}^m \eta_i b_i b_i^T & [a_1[x]; a_2[x]; \ldots; a_m[x]] \\ [a_1[x]; a_2[x]; \ldots; a_m[x]]^T & \mathrm{Diag}(\eta_1, \ldots, \eta_m) \end{pmatrix} \succeq 0, \\ Y \preceq \lambda A_0[x] \end{array} \right\}. \tag{4.4.52}$$

According to (4.4.48), we have

$$\frac{1}{\rho_*} = \inf_{\lambda,Y,x,u,\{\eta_i\}} \left\{ \lambda \;\middle|\; \begin{array}{c} Px + Qu + r \succeq 0, \\ \begin{pmatrix} Y - \sum_{i=1}^m \eta_i b_i b_i^T & [a_1[x]; a_2[x]; \ldots; a_m[x]] \\ [a_1[x]; a_2[x]; \ldots; a_m[x]]^T & \mathrm{Diag}(\eta_1, \ldots, \eta_m) \end{pmatrix} \succeq 0, \\ A_0[x] \succ 0, \\ Y \preceq \lambda A_0[x] \end{array} \right\}. \tag{4.4.53}$$

The optimization problem in the left-hand side of the resulting representation of ρ_* is not exactly a semidefinite problem (due to the bilinear term $\lambda A_0[x]$ in the right-hand side of one of the constraints). This is what is called a *generalized eigenvalue problem*. Note, however, that the problem can be easily reduced to a small series of semidefinite programs. Indeed, let λ_* be the (unknown) optimal value in the problem. Given a candidate value of λ, to verify that $\lambda > \lambda_*$ is the same as to verify that the optimal value in the semidefinite program

$$\max_{s,Y,x,u,\{\eta_i\}} \left\{ s \;\middle|\; \begin{array}{ll} \text{(a)} & Px + Qu + r \succeq 0, \\ \text{(b)} & \begin{pmatrix} Y - \sum_{i=1}^m \eta_i b_i b_i^T & [a_1[x]; a_2[x]; \ldots; a_m[x]] \\ [a_1[x]; a_2[x]; \ldots; a_m[x]]^T & \mathrm{Diag}(\eta_1, \ldots, \eta_m) \end{pmatrix} \succeq 0, \\ \text{(c)} & sI \preceq A_0[x], \\ \text{(d)} & Y + sI \preceq \lambda A_0[x] \end{array} \right\} \tag{P$^\lambda$}$$

is positive. With this observation in mind, we can find λ_* applying bisection in λ. At every step of this process, we solve problem (P$^\lambda$) corresponding to the current value of λ in order to check whether this value is or is not $> \lambda_*$. This approach allows us to build a high-accuracy approximation of $\lambda_* = \frac{1}{\rho_*}$ at the cost of solving a small series of problems (P$^\lambda$).

Let us look how the outlined approach works in the problem of computing $\rho[A^*, D]$. Here $x = X$ is a symmetric $n \times n$ matrix, $Px + Qu + r \equiv X - I$, and $m = \mathrm{Card}(\mathcal{D})$ is the total number of uncertain entries in our uncertain interval matrix. Consequently, problem (P$^\lambda$) has two symmetric matrix variables X, Y, a single scalar variable s, and $m \le n^2$ scalar variables η_i, i.e., totally at most $2n^2 + n + 2$ scalar design variables. As about LMIs, (P$^\lambda$) has three large $(n \times n)$ LMIs (a), (c), (d) and one very large $((n+m) \times (n+m))$ LMI (b); note, however, that this very large LMI is of a very simple structure. Thus, (P$^\lambda$) seems to be much better suited for numerical processing than our original system (A$[\rho]$), where we have totally $\frac{(m+1)n(n+1)}{2}$ scalar design variables and $m+1$ LMIs of size $n \times n$ each.

Remark 4.4.1. Note that our results on the Matrix Cube problem can be applied to the interval version of the Lyapunov stability synthesis problem, where we want to find the supremum R of those ρ for which an uncertain controllable system

$$\frac{d}{dt}x(t) = A(t)x(t) + B(t)u(t)$$

with interval uncertainty

$$(A(t), B(t)) \in \mathcal{U}_\rho = \left\{(A, B) : |A_{ij} - A^*_{ij}| \le \rho D_{ij},\ |B_{i\ell} - B^*_{i\ell}| \le \rho C_{i\ell}\ \ \forall i, j, \ell\right\}$$

admits a linear feedback

$$u(t) = Kx(t)$$

such that all instances $A(t) + B(t)K$ of the resulting closed-loop system share a common quadratic Lyapunov function. Here our constructions should be applied to the semi-infinite system of LMIs

$$Y \succeq I, \quad BL + AY + L^T B^T + YA^T \preceq -I \quad \forall (A, B) \in \mathcal{U}_\rho$$

in variables L, Y (see Proposition 4.4.4) and then yield an efficiently computable lower bound on R which is at most $\frac{\pi}{2}$ times less than R.

Nesterov's theorem revisited. Our results on the Matrix Cube problem give an alternative proof of the Nesterov theorem (Theorem 4.3.2). Recall that in this theorem we compare the true maximum

$$OPT = \max_d \{d^T Ad \mid \|d\|_\infty \leq 1\}$$

of a positive semidefinite ($A \succeq 0$) quadratic form on the unit n-dimensional cube and the semidefinite upper bound

$$SDP = \max_X \{\mathrm{Tr}(AX) \mid X \succeq 0, X_{ii} \leq 1, \ i = 1, \ldots, n\} \tag{4.4.54}$$

on OPT; the theorem says that

$$OPT \leq SDP \leq \frac{\pi}{2} OPT. \tag{4.4.55}$$

To derive (4.4.55) from the Matrix Cube–related considerations, assume that $A \succ 0$ rather than $A \succeq 0$ (by continuity reasons, to prove (4.4.55) for the case of $A \succ 0$ is the same as to prove the relation for all $A \succeq 0$) and let us start with the following simple observation.

Lemma 4.4.6. *Let $A \succ 0$ and*

$$OPT = \max_d \left\{d^T Ad \mid \|d\|_\infty \leq 1\right\}.$$

Then

$$\frac{1}{OPT} = \max \left\{\rho : \begin{pmatrix} 1 & d^T \\ d & A^{-1} \end{pmatrix} \succeq 0 \quad \forall (d : \|d\|_\infty \leq \rho^{1/2})\right\} \tag{4.4.56}$$

and

$$\frac{1}{OPT} = \max \left\{\rho : A^{-1} \succeq X \quad \forall (X \in \mathbf{S}^n : |X_{ij}| \leq \rho \ \forall i, j)\right\}. \tag{4.4.57}$$

Proof. To get (4.4.56), note that by the lemma on the Schur complement, all matrices of the form $\begin{pmatrix} 1 & d^T \\ d & A^{-1} \end{pmatrix}$ with $\|d\|_\infty \leq \rho^{1/2}$ are $\succeq 0$ if and only if $d^T(A^{-1})^{-1}d = d^T A d \leq 1 \; \forall d$, $\|d\|_\infty \leq \rho^{1/2}$, i.e., if and only if $\rho \cdot OPT \leq 1$; we have derived (4.4.56). We now have

$$
\begin{array}{lcr}
\text{(a)} & \frac{1}{OPT} \geq \rho & \\
& \Updownarrow & \text{[by (4.4.56)]} \\
& \begin{pmatrix} 1 & d^T \\ d & A^{-1} \end{pmatrix} \succeq 0 \quad \forall (d : \|d\|_\infty \leq \rho^{1/2}) & \\
& \Updownarrow & \text{[the Schur complement lemma]} \\
& A^{-1} \succeq \rho d d^T \quad \forall (d, \|d\|_\infty \leq 1) & \\
& \Updownarrow & \\
& x^T A^{-1} x \geq \rho (d^T x)^2 \quad \forall x \, \forall (d : \|d\|_\infty \leq 1) & \\
& \Updownarrow & \\
& x^T A^{-1} x \geq \rho \|x\|_1^2 \quad \forall x & \\
& \Updownarrow & \\
\text{(b)} & A^{-1} \succeq \rho Y \quad \forall (Y = Y^T : |Y_{ij}| \leq 1 \; \forall i, j), &
\end{array}
$$

where the concluding $\Updownarrow$ is given by the evident relation

$$\|x\|_1^2 = \max_Y \left\{ x^T Y x : Y = Y^T, |Y_{ij}| \leq 1 \; \forall i, j \right\}.$$

The equivalence (a) $\Leftrightarrow$ (b) is exactly (4.4.57). □

By (4.4.57), $\frac{1}{OPT}$ is exactly the maximum R of those ρ for which the matrix cube

$$\mathcal{C}_\rho = \left\{ A^{-1} + \sum_{1 \leq i \leq j \leq n} z_{ij} S^{ij} \,\Big|\, \max_{i,j} |z_{ij}| \leq \rho \right\}$$

is contained in $\mathbf{S}^n_+$. Here S^{ij} are the basic symmetric matrices. (S^{ii} has a single nonzero entry, equal to 1, in the cell ii, and S^{ij}, $i < j$, has exactly two nonzero entries, equal to 1, in the cells ij and ji.) Since the ranks of the matrices S^{ij} do not exceed 2, Proposition 4.4.5 and Theorem 4.4.1 say that the optimal value in the semidefinite program

$$\rho(A) = \max_{\rho, X^{ij}} \left\{ \rho \,\middle|\, \begin{array}{c} X^{ij} \succeq \rho S^{ij}, \; X^{ij} \succeq -\rho S^{ij}, \; 1 \leq i \leq j \leq n, \\ \sum\limits_{i \leq j} X^{ij} \preceq A^{-1} \end{array} \right\} \tag{S}$$

is a lower bound for R, and this bound coincides with R up to the factor $\frac{\pi}{2}$. Consequently, $\frac{1}{\rho(A)}$ is an upper bound on OPT, and this bound is at most $\frac{\pi}{2}$ times larger than OPT. It remains to note that a direct computation (completely similar to the one that led us from (P[x]) to (4.4.50)) demonstrates that $\frac{1}{\rho(A)}$ is exactly the quantity SDP given by (4.4.54).

4.5 Applications III: Robust quadratic programming

The concept of a robust counterpart of an optimization problem with uncertain data (see section 3.4.2) is in no sense restricted to LP. Whenever we have an optimization problem depending on certain data, we may ask what happens when the data are uncertain and all we know is an uncertainty set the data belong to. Given such an uncertainty set, we may require candidate solutions to be robust feasible—to satisfy the realizations of the constraints for all data belonging through the uncertainty set. The robust counterpart of an uncertain problem is the problem of minimizing the objective[31] over the set of robust feasible solutions.

Now, we have seen in section 3.4.2 that the robust form of an uncertain linear inequality with the coefficients varying in an ellipsoid is a conic quadratic inequality; as a result, the robust counterpart of an uncertain LP problem with ellipsoidal uncertainty is a conic quadratic problem. What is the robust form of an uncertain CQI

$$\|Ax + b\|_2 \le c^T x + d \qquad \left[A \in \mathbf{M}^{m,n}, b \in \mathbf{R}^m, c \in \mathbf{R}^n, d \in \mathbf{R}\right] \tag{4.5.58}$$

with uncertain data $(A, b, c, d) \in \mathcal{U}$? We want to know how to describe the set of all robust feasible solutions of this inequality, i.e., the set of x's such that

$$\|Ax + b\|_2 \le c^T x + d \quad \forall (A, b, c, d) \in \mathcal{U}. \tag{4.5.59}$$

We are about to demonstrate that in the case when the data (P, p) of the left-hand side and the data (q, r) of the right-hand side of the inequality (4.5.58) independently of each other run through respective ellipsoids, i.e., the uncertainty set is of the form

$$\begin{aligned}\mathcal{U} = \Bigg\{ & (A, b, c, d) \mid \exists (u \in \mathbf{R}^l, u^T u \le 1, v \in \mathbf{R}^r, v^T v \le 1) : \\ & [A; b] = [A^0; b^0] + \sum_{i=1}^{l} u_i [A^i; b^i], (c, d) = (c^0, d^0) + \sum_{i=1}^{r} v_i (c^i, d^i) \Bigg\},\end{aligned} \tag{4.5.60}$$

then the robust version (4.5.59) of the uncertain inequality (4.5.58) can be expressed via LMIs.

Proposition 4.5.1. Robust counterpart of a conic quadratic inequality with simple ellipsoidal uncertainty. *In the case of uncertain conic inequality with uncertainty* (4.5.60)*, the set of robust feasible solutions of the inequality* (4.5.59) *is SDr with the following SDR:*

[31] Without loss of generality, we may assume that the objective is certain—is not affected by the data uncertainty. Indeed, we always may ensure this situation by passing to an equivalent problem with linear (and standard) objective:

$$\min_x \{ f(x) : x \in X \} \mapsto \min_{t,x} \{ t : f(x) - t \le 0, x \in X \}.$$

x satisfies (4.5.59) *if and only if there exist real s, μ such that the triple (x, s, μ) satisfies the following LMIs:*

(a)
$$\left(\begin{array}{c|ccccc} (c^0)^Tx+d^0-s & (c^1)^Tx+d^1 & (c^2)^Tx+d^2 & \cdots & (c^r)^Tx+d^r \\ \hline (c^1)^Tx+d^1 & (c^0)^Tx+d^0-s & & & \\ (c^2)^Tx+d^2 & & (c^0)^Tx+d^0-s & & \\ \vdots & & & \ddots & \\ (c^r)^Tx+d^r & & & & (c^0)^Tx+d^0-s \end{array}\right) \succeq 0,$$

(b)
$$\left(\begin{array}{c|c|ccc} sI_m & A^0x+b^0 & A^1x+b^1 & \cdots & A^lx+b^l \\ \hline [A^0x+b^0]^T & s-\mu & & & \\ \hline [A^1x+b^1]^T & & \mu & & \\ \vdots & & & \ddots & \\ [A^lx+b^l]^T & & & & \mu \end{array}\right) \succeq 0. \tag{4.5.61}$$

Proof. Since the uncertain data of the left and the right sides in (4.5.58) run independently through their respective ellipsoids, x satisfies (4.5.59) (let us call this Property 0) if and only if the following holds.

(Property 1) There exists $s \in \mathbf{R}$ such that

$$\begin{array}{lrcl} \text{(a)} & s & \le & \left[c^0+\sum_{i=1}^r v_i c^i\right]^T x + d^0 + \sum_{i=1}^r v_i d^i \quad \forall v : v^Tv \le 1, \\ \text{(b)} & s & \ge & \left\|\left[A^0+\sum_{i=1}^l u_i A^i\right]x + b^0 + \sum_{i=1}^l u_i b^i\right\|_2 \quad \forall u : u^Tu \le 1. \end{array} \tag{4.5.62}$$

Now, the relation (4.5.62)(a) is equivalent to the conic quadratic inequality

$$[c^0]^Tx + d^0 - \left\|\begin{pmatrix} [c^1]^Tx+d^1 \\ [c^2]^Tx+d^2 \\ \cdots \\ [c^r]^Tx+d^r \end{pmatrix}\right\|_2 \ge s$$

(why?) or, which is the same (see (4.2.1)), to the LMI (4.5.61)(a).

Now let us set

$$\begin{array}{rcl} p(x) & = & A^0x+b^0 \in \mathbf{R}^m; \\ P(x) & = & [A^1x+b^1; A^2x+b^2; \ldots; A^lx+b^l] \in \mathbf{M}^{m,l}. \end{array}$$

The relation (4.5.62)(b) is nothing but

$$s \ge \|P(x)u + p(x)\|_2 \quad \forall u : u^Tu \le 1;$$

thus, it is equivalent to the fact that $s \geq 0$ and the quadratic form of u

$$s^2 - p^T(x)p(x) - 2p^T(x)P(x)u - u^T P^T(x)P(x)u$$

is nonnegative whenever $u^T u \leq 1$. This, in turn, is equivalent to the fact that the homogeneous quadratic form

$$(s^2 - p^T(x)p(x))t^2 - 2tp^T(x)P(x)u - u^T P^T(x)P(x)u$$

of u, t ($t \in \mathbf{R}$) is nonnegative whenever $u^T u \leq t^2$. Applying the $\mathcal{S}$-lemma, we conclude that

(!) *Relation* (4.5.62)(b) *is equivalent to the facts that* $s \geq 0$ *and that there exists* $\nu \geq 0$ *such that*

$$(s^2 - p^T(x)p(x))t^2 - 2tp^T(x)P(x)u - u^T P^T(x)P(x)u - \nu[t^2 - u^T u] \geq 0 \quad \forall (u,t) \in \mathbf{R}^l \times \mathbf{R}. \tag{4.5.63}$$

We now claim that the quantity ν in (!) can be represented as μs with some nonnegative μ. There is nothing to prove when $s > 0$. Now assume that $s = 0$ and that (4.5.63) is satisfied by some $\nu \geq 0$. Then $\nu = 0$ (look what happens when $t = 1, u = 0$), and so it can be represented as μs with, say, $\mu = 0$. Thus, we have demonstrated that

(!!) *Relation* (4.5.62)(b) *is equivalent to the facts that* $s \geq 0$ *and that there exists* $\mu \geq 0$ *such that*

$$s\left[(s-\mu)t^2 + \mu u^T u\right] - \left[p^T(x)p(x)t^2 + 2tp^T(x)P(x)u + u^T P^T(x)P(x)u\right] \geq 0 \quad \forall (u,t)$$

or, which is the same, such that

$$s\begin{pmatrix} s-\mu & & \\ & \ddots & \\ & & \mu I_l \end{pmatrix} - ([p(x);\, P(x)]]^T\,[p(x);\, P(x)] \succeq 0. \tag{4.5.64}$$

Now note that when $s > 0$, (4.5.64) says exactly that the Schur complement to the northwestern block in the matrix

$$\left(\begin{array}{c|c} sI_m & [p(x);\, P(x)] \\ \hline [p(x);\, P(x)]^T & \begin{matrix} s-\mu & & \\ & \ddots & \\ & & \mu I_l \end{matrix} \end{array}\right) \tag{$*$}$$

is positive semidefinite. By the Schur complement lemma, this is exactly the same as to say that $s > 0$ and the matrix (*) is positive semidefinite. Thus, in the case of $s > 0$, relation (4.5.62) says precisely that there exists μ such that (*) is positive semidefinite. In the case of $s = 0$, relation (4.5.62) can be satisfied if and only if $p(x) = 0$ and $P(x) = 0$ (see (!!)). This again is exactly the case when there exists a μ such that (*) is positive semidefinite. Since (*) can be positive semidefinite only when $s \geq 0$, we come to the following conclusion:

Relation (4.5.62)(b) *is satisfied if and only if there exists* μ *such that the matrix* (*) *is positive semidefinite.*

It remains to notice that (*) is exactly the matrix in (4.5.61)(b). □

Remark 4.5.1. We have built an explicit semidefinite representation of the robust version of a conic quadratic inequality in the case of simple ellipsoidal uncertainty. In more complicated cases (e.g., when b, c, d in (4.5.58) are not affected by uncertainty and the matrix A is affected by interval uncertainty: all its entries, independently of each other, run through given intervals), it may be computationally intractable already to check whether a given x is robust feasible.

Example: Robust synthesis of antenna array. We have already considered the problem of antenna array synthesis in the Tschebyshev setting, i.e., when the discrepancy between the target diagram and the designed one is measured in the uniform norm (section 1.2.4). We have also seen that the solution to the resulting LP problem may be extremely unstable with respect to small implementation errors, and furthermore we have shown how to overcome this difficulty by switching from the nominal design to the one given by the robust counterpart methodology (section 3.4.2). Now, what happens if we replace the uniform norm of the discrepancy with the $\|\cdot\|_2$-norm, i.e., define the optimal design as the optimal solution x_j^* to the usual least squares problem

$$\min_x \left\{ \|Z_* - \sum_j x_j Z_j\|_2 \equiv \sqrt{\frac{1}{N}\sum_{\theta\in T}(Z_*(\theta) - \sum_j x_j Z_j(\theta))^2} \right\}, \qquad \text{(LS)}$$

where N is the cardinality of the grid T?

Note that the factor $\frac{1}{N}$ under the square root does not influence the least squares solution. The only purpose of it is to make the figures comparable with those related to the case of the best uniform approximation: with our normalization, we have

$$\|Z_* - \sum_j x_j Z_j\|_2 \le \|Z_* - \sum_j x_j Z_j\|_\infty$$

for every x.

(LS) is just a linear algebra problem; assuming $Z_*(\cdot)$, $Z_j(\cdot)$ real, its optimal solution x^* is exactly the solution to the normal system of equations

$$(A^T A)x = A^T b,$$

where $b = \frac{1}{\sqrt{N}}(Z_*(\theta_1), \dots Z_*(\theta_N))^T$ and A is the matrix with the columns $\frac{1}{\sqrt{N}}(Z_j(\theta_1), \dots, Z_j(\theta_N))^T$.

Now let us check the stability properties of the least squares solution. Consider exactly the same design data as in sections 1.2.4 and 3.4.2 and assume that the actual amplification coefficients x_j are obtained from their nominal values x_j^* by random perturbations $x_j^* \mapsto x_j = p_j x_j^*$, where p_j are independent random factors with expectations 1 taking values in the segment $[0.999, 1.001]$. Based on our previous experience (see section 3.4.2), we should not be too surprised by the fact that these stability properties are extremely poor, as seen in Fig. 4.7. The reason for instability of the nominal least squares solution is, basically, the same as in the case of the nominal Tschebyshev solution. The system of basic functions Z_j is nearly linearly dependent. (This unpleasant phenomenon is met in the majority of approximation problems arising in applications.) As a result, the normal system of equations

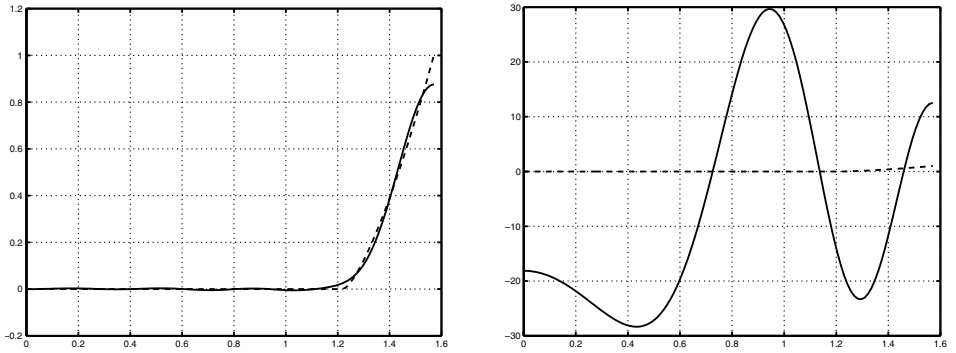


Figure 4.7. *Dream and reality: the nominal diagram (left, solid line) and an actual diagram (right, solid line). Dashed lines are the target diagram. The target diagram varies from 0 to 1, and the nominal diagram (the one corresponding to $x_j = x_j^*$) is at the $\|\cdot\|_2$-distance* 0.0178 *from the target diagram. An actual diagram varies from ≈ -30 to ≈ 30 and is at the $\|\cdot\|_2$-distance 20.0 (1124 times larger!) from the target.*

becomes ill conditioned, and its solution has large entries. And of course even small relative perturbations of very large nominal amplification coefficients may cause and do cause huge perturbations in the actual diagram.

To resolve the difficulty, let us use the robust counterpart methodology. What we are solving now is a conic quadratic inequality, so that we may use the results of this section. The question, however, is how to define a reasonable uncertainty set. Let us look at this question for a general CQI

$$\|Ax + b\|_2 \le c^T x + d \tag{CQI}$$

with $m \times n$ matrix A, assuming, as is the case in our antenna example, that the uncertainty comes from the fact that the entries x_j of a candidate solution are affected by noise:

$$x_j \mapsto x_j(1 + \kappa_j \epsilon_j), \tag{4.5.65}$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent random variables with zero means and unit standard deviations, and $\kappa_j \ge 0$ are (deterministic) relative implementation errors.

What is a reliable version of the inequality (CQI) in the case of random perturbations (4.5.65) in x? Note that it is the same—to view the data A, b, c, d in (CQI) as fixed and to consider perturbations in x, and to think that there are no perturbations in x, but the data in (CQI) are perturbed equivalently. With this latter viewpoint, how could we define an uncertainty set $\mathcal{U}$ so that the robust counterpart of the uncertain conic inequality

$$\|A'x + b'\|_2 \le (c')^T x + d' \quad \forall (A', b', c', d') \in \mathcal{U}$$

would be a reliable version of (CQI)?

The question we have posed is not a purely mathematical question. It has to do with modeling, and modeling—description of a real-world situation in mathematical terms—is always beyond the scope of the mathematics itself. It follows that in our current situation we

are free to use whatever arguments we want—detailed (and time-consuming) mathematical analysis of the random variables we are dealing with common sense, spiritualism, and so forth. Proof of our pudding will be in the eating—testing the quality of the resulting robust solution.

Since we do not have much experience with spiritualism, and a detailed mathematical analysis of the situation does not seem to be simple, we prefer to rely on common sense, namely, to choose somehow a safety parameter ω of order of 1 and to utilize the following principle:

A nonnegative random variable is never larger than ω times its expected value.

Put more mildly, we intend to ignore rare events—those where the above principle is violated, and to take on ourselves full responsibility for the remaining events.

Equipped with our principle, let us build a reliable version of (CQI) as follows. First, we separate the influence of the perturbations on the left and right sides of our inequality. Namely, the value of the right-hand side $c^T y + d$ at a randomly perturbed x—i.e., at a random vector y with the coordinates

$$y_j = x_j + \kappa_j x_j \epsilon_j$$

—is a random variable of the form

$$c^T x + d + \eta,$$

η being a zero mean random variable with standard deviation $V^{1/2}(x)$, where

$$V(x) = \sum_{j=1}^{n} c_j^2 \kappa_j^2 x_j^2. \tag{4.5.66}$$

According to our principle, the value of the right-hand side in (CQI) is never less than the quantity

$$R(x) = c^T x + d - \omega V^{1/2}(x). \tag{4.5.67}$$

It follows that if we ensure that the value of the left-hand side in (CQI) is never larger than $R(x)$, then the perturbations in x never result in violating of (CQI). This scheme is a bit conservative (it may happen that a perturbation that increases the left-hand side of (CQI) increases the right-hand side as well), but this is life—we want to get something tractable and thus can afford to be conservative. Recall that at the moment we are not obliged to be rigorous!

Now we came to the following situation. We would like $R(x)$ to be a reliable upper bound on the values of the left-hand side in (CQI), and this requirement on x will be the reliable version of (CQI). Now, what are typical values of the left-hand side of (CQI)? These are the Euclidean norms of the random vector

$$z \equiv z(x) + \zeta = Ax + b + \zeta,$$

where

$$\zeta = \sum_{j=1}^{n} \kappa_j x_j \epsilon_j A_j$$

(A_j are the columns of A). Note that the vector ζ is a random vector with zero mean. For a given x, let e be the unit vector collinear to the vector $Ax + b$. We have

$$\|z\|_2^2 = l_x^2 + \|\zeta_x\|_2^2,$$

where l_x is the length of the projection of z on the line spanned by e and ζ_x is the projection of z (or, which is the same, of ζ) onto the orthogonal complement to this line.

We have $\|\zeta_x\|_2^2 \le \|\zeta\|_2^2$, and the expected value of $\|\zeta\|_2^2$ is

$$S(x) = \sum_{j=1}^{n} S_{jj} x_j^2, \quad S_{jj} = \kappa_j^2 A_j^T A_j. \tag{4.5.68}$$

According to our principle, $\|\zeta_x\|_2$ is never greater than $\omega S^{1/2}(x)$.

Now let us find a never-type upper bound for l_x—for the length of the projection of z onto the line spanned by the vector $Ax + b$ (or, which is the same, by the unit vector e). We have, of course,

$$|l_x| \le \|Ax + b\|_2 + |e^T \zeta|.$$

Now, $e^T \zeta$ is the random variable

$$e^T \zeta = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} e_i \kappa_j x_j \epsilon_j = \sum_{j=1}^{n} \left[\sum_{i=1}^{m} a_{ij} e_i \right] \kappa_j x_j \epsilon_j$$

with zero mean and the variance

$$v = \sum_{j=1}^{n} \left[\kappa_j \sum_{i=1}^{m} a_{ij} e_i \right]^2 x_j^2. \tag{4.5.69}$$

To end up with tractable formulae, it is desirable to bound from above the latter quantity by a simple function of x. (This is not the case for the quantity itself, since e depends on x in a not that pleasant way!) A natural bound of this type is

$$v \le \sigma^2 \|x\|_\infty^2,$$

where

$$\sigma = |A \mathrm{Diag}(\kappa_1, \ldots, \kappa_n)| \tag{4.5.70}$$

and $|\cdot|$ is the operator norm. Indeed, the coefficients of x_j^2 in (4.5.69) are the squared entries of the vector $\mathrm{Diag}(\kappa_1, \ldots, \kappa_n) A^T e$, and since e is unit, the sum of these squared entries does not exceed σ^2.

According to our principle, the absolute value of $e^T \zeta$ never exceeds the quantity

$$\omega \sigma \|x\|_\infty.$$

Combining all our observations, we conclude that $\|z\|_2$ never exceeds the quantity

$$L(x) = \sqrt{[\|Ax + b\|_2 + \omega \sigma \|x\|_\infty]^2 + \omega^2 S(x)}.$$

Consequently, the reliable version of (CQI) is the inequality $L(x) \le R(x)$, i.e., the inequality

$$\omega V^{1/2}(x) + \sqrt{[\|Ax+b\|_2 + \omega\sigma\|x\|_\infty]^2 + \omega^2 S(x)} \le c^T x + d,$$
$$\left[\begin{array}{rcl} \sigma & = & |A\mathrm{Diag}(\kappa_1,\dots,\kappa_n)|, \\ S(x) & = & \sum_{j=1}^n [\kappa_j^2 A_j^T A_j] x_j^2, \\ V(x) & = & \sum_{j=1}^n \kappa_j^2 c_j^2 x_j^2 \end{array}\right]. \tag{4.5.71}$$

The resulting inequality is indeed tractable—it can be represented by the following system of linear and CQIs:

$$\begin{array}{rcl} t_1 + t_2 & \le & c^T x + d; \\ \omega\|Wx\|_2 & \le & t_1, \\ & & W = \mathrm{Diag}(\kappa_1 c_1, \kappa_2 c_2, \dots, \kappa_n c_n); \\ \|Ax+b\|_2 & \le & s_1; \\ |x_i| & \le & s_2,\ i = 1,\dots,n; \\ \left\|\begin{pmatrix} s_1 + \omega\sigma s_2 \\ \omega D x \end{pmatrix}\right\|_2 & \le & t_2, \\ & & D = \mathrm{Diag}(|\kappa_1|\|A_1\|_2, |\kappa_2|\|A_2\|_2, \dots, |\kappa_n|\|A_n\|_2), \end{array} \tag{4.5.72}$$

where t_1, t_2, s_1, s_2 are additional design variables.

It is worth mentioning—just for fun!—that problem (4.5.72) is, in a sense, the robust counterpart of (CQI) associated with a specific ellipsoidal uncertainty.

Indeed, we can rewrite (CQI) equivalently as the following crazy system of inequalities with respect to x and additional variables t_1, t_2, s_1, s_2:

$$\begin{array}{rcl} t_1 + t_2 & \le & c^T x + d; \\ \|\mathrm{Diag}(\alpha_1,\dots,\alpha_n)x\|_2 & \le & t_1, \\ & & \alpha_1 = 0, \alpha_2 = 0, \dots, \alpha_n = 0; \\ \|Ax+b\|_2 & \le & s_1; \\ |x_i| & \le & s_2,\ i = 1,\dots,n; \\ \left\|\begin{pmatrix} s_1 + \beta_0 s_2 \\ \mathrm{Diag}(\beta_1,\beta_2,\dots,\beta_n)x \end{pmatrix}\right\|_2 & \le & t_2, \\ & & \beta_0 = 0, \beta_1 = 0, \dots, \beta_n = 0. \end{array} \tag{4.5.73}$$

Now assume that the data α_i, β_i in this system are uncertain, namely, linearly depend on perturbation u varying in the segment $[-1, 1]$ (ellipsoidal uncertainty!):

$$\begin{array}{rcl} \alpha_i & = & [\omega\kappa_i c_i]u,\ i = 1,\dots,n, \\ \beta_0 & = & [\omega\sigma]u, \\ \beta_i & = & [\omega|\kappa_i|\|A_i\|_2]u,\ i = 1,\dots,n. \end{array}$$

It is easily seen that the robust counterpart of (4.5.73)—which is, basically, our original conic inequality (CQI)—is exactly (4.5.72). Thus, (4.5.72) is the robust counterpart of (CQI) corresponding to an ellipsoidal uncertainty, and this uncertainty affects the data that are not present in (CQI) at all!

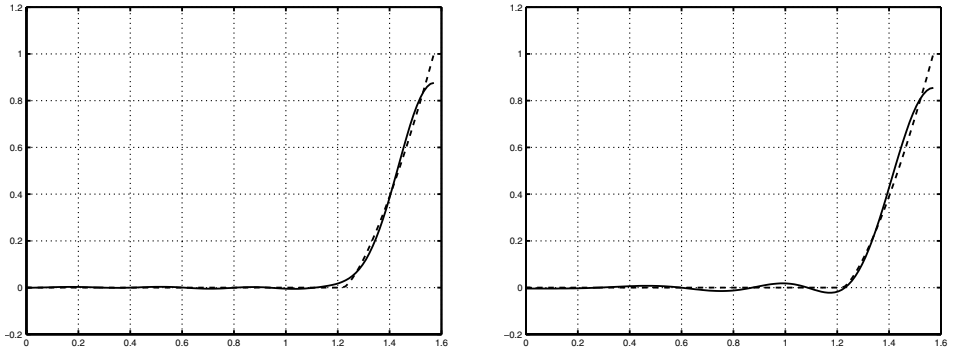


Figure 4.8. *Dream and reality: the nominal least squares diagram (left, solid line) and an actual diagram yielded by robust least squares (right, solid line). Dashed lines are the target diagram.*

What about the pudding we have cooked? Is this approach working for our antenna synthesis problem? It works fine! Look at Fig. 4.8 (safety parameter $\omega = 1$). The robust optimal value in our uncertain least squares problem is 0.0236 (approximately 30% larger than the nominal optimal value 0.0178—the one corresponding to the usual least squares with no implementation errors). The $\|\cdot\|_2$-distance between the target diagram and the actual diagram shown on the picture is the same 0.0236. When generating a sample of random diagrams yielded by our robust least squares design, this distance varies only in the fourth digit after the dot. In a sample of 40 diagrams, the distances to the target varied from 0.0236 to 0.0237. And what happens when in the course of our design we thought that the implementation errors would be 0.1%, while in reality they are 1% (10 times larger)? Nothing bad: now the $\|\cdot\|_2$-distances from the target in a sample of 40 diagrams vary from 0.0239 to 0.0384.

4.6 Applications IV: Synthesis of filters and antennae arrays

Consider a discrete time linear time invariant SISO (single input–single output) dynamic system (cf. section 1.2.3).[32] Such a system $\mathcal{H}$ takes as input a two-sided sequence of reals $u(\cdot) = \{u(k)\}_{k=-\infty}^{\infty}$ and converts it into an output sequence $\mathcal{H}u(\cdot)$ according to

$$\mathcal{H}u(k) = \sum_{l=-\infty}^{\infty} u(l)h(k-l),$$

where $h = \{h(k)\}_{k=-\infty}^{\infty}$ is a real sequence characteristic for $\mathcal{H}$—the impulse response of $\mathcal{H}$. Let us focus on the case of a filter—a causal system with finite memory. Causality means

[32] The models to be presented in this section originate from S.-P. Wu, S. Boyd, and L. Vandenberghe, FIR filter design via spectral factorization and convex optimization, Biswa Datta, ed., in *Applied and Computational Control, Signal and Circuits*, Birkhauser, Basel, 1997, pp. 51–81.

that $h(k) = 0$ for $k < 0$, so that the output $\mathcal{H}u$ at every time k is independent of the input after this time, while the property to have memory n means that $h(k) = 0$ for $k \geq n$, so that $\mathcal{H}u(k)$, for every k, depends on the n inputs $u(k), u(k-1), \ldots, u(k-n+1)$ only. Thus, a filter of order n is just a sequence $h = \{h(k)\}_{k=-\infty}^{\infty}$ with $h(k) = 0$ for all negative k and all $k \geq n$. Of course, a filter $\{h(k)\}_{k=-\infty}^{\infty}$ of order n can be identified with the vector $h = (h(0), \ldots, h(n-1))^T \in \mathbf{R}^n$.

A natural way to look at a filter $h(\cdot)$ of order n is to associate with it the polynomial

$$\hat{h}(z) = \sum_{l=0}^{n-1} h(l) z^l.$$

As any other polynomial on the complex plane, $\hat{h}$ is completely determined by its restriction on the unit circumference $|z| = 1$. This restriction, regarded as 2π-periodic function of real variable ω,

$$H(\omega) = \hat{h}(\exp\{i\omega\}) = \sum_{l=0}^{n-1} h(l) \exp\{il\omega\}$$

is called the *frequency response* of the filter $h(\cdot)$. The frequency response is just a trigonometric polynomial (with complex coefficients) of ω of degree $\leq n-1$.

The meaning of the frequency response is quite transparent: if the input to the filter is a harmonic oscillation

$$u(k) = \Re(a \exp\{i\omega k\}) = |a| \cos(\omega k + \arg(a)) \quad [a \in \mathbf{C} \text{ is the complex amplitude}],$$

then the output is

$$\begin{aligned} \mathcal{H}u(k) &= \sum_{l=-\infty}^{\infty} u(l)h(k-l) = \sum_{l=0}^{n-1} h(l)u(k-l) \\ &= \Re\left(\sum_{l=0}^{n-1} h(l) a \exp\{i\omega(k-l)\}\right) = \Re\left(H(-\omega) a \exp\{i\omega k\}\right). \end{aligned}$$

Thus, the output is a harmonic oscillation of the same frequency as the input, and the complex amplitude of the output is $H(-\omega)$ times the complex amplitude of the input. Thus, the frequency response says how the filter affects a harmonic oscillation of certain frequency ω: the filter multiplies the real amplitude of the oscillation by $|H(-\omega)|$ and shifts the initial phase of the oscillation by $\arg(H(-\omega))$. Since typical inputs of interest can be decomposed into sums of harmonic oscillations, typical design specifications in filter synthesis problems have to do with the frequency response—they prescribe its behavior on a segment $\Delta \in [-\pi, \pi]$. Note that the coefficients $h(l)$ of the filter are real, so that the frequency response possesses an evident symmetry:

$$H(-\omega) = H^*(\omega),$$

where z^* denotes the complex conjugate of a complex number z. Consequently, it suffices to specify the behavior of a frequency response on the segment $[0, \pi]$ only.

The simplest type of design specifications would be to fix a target function $F(\omega)$, $0 \le \omega \le \pi$, and to require H to be as close as possible (e.g., in the uniform metric) to the target. This would result in a Tschebyshev-type problem

$$\min_h \left\{ \max_{0\le\omega\le\pi} |F(\omega) - \sum_{l=0}^{n-1} h(l)\exp\{il\omega\}| \right\}$$

in the (real) design variables $h(0), \dots, h(n-1)$. After discretization in ω, we end up with a simple conic quadratic (or even an LP) program.

The outlined design specifications are aimed at prescribing both what a filter does with the real amplitudes of harmonic oscillations and their initial phases. However, in most applications the only issue of interest is how the filter affects the real amplitudes of harmonic oscillations of different frequencies, not how it shifts the phases. Consequently, typical design specifications prescribe the behavior of $|H(\omega)|$ only, e.g., require from this function to be between two given bounds:

$$L(\omega) \le |H(\omega)| \le U(\omega), \ \ 0 \le \omega \le \pi. \tag{B}$$

For example, when designing a low-pass filter, we wish to reproduce exactly the amplitudes of oscillations with frequencies below a certain level and to suppress oscillations with frequencies higher than another prescribed level, i.e., the specifications are like

$$1-\epsilon \le |H(\omega)| \le 1+\epsilon, 0 \le \omega \le \underline{\omega}, \quad |H(\omega)| \le \epsilon, \overline{\omega} \le \omega \le \pi.$$

When trying to process the constraint of the latter type, we meet with a severe difficulty: $|H(\omega)|$ is not a convex function of our natural design parameters $h(0), \dots, h(n-1)$. There is, however, a way to overcome the difficulty. It turns out that the function $|H(\omega)|^2$ can be linearly parameterized by properly chosen new design parameters, so that lower and upper bounds on $|H(\omega)|^2$ become linear (and thus tractable) constraints on the new design parameters. And of course it is the same to impose bounds on $|H(\omega)|$ or on $|H(\omega)|^2$.

A proper parameterization of the function $R(\omega) \equiv |H(\omega)|^2$ is very simple. We have

$$H(\omega) = \sum_{l=0}^{n-1} h(l)\exp\{il\omega\}$$

$$\Downarrow$$

$$R(\omega) = \left(\sum_{p=0}^{n-1} h(p)\exp\{ip\omega\}\right)\left(\sum_{q=0}^{n-1} h(q)\exp\{-iq\omega\}\right) = \sum_{l=-(n-1)}^{n-1} r(l)\exp\{il\omega\},$$

$$r(l) = \sum_q h(l+q)h(q).$$

The reals $\{r(l)\}_{l=-(n-1)}^{n-1}$ (they are called the *autocorrelation coefficients* of the filter h) are exactly the parameters we need. Note that

$$r(-l) = \sum_q h(\underbrace{-l+q}_{p})h(q) = \sum_p h(p)h(p+l) = r(l),$$

so that

$$R(\omega) = \sum_{l=-(n-1)}^{n-1} r(l)\exp\{il\omega\} = r(0) + 2\sum_{l=1}^{n-1} r(l)\cos(l\omega).$$

Thus, $R(\cdot)$ is just an even trigonometric polynomial of degree $\le n-1$ with real coefficients, and $r(\cdot)$ are, essentially, the coefficients of this trigonometric polynomial.

The function $R(\omega) = |H(\omega)|^2$ is linearly parameterized by the coefficients $r(\cdot)$, which is fine. These coefficients, however, cannot be arbitrary: not every even trigonometric polynomial of a degree $\le n-1$ can be represented as $|H(\omega)|^2$ for certain $H(\omega) = \sum_{l=0}^{n-1} h(l)\exp\{il\omega\}$! The coefficients $r = (r(0), 2r(1), 2r(2), \ldots, 2r(n-1))^T$ of proper even trigonometric polynomials $R(\cdot)$—those that are squares of modulae of frequency responses—form a proper subset $\mathcal{R}$ in $\mathbf{R}^n$, and to handle constraints of the type (B), we need a tractable representation of $\mathcal{R}$. Such a representation does exist, due to the following fundamental fact.

Proposition 4.6.1. Spectral factorization theorem. *A trigonometric polynomial*

$$R(\omega) = a_0 + \sum_{l=1}^{n-1}(a_l\cos(l\omega) + b_l\sin(l\omega))$$

with real coefficients $a_0, \ldots, a_{n-1}, b_1, \ldots, b_{n-1}$ can be represented as

$$\left|\sum_{l=0}^{n-1} h(l)\exp\{il\omega\}\right|^2 \qquad (*)$$

for properly chosen complex $h(0), h(1), \ldots, h(n-1)$ if and only if $R(\omega)$ is nonnegative on $[-\pi, \pi]$. An even trigonometric polynomial $R(\omega)$ of degree $\le n-1$ can be represented in the form (*) with real $h(0), \ldots, h(n-1)$ if and only if it is nonnegative on $[-\pi, \pi]$ (or, which is the same, on $[0, \pi]$).

Postponing the proof of Proposition 4.6.1 until the end of this section, let us look first at the consequences. The proposition says that the set $\mathcal{R} \in \mathbf{R}^n$ of the coefficients of those even trigonometric polynomials of degree $\le n-1$ that are squares of modulae of frequency responses of filters of order n is exactly the set of coefficients of those even trigonometric polynomials of degree $\le n-1$ that are nonnegative on $[-\pi, \pi]$. Consequently, this set is SDr with an explicit semidefinite representation (example 21c, section 4.2). Thus, passing from our original design variables $h(0), \ldots, h(n-1)$ to the new design variables $r \in \mathcal{R}$, we make the design specifications of the form (B) a (semi-infinite) system of linear constraints on the design variables varying in a SDr set. As a result, we get a possibility to handle numerous filter synthesis problems with design specifications of the type (B) via SDP. Let us look at a couple of examples.

Example 1: Low-pass filter. Assume we are given a number of (possibly overlapping) segments $\Delta_k \subset [0, \pi]$, $k = 1, \ldots, K$, along with nonnegative continuous functions $S_k(\omega)$,

$T_k(\omega)$ ($S_k(\omega) \le T_k(\omega)$) defined on these segments, and our goal is to design a filter of a given order n with $|H(\omega)|^2$ being at every segment Δ_k as close as possible to the strip between S_k and T_k. Since a natural measure of closeness in filter synthesis problems is the relative closeness, we can pose the problem as

$$\min_{\epsilon, H(\cdot)} \left\{ \epsilon : \frac{1}{(1+\epsilon)} S_k(\omega) \le |H(\omega)|^2 \le (1+\epsilon) T_k(\omega) \quad \forall \omega \in \Delta_k \quad \forall k = 1, \ldots, K \right\}. \quad \text{(P)}$$

For example, when dealing with two nonoverlapping segments $\Delta_1 = [0, \underline{\omega}]$ and $\Delta_2 = [\overline{\omega}, \pi]$ and setting $S_1 \equiv T_1 \equiv 1$, $S_2 \equiv 0$, $T_2 \equiv \beta$, with small positive β, we come to the problem of designing a low-pass filter: $|H(\omega)|$ should be as close to 1 as possible in Δ_1 and should be small in Δ_2.

In terms of the autocorrelation coefficients, problem (P) reads

$$\begin{array}{lll}
\text{minimize} & \epsilon & \\
\text{s.t.} & & \\
\text{(a)} & \delta S_k(\omega) \le R(\omega) \equiv r(0) + 2\sum_{l=1}^{n-1} r(l)\cos(l\omega) \le (1+\epsilon) T_k(\omega) & \forall \omega \in \Delta_k, \\
 & & k = 1, \ldots, K; \\
\text{(b)} & \delta(1+\epsilon) \ge 1, & \\
\text{(c)} & \delta, \epsilon \ge 0, & \\
\text{(d)} & r \in \mathcal{R}. &
\end{array} \quad \text{(P}')$$

Indeed, (a)–(c) say that

$$\frac{1}{1+\epsilon} S_k(\omega) \le R(\omega) \le (1+\epsilon) T_k(\omega), \ \omega \in \Delta_k, \ k = 1, \ldots, K,$$

while the role of the constraint (d) is to express the fact that $R(\cdot)$ comes from certain filter of order n.

Problem (P′) is not exactly a semidefinite program—the obstacle is that the constraints (a) are semi-infinite. To overcome this difficulty, we can use discretization in ω (i.e., can replace each segment Δ_k by a dense finite set of its points), thus approximating (P) by a semidefinite program. In many cases we can even avoid approximation. One such case is when all S_k and T_k are trigonometric polynomials. As we know from example 21c, section 4.2, the restriction that a trigonometric polynomial $R(\omega)$ majorates (or is majorated by) another trigonometric polynomial is an SDr constraint on the coefficients of the polynomials, so that in the case in question the constraints (a) are SDr restrictions on r, δ, ϵ.

Another formulation of the low-pass filter problem is obtained when instead of minimizing the relative uniform distance between $|H(\omega)|^2$ and given targets we minimize the relative $\|\cdot\|_2$-distance. A natural form of the latter problem is

$$\begin{array}{lll}
\text{minimize} & \epsilon & \\
\text{s.t.} & & \\
 & \frac{1}{1+\epsilon_k(\omega)} S_k(\omega) \le |H(\omega)|^2 \le (1+\epsilon_k(\omega)) T_k(\omega), \ \omega \in \Delta_k, & \\
 & k = 1, \ldots, K; & \\
 & \sqrt{\frac{1}{|\Delta_k|} \int_{\Delta_k} \epsilon_k^2(\omega) d\omega} \le \epsilon, \ k = 1, \ldots, K. &
\end{array}$$

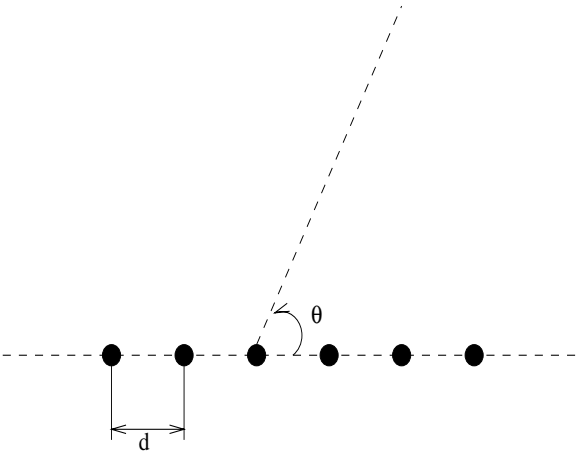


Figure 4.9. *Linear array of equidistant harmonic oscillators.*

After discretization in ω—replacing Δ_k by a finite set $\Omega_k \subset \Delta_k$—we can pose the problem as the semidefinite program

$$
\begin{array}{l}
\text{minimize} \quad \epsilon \\
\text{s.t.} \\
\qquad \delta_k(\omega) S_k(\omega) \le R(\omega) \equiv r(0) + 2\sum_{l=1}^{n-1} r(l)\cos(l\omega) \;\le\; (1+\epsilon_k(\omega)) T_k(\omega)\ \forall \omega \in \Omega_k, \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k = 1, \ldots, K; \\
\qquad\qquad\qquad\qquad\qquad \delta_k(\omega)(1+\epsilon_k(\omega)) \;\ge\; 1,\ \forall \omega \in \Omega_k, \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k = 1, \ldots, K; \\
\qquad\qquad\qquad\qquad\qquad \delta_k(\omega),\ (1+\epsilon_k(\omega)) \;\ge\; 0,\ \forall \omega \in \Omega_k, \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k = 1, \ldots, K; \\
\qquad\qquad\qquad\qquad \sqrt{\frac{1}{\mathrm{Card}(\Omega_k)} \sum_{\omega \in \Omega_k} \epsilon_k^2(\omega)} \;\le\; \epsilon,\ k = 1, \ldots, K; \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad r \;\in\; \mathcal{R}.
\end{array}
$$

Example 2. Synthesis of array of antennae. Consider a linear array of antennae (see section 1.2.4) made up of n equidistantly placed harmonic oscillators in the plane XY (Fig. 4.9). We can easily see that the diagram of the array depends on the angle θ between the direction in question and the line where the oscillators are placed and is given by

$$
Z(\theta) = \sum_{l=0}^{n-1} z_l \exp\{-il\Omega(\theta)\}, \quad \Omega(\theta) = -\frac{2\pi d}{\lambda}\cos\theta,
$$

where $z_0, z_1, \ldots, z_{n-1}$ are the (complex) amplification coefficients of the oscillators and λ is the wavelength.

In our previous antenna synthesis considerations, we were interested in the case when the design specifications were aimed to get a diagram as close as possible to a given target diagram $Z_*(\theta)$. In fact, what is of interest in many antenna synthesis problems is only the modulus $|Z(\theta)|$ of the resulting diagram ($|Z(\theta)|^2$ is responsible for the energy sent by antenna in a direction θ). In these situations we are interested in a prescribed behavior of the function $|Z(\theta)|$. Here again Proposition 4.6.1 is the key for handling the problem via convex optimization. Indeed, defining the function

$$H(\omega) = \sum_{l=0}^{n-1} z_l \exp\{il\omega\},$$

we get a frequency response of a complex filter $h = \{h(l) = z_l\}_{l=0}^{n-1}$ such that

$$Z(\theta) = H(\Omega(\theta)).$$

It follows that to impose restrictions, like upper and lower bounds, on the function $|Z(\theta)|$, $0 \le \theta \le \pi$, is the same as to impose bounds of the same type on the function $|H(\omega)|$ in the segment Δ of values taken by $\Omega(\theta)$ when θ varies from 0 to π. Assuming (this is normally the case) that $\lambda > 2d$, we observe that the mapping $\theta \mapsto \Omega(\theta)$ is a one-to-one from the segment $[0, \pi]$ to a certain segment $\Delta \subset [-\pi, \pi]$, so that design specifications on $|Z(\theta)|$ can be easily converted to design specifications on $|H(\theta)|$. For example, building a diagram with $|Z(\theta)|$ as close as possible to given stripes can be formulated as the problem

$$\begin{aligned} &\min_{\epsilon, r(\cdot)} \left\{\epsilon : \tfrac{1}{1+\epsilon} S_k(\omega) \le R(\omega) \le (1+\epsilon) T_k(\omega) \quad \forall \omega \in \Delta_k,\ k = 1, \ldots, K\right\}, \\ &R(\omega) = |H(\omega)|^2 = \sum_{l=-(n-1)}^{n-1} r(l) \exp\{il\omega\}. \end{aligned} \tag{$\mathrm{P_C}$}$$

The only differences between ($\mathrm{P_C}$) and problem (P) we investigated in Example 1 is that now the autocorrelation coefficients correspond to *complex* amplification coefficients z_l—the actual design variables—through the relation

$$r(k) = \sum_{l=0}^{n-1} z_l^* z_{l+k}$$

and are therefore complex; clearly, these complex coefficients possess the symmetry

$$r(-k) = r^*(k), \ |k| \le n-1,$$

reflecting the fact that the function $R(\omega)$ is real-valued. This function, as it is immediately seen, is just a real trigonometric polynomial (now not necessarily even) of degree $\le n-1$:

$$R(\omega) = \rho(0) + \sum_{l=1}^{n-1} (\rho(2l-1)\cos(l\omega) + \rho(2l)\sin(l\omega))$$

with real vector of coefficients $\rho = (\rho(0), \dots, \rho(2n-2))^T \in \mathbf{R}^{2n-1}$. The vector of these coefficients can be treated as our new design vector. Invoking again Proposition 4.6.1, we see that such a vector gives rise to a function $R(\omega)$ which indeed is of the form $|H(\omega)|^2$, $H(\omega) = \sum_{l=0}^{n-1} r_l \exp\{il\omega\}$, if and only if the trigonometric polynomial $R(\cdot)$ is nonnegative on $[-\pi, \pi]$. As we remember from example 21c, section 4.2, the set $\mathcal{C}$ of the vectors of coefficients ρ of this type is SDr.

In view of the outlined observations, problem $(\mathrm{P_C})$ can be posed as a semi-infinite semidefinite program in exactly the same way as problem (P), and this semi-infinite program can be approximated by (or sometimes is equivalent to) a usual semidefinite program. For example, approximating segments Δ_k by finite grids Ω_k, we approximate $(\mathrm{P_C})$ by the semidefinite program

$$
\begin{array}{ll}
\text{minimize} & \epsilon \\
\text{s.t.} & \\
\multicolumn{2}{c}{\delta S_k(\omega) \le R(\omega) \equiv \rho(0) + \sum_{l=1}^{n-1} (\rho(2l-1)\cos(l\omega) + \rho(2l)\sin(l\omega)) \quad \le \quad (1+\epsilon)T_k(\omega)} \\
\multicolumn{2}{r}{\forall \omega \in \Omega_k, k = 1, \dots, K;} \\
\multicolumn{2}{c}{\delta(1+\epsilon) \ge 1,} \\
\multicolumn{2}{c}{\delta, \epsilon \ge 0,} \\
\multicolumn{2}{c}{\rho \in \mathcal{C},}
\end{array}
$$

in the design variables δ, ϵ and $\rho = (\rho(0), \dots, \rho(2n-2))^T \in \mathbf{R}^{2n-1}$.

Proof of Proposition 4.6.1. Let us first prove that a real trigonometric polynomial

$$R(\omega) = c_0 + \sum_{l=1}^{n-1} (a_l \cos(l\omega) + b_l \sin(l\omega))$$

can be represented as $|\sum_{l=0}^{n-1} h(l)\exp\{il\omega\}|^2$ with some complex coefficients $h(l)$ if and only if $R(\cdot)$ is nonnegative on $[-\pi, \pi]$. The necessity is evident, so let us focus on the sufficiency. Thus, assume that R is nonnegative, and let us prove that R admits the required decomposition.

1. It suffices to prove the announced statement in the case when $R(\omega)$ is strictly positive on $[-\pi, \pi]$ rather than merely nonnegative. Indeed, assume that our decomposition is possible for positive trigonometric polynomials. Given a nonnegative polynomial R, let us apply our assumption to the positive trigonometric polynomial $R(\omega) + \epsilon$, $\epsilon > 0$:

$$R(\omega) + \epsilon = \left| \sum_{l=0}^{n-1} h_\epsilon(l) \exp\{il\omega\} \right|^2 .$$

From this representation it follows that

$$c_0 + \epsilon = \sum_{l=0}^{n-1} |h_\epsilon(l)|^2,$$

whence the coefficients $h_\epsilon(l)$ remain bounded as $\epsilon \to +0$. Taking as h an accumulation point of the vectors h_ϵ as $\epsilon \to +0$, we get

$$R(\omega) = \left| \sum_{l=0}^{n-1} h(l) \exp\{il\omega\} \right|^2,$$

as required.

2. Thus, it suffices to consider the case when R is a positive trigonometric polynomial. And of course we may assume that the degree of R is exactly $n-1$, i.e., that $a_{n-1}^2 + b_{n-1}^2 > 0$.

We can rewrite R in the form

$$R(\omega) = \sum_{l=-(n-1)}^{n-1} r(l) \exp\{il\omega\}; \tag{4.6.74}$$

since R is real-valued, we have

$$r(l) = r^*(-l), \quad |l| \le n-1. \tag{4.6.75}$$

Now consider the polynomial

$$P(z) = z^{(n-1)} \left(\sum_{l=-(n-1)}^{n-1} r(l) z^l \right).$$

This polynomial is of degree $2(n-1)$, is nonzero at $z = 0$, and has no zeros on the unit circumference (since $|P(\exp\{i\omega\})| = R(\omega)$). Moreover, from (4.6.75) it immediately follows that if λ is a root of $P(z)$, then also $(\lambda^*)^{-1}$ is a root of the polynomial of exactly the same multiplicity as λ. It follows that the roots of the polynomial P can be separated into two nonintersecting groups: $(n-1)$ roots λ_l, $l = 1, \dots, n-1$, inside the unit circle and $(n-1)$ roots $1/\lambda_l^*$ outside the circle. Thus,

$$P(z) = \alpha \left[\prod_{l=1}^{n-1} (z - \lambda_l) \right] \left[\prod_{l=1}^{n-1} (z - 1/\lambda_l^*) \right].$$

Moreover, we have

$$R(0) \quad = \quad P(1) = \alpha \prod_{l=1}^{n-1} \left[(1-\lambda_l)(1-1/\lambda_l^*) \right] = \alpha(-1)^{n-1} \left[\prod_{l=1}^{n-1} |1-\lambda_l|^2 \right] \left[\prod_{l=1}^{n-1} \lambda_l^* \right]^{-1},$$

and since $R(0) > 0$, the number

$$\alpha(-1)^{n-1} \left[\prod_{l=1}^{n-1} \lambda_l^* \right]^{-1}$$

is positive. Denoting this number by β^2, let us set

$$H(\omega) = \beta \prod_{l=1}^{n-1} (\exp\{i\omega\} - \lambda_l) \equiv \sum_{l=0}^{n-1} h(l) \exp\{il\omega\}.$$

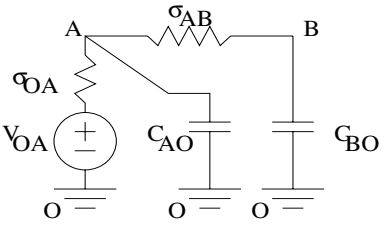


Figure 4.10. *A simple circuit. Element OA: outer supply of voltage V_{OA} and resistor with conductance σ_{OA}. Element AO: capacitor with capacitance C_{AO}. Element AB: resistor with conductance σ_{AB}. Element BO: capacitor with capacitance C_{BO}.*

Then

$$
\begin{aligned}
&|H(\omega)|^2 \\
&\quad = \beta^2 \left| \prod_{l=1}^{n-1} (\exp(i\omega) - \lambda_l)(\exp\{-i\omega\} - \lambda_l^*) \right| \\
&\quad = \beta^2 \left| \exp\{-i(n-1)\omega\}(-1)^{n-1} \left[\prod_{l=1}^{n-1} \lambda_l^* \right] \left[\prod_{l=1}^{n-1} \left[(\exp\{i\omega\} - \lambda_l)(\exp\{i\omega\} - 1/\lambda_l^*) \right] \right] \right| \\
&\quad = \beta^2 \left| \exp\{-i(n-1)\omega\}(-1)^{n-1} \alpha^{-1} \left[\prod_{l=1}^{n-1} \lambda_l^* \right] P(\exp\{i\omega\}) \right| \\
&\quad = |\exp\{-i(n-1)\omega\} P(\exp\{i\omega\})| \\
&\quad = R(\omega),
\end{aligned}
$$

as required.

3. To complete the proof of Proposition 4.6.1, it suffices to verify that if $R(\omega)$ is an even nonnegative trigonometric polynomial, then the coefficients $h(l)$ in the representation $R(\omega) = |\sum_{l=1}^{n-1} h(l) \exp\{il\omega\}|^2$ can be chosen real. But this is immediate: if $R(\cdot)$ is even, the coefficients $\rho(l)$ in (4.6.74) are real, so that $P(z)$ is a polynomial with real coefficients. Consequently, the complex numbers met among the roots $\lambda_1, \ldots, \lambda_{n-1}$ are met only in conjugate pairs, both members of a pair being roots of the same multiplicity. Consequently, the function $H(\omega)$ is $\hat{h}(\exp\{i\omega\})$, where $\hat{h}(\cdot)$ is a real polynomial, as claimed. □

4.7 Applications V: Design of chips

Consider an RC-electric circuit, i.e., a circuit comprising three types of elements: resistors, capacitors, and resistors in a series combination with outer sources of voltage (see Fig. 4.10).[33] For example, a chip is, electrically, a complicated circuit comprising elements of the indicated type. When designing chips, the following characteristics are of primary importance:

[33] The model presented in this section originates from L. Vanderberghe, S. Boyd, and A. El Gamal, *Optimizing dominant time constant in RC circuits*, IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, 17 (1998), pp. 110–125.

• *Speed.* In a chip, the outer voltages are switching at a certain frequency from one constant value to another. Every switch is accompanied by a transition period. During this period, the potentials and currents in the elements are moving from their previous values (corresponding to the static steady state for the old outer voltages) to the values corresponding to the new static steady state. Since there are elements with inertia—capacitors—this transition period takes some time.[34] To ensure stable performance of the chip, the transition period should be much less than the time between subsequent switches in the outer voltages. Thus, the duration of the transition period is responsible for the speed at which the chip can perform.

• *Dissipated heat.* Resistors in the chip dissipate heat that should be eliminated; otherwise the chip will not function. This requirement is very serious for modern high-density chips. Thus, a characteristic of vital importance is the dissipated heat power.

The two objectives—high speed (i.e., a small transition period) and small dissipated heat—usually are conflicting. As a result, a chip designer faces the tradeoff problem of how to get a chip with a given speed and with the minimal dissipated heat. We are about to demonstrate that the ensuing optimization problem belongs to the semidefinite universe.

4.7.1 Building the model

A circuit

Mathematically, a circuit can be represented as a graph; the nodes of the graph correspond to the points where elements of the circuit are linked to each other, and the arcs correspond to the elements themselves. We may assume that the nodes are enumerated, $1, 2, \ldots, N$, and that the arcs are (somehow) oriented, so that every arc γ links its origin node $s(\gamma)$ with its destination node $d(\gamma)$. Note that we do not forbid parallel arcs—distinct arcs linking the same pairs of nodes. For example, for the circuit depicted in Fig. 4.10 we could orient both arcs linking the ground O with the point A (one with resistor and one with capacitor) in the same way, thus creating two parallel arcs. Let us denote by Γ the set of all arcs of our graph (all elements of our circuit), and let us equip an arc $\gamma \in \Gamma$ with three parameters $v_\gamma, c_\gamma, \sigma_\gamma$ (outer voltage, capacitance, conductance) as follows:

- For an arc γ representing a resistor, σ_γ is the conductance of the resistor, $c_\gamma = v_\gamma = 0$.
- For an arc γ representing a capacitor, c_γ is the capacitance of the capacitor, $v_\gamma = \sigma_\gamma = 0$.
- For an arc γ of the type "outer source of voltage—resistor," σ_γ is the conductance of the resistor, v_γ is the outer voltage, and $c_\gamma = 0$.

Transition period

Let us build a model for the duration of a transition period. The question we are addressing is, Assume that before instant 0 the outer voltages were certain constants and the circuit was

[34] From a purely mathematical viewpoint, the transition period takes infinite time—the currents or voltages approach asymptotically the new steady state but never actually reach it. From the engineering viewpoint, however, we may think that the transition period is over when the currents or voltages become close enough to the new static steady state.

in the corresponding static steady state. At instant $t = 0$ the outer voltages jump to new values v_γ and remain at these values. What will happen with the circuit? The answer is given by the Kirchoff laws and is as follows. Let $u_i(t)$, $t \geq 0$, be the potentials at the nodes $i = 1, \ldots, N$, and let $I_\gamma(t)$ be the currents in arcs $\gamma \in \Gamma$ at time t.

The first law of Kirchoff says that

$$\begin{array}{lcll} I_\gamma(t) & = & \sigma_\gamma[u_{s(\gamma)}(t) - u_{d(\gamma)}(t)] & \text{if } \gamma \text{ is a resistor;} \\ I_\gamma(t) & = & c_\gamma \frac{d}{dt}[u_{s(\gamma)}(t) - u_{d(\gamma)}(t)] & \text{if } \gamma \text{ is a capacitor;} \\ I_\gamma(t) & = & \sigma_\gamma[u_{s(\gamma)}(t) - u_{d(\gamma)}(t) - v_\gamma] & \text{if } \gamma \text{ is an outer voltage followed by a resistor.} \end{array}$$

With our rule for assigning parameters to the arcs, we can write these relations in the unified form

$$I_\gamma(t) = \sigma_\gamma[u_{s(\gamma)}(t) - u_{d(\gamma)}(t) - v_\gamma] - c_\gamma \frac{d}{dt}[u_{s(\gamma)}(t) - u_{d(\gamma)}(t)]. \tag{4.7.76}$$

The second law of Kirchoff says that for every node i, the sum of currents in the arcs entering the node should be equal to the sum of currents in the arcs leaving the node. To represent this law conveniently, we introduce the *incidence matrix* P of our circuit as follows:

> *Incidence matrix.* The columns of the matrix P are indexed by the nodes $1, \ldots, N$, and the rows are indexed by the arcs $\gamma \in \Gamma$. The row P_γ corresponding to an arc γ is filled with zeros, except for two entries: the one in column $s(\gamma)$ $(+1)$ and the one in column $d(\gamma)$ (-1).

With this formalism, the second law of Kirchoff is

$$P^T I(t) = 0, \tag{4.7.77}$$

where $I(t)$ is the vector with the entries $I_\gamma(t)$, $\gamma \in \Gamma$. With the same formalism, (4.7.76) can be written as

$$I(t) = \Xi P u(t) + \vartheta P \frac{d}{dt} u(t) - \Xi v, \tag{4.7.78}$$

where

- $u(t)$ is the N-dimensional vector comprising the potentials of the nodes $u_i(t)$, $i = 1, \ldots, N$;
- Ξ is the diagonal $M \times M$ matrix (M is the number of arcs) with diagonal entries σ_γ, $\gamma \in \Gamma$;
- v is the M-dimensional vector comprised of outer voltages v_γ, $\gamma \in \Gamma$; and
- ϑ is the diagonal $M \times M$ matrix with the diagonal entries c_γ, $\gamma \in \Gamma$.

Multiplying (4.7.78) by P^T and taking into account (4.7.77), we get

$$\left[P^T \vartheta P\right] \frac{d}{dt} u(t) = -P^T \Xi P u(t) + P^T \Xi v. \tag{4.7.79}$$

Now, potentials are quantities defined up to a common additive constant. What makes physical sense are not the potentials themselves but their differences. To avoid the resulting nonuniqueness of our solutions, we may enforce one of the potentials, say, $u_N(t)$, to be identically zero (node N is the ground). Let

$$\begin{array}{rcl} C & = & ([P^T\vartheta P]_{ij})_{i,j\le N-1}, \\ S & = & ([P^T\Xi P]_{ij})_{i,j\le N-1}, \\ R & = & ([P^T\Xi])_{i\gamma})_{i\le N-1,\gamma\in\Gamma} \end{array} \tag{4.7.80}$$

be the corresponding submatrices of the matrices participating in (4.7.79). Denoting by $w(t)$ the $(N-1)$-dimensional vector comprising the first $N-1$ entries of the vector of potentials $u(t)$ (recall that the latter vector is normalized by $u_N(t) \equiv 0$), we can rewrite (4.7.80) as

$$C\frac{d}{dt}w(t) = -Sw(t) + Rv. \tag{4.7.81}$$

Note that due to their origin, the matrices Ξ and ϑ are diagonal with nonnegative diagonal entries, i.e., they are symmetric positive semidefinite. Consequently, the matrices C, S also are symmetric positive semidefinite. In the sequel, we make the following assumption:

Assumption (W). *The matrices C and S are positive definite.*

The assumption in fact is a reasonable restriction on the topology of the circuit: when deleting all capacitors, the resulting net made up of resistors should be connected, and similarly for the net of capacitors obtained after the resistors are deleted. With this assumption, (4.7.81) is equivalent to a system of linear ordinary differential equations with constant coefficients and a constant right-hand side:

$$\frac{d}{dt}w(t) = -C^{-1}Sw(t) + C^{-1}Rv. \tag{4.7.82}$$

Now, the matrix of the system is similar to the negative definite matrix:

$$-C^{-1}S = C^{-1/2}[-C^{-1/2}SC^{-1/2}]C^{1/2}.$$

Consequently, the eigenvalues $(-\lambda_i)$, $i = 1, \dots, N-1$, of the matrix $C^{-1}S$ of the system (4.7.82) are negative, there exists a system $e_1, \dots, e_{N-1}$ of linearly independent eigenvectors associated with these eigenvalues, and the solution to (4.7.82) is of the form

$$w(t) = w_*(v) + \sum_{i=1}^{N-1} \kappa_i \exp\{-\lambda_i t\} e_i,$$

where

$$w_*(v) = S^{-1}Rv \tag{4.7.83}$$

is the vector comprising the static steady-state potentials associated with the outer voltages v, and κ_i are certain constants coming from the initial state $w(0)$ of (4.7.82). From (4.7.78) we get a representation of the same structure also for the currents,

$$I(t) = I_*(v) + \sum_{i=1}^{N-1} \chi_i \exp\{-\lambda_i\} f_i, \tag{4.7.84}$$

with certain M-dimensional vectors f_i.

We see that during the transition period, the potentials and the currents approach the steady state exponentially fast, the rate of convergence being governed by the quantities λ_i. The most unfavorable rate of convergence to the steady state is given by the smallest of the λ_i's:

$$\hat{\lambda} = \min_i \lambda_i(C^{-1/2}SC^{-1/2}) \tag{4.7.85}$$

The quantity $\hat{\lambda}$ can be treated as a (perhaps rough) measure of the speed of the circuit. The larger the quantity, the shorter the transition period. For reasonable initial conditions, the potentials and the currents in the circuit will become very close to their steady-state values after a period which is a moderate constant times the quantity $1/\hat{\lambda}$. It was proposed by Boyd to treat $1/\hat{\lambda}$ as the characteristic time constant of the underlying circuit and to model restrictions such as "the duration of a transition period in a circuit should be at most certain bound" as "the time constant of the circuit should be at most certain bound."[35] Now, it is easy to understand what is, mathematically, $\hat{\lambda}$—it is nothing but the smallest eigenvalue $\lambda_{\min}(S : C)$ of the pencil $[C, S]$ (cf. section 4.4.1). Consequently, in Boyd's methodology the requirement "the speed of the circuit to be designed should be not less than..." is modeled as the restriction $1/\lambda_{\min}(S : C) \leq T$ or, equivalently, as the matrix inequality

$$S - \kappa C \succeq 0 \qquad [\kappa = T^{-1}]. \tag{4.7.86}$$

As we shall see in a while, in typical chip design problems S and C are affine functions of the design variables, so that a design requirement on the speed of the chip can be expressed by an LMI.

Dissipated heat. Concerning issues related to dissipated heat, one could be interested in the heat power dissipated in the steady state corresponding to given outer voltages and the heat dissipated during a transition.

We shall see that imposing restrictions on the steady-state dissipated heat power leads to an intractable computational problems, while restrictions on the heat dissipated in a transition period in some meaningful cases (although not always) lead to semidefinite programs.

Bad news on steady-state dissipated heat power. Physics says that the dissipated heat power in the steady state corresponding to outer voltages v is

$$H = \sum_{\gamma\in\Gamma}(I_*(v))_\gamma[(u_*(v))_{s(\gamma)} - (u_*(v))_{d(\gamma)} - v_\gamma] = [I_*(v)]^T(Pu_*(v) - v),$$

where $I_*(v)$ and $u_*(v)$ are the steady-state currents and potentials associated with v.[36] The formula expresses the very well known rule: The heat power dissipated by a resistor is the product of the current and the voltage applied to the resistor.

[35]For many years, engineers were (and they still are) using a more precise measure of speed—the Elmore constant. A disadvantage of the Elmore constant is that it can be efficiently computed only for a restricted family of circuits. In contrast to this, Boyd's time constant is computationally tractable.

[36]We assume that the potential of the ground (node N) is 0; with this convention, the notion of steady-state potentials becomes well defined.

In fact H can be derived from the following variational principle.

Given a circuit satisfying Assumption (W) *and a vector of outer voltages v, consider the quadratic form*

$$G(u) = (Pu - v)^T \Xi (Pu - v)$$

of N-dimensional vector u. The heat power dissipated by the circuit at the static steady state associated with v is the minimum value of this quadratic form over $u \in \mathbf{R}^N$.

Indeed, G depends on the differences of coordinates of u only, so that its minimum over all u is the same as its minimum over u of the form $u = \binom{w}{0}$. Regarded as a function of $(N-1)$-dimensional vector w rather than of u, the quadratic form becomes

$$\hat{G}(w) = w^T S w - 2w^T R v + v^T \Xi v.$$

The minimizer of $\hat{G}(\cdot)$ is given by $w_* = S^{-1} R v$, which is exactly the vector of steady-state potentials at the nodes (cf. (4.7.83)). Thus, the vector of steady-state potentials $u_*(v)$ is a minimizer of $G(\cdot)$. The value of G at this minimizer is $[\Xi(Pu_*(v) - v)]^T (Pu_*(v) - v)$, and the vector $\Xi(Pu_*(v) - v)$ is exactly the vector of steady-state currents; see (4.7.76). Thus, the optimal value of G is $I_*^T(v)(Pu_*(v) - v)$, which is precisely H.

The variational principle says that an upper bound restriction $H \le h$ on the steady-state dissipated heat power is

$$\mathcal{H}(S, v) \equiv \min_u (Pu - v)^T S (Pu - v) \le h.$$

The left-hand side in this inequality is a concave function of S (as a minimum of linear functions of S). Therefore the above restriction defines a nonconvex set $\mathcal{F}_v$ of feasible matrices S. If S depends affinely on free design parameters, as is normally the case, the nonconvexity of $\mathcal{F}_v$ implies the nonconvexity of the feasible set in the space of design parameters and hence leads to an intractable optimization problem.

Note that if we were designing a heater rather than a chip (i.e., were interested to get at least a prescribed heat power dissipation), the restriction would fall into the realm of convex (specifically, semidefinite) programming (cf. simple lemma).

Good news on heat dissipated in transition period. The heat dissipated during a transition from the old steady state associated with outer voltages v_- to the new steady state associated with outer voltages v_+ is, in general, a senseless notion. Indeed, the transition period, rigorously speaking, is infinite. If the new steady state is active (i.e., not all of the corresponding steady-state currents are zero), then the heat dissipation power during the transition will approach a positive quantity (the steady-state dissipated heat power for the new steady state), and the entire power energy dissipated during the (infinite!) transition period will be $+\infty$. There is, however, a case where this difficulty does not occur and we may speak about the heat energy dissipated during the transition—this is the case when the new steady-state currents are zero. In this case, the dissipated heat power decays to 0 exponentially fast, the decay rate being (bounded by) Boyd's time constant, and so it makes

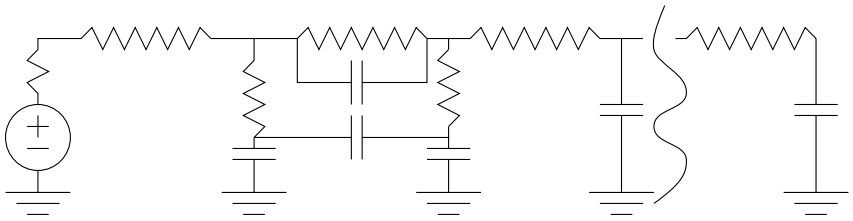


Figure 4.11. *A simple RC-circuit.*

sense to speak about the heat dissipated during the transition. Now, there is a particular (but quite important) class of simple RC-circuits in which the currents at a static steady state indeed are zero—circuits of the type shown in Fig. 4.11. In such a circuit, there is a single source of outer voltage, the resistors form a connected net that starts at one of the poles of the source and does not reach the other pole of the source (the ground), and each capacitor either links a node incident to a resistor and the ground (capacitor of type I) or links two nodes incident to resistors (capacitor of type II). Here the steady-state currents are clearly zero, whence also

$$Pu_*(v) = v \quad \forall v. \tag{4.7.87}$$

Moreover, the steady-state potentials at all nodes incident to resistors are equal to the magnitude of the outer voltage, and the voltage at a capacitor either equals to the magnitude of the outer voltage (for capacitors of type I) or equals zero (for capacitors of type II).

For a simple circuit, the heat energy dissipated during a transition can be found as follows. Assume that the outer voltage switches from its old value v_- of a magnitude μ_- to its new value v_+ of magnitude μ_+.[37] Let us compute the heat dissipated during the transition period starting at time 0. Denoting by $u(t)$, $I(t)$, $H(t)$ the potentials, the currents, and the dissipated heat, respectively, at time $t \geq 0$, and applying (4.7.76), we get

$$\begin{aligned} H(t) &= (Pu(t) - v_+)^T \Xi (Pu(t) - v_+)^T \\ &= [P(u(t) - u_*(v_+))]^T \Xi [P(\underbrace{u(t) - u_*(v_+)}_{\Delta(t)})] \\ &\qquad\qquad \text{(see (4.7.87))} \\ &= \Delta^T(t) P^T \Xi P \Delta(t). \end{aligned}$$

Recalling that $\Delta(t) = \binom{\delta(t)}{0}$ (all our potentials are normalized by the requirement that the potential of the Nth node is zero), we can rewrite the expression for $H(t)$ as

$$H(t) = \delta^T(t) S \delta(t).$$

By its origin, $\delta(t)$ satisfies the homogeneous version of (4.7.82), whence

$$\frac{1}{2}\frac{d}{dt}\left[\delta^T(t) C \delta(t)\right] = -\delta^T(t) S \delta(t),$$

[37] Although we now are speaking about a single-source circuit, it would be bad to identify the magnitude of the outer voltage and the voltage itself. According to our formalism, an outer voltage is a vector with the coordinates indexed by arcs of the circuit. A coordinate of this vector is the physical magnitude of the outer source inside an arc. Thus, μ_- is a number and v_- is a vector with all but one zero coordinates; the only nonzero coordinate is equal to μ_- and corresponds to the arc containing the outer source.

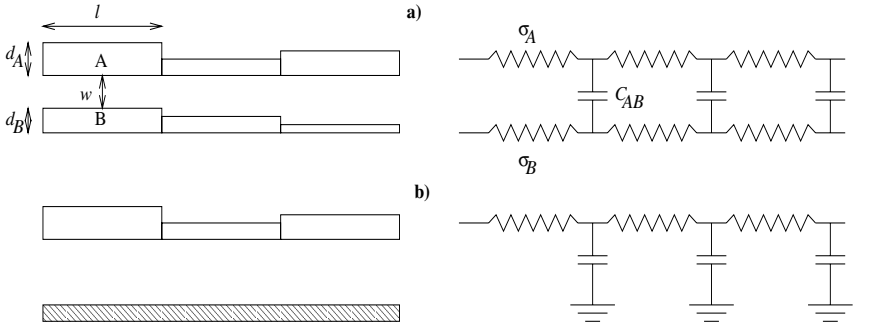


Figure 4.12. *Wires (left) and the equivalent RC-structure (right). (a) a pair of two neighboring wires; (b) a wire and the substrate.*

so that

$$H(t) = -\frac{1}{2}\frac{d}{dt}\left[\delta^T(t)C\delta(t)\right];$$

therefore the heat **H** dissipated during the transition is

$$\mathbf{H} = -\frac{1}{2}\int_0^\infty \frac{1}{2}\frac{d}{dt}\left[\delta^T(t)C\delta(t)\right]dt = \frac{1}{2}\delta^T(0)C\delta(0).$$

From the definition of $\delta(\cdot)$ it is clear that $\delta(0) = (\mu_- - \mu_+)e$, where e is the $(N-1)$-dimensional vector of ones. Thus, in the case of a simple circuit the heat dissipated during the transition is

$$\mathbf{H} = \frac{(\mu_- - \mu_+)^2}{2}e^T Ce, \quad e = (1, \ldots, 1)^T \in \mathbf{R}^{N-1}, \tag{4.7.88}$$

which is a linear function of C.

4.7.2 Wire sizing

Modern submicron chips can be modeled as RC-circuits. In these circuits, the resistors, physically, are the interconnecting wires (and the transistors), and the capacitors model the capacitances between pairs of wires or a wire and the substrate. After the topology of a chip and the placement of its elements on a substrate are designed, engineers start to define the widths of the wires, and this is the stage where the outlined models could be used. To pose the wire sizing problem as an optimization program, one may think of a wire as being partitioned into rectangular segments of a prescribed length and then treat the widths of these segments as design variables. A pair of two neighboring wires (or a wire and the substrate) can be modeled by an RC-structure, as shown in Fig. 4.12. A nice feature of this model is that both the conductances of the resulting resistors and the capacitances of the resulting capacitors turn out to be linear functions of our design parameters—the widths of

the segments or, which is the same, of the areas of the segments (the lengths of the segments are fixed!). For example, for the RC-structure depicted in Fig. 4.12(a) one has

$$\begin{array}{rcl} c_{AB} & = & \kappa_{A,B}(s_A + s_B), \\ \sigma_A & = & \kappa_A s_A, \end{array}$$

where s_A, s_B are the areas of the corresponding segments. The coefficients $\kappa_{A,B}, \kappa_A$ depend on several parameters (e.g., on distances between the wires), but all these parameters are already set at the stage of design we are speaking about. Thus, in the wire sizing problem the matrices Ξ and ϑ, and therefore the matrices S, C, R as well, are affine functions of the design vector x comprising the areas

$$C = \mathcal{C}(x), S = \mathcal{S}(x), R = \mathcal{R}(x).$$

As a result, we may pose numerous sizing-related problems as semidefinite programs. Here are some examples:

• We can minimize the total area occupied by the wires under the restriction that the speed (i.e., the time constant) of the circuit should be above a certain bound. The ensuing semidefinite program is

$$\begin{array}{ll} \text{minimize} & \sum_i x_i \\ \text{s.t.} & \\ & \begin{array}{rcl} \mathcal{S}(x) - \kappa\mathcal{C}(x) & \succeq & 0, \\ x & \geq & 0 \end{array} \end{array}$$

($\kappa > 0$ is fixed).

• In the case of a simple circuit (which in fact is a common case in chip design), we can minimize the heat dissipated during a transition, under the restriction that the speed is above a certain bound. The ensuing semidefinite program is

$$\begin{array}{ll} \text{minimize} & e^T[\mathcal{C}(x)]e \\ \text{s.t.} & \\ & \begin{array}{rcl} \mathcal{S}(x) - \kappa\mathcal{C}(x) & \succeq & 0, \\ x & \geq & 0. \end{array} \end{array}$$

• We can add to the above programs upper and lower bounds on the areas of segments, as well as other linear constraints on the design variables, etc.

4.8 Applications VI: Structural design

Structural design is an engineering area dealing with mechanical constructions like trusses and plates. We already know what a truss is—a construction made up of thin elastic bars linked to each other. A plate is a construction made up of a material occupying a given domain, the mechanical properties of the material varying continuously from point to point. In engineering, design of plates is called shape design; in what follows we call these objects of our interest shapes instead of plates.

A typical structural design problem is, "Given the type of material to be used, a resource (an upper bound on the amount of material to be used) and a set of loading scenarios—external loads operating on the construction—find an optimal truss or shape, one able to withstand best of all the loads in question." It turns out that numerous problems of this type can be cast as semidefinite programs, which offers a natural way to model and process them analytically and numerically. The purpose of this section is to develop a unified semidefinite-programming-based approach to these structural design problems.

4.8.1 Building a model

The mechanical constructions we are considering (the so-called constructions with linear elasticity) can be described as follows.

I. A construction C can be characterized by

I.1. A linear space $\mathbf{V} = \mathbf{R}^m$ of virtual displacements of C.

I.2. A positive semidefinite quadratic form

$$E_C(v) = \frac{1}{2} v^T A_C v$$

on the space of displacements. The value of this form at a displacement v is the potential energy stored by the construction as a result of the displacement. The (positive semidefinite symmetric) $m \times m$ matrix A_C of this form is called the stiffness matrix of C.

Example. A truss fits I.1–I.2; see sections 1.3.5 and 3.4.3.

I.3. A closed convex subset $\mathcal{V} \subset \mathbf{R}^m$ of kinematically admissible displacements.

Example, continued. In our previous discussions of trusses, there was no specific set of kinematically admissible displacements—we assumed that in principle every virtual displacement $v \in \mathbf{R}^m$ may become an actual displacement, provided that an external load is chosen accordingly. However, sometimes the tentative displacements of the nodes are restricted by external obstacles, like the one in Fig. 4.13.

II. An external load applied to the construction C can be represented by a vector $f \in \mathbf{R}^m$. The static equilibrium of C loaded by f is given by the following variational principle.

A construction C is able to carry an external load f if and only if the quadratic form

$$E_C^f(v) = \frac{1}{2} v^T A_C v - f^T v \tag{4.8.89}$$

of displacements v attains its minimum on the set $\mathcal{V}$ of kinematically admissible displacements, and a displacement yielding (static) equilibrium is a minimizer of $E_C^f(\cdot)$ on $\mathcal{V}$.

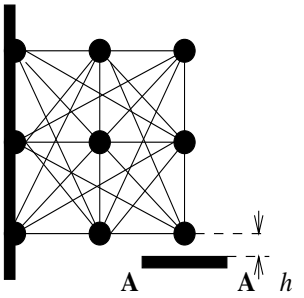


Figure 4.13. *An obstacle. What we see is a nine-node planar ground structure with 33 tentative bars and a rigid obstacle AA. This obstacle does not allow the southeastern node to move down more than by h and thus induces a linear inequality constraint on the vector of virtual displacements of the nodes.*

The minus minimum value of E_C^f on $\mathcal{V}$ is called the compliance of the construction C with respect to the load f:

$$\mathrm{Compl}_f(C) = \sup_{v\in\mathcal{V}} \left[f^T v - \frac{1}{2} v^T A_C v \right].$$

Example, continued.[38] We saw in section 3.4.3 that the variational principle does work for a truss. At that moment, however, we dealt with the particular obstacle-free case: $\mathcal{V} = \mathbf{V}$. What happens when there are obstacles? Assume that the obstacles are absolutely rigid and frictionless. When in the course of truss deformation a moving node meets an obstacle, a contact force occurs and comes into play—it becomes a part of the external load. As a result, the equilibrium displacement is given by the equations

$$Av = f + \sum_l f_l, \qquad (*)$$

where Av, as we remember, is the vector of reaction forces caused by the deformation of the truss and f_l's represent the contact forces coming from obstacles. Nature is free to choose these forces, with only the restriction that a contact force should be normal to the boundary of the corresponding obstacle (there is no friction!) and should point toward the truss. With these remarks in mind, one can easily recognize in (*) the usual KKT conditions for constrained minimum of E_C^f, the constraints being given by the obstacles. Thus, an equilibrium displacement is a KKT point of the problem of minimizing E_C^f over $\mathcal{V}$. Since the problem is convex, its KKT points are the same as the minimizers of E_C^f over $\mathcal{V}$.

The part of the story we have told so far relates to a particular system and does not address the question of what is affected by the design: May we play with the space of virtual

[38] A reader not acquainted with the KKT optimality conditions may skip this paragraph.

displacements? or with $\mathcal{V}$? or with the stiffness matrix? We shall focus on the case when the only element affected by the design is the stiffness matrix. Specifically, we assume the following.

III. The stiffness matrix A_C depends on mechanical characteristics $t_1, \ldots, t_n$ of elements $E_1, \ldots, E_n$ making up the construction, and these characteristics t_i are positive semidefinite symmetric $d \times d$ matrices, with d given by the type of the construction. Specifically,

$$A_C = \sum_{i=1}^{n} \sum_{s=1}^{S} b_{is} t_i b_{is}^T, \tag{4.8.90}$$

where b_{is} are given $m \times d$ matrices.

At first glance, III looks very strange: where are the positive semidefinite matrices t_i coming from? Well, this is what mechanics says on both trusses and shapes.

Indeed, we know from section 1.3.5 that the stiffness matrix of a truss is

$$A(t) = \sum_{i=1}^{n} t_i b_i b_i^T,$$

where $t_i \geq 0$ are bar volumes and b_i are certain vectors given by the geometry of the nodal set and where nonnegative reals t_i may be viewed as positive semidefinite 1×1 matrices.

Now, what about shapes? To see that III holds in this case as well, it requires an additional excursion to mechanics. (This can be omitted by noninterested readers.)

As mentioned, a shape is made up of material occupying a given 2D or 3D domain Ω, the mechanical properties of the material varying from point to point. Such a construction is infinite dimensional: its virtual displacements are vector fields on Ω and, taken together, form certain linear space $\mathbf{V}$ of vector fields on Ω. ($\mathbf{V}$ should not necessarily comprise all vector fields; e.g., some parts of the boundary of Ω may be fixed, so that the displacement fields must vanish at these parts of the boundary.)

The elasticity properties of the material at a point $P \in \Omega$ are represented by the rigidity tensor $E(P)$, which, mathematically, is a symmetric positive semidefinite $d \times d$ matrix, where $d = 3$ for planar and $d = 6$ for spatial shapes. Mechanics says that the density, at a point P, of the potential energy stored by a shape as a reaction on a displacement $v(\cdot)$, is

$$\frac{1}{2} s_P^T[v] E(P) s_P[v], \tag{4.8.91}$$

so that the total potential energy stored in a deformated shape is

$$\frac{1}{2} \int_\Omega s_P^T[v] E(P) s_P[v] dP.$$

Here for a 2D shape

$$s_P[v] = \begin{pmatrix} \frac{\partial v_x(P)}{\partial x} \\ \frac{\partial v_y(P)}{\partial y} \\ \frac{1}{\sqrt{2}} \left[\frac{\partial v_x(P)}{\partial y} + \frac{\partial v_y(P)}{\partial x} \right] \end{pmatrix},$$

where v_x and v_y are the x- and the y-components of the 2D vector field $v(\cdot)$. Note that $s_P[v]$ can be obtained as follows. We first build the Jacobian of the vector field v at P, the matrix

$$J(P) = \begin{pmatrix} \frac{\partial v_x(P)}{\partial x} & \frac{\partial v_x(P)}{\partial y} \\ \frac{\partial v_y(P)}{\partial x} & \frac{\partial v_y(P)}{\partial y} \end{pmatrix},$$

and then symmeterize the matrix—build the symmetric 2×2 matrix

$$J^s(P) = \frac{1}{2}[J(P) + J^T(P)].$$

$s_P[v]$ is nothing but the vector of the coordinates of $J^s(P) \in \mathbf{S}^2$ in the natural orthonormal basis $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 0 & 2^{-1/2} \\ 2^{-1/2} & 0 \end{smallmatrix}\right)$ of the 3D space $\mathbf{S}^2$.

For a 3D shape, $s_P[v]$ is given by a completely similar construction. We build the 3×3 Jacobian of the 3D vector fields $v(\cdot)$ at the point P, symmeterize it, and then pass from this 3×3 symmetric matrix to the vector of coordinates of this matrix in the natural basis of the six-dimensional space $\mathbf{S}^3$. We skip the corresponding explicit formulas.

An external load acting on a shape can be represented by a linear form $f[v]$ on the space of displacements. This form measures the work carried out by the load at a displacement. In typical cases, this functional looks like $\int_{\partial\Omega} f^T(P)v(P)dS(P)$, $f(P)$ being the field of external forces acting at the boundary. Mechanics says that the equilibrium displacement field in the loaded shape minimizes the energy functional

$$\frac{1}{2}\int_\Omega s_P^T[v]E(P)s_P[v]dP - f[v]$$

over the set of kinematically admissible vector fields $v(\cdot)$. The minus minimum value of this functional is called the compliance of the shape with respect to the load in question.

As we see, the true model of a shape is infinite dimensional. To get a computationally tractable model, a finite element approximation is used, namely,

1. The domain Ω is partitioned into finitely many nonoverlapping cells $\Omega_1, \ldots, \Omega_n$, and the properties of the material are assumed to be constant within the cells:

$$E(P) = E_i \text{ for } P \in \Omega_i.$$

2. The infinite dimensional space $\mathbf{V}$ of vector fields on Ω is approximated by its finite dimensional subspace $\mathbf{V}^m$ spanned by m basic continuously differentiable displacement fields $w_1(\cdot), \ldots, w_m(\cdot)$. With this approximation, the set of kinematically admissible displacement fields shrinks to a set in $\mathbf{V}^m$.

With this approximation, the finite element model of a shape becomes as follows:

• A virtual displacement v becomes a vector from $\mathbf{R}^m$ (the actual displacement field corresponding to a vector $v = (v_1, \ldots, v_m)^T$ is, of course, $\sum_{i=1}^m v_i w_i(\cdot)$).

• The potential energy stored by the shape, the displacement being v, is

$$\frac{1}{2}\sum_{i=1}^n \int_{\Omega_i} v^T[s(P)E_i s^T(P)]vdP, \qquad s^T(P) = [s_P[w_1]; s_P[w_2]; \ldots; s_P[w_m]] \in \mathbf{M}^{d,m}.$$

• The linear functional $f[\cdot]$ representing a load becomes a usual linear form $f^T v$ on $\mathbf{R}^m$ (so that we can treat the vector f of the coefficients of this form as the load itself).

• The equilibrium displacement of the shape under a load f is the minimizer of the quadratic form

$$\frac{1}{2}v^T\left[\sum_{i=1}^{n}\int_{\Omega_i} s(P)E_i s^T(P)dP\right]v - f^T v$$

on a given set $\mathcal{V} \subset \mathbf{R}^m$ of kinematically admissible displacements, and the compliance is minus the minimum value of this form on $\mathcal{V}$.

It remains to note that, as stated in Exercise 1.17, there exist a positive integer S and cubature formulas such that

$$\int_{\Omega_i} s_P^T[w_p]Es_P[w_q]dP = \sum_{s=1}^{S}\alpha_{is}s_{P_{is}}^T[w_p]Es_{P_{is}}[w_q] \quad \forall E \in \mathbf{S}^d \quad \forall p,q = 1,\dots,m$$

with nonnegative weights α_{is}. Denoting by ω_i the measures of the cells Ω_i and setting

$$t_i = \omega_i E_i, \quad b_{is} = \alpha_{is}^{1/2}\omega_i^{-1/2}s(P_{is}),$$

we get

$$\int_{\Omega_i} s(P)E_i s^T(P)dP = \sum_{s=1}^{S} b_{is}t_i b_{is}.$$

Thus, we have represented a shape by a collection $t_1,\dots,t_n$ of positive semidefinite $d \times d$ matrices, and the potential energy of a deformated shape now becomes

$$\frac{1}{2}v^T\left[\sum_{i=1}^{n}\int_{\Omega_i} s(P)E_i s^T(P)\right]v = \frac{1}{2}v^T\left[\sum_{i=1}^{n}\sum_{s=1}^{S} b_{is}t_i b_{is}^T\right]v,$$

where v is the displacement. We see that a shape, after finite element discretization, fits the requirement III.

The concluding chapter of our story outlines the way we measure the amount of material used to build the construction.

IV. The amount of material consumed by a construction C is completely characterized by the vector $(\mathrm{Tr}(t_1),\dots,\mathrm{Tr}(t_n))$ of the traces of the matrices $t_1,\dots,t_n$ mentioned in III.

For a truss, the indicated traces are exactly the same as t_i's themselves and are the volumes of the bars constituting the truss, so that IV is quite reasonable. For a shape, $\mathrm{Tr}(E(P))$ is a natural measure of the material density of the shape at a point $P \in \Omega$, so that IV again is quite reasonable.

Now we can formulate the general structural design problem we are interested in.

Problem 4.8.1. Static structural design. *Given*

1. *a ground structure, i.e.,*

   • *the space* $\mathbf{R}^m$ *of virtual displacements along with its closed convex subset* $\mathcal{V}$ *of kinematically admissible displacements,*

   • *a collection* $\{b_{is}\}_{\substack{i=1,\dots,n\\ s=1,\dots,S}}$ *of* $m \times d$ *matrices;*

2. *a set* $T = \{(t_1, \ldots, t_n) \mid t_i \in \mathbf{S}^d_+ \quad \forall i\}$ *of admissible designs; and*

3. *a set* $\mathcal{F} \subset \mathbf{R}^m$ *of loading scenarios,*

find an admissible construction that is stiffest, with respect to $\mathcal{F}$*; i.e., find a collection* $t \in T$ *that minimizes the worst, over* $f \in \mathcal{F}$*, compliance of the construction with respect to* f*:*

$$\min_{t\in T}\left\{\mathrm{Compl}_{\mathcal{F}}(t) \equiv \sup_{f\in\mathcal{F}}\sup_{v\in\mathcal{V}}\left[f^T v - \frac{1}{2}v^T\left[\sum_{i=1}^n\sum_{s=1}^S b_{is}t_i b_{is}^T\right]v\right] : t \in T\right\}.$$

4.8.2 Standard case

The static structural design Problem 4.8.1 in its general form is a little bit diffuse—we did not name the geometries of the set $\mathcal{V}$ of kinematically admissible displacements, or the set T of admissible designs, or the set $\mathcal{F}$ of loading scenarios. For all applications known to us these geometries can be specialized as follows.

S.1. The set $\mathcal{V}$ of kinematically admissible displacements is a polyhedral set,

$$\mathcal{V} = \{v \in \mathbf{R}^m \mid Rv \le r\} \qquad [R \in \mathbf{M}^{q,m}], \tag{4.8.92}$$

and the system of linear inequalities $Rv \le r$ satisfies the Slater condition: there exists $\bar{v}$ such that $R\bar{v} < r$.

S.2. The set T of admissible designs is given by simple linear constraints on the traces of the positive semidefinite rigidity matrices t_i, namely, upper and lower bounds on $\mathrm{Tr}(t_i)$ and an upper bound on the total material resource $\sum_{i=1}^n \mathrm{Tr}(t_i)$:

$$T = \{t = (t_1, \ldots, t_n) \mid t_i \in \mathbf{S}^d_+, \underline{\rho}_i \le \mathrm{Tr}(t_i) \le \overline{\rho}_i, \sum_{i=1}^n \mathrm{Tr}(t_i) \le w\} \tag{4.8.93}$$

with given parameters

$$0 \le \underline{\rho}_i < \overline{\rho}_i < \infty, \sum_{i=1}^n \underline{\rho}_i < w.$$

S.3. The set $\mathcal{F}$ of loading scenarios is either a finite set,

$$\mathcal{F} = \{f_1, \ldots, f_k\} \tag{4.8.94}$$

(multiload structural design), or an ellipsoid,

$$\mathcal{F} = \{f = Qu \mid u^T u \le 1\} \quad [Q \in \mathbf{M}^{m,k}] \tag{4.8.95}$$

(robust structural design).

The interpretation of the multiload setting is quite clear: the construction is supposed to work under different nonsimultaneous loading scenarios, and we intend to control its stiffness with regard to these scenarios. To motivate the robust setting, consider the following example.

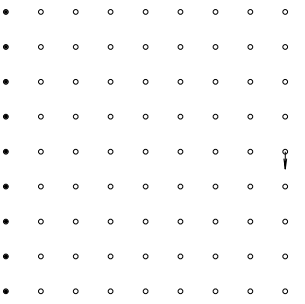


Figure 4.14. 9×9 *ground structure and the load of interest.*

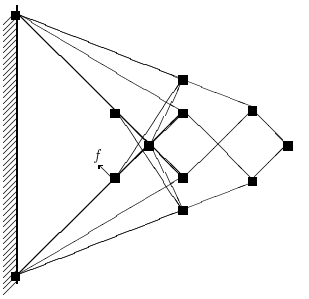


Figure 4.15. *Optimal cantilever (single-load design); the compliance is 1.000.*

Example. Assume we are designing a planar truss—a cantilever. The 9×9 nodal structure and the only load of interest f^* are as shown in Fig. 4.14. The optimal single-load design yields a nice truss shown in Fig. 4.15. The compliance of the optimal truss with respect to the load of interest is 1.000.

Now, what happens if, instead of the load f^*, the truss is affected by a small occasional load f shown in Fig. 4.15, the magnitude ($\equiv$ the Euclidean length) of f being just 0.5% of the magnitude of f^*? The results are disastrous: the compliance is increased by factor 8.4 (!) In fact, our optimal cantilever is highly unstable: it may collapse when a bird tries to build a nest in a badly placed node of the construction.

To ensure stability of our design, we should control the compliance not only with respect to a restricted set of loads of interest, but also with respect to all relatively small occasional loads somehow distributed along the nodes. The simplest way to do it is to add to the original finite set of loads of interest the ball comprising all occasional loads of magnitude not exceeding some level. There are, however, two difficulties in this approach:

• From the viewpoint of mathematics, it is not that easy to deal with the set of loading scenarios of the form "the union of a ball and a finite set." It would be easier to handle either a finite set or an ellipsoid.

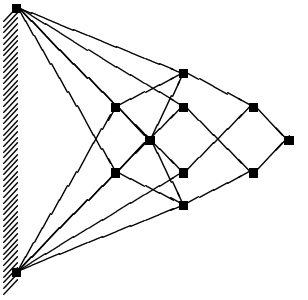


Figure 4.16. *"Robust" cantilever.*

• From the engineering viewpoint, the difficulty is to decide where the occasional loads should be applied. If we allow them to be distributed along all 9×9 nodes of the original ground structure, the resulting design will incorporate all these nodes (otherwise its compliance with respect to some occasional loads will be infinite), which makes no sense. What we actually are concerned with are occasional loads distributed along the nodes that will be used by the resulting construction, but how could we know these nodes in advance?

Regarding the first difficulty, a natural way to overcome it is to take, as $\mathcal{F}$, the ellipsoidal envelope of the original—finite—set of loads and a small ball, i.e., to choose as $\mathcal{F}$ the ellipsoid of the smallest volume centered at the origin which contains the original loading scenarios and the ball.

To overcome, to some extent, the second difficulty, we could use a two-stage scheme. At the first stage, we take into consideration the loads of interest only and solve the corresponding single- or multiload problem, thus getting a certain preliminary truss. At the second stage, we treat the set of nodes actually used by the preliminary truss as our new nodal set and take as $\mathcal{F}$ the ellipsoidal envelope of the loads of interest and the ball comprising all small occasional loads distributed along this reduced nodal set.

Let us look at what this approach yields in our cantilever example. The cantilever depicted in Fig. 4.15 uses 12 nodes from the original 81-node grid; two of these 12 nodes are fixed. Taking the resulting 12 nodes as our new nodal set, and allowing all pair-connections of these nodes, we get a new—reduced—ground structure with 20 degrees of freedom. Now let us define $\mathcal{F}$ as the ellipsoidal envelope of f^* and the ball comprising all loads with the Euclidean norm not exceeding 10% of the norm of f^*. This 20-dimensional ellipsoid $\mathcal{F}$ is very simple: one of its principal half-axes is f^*, and the remaining 19 half-axes are of the length $0.1\|f^*\|_2$ each, the directions of these half-axes forming a basis in the orthogonal complement to f^* in the 20-dimensional space of virtual displacements of our 12 nodes. Minimizing the worst, with respect to the ellipsoid of loads $\mathcal{F}$, compliance under the original design constraints, we come to a new cantilever depicted in Fig. 4.16. The maximum, over the ellipsoid of loads $\mathcal{F}$, compliance of the robust cantilever is 1.03, and its compliance with

respect to the load of interest f^* is 1.0024—only 0.24% larger than the optimal compliance given by the single-load design! We see that when passing from the nominal single-load design to the robust one we lose basically nothing in optimality and at the same time get dramatic improvement in the stability. (For the nominal design, the compliance with respect to a badly chosen occasional load of magnitude $0.1\|f^*\|_2$ may be as large as 32,000!)

The above example is a good argument for considering ellipsoidal sets of loads.

We shall refer to the static structural design problem with the data satisfying **S.1**—**S.3** as to the standard SSD problem. We always assume that the data of the standard SSD problem satisfies the following assumption:

S.4. $\sum_{i=1}^n \sum_{s=1}^S b_{is} t_i b_{is}^T \succ 0$ whenever $t_i \succ 0$, $i = 1, \ldots, n$.

This is a physically meaningful assumption that excludes rigid body motions of the ground structure: if all rigidities are positive definite, then the potential energy stored by the construction under any nontrivial displacement is strictly positive.

4.8.3 Semidefinite reformulation of the standard SSD problem

To get a semidefinite reformulation of the standard SSD problem, we start with building semidefinite representation of the compliance. Thus, our goal is to get an SDR for the set

$$\mathcal{C} = \left\{(t, f, \tau) \in \left(\mathbf{S}_+^d\right)^n \times \mathbf{R}^m \times \mathbf{R} \mid \mathrm{Compl}_f(t) \equiv \sup_{v \in \mathbf{R}^m : Rv \le r} \left[f^T v - \frac{1}{2} v^T \left[\sum_{i=1}^n \sum_{s=1}^S b_{is} t_i b_{is}^T \right] v \right] \le \tau \right\}.$$

The required SDR is given by the following proposition.

Proposition 4.8.1. *Let $t = (t_1, \ldots, t_n) \in \left(\mathbf{S}_+^d\right)^n$ and $f \in \mathbf{R}^m$. Then the inequality*

$$\mathrm{Compl}_f(t) \le \tau$$

is satisfied if and only if there exists a nonnegative vector μ of the dimension q equal to the number of linear inequalities defining the set of virtual displacements (see (4.8.92)*) such that the matrix*

$$\mathcal{A}(t, f, \tau, \mu) = \begin{pmatrix} 2\tau - 2r^T\mu & -f^T + \mu^T R \\ -f + R^T\mu & \sum_{i=1}^n \sum_{s=1}^S b_{is} t_i b_{is}^T \end{pmatrix}$$

is positive semidefinite. Thus, the epigraph of $\mathrm{Compl}_f(t)$ *(regarded as a function of $t \in \left(\mathbf{S}_+^d\right)^n$ and $f \in \mathbf{R}^m$) admits the SDR*

$$\begin{array}{rcl} \begin{pmatrix} 2\tau - 2r^T\mu & -f^T + \mu^T R \\ -f + R^T\mu & \sum_{i=1}^n \sum_{s=1}^S b_{is} t_i b_{is}^T \end{pmatrix} & \succeq & 0, \\ t_i & \succeq & 0, \ i = 1, \ldots, n, \\ \mu & \succeq & 0. \end{array} \tag{4.8.96}$$

Proof. First let us explain where the result comes from. By definition, $\mathrm{Compl}_f(t) \le \tau$ if

$$\sup_{v:Rv\le r}\left[f^Tv-\frac{1}{2}v^TA(t)v\right]\le\tau \qquad \left[A(t)=\sum_{i=1}^n\sum_{s=1}^S b_{is}t_ib_{is}^T\right].$$

The supremum in the left-hand side is taken over v varying in a set given by linear constraints $Rv \le r$. If we add penalized constraints $\mu^T(r-Rv)$ to the objective, μ being a nonnegative weight vector, and then remove the constraints, passing to the supremum of the penalized objective over the entire space, i.e., to the quantity

$$\phi_f(t,\mu)\equiv\sup_{v\in\mathbf{R}^m}\left[f^Tv-\frac{1}{2}v^TA(t)v+\mu^T(r-Rv)\right],$$

then we end up with something that is $\ge \mathrm{Compl}_f(t)$. Consequently, if there exists $\mu \ge 0$ such that

$$f^Tv-\frac{1}{2}v^TA(t)v+\mu^T(r-Rv)\le\tau \quad \forall v\in\mathbf{R}^m,$$

then we have $\tau \ge \mathrm{Compl}_f(t)$. On the other hand, the Lagrange duality says that under the Slater condition (see assumption **S.1**) the quantity $\phi_f(t,\mu)$ for properly chosen $\mu \ge 0$ is exactly the supremum of $f^Tv-\frac{1}{2}v^TA(t)v$ over v satisfying $Rv \le r$: if $\tau \ge \mathrm{Compl}_f(t)$, then $\tau \ge \phi_f(t,\mu)$ for some $\mu \ge 0$. Thus, believing in the Lagrange duality, we come to the following observation:

> (!) *The inequality* $\mathrm{Compl}_f(t) \le \tau$ *is equivalent to the existence of a* $\mu \ge 0$ *such that* $\phi_f(t,\mu) \le \tau$.

It remains to note that the inequality $\phi_f(t,\mu) \le \tau$ says that the unconstrained minimum of the quadratic form

$$Q(v)=[\tau-r^T\mu]+\frac{1}{2}v^TA(t)v+(-f+R^T\mu)^Tv$$

is nonnegative. By the simple lemma (see section 4.3.1), the latter fact is equivalent to the positive semidefiniteness of the matrix $\mathcal{A}(t,f,\tau,\mu)$.

To be self-sufficient, let us derive (!) from conic duality.

Our first observation is as follows:

> (∗) *Let* $t_1,\dots,t_n \in \mathbf{S}^d_+$, $f \in \mathbf{R}^m$, *and* $\tau \in \mathbf{R}$ *be fixed. Consider the following system of inequalities with variables* $v \in \mathbf{R}^m$, $\sigma \in \mathbf{R}$*:*
>
> $$\begin{array}{rcl}\frac{1}{2}v^TA(t)v+\sigma-f^Tv &\le& 0\\ R^Tv &\le& r.\end{array}\qquad\left[A(t)=\sum_{i=1}^n\sum_{s=1}^S b_{is}t_ib_{is}^T\right], \qquad (S(t,f))$$
>
> *Then* $(t,f,\tau)\in\mathcal{C}$ *if and only if* $\tau \ge \sigma$ *for all solutions to* $(S(t,f))$*, i.e., if and only if the linear inequality* $\tau \ge \sigma$ *is a consequence of the system* $(S(t,f))$.

Indeed, the σ-components of the solutions to $(S(t, f))$ are exactly those σ's that do not exceed the value of the quadratic form

$$E_t^f(v) = f^T v - \frac{1}{2} v^T A(t) v$$

at a certain point of the set $\mathcal{V} = \{v \mid Rv \leq r\}$. Consequently, to say that a given τ is $\geq$ all such σ's is exactly the same as to say that τ is $\geq$ the supremum of the form $E_t^f(v)$ over $v \in \mathcal{V}$, or equivalently that $\tau \geq \mathrm{Compl}_f(t)$.

Now, $(S(t, f))$ is nothing but a linear vector inequality. Indeed, the quadratic inequality in $(S(t, f))$ is a conic quadratic inequality:

$$\begin{array}{c} \frac{1}{2} v^T A(t) v + \sigma - f^T v \leq 0, \\ \Updownarrow \\ v^T A(t) v + 2\sigma - 2 f^T v \leq 0, \\ \Updownarrow \\ \|B^T(t) v\|_2^2 - 2 f^T v + 2\sigma \leq 0, \end{array}$$

where

$$B(t) = [b_{11} t_1^{1/2}, b_{12} t_1^{1/2}, \ldots, b_{1S} t_1^{1/2}; \ldots; b_{n1} t_n^{1/2}, \ldots, b_{nS} t_n^{1/2}] \quad [A(t) = B(t) B^T(t)]$$

$$\Updownarrow$$

$$\begin{pmatrix} B^T(t) v \\ \frac{1}{2} + \sigma - f^T v \\ \frac{1}{2} - \sigma + f^T v \end{pmatrix} \geq_{\mathbf{L}} 0$$

so that $(S(t, f))$ is the linear vector inequality in variables v, σ, τ,

$$Q \begin{pmatrix} v \\ \sigma \end{pmatrix} - q \equiv \begin{pmatrix} B^T(t) v \\ \frac{1}{2} + \sigma - f^T v \\ \frac{1}{2} - \sigma + f^T v \\ r - Rv \end{pmatrix} \geq_{\mathbf{K}} 0,$$

where $\mathbf{K}$ is the direct product of the ice cream cone and the nonnegative orthant of appropriate dimensions.

Note that the resulting linear vector inequality is strictly feasible. Indeed, due to the Slater assumption in **S.1**, we may choose v such that $r - Rv > 0$. After v is chosen, we may choose σ to be negative enough to make strict also the conic quadratic part

$$\begin{pmatrix} B^T(t) v \\ \frac{1}{2} + \sigma - f^T v \\ \frac{1}{2} - \sigma + f^T v \end{pmatrix} \geq_{\mathbf{L}} 0$$

of our vector inequality.

Now let us use the necessary and sufficient conditions for a linear inequality to be a consequence of a strictly feasible linear vector inequality (see Proposition 2.4.3). Here these

conditions are equivalent to the existence of a nonnegative vector μ, of the same dimension q as the vector r, and a vector $\zeta \in \mathbf{L}$ such that

$$[v^T;\sigma]Q^T\begin{pmatrix}\zeta\\ \mu\end{pmatrix} = -\sigma \quad \forall (v,\sigma)\in \mathbf{R}^m\times\mathbf{R},$$
$$[\zeta^T;\mu^T]q \geq -\tau.$$

Recalling the origin of Q and q, we come to the following conclusion:

(∗∗) *Let $t \in \left(\mathbf{S}^d_+\right)^n$, $f \in \mathbf{R}^m$, and $\tau \in \mathbf{R}$. Then* $\mathrm{Compl}_f(t) \leq \tau$ *if and only if there exist d-dimensional vectors ξ_{is}, reals α, β and a vector μ such that*

$$\begin{array}{lrcl}
\text{(a)} & (\beta-\alpha)f - R^T\mu + \sum_{i=1}^n\sum_{s=1}^S b_{is}t_i^{1/2}\xi_{is} & = & 0,\\
\text{(b)} & \alpha-\beta & = & -1,\\
\text{(c)} & \frac{1}{2}(\alpha+\beta) + r^T\mu & \leq & \tau,\\
\text{(d)} & \mu & \geq & 0,\\
\text{(e)} & \beta & \geq & \sqrt{\alpha^2 + \sum_{i=1}^n\sum_{s=1}^S \xi_{is}^T\xi_{is}}.
\end{array} \tag{4.8.97}$$

Now consider the CQI

$$\begin{pmatrix} B^T(t)v\\ \frac{1}{2} - r^T\mu + \sigma + [-f+R^T\mu]^Tv\\ \frac{1}{2} + r^T\mu - \sigma - [-f+R^T\mu]^Tv\end{pmatrix} \geq_{\mathbf{L}} 0, \tag{4.8.98}$$

where v,σ are the variables and $\mu \geq 0$ is a vector of parameters, and let us ask ourselves when the inequality $\sigma \leq \tau$ is a consequence of this (clearly strictly feasible) vector inequality. According to Proposition 2.4.3 this is the case if and only if there exist vectors $\xi_{is} \in \mathbf{R}^d$ and reals α, β such that

$$\begin{array}{lrcl}
\text{(a)} & (\alpha-\beta)[-f+R^T\mu] + \sum_{i=1}^n\sum_{s=1}^S b_{is}t_i^{1/2}\xi_{is} & = & 0,\\
\text{(b)} & \alpha-\beta & = & -1,\\
\text{(c)} & \frac{1}{2}(\alpha+\beta) + (\beta-\alpha)r^T\mu & \leq & \tau,\\
\text{(d)} & \beta & \geq & \sqrt{\alpha^2 + \sum_{i=1}^n\sum_{s=1}^S \xi_{is}^T\xi_{is}}.
\end{array} \tag{4.8.99}$$

Comparing (4.8.97) and (4.8.99), we come to the following conclusion:

(∗∗∗) *Let $t \in \left(\mathbf{S}^d_+\right)^n$, $f \in \mathbf{R}^m$, and $\tau \in \mathbf{R}$. Then* $\mathrm{Compl}_f(t) \leq \tau$ *if and only if there exists $\mu \geq 0$ such that the inequality $\sigma \leq \tau$ is a consequence of the CQI* (4.8.98).

It remains to note that the CQI (4.8.98) is equivalent to the scalar quadratic inequality

$$v^T A(t)v - 2r^T\mu + 2\sigma + 2[-f + R^T\mu]v \le 0.$$

Consequently, (***) says that $\mathrm{Compl}_f(t) \le \tau$ if and only if there exists $\mu \ge 0$ such that the following implication holds true:

$$\forall(v,\sigma): \quad \sigma \le [f - R^T\mu]^T v - \frac{1}{2}v^T A(t)v + r^T\mu \Rightarrow \sigma \le \tau.$$

But the latter implication clearly holds true if and only if

$$\tau \ge \max_{v\in\mathbf{R}^m}\left[[f - R^T\mu]^T v - \frac{1}{2}v^T Av + r^T\mu\right]. \tag{4.8.100}$$

Thus, $\tau \ge \mathrm{Compl}_f(t)$ if and only if there exists $\mu \ge 0$ such that (4.8.100) holds, which is exactly the statement (!) we need. □

An SDR of the epigraph of the compliance immediately implies a semidefinite reformulation of the multiload standard SSD problem with k loading scenarios $f_1, \ldots, f_k$:

$$\begin{array}{lll}
\text{minimize} & \tau & \\
\text{s.t.} & & \\
\text{(a)} & \begin{pmatrix} 2\tau - 2r^T\mu_l & -f_l^T + \mu_l^T R \\ -f_l + R^T\mu_l & \sum_{i=1}^n\sum_{s=1}^S b_{is}t_i b_{is}^T \end{pmatrix} \succeq 0, & l = 1,\ldots,k, \\
\text{(b)} & t_i \succeq 0, & i = 1,\ldots,n, \\
\text{(c)} & \sum_{i=1}^n \mathrm{Tr}(t_i) \le w, & \\
\text{(d)} & \underline{\rho}_i \le \mathrm{Tr}(t_i) \le \overline{\rho}_i, & i = 1,\ldots,n, \\
\text{(e)} & \mu_l \ge 0, & l = 1,\ldots,k,
\end{array} \tag{4.8.101}$$

where the design variables are $t_i \in \mathbf{S}^d$, $i = 1,\ldots,n$, μ_l (vectors of the dimension q equal to the number of linear inequalities in (4.8.92)), $l = 1,\ldots,k$, and $\tau \in \mathbf{R}$. Indeed, the LMIs (a) along with nonnegativity constraints (e) express the fact that the worst, over the loads $f_1,\ldots,f_k$, compliance of the construction yielded by the rigidities $t_1,\ldots,t_n$ does not exceed τ (see Proposition 4.8.1), while the remaining constraints (b), (c), (d) express the fact that $t = (t_1,\ldots,t_n)$ is an admissible design.

Consider next the robust standard SSD problem, where $\mathcal{F}$ is an ellipsoid:

$$\mathcal{F} = \{f = Qu \mid u^T u \le 1\}.$$

Here we meet with a difficulty not present in the case of finite $\mathcal{F}$: our objective now is

$$\mathrm{Compl}_{\mathcal{F}}(t) = \sup_{f\in\mathcal{F}} \mathrm{Compl}_f(t),$$

i.e., it is the supremum of infinitely many SDr functions. Our calculus does not offer tools to build an SDR for such an aggregate. This difficulty reflects the essence of the matter:

for an SSD problem with obstacles, the robust version of the SSD problem is extremely difficult (at least as difficult as an NP-complete combinatorial problem). Fortunately, in the obstacle-free case it is easy to get an SDR for $\mathrm{Compl}_{\mathcal{F}}(t)$, provided that $\mathcal{F}$ is an ellipsoid.

Proposition 4.8.2. *Let the set of kinematically admissible displacements coincide with the space $\mathbf{R}^m$ of all virtual displacements: $\mathcal{V} = \mathbf{R}^m$, and let $\mathcal{F}$ be an ellipsoid:*

$$\mathcal{F} = \{f = Qu \mid u^T u \le 1\} \quad [Q \in \mathbf{M}^{m,k}].$$

Then the function $\mathrm{Compl}_{\mathcal{F}}(t)$, regarded as a function of $t = (t_1, \ldots, t_n) \in \left(\mathbf{S}^d_+\right)^n$, is SDr: for $t \in \left(\mathbf{S}^d_+\right)^n$,

$$\mathrm{Compl}_{\mathcal{F}}(t) \le \tau \Leftrightarrow \begin{pmatrix} 2\tau I_k & Q^T \\ Q & \sum_{i=1}^n \sum_{s=1}^S b_{is} t_i b_{is}^T \end{pmatrix} \succeq 0. \tag{4.8.102}$$

Consequently, the robust obstacle-free standard SSD problem can be posed as the following semidefinite program:

$$\begin{array}{ll} \textit{minimize} & \tau \\ \textit{s.t.} & \\ & \begin{pmatrix} 2\tau I_k & Q^T \\ Q & \sum_{i=1}^n \sum_{s=1}^S b_{is} t_i b_{is}^T \end{pmatrix} \succeq 0; \\ & t_i \succeq 0, \ i = 1, \ldots, n, \\ & \sum_{i=1}^n \mathrm{Tr}(t_i) \le w, \\ & \underline{\rho}_i \le \mathrm{Tr}(t_i) \le \overline{\rho}_i, \ i = 1, \ldots, n. \end{array} \tag{4.8.103}$$

Proof. Let, as always, $A(t) = \sum_{i=1}^n \sum_{s=1}^S b_{is} t_i b_{is}^T$. We have

$$\begin{array}{rcll} \mathrm{Compl}_{\mathcal{F}}(t) & \le & \tau & \Leftrightarrow \\ (Qu)^T v - \frac{1}{2} v^T A(t) v & \le & \tau \quad \forall v \quad \forall (u : u^T u \le 1) & \Leftrightarrow \\ (Qu)^T v - \frac{1}{2} v^T A(t) v & \le & \tau \quad \forall v \quad \forall (u : u^T u = 1) & \Leftrightarrow \\ (Q\|w\|_2^{-1} w)^T v - \frac{1}{2} v^T A(t) v & \le & \tau \quad \forall v \forall (w \ne 0) & \Leftrightarrow \\ (Qw)^T \underbrace{(\|w\|_2 v)}_{-y} - \frac{1}{2} (\|w\|_2 v)^T A(t) (\|w\|_2 v) & \le & \tau w^T w \quad \forall v \quad \forall (w \ne 0) & \Leftrightarrow \\ 2\tau w^T w + 2 w^T Q^T y + y^T A(t) y & \ge & 0 \quad \forall y \forall (w \ne 0) & \Leftrightarrow \\ 2\tau w^T w + 2 w^T Q^T y + y^T A(t) y & \ge & 0 \quad \forall y \in \mathbf{R}^m, w \in \mathbf{R}^k & \Leftrightarrow \\ \begin{pmatrix} 2\tau I_k & Q^T \\ Q & A(t) \end{pmatrix} & \succeq & 0. \quad \square & \end{array}$$

Universal semidefinite form of the standard SSD problem. Both the multiload standard SSD problem (4.8.101) and the robust obstacle-free problem (4.8.103) are particular cases of the following generic semidefinite program:

$$
\begin{array}{lll}
\text{minimize} & \tau & \\
\text{s.t.} & & \\
\text{(a)} & \begin{pmatrix} 2\tau I_p + \mathcal{D}_l z + D_l & [\mathcal{E}_l z + E_l]^T \\ [\mathcal{E}_l z + E_l] & \sum_{i=1}^{n}\sum_{s=1}^{S} b_{is} t_i b_{is}^T \end{pmatrix} \succeq 0,\ l = 1, \ldots, K, & \\
\text{(b)} & t_i \succeq 0,\ i = 1, \ldots, n, & \\
\text{(c)} & \sum_{i=1}^{n} \mathrm{Tr}(t_i) \leq w, & \\
\text{(d)} & \underline{\rho}_i \leq \mathrm{Tr}(t_i) \leq \overline{\rho}_i,\ i = 1, \ldots, n, & \\
\text{(e)} & z \geq 0, &
\end{array}
\qquad \text{(Pr)}
$$

where

- the design variables are $t_i \in \mathbf{S}^d$, $i = 1, \ldots, n$, $z \in \mathbf{R}^N$, $\tau \in \mathbf{R}$, and
- the data are given by the $m \times d$ matrices b_{is}, affine mappings

$$z \mapsto \mathcal{D}_l z + D_l : \mathbf{R}^N \to \mathbf{S}^d, \quad z \mapsto \mathcal{E}_l z + E_l : \mathbf{R}^N \to \mathbf{M}^{m,p},\ l = 1, \ldots, K,$$

and the reals $\underline{\rho}_i, \overline{\rho}_i$, $i = 1, \ldots, n$, $w > 0$.

Indeed,

• The multiload problem (4.8.101) corresponds to the case of $p = 1$, $K = k$ (the number of loading scenarios),

$$z = (\mu^1, \ldots, \mu^k) \in \mathbf{R}^q \times \cdots \times \mathbf{R}^q, \mathcal{D}_l z + D_l = -2r^T \mu^l, \quad \mathcal{E}_l z + E_l = -f_l + R^T \mu^l.$$

• The robust problem (4.8.103) corresponds to the case of $K = 1$, $p = k$ (the dimension of the loading ellipsoid). In fact, in the robust problem there should be no z-variable at all; however, to avoid extra comments, we introduce a 1D redundant variable z and set

$$\mathcal{E}_1 = 0, E_1 = Q; \mathcal{D}_1 z + D_1 = -2z I_k.$$

It is immediately seen that problem (Pr) reduces for this case to the problem

$$\min_{\tau,z,t} \left\{ \tau : \begin{pmatrix} 2(\tau - z) I_k & Q^T \\ Q & \sum_{i=1}^{n}\sum_{s=1}^{S} b_{is} t_i b_{is}^T \end{pmatrix} \succeq 0 \ \&\ (\mathrm{Pr}.b - e) \right\},$$

which is equivalent to (4.8.103).

Note that when converting the problems of our actual interest (4.8.101) and (4.8.103) to the generic form (Pr), we ensure the following property of the resulting problem.

S.5. For every $l = 1, \ldots, K$, there exists $\alpha_l \in \mathbf{S}^p_{++}$ and $V_l \in \mathbf{M}^{m,p}$ such that

$$\sum_{l=1}^{K} \left[\mathcal{D}_l^* \alpha_l + 2\mathcal{E}_l^* V_l\right] < 0.$$

Indeed, in the case when the original problem is the multiload problem (4.8.101), we have $\mathbf{S}^p = \mathbf{R}$, $\mathbf{M}^{m,p} = \mathbf{R}^m$, and

$$\sum_{l=1}^{K} \left[\mathcal{D}_l^* \alpha_l + 2\mathcal{E}_l^* V_l\right] = \begin{pmatrix} -2\alpha_1 r + 2RV_1 \\ -2\alpha_2 r + 2RV_2 \\ \cdots \\ -2\alpha_K r + 2RV_K \end{pmatrix};$$

the latter vector is negative when all α_l are equal to 1 and all V_l are equal to a strictly feasible solution of the system $Rv \leq r$. Such a solution exists by **S.1.**

In the case when the original problem is the obstacle-free robust problem (4.8.103), we have

$$\mathcal{D}_1^* \alpha_1 + 2\mathcal{E}_1^* V_1 = -2\mathrm{Tr}(\alpha_1),$$

and to guarantee the validity of **S.5**, it suffices to set $\alpha_1 = I_p$.

From now on we assume that the data of (Pr) satisfy **S.3**, **S.4**, and **S.5.**

Remark 4.8.1. *Problem* (Pr) *is strictly feasible.*

Indeed, let us choose somehow $z > 0$. By **S.3**, we can choose $t_i \succ 0$, $i = 1, \ldots, n$, to satisfy the strict versions of the inequalities (Pr)(b), (c), (d). By **S.4** the matrix $\sum_{i=1}^{n} \sum_{s=1}^{S} b_{is} t_i b_{is}^T$ is positive definite. But then, by the Schur complement lemma, LMIs (Pr)(a) are satisfied strictly for all large-enough values of τ.

4.8.4 From primal to dual

From the viewpoint of numerical processing, a disadvantage of the problem (Pr) is its huge design dimension. Consider, e.g., the case of a multiload design of an obstacle-free truss with M-node ground structure. In this case (Pr) is a semidefinite program of design dimension $n + 1$, $n = O(M^2)$ being the number of tentative bars. The program includes k (k is the number of loading scenarios) big LMIs (each of the row size $m + 1$, where m is the number of degrees of freedom of the nodal set; $m \approx 2M$ for planar and $m \approx 3M$ for spatial constructions) and a number of scalar linear inequality constraints. For a 15×15 planar nodal grid with the leftmost nodes fixed, we get $M = 225$, $n + 1 = 25096$, $m = 420$. Even an LP program with 25,000 variables should not be treated as a small one; a semidefinite program of such a design dimension is definitely not accessible for existing software. The situation, however, is not hopeless, and the way to overcome the difficulty is offered by duality; we shall show that the problem dual to (Pr) can be greatly simplified by analytical elimination of most of the variables. For example, the dual to the outlined multiload truss problem can be converted to a semidefinite program with nearly mk design variables; for the above 15×15 ground structure and three scenarios, its design dimension is about 1300, which is within the range of applicability of the existing solvers.

We are about to build the problem dual to (Pr) and to process it analytically.

Step 0. Building the dual. Applying to (Pr) our formalism for passing from a semidefinite program to its dual, we obtain the following semidefinite program:

$$
\begin{array}{lrcl}
\text{maximize} & -\phi \equiv -\sum_{l=1}^{K} \mathrm{Tr}(D_l\alpha_l + 2E_l^T V_l) - \sum_{i=1}^{n} [\overline{\rho}_i \sigma_i^+ - \underline{\rho}_i \sigma_i^-] - w\gamma & & \\
\text{s.t.} & & & \\
& \begin{pmatrix} \alpha_l & V_l^T \\ V_l & \beta_l \end{pmatrix} & \succeq & 0,\ l = 1, \ldots, K \\
& & & [\alpha_l \in \mathbf{S}^p,\ \beta_l \in \mathbf{S}^m,\ V_l \in \mathbf{M}^{m,p}], \\
& \tau_i & \geq & 0,\ i = 1, \ldots, n \\
& & & [\tau_l \in \mathbf{S}^{D_l}], \\
& \sigma_i^+,\ \sigma_i^- & \geq & 0,\ i = 1, \ldots, n \\
& & & [\sigma_i^+,\ \sigma_i^- \in \mathbf{R}], \\
& \gamma & \geq & 0 \\
& & & [\gamma \in \mathbf{R}], \\
& \eta & \geq & 0 \\
& & & [\eta \in \mathbf{R}^N], \\
& 2\sum_{l=1}^{K} \mathrm{Tr}(\alpha_l) & = & 1, \\
& \sum_{l=1}^{K} [\mathcal{D}_l^* \alpha_l + 2\mathcal{E}_l^* V_l] + \eta & = & 0, \\
& \sum_{l=1}^{K} \sum_{s=1}^{S} b_{is}^T \beta_l b_{is} + \tau_i + [\sigma_i^- - \sigma_i^+ - \gamma] I_d & = & 0,\ i = 1, \ldots, n,
\end{array}
\tag{$\mathrm{D}_{\mathrm{ini}}$}
$$

with design variables $\{\alpha_l, \beta_l, V_l\}_{l=1}^K$, $\{\sigma_i^+.\sigma_i^-, \tau_i\}_{i=1}^n$, γ, η.

Step 1. Eliminating η and $\{\tau_i\}_{i=1}^n$. ($\mathrm{D}_{\mathrm{ini}}$) clearly is equivalent to the problem

$$
\begin{array}{lrcl}
\text{minimize} & \phi \equiv \sum_{l=1}^{K} \mathrm{Tr}(D_l\alpha_l + 2E_l^T V_l) + \sum_{i=1}^{n} [\overline{\rho}_i \sigma_i^+ - \underline{\rho}_i \sigma_i^-] + w\gamma & & \\
\text{s.t.} & & & \\
\text{(a)} & \begin{pmatrix} \alpha_l & V_l^T \\ V_l & \beta_l \end{pmatrix} & \succeq & 0,\ l = 1, \ldots, K, \\
\text{(b)} & \sigma_i^+,\ \sigma_i^- & \geq & 0,\ l = 1, \ldots, n, \\
\text{(c)} & \gamma & \geq & 0, \\
\text{(d)} & 2\sum_{l=1}^{K} \mathrm{Tr}(\alpha_l) & = & 1, \\
\text{(e)} & \sum_{l=1}^{K} [\mathcal{D}_l^* \alpha_l + 2\mathcal{E}_l^* V_l] & \preceq & 0, \\
\text{(f)} & \sum_{l=1}^{K} \sum_{s=1}^{S} b_{is}^T \beta_l b_{is} & \preceq & [\gamma + \sigma_i^+ - \sigma_i^-] I_d,\ i = 1, \ldots, n,
\end{array}
\tag{D$'$}
$$

in variables $\{\alpha_l, \beta_l, V_l\}_{l=1}^K$, $\{\sigma_i^+.\sigma_i^-\}_{i=1}^n$, γ. Note that passing from ($\mathrm{D}_{\mathrm{ini}}$) to (D′), we have switched from maximization of $-\phi$ to minimization of ϕ, so that the optimal value in (D′) is minus the one of ($\mathrm{D}_{\mathrm{ini}}$).

Step 2. Eliminating $\{\beta_l\}_{l=1}^K$. We start with observing that (D′) is strictly feasible. Indeed, let us set, say, $\sigma_i^{\pm} = 1$. By **S.5** there exist positive definite matrices α_l and rectangular matrices V_l of appropriate sizes satisfying the strict version of (D′)(e); by normalization, we may enforce these matrices to satisfy (D′)(d). Given indicated α_l, V_l and choosing large-enough β_l, we enforce validity of the strict versions of (D′)(a). Finally, choosing large-enough $\gamma > 0$, we enforce strict versions of (D′)(c) and (D′)(f).

Note that the same arguments demonstrate the following remark.

Remark 4.8.2. Problem ($\mathrm{D}_{\mathrm{ini}}$) is strictly feasible.

Since (D′) is strictly feasible, its optimal value is the same as in problem (D″) obtained from (D′) by adding the constraints

$$\text{(g)} \quad \alpha_l \succ 0, \ l = 1, \ldots, K.$$

Now note that if a collection

$$(\alpha = \{\alpha_l\}_{l=1}^K, V = \{V_l\}_{l=1}^K, \beta = \{\beta_l\}_{l=1}^K, \sigma = \{\sigma_i^{\pm}\}_{i=1}^n, \gamma)$$

is a feasible solution to (D″), then the collection

$$(\alpha, V, \beta(\alpha, V) = \{\beta_l(\alpha, V) = V_l \alpha_l^{-1} V_l^T\}_{l=1}^K, \sigma, \gamma)$$

is also a feasible solution to (D″) with the same value of the objective. Indeed, from the LMI (D′)(a) by the Schur complement lemma it follows that $\beta_l(\alpha, V) \preceq \beta_l$, so that replacing β_l with $\beta_l(\alpha, V)$ we preserve validity of the LMIs (D′)(f) as well as (D′)(a). Consequently, (D″) is equivalent to the problem

$$\begin{array}{lrcl}
\text{minimize} & \phi \equiv \displaystyle\sum_{l=1}^K \mathrm{Tr}(D_l \alpha_l + 2E_l^T V_l) + \sum_{i=1}^n [\overline{\rho}_i \sigma_i^+ - \underline{\rho}_i \sigma_i^-] + w\gamma & & \\
\text{s.t.} & & & \\
\text{(b)} & \sigma_i^+, \sigma_i^- & \geq & 0, \ i = 1, \ldots, n, \\
\text{(c)} & \gamma & \geq & 0, \\
\text{(d)} & 2\displaystyle\sum_{l=1}^K \mathrm{Tr}(\alpha_l) & = & 1, \\
\text{(e)} & \displaystyle\sum_{l=1}^K [\mathcal{D}_l^* \alpha_l + 2\mathcal{E}_l^* V_l] & \preceq & 0, \\
\text{(f′)} & \displaystyle\sum_{l=1}^K \sum_{s=1}^S b_{is}^T V_l \alpha_l^{-1} V_l^T b_{is} & \preceq & [\gamma + \sigma_i^+ - \sigma_i^-] I_d, \ i = 1, \ldots, n, \\
\text{(g)} & \alpha_l & \succ & 0, \ l = 1, \ldots, K,
\end{array} \qquad (\mathrm{D}''')$$

in variables $\alpha = \{\alpha_l\}_{l=1}^K$, $V = \{V_l\}_{l=1}^K$, $\sigma = \{\sigma_i^{\pm}\}_{i=1}^n$, γ.

Now note that the system of LMIs (D‴)(g)–(D‴)(f′) is equivalent to the system of LMIs

$$\begin{array}{lll} \text{(a)} & \begin{pmatrix} A(\alpha), & B_i^T(V) \\ B_i(V), & (\gamma + \sigma_i^+ - \sigma_i^-) I_d \end{pmatrix} & \succeq \ 0, \ i = 1, \ldots, n, \\ \text{(b)} & A(\alpha) & \succ \ 0, \end{array} \tag{4.8.104}$$

where

$$\begin{aligned} \alpha &= \{\alpha_l \in \mathbf{S}^p\}_{l=1}^K, \\ V &= \{V_l \in \mathbf{M}^{p,m}\}_{l=1}^K, \\ A(\alpha) &= \mathrm{Diag}(\overbrace{\alpha_1, \ldots, \alpha_1}^{S \text{ times}}, \overbrace{\alpha_2, \ldots, \alpha_2}^{S \text{ times}}, \ldots, \overbrace{\alpha_K, \ldots, \alpha_K}^{S \text{ times}}), \\ B_i(V) &= [b_{i1}^T V_1, b_{i2}^T V_1, \ldots, b_{iS}^T V_1; b_{i1}^T V_2, b_{i2}^T V_2, \ldots, b_{iS}^T V_2; \ldots; b_{i1}^T V_K, b_{i2}^T V_K, \ldots, b_{iS}^T V_K]. \end{aligned} \tag{4.8.105}$$

Indeed, the difference of the right- and left-hand sides of (D‴)(f′) is the Schur complement of the angular block of the left-hand-side matrix in (4.8.104)(a), and it remains to apply the lemma on the Schur complement. Consequently, (D‴) is equivalent to the problem

$$\text{minimize} \qquad \phi \equiv \sum_{l=1}^K \mathrm{Tr}(D_l \alpha_l + 2E_l^T V_l) + \sum_{i=1}^n [\overline{\rho}_i \sigma_i^+ - \underline{\rho}_i \sigma_i^-] + w\gamma$$

s.t.

$$\begin{array}{lll} \text{(a)} & \begin{pmatrix} A(\alpha), & B_i^T(V) \\ B_i(V), & (\gamma + \sigma_i^+ - \sigma_i^-) I_d \end{pmatrix} & \succeq \ 0, \ i = 1, \ldots, n, \\ \text{(b)} & \sigma_i^+, \sigma_i^- & \geq \ 0, \ i = 1, \ldots, n, \\ \text{(c)} & \gamma & \geq \ 0, \\ \text{(d)} & 2\sum_{l=1}^K \mathrm{Tr}(\alpha_l) & = \ 1, \\ \text{(e)} & \alpha_l & \succ \ 0, \ l = 1, \ldots, K, \\ \text{(f)} & \sum_{l=1}^K [\mathcal{D}_l^* \alpha_l + 2\mathcal{E}_l^* V_l] & \leq \ 0 \end{array}$$

in variables $\alpha, V, \sigma, \gamma$.

The resulting problem is strictly feasible along with (D′), so that its optimal value remains unchanged when we replace the strict LMIs (e) with their nonstrict counterparts $\alpha_l \succeq 0$; the latter nonstrict LMIs are already implied by (a). Eliminating the LMIs (e), we come to the final form of the problem dual to (Pr):

$$
\begin{array}{ll}
\text{minimize} & \phi \equiv \sum_{l=1}^{K} \mathrm{Tr}(D_l\alpha_l + 2E_l^T V_l) + \sum_{i=1}^{n} [\overline{\rho}_i\sigma_i^+ - \underline{\rho}_i\sigma_i^-] + w\gamma \\
\text{s.t.} & \\
& \begin{pmatrix} A(\alpha), & B_i^T(V) \\ B_i(V), & (\gamma + \sigma_i^+ - \sigma_i^-)I_d \end{pmatrix} \succeq 0,\ l = 1, \ldots, N, \\
& \sigma_i^+, \sigma_i^- \geq 0,\ i = 1, \ldots, n, \\
& \gamma \geq 0, \\
& 2\sum_{l=1}^{K} \mathrm{Tr}(\alpha_l) = 1, \\
& \sum_{l=1}^{K} [\mathcal{D}_l^*\alpha_l + 2\mathcal{E}_l^* V_l] \leq 0,
\end{array}
\tag{Dl}
$$

in design variables

$$
\alpha = \{\alpha_l \in \mathbf{S}^p\}_{l=1}^K,\ V = \{V_i \in \mathbf{M}^{m,p}\}_{l=1}^K,\ \sigma = \{\sigma_i^{\pm} \in \mathbf{R}\}_{i=1}^n,\ \gamma \in \mathbf{R}.
$$

As we have seen, both the primal problem (Pr) and its dual ($\mathrm{D}_{\mathrm{ini}}$) are strictly feasible (Remarks 4.8.1, 4.8.2). Consequently, both (Pr) and ($\mathrm{D}_{\mathrm{ini}}$) are solvable with equal optimal values (the conic duality theorem) and with bounded level sets (see Exercise 2.12). By its origin, the optimal value in the problem (Dl) is minus the optimal value in ($\mathrm{D}_{\mathrm{ini}}$), and of course (Dl) inherits from ($\mathrm{D}_{\mathrm{ini}}$) the property to have bounded level sets and is therefore solvable. Thus, we get the following proposition.

Proposition 4.8.3. *Both problems* (Pr), (Dl) *are strictly feasible and solvable and possess bounded level sets. The optimal values in these problems are negations of each other.*

Case of simple bounds. In the case when there are no actual bounds on $\mathrm{Tr}(t_i)$ (formally it means that $\underline{\rho}_i = 0, \overline{\rho}_i = w\ \forall i$), the dual problem (Dl) can be further simplified, namely, we can eliminate the σ-variables. Indeed, consider a feasible solution to (Dl). When replacing all $\sigma_i^{\pm}$ with zeros, simultaneously increasing γ by $\delta = \max[0, \max_i(\sigma_i^+ - \sigma_i^-)]$, we clearly preserve feasibility and add to the objective the quantity

$$
\begin{aligned}
w\delta - \sum_{i=1}^{n} \left[\overline{\rho}_i\sigma_i^+ - \underline{\rho}_i\sigma_i^-\right] &= w\delta - \sum_{i=1}^{n} \overline{\rho}_i\sigma_i^+ && [\text{since } \underline{\rho}_i = 0] \\
&\leq w\delta - w\sum_{i=1}^{n} \sigma_i^+ && [\text{since } \overline{\rho}_i \geq w] \\
&\leq w \max_i \sigma_i^+ - w\sum_{i=1}^{n} \sigma_i^+ && [\text{since } \delta \leq \max_i \sigma_i^+ \text{ due to } \sigma_i^{\pm} \geq 0] \\
&\leq 0 && [\text{since } \sigma_i^+ \geq 0],
\end{aligned}
$$

i.e., we gain in the objective value. Thus, we loose nothing when setting in (Dl) $\sigma_i^{\pm} = 0$, thus coming to the problem

$$\begin{array}{ll} \text{minimize} & \phi = \sum_{l=1}^{K} \mathrm{Tr}(D_l\alpha_l + 2E_l^T V_l) + w\gamma \\ \text{s.t.} & \\ & \begin{pmatrix} A(\alpha) & B_i^T(V) \\ B_i(V) & \gamma I_d \end{pmatrix} \succeq 0, \ i = 1, \ldots, n, \\ & \sum_{l=1}^{K} [\mathcal{D}_l^*\alpha_l + 2\mathcal{E}_l^* V_l] \leq 0, \\ & 2\sum_{l=1}^{K} \mathrm{Tr}(\alpha_l) = 1, \\ & \gamma \geq 0, \end{array} \qquad (\mathrm{Dl}_{\mathrm{sb}})$$

in design variables $\alpha = \{\alpha_l \in \mathbf{S}^p\}_{l=1}^K$, $V = \{V_l \in \mathbf{M}^{p,m}\}_{l=1}^K$, $\gamma \in \mathbf{R}$.

To understand how fruitful our effort was, let us compare the sizes of the problem ($\mathrm{Dl}_{\mathrm{sb}}$) with those of the original problem (Pr) in the simplest case of a planar k-load obstacle-free truss design problem with M nodes and simple bounds. Note that in this case $n \approx 0.5M^2$ and $m \approx 2M$. Assuming $k << M$, here are the sizes of (Pr) and ($\mathrm{Dl}_{\mathrm{sb}}$):

Size	(Pr)	($\mathrm{Dl}_{\mathrm{sb}}$)
Design dimension	$n + 1 \approx 0.5M^2$	$mk + k + 1 \approx 2kM$
# and sizes of LMIs	k of $(2m + 1) \times (2m + 1)$ LMIs	$n \approx 0.5M^2$ of $(k + 1) \times (k + 1)$ LMIs
# of linear constraints	$n + 1 \approx 0.5M^2$	$k + 1$

We see that if the number of loading scenarios k is a small integer (which normally is the case), the design dimension of the dual problem is by orders of magnitude less than the design dimension of the primal problem (Pr). As a kind of penalization, the dual problem involves a lot ($\approx 0.5M^2$) of nonscalar LMIs instead of just k nonscalar LMIs in (Pr), but all LMIs in (Dl) are small—of row size $k + 1$ each—while the nonscalar LMIs in (Pr) are large—of row size $\approx 2M$ each. As a result, when solving (Pr) and ($\mathrm{Dl}_{\mathrm{sb}}$) by the best-known numerical techniques so far (the interior-point algorithms), the computational effort for (Pr) turns out to be $O(M^6)$, while for ($\mathrm{Dl}_{\mathrm{sb}}$) it is only $O(k^3M^3)$. For large M and small k, this does make a difference!

Of course, there is an immediate concern about the dual problem: The actual design variables are not seen in it at all. How do we recover a (nearly) optimal construction from a (nearly) optimal solution to the dual problem? In fact, however, there is no reason to be concerned: the required recovering routines exist and are cheap computationally.

4.8.5 Back to primal

Problem (Dl) is not exactly the dual of (Pr)—it is obtained from this dual by eliminating part of the variables. What happens when we pass from (Dl) to its dual? It turns out that we end up with a nontrivial (and instructive) equivalent reformulation of (Pr), namely, with the problem

$$
\begin{array}{ll}
\text{minimize} & \tau \\
\text{s.t.} & \\
\text{(a)} & \left(\begin{array}{c|ccc|c|ccc}
2\tau I_p + \mathcal{D}_l z + D_l & [q^l_{11}]^T & \cdots & [q^l_{1S}]^T & \cdots & [q^l_{n1}]^T & \cdots & [q^l_{nS}]^T \\ \hline
q^l_{11} & t_1 & & & & & & \\
\vdots & & \ddots & & & & & \\
q^l_{1S} & & & t_1 & & & & \\ \hline
\vdots & & & & \ddots & & & \\ \hline
q^l_{n1} & & & & & t_n & & \\
\vdots & & & & & & \ddots & \\
q^l_{nS} & & & & & & & t_n
\end{array}\right) \succeq 0, \\
& \qquad l = 1, \dots, K, \\
\text{(b)} & \underline{\rho}_i \le \mathrm{Tr}(t_i) \le \overline{\rho}_i, \ \ i = 1, \dots, n, \\
\text{(c)} & \sum_{i=1}^{n} \mathrm{Tr}(t_i) \le w, \\
\text{(d)} & \sum_{i=1}^{n} \sum_{s=1}^{S} b_{is} q^l_{is} = \mathcal{E}_l z + E_l, \ \ l = 1, \dots, k, \\
\text{(e)} & z \ge 0,
\end{array}
\qquad (\mathrm{Pr}^+)
$$

in the design variables which are symmetric $d \times d$ matrices t_i, $i = 1, \dots, n$, $d \times p$ matrices q^l_{is}, $l = 1, \dots, K$, $i = 1, \dots, n$, $s = 1, \dots, S$, real τ, and $z \in \mathbf{R}^N$. Problem (Pr^+) is not the straightforward dual of (Dl); it is obtained from this dual by eliminating part of the variables. Instead of boring derivation of (Pr^+) via duality, we prefer to give a direct proof of equivalence between (Pr) and (Pr^+).

Proposition 4.8.4. *A collection* $(\{t_i\}_{i=1}^n, z, \tau)$ *is a feasible solution to* (Pr) *if and only if it can be extended by properly chosen* $\{q^l_{is} \mid l = 1, \dots, K, i = 1, \dots, n, s = 1, \dots, S\}$ *to a feasible solution to* (Pr^+).

Proof. "If" part. Let a collection

$$(\{t_i\}_{i=1}^n, z, \tau, \{q^l_{is} \mid l = 1, \dots, K, i = 1, \dots, n, s = 1, \dots, S\})$$

be a feasible solution to (Pr^+); all we should prove is the validity of the LMIs (Pr.a). Let us fix $l \le K$. We should prove that for every pair (x, y) of vectors of appropriate dimensions we have

$$x^T [2\tau I_p + \mathcal{D}_l z + D_l] x + 2x^T [E_i + \mathcal{E}_l z]^T y + y^T \left[\sum_{i=1}^{n} \sum_{s=1}^{S} b_{is}^T t_i b_{is}^T \right] y \ge 0. \qquad (4.8.106)$$

Indeed, in view of $(\mathrm{Pr}^+.d)$, the left-hand side of (4.8.106) is equal to

$$
\begin{array}{rcl}
x^T[2\tau I_p + \mathcal{D}_l z + D_l]x & & \\
+2x^T\left[\sum_{i=1}^n\sum_{s=1}^S b_{is}q_{is}^l\right]^T y + y^T\left[\sum_{i=1}^n\sum_{s=1}^S b_{is}t_i b_{is}^T\right]y & = & x^T[2\tau I_p + \mathcal{D}_l z + D_l]x \\
& & + 2\sum_{i=1}^n\sum_{s=1}^S x^T[q_{is}^l]^T y_{is} \\
& & + \sum_{i=1}^n\sum_{s=1}^S y_{is}^T t_i y_{is}, \\
& & y_{is} = b_{is}^T y.
\end{array}
$$

The resulting expression is nothing but the value of the quadratic form with the matrix from the left-hand side of the corresponding LMI (Pr^+)(a) at the vector comprising x and $\{y_{is}\}_{i,s}$, and therefore it is nonnegative, as claimed.

"Only if" part. Let

$$(\{t_i\}_{i=1}^n, z, \tau)$$

be a feasible solution to (Pr). Let us fix l, $1 \le l \le K$, and let us set

$$f_l = \mathcal{E}_l z + E_l.$$

For every $x \in \mathbf{R}^p$ the quadratic form of $y \in \mathbf{R}^M$:

$$x^T[2\tau I_p + \mathcal{D}_l z + d_l] + 2x^T f_l^T y + y^T A(t) y \quad \left[A(t) = \sum_{i=1}^n\sum_{s=1}^S b_{is}t_i b_{is}^T\right]$$

is nonnegative, i.e., the equation

$$A(t)y = f_l x$$

is solvable for every x. Of course, we can choose its solution to be linear in x:

$$y = Y_l x.$$

Note that then

$$A(t)Y_l x = f_l x \quad \forall x,$$

i.e.,

$$A(t)Y_l = f_l.$$

Let us now set

$$[q_{is}^l]^T = Y_l^T b_{is} t_i. \tag{4.8.107}$$

Then

$$\sum_{i=1}^{n}\sum_{s=1}^{S} b_{is}q_{is}^{l} = \sum_{i=1}^{n}\sum_{s=1}^{S} b_{is}t_i b_{is}^T Y_l = A(t)Y_l = f_l.$$

Recalling the definition of f_l, we see that extending $(\{t_i\}, z, \tau)$ by $\{q_{is}^l\}$ we ensure the validity of (Pr^+)(d). It remains to verify that the indicated extensions ensure the validity of LMIs (Pr^+)(a) as well. What we should verify is that for every collection $\{y_{is}\}$ of vectors of appropriate dimension and for every $x \in \mathbf{R}^p$ we have

$$F(x,\{y_{is}\}) \equiv x^T[2\tau I_p + \mathcal{D}_l z + D_l]x + 2x^T \sum_{i=1}^{n}\sum_{s=1}^{S}[q_{is}^l]^T y_{is} + \sum_{i=1}^{n}\sum_{s=1}^{S} y_{is}^T t_i y_{is} \geq 0. \tag{4.8.108}$$

Given x, let us set

$$y_{is}^* = -b_{is}^T Y_l x,$$

and let us prove that the collection $\{y_{is}^*\}$ minimizes $F(x,\cdot)$, which is immediate: $F(x,\cdot)$ is convex quadratic form, and its partial derivative with respect to y_{is} at the point $\{y_{is}^*\}$ is equal to (see (4.8.107))

$$2q_{is}^l x + 2t_i y_{is}^* = 2[t_i b_{is}^T Y_l x - t_i b_{is}^T Y_l x] = 0$$

$\forall i, s$. It remains to note that

$$\begin{aligned}
F(x,\{y_{is}^*\}) &= x^T[2\tau I_p + \mathcal{D}_l z + D_l]x - 2x^T \sum_{i=1}^{n}\sum_{s=1}^{S}[q_{is}^l]^T b_{is}^T Y_l x \\
&\quad + \sum_{i=1}^{n}\sum_{s=1}^{S} x^T Y_l^T b_{is} t_i b_{is}^T Y_l x \\
&= x^T[2\tau I_p + \mathcal{D}_l z + D_l]x - 2x^T\left[\sum_{i=1}^{n}\sum_{s=1}^{S} b_{is}q_{is}^l\right]^T Y_l x \\
&\quad + x^T Y_l^T A(t) Y_l x \\
&= x^T[2\tau I_p + \mathcal{D}_l z + D_l]x - 2x^T[\mathcal{E}_l z + E_l]^T Y_l x + x^T Y_l^T A(t) Y_l x \\
&\quad \text{[due to already proved } (\mathrm{Pr}^+)\text{(d)]} \\
&= (x^T; -x^T Y_l^T)\begin{pmatrix} \tau I_p + \mathcal{D}_l z + D_l & [\mathcal{E}_l z + E_l]^T \\ \mathcal{E}_l z + E_l & A(t) \end{pmatrix}\begin{pmatrix} x \\ -Y_l x \end{pmatrix} \\
&\geq 0 \\
&\quad \text{[since } (\{t_i\}, z, \tau) \text{ is feasible for (Pr)].}
\end{aligned}$$

Thus, the minimum of $F(x,\{y_{is}\})$ in $\{y_{is}\}$ is nonnegative, and therefore (4.8.108) indeed is valid. □

4.8.6 Explicit forms of the standard truss and shape problems

Let us list the explicit forms of problems (Pr), (Dl), (Pr$^+$) for the standard cases of the multiload and robust static truss and shape design.

Multiload static truss design. Here

$$\mathcal{V} = \{v \in \mathbf{R}^m \mid Rv \le r\} \quad [\dim(r) = q]; \qquad \mathcal{F} = \{f_1, \dots, f_k\}.$$

The settings are

- (Pr):

$$\begin{array}{ll}\text{minimize} & \tau \\ \text{s.t.} & \\ & \begin{pmatrix} 2\tau - 2r^T\mu_l & -f_l^T + \mu_l^T R \\ -f_l + R^T\mu_l & \sum_{i=1}^n b_i b_i^T t_i \end{pmatrix} \succeq 0,\ l = 1, \dots, k, \\ & \underline{\rho}_i \le t_i \le \overline{\rho}_i, i = 1, \dots, n, \\ & \sum_{i=1}^n t_i \le w, \\ & \mu_l \ge 0,\ l = 1, \dots, k; \\ & [\tau, t_i, \mu_l \in \mathbf{R}^q];\end{array}$$

- (Dl):

$$\begin{array}{ll}\text{minimize} & -2\sum_{l=1}^k f_l^T v_l + \sum_{i=1}^n [\overline{\rho}_i \sigma_i^+ - \underline{\rho}_i \sigma_i^-] + w\gamma \\ \text{s.t.} & \\ & \left(\begin{array}{ccc|c} \alpha_1 & & & b_i^T v_1 \\ & \ddots & & \vdots \\ & & \alpha_k & b_i^T v_k \\ \hline b_i^T v_1 & \cdots & b_i^T v_k & \gamma + \sigma_i^+ - \sigma_i^- \end{array}\right) \succeq 0,\ i = 1, \dots, n, \\ & \sigma_i^\pm \ge 0,\ i = 1, \dots, n, \\ & \gamma \ge 0, \\ & Rv_l \le \alpha_l r,\ l = 1, \dots, k, \\ & 2\sum_{l=1}^k \alpha_l = 1 \\ & [\alpha_l, \sigma_i^\pm, \gamma \in \mathbf{R}, v_l \in \mathbf{R}^m];\end{array}$$

- (Pr$^+$):

$$
\begin{array}{ll}
\text{minimize} & \tau \\
\text{s.t.} & \\
& \left(\begin{array}{c|ccc} 2\tau - 2r^T\mu_l & q_1^l & \cdots & q_n^l \\ \hline q_1^l & t_1 & & \\ \vdots & & \ddots & \\ q_n^l & & & t_n \end{array}\right) \succeq 0, \ l = 1, \ldots, k, \\
& \underline{\rho}_i \le t_i \le \overline{\rho}_i, \ i = 1, \ldots, n, \\
& \sum_{i=1}^n t_i \le w, \\
& \sum_{i=1}^n q_i^l b_i = f_l - R^T\mu_l, \ l = 1, \ldots, k, \\
& \mu_l \ge 0, \ l = 1, \ldots, k \\
& [\tau, t_i, q_i^l \in \mathbf{R}, \mu_l \in \mathbf{R}^q].
\end{array}
$$

Robust obstacle-free static truss design. Here

$$\mathcal{V} = \mathbf{R}^m; \quad \mathcal{F} = \{f = Qu \mid u^T u \le 1\} \quad [Q \in \mathbf{M}^{m,k}].$$

The settings are

- (Pr):

$$
\begin{array}{ll}
\text{minimize} & \tau \\
\text{s.t.} & \\
& \left(\begin{array}{cc} 2\tau I_k & Q^T \\ Q & \sum_{i=1}^n b_i b_i^T t_i \end{array}\right) \succeq 0, \\
& \underline{\rho}_i \le t_i \le \overline{\rho}_i, i = 1, \ldots, n, \\
& \sum_{i=1}^n t_i \le w \\
& [\tau, t_i \in \mathbf{R}];
\end{array}
$$

- (Dl):

$$
\begin{array}{ll}
\text{minimize} & 2\mathrm{Tr}(Q^T V) + \sum_{i=1}^n [\overline{\rho}_i \sigma_i^+ - \underline{\rho}_i \sigma_i^-] + w\gamma \\
\text{s.t.} & \\
& \left(\begin{array}{cc} \alpha & V^T b_i \\ b_i^T V & \gamma + \sigma_i^+ - \sigma_i^- \end{array}\right) \succeq 0, \ i = 1, \ldots, n, \\
& \sigma_i^{\pm} \ge 0, \ i = 1, \ldots, n, \\
& \gamma \ge 0, \\
& 2\mathrm{Tr}(\alpha) = 1 \\
& [\alpha \in \mathbf{S}^k, \sigma_i^{\pm}, \gamma \in \mathbf{R}, V \in \mathbf{M}^{m,k}];
\end{array}
$$

- (Pr$^+$):

$$
\begin{array}{ll}
\text{minimize} & \tau \\
\text{s.t.} & \\
& \left(\begin{array}{c|ccc} 2\tau I_k & q_1^T & \cdots & q_n^T \\ \hline q_1 & t_1 & & \\ \vdots & & \ddots & \\ q_n & & & t_n \end{array}\right) \succeq 0, \\
& \underline{\rho}_i \le t_i \le \overline{\rho}_i,\ i = 1, \ldots, n, \\
& \sum_{i=1}^{n} t_i \le w, \\
& \sum_{i=1}^{N} b_i q_i = Q \\
& [t_i \in \mathbf{R}, q_i^T \in \mathbf{R}^k].
\end{array}
$$

Multiload static shape design. Here

$$
\begin{aligned}
t_i &\in \mathbf{S}^d,\ d = \begin{cases} 3, & \text{planar shape,} \\ 6, & \text{spatial shape,} \end{cases} \\
\mathcal{F} &= \{f_1, \ldots, f_k\}, \\
\mathcal{V} &= \{v \in \mathbf{R}^m \mid Rv \le r\} \quad [\dim(r) = q].
\end{aligned}
$$

The settings are

- (Pr):

$$
\begin{array}{ll}
\text{minimize} & \tau \\
\text{s.t.} & \\
& \left(\begin{array}{cc} 2\tau - 2r^T\mu_l & -f_l^T + \mu_l^T R \\ -f_l + R^T\mu_l & \sum_{i=1}^{n}\sum_{s=1}^{S} b_{is} t_i b_{is}^T \end{array}\right) \succeq 0,\ l = 1, \ldots, k, \\
& t_i \succeq 0,\ i = 1, \ldots, n, \\
& \underline{\rho}_i \le \operatorname{Tr}(t_i) \le \overline{\rho}_i, i = 1, \ldots, n, \\
& \sum_{i=1}^{n} \operatorname{Tr}(t_i) \le w, \\
& \mu_l \ge 0,\ l = 1, \ldots, k \\
& [\tau \in \mathbf{R}, t_i \in \mathbf{S}^d, \mu_l \in \mathbf{R}^q];
\end{array}
$$

• (Dl):

$$\text{minimize} \qquad -2\sum_{l=1}^{k} f_l^T v_l + \sum_{i=1}^{n} [\overline{\rho}_i \sigma_i^+ - \underline{\rho}_i \sigma_i^-] + w\gamma$$

s.t.

$$\left(\begin{array}{ccc|c|ccc|c}
\alpha_1 & & & & & & & v_1^T b_{i1} \\
& \ddots & & & & & & \vdots \\
& & \alpha_1 & & & & & v_1^T b_{iS} \\
\hline
& & & \ddots & & & & \vdots \\
\hline
& & & & \alpha_k & & & v_k^T b_{i1} \\
& & & & & \ddots & & \vdots \\
& & & & & & \alpha_k & v_k^T b_{iS} \\
\hline
b_{i1}^T v_1 & \cdots & b_{iS}^T v_1 & \cdots & b_{i1}^T v_k & \cdots & b_{iS}^T v_k & (\gamma + \sigma_i^+ - \sigma_i^-) I_d
\end{array}\right) \succeq 0, \ i = 1, \ldots, n;$$

$$\begin{aligned}
\sigma_i^{\pm} &\geq 0, \ i = 1, \ldots, n, \\
\gamma &\geq 0, \\
Rv_l &\leq \alpha_l r, \ l = 1, \ldots, k, \\
2\sum_{l=1}^{k} \alpha_l &= 1 \\
[\alpha_l, \sigma_i^{\pm}, \gamma \in \mathbf{R}, v_l \in \mathbf{R}^m];&
\end{aligned}$$

• (Pr^+):

$$\text{minimize} \qquad \tau$$

s.t.

$$\left(\begin{array}{c|ccc|c|ccc}
2\tau - 2r^T \mu_l & [q_{11}^l]^T & \cdots & [q_{1S}^l]^T & \cdots & [q_{n1}^l]^T & \cdots & [q_{nS}^l]^T \\
\hline
q_{11}^l & t_1 & & & & & & \\
\vdots & & \ddots & & & & & \\
q_{1S}^l & & & t_1 & & & & \\
\hline
\vdots & & & & \ddots & & & \\
\hline
q_{n1}^l & & & & & t_n & & \\
\vdots & & & & & & \ddots & \\
q_{nS}^l & & & & & & & t_n
\end{array}\right) \succeq 0, \ l = 1, \ldots, k,$$

$$\begin{aligned}
\underline{\rho}_i \leq \mathrm{Tr}(t_i) \leq \overline{\rho}_i, \ i = 1, \ldots, n, \\
\sum_{i=1}^{n} \mathrm{Tr}(t_i) \leq w, \\
\sum_{i=1}^{n} \sum_{s=1}^{S} b_{is} q_{is}^l = f_l - R^T \mu_l, \ l = 1, \ldots, k, \\
\mu_l \geq 0, \ l = 1, \ldots, k \\
[\tau \in \mathbf{R}, t_i \in \mathbf{S}^d, q_{is}^l \in \mathbf{R}^d, \mu_l \in \mathbf{R}^q].
\end{aligned}$$

Robust obstacle-free static shape design. Here

$$\begin{array}{rcl} t_i & \in & \mathbf{S}^d,\ d = \begin{cases} 3, & \text{planar shape,} \\ 6, & \text{spatial shape,} \end{cases} \\ \mathcal{F} & = & \{f = Qu \mid u^T u \le 1\} \qquad [Q \in \mathbf{M}^{m,k}], \\ \mathcal{V} & = & \mathbf{R}^m. \end{array}$$

The settings are

- (Pr):

$$\begin{array}{ll} \text{minimize} & \tau \\ \text{s.t.} & \\ & \begin{pmatrix} 2\tau I_k & Q^T \\ Q & \sum_{i=1}^{n}\sum_{s=1}^{S} b_{is} t_i b_{is}^T \end{pmatrix} \succeq 0, \\ & t_i \succeq 0,\ i = 1, \ldots, n, \\ & \underline{\rho}_i \le \mathrm{Tr}(t_i) \le \overline{\rho}_i, i = 1, \ldots, n, \\ & \sum_{i=1}^{n} \mathrm{Tr}(t_i) \le w \\ & [\tau \in \mathbf{R}, t_i \in \mathbf{S}^d]; \end{array}$$

- (Dl):

$$\begin{array}{ll} \text{minimize} & 2\mathrm{Tr}(Q^T V) + \sum_{i=1}^{n} [\overline{\rho}_i \sigma_i^+ - \underline{\rho}_i \sigma_i^-] + w\gamma \\ \text{s.t.} & \\ & \left(\begin{array}{ccc|c} \alpha & & & V^T b_{i1} \\ & \ddots & & \vdots \\ & & \alpha & V^T b_{iS} \\ \hline b_{i1}^T V & \cdots & b_{iS}^T V & (\gamma + \sigma_i^+ - \sigma_i^-) I_d \end{array}\right) \succeq 0,\ i = 1, \ldots, n, \\ & \sigma_i^{\pm} \ge 0,\ i = 1, \ldots, n, \\ & \gamma \ge 0, \\ & 2\mathrm{Tr}(\alpha) = 1 \\ & [\alpha \in \mathbf{S}^k, \sigma_i^{\pm}, \gamma \in \mathbf{R}, V \in \mathbf{M}^{m,k}]; \end{array}$$

• (Pr^+):

$$
\begin{array}{ll}
\text{minimize} & \tau \\
\text{s.t.} & \\
& \left(\begin{array}{c|ccc|c|ccc}
2\tau I_k & [q_{11}]^T & \cdots & [q_{1S}]^T & \cdots & [q_{n1}]^T & \cdots & [q_{nS}]^T \\
\hline
q_{11} & t_1 & & & & & & \\
\vdots & & \ddots & & & & & \\
q_{1S} & & & t_1 & & & & \\
\hline
\vdots & & & & \ddots & & & \\
\hline
q_{n1} & & & & & t_n & & \\
\vdots & & & & & & \ddots & \\
q_{nS} & & & & & & & t_n
\end{array}\right) \succeq 0, \\
& \underline{\rho}_i \le \mathrm{Tr}(t_i) \le \overline{\rho}_i,\ i = 1, \ldots, n, \\
& \sum_{i=1}^{n} \mathrm{Tr}(t_i) \le w, \\
& \sum_{i=1}^{n} \sum_{s=1}^{S} b_{is} q_{is} = Q \\
& [\tau \in \mathbf{R}, t_i \in \mathbf{S}^d, q_{is} \in \mathbf{M}^{d,k}].
\end{array}
$$

4.9 Applications VII: Extremal ellipsoids

We already have met, on different occasions, with the notion of an ellipsoid—a set E in $\mathbf{R}^n$ that can be represented as the image of the unit Euclidean ball under an affine mapping:

$$E = \{x = Au + c \mid u^T u \le 1\} \quad [A \in \mathbf{M}^{n,q}]. \tag{Ell}$$

Ellipsoids are very convenient mathematical entities:

• It is easy to specify an ellipsoid—just indicate the corresponding matrix A and vector c.

• The family of ellipsoids is closed with respect to affine transformations: the image of an ellipsoid under an affine mapping again is an ellipsoid.

• There are many operations, like minimization of a linear form and computation of volume that are easy to carry out when the set in question is an ellipsoid and are difficult to carry out for more general convex sets.

By the indicated reasons, ellipsoids play an important role in different areas of applied mathematics. In particular, ellipsoids are used to approximate more complicated sets. As a simple motivating example, consider a discrete time linear time invariant controlled system:

$$
\begin{array}{rcl}
x(t+1) & = & Ax(t) + Bu(t),\ t = 0, 1, \ldots, \\
x(0) & = & 0
\end{array}
$$

and assume that the control is norm-bounded:

$$\|u(t)\|_2 \le 1 \quad \forall t.$$

The question is, What is the set X_T of all states reachable in a given time T, i.e., the set of all possible values of $x(T)$? We can easily write down the answer:

$$X_T = \{x = Bu_{T-1} + ABu_{T-2} + A^2Bu_{T-3} + \cdots + A^{T-1}Bu_0 \mid \|u_t\|_2 \le 1,\ t = 0, \ldots, T-1\},$$

but this answer is not explicit. To check whether a given vector x belongs to X_T requires solving a nontrivial conic quadratic problem; the larger T, the greater the complexity of the problem. In fact, the geometry of X_T may be very complicated, so that there is no possibility to get a tractable explicit description of the set. This is why in many applications it makes sense to use simple—ellipsoidal—approximations of X_T. As we shall see, approximations of this type can be computed in a recurrent and computationally efficient fashion.

It turns out that the natural framework for different problems of the best possible approximation of convex sets by ellipsoids is given by semidefinite programming. In this section we consider a number of basic problems of this type.

Preliminaries on ellipsoids. According to our definition, an ellipsoid in $\mathbf{R}^n$ is the image of the unit Euclidean ball in certain $\mathbf{R}^q$ under an affine mapping; e.g., for us a segment in $\mathbf{R}^{100}$ is an ellipsoid. Indeed, it is the image of a 1D Euclidean ball under affine mapping. In contrast to this, in geometry an ellipsoid in $\mathbf{R}^n$ is usually defined as the image of the n-dimensional unit Euclidean ball under an invertible affine mapping, i.e., as the set of the form (Ell) with additional requirements that $q = n$, i.e., that the matrix A is square, and that it is nonsingular. To avoid confusion, let us call these true ellipsoids *full-dimensional*. Note that a full-dimensional ellipsoid E admits two nice representations:

- First, E can be represented in the form (Ell) with positive definite symmetric A:

$$E = \{x = Au + c \mid u^T u \le 1\} \quad [A \in \mathbf{S}^n_{++}]. \tag{4.9.109}$$

It is clear that if a matrix A represents, via (Ell), a given ellipsoid E, the matrix AU, U being an orthogonal $n \times n$ matrix, represents E as well. It is known from linear algebra that by multiplying a nonsingular square matrix from the right by a properly chosen orthogonal matrix, we get a positive definite symmetric matrix, so that we always can parameterize a full-dimensional ellipsoid by a positive definite symmetric A.

- Second, E can be given by a strictly convex quadratic inequality:

$$E = \{x \mid (x - c)^T D(x - c) \le 1\} \quad [D \in \mathbf{S}^n_{++}]. \tag{4.9.110}$$

One may take $D = A^{-2}$, where A is the matrix from the representation (4.9.109).

Note that the set (4.9.110) makes sense and is convex when the matrix D is positive semidefinite rather than positive definite. When $D \succeq 0$ is not positive definite, the set (4.9.109) is, geometrically, an elliptic cylinder—a shift of the direct product of a full-dimensional ellipsoid in the range space of D and the complementary to this range linear subspace—the kernel of D.

In the sequel we deal a lot with volumes of full-dimensional ellipsoids. Since an invertible affine transformation $x \mapsto Ax + b : \mathbf{R}^n \to \mathbf{R}^n$ multiplies the volumes of n-dimensional domains by $|\mathrm{Det}A|$, the volume of a full-dimensional ellipsoid E given by (4.9.109) is $\kappa_n \mathrm{Det}A$, where κ_n is the volume of the n-dimensional unit Euclidean ball. To

avoid meaningless constant factors, it makes sense to pass from the usual n-dimensional volume $\mathrm{mes}_n(G)$ of a domain G to its normalized volume

$$\mathrm{Vol}(G) = \kappa_n^{-1}\mathrm{mes}_n(G),$$

i.e., to choose, as the unit of volume, the volume of the unit ball rather than the one of the cube with unit edges. From now on, speaking about volumes of n-dimensional domains, we always mean their normalized volume (and omit the word normalized). With this convention, the volume of a full-dimensional ellipsoid E given by (4.9.109) is just

$$\mathrm{Vol}(E) = \mathrm{Det}A,$$

while for an ellipsoid given by (4.9.109) the volume is

$$\mathrm{Vol}(E) = [\mathrm{Det}D]^{-1/2}.$$

Outer and inner ellipsoidal approximations. It already has been mentioned that our current goal is to realize how to solve basic problems of the best ellipsoidal approximation E of a given set S. There are two types of these problems:

- *outer approximation*, where we look for the smallest ellipsoid E containing the set S, and
- *inner approximation*, where we look for the largest ellipsoid E contained in the set S.

In both these problems, a natural way to say when one ellipsoid is smaller than another is to compare the volumes of the ellipsoids. The main advantage of this viewpoint is that it results in affine-invariant constructions: an invertible affine transformation multiplies volumes of all domains by the same constant and therefore preserves ratios of volumes of the domains.

Thus, we are interested in the largest volume ellipsoid(s) contained in a given set S and the smallest volume ellipsoid(s) containing a given set S. In fact, these extremal ellipsoids are unique, provided that S is a solid—a closed and bounded convex set with a nonempty interior, and are not too bad approximations of the set.

Theorem 4.9.1. Löwner–Fritz John theorem. *Let $S \subset \mathbf{R}^n$ be a solid. Then*

(i) *there exists and is uniquely defined the largest volume full-dimensional ellipsoid E_{in} contained in S. The concentric to E_{in} n times larger (in linear sizes) ellipsoid contains S. If S is central-symmetric, then already $\sqrt{n}$ times larger than E_{in} concentric to E_{in} ellipsoid contains S.*

(ii) *there exists and is uniquely defined the smallest volume full-dimensional ellipsoid E_{out} containing S. The concentric to E_{out} n times smaller (in linear sizes) ellipsoid is contained in S. If S is central-symmetric, then already $\sqrt{n}$ times smaller than E_{out} concentric to E_{out} ellipsoid is contained in S.*

The proof is the subject of Exercise 4.71.

The existence of extremal ellipsoids is, of course, good news. But how do we compute these ellipsoids? The possibility to compute efficiently (nearly) extremal ellipsoids heavily depends on the description of S. Let us start with two simple examples.

Example: Inner ellipsoidal approximation of a polytope. Let S be a polyhedral set given by a number of linear equalities:

$$S = \{x \in \mathbf{R}^n \mid a_i^T x \le b_i,\ i = 1, \dots, m\}.$$

Proposition 4.9.1. *Assume that S is a full-dimensional polytope (i.e., is bounded and possesses a nonempty interior). Then the largest-volume ellipsoid contained in S is*

$$E = \{x = Z_* u + z_* \mid u^T u \le 1\},$$

where Z_*, z_* are given by an optimal solution to the following semidefinite program:

$$\begin{array}{lrcl} \textit{maximize} & t & & \\ \textit{s.t.} & & & \\ \text{(a)} & t & \le & (\mathrm{Det}Z)^{1/n}, \\ \text{(b)} & Z & \succeq & 0, \\ \text{(c)} & \|Za_i\|_2 & \le & b_i - a_i^T z,\ i = 1, \dots, m, \end{array} \qquad \text{(In)}$$

with the design variables $Z \in \mathbf{S}^n, z \in \mathbf{R}^n, t \in \mathbf{R}$.

Note that (In) *indeed is a semidefinite program: both* (In)(a) *and* (In)(c) *can be represented by LMIs; see examples* 18d *and* **1-17** *in section* 4.2.

Proof. Indeed, an ellipsoid (4.9.109) is contained in S if and only if

$$a_i^T (Au + c) \le b_i \quad \forall u : u^T u \le 1$$

or, which is the same, if and only if

$$\|Aa_i\|_2 + a_i^T c = \max_{u: u^T u \le 1} [a_i^T Au + a_i^T c] \le b_i.$$

Thus, (In)(b)–(c) just express the fact that the ellipsoid $\{x = Zu + z \mid u^T u \le 1\}$ is contained in S, so that (In) is nothing but the problem of maximizing (a positive power of) the volume of an ellipsoid over ellipsoids contained in S. □

We see that if S is a polytope given by a set of linear inequalities, then the problem of the best inner ellipsoidal approximation of S is an explicit semidefinite program and as such can be efficiently solved. In contrast to this, if S is a polytope given as a convex hull of finite set,

$$S = \mathrm{Conv}\{x_1, \dots, x_m\},$$

then the problem of the best inner ellipsoidal approximation of S is computationally intractable—in this case, it is difficult just to check whether a given candidate ellipsoid is contained in S.

Example: Outer ellipsoidal approximation of a finite set. Let S be a polyhedral set given as a convex hull of a finite set of points:

$$S = \mathrm{Conv}\{x_1, \dots, x_m\}.$$

Proposition 4.9.2. *Assume that S is a full-dimensional polytope (i.e., possesses a nonempty interior). Then the smallest volume ellipsoid containing S is*

$$E = \{x \mid (x - c_*)^T D_* (x - c_*) \leq 1\},$$

where c_*, D_* are given by an optimal solution (t_*, Z_*, z_*, s_*) to the semidefinite program

$$\begin{array}{lrcl} \textit{maximize} & t & & \\ \textit{s.t.} & & & \\ \text{(a)} & t & \leq & (\mathrm{Det}Z)^{1/n}, \\ \text{(b)} & Z & \succeq & 0, \\ \text{(c)} & \begin{pmatrix} s & z^T \\ z & Z \end{pmatrix} & \succeq & 0, \\ \text{(d)} & x_i^T Z x_i - 2x_i^T z + s & \leq & 1, \ i = 1, \ldots, m, \end{array} \qquad \text{(Out)}$$

with the design variables $Z \in \mathbf{S}^n$, $z \in \mathbf{R}^n$, $t, s \in \mathbf{R}$ via the relations

$$D_* = Z_*; \, c_* = Z_*^{-1} z_*.$$

Note that (Out) *indeed is a semidefinite program; cf. Proposition* 4.9.1.

Proof. Let us pass in the description (4.9.110) from the parameters D, c to the parameters $Z = D$, $z = Dc$, thus coming to the representation

$$E = \{x \mid x^T Z x - 2x^T z + z^T Z^{-1} z \leq 1\}. \qquad (!)$$

The ellipsoid of the latter type contains the points $x_1, \ldots, x_m$ if and only if

$$x_i^T Z x_i - 2x_i^T z + z^T Z^{-1} z \leq 1, \ i = 1, \ldots, m,$$

or, which is the same, if and only if there exists $s \geq z^T Z^{-1} z$ such that

$$x_i^T Z x_i - 2x_i^T z + s \leq 1, \ i = 1, \ldots, m.$$

Recalling the lemma on the Schur complement, we see that the constraints (Out)(b)–(d) say exactly that the ellipsoid (!) contains the points $x_1, \ldots, x_m$. Since the volume of such an ellipsoid is $(\mathrm{Det}Z)^{-1/2}$, (Out) is the problem of maximizing a negative power of the volume of an ellipsoid containing the finite set $\{x_1, \ldots, x_m\}$, i.e., the problem of finding the smallest-volume ellipsoid containing this finite set. It remains to note that an ellipsoid is convex, so that it is exactly the same—to say that it contains a finite set $\{x_1, \ldots, x_m\}$ and to say that it contains the convex hull of this finite set. □

We see that if S is a polytope given as a convex hull of a finite set, then the problem of the best outer ellipsoidal approximation of S is an explicit semidefinite program and as such can be efficiently solved. In contrast to this, if S is a polytope given by a list of inequality constraints, then the problem of the best outer ellipsoidal approximation of S is computationally intractable—in this case, it is difficult just to check whether a given candidate ellipsoid contains S.

4.9.1 Ellipsoidal approximations of unions and intersections of ellipsoids

Speaking informally, Proposition 4.9.1 deals with inner ellipsoidal approximation of the intersection of degenerate ellipsoids, namely, half-spaces. (A half-space is just a very large Euclidean ball!) Similarly, Proposition 4.9.2 deals with the outer ellipsoidal approximation of the union of degenerate ellipsoids, namely, points. (A point is just a ball of zero radius!) We are about to demonstrate that when passing from degenerate ellipsoids to the normal ones, we still have a possibility to reduce the corresponding approximation problems to explicit semidefinite programs. The key observation here is as follows.

Proposition 4.9.3.[39] *An ellipsoid*

$$E = E(Z, z) \equiv \{x = Zu + z \mid u^T u \le 1\} \quad [Z \in \mathbf{M}^{n,q}]$$

is contained in the full-dimensional ellipsoid

$$W = W(Y, y) \equiv \{x \mid (x - y)^T Y^T Y (x - y) \le 1\} \quad [Y \in \mathbf{M}^{n,n}, \mathrm{Det} Y \neq 0]$$

if and only if there exists λ such that

$$\begin{pmatrix} I_n, & Y(z-y), & YZ \\ (z-y)^T Y^T, & 1-\lambda, & \\ Z^T Y^T, & & \lambda I_q \end{pmatrix} \succeq 0 \tag{4.9.111}$$

as well as if and only if there exists λ such that

$$\begin{pmatrix} Y^{-1}(Y^{-1})^T, & z-y, & Z \\ (z-y)^T, & 1-\lambda, & \\ Z^T, & & \lambda I_q \end{pmatrix} \succeq 0. \tag{4.9.112}$$

Proof. We clearly have

$$\begin{array}{c} E \subset W \\ \Updownarrow \\ u^T u \le 1 \Rightarrow (Zu + z - y)^T Y^T Y (Zu + z - y) \le 1 \\ \Updownarrow \\ u^T u \le t^2 \Rightarrow (Zu + t(z-y))^T Y^T Y (Zu + t(z-y)) \le t^2 \\ \Updownarrow \mathcal{S}\text{-lemma} \\ \exists \lambda \ge 0 : [t^2 - (Zu + t(z-y))^T Y^T Y (Zu + t(z-y))] - \lambda[t^2 - u^T u] \ge 0 \quad \forall (u, t) \\ \Updownarrow \\ \exists \lambda \ge 0 : \begin{pmatrix} 1 - \lambda - (z-y)^T Y^T Y (z-y), & -(z-y)^T Y^T YZ \\ -Z^T Y^T Y (z-y), & \lambda I_q - Z^T Y^T YZ \end{pmatrix} \succeq 0 \\ \Updownarrow \\ \exists \lambda \ge 0 : \begin{pmatrix} 1-\lambda & \\ & \lambda I_q \end{pmatrix} - \begin{pmatrix} (z-y)^T Y^T \\ Z^T Y^T \end{pmatrix} \begin{pmatrix} Y(z-y) & YZ \end{pmatrix} \succeq 0. \end{array}$$

[39] S. Boyd et al., *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, 1994.

Now note that in view of the lemma on the Schur complement the matrix

$$\begin{pmatrix} 1-\lambda & \\ & \lambda I_q \end{pmatrix} - \begin{pmatrix} (z-y)^T Y^T \\ Z^T Y^T \end{pmatrix} (\, Y(z-y), \quad YZ\,)$$

is positive semidefinite if and only if the matrix in (4.9.111) is so. Thus, $E \subset W$ if and only if there exists a nonnegative λ such that the matrix in (4.9.111), let it be called $P(\lambda)$, is positive semidefinite. Since the latter matrix can be positive semidefinite only when $\lambda \geq 0$, we have proved the first statement of the proposition. To prove the second statement, note that the matrix in (4.9.112), let it be called $Q(\lambda)$, is closely related to $P(\lambda)$:

$$Q(\lambda) = SP(\lambda)S^T, \quad S = \begin{pmatrix} Y^{-1}, & & \\ & 1, & \\ & & I_q \end{pmatrix} \succ 0,$$

so that $Q(\lambda)$ is positive semidefinite if and only if $P(\lambda)$ is so. □

Here are some consequences of Proposition 4.9.3.

Inner ellipsoidal approximation of the intersection of full-dimensional ellipsoids. Let

$$W_i = \{x \mid (x-c_i)^T B_i^2 (x-c_i) \leq 1\} \quad [B_i \in \mathbf{S}^n_{++}],$$

$i = 1, \ldots, m$, be given full-dimensional ellipsoids in $\mathbf{R}^n$; assume that the intersection W of these ellipsoids possesses a nonempty interior. Then the problem of the best inner ellipsoidal approximation of W is the explicit semidefinite program

$$\begin{array}{ll} \text{maximize} & t \\ \text{s.t.} & \\ & \begin{array}{rcl} t & \leq & (\mathrm{Det}Z)^{1/n}, \\ \begin{pmatrix} I_n, & B_i(z-c_i), & B_i Z \\ (z-c_i)^T B_i, & 1-\lambda_i, & \\ Z B_i, & & \lambda_i I_n \end{pmatrix} & \succeq & 0, \ i = 1, \ldots, m, \\ Z & \succeq & 0 \end{array} \end{array} \qquad \text{(InEll)}$$

with the design variables $Z \in \mathbf{S}^n, z \in \mathbf{R}^n, \lambda_i, t \in \mathbf{R}$. The largest ellipsoid contained in $W = \bigcap_{i=1}^m W_i$ is given by an optimal solution $Z_*, z_*, t_*, \{\lambda_i^*\}$) of (InEll) via the relation

$$E = \{x = Z_* u + z_* \mid u^T u \leq 1\}.$$

Indeed, by Proposition 4.9.3 the LMIs

$$\begin{pmatrix} I_n, & B_i(z-c_i), & B_i Z \\ (z-c_i)^T B_i, & 1-\lambda_i, & \\ Z B_i, & & \lambda_i I_n \end{pmatrix} \succeq 0, \ i = 1, \ldots, m,$$

express the fact that the ellipsoid $\{x = Zu + z \mid u^T u \leq 1\}$ with $Z \succeq 0$ is contained in every one of the ellipsoids W_i, i.e., is contained in the intersection W of these ellipsoids. Consequently, (InEll) is exactly the problem of maximizing (a positive power of) the volume of an ellipsoid over the ellipsoids contained in W.

Outer ellipsoidal approximation of the union of ellipsoids. Let

$$W_i = \{x = A_i u + c_i \mid u^T u \le 1\} \quad [A_i \in \mathbf{M}^{n,k_i}],$$

$i = 1, \ldots, m$, be given ellipsoids in $\mathbf{R}^n$; assume that the convex hull W of the union of these ellipsoids possesses a nonempty interior. Then the problem of the best outer ellipsoidal approximation of W is the explicit semidefinite program

$$\begin{array}{ll} \text{maximize} & t \\ \text{s.t.} & \\ & \begin{array}{rcl} t & \le & (\mathrm{Det}Y)^{1/n}, \\ \begin{pmatrix} I_n, & Yc_i - z, & YA_i \\ (Yc_i - z)^T, & 1 - \lambda_i, & \\ A_i^T Y, & & \lambda_i I_{k_i} \end{pmatrix} & \succeq & 0,\ i = 1, \ldots, m, \\ Y & \succeq & 0 \end{array} \end{array} \qquad \text{(OutEll)}$$

with the design variables $Y \in \mathbf{S}^n, z \in \mathbf{R}^n, \lambda_i, t \in \mathbf{R}$. The smallest ellipsoid containing $W = \mathrm{Conv}(\bigcup_{i=1}^m W_i)$ is given by an optimal solution $(Y_*, z_*, t_*, \{\lambda_i^*\})$ of (OutEll) via the relation

$$E = \{x \mid (x - y_*)Y_*^2(x - y_*) \le 1\}, \quad y_* = Y_*^{-1} z_*.$$

Indeed, by Proposition 4.9.3 for $Y \succ 0$ the LMIs

$$\begin{pmatrix} I_n, & Yc_i - z, & YA_i \\ (Yc_i - z)^T, & 1 - \lambda_i, & \\ A_i^T Y, & & \lambda_i I_{k_i} \end{pmatrix} \succeq 0,\ i = 1, \ldots, m,$$

express the fact that the ellipsoid $E = \{x \mid (x - Y^{-1}z)^T Y^2 (x - Y^{-1}y) \le 1\}$ contains every one of the ellipsoids W_i, i.e., contains the convex hull W of the union of these ellipsoids. The volume of the ellipsoid E is $(\mathrm{Det}Y)^{-1}$. Consequently, (OutEll) is exactly the problem of maximizing a negative power (i.e., of minimizing a positive power) of the volume of an ellipsoid over the ellipsoids containing W.

4.9.2 Approximating sums of ellipsoids

Let us come back to our motivating example, where we wanted to build ellipsoidal approximation of the set X_T of all states $x(T)$ where a given discrete time invariant linear system

$$\begin{array}{rcl} x(t+1) & = & Ax(t) + Bu(t),\ t = 0, \ldots, T-1, \\ x(0) & = & 0 \end{array}$$

can be driven in time T by a control $u(\cdot)$ satisfying the norm bound

$$\|u(t)\|_2 \le 1,\ t = 0, \ldots, T-1.$$

How could we build such an approximation recursively? Let X_t be the set of all states where the system can be driven in time $t \le T$, and assume that we have already built inner and outer ellipsoidal approximations E_{in}^t and E_{out}^t of the set X_t:

$$E_{\mathrm{in}}^t \subset X_t \subset E_{\mathrm{out}}^t.$$

Let also

$$E = \{x = Bu \mid u^T u \le 1\}.$$

Then the set

$$F_{\text{in}}^{t+1} = AE_{\text{in}}^t + E \equiv \{x = Ay + z \mid y \in E_{\text{in}}^t, z \in E\}$$

clearly is contained in X_{t+1}, so that a natural recurrent way to define an inner ellipsoidal approximation of X_{t+1} is to take as E_{in}^{t+1} the largest-volume ellipsoid contained in F_{in}^{t+1}. Similarly, the set

$$F_{\text{out}}^{t+1} = AE_{\text{out}}^t + E \equiv \{x = Ay + z \mid y \in E_{\text{out}}^t, z \in E\}$$

clearly covers X_{t+1}, and the natural recurrent way to define an outer ellipsoidal approximation of X_{t+1} is to take as E_{out}^{t+1} the smallest-volume ellipsoid containing F_{out}^{t+1}.

Note that the sets F_{in}^{t+1} and F_{out}^{t+1} are of the same structure: each of them is the arithmetic sum $\{x = v + w \mid v \in V, w \in W\}$ of two ellipsoids V and W. Thus, we come to the problem as follows: Given two ellipsoids W, V, find the best inner and outer ellipsoidal approximations of their arithmetic sum $W + V$. In fact, it makes sense to consider a more general problem:

> *Given m ellipsoids $W_1, \ldots, W_m$ in $\mathbf{R}^n$, find the best inner and outer ellipsoidal approximations of the arithmetic sum*
>
> $$W = \{x = w_1 + w_1 + \cdots + w_m \mid w_i \in W_i,\ i = 1, \ldots, m\}$$
>
> *of the ellipsoids $W_1, \ldots, W_m$.*

We have posed two different problems: the one of inner approximation of W (let this problem be called (I)) and the other one (let it be called (O)) of outer approximation. It seems that in general both these problems are difficult (at least when m is not once for ever fixed). There exist, however, computationally tractable approximations of both (I) and (O) we are about to consider.

In the considerations to follow, we assume, for the sake of simplicity, that the ellipsoids $W_1, \ldots, W_m$ are full-dimensional (which is not a severe restriction—a flat ellipsoid can be easily approximated by a nearly flat full-dimensional ellipsoid). Besides this, we may assume w.l.o.g. that all our ellipsoids W_i are centered at the origin. Indeed, we have $W_i = c_i + V_i$, where c_i is the center of W_i and $V_i = W_i - c_i$ is centered at the origin. Consequently,

$$W_1 + \cdots + W_m = (c_1 + \cdots + c_m) + (V_1 + \cdots + V_m),$$

so that the problems (I) and (O) for the ellipsoids $W_1, \ldots, W_m$ can be straightforwardly reduced to similar problems for the centered at the origin ellipsoids $V_1, \ldots, V_m$.

Problem (O). Let the ellipsoids $W_1, \ldots, W_m$ be represented as

$$W_i = \{x \in \mathbf{R}^n \mid x^T B_i x \le 1\} \qquad [B_i \succ 0].$$

Our strategy to approximate (O) is very natural: we intend to build a parametric family of ellipsoids in such a way that, first, every ellipsoid from the family contains the arithmetic sum $W_1 + \cdots + W_m$ of given ellipsoids, and, second, the problem of finding the smallest volume ellipsoid within the family is a computationally tractable problem (specifically, is an explicit semidefinite program).[40] The seemingly simplest way to build the desired family was proposed in Boyd et al. (1994) and is based on the idea of semidefinite relaxation. Let us start with the observation that an ellipsoid

$$W[Z] = \{x \mid x^T Z x \le 1\} \qquad [Z \succ 0]$$

contains $W_1 + \cdots + W_m$ if and only if the following implication holds:

$$\left\{\{x^i \in \mathbf{R}^n\}_{i=1}^m, [x^i]^T B_i x^i \le 1,\ i = 1, \ldots, m\right\} \Rightarrow (x^1 + \cdots + x^m)^T Z (x^1 + \cdots + x^m) \le 1. \qquad (*)$$

Now let B^i be $(nm) \times (nm)$ block-diagonal matrix with m diagonal blocks of the size $n \times n$ each, such that all diagonal blocks, except the ith one, are zero, and the ith block is the $n \times n$ matrix B_i. Let also $M[Z]$ denote $(mn) \times (mn)$ block matrix with m^2 blocks of the size $n \times n$ each, every block being the matrix Z. This is how B^i and $M[Z]$ look in the case of $m = 2$:

$$B^1 = \begin{bmatrix} B_1 & \\ & \end{bmatrix}, \quad B^2 = \begin{bmatrix} & \\ & B_2 \end{bmatrix}, \quad M[Z] = \begin{bmatrix} Z & Z \\ Z & Z \end{bmatrix}.$$

Validity of implication $(*)$ clearly is equivalent to the following fact:

($*$.1) *For every (mn)-dimensional vector x such that*

$$x^T B^i x \equiv \mathrm{Tr}(B^i \underbrace{x x^T}_{X[x]}) \le 1,\ i = 1, \ldots, m,$$

one has

$$x^T M[Z] x \equiv \mathrm{Tr}(M[Z] X[x]) \le 1.$$

Now we can use the standard trick: the rank one matrix $X[x]$ is positive semidefinite, so that we for sure enforce the validity of the above fact when enforcing the following stronger fact:

($*$.2) *For every $(mn) \times (mn)$ symmetric positive semidefinite matrix X such that*

$$\mathrm{Tr}(B^i X) \le 1,\ i = 1, \ldots, m,$$

one has

$$\mathrm{Tr}(M[Z] X) \le 1.$$

[40] Note that we, in general, do not pretend that our parametric family includes all ellipsoids containing $W_1 + \cdots + W_m$, so that the ellipsoid we end with should be treated as nothing more than a computable surrogate of the smallest-volume ellipsoid containing the sum of W_i's.

We have arrived at the following result.

> **(D)** *Let a positive definite $n \times n$ matrix Z be such that the optimal value in the semidefinite program*
>
> $$\max_X \left\{ \mathrm{Tr}(M[Z]X) \middle| \mathrm{Tr}(B^i X) \le 1,\ i = 1, \dots, m,\ X \succeq 0 \right\} \qquad \text{(SDP)}$$
>
> *is ≤ 1. Then the ellipsoid*
>
> $$W[Z] = \{x \mid x^T Z x \le 1\}$$
>
> *contains the arithmetic sum $W_1 + \cdots + W_m$ of the ellipsoids $W_i = \{x \,|\, x^T B_i x \le 1\}$.*

We are basically done. The set of those symmetric matrices Z for which the optimal value in (SDP) is ≤ 1 is SDr; indeed, the problem is clearly strictly feasible, and Z affects, in a linear fashion, the objective of the problem only. On the other hand, the optimal value in a strictly feasible semidefinite maximization program is an SDr function of the objective (semidefinite version of Proposition 3.4.3). Consequently, the set of those Z for which the optimal value in (SDP) is ≤ 1 is SDr (as the inverse image, under affine mapping, of the level set of an SDr function). Thus, the parameter Z of those ellipsoids $W[Z]$ that satisfy the premise in **(D)** and thus contain $W_1 + \cdots + W_m$ varies in an SDr set $\mathcal{Z}$. Consequently, the problem of finding the smallest-volume ellipsoid in the family $\{W[Z]\}_{Z \in \mathcal{Z}}$ is equivalent to the problem of maximizing a positive power of $\mathrm{Det}(Z)$ over the SDr set $\mathcal{Z}$, i.e., is equivalent to a semidefinite program.

It remains to build the aforementioned semidefinite program. By the conic duality theorem the optimal value in the (clearly strictly feasible) maximization program (SDP) is ≤ 1 if and only if the dual problem

$$\min_\lambda \left\{ \sum_{i=1}^m \lambda_i \middle| \sum_i \lambda_i B^i \succeq M[Z],\ \lambda_i \ge 0,\ i = 1, \dots, m \right\}$$

admits a feasible solution with the value of the objective ≤ 1 or, which is clearly the same (why?), admits a feasible solution with the value of the objective equal 1. In other words, whenever $Z \succeq 0$ is such that $M[Z]$ is $\preceq$ a convex combination of the matrices B^i, the set

$$W[Z] = \{x \mid x^T Z x \le 1\}$$

(which is an ellipsoid when $Z \succ 0$) contains the set $W_1 + \cdots + W_m$. We have arrived at the following result (see footnote 23, section 3.7.4).

Proposition 4.9.4. *Given m centered at the origin full-dimensional ellipsoids*

$$W_i = \{x \in \mathbf{R}^n \mid x^T B_i x \le 1\} \quad [B_i \succ 0],$$

$i = 1, \ldots, m$, in $\mathbf{R}^n$, let us associate with these ellipsoids the semidefinite program

$$\max_{t,Z,\lambda} \left\{ t \;\middle|\; \begin{array}{c} t \le \mathrm{Det}^{1/n}(Z) \\ \sum_{i=1}^m \lambda_i B^i \succeq M[Z] \\ \lambda_i \ge 0,\ i = 1, \ldots, m \\ Z \succeq 0 \\ \sum_{i=1}^m \lambda_i = 1 \end{array} \right\}, \tag{$\tilde{\mathrm{O}}$}$$

where B^i is the $(mn) \times (mn)$ block-diagonal matrix with blocks of the size $n \times n$ and the only nonzero diagonal block (the ith one) equal to B_i, and $M[Z]$ is the $(mn) \times (mn)$ matrix partitioned into m^2 blocks, every one of them being Z. Every feasible solution $(Z, \ldots)$ to this program with positive value of the objective produces ellipsoid

$$W[Z] = \{x \mid x^T Z x \le 1\},$$

which contains $W_1 + \cdots + W_m$, and the volume of this ellipsoid is at most $t^{-n/2}$. The smallest-volume ellipsoid that can be obtained in this way is given by (any) optimal solution of ($\tilde{\mathrm{O}}$).

How conservative is ($\tilde{\mathrm{O}}$)? The ellipsoid $W[Z^*]$ given by the optimal solution of ($\tilde{\mathrm{O}}$) contains the arithmetic sum W of the ellipsoids W_i but is not necessarily the smallest-volume ellipsoid containing W. All we know is that this ellipsoid is the smallest-volume one in a certain subfamily of the family of all ellipsoids containing W. In nature there exists the true smallest-volume ellipsoid $W[Z^{**}] = \{x \mid x^T Z^{**} x \le 1\}$, $Z^{**} \succ 0$, containing W. It is natural to ask how large the ratio

$$\vartheta = \frac{\mathrm{Vol}(W[Z^*])}{\mathrm{Vol}(W[Z^{**}])}$$

could be. The answer is as follows.

Proposition 4.9.5. *One has $\vartheta \le \left(\frac{\pi}{2}\right)^{n/2}$.*

Note that the bound stated by Proposition 4.9.5 is not as bad as it looks. The natural way to compare the sizes of two n-dimensional bodies E', E'' is to look at the ratio of their average linear sizes $\left(\frac{\mathrm{Vol}(E')}{\mathrm{Vol}(E'')}\right)^{1/n}$. (It is natural to assume that by shrinking a body by a certain factor, say, 2, we reduce the size of the body exactly by this factor, and not by 2^n.) With this approach, the level of nonoptimality of $W[Z^*]$ is no more than $\sqrt{\pi/2} = 1.253\ldots$, i.e., is within a 25% margin.

Proof of Proposition 4.9.5. Since Z^{**} contains W, the implication (*.1) holds true, i.e., one has

$$\max_{x \in \mathbf{R}^{mn}} \{x^T M[Z^{**}] x \mid x^T B^i x \le 1,\ i = 1, \ldots, m\} \le 1.$$

Since the matrices B^i, $i = 1, \ldots, m$, commute and $M[Z^{**}] \succeq 0$, we can apply Proposition 4.10.5 (see section 4.10.5) to conclude that there exist nonnegative μ_i, $i = 1, \ldots, m$, such that

$$M[Z^{**}] \preceq \sum_{i=1}^{m} \mu_i B^i, \quad \sum_i \mu_i \leq \frac{\pi}{2}.$$

It follows that setting

$$\lambda_i = \left(\sum_j \mu_j\right)^{-1} \mu_i, \; Z = \left(\sum_j \mu_j\right)^{-1} Z^{**}, \; t = \mathrm{Det}^{1/n}(Z),$$

we get a feasible solution of ($\tilde{\mathrm{O}}$). Recalling the origin of Z^*, we come to

$$\mathrm{Vol}(W[Z^*]) \leq \mathrm{Vol}(W[Z]) = \left(\sum_j \mu_j\right)^{n/2} \mathrm{Vol}(W[Z^{**}]) \leq (\pi/2)^{n/2} \mathrm{Vol}(W[Z^{**}]),$$

as claimed. □

Problem (O), the case of coaxial ellipsoids. Consider the coaxial case—the one in which there exist coordinates (not necessarily orthogonal) such that all m quadratic forms defining the ellipsoids W_i are diagonal in these coordinates, or, which is the same, there exists a nonsingular matrix C such that all the matrices $C^T B_i C$, $i = 1, \ldots, m$, are diagonal. Note that the case of $m = 2$ always is coaxial—linear algebra says that every two homogeneous quadratic forms, at least one of the forms being positive outside of the origin, become diagonal in a properly chosen coordinates.

We are about to prove that

(E) *In the coaxial case,* ($\tilde{\mathrm{O}}$) *yields the smallest-volume ellipsoid containing* $W_1 + \cdots + W_m$.

Consider the coaxial case. Since we are interested in volume-related issues, and the ratio of volumes remains unchanged under affine transformations, we can assume w.l.o.g. that the matrices B_i defining the ellipsoids $W_i = \{x \mid x^T B_i x \leq 1\}$ are positive definite and diagonal; let b_ℓ^i be the ℓth diagonal entry of B_i, $\ell = 1, \ldots, n$. By the Fritz John theorem, in nature there exists a unique smallest-volume ellipsoid W_* which contains $W_1 + \cdots + W_m$. From uniqueness combined with the fact that the sum of our ellipsoids is symmetric with respect to the origin, it follows that this optimal ellipsoid W_* is centered at the origin:

$$W_* = \{x \mid x^T Z_* x \leq 1\}$$

with certain positive definite matrix Z_*.

Our next observation is that the matrix Z_* is diagonal. Indeed, let E be a diagonal matrix with diagonal entries ± 1. Since all B_i's are diagonal, the sum $W_1 + \cdots + W_m$ remains invariant under multiplication by E:

$$x \in W_1 + \cdots + W_m \Leftrightarrow Ex \in W_1 + \cdots + W_m.$$

It follows that the ellipsoid $E(W_*) = \{x \mid x^T(E^T Z_* E)x \le 1\}$ covers $W_1 + \cdots + W_m$ along with W_* and of course has the same volume as W_*. From the uniqueness of the optimal ellipsoid it follows that $E(W_*) = W_*$, whence $E^T Z_* E = Z_*$ (why?). Since the concluding relation should be valid for all diagonal matrices E with diagonal entries ± 1, Z_* must be diagonal.

Now assume that the set

$$W(z) = \{x \mid x^T \mathrm{Diag}(z) x \le 1\} \tag{4.9.113}$$

given by a nonnegative vector z contains $W_1 + \cdots + W_m$. Then the following implication holds true:

$$\forall \{x_\ell^i\}_{\substack{i=1,\dots,m\\ \ell=1,\dots,n}} : \quad \sum_{\ell=1}^n b_\ell^i (x_\ell^i)^2 \le 1,\ i = 1,\dots,m \ \Rightarrow\ \sum_{\ell=1}^n z_\ell (x_\ell^1 + x_\ell^2 + \ldots + x_\ell^m)^2 \le 1. \tag{4.9.114}$$

Denoting $y_\ell^i = (x_\ell^i)^2$ and taking into account that $z_\ell \ge 0$, we see that the validity of (4.9.114) implies the validity of the implication

$$\forall \{y_\ell^i \ge 0\}_{\substack{i=1,\dots,m\\ \ell=1,\dots,n}} : \sum_{\ell=1}^n b_\ell^i y_\ell^i \le 1,\ i = 1,\dots,m \ \Rightarrow\ \sum_{\ell=1}^n z_\ell \left(\sum_{i=1}^m y_\ell^i + 2 \sum_{1 \le i < j \le m} \sqrt{y_\ell^i y_\ell^j} \right) \le 1. \tag{4.9.115}$$

Now let Y be an $(mn) \times (mn)$ symmetric matrix satisfying the relations

$$Y \succeq 0;\ \mathrm{Tr}(Y B^i) \le 1,\ i = 1,\dots,m. \tag{4.9.116}$$

Let us partition Y into m^2 square blocks, and let Y_ℓ^{ij} be the ℓth diagonal entry of the ijth block of Y. For all i, j with $1 \le i < j \le m$ and all ℓ, $1 \le \ell \le n$, the 2×2 matrix $\begin{pmatrix} Y_\ell^{ii} & Y_\ell^{ij} \\ Y_\ell^{ij} & Y_\ell^{jj} \end{pmatrix}$ is a principal submatrix of Y and therefore is positive semidefinite along with Y, whence

$$Y_\ell^{ij} \le \sqrt{Y_\ell^{ii} Y_\ell^{jj}}. \tag{4.9.117}$$

In view of (4.9.116), the numbers $y_\ell^i \equiv Y_\ell^{ii}$ satisfy the premise in the implication (4.9.115), so that

$$\begin{aligned}
1 &\ge \sum_{\ell=1}^n z_\ell \left[\sum_{i=1}^m Y_\ell^{ii} + 2 \sum_{1 \le i < j \le m} \sqrt{Y_\ell^{ii} Y_\ell^{jj}} \right] && \text{[by (4.9.115)]} \\
&\ge \sum_{\ell=1}^n z_\ell \left[\sum_{i=1}^m Y_\ell^{ii} + 2 \sum_{1 \le i < j \le m} Y_\ell^{ij} \right] && \text{[since } z \ge 0 \text{ and by (4.9.117)]} \\
&= \mathrm{Tr}(Y M[\mathrm{Diag}(z)]).
\end{aligned}$$

Thus, (4.9.116) implies the inequality $\mathrm{Tr}(Y M[\mathrm{Diag}(z)]) \le 1$, i.e., the implication

$$Y \succeq 0,\ \mathrm{Tr}(Y B^i) \le 1,\ i = 1,\dots,m \ \Rightarrow\ \mathrm{Tr}(Y M[\mathrm{Diag}(z)]) \le 1$$

holds true. Since the premise in this implication is strictly feasible, the validity of the implication, by semidefinite duality, implies the existence of nonnegative λ_i, $\sum_i \lambda_i \leq 1$, such that

$$M[\mathrm{Diag}(z)] \preceq \sum_i \lambda_i B^i.$$

Combining our observations, we come to the conclusion as follows:

> *In the case of diagonal matrices B_i, if the set* (4.9.113)*, given by a nonnegative vector z, contains $W_1 + \cdots + W_m$, then the matrix* $\mathrm{Diag}(z)$ *can be extended to a feasible solution of the problem* ($\tilde{\mathrm{O}}$)*. Consequently, in the case in question the approximation scheme given by* ($\tilde{\mathrm{O}}$) *yields the minimum volume ellipsoid containing $W_1 + \cdots + W_m$* (since the latter ellipsoid, as we have seen, is of the form (4.9.113) with $z \geq 0$).

It remains to note that the approximation scheme associated with ($\tilde{\mathrm{O}}$) is affine invariant, so that the above conclusion remains valid when we replace in its premise "the case of diagonal matrices B_i" with "the coaxial case."

Remark 4.9.1. In fact, **(E)** is an immediate consequence of the following fact (which, essentially, is proved in the above reasoning).

Let $A_1, \ldots, A_m$, B be symmetric matrices such that the off-diagonal entries of all A_i's are nonpositive, and the off-diagonal entries of B are nonnegative. Assume also that the system of inequalities

$$x^T A_i x \leq a_i, \ i = 1, \ldots, m, \tag{S}$$

is strictly feasible. Then the inequality

$$x^T B x \leq b$$

is a consequence of the system (S) if and only if it is a linear consequence of (S), i.e., if and only if there exist nonnegative weights λ_i such that

$$B \preceq \sum_i \lambda_i A_i, \quad \sum_i \lambda_i a_i \leq b.$$

In other words, in the case in question the optimization program

$$\max_x \left\{ x^T B x \mid x^T A_i x \leq a_i, \ i = 1, \ldots, m \right\}$$

and its standard semidefinite relaxation

$$\max_X \{ \mathrm{Tr}(BX) \mid X \succeq 0, \ \mathrm{Tr}(A_i X) \leq a_i, \ i = 1, \ldots, m \}$$

share the same optimal value.

Problem (I). Let us represent the given centered at the origin ellipsoids W_i as

$$W_i = \{x \mid x = A_i u \mid u^T u \le 1\} \qquad [\mathrm{Det}(A_i) \neq 0].$$

We start from the following observation:

> **(F)** *An ellipsoid $E[Z] = \{x = Zu \mid u^T u \le 1\}$ ($[\mathrm{Det}(Z) \neq 0]$) is contained in the sum $W_1 + \cdots + W_m$ of the ellipsoids W_i if and only if one has*
>
> $$\forall x: \quad \|Z^T x\|_2 \le \sum_{i=1}^m \|A_i^T x\|_2. \tag{4.9.118}$$

Indeed, assume, first, that there exists a vector x_* such that the inequality in (4.9.118) is violated at $x = x_*$, and let us prove that in this case $W[Z]$ is not contained in the set $W = W_1 + \cdots + W_m$. We have

$$\max_{x \in W_i} x_*^T x = \max\left[x_*^T A_i u \mid u^T u \le 1\right] = \|A_i^T x_*\|_2, \ i = 1, \ldots, m,$$

and similarly

$$\max_{x \in E[Z]} x_*^T x = \|Z^T x_*\|_2,$$

whence

$$\begin{aligned} \max_{x \in W} x_*^T x &= \max_{x^i \in W_i} x_*^T (x^1 + \cdots + x^m) = \sum_{i=1}^m \max_{x^i \in W_i} x_*^T x^i \\ &= \sum_{i=1}^m \|A_i^T x_*\|_2 < \|Z^T x_*\|_2 = \max_{x \in E[Z]} x_*^T x, \end{aligned}$$

and we see that $E[Z]$ cannot be contained in W. Conversely, assume that $E[Z]$ is not contained in W, and let $y \in E[Z] \backslash W$. Since W is a convex compact set and $y \notin W$, there exists a vector x_* such that $x_*^T y > \max_{x \in W} x_*^T x$, whence, due to the previous computation,

$$\|Z^T x_*\|_2 = \max_{x \in E[Z]} x_*^T x \ge x_*^T y > \max_{x \in W} x_*^T x = \sum_{i=1}^m \|A_i^T x_*\|_2,$$

and we have found a point $x = x_*$ at which the inequality in (4.9.118) is violated. Thus, $E[Z]$ is not contained in W if and only if (4.9.118) is not true, which is exactly what should be proved.

A natural way to generate ellipsoids satisfying (4.9.118) is to note that whenever X_i are $n \times n$ matrices of spectral norms

$$|X_i| \equiv \sqrt{\lambda_{\max}(X_i^T X_i)} = \sqrt{\lambda_{\max}(X_i X_i^T)} = \max_x \{\|X_i x\|_2 \mid \|x\|_2 \le 1\}$$

not exceeding 1, the matrix

$$Z = Z(X_1, \ldots, X_m) = A_1 X_1 + A_2 X_2 + \cdots + A_m X_m$$

satisfies (4.9.118):

$$\|Z^T x\|_2 = \|[X_1^T A_1^T + \cdots + X_m^T A_m^T]x\|_2 \le \sum_{i=1}^m \|X_i^T A_i^T x\|_2 \le \sum_{i=1}^m |X_i^T| \|A_i^T x\|_2 \le \sum_{i=1}^m \|A_i^T x\|_2.$$

Thus, every collection of square matrices X_i with spectral norms not exceeding 1 produces an ellipsoid satisfying (4.9.118) and thus contained in W, and we could use the largest volume ellipsoid of this form (i.e., the one corresponding to the largest $|\mathrm{Det}(A_1 X_1 + \cdots + A_m X_m)|$) as a surrogate of the largest volume ellipsoid contained in W. Recall that we know how to express a bound on the spectral norm of a matrix via LMI:

$$|X| \le t \Leftrightarrow \begin{pmatrix} tI_n & -X^T \\ -X & tI_n \end{pmatrix} \succeq 0 \quad [X \in \mathbf{M}^{n,n}]$$

(item 16 of section 4.2). The difficulty, however, is that the matrix $\sum_{i=1}^m A_i X_i$ specifying the ellipsoid $E(X_1, \ldots, X_m)$, although being linear in the design variables X_i, is not necessarily symmetric positive semidefinite, and we do not know how to maximize the determinant over general-type square matrices. We may, however, use the following fact from linear algebra.

Lemma 4.9.1. *Let $Y = S + C$ be a square matrix represented as the sum of a symmetric matrix S and a skew-symmetric (i.e., $C^T = -C$) matrix C. Assume that S is positive definite. Then*

$$|\mathrm{Det}(Y)| \ge \mathrm{Det}(S).$$

Proof. We have $Y = S + C = S^{1/2}(I + \Sigma)S^{1/2}$, where $\Sigma = S^{-1/2} C S^{-1/2}$ is skew-symmetric along with C. We have $|\mathrm{Det}(Y)| = \mathrm{Det}(S)|\mathrm{Det}(I + \Sigma)|$; it remains to note that all eigenvalues of the skew-symmetric matrix Σ are purely imaginary, so that the eigenvalues of $I + \Sigma$ are ≥ 1 in absolute value, whence $|\mathrm{Det}(I + \Sigma)| \ge 1$. □

In view of the lemma, it makes sense to impose on $X_1, \ldots, X_m$, in addition to the requirement that their spectral norms be ≤ 1, also the requirement that the symmetric part

$$S(X_1, \ldots, X_m) = \frac{1}{2}\left[\sum_{i=1}^m A_i X_i + \sum_{i=1}^m X_i^T A_i\right]$$

of the matrix $\sum_i A_i X_i$ be positive semidefinite, and to maximize under these constraints the quantity $\mathrm{Det}(S(X_1, \ldots, X_m))$—a lower bound on the volume of the ellipsoid $E[Z(X_1, \ldots, X_m)]$. With this approach, we come to the following result.

Proposition 4.9.6. *Let $W_i = \{x = A_i u \mid u^T u \le 1\}$, $A_i \succ 0$, $i = 1, \ldots, m$. Consider the semidefinite program*

$$
\begin{array}{lll}
\textit{maximize} & t & \\
\textit{s.t.} & & \\
\text{(a)} & t \le \left(\mathrm{Det}\left(\frac{1}{2}\sum_{i=1}^m [X_i^T A_i + A_i X_i]\right)\right)^{1/n}, & \\
\text{(b)} & \sum_{i=1}^m [X_i^T A_i + A_i X_i] \succeq 0, & \\
\text{(c)} & \begin{pmatrix} I_n & -X_i^T \\ -X_i & I_n \end{pmatrix} \succeq 0,\ i = 1, \ldots, m, &
\end{array}
\qquad (\tilde{I})
$$

with design variables $X_1, \ldots, X_m \in \mathbf{M}^{n,n}, t \in \mathbf{R}$. *Every feasible solution* $(\{X_i\}, t)$ *to this problem produces the ellipsoid*

$$E(X_1, \ldots, X_m) = \left\{x = \left(\sum_{i=1}^m A_i X_i\right) u \mid u^T u \le 1\right\}$$

contained in the arithmetic sum $W_1 + \cdots + W_m$ *of the original ellipsoids, and the volume of this ellipsoid is at least* t^n. *The largest-volume ellipsoid which can be obtained in this way is associated with (any) optimal solution to* (Ĩ).

In fact, problem (I) is equivalent to the problem we started with,

$$\left|\mathrm{Det}\left(\sum_{i=1}^m A_i X_i\right)\right| \to \max \mid |X_i| \le 1,\ i = 1, \ldots, m, \qquad (4.9.119)$$

since the latter problem always has an optimal solution $\{X_i^*\}$ with positive semidefinite symmetric matrix $G_* = \sum_{i=1}^m A_i X_i^*$. Indeed, let $\{X_i^+\}$ be an optimal solution of the problem. The matrix $G_+ = \sum_{i=1}^m A_i X_i^+$, as every $n \times n$ square matrix, admits a representation $G_+ = G_* U$, where G_+ is a positive semidefinite symmetric, and U is an orthogonal matrix. Setting $X_i^* = X_i U^T$, we convert $\{X_i^+\}$ into a new feasible solution of (4.9.119). For this solution $\sum_{i=1}^m A_i X_i^* = G_* \succeq 0$, and $\mathrm{Det}(G_+) = \mathrm{Det}(G_*)$, so that the new solution is optimal along with $\{X_i^+\}$.

Problem (I), the coaxial case. We are about to demonstrate that in the coaxial case, when in properly chosen coordinates in $\mathbf{R}^n$ the ellipsoids W_i can be represented as

$$W_i = \{x = A_i u \mid u^T u \le 1\}$$

with positive definite diagonal matrices A_i, the above scheme yields the best (the largest-volume) ellipsoid among those contained in $W = W_1 + \cdots + W_m$. Moreover, this ellipsoid can be pointed out explicitly—it is exactly the ellipsoid $E[Z]$ with $Z = Z(I_n, \ldots, I_n) = A_1 + \cdots + A_m$!

The announced fact is nearly evident. Assuming that A_i are positive definite and diagonal, consider the parallelotope

$$\widehat{W} = \left\{x \in \mathbf{R}^n \mid |x_j| \le \ell_j = \sum_{i=1}^m [A_i]_{jj},\ j = 1, \ldots, n\right\}.$$

This parallelotope clearly contains W (why?), and the largest-volume ellipsoid contained in $\widehat{W}$ clearly is the ellipsoid

$$\left\{x \mid \sum_{j=1}^{n} \ell_j^{-2} x_j^2 \leq 1\right\},$$

i.e., is nothing else but the ellipsoid $E[A_1 + \cdots + A_m]$. As we know from our previous considerations, the latter ellipsoid is contained in W, and since it is the largest-volume ellipsoid among those contained in the set $\widehat{W} \supset W$, it is the largest-volume ellipsoid contained in W as well.

Example. In the example to follow we want to understand what domain D_T on the 2D plane which can be reached by a trajectory of the differential equation

$$\frac{d}{dt}\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \underbrace{\begin{pmatrix} -0.8147 & -0.4163 \\ 0.8167 & -0.1853 \end{pmatrix}}_{A} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} + \begin{pmatrix} u_1(t) \\ 0.7071 u_2(t) \end{pmatrix}, \quad \begin{pmatrix} x_1(0) \\ x_2(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

in T seconds under a piecewise-constant control $u(t) = \binom{u_1(t)}{u_2(t)}$ which switches from one constant value to another one every $\Delta t = 0.01$ second and is subject to the norm bound

$$\|u(t)\|_2 \leq 1 \quad \forall t.$$

The system is stable (the eigenvalues of A are $-0.5 \pm 0.4909i$). To build D_T, note that the states of the system at time instants $k\Delta t$, $k = 0, 1, 2, \ldots$, are the same as the states $x[k] = \binom{x_1(k\Delta t)}{x_2(k\Delta t)}$ of the discrete time system

$$x[k+1] = \underbrace{\exp\{A\Delta t\}}_{S} x[k] + \underbrace{\left[\int_0^{\Delta t} \exp\{As\} \begin{pmatrix} 1 & 0 \\ 0 & 0.7071 \end{pmatrix} ds\right]}_{B} u[k], \;\; x[0] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{4.9.120}$$

where $u[k]$ is the value of the control on the continuous time interval $(k\Delta t, (k+1)\Delta t)$.

We build the inner $\mathcal{I}_k$ and the outer $\mathcal{O}_k$ ellipsoidal approximations of the domains $D^k = D_{k\Delta t}$ in a recurrent manner:

- The ellipses $\mathcal{I}_0$ and $\mathcal{O}_0$ are just the singletons (the origin).
- $\mathcal{I}_{k+1}$ is the best (the largest in the area) ellipsis contained in the set

  $$S\mathcal{I}_k + BW, \quad W = \{u \in \mathbf{R}^2 \mid \|u\|_2 \leq 1\},$$

  which is the sum of two ellipses.
- $\mathcal{O}_{k+1}$ is the best (the smallest in the area) ellipsis containing the set

  $$S\mathcal{O}_k + BW,$$

  which again is the sum of two ellipses.

The picture we get is shown in Fig. 4.17.

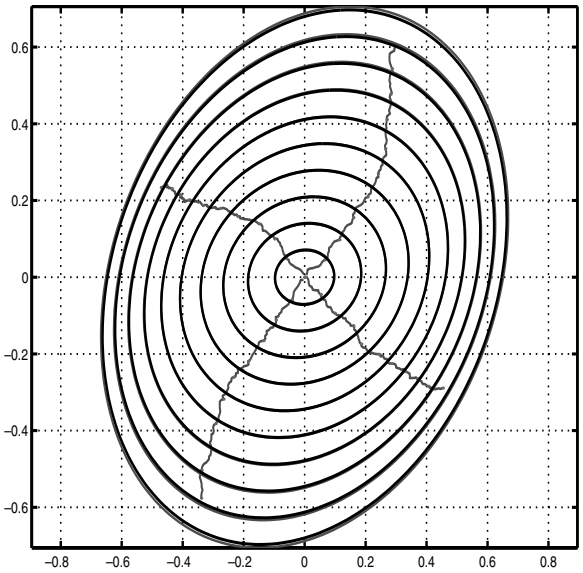


Figure 4.17. *Outer and inner approximations of the reachability domains* $D^{10\ell} = D_{0.1\ell \text{ sec}}$, $\ell = 1, 2, \ldots, 10$, *for system* (4.9.120). *Ten pairs of ellipses are the outer and inner approximations of the domains* $D^1, \ldots, D^{10}$. *(Look how close the ellipses from a pair are to each other!) Four curves are sample trajectories of the system. (Dots correspond to time instants* 0.1ℓ *seconds in continuous time, i.e., time instants* 10ℓ *in discrete time,* $\ell = 0, 1, \ldots, 10$.*)*

4.10 Exercises to Lecture 4

4.10.1 Around positive semidefiniteness, eigenvalues, and $\succeq$-ordering

Criteria for positive semidefiniteness

Recall the criterion of positive definiteness of a symmetric matrix:

> [Sylvester's Rule] *A symmetric* $m \times m$ *matrix* $A = [a_{ij}]_{i,j=1}^m$ *is positive definite if and only if all angular minors*
>
> $$\text{Det}\left([a_{ij}]_{i,j=1}^k\right), \quad k = 1, \ldots, m,$$
>
> *are positive.*

Exercise 4.1. *Prove that a symmetric* $m \times m$ *matrix* A *is positive semidefinite if and only if all its principal minors (i.e., determinants of square submatrices symmetric with respect to the diagonal) are nonnegative.*

Hint. Look at the angular minors of the matrices $A + \epsilon I_n$ for small positive ϵ.

Demonstrate by an example that nonnegativity of angular minors of a symmetric matrix is not sufficient for the positive semidefiniteness of the matrix.

Exercise 4.2. Diagonal-dominant matrices. *Let a symmetric matrix* $A = [a_{ij}]_{i,j=1}^m$ *satisfy the relation*

$$a_{ii} \geq \sum_{j \neq i} |a_{ij}|, \ i = 1, \ldots, m.$$

Prove that A is positive semidefinite.

Diagonalization

Exercise 4.3. *Prove the following standard facts from linear algebra:*

1. *If A is a symmetric positive semidefinite* $m \times m$ *matrix and P is an* $n \times m$ *matrix, then the matrix* PAP^T *is positive semidefinite.*

2. *A symmetric* $m \times m$ *matrix A is positive semidefinite if and only if it can be represented as* $A = Q\Lambda Q^T$*, where Q is orthogonal* ($Q^T Q = I$) *and* Λ *is diagonal with nonnegative diagonal entries. What are these entries? What are the columns of Q?*

3. *Let A, B be two symmetric matrices of the same size and let A be positive definite. Then there exist nonsingular square matrix Q and diagonal matrix* Λ *such that*

$$A = QQ^T, B = Q\Lambda Q^T.$$

Variational characterization of eigenvalues

The basic fact about eigenvalues of a symmetric matrix is the following.

> **Variational description of eigenvalues.** *Let A be a symmetric* $m \times m$ *matrix and* $\lambda(A) = (\lambda_1(A), \ldots, \lambda_m(A))$ *be the vector of eigenvalues of A taken with their multiplicities and arranged in nonascending order:*
>
> $$\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_m(A).$$
>
> *Then for every* $i = 1, \ldots, m$ *one has*
>
> $$\lambda_i(A) = \min_{E \in \mathcal{E}_i} \max_{v \in E, v^T v = 1} v^T A v,$$
>
> *where* $\mathcal{E}_i$ *is the family of all linear subspaces of* $\mathbf{R}^m$ *of dimension* $m - i + 1$.

Singular values of rectangular matrices also admit variational description:

> **Variational description of singular values.** *Let A be an* $m \times n$ *matrix,* $m \leq n$*, and let* $\sigma(A) = \lambda((AA^T)^{1/2})$ *be the vector of singular values of A. Then for every* $i = 1, \ldots, m$ *one has*
>
> $$\sigma_i(A) = \min_{E \in \mathcal{E}_i} \max_{v \in E, v^T v = 1} \|Av\|_2,$$
>
> *where* $\mathcal{E}_i$ *is the family of all linear subspaces of* $\mathbf{R}^n$ *of dimension* $n - i + 1$.

Exercise 4.4. *Prove the variational description of eigenvalues and the variational description of singular values.*

Exercise 4.5. *Derive from the variational description of eigenvalues the following facts:*

1. (monotonicity of the vector of eigenvalues) *If $A \succeq B$, then $\lambda(A) \geq \lambda(B)$.*
2. (interlacing of eigenvalues) *Let $A \in \mathbf{S}^m$, and let E be a linear subspace of $\mathbf{R}^m$ of codimension $k < m$ (i.e., of dimension $m - k$). Let A_E be the restriction of the operator $x \mapsto Ax$ onto E, i.e., the operator $x \mapsto PAx : E \to E$, where P is the orthoprojector onto E. (In terms of matrices: let $e_1, \ldots, e_{n-k}$ be an orthonormal basis in E; you may think of A_E as of the $(n-k) \times (n-k)$ matrix with the entries $e_i^T A e_j$, $i, j = 1, \ldots, n-k$.) Then for every $i \leq n - k$ one has*

$$\lambda_i(A) \geq \lambda_i(A_E) \geq \lambda_{i+k}(A).$$

3. *The functions $\lambda_1(X)$, $\lambda_m(X)$ of $X \in \mathbf{S}^m$ are convex and concave, respectively.*
4. *If Δ is a convex subset of the real axis, then the set of all matrices $X \in \mathbf{S}^m$ with spectrum from Δ is convex.*

Recall now the definition of a function of symmetric matrix. Let A be a symmetric $m \times m$ matrix and

$$p(t) = \sum_{i=0}^{k} p_i t^i$$

be a real polynomial on the axis. By definition,

$$p(A) = \sum_{i=0}^{k} p_i A^i \in \mathbf{S}^m.$$

This definition is compatible with the arithmetic of real polynomials: when you add or multiply polynomials, you add or multiply the values of these polynomials at every fixed symmetric matrix:

$$(p+q)(A) = p(A) + q(A); \; (p \cdot q)(A) = p(A)q(A).$$

A nice feature of this definition is that

> **(A)** *For $A \in \mathbf{S}^m$, the matrix $p(A)$ depends only on the restriction of p on the spectrum (set of eigenvalues) of A: if p and q are two polynomials such that $p(\lambda_i(A)) = q(\lambda_i(A))$ for $i = 1, \ldots, m$, then $p(A) = q(A)$.*

Indeed, we can represent a symmetric matrix A as $A = U^T \Lambda U$, where U is orthogonal and Λ is diagonal with the eigenvalues of A on its diagonal. Since $UU^T = I$, we have $A^i = U^T \Lambda^i U$; consequently,

$$p(A) = U^T p(\Lambda) U,$$

and since the matrix $p(\Lambda)$ depends on the restriction of p on the spectrum of A only, the result follows.

As a byproduct of our reasoning, we get an explicit representation of $p(A)$ in terms of the spectral decomposition $A = U^T \Lambda U$ (U is orthogonal, Λ is diagonal with the diagonal $\lambda(A)$):

(B) *The matrix $p(A)$ is just $U^T \mathrm{Diag}(p(\lambda_1(A)), \ldots, p(\lambda_n(A)))U$.*

(A) allows us to define arbitrary functions of matrices, not necessarily polynomials:

> *Let A be a symmetric matrix and f be a real-valued function defined at least at the spectrum of A. By definition, the matrix $f(A)$ is defined as $p(A)$, where p is a polynomial coinciding with f on the spectrum of A.* (The definition makes sense, since by **(A)** $p(A)$ depends only on the restriction of p on the spectrum of A, i.e., every polynomial continuation $p(\cdot)$ of f from the spectrum of A to the entire axis results in the same $p(A)$).)

The calculus of functions of a symmetric matrix is fully compatible with the usual arithmetic of functions, e.g.,

$$(f+g)(A) = f(A) + g(A);\ (\mu f)(A) = \mu f(A);\ (f \cdot g)(A) = f(A)g(A);\ (f \circ g)(A) = f(g(A)),$$

provided that the functions in question are well defined on the spectrum of the corresponding matrix. And of course the spectral decomposition of $f(A)$ is just $f(A) = U^T \mathrm{Diag}(f(\lambda_1(A)), \ldots, f(\lambda_m(A)))U$, where $A = U^T \mathrm{Diag}(\lambda_1(A), \ldots, \lambda_m(A))U$ is the spectral decomposition of A.

Note that calculus of functions of symmetric matrices becomes very unusual when we are trying to operate with functions of several (noncommuting) matrices. For example, it is generally not true that $\exp\{A+B\} = \exp\{A\}\exp\{B\}$ (the right-hand side matrix may even be nonsymmetric!). It is also generally not true that if f is monotone and $A \succeq B$, then $f(A) \succeq f(B)$, etc.

Exercise 4.6. *Demonstrate by an example that the relation $0 \preceq A \preceq B$ does not necessarily imply that $A^2 \preceq B^2$.*

By the way, the relation $0 \preceq A \preceq B$ does imply that $0 \preceq A^{1/2} \preceq B^{1/2}$. Sometimes, however, we can get weak matrix versions of usual arithmetic relations, as in the following exercise.

Exercise 4.7. *Let f be a nondecreasing function on the real line, and let $A \succeq B$. Prove that $\lambda(f(A)) \geq \lambda(f(B))$.*

The strongest (and surprising) weak matrix version of a usual (scalar) inequality is as follows.

Let $f(t)$ be a closed convex function on the real line; by definition, it means that f is a function on the axis taking real values and the value $+\infty$ such that the set $\mathrm{Dom} f$ of the values of argument where f is finite is convex and nonempty, and, if a sequence $\{t_i \in \mathrm{Dom} f\}$ converges to a point t and the sequence $f(t_i)$ has a limit, then $t \in \mathrm{Dom} f$ and $f(t) \leq \lim_{i\to\infty} f(t_i)$. (This property is called lower semicontinuity.)

For example, the function

$$f(x) = \begin{cases} 0, & 0 \le t \le 1 \\ +\infty & \text{otherwise} \end{cases}$$

is closed. In contrast to this, the functions

$$g(x) = \begin{cases} 0, & 0 < t \le 1, \\ 1, & t = 0, \\ +\infty & \text{for all remaining } t \end{cases}$$

and

$$h(x) = \begin{cases} 0, & 0 < t < 1, \\ +\infty & \text{otherwise} \end{cases}$$

are not closed, although they are convex: a closed function cannot jump up at an endpoint of its domain, as is the case for g, and it cannot take value $+\infty$ at a point, if it takes values $\le a < \infty$ in a neighborhood of the point, as is the case for h.

For a convex function f, its Legendre transformation f_* (also called the conjugate or the Fenchel dual of f) is defined as

$$f_*(s) = \sup_t [ts - f(t)].$$

It turns out that the Legendre transformation of a closed convex function also is closed and convex, and that the twice-taken Legendre transformation of a closed convex function is this function.

The Legendre transformation (which, by the way, can be defined for convex functions on $\mathbf{R}^n$ as well) underlies many standard inequalities. Indeed, by definition of f_* we have

$$f_*(s) + f(t) \ge st \quad \forall s, t. \tag{L}$$

For specific choices of f, we can derive from the general inequality (L) many useful inequalities. For example,

- If $f(t) = \frac{1}{2}t^2$, then $f_*(s) = \frac{1}{2}s^2$, and (L) becomes the standard inequality

$$st \le \frac{1}{2}t^2 + \frac{1}{2}s^2 \quad \forall s, t \in \mathbf{R}.$$

- If $1 < p < \infty$ and

$$f(t) = \begin{cases} \frac{t^p}{p}, & t \ge 0 \\ +\infty, & t < 0 \end{cases},$$

then

$$f_*(s) = \begin{cases} \frac{s^q}{q}, & s \ge 0 \\ +\infty, & s < 0 \end{cases},$$

with q given by $\frac{1}{p} + \frac{1}{q} = 1$, and (L) becomes the Young inequality

$$\forall (s, t \ge 0): \quad ts \le \frac{t^p}{p} + \frac{s^q}{q}, \ 1 < p, q < \infty, \frac{1}{p} + \frac{1}{q} = 1.$$

Now, what happens with (L) if s, t are symmetric matrices? Of course, both sides of (L) still make sense and are matrices, but we have no hope to say something reasonable about the relation between these matrices (e.g., the right-hand side in (L) is not necessarily symmetric). However, we come to the next exercise.

Exercise 4.8. *Let f_* be a closed convex function with the domain* $\mathrm{Dom} f_* \subset \mathbf{R}_+$*, and let f be the Legendre transformation of f_*. Then for every pair of symmetric matrices X, Y of the same size with the spectrum of X belonging to* $\mathrm{Dom} f$ *and the spectrum of Y belonging to* $\mathrm{Dom} f_*$ *one has*[41]

$$\lambda(f(X)) \geq \lambda\left(Y^{1/2} X Y^{1/2} - f_*(Y)\right).$$

Birkhoff's theorem

Surprisingly enough, one of the most useful facts about eigenvalues of symmetric matrices is the following, essentially combinatorial, statement (it does not mention the word eigenvalue at all).

Birkhoff's theorem. *Consider the set $\mathcal{S}_m$ of double-stochastic $m \times m$ matrices, i.e., square matrices $[p_{ij}]_{i,j=1}^m$ satisfying the relations*

$$\begin{aligned} p_{ij} &\geq 0, \ i, j = 1, \ldots, m, \\ \sum_{i=1}^m p_{ij} &= 1, \ j = 1, \ldots, m, \\ \sum_{j=1}^m p_{ij} &= 1, \ i = 1, \ldots, m. \end{aligned}$$

A matrix P belongs to $\mathcal{S}_m$ if and only if it can be represented as a convex combination of $m \times m$ permutation matrices:

$$P \in \mathcal{S}_m \Leftrightarrow \exists \left(\lambda_i \geq 0, \sum_i \lambda_i = 1 \right) : P = \sum_i \lambda_i \Pi^i,$$

where all Π^i are permutation matrices (i.e., with exactly one nonzero element, equal to 1*, in every row and every column).*

An immediate corollary of the Birkhoff theorem is the following fact:

(C) *Let $f : \mathbf{R}^m \to \mathbf{R} \cup \{+\infty\}$ be a convex symmetric function (symmetry means that the value of the function remains unchanged when we permute the coordinates in an argument), let $x \in \mathrm{Dom} f$, and let $P \in \mathcal{S}_m$. Then*

$$f(Px) \leq f(x).$$

[41] In the scalar case, our inequality reads $f(x) \geq y^{1/2} x y^{1/2} - f_*(y)$, which is an equivalent form of (L) when $\mathrm{Dom} f_* \subset \mathbf{R}_+$.

The proof is immediate. By Birkhoff's theorem, Px is a convex combination of a number of permutations x^i of x. Since f is convex, we have

$$f(Px) \leq \max_i f(x^i) = f(x),$$

the concluding equality resulting from the symmetry of f.

The role of **(C)** in numerous questions related to eigenvalues is based on the following simple observation.

> *Let A be a symmetric $m \times m$ matrix. Then the diagonal* $\mathrm{Dg}(A)$ *of the matrix A is the image of the vector* $\lambda(A)$ *of the eigenvalues of A under multiplication by a double stochastic matrix:*
>
> $$\mathrm{Dg}(A) = P\lambda(A) \text{ for some } P \in \mathcal{S}_m$$

Indeed, consider the spectral decomposition of A:

$$A = U^T \mathrm{Diag}(\lambda_1(A), \ldots, \lambda_m(A))U$$

with orthogonal $U = [u_{ij}]$. Then

$$A_{ii} = \sum_{j=1}^m u_{ji}^2 \lambda_j(A) \equiv (P\lambda(A))_i,$$

where the matrix $P = [u_{ji}^2]_{i,j=1}^m$ is double stochastic.

Combining the observation and **(C)**, we conclude that if f is a convex symmetric function on $\mathbf{R}^m$, then for every $m \times m$ symmetric matrix A one has

$$f(\mathrm{Dg}(A)) \leq f(\lambda(A)).$$

Moreover, let $\mathcal{O}_m$ be the set of all orthogonal $m \times m$ matrices. For every $V \in \mathcal{O}_m$, the matrix $V^T AV$ has the same eigenvalues as A, so that for a convex symmetric f one has

$$f(\mathrm{Dg}(V^T AV)) \leq f(\lambda(V^T AV)) = f(\lambda(A)),$$

whence

$$f(\lambda(A)) \geq \max_{V \in \mathcal{O}_m} f(\mathrm{Dg}(V^T AV)).$$

In fact, the inequality here is equality, since for properly chosen $V \in \mathcal{O}_m$ we have $\mathrm{Dg}(V^T AV) = \lambda(A)$. We have arrived at the following result:

> **(D)** *Let f be a symmetric convex function on $\mathbf{R}^m$. Then for every symmetric $m \times m$ matrix A one has*
>
> $$f(\lambda(A)) = \max_{V \in \mathcal{O}_m} f(\mathrm{Dg}(V^T AV)),$$
>
> *where $\mathcal{O}_m$ is the set of all $m \times m$ orthogonal matrices.*

In particular, the function

$$F(A) = f(\lambda(A))$$

is convex in $A \in \mathbf{S}^m$ (as the maximum of a family of convex in A functions $F_V(A) = f(\mathrm{Dg}(V^T AV))$, $V \in \mathcal{O}_m$).

Exercise 4.9. *Let $g(t) : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ be a convex function, and let $\mathcal{F}_n$ be the set of all matrices $X \in \mathbf{S}^n$ with the spectrum belonging to* Dom$g$. *Prove that the function* $\mathrm{Tr}(g(X))$ *is convex on $\mathcal{F}_n$.*

Hint. Apply **(D)** to the function $f(x_1, \ldots, x_n) = g(x_1) + \cdots + g(x_n)$.

Exercise 4.10. *Let $A = [a_{ij}]$ be a symmetric $m \times m$ matrix. Prove that*

1. *Whenever $p \geq 1$, one has $\sum_{i=1}^m |a_{ii}|^p \leq \sum_{i=1}^m |\lambda_i(A)|^p$.*
2. *Whenever A is positive semidefinite, $\prod_{i=1}^m a_{ii} \geq \mathrm{Det}(A)$.*
3. *For $x \in \mathbf{R}^m$, let the function $S_k(x)$ be the sum of k largest entries of x (i.e., the sum of the first k entries in the vector obtained from x by writing the coordinates of x in the nonascending order). Prove that $S_k(x)$ is a convex symmetric function of x and derive from this observation that*

$$S_k(\mathrm{Dg}(A)) \leq S_k(\lambda(A)).$$

Hint. Note that $S_k(x) = \max_{1 \leq i_1 < i_2 < \ldots < i_k \leq m} \sum_{l=1}^k x_{i_l}$.

4. *(Trace inequality.) Whenever $A, B \in \mathbf{S}^m$, one has*

$$\lambda^T(A)\lambda(B) \geq \mathrm{Tr}(AB).$$

Exercise 4.11. *Prove that if $A \in \mathbf{S}^m$ and $p, q \in [1, \infty]$ are such that $\frac{1}{p} + \frac{1}{q} = 1$, then*

$$\max_{B \in \mathbf{S}^m : \|\lambda(B)\|_q = 1} \mathrm{Tr}(AB) = \|\lambda(A)\|_p.$$

In particular, $\|\lambda(\cdot)\|_p$ is a norm on $\mathbf{S}^m$, and the conjugate of this norm is $\|\lambda(\cdot)\|_q$, $\frac{1}{p} + \frac{1}{q} = 1$.

Exercise 4.12. *Let*

$$X = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1m} \\ X_{12}^T & X_{22} & \ldots & X_{2m} \\ \vdots & \vdots & \ddots & \ldots \\ X_{1m}^T & X_{2m}^T & \ldots & X_{mm} \end{pmatrix}$$

be an $n \times n$ symmetric matrix that is partitioned into m^2 blocks X_{ij} in a symmetric, with respect to the diagonal, fashion (so that the blocks X_{jj} are square), and let

$$\widehat{X} = \begin{pmatrix} X_{11} & & & \\ & X_{22} & & \\ & & \ddots & \\ & & & X_{mm} \end{pmatrix}.$$

1. *Let $F : \mathbf{S}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex rotation-invariant function: for all $Y \in \mathbf{S}^n$ and all orthogonal matrices U one has $F(U^T Y U) = F(Y)$. Prove that*

$$F(\widehat{X}) \leq F(X).$$

Hint. Represent the matrix $\widehat{X}$ as a convex combination of the rotations $U^T X U$, $U^T U = I$, of X.

2. *Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex symmetric with respect to permutations of the entries in the argument function, and let $F(Y) = f(\lambda(Y))$, $Y \in \mathbf{S}^n$. Prove that*

$$F(\widehat{X}) \leq F(X).$$

3. *Let $g : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ be a convex function on the real line which is finite on the set of eigenvalues of X, and let $\mathcal{F}_n \subset \mathbf{S}^n$ be the set of all $n \times n$ symmetric matrices with all eigenvalues belonging to the domain of g. Assume that the mapping*

$$Y \mapsto g(Y) : \mathcal{F}_n \to \mathbf{S}^n$$

is $\succeq$-convex:

$$g(\lambda' Y' + \lambda'' Y'') \preceq \lambda' g(Y') + \lambda'' g(Y'') \quad \forall (Y', Y'' \in \mathcal{F}_n, \lambda', \lambda'' \geq 0, \lambda' + \lambda'' = 1).$$

Prove that

$$(g(X))_{ii} \succeq g(X_{ii}), \ i = 1, \ldots, m,$$

where the partition of $g(X)$ into the blocks $(g(X))_{ij}$ is identical to the partition of X into the blocks X_{ij}.

Exercise 4.12 gives rise to a number of interesting inequalities. Let X, $\widehat{X}$ be the same as in the exercise, and let $[Y]$ denote the northwest block, of the same size as X_{11}, of an $n \times n$ matrix Y. Then

1. $(\sum_{i=1}^m \|\lambda(X_{ii})\|_p^p)^{1/p} \leq \|\lambda(X)\|_p$, $1 \leq p < \infty$
   [Exercise 4.12.2, $f(x) = \|x\|_p$].
2. If $X \succ 0$, then $\mathrm{Det}(X) \leq \prod_{i=1}^m \mathrm{Det}(X_{ii})$
   [Exercise 4.12.2, $f(x) = -(x_1 \ldots x_n)^{1/n}$ for $x \geq 0$].
3. $[X^2] \succeq X_{11}^2$
   [This inequality is nearly evident. It follows also from exercise 4.12.3 with $g(t) = t^2$. The $\succeq$-convexity of $g(Y)$ is stated in Exercise 4.22.1.]
4. If $X \succ 0$, then $X_{11}^{-1} \preceq [X^{-1}]$
   [Exercise 4.12.3 with $g(t) = t^{-1}$ for $t > 0$; the $\succeq$-convexity of $g(Y)$ on $\mathbf{S}^n_{++}$ is stated by Exercise 4.22.2].

5. For every $X \succeq 0$, $[X^{1/2}] \preceq X_{11}^{1/2}$

   [Exercise 4.12.3 with $g(t) = -\sqrt{t}$; the $\succeq$-convexity of $g(Y)$ is stated by Exercise 4.22.4].

   Extension: If $X \succeq 0$, then for every $\alpha \in (0, 1)$ one has $[X^\alpha] \preceq X_{11}^\alpha$

   [Exercise 4.12.3 with $g(t) = -t^\alpha$; the function $-Y^\alpha$ of $Y \succeq 0$ is known to be $\succeq$-convex].

6. If $X \succ 0$, then $[\ln(X)] \preceq \ln(X_{11})$

   [Exercise 4.12.3 with $g(t) = -\ln t$, $t > 0$; the $\succeq$-convexity of $g(Y)$ is stated by Exercise 4.22.5].

Exercise 4.13. 1. *Let $A = [a_{ij}]_{i,j} \succeq 0$, let $\alpha \geq 0$, and let $B \equiv [b_{ij}]_{i,j} = A^\alpha$. Prove that*

$$b_{ii} \begin{cases} \leq a_{ii}^\alpha, & \alpha \leq 1, \\ \geq a_{ii}^\alpha, & \alpha \geq 1. \end{cases}$$

2. *Let $A = [a_{ij}]_{i,j} \succ 0$, and let $B \equiv [b_{ij}]_{i,j} = A^{-1}$. Prove that $b_{ii} \geq a_{ii}^{-1}$.*

3. *Let $[A]$ denote the northwest 2×2 block of a square matrix. Which of the implications*

$$\begin{array}{ll} \text{(a)} & A \succeq 0 \Rightarrow [A^4] \succeq [A]^4, \\ \text{(b)} & A \succeq 0 \Rightarrow [A^4]^{1/4} \succeq [A] \end{array}$$

are true?

Semidefinite representations of functions of eigenvalues

The goal of the subsequent series of exercises is to prove Proposition 4.2.1.

We start with a description (important in its own right) of the convex hull of permutations of a given vector. Let $x \in \mathbf{R}^m$, and let $X[x]$ be the set of all convex combinations of $m!$ vectors obtained from x by all permutations of the coordinates.

Claim. (majorization principle) *$X[x]$ is exactly the solution set of the following system of inequalities in variables $y \in \mathbf{R}^m$:*

$$\begin{array}{c} S_j(y) \leq S_j(x), \quad j = 1, \ldots, m-1, \\ y_1 + \cdots + y_m = x_1 + \cdots + x_m. \end{array} \qquad (+)$$

(Recall that $S_j(y)$ is the sum of the largest j entries of a vector y.)

Exercise 4.14. Easy part of the claim. *Let Y be the solution set of* (+)*. Prove that $Y \supset X[x]$.*

Hint. Use **(C)** and the convexity of the functions $S_j(\cdot)$.

Exercise 4.15. Difficult part of the claim. *Let Y be the solution set of* (+)*. Prove that $Y \subset X[x]$.*

Sketch of the proof. Let $y \in Y$. We should prove that $y \in X[x]$. By symmetry, we may assume that the vectors x and y are ordered: $x_1 \geq x_2 \geq \cdots \geq x_m$, $y_1 \geq y_2 \geq \cdots \geq y_m$. Assume that $y \notin X[x]$, and let us lead this assumption to a contradiction.

1. Since $X[x]$ clearly is a convex compact set and $y \notin X[x]$, there exists a linear functional $c(z) = \sum_{i=1}^m c_i z_i$ which separates y and $X[x]$:

$$c(y) > \max_{z \in X[x]} c(z).$$

Prove that such a functional can be chosen to be ordered: $c_1 \geq c_2 \geq \cdots \geq c_m$.

2. Verify that

$$c(y) \equiv \sum_{i=1}^m c_i y_i = \sum_{i=1}^{m-1} (c_i - c_{i+1}) \sum_{j=1}^i y_j + c_m \sum_{j=1}^m y_j$$

(Abel's formula—a discrete version of integration by parts). Use this observation along with orderedness of $c(\cdot)$ and the inclusion $y \in Y$ to conclude that $c(y) \leq c(x)$, thus coming to the desired contradiction. □

Exercise 4.16. *Use the majorization principle to prove Proposition* 4.2.1.

The next pair of exercises is aimed at proving Proposition 4.2.2.

Exercise 4.17. *Let $x \in \mathbf{R}^m$ and let $X_+[x]$ be the set of all vectors x' dominated by a vector form $X[x]$:*

$$X_+[x] = \{y \mid \exists z \in X[x] : y \leq z\}.$$

1. *Prove that $X_+[x]$ is a closed convex set.*
2. *Prove the following characterization of $X_+[x]$:*

$X_+[x]$ is exactly the set of solutions of the system of inequalities $S_j(y) \leq S_j(x)$, $j = 1, \ldots, m$, in variables y.

Hint. Modify appropriately the constriction outlined in Exercise 4.15.

Exercise 4.18. *Derive Proposition* 4.2.2 *from the result of Exercise* 4.17.2.

Cauchy's inequality for matrices

The standard Cauchy's inequality says that

$$\left| \sum_i x_i y_i \right| \leq \sqrt{\sum_i x_i^2} \sqrt{\sum_i y_i^2} \tag{4.10.121}$$

for reals x_i, y_i, $i = 1, \ldots, n$. This inequality is exact in the sense that for every collection $x_1, \ldots, x_n$ there exists a collection $y_1, \ldots, y_n$ with $\sum_i y_i^2 = 1$ which makes (4.10.121) an equality.

Exercise 4.19. 1. *Prove that whenever $X_i, Y_i \in \mathbf{M}^{p,q}$, one has*

$$\sigma\left(\sum_i X_i^T Y_i\right) \le \lambda\left(\left[\sum_i X_i^T X_i\right]^{1/2}\right) \| \lambda\left(\sum_i Y_i^T Y_i\right) \|_\infty^{1/2}, \qquad (*)$$

where $\sigma(A) = \lambda([AA^T]^{1/2})$ is the vector of singular values of a matrix A arranged in the nonascending order.

Prove that for every collection $X_1, \dots, X_n \in \mathbf{M}^{p,q}$ there exists a collection $Y_1, \dots, Y_n \in \mathbf{M}^{p,q}$ with $\sum_i Y_i^T Y_i = I_q$ which makes $(*)$ an equality.

2. *Prove the following matrix version of the Cauchy inequality: whenever $X_i, Y_i \in \mathbf{M}^{p,q}$, one has*

$$\left|\sum_i \mathrm{Tr}(X_i^T Y_i)\right| \le \mathrm{Tr}\left(\left[\sum_i X_i^T X_i\right]^{1/2}\right) \| \lambda\left(\sum_i Y_i^T Y_i\right) \|_\infty^{1/2}, \qquad (**)$$

and for every collection $X_1, \dots, X_n \in \mathbf{M}^{p,q}$ there exists a collection $Y_1, \dots, Y_n \in \mathbf{M}^{p,q}$ with $\sum_i Y_i^T Y_i = I_q$ which makes $(**)$ an equality.

Here is another exercise of the same flavor.

Exercise 4.20. *For nonnegative reals $a_1, \dots, a_m$ and a real $\alpha > 1$ one has*

$$\left(\sum_{i=1}^m a_i^\alpha\right)^{1/\alpha} \le \sum_{i=1}^m a_i.$$

Both sides of this inequality make sense when the nonnegative reals a_i are replaced with positive semidefinite $n \times n$ matrices A_i. What happens with the inequality in this case?

Consider the following four statements (where $\alpha > 1$ is a real and $m, n > 1$):

$$1. \quad \forall (A_i \in \mathbf{S}_+^n): \quad \left(\sum_{i=1}^m A_i^\alpha\right)^{1/\alpha} \preceq \sum_{i=1}^m A_i.$$

$$2. \quad \forall (A_i \in \mathbf{S}_+^n): \quad \lambda_{\max}\left(\left(\sum_{i=1}^m A_i^\alpha\right)^{1/\alpha}\right) \le \lambda_{\max}\left(\sum_{i=1}^m A_i\right).$$

$$3. \quad \forall (A_i \in \mathbf{S}_+^n): \quad \mathrm{Tr}\left(\left(\sum_{i=1}^m A_i^\alpha\right)^{1/\alpha}\right) \le \mathrm{Tr}\left(\sum_{i=1}^m A_i\right).$$

$$4. \quad \forall (A_i \in \mathbf{S}_+^n): \quad \mathrm{Det}\left(\left(\sum_{i=1}^m A_i^\alpha\right)^{1/\alpha}\right) \le \mathrm{Det}\left(\sum_{i=1}^m A_i\right).$$

Of these four statements, exactly two are true. Identify and prove the true statements.

$\succeq$-convexity of some matrix-valued functions

Consider a function $F(x)$ defined on a convex set $X \subset \mathbf{R}^n$ and taking values in $\mathbf{S}^m$. We say that such a function is $\succeq$-convex if

$$F(\alpha x + (1-\alpha)y) \preceq \alpha F(x) + (1-\alpha)F(y)$$

$\forall x, y \in X$ and $\forall \alpha \in [0,1]$. F is called $\succeq$-concave if $-F$ is $\succeq$-convex.

A function $F : \mathrm{Dom}F \to \mathbf{S}^m$ defined on a set $\mathrm{Dom}F \subset \mathbf{S}^k$ is called $\succeq$-monotone if

$$x, y \in \mathrm{Dom}F, x \succeq y \Rightarrow F(x) \succeq F(y);$$

F is called $\succeq$-antimonotone if $-F$ is $\succeq$-monotone.

Exercise 4.21. 1. *Prove that a function $F : X \to \mathbf{S}^m$, $X \subset \mathbf{R}^n$, is $\succeq$-convex if and only if its epigraph*

$$\{(x, Y) \in \mathbf{R}^n \to \mathbf{S}^m \mid x \in X, F(x) \preceq Y\}$$

is a convex set.

2. *Prove that a function $F : X \to \mathbf{S}^m$ with convex domain $X \subset \mathbf{R}^n$ is $\succeq$-convex if and only if for every $A \in \mathbf{S}^m_+$ the function* $\mathrm{Tr}(AF(x))$ *is convex on X.*

3. *Let $X \subset \mathbf{R}^n$ be a convex set with a nonempty interior and $F : X \to \mathbf{S}^m$ be a function continuous on X which is twice differentiable in* $\mathrm{int}X$. *Prove that F is $\succeq$-convex if and only if the second directional derivative of F*

$$D^2F(x)[h,h] \equiv \left.\frac{d^2}{dt^2}\right|_{t=0} F(x+th)$$

is $\succeq 0$ for every $x \in \mathrm{int}X$ and every direction $h \in \mathbf{R}^n$.

4. *Let $F : \mathrm{dom}F \to \mathbf{S}^m$ be defined and continuously differentiable on an open convex subset of $\mathbf{S}^k$. Prove that the necessary and sufficient condition for F to be $\succeq$-monotone is the validity of the implication*

$$h \in \mathbf{S}^k_+, x \in \mathrm{Dom}F \Rightarrow DF(x)[h] \succeq 0.$$

5. *Let F be $\succeq$-convex and $\mathcal{S} \subset \mathbf{S}^m$ be a convex set that is $\succeq$-antimonotone, i.e., whenever $Y' \preceq Y$ and $Y \in \mathcal{S}$, one has $Y' \in \mathcal{S}$. Prove that the set $F^{-1}(\mathcal{S}) = \{x \in X \mid F(x) \in \mathcal{S}\}$ is convex.*

6. *Let $G : \mathrm{Dom}G \to \mathbf{S}^k$ and $F : \mathrm{Dom}F \to \mathbf{S}^m$, let $G(\mathrm{Dom}G) \subset \mathrm{Dom}F$, and let $H(x) = F(G(x)) : \mathrm{Dom}G \to \mathbf{S}^m$.*

(a) *Prove that if G and F are $\succeq$-convex and F is $\succeq$-monotone, then H is $\succeq$-convex.*

(b) *Prove that if G and F are $\succeq$-concave and F is $\succeq$-monotone, then H is $\succeq$-concave.*

7. *Let $F_i : G \to \mathbf{S}^m$, and assume that for every $x \in G$ exists*

$$F(x) = \lim_{i\to\infty} F_i(x).$$

Prove that if all functions from the sequence $\{F_i\}$ are $\succeq$-convex or $\succeq$-concave, or $\succeq$-monotone or $\succeq$-antimonotone, then so is F.

The goal of the next exercise is to establish the $\succeq$-convexity of several matrix-valued functions.

Exercise 4.22. *Prove that the following functions are $\succeq$-convex:*

1. $F(x) = xx^T : \mathbf{M}^{p,q} \to \mathbf{S}^p$.
2. $F(x) = x^{-1} : \mathrm{int}\mathbf{S}^m_+ \to \mathrm{int}\mathbf{S}^m_+$.
3. $F(u, v) = u^T v^{-1} u : \mathbf{M}^{p,q} \times \mathrm{int}\mathbf{S}^p_+ \to \mathbf{S}^q$.

Prove that the following functions are $\succeq$-concave and monotone:

4. $F(x) = x^{1/2} : \mathbf{S}^m_+ \to \mathbf{S}^m$.
5. $F(x) = \ln x : \mathrm{int}\mathbf{S}^m_+ \to \mathbf{S}^m$.
6. $F(x) = \left(Ax^{-1}A^T\right)^{-1} : \mathrm{int}\mathbf{S}^n_+ \to \mathbf{S}^m$, *provided that A is an $m \times n$ matrix of rank m.*

Minors of positive semidefinite matrices

Exercise 4.23. *Let A, B be two $n \times n$ symmetric positive definite matrices. Prove that*

$$\mathrm{Det}^{1/n}(A + B) \geq \mathrm{Det}^{1/n}(A) + \mathrm{Det}^{1/n}(B).$$

Hint. Reduce the general case to the one of $A = I$.

A well-known fact of linear algebra is that a symmetric $m \times m$ matrix (A_{ij}) is positive semidefinite if and only if it is a Gram matrix, i.e.,

there exists a system of m vectors f_i such that

$$A_{ij} = f_i^T f_j$$

$\forall i, j$.

The goal of the subsequent exercises is to prove the following nice extension of the "if" part of this result:

(E) *Let $F_1, \ldots, F_m$ be $p \times q$ rectangular matrices. Then the $m \times m$ matrix with the entries*

$$A_{ij} = \mathrm{Det}(F_i^T F_j), \ i, j = 1, \ldots, m,$$

is positive semidefinite.

(E) is an immediate consequence of the following fact:

(F) *Let B be $pm \times pm$ positive semidefinite symmetric matrix partitioned into m^2 blocks B^{ij} of sizes $p \times p$ each, as in the following example with $m = 3$ and $p = 2$:*

$$B = \left(\begin{array}{cc|cc|cc} b_{11} & b_{12} & b_{13} & b_{14} & b_{15} & b_{16} \\ b_{12} & b_{22} & b_{23} & b_{24} & b_{25} & b_{26} \\ \hline b_{13} & b_{23} & b_{33} & b_{34} & b_{35} & b_{36} \\ b_{14} & b_{24} & b_{34} & b_{44} & b_{45} & b_{46} \\ \hline b_{15} & b_{25} & b_{35} & b_{45} & b_{55} & b_{56} \\ b_{16} & b_{26} & b_{36} & b_{46} & b_{56} & b_{66} \end{array}\right).$$

Then the $m \times m$ matrix $B^{(p),\det}$:

$$B^{(p),\det}_{ij} = \mathrm{Det}(B^{ij}),\ \ i, j = 1, \ldots, m,$$

is positive semidefinite.

Moreover, if $0 \preceq B \preceq C$, then

$$0 \preceq B^{(p),\det} \preceq C^{(p),\det}.$$

Exercise 4.24. *Prove the implication* **(F)⇒(E)**.

The proof of **(F)** is based on the following construction. Let us fix a positive integer k. For an integer $n \geq k$, let $\mathcal{I}_n$ denote the family of all k-element subsets $\bar{i} = \{i_1 < i_2 < \cdots < i_k\}$ of the index set $\{1, 2, \ldots, n\}$. Now, given an $n \times n$ matrix A (not necessarily symmetric), let us define $\binom{n}{k} \times \binom{n}{k}$ matrix $\overline{A}$ as follows: the rows and the columns of $\overline{A}$ are indexed by elements of $\mathcal{I}_n$, and $\overline{A}_{\bar{i}\bar{j}}$ is the $k \times k$ minor of A formed by elements of the rows from the set $\bar{i}$ and the columns from the set $\bar{j}$. A nice property of the mapping $A \mapsto \overline{A}$ is that it preserves multiplication and transposition:

$$\overline{A} \cdot \overline{B} = \overline{AB}; \quad \overline{A^T} = (\overline{A})^T. \tag{$*$}$$

Exercise 4.25. 1. *Verify* (*).

2. *Derive* **(F)** *from* (*).

Hint. Use the result of Exercise 4.3.3.

Recall that if A is an $m \times m$ matrix, then its adjoint A^{adj} is the $m \times m$ matrix with ijth entry A^{adj}_{ij} being the algebraic complement (in A) to the cell i, j.

Exercise 4.26. 1. *Prove that* $\frac{d}{dt}\big|_{t=0} \mathrm{Det}(A + tB) = \mathrm{Tr}(B^T A^{\mathrm{adj}})$.

2. *Derive from* **(F)** *and* 1 *that if B is symmetric $pm \times pm$ positive semidefinite matrix partitioned into m^2 blocks B^{ij} of sizes $p \times p$ each, then the $pm \times pm$ matrix*

$$B^{(p),\mathrm{adj}} = \begin{pmatrix} (B^{11})^{\mathrm{adj}} & \cdots & (B^{1m})^{\mathrm{adj}} \\ \vdots & \cdots & \vdots \\ (B^{m1})^{\mathrm{adj}} & \cdots & (B^{mm})^{\mathrm{adj}} \end{pmatrix}$$

is symmetric positive semidefinite.

3. *Derive* 2 *directly from* **(F)** *and from the following observation:*

For a square matrix A, let A^+ be the matrix with ijth entry A^+_{ij} being the determinant of the matrix obtained from A by eliminating the row i and the column j. Then, in the notation of 2, there exists a diagonal matrix D, independent of B, such that

$$B^{(p),\mathrm{adj}} = D^T B^{(p),+} D,$$

where

$$B^{(p),+} = \begin{pmatrix} (B^{11})^+ & \cdots & (B^{1m})^+ \\ \vdots & \cdots & \vdots \\ (B^{m1})^+ & \cdots & (B^{mm})^+ \end{pmatrix}.$$

4. *Using the same line of argument as in* 3, *prove that the mappings*

$$\begin{array}{rcl} B & \mapsto & B^{(p),+}, \\ B & \mapsto & B^{(p),\mathrm{adj}}. \end{array}$$

are $\succeq$*-monotone mappings of* $\mathbf{S}^{pm}_+$ *into* $\mathbf{S}^{pm}_+$*:*

$$0 \preceq B \preceq C \Rightarrow \left\{ \begin{array}{rcl} 0 \preceq B^{(p),+} & \preceq & C^{(p),+}, \\ 0 \preceq B^{(p),\mathrm{adj}} & \preceq & C^{(p),\mathrm{adj}}. \end{array} \right.$$

Note that the $\succeq$-monotonicity of the mapping

$$A \mapsto A^{\mathrm{adj}} : \mathbf{S}^p_+ \to \mathbf{S}^p_+$$

(stated in 4 with $m = 1$) is a rather surprising fact. Indeed, the mapping $A \mapsto A^{-1} : \mathrm{int}\mathbf{S}^p_+ \to \mathrm{int}\mathbf{S}^p_+$ which is very close to the mapping $A \mapsto A^{\mathrm{adj}}$ (since $A^{-1} = [\mathrm{Det}(A)]^{-1}(A^{\mathrm{adj}})^T$) is $\succeq$-antimonotone:

$$0 \prec A \preceq B \Rightarrow A^{-1} \succeq B^{-1}.$$

4.10.2 Semidefinite representations of epigraphs of convex polynomials

Mathematically speaking, the central question concerning the expressive abilities of SDP is how wide is the family of convex sets that are SDr. By definition, an SDr set is the projection of the inverse image of $\mathbf{S}^m_+$ under affine mapping. In other words, every SDr set is a projection of a convex set given by a number of polynomial inequalities. (Indeed, the cone $\mathbf{S}^m_+$ is a convex set given by polynomial inequalities saying that all principal minors of matrix are nonnegative.) Consequently, the inverse image of $\mathbf{S}^m_+$ under an affine mapping is also a convex set given by a number of (nonstrict) polynomial inequalities. And it is known that every projection of such a set is also given by a number of polynomial inequalities (both strict and nonstrict). We conclude that

> *An SDr set is always a convex set given by finitely many polynomial inequalities (strict and nonstrict).*

A natural (and seemingly very difficult) question is whether the inverse is true—is a convex set given by a number of polynomial inequalities always SDr? This question can be simplified in many ways—we may fix the dimension of the set, we may assume the polynomials participating in inequalities to be convex, we may fix the degrees of the polynomials, etc. To the best of our knowledge, all these questions are open.

The goal of the subsequent exercises is to answer affirmatively the simplest question of the above series:

Let $\pi(x)$ be a convex polynomial of one variable. Then its epigraph

$$\{(t,x) \in \mathbf{R}^2 \mid t \geq \pi(x)\}$$

is SDr.

Let us fix a nonnegative integer k and consider the curve

$$p(x) = (1, x, x^2, \ldots, x^{2k})^T \in \mathbf{R}^{2k+1}.$$

Let Π_k be the closure of the convex hull of values of the curve. How can one describe Π_k?

A convenient way to answer this question is to pass to a matrix representation of all objects involved. Namely, let us associate with a vector $\xi = (\xi_0, \xi_1, \ldots, \xi_{2k}) \in \mathbf{R}^{2k+1}$ the $(k+1) \times (k+1)$ symmetric matrix

$$\mathcal{M}(\xi) = \begin{pmatrix} \xi_0 & \xi_1 & \xi_2 & \xi_3 & \cdots & \xi_k \\ \xi_1 & \xi_2 & \xi_3 & \xi_4 & \cdots & \xi_{k+1} \\ \xi_2 & \xi_3 & \xi_4 & \xi_5 & \cdots & \xi_{k+2} \\ \xi_3 & \xi_4 & \xi_5 & \xi_6 & \cdots & \xi_{k+3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \xi_k & \xi_{k+1} & \xi_{k+2} & \xi_{k+3} & \cdots & \xi_{2k} \end{pmatrix},$$

so that

$$[\mathcal{M}(\xi)]_{ij} = \xi_{i+j}, \ i, j = 0, \ldots, k.$$

The transformation $\xi \mapsto \mathcal{M}(\xi) : \mathbf{R}^{2k+1} \to \mathbf{S}^{k+1}$ is a linear embedding; the image of Π_k under this embedding is the closure of the convex hull of values of the curve

$$P(x) = \mathcal{M}(p(x)).$$

It follows that the image $\widehat{\Pi}_k$ of Π_k under the mapping $\mathcal{M}$ possesses the following properties:

(i) $\widehat{\Pi}_k$ belongs to the image of $\mathcal{M}$, i.e., to the subspace H_k of $\mathbf{S}^{2k+1}$ comprising Hankel matrices—matrices with entries depending on the sum of indices only:

$$H_k = \left\{X \in \mathbf{S}^{2k+1} | i + j = i' + j' \Rightarrow X_{ij} = X_{i'j'}\right\}.$$

(ii) $\widehat{\Pi}_k \subset \mathbf{S}^{k+1}_+$ (indeed, every matrix $\mathcal{M}(p(x))$ is positive semidefinite).

(iii) For every $X \in \widehat{\Pi}_k$ one has $X_{00} = 1$.

It turns out that properties (i)–(iii) characterize $\widehat{\Pi}_k$:

(G) *A symmetric $(k+1) \times (k+1)$ matrix X belongs to $\widehat{\Pi}_k$ if and only if it possesses the properties* (i)–(iii)*: its entries depend on sum of indices only (i.e., $X \in H_k$), X is positive semidefinite and $X_{00} = 1$.*

(G) is a particular case of the classical results related to the so called moment problem. The goal of the subsequent exercises is to give a simple alternative proof of this statement.

Note that the mapping $\mathcal{M}^* : \mathbf{S}^{k+1} \to \mathbf{R}^{2k+1}$ conjugate to the mapping $\mathcal{M}$ is as follows:

$$(\mathcal{M}^* X)_l = \sum_{i=0}^{l} X_{i,l-i}, \ l = 0, 1, \ldots, 2k,$$

and we know something about this mapping: example 21.a (Lecture 4, p. 157) says that

(H) *The image of the cone $\mathbf{S}_+^{k+1}$ under the mapping $\mathcal{M}^*$ is exactly the cone of coefficients of polynomials of degree $\leq 2k$ which are nonnegative on the entire real line.*

Exercise 4.27. *Derive* **(G)** *from* **(H)**.

(G), among other useful things, implies the result we need:

(I) *Let $\pi(x) = \pi_0 + \pi_1 x + \pi_2 x^2 + \cdots + \pi_{2k} x^{2k}$ be a convex polynomial of degree $2k$. Then the epigraph of π is SDr:*

$$\{(t,x) \in \mathbf{R}^2 \mid t \geq p(x)\} = \mathcal{X}[\pi],$$

where

$$\mathcal{X}[\pi] = \left\{ (t,x) \;\middle|\; \exists x_2, \ldots, x_{2k} : \begin{pmatrix} 1 & x & x_2 & x_3 & \ldots & x_k \\ x & x_2 & x_3 & x_4 & \ldots & x_{k+1} \\ x_2 & x_3 & x_4 & x_5 & \ldots & x_{k+2} \\ x_3 & x_4 & x_5 & x_6 & \ldots & x_{k+3} \\ \ldots & \ldots & \ldots & \ldots & \ddots & \ldots \\ x_k & x_{k+1} & x_{k+2} & x_{k+3} & \ldots & x_{2k} \end{pmatrix} \succeq 0, \right.$$
$$\left. \pi_0 + \pi_1 x + \pi_2 x_2 + \pi_3 x_3 + \ldots + \pi_{2k} x_{2k} \leq t \right\}$$

Exercise 4.28. *Prove* **(I)**.

Note that the set $\mathcal{X}[\pi]$ makes sense for an arbitrary polynomial π, not necessary for a convex one. What is the projection of this set onto the (t,x)-plane? The answer is surprisingly nice: this is the convex hull of the epigraph of the polynomial π!

Exercise 4.29. *Let $\pi(x) = \pi_0 + \pi_1 x + \cdots + \pi_{2k} x^{2k}$ with $\pi_{2k} > 0$, and let*

$$G[\pi] = \mathrm{Conv}\{(t,x) \in \mathbf{R}^2 \mid t \geq p(x)\}$$

be the convex hull of the epigraph of π (the set of all convex combinations of points from the epigraph of π).

1. *Prove that $G[\pi]$ is a closed convex set.*
2. *Prove that*

$$G[\pi] = \mathcal{X}[\pi].$$

4.10.3 Lovasz capacity number and semidefinite relaxations of combinatorial problems

Recall that the Lovasz capacity number $\Theta(\Gamma)$ of an n-node graph Γ is the optimal value of the following semidefinite program:

$$\min_{\lambda, x} \{\lambda : \lambda I_n - \mathcal{L}(x) \succeq 0\}, \tag{L}$$

where the symmetric $n \times n$ matrix $\mathcal{L}(x)$ is defined as follows:

- The dimension of x is equal to the number of arcs in Γ, and the coordinates of x are indexed by these arcs.
- The element of $\mathcal{L}(x)$ in an empty cell ij (one for which the nodes i and j are not linked by an arc in Γ) is 1.
- The elements of $\mathcal{L}(x)$ in a pair of symmetric nonempty cells ij, ji (those for which the nodes i and j are linked by an arc) are equal to the coordinate of x indexed by the corresponding arc.

As we remember, the importance of $\Theta(\Gamma)$ comes from the fact that $\Theta(\Gamma)$ is a computable upper bound on the stability number $\alpha(\Gamma)$ of the graph. We have seen also that the Shor semidefinite relaxation of the problem of finding the stability number of Γ leads to a seemingly stronger upper bound on $\alpha(\Gamma)$, namely, the optimal value $\sigma(\Gamma)$ in the semidefinite program

$$\min_{\lambda,\mu,\nu}\left\{\lambda : \begin{pmatrix} \lambda & -\frac{1}{2}(e+\mu)^T \\ -\frac{1}{2}(e+\mu) & A(\mu,\nu)\end{pmatrix} \succeq 0\right\}, \tag{Sh}$$

where $e = (1,\dots,1)^T \in \mathbf{R}^n$ and $A(\mu,\nu)$ is the matrix as follows:

- The dimension of ν is equal to the number of arcs in Γ, and the coordinates of ν are indexed by these arcs.
- The diagonal entries of $A(\mu,\nu)$ are $\mu_1,\dots,\mu_n$.
- The off-diagonal entries of $A(\mu,\nu)$ corresponding to empty cells are zeros.
- The off-diagonal entries of $A(\mu,\nu)$ in a pair of symmetric nonempty cells ij, ji are equal to the coordinate of ν indexed by the corresponding arc.

We have seen that (L) can be obtained from (Sh) when the variables μ_i are set to 1, so that $\sigma(\Gamma) \le \Theta(\Gamma)$. Thus,

$$\alpha(\Gamma) \le \sigma(\Gamma) \le \Theta(\Gamma). \tag{4.10.122}$$

Exercise 4.30. 1. *Prove that if* (λ,μ,ν) *is a feasible solution to* (Sh)*, then there exists a symmetric* $n\times n$ *matrix* A *such that* $\lambda I_n - A \succeq 0$*, and at the same time the diagonal entries of* A *and the off-diagonal entries in the empty cells are* ≥ 1*. Derive from this observation that the optimal value in* (Sh) *is not less than the optimal value* $\Theta'(\Gamma)$ *in the following semidefinite program:*

$$\min_{\lambda,x}\left\{\lambda : \lambda I_n - X \succeq 0,\ X_{ij} \ge 1 \text{ whenever } i,\ j \text{ are not adjacent in } \Gamma\right\}. \tag{Sc}$$

2. *Prove that* $\Theta'(\Gamma) \ge \alpha(\Gamma)$.

Hint. Demonstrate that if all entries of a symmetric $k\times k$ matrix are ≥ 1, then the maximum eigenvalue of the matrix is at least k. Derive from this observation and the interlacing eigenvalues theorem (Exercise 4.5.2) that if a symmetric matrix contains a principal $k\times k$ submatrix with entries ≥ 1, then the maximum eigenvalue of the matrix is at least k.

The upper bound $\Theta'(\Gamma)$ on the stability number of Γ is called the Schrijver capacity of graph Γ. Note that we have

$$\alpha(\Gamma) \le \Theta'(\Gamma) \le \sigma(\Gamma) \le \Theta(\Gamma).$$

Table 4.1.

Graph #	# of nodes	α	Θ'	σ	Θ
1	20	?	4.373	4.378	4.378
2	20	?	5.062	5.068	5.068
3	20	?	4.383	4.389	4.389
4	20	?	4.216	4.224	4.224
5	13	?	4.105	4.114	4.114
6	20	?	5.302	5.312	5.312
7	20	?	6.105	6.115	6.115
8	20	?	5.265	5.280	5.280
9	9	3	3.064	3.094	3.094
10	12	4	4.197	4.236	4.236
11	8	3	3.236	3.302	3.302
12	12	4	4.236	4.338	4.338
13	10	3	3.236	3.338	3.338

A natural question is, Which inequalities in this chain may be strict? To answer it, we have computed the quantities in question for about 2000 random graphs with the number of nodes varying from 8 to 20. In our experiments, the stability number was computed—by brute force—for graphs with ≤ 12 nodes. For all these graphs, the integral part of $\Theta(\Gamma)$ was equal to $\alpha(\Gamma)$. Furthermore, $\Theta(\Gamma)$ was noninteger in 156 of our 2000 experiments, and in 27 of these 156 cases the Schrijver capacity number $\Theta'(\Gamma)$ was strictly less than $\Theta(\Gamma)$. The quantities $\Theta'(\cdot)$, $\sigma(\cdot)$, $\Theta(\cdot)$ for 13 of these 27 cases are listed in Table 4.1.

Exercise 4.31. *Compute the stability numbers of graphs* 8 *and* 13*—see Fig.* 4.18.

Exercise 4.32. *Prove that* $\sigma(\Gamma) = \Theta(\Gamma)$.

The chromatic number $\xi(\Gamma)$ of a graph Γ is the minimal number of colors such that one can color the nodes of the graph in such a way that no two adjacent (i.e., linked by an arc) nodes get the same color.[42] The complement $\bar{\Gamma}$ of a graph Γ is the graph with the same set of nodes, and two distinct nodes in $\bar{\Gamma}$ are linked by an arc if and only if they are not linked by an arc in Γ.

Lovasz proved that for every graph

$$\Theta(\Gamma) \leq \xi(\bar{\Gamma}), \qquad (*)$$

[42] For example, when coloring a geographic map, it is convenient not to use the same color for a pair of countries with a common border. It was observed that to meet this requirement for actual maps, four colors are sufficient. The famous "four-color" conjecture claims that this is so for every geographic map. Mathematically, you can represent a map by a graph, where the nodes represent the countries, and two nodes are linked by an arc if and only if the corresponding countries have a common border. A characteristic feature of such a graph is that it is planar—you may draw it on a 2D plane in such a way that the arcs will not cross each other, meeting only at the nodes. Thus, the mathematical form of the four-color conjecture is that the chromatic number of any planar graph is at most 4. This is indeed true, but it took about 100 years to prove the conjecture!

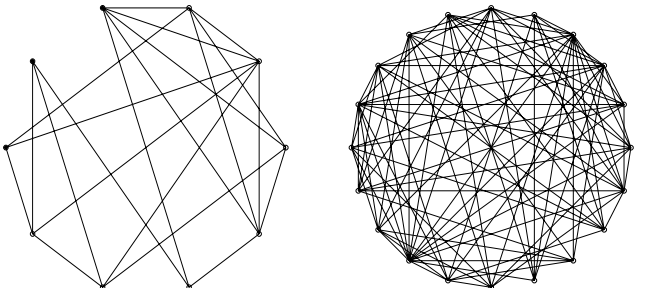


Figure 4.18. *Graphs 13 (left) and 8 (right); all nodes are on circumferences.*

so that

$$\alpha(\Gamma) \leq \Theta(\Gamma) \leq \xi(\bar{\Gamma})$$

(Lovasz's sandwich theorem).

Exercise 4.33. *Prove* (*).

Hint. Let us color the vertices of Γ in $k = \xi(\bar{\Gamma})$ colors in such a way that no two vertices of the same color are adjacent in $\bar{\Gamma}$, i.e., every two nodes of the same color are adjacent in Γ. Set $\lambda = k$, and let x be such that

$$[\mathcal{L}(x)]_{ij} = \begin{cases} -(k-1), & i \neq j,\ i, j \text{ are of the same color,} \\ 1 & \text{otherwise.} \end{cases}$$

Prove that (λ, x) is a feasible solution to (L). □

Now let us switch from the Lovasz capacity number to semidefinite relaxations of combinatorial problems, specifically to those of maximizing a quadratic form over the vertices of the unit cube, and over the entire cube:

$$\begin{array}{ll} \text{(a)} & \max\limits_x \left\{x^T A x : x \in \mathrm{Vrt}(C_n) = \{x \in \mathbf{R}^n \mid x_i = \pm 1\ \forall i\}\right\}, \\ \text{(b)} & \max\limits_x \left\{x^T A x : x \in C_n = \{x \in \mathbf{R}^n \mid -1 \leq x_i \leq 1,\ \forall i\}\right\}. \end{array} \tag{4.10.123}$$

The standard semidefinite relaxations of the problems are, respectively, the problems

$$\begin{array}{ll} \text{(a)} & \max\limits_X \{\mathrm{Tr}(AX) : X \succeq 0, X_{ii} = 1,\ i = 1, \ldots, n\}, \\ \text{(b)} & \max\limits_X \{\mathrm{Tr}(AX) : X \succeq 0, X_{ii} \leq 1,\ i = 1, \ldots, n\}; \end{array} \tag{4.10.124}$$

the optimal value of a relaxation is an upper bound for the optimal value of the respective original problem.

Exercise 4.34. *Let $A \in \mathbf{S}^n$. Prove that*

$$\max_{x: x_i = \pm 1,\, i = 1, \ldots, n} x^T A x \geq \mathrm{Tr}(A).$$

Develop an efficient algorithm that, given A, generates a point x with coordinates ± 1 such that $x^T Ax \geq \mathrm{Tr}(A)$.

Exercise 4.35. *Prove that if the diagonal entries of A are nonnegative, then the optimal values in* (4.10.124)(a) *and* (4.10.124)(b) *are equal to each other. Thus, in the case in question, the relaxations do not understand whether we are maximizing over the vertices of the cube or over the entire cube.*

Exercise 4.36. *Prove that the problems dual to* (4.10.124)(a),(b) *are, respectively,*

$$\begin{array}{ll} \text{(a)} & \min\limits_{\Lambda} \{\mathrm{Tr}(\Lambda) : \Lambda \succeq A,\ \Lambda \textit{ is diagonal}\}, \\ \text{(b)} & \min\limits_{\Lambda} \{\mathrm{Tr}(\Lambda) : \Lambda \succeq A,\ \Lambda \succeq 0,\ \Lambda \textit{ is diagonal}\}. \end{array} \tag{4.10.125}$$

The optimal values in these problems are equal to those of the respective problems in (4.10.124) *and are therefore upper bounds on the optimal values of the respective combinatorial problems from* (4.10.123)*.*

The latter claim is quite transparent, since the problems (4.10.125) can be obtained as follows:

• To bound from above the optimal value of a quadratic form $x^T Ax$ on a given set S, we look at those quadratic forms $x^T \Lambda x$ that can be easily maximized over S. For the case of $S = \mathrm{Vrt}(C_n)$ these are quadratic forms with diagonal matrices Λ, and for the case of $S = C_n$ these are quadratic forms with diagonal and positive semidefinite matrices Λ; in both cases, the respective maxima are merely $\mathrm{Tr}(\Lambda)$.

• Having specified a family $\mathcal{F}$ of quadratic forms $x^T \Lambda x$ easily optimizable over S, we then look at those forms from $\mathcal{F}$ that majorate everywhere the original quadratic form $x^T Ax$, and take among these forms the one with the minimal $\max_{x \in S} x^T \Lambda x$, thus coming to the problem

$$\min_{\Lambda} \left\{ \max_{x \in S} x^T \Lambda x : \Lambda \succeq A,\ \Lambda \in \mathcal{F} \right\}. \tag{!}$$

It is evident that the optimal value in this problem is an upper bound on $\max_{x \in S} x^T Ax$. It is also immediately seen that in the case of $S = \mathrm{Vrt}(C_n)$ the problem (!), with $\mathcal{F}$ specified as the set $\mathcal{D}$ of all diagonal matrices, is equivalent to (4.10.125)(a), while in the case of $S = C_n$ (!), with $\mathcal{F}$ specified as the set $\mathcal{D}_+$ of positive semidefinite diagonal matrices, is nothing but (4.10.125)(b).

Given the direct and quite transparent road leading to (4.10.125)(a),(b), we can try to move a little bit further along this road. To this end observe that there are trivial upper bounds on the maximum of an arbitrary quadratic form $x^T \Lambda x$ over $\mathrm{Vrt}(C_n)$ and C_n, specifically:

$$\max_{x \in \mathrm{Vrt}(C_n)} x^T \Lambda x \leq \mathrm{Tr}(\Lambda) + \sum_{i \neq j} |\Lambda_{ij}|, \quad \max_{x \in C_n} x^T \Lambda x \leq \sum_{i,j} |\Lambda_{ij}|.$$

For the above families $\mathcal{D}$, $\mathcal{D}_+$ of matrices Λ for which $x^T \Lambda x$ is easily optimizable over $\mathrm{Vrt}(C_n)$, respectively, C_n, the above bounds are equal to the precise values of the respective maxima. Now let us update (!) as follows: we eliminate the restriction $\Lambda \in \mathcal{F}$, replacing

simultaneously the objective $\max_{x\in S} x^T\Lambda x$ with its upper bound, thus coming to the pair of problems

$$\begin{array}{lll} \text{(a)} & \min\limits_{\Lambda}\left\{\mathrm{Tr}(\Lambda)+\sum\limits_{i\neq j}|\Lambda_{ij}| : \Lambda\succeq A\right\} & [S=\mathrm{Vrt}(C_n)], \\ \text{(b)} & \min\limits_{\Lambda}\left\{\sum\limits_{i,j}|\Lambda_{ij}| : \Lambda\succeq A\right\} & [S=C_n]. \end{array} \tag{4.10.126}$$

From the origin of the problems it is clear that they still yield upper bounds on the optimal values of the respective problems (4.10.123)(a),(b) and that these bounds are at least as good as the bounds yielded by the standard relaxations (4.10.124)(a),(b):

$$\begin{array}{ll} \text{(a)} & \mathrm{Opt}(4.10.123.a)\le \mathrm{Opt}(4.10.126.a)\underbrace{\le}_{(*)}\mathrm{Opt}(4.10.124.a)=\mathrm{Opt}(4.10.125.a), \\ \text{(b)} & \mathrm{Opt}(4.10.123.b)\le \mathrm{Opt}(4.10.126.b)\underbrace{\le}_{(**)}\mathrm{Opt}(4.10.124.b)=\mathrm{Opt}(4.10.125.b), \end{array} \tag{4.10.127}$$

where $\mathrm{Opt}(\cdot)$ means the optimal value of the corresponding problem.

Indeed, consider the problem (4.10.126)(a). Whenever Λ is a feasible solution of this problem, the quadratic form $x^T\Lambda x$ majorates everywhere the form x^TAx, so that $\max_{x\in\mathrm{Vrt}(C_n)} x^TAx \le \max_{x\in\mathrm{Vrt}(C_n)} x^T\Lambda x$; the latter quantity, in turn, is majorated by $\mathrm{Tr}(\Lambda)+\sum_{i\neq j}|\Lambda_{ij}|$, whence the value of the objective of the problem (4.10.126)(a) at every feasible solution of the problem majorates the quantity $\max_{x\in\mathrm{Vrt}(C_n)} x^TAx$. Thus, the optimal value in (4.10.126)(a) is an upper bound on the maximum of x^TAx over the vertices of the cube C_n. At the same time, when passing from the (dual form of the) standard relaxation (4.10.125)(a) to our new bounding problem (4.10.126)(a), we only extend the feasible set and do not vary the objective at the old feasible set; as a result of such a modification, the optimal value may only decrease. Thus, the upper bound on the maximum of x^TAx over $\mathrm{Vrt}(C_n)$ yielded by (4.10.126)(a) is at least as good as those (equal to each other) bounds yielded by the standard relaxations (4.10.124)(a), (4.10.125)(a), as required in (4.10.127)(a). Similar reasoning proves (4.10.127)(b).

Note that problems (4.10.126) are equivalent to semidefinite programs and thus are of the same status of computational tractability as the standard SDP relaxations (4.10.125) of the combinatorial problems in question. At the same time, our new bounding problems are more difficult than the standard SDP relaxations. Can we justify this by getting an improvement in the quality of the bounds?

Exercise 4.37. *Find out whether the problems* (4.10.126)(a),(b) *yield better bounds than the respective problems* (4.10.125)(a),(b)*, i.e., whether the inequalities* (*), (**) *in* (4.10.127) *can be strict.*

Hint. Look at the problems dual to (4.10.126)(a),(b).

Exercise 4.38. *Let D be a given subset of $\mathbf{R}^n_+$. Consider the following pair of optimization problems:*

$$\max_x \left\{x^T Ax : (x_1^2, x_2^2, \ldots, x_n^2)^T \in D\right\}, \qquad (P)$$
$$\max_X \{\mathrm{Tr}(AX) : X \succeq 0, \mathrm{Dg}(X) \in D\} \qquad (R)$$

($\mathrm{Dg}(X)$ is the diagonal of a square matrix X). Note that when $D = \{(1, \ldots, 1)^T\}$, (P) is the problem of maximizing a quadratic form over the vertices of C_n, while (R) is the standard semidefinite relaxation of (P). When $D = \{x \in \mathbf{R}^n \mid 0 \le x_i \le 1 \ \forall i\}$, (P) is the problem of maximizing a quadratic form over the cube C_n, and (R) is the standard semidefinite relaxation of the latter problem.

1. *Prove that if D is SDr, then (R) can be reformulated as a semidefinite program.*
2. *Prove that (R) is a relaxation of (P), i.e., that*

$$\mathrm{Opt}(P) \le \mathrm{Opt}(R).$$

3. (Nesterov, Ye) *Let $A \succeq 0$. Prove that then*

$$\mathrm{Opt}(P) \le \mathrm{Opt}(R) \le \frac{\pi}{2}\mathrm{Opt}(P).$$

Hint. Use Nesterov's theorem (Theorem 4.3.2).

Exercise 4.39. *Let $A \in \mathbf{S}^m_+$. Prove that*

$$\max\{x^T Ax \mid x_i = \pm 1, \ i = 1, \ldots, m\} = \max\left\{\frac{2}{\pi}\sum_{i,j=1}^m a_{ij}\mathrm{asin}(X_{ij}) \mid X \succeq 0, X_{ii} = 1, \ i = 1, \ldots, m\right\}.$$

4.10.4 Lyapunov stability analysis

A natural mathematical model of a swing is the linear time invariant dynamic system

$$y''(t) = -\omega^2 y(t) - 2\mu y'(t) \qquad (S)$$

with positive ω^2 and $0 \le \mu < \omega$. (The term $2\mu y'(t)$ represents friction.) A general solution to this equation is

$$y(t) = a\cos(\omega' t + \phi_0)\exp\{-\mu t\}, \ \omega' = \sqrt{\omega^2 - \mu^2},$$

with free parameters a and ϕ_0, i.e., this is a decaying oscillation. Note that the equilibrium

$$y(t) \equiv 0$$

is stable—every solution to (S) converges to 0, along with its derivative, exponentially fast.

After stability is observed, an immediate question arises: How is it possible to swing on a swing? Everybody knows from practice that it is possible. On the other hand, since the equilibrium is stable, it looks as if it was impossible to swing, without somebody's

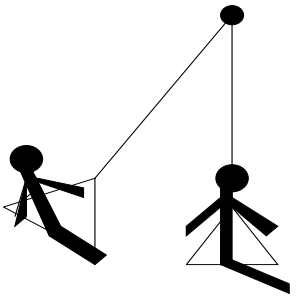


Figure 4.19. *A swinging child.*

assistance, for a long time. Swinging is possible for a highly nontrivial reason—*parametric resonance*. A swinging child does not sit on the swing in a forever-fixed position; what he does is shown in Fig. 4.19. As a result, the effective length of the swing l—the distance from the point where the rope is fixed to the center of gravity of the system—is varying with time: $l = l(t)$. Basic mechanics says that $\omega^2 = g/l$, g being the gravity acceleration. Thus, the actual swing is a time-varying linear dynamic system,

$$y''(t) = -\omega^2(t)y(t) - 2\mu y'(t), \qquad (\mathrm{S}')$$

and it turns out that for properly varied $\omega(t)$ the equilibrium $y(t) \equiv 0$ is not stable. A swinging child is just varying $l(t)$ in a way that results in an unstable dynamic system (S′), and this instability is in fact what the child enjoys; see Fig. 4.20.

Exercise 4.40. *Assume that you are given parameters l ("nominal length of the swing rope"), $h > 0$, and $\mu > 0$, and it is known that a swinging child can vary the effective length of the rope within the bounds $l \pm h$, i.e., his or her movement is governed by the uncertain linear time-varying system*

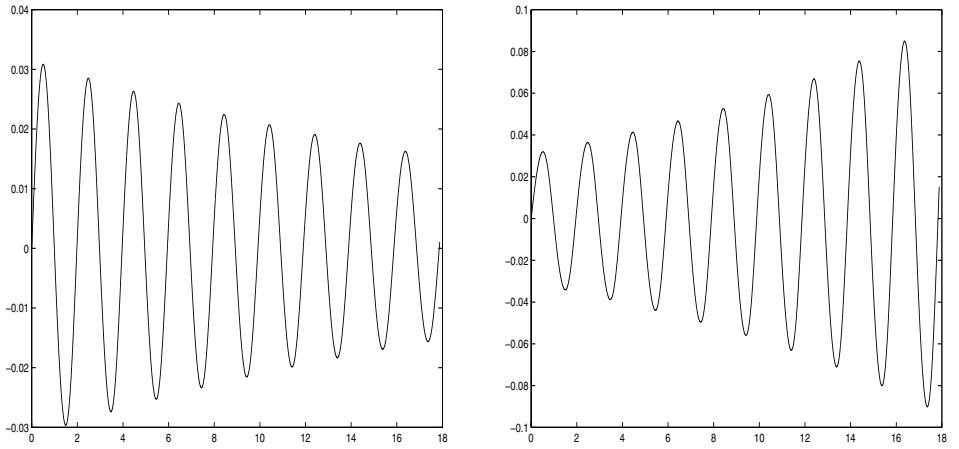
$$y''(t) = -a(t)y(t) - 2\mu y'(t), \quad a(t) \in \left[\frac{g}{l+h}, \frac{g}{l-h}\right].$$

Try to identify the domain in the 3D-space of parameters l, μ, h where the system is stable, as well as the domain where its stability can be certified by a quadratic Lyapunov function. What is the difference between these two domains?

4.10.5 $\mathcal{S}$-lemma

The $\mathcal{S}$-lemma is a kind of theorem on the alternative, more specifically, a quadratic analog of the homogeneous Farkas lemma:

> *Homogeneous Farkas lemma. A homogeneous linear inequality $a^T x \geq 0$ is a consequence of a system of homogeneous linear inequalities $b_i^T x \geq 0$, $i = 1, \ldots, m$, if and only if it is a linear consequence of the system, i.e., if and only if*
>
> $$\exists(\lambda \geq 0): \quad a = \sum_i \lambda_i b_i.$$



$$y''(t) = -\frac{g}{l+h\sin(2\omega t)}y(t) - 2\mu y'(t), \quad y(0) = 0,\ y'(0) = 0.1$$

$$\left[l = 1\,[\mathrm{m}],\ g = 10\,[\tfrac{\mathrm{m}}{\mathrm{sec}^2}],\ \mu = 0.15[\tfrac{1}{\mathrm{sec}}],\ \omega = \sqrt{g/l}\right]$$

Figure 4.20. *Graph of $y(t)$. Left: $h = 0.125$; this child is too small; he should grow up. Right: $h = 0.25$; this child can already swing.*

$\mathcal{S}$-lemma. A homogeneous quadratic inequality $x^T Ax \geq 0$ is a consequence of a strictly feasible system of homogeneous quadratic inequalities $x^T B_i x \geq 0$, $i = 1, \ldots, m$, with $m = 1$ if and only if it is a linear consequence of the system and a trivial—identically true—quadratic inequality, i.e., if and only if

$$\exists(\lambda \geq 0, \Delta \succeq 0): \quad A = \sum_i \lambda_i B_i + \Delta.$$

We see that the $\mathcal{S}$-lemma is indeed similar to the Farkas lemma, up to a (severe!) restriction that now the system in question must contain a single quadratic inequality (and up to the mild "regularity assumption" of strict feasibility).

The homogeneous Farkas lemma gives rise to the theorem on the alternative for systems of linear inequalities. As a matter of fact, this lemma is the basis of the entire convex analysis and the reason why convex programming problems are easy (see Lecture 5). The fact that a similar statement for quadratic inequalities—i.e., $\mathcal{S}$-lemma—fails to be true for a multi-inequality system is very unpleasant and finally is the reason for the existence of simple-looking computationally intractable (NP-complete) optimization problems.

Given the crucial role of different theorems on the alternative in optimization, it is vitally important to understand the extent to which the linear theorem on the alternative can be generalized onto the case of nonlinear inequalities. The standard generalization of this type is as follows:

Lagrange duality theorem (LDT). Let f_0 be a convex function and $f_1, \ldots, f_m$ be concave functions on $\mathbf{R}^n$ such that the system of inequalities

$$f_i(x) \geq 0, \ i = 1, \ldots, m, \tag{S}$$

is strictly feasible (i.e., $f_i(\bar{x}) > 0$ for some $\bar{x}$ and all $i = 1, \ldots, m$). The inequality

$$f_0(x) \geq 0$$

is a consequence of the system (S) *if and only if it can be obtained, in a linear fashion, from* (S) *and a trivially true—valid on the entire* $\mathbf{R}^n$*—inequality, i.e., if and only if there exist m nonnegative weights* λ_i *such that*

$$f_0(x) \geq \sum_{i=1}^{m} \lambda_i f_i(x) \quad \forall x.$$

The LDT plays the central role in computationally tractable optimization, i.e., in convex programming. (For example, the conic duality theorem from Lecture 2 is just a reformulation of the LDT.) This theorem, however, imposes severe convexity-concavity restrictions on the inequalities in question. In the case when all the inequalities are homogeneous quadratic, LDT is empty. Indeed, a homogeneous quadratic function $x^T Bx$ is concave if and only if $B \preceq 0$ and is convex if and only if $B \succeq 0$. It follows that in the case of $f_i = x^T A_i x$, $i = 0, \ldots, m$, the premise in the LDT is empty (a system of homogeneous quadratic inequalities $x^T A_i x \geq 0$ with $A_i \preceq 0$, $i = 1, \ldots, m$, simply cannot be strictly feasible), and the conclusion in the LDT is trivial (if $f_0(x) = x^T A_0 x$ with $A_0 \succeq 0$, then $f_0(x) \geq \sum_{i=1}^{m} 0 \times f_i(x)$, whatever are f_i's). Comparing the $\mathcal{S}$-lemma to the LDT, we see that the former statement is, in a sense, complementary to the second one: the $\mathcal{S}$-lemma, when applicable, provides us with information that definitely cannot be extracted from the LDT. Given this unique role of the $\mathcal{S}$-lemma, it surely deserves the effort to understand the possibilities and limitations of extending the lemma to the case of a multi-inequality system, i.e., to address the question as follows:

(**SL**.?) *We are given a homogeneous quadratic inequality*

$$x^T Ax \geq 0 \tag{I}$$

along with a strictly feasible system of homogeneous quadratic inequalities

$$x^T B_i x \geq 0, \quad i = 1, \ldots, m. \tag{S}$$

Consider the following two statements:

(i) (I) *is a consequence of* (S)*, i.e.,* (I) *is satisfied at every solution of* (S)*.*

(ii) (I) *is a linear consequence of* (S) *and a trivial—identically true—homogeneous quadratic inequality:*

$$\exists(\lambda \geq 0, \Delta \succeq 0): \quad A = \sum_{i=1}^{m} \lambda_i B_i + \Delta.$$

What is the gap between (i) *and* (ii)*?*

One obvious fact is expressed in the following exercise.

Exercise 4.41. Inverse $\mathcal{S}$-lemma. *Prove the implication* (ii)⇒(i).

In what follows, we focus on less trivial results concerning the aforementioned gap.

Exercise 4.42. *Prove the following.*

Proposition 4.10.1. Inhomogeneous $\mathcal{S}$-lemma. *Let*

$$\begin{array}{rcll} f(x) & = & x^T Ax + 2a^T x + \alpha & [A = A^T], \\ g(x) & = & x^T Bx + 2b^T x + \beta & [B = B^T] \end{array} \tag{4.10.128}$$

be two quadratic forms such that the inequality $f(x) \leq 0$ is strictly feasible. Then the implication

$$f(x) \leq 0 \Rightarrow g(x) \leq 0$$

holds true if and only if there exists $\lambda \geq 0$ such that

$$g(x) \leq \lambda f(x) \quad \forall x$$

or, which is the same, if and only if there exists $\lambda \geq 0$ such that

$$\begin{pmatrix} \lambda\alpha & [\lambda a - b]^T \\ \lambda a - b & \lambda A - B \end{pmatrix} \succeq 0.$$

Under the additional assumption that the inequality $g(x) \geq 0$ is strictly feasible, the quantity λ in the above "if and only if" is strictly positive.

Hint. The "if" part is evident. To prove the "only if" part, act as follows.
1. Verify that there exist a sequence $\{\gamma_i > 0\}$ and $\delta > 0$ such that

   (i) $\gamma_i \to 0, i \to \infty$;

   (ii) all the matrices $A_i \equiv A + \gamma_i I$ are nonsingular; and

   (iii) $\bar{x}^T A_i \bar{x} + 2a^T \bar{x} + \alpha \leq -\delta \ \forall i$.

2. Assuming the validity of the implication $f(x) \leq 0 \Rightarrow g(x) \leq 0$ and setting

$$f_i(x) = x^T A_i x + 2a^T x + \alpha,$$

prove that for every i one has

$$\forall y: \quad f_i(y - A_i^{-1}a) \leq 0 \Rightarrow g(y - A_i^{-1}a) \leq 0,$$

i.e., that

$$\forall y: \quad y^T A_i y + [\alpha - a^T A_i^{-1} a] \leq 0 \Rightarrow g(y - A_i^{-1}a) \equiv y^T By + 2b_i^T y + \beta_i \leq 0.$$

Derive from the latter fact that

$$\forall(y,t): \quad y^T A_i y + [\alpha - a^T A_i^{-1} a]t^2 \le 0 \Rightarrow y^T B y + 2t b_i^T y + \beta_i t^2 \le 0,$$

whence, by the usual $\mathcal{S}$-lemma,

$$\exists \lambda \ge 0: \quad g(x) \le \lambda_i f_i(x).$$

Prove that the sequence $\{\lambda_i\}$ is bounded, and look at limiting points of this sequence.

In view of Proposition 4.10.1, if $f(\cdot)$, $g(\cdot)$ are as in (4.10.128) and the inequality $f(x) < 0$ is solvable, then the inclusion

$$\{x \mid f(x) \le 0\} \subset \{x \mid g(x) \le 0\} \tag{4.10.129}$$

holds true if and only if the system of LMIs

$$\lambda \ge 0, \quad \lambda F - G \equiv \begin{pmatrix} \lambda\alpha & [\lambda a - b]^T \\ \lambda a - b & \lambda A - B \end{pmatrix} \succeq 0 \tag{4.10.130}$$

in a single variable λ is solvable.

Assuming that we can check efficiently whether a system of LMIs is feasible and we are able to find a solution of this system when the system is feasible,[43] we see that the inclusion (4.10.129) admits an efficient verification. In some cases we meet with the following situation: we have checked whether the system (4.10.130) is feasible and have found that it is not the case, i.e., we know that the inclusion (4.10.129) is not valid. How could we efficiently build an evident certificate of this conclusion, i.e., a point x_* such that $f(x_*) \le 0$ and $g(x_*) > 0$?

Exercise 4.43. *Let f, g be as in* (4.10.1)*, let $f(\bar{x}) < 0$ for certain $\bar{x}$, and let the system* (4.10.130) *be infeasible.*

1. *Prove that the system of LMIs*

$$X \succeq 0, \ \mathrm{Tr}(XF) \le -1, \ \mathrm{Tr}(XG) \ge 1 \tag{4.10.131}$$

is feasible.

2. *Let $X_* = D_* D_*^T$ be a solution of* (4.10.131)*. Set*

$$\widehat{F} = D_*^T F D_*, \quad \widehat{G} = D_*^T G D_*$$

and consider the eigenvalue decomposition of $\widehat{G}$

$$\widehat{G} = U^T \widetilde{G} U$$

with diagonal $\widetilde{G}$ and orthogonal U. Verify that

$$\mathrm{Tr}(\underbrace{U \widehat{F} U^T}_{\widetilde{F}}) \le -1, \quad \mathrm{Tr}(\widetilde{G}) \ge 1. \tag{4.10.132}$$

[43] This assumption nearly fits the reality; see Lecture 5.

3. *Build an algorithm that allows one to find efficiently a vector ξ_* with coordinates ± 1 such that*

$$\xi_*^T \widetilde{F} \xi_* \le \mathrm{Tr}(\widetilde{F}).$$

Verify that the vector $\eta_* = D_* U^T \xi_*$ satisfies the relation

$$\eta_*^T F \eta_* \le -1, \ \ \eta_*^T G \eta_* \ge 1.$$

Think how to convert the vector η_* efficiently to a vector x_* such that $f(x_*) \le 0$, $g(x_*) > 0$.

Straightforward proof of the standard $\mathcal{S}$-lemma. The goal of the subsequent exercises is to work out a straightforward proof of the $\mathcal{S}$-lemma instead of the tricky, although elegant, proof presented in Lecture 4. The "if" part of the lemma is evident, and we focus on the "only if" part. Thus, we are given two quadratic forms $x^T Ax$ and $x^T Bx$ with symmetric matrices A, B such that $\bar{x}^T A \bar{x} > 0$ for some $\bar{x}$ and the implication

$$x^T Ax \ge 0 \Rightarrow x^T Bx \ge 0 \qquad (\Rightarrow)$$

is true. Our goal is to prove that

(**SL**.A) *There exists $\lambda \ge 0$ such that $B \succeq \lambda A$.*

The main tool we need is the following theorem.

Theorem 4.10.1. General Helley theorem. *Let $\{A_\alpha\}_{\alpha \in I}$ be a family of closed convex sets in $\mathbf{R}^n$ such that*

1. *every $n+1$ sets from the family have a point in common;*
2. *there is a finite subfamily of the family such that the intersection of the sets from the subfamily is bounded.*

Then all sets from the family have a point in common.

Exercise 4.44. *Prove the general Helley theorem.*

Exercise 4.45. *Show that* (**SL**.A) *is a corollary of the following statement:*

(**SL**.B) Let $x^T Ax$, $x^T Bx$ be two quadratic forms such that $\bar{x}^T A \bar{x} > 0$ for certain $\bar{x}$ and

$$x^T Ax \ge 0, x \ne 0 \Rightarrow x^T Bx > 0. \qquad (\Rightarrow')$$

Then there exists $\lambda \ge 0$ such that $B \succeq \lambda A$.

Exercise 4.46. *Given data A, B satisfying the premise of* (**SL**.B)*, define the sets*

$$Q_x = \{\lambda \ge 0 : x^T Bx \ge \lambda x^T Ax\}.$$

1. *Prove that every one of the sets Q_x is a closed nonempty convex subset of the real line.*
2. *Prove that at least one of the sets Q_x is bounded.*
3. *Prove that every two sets $Q_{x'}$, $Q_{x''}$ have a point in common.*
4. *Derive* (**SL**.B) *from* 1–3*, thus concluding the proof of the $\mathcal{S}$-lemma.*

$\mathcal{S}$-lemma with a multi-inequality premise. The goal of the subsequent exercises is to present a number of cases when, under appropriate additional assumptions on the data (I), (S), of the question (**SL**.?), statements (i) and (ii) are equivalent, even if the number m of homogeneous quadratic inequalities in (S) is > 1.

Our first exercise demonstrates that certain additional assumptions are definitely necessary.

Exercise 4.47. *Demonstrate by example that if x^TAx, x^TBx, x^TCx are three quadratic forms with symmetric matrices such that*

$$\begin{array}{c}\exists \bar{x} : \bar{x}^TA\bar{x} > 0, \bar{x}^TB\bar{x} > 0,\\ x^TAx \geq 0, x^TBx \geq 0 \Rightarrow x^TCx \geq 0,\end{array} \tag{4.10.133}$$

then not necessarily there exist $\lambda, \mu \geq 0$ such that $C \succeq \lambda A + \mu B$.

Hint. Clearly there do not exist nonnegative λ, μ such that $C \succeq \lambda A + \mu B$ when

$$\mathrm{Tr}(A) \geq 0, \quad \mathrm{Tr}(B) \geq 0, \quad \mathrm{Tr}(C) < 0. \tag{4.10.134}$$

Thus, to build the required example it suffices to find A, B, C satisfying both (4.10.133) and (4.10.134).

Seemingly the simplest way to ensure (4.10.133) is to build 2×2 matrices A, B, C such that the associated quadratic forms $f_A(x) = x^TAx$, $f_B(x) = x^TBx$, $f_C(x) = x^TCx$ are as follows:

- The set $X_A = \{x \mid f_A(x) \geq 0\}$ is the union of an angle D symmetric with regard to the x_1-axis and the angle $-D$: $f_A(x) = \lambda^2x_1^2 - x_2^2$ with $\lambda > 0$.
- The set $X_B = \{x \mid f_B(x) \geq 0\}$ looks like a clockwise rotation of X_A by a small angle: $f_B(x) = (\mu x - y)(\nu x + y)$ with $0 < \mu < \lambda$ and $\nu > \lambda$.
- The set $X_C = \{x \mid x^TCx \geq 0\}$ is the intersection of X_A and X_B: $f_C(x) = (\mu x - y)(\nu x + y)$.

Surprisingly, there exists a semiextension of the $\mathcal{S}$-lemma to the case of $m = 2$ in (**SL**.?):

(**SL**.C) *Let $n \geq 3$, and let A, B, C be three symmetric $n \times n$ matrices such that*

(i) *a certain linear combination of the matrices A, B, C is positive definite, and*

(ii) *the system of inequalities*

$$\left\{\begin{array}{l} x^TAx \geq 0,\\ x^TBx \geq 0\end{array}\right. \tag{4.10.135}$$

is strictly feasible, i.e., $\exists \bar{x}$: $\bar{x}^TA\bar{x} > 0$, $\bar{x}^TB\bar{x} > 0$.

Then the inequality

$$x^TCx \geq 0$$

is a consequence of the system (4.10.135) *if and only if there exist nonnegative* λ, μ *such that*

$$C \succeq \lambda A + \mu B.$$

The proof of (**SL**.C) uses a nice convexity result that is interesting by its own right:

(**SL**.D) (B. T. Polyak) *Let* $n \geq 3$*, and let* $f_i(x) = x^T A_i x$*,* $i = 1, 2, 3$*, be three homogeneous quadratic forms on* $\mathbf{R}^n$ *(here* A_i*,* $i = 1, 2, 3$*, are symmetric* $n \times n$ *matrices). Assume that a certain linear combination of the matrices* A_i *is positive definite. Then the image of* $\mathbf{R}^n$ *under the mapping*

$$F(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ f_3(x) \end{pmatrix}$$

is a closed convex set.

Exercise 4.48. *Derive* (**SL**.C) *from* (**SL**.D).

Hint. For the only nontrivial "only if" part of (**SL**.C): By (**SL**.C)(i) and (**SL**.D), the set

$$Y = \left\{ y \in \mathbf{R}^3 \mid \exists x : y = \begin{pmatrix} x^T A x \\ x^T B x \\ x^T C x \end{pmatrix} \right\}$$

is a closed and convex set in $\mathbf{R}^3$. Prove that if $x^T B x \geq 0$ for every solution of (4.10.135), then Y does not intersect the convex set $Z = \{(y = (y_1, y_2, y_3)^T \mid y_1, y_2 \geq 0, y_3 < 0\}$. Applying the separation theorem, conclude that there exist nonnegative weights θ, λ, μ, not all of them zero, such that the matrix $\theta C - \lambda A - \mu B$ is positive definite. Use (**SL**.C)(ii) to demonstrate that $\theta > 0$.

Now let us prove (**SL**.D). We start with a number of simple topological facts. Recall that a metric space X is called connected if there does not exist a pair of nonempty open sets $V, U \subset X$ such that $U \cap V = \emptyset$ and $U \cup V = X$. The simplest facts about connectivity are as follows:

(C.1) If a metric space Y is linearly connected: for every two points $x, y \in Y$ there exists a continuous curve linking x and y, i.e., a continuous function $\gamma : [0, 1] \to Y$ such that $\gamma(0) = x$ and $\gamma(1) = y$, then Y is connected. In particular, a line segment in $\mathbf{R}^k$, same as every other convex subset of $\mathbf{R}^k$ is connected. (From now on, a set $Y \subset \mathbf{R}^k$ is treated as a metric space, the metric coming from the standard metric on $\mathbf{R}^k$.)

(C.2) Let $F : Y \to Z$ be a continuous mapping from a connected metric space to a metric space Z. Then the image $F(Y)$ of the mapping (regarded as a metric space, the metric coming from Z) is connected.

We see that the connectivity of a set $Y \in \mathbf{R}^n$ is a much weaker property than the convexity. There is, however, a simple case where these properties are equivalent: the one-dimensional case $k = 1$.

Exercise 4.49. *Prove that a set $Y \subset \mathbf{R}$ is connected if and only if it is convex.*

To proceed, recall the notion of the n-dimensional projective space $\mathbf{P}^n$. A point in this space is a line in $\mathbf{R}^{n+1}$ passing through the origin. To define the distance between two points of this type, i.e., between two lines ℓ, ℓ' in $\mathbf{R}^{n+1}$ passing through the origin, we take the intersections of the lines with the unit Euclidean sphere in $\mathbf{R}^{n+1}$; let the first intersection comprise the points $\pm e$ and the second the points $\pm e'$. The distance between ℓ and ℓ' is, by definition, $\min\{\|e + e'\|_2, \|e - e'\|_2\}$. (It is clear that the resulting quantity is well-defined and that it is a metric.) Note that there exists a natural mapping Φ (the canonical projection) of the unit sphere $S^n \subset \mathbf{R}^{n+1}$ onto $\mathbf{P}^n$—the mapping that maps a unit vector $e \in S^n$ onto the line spanned by e. It is immediately seen that this mapping is continuous and maps points $\pm e$, $e \in S^n$, onto the same point of $\mathbf{P}^n$. In what follows we will make use of the following simple facts.

Proposition 4.10.2. *Let $Y \subset S^n$ be a set with the following property: for every two points $x, x' \in Y$ there exists a point $y \in Y$ such that both x, x' can be linked by continuous curves in Y with the set $\{y; -y\}$ (i.e., we can link in Y both x and x' with y, or x with y, and x' with $-y$, or both x and x' with $-y$, or x with $-y$, and x' with y). Then the set $\Phi(Y) \subset \mathbf{P}^n$ is linearly connected (and thus connected).*

Proposition 4.10.3. *Let $F : Y \to \mathbf{R}^k$ be a continuous mapping defined on a central-symmetric subset ($Y = -Y$) of the unit sphere $S^n \subset \mathbf{R}^{n+1}$, and let the mapping be even: $F(y) = F(-y)$ for every $y \in Y$. Let $Z = \Phi(Y)$ be the image of Y in $\mathbf{P}^n$ under the canonical projection, and let the mapping $G : Z \to \mathbf{R}^k$ be defined as follows: to specify $G(z)$ for $z \in Z$, we choose somehow a point $y \in Y$ such that $\Phi(y) = z$ and set $G(z) = F(y)$. Then the mapping G is well defined and continuous on Z.*

Exercise 4.50. *Prove Proposition* 4.10.2.

Exercise 4.51. *Prove Proposition* 4.10.3.

The key argument in the proof of (**SL**.D) is the following fact.

Proposition 4.10.4. *Let $f(x) = x^T Qx$ be a homogeneous quadratic form on $\mathbf{R}^n$, $n \geq 3$. Assume that the set $Y = \{x \in S^{n-1} : f(x) = 0\}$ is nonempty. Then the set Y is central symmetric, and its image Z under the canonical projection $\Phi : S^{n-1} \to \mathbf{P}^{n-1}$ is connected.*

The goal of the next exercise is to prove Proposition 4.10.4. In what follows f, Y, Z are as in the proposition. The relation $Y = -Y$ is evident, so that all we need to prove is the connectedness of Z. Without loss of generality we may assume that

$$f(x) = \sum_{i=1}^{n} \lambda_i x_i^2, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n;$$

since Y is nonempty, we have $\lambda_1 \geq 0$, $\lambda_n \leq 0$. Replacing, if necessary, f with $-f$ (which does not vary Y), we may further assume that $\lambda_1 \geq |\lambda_n|$. The case $\lambda_1 = \lambda_n = 0$ is trivial, since in this case $f \equiv 0$, whence $Y = S^{n-1}$; thus, Y (and therefore Z; see (C.2)) is connected. Thus, we may assume that $\lambda_1 \geq |\lambda_n|$ and $\lambda_1 > 0 \geq \lambda_n$. Finally, it is convenient to set $\theta_1 = \lambda_1$, $\theta_2 = -\lambda_n$. Reordering the coordinates of x, we come to the situation as follows:

$$\begin{array}{ll} \text{(a)} & f(x) = \theta_1 x_1^2 - \theta_2 x_2^2 + \sum_{i=3}^{n} \theta_i x_i^2, \\ \text{(b)} & \theta_1 \geq \theta_2 \geq 0,\ \theta_1 + \theta_2 > 0, \\ \text{(c)} & -\theta_2 \leq \theta_i \leq \theta_1,\ i = 3, \dots, n. \end{array} \tag{4.10.136}$$

Exercise 4.52. 1. *Let $x \in Y$. Prove that x can be linked in Y by a continuous curve with a point x' such that the coordinates of x' with indices* $3, 4, \dots, n$ *vanish.*

Hint. Setting $d = (0, 0, x_3, \dots, x_n)^T$, prove that there exists a continuous curve $\mu(t)$, $0 \leq t \leq 1$, in Y such that

$$\mu(t) = (x_1(t), x_2(t), 0, 0, \dots, 0)^T + (1-t)d,\ \ 0 \leq t \leq 1,$$

and $x_1(0) = x_1$, $x_2(0) = x_2$.

2. *Prove that there exists a point $z^+ = (z_1, z_2, z_3, 0, 0, \dots, 0)^T \in Y$ such that*
(i) $z_1 z_2 = 0$.
(ii) *Given a point $u = (u_1, u_2, 0, 0, \dots, 0)^T \in Y$, you can either* (ii)(a) *link u by continuous curves in Y both to z^+ and to $\bar{z}^+ = (z_1, z_2, -z_3, 0, 0, \dots, 0)^T \in Y$, or* (ii)(b) *link u both to $z^- = (-z_1, -z_2, z_3, 0, 0, \dots, 0)^T$ and $\bar{z}^- = (-z_1, -z_2, -z_3, 0, 0, \dots, 0)^T$. (Note that $z^+ = -\bar{z}^-$, $\bar{z}^+ = -z^-$.)*

Hint. Given a point $u \in Y$, $u_3 = u_4 = \dots = u_n = 0$, build a continuous curve $\mu(t) \in Y$ of the type

$$\mu(t) = (x_1(t), x_2(t), t, 0, 0, \dots, 0)^T \in Y$$

such that $\mu(0) = u$, and look what can be linked with u by such a curve.

3. *Conclude from* 1 *and* 2 *that Y satisfies the premise of Proposition* 4.10.2, *and thus complete the proof of Proposition* 4.10.4.

Now we are ready to prove (**SL**.D).

Exercise 4.53. *Let A_i, $i = 1, 2, 3$, satisfy the premise of* (**SL**.D).
1. *Demonstrate that in order to prove* (**SL**.D), *it suffices to prove the statement in the particular case $A_1 = I$.*

Hint. The validity status of the conclusion in (**SL**.D) remains the same when we replace our initial quadratic forms $f_i(x)$, $i = 1, 2, 3$, by the forms $g_i(x) = \sum_{j=1}^{3} c_{ij} f_j(x)$, $i = 1, 2, 3$, provided that the matrix $[c_{ij}]$ is nonsingular. Taking

into account the premise in (**SL**.D), we can choose such a transformation to get, as g_1, a positive definite quadratic form. Without loss of generality we may therefore assume from the very beginning that $A_1 \succ 0$. Now, passing from the quadratic forms given by the matrices A_1, A_2, A_3 to those given by the matrices I, $A_1^{-1/2}A_2A_1^{-1/2}$, $A_1^{-1/2}A_3A_1^{-1/2}$, we do not vary the set H at all. Thus, we can restrict ourselves with the case $A_1 = I$.

2. *Assuming $A_1 = I$, prove that the set*

$$H_1 = \{(v_1, v_2)^T \in \mathbf{R}^2 \mid \exists x \in \mathbf{S}^{n-1} : v_1 = f_2(x), v_3 = f_3(x)\}$$

is convex.

Hint. Prove that the intersection of H_1 with every line $\ell \subset \mathbf{R}^2$ is the image of a connected set in $\mathbf{P}^{n-1}$ under a continuous mapping and is therefore connected by (C.2). Then apply the result of Exercise 4.49.

3. *Assuming $A_1 = I$, let $\widetilde{H}_1 = \{(1, v_1, v_2)^T \in \mathbf{R}^3 \mid (v_1, v_2)^T \in H_1\}$, and let $H = F(\mathbf{R}^n)$, $F(x) = (f_1(x), f_2(x), f_3(x))^T$. Prove that H is the closed convex hull of $\widetilde{H}_1$:*

$$H = \mathrm{cl}\{y \mid \exists t > 0, u \in \widetilde{H}_1 : y = tu\}.$$

Use this fact and the result of 2 *to prove that H is closed and convex, thus completing the proof of* (**SL**.D).

Note that the restriction $n \geq 3$ in (**SL**.D) and (**SL**.C) is essential.

Exercise 4.54. *Demonstrate by example that* (**SL**.C) *not necessarily remains valid when skipping the assumption $n \geq 3$ in the premise.*

Hint. An example can be obtained via the construction outlined in the hint to Exercise 4.47.

To extend (**SL**.C) to the case of 2×2 matrices, it suffices to strengthen the premise a bit:

(**SL**.E) *Let A, B, C be three 2×2 symmetric matrices such that*

(i) *a certain linear combination of the matrices A, B is positive definite,*

and

(ii) *the system of inequalities* (4.10.135) *is strictly feasible.*

Then the inequality

$$x^T C x \geq 0$$

is a consequence of the system (4.10.135) *if and only if there exist nonnegative λ, μ such that*

$$C \succeq \lambda A + \mu B.$$

Exercise 4.55. *Let A, B, C be three 2×2 symmetric matrices such that the system of inequalities $x^T Ax \geq 0$, $x^T Bx \geq 0$ is strictly feasible and the inequality $x^T Cx$ is a consequence of the system.*

1. *Assume that there exists a nonsingular matrix Q such that both the matrices QAQ^T and QBQ^T are diagonal. Prove that then there exist $\lambda, \mu \geq 0$ such that $C \succeq \lambda A + \mu B$.*

2. *Prove that if a linear combination of two symmetric matrices A, B (not necessarily 2×2 ones) is positive definite, then there exists a system of (not necessarily orthogonal) coordinates where both quadratic forms $x^T Ax$, $x^T Bx$ are diagonal or, equivalently, that there exists a nonsingular matrix Q such that both QAQ^T and QBQ^T are diagonal matrices. Combine this fact with* 1 *to prove* (**SL**.E).

We have seen that the two-inequality-premise version of the $\mathcal{S}$-lemma is valid (under an additional mild assumption that a linear combination of the three matrices in question is positive definite). In contrast, the three-inequality-premise version of the lemma is hopeless.

Exercise 4.56. *Consider four matrices*

$$A_1 = \begin{pmatrix} 2+\epsilon & & \\ & -1 & \\ & & -1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} -1 & & \\ & 2+\epsilon & \\ & & -1 \end{pmatrix}, \quad A_3 = \begin{pmatrix} -1 & & \\ & -1 & \\ & & 2+\epsilon \end{pmatrix},$$

$$B = \begin{pmatrix} 1 & 1.1 & 1.1 \\ 1.1 & 1 & 1.1 \\ 1.1 & 1.1 & 1 \end{pmatrix}.$$

1. *Prove that if $\epsilon > 0$ is small enough, then the matrices satisfy the conditions*

$$\begin{array}{ll} \text{(a)} & \forall (x, x^T A_i x \geq 0,\ i = 1,2,3): \quad x^T Bx \geq 0, \\ \text{(b)} & \exists \bar{x}: \quad \bar{x}^T A_i \bar{x} > 0. \end{array}$$

2. *Prove that whenever $\epsilon \geq 0$, there do not exist nonnegative λ_i, $i = 1, 2, 3$, such that $B \succeq \sum_{i=1}^3 \lambda_i A_i$.*

Thus, an attempt to extend the $\mathcal{S}$-lemma to the case of three quadratic inequalities in the premise fails already when the matrices of these three quadratic forms are diagonal.

> *Hint*. Note that if there exists a collection of nonnegative weights $\lambda_i \geq 0$, $i = 1, 2, 3$, such that $B \succeq \sum_{i=1}^3 \lambda_i A_i$, then the same property is shared by any collection of weights obtained from the original one by a permutation. Conclude that under the above "if" one should have $B \succeq \theta \sum_{i=1}^3 A_i$ with some $\theta \geq 0$, which is not the case.

Exercise 4.56 demonstrates that when considering (**SL**.?) for $m = 3$, even the assumption that all quadratic forms in (S) are diagonal not necessarily implies an equivalence between (i) and (ii). Note that under a stronger assumption that all quadratic forms in question are diagonal, (i) is equivalent to (ii) for all m.

Exercise 4.57. *Let A, $B_1, \ldots, B_m$ be diagonal matrices. Prove that the inequality*

$$x^T Ax \geq 0$$

is a consequence of the system of inequalities

$$x^T B_i x \geq 0, \ \ i = 1, \ldots, m,$$

if and only if it is a linear consequence of the system and an identically true quadratic inequality, i.e., if and only if

$$\exists(\lambda \geq 0, \Delta \succeq 0): \quad A = \sum_i \lambda_i B_i + \Delta.$$

Hint. Pass to new variables $y_i = x_i^2$ and apply the homogeneous Farkas lemma.

Relaxed versions of the $\mathcal{S}$-lemma. In Exercises 4.48–4.57 we wanted to understand under which additional assumptions on the data of (**SL**.?) we can be sure that (i) is equivalent to (ii). In the exercises to follow, we want to understand the gap between (i) and (ii) in general. An example of such a gap statement is as follows:

(**SL**.F) *Consider the situation of* (**SL**.?) *and assume that* (i) *holds. Then* (ii) *is valid on a subspace of codimension* $\leq m - 1$*, i.e., there exist nonnegative weights* λ_i *such that the symmetric matrix*

$$\Delta = A - \sum_{i=1}^m \lambda_i B_i$$

has at most $m - 1$ *negative eigenvalues (counted with their multiplicities).*

Note that in the case $m = 1$ this statement becomes exactly the $\mathcal{S}$-lemma.

The idea of the proof of (**SL**.F) is very simple. To say that (I) is a consequence of (S) is basically the same as to say that the optimal value in the optimization problem

$$\min_x \left\{ f_0(x) \equiv x^T A x : f_i(x) \equiv x^T B_i x \geq \epsilon, \ \ i = 1, \ldots, m \right\} \qquad (P_\epsilon)$$

is positive whenever $\epsilon > 0$. Assume that the problem is solvable with an optimal solution x_ϵ, and let $I_\epsilon = \{i \geq 1 \mid \ f_i(x_\epsilon) = \epsilon\}$. Assume, in addition, that the gradients $\{\nabla f_i(x_\epsilon) \mid i \in I_\epsilon\}$ are linearly independent. Then the second-order necessary optimality conditions are satisfied at x_ϵ, i.e., there exist nonnegative Lagrange multipliers λ_i^ϵ, $i \in I_\epsilon$, such that for the function

$$L_\epsilon(x) = f_0(x) - \sum_{i \in I_\epsilon} \lambda_i^\epsilon f_i(x)$$

one has

$$\begin{gathered} \nabla L_\epsilon(x_\epsilon) = 0, \\ \forall (d \in E = \{d : d^T \nabla f_i(x_\epsilon) = 0, i \in I_\epsilon\}): \quad d^T \nabla^2 L_\epsilon(x_\epsilon) d \geq 0. \end{gathered}$$

In other words, setting $D = A - \sum_{i \in I_\epsilon} \lambda_i^\epsilon B_i$, we have

$$D x_\epsilon = 0; \quad d^T D d \geq 0 \quad \forall d \in E.$$

We conclude that $d^T Dd \geq 0 \; \forall d \in E^+ = E + \mathbf{R}x_\epsilon$, and it is easily seen that the codimension of E^+ is at most $m-1$. Consequently, the number of negative eigenvalues of D, counted with their multiplicities, is at most $m-1$.

The outlined proof is, of course, incomplete: we should justify all the assumptions made along the way. This indeed can be done. (The driving force of the justification is the Sard theorem: if $f : \mathbf{R}^n \to \mathbf{R}^k$, $n \geq k$, is a C^∞ mapping, then the image under f of the set of points x where the rank of $f'(x)$ is $< k$, is of the k-dimensional Lebesque measure 0.)

We should confess that we do not know any useful applications of (**SL**.F), which is not the case for other relaxations of the $\mathcal{S}$-lemma we are about to consider. All these relaxations have to do with inhomogeneous versions of the lemma, like the one that follows.

(**SL**.??) *Consider a system of quadratic inequalities of the form*

$$x^T B_i x \leq d_i, \;\; i = 1, \ldots, m, \tag{S$'$}$$

where all d_i are ≥ 0, and a homogeneous quadratic form

$$f(x) = x^T Ax.$$

We want to evaluate the maximum f^* of the latter form on the solution set of (S$'$).

The standard semidefinite relaxation of the optimization problem

$$\max_x \left\{ f(x) : x^T B_i x \leq d_i \right\} \tag{P}$$

is the problem

$$\max_X \left\{ F(X) \equiv \mathrm{Tr}(AX) : \mathrm{Tr}(B_i X) \leq d_i, \;\; i = 1, \ldots, m, \;\; X = X^T \succeq 0 \right\}, \tag{SDP}$$

and the optimal value F^* in this problem is an upper bound on f^* (why?). *How large can the difference $F^* - f^*$ be?*

The relation between (**SL**.??) and (**SL**.?) is as follows. Assume that the only solution to the system of inequalities

$$x^T B_i x \leq 0, \;\; i = 1, \ldots, m,$$

is $x = 0$. Then (P) is equivalent to the optimization problem

$$\min_{\theta, x} \left\{ \theta : x^T Ax \leq \theta t^2, \;\; x^T B_i x \leq d_i t^2, \;\; i = 1, \ldots, m \right\} \tag{P$'$}$$

in the sense that both problems have the same optimal value f^* (why?). In other words,

> **(J)** *f^* is the smallest value of a parameter θ such that the homogeneous quadratic inequality*
>
> $$\begin{bmatrix} x \\ t \end{bmatrix}^T \widehat{A}_\theta \underbrace{\begin{bmatrix} x \\ t \end{bmatrix}}_{z} \geq 0, \quad \widehat{A}_\theta = \begin{pmatrix} -A & \\ & \theta \end{pmatrix}, \tag{C}$$

is a consequence of the system of homogeneous quadratic inequalities

$$z^T \widehat{B}_i z \geq 0, \ i = 1, \ldots, m, \quad \widehat{B}_i = \begin{pmatrix} -B_i & \\ & d_i \end{pmatrix}. \qquad \text{(H)}$$

Now let us assume that (P) is strictly feasible, so that (SDP) is also strictly feasible (why?), and that (SDP) is bounded above. By the conic duality theorem, the semidefinite dual of (SDP)

$$\min_{\lambda} \left\{ \sum_{i=1}^{m} \lambda_i d_i : \sum_{i=1}^{m} \lambda_i B_i \succeq A, \ \lambda \geq 0 \right\} \qquad \text{(SDD)}$$

is solvable and has the same optimal value F^* as (SDP). On the other hand, it is immediately seen that the optimal value in (SDD) is the smallest θ such that there exist nonnegative weights λ_i satisfying the relation

$$\widehat{A}_\theta \succeq \sum_{i=1}^{m} \lambda_i \widehat{B}_i.$$

Thus,

(K) *F^* is the smallest value of θ such that $\widehat{A}_\theta$ is $\succeq$ a combination, with nonnegative weights, of $\widehat{B}_i$'s, or, which is the same, F^* is the smallest value of the parameter θ for which* (C) *is a linear consequence of* (H).

Comparing **(J)** and **(K)**, we see that our question (**SL**.??) is closely related to the question, What is the gap between (i) and (ii) in (**SL**.?). In (**SL**.??), we are considering a parameterized family $z^T \widehat{A}_\theta z \geq 0$ of quadratic inequalities and we ask ourselves what the gap is between

(a) the smallest value f^* of the parameter θ for which the inequality $z^T \widehat{A}_\theta z \geq 0$ is a consequence of the system (H) of homogeneous quadratic inequalities,

and

(b) the smallest value F^* of θ for which the inequality $z^T \widehat{A}_\theta z \geq 0$ is a linear consequence of (H).

The goal of the subsequent exercises is to establish the following two results related to (**SL**.??).

Proposition 4.10.5. (Nesterov, Ye) *Consider* (**SL**.??) *and assume that*

1. *The matrices $B_1, \ldots, B_m$ commute with each other.*
2. *System* (S′) *is strictly feasible, and there exists a combination of the matrices B_i with nonnegative coefficients which is positive definite.*
3. *$A \succeq 0$.*

Then $f^* \geq 0$, (SDD) *is solvable with the optimal value F^*, and*

$$F^* \leq \frac{\pi}{2} f^*. \tag{4.10.137}$$

Proposition 4.10.6. (Nemirovski, Roos, Telaky) *Consider* (**SL**.??) *and assume that $B_i \succeq 0$, $d_i > 0$, $i = 1, \ldots, m$, and $\sum_i B_i \succ 0$. Then $f^* \geq 0$,* (SDD) *is solvable with the optimal value F^*, and*

$$F^* \leq 2 \ln \left(2 \sum_{i=1}^{m} \mathrm{Rank}(B_i) \right) f^*. \tag{4.10.138}$$

Exercise 4.58. *Derive Proposition* 4.10.5 *from the result of Exercise* 4.38.

> *Hint*. Observe that since B_i are commuting symmetric matrices, they share a common orthogonal eigenbasis, so that w.l.o.g. we can assume that all B_i's are diagonal.

The remaining exercises of this section are aimed at proving Proposition 4.10.6. We start with a simple technical result.

Exercise 4.59. 1. *Let $b \in \mathbf{R}^n$, and let $\xi = (\xi_1, \ldots, \xi_n)^T$ be a random vector with independent identically distributed coordinates each taking values ± 1 with probabilities* 0.5. *Prove that*

$$\forall (t \geq 0): \quad \mathrm{Prob}\{|b^T \xi| \geq t \|b\|_2\} \leq 2 \exp\{-t^2/2\}.$$

> *Hint*. By symmetry reasons, it suffices to demonstrate that
>
> $$\forall (t \geq 0): \quad \mathrm{Prob}\{b^T \xi \geq t \|b\|_2\} \leq \exp\{-t^2/2\}.$$
>
> To prove the latter inequality, observe that whenever $\rho \geq 0$, one has (why?)
>
> $$\mathrm{Prob}\{b^T \xi \geq t \|b\|_2\} \leq \mathbf{E}\left\{\exp\{\rho b^T \xi\}\right\} \exp\{-\rho t \|b\|_2\},$$
>
> where $\mathbf{E}$ stands for the expectation with respect to the distribution of ξ.
>
> Prove that
>
> $$\mathbf{E}\left\{\exp\{\rho b^T \xi\}\right\} = \prod_{i=1}^{n} \cosh(\rho b_i) \leq \prod_{i=1}^{n} \exp\{\rho^2 b_i^2/2\},$$
>
> thus coming to the bound
>
> $$\mathrm{Prob}\{b^T \xi \geq t \|b\|_2\} \leq \exp\{\rho^2 \|b\|_2^2/2 - \rho t \|b\|_2\} \quad \forall \rho \geq 0,$$
>
> and optimize the resulting bound in ρ.

Remark. The outlined reasoning is one of the standard ways (going back to Bernshtein) of bounding the probabilities of large deviations.

2. *Derive from* 1 *that if* B *is a positive semidefinite matrix and* ξ *is a random vector as in* 1*, then*

$$\mathrm{Prob}\{\xi^T B\xi \geq t\mathrm{Tr}(B)\} \leq 2\mathrm{Rank}(B)\exp\{-t/2\}.$$

Now we are ready to prove Proposition 4.10.6. Below it is assumed that the data of (**SL**.??) satisfy the premise of the Proposition 4.10.6.

Exercise 4.60. *Prove that* $F^* \geq f^* \geq 0$*,* $f^* = 0$ *if and only if* $A \preceq 0$*, and in the latter case* $F^* = 0$ *as well.*

In view of Exercise 4.60, the conclusion of Proposition 4.10.6 is valid in the case $f^* \leq 0$. In what follows, we assume that $f^* > 0$.

Exercise 4.61. 1. *Prove that* (SDP) *is solvable.*

2. *Let* X_* *be an optimal solution of* (SDP)*, and let*

$$\widehat{A} \equiv X_*^{1/2} A X_*^{1/2} = U\widetilde{A}U^T,$$

where U *is orthogonal and* $\widetilde{A}$ *is diagonal. Let us set*

$$\widetilde{B}_i = U X_*^{1/2} B_i X_*^{1/2} U^T,\ i = 1, \ldots, m.$$

2.a. *Prove that the optimal value* f_* *of the problem*

$$\max_x \left\{x^T \widetilde{A} x : x^T \widetilde{B}_i x \leq d_i,\ i = 1, \ldots, m\right\} \qquad (\widehat{\mathrm{P}})$$

(which is a properly scaled version of (P)*) is* $\leq f^*$ *(and is equal to* f^**, provided that* $X_* \succ 0$*).*

2.b. *Prove that* $\mathrm{Tr}(\widetilde{A}) = F^*$*,* $\widetilde{B}_i \succeq 0$ *and* $\mathrm{Tr}(\widetilde{B}_i) \leq d_i$*,* $i = 1, \ldots, m$*.*

2.c. *Let* ξ *be a random vector in* $\mathbf{R}^n$ *(*n *is the row size of the matrices* $A, B_1, \ldots, B_m$*) with independent coordinates taking values* ± 1 *with probabilities* $1/2$*. Derive from* 2.b *that*

$$\mathrm{Prob}\{\xi^T \widetilde{A}\xi = F^*\} = 1;$$
$$\mathrm{Prob}\{\exists i : \xi^T \widetilde{B}_i \xi > t d_i\} \leq 2\exp\{-t/2\}\sum_{i=1}^m \mathrm{Rank}(B_i).$$

Hint. Use the result of Exercise 4.59.2.

3. *Derive from* 2.a, c *that* $F^* \leq 2\ln\left(\sum_{i=1}^m \mathrm{Rank}(B_i)\right) f^*$*, thus completing the proof of Proposition* 4.10.6.

$\mathcal{S}$-lemma, matrix cube, Nesterov's theorem, Proposition 4.10.6, and more

A careful reader already will have discovered a common denominator of the important results given in the heading. This common denominator can be outlined as follows:

• Our goal is to build a computationally tractable upper bound on the optimal value of a seemingly difficult optimization problem

$$f^* = \max_x \{f(x) : x \in X\} \tag{4.10.139}$$

and to evaluate the quality of the bound.

• To achieve our goal, we act as follows:

1. **Build an approximation.** We build somehow an approximation of the problem of interest—a semidefinite program

$$\max_x \left\{ c^T x : \mathcal{A}(x) \equiv \sum_i x_i A_i \succeq B \right\} \tag{4.10.140}$$

such that the optimal value c^* in (4.10.140) is an upper bound on f^*.

2. **Write optimality conditions.** We write the optimality conditions for an optimal solution x_* of the approximation (4.10.140):

$$\begin{array}{rl} \mathcal{A}(x_*) \succeq B & \text{[primal feasibility]}, \\ \Xi_* \succeq 0,\ \mathcal{A}^*\Xi_* = c & \text{[dual feasibility]}, \\ c^T x_* = \mathrm{Tr}(B\Xi_*) & \text{[zero duality gap]}. \end{array} \tag{4.10.141}$$

3. **Apply stochastic interpretation.** We treat the positive semidefinite matrices appearing in (4.10.141) as the covariance matrices of certain distributions (usually just Gaussian ones), and interpret the optimality conditions (4.10.141) as relations between expectations of appropriate random variables associated with these distributions. In all examples we are speaking about, these relations between the expectations say something about the original problem (4.10.139), and this something straightforwardly yields a feasible solution of the original problem with the value of the objective of order of c^*, and thus yields a bound on the quality of our approximation.

Of course, this description is too diffuse to be a recipe (and there hardly could exist a recipe for getting nontrivial results). But this diffuse recipe allows us to prove the $\mathcal{S}$-lemma, the matrix cube theorem, Theorem 4.4.1, the Nesterov theorem, Theorem 4.3.2, and others. In what follows, we give two more applications of the recipe. Both deal with the bounds proposed and studied (via techniques completely different from ours) by Nesterov.[44] In these applications, we need the following fundamental fact. (We have already mentioned it on different occasions; now it is time to prove it.)

Theorem 4.10.2. Lagrange duality theorem for convex programming problem with LMI constraints. *Let $g_i(x) : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$, $i = 0, 1, \ldots, k$, be convex functions. Consider the optimization program*

$$\min_x \left\{ g_0(x) : g_i(x) \le 0,\ i = 1, \ldots, k, \mathcal{A}(x) \equiv A_0 + \sum_{j=1}^n x_j A_j \succeq 0 \right\} \quad [x \in \mathbf{R}^n,\ A_j \in \mathbf{S}^m]. \tag{4.10.142}$$

[44] Y. Nesterov, *Semidefinite relaxation and non-convex quadratic optimization*, Optim. Methods Software, 12 (1997), pp. 1–20. Y. Nesterov, Nonconvex quadratic optimization via conic relaxation, in *Handbook on Semidefinite Programming*, R. Saigal, H. Wolkowicz, and L. Vandenberghe, eds., Kluwer Academic Publishers, Dordrecht, the Netherlands, 2000, pp. 363–387.

Assume that the problem is strictly feasible:

$$\exists \bar{x}: \quad g_i(\bar{x}) < 0,\ i = 1, \ldots, k,\ \mathcal{A}(\bar{x}) \succ 0.$$

Let x_* be a feasible solution of (4.10.142) *such that all functions g_i, $i = 0, 1, \ldots, k$, are finite in a neighborhood of x_* and are differentiable at x_*. Then x_* is an optimal solution of* (4.10.142) *if and only if there exist $X_* \in \mathbf{S}^m_+$ and $\lambda_* \in \mathbf{R}^k_+$ such that the point $(x_*; X_*, \lambda_*)$ is a saddle point of the Lagrange function*

$$L(x; X, \lambda) = g_0(x) + \sum_{i=1}^{k} \lambda_i g_i(x) - \mathrm{Tr}(X\mathcal{A}(x))$$

on the set $\mathbf{R}^n \times (\mathbf{S}^m_+ \times \mathbf{R}^k_+)$, i.e.,

$$\forall (x \in \mathbf{R}^n, X \in \mathbf{S}^m_+, \lambda \in \mathbf{R}^k_+): \quad L(x_*; X, \lambda) \le L(x_*; X_*, \lambda_*) \le L(x; X_*, \lambda_*).$$

Exercise 4.62. *Prove Theorem* 4.10.2.

Hint. Prove that under the premise of the theorem x_* is an optimal solution of (4.10.142) if and only if x_* is an optimal solution in the linearized problem

$$\min_x \big\{ g_0(x_*) + (x - x_*)^T \nabla g_0(x_*) : g_i(x_*) + (x - x_*)^T \nabla g_i(x_*) \le 0,\ i = 1, \ldots, k,\ \mathcal{A}(x) \preceq 0 \big\}$$

and apply the conic duality theorem.

Bounding maximum of a convex quadratic form over $\|\cdot\|_p$-ball. Let $2 < p \le \infty$, and let $Q \succ 0$, $Q \in \mathbf{S}^m$. Consider the problem

$$\rho(Q) = \max_x \left\{ x^T Q x : \|x\|_p \le 1 \right\}, \tag{4.10.143}$$

which is a natural extension of the problem of maximizing a convex quadratic form over the cube $\|\cdot\|_\infty \le 1$. Nesterov proposed the following approximation of (4.10.143):

$$R(Q) = \min_{\lambda, \rho} \left\{ \rho : \rho Q^{-1} \succeq (\mathrm{Diag}(\lambda))^{-1},\ \lambda > 0,\ \|\lambda\|_r \le 1 \right\}, \quad r = \frac{p}{p-2}. \tag{4.10.144}$$

Our first goal is to verify that (4.10.144) yields an upper bound on $\rho(Q)$.

Exercise 4.63. 1. *Prove the following equivalences:*

$$\begin{array}{c} \rho(Q) \le \rho \\ \Updownarrow \\ \begin{pmatrix} \rho & d^T \\ d & Q^{-1} \end{pmatrix} \succeq 0 \quad \forall (d: \|d\|_p \le 1) \\ \Updownarrow \\ \rho Q^{-1} \succeq d d^T \quad \forall (d, \|d\|_p \le 1) \\ \Updownarrow \\ \rho x^T Q^{-1} x \ge \|x\|_q^2 \quad \forall x \quad [q = \frac{p}{p-1}]. \end{array} \tag{4.10.145}$$

Hint. Look at the proof of Lemma 4.4.6.

2. *Verify that if* $\lambda > 0$, $\|\lambda\|_r \leq 1$, *and* $\rho > 0$ *are such that* $\rho Q^{-1} \succeq (\mathrm{Diag}(\lambda))^{-1}$, *then*

$$\rho x^T Q^{-1} x \geq \|x\|_q^2 \quad \forall x.$$

Derive from 1 *and* 2 *that* $\rho(Q) \leq R(Q)$.

The challenge is to quantify the quality of the bound $R(Q)$, and here is the answer.

Proposition 4.10.7. *One has*

$$R(Q) \leq \omega_p^2 \rho(Q), \quad \omega_p = \left(\int_{-\infty}^{\infty} |t|^{\frac{p}{p-1}} \mathcal{G}(t) dt \right)^{-\frac{p-1}{p}}, \tag{4.10.146}$$

where

$$\mathcal{G}(t) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{t^2}{2} \right\}$$

is the standard Gaussian density. When p *grows from* 2 *to* ∞, *the factor* ω_p^2 *grows from* $\omega_2^2 = 1$ *to* $\omega_\infty^2 = \frac{\pi}{2}$.

To prove Proposition 4.10.7, observe that

$$R^r(Q) = \min_\mu \left\{ \sum_i \frac{1}{\mu_i^r} : \mu > 0,\ Q^{-1} \succeq \mathrm{Diag}(\mu) \right\}$$

(why?). Thus, it suffices to lead to a contradiction the assumption that

$$(\omega_p^2 \rho(Q))^r < \min_\mu \left\{ \sum_i \frac{1}{\mu_i^r} : \mu > 0,\ Q^{-1} \succeq \mathrm{Diag}(\mu) \right\}. \tag{4.10.147}$$

Exercise 4.64. *Assume that* (4.10.147) *takes place.*

1. *Prove that there exists* $X \succeq 0$ *such that*

$$\begin{array}{rcl} (\omega_p^2 \rho(Q))^r & < & \min_\mu \left\{ \sum_i \left[\frac{1}{\mu_i^r} + X_{ii} \mu_i \right] - \mathrm{Tr}(X Q^{-1}) : \mu \geq 0 \right\} \\ & = & \frac{2}{q^{\frac{q}{2}} (2-q)^{\frac{2-q}{2}}} \sum_i X_{ii}^{\frac{q}{2}} - \mathrm{Tr}(X Q^{-1}), \\ q & = & \frac{p}{p-1}. \end{array} \tag{4.10.148}$$

Hint. Use Theorem 4.10.2

2. *Let* X *be given by* 1, $Y = X^{1/2}$, *and let* ξ *be a Gaussian random vector with zero mean and unit covariance matrix. Verify that* (4.10.148) *says exactly that*

$$(\omega_p^2 \rho(Q))^r + \mathbf{E}\left\{ \xi^T Y Q^{-1} Y \xi \right\} < \frac{2}{q^{\frac{q}{2}} (2-q)^{\frac{2-q}{2}}} \omega_p^q \mathbf{E}\left\{ \|Y\xi\|_q^q \right\}$$

and use the result of Exercise 4.63 *to derive from this inequality that*

$$(\omega_p^2\rho(Q))^r + \rho^{-1}(Q)\mathbf{E}\left\{\|Y\xi\|_q^2\right\} - \frac{2}{q^{\frac{q}{2}}(2-q)^{\frac{2-q}{2}}}\omega_p^q \sum_i \mathbf{E}\left\{\|Y\xi\|_q^q\right\} < 0.$$

Conclude from the latter relation that

$$\min_{t>0}\left[(\omega_p^2\rho(Q))^r + \rho^{-1}(Q)t^2 - \frac{2}{q^{\frac{q}{2}}(2-q)^{\frac{2-q}{2}}}\omega_p^q t^q\right] < 0,$$

which in fact is not the case. Conclude that (4.10.147) *is impossible, thus completing the proof of Proposition* 4.10.7.

Bounding maximum of a quadratic form over a central cross section of a $\|\cdot\|_p$-ball. Let $2 < p \le \infty$, $Q \in \mathbf{S}^m$, and let A be an $n \times m$ matrix of rank m. Consider the problem

$$\rho_A(Q) = \max_x \left\{x^T Q x : \|Ax\|_p \le 1\right\}. \tag{4.10.149}$$

Geometrically we are maximizing a quadratic form on the intersection of a $\|\cdot\|_p$-ball in $\mathbf{R}^n$ and an m-dimensional linear subspace (the image space of A), while in (4.10.143) we were maximizing the form on the entire ball. Note also that now Q can be indefinite. Nesterov proposed the following approximation of (4.10.149):

$$R_A(Q) = \min_{\lambda,\rho}\left\{\rho : Q \preceq A^T\mathrm{Diag}(\lambda)A,\ \lambda \ge 0,\ \|\lambda\|_r \le \rho\right\}, \quad r = \frac{p}{p-2}. \tag{4.10.150}$$

Our first goal is to verify that (4.10.150) yields an upper bound on $\rho_A(Q)$.

Exercise 4.65. 1. *Prove the following equivalences:*

$$\begin{array}{c} \rho_A(Q) \le \rho \\ \Updownarrow \\ \rho \ge x^T Q x \quad \forall (x : \|Ax\|_p \le 1) \\ \Updownarrow \\ \rho\|Ax\|_p^2 \ge x^T Q x \quad \forall x. \end{array} \tag{4.10.151}$$

2. *Verify that if* $\lambda > 0$, $\|\lambda\|_r \le 1$, *and* $\rho \ge 0$ *are such that* $\rho Q \preceq \mathrm{Diag}(\lambda)$, *then*

$$\rho x^T Q x \le \|Ax\|_p^2 \quad \forall x.$$

Derive from 1 *and* 2 *that* $\rho_A(Q) \le R_A(Q)$.

The challenge is to quantify the quality of the bound $R_A(Q)$, and here is the answer.

Proposition 4.10.8. *One has*

$$R_A(Q) \le \sigma_p^2\rho_A(Q), \quad \sigma_p = \left(\int_{-\infty}^{\infty}|t|^p\mathcal{G}(t)dt\right)^{\frac{1}{p}}. \tag{4.10.152}$$

Note that

$$\sigma_p^2 \le 1 + \exp\{-1\}(p-2), \;\; p \ge 2.$$

To prove Proposition 4.10.8, observe that

$$R_A^r(Q) = \min_\lambda \left\{ \sum_i |\lambda_i|^r : Q \preceq A^T \mathrm{Diag}(\lambda) A \right\}.$$

(Why?) Thus, it suffices to lead to a contradiction the assumption that

$$(\sigma_p^2 \rho_A(Q))^r < \min_\lambda \left\{ \sum_i |\lambda_i|^r : Q \preceq A^T \mathrm{Diag}(\lambda) A \right\}. \tag{4.10.153}$$

Exercise 4.66. *Assume that* (4.10.153) *takes place.*

1. *Prove that there exists $X \succeq 0$ such that*

$$\begin{array}{rcl} (\sigma_p^2 \rho_A(Q))^r & < & \min\limits_\lambda \left\{ \sum\limits_i \left[|\lambda_i|^r - (AXA^T)_{ii} \lambda_i \right] + \mathrm{Tr}(XQ) \right\} \\ & = & -(r-1) r^{-\frac{r}{r-1}} \sum\limits_i ((AXA^T)_{ii})^{\frac{p}{2}} + \mathrm{Tr}(XQ). \end{array} \tag{4.10.154}$$

Hint. Use Theorem 4.10.2.

2. *Let X be given by* 1, *$Y = X^{1/2}$, $B = AY$, and let ξ be a Gaussian random vector with zero mean and unit covariance matrix. Verify that* (4.10.154) *says exactly that*

$$(\sigma_p^2 \rho_A(Q))^r + (r-1) r^{-\frac{r}{r-1}} \sigma_p^{-p} \mathbf{E}\left\{ \|B\xi\|_p^p \right\} < \mathbf{E}\left\{ \xi^T Y Q Y \xi \right\}$$

and use the result of Exercise 4.65 *to derive from this inequality that*

$$(\sigma_p^2 \rho_A(Q))^r + (r-1) r^{-\frac{r}{r-1}} \sigma_p^{-p} \mathbf{E}\left\{ \|B\xi\|_p^p \right\} < \rho_A(Q) \mathbf{E}\left\{ \|B\xi\|_p^2 \right\}.$$

Conclude from the latter relation that

$$\min_{t>0} \left[(\sigma_p^2 \rho_A(Q))^r + (r-1) r^{-\frac{r}{r-1}} \sigma_p^{-p} t^p - \rho_A(Q) t^2 \right] < 0,$$

which in fact is not the case. Conclude that (4.10.153) *is impossible, thus completing the proof of Proposition* 4.10.8.

Exercise 4.67. *Let $Q \in \mathbf{S}^m$ be a positive definite matrix, and let $L = \{x \in \mathbf{R}^m : e^T x = 0\}$, where $e \neq 0$. Assume that we want to bound from above the quantity*

$$\rho_e(Q) = \max_x \left\{ x^T Q x : x \in L, \|x\|_\infty \le 1 \right\}.$$

1. *Prove that the optimal value $R_e(Q)$ in the optimization problem*

$$R_e(Q) = \min_{\lambda, f} \left\{ \sum_i \lambda_i : \lambda \ge 0, (I - ef^T) Q (I - fe^T) \preceq \mathrm{Diag}(\lambda), \, f^T e = 1 \right\}$$

is an upper bound on $\rho_e(Q)$.

Hint. Observe that whenever λ, f is a feasible solution of the problem, one has

$$\max_x \left\{x^T Q x : x \in L, \|x\|_\infty \le 1\right\} \le \max_x \left\{x^T (I - ef^T) Q (I - fe^T) x : \|x\|_\infty \le 1\right\} \le \sum_i \lambda_i.$$

2. *Prove that the quantity $R_e(Q)$ is at most 2π times larger than $\rho_e(Q)$.*

Hint. Choose as f the vector with coordinates $\mathrm{sign}(e_i)(\sum_i |e_i|)^{-1}$ and prove that with this choice of f, the ratio

$$\frac{\max_x \left\{x^T (I - ef^T) Q (I - fe^T) x : \|x\|_\infty \le 1\right\}}{\max_x \left\{x^T Q x : x \in L, \|x\|_\infty \le 1\right\}}$$

does not exceed 4. Further apply the Nesterov Theorem 4.3.2.

3. *Prove that the quantity $R_e(Q)$ is computable via SDP, namely,*

$$R_e(Q) = \min_{\lambda, f} \left\{ \sum_i \lambda_i : \begin{pmatrix} \mathrm{Diag}(\lambda) & I - ef^T \\ I - fe^T & Q^{-1} \end{pmatrix} \succeq 0,\ f^T e = 1 \right\}.$$

The exercise to follow deals with a simple combinatorial applications of Exercise 4.67. Consider the following combinatorial problem:

> MAXCUT(n, p). You are given an n-node graph and an integer $p \ge n/2$. The arcs of the graph are assigned nonnegative weights. You are allowed to color the nodes of the graph in two colors, red and green, in such a way that there will be at most p and at least $n/2$ red nodes. Under this restriction, you are want to maximize the total weight of the arcs linking nodes of different colors.

Note that MAXCUT(n, n) is the usual MAXCUT problem.

Exercise 4.68. *Build a computable upper bound on the optimal value in* MAXCUT(p, n) *that coincides with the true optimal value within the factor 2π. Think how to produce efficiently a 2π-suboptimal coloring.*

Here are the numerical results we have obtained for the *equipartition* problem MAXCUT$(n/2, n)$ in a sample of 10 randomly generated 40-node graphs:

#	Bound	True value $\ge$	$\frac{\text{Bound}}{\text{True value}}$ $\le$	#	Bound	True value $\ge$	$\frac{\text{Bound}}{\text{True value}}$ $\le$
1	60.86	58.52	1.04	6	62.09	59.83	1.04
2	48.70	47.70	1.02	7	58.16	55.15	1.05
3	69.34	65.60	1.06	8	59.49	56.92	1.05
4	57.71	55.24	1.04	9	61.51	59.77	1.03
5	64.82	61.88	1.05	10	60.43	59.00	1.02

Note that the brute force search for the optimal equipartition of a 40-node graph requires us to look through $\frac{1}{2}\binom{40}{20} \approx 6.89 \times 10^{10}$ variants.

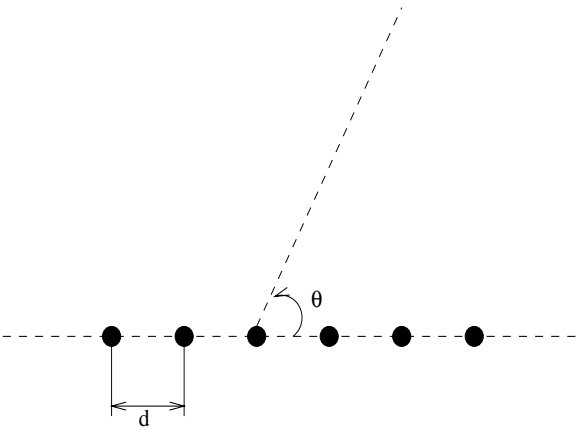


Figure 4.21. *Linear array of harmonic oscillators.*

4.10.6 Antenna synthesis

Consider the antenna synthesis problem from Example 2, section 4.6:

> *Given an equidistant array of n harmonic oscillators placed along the X-axis in the XY-plane* (*Fig.* 4.21), *choose complex weights* z_j, $j = 1, \ldots, n$, *to get the modulus* $|Z(\cdot)|$ *of the corresponding diagram*
>
> $$Z(\theta) = \sum_{l=0}^{n-1} z_l \exp\{-il\Omega(\theta)\}, \quad \Omega(\theta) = \frac{2\pi d \cos(\theta)}{\lambda},$$
>
> *as close as possible to a given target.*

Assume we want to get a diagram that is concentrated in the beam $-\phi_+ \le \theta \le \phi_+$. The natural design specifications in this case are as follows: we fix a ripple $\alpha > 1$, require the diagram to satisfy the inequality

$$\frac{1}{\alpha} \le |Z(\theta)| \le \alpha, \ -\phi_+ \le \theta \le \phi_+,$$

and minimize under this restriction the sidelobe attenuation level

$$\max_{\phi_- \le |\theta| \le \pi} |Z(\theta)|,$$

where $\phi_- > \phi_+$ is a given sidelobe angle.

To solve the problem, one may use a simplified version of the approach presented in section 4.6. Specifically, the problem is posed as

$$
\begin{array}{lrcl}
\text{minimize} & \epsilon & & \\
\text{s.t.} & & & \\
\text{(a)} & 0 \le R(\omega) & \equiv & r(0) + \sum_{l=1}^{n-1} (r(2l-1)\cos(l\omega) + r(2l)\sin(l\omega)),\ \omega \in \Gamma_1, \\
\text{(b)} & 0 \le R(\omega) & \le & \epsilon,\ \omega \in \Gamma_2, \\
\text{(c)} & 0 & \le & R(\omega),\ \omega \in \Gamma_3, \\
\text{(d)} & \alpha^{-2} \le R(\omega) & \le & \alpha^2,\ \omega \in \Gamma_4, \\
\text{(e)} & 0 & \le & R(\omega),\ \omega \in \Gamma_5,
\end{array}
$$

where Γ_j, $j = 1, 2, 3, 4, 5$, are fine finite grids in the segments

$$
\begin{array}{rcl}
\Delta_1 & = & [-\pi, \omega_{\min}],\ \omega_{\min} = -\frac{2\pi d}{\lambda}, \\
\Delta_2 & = & [\omega_{\min}, \omega_-],\ \omega_- = \frac{2\pi d \cos(\phi_-)}{\lambda}, \\
\Delta_3 & = & [\omega_-, \omega_+],\ \omega_+ = \frac{2\pi d \cos(\phi_+)}{\lambda}, \\
\Delta_4 & = & [\omega_+, \omega_{\max}],\ \omega_{\max} = \frac{2\pi d}{\lambda}, \\
\Delta_5 & = & [\omega_{\max}, \pi].
\end{array}
\qquad (*)
$$

Note that the lower bounds in (a), (b), (c), (e) are aimed at ensuring that $R(\omega)$ is nonnegative on $[-\pi, \pi]$, which is a necessary and sufficient condition for $R(\Omega(\theta))$ to be of the form $|Z(\theta)|^2$ for some $Z(\theta)$. Of course, the indicated lower bounds ensure nonnegativity of R only on the grid $\Gamma = \bigcup_{j=1}^5 \Gamma_j$ in the segment $[-\pi, \pi]$, not on the segment itself, so that a solution to (*) sometimes should be modified to yield an actual diagram. Note that the settings we dealt with in section 4.6 were free of this drawback: there we were ensuring nonnegativity of $R(\cdot)$ by restricting the coefficient vector r to belong to the SDr set of coefficients of nonnegative trigonometric polynomials. The approximate setting (*), however, has a definite advantage—this is just an LP program, not a semidefinite one. To utilize this advantage, we should know how to modify a solution to (*) in order to make the corresponding $R(\cdot)$ nonnegative on the entire segment.

Exercise 4.69. *Assume that $n \ge 2$, Γ is an M-point equidistant grid on $[-\pi, \pi]$,*

$$\Gamma = \left\{ -\pi + \frac{2\pi j}{M} \mid j = 1, 2, \ldots, M \right\},$$

and

$$M \ge \sqrt{\pi^3 (n-1)^3}.$$

Prove that if (ϵ, r) is a feasible solution to (*), then $r(0) \ge 0$ and the regularized $R(\cdot)$, the trigonometric polynomial

$$R(\omega) + \delta \equiv r(0) + \delta + \sum_{l=1}^{n-1} (r(2l-1)\cos(l\omega) + r(2l)\sin(l\omega)),$$

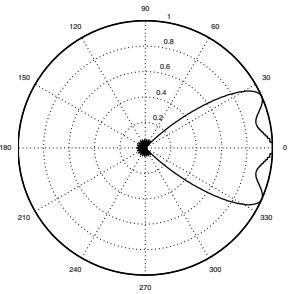


Figure 4.22. *The dream.*

where the regularization δ is given by

$$\delta = \frac{\pi^3(n-1)^3}{M^2} r(0),$$

is nonnegative on $[-\pi, \pi]$.

Consider now the particular data as follows:

- The number of oscillators in array is $n = 12$.
- The interelement distance is $d = 0.25\lambda$.
- The (half-) width of the beam $\phi_+ = 30^o$.
- The sidelobe angle $\phi_- = 45^o$.
- The ripple $\alpha = 1\text{dB} = 10^{1/20} \approx 1.1220$.

When solving (*) with Γ chosen as the equidistant 2048-point grid, one obtains the pattern function $|Z^{\text{nom}}(\theta)|$ shown in Fig. 4.22. The resulting sidelobe attenuation level $\sqrt{\epsilon}$ is $-23.42\text{dB} = 10^{-23.42/20} \approx 0.068$. The result, however, is completely unstable with respect to implementation errors: when the weights z_j^{nom} of the resulting diagram $Z^{\text{nom}}(\theta)$ are perturbed as

$$z_j^{\text{nom}} \mapsto (1 + \epsilon_j) z_j^{\text{nom}},$$

where ϵ_j are independent random (complex-valued) perturbations with zero mean and norm not exceeding $\epsilon\sqrt{2}$, the pattern function of an actual (random) diagram may look as shown in Fig. 4.23. The pattern functions in Fig. 4.24 are given by a robust design.

Exercise 4.70. *Think how to build a robust setting of the antenna synthesis problem from Example* 2, *section* 4.6.

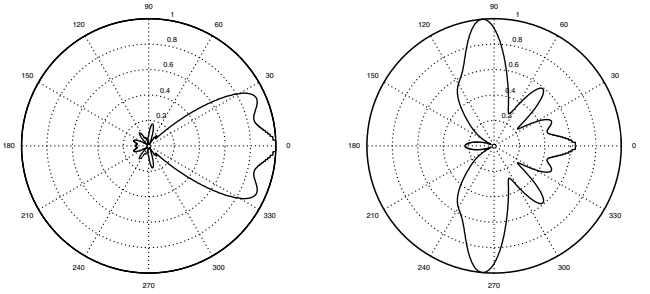


Figure 4.23. *The reality. Left: perturbation level* $\epsilon = 0.005$*, sidelobe attenuation level* -15.12dB ≈ 0.18*, actual ripple* 1.65dB ≈ 1.21*. Right: perturbation level* $\epsilon = 0.025$*, sidelobe attenuation level* 3.89dB ≈ 1.56*, actual ripple* 4.45dB ≈ 1.67.

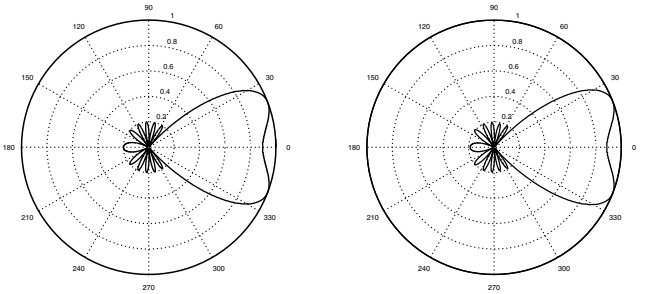


Figure 4.24. *Robust design, actual pattern functions. Left: perturbation level* $\epsilon = 0.005$*, sidelobe attenuation level* -14.07dB ≈ 0.20*, actual ripple* 0.97dB ≈ 1.12*. Right: perturbation level* $\epsilon = 0.05$*, sidelobe attenuation level* -13.43dB ≈ 0.21*, actual ripple* 1.04dB ≈ 1.13.

4.10.7 Ellipsoidal approximations

Exercise 4.71. *Prove the Löwner–Fritz John theorem* (*Theorem* 4.9.1).

More on ellipsoidal approximations of sums of ellipsoids. The goal of the two subsequent exercises is to get in an alternative way the problem (Õ) generating a parametric family of ellipsoids containing the arithmetic sum of m given ellipsoids (section 4.9.2).

Exercise 4.72. *Let* P_i *be nonsingular and* Λ_i *be positive definite* $n \times n$ *matrices,* $i = 1, \ldots, m$*. Prove that for every collection* $x^1, \ldots, x^m$ *of vectors from* $\mathbf{R}^n$ *one has*

$$[x^1 + \cdots + x^m]^T \left[\sum_{i=1}^m [P_i^T]^{-1} \Lambda_i^{-1} P_i^{-1} \right]^{-1} [x^1 + \cdots + x^m] \leq \sum_{i=1}^m [x^i]^T P_i \Lambda_i P_i^T x^i. \tag{4.10.155}$$

Hint. Consider the $(nm+n)\times(nm+n)$ symmetric matrix

$$A=\left[\begin{array}{ccc|c} P_1\Lambda_1P_1^T & & & I_n \\ & \ddots & & \vdots \\ & & P_m\Lambda_mP_m^T & I_n \\ \hline I_n & \cdots & I_n & \sum_{i=1}^m [P_i^T]^{-1}\Lambda_i^{-1}P_i^{-1} \end{array}\right]$$

and apply twice the Schur complement lemma: the first time to prove that the matrix is positive semidefinite, and the second time to get from the latter fact the desired inequality.

Exercise 4.73. *Assume you are given m full-dimensional ellipsoids centered at the origin*

$$W_i=\{x\in\mathbf{R}^n \mid x^TB_ix\le 1\},\ i=1,\dots,m, \qquad [B_i\succ 0]$$

in $\mathbf{R}^n$.

1. *Prove that for every collection Λ of positive definite $n\times n$ matrices Λ_i such that*

$$\sum_i \lambda_{\max}(\Lambda_i)\le 1$$

the ellipsoid

$$E_\Lambda=\{x \mid x^T\left[\sum_{i=1}^m B_i^{-1/2}\Lambda_i^{-1}B_i^{-1/2}\right]^{-1}x\le 1\}$$

contains the sum $W_1+\cdots+W_m$ of the ellipsoids W_i.

2. *Prove that in order to find the smallest-volume ellipsoid in the family $\{E_\Lambda\}_\Lambda$ defined in* 1 *it suffices to solve the semidefinite program*

maximize t

s.t.

$$\begin{array}{ll} \text{(a)} & t\le \mathrm{Det}^{1/n}(Z), \\ \text{(b)} & \left[\begin{array}{cccc} \Lambda_1 & & & \\ & \Lambda_2 & & \\ & & \ddots & \\ & & & \Lambda_m \end{array}\right] \succeq \left[\begin{array}{cccc} B_1^{-1}ZB_1^{-1} & B_1^{-1/2}ZB_2^{-1/2} & \dots & B_1^{-1/2}ZB_m^{-1/2} \\ B_2^{-1/2}ZB_1^{-1/2} & B_2^{-1/2}ZB_2^{-1/2} & \dots & B_2^{-1/2}ZB_m^{-1/2} \\ \vdots & \vdots & \ddots & \vdots \\ B_m^{-1/2}ZB_1^{-1/2} & B_m^{-1/2}ZB_2^{-1/2} & \dots & B_m^{-1/2}ZB_m^{-1/2} \end{array}\right], \\ \text{(c)} & Z\succeq 0, \\ \text{(d)} & \Lambda_i\preceq\lambda_iI_n,\ i=1,\dots,m, \\ \text{(e)} & \sum_{i=1}^m\lambda_i\le 1 \end{array} \qquad (4.10.156)$$

in variables $Z,\Lambda_i\in\mathbf{S}^n, t,\lambda_i\in\mathbf{R}$. The smallest-volume ellipsoid in the family $\{E_\Lambda\}_\Lambda$ is E_{Λ^*}, where Λ^* is the Λ-part of an optimal solution of the problem.

Hint. Use example 20.c from this lecture, page 155.

3. *Demonstrate that the optimal value in* (4.10.156) *remains unchanged when the matrices* Λ_i *are further restricted to be scalar:* $\Lambda_i = \lambda_i I_n$. *Prove that with this additional constraint problem* (4.10.156) *becomes equivalent to problem* ($\tilde{\text{O}}$) *from section* 4.9.2.

Remark 4.10.1. Exercise 4.73 demonstrates that the approximating scheme for solving problem (O) presented in Proposition 4.9.4 is equivalent to the following one:

> Given m positive reals λ_i with unit sum, one defines the ellipsoid $E(\lambda) = \{x \mid x^T \left[\sum_{i=1}^m \lambda_i^{-1} B_i^{-1}\right]^{-1} x \leq 1\}$. This ellipsoid contains the arithmetic sum W of the ellipsoids $\{x \mid x^T B_i x \leq 1\}$, and to approximate the smallest volume ellipsoid containing W, we merely minimize $\text{Det}(E(\lambda))$ over λ varying in the standard simplex $\{\lambda \geq 0, \sum_i \lambda_i = 1\}$.

In this form, the approximation scheme in question was proposed by Schweppe (1975).

Exercise 4.74. *Let* A_i *be nonsingular* $n \times n$ *matrices,* $i = 1, \ldots, m$, *and let* $W_i = \{x = A_i u \mid u^T u \leq 1\}$ *be the associated ellipsoids in* $\mathbf{R}^n$. *Let* $\Delta_m = \{\lambda \in \mathbf{R}^m_+ \mid \sum_i \lambda_i = 1\}$. *Prove that*

1. *whenever* $\lambda \in \Delta_m$ *and* $A \in \mathbf{M}^{n,n}$ *is such that*

$$AA^T \succeq F(\lambda) \equiv \sum_{i=1}^m \lambda_i^{-1} A_i A_i^T,$$

the ellipsoid $E[A] = \{x = Au \mid u^T u \leq 1\}$ *contains* $W = W_1 + \cdots + W_m$;

Hint. Use the result of Exercise 4.73.1.

2. *whenever* $A \in \mathbf{M}^{n,n}$ *is such that*

$$AA^T \preceq F(\lambda) \quad \forall \lambda \in \Delta_m,$$

the ellipsoid $E[A]$ *is contained in* $W_1 + \cdots + W_m$, *and vice versa.*

Hint. Note that

$$\left(\sum_{i=1}^m |\alpha_i|\right)^2 = \min_{\lambda \in \Delta_m} \sum_{i=1}^m \frac{\alpha_i^2}{\lambda_i}$$

and use statement **(F)** from section 4.9.2.

Simple ellipsoidal approximations of sums of ellipsoids. Let $W_i = \{x = A_i u \mid u^T u \leq 1\}$, $i = 1, \ldots, m$, be full-dimensional ellipsoids in $\mathbf{R}^n$ (so that A_i are nonsingular $n \times n$ matrices), and let $W = W_1 + \cdots + W_m$ be the arithmetic sum of these ellipsoids. Observe that W is the image of the set

$$\mathcal{B} = \left\{ u = \begin{bmatrix} u[1] \\ \vdots \\ u[m] \end{bmatrix} \in \mathbf{R}^{nm} \mid u^T[i] u[i] \leq 1, \ i = 1, \ldots, m \right\}$$

under the linear mapping

$$u \mapsto \mathcal{A}u = \sum_{i=1}^{m} A_i u[i] : \mathbf{R}^{nm} \to \mathbf{R}^n.$$

It follows that

> *Whenever an nm-dimensional ellipsoid $\mathcal{W}$ contains $\mathcal{B}$, the set $\mathcal{A}(\mathcal{W})$, which is an n-dimensional ellipsoid* (why?) *contains W, and whenever $\mathcal{W}$ is contained in $\mathcal{B}$, the ellipsoid $\mathcal{A}(\mathcal{W})$ is contained in W.*

In view of this observation, we can try to approximate W from inside and from outside by the ellipsoids $W_- \equiv \mathcal{A}(\mathcal{W}_-)$ and $W^+ = \mathcal{A}(\mathcal{W}^+)$, where $\mathcal{W}_-$ and $\mathcal{W}^+$ are, respectively, the largest and the smallest volume nm-dimensional ellipsoids contained in or containing $\mathcal{B}$.

Exercise 4.75. 1. *Prove that*

$$\begin{aligned}
\mathcal{W}_- &= \left\{ u \in \mathbf{R}^{nm} \mid \sum_{i=1}^{m} u^T[i]u[i] \le 1 \right\}, \\
\mathcal{W}^+ &= \left\{ u \in \mathbf{R}^{nm} \mid \sum_{i=1}^{m} u^T[i]u[i] \le m \right\},
\end{aligned}$$

so that

$$\begin{aligned}
W \supset W_- &\equiv \left\{ x = \sum_{i=1}^{m} A_i u[i] \mid \sum_{i=1}^{m} u^T[i]u[i] \le 1 \right\}, \\
W \subset W_+ &\equiv \left\{ x = \sum_{i=1}^{m} A_i u[i] \mid \sum_{i=1}^{m} u^T[i]u[i] \le m \right\} = \sqrt{m} W_-.
\end{aligned}$$

2. *Prove that W_- can be represented as*

$$W_- = \{x = Bu \mid u \in \mathbf{R}^n, u^T u \le 1\}$$

with matrix $B \succ 0$ representable as

$$B = \sum_{i=1}^{m} A_i X_i$$

with square matrices X_i of norms $|X_i| \le 1$.

Derive from this observation that the level of conservativeness of the inner ellipsoidal approximation of W given by Proposition 4.9.6 *is at most $\sqrt{m}$: if W_* is this inner ellipsoidal approximation and W_{**} is the largest volume ellipsoid contained in W, then*

$$\left(\frac{\mathrm{Vol}(W_{**})}{\mathrm{Vol}(W_*)} \right)^{1/n} \le \left(\frac{\mathrm{Vol}(W)}{\mathrm{Vol}(W_*)} \right)^{1/n} \le \sqrt{m}.$$

Invariant ellipsoids

Exercise 4.76. *Consider a discrete time controlled dynamic system*

$$\begin{array}{rcl} x(t+1) & = & Ax(t)+bu(t),\ t=0,1,2,\ldots, \\ x(0) & = & 0, \end{array}$$

where $x(t)\in\mathbf{R}^n$ is the state vector and $u(t)\in[-1,1]$ is the control at time t. An ellipsoid centered at the origin

$$W=\{x\mid x^TZx\le 1\}\quad [Z\succ 0]$$

is called invariant if

$$x\in W\Rightarrow Ax\pm b\in W.$$

Prove the following:

1. *If W is an invariant ellipsoid and $x(t)\in W$ for some t, then $x(t')\in W\ \forall t'\ge t$.*
2. *Assume that the vectors $b, Ab, A^2b,\ldots,A^{n-1}b$ are linearly independent. Prove that an invariant ellipsoid exists if and only if A is stable (the absolute values of all eigenvalues of A are <1).*
3. *Assuming that A is stable, prove that an ellipsoid $\{x\mid x^TZx\le 1\}$ $[Z\succ 0]$ is invariant if and only if there exists $\lambda\ge 0$ such that*

$$\left(\begin{array}{cc} 1-b^TZb-\lambda, & -b^TZA \\ -A^TZb, & \lambda Z-A^TZA \end{array}\right)\succeq 0.$$

How could one use this fact to approximate numerically the smallest-volume invariant ellipsoid?

Greedy infinitesimal ellipsoidal approximations. Consider a linear time-varying controlled system

$$\frac{d}{dt}x(t)=A(t)x(t)+B(t)u(t)+v(t) \tag{4.10.157}$$

with continuous matrix-valued functions $A(t)$, $B(t)$, continuous vector-valued function $v(\cdot)$, and norm-bounded control:

$$\|u(\cdot)\|_2\le 1. \tag{4.10.158}$$

Assume that the initial state of the system belongs to a given ellipsoid:

$$x(0)\in E(0)=\{x\mid (x-x^0)^TG^0(x-x^0)\le 1\}\quad [G^0=[G^0]^T\succ 0]. \tag{4.10.159}$$

Our goal is to build, in an on-line fashion, a system of ellipsoids

$$E(t)=\{x\mid (x-x_t)^TG_t(x-x_t)\le 1\}\quad [G_t=G_t^T\succ 0] \tag{4.10.160}$$

in such a way that if $u(\cdot)$ is a control satisfying (4.10.158) and $x(0)$ is an initial state satisfying (4.10.159), then for every $t\ge 0$ it holds that

$$x(t)\in E(t).$$

We want to minimize the volumes of the resulting ellipsoids.

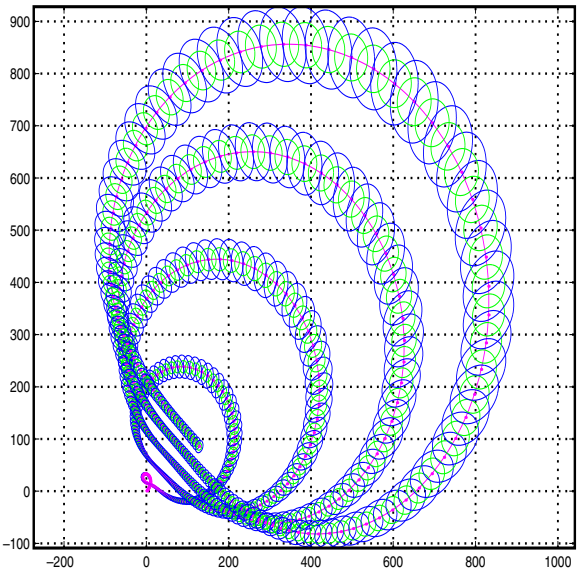


Figure 4.25. *Spiral.*

$$\frac{d}{dt}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} \cos(t) & -\sin(t) \\ \sin(t) & \cos(t) \end{bmatrix}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + u(t)\begin{bmatrix} \cos(t) \\ \sin(t) \end{bmatrix} + \begin{bmatrix} 10 \\ 10 \end{bmatrix}$$

$$x(0) = 0, \quad |u(\cdot)| \le 1, \quad 0 \le t \le 30$$

There is no difficulty with the path x_t of centers of the ellipsoids: it obviously should satisfy the requirements

$$\frac{d}{dt}x_t = A(t)x_t + v(t), \ t \ge 0; \quad x_0 = x^0. \tag{4.10.161}$$

Let us take this choice for granted and focus on how we should define the positive definite matrices G_t. Let us look for a continuously differentiable matrix-valued function G_t, taking values in the set of positive definite symmetric matrices, with the following property:

(L) *For every $t \ge 0$ and every point $x^t \in E(t)$* (see (4.10.160)), *every trajectory $x(\tau)$, $\tau \ge t$, of the system*

$$\frac{d}{d\tau}x(\tau) = A(\tau)x(\tau) + B(\tau)u(\tau) + v(\tau), \quad x(t) = x^t,$$

with $\|u(\cdot)\|_2 \le 1$ satisfies $x(\tau) \in E(\tau)\ \forall \tau \ge t$.

Note that **(L)** is a sufficient (but in general not necessary) condition for the system of ellipsoids $E(t)$, $t \ge 0$, to cover all trajectories of (4.10.157)–(4.10.158). Indeed, when formulating **(L)**, we act as if we were sure that the states $x(t)$ of our system run through the entire ellipsoid $E(t)$, which is not necessarily the case. The advantage of **(L)** is that this condition can be converted into an infinitesimal form.

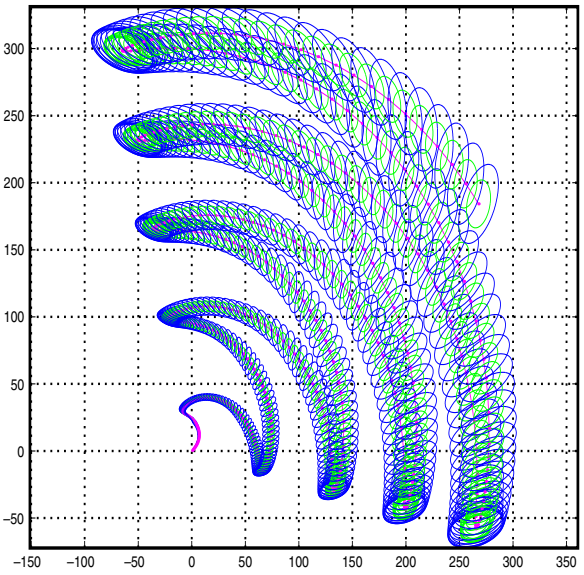


Figure 4.26. *Snake.*

$$\frac{d}{dt}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 0 & -\sin(t) \\ \sin(t) & 0 \end{bmatrix}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + u(t)\begin{bmatrix} \cos(t) \\ \sin(t) \end{bmatrix} + \begin{bmatrix} 10 \\ 10 \end{bmatrix}$$
$$x(0) = 0, \quad |u(\cdot)| \le 1, \quad 0 \le t \le 30$$

Exercise 4.77. *Prove that if $G_t \succ 0$ is continuously differentiable and satisfies* **(L)**, *then*

$$\forall \left(t \ge 0, x, u : x^T G_t x = 1, u^T u \le 1\right) : \\ 2u^T B^T(t) G_t x + x^T \left[\frac{d}{dt} G_t + A^T(t) G_t + G_t A(t)\right] x \le 0. \tag{4.10.162}$$

Conversely, if G_t is a continuously differentiable function taking values in the set of positive definite symmetric matrices and satisfying (4.10.162) *and the initial condition $G_0 = G^0$, then the associated system of ellipsoids $\{E(t)\}$ satisfies* **(L)**.

The result of Exercise 4.77 provides us with a kind of description of the families of ellipsoids $\{E(t)\}$ we are interested in. Now let us take care of the volumes of these ellipsoids. The latter can be done via a "greedy" (locally optimal) policy: given $E(t)$, let us try to minimize, under restriction (4.10.162), the derivative of the volume of the ellipsoid at time t. Note that this locally optimal policy does not necessary yield the smallest volume ellipsoids satisfying **(L)** (achieving "instant reward" is not always the best way to happiness); nevertheless, this policy makes sense.

We have $2 \ln \mathrm{vol}(E_t) = -\ln \mathrm{Det}(G_t)$, whence

$$2\frac{d}{dt} \ln \mathrm{vol}(E(t)) = -\mathrm{Tr}\left(G_t^{-1} \frac{d}{dt} G_t\right);$$

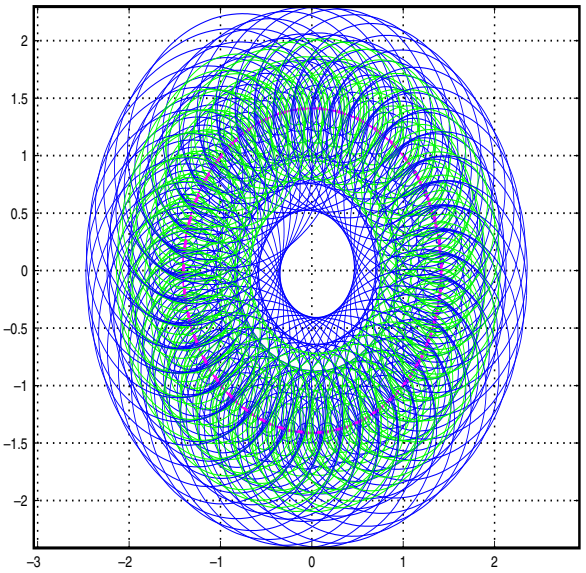


Figure 4.27. *Pendulum.*

$$\frac{d}{dt}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + u(t)\begin{bmatrix} 0 \\ 0.05 \end{bmatrix} \quad \left[\Leftrightarrow \begin{cases} \frac{d^2}{dt^2}x_1(t) &= -x_1(t) + 0.05u(t) \\ x_2(t) &= \frac{d}{dt}x_1(t) \end{cases}\right]$$

$$x_1(0) = 0,\ x_2(0) = 1, \quad |u(\cdot)| \le 1, \quad 0 \le t \le 30$$

thus, our greedy policy requires us to choose $H_t \equiv \frac{d}{dt}G_t$ as a solution to the optimization problem

$$\max_{H=H^T}\Big\{\mathrm{Tr}(G_t^{-1}H) : 2u^TB^T(t)G_tx + x^T[\tfrac{d}{dt}G_t - A^T(t)G_t - G_tA(t)]x \le 0 \\ \forall\,(x,u : x^TG_tx = 1, u^Tu \le 1)\Big\}.$$

Exercise 4.78. *Prove that the outlined greedy policy results in the solution G_t to the differential equation*

$$\frac{d}{dt}G_t = -A^T(t)G_t - G_tA(t) - \sqrt{\frac{n}{\mathrm{Tr}(G_tB(t)B^T(t))}}G_tB(t)B^T(t)G_t \\ - \sqrt{\frac{\mathrm{Tr}(G_tB(t)B^T(t))}{n}}\,G_t,\ \ t \ge 0; \\ G_0 = G^0.$$

Prove that the solution to this equation is symmetric and positive definite for all $t > 0$, provided that $G^0 = [G^0]^T \succ 0$.

Exercise 4.79. *Modify the previous reasoning to demonstrate that the locally optimal policy for building inner ellipsoidal approximation of the set*

$$X(t) = \left\{x(t) \mid \exists x^0 \in E(0) \equiv \{x \mid (x-x^0)^T G^0 (x-x^0) \le 1\}, \exists u(\cdot), \|u(\cdot)\|_2 \le 1 : \frac{d}{d\tau}x(\tau) = A(\tau)x(\tau) + B(\tau)u(\tau) + v(\tau),\ 0 \le \tau \le t,\ x(0) = x^0\right\}$$

results in the family of ellipsoids

$$\underline{E}(t) = \{x \mid (x - x_t)^T W_t (x - x_t) \le 1\},$$

where x_t is given by (4.10.161) *and W_t is the solution of the differential equation*

$$\frac{d}{dt}W_t = -A^T(t)W_t - W_t A(t) - 2W_t^{1/2}(W_t^{1/2}B(t)B^T(t)W_t^{1/2})^{1/2}W_t^{1/2},\ t \ge 0; \quad W_0 = G^0.$$

Illustrations of the results of Exercises 4.78 and 4.79 are given in Figs. 4.25, 4.26, and 4.27. The pictures relate to three distinct 2D systems ($\dim x = 2$). The dotted path is the path x_t of centers of the ellipses $E(\cdot)$ (the larger ones) and $\underline{E}(\cdot)$ (the smaller ones).

Lecture 5

Computational Tractability of Convex Programs

Until now, we did not consider the question of how to solve optimization problems of the types we have encountered. This is the issue we address in this and in the next lectures.

5.1 Numerical solution of optimization programs—preliminaries

5.1.1 Mathematical programming programs

All optimization programs we dealt with are covered by the following universal form of a mathematical programming program:

$$\min_x \left\{ p_0(x) : x \in X(p) \subset \mathbf{R}^{n(p)} \right\}, \tag{p}$$

where

- $n(p)$ is the design dimension of problem (p),
- $X(p) \subset \mathbf{R}^{n(p)}$ is the feasible domain of the problem, and
- $p_0(x) : \mathbf{R}^{n(p)} \to \mathbf{R}$ is the objective of (p).

The mathematical programming form (p) of an optimization program is most convenient for investigating solvability issues in optimization, so at this point we switch to this form. Note that the optimization programs we dealt with in the previous lectures—the conic programs

$$\min_x \left\{ c^T x : Ax - b \in \mathbf{K} \right\},$$

where $\mathbf{K} \subset \mathbf{R}^m$ is closed convex pointed cone with a nonempty interior—can be easily written in the MP form with

$$X(p) = \{x \mid Ax - b \in \mathbf{K}\}, \;\; p_0(x) = c^T x. \tag{5.1.1}$$

The MP programs obtained from the conic ones possess a very important characteristic feature: they are *convex*.

Definition 5.1.1. *A mathematical programming program (p) is called convex if*

- *the domain $X(p)$ of the program is a convex set: whenever $x, x' \in X(p)$, the segment $\{y = \lambda x + (1-\lambda)x' \mid 0 \le \lambda \le 1\}$ linking the points x, x' is contained in $X(p)$;*
- *the objective $p_0(x)$ is convex on $\mathbf{R}^{n(p)}$:*

$$\forall x, x' \ \forall \lambda \in [0,1]: \qquad p_0(\lambda x + (1-\lambda)x') \le \lambda p_0(x) + (1-\lambda)p_0(x').$$

One can immediately verify that (5.1.1) indeed defines a convex optimization problem.

Our interest in convex optimization programs, as opposed to other mathematical programming programs, comes from the following state of affairs:

> (!) *Convex optimization programs are computationally tractable: there exist solution methods that efficiently solve every convex optimization program satisfying very mild computability restrictions.*
>
> (!!) *In contrast to this, no efficient universal solution methods for nonconvex mathematical programming programs are known, and there are strong reasons to expect that no such methods exist.*

Note that even the first—the positive—statement (!) is at the moment just a claim, not a theorem. We have not yet defined what a solution method is and what efficiency means. The goal of this lecture is to clarify all these notions and to convert (!) to a theorem.

5.1.2 Solution methods and efficiency

Intuitively, a (numerical) solution method is a computer code. When solving a particular optimization program, a computer loaded with this code inputs the data of the program, executes the code, and outputs the result—an array of reals representing the solution, or the message "no solution exists". It is natural to measure the efficiency of such a solution method, for a particular program, by the running time of the code as applied to the data of the program, i.e., by the number of elementary operations performed by the computer when executing the code: the shorter the running time, the higher the efficiency.

When formalizing these intuitive considerations, we should specify a number of key ingredients, namely,

- *Model of computations.* What can our computer do; in particular, what are its elementary operations?
- *Encoding of program instances.* What are the programs we intend to solve, and what are the data of particular programs the computer works with?
- *Quality of solution.* What kind of solution do we expect to get? An exactly optimal or an approximate one? After all, even for simple convex programs, it is unrealistic

to expect that the data can be converted to an *exact* optimal solution in finitely many elementary operations. So, realistically we are content with only an approximate solution. This requires us to decide how to measure the quality of an approximate solution and makes it necessary to inform the computer on the quality of an approximation we wish to obtain.

We shall specify these elements in a way that is most convenient for our subject domain—optimization programs like linear, conic quadratic, and semidefinite ones.

Model of computations. This is what is known as real arithmetic model of computations. To avoid tedious formalities, we restrict ourselves with a kind of a semiformal description. We assume that the computations are carried out by an idealized version of the usual computer which is able to store countably many reals and can perform with them the standard exact real arithmetic operations—the four basic arithmetic operations, evaluating elementary functions, like cos and exp, and making comparisons. Thus, as far as the logical part of executing a code is concerned, we deal with the usual von Neumann computer, and the idealization is in the assumption that the data stored in memory are actual reals (not their floating point approximations) and that the operations with these reals are free of any rounding.

Families of optimization programs. We want to speak about methods for solving optimization programs (p) of a given structure, like linear, conic quadratic, or semidefinite ones. All programs (p) of a given structure form certain family $\mathcal{P}$, and we assume that every particular program in this family—every instance (p) of $\mathcal{P}$—is specified by its particular data $\mathrm{Data}(p)$ which is a finite-dimensional real vector. One may think about the entries of this data vector as about particular values of coefficients of generic (specific for $\mathcal{P}$) analytic expressions for $p_0(x)$ and $X(p)$. This approach is in full accordance with our intuition, as is seen from the following examples.

1. $\mathcal{LP}$. Here instances (p) are all possible LP programs

$$\min_x \left\{ p_0^T x : x \in X(p) = \{x : Ax - b \geq 0\} \right\}.$$

The data vector specifying a particular LP program (p) can be obtained by writing successively the dimensions n (number of variables) and m (number of constraints), next the n entries of the objective p_0, next the mn entries of the constraint matrix A (say, row by row), and finally the m entries of b.

2. $\mathcal{CQP}$. Here instances are all possible conic quadratic programs

$$\min_x \left\{ p_0^T x : x \in X(p) = \{x \mid \|D_i x - d_i\|_2 \leq e_i^T x - c_i,\ i = 1, \ldots, k\} \right\},$$

where D_i are $m_i \times \dim x$ matrices. A natural way to encode the data of a particular instance by a finite-dimensional data vector is to write successively the sizes $n = \dim x$ (design dimension), k (number of conic quadratic inequality constraints), next the integers $m_1, \ldots, m_k$ (image dimensions of the constraints), next the n entries of p_0, and finally the entries of (D_i, d_i, e_i, c_i), $i = 1, \ldots, k$.

3. $\mathcal{SDP}$. Here instances are all possible semidefinite programs

$$\min_x \left\{ p_0^T x : x \in X(p) = \left\{ x : \sum_{i=1}^n x_i A_i - B \succeq 0 \right\} \right\}$$

with $m \times m$ symmetric matrices $A_1, \ldots, A_n, B$. To encode the data of an instance by a finite-dimensional vector, we write successively n, m, then the n entries of p_0, and finally, row by row, the entries of the matrices $A_1, \ldots, A_n, B$.

We always assume that the first entry in Data(p) is the design dimension $n(p)$ of the instance. The dimension of the vector Data(p) will be called the *size* of the instance:

$$\text{Size}(p) = \dim \text{Data}(p).$$

Accuracy of approximate solutions. Consider a generic problem $\mathcal{P}$[45] and assume that we have somehow fixed an infeasibility measure of a vector $x \in \mathbf{R}^{n(p)}$ as a solution to an instance $(p) \in \mathcal{P}$; let this measure be denoted by $\text{Infeas}_{\mathcal{P}}(x, p)$. In our general considerations, all we require from this measure is that

- $\text{Infeas}_{\mathcal{P}}(x, p) \geq 0$, and $\text{Infeas}_{\mathcal{P}}(x, p) = 0$ when x is feasible for (p) (i.e., when $x \in X(p)$);
- $\text{Infeas}_{\mathcal{P}}(x, p)$ is a real-valued convex function of $x \in \mathbf{R}^{n(p)}$.

Examples are as follows:

1. $\mathcal{LP}$ *(continued).* A natural way to measure infeasibility of an $x \in \mathbf{R}^n$ as a candidate solution to an LP instance

$$(p): \qquad \min_x \left\{ p_0^T x : Ax - b \geq 0 \right\}$$

is to set

$$\text{Infeas}_{\mathcal{LP}}(x, p) = \min \{ t \geq 0 \mid Ax + te - b \geq 0 \}, \tag{5.1.2}$$

where e is the vector of ones of appropriate dimension. It is immediately seen that

$$\text{Infeas}_{\mathcal{LP}}(x, p) = \max \left[0, \max_{i=1,\ldots,m} [b_i - (Ax)_i] \right], \tag{5.1.3}$$

m being the dimension of the right-hand-side vector b. Thus, our infeasibility measure is just the maximum of violations of the linear constraints of the program at x.

2. $\mathcal{CQP}$ *(continued).* A natural way to measure infeasibility of an $x \in \mathbf{R}^n$ as a candidate solution to a CQP instance

$$(p): \qquad \min_x \left\{ c^T x : \|D_i x - d_i\|_2 \leq e_i^T x - c_i, \ i = 1, \ldots, k \right\}$$

[45] We use the words "generic program" as a synonym of "family of optimization programs."

is to set

$$\begin{aligned}\mathrm{Infeas}_{\mathcal{CQP}}(x,\,p) &= \min\left\{t\geq 0:\|D_ix-d_i\|_2\leq e_i^Tx-c_i+t,\ i=1,\ldots,k\right\}\\ &= \max\left[0,\,\max_{i=1,\ldots,k}[\|D_ix-d_i\|_2-e_i^Tx+c_i]\right].\end{aligned}\tag{5.1.4}$$

Note that geometrically this definition is very similar to the one we used for LP: to measure the violation of a vector inequality

$$Ax-b\geq_{\mathbf{K}}0\tag{V}$$

at a given point, we take a central interior point e of $\mathbf{K}$ and see what is the smallest nonnegative coefficient t such that after we add te to the left-hand side of (V), we get a valid vector inequality. In the case of LP we have $\mathbf{K}=\mathbf{R}^m_+$, the natural central point of the cone is the vector of ones, and we come to (5.1.2). In the case of $\mathbf{K}=\mathbf{L}^{m_1+1}\times\cdots\times\mathbf{L}^{m_k+1}$ the natural choice of e is $e=(e^1,\ldots,e^k)$, where $e^i\in\mathbf{R}^{m_i+1}$ has the first m_i coordinates equal to 0 and the last coordinate equal to 1 (i.e., e^i is the direction of the axis of symmetry of the corresponding Lorentz cone), and we come to (5.1.4).

3. *$\mathcal{SDP}$ (continued).* A natural way to measure infeasibility of an $x\in\mathbf{R}^n$ as a candidate solution to an instance

$$(p):\qquad \min_x\left\{c^Tx:\mathcal{A}x-B\equiv\sum_{i=1}^n x_iA_i-B\succeq 0\right\}$$

is to set

$$\mathrm{Infeas}_{\mathcal{SDP}}(x,\,p)=\min\{t\geq 0:\mathcal{A}x-B+tI\succeq 0\},\tag{5.1.5}$$

where I is the unit matrix of appropriate size. (We again have used the above construction, taking, as the "central interior point" of the semidefinite cone $\mathbf{S}^m_+$, the unit matrix of the corresponding size.)

4. *General convex programming problems.* Assume that the instances of a generic problem in question are of the form

$$(p):\quad \min_x\left\{p_0(x):x\in X(p)=\{x\in\mathbf{R}^{n(p)}:p_i(x)\leq 0,\ i=1,\ldots,m(p)\}\right\},$$

where $p_i(x):\mathbf{R}^{n(p)}\to\mathbf{R}$, $i=0,\ldots,m(p)$, are convex functions. Here a natural infeasibility measure is the maximum constraint violation

$$\mathrm{Infeas}(x,\,p)=\min\{t\geq 0:p_i(x)\leq t,\ i=1,\ldots,m(p)\}=\max\left[0,\,\max_{i=1,\ldots,m(p)}p_i(x)\right];\tag{5.1.6}$$

cf. (5.1.2), (5.1.4).

Given an infeasibility measure, we can proceed to define the notion of an ϵ-solution to an instance $(p)\in\mathcal{P}$, namely, as follows. Let $\mathrm{Opt}(p)\in\{-\infty\}\cup\mathbf{R}\cup\{+\infty\}$ be the optimal value of the instance (i.e., the infimum of the values of the objective on the feasible set, if

the instance is feasible, and $+\infty$ otherwise). A point $x \in \mathbf{R}^{n(p)}$ is called an ϵ-solution to (p) if

$$\mathrm{Infeas}_{\mathcal{P}}(x, p) \leq \epsilon \text{ and } p_0(x) - \mathrm{Opt}(p) \leq \epsilon,$$

i.e., if x is both ϵ-feasible and ϵ-optimal for the instance.

Solution methods. We are ready to name a solution method $\mathcal{M}$ for a given family $\mathcal{P}$ of optimization programs. By definition, this is a code for our idealized real arithmetic computer. When solving an instance $(p) \in \mathcal{P}$, the computer first inputs the data vector $\mathrm{Data}(p)$ of the instance and a real $\epsilon > 0$, the accuracy to which the instance should be solved, and then executes the code $\mathcal{M}$ on this input. We assume that the execution, on every input $(\mathrm{Data}(p), \epsilon > 0)$ with $(p) \in \mathcal{P}$, takes finitely many elementary operations of the computer, let this number be denoted by $\mathrm{Compl}_{\mathcal{M}}(p, \epsilon)$, and results in one of the following three possible outputs:

- an $n(p)$-dimensional vector $\mathrm{Res}_{\mathcal{M}}(p, \epsilon)$ that is an ϵ-solution to (p),
- a correct message "(p) is infeasible,"
- a correct message "(p) is unbounded below."

Now let us define the central notion of the complexity of a method $\mathcal{M}$. We have agreed to measure the efficiency of a method on the basis of the running time $\mathrm{Compl}_{\mathcal{M}}(p, \epsilon)$—the number of elementary operations performed by the method when solving instance (p) within accuracy ϵ. This characteristic, however, depends on a particular instance (p) and on ϵ. The crucial step in our formalization is to clarify what it means that $\mathcal{M}$ is an efficient (a polynomial time) on $\mathcal{P}$. By definition, it means that there exists a polynomial $\pi(s, \tau)$ such that

$$\mathrm{Compl}_{\mathcal{M}}(p, \epsilon) \leq \pi\left(\mathrm{Size}(p), \ln\left(\frac{\mathrm{Size}(p) + \|\mathrm{Data}(p)\|_1 + \epsilon^2}{\epsilon}\right)\right) \quad \forall (p) \in \mathcal{P}\ \forall \epsilon > 0; \tag{5.1.7}$$

here $\|u\|_1 = \sum_{i=1}^{\dim u} |u_i|$. This definition is by no means a self-evident way to formalize the common sense notion of an efficient computational method; however, there are strong reasons (taking their roots in combinatorial optimization) to use just this notion. Let us present a transparent common sense interpretation of our definition. The quantity

$$\mathrm{Digits}(p, \epsilon) = \ln\left(\frac{\mathrm{Size}(p) + \|\mathrm{Data}(p)\|_1 + \epsilon^2}{\epsilon}\right) \tag{5.1.8}$$

may be thought of as a number of accuracy digits in an ϵ-solution; at least this is the case when (p) is fixed and $\epsilon \to +0$, so that the numerator in the fraction in (5.1.8) becomes unimportant. With this interpretation, polynomiality of $\mathcal{M}$ means a very simple thing: when we increase the size of an instance and the required number of accuracy digits by absolute constant factors (say, by factor 2), the running time increases by no more than another absolute constant factor. Roughly speaking, when $\mathcal{M}$ is polynomial time, then

an improvement, by a constant factor, in the performance of our real arithmetic computer results in the possibility to increase by another constant factors the sizes of the instances we can process and the number of accuracy digits we can obtain in a fixed time (say, in 24 hours). In contrast to this, for a nonpolynomial-time method, say, one with complexity

$$\mathrm{Compl}_{\mathcal{M}}(p,\epsilon) \geq O\left(\exp\{\mathrm{Size}(p)\} f(\epsilon)\right),$$

no such conclusions can be deduced. Say, if our old computer was able to solve in 24 hours to three accuracy digits every instance of size 100 (or 1000), and now we get a computer that is 10 times faster, then all we can hope for is to solve in the same 24 hours to the same three accuracy digits all instances of the size 102 (respectively, 1002). Similarly, if we are using a method with complexity,

$$\mathrm{Compl}_{\mathcal{M}}(p,\epsilon) = O\left(f(\mathrm{Size}(p))\frac{1}{\epsilon}\right),$$

and with our old computer were able to get a number of digits in 24 hours on all instances of size 100, and now we get a computer that is 10 times faster, we achieve, on the same instances and in the same time, just one accuracy digit more. In both these examples, constant times progress in computer's performance improves just by additive constant the size of the instances we can process, or the number of accuracy digits we can obtain in a given time.

For typical polynomial time methods the upper bound (5.1.7) is in fact linear in the number of accuracy digits:

$$\mathrm{Compl}_{\mathcal{M}}(p,\epsilon) \leq \pi\left(\mathrm{Size}(p)\right)\mathrm{Digits}(p,\epsilon) \quad \forall (p) \in \mathcal{P}\ \forall \epsilon > 0.$$

Such a complexity bound admits even more transparent interpretation: the computational effort in solving problems from $\mathcal{P}$ is proportional to the number of accuracy digits we want to get, the proportionality coefficient (the price of an accuracy digit) being polynomial in the size of an instance.

The final point in our formalization is the notion of a polynomially solvable family $\mathcal{P}$ of optimization programs: $\mathcal{P}$ is called polynomially solvable if it admits a polynomial time solution method. Polynomial solvability of a generic optimization problem $\mathcal{P}$ is a theoretical synonym of computational tractability of $\mathcal{P}$. As far as computational practice is concerned, polynomiality of $\mathcal{P}$ is neither a necessary nor a sufficient condition for practical tractability of the instances of $\mathcal{P}$ (simply because there cannot exist conditions of this type—whatever the complexity bounds, the sizes of problems we can solve in practice are limited). Not every polynomial time method can be used in practice (think about a polynomial time method for solving LP programs with complexity proportional to $\mathrm{Size}^8(p)$). On the other hand, a theoretically nonpolynomial method is not necessarily bad in practice (the most famous example of this type is the Simplex method for LP)—the complexity is a worst-case-oriented notion, and perhaps we should not bother so much about the worst case in practice. The historical fact, however, is that for those generic optimization problems (first and foremost for LP) which were for a long time routinely solved by theoretically bad, although practically efficient, methods, eventually theoretically good *and* practically efficient methods were discovered. Thus, theoretical complexity studies finally do have strong impact on computational practice.

We are back on the route to our goal: to demonstrate that convex optimization problems are computationally tractable. At this point we better understand what should be proved. The proof itself, however, comes from a quite unexpected side—from considerations that bear no straightforward connection with the above chain of complexity-related definitions.

5.2 Black box–represented convex programs

Consider a convex program

$$\min_x \left\{ f(x) : x \in X \subset \mathbf{R}^n \right\}, \tag{5.2.9}$$

where $f : \mathbf{R}^n \to \mathbf{R}$ is a convex function and X is a closed and bounded convex set with a nonempty interior. Moreover, assume that we know in advance that X is neither too large nor too flat, namely, the following.

A. We are given in advance reals r, $R \in (0, \infty)$ such that X is contained in the center at the origin ball $\{x \mid \|x\|_2 \le R\}$ and contains a Euclidean ball $\{x \mid \|x - \bar{x}\|_2 \le r\}$. (Note that only the radius r of the small ball is known, not the center of the ball!)[46]

To proceed, we need to recall two important elements of convex analysis.

The separation theorem states that if X is a nonempty convex set in $\mathbf{R}^n$ and $x \in \mathbf{R}^n \setminus X$, then x can be separated from X by a hyperplane: there exists $a \neq 0$ such that

$$a^T x \ge \sup_{y \in X} a^T y. \tag{5.2.10}$$

A separation oracle $\mathrm{Sep}(X)$ for X is a routine (a black box) which, given as input a point $x \in \mathbf{R}^n$, checks whether this point belongs to X. If this is the case, the oracle reports that $x \in X$, and if $x \notin X$, $\mathrm{Sep}(X)$ reports that this is the case and returns a nonzero vector a which separates x and X in the sense of (5.2.10).

Subgradient. Let $f : \mathbf{R}^n \to \mathbf{R}$ be a convex function. A vector $\eta \in \mathbf{R}^n$ is called a subgradient of f at a point $x \in \mathbf{R}^n$ if

$$f(y) \ge f(x) + \eta^T (y - x) \quad \forall y \in \mathbf{R}^n. \tag{5.2.11}$$

In other words, a subgradient of f at x is the slope of an affine function which is $\le f$ everywhere and is equal to f at x. The set of all subgradients of f at a point x is denoted by $\partial f(x)$ and is called the *subdifferential* of f at x. A fundamental result of convex analysis is that *if* $f : \mathbf{R}^n \to \mathbf{R}$ is convex, then $\partial f(x) \neq \emptyset \ \forall x$.[47] It is easy to verify that if f is differentiable at x, then $\partial f(x)$ is a singleton comprised of the usual gradient of f at x.

[46] We could weaken to some extent our a priori knowledge; however, in our further applications the strong assumption **A** will be automatically satisfied.

[47] In fact, if f is an only partially defined convex function, then $\partial f(x)$ is nonempty at every point from the relative interior of the domain of f, and you can easily prove that statement by applying the separation theorem to the point $(x, f(x))$ and the convex set $\{(x, t) \mid t > f(x)\}$ in $\mathbf{R}^{n+1}$. In our context, however, we have no need to consider the case of a partially defined f.

A first order oracle $\mathcal{O}(f)$ for f is a routine (a black box) which, given as input a point $x \in \mathbf{R}^n$, returns as output the value $f(x)$ and a subgradient $f'(x) \in \partial f(x)$ of f at x.

Assume that we want to solve the convex program (5.2.9) and we have an access to a separation oracle $\mathrm{Sep}(X)$ for the feasible domain of (5.2.9) and to a first order oracle $\mathcal{O}(f)$ for the objective of (5.2.9). How could we solve the problem with these tools? An extremely transparent way is given by the ellipsoid method, which can be viewed as a multidimensional extension of the usual bisection.

Ellipsoid method: The idea. Assume that we have already found an n-dimensional ellipsoid

$$E = \{x = c + Bu \mid u^T u \le 1\} \qquad [B \in \mathbf{M}^{n,n}, \mathrm{Det}B \ne 0],$$

which contains the optimal set X_* of (5.2.9) (note that $X_* \ne \emptyset$, since the feasible set X of (5.2.9) is assumed to be compact and the objective f to be convex on the entire $\mathbf{R}^n$ and therefore continuous).[48] How could we construct a smaller ellipsoid containing X_* ?

The answer is immediate.

1. Let us call the separation oracle $\mathrm{Sep}(X)$, the center c of the current ellipsoid being the input. There are two possible cases:

1.a. $\mathrm{Sep}(X)$ reports that $c \notin X$ and returns a separator a:

$$a \ne 0, \quad a^T c \ge \sup_{y \in X} a^T y. \tag{5.2.12}$$

In this case we can replace our current localizer E of the optimal set X_* by a smaller one, namely, by the half-ellipsoid

$$\widehat{E} = \{x \in E \mid a^T x \le a^T c\}.$$

Indeed, by assumption $X_* \subset E$; when passing from E to $\widehat{E}$, we cut off all points x of E where $a^T x > a^T c$, and by (5.2.12) all these points are outside of X and therefore outside of $X_* \subset X$. Thus, $X_* \subset \widehat{E}$.

1.b. $\mathrm{Sep}(X)$ reports that $c \in X$. In this case we call the first order oracle $\mathcal{O}(f)$, c being the input; the oracle returns the value $f(c)$ and a subgradient $a \in \partial f(c)$ of f at c. Again, two cases are possible:

1.b.1. $a = 0$. In this case we are done—c is a minimizer of f on X. Indeed, $c \in X$, and (5.2.11) now reads

$$f(y) \ge f(c) + 0^T(y - c) = f(c) \quad \forall y \in \mathbf{R}^n.$$

Thus, c is a minimizer of f on $\mathbf{R}^n$, and since $c \in X$, c minimizes f on X as well.

1.b.2. $a \ne 0$. In this case (5.2.11) reads

$$a^T(x - c) > 0 \Rightarrow f(x) > f(c),$$

[48] A simple fact (try to prove it) is that a function that is convex in a neighborhood of a point x is continuous in this neighborhood.

so that replacing the ellipsoid E with the half-ellipsoid

$$\widehat{E} = \{x \in E \mid a^T x \le a^T c\}$$

we ensure the inclusion $X_* \subset \widehat{E}$. Indeed, $X_* \subset E$ by assumption, and when passing from E to $\widehat{E}$, we cut off all points of E where $a^T x > a^T c$ and, consequently, where $f(x) > f(c)$; since $c \in X$, no one of these points can belong to the set X_* of minimizers of f on X.

2. We have seen that as a result of operations described in 1.a-b we either terminate with an exact minimizer of f on X or obtain a half-ellipsoid

$$\widehat{E} = \{x \in E \mid a^T x \le a^T c\} \qquad [a \ne 0]$$

containing X_*. It remains to use the following simple geometric fact:

(*) *Let $E = \{x = c + Bu \mid u^T u \le 1\}$ ($\mathrm{Det} B \ne 0$) be an n-dimensional ellipsoid and $\widehat{E} = \{x \in E \mid a^T x \le a^T c\}$ ($a \ne 0$) be a half of E. If $n > 1$, then $\widehat{E}$ is contained in the ellipsoid*

$$\begin{gathered} E^+ = \{x = c^+ + B^+ u \mid u^T u \le 1\}, \\ c^+ = c - \tfrac{1}{n+1} Bp, \\ B^+ = B\left(\tfrac{n}{\sqrt{n^2-1}}(I_n - pp^T) + \tfrac{n}{n+1} pp^T\right) = \tfrac{n}{\sqrt{n^2-1}} B + \left(\tfrac{n}{n+1} - \tfrac{n}{\sqrt{n^2-1}}\right)(Bp)p^T, \\ p = \tfrac{B^T a}{\sqrt{a^T B B^T a}}, \end{gathered} \tag{5.2.13}$$

and if $n = 1$, then the set $\widehat{E}$ is contained in the ellipsoid (which now is just a segment)

$$\begin{gathered} E^+ = \{x \mid c^+ B^+ u \mid |u| \le 1\}, \\ c^+ = c - \tfrac{1}{2} \tfrac{Ba}{|Ba|}, \\ B_+ = \tfrac{1}{2} B. \end{gathered}$$

In all cases, the n-dimensional volume $\mathrm{Vol}(E^+)$ of the ellipsoid E^+ is less than the one of E:

$$\mathrm{Vol}(E^+) = \left(\frac{n}{\sqrt{n^2-1}}\right)^{n-1} \frac{n}{n+1} \mathrm{Vol}(E) \le \exp\{-1/(2n)\} \mathrm{Vol}(E) \tag{5.2.14}$$

(in the case of $n = 1$, $\left(\tfrac{n}{\sqrt{n^2-1}}\right)^{n-1} = 1$).

(*) says that there exists (and can be explicitly specified) an ellipsoid $E^+ \supset \widehat{E}$ with the volume constant times less than the one of E. Since E^+ covers $\widehat{E}$, and the latter set, as we have seen, covers X_*, E^+ covers X_*. Now we can iterate the above construction, thus obtaining a sequence of ellipsoids $E, E^+, (E^+)^+, \ldots$ with volumes going to 0 at a linear rate (depending on the dimension n only) which collapses to the set X_* of optimal solutions of our problem—exactly as in the usual bisection!

Note that (*) is just an exercise in elementary calculus. Indeed, the ellipsoid E is given as an image of the unit Euclidean ball $W = \{u \mid u^T u \le 1\}$ under the one-to-one affine

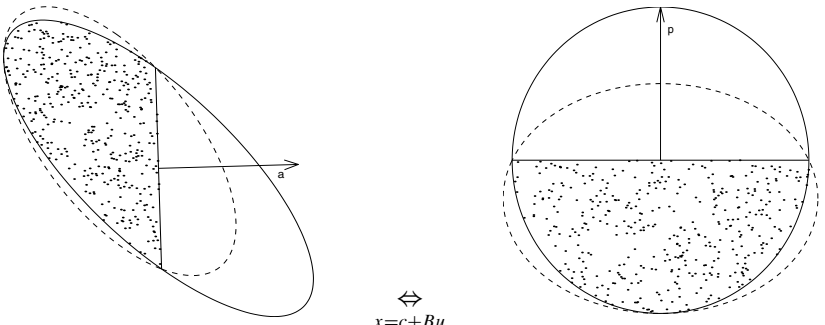


Figure 5.1.

mapping $u \mapsto c + Bu$; the half-ellipsoid $\widehat{E}$ is then the image, under the same mapping, of the half-ball

$$\widehat{W} = \{u \in W \mid p^T u \leq 0\},$$

where p is the unit vector from (5.2.13). Indeed, if $x = c + Bu$, then $a^T x \leq a^T c$ if and only if $a^T Bu \leq 0$ or, which is the same, if and only if $p^T u \leq 0$. Now, instead of covering $\widehat{E}$ by a small-in-volume ellipsoid E^+, we may cover by a small ellipsoid W^+ the half-ball $\widehat{W}$ and then take E^+ to be the image of W^+ under our affine mapping:

$$E^+ = \{x = c + Bu \mid u \in W^+\}. \tag{5.2.15}$$

Indeed, if W^+ contains $\widehat{W}$, then the image of W^+ under our affine mapping $u \mapsto c + Bu$ contains the image of $\widehat{W}$, i.e., contains $\widehat{E}$. And since the ratio of volumes of two bodies remain invariant under affine mapping (passing from a body to its image under an affine mapping $u \mapsto c + Bu$, we just multiply the volume by $|\mathrm{Det}B|$), we have

$$\frac{\mathrm{Vol}(E^+)}{\mathrm{Vol}(E)} = \frac{\mathrm{Vol}(W^+)}{\mathrm{Vol}(W)}.$$

Thus, the problem of finding a small ellipsoid E^+ containing the half-ellipsoid $\widehat{E}$ can be reduced to the one of finding a small ellipsoid W^+ containing the half-ball $\widehat{W}$, as shown in Fig. 5.1. Now, the problem of finding a small ellipsoid containing $\widehat{W}$ is very simple: our geometric data are invariant with respect to rotations around the p-axis, so that we may look for W^+ possessing the same rotational symmetry. Such an ellipsoid W^+ is given by just three parameters: its center should belong to our symmetry axis, i.e., should be $-hp$ for certain h, one of the half-axes of the ellipsoid (let its length be μ) should be directed along p, and the remaining $n-1$ half-axes should be of the same length λ and be orthogonal to p. For our three parameters h, μ, λ we have two equations expressing the fact that the boundary of W^+ should pass through the South Pole $-p$ of W and through the equator $\{u \mid u^T u = 1, p^T u = 0\}$; indeed, W^+ should contain $\widehat{W}$ and thus both the pole and the equator, and since we are looking for W^+ with the smallest possible volume, both the pole and the equator should be on the boundary of W^+. Using our two equations to

express μ and λ via h, we end up with a single free parameter h, and the volume of W^+ (i.e., $\mathrm{const}(n)\mu\lambda^{n-1}$) becomes an explicit function of h. Minimizing this function in h, we find the optimal ellipsoid W^+, check that it indeed contains $\widehat{W}$ (i.e., that our geometric intuition was correct), and then convert W^+ into E^+ according to (5.2.15), thus coming to the explicit formulas (5.2.13)–(5.2.14). Implementation of the outlined scheme takes from 10 to 30 minutes, depending on how many miscalculations are made.

It should be mentioned that although the indicated scheme is quite straightforward and elementary, the fact that it works is not evident a priori: it might happen that the smallest-volume ellipsoid containing a half-ball is just the original ball! This would be the death of our idea—instead of a sequence of ellipsoids collapsing to the solution set X_*, we would get a stationary sequence E, E, E. Fortunately, it is not happening, and this is a great favor Nature does for convex optimization.

Ellipsoid method: The construction. There is a small problem with implementing our idea of trapping the optimal set X_* of (5.2.9) by a collapsing sequence of ellipsoids. The only thing we can ensure is that all our ellipsoids contain X_* and that their volumes rapidly (at a linear rate) converge to 0. However, the linear sizes of the ellipsoids should not necessarily go to 0—it may happen that the ellipsoids are shrinking in some directions and are not shrinking (or even become larger) in other directions. (Look what happens if we apply the construction to minimizing a function of two variables which in fact depends only on the first coordinate.) Thus, for the moment it is unclear how to build a sequence of points converging to X_*. This difficulty, however, can be easily resolved: as we shall see, we can form this sequence from the best feasible solutions generated so far. Another issue that remains open at the moment is when to terminate the method. As we shall see in a while, this issue also can be settled satisfactory.

The precise description of the ellipsoid method as applied to (5.2.9) is as follows (in this description, we assume that $n \geq 2$, which of course does not restrict generality):

The Ellipsoid Method

Initialization. Recall that when formulating (5.2.9) it was assumed that the feasible set X of the problem is contained in the ball $E_0 = \{x \mid \|x\|_2 \leq R\}$ of a given radius R and contains an (unknown) Euclidean ball of a known radius $r > 0$. The ball E_0 will be our initial ellipsoid. Thus, we set

$$c_0 = 0, \;\; B_0 = RI, \;\; E_0 = \{x = c_0 + B_0 u \mid u^T u \leq 1\};$$

note that $E_0 \supset X$.

We also set

$$\rho_0 = R, \;\; L_0 = 0.$$

The quantities ρ_t will be the radii of the ellipsoids E_t to be built, i.e., the radii of the Euclidean balls of the same volumes as E_t's. The quantities L_t will be our guesses for the variation

$$\mathrm{Var}_R(f) = \max_{x \in E_0} f(x) - \min_{x \in E_0} f(x)$$

of the objective on the initial ellipsoid E_0. We shall use these guesses in the termination test.

Finally, we input the accuracy $\epsilon > 0$ to which we want to solve the problem.

Step t, $t = 1, 2, \ldots$. At the beginning of step t, we have the previous ellipsoid

$$E_{t-1} = \{x = c_{t-1} + B_{t-1}u \mid u^T u \le 1\} \qquad [c_{t-1} \in \mathbf{R}^n,\ B_{t-1} \in \mathbf{M}^{n,n},\ \mathrm{Det}B_{t-1} \neq 0]$$

(i.e., have c_{t-1}, B_{t-1}) along with the quantities $L_{t-1} \ge 0$ and

$$\rho_{t-1} = |\mathrm{Det}B_{t-1}|^{1/n}.$$

At step t, we act as follows (cf. the preliminary description of the method):

1. We call the separation oracle $\mathrm{Sep}(X)$, c_{t-1} the input. It is possible that the oracle reports that $c_{t-1} \notin X$ and provides us with a separator

$$a \neq 0: \quad a^T c_{t-1} \ge \sup_{y \in X} a^T y.$$

In this case we call step t nonproductive; set

$$a_t = a,\ L_t = L_{t-1}$$

and go to rule 3 below. Otherwise—i.e., when $c_{t-1} \in X$—we call step t productive and go to rule 2.

2. We call the first order oracle $\mathcal{O}(f)$, c_{t-1} the input, and we get the value $f(c_{t-1})$ and a subgradient $a \equiv f'(c_{t-1}) \in \partial f(c_{t-1})$ of f at the point c_{t-1}. It is possible that $a = 0$. In this case we terminate and claim that c_{t-1} is an optimal solution to (5.2.9). In the case of $a \neq 0$ we set

$$a_t = a,$$

compute the quantity

$$\ell_t = \max_{y \in E_0}[a_t^T y - a_t^T c_{t-1}] = R\|a_t\|_2 - a_t^T c_{t-1},$$

update L by setting

$$L_t = \max\{L_{t-1}, \ell_t\},$$

and go to rule 3.

3. We set

$$\widehat{E}_t = \{x \in E_{t-1} \mid a_t^T x \le a_t^T c_{t-1}\}$$

(cf. the definition of $\widehat{E}$ in our preliminary description of the method) and define the new ellipsoid

$$E_t = \{x = c_t + B_t u \mid u^T u \le 1\}$$

by setting (see (5.2.13))

$$\begin{array}{rcl} p_t & = & \frac{B_{t-1}^T a_t}{\sqrt{a_t^T B_{t-1} B_{t-1}^T a_t}}, \\ c_t & = & c_{t-1} - \frac{1}{n+1} B_{t-1} p_t, \\ B_t & = & \frac{n}{\sqrt{n^2-1}} B_{t-1} + \left(\frac{n}{n+1} - \frac{n}{\sqrt{n^2-1}} \right) (B_{t-1} p_t) p_t^T. \end{array} \tag{5.2.16}$$

We also set

$$\rho_t = |\mathrm{Det} B_t|^{1/n} = \left(\frac{n}{\sqrt{n^2-1}} \right)^{(n-1)/n} \left(\frac{n}{n+1} \right)^{1/n} \rho_{t-1}$$

(see (5.2.14)) and go to rule 4.

4. (termination test) We check whether the inequality

$$\frac{\rho_t}{r} < \frac{\epsilon}{L_t + \epsilon} \tag{5.2.17}$$

is satisfied. If it is the case, we terminate and output, as the result of the solution process, the best (i.e., with the smallest value of f) of the search points $c_{\tau-1}$ associated with productive steps $\tau \le t$. (We shall see that these productive steps indeed exist, so that the result of the solution process is well defined.) If (5.2.17) is not satisfied, we go to step $t+1$.

To get some feeling how the method works, here is a 2D illustration. The problem is

$$\begin{array}{rcl} & & \min\limits_{-1 \le x_1, x_2 \le 1} f(x), \\ f(x) & = & \frac{1}{2}(1.443508244 x_1 + 0.623233851 x_2 - 7.957418455)^2 \\ & & +5(-0.350974738 x_1 + 0.799048618 x_2 + 2.877831823)^4, \end{array}$$

the optimal solution is $x_1^* = 1$, $x_2^* = -1$, and the exact optimal value is 70.030152768.

The values of f at the best (i.e., with the smallest value of the objective) feasible solutions found in course of first t steps of the method, $t = 1, 2, \ldots, 256$, are shown in the following table:

t	Best value	t	Best value
1	374.61091739	16	76.838253451
2	216.53084103	…	…
3	146.74723394	32	70.901344815
4	112.42945457	…	…
5	93.84206347	64	70.031633483
6	82.90928589	…	…
7	82.90928589	128	70.030154192
8	82.90928589	…	…
…	…	256	70.030152768

The initial phase of the process looks as shown in Fig. 5.2.

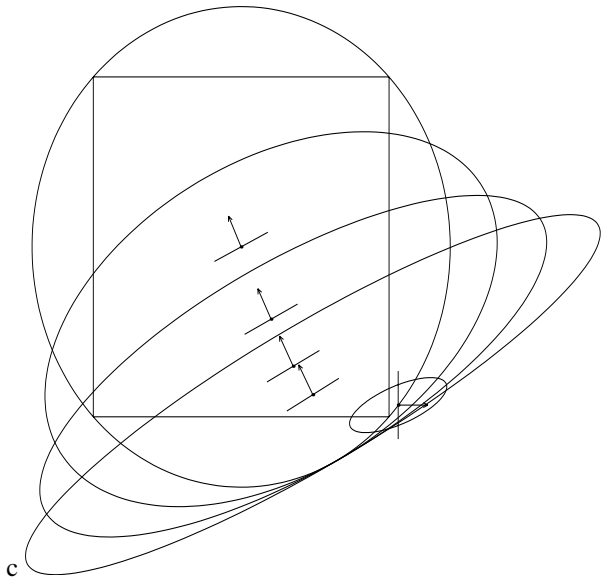


Figure 5.2. *Ellipses E_{t-1} and search points c_{t-1}, $t = 1, 2, 3, 4, 16$. Arrows: gradients of the objective $f(x)$; unmarked segments: tangents to the level lines of $f(x)$.*

Ellipsoid method: Complexity analysis. We are about to establish our key result.

Theorem 5.2.1. *Let the ellipsoid method be applied to convex program* (5.2.9) *of dimension $n \geq 2$ such that the feasible set X of the problem contains a Euclidean ball of a given radius $r > 0$ and is contained in the ball $E_0 = \{\|x\|_2 \leq R\}$ of a given radius R. For every input accuracy $\epsilon > 0$, the ellipsoid method terminates after no more than*

$$N(\epsilon) = \mathrm{Ceil}\left(2n^2\left[\ln\left(\frac{R}{r}\right) + \ln\left(\frac{\epsilon + \mathrm{Var}_R(f)}{\epsilon}\right)\right]\right) + 1 \tag{5.2.18}$$

steps, where

$$\mathrm{Var}_R(f) = \max_{E_0} f - \min_{E_0} f,$$

and $\mathrm{Ceil}(a)$ *is the smallest integer $\geq a$. Moreover, the result $\widehat{x}$ generated by the method is a feasible ϵ-solution to* (5.2.9):

$$\widehat{x} \in X \text{ and } f(x) - \min_X f \leq \epsilon. \tag{5.2.19}$$

Proof. We should prove the following pair of statements:

(i) The method terminates in course of the first $N(\epsilon)$ steps.

(ii) The result $\widehat{x}$ is a feasible ϵ-solution to the problem.

1*. Comparing the preliminary and the final description of the ellipsoid method and taking into account the initialization rule, we see that if the method does not terminate before step t or terminates at this step according to rule 4, then

$$\begin{array}{lrcl}
\text{(a)} & E_0 & \supset & X; \\
\text{(b)} & E_\tau & \supset & \widehat{E}_\tau = \left\{x \in E_{\tau-1} \mid a_\tau^T x \le a_\tau^T c_{\tau-1}\right\},\ \tau = 1, \dots, t; \\
\text{(c)} & \mathrm{Vol}(E_\tau) & = & \rho_\tau^n \mathrm{Vol}(E_0) = \left(\frac{n}{\sqrt{n^2-1}}\right)^{n-1} \frac{n}{n+1} \mathrm{Vol}(E_{\tau-1}) \\
 & & \le & \exp\{-1/(2n)\}\mathrm{vol}(E_{\tau-1}),\ \tau = 1, \dots, t.
\end{array} \tag{5.2.20}$$

Note that from (c) it follows that

$$\rho_\tau \le \exp\{-\tau/(2n^2)\}R,\ \tau = 1, \dots, t. \tag{5.2.21}$$

2^*. We claim that

If the ellipsoid method terminates at certain step t, then the result $\widehat{x}$ is well defined and is a feasible ϵ-solution to (5.2.9).

Indeed, there are only two possible reasons for termination. First, it may happen that $c_{t-1} \in X$ and $f'(c_{t-1}) = 0$ (see rule 2)). From our preliminary considerations we know that in this case c_{t-1} is an optimal solution to (5.2.9), which is even more than what we have claimed. Second, it may happen that at step t relation (5.2.17) is satisfied. Let us prove that the claim of 2^* takes place in this case as well.

2^*a. Let us set

$$\nu = \frac{\epsilon}{\epsilon + L_t} \in (0, 1].$$

By (5.2.17), we have $\rho_t/r < \nu$, so that there exists ν' such that

$$\frac{\rho_t}{r} < \nu' < \nu \quad [\le 1]. \tag{5.2.22}$$

Let x_* be an optimal solution to (5.2.9) and let X^+ be the ν'-shrinkage of X to x_*:

$$X^+ = x_* + \nu'(X - x_*) = \{x = (1 - \nu')x_* + \nu' z \mid z \in X\}. \tag{5.2.23}$$

We have

$$\mathrm{Vol}(X^+) = (\nu')^n \mathrm{Vol}(X) \ge \left(\frac{r\nu'}{R}\right)^n \mathrm{Vol}(E_0) \tag{5.2.24}$$

(the last inequality is given by the fact that X contains a Euclidean ball of the radius r), while

$$\mathrm{Vol}(E_t) = \left(\frac{\rho_t}{R}\right)^n \mathrm{Vol}(E_0) \tag{5.2.25}$$

by definition of ρ_t. Comparing (5.2.24), (5.2.25) and taking into account that $\rho_t < r\nu'$ by (5.2.22), we conclude that $\mathrm{Vol}(E_t) < \mathrm{Vol}(X^+)$ and, consequently, X^+ cannot be contained in E_t. Thus, there exists a point y which belongs to X^+:

$$y = (1 - \nu')x_* + \nu' z \qquad [z \in X], \tag{5.2.26}$$

and does *not* belong to E_t.

2*b. Since y does not belong to E_t and at the same time belongs to $X \subset E_0$ along with x_* and z (X is convex!), we see that there exists a $\tau \le t$ such that $y \in E_{\tau-1}$ and $y \notin E_\tau$. By (5.2.20)(b), every point x from the complement of E_τ in $E_{\tau-1}$ satisfies the relation $a_\tau^T x > a_\tau^T c_{\tau-1}$. Thus, we have

$$a_\tau^T y > a_\tau^T c_{\tau-1}. \tag{5.2.27}$$

2*c. Observe that the step τ is surely productive. Indeed, otherwise, by construction of the method, a_t would separate X from $c_{\tau-1}$, and (5.2.27) would be impossible (we know that $y \in X$!). Notice that in particular we have just proved that if the method terminates at a step t, then at least one of the steps $1, \dots, t$ is productive, so that the result is well defined.

Since step τ is productive, a_τ is a subgradient of f at $c_{\tau-1}$ (description of the method!), so that

$$f(u) \ge f(c_{\tau-1}) + a_\tau^T (u - c_{\tau-1})$$

$\forall u \in X$, and in particular for $u = x_*$. On the other hand, $z \in X \subset E_0$, so that by the definition of ℓ_τ and L_τ we have

$$a_\tau^T (z - c_{\tau-1}) \le \ell_\tau \le L_\tau.$$

Thus,

$$\begin{array}{l} f(x_*) \ge f(c_{\tau-1}) + a_\tau^T (x_* - c_{\tau-1}), \\ L_\tau \ge a_\tau^T (z - c_{\tau-1}). \end{array}$$

Multiplying the first inequality by $(1-\nu')$, the second – by ν', and adding the results, we get

$$\begin{array}{rcl} (1-\nu') f(x_*) + \nu' L_\tau & \ge & (1-\nu') f(c_{\tau-1}) + a_\tau^T ([(1-\nu')x_* + \nu' z] - c_{\tau-1}) \\ & = & (1-\nu') f(c_{\tau-1}) + a_\tau^T (y - c_{\tau-1}) \\ & & \text{[see (5.2.26)]} \\ & \ge & (1-\nu') f(c_{\tau-1}) \\ & & \text{[see (5.2.27)]} \end{array}$$

and we come to

$$\begin{array}{rcl} f(c_{\tau-1}) & \le & f(x_*) + \frac{\nu' L_\tau}{1-\nu'} \\ & \le & f(x_*) + \frac{\nu' L_t}{1-\nu'} \\ & & \text{[since } L_\tau \le L_t \text{ in view of } \tau \le t\text{]} \\ & \le & f(x_*) + \epsilon \\ & & \text{[by definition of } \nu \text{ and since } \nu' < \nu\text{]} \\ & = & \mathrm{Opt(C)} + \epsilon. \end{array}$$

We see that there exists a productive (i.e., with feasible $c_{\tau-1}$) step $\tau \le t$ such that the corresponding search point $c_{\tau-1}$ is ϵ-optimal. Since we are in the situation where the result $\widehat{x}$ is the best of the feasible search points generated in course of the first t steps, $\widehat{x}$ is also feasible and ϵ-optimal, as claimed in 2*.

3*. It remains to verify that the method does terminate in course of the first $N = N(\epsilon)$ steps. Assume, on the contrary, that it is not the case, and let us lead this assumption to a contradiction.

First, observe that for every productive step t we have

$$c_{t-1} \in X \text{ and } a_t = f'(c_{t-1}),$$

whence, by the definition of a subgradient and the variation $\mathrm{Var}_R(f)$,

$$u \in E_0 \Rightarrow \mathrm{Var}_R(f) \geq f(u) - f(c_{t-1}) \geq a_t^T(u - c_{t-1}),$$

whence

$$\ell_t \equiv \max_{u \in E_0} a_t^T(u - c_{t-1}) \leq \mathrm{Var}_R(f).$$

Looking at the description of the method, we conclude that

$$L_t \leq \mathrm{Var}_R(f) \qquad \forall t. \tag{5.2.28}$$

Since we have assumed that the method does not terminate in course of the first N steps, we have

$$\frac{\rho_N}{r} \geq \frac{\epsilon}{\epsilon + L_N}. \tag{5.2.29}$$

The right-hand side in this inequality is $\geq \epsilon/(\epsilon + \mathrm{Var}_R(f))$ by (5.2.28), while the left-hand side is $\leq \exp\{-N/(2n^2)\}R$ by (5.2.21). We get

$$\exp\{-N/(2n^2)\}R/r \geq \frac{\epsilon}{\epsilon + \mathrm{Var}_R(f)} \Rightarrow N \leq 2n^2\left[\ln\left(\frac{R}{r}\right) + \ln\left(\frac{\epsilon + \mathrm{Var}_R(f)}{\epsilon}\right)\right],$$

which is the desired contradiction (see the definition of $N = N(\epsilon)$ in (5.2.18)). □

5.3 Polynomial solvability of convex programming

Equipped with the ellipsoid method, we are ready to formulate the mild assumptions under which a family $\mathcal{P}$ of convex optimization programs is polynomially solvable. Our assumptions are those of polynomial computability, polynomial growth, and polynomial boundedness of feasible sets. When formulating these assumptions, we shall associate with $\mathcal{P}$ a number of positive characteristic constants; their particular values are of no importance—the only thing that counts is their existence. To simplify notation, we denote all these constants by the same symbol χ, so that this symbol in different places of even the same equation may have different values (cf. the usual conventions on how one interprets symbols like $o(1)$).

Polynomial computability. Let $\mathcal{P}$ be a family of convex optimization programs, and let $\mathrm{Infeas}_{\mathcal{P}}(x, p)$ be the corresponding measure of infeasibility of candidate solutions. We say that our family is polynomially computable if there exist two codes $\mathcal{C}_{\mathrm{obj}}$, $\mathcal{C}_{\mathrm{cons}}$ for the real arithmetic computer such that

1. For every instance $(p) \in \mathcal{P}$, the computer, when given as input the data vector of the instance (p) and a point $x \in \mathbf{R}^{n(p)}$ and executing the code $\mathcal{C}_{\rm obj}$, outputs the value $p_0(x)$ and a subgradient $e(x) \in \partial p_0(x)$ of the objective p_0 of the instance at the point x, and the running time (i.e., total number of operations) of this computation $T_{\rm obj}(x, p)$ is bounded from above by a polynomial of the size of the instance

$$\forall \left((p) \in \mathcal{P}, x \in \mathbf{R}^{n(p)}\right): \quad T_{\rm obj}(x, p) \le \chi \,{\rm Size}^{\chi}(p) \quad [{\rm Size}(p) = \dim {\rm Data}(p)]. \tag{5.3.30}$$

2. For every instance $(p) \in \mathcal{P}$, the computer, when given as input the data vector of the instance (p), a point $x \in \mathbf{R}^{n(p)}$, and an $\epsilon > 0$ and executing the code $\mathcal{C}_{\rm cons}$, reports as output whether ${\rm Infeas}_{\mathcal{P}}(x, p) \le \epsilon$ and, if it is not the case, outputs a linear form a that separates the point x from all those points y where ${\rm Infeas}_{\mathcal{P}}(y, p) \le \epsilon$:

$$\forall\, (y, {\rm Infeas}_{\mathcal{P}}(y, p) \le \epsilon): \qquad a^T x > a^T y, \tag{5.3.31}$$

and the running time $T_{\rm cons}(x, \epsilon, p)$ of the computation is bounded by a polynomial of the size of the instance and of the number of accuracy digits:

$$\forall \left((p) \in \mathcal{P}, x \in \mathbf{R}^{n(p)}, \epsilon > 0\right): \quad T_{\rm cons}(x, \epsilon, p) \le \chi\, ({\rm Size}(p) + {\rm Digits}(p, \epsilon))^{\chi}\,. \tag{5.3.32}$$

Note that the vector a in (5.3.31) is not supposed to be nonzero; when it is 0, (5.3.31) simply says that there are no points y with ${\rm Infeas}_{\mathcal{P}}(y, p) \le \epsilon$.

Polynomial growth. We say that a family $\mathcal{P}$ of convex programs equipped with an infeasibility measure ${\rm Infeas}_{\mathcal{P}}(x, p)$ is a family with polynomial growth if the objectives and the infeasibility measures, as functions of x, grow polynomially with $\|x\|_1$, the degree of the polynomial being a power of ${\rm Size}(p)$:

$$\begin{aligned} &\forall \left((p) \in \mathcal{P}, x \in \mathbf{R}^{n(p)}\right): \\ &|p_0(x)| + {\rm Infeas}_{\mathcal{P}}(x, p) \le (\chi\, [{\rm Size}(p) + \|x\|_1 + \|{\rm Data}(p)\|_1])^{\left(\chi {\rm Size}^{\chi}(p)\right)}\,. \end{aligned} \tag{5.3.33}$$

Examples. Let us verify that the families of linear, conic quadratic, and semidefinite programs equipped with the corresponding infeasibility measures (see section 5.1.2) are polynomially computable with polynomial growth.

1. *$\mathcal{LP}$ (continued).* Polynomial computability means that given the data (n, m, p_0, A, b) of an LP instance

$$(p): \qquad \min_x \left\{p_0^T x : Ax - b \ge 0\right\} \quad [A : m \times n]$$

and given $x \in \mathbf{R}^n$, $\epsilon > 0$, we are able to compute efficiently, in the aforementioned sense,

(a) the value $p_0(x) = p_0^T x$ of the objective at x,

(b) the subgradient of the objective $p_0(\cdot)$ at x (which is just p_0 !),

(c) whether

$$\mathrm{Infeas}_{\mathcal{LP}}(x,\,p)=\max\left[0,\,\max_{i=1,\ldots,m}[b_i-(Ax)_i]\right]\le\epsilon,$$

and, if it is not the case, to compute a vector a such that

$$a^Tx>a^Ty\quad\forall\,(y:\mathrm{Infeas}_{\mathcal{LP}}(y)\le\epsilon\}\,. \tag{5.3.34}$$

The straightforward implementation of (a) and (b) requires χn operations, and, of course,

$$n\le\mathrm{Size}(p)=\dim\mathrm{Data}(P)=1+n+nm+m.$$

Thus, we have no problems with $\mathcal{C}_{\mathrm{obj}}$.

To build $\mathcal{C}_{\mathrm{cons}}$, let us compute straightforwardly the value of $\mathrm{Infeas}_{\mathcal{LP}}(x,\,p)$ according to the explicit formula for this particular infeasibility measure. This computation requires just a number of arithmetic operations that is linear in $\mathrm{Size}(p)$. If the resulting value of the infeasibility measure is $>\epsilon$, so that we should compute a separator a, we also have not that much to do: we are in a situation in which the residual $b_{\widehat{i}}-(Ax)_{\widehat{i}}$ of one of our constraints is $>\epsilon$, and we can take as a the minus $\widehat{i}$th row of the constraint matrix. Indeed, with this choice

$$a^Tx=-(Ax)_{\widehat{i}}>\epsilon-b_{\widehat{i}},$$

while for every candidate solution y with $\mathrm{Infeas}_{\mathcal{LP}}(y,\,p)\le\epsilon$ we have

$$a^Ty=-(Ay)_{\widehat{j}}\le\epsilon-b_{\widehat{i}}.$$

Thus, both $\mathcal{C}_{\mathrm{obj}}$ and $\mathcal{C}_{\mathrm{cons}}$ can be chosen to have running time just $\chi\,\mathrm{Size}(p)$.

The fact that $\mathcal{LP}$ is of polynomial growth is evident.

2. *$\mathcal{CQP}$ (continued).* Here the instances are

$$(p):\qquad\min_x\left\{p_0^Tx:\|D_ix-d_i\|_2\le e_i^Tx-c_i,\;\;i=1,\ldots,k\right\},$$

and the infeasibility measure is

$$\begin{array}{rcl}\mathrm{Infeas}_{\mathcal{CQP}}(x,\,p) & = & \max\left[0,\,\max\limits_{i=1,\ldots,k}[\|D_ix-d_i\|_2-e_i^Tx+c_i]\right]\\ & = & \max\left[0,\,\max\limits_{i=1,\ldots,k}p_i(x)\right],\\ p_i(x) & = & \|D_ix-d_i\|_2-e_i^Tx+c_i,\;\;i=1,\ldots,k.\end{array} \tag{5.3.35}$$

To verify that the family is polynomially computable, let us denote by m_i the number of rows in D_i and by n the dimension of x. Observe that

$$\mathrm{Size}(p)\ge n+\sum_{i=1}^k(m_i+1)(n+1).$$

(The right-hand side is the total number of entries in p_0 and all collections (D_i, d_i, e_i, c_i), $i = 1, \ldots, k$.) Now we can build $\mathcal{C}_{\text{obj}}$ and $\mathcal{C}_{\text{cons}}$ as follows. Given x and ϵ, we can straightforwardly compute the value $p_0(x) = p_0^T x$ of the objective at x, the subgradient p_0 of the objective, and the value of the infeasibility measure $\text{Infeas}_{\mathcal{CQP}}(x)$ at x in $\chi \text{Size}(p)$ operations. After $\text{Infeas}_{\mathcal{CQP}}(x, p)$ is computed, we check whether this quantity is $> \epsilon$. If it is the case, we should also build a separator a. To this end let us look at the largest (at x) of the constraints $p_i(x)$ (see (5.3.35)); let $\widehat{i}$ be its index. By (5.3.35), the relation $\text{Infeas}_{\mathcal{CQP}}(x, p) > \epsilon$ means exactly that $p_{\widehat{i}}(x) > \epsilon$, while for every y with $\text{Infeas}_{\mathcal{CQP}}(y, p) \le \epsilon$ we have $p_{\widehat{i}}(y) \le \epsilon$. It follows that we can choose as a any subgradient of $p_{\widehat{i}}(\cdot)$ at the point x, since then

$$\text{Infeas}_{\mathcal{CQP}}(y, p) \le \epsilon \Rightarrow p_{\widehat{i}}(y) < p_{\widehat{i}}(x) \Rightarrow a^T y < a^T x,$$

the last $\Rightarrow$ being given by the definition of a subgradient:

$$a \in \partial f(y) \Rightarrow f(y) \ge f(x) + a^T(y - x) \ \forall y \Leftrightarrow a^T(x - y) \ge f(x) - f(y) \ \forall y.$$

On the other hand, a subgradient of $p_i(\cdot)$ is easy to compute. Indeed, we have $p_i(x) = \|D_i x - d_i\|_2 - e_i^T x + c_i$. If x is such that $D_i x - d_i \neq 0$, then p_i is differentiable at x, so that its subgradient at x is the usual gradient

$$\nabla p_i(x) = -e_i + \frac{1}{\|D_i x - d_i\|_2} D_i^T (D_i x - d_i),$$

and it can be computed in $\chi m_i n \le \chi \text{Size}(p)$ operations. And if $D_i x - d_i = 0$, then, as it is immediately seen, $-e_i$ is a subgradient of $p_i(x)$ at x.

Thus, $\mathcal{CQP}$ is easily polynomially computable—$\mathcal{C}_{\text{obj}}$ and $\mathcal{C}_{\text{cons}}$ can be built to have running times $\chi \text{Size}(p)$.

The fact that $\mathcal{CQP}$ is a family with polynomial growth is evident.

3. *$\mathcal{SDP}$ (continued).* Here the instances are semidefinite programs

$$\min_x \left\{ p_0^T x : x \in X(p) = \{x \mid \mathcal{A}x - B \equiv \sum_{j=1}^n x_j A_j - B \succeq 0\} \right\},$$

and the infeasibility measure is

$$\text{Infeas}_{\mathcal{SDP}}(x, p) = \min \{t \ge 0 : \mathcal{A}x - B + tI \succeq 0\}.$$

To verify that the family if polynomially computable, observe first that if m is the row size of the matrices A_j, B, then

$$\text{Size}(p) = \dim \text{Data}(p) \ge n + (n+1)m^2. \tag{5.3.36}$$

(The right-hand side is the total number of entries in $p_0, A_1, \ldots, A_n, B$.) As in the previous cases, given $\text{Data}(p)$ and x, we have no problems with computing the value and the subgradient (which is just p_0) of our linear objective $p_0(x) = p_0^T x$ in $\chi n \le \chi \text{Size}(p)$ operations, so that there is no problem with $\mathcal{C}_{\text{obj}}$.

Regarding $\mathcal{C}_{\text{cons}}$, let us start with the observation that there exists a linear algebra algorithm $\mathcal{S}$ that, given on input a symmetric $m \times m$ matrix A, checks in $O(m^3)$ operations whether A is positive semidefinite and, if it is not the case, generates a vector ξ such that $\xi^T A\xi < 0$.

As a simple example of such an algorithm $\mathcal{S}$, we may use the Lagrange scheme (explained in every linear algebra textbook) of representing a quadratic form $\eta^T A\eta$ as a weighted sum of squares of (linearly independent) linear forms,

$$\eta^T A\eta = \sum_{j=1}^{m} \lambda_j (q_j^T \eta)^2,$$

with m linearly independent vectors $q_1, \ldots, q_m$. This scheme is a simple algorithm (with running time $O(m^3)$) that converts A into the collection of weights λ_j and vectors q_j, $j = 1, \ldots, m$. To check whether a given symmetric $m \times m$ matrix A is positive semidefinite, we may run this Lagrange algorithm on A. If all resulting λ_j are nonnegative, A is positive semidefinite. And if one of λ_j, say, λ_1, turns out to be negative, we can find a vector ξ such that $q_1^T \xi = 1$, $q_j^T \xi = 0$, $j = 2, \ldots, m$, to get a certificate of the fact that A is not positive semidefinite:

$$\xi^T A\xi = \lambda_1 (q_i^T \xi)^2 = \lambda_1 < 0.$$

Note that to find ξ is the same as to solve the linear system

$$q_j^T \xi = \begin{cases} 1, & j = 1 \\ 0, & j = 2, \ldots, m \end{cases}$$

with a nonsingular matrix, i.e., this computation requires just $O(m^3)$ operations.

Equipped with $\mathcal{S}$, let us implement $\mathcal{C}_{\text{cons}}$ as follows. Given x and $\epsilon > 0$, we compute the matrix

$$A = \sum_{j=1}^{n} x_j A_j - B + \epsilon I.$$

Note that by the definition of our infeasibility measure, $\text{Infeas}_{\mathcal{SDP}}(x, p) \leq \epsilon$ if and only if A is positive semidefinite. To check whether this indeed is the case, we apply to A the algorithm $\mathcal{S}$. If $\mathcal{S}$ reports that $A \succeq 0$, we conclude that $\text{Infeas}_{\mathcal{SDP}}(x, p) \leq \epsilon$ and stop. If A is not positive semidefinite, $\mathcal{S}$ returns a corresponding certificate—a vector ξ such that $\xi^T A\xi < 0$. Let us set

$$a = (-\xi^T A_1 \xi, \ldots, -\xi^T A_n \xi)^T;$$

we claim that a can be used as the separator $\mathcal{C}_{\text{cons}}$ should return in the case of $\text{Infeas}_{\mathcal{SDP}}(x, p) > \epsilon$. Indeed, we have

$$0 > \xi^T A\xi = \xi^T \left[\sum_j x_j A_j - B + \epsilon I \right] \xi,$$

i.e.,

$$a^T x > \xi^T[-B + \epsilon I]\xi.$$

On the other hand, for every y with $\mathrm{Infeas}_{\mathcal{SDP}}(y, p) \le \epsilon$ the matrix $\sum_j y_j A_j - B + \epsilon I$ is positive semidefinite, so that

$$0 \le \xi^T \left[\sum_j y_j A_j - B + \epsilon I \right] \xi,$$

whence

$$a^T y \le \xi^T[-B + \epsilon I]\xi.$$

Thus,

$$\mathrm{Infeas}_{\mathcal{SDP}}(y, p) \le \epsilon \Rightarrow a^T y \le \xi^T[-B + \epsilon I]\xi < a^T x,$$

and a is indeed a required separator.

It remains to note that the running time of the routine $\mathcal{C}_{\mathrm{cons}}$ we have built is χnm^2 operations to compute A, χm^3 operations more to run $\mathcal{S}$, and χnm^2 operations to convert ξ into a. Thus, the running time of $\mathcal{C}_{\mathrm{cons}}$ as applied to $\mathrm{Data}(p), x, \epsilon$ does not exceed $\chi(n+m)m^2 \le \chi\,\mathrm{Size}^{3/2}(p)$ (see (5.3.36)).

We have seen that $\mathcal{SDP}$ is polynomially computable. The fact that the family is of polynomial growth is evident.

4. *General convex programming problems (continued).* Consider a family $\mathcal{P}$ of convex optimization programs with instances of the form

$$(p) \qquad \min_x \{p_0(x) : x \in X(p) = \{x \mid p_i(x) \le 0,\ i = 1, .., m(p)\}\}$$

($p_i(\cdot) : \mathbf{R}^{n(p)} \to \mathbf{R}$ are convex, $i = 0, \ldots, m(p)$). The infeasibility measure here (see (5.1.6)) is

$$\begin{aligned} \mathrm{Infeas}_{\mathcal{P}}(x, p) &= \min\left\{t \ge 0 : p_j(x) - t \le 0,\ j = 1, \ldots, m(p)\right\} \\ &= \max\left[0, \max_{j=1,\ldots,m(p)} p_j(x)\right]. \end{aligned}$$

Assume that

I. The vector-function $p(x) = (p_0(x), \ldots, p_{m(p)}(x))^T$, $(p) \in \mathcal{P}$, is polynomially computable: there exists a code $\mathcal{C}$ that, given as input the data vector $\mathrm{Data}(p)$ of an instance (p) and a point $x \in \mathbf{R}^{n(p)}$, returns the values $p_i(x)$ and subgradients $p_i'(x)$ of all components of the function at x, the running time $T(p)$ of the computation being bounded by a polynomial of the size of the instance:

$$\forall (p) \in \mathcal{P} : \quad T(p) \le \chi\,\mathrm{Size}^{\chi}(p).$$

II. The vector-function $p(x)$ is of polynomial growth:

$$\begin{aligned} &\forall \left((p) \in \mathcal{P}, x \in \mathbf{R}^{n(p)}\right) : \\ &\|p(x)\|_1 \le \left(\chi\,[\mathrm{Size}(p) + \|x\|_1 + \|\mathrm{Data}(p)\|_1]\right)^{\left(\chi \mathrm{Size}^{\chi}(p)\right)}. \end{aligned}$$

We claim that under these assumptions $\mathcal{P}$ is polynomially computable and is of polynomial growth. The second of these statements is evident. To verify the first, note that $\mathcal{C}_{\rm obj}$ is readily given by $\mathcal{C}$. The code $\mathcal{C}_{\rm cons}$ can be implemented in the same way as in the cases of linear and conic quadratic programs, as follows. Given Data(p), $x \in \mathbf{R}^{n(p)}$ and an $\epsilon > 0$, we first run $\mathcal{C}$ on Data(p), x to get $p(x)$ and $p'(x) = \{p_i'(x)\}_{i=0}^{m(p)}$. Note that this computation, as a byproduct, gives us the number $m(p)$ of constraints in (p) (since $m(p)+1$ is the number of entries in the vector $p(x)$ we get); besides this, we may be sure that

$$\max[n(p), m(p)] \le \chi \mathrm{Size}^{\chi}(p). \tag{5.3.37}$$

Indeed, the running time of executing $\mathcal{C}$—which is $\chi \mathrm{Size}^{\chi}(p)$—cannot be less than the time required to read the $n(p)$ entries of x and to write the $m(p)+1$ entries of $p(x)$.

After $p(x)$ is obtained, we compute the quantity $g(x) = \max_{i=1,\dots,m(p)} p_i(x)$ and check whether this quantity is $\le \epsilon$. If it is the case, we report that $\mathrm{Infeas}_{\mathcal{P}}(x, p) = \max[0, g(x)]$ is $\le \epsilon$ and stop. In the case of $g(x) > \epsilon$, we find the largest $\widehat{i} \in \{1, \dots, m(p)\}$ such that $p_{\widehat{i}}(x) = g(x)$ and report, as the required separator a, the vector $p'_{\widehat{i}}(x)$. The fact that it indeed is a separator is immediate: if $\mathrm{Infeas}_{\mathcal{P}}(y, p) \le \epsilon$, then $p_{\widehat{i}}(y) \le \epsilon < p_{\widehat{i}}(x)$, whence, by the definition of a subgradient,

$$a^T(y - x) \le p_{\widehat{i}}(y) - p_{\widehat{i}}(x) < 0.$$

It remains to note that apart from the operations needed to run $\mathcal{C}$, all our additional manipulations require $O(m(p))$ operations, and the latter quantity is $\le \chi \mathrm{Size}^{\chi}(p)$ in view of (5.3.37).

The last assumption we need is as follows.

Polynomial boundedness of feasible sets. We say that a family of convex optimization problems $\mathcal{P}$ has polynomially bounded feasible sets if the feasible set $X(p)$ of every instance $(p) \in \mathcal{P}$ is bounded and is contained in the Euclidean ball, centered at the origin, of a not-too-large radius:

$$\begin{array}{l} \forall (p) \in \mathcal{P}: \\ X(p) \subset \left\{ x \in \mathbf{R}^{n(p)} : \|x\|_2 \le \left(\chi\,[\mathrm{Size}(p) + \|\mathrm{Data}(p)\|_1]\right)^{\chi \mathrm{Size}^{\chi}(p)} \right\}. \end{array} \tag{5.3.38}$$

Note that this assumption is not satisfied for typical families of convex optimization programs. Given the data of an LP instance, we cannot bound the size of the feasible set by a function of the norm of the data vector. To see this, look at the subfamily of $\mathcal{LP}$ comprising (1D) instances

$$\min_x \{x : \delta x \ge 1\}$$

with $\delta > 0$.

However, we can impose the property of polynomial boundedness of feasible sets by brute force: just assume that the description of $X(p)$ includes an explicit box constraint

$$|x_j| \le R(p) n^{-1/2}(p), \;\; j = 1, \dots, n(p),$$

where $R(p)$ is an element of the data vector $\mathrm{Data}(p)$. Thus, given a family $\mathcal{P}$ of convex programs, we may pass from it to the family $\mathcal{P}^+$ described as follows: instances (p^+) of $\mathcal{P}^+$ are pairs $((p), R)$, (p) being an instance of $\mathcal{P}$ and R being a positive real; if (p) is the optimization program

$$(p): \quad \min_x \left\{ p_0(x) : x \in X(p) \subset \mathbf{R}^{n(p)} \right\},$$

then $(p^+) = ((p), R)$ is the optimization program

$$(p^+): \quad \min_x \left\{ p_0(x) : x \in X(p^+) = \{x \in X(p) \mid |x_j| \le Rn^{-1/2}(p),\ j = 1, \ldots, n(p)\} \right\},$$

and

$$\mathrm{Data}(p^+) = (\mathrm{Data}^T(p), R)^T.$$

Note that the resulting family $\mathcal{P}^+$ has polynomially bounded feasible sets: by construction, for every $(p^+) = ((p), R) \in \mathcal{P}^+$ we have

$$X(p^+) \subset \{x \in \mathbf{R}^{n(p)} \mid \|x\|_2 \le R \le \|\mathrm{Data}(p^+)\|_1\}.$$

For the families of linear, conic quadratic, and semidefinite programming, the outlined brute force way to ensure the polynomial boundedness of feasible sets shrinks the family: adding box constraints to a linear, conic quadratic, or semidefinite program, we again get a program of the same structure. It follows that if we insist on the property of polynomial boundedness of feasible sets (which is crucial for polynomial time solvability), we cannot deal with the entire families $\mathcal{LP}, \mathcal{CQP}, \mathcal{SDP}$, etc., only with their subfamilies $\mathcal{LP}^+, \mathcal{CQP}^+, \mathcal{SDP}^+, \ldots$, thus restricting our universe. Nevertheless, from the viewpoint of practical computations there is no restriction at all. Indeed, when solving a real-world optimization problem, we never lose much when adding to the original formulation of the problem box constraints like $|x_j| \le 10^{400}$, or even $|x_j| \le 10^{12}$, because in actual computations there is no possibility to get a solution of such a huge magnitude, and moreover such a huge solution could hardly make practical sense.

5.3.1 Polynomial solvability of convex programming

We are about to establish our central result (which is the exact form of the claim (!)).

Theorem 5.3.1. *Let $\mathcal{P}$ be a family of convex optimization programs equipped with infeasibility measure* $\mathrm{Infeas}_{\mathcal{P}}(\cdot, \cdot)$. *Assume that the family is polynomially computable, with polynomial growth and with polynomially bounded feasible sets. Then $\mathcal{P}$ is polynomially solvable.*

Proof. We shall show that polynomial time solvability can be achieved by the ellipsoid method. Our plan is as follows. Assume we are given a positive ϵ and the data vector of an instance $(p) \in \mathcal{P}$,

$$(p): \quad \min_x \left\{ p_0(x) : x \in X(p) \subset \mathbf{R}^{n(p)} \right\},$$

and we wish to compute an ϵ-solution to the instance or to conclude that the instance is infeasible.[49] Since our instances are with polynomially bounded feasible sets, we can extract from Data(p) an a priori upper bound $R(p)$ on the Euclidean norms of feasible solutions to the instance and thus convert (p) into an equivalent program

$$\overline{(p)}: \quad \min_x \left\{ p_0(x) : x \in \bar{X}(p) = \{x \in X(p) \mid \|x\|_2 \le R(p)\} \right\}.$$

Now, to find an ϵ-solution to the latter problem, it suffices to find a feasible ϵ-solution to the augmented problem

$$(p_\epsilon): \quad \min_x \{p_0(x) : x \in X = \{x \mid \text{Infeas}_{\mathcal{P}}(x, p) \le \epsilon, \|x\|_2 \le R(p)\}\}. \tag{5.3.39}$$

A feasible ϵ-solution to the latter problem can be found by the ellipsoid method, provided that we can equip the problem with a separation oracle for X and a first order oracle for $p_0(x)$ and we can point out $r = r(p, \epsilon) > 0$ such that X contains an Euclidean ball of the radius r. As we shall see in a while, our a priori assumptions on the family allow us to build all these entities so that the resulting ellipsoid-method-based solution routine will be a polynomial time one.

Let us implement our plan.

Specifying $R(p)$. Since the problems of the family have polynomially bounded feasible sets, $X(p)$ is contained in the Euclidean ball E_0, centered at the origin, of the radius

$$R(p) = (\chi\,[\text{Size}(p) + \|\text{Data}(p)\|_1])^{\chi \text{Size}^{\chi}(p)}, \tag{5.3.40}$$

where $\chi > 0$ is a certain characteristic constant of $\mathcal{P}$, and therefore is known to us a priori. Given Data(p), we compute $R(p)$ according to (5.3.40), which requires a number of real arithmetic operations which is polynomial in Size(p).

Specifying $r(p, \epsilon)$. We now need to find an $r(p, \epsilon) > 0$ in such a way that the feasible set X of the augmented problem (5.3.39) contains a ball of the radius $r(p, \epsilon)$. Interpreting this target literally, we immediately conclude that it is unachievable, since X can be empty. (This is the case when (p) is heavily infeasible—it does not admit even ϵ-feasible solutions.) However, we can define an appropriate $r(p, \epsilon)$ for the case when (p) is feasible, namely, as follows. To save notation, let us set

$$g(x) = \text{Infeas}_{\mathcal{P}}(x, p).$$

From the polynomial growth property we know that both $p_0(x)$ and $g(x)$ are not very large in $E_0 = \{x \mid \|x\|_2 \le R(p)\}$, namely,

$$\begin{aligned}
&\text{(a)} \quad \text{Var}_{R(p)}(p_0) \le V(p), \\
&\text{(b)} \quad g(x) \le V(p) \quad \forall (x, \|x\|_2 \le R(p)), \\
&\text{(c)} \quad V(p) = (\chi\,[\text{Size}(p) + \max[\|x\|_1 \mid \|x\|_2 \le R(p)] + \|\text{Data}(p)\|_1])^{\left(\chi \text{Size}^{\chi}(p)\right)} \\
&\qquad\quad = \left(\chi\left[\text{Size}(p) + n^{1/2}(p)R(p) + \|\text{Data}(p)\|_1\right]\right)^{\left(\chi \text{Size}^{\chi}(p)\right)},
\end{aligned} \tag{5.3.41}$$

[49] Since all our instances are with bounded feasible sets, we should not bother about the possibility of (p) being unbounded below.

where χ is a characteristic constant of $\mathcal{P}$ and therefore is a priori known. We compute $V(p)$ according to (5.3.41) (which again takes a number of operations polynomial in Size(p)) and set

$$r(p,\epsilon) = \frac{\epsilon}{V(p)+\epsilon} R(p). \tag{5.3.42}$$

We claim that

> (*) *If (p) is feasible, then the feasible set X of problem* (5.3.39) *contains a Euclidean ball of the radius $r(p,\epsilon)$.*

Indeed, by definition of an infeasibility measure, $g(x)$ is a convex nonnegative real-valued function on $\mathbf{R}^{n(p)}$; if (p) is feasible, then g attains value 0 at certain point $\bar{x} \in E_0 \equiv \{x \mid \|x\|_2 \le R(p)\}$. Consider the shrinkage of E_0 to $\bar{x}$ with coefficient $\nu = r(p,\epsilon)/R(p)$ (note that $\nu \in (0,1)$ by (5.3.42)):

$$Y = (1-\nu)\bar{x} + \nu E_0 = \{x = (1-\nu)\bar{x} + \nu z \mid \|z\|_2 \le R(p)\}.$$

On one hand, Y is a Euclidean ball of radius $\nu R(p) = r(p,\epsilon)$. On the other hand, for every $x = (1-\nu)\bar{x} + \nu z \in Y$ ($\|z\|_2 \le R(p)$) we have

$$g(x) \le (1-\nu)g(\bar{x}) + \nu g(z) \le \nu g(z) \le \nu V(p) \le \epsilon.$$

(Recall that g is convex and satisfies (5.3.41)(b).)

Mimicking the oracles. The separation oracle Sep(X) for the feasible set X can be built as follows. Given x, we first check whether $\|x\|_2 \le R(p)$. If this is not the case, then clearly $a = x$ separates x from E_0 (and therefore from $X \subset E_0$), and Sep(x) reports that $x \notin X$ and is separated from X by $a = x$. If $\|x\|_2 \le R(p)$, the oracle calls $\mathcal{C}_{\text{cons}}$, forwarding to it Data$(p), x, \epsilon$. If the result returned to Sep(X) by $\mathcal{C}_{\text{cons}}$ is the claim that $g(x) = \text{Infeas}_{\mathcal{P}}(x,p)$ is $\le \epsilon$ (i.e., if $x \in X$), Sep(X) reports that $x \in X$ and stops. If the result returned to Sep(X) by $\mathcal{C}_{\text{cons}}$ is the claim that $g(x) > \epsilon$ along with a vector a such that

$$g(y) \le \epsilon \Rightarrow a^T y < a^T x,$$

Sep(X) reports that $x \notin X$ and outputs, as the required separator, either a (if $a \neq 0$), or an arbitrary vector $a' \neq 0$ (if $a = 0$). It is clear that Sep(X) works correctly (in particular, the case of $a = 0$ can arise only when X is empty, and in this case every nonzero vector separates x from X). Note that the running time T_{Sep} of Sep(X) (per a single call to the oracle) does not exceed $O(n(p))$ plus the running time of $\mathcal{C}_{\text{cons}}$, i.e., it does not exceed $T_{\text{cons}}(x,\epsilon,p) + O(n(p))$. Since $\mathcal{P}$ is polynomially computable, we have

$$n(p) \le \chi \,\text{Size}^{\chi}(p). \tag{5.3.43}$$

(Indeed, $n(p) \le T_{\text{obj}}(x,p)$, since $\mathcal{C}_{\text{obj}}$ should at least read $n(p)$ entries of an input value of x.) Combining (5.3.43) and (5.3.32), we conclude that

$$T_{\text{Sep}} \le \chi\,(\text{Size}(p) + \text{Digits}(p,\epsilon))^{\chi}. \tag{5.3.44}$$

The first order oracle $\mathcal{O}(p_0)$ is readily given by $\mathcal{C}_{\rm obj}$, and its running time $T_{\mathcal{O}}$ (per a single call to the oracle) can be bounded as

$$T_{\mathcal{O}} \le \chi \mathrm{Size}^{\chi}(p); \tag{5.3.45}$$

see (5.3.30).

Running the ellipsoid method. After we have built $R(p)$, $r(p,\epsilon)$, $\mathrm{Sep}(X)$, and $\mathcal{O}(p_0)$, we can apply to problem (5.3.39) the ellipsoid method as defined above. The only precaution we should take deals with the case when X does not contain a ball of the radius $r(p,\epsilon)$; this may happen only in the case when (p) is infeasible (see (*)), but how could we know whether (p) is or is not feasible? To resolve the difficulty, let us act as follows. If (p) is feasible, then, by Theorem 5.2.1, the ellipsoid method would terminate after no more than

$$\mathrm{Ceil}\left(2n^2(p)\left[\ln\left(\frac{R(p)}{r(p,\epsilon)}\right)+\ln\left(\frac{\epsilon+\mathrm{Var}_{R(p)}(p_0)}{\epsilon}\right)\right]\right)+1$$

steps and will produce a feasible ϵ-solution to (5.3.39), i.e., an ϵ-solution to (p). The indicated number of steps can be bounded from above by the quantity

$$N \equiv N(p,\epsilon) = \mathrm{Ceil}\left(2n^2(p)\left[\ln\left(\frac{R(p)}{r(p,\epsilon)}\right)+\ln\left(\frac{\epsilon+V(p)}{\epsilon}\right)\right]\right)+1, \tag{5.3.46}$$

since $\mathrm{Var}_{R(p)}(p_0) \le V(p)$ by (5.3.41)(a). Let us terminate the ellipsoid method by force if it intends to perform more than $N = N(p,\epsilon)$ steps. When using this emergency stop, we define the result generated by the method as the best (with the smallest value of $p_0(\cdot)$) of the search points c_{t-1} associated with productive steps $t \le N$, if there were productive steps. If no productive steps in the course of our run are encountered, the result of the solution process is the conclusion that (p) is infeasible.

Correctness. We claim that the outlined implementation of the ellipsoid method is correct—i.e., when the corresponding result is an approximate solution $\widehat{x}$, this is an ϵ-solution to (p), and when the result is the conclusion "(p) is infeasible", this conclusion is true. Indeed, if (p) is feasible, then the arguments used in the previous paragraph demonstrate that $\widehat{x}$ is well defined and is an ϵ-solution of (p). If (p) is infeasible, then the result, by construction, is either the correct conclusion that (p) is infeasible or a point $\widehat{x}$ such that $\mathrm{Infeas}_{\mathcal{P}}(\widehat{x}, p) \le \epsilon$. Such a point, in the case of infeasible (p), is an ϵ-solution of (p), since in the case in question $\mathrm{Opt}(p) = +\infty$ and therefore $p_0(x) \le \mathrm{Opt}(p) + \epsilon$ for every x.

Polynomiality. It remains to verify that our solution method is indeed a polynomial time one. Observe, first, that all preliminary computations—those needed to specify $R(p)$, $V(p)$, $r(p,\epsilon)$, $N(p,\epsilon)$—require no more than $\chi\mathrm{Size}^{\chi}(p)$ operations (we have already seen that this is the case for $R(p)$, $V(p)$ and $r(p,\epsilon)$. Given these quantities, it takes just χ operations to compute $N(p,\epsilon)$). It remains to show that the running time of the ellipsoid method admits a polynomial time bound. This is immediate: the method performs no more than $N(p,\epsilon)$ steps, and the arithmetic cost of a step does not exceed the quantity

$$T = T_{\rm Sep} + T_{\mathcal{O}} + \chi n^2(p),$$

where the rightmost term represents the arithmetic cost of updating (5.2.16), computing ℓ_t and all other out-of-oracles operations required by the method. Thus, the overall running time $T(p,\epsilon)$ of our solution method can be bounded as

$$\begin{array}{rcl} T(p,\epsilon) & \leq & \chi\,\mathrm{Size}^\chi(p) + N(p,\epsilon)\left[T_{\mathrm{Sep}} + T_{\mathcal{O}} + \chi n^2(p)\right] \\ & \leq & \chi\,\mathrm{Size}^\chi(p) + N(p,\epsilon)\left[\chi\,(\mathrm{Size}(p) + \mathrm{Digits}(p,\epsilon))^\chi + \chi\,\mathrm{Size}^\chi(p)\right] \\ & & \text{[we have used (5.3.44), (5.3.45) and (5.3.43)]} \\ & \leq & \chi\,\mathrm{Size}^\chi(p)\mathrm{Digits}^\chi(p,\epsilon) \\ & & \text{[see (5.1.8), (5.3.46), (5.3.42), (5.3.41)],} \end{array}$$

so the method is indeed a polynomial time one.

5.4 Difficult problems and NP-completeness

The fundamental motivation for our convexity-oriented approach to optimization, as was announced in the preface and as we are well aware by now, is that convex optimization programs are computationally tractable. On several occasions we also claimed that such-and-such problems are hard or computationally intractable. What do these words actually mean? Without answering this question, a lot of our activity would become seemingly senseless: e.g., why should we bother about semidefinite relaxations of combinatorial problems like MAXCUT? What is wrong with these problems as they are? If we claim that something—e.g., convex programming—is good, we should understand what "bad" means: "good," at least on Earth, and particularly in science, is a relative notion.

To understand what "computational intractability" means, let us outline briefly the basic results of combinatorial complexity theory (CCT).

5.4.1 CCT—a quick introduction

A generic combinatorial problem is a special case of the generic optimization problem. It is a family $\mathcal{P}$ of problem instances, where every instance $(p) \in \mathcal{P}$ is specified by a finite dimensional data vector $\mathrm{Data}(p)$ which now is a Boolean vector, i.e., with entries taking values 0, 1 only (so that the data vectors are, actually, finite binary words).

The model of computations in CCT is also more restrictive (and, in a sense, more realistic) than the real arithmetic model we have dealt with. Now our computer is able to store only integers (i.e., finite binary words), and its operations are bitwise: we are allowed to multiply, add, and compare integers, but now the cost of a single operation of this type depends on the bit length of the operands. To add and to compare two ℓ-bit integers takes $O(\ell)$ bitwise elementary operations, and to multiply a pair of ℓ-bit integers costs $O(\ell^2)$ elementary operations.[50]

In CCT, a solution to an instance (p) of a generic problem $\mathcal{P}$ is a finite binary word y such that the pair $(\mathrm{Data}(p), y)$ satisfies certain verifiable condition $\mathcal{A}(\cdot,\cdot)$. Namely, it is assumed that there exists a code $\mathcal{M}$ for the above integer arithmetic computer such that executing the code on every input pair x, y of finite binary words, the computer terminates

[50] In fact, two ℓ-bit integers can be multiplied in $O(\ell \ln \ell)$ bitwise operations, but for us it makes no difference; the only fact we need is that the bitwise cost of an operation with integers is at least the bit size and at most a fixed polynomial of the bit size of the operands.

after finitely many elementary operations and outputs either "yes," if $\mathcal{A}(x, y)$ is satisfied, or "no," if $\mathcal{A}(x, y)$ is not satisfied. Thus, $\mathcal{P}$ is the problem

given x, find y such that

$$\mathcal{A}(x, y) = \texttt{true}, \tag{5.4.47}$$

or detect that no such y exists.

Two typical examples of generic combinatorial problem are the shortest path problem,

Given a graph with arcs assigned nonnegative integer lengths, two nodes a, b in the graph, and a positive integer d, find a path from a to b of total length not exceeding d, or detect that no such path exists

and the stones problem from Lecture 4,

Given n positive integers $a_1, \ldots, a_n$, find a vector $x = (x_1, \ldots, x_n)^T$ with coordinates ± 1 such that $\sum_i x_i a_i = 0$, or detect that no such vector exists.

Indeed, the data of instances of both problems and candidate solutions to the instances can be naturally encoded by finite sequences of integers. In turn, finite sequences of integers can be easily encoded by finite binary words—you just encode binary digit `0` of an integer as `00`, binary digit `1` as `11`, use `01` to represent the commas separating integers in the sequence from each other, and use `10` to represent the minus sign:

$$5, -3, 7 \Rightarrow \underbrace{110011}_{101_2=5}\ \underbrace{01}_{,}\ \underbrace{10}_{-}\ \underbrace{1111}_{11_2=3}\ \underbrace{01}_{,}\ \underbrace{111111}_{111_2=7}.$$

Clearly, for both the shortest path and stones problems you can easily point out a code for the integer arithmetic computer, which, given on input two binary words, $x = \mathrm{Data}(p)$ encoding the data vector of an instance (p) and y encoding a candidate solution, verifies in finitely many bit operations whether y represents a solution to (p).

A solution algorithm for a generic problem $\mathcal{P}$ is a code $\mathcal{S}$ for the integer arithmetic computer which, given on input the data vector $\mathrm{Data}(p)$ of an instance $(p) \in \mathcal{P}$, terminates after finitely many operations and returns either a solution to the instance or a (correct!) claim that no solution exists. The running time $T_{\mathcal{S}}(p)$ of the solution algorithm on instance (p) is exactly the number of elementary (i.e., bit) operations performed in course of executing $\mathcal{S}$ on $\mathrm{Data}(p)$.

A solvability test for a generic problem $\mathcal{P}$ is defined similarly to a solution algorithm, but now all we want from the code is to say (correctly!) whether the input instance is or is not solvable, i.e., to say just "yes" or "no," without constructing a solution in the case of the "yes" answer.

The complexity of a solution algorithm or solvability test $\mathcal{S}$ is defined as

$$\mathrm{Compl}_{\mathcal{S}}(\ell) = \max\{T_{\mathcal{S}}(p) \mid (p) \in \mathcal{P}, \mathrm{length}(\mathrm{Data}(p)) \le \ell\},$$

where $\mathrm{length}(x)$ is the bit length (i.e., number of bits) of a finite binary word x. The algorithm or test is called polynomial time if its complexity is bounded from above by a polynomial of ℓ.

Finally, a generic problem $\mathcal{P}$ is said to be polynomially solvable if it admits a polynomial time solution algorithm. If $\mathcal{P}$ admits a polynomial time solvability test, we say that $\mathcal{P}$ is polynomially verifiable.

Classes P and NP. A generic problem $\mathcal{P}$ is said to belong to the class NP if the corresponding condition $\mathcal{A}$, see (5.4.47), possesses the following two properties:

I. $\mathcal{A}$ is polynomially computable, i.e., the running time $T(x, y)$ (measured, of course, in elementary bit operations) of the associated code $\mathcal{M}$ is bounded from above by a polynomial of the binary length of the input:

$$T(x, y) \le \chi(\text{length}(x) + \text{length}(y))^{\chi} \quad \forall (x, y).$$

Thus, the first property of an NP problem states that given data Data(p) of a problem instance p and a candidate solution y, it is easy to check whether y is an actual solution of (p). To verify this fact, it suffices to compute $A(\text{Data}(p), y)$, and the time of this computation is polynomial in length(Data(p)) + length(y).

The second property of an NP problem makes its instances even easier:

II. A solution to an instance (p) of a problem cannot be too long as compared to the data of the instance: there exists χ such that

$$\text{length}(y) > \pi(\text{length}(x)) \equiv \chi \text{length}^{\chi}(x) \Rightarrow \mathcal{A}(x, y) = \text{"no"}.$$

A generic problem $\mathcal{P}$ is said to belong to the class P, if it belongs to the class NP and is polynomially solvable.

Note that there is no problem in building a brute force solution algorithm for an NP problem: given Data(p), you just look successively at finite binary words 0,1,00,01,10,11, ... and compute $A(\text{Data}(p), y)$, y being the current candidate solution, to check whether y is a solution to (p). If this is the case, you stop; otherwise, pass to the next candidate solution. After all candidates of length $\le \pi(\text{length}(x))$ are checked and no solution is found, you conclude that (p) has no solution and terminate. Of course, this brute force algorithm is not polynomial—its complexity exceeds $2^{\pi(\ell)}$.

As is immediately seen, both the shortest path and the stones problem belong to NP. There is, however, a dramatic difference between these two problems: the first is polynomially solvable, but for the second no polynomial time solution algorithms are known. Moreover, the second problem is as difficult as a problem from NP can be—it is NP-complete.

Definition 5.4.1. (i) *Let $\mathcal{P}$, $\mathcal{Q}$ be two problems from NP. Problem $\mathcal{Q}$ is called polynomially reducible to $\mathcal{P}$ if there exists a polynomial time algorithm $\mathcal{M}$ (i.e., a code for the integer arithmetic computer with the running time bounded by a polynomial of the length of the input) with the following property. Given on input the data vector* Data(q) *of an instance $(q) \in \mathcal{Q}$, $\mathcal{M}$ converts this data vector to the data vector* Data$(p[q])$ *of an instance of $\mathcal{P}$ such that $(p[q])$ is solvable if and only if (q) is solvable.*

(ii) *A generic problem $\mathcal{P}$ from NP is called NP-complete, if every other problem $\mathcal{Q}$ from NP is polynomially reducible to $\mathcal{P}$.*

One of the most basic results of theoretical computer science is that NP-complete problems do exist (the stones problem is an example).

The importance of the notion of an NP-complete problem comes from the following fact:

(!!!) *If a particular NP-complete problem is polynomially verifiable (i.e., admits a polynomial time solvability test), then every problem from NP is polynomially solvable: $P = NP$.*

(!!!) is an immediate consequence of the following two observations:

(A) *If there exists an NP-complete problem $\mathcal{P}$ that is polynomially verifiable, then every problem from NP is polynomially verifiable.*

Indeed, under the premise of our statement, we can build a polynomial time solvability test for a problem $\mathcal{Q}$ from NP as follows: given an instance $(q) \in \mathcal{Q}$, we first apply the corresponding polynomial time reduction algorithm $\mathcal{M}$ to convert Data(q) into Data$(p[q])$ (see item (i) of Definition 5.4.1). Let $\theta(\cdot)$ be a polynomial such that the running time of algorithm $\mathcal{M}$ on input of length $l = 1, 2, \ldots$ does not exceed $\theta(l)$. The quantity length(Data$(p[q])$) is bounded by the running time of $\mathcal{M}$ on the input Data(q)—it takes one bit operation just to write a single output bit. Since $\mathcal{M}$ is polynomial, we conclude that length(Data$(p[q])$) $\leq \theta(\ell)$, where $\ell =$ length(Data(q)). After Data$(p[q])$ is built, we run the polynomial time solvability test associated with $\mathcal{P}$ (we have assumed that it exists!) to check whether $(p[q])$ is solvable, thus detecting solvability or unsolvability of (q). The running time of this latter computation is bounded by π(length(Data$(p[q])$)) $\leq \pi(\theta(\ell))$, π being a polynomial. Thus, the overall running time does not exceed $\theta(\ell) + \pi(\theta(\ell))$, which is a polynomial in $\ell =$ length(Data(q)).

(B) *If every problem from NP is polynomially verifiable, then every problem from* NP *is polynomially solvable.*

The idea of the proof (we skip the technical details) is as follows: assume we want to solve an instance $(p) \in \mathcal{P}$ ($\mathcal{P}$ is in NP) and have a polynomial time solvability test for the problem. We first check in polynomial time whether (p) is solvable. If the answer is "no," this is all we need; if the answer is "yes," we should find a solution itself. Applying our solvability test, we can decide in polynomial time whether (p) has a solution with the first bit 0. Whatever the answer will be, we will get the first bit y_1 of (some) solution: if the answer is "yes," this bit is 0; otherwise it is 1. Since $\mathcal{P}$ is in NP, we can check in polynomial time whether the single-bit word $y = y_1$ is a solution. If the answer is "no," we proceed in the same way: run our solvability test to check whether (p) has a solution with the first bit y_1 and the second bit 0, thus getting the first *two* bits of a solution, check whether the resulting two-bit word is a solution, and so on. Since the length of all possible solutions to (p) is bounded from above by a polynomial of the length of Data(p) ($\mathcal{P}$ is in NP!), we shall build a solution in a polynomial in $\ell =$ length(Data(p)) number of calls to the solvability test and the verification test (the latter verifies whether a given y solves (p)), running time per call being bounded by a polynomial in ℓ, and the overall running time of building a solution to (p) turns out to be polynomial in ℓ.

As we have mentioned, NP-complete problems do exist. Moreover, a detailed investigation of combinatorial problems carried out during the last three decades (i.e., after the notion of an NP-complete problem and existence of these problems were discovered) demonstrates that

Basically all interesting combinatorial problems are in NP, and nearly all of those that were thought to be difficult (with no known polynomial time solution algorithms) are NP-complete. The list of NP-complete problems includes integer and boolean LP, the traveling salesman problem, MAXCUT, *and hundreds*

of other combinatorial problems, including ones which at first glance appear simple, such as the stones problem.

There are, essentially, just two important generic combinatorial problems from NP that were not known to be polynomially solvable in 1970 and still are not in the list of NP-complete problems. The first of them is the graph isomorphism problem—given two graphs, decide whether they are isomorphic, whether there exist one-to-one correspondences between the set of nodes and the set of arcs of the graphs which preserve the node-arc incidence. The complexity status of this problem is still unknown. The second one is LP over rationals (i.e., LP with rational data). An LP program with rational data, if solvable, admits a rational solution, so that this family of problems can be treated as a generic combinatorial problem. In 1978 Khachiyan proved that LP over rationals is polynomially solvable.

It still is unknown whether P = NP, i.e., whether NP-complete problems are polynomially solvable; this question (viewed by many as the most important open problem in theoretical computer science) has remained open for 30 years.

Although we do not know whether P = NP, i.e., whether there indeed exist difficult combinatorial problems, at the practical level the answer seems to be clear. Indeed, there are a lot of NP-complete problems; some of them, like integer and Boolean linear programming programs and their numerous special cases, are of immense practical importance and therefore for decades have been the subject of intensive studies of thousands of excellent researchers in academia and in industry. However, no polynomial time algorithm for any one of these problems was found. With the discovery of the theory of NP-completeness it became clear that in a sense all research in the area of solution methods for combinatorial programs deals with a *single* problem (since polynomial time solvability of a particular NP-complete problem would automatically imply polynomial time solvability of all such problems). Given the huge total effort invested in this research, we should conclude that it is highly unprobable that NP-complete problems are polynomially solvable. Thus, at the practical level the fact that a certain problem is NP-complete is sufficient to qualify the problem as computationally intractable, at least at our present level of knowledge.

5.4.2 From the real arithmetic complexity theory to the CCT and back

We have outlined two complexity theories, one based on the real arithmetic computer and interested in finding ϵ-solutions for problems with real data, and the other based on the integer arithmetic computer and interested in finding exact solutions to problems with binary data. The theories are similar, but in no sense identical. To stress the difference, consider a simple computational problem—just one of solving a square system of linear equations

$$Ax = b \qquad (A, b)$$

with $n = n(A, b)$ unknowns and a nonsingular matrix A. The real arithmetic complexity theory will qualify the problem as follows:

(R) *The family $\mathcal{L}$ of instances (A, b) with real entries in A, b and nonsingular A is polynomially solvable: there exists an algorithm (e.g., the Gauss elimination method) that, as applied to any instance, finds the exact solution of the instance*

in no more than $O(n^3(A,b))$ operations of exact real arithmetic, which is a polynomial in the dimension $\mathrm{Size}(A,b) = n^2(A,b)+n(A,b)$ *of the data vector of the instance.*

The CCT first cannot handle systems of linear equations with real data, since in this case neither the data nor candidate solutions can be encoded by finite binary words. However, CCT can handle systems with rational data and will qualify these systems as follows:

(C) *The family $\mathcal{L}$ of instances (A,b) with rational entries in A,b and nonsingular A is polynomially solvable: there exists a solution algorithm (e.g., the Gauss elimination method) that, as applied to any instance, finds an exact solution of the instance in no more than a polynomial in ℓ number $\pi(\ell)$ of bit operations, ℓ being the length of the instance data, i.e., the total number of bits in binary representations of the numerators and denominators of all entries of the data.*

We see that these two statements say different things about different entities, and neither is a consequence of the other; a priori it might happen that one of these statements is true and another is false, or both are false, or both are true (which is indeed the case here). The essence of the difference comes not from the fact that (R) speaks about all systems, while (C) addresses only systems with rational data; this is a minor point. The essence of the difference is that in (C) an elementary operation is a *bit* operation, while in (R) it is an operation with reals; thus, (C), as compared to (R), deals with a much more restricted set of elementary operations and therefore with a much more strict (and more realistic) notion of computational effort. As a kind of compensation, (C) uses a less-strict notion of a polynomial time method than (R): for (C) a method is polynomial if its running time (measured in bit operations) is bounded by a polynomial of the total binary length of the data, while for (R) this running time (measured in the number of real arithmetic operations) should be bounded by a polynomial of the number of data entries—a quantity that is definitely less than the binary length of the data. For example, when (C) says that systems of linear equations are polynomially solvable, it says nothing definite about the complexity of solving a system of two linear equations with two variables: the bitwise complexity of this simplest computational problem can be as large as you wish, provided that the coefficients are rationals with large numerators and denominators. In contrast to this, when (R) says that systems of linear equations are polynomially solvable, it says, in particular, that a system of two linear equations with two unknowns can be solved in $O(1)$ operations of real arithmetic.

Although the two outlined complexity theories deal with different setups, each one can nevertheless utilize (and in fact does!) some results of its counterpart. Sometimes a real arithmetic polynomial time method, as restricted to a family of problems with rational data, can be converted to a CCT-polynomial time algorithm for a combinatorial problem, thus yielding CCT-polynomial solvability of this problem. Borrowing efficient algorithms in the opposite direction—from combinatorial problems to those with real data—does not make much sense; the real arithmetic complexity theory does borrow from CCT the techniques for recognizing computationally intractable problems. In this book we are at the side of optimization programs with real data, hence our primary interest is what can be borrowed from the CCT and not what can be given to it. However, we start with an example of the latter.

Combinatorial complexity theory polynomial solvability of linear programming over rationals

Let us start with some historical remarks. Linear programming in most of its methodological and computational aspects was discovered by Dantzig in the late 1940s. He proposed the simplex method for LP (1948), which for about 40 years was the only practical (and extremely efficient) tool for solving LP programs. It is still one of the most efficient computational techniques known for LP, and is the most frequently used. The theoretical justification of the method is that in the real arithmetic model of computations it finds an exact solution to any LP program (or detects correctly that no solution exists) in finitely many arithmetic operations, while its practical justification is that this number of operations is typically quite moderate. However, it was discovered that the worst-case complexity of the simplex method is very bad: Klee and Minty (1964) built a subfamily of LP programs $\{(p_n)\}_{n=1}^{\infty}$ with the following property: (p_n) has rational data, and the size of the instance (p_n) is polynomial in n, whether we measure the size as the number of data entries or as their total bit length. The number of arithmetic operations performed by the simplex method as applied to (p_n) is more than 2^n. Thus, the simplex method is not a polynomial time method in the sense of real arithmetic complexity theory, same as in the sense of the CCT. For about 15 years, the question whether LP is polynomially solvable was one of the major challenges in the area of computational complexity. The CCT-version of this question was answered affirmatively by Khachiyan in 1978, and the tool he used was borrowed from convex optimization with real data—it was the ellipsoid method (1976). The sketch of Khachiyan's construction is as follows.

Linear feasibility problem. Let us start with the LP feasibility problem:

(FeasLP) *Given a system of linear inequalities*

$$Ax \leq b \tag{S}$$

with rational coefficients, check whether the system has a solution.

The polynomial time (in the sense of CCT!) algorithm for (FeasLP) proposed by Khachiyan is as follows.

We may assume without loss of generality that the columns of A are linearly independent. (Otherwise we can eliminate the columns that are linear combinations of the remaining ones, which does not affect the feasibility. This linear algebra operation takes time which is polynomial in the total bit length L of the data and results in a system of the same structure and the total bit length of the data which is polynomial in L.) It is also convenient to assume that the data is integer. (We may multiply all coefficients involved in a particular inequality by their common denominator. This equivalent transformation results in a problem with the total bit length of the data being polynomial in the bit length of the original data.) Thus, we may assume that the data in (S) are integer and the columns in A are linearly independent. Let L be the total bit length of the data, m be the number of inequalities, and n be the number of unknowns in (S).

The first step is to get an a priori upper bound on the norm of a solution to (S), assuming that such a solution exists. This can be done as follows.

It is a well-known fact of LP and convex analysis that if A has linearly independent columns and (S) has a solution, then (S) has an extreme point solution x as follows: n of the m inequalities of (S) at x are equalities, and the rows of A corresponding to these inequalities are linearly independent. In other words, x is the unique solution of the square system of linear equations

$$\widehat{A}x = \widehat{b},$$

where $\widehat{A}$ is a nonsingular $n \times n$ submatrix of A and $\widehat{b}$ is the corresponding subvector of b. From the Cramer rules it follows that every coordinate in x is the ratio $\frac{\Delta'}{\Delta}$ of two determinants: $\Delta = \mathrm{Det}\widehat{A}$ and Δ' is the determinant of an $n \times n$ submatrix of the matrix $[\widehat{A}; \widehat{b}]$. Now, Δ is a nonzero integer (all entries in A, b are integer!), and Δ' is not too large—the absolute value of a determinant does not exceed the product of Euclidean lengths of its rows.[51] Since the sum of binary lengths of all entries in A, b does not exceed L, the above product cannot be very large; a simple computation demonstrates that it does not exceed $2^{O(1)L}$ (all $O(1)$'s are absolute constants). We conclude that the absolute values of all entries in x do not exceed $2^{O(1)L}$, so that $\|x\|_2 \leq 2^{O(1)L}\sqrt{n} \leq 2^{O(1)L}$. (We have used the evident fact that both n and m do not exceed L—it takes at least one bit to represent every one of the mn entries of A.)

The second step is to look at the minimum value g_* of the residual

$$g(x) = \max\left[0, \max_{i=1,\ldots,m}[(Ax)_i - b_i]\right];$$

note that this minimum value is 0 if (S) is solvable and is > 0 otherwise. What we are interested in is to understand whether g_* can be positive but very small. The answer is "no": if $g_* > 0$, then $g_* > 2^{-O(1)L}$.

Indeed, g_* is the optimal value of the feasible LP program

$$\min_{t,x}\{t : t \geq 0, (Ax)_i - b_i \leq t,\ i = 1, \ldots, m\}. \qquad (\mathrm{S}')$$

The binary length L' of the data in this problem is of order L, and the problem for sure is solvable. From the theory of LP it is well known that if an LP has a solution and the feasible set of the problem does not contain lines, then the problem has a solution that is an extreme point of the feasible set. Since the columns of A are linearly independent and (S′) is solvable, the feasible set of the problem does not contain lines (why?); consequently, the problem admits an extreme point solution. The coordinates of the latter solution, in particular, its t-coordinate (i.e., the optimal value in (S′), i.e., g_*), again are ratios of determinants, now coming from the matrix $[A; b; e]$, e being the vector of ones. Thus, g_* is the ratio of two integers and the absolute values of these integers, same as above, do not exceed $2^{O(1)L'} = 2^{O(1)L}$. It follows that if $g_* > 0$, then $g_* \geq 2^{-O(1)L}$—the numerator in the ratio representing g_*, being a nonzero integer, should be at least one, and the denominator cannot be larger than $2^{O(1)L}$.

The third step—we already know that if (S) is feasible, the minimum of g in the ball $E_0 = \{x \mid \|x\|_2 \leq 2^{O(1)L}\}$ is zero (since then this ball contains a solution to (S)). We know also that if (S) is infeasible, then the minimum of g in E_0 is at least $2\epsilon = 2^{-O(1)L}$, since

[51] Hadamard's inequality expressing an evident geometric fact: the volume of a parallelotope does not exceed the product of its edges.

in this case already the minimum of g on the entire $\mathbf{R}^n$ admits the indicated lower bound. It follows that in order to check feasibility of (S) it suffices to find an ϵ-solution x_ϵ to the optimization problem

$$\min_x \{g(x) : x \in E_0\}; \tag{C}$$

if the value of g at x_ϵ is $\leq \epsilon$, we are sure that the true minimum value of g is less than ϵ, which, in view of the origin of ϵ, is possible only if the true optimal value in (C) is 0 and (S) is solvable. And if $g(x_\epsilon) > \epsilon$, the optimal value in (C) is > 0 (since x_ϵ is an ϵ-solution to (C)), and (S) is infeasible.

Now, g clearly is a convex function with easily (in $O(1)mn$ arithmetic operations) computable value and subgradient. It follows that an ϵ-solution to (C) can be found by the ellipsoid method. Let us evaluate the complexity of this process. In the notation from Theorem 5.2.1 our case is the one of $X = E_0$ (i.e., $r = R = 2^{O(1)L}$) and, as is easily seen, $\mathrm{Var}_R(g) \leq R2^L$. Theorem 5.2.1 therefore says that the number of steps in the ellipsoid method is bounded from above by $O(1)n^2 \ln(\frac{\epsilon+\mathrm{Var}_R(g)}{\epsilon}) \leq O(1)n^2L$ (note that both ϵ^{-1} and $\mathrm{Var}_R(g)$ are of order of $2^{O(1)L}$). The number of arithmetic operations per step is $O(1)(n^2 + mn)$, where the n^2-term represents the operation cost of the method itself and the mn-term represents the computational expenses of computing $g(x)$, $g'(x)$ at a given x and mimicking the separation oracle for the Euclidean ball E_0. (When proving Theorem 5.3.1, we have built such an oracle and have evaluated its running time—it is just $O(1)n$.) Thus, the overall number of arithmetic operations required to find an ϵ-solution to (C) is $O(1)(n^2 + mn)n^2L$, which is a polynomial in L (recall that $mn \leq L$).

We are almost done. The only remaining problem is that the ellipsoid method is a real arithmetic procedure, so that the polynomial in L complexity bound of checking feasibility of (S) counts the number of operations of real arithmetic, while what we need is an integer arithmetic computer routine and a bound on the number of bit operations. Well, a quite straightforward (although tedious) analysis demonstrates that we can obtain the same accuracy guarantees when implementing the ellipsoid method on an inexact arithmetic computer, where every elementary operation $+, -, /, \times, \sqrt{}$ is applied to $O(1)nL$-digit operands and rounds the exact result to the same $O(1)nL$ digits. Now every arithmetic operation costs a number of bit operations which is polynomial in L, and thus the overall bit complexity of the computation is also polynomial in L.

From checking feasibility to finding optimal solutions. It remains to explain how a CCT-polynomial time algorithm for checking feasibility of systems of linear inequalities with rational data can be converted into a CCT-polynomial time algorithm for solving LP programs with rational data. Observe, first, that to solve an LP problem is the same as to solve certain system of linear inequalities (S) (write the constraints of the primal problem along with those of the dual and the linear equation saying that the duality gap is 0; of course, a linear equation can be written as a pair of opposite linear inequalities). We already know how to check in polynomial time the feasibility of (S), so all we need is to figure out how to find a feasible solution to (S) given that the system is feasible. The simplest way to do it is as follows. Let us take the first inequality $a^T x \leq b$ in (S), replace it with the equality $a^T x = b$, and check whether the modified system we obtain is feasible. If not, we know that the hyperplane $a^T x = b$ does not intersect the solution set of (S); since this set

is nonempty and convex, we conclude that every solution to the system (S_1) obtained from (S) by eliminating the first inequality is a solution to (S) as well. If the modified system is feasible, let (S_1) be this modified system. Note that in both cases (S_1) is solvable, and every solution to (S_1) is a solution to (S). Now let us look at (S_1); this system can have both inequalities and equalities. Let us take the first inequality of the system, if it exists, make it equality, and check whether the modified system is feasible. If it is, this modified system will be our (S_2); otherwise (S_2) will be obtained from (S_1) by eliminating the first inequality in (S_1). Note that in both cases (S_2) is solvable, and every solution to it is a solution to (S_1) and therefore to (S). Note also that the number of inequalities in (S_2) is less than that one in (S) by 2. Proceeding in the same way, we look in turn at all inequalities of the original system, check feasibility of certain intermediate system of equations and inequalities, and, as a result, either make the inequality an equality or eliminate it, thus getting a new intermediate system. By construction, this system is solvable, and all its solutions are solutions to (S). After m steps of this process (m is the number of inequalities in (S)) we terminate with a solvable system (S_m) of equations, and every solution to this system is a solution to (S). Note that all our intermediate systems are of the same total data length L as (S), so that the overall CCT-complexity of the outlined process is polynomial in L. It remains to note that we can use the standard linear algebra routines to find a solution to the solvable system of equations (S_m) in time that is polynomial in L, thus getting in polynomial time a solution to (S).

Pay attention to the intrinsic mechanics of the outlined construction: its nucleus is a simple real arithmetic polynomial time routine. This nucleus is spoiled by replacing real arithmetic operations with their inexact counterparts and is equipped with a completely exterior termination rules based on the fact that we are dealing with percolated—rational—data. Note that the more natural version of the question whether LP is polynomially solvable—namely, the question whether an LP program with rational or real data can be solved exactly in a number of real arithmetic operations which is polynomial in the size $\mathrm{Size}(p) = \dim \mathrm{Data}(p)$ of the instance—still remains open.

Difficult convex optimization problems

As mentioned, what real arithmetic complexity theory can borrow from CCT are techniques for detecting computationally intractable problems. Consider the situation as follows: we are given a family $\mathcal{P}$ of convex optimization programs[52] and we want to understand whether the family is polynomially solvable. Theorem 5.3.1 gives us sufficient conditions for polynomial solvability of $\mathcal{P}$; what to do if one of these conditions is not satisfied? To be more concrete, let us look at the following family of convex optimization programs:

$$\min_{t,x}\left\{t : x \in X = \left\{x \in \mathbf{S}^n \mid A \succeq x \succeq B, \max_{u\in C_n} u^T x u \le t\right\}\right\}, \tag{5.4.48}$$

where $C_n = \{u \in \mathbf{R}^n \mid |u_i| \le 1, i = 1, \ldots, n\}$ is the n-dimensional unit cube and A, B are symmetric matrices. Note that this problem is of essential interest for the robust conic

[52] We could speak about other computational problems with real data, in particular, nonconvex optimization ones, but recall that our subject is convex optimization.

quadratic optimization we mentioned in Lecture 4. Here we can specify in a natural way the data vectors of instances and the associated infeasibility measure:

$$\mathrm{Infeas}(x,\,p) = \min\left\{\tau \geq 0 : x \preceq A + \tau I,\, x \succeq B - \tau I,\, \max_{u\in C_n} u^T x u \leq t + \tau\right\},$$

thus coming to a family of polynomial growth and with polynomially bounded feasible sets. The difficulty is with polynomial computability: we do not see an easy way to implement $\mathcal{C}_{\mathrm{cons}}$. Indeed, a direct way—to compute $\mathrm{Infeas}(x,\,p)$ according to the definition of this function—fails, since no algorithms for computing the maximum of $g_x(u) = u^T x u$ over the unit cube with complexity less than $2^{O(n)}$ operations are known, while $\mathrm{Size}(p)$—the dimension of the data vector—is only of order n^2.

How should we proceed? Perhaps we just do not see a proper way to implement $\mathcal{C}_{\mathrm{cons}}$ and should think more on this subject? For how long? Fortunately (or unfortunately, depending on viewpoint), we can easily understand that our problem hardly is polynomially solvable. To explain the reason, let us forget for the moment about our particular family of convex programs and ask

> (?) *How could we convince ourselves that a given generic program $\mathcal{P}$ is computationally intractable?*

One way to answer (?) is as follows. Assume that the objectives of the instances of $\mathcal{P}$ are polynomially computable and that we can point out a generic combinatorial problem $\mathcal{Q}$ known to be NP-complete which can be reduced to $\mathcal{P}$ in the following sense:

> *There exists a CCT-polynomial time algorithm $\mathcal{M}$ that, given on input the data vector* $\mathrm{Data}(q)$ *of an instance $(q) \in \mathcal{Q}$, converts it into a triple* $\mathrm{Data}(p[q])$, $\epsilon(q)$, $\mu(q)$ *comprising the data vector of an instance $(p[q]) \in \mathcal{P}$, positive rational $\epsilon(q)$, and rational $\mu(q)$ such that $(p[q])$ is solvable and*
>
> • *if (q) is unsolvable, then the value of the objective of $(p[q])$ at every $\epsilon(q)$-solution to this problem is $\leq \mu(q) - \epsilon(q)$;*
>
> • *if (q) is solvable, then the value of the objective of $(p[q])$ at every $\epsilon(q)$-solution to this problem is $\geq \mu(q) + \epsilon(q)$.*

We claim that in the case in question we have all reasons to qualify $\mathcal{P}$ as a computationally intractable problem. Assume, to the contrary, that $\mathcal{P}$ admits a polynomial time solution method $\mathcal{S}$, and let us look what happens if we apply this algorithm to solve $(p[q])$ within accuracy $\epsilon(q)$. Since $(p[q])$ is solvable, the method must produce an $\epsilon(q)$-solution $\widehat{x}$ to $(p[q])$. With additional polynomial time effort we may compute the value of the objective of $(p[q])$ at $\widehat{x}$. (Recall that the objectives of instances from $\mathcal{P}$ are assumed to be polynomially computable.) Now we can compare the resulting value of the objective with $\mu(q)$. By the definition of reducibility, if this value is $\leq \mu(q)$, then q is unsolvable; otherwise q is solvable. Thus, we get a correct real arithmetic solvability test for $\mathcal{Q}$. What is the (real arithmetic) running time of this test? By definition of a real arithmetic polynomial time algorithm, it is bounded by a polynomial of $s(q) = \mathrm{Size}(p[q])$ and

$$d(q) = \mathrm{Digits}((p[q]), \epsilon(q)) = \ln\left(\frac{\mathrm{Size}(p[q]) + \|\mathrm{Data}(p[q])\|_1 + \epsilon^2(q)}{\epsilon(q)}\right).$$

Now note that if $\ell = \text{length}(\text{Data}(q))$, then the total number of bits in $\text{Data}(p[q])$ and in $\epsilon(q)$ is bounded by a polynomial of ℓ (since the transformation $\text{Data}(q) \mapsto (\text{Data}(p[q]), \epsilon(q), \mu(q))$ takes CCT-polynomial time). It follows that both $s(q)$ and $d(q)$ are bounded by polynomials in ℓ, so that our real arithmetic solvability test for $\mathcal{Q}$ takes a number of arithmetic operations that is polynomial in $\text{length}(\text{Data}(q))$.

Recall that $\mathcal{Q}$ was assumed to be an NP-complete generic problem, so that it would be highly improbable to find a CCT-polynomial time solvability test for this problem, while we have managed to build such a test, with the only (but important!) difference that our test is a real arithmetic one—it uses incomparably more powerful elementary operations. Well, a reasonable real arithmetic algorithm—one that can be used in actual computations—must be tolerant to small rounding errors (cf. what was said about the ellipsoid algorithm in the context of LP). Specifically, such an algorithm, as applied to a pair $((p), \epsilon)$ should be capable to say to the computer "I need to work with reals with such and such number of binary digits before and after the dot, and I need all elementary operations with these reals to be precise within the same number of accuracy digits," and the algorithm should preserve its performance and accuracy guarantees, provided that the computer meets the indicated requirement. Moreover, for a reasonable real arithmetic algorithm the aforementioned number of digits before and after the dot must be polynomial in $\text{Size}(p)$ and $\text{Digits}(p, \epsilon)$.[53] With these assumptions, our polynomial time real arithmetic solvability test can be easily converted into a CCT-polynomial time solvability test for $\mathcal{Q}$, which—once again—hardly could exist. Thus, a real arithmetic polynomial time algorithm for $\mathcal{P}$ hardly could exist as well.

Since we do not know whether NP-complete problems are computationally intractable, the above reasoning does not prove that if you can reduce a NP-complete combinatorial problem to a generic program $\mathcal{P}$ with real data, the latter program is not polynomially solvable in the sense of real arithmetic complexity theory; note, however, that (?) asks about convince, not prove.

As an illustration, let us demonstrate that the NP-complete stones problem can be reduced to the generic convex program $\mathcal{P}_0$ with instances (5.4.48), so that $\mathcal{P}_0$ is computationally intractable. Indeed, let $(n, a = (a_1, \ldots, a_n)^T)$ be the data of an instance (q) of the stones problem; recall that the instance is solvable if and only if there exist $u_i = \pm 1$ such that $\sum_i a_i u_i = 0$. Given (n, a), let us define the data of the instance $(p[q]) \in \mathcal{P}_0$ as

$$A = B = \|a\|_2^2 I_n - aa^T$$

and set

$$\epsilon(q) = \frac{1}{2(n+2)}, \mu(q) = n\|a\|_2^2 - \frac{1}{2}.$$

Let us demonstrate that this is indeed a reduction. Observe, first, that the conversion $\text{Data}(q) \mapsto (\text{Data}(p[q]), \epsilon(q), \mu(q))$ is clearly CCT-polynomial time. Now, since $A = B$, the feasible set of $(p[q])$ is

$$\{x = A = B, t \geq \max_{u \in C_n} u^T x u\},$$

[53] In fact, this property normally is included into the very definition of a real arithmetic polynomial time algorithm; we prefer to skip these boring technicalities and to work with a simplified definition.

and the optimal value in $(p[q])$ is

$$\mathrm{Opt}(p[q]) = \max_{u \in C_n} g(u), \quad g(u) = u^T A u.$$

Since $A \succeq 0$ (check it!), the quadratic form g is convex, and therefore its maximum over C_n is the same as its maximum over the set of vertices of C_n (why?). If u is a vertex of C_n (i.e., a vector with coordinates ± 1), then the value of g at u is $n\|a\|_2^2 - (a^T u)^2$. Thus,

$$\mathrm{Opt}(p[q]) = \max \left\{ n\|a\|_2^2 - (a^T u)^2 \mid u_i = \pm 1, \ i = 1, \ldots, n \right\}.$$

If (q) is unsolvable—i.e., $a^T u \neq 0 \ \forall u$ with coordinates ± 1—then $(a^T u)^2 \geq 1$ for the indicated u (since for these u $a^T u$ is an integer), and we see that $\mathrm{Opt}(p[q]) \leq n\|a\|_2^2 - 1 = \mu(q) - 1/2$. On the other hand, if (q) is solvable, then the optimal value of the objective in $(p[q])$ is equal to $n\|a\|_2^2 = \mu(q) + 1/2$. Thus, the exact optimal value of $(p[q])$ is quite sensitive to solvability or unsolvability of (q). This is nearly what we need, but not exactly: we should prove that already the value of the objective of $(p[q])$ at any $\epsilon(q)$-solution to the problem is sensitive to the solvability status of (q). Let $(\widehat{x}, \widehat{t})$ be an $\epsilon(q)$-solution to $(p[q])$. In the case of unsolvable (q) we should have

$$\widehat{t} \leq \mathrm{Opt}(p[q]) + \epsilon(q) \leq \mu(q) - 1/2 + \epsilon(q) \leq \mu(q) - \epsilon(q). \tag{5.4.49}$$

Now assume that (q) is solvable. By definition of the infeasibility measure we have

$$\begin{array}{rcl} \widehat{x} & \succeq & A - \epsilon(q) I_n, \\ \widehat{t} & \geq & \max\limits_{u \in C_n} u^T \widehat{x} u - \epsilon(q). \end{array}$$

Recalling that if (q) is solvable, then

$$\mathrm{Opt}(p[q]) \equiv \max_{u \in C_n} u^T A u = \mu(q) + 1/2,$$

we have

$$\begin{array}{rcl} \widehat{t} & \geq & \max\limits_{u \in C_n} u^T \widehat{x} u - \epsilon(q) \\ & \geq & \max\limits_{u \in C_n} \left[u^T A u - \epsilon(q) u^T u \right] - \epsilon(q) \\ & \geq & \max\limits_{u \in C_n} u^T A u - n\epsilon(q) - \epsilon(q) \\ & = & \mathrm{Opt}(p[q]) - (n+1)\epsilon(q) \\ & = & \mu(q) + 1/2 - (n+1)\epsilon(q) \\ & \geq & \mu(q) + \epsilon(q). \end{array}$$

Combining the resulting inequality with (5.4.49), we see that the outlined construction indeed is a reduction of the stones problem to $\mathcal{P}_0$.

The generic convex program (5.4.48) illustrates the most typical source of intractable convex programs—semi-infiniteness. If we write (5.4.48) explicitly, we get the problem

$$\min_{t,x} \left\{ t : A \succeq x \succeq B, u^T x u \leq t \quad \forall u \in C_n \right\},$$

which has infinitely many simple convex constraints parameterized by a point $u \in C_n$. Computational tractability of a problem of this type depends on the geometry of the parameter set. For example, if we replace the cube C_n by a simplex or an Euclidean ball, we get a polynomially computable (and polynomially solvable) generic program.

Lecture 6

Interior Point Polynomial Time Methods for Linear Programming, Conic Quadratic Programming, and Semidefinite Programming

6.1 Motivation

We have seen that generic convex problems, under mild computability and boundedness assumptions, are polynomially solvable, e.g., by the ellipsoid method. This result is extremely important theoretically; however, from the practical viewpoint it is essentially no more than an existence theorem. Indeed, the complexity bounds for the ellipsoid method, although polynomial, are not very attractive. By Theorem 5.2.1, when solving problem (5.2.9) with n design variables, the price of an accuracy digit (the cost of reducing the current inaccuracy ϵ by factor 2) is $O(n^2)$ calls to the first order and the separation oracles, plus $O(n^4)$ arithmetic operations to process the answers of the oracles. Thus, even for problems with very simple objectives and feasible sets, the arithmetic price of an accuracy digit is $O(n^4)$. Imagine what it takes then to solve a problem with, say, 1000 variables (which is still a small size for many applications). One could hope, of course, that the efficiency estimate stated in Theorem 5.2.1 is no more than a worst-case theoretical upper bound, while the actual behavior of the ellipsoid method is typically much better than the bound says. A priori this is not entirely hopeless; for example, the LP simplex method is an extremely powerful computational tool, despite its disastrously bad (exponential in the size of an instance) worst-case efficiency estimate. Unfortunately, practice demonstrates that the ellipsoid method does work according to its theoretical efficiency estimate, and thus it is unable to solve in reasonable time even medium-scale ($> 10^2$ variables) convex programs.

It turns out that both the strong (universality) and the weak (poor practical performance) features of the ellipsoid method have the same origin—the fact that the method deals with black box–represented problems. When starting the solution process, the method has no idea what problem it is working with. All it knows is how many design variables are in the problem and how large and thin its feasible set can be (i.e., knows the radii R

and r; see the description of the method). All the rest should be learned in the course of solving the problem via the answers of the separation and the first order oracles, and these answers provide local information only. Intuitively it seems quite natural that a method that is restricted in its information sources can hardly be expected to perform well[54]—it is a bad idea to always look at the world through a magnifying glass. On the other hand, the need to solve black box–represented convex problems rarely occurs in practice. In practice, there are no entities like "a convex objective f" and "a convex feasible set $\mathcal{X}$," with all our a priori knowledge of f and $\mathcal{X}$ being expressed by the word "convex." Normally we deal with instances of a known-in-advance generic convex program, like LP, CQP, or SDP, with known data and therefore we possess from the very beginning complete global information on the problem to be processed.[55] Of course it seems ridiculous to use our complete global knowledge of the instance just to mimic the local in their nature first order and separation oracles. What we would like to have is an optimization technique able to utilize efficiently our global knowledge of the instance and thus able to generate a solution much faster than a nearly blind algorithm like the ellipsoid method. The major event in the recent history of convex optimization, called sometimes interior point revolution, was the invention of such smart techniques.

6.1.1 Interior point methods

The interior point revolution was started by the seminal work of Karmarkar (1984), where the first interior point method for LP was proposed; 15 years since then, interior point polynomial time methods have become an extremely deep and rich, theoretically, and highly promising, computationally, area of convex optimization. A detailed overview of the history and the current state of area is beyond the scope of our book.[56] All we intend to do is to give an idea of what the interior point methods are, for the sake of those using optimization techniques (not to those *developing* them!). In particular, we skip nearly all (sometimes highly instructive and nontrivial) technicalities.

[54] And this intuition is valid: it can be proved that when minimizing within accuracy ϵ, in a black box fashion, an arbitrary convex function f of n variables over the unit n-dimensional cube, the variation of the function on the cube being ≤ 1, the number of calls to the first order oracle cannot be less than $0.48n \ln \frac{1}{\epsilon}$ when $\epsilon \le 0.5$. In other words, for every $\epsilon < 0.5$ and every optimization method using solely the first order oracle, there exists a function f in the aforementioned class such that the method is unable to find an ϵ-minimizer of f in less than $0.48n \ln \frac{1}{\epsilon}$ computations of f and f'. This result remains true even when we further restrict our convex objectives to be C^∞-smooth and replace the first order oracle by an oracle that returns the values of the objective and all its derivatives (the first, the second, etc.) at the point the oracle gets on input.

We see that the performance of the ellipsoid method is not that poor: it is merely $O(n)$ times worse than the best possible performance of a method dealing with black box–represented convex programs!

[55] There are, of course, optimization programs for which our complete global knowledge is too complicated to be useful for something more reasonable than mimicking the local oracles. Imagine, e.g., an unconstrained optimization program where the objective is the value, at a given point, of a solution to a messy system of differential equations affected somehow by the design variables. Such a case, however, is perhaps of interest for mathematical programming in general, but it is of no interest for convex programming in particular: if your objective is as complicated as if it had no understandable structure, how do you know that it is convex?

[56] An interested reader is referred to the following books: Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming,* SIAM, Philadelphia, 1994; C. Roos, T. Terlaky, and J.-P. Vial, *Theory and Algorithms for Linear Optimization: An Interior Point Approach*, John Wiley, New York, 1997; and Y. Ye, *Interior Point Algorithms: Theory and Analysis*, John Wiley, New York, 1997.

The simplest way to get a proper impression of the (most of) interior point methods is to start with a quite traditional interior penalty scheme for solving optimization problems.

6.2 Newton method and the interior penalty scheme

6.2.1 Unconstrained minimization and the Newton method

Seemingly the simplest convex optimization problem is the one of unconstrained minimization of a smooth strongly convex objective:

$$\min_x \left\{ f(x) : x \in \mathbf{R}^n \right\}. \tag{UC}$$

Smooth strongly convex in this context means a three-times continuously differentiable convex function f such that $f(x) \to \infty$ as $\|x\|_2 \to \infty$ and such that the Hessian matrix $f''(x) = [\frac{\partial^2 f(x)}{\partial x_i \partial x_j}]$ of f is positive definite at every point x. Among numerous techniques for solving (UC), the most promising is the Newton method. In its pure form, the Newton method is extremely transparent and natural: given a current iterate x, we approximate our objective f by its second order Taylor expansion at the iterate, i.e., by the quadratic function

$$f_x(y) = f(x) + (y - x)^T f'(x) + \frac{1}{2}(y - x)^T f''(x)(y - x),$$

and choose as the next iterate x_+ the minimizer of this quadratic approximation. Thus, the Newton method merely iterates the updating

$$x \mapsto x_+ = x - [f''(x)]^{-1} f'(x). \tag{Nwt}$$

In the case of a (strongly convex) quadratic objective, the approximation coincides with the objective itself, and the method reaches the exact solution in one step. It is natural to guess (and indeed it is true) that if the objective is smooth and strongly convex (although not necessary quadratic) and the current iterate x is close enough to the minimizer x_* of f, then the next iterate x_+, although not being x_* exactly, will be much closer to x_* than x. The precise (and easy) result is that the Newton method converges *locally quadratically*, i.e., that

$$\|x_+ - x_*\|_2 \le C\|x - x_*\|_2^2,$$

provided that $\|x - x_*\|_2 \le r$ with a small enough value of $r > 0$ (both r and C depend on f). Quadratic convergence means essentially that, eventually, every new step of the process increases by a constant factor the number of accuracy digits in the approximate solution.

When started not close enough to the minimizer, the pure Newton method (Nwt) can demonstrate weird behavior. (Look, e.g., at what happens when the method is applied to the univariate function $f(x) = x^2 + x^4$.) The simplest way to overcome this drawback is to pass from the pure Newton method to its *damped* version

$$x \mapsto x_+ = x - \gamma(x)[f''(x)]^{-1} f'(x), \tag{NwtD}$$

where the stepsize $\gamma(x) > 0$ is chosen in a way that, on one hand, ensures global convergence of the method and, on the other hand, enforces $\gamma(x) \to 1$ as $x \to x_*$, thus ensuring

fast (essentially the same as for the pure Newton method) asymptotic convergence of the process.[57] Practitioners consider the (properly modified) Newton method as the fastest routine, in terms of the iteration count, for smooth (not necessarily convex) unconstrained minimization, although sometimes it is too heavy for practical use. The practical drawbacks of the method are both the necessity to invert the Hessian matrix at each step, which is computationally costly in the large-scale case, and especially the necessity to compute this matrix. (Think how difficult it is to write a code computing 5050 second order derivatives of a messy function of 100 variables.)

6.2.2 Classical interior penalty scheme: Construction

Now consider a constrained convex optimization program. As we remember, one can w.l.o.g. make its objective linear, moving, if necessary, the actual objective to the list of constraints. Thus, let the problem be

$$\min_x \left\{c^T x : x \in \mathcal{X} \subset \mathbf{R}^n\right\}, \tag{C}$$

where $\mathcal{X}$ is a closed convex set, which we assume to possess a nonempty interior. How could we solve the problem?

Traditionally it was thought that the problems of smooth convex unconstrained minimization are easy; thus, a quite natural scheme was to reduce the constrained problem (C) to a series of smooth unconstrained optimization programs. To this end, let us choose somehow a barrier (also called an interior penalty function) $F(x)$ for the feasible set $\mathcal{X}$—a function that is well defined, smooth, strongly convex on the interior of $\mathcal{X}$, and that blows up as a point from $\mathrm{int}\mathcal{X}$ approaches a boundary point of $\mathcal{X}$:

$$x_i \in \mathrm{int}\mathcal{X},\ x \equiv \lim_{i\to\infty} x_i \in \partial\mathcal{X} \Rightarrow \lim_{i\to\infty} F(x_i) = \infty.$$

We next consider the one-parametric family of functions generated by our objective and the barrier:

$$F_t(x) = tc^T x + F(x) : \mathrm{int}\mathcal{X} \to \mathbf{R}.$$

Here t is the penalty parameter, which is assumed to be nonnegative.

It is easily seen that under mild regularity assumptions (e.g., in the case of bounded $\mathcal{X}$, which we assume from now on)

- Every function $F_t(\cdot)$ attains its minimum over the interior of $\mathcal{X}$, the minimizer $x_*(t)$ being unique.
- The central path $x_*(t)$ is a smooth curve, and all its limiting, $t \to \infty$, points belong to the set of optimal solutions of (C).

This fact is quite clear intuitively. To minimize $F_t(\cdot)$ for large t is the same as minimizing the function $f_\rho(x) = f(x) + \rho F(x)$ for small $\rho = \frac{1}{t}$. When ρ is small, the

[57] There are many ways to provide the required behavior of $\gamma(x)$; e.g., choose $\gamma(x)$ by a linesearch in the direction $e(x) = -[f''(x)]^{-1} f'(x)$ of the Newton step: $\gamma(x) = \mathrm{argmin}_t\, f(x + te(x))$.

function f_ρ is very close to f everywhere in $\mathcal{X}$ except a narrow stripe along the boundary of $\mathcal{X}$, the stripe becoming thinner and thinner as $\rho \to 0$. Therefore we have good reasons to believe that the minimizer of F_t for large t (i.e., the minimizer of f_ρ for small ρ) must be close to the set of minimizers of f on $\mathcal{X}$.

We see that the central path $x_*(t)$ is a kind of Ariadne's thread that leads to the solution set of (C). On the other hand, given a value $t \geq 0$ of the penalty parameter, to reach the point $x_*(t)$ on this path is the same as minimizing a smooth strongly convex function $F_t(\cdot)$ which attains its minimum at an interior point of $\mathcal{X}$. The latter problem is nearly unconstrained, up to the fact that its objective is not everywhere defined. However, we can easily adapt the methods of unconstrained minimization, including the Newton one, to handle nearly unconstrained problems. We see that constrained convex optimization in a sense can be reduced to the easy unconstrained one. The conceptually simplest way to make use of this observation is to choose a very large value $\bar{t}$ of the penalty parameter, like $\bar{t} = 10^6$ or $\bar{t} = 10^{10}$, and to run an unconstrained minimization routine, say, the Newton method, on the function $F_{\bar{t}}$, thus getting a good approximate solution to (C) in one shot. This policy, however, is impractical: since we have no idea where $x_*(\bar{t})$ is, normally we will start minimizing $F_{\bar{t}}$ very far from the minimizer of this function, and thus for a long time will be unable to exploit fast local convergence of the unconstrained minimization method we have chosen. A smarter way to use our Ariadne's thread is exactly the one used by Theseus: to follow the thread. Assume, e.g., that we know in advance the minimizer of $F_0 \equiv F$, i.e., the point $x_*(0)$.[58] Thus, we know where the central path starts. Now let us follow this path: at ith step, standing at a point x_i close enough to some point $x_*(t_i)$ of the path, we

- first, increase a bit the current value t_i of the penalty parameter, thus getting a new target point $x_*(t_{i+1})$ on the path; and
- second, approach our new target point $x_*(t_{i+1})$ by running, say, the Newton method on the function $F_{t_{i+1}}$ and starting the method at our current iterate x_i, until a new iterate x_{i+1} close enough to $x_*(t_{i+1})$ is generated.

As a result of these steps, we restore the initial situation—we again stand at a point close to a point on the central path, but this latter point has been moved along the central path toward the optimal set of (C). Iterating this updating and strengthening appropriately our closeness requirements as the process goes on, we approach the optimal set along the central path. A conceptual advantage of this path-following policy as compared to the brute force attempt to reach a target point $x_*(\bar{t})$ with large $\bar{t}$ is that now we have a hope to exploit all the time the strongest feature of our working horse (the Newton method)—its fast local convergence. Indeed, assuming that x_i is close to $x_*(t_i)$ and that we do not increase the penalty parameter too rapidly, so that $x_*(t_{i+1})$ is close to $x_*(t_i)$ (recall that the central path is smooth!), we conclude that x_i is close to our new target point $x_*(t_{i+1})$. If all our requirements of "close enough" and "not too rapidly" are properly controlled, we may ensure x_i to be in the domain of quadratic convergence of the Newton method as applied to $F_{t_{i+1}}$, which means that it will take a quite small number of steps to recover closeness to our new target point.

[58] It is not difficult to ensure this assumption: given an arbitrary barrier F and an arbitrary starting point $\bar{x} \in \mathrm{int}\mathcal{X}$, we can pass from F to a new barrier $\bar{F} = F(x) - (x - \bar{x})^T F'(\bar{x})$ which attains its minimum exactly at $\bar{x}$, and then use the new barrier $\bar{F}$ instead of our original barrier F. For the traditional approach we are following for the time being, F has absolutely no advantages over $\bar{F}$.

6.2.3 Classical interior penalty scheme: Drawbacks

At a qualitative common sense level, the interior penalty scheme looks quite attractive and extremely flexible. For the majority of optimization problems treated by the classical optimization, there are plenty of ways to build a relatively simple barrier meeting all the requirements imposed by the scheme, there is vast room to play with the policies for increasing the penalty parameter and controlling closeness to the central path, etc. The theory says that under quite mild and general assumptions on the choice of the numerous free parameters of our construction, convergence to the optimal set is guaranteed. All looks wonderful, until we realize that the convergence ensured by the theory is completely unqualified. It is a purely asymptotical phenomenon: we are promised to reach eventually a solution of a whatever accuracy we wish, but how long will it take to reach a given accuracy? This is the question the classical optimization theory with its asymptotic linear or superlinear or quadratic convergence neither posed nor answered. Unfortunately, our life in this world is finite (usually more finite than we would like it to be), and so asymptotical promises are perhaps better than nothing but are definitely not the crux of the matter. What is vitally important for us in theory (and to some extent also in practice) is the issue of *complexity*: given an instance of such-and-such generic optimization problem and a desired accuracy ϵ, how large is the computational effort (number of arithmetic operations) needed to get an ϵ-solution of the instance? Certainly, we would like the answer to be a kind of a polynomial time complexity bound and not a quantity depending on unobservable and uncontrollable properties of the instance, like the level of regularity of the boundary of $\mathcal{X}$ at the (unknown!) optimal solution.

It turns out that the intuitively nice classical theory we have outlined is unable to say a single word on the complexity issues. What else could we expect—a reasoning in purely qualitative terms (like "smooth," "strongly convex," etc.) surely cannot yield a quantitative result. Moreover, from the complexity viewpoint the very philosophy of the classical convex optimization is in fact wrong:

• As far as complexity is concerned, for nearly all black box–represented classes of unconstrained convex optimization problems (those where all we know is that the objective is called $f(x)$, it is (strongly) convex, and 2 (3, 4, 5...) times continuously differentiable, and it can be computed, along with its derivatives up to order 1 (2, 3, 4, ...) at every given point), there is no such phenomenon as local quadratic convergence, the Newton method (which uses the second order derivatives) has no advantages over the methods that use only first order derivatives, etc.

• The very idea to reduce black box–represented constrained convex problems to unconstrained ones makes no sense—both classes have the same complexity, and the methods underlying this complexity can work equally well with constrained and unconstrained problems.

6.2.4 But...

Luckily, the pessimistic analysis of the classical interior penalty scheme is not the final truth. What prevents this scheme from yielding a polynomial time method is not the structure of the scheme but the complete freedom it allows for some of its elements. (Too much freedom is another word for anarchy.) After some order is added, the scheme becomes a polynomial time one! Specifically, it was found that

1. there is a (completely nontraditional) class of good (self-concordant[59]) barriers. Every barrier F of this type is associated with a self-concordance parameter $\theta(F)$, which is a real ≥ 1.

2. whenever a barrier F underlying the interior penalty scheme is self-concordant, one can specify the notion of closeness to the central path and the policy for updating the penalty parameter in such a way that a single Newton step

$$x_i \mapsto x_{i+1} = x_i - [\nabla^2 F_{t_{i+1}}(x_i)]^{-1} \nabla F_{t_{i+1}}(x_i) \tag{6.2.1}$$

suffices to update an iterate x_i close to $x_*(t_i)$ into a new iterate x_{i+1} which is close, in the same sense, to $x_*(t_{i+1})$. All these points close to the central path belong to $\operatorname{int}\mathcal{X}$, so that the scheme keeps all the iterates strictly feasible.

3. The penalty updating policy mentioned in the previous item is quite simple:

$$t_i \mapsto t_{i+1} = \left(1 + \frac{0.1}{\sqrt{\theta(F)}}\right) t_i.$$

In particular, it does not slow down as t_i grows and ensures linear growth of the penalty with the ratio $(1 + \frac{0.1}{\sqrt{\theta(F)}})$. This is vitally important due to the next fact.

4. If x is close to a point $x_*(t)$ of the central path, then the inaccuracy of x as an approximate solution to (C) is inverse proportional to t:

$$c^T x - \min_{y \in \mathcal{X}} c^T y \leq \frac{2\theta(F)}{t}.$$

It follows that

> (!) *once we managed to get close to the central path, i.e., we got a point x_0 which is close to a point $x(t_0)$, $t_0 > 0$, on the path, then every $O(\sqrt{\theta(F)})$ steps of the scheme improve the quality of the approximate solutions generated by the scheme by an absolute constant factor. In particular, it takes no more than*
>
> $$O(1)\sqrt{\theta(F)} \ln\left(1 + \frac{\theta(F)}{t_0 \epsilon}\right)$$
>
> *steps to generate a strictly feasible ϵ-solution to* (C).

Note that with our simple penalty updating policy all that is needed to perform a step of the interior penalty scheme is to compute the gradient and the Hessian of the underlying barrier at a single point and to invert the resulting Hessian.

Items 3 and 4 say that essentially all we need to derive a polynomial time method for a generic convex optimization problem is the ability to equip every instance (p) of the problem with a good barrier $F = F_{(p)}$ in such a way that both the self-concordance parameter $\theta(F)$ of F and the arithmetic cost of computing the gradient and the Hessian of F at a given point are polynomial in the size of the instance.[60] It turns out that we can

[59] We do not intend to explain here what a "self-concordant barrier" is. For our purposes it suffices to say that this is a three-times continuously differentiable convex barrier F satisfying a pair of specific differential inequalities linking the first, second, and third directional derivatives of F.

[60] Another requirement is to be able to initialize the method with a point X_0 close to a point $x_*(t_0)$ on the central path, with t_0 being not disastrously small. It turns out that such an initialization is a minor problem which can be resolved via the same path-following technique, provided we are given in advance a strictly feasible solution to our problem.

meet the latter requirement for all interesting well-structured generic convex programs, in particular, for linear, conic quadratic, and semidefinite programming. Moreover, these are particularly nice application fields for the general theory of interior point polynomial time methods, since for them the theory can be both simplified and strengthened.

6.3 Interior point methods for linear programming, conic quadratic programming, and semidefinite programming: Building blocks

We are about to explain what the interior point methods for LP, CQP, SDP look like.

6.3.1 Canonical cones and canonical barriers

We will be interested in a generic conic problem

$$\min_x \left\{c^T x : \mathcal{A}x - B \in \mathbf{K}\right\} \tag{CP}$$

associated with a cone $\mathbf{K}$ given as a direct product of m basic cones, each being either a Lorentz or a semidefinite cone:

$$\mathbf{K} = \mathbf{S}^{k_1}_+ \times \cdots \times \mathbf{S}^{k_p}_+ \times \mathbf{L}^{k_{p+1}} \times \cdots \times \mathbf{L}^{k_m} \subset E = \mathbf{S}^{k_1} \times \cdots \times \mathbf{S}^{k_p} \times \mathbf{R}^{k_{p+1}} \times \cdots \times \mathbf{R}^{k_m}. \tag{Cone}$$

Of course, the generic problem in question covers LP (no Lorentz cones in the right-hand side, and all semidefinite cones are of dimension 1), CQP (no semidefinite cones), and SDP (no Lorentz cones).

We equip the semidefinite and the Lorentz cones with canonical barriers:

- The canonical barrier for a semidefinite cone $\mathbf{S}^k_+$ is

$$S_k(X) = -\ln \mathrm{Det}(X) : \mathrm{int}\mathbf{S}^k_+ \to \mathbf{R};$$

the parameter of this barrier, by definition, is $\theta(S_k) = k$.[61]

- The canonical barrier for a Lorentz cone $\mathbf{L}^k = \left\{x \in \mathbf{R}^k \mid x_k \geq \sqrt{x_1^2 + \cdots + x_{k-1}^2}\right\}$ is

$$L_k(x) = -\ln(x_k^2 - x_1^2 - \cdots - x_{k-1}^2) = -\ln(x^T J_k x), \quad J_k = \begin{pmatrix} -1 & & & & \\ & -1 & & & \\ & & \ddots & & \\ & & & -1 & \\ & & & & 1 \end{pmatrix};$$

the parameter of this barrier is $\theta(L_k) = 2$.

[61] The barrier S_k, same as the canonical barrier L_k for the Lorentz cone $\mathbf{L}^k$, are self-concordant (whatever it means), and the parameters they are assigned here by definition are exactly their parameters of self-concordance. However, we do not use these facts explicitly in our derivations.

• The canonical barrier $K(\cdot)$ for the cone $\mathbf{K}$ given by (Cone), is by definition the direct sum of the canonical barriers of the factors:

$$K(X) = S_{k_1}(X_1) + \cdots + S_{k_p}(X_p) + L_{k_{p+1}}(X_{p+1}) + \cdots + L_{k_m}(X_m),$$

$$X_i \in \begin{cases} \mathrm{int}\mathbf{S}^{k_i}_+, & i \le p, \\ \mathrm{int}\mathbf{L}^{k_i}, & p < i \le m. \end{cases}$$

From now on, we use uppercase Latin letters, like X, Y, Z, to denote elements of the space E; for such an element X, X_i denotes the projection of X onto ith factor in the direct product representation of E as shown in (Cone).

The parameter of the barrier $K(\cdot)$, again by definition, is the sum of parameters of the basic barriers involved:

$$\theta(K) = \theta(S_{k_1}) + \cdots + \theta(S_{k_p}) + \theta(L_{k_{p+1}}) + \cdots + \theta(L_{k_m}) = \sum_{i=1}^{p} k_i + 2(m-p).$$

Recall that all direct factors in the direct product representation (Cone) of our universe E are Euclidean spaces. The matrix factors $\mathbf{S}^{k_i}$ are endowed with the Frobenius inner product

$$\langle X_i, Y_i \rangle_{\mathbf{S}^k} = \mathrm{Tr}(X_i Y_i),$$

while the arithmetic factors $\mathbf{R}^{k_i}$ are endowed with the usual inner product

$$\langle X_i, Y_i \rangle_{\mathbf{R}^k} = X_i^T Y_i.$$

E itself will be regarded as a Euclidean space endowed with the inner product, which is the direct sum of the inner products on the factors:

$$\langle X, Y \rangle_E = \sum_{i=1}^{p} \mathrm{Tr}(X_i Y_i) + \sum_{i=p+1}^{m} X_i^T Y_i.$$

It is clearly seen that our basic barriers (and therefore their direct sum $K(\cdot)$) are indeed barriers for the corresponding cones: they are C^∞-smooth on the interiors of their domains, blow up to ∞ along every sequence of points from these interiors converging to a boundary point of the corresponding domain, and are strongly convex. To verify the latter property, it makes sense to compute explicitly the first and second directional derivatives of these barriers (we need the corresponding formulae in any case); to simplify notation, we write the derivatives of the basic functions S_k, L_k taken at a point x from their domain along a direction h (you should remember that in the case of S_k both x and h, despite their lowercase denotation, are $k \times k$ symmetric matrices):

$$
\begin{array}{|rcl|}
\hline
DS_k(x)[h] \equiv \frac{d}{dt}\Big|_{t=0} S_k(x+th) & = & -\mathrm{Tr}(x^{-1}h) = -\langle x^{-1}, h\rangle_{\mathbf{S}^k}, \\
\text{i.e.,} & & \\
\nabla S_k(x) & = & -x^{-1}; \\
D^2S_k(x)[h,h] \equiv \frac{d^2}{dt^2}\Big|_{t=0} S_k(x+th) & = & \mathrm{Tr}(x^{-1}hx^{-1}h) = \langle x^{-1}hx^{-1}, h\rangle_{\mathbf{S}^k}, \\
\text{i.e.,} & & \\
[\nabla^2 S_k(x)]h & = & x^{-1}hx^{-1}; \\
\hline
DL_k(x)[h] \equiv \frac{d}{dt}\Big|_{t=0} L_k(x+th) & = & -2\frac{h^TJ_kx}{x^TJ_kx}, \\
\text{i.e.,} & & \\
\nabla L_k(x) & = & -\frac{2}{x^TJ_kx}J_kx; \\
D^2L_k(x)[h,h] \equiv \frac{d^2}{dt^2}\Big|_{t=0} L_k(x+th) & = & 4\frac{[h^TJ_kx]^2}{[x^TJ_kx]^2} - 2\frac{h^TJ_kh}{x^TJx}, \\
\text{i.e.,} & & \\
\nabla^2 L_k(x) & = & \frac{4}{[x^TJ_kx]^2}J_kxx^TJ_k - \frac{2}{x^TJ_kx}J_k. \\
\hline
\end{array}
\tag{6.3.2}
$$

Comments on computations for $S_k(\cdot)$: This is a rare case in our adult life when the best way to compute a derivative is to recall the definition of this notion, not just to blindly use calculus rules. Here is the computation:

$$
\begin{array}{rcl}
\ln\mathrm{Det}(x+th) - \ln\mathrm{Det}(x) & = & \ln\mathrm{Det}(x(I+tx^{-1}h)) - \ln\mathrm{Det}(x) \\
& = & \ln\big(\mathrm{Det}(x)\mathrm{Det}(I+tx^{-1}h)\big) - \ln\mathrm{Det}(x) \\
& = & \ln\mathrm{Det}(I+tx^{-1}h) \\
& \underbrace{=}_{(*)} & \ln(1+t\mathrm{Tr}(x^{-1}h)+O(t^2)),\ t\to 0, \\
& = & t\mathrm{Tr}(x^{-1}h)+O(t^2),\ t\to 0, \\
\Rightarrow & & \\
\frac{d}{dt}\Big|_{t=0}\ln\mathrm{Det}(x+th) & = & \mathrm{Tr}(x^{-1}h) = \langle x^{-1}, h\rangle_{\mathbf{S}^k}.
\end{array}
$$

The crucial relation $(*)$ becomes absolutely clear when you think of the decomposition of $\mathrm{Det}(I+tQ)$, for small t, into the alternated sum of products of entries of $I+tQ$ taken along all diagonals of the matrix (definition of determinant!): all terms of the zero and the first order in t come from the product $(1+tQ_{11})(1+tQ_{22})\ldots(1+tQ_{kk})$, and the sum of these zero and first order terms is exactly $1+\mathrm{Tr}(Q)t$.

We have seen that $\frac{d}{dt}S_k(x+th) = -\mathrm{Tr}((x+th)^{-1}h)$. To proceed with differentiation, we should know how to differentiate the matrix $(x+th)^{-1}$ with respect to t at $t=0$. Here is the computation:

$$
\begin{array}{rcl}
(x+th)^{-1} & = & (x(I+tx^{-1}h))^{-1} = (I+tx^{-1}h)^{-1}x^{-1} = (I-tx^{-1}h+O(t^2))x^{-1} \\
& = & x^{-1} - tx^{-1}hx^{-1} + O(t^2),\ t\to 0.
\end{array}
$$

From the expression for $D^2S_k(x)[h,h]$ we see that

$$
D^2S_k(x)[h,h] = \mathrm{Tr}(x^{-1}hx^{-1}h) = \mathrm{Tr}([x^{-1/2}hx^{-1/2}]^2),
$$

so that $D^2S_k(x)[h,h]$ is positive whenever $h \neq 0$. It is not difficult to prove that the same is true for $D^2L_k(x)[h,h]$ (see Exercise 6.1). Thus, the canonical barriers for semidefinite and Lorentz cones are strongly convex, and so is their direct sum $K(\cdot)$.

In what follows, it makes sense to illustrate the general (**K** is given by (Cone)) concepts and results by their particular forms corresponding to the SDP case, where **K** is the semidefinite cone $\mathbf{S}^k_+$. The essence of the matter in our general case is exactly the same as in this particular one, but straightforward computations that are easy in the SDP case become nearly impossible in the general case. And we have no chance to explain here how it is possible (it is!) to get the desired results with the minimum amount of computations.

Due to the role played by the SDP case in our exposition, we use for this case special notation, along with the just-introduced general one. Specifically, we denote the Frobenius inner product on $E = \mathbf{S}^k$ as $\langle \cdot, \cdot \rangle_F$, although feel free, if necessary, to use our general notation $\langle \cdot, \cdot \rangle_E$ as well; the associated norm is denoted by $\| \cdot \|_2$, so that $\|X\|_2 = \sqrt{\mathrm{Tr}(X^2)}$, X being a symmetric matrix.

6.3.2 Elementary properties of canonical barriers

Let us establish a number of simple and useful properties of canonical barriers.

Proposition 6.3.1. *A canonical barrier F (F can be S_k, L_k, or the direct sum K of several copies of these elementary barriers) possesses the following properties:*

(i) *F is logarithmically homogeneous with the parameter of logarithmic homogeneity equal to $-\theta(F)$, i.e., the following identity holds:*

$$t > 0,\ x \in \mathrm{Dom}F \Rightarrow F(tx) = F(x) - \theta(F)\ln t.$$

• In the SDP case, i.e., when $F(x) = S_k(x) = -\ln \mathrm{Det}(x)$ and x is a $k \times k$ positive definite matrix, (i) claims that

$$-\ln \mathrm{Det}(tx) = -\ln \mathrm{Det}(x) - k \ln t,$$

which of course is true.

(ii) *Consequently, the following two equalities hold identically in $x \in \mathrm{Dom}F$:*

$$\begin{array}{ll} \text{(a)} & \langle \nabla F(x), x \rangle = -\theta(F), \\ \text{(b)} & [\nabla^2 F(x)]x = -\nabla F(x). \end{array}$$

• In the SDP case, $\nabla F(x) = \nabla S_k(x) = -x^{-1}$ and $[\nabla^2 F(x)]h = \nabla^2 S_k(x)h = x^{-1}hx^{-1}$ (see (6.3.2)), so (a) becomes the identity $\langle x^{-1}, x \rangle_F \equiv \mathrm{Tr}(x^{-1}x) = k$, and (b) kindly informs us that $x^{-1}xx^{-1} = x^{-1}$.

(iii) *Consequently, ℓth differential $D^\ell F(x)$ of F, $\ell \geq 1$, is homogeneous of degree $-\ell$ in $x \in \mathrm{Dom}F$:*

$$\begin{array}{l} \forall (x \in \mathrm{Dom}F, t > 0, h_1, \ldots, h_\ell): \\ D^\ell F(tx)[h_1, \ldots, h_\ell] \equiv \left. \frac{\partial^\ell F(tx + s_1h_1 + \cdots + s_\ell h_\ell)}{\partial s_1 \partial s_2 \cdots \partial s_\ell} \right|_{s_1 = \cdots = s_\ell = 0} = t^{-\ell} D^\ell F(x)[h_1, \ldots, h_\ell]. \end{array} \tag{6.3.3}$$

Proof. (i) it is immediately seen that S_k and L_k are logarithmically homogeneous with the parameters of logarithmic homogeneity $-\theta(S_k)$, $-\theta(L_k)$, respectively. And of course the property of logarithmic homogeneity is stable with respect to taking direct sums of functions: if $\mathrm{Dom}\Phi$ and $\mathrm{Dom}\Psi$ are closed with respect to the operation of multiplying a vector by a positive scalar, and both Φ and Ψ are logarithmically homogeneous with parameters α, β, respectively, then the function $\Phi(u) + \Psi(v)$ is logarithmically homogeneous with the parameter $\alpha + \beta$.

(ii) To get (ii)(a), it suffices to differentiate the identity

$$F(tx) = F(x) - \theta(F)\ln t$$

in t at $t = 1$,

$$F(tx) = F(x) - \theta(F)\ln t \Rightarrow \langle \nabla F(tx), x\rangle = \frac{d}{dt}F(tx) = -\theta(F)t^{-1},$$

and it remains to set $t = 1$ in the concluding identity.

Similarly, to get (ii)(b), it suffices to differentiate the identity

$$\langle \nabla F(x + th), x + th\rangle = -\theta(F)$$

(which is just (ii)(a)) in t at $t = 0$, thus arriving at

$$\langle [\nabla^2 F(x)]h, x\rangle + \langle \nabla F(x), h\rangle = 0;$$

since $\langle [\nabla^2 F(x)]h, x\rangle = \langle [\nabla^2 F(x)]x, h\rangle$ (symmetry of partial derivatives!) and since the resulting equality

$$\langle [\nabla^2 F(x)]x, h\rangle + \langle \nabla F(x), h\rangle = 0$$

holds true identically in h, we come to $[\nabla^2 F(x)]x = -\nabla F(x)$.

(iii) Differentiating ℓ times the identity

$$F(tx) = F(x) - \theta\ln t$$

in x, we get

$$t^\ell D^\ell F(tx)[h_1, \ldots, h_\ell] = D^\ell F(x)[h_1, \ldots, h_\ell]. \qquad \square$$

An especially nice specific feature of canonical barriers is their self-duality:

Proposition 6.3.2. *A canonical barrier F (F can be S_k, L_k, or the direct sum K of several copies of these elementary barriers) possesses the following property: for every $x \in \mathrm{Dom}F$, $-\nabla F(x)$ belongs to $\mathrm{Dom}F$ as well, and the mapping $x \mapsto -\nabla F(x) : \mathrm{Dom}F \to \mathrm{Dom}F$ is self-inverse,*

$$-\nabla F(-\nabla F(x)) = x \quad \forall x \in \mathrm{Dom}F, \tag{6.3.4}$$

and is homogeneous of degree -1:

$$t > 0, x \in \mathrm{intdom}F \Rightarrow -\nabla F(tx) = -t^{-1}\nabla F(x). \tag{6.3.5}$$

In the SDP case $F = S_k$, x is a $k \times k$ positive definite matrix, and $\nabla F(x) = \nabla S_k(x) = -x^{-1}$ (see (6.3.2)) so that the above statements merely say that the mapping $x \mapsto x^{-1}$ is a self-inverse one-to-one mapping of the interior of the semidefinite cone onto itself, and that $-(tx)^{-1} = -t^{-1}x^{-1}$, both claims being trivially true.

The proof of this proposition for the general case is the subject of Exercise 6.2.

6.4 Primal-dual pair of problems and the primal-dual central path

6.4.1 Problem(s)

It is reasonable to consider simultaneously the problem of interest (CP) and its conic dual. Since $\mathbf{K}$ is a direct product of self-dual cones, this dual is a conic problem on the same cone $\mathbf{K}$. As we remember from Lecture 2, the primal-dual pair associated with (CP) is

$$\min_x \left\{c^T x : \mathcal{A}x - B \in \mathbf{K}\right\}, \qquad \text{(CP)}$$
$$\max_S \left\{\langle B, S\rangle_E : \mathcal{A}^* S = c,\ S \in \mathbf{K}\right\}. \qquad \text{(CD)}$$

Assuming that the linear mapping $x \mapsto \mathcal{A}x$ is an embedding (i.e., that $\mathrm{Null}(\mathcal{A}) = \{0\}$)—this is Assumption **A** from Lecture 2—we can write down our primal-dual pair in a symmetric geometric form (Lecture 2, section 2.3.1):

$$\min_X \left\{\langle C, X\rangle_E : X \in (\mathcal{L} - B) \cap \mathbf{K}\right\}, \qquad \text{(P)}$$
$$\max_S \left\{\langle B, S\rangle_E : S \in (\mathcal{L}^\perp + C) \cap \mathbf{K}\right\}, \qquad \text{(D)}$$

where $\mathcal{L}$ is a linear subspace in E (the image space of the linear mapping $x \mapsto \mathcal{A}x$), $\mathcal{L}^\perp$ is the orthogonal complement to $\mathcal{L}$ in E, and $C \in E$ satisfies $\mathcal{A}^*C = c$, i.e., $\langle C, \mathcal{A}x\rangle_E \equiv c^T x$.

To simplify things, from now on we assume that both problems (CP) and (CD) are strictly feasible. In terms of (P) and (D) this assumption means that both the primal feasible plane $\mathcal{L} - B$ and the dual feasible plane $\mathcal{L}^\perp + C$ intersect the interior of the cone $\mathbf{K}$.

Remark 6.4.1. By the conic duality theorem (Lecture 2), both (CP) and (D) are solvable with equal optimal values:

$$\mathrm{Opt(CP)} = \mathrm{Opt(D)}.$$

(Recall that we have assumed strict primal-dual feasibility.) Since (P) is equivalent to (CP), (P) is solvable as well, and the optimal value of (P) differs from the one of (CP) by $\langle C, B\rangle_E$.[62] It follows that the optimal values of (P) and (D) are linked by the relation

$$\mathrm{Opt(P)} - \mathrm{Opt(D)} + \langle C, B\rangle_E = 0. \qquad (6.4.6)$$

[62] Indeed, the values of the respective objectives $c^T x$ and $\langle C, \mathcal{A}x - B\rangle_E$ at the feasible solutions x of (CP) and $X = \mathcal{A}x - B$ of (P) corresponding to each other differ from each other by exactly $\langle C, B\rangle_E$:

$$c^T x - \langle C, X\rangle_E = c^T x - \langle C, \mathcal{A}x - B\rangle_E = \underbrace{c^T x - \langle \mathcal{A}^* C, x\rangle_E}_{=0 \text{ due to } \mathcal{A}^*C = c} + \langle C, B\rangle_E.$$

6.4.2 Central path(s)

The canonical barrier K of $\mathbf{K}$ induces a barrier for the feasible set $\mathcal{X} = \{x \mid \mathcal{A}x - B \in \mathbf{K}\}$ of the problem (CP) written down in the form of (C), i.e., as

$$\min_x \left\{c^T x : x \in \mathcal{X}\right\};$$

this barrier is

$$\widehat{K}(x) = K(\mathcal{A}x - B) : \mathrm{int}X \to \mathbf{R} \tag{6.4.7}$$

and is indeed a barrier. Now we can apply the interior penalty scheme to trace the central path $x_*(t)$ associated with this barrier. With some effort it can be derived from the primal-dual strict feasibility that the central path is well defined, i.e., that the minimizer of

$$\widehat{K}_t(x) = tc^T x + \widehat{K}(x)$$

on $\mathrm{int}X$ exists for every $t \geq 0$ and is unique.[63] What is important for us at the moment is the central path itself, not how to trace it. Moreover, it is highly instructive to pass from the central path $x_*(t)$ in the space of design variables to its image

$$X_*(t) = \mathcal{A}x_*(t) - B$$

in E. The resulting curve is called the *primal central path of the primal-dual pair* (P), (D); by its origin, it is a curve comprising strictly feasible solutions of (P) (since it is the same to say that x belongs to the (interior of) the set $\mathcal{X}$ and to say that $X = \mathcal{A}x - B$ is a (strictly) feasible solution of (P)). A simple yet very useful observation is that the primal central path can be defined solely in terms of (P), (D) and thus it is a geometric entity—it is independent of a particular parameterization of the primal feasible plane $\mathcal{L} - B$ by the design vector x:

(∗) *A point $X_*(t)$ of the primal central path is the minimizer of the aggregate*

$$P_t(X) = t\langle C, X\rangle_E + K(X)$$

on the set $(\mathcal{L} - B) \cap \mathrm{int}\mathbf{K}$ of strictly feasible solutions of (P).

This observation is just a tautology: $x_*(t)$ is the minimizer on $\mathrm{int}\mathcal{X}$ of the aggregate

$$\widehat{K}_t(x) \equiv tc^T x + \widehat{K}(x) = t\langle C, \mathcal{A}x\rangle_E + K(\mathcal{A}x - B) = P_t(\mathcal{A}x - B) + t\langle C, B\rangle_E;$$

we see that for $x \in \mathrm{int}\mathcal{X}$ the function $\widehat{P}_t(x) = P_t(\mathcal{A}x - B)$ differs from the function $\widehat{K}_t(x)$ by a constant (depending on t) and has therefore the same minimizer $x_*(t)$ as the function $\widehat{K}_t(x)$. Now, when x runs through $\mathrm{int}\mathcal{X}$, the point $X = \mathcal{A}x - B$ runs exactly through the set of strictly feasible solutions of (P), so that the minimizer X_* of P_t on the latter set and the minimizer $x_*(t)$ of the function $\widehat{P}_t(x) = P_t(\mathcal{A}x - B)$ on $\mathrm{int}\mathcal{X}$ are linked by the relation $X_* = \mathcal{A}x_*(t) - B$.

[63] In section 6.1, there was no problem with the existence of the central path, since there, $\mathcal{X}$ was assumed to be bounded. In our present context, $\mathcal{X}$ is not necessarily bounded.

The analytic translation of the above observation is as follows:

($*'$) *A point $X_*(t)$ of the primal central path is exactly the strictly feasible solution X to* (P) *such that the vector $tC + \nabla K(X) \in E$ is orthogonal to $\mathcal{L}$ (i.e., belongs to $\mathcal{L}^\perp$).*

Indeed, we know that $X_*(t)$ is the unique minimizer of the smooth convex function $P_t(X) = t\langle C, X\rangle_E + K(X)$ on $(\mathcal{L} - B) \cap \mathrm{int}\mathbf{K}$. A necessary and sufficient condition for a point X of this intersection to minimize P_t over the intersection is that ∇P_t must be orthogonal to $\mathcal{L}$.

In the SDP case, a point $X_*(t)$, $t > 0$, of the primal central path is uniquely defined by the following two requirements: $X_*(t) \succ 0$ should be feasible for (P), and the $k \times k$ matrix

$$tC - X_*^{-1}(t) = tC + \nabla S_k(X_*(t))$$

(see (6.3.2)) should belong to $\mathcal{L}^\perp$, i.e., should be orthogonal (with respect to the Frobenius inner product) to every matrix of the form $\mathcal{A}x$.

The dual problem (D) is in no sense worse than the primal problem (P) and thus also possesses the central path, now called the *dual central path* $S_*(t)$, $t \geq 0$, of the primal-dual pair (P), (D). Similar to (*), ($*'$), the dual central path can be characterized as follows:

($**'$) *A point $S_*(t)$, $t \geq 0$, of the dual central path is the unique minimizer of the aggregate*

$$D_t(S) = -t\langle B, S\rangle_E + K(S)$$

on the set of strictly feasible solutions of (D).[64] *$S_*(t)$ is exactly the strictly feasible solution S to* (D) *such that the vector $-tB + \nabla F(S)$ is orthogonal to $\mathcal{L}^\perp$ (i.e., belongs to $\mathcal{L}$).*

In the SDP case, a point $S_*(t)$, $t > 0$, of the dual central path is uniquely defined by the following two requirements: $S_*(t) \succ 0$ should be feasible for (D), and the $k \times k$ matrix

$$-tB - S_*^{-1}(t) = -tB + \nabla S_k(S_*(t))$$

(see (6.3.2)) should belong to $\mathcal{L}$, i.e., should be representable in the form $\mathcal{A}x$ for some x.

From Proposition 6.3.2 we can derive a wonderful connection between the primal and the dual central paths.

Theorem 6.4.1. *For $t > 0$, the primal and the dual central paths $X_*(t)$, $S_*(t)$ of a (strictly feasible) primal-dual pair* (P), (D) *are linked by the relations*

$$\begin{aligned} S_*(t) &= -t^{-1}\nabla K(X_*(t)), \\ X_*(t) &= -t^{-1}\nabla K(S_*(t)). \end{aligned} \tag{6.4.8}$$

[64] Note the slight asymmetry between the definitions of the primal aggregate P_t and the dual aggregate D_t: in the former, the linear term is $t\langle C, X\rangle_E$, while in the latter it is $-t\langle B, S\rangle_E$. This asymmetry is in complete accordance with the fact that we write (P) as a minimization, and (D)—as a maximization problem. To write (D) in exactly the same form as (P), we were supposed to replace B with $-B$, thus getting the formula for D_t completely similar to the one for P_t.

Proof. By (*′), the vector $tC+\nabla K(X_*(t))$ belongs to $\mathcal{L}^\perp$, so the vector $S = -t^{-1}\nabla K(X_*(t))$ belongs to the dual feasible plane $\mathcal{L}^\perp + C$. On the other hand, by Proposition 6.4.8 the vector $-\nabla K(X_*(t))$ belongs to $\mathrm{Dom}K$, i.e., to the interior of $\mathbf{K}$; since $\mathbf{K}$ is a cone and $t > 0$, the vector $S = -t^{-1}\nabla K(X_*(t))$ belongs to the interior of $\mathbf{K}$ as well. Thus, S is a strictly feasible solution of (D). Now let us compute the gradient of the aggregate D_t at the point S:

$$\begin{array}{rcl}
\nabla D_t(S) & = & -tB + \nabla K(-t^{-1}\nabla K(X_*(t))) \\
& = & -tB + t\nabla K(-\nabla K(X_*(t))) \\
& & \text{[we have used (6.3.5)]} \\
& = & -tB - tX_*(t) \\
& & \text{[we have used (6.3.4)]} \\
& = & -t(B + X_*(t)) \\
& \in & \mathcal{L} \\
& & \text{[since } X_*(t) \text{ is primal feasible]}.
\end{array}$$

Thus, S is strictly feasible for (D) and $\nabla D_t(S) \in \mathcal{L}$. But by (**′) these properties characterize $S_*(t)$; thus, $S_*(t) = S \equiv -t^{-1}\nabla K(X_*(t))$. In view of Proposition 6.3.2, this implies that $X_*(t) = -t^{-1}\nabla K(S_*(t))$. Another way to get the latter relation from the relation $S_*(t) = -t^{-1}\nabla K(X_*(t))$ is just to refer to the primal-dual symmetry. □

In fact, the connection between the primal and the dual central path stated by Theorem 6.4.1 can be used to characterize both the paths.

THEOREM 6.4.2. *Let* (P), (D) *be a strictly feasible primal-dual pair.*

For every $t > 0$, there exists a unique strictly feasible solution X of (P) *such that $-t^{-1}\nabla K(X)$ is a feasible solution to* (D)*, and this solution X is exactly $X_*(t)$.*

Similarly, for every $t > 0$, there exists a unique strictly feasible solution S of (D) *such that $-t^{-1}\nabla K(S)$ is a feasible solution of* (P)*, and this solution S is exactly $S_*(t)$.*

Proof. By primal-dual symmetry, it suffices to prove the first claim. We already know (Theorem 6.4.1) that $X = X_*(t)$ is a strictly feasible solution of (P) such that $-t^{-1}\nabla K(X)$ is feasible for (D). All we need to prove is that $X_*(t)$ is the only point with these properties, which is immediate: if X is a strictly feasible solution of (P) such that $-t^{-1}\nabla K(X)$ is dual feasible, then $-t^{-1}\nabla K(X) \in \mathcal{L}^\perp + C$, whence $\nabla K(X) \in \mathcal{L}^\perp - tC$, or, which is the same, $\nabla P_t(X) = tC + \nabla K(X) \in \mathcal{L}^\perp$. We already know from (*′) that the latter property, together with the strict primal feasibility, is characteristic for $X_*(t)$. □

On the central path

As we have seen, the primal and dual central paths are intrinsically linked one to another, and it makes sense to treat them as a unique entity—the primal-dual central path of the primal-dual pair (P), (D). The primal-dual central path is just a curve $(X_*(t), S_*(t))$ in $E \times E$ such that its projection onto the primal space is the primal central path, and its projection onto the dual space is the dual central path.

To save words, from now on we refer to the primal-dual central path simply as the *central path*.

The central path possesses a number of extremely nice properties; let us list some of them.

Characterization of the central path. By Theorem 6.4.2, the points $(X_*(t), S_*(t))$ of the central path possess the following properties:

- CP_1 (primal feasibility) *The point $X_*(t)$ is strictly primal feasible.*
- CP_2 (dual feasibility) *The point $S_*(t)$ is dual feasible.*
- CP_3 (augmented complementary slackness) *The points $X_*(t)$ and $S_*(t)$ are linked by the relation*

$$S_*(t) = -t^{-1}\nabla K(X_*(t)) \quad [\Leftrightarrow X_*(t) = -t^{-1}\nabla K(S_*(t))].$$

In the SDP case, $\nabla K(U) = \nabla S_k(U) = -U^{-1}$ (see (6.3.2)), and the augmented complementary slackness relation takes the nice form

$$X_*(t)S_*(t) = t^{-1}I. \tag{6.4.9}$$

In fact, properties CP_1, CP_2, CP_3 fully characterize the central path: if two points X, S possess these properties with respect to some $t > 0$, then $X = X_*(t)$ and $S = S_*(t)$ (Theorem 6.4.2).

Duality gap along the central path. Recall that for an arbitrary primal-dual feasible pair (X, S) of the (strictly feasible!) primal-dual pair of problems (P), (D), the duality gap

$$\begin{aligned}\text{DualityGap}(X, S) &\equiv [\langle C, X\rangle_E - \text{Opt(P)}] + [\text{Opt(D)} - \langle B, S\rangle_E] \\ &= \langle C, X\rangle_E - \langle B, S\rangle_E + \langle C, B\rangle_E\end{aligned}$$

(see (6.4.6)), which measures the total inaccuracy of X, S as approximate solutions of the respective problems, can be written equivalently as $\langle S, X\rangle_E$ (see Proposition 2.4.1 in section 2.4). Now, what is the duality gap along the central path? The answer is immediate:

$$\begin{aligned}\text{DualityGap}(X_*(t), S_*(t)) &= \langle S_*(t), X_*(t)\rangle_E \\ &= \langle -t^{-1}\nabla K(X_*(t)), X_*(t)\rangle_E \\ &\quad \text{[see (6.4.8)]} \\ &= t^{-1}\theta(K) \\ &\quad \text{[see Proposition 6.3.1(ii)]}\end{aligned}$$

We have arrived at a wonderful result.[65]

[65] Which, among other, much more important consequences, explains the name "augmented complementary slackness" of the property CP_3: at the primal-dual pair of *optimal* solutions X^*, S^* the duality gap should be zero—$\langle S^*, X^*\rangle_E = 0$. Property CP_3, as we just have seen, implies that the duality gap at a primal-dual pair $(X_*(t), S_*(t))$ on the central path, although nonzero, is controllable (namely, is equal to $\frac{\theta(K)}{t}$) and becomes small as t grows.

Proposition 6.4.1. *Under assumption of primal-dual strict feasibility, the duality gap along the central path is inverse proportional to the penalty parameter, and the proportionality coefficient is the parameter $\theta(K)$ of the canonical barrier K:*

$$\mathrm{DualityGap}(X_*(t), S_*(t)) = \frac{\theta(K)}{t}.$$

In particular, both $X_*(t)$ and $S_*(t)$ are strictly feasible $(\frac{\theta(K)}{t})$-approximate solutions to their respective problems:

$$\begin{aligned} \langle C, X_*(t)\rangle_E - \mathrm{Opt}(P) &\leq \tfrac{\theta(K)}{t}, \\ \mathrm{Opt}(D) - \langle B, S_*(t)\rangle_E &\leq \tfrac{\theta(K)}{t}. \end{aligned}$$

In the SDP case, $\mathbf{K} = \mathbf{S}^k_+$ and $\theta(K) = \theta(S_k) = k$.
We see that

all we need to get quickly good primal and dual approximate solutions is to trace fast the central path. If we were interested in solving only one of the problems (P), (D), it would be sufficient to trace fast the associated—primal or dual—component of this path. The quality guarantees we get in such a process depend—in a completely universal fashion!—solely on the value t of the penalty parameter and on the value of the parameter $\theta(K)$ of the canonical barrier K and are completely independent of other elements of the data.

Near the central path

The conclusion we have just made is a bit too optimistic: Well, our life when moving along the central path would be just fine (at the very least, we would know how good the solutions we already have are), but how could we move *exactly* along the path? Among the relations CP_1–CP_3 defining the path, the first two are simple—just linear—but the third is in fact a system of *nonlinear* equations, and we have no hope to satisfy these equations exactly. Thus, we arrive at the crucial question, which, a bit informally, is as follows:

How close should we be to the path (and in what sense close) for our life to be as nice as if we were exactly on the path?

There are several ways to answer this question; we will present the simplest one.

Distance to the central path. Our canonical barrier $K(\cdot)$ is a strongly convex smooth function on $\mathrm{int}\mathbf{K}$; in particular, its Hessian matrix $\nabla^2 K(Y)$, taken at a point $Y \in \mathrm{int}\mathbf{K}$, is positive definite. We can use the inverse of this matrix to measure the distances between points of E, thus arriving at the norm

$$\|H\|^*_Y = \sqrt{\langle [\nabla^2 K(Y)]^{-1} H, H\rangle_E}.$$

It turns out that

> *a good measure of proximity of a strictly feasible primal-dual pair $Z = (X, S)$ to a point $Z_*(t) = (X_*(t), S_*(t))$ from the primal-dual central path is the quantity*
>
> $$\mathrm{dist}(Z, Z_*(t)) \equiv \|tS+\nabla K(X)\|_X^* \equiv \sqrt{\langle [\nabla^2 K(X)]^{-1}(tS + \nabla K(X)), tS + \nabla K(X)\rangle_E}.$$
>
> *Although written in a form not symmetric with respect to X, S, this quantity is in fact symmetric in X, S: it turns out that*
>
> $$\|tS + \nabla K(X)\|_X^* = \|tX + \nabla K(S)\|_S^* \tag{6.4.10}$$
>
> *$\forall t > 0$ and $S, X \in \mathrm{int}\mathbf{K}$.*

Observe that $\mathrm{dist}(Z, Z_*(t)) \geq 0$, and $\mathrm{dist}(Z, Z_*(t)) = 0$ if and only if $S = -t^{-1}\nabla K(X)$, which, for a strictly primal-dual feasible pair $Z = (X, S)$, means that $Z = Z_*(t)$ (see the characterization of the primal-dual central path); thus, $\mathrm{dist}(Z, Z_*(t))$ can indeed be viewed as a kind of distance from Z to $Z_*(t)$.

In the SDP case X, S are $k \times k$ symmetric matrices, and

$$\begin{array}{rcl}
\mathrm{dist}^2(Z, Z_*(t)) & = & (\|tS + \nabla S_k(X)\|_X^*)^2 \\
& = & \langle [\nabla^2 S_k(X)]^{-1}(tS + \nabla S_k(X)), tS + \nabla S_k(X)\rangle_F \\
& = & \mathrm{Tr}\left(X(tS - X^{-1})X(tS - X^{-1})\right) \\
& & [\text{see (6.3.2)}] \\
& = & \mathrm{Tr}([tX^{1/2}SX^{1/2} - I]^2),
\end{array}$$

so that

$$\mathrm{dist}^2(Z, Z_*(t)) = \mathrm{Tr}\left(X(tS - X^{-1})X(tS - X^{-1})\right) = \|tX^{1/2}SX^{1/2} - I\|_2^2. \tag{6.4.11}$$

To prove the symmetry claimed in (6.4.10), note that

$$\begin{array}{rcl}
\|tX^{1/2}SX^{1/2} - I\|_2^2 & = & \mathrm{Tr}\left([tX^{1/2}SX^{1/2} - I]^2\right) \\
& = & \mathrm{Tr}\left(t^2X^{1/2}SX^{1/2}X^{1/2}SX^{1/2} - 2tX^{1/2}SX^{1/2} + I\right) \\
& = & \mathrm{Tr}(t^2X^{1/2}SXSX^{1/2}) - 2t\mathrm{Tr}(X^{1/2}SX^{1/2}) + \mathrm{Tr}(I) \\
& = & \mathrm{Tr}(t^2XSXS - 2tXS + I) \\
& = & \mathrm{Tr}(t^2SXSX - 2tSX + I) \\
& = & \mathrm{Tr}(t^2S^{1/2}XS^{1/2}S^{1/2}XS^{1/2} - 2tS^{1/2}XS^{1/2} + I) \\
& = & \mathrm{Tr}([tS^{1/2}XS^{1/2} - I]^2) = \|tS^{1/2}XS^{1/2} - I\|_2^2.
\end{array}$$

In a moderate $\mathrm{dist}(\cdot, Z_*(\cdot))$-neighborhood of the central path. It turns out that in such a neighborhood all is essentially as fine as on the central path itself.

Proposition 6.4.2. *If $Z = (X, S)$ is a pair of primal-dual strictly feasible solutions to* (P), (D) *such that*

$$\mathrm{dist}(Z, Z_*(t)) \leq 1, \tag{Close}$$

then Z is essentially as good as $Z_*(t)$, namely, the duality gap at (X, S) is essentially as small as at the point $Z_*(t)$:

$$\text{DualityGap}(X, S) = \langle S, X\rangle_E \le 2\,\text{DualityGap}(Z_*(t)) = \frac{2\theta(K)}{t}. \tag{6.4.12}$$

Let us verify Proposition 6.4.2 in the SDP case. Suppose that (t, X, S) satisfies (Close). The duality gap at the pair (X, S) of strictly primal-dual feasible solutions is

$$\text{DualityGap}(X, S) = \langle X, S\rangle_F = \text{Tr}(XS),$$

while by (6.4.11) the relation $\text{dist}((S, X), Z_*(t)) \le 1$ means that

$$\|tX^{1/2}SX^{1/2} - I\|_2 \le 1,$$

hence

$$\|X^{1/2}SX^{1/2} - t^{-1}I\|_2 \le \frac{1}{t}.$$

Denoting by δ the vector of eigenvalues of the symmetric matrix $X^{1/2}SX^{1/2}$, we conclude that $\sum_{i=1}^k(\delta_i - t^{-1})^2 \le t^{-2}$, whence

$$\begin{aligned}
\text{DualityGap}(X, S) &= \text{Tr}(XS) = \text{Tr}(X^{1/2}SX^{1/2}) = \sum_{i=1}^k \delta_i \\
&\le kt^{-1} + \sum_{i=1}^k |\delta_i - t^{-1}| \le kt^{-1} + \sqrt{k}\sqrt{\sum_{i=1}^k (\delta_i - t^{-1})^2} \\
&\le kt^{-1} + \sqrt{k}t^{-1},
\end{aligned}$$

and (6.4.12) follows.

It follows from Proposition 6.4.2 that

For our purposes, it is essentially the same—to move along the primal-dual central path, or to trace this path, staying in its time-space neighborhood

$$\mathcal{N}_\kappa = \{(t, X, S) \mid X \in \mathcal{L} - B, S \in \mathcal{L}^\perp + C, t > 0, \text{dist}((X, S), (X_*(t), S_*(t))) \le \kappa\} \tag{6.4.13}$$

with certain $\kappa \le 1$.

Most of the interior point methods for LP, CQP, and SDP, including those most powerful in practice, solve the primal-dual pair (P), (D) by tracing the central path,[66] although not all of them keep the iterates in $\mathcal{N}_{O(1)}$; some of the methods work in much wider neighborhoods of the central path, in order to avoid slowing down when passing highly curved segments of the path. At the level of ideas, these long-step path following methods do not differ essentially from the short-step ones (those keeping the iterates in $\mathcal{N}_{O(1)}$). This is why in the analysis of our forthcoming presentation we restrict ourselves to the short-step methods. It should be added that as far as the theoretical efficiency estimates are concerned, the short-step methods yield the best complexity bounds known so far for LP, CQP and SDP, although in practice the long-step methods usually outperform their short-step counterparts.

[66] There exist also potential reduction interior point methods that do not take explicit care of tracing the central path; an example is the very first interior point method for LP—the method of Karmarkar. The potential reduction IP methods are beyond the scope of our course, which is not a big loss for a practically oriented reader since, as a practical tool, these methods are considered almost obsolete.

6.5 Tracing the central path

6.5.1 Path-following scheme

Assume we are solving a strictly feasible primal-dual pair of problems (P), (D) and intend to trace the associated central path. For this purpose, we just need a mechanism for updating a current iterate $(\bar{t}, \bar{X}, \bar{S})$ such that $\bar{t} > 0$, $\bar{X}$ is strictly primal feasible, $\bar{S}$ is strictly dual feasible, and $(\bar{X}, \bar{S})$ is a good (in certain precise sense) approximation of the point $Z_*(\bar{t}) = (X_*(\bar{t}), S_*(\bar{t}))$ on the central path, into a new iterate (t_+, X_+, S_+) with similar properties and a larger value $t_+ > \bar{t}$ of the penalty parameter. Given such an updating mechanism and iterating it, we indeed shall trace the central path, with all the benefits (see above) coming from the latter fact.[67] How could we construct the required updating? Recalling the description of the central path, we rephrase our question as follows:

Given a triple $(\bar{t}, \bar{X}, \bar{S})$ *that satisfies the relations*

$$\begin{array}{rcl} X & \in & \mathcal{L} - B, \\ S & \in & \mathcal{L}^\perp + C \end{array} \tag{6.5.14}$$

(which is in fact a system of linear equations) and approximately satisfies the system of nonlinear equations

$$G_t(X, S) \equiv S + t^{-1}\nabla K(X) = 0, \tag{6.5.15}$$

update it into a new triple (t_+, X_+, S_+) *with the same properties and such that* $t_+ > \bar{t}$.

Since the left-hand side $G(\cdot)$ in our system of nonlinear equations is smooth around $(\bar{t}, \bar{X}, \bar{S})$ (recall that $\bar{X}$ was assumed to be strictly primal feasible), the most natural way, from the viewpoint of computational mathematics, to achieve our target is as follows:

1. Choose somehow a desired new value $t_+ > \bar{t}$ of the penalty parameter.

2. Linearize the left-hand-side $G_{t_+}(X, S)$ of the system of nonlinear equations (6.5.15) at the point $(\bar{X}, \bar{S})$, and replace (6.5.15) with the linearized system of equations

$$G_{t_+}(\bar{X}, \bar{S}) + \frac{\partial G_{t_+}(\bar{X}, \bar{S})}{\partial X}(X - \bar{X}) + \frac{\partial G_{t_+}(\bar{X}, \bar{S})}{\partial S}(S - \bar{S}) = 0. \tag{6.5.16}$$

3. Compute the corrections ΔX, ΔS from the requirement that the updated pair $X_+ = \bar{X} + \Delta X$, $S_+ = \bar{S} + \Delta S$ must satisfy (6.5.14) and (6.5.16) (a linearized version of (6.5.15)). In other words, the corrections should solve the system

$$\begin{array}{c} \Delta X \in \mathcal{L}, \\ \Delta S \in \mathcal{L}^\perp, \\ G_{t_+}(\bar{X}, \bar{S}) + \frac{\partial G_{t_+}(\bar{X}, \bar{S})}{\partial X}\Delta X + \frac{\partial G_{t_+}(\bar{X}, \bar{S})}{\partial S}\Delta S = 0. \end{array} \tag{6.5.17}$$

[67] Of course, besides knowing how to trace the central path, we should also know how to initialize this process. There are different techniques to resolve this initialization difficulty, and all of them achieve the goal by using the same path-tracing technique, now applied to an appropriate auxiliary problem where the initialization difficulty does not arise at all. Thus, at the level of ideas the initialization techniques do not add something essentially new, and so we skip in our presentation all initialization-related issues.

4. Complete updating the current solution by setting

$$\begin{array}{rcl} X_+ & = & \bar{X} + \Delta X, \\ S_+ & = & \bar{S} + \Delta S. \end{array} \tag{6.5.18}$$

The primal-dual interior point methods we are describing basically fit the outlined scheme, up to the following two important points:

• If the current iterate $(\bar{X}, \bar{S})$ is not close enough to $Z_*(\bar{t})$, or if the desired improvement $t_+ - \bar{t}$ is too large, the corrections given by the outlined scheme may be too large. As a result, the updating (6.5.18) may be inappropriate, e.g., X_+, or S_+, or both, may be kicked out of the cone $\mathbf{K}$. (After all, the linearized system (6.5.16) approximates well the true system (6.5.15) only locally, and we have no reasons to trust in corrections coming from the linearized system, when these corrections are large.)

There is a standard way to overcome the outlined difficulty—to use the corrections in a damped fashion, namely, to replace the updating (6.5.18) with

$$\begin{array}{rcl} X_+ & = & \bar{X} + \alpha \Delta X, \\ S_+ & = & \bar{S} + \beta \Delta S, \end{array} \tag{6.5.19}$$

and to choose the stepsizes $\alpha > 0$, $\beta > 0$ from additional safety considerations, like ensuring the updated pair (X_+, S_+) to reside in the interior of $\mathbf{K}$, or enforcing it to stay in a desired neighborhood of the central path, etc. In interior point methods, the solution $(\Delta X, \Delta S)$ of (6.5.17) plays the role of a search direction, and the actual corrections are proportional to ΔX, ΔS rather than equal to these directions. In this sense the situation is completely similar to the one with the Newton method from section 6.2.1 (which is natural: the latter method is exactly the linearization method for solving the Fermat equation $\nabla f(x) = 0$).

• The augmented complementary slackness system (6.5.15) can be written in many different forms that are equivalent to each other in the sense that they share a common solution set. For example, we have the same reasons to express the augmented complementary slackness requirement by the nonlinear system (6.5.15) as to express it by the system

$$\widehat{G}_t(X, S) \equiv X + t^{-1} \nabla K(S) = 0,$$

not speaking about other possibilities. Note that although all systems of nonlinear equations

$$H_t(X, S) = 0$$

expressing the augmented complementary slackness are equivalent in the sense that they share a common solution set and their linearizations are different and thus lead to different search directions and finally to different path-following methods. Choosing appropriate analytic representation of the augmented complementary slackness requirement (perhaps varying from iteration to iteration), one can gain substantially in the performance of the resulting path-following method. The interior point machinery indeed facilitates this flexibility (see "SDP case examples" below).

6.5.2 Speed of path-tracing

In the LP-CQP-SDP situation, the speed at which the best, from the theoretical viewpoint, path-following methods manage to trace the path is inversely proportional to the square root

of the parameter $\theta(K)$ of the underlying canonical barrier. This means that when started at a point (t^0, X^0, S^0) from the neighborhood $\mathcal{N}_{0.1}$ of the central path, the method after $O(1)\sqrt{\theta(K)}$ steps reaches the point $(t^1 = 2t^0, X^1, S^1) \in \mathcal{N}_{0.1}$, after the same $O(1)\sqrt{\theta(K)}$ steps more reaches the point $(t^2 = 2^2 t^0, X^2, S^2) \in \mathcal{N}_{0.1}$, and so on. Thus, it takes a fixed number $O(1)\sqrt{\theta(K)}$ steps to increase the current value of the penalty parameter by factor 2, staying all the time in $\mathcal{N}_{0.1}$. From (6.4.12) it then follows that every $O(1)\sqrt{\theta(K)}$ steps of the method reduce the (upper bound on the) inaccuracy of current approximate solutions by factor 2 or, which is the same, add a fixed number of accuracy digits to these solutions. Thus, the cost of an accuracy digit for the (best) path-following methods is $O(1)\sqrt{\theta(K)}$ steps. To realize what the implications are, we should, of course, know how heavy a step is—what its arithmetic cost is. The cost of a step in the cheapest among the fastest interior point methods as applied to (CP) is as if all operations at a step were those required by

1. Assembling, for a given $X \in \mathrm{int}\mathbf{K}$, the symmetric $n \times n$ matrix ($n = \dim x$)

$$\mathcal{H} = \mathcal{A}^*[\nabla^2 K(X)]\mathcal{A};$$

2. Subsequent Choleski factorization of the matrix $\mathcal{H}$ (which, due to its origin, is symmetric positive definite and thus admits Choleski decomposition $\mathcal{H} = DD^T$ with lower triangular D).

Looking at (Cone), (CP), and (6.3.2), we immediately conclude that the arithmetic cost of assembling and factorizing $\mathcal{H}$ is polynomial in the dimension of the data vector defining (CP) and that the parameter $\theta(K)$ is also polynomial in this size. Thus, the cost of an accuracy digit for the methods in question is polynomial in the size of the data, as required from polynomial time methods.[68] To get an impression of the power of interior point methods, let us compare their complexity characteristics with those of the ellipsoid method. To simplify things, we restrict the comparison to the case when all the basic cones participating in the decomposition (Cone) are identical. Besides this, we assume that we multiply and factorize matrices by the standard linear algebra techniques, so that it takes $O(p^3)$ arithmetic operations to multiply two $p \times p$ matrices, to factorize a symmetric positive semidefinite $p \times p$ matrix, to check whether a given $p \times p$ symmetric matrix is positive (semi)definite, etc.[69]

$\mathbf{K}$ is a direct product of m semidefinite cones $\mathbf{S}^k_+$. In this case, $\theta(K) = mk$. From (6.3.2) it is easily seen that the cost of assembling $\mathcal{H}$ is

$$\mathcal{C}_{\mathrm{ass}} = O(1)mnk^2(n+k). \tag{6.5.20}$$

Indeed, writing down $\mathcal{A}x$ as $\sum_{i=1}^n x_i A_i$, we see that the elements of $\mathcal{H}$ are

$$\mathcal{H}_{ij} = \mathrm{Tr}(A_i X^{-1} A_j X^{-1}). \tag{6.5.21}$$

Taking into account the block-diagonal structure of the matrices A_i and X (m diagonal blocks, $k \times k$ each), we see that it takes $O(mk^3)$ arithmetic operations to compute X^{-1}.

[68] Strictly speaking, the outlined complexity considerations are applicable to the highway phase of the solution process, once we have reached the neighborhood $\mathcal{N}_{0.1}$ of the central path. However, the results of our considerations remain unchanged when the initialization expenses are also taken into account; see section 6.6.

[69] This indeed is an assumption. There exist, at least in theory, fast linear algebra routines; e.g., two $p \times p$ matrices can be multiplied in less than $O(p^{2.4})$ operations. These fast routines, however, for the time being are of no practical use because of large absolute constant factors hidden in $O(\cdot)$'s

After we have X^{-1}, it takes $O(mk^3)$ arithmetic operations to compute the matrix $X^{-1}A_jX^{-1}$ for every particular j. After we have $X^{-1}A_jX^{-1}$, it takes $O(mk^2)$ arithmetic operations to compute every particular entry in the jth column of $\mathcal{H}$. Thus, the arithmetic cost of a column in $\mathcal{H}$ is $O(nmk^2 + mk^3)$, and the arithmetic cost of $\mathcal{H}$ itself is $O(n^2mk^2 + nmk^3)$.

Finally, the arithmetic cost of factorizing $\mathcal{H}$ after it is computed is

$$\mathcal{C}_{\text{fctr}} = O(n^3). \tag{6.5.22}$$

We see that with the interior point methods, the arithmetic cost of an accuracy digit is

$$\mathcal{C}_{\text{IP}} = \sqrt{\theta(K)}(\mathcal{C}_{\text{ass}} + \mathcal{C}_{\text{fctr}}) = O(1)\sqrt{mk}n(n^2 + mnk^2 + mk^3). \tag{6.5.23}$$

Warning! *You should never blindly trust in* (6.5.23)*!*

The above calculations say that the cost of factorizing $\mathcal{H}$ is dominated by the cost of assembling the matrix (indeed, due to $\text{Null}(\mathcal{A}) = \{0\}$ we have $n \le \dim E = m\frac{k(k+1)}{2}$), which is not true in reality. The data of an SDP arising in applications normally has a lot of structure (as it is the case with all SDPs we have considered so far). Exploiting this structure, we can reduce the cost of assembling $\mathcal{H}$ by orders of magnitude (and sometimes manage to reduce the cost of factorizing $\mathcal{H}$ as well). Consider, e.g., the semidefinite program yielding the Lovasz capacity number of a graph:

$$\min_{\lambda, Y} \{\lambda : \lambda I - Y - B \succeq 0, \quad Y \in \mathcal{G}\},$$

where $\mathcal{G}$ is the linear space of all symmetric matrices with a given pattern of zeros (for a matrix from $\mathcal{G}$, nonzero entries occupy the cells ij indexed by adjacent pairs of nodes i, j of the graph). The size k of our LMI constraint is the number of nodes in the graph, and the design dimension n of the problem is the number of free entries in Y (i.e., the number of pairs of adjacent nodes) plus one (for λ). When assembling $\mathcal{H}$, the major effort is to build the $(n-1) \times (n-1)$ principal submatrix $\mathcal{H}'$ of $\mathcal{H}$ associated with the design variables distinct from λ. (Assembling the single column associated with the variable λ is not a big deal.) It is convenient to denote the design variables distinct from λ by y_{pq} (y_{pq} is the entry of Y in the cell (p, q)), where the index $i = (p, q)$ runs through the set

$$J = \{(p, q) : 1 \le p < q \le k, \text{ the nodes } p, q \text{ are adjacent}\}.$$

The column A_j, $j = (p, q)$, of the mapping $\mathcal{A}$ associated with our problem is the $k \times k$ matrix with just two nonzero entries (both equal to 1) occupying the cells (p, q) and (q, p); thus, $A_j = e_pe_q^T + e_qe_p^T$, where e_p are the standard basic orths of $\mathbf{R}^k$. According to (6.5.21), the entries of the jth column of $\mathcal{H}'$ are

$$\mathcal{H}'_{ij} = \text{Tr}(A_iZA_jZ), \quad Z = X^{-1},$$

i.e., $\mathcal{H}_{ij}$ are just twice the entries, in the cells from J, of the $k \times k$ matrix

$$ZA_jZ = Z(e_pe_q^T + e_qe_p^T)Z = z_pz_q^T + z_qz_p^T,$$

where $z_1, \ldots, z_k$ are the columns of Z. We see that after Z is computed (when assembling $\mathcal{H}$, this computation should be done just once!), it takes $O(1)$ arithmetic operations to build

an entry of the jth column in $\mathcal{H}'$, i.e., it takes $O(n)$ arithmetic operations to build the entire column. Consequently, the cost of computing $\mathcal{H}'$ after Z is computed is $O(n^2)$ (as if we were just writing the matrix!). When taking into account the cost of computing Z, the computational price of $\mathcal{H}'$ becomes $O(k^3+n^2)$, and this is also the cost of computing the entire matrix $\mathcal{H}$. (The price of the λ-column is dominated by the price of $\mathcal{H}'$.) Thus, in the case in question $\mathcal{C}_{\rm ass} = O(k^3+n^2)$, which is far less than stated by (6.5.20). Moreover, for typical graphs, the number n of nonadjacent pairs of vertices is of order of the total number of pairs of vertices, i.e., $n = O(k^2)$. In this case, $\mathcal{C}_{\rm ass} = O(k^4)$ (while (6.5.20) yields $O(k^6)$—quite a difference!), which is negligible compared to the cost $O(n^3) = O(k^6)$ of factorizing $\mathcal{H}$.

For the ellipsoid method the cost of an accuracy digit comes from $O(n^2)$ calls to the first order and the separation oracles, and $O(n^4)$ arithmetic operations to process their answers. A call to the first order oracle is costless—$O(n)$ arithmetic operations (our objective is linear!), while a call to the separation oracle built in Lecture 5 costs $O(mk^2(n+k))$ arithmetic operations.

Indeed, to mimic the separation oracle at a point $x \in \mathbf{R}^n$, we should compute $X = \mathcal{A}x - B$ ($O(mnk^2)$ arithmetic operations) and then check whether X is positive semidefinite ($O(mk^3)$ arithmetic operations). If the result is positive, x is feasible for (CP); otherwise our check yields a vector ξ such that $\xi^T X \xi < 0$, so that the vector $\eta = \mathcal{A}^*[\xi\xi^T]$ separates x from the feasible set of (CP). Given ξ, it takes no more than $O(mnk^2)$ arithmetic operations to compute η.

Thus, the arithmetic cost per accuracy digit for the ellipsoid method is

$$\mathcal{C}_{\rm Ell} = O(1)n^2(n^2+mnk^2+mk^3).$$

We see that

$$\frac{\mathcal{C}_{\rm Ell}}{\mathcal{C}_{\rm IP}} = O(1)\frac{n}{\sqrt{mk}}; \tag{6.5.24}$$

whether this ratio is large or small depends on the relations between the sizes of the problem. In general, all we can say is that the ratio is within the bounds $[O(1)\frac{1}{\sqrt{mk}}, O(1)k\sqrt{mk}]$. The lower bound is trivial, since $n \geq 1$. The upper bound comes from the fact that $n \leq \dim E = m\frac{k(k+1)}{2}$ (recall our basic assumption $\mathrm{Null}(\mathcal{A}) = \{0\}$).

To get a better insight, let us consider two extreme cases of the structure of $\mathbf{K}$—one with $k = 1$ and the other with $m = 1$.

The case of $k = 1$. This case is especially interesting—this is LP! Indeed, the direct product of m one-dimensional semidefinite cones is exactly the nonnegative orthant $\mathbf{R}^m_+$. Thus, what we are speaking about now are LP programs

$$\min_x \left\{c^T x : Ax \geq b\right\}$$

with n variables and m inequality constraints. In this case,

$$\mathcal{C}_{\rm IP} = O(1)(m^{3/2}n^2), \quad \mathcal{C}_{\rm Ell} = O(1)n^3m$$

(note that $n \leq m = \dim E$), and the ratio (6.5.24) becomes $\frac{O(n)}{\sqrt{m}}$. Thus, the ellipsoid method is outperformed by the interior point ones whenever $m \leq n^2$ and beats the interior point

methods only when $m >> n^2$ ("few variables, plenty of constraints"). In practical LPs, m and n are of the same order, and the ellipsoid method is outperformed by the interior point methods by a factor of order of $n^{1/2}$.

From the viewpoint of LP practice, the ellipsoid method *always* is outperformed by the interior point methods. The practical performance of the ellipsoid method is fairly close to the theoretical bound, which in the LP case is $O(n^3m)$ arithmetic operations per accuracy digit, and the best of the existing IP methods for LP work in practice much better than it is said by the theoretical complexity analysis. First of all, practice demonstrates that the iteration count per accuracy digit is not proportional to $\sqrt{\theta(K)} = \sqrt{m}$; it is a quite moderate constant, so that to get a reasonably high accuracy, it takes just a fixed number of steps, typically 40–50. Second, recall the Warning above: practical large-scale LPs almost always are structured, and a part of this structure is inherited by $\mathcal{H}$. As a result, with smart implementation the cost of assembling and factorizing $\mathcal{H}$ for practical LPs is significantly less than the one given by our theoretical analysis. Thus, our analysis treats properly the numerator in the fraction $\frac{\mathcal{C}_{\text{Ell}}}{\mathcal{C}_{\text{IP}}}$ and overestimates dramatically the denominator.

A practical comparison of the ellipsoid method and the interior point methods as applied to LP is quite simple: you just *cannot use* the ellipsoid method for LP, unless when solving a toy problem with something like 20–40 variables. With interior point methods, you can solve in a few seconds LPs with thousands of variables, in a few hours LPs with many tens of thousands of variables, and in still realistic time LPs with few millions of variables (if, of course, the LPs are well structured, which normally is the case).

The case of $m = 1$. This case of pure SDP with a single LMI constraint is not too typical: among the numerous applications of SDP discussed in Lecture 4, we met with such a situation only in connection with relaxations of combinatorial problems and in the dynamic stability analysis. However, this is a meaningful case. The ratio (6.5.24) now becomes

$$\frac{\mathcal{C}_{\text{Ell}}}{\mathcal{C}_{\text{IP}}} = O(1)\frac{n}{\sqrt{k}}, \tag{6.5.25}$$

and the ellipsoid method is beaten by the interior point methods whenever $k << n^2$. We do not know of a single application of SDP when the latter relation is not satisfied. For example, in the SDP relaxations of combinatorial problems we have $n = O(k^2)$ (computing Lovasz θ-function for a typical graph, where the numbers of adjacent and nonadjacent pairs of vertices are of the same order) or, as an extreme, $n = k$ (SDP relaxation of MAXCUT in its dual form). Thus, in all applications known to us, the ellipsoid method is by far inferior compared with the interior point ones.

6.5.3 Primal and dual path-following methods

The simplest way to implement the path-following scheme from section 6.5.1 is to linearize the augmented complementary slackness equations (6.5.15) as they are, ignoring the option to rewrite these equations equivalently before linearization. Let us look at the resulting method in more details. Linearizing (6.5.15) at a current iterate $\bar{X}$, $\bar{S}$, we get the vector equation

$$t_+(\bar{S} + \Delta S) + \nabla K(\bar{X}) + [\nabla^2 K(\bar{X})]\Delta X = 0,$$

where t_+ is the target value of the penalty parameter. The system (6.5.17) now becomes

$$\begin{array}{ll}
\text{(a)} & \Delta X \in \mathcal{L} \\
 & \Updownarrow \\
\text{(a}')\ & \Delta X = \mathcal{A}\Delta x \quad [\Delta x \in \mathbf{R}^n], \\
\text{(b)} & \Delta S \in \mathcal{L}^\perp \\
 & \Updownarrow \\
\text{(b}')\ & \mathcal{A}^*\Delta S = 0, \\
\text{(c)} & t_+[\bar{S} + \Delta S] + \nabla K(\bar{X}) + [\nabla^2 K(\bar{X})]\Delta X = 0;
\end{array} \tag{6.5.26}$$

the unknowns here are ΔX, ΔS, and Δx. To process the system, we eliminate ΔX via (a′) and multiply both sides of (c) by $\mathcal{A}^*$, thus getting the equation

$$\underbrace{\mathcal{A}^*[\nabla^2 K(\bar{X})]\mathcal{A}}_{\mathcal{H}}\Delta x + [t_+\mathcal{A}^*[\bar{S} + \Delta S] + \mathcal{A}^*\nabla K(\bar{X})] = 0. \tag{6.5.27}$$

Note that $\mathcal{A}^*[\bar{S} + \Delta S] = c$ is the objective of (CP) (indeed, $\bar{S} \in \mathcal{L}^\perp + C$, i.e., $\mathcal{A}^*\bar{S} = c$, while $\mathcal{A}^*\Delta S = 0$ by (b′)). Consequently, (6.5.27) becomes the primal Newton system

$$\mathcal{H}\Delta x = -[t_+c + \mathcal{A}^*\nabla K(\bar{X})]. \tag{6.5.28}$$

Solving this system (which is possible since, as it is easily seen, the $n \times n$ matrix $\mathcal{H}$ is positive definite), we get Δx and then set

$$\begin{array}{rcl}
\Delta X & = & \mathcal{A}\Delta x, \\
\Delta S & = & -t_+^{-1}[\nabla K(\bar{X}) + [\nabla^2 K(\bar{X})\Delta X] - \bar{S},
\end{array} \tag{6.5.29}$$

thus getting a solution to (6.5.26). Restricting ourselves with the stepsizes $\alpha = \beta = 1$ (see (6.5.19)), we come to the closed-form description of the method:

$$\begin{array}{ll}
\text{(a)} & t \mapsto t_+ > t, \\
\text{(b)} & x \mapsto x_+ = x + \underbrace{\left(-[\mathcal{A}^*(\nabla^2 K(X))\mathcal{A}]^{-1}[t_+c + \mathcal{A}^*\nabla K(X)]\right)}_{\Delta x}, \\
\text{(c)} & S \mapsto S_+ = -t_+^{-1}[\nabla K(X) + [\nabla^2 K(X)]\mathcal{A}\Delta x],
\end{array} \tag{6.5.30}$$

where x is the current iterate in the space $\mathbf{R}^n$ of design variables and $X = \mathcal{A}x - B$ is its image in the space E.

The resulting scheme admits a natural explanation. Consider the function

$$F(x) = K(\mathcal{A}x - B);$$

you can immediately verify that this function is a barrier for the feasible set of (CP). Let also

$$F_t(x) = tc^T x + F(x)$$

be the associated barrier-generated family of penalized objectives. Relation (6.5.30)(b) says that the iterates in the space of design variables are updated according to

$$x \mapsto x_+ = x - [\nabla^2 F_{t_+}(x)]^{-1}\nabla F_{t_+}(x),$$

i.e., the process in the space of design variables is exactly the process (6.2.1) from section 6.2.4.

Note that (6.5.30) is, essentially, a purely primal process (this is where the name of the method comes from). Indeed, the dual iterates S, S_+ just do not appear in formulas for x_+, X_+, and in fact the dual solutions are no more than shadows of the primal ones.

Remark 6.5.1. When constructing the primal path-following method, we started with the augmented slackness equations in form (6.5.15). Needless to say, we could start our developments with the same conditions written in the swapped form

$$X + t^{-1}\nabla K(S) = 0,$$

thus coming to what is called dual path-following method. Of course, as applied to a given pair (P), (D), the dual path-following method differs from the primal one. However, the constructions and results related to the dual path-following method require no special care—they can be obtained from their primal counterparts just by swapping primal and dual entities.

The complexity analysis of the primal path-following method can be summarized in the following theorem.

Theorem 6.5.1. *Let $0 < \chi \le \kappa \le 0.1$. Assume that we are given a starting point (t_0, x_0, S_0) such that $t_0 > 0$ and the point*

$$(X_0 = \mathcal{A}x_0 - B, S_0)$$

is κ-close to $Z_*(t_0)$:

$$\mathrm{dist}((X_0, S_0), Z_*(t_0)) \le \kappa.$$

Starting with (t_0, x_0, X_0, S_0), let us iterate process (6.5.30) *equipped with the penalty updating policy*

$$t_+ = \left(1 + \frac{\chi}{\sqrt{\theta(K)}}\right) t, \tag{6.5.31}$$

i.e., let us build the iterates (t_i, x_i, X_i, S_i) according to

$$\begin{aligned} t_i &= \left(1 + \tfrac{\chi}{\sqrt{\theta(K)}}\right) t_{i-1}, \\ x_i &= x_{i-1} + \underbrace{\left(-[\mathcal{A}^*(\nabla^2 K(X_{i-1}))\mathcal{A}]^{-1}[t_i c + \mathcal{A}^*\nabla K(X_{i-1})]\right)}_{\Delta x_i}, \\ X_i &= \mathcal{A}x_i - B, \\ S_i &= -t_i^{-1}[\nabla K(X_{i-1}) + [\nabla^2 K(X_{i-1})]\mathcal{A}\Delta x_i]. \end{aligned}$$

The resulting process is well defined and generates strictly primal-dual feasible pairs (X_i, S_i) such that (t_i, X_i, S_i) stay in the neighborhood $\mathcal{N}_\kappa$ of the primal-dual central path.

The theorem says that, with properly chosen κ, χ (e.g., $\kappa = \chi = 0.1$), after we have somehow reached the $\mathcal{N}_\kappa$-neighborhood of the primal-dual central path, we can trace it by the primal path-following method, keeping the iterates in $\mathcal{N}_\kappa$ and increasing the penalty

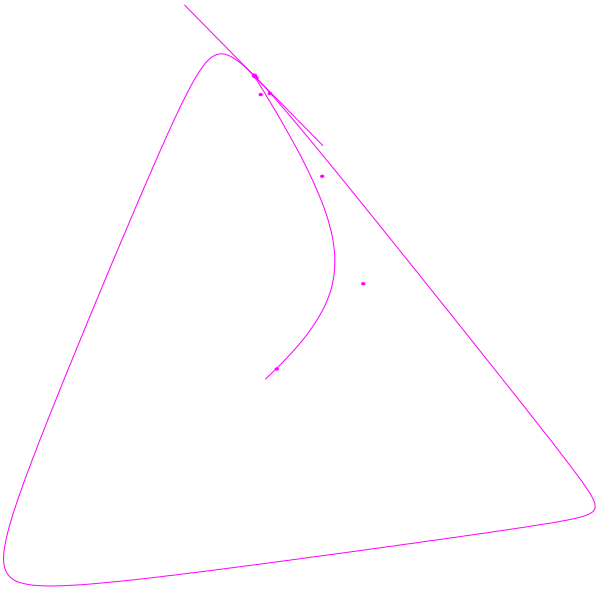


Figure 6.1. *Primal path-following method. What you see is the 2D feasible set of a toy SDP ($\mathbf{K} = \mathbf{S}^3_+$). The continuous curve is the primal central path; dots are iterates x_i of the algorithm. We cannot draw the dual solutions, since they live in four-dimensional space (*$\dim \mathcal{L}^\perp = \dim \mathbf{S}^3 - \dim \mathcal{L} = 6 - 2 = 4$*).*

parameter by an absolute constant factor in every $O(\sqrt{\theta(K)})$ steps—exactly as claimed in sections 6.2.4 and 6.5.2. This fact is extremely important theoretically; in particular, it underlies the polynomial time complexity bounds for LP, CQP, and SDP from section 6.6. As a practical tool, the primal and the dual path-following methods, at least in their short-step form presented above, are not that attractive. The computational power of the methods can be improved by passing to appropriate large-step versions of the algorithms, but even these versions are thought to be inferior to true primal-dual path-following methods (those that truly work simultaneously with (P) and (D); see below).

To get an impression of how the primal path-following method works, look at Fig. 6.1. Here are the corresponding numbers:

Itr#	Objective	Duality gap	Itr#	Objective	Duality gap
1	−0.100000	2.96	7	−1.359870	8.4e−4
2	−0.906963	0.51	8	−1.360259	2.1e−4
3	−1.212689	0.19	9	−1.360374	5.3e−5
4	−1.301082	6.9e-2	10	−1.360397	1.4e−5
5	−1.349584	2.1e-2	11	−1.360404	3.8e−6
6	−1.356463	4.7e-3	12	−1.360406	9.5e−7

6.5.4 Semidefinite programming case

In what follows, we specialize the primal-dual path-following scheme in the SDP case and carry out its complexity analysis.

Path-following scheme in SDP

In the SDP case, the system of nonlinear equations (6.5.15) becomes (see (6.3.2))

$$G_t(X, S) \equiv S - t^{-1}X^{-1} = 0, \tag{6.5.32}$$

where X, S are positive definite $k \times k$ symmetric matrices.

Recall that our generic scheme of a path-following interior point method suggests, given a current triple $(\bar{t}, \bar{X}, \bar{S})$ with positive $\bar{t}$ and strictly primal, respectively, dual feasible $\bar{X}$ and $\bar{S}$, to update the triple into a new triple (t_+, X_+, S_+) of the same type as follows:

(i) We somehow rewrite the system (6.5.32) as an equivalent system

$$\bar{G}_t(X, S) = 0. \tag{6.5.33}$$

(ii) We choose somehow a new value $t_+ > \bar{t}$ of the penalty parameter and linearize the system (6.5.33) (with t set to t_+) at the point $(\bar{X}, \bar{S})$, thus coming to the system of linear equations

$$\frac{\partial \bar{G}_{t_+}(\bar{X}, \bar{S})}{\partial X}\Delta X + \frac{\partial \bar{G}_{t_+}(\bar{X}, \bar{S})}{\partial S}\Delta S = -\bar{G}_{t_+}(\bar{X}, \bar{S}) \tag{6.5.34}$$

for the search direction $(\Delta X, \Delta S)$.

We add to (6.5.34) the system of linear equations on ΔX, ΔS expressing the requirement that a shift of $(\bar{X}, \bar{S})$ in the direction $(\Delta X, \Delta S)$ preserve the validity of the linear constraints in (P), (D), i.e., the equations saying that $\Delta X \in \mathcal{L}$, $\Delta S \in \mathcal{L}^\perp$:

$$\begin{array}{rcll} \Delta X & = & \mathcal{A}\Delta x & [\Leftrightarrow \Delta X \in \mathcal{L}], \\ \mathcal{A}^*\Delta S & = & 0 & [\Leftrightarrow \Delta S \in \mathcal{L}^\perp]. \end{array} \tag{6.5.35}$$

(iii) We solve the system of linear equations (6.5.34), (6.5.35), thus obtaining a primal-dual search direction $(\Delta X, \Delta S)$, and update the current iterates:

$$X_+ = \bar{X} + \alpha\Delta x, \quad S_+ = \bar{S} + \beta\Delta S,$$

where the primal and the dual stepsizes α, β are given by certain side requirements.

The major degree of freedom of the scheme comes from (i), i.e., from how we construct the system (6.5.33). A very popular way to handle (i) which indeed leads to primal-dual methods, starts from rewriting (6.5.32) in a form which is symmetric with respect to X and S. To this end we first observe that (6.5.32) is equivalent to each one of the following two matrix equations:

$$XS = t^{-1}I; \quad SX = t^{-1}I.$$

Adding these equations, we get a symmetric matrix equation

$$XS + SX = 2t^{-1}I, \tag{6.5.36}$$

which, by its origin, is a consequence of (6.5.32). A closer inspection reveals that (6.5.36), regarded as a matrix equation with positive definite symmetric matrices, is equivalent to (6.5.32). It is possible to use in the role of (6.5.33) the matrix equation (6.5.36) as it is; this

policy leads to the so-called Alizadeh–Overton–Haeberly (AHO) search direction and the $XS + SX$ primal-dual path-following method.

It is also possible to use a scaled version of (6.5.36). Namely, let us choose somehow a positive definite scaling matrix Q and observe that our original matrix equation (6.5.32) says that $S = t^{-1}X^{-1}$, which is the same as $Q^{-1}SQ^{-1} = t^{-1}(QXQ)^{-1}$. The latter, in turn, is equivalent to each one of the matrix equations

$$QXSQ^{-1} = t^{-1}I, \quad Q^{-1}SXQ = t^{-1}I.$$

Adding these equations, we get the scaled version of (6.5.36),

$$QXSQ^{-1} + Q^{-1}SXQ = 2t^{-1}I, \tag{6.5.37}$$

which, as before, is equivalent to (6.5.32).

With (6.5.37) playing the role of (6.5.33), we get a quite flexible scheme with complete freedom for choosing the scaling matrix Q, which in particular can be varied from iteration to iteration. As we shall see in a while, this freedom reflects the intrinsic (and extremely important in the interior-point context) symmetries of the semidefinite cone.

Analysis of the path-following methods based on search directions coming from (6.5.37) (Zhang's family of search directions) is simplified considerably when the policy for choosing the scaling matrix at an iteration ensures that the matrices

$$\widetilde{S} = Q^{-1}\bar{S}Q^{-1}, \ \widehat{X} = Q\bar{X}Q$$

commute ($\bar{X}$, $\bar{S}$ are the iterates to be updated). Such a policy is called a commutative scaling. Popular commutative scalings are

1. $Q = \bar{S}^{1/2}$ ($\widetilde{S} = I$, $\widehat{X} = \bar{S}^{1/2}\bar{X}\bar{S}^{1/2}$) (the XS method);
2. $Q = \bar{X}^{-1/2}$ ($\widetilde{S} = \bar{X}^{1/2}\bar{S}\bar{X}^{1/2}$, $\widehat{X} = I$) (the SX method);
3. Q is such that $\widetilde{S} = \widehat{X}$ (the Nesterov–Todd method, extremely attractive and deep)

If $\bar{X}$ and $\bar{S}$ were just positive reals, the formula for Q would be simply $Q = (\frac{\bar{S}}{\bar{X}})^{1/4}$. In the matrix case this simple formula becomes a bit more complicated (to simplify notation, below we write X instead of $\bar{X}$ and S instead of $\bar{S}$):[70]

$$Q = P^{1/2}, \text{ where } P = X^{-1/2}(X^{1/2}SX^{1/2})^{-1/2}X^{1/2}S.$$

We should verify that (a) P is symmetric positive definite, so that Q is well defined, and that (b) $Q^{-1}SQ^{-1} = QXQ$.

[70] You should not think that Nesterov and Todd guessed the formula for this scaling matrix. They did much more: they developed a deep theory (covering the general LP-CQP-SDP case, not just the SDP one!) that, among other things, guarantees that the desired scaling matrix exists (and even is unique). After the existence is established, it becomes easier to find an explicit formula for Q.

(a) Let us first verify that P is symmetric:

$$\begin{array}{c} P \;?\!=\!?\; P^T \\ \Updownarrow \\ X^{-1/2}(X^{1/2}SX^{1/2})^{-1/2}X^{1/2}S \;?\!=\!?\; SX^{1/2}(X^{1/2}SX^{1/2})^{-1/2}X^{-1/2} \\ \Updownarrow \\ \left(X^{-1/2}(X^{1/2}SX^{1/2})^{-1/2}X^{1/2}S\right)\left(X^{1/2}(X^{1/2}SX^{1/2})^{1/2}X^{-1/2}S^{-1}\right) \;?\!=\!?\; I \\ \Updownarrow \\ X^{-1/2}(X^{1/2}SX^{1/2})^{-1/2}(X^{1/2}SX^{1/2})(X^{1/2}SX^{1/2})^{1/2}X^{-1/2}S^{-1} \;?\!=\!?\; I \\ \Updownarrow \\ X^{-1/2}(X^{1/2}SX^{1/2})X^{-1/2}S^{-1} \;?\!=\!?\; I \end{array}$$

and the last $?\!=\!?$ is indeed $=$.

To verify that P is positive definite, recall that the spectrum of the product of two square matrices (symmetric or not) remains unchanged when swapping the factors. Therefore, denoting $\sigma(A)$ the spectrum of A, we have

$$\begin{array}{rcl} \sigma(P) & = & \sigma\left(X^{-1/2}(X^{1/2}SX^{1/2})^{-1/2}X^{1/2}S\right) \\ & = & \sigma\left((X^{1/2}SX^{1/2})^{-1/2}X^{1/2}SX^{-1/2}\right) \\ & = & \sigma\left((X^{1/2}SX^{1/2})^{-1/2}(X^{1/2}SX^{1/2})X^{-1}\right) \\ & = & \sigma\left((X^{1/2}SX^{1/2})^{1/2}X^{-1}\right) \\ & = & \sigma\left(X^{-1/2}(X^{1/2}SX^{1/2})^{1/2}X^{-1/2}\right), \end{array}$$

and the argument of the last $\sigma(\cdot)$ is clearly a positive definite symmetric matrix. Thus, the spectrum of symmetric matrix P is positive, i.e., P is positive definite.

(b) To verify that $QXQ = Q^{-1}SQ^{-1}$, i.e., that $P^{1/2}XP^{1/2} = P^{-1/2}SP^{-1/2}$, is the same as to verify that $PXP = S$. The latter equality is given by the following computation:

$$\begin{array}{rcl} PXP & = & \left(X^{-1/2}(X^{1/2}SX^{1/2})^{-1/2}X^{1/2}S\right)X\left(X^{-1/2}(X^{1/2}SX^{1/2})^{-1/2}X^{1/2}S\right) \\ & = & X^{-1/2}(X^{1/2}SX^{1/2})^{-1/2}(X^{1/2}SX^{1/2})(X^{1/2}SX^{1/2})^{-1/2}X^{1/2}S \\ & = & X^{-1/2}X^{1/2}S \\ & = & S. \end{array}$$

Complexity analysis

We are about to carry out the complexity analysis of the primal-dual path-following methods based on commutative Zhang's scalings. This analysis is more technical than whatever else is in the book, and an uninterested reader may skip it without much harm.

Scalings. We already have mentioned what a scaling of $\mathbf{S}^k_+$ is: this is the linear one-to-one transformation of $\mathbf{S}^k$ given by the formula

$$H \mapsto QHQ^T, \tag{Scl}$$

where Q is a nonsingular scaling matrix. It is immediately seen that (Scl) is a symmetry of the semidefinite cone $\mathbf{S}^k_+$—it maps the cone onto itself. This family of symmetries is quite

rich: for every pair of points A, B from the interior of the cone, there exists a scaling that maps A onto B, e.g., the scaling

$$H \mapsto (\underbrace{B^{1/2}A^{-1/2}}_{Q})H(\underbrace{A^{-1/2}B^{1/2}}_{Q^T}).$$

In fact, this is the existence of a rich family of symmetries of the underlying cones which makes SDP (along with LP and CQP, where the cones also are perfectly symmetric) especially well suited for interior point methods.

In what follows we will be interested in scalings associated with positive definite scaling matrices. The scaling given by such a matrix Q (X,S,...) will be denoted by $\mathcal{Q}$ (respectively, $\mathcal{X}$,$\mathcal{S}$,...):

$$\mathcal{Q}[H] = QHQ.$$

Given a problem of interest (CP) (where $\mathbf{K} = \mathbf{S}^k_+$) and a scaling matrix $Q \succ 0$, we can scale the problem, i.e., pass to the equivalent problem

$$\min_x \left\{c^T x : \mathcal{Q}[\mathcal{A}x - B] \succeq 0\right\} \qquad \mathcal{Q}(\mathrm{CP})$$

(recall that $\mathcal{Q}[H]$ is positive semidefinite if and only if H is so). In terms of the geometric reformulation (P) of (CP), this transformation is nothing but the substitution of variables

$$QXQ = Y \Leftrightarrow X = Q^{-1}YQ^{-1}.$$

In the Y-variables, (P) is the problem

$$\min_Y \left\{\mathrm{Tr}(C[Q^{-1}YQ^{-1}]) : Y \in \mathcal{Q}(\mathcal{L}) - \mathcal{Q}[B],\ \ Y \succeq 0\right\},$$

i.e., the problem

$$\begin{array}{c}\min\limits_Y \left\{\mathrm{Tr}(\widehat{C}Y) : Y \in \widehat{\mathcal{L}} - \widehat{B},\ \ Y \succeq 0\right\}\\ \left[\widehat{C} = Q^{-1}CQ^{-1}, \widehat{B} = QBQ, \widehat{\mathcal{L}} = \mathrm{Im}(\mathcal{Q}\mathcal{A}) = \mathcal{Q}(\mathcal{L})\right].\end{array} \qquad (\widehat{\mathrm{P}})$$

The problem dual to $(\widehat{\mathrm{P}})$ is

$$\max_Z \left\{\mathrm{Tr}(\widehat{B}Z) : Z \in \widehat{\mathcal{L}}^\perp + \widehat{C},\ \ Z \succeq 0\right\}. \qquad (\widetilde{\mathrm{D}})$$

To realize what is $\widehat{\mathcal{L}}^\perp$, note that

$$\langle Z, QXQ\rangle_F = \mathrm{Tr}(ZQXQ) = \mathrm{Tr}(QZQX) = \langle QZQ, X\rangle_F;$$

thus, Z is orthogonal to every matrix from $\widehat{\mathcal{L}}$, i.e., to every matrix of the form QXQ with $X \in \mathcal{L}$ if and only if the matrix QZQ is orthogonal to every matrix from $\mathcal{L}$, i.e., if and only if $QZQ \in \mathcal{L}^\perp$. It follows that

$$\widehat{\mathcal{L}}^\perp = \mathcal{Q}^{-1}(\mathcal{L}^\perp).$$

Thus, when acting on the primal-dual pair (P), (D) of SDPs, a scaling, given by a matrix $Q \succ 0$, converts it into another primal-dual pair of problems, and this new pair is as follows:

• The primal geometric data—the subspace $\mathcal{L}$ and the primal shift B (which has a part-time job to be the dual objective as well)—are replaced by their images under the mapping $\mathcal{Q}$.

• The dual geometric data—the subspace $\mathcal{L}^\perp$ and the dual shift C (it is the primal objective as well)—are replaced by their images under the mapping $\mathcal{Q}^{-1}$ inverse to $\mathcal{Q}$; this inverse mapping is the scaling given by the matrix Q^{-1}.

We see that it makes sense to speak about primal-dual scaling which acts on both the primal and the dual variables and maps a primal variable X onto QXQ, and a dual variable S onto $Q^{-1}SQ^{-1}$. Formally speaking, the primal-dual scaling associated with a matrix $Q \succ 0$ is the linear transformation $(X, S) \mapsto (QXQ, Q^{-1}SQ^{-1})$ of the direct product of two copies of $\mathbf{S}^k$ (the primal and the dual ones). A primal-dual scaling acts naturally on different entities associated with a primal-dual pair (P), (S), in particular, at

- the pair (P), (D) itself—it is converted into another primal-dual pair of problems $(\widehat{\mathrm{P}})$, $(\widetilde{\mathrm{D}})$.

- a primal-dual feasible pair (X, S) of solutions to (P), (D)—it is converted to the pair $(\widehat{X} = QXQ, \widetilde{S} = Q^{-1}SQ^{-1})$ of feasible solutions to $(\widehat{\mathrm{P}})$, $(\widetilde{\mathrm{D}})$. Note that the primal-dual scaling preserves strict feasibility and the duality gap:

  $$\mathrm{DualityGap}_{\mathrm{P,D}}(X, S) = \mathrm{Tr}(XS) = \mathrm{Tr}(QXSQ^{-1}) = \mathrm{Tr}(\widehat{X}\widetilde{S}) = \mathrm{DualityGap}_{\widehat{\mathrm{P}},\widetilde{\mathrm{D}}}(\widehat{X}, \widetilde{S}).$$

- the primal-dual central path $Z_*(\cdot) = (X_*(\cdot), S_*(\cdot))$ of (P), (D)—it is converted into the curve $(\widehat{X}_*(t) = QX_*(t)Q, \widetilde{S}_*(t) = Q^{-1}S_*(t)Q^{-1})$, which is nothing but the primal-dual central path $\overline{Z}_*(t)$ of the primal-dual pair $(\widehat{\mathrm{P}})$, $(\widetilde{\mathrm{D}})$.

  The latter fact can be easily derived from the characterization of the primal-dual central path. A more instructive derivation is based on the fact that our hero—the barrier $S_k(\cdot)$—is semi-invariant with respect to scaling:

  $$S_k(\mathcal{Q}(X)) = -\ln\mathrm{Det}(QXQ) = -\ln\mathrm{Det}(X) - 2\ln\mathrm{Det}(Q) = S_k(X) + \mathrm{const}(Q).$$

  Now, a point $Y(t)$ on the primal central path of the problem $(\widehat{\mathrm{P}})$ is the unique minimizer of the aggregate

  $$S_k^t(Y) = t\langle Q^{-1}CQ^{-1}, Y\rangle_F + S_k(Y) \equiv t\mathrm{Tr}(Q^{-1}CQ^{-1}Y) + S_k(Y)$$

  over the set of strictly feasible solutions of $(\widehat{\mathrm{P}})$. The latter set is exactly the image of the set of strictly feasible solutions of (P) under the transformation $\mathcal{Q}$, so that $Y(t)$ is the image, under the same transformation, of the point $X(t)$ which minimizes the aggregate

  $$S_k^t(QXQ) = t\mathrm{Tr}((Q^{-1}CQ^{-1})(QXQ)) + S_k(QXQ) = t\mathrm{Tr}(CX) + S_k(X) + \mathrm{const}(Q)$$

  over the set of strictly feasible solutions to (P). We see that $X(t)$ is exactly the point $X_*(t)$ on the primal central path associated with problem (P). Thus, the point $Y(t)$ of the primal central path associated with $(\widehat{\mathrm{P}})$ is nothing but $\widehat{X}_*(t) = QX_*(t)Q$. Similarly, the points of the central path associated with the problem $(\widetilde{\mathrm{D}})$ are exactly the points $\widetilde{S}_*(t) = Q^{-1}S_*(t)Q^{-1}$.

- the neighborhood $\mathcal{N}_\kappa$ of the primal-dual central path $Z_*(\cdot)$ associated with the pair of problems (P), (D) (see (6.4.13)). As you can guess, the image of $\mathcal{N}_\kappa$ is exactly the neighborhood $\overline{\mathcal{N}}_\kappa$, given by (6.4.13), of the primal-dual central path $\overline{Z}_*(\cdot)$ of $(\widehat{\mathrm{P}})$, $(\widetilde{\mathrm{D}})$.

  The latter fact is immediate: for a pair (X, S) of strictly feasible primal and dual solutions to (P), (D) and a $t > 0$ we have (see (6.4.11))

$$\begin{aligned}
&\mathrm{dist}^2((\widehat{X}, \widetilde{S}), \overline{Z}_*(t)) \\
&\quad = \mathrm{Tr}\left([QXQ](tQ^{-1}SQ^{-1} - [QXQ]^{-1})[QXQ](tQ^{-1}SQ^{-1} - [QXQ]^{-1})\right) \\
&\quad = \mathrm{Tr}\left(QX(tS - X^{-1})X(tS - X^{-1})Q^{-1}\right) \\
&\quad = \mathrm{Tr}\left(X(tS - X^{-1})X(tS - X^{-1})\right) \\
&\quad = \mathrm{dist}^2((X, S), Z_*(t)).
\end{aligned}$$

Primal-dual short-step path-following methods based on commutative scalings. Path-following methods we are about to consider trace the primal-dual central path of (P), (D), staying in $\mathcal{N}_\kappa$-neighborhood of the path; here $\kappa \le 0.1$ is fixed. The path is traced by iterating the following updating:

(U) *Given a current pair of strictly feasible primal and dual solutions* $(\bar{X}, \bar{S})$ *such that the triple*

$$\left(\bar{t} = \frac{k}{\mathrm{Tr}(\bar{X}\bar{S})}, \bar{X}, \bar{S}\right) \tag{6.5.38}$$

belongs to $\mathcal{N}_\kappa$*, i.e. (see* (6.4.11)),

$$\|\bar{t}\bar{X}^{1/2}\bar{S}\bar{X}^{1/2} - I\|_2 \le \kappa, \tag{6.5.39}$$

we

1. *choose the new value* t_+ *of the penalty parameter according to*

$$t_+ = \left(1 - \frac{\chi}{\sqrt{k}}\right)^{-1} \bar{t}, \tag{6.5.40}$$

   where $\chi \in (0, 1)$ *is a parameter of the method;*

2. *choose somehow the scaling matrix* $Q \succ 0$ *such that the matrices* $\widehat{X} = Q\bar{X}Q$ *and* $\widetilde{S} = Q^{-1}\bar{S}Q^{-1}$ *commute with each other;*

3. *linearize the equation*

$$QXSQ^{-1} + Q^{-1}SXQ = \frac{2}{t_+}I$$

   at the point $(\bar{X}, \bar{S})$*, thus coming to the equation*

$$\begin{aligned}
&Q[\Delta X \cdot S + X \cdot \Delta S]Q^{-1} + Q^{-1}[\Delta S \cdot X + S \cdot \Delta X]Q \\
&\quad = \tfrac{2}{t_+}I - [Q\bar{X}\bar{S}Q^{-1} + Q^{-1}\bar{S}\bar{X}Q];
\end{aligned} \tag{6.5.41}$$

4. *add to* (6.5.41) *the linear equations*

$$\begin{array}{rcl} \Delta X & \in & \mathcal{L}, \\ \Delta S & \in & \mathcal{L}^{\perp}; \end{array} \tag{6.5.42}$$

5. *solve system* (6.5.41), (6.5.42), *thus getting primal-dual search direction* $(\Delta X, \Delta S)$*; and*
6. *update current primal-dual solution* $(\bar{X}, \bar{S})$ *into a new pair* (X_+, S_+) *according to*

$$X_+ = \bar{X} + \Delta X, \quad S_+ = \bar{S} + \Delta S.$$

We already have explained the ideas underlying (U), up to the fact that in our previous explanations we dealt with three independent entities: $\bar{t}$ (current value of the penalty parameter), $\bar{X}$, $\bar{S}$ (current primal and dual solutions), while in (U) $\bar{t}$ is a function of $\bar{X}$, $\bar{S}$:

$$\bar{t} = \frac{k}{\mathrm{Tr}(\bar{X}\bar{S})}. \tag{6.5.43}$$

The reason for establishing this dependence is very simple: if (t, X, S) were on the primal-dual central path: $XS = t^{-1}I$, then, taking traces, we indeed would get $t = \frac{k}{\mathrm{Tr}(XS)}$. Thus, (6.5.43) is a reasonable way to reduce the number of independent entities we deal with.

Note also that (U) is a pure Newton scheme—here the primal and the dual stepsizes are equal to 1 (cf. (6.5.19)).

The major element of the complexity analysis of path-following polynomial time methods for SDP is as follows.

Theorem 6.5.2. *Let the parameters* κ, χ *of* (U) *satisfy the relations*

$$0 < \chi \leq \kappa \leq 0.1. \tag{6.5.44}$$

Let, further, $(\bar{X}, \bar{S})$ *be a pair of strictly feasible primal and dual solutions to* (P), (D) *such that the triple* (6.5.38) *satisfies* (6.5.39)*. Then the updated pair* (X_+, S_+) *is well defined (i.e., system* (6.5.41), (6.5.42) *is solvable with a unique solution),* X_+, S_+ *are strictly feasible solutions to* (P), (D)*, respectively,*

$$t_+ = \frac{k}{\mathrm{Tr}(X_+S_+)},$$

and the triple (t_+, X_+, S_+) *belongs to* $\mathcal{N}_\kappa$.

The theorem says that with properly chosen κ, χ (say, $\kappa = \chi = 0.1$), updating (U) converts a strictly primal-dual feasible iterate $(\bar{X}, \bar{S})$ (which is close, in the sense of (6.5.39), (6.5.38), to the primal-dual central path) into a new strictly primal-dual feasible iterate with the same closeness-to-the-path property and larger, by factor $(1 + O(1)k^{-1/2})$, value of the penalty parameter. Thus, after we once reach $\mathcal{N}_{0.1}$, we are able to trace the primal-dual central path, staying in $\mathcal{N}_{0.1}$ and increasing the penalty parameter by an absolute constant factor in $O(\sqrt{k}) = O(\sqrt{\theta(K)})$ steps, exactly as announced in section 6.5.2.

Proof of Theorem 6.5.2. 1*. Observe, first (this observation is crucial!) that it suffices to prove our theorem in the particular case when $\bar{X}$, $\bar{S}$ commute with each other and $Q = I$. Indeed, it is immediately seen that the updating (U) can be represented as follows:

1. We first scale by Q the input data of (U)—the primal-dual pair of problems (P), (D) and the strictly feasible pair $\bar{X}$, $\bar{S}$ of primal and dual solutions to these problems, as explained in the section on scaling. Note that the resulting entities—a pair of primal-dual problems and a strictly feasible pair of primal-dual solutions to these problems—are linked with each other exactly in the same fashion as the original entities, due to scaling invariance of the duality gap and the neighborhood $\mathcal{N}_\kappa$. In addition, the scaled primal and dual solutions commute.

2. We apply to the scaled input data yielded by the previous step the updating ($\widehat{\text{U}}$) completely similar to (U), but using the unit matrix in the role of Q.

3. We scale back the result of the previous step, i.e., subject this result to the scaling associated with Q^{-1}, thus obtaining the updated iterate (X_+, S_+).

Given that the second step of this procedure preserves primal-dual strict feasibility of the updated iterate with respect to the scaled primal-dual pair of problems and keeps the iterate in the κ-neighborhood $\mathcal{N}_\kappa$ of the corresponding central path, we could use once again the scaling invariance reasoning to conclude that the result (X_+, S_+) of (U) is well defined, is strictly feasible for (P), (D), and is close to the original central path, as claimed in the theorem. Thus, all we need is to justify the above "given," which is exactly the same as to prove the theorem in the particular case of $Q = I$ and commuting $\bar{X}$, $\bar{S}$. Thus, in the rest of the proof we assume that $Q = I$ and that the matrices $\bar{X}$, $\bar{S}$ commute with each other. Due to the latter property, $\bar{X}$, $\bar{S}$ are diagonal in a properly chosen orthonormal basis; representing all matrices from $\mathbf{S}^k$ in this basis, we can reduce the situation to the case when $\bar{X}$ and $\bar{S}$ are diagonal. Thus, we may (and do) assume in the sequel that $\bar{X}$ and $\bar{S}$ are diagonal, with diagonal entries x_i,s_i, $i = 1, \ldots, k$, respectively, and that $Q = I$. Finally, to simplify notation, we write t, X, S instead of $\bar{t}$, $\bar{X}$, $\bar{S}$, respectively.

2*. Our situation and goals now are as follows. We are given orthogonal to each other affine planes $\mathcal{L} - B$, $\mathcal{L}^\perp + C$ in $\mathbf{S}^k$ and two positive definite diagonal matrices $X = \text{Diag}(\{x_i\}) \in \mathcal{L} - B$, $S = \text{Diag}(\{s_i\}) \in \mathcal{L}^\perp + C$. We set

$$\mu = \frac{1}{t} = \frac{\text{Tr}(XS)}{k}$$

and know that

$$\|tX^{1/2}SX^{1/2} - I\|_2 \leq \kappa.$$

We further set

$$\mu_+ = \frac{1}{t_+} = (1 - \chi k^{-1/2})\mu \tag{6.5.45}$$

and consider the system of equations with respect to unknown symmetric matrices ΔX, ΔS:

$$\begin{array}{ll} \text{(a)} & \Delta X \in \mathcal{L}, \\ \text{(b)} & \Delta S \in \mathcal{L}^\perp, \\ \text{(c)} & \Delta X \cdot S + X\Delta \cdot S + \Delta S \cdot X + S \cdot \Delta X = 2\mu_+ I - 2XS. \end{array} \tag{6.5.46}$$

We should prove that the system has a unique solution $(\Delta X, \Delta S)$ and that the matrices

$$X_+ = X + \Delta X, \;\; S_+ = S + \Delta S$$

are

(i) positive definite;

(ii) belong to $\mathcal{L} - B$, $\mathcal{L}^\perp + C$, respectively, and satisfy the relation

$$\mathrm{Tr}(X_+ S_+) = \mu_+ k; \tag{6.5.47}$$

(iii) satisfy the relation

$$\Omega \equiv \|\mu_+^{-1} X_+^{1/2} S_+ X_+^{1/2} - I\|_2 \le \kappa. \tag{6.5.48}$$

Observe that the situation can be reduced to the one with $\mu = 1$. Indeed, let us pass from the matrices $X, S, \Delta X, \Delta S, X_+, S_+$ to $X, S' = \mu^{-1} S, \Delta X, \Delta S' = \mu^{-1}\Delta S, X_+, S'_+ = \mu^{-1} S_+$. Now the "we are given" part of our situation becomes as follows: we are given two diagonal positive definite matrices X, S' such that $X \in \mathcal{L} - B$, $S' \in \mathcal{L}^\perp + C'$, $C' = \mu^{-1} C$,

$$\mathrm{Tr}(X S') = k,$$

and

$$\|X^{1/2} S' X^{1/2} - I\|_2 = \|\mu^{-1} X^{1/2} S X^{1/2} - I\|_2 \le \kappa.$$

The "we should prove" part becomes as follows: to verify that the system of equations with symmetric unknowns $\Delta X, \Delta S'$

(a) $\Delta X \in \mathcal{L}$,

(b) $\Delta S' \in \mathcal{L}^\perp$,

(c) $\Delta X \cdot S' + X \cdot \Delta S' + \Delta S' \cdot X + S' \cdot \Delta X = 2(1 - \chi k^{-1/2}) I - 2 X S'$

has a unique solution and that the matrices $X_+ = X + \Delta X$, $S'_+ = S' + \Delta S'$ are positive definite, belong to $\mathcal{L} - B$, $\mathcal{L}^\perp + C'$, respectively, and satisfy the relations

$$\mathrm{Tr}(X_+ S'_+) = \frac{\mu_+}{\mu} = 1 - \chi k^{-1/2}$$

and

$$\|(1 - \chi k^{-1/2})^{-1} X_+^{1/2} S'_+ X_+^{1/2} - I\|_2 \le \kappa.$$

Thus, the general situation indeed can be reduced to the one with $\mu = 1$, $\mu_+ = 1 - \chi k^{-1/2}$, and we loose nothing, assuming, in addition to what was already postulated, that

$$\mu \equiv t^{-1} \equiv \frac{\mathrm{Tr}(XS)}{k} = 1, \quad \mu_+ = 1 - \chi k^{-1/2},$$

whence

$$[\mathrm{Tr}(XS) =] \quad \sum_{i=1}^{k} x_i s_i = k \tag{6.5.49}$$

and

$$[\|tX^{1/2}SX^{1/2} - I\|_2^2 \equiv] \quad \sum_{i=1}^{n}(x_i s_i - 1)^2 \le \kappa^2. \tag{6.5.50}$$

3*. We start with proving that (6.5.46), regarded as a system with symmetric unknowns ΔX, ΔS, has a unique solution. It is convenient to pass in (6.5.46) from the unknowns ΔX, ΔS to the unknowns

$$\begin{array}{rcl} \delta X = X^{-1/2}\cdot\Delta X\cdot X^{-1/2} & \Leftrightarrow & \Delta X = X^{1/2}\cdot\delta X\cdot X^{1/2}, \\ \delta S = X^{1/2}\cdot\Delta S\cdot X^{1/2} & \Leftrightarrow & \Delta S = X^{-1/2}\cdot\delta S\cdot X^{-1/2}. \end{array} \tag{6.5.51}$$

With respect to the new unknowns, (6.5.46) becomes

$$\begin{array}{ll} \text{(a)} & X^{1/2}\cdot\delta X\cdot X^{1/2}\in\mathcal{L}, \\ \text{(b)} & X^{-1/2}\cdot\delta S\cdot X^{-1/2}\in\mathcal{L}^\perp, \\ \text{(c)} & X^{1/2}\cdot\delta X\cdot X^{1/2}S + X^{1/2}\cdot\delta S\cdot X^{-1/2} \\ & \quad +X^{-1/2}\cdot\delta S\cdot X^{1/2} + SX^{1/2}\cdot\delta X\cdot X^{1/2} = 2\mu_+ I - 2XS \\ & \qquad\qquad \Updownarrow \\ \text{(c}')\quad & L(\delta X,\delta S)\equiv\Bigg[\underbrace{\sqrt{x_i x_j}(s_i+s_j)}_{\phi_{ij}}(\delta X)_{ij} \\ & \quad +\underbrace{\left(\sqrt{\frac{x_i}{x_j}}+\sqrt{\frac{x_j}{x_i}}\right)}_{\psi_{ij}}(\delta S)_{ij}\Bigg]_{i,j=1}^{k} = 2\left[(\mu_+ - x_i s_i)\delta_{ij}\right]_{i,j=1}^{k}, \end{array} \tag{6.5.52}$$

where

$$\delta_{ij} = \begin{cases} 0, & i\ne j, \\ 1, & i = j, \end{cases}$$

are the Kronecker symbols.

We first claim that (6.5.52), regarded as a system with unknown symmetric matrices δX, δS, has a unique solution. Observe that (6.5.52) is a system with $2\dim\mathbf{S}^k \equiv 2N$ scalar unknowns and $2N$ distinct scalar linear equations. Indeed, (6.5.52)(a) is a system of $N' \equiv N - \dim\mathcal{L}$ linear equations, (6.5.52)(b) is a system of $N'' = N - \dim\mathcal{L}^\perp = \dim\mathcal{L}$ linear equations, and (6.5.52)(c) has N distinct equations, so that the total number of distinct linear equations in our system is $N' + N'' + N = (N - \dim\mathcal{L}) + \dim\mathcal{L} + N = 2N$. Now, to verify that the square system of linear equations (6.5.52) has exactly one solution, it suffices to prove that the homogeneous system

$$X^{1/2}\cdot\delta X\cdot X^{1/2}\in\mathcal{L},\ \ X^{-1/2}\cdot\delta S\cdot X^{-1/2}\in\mathcal{L}^\perp,\ \ L(\delta X,\delta S) = 0$$

has only trivial solution. Let $(\delta X,\delta S)$ be a solution to the homogeneous system. Relation $L(\delta X,\Delta S) = 0$ means that

$$(\delta X)_{ij} = -\frac{\psi_{ij}}{\phi_{ij}}(\delta S)_{ij}, \tag{6.5.53}$$

whence

$$\mathrm{Tr}(\delta X \cdot \delta S) = -\sum_{i,j} \frac{\psi_{ij}}{\phi_{ij}} (\Delta S)_{ij}^2. \tag{6.5.54}$$

Representing δX, δS via ΔX, ΔS according to (6.5.51), we get

$$\mathrm{Tr}(\delta X \cdot \delta S) = \mathrm{Tr}(X^{-1/2} \cdot \Delta X \cdot X^{-1/2} X^{1/2} \cdot \Delta S \cdot X^{1/2}) = \mathrm{Tr}(X^{-1/2} \cdot \Delta X \cdot \Delta S \cdot X^{1/2}) = \mathrm{Tr}(\Delta X \cdot \Delta S),$$

and the latter quantity is 0 due to $\Delta X = X^{1/2} \cdot \delta X \cdot X^{1/2} \in \mathcal{L}$ and $\Delta S = X^{-1/2} \cdot \delta S \cdot X^{-1/2} \in \mathcal{L}^\perp$. Thus, the left-hand side in (6.5.54) is 0; since $\phi_{ij} > 0$, $\psi_{ij} > 0$, (6.5.54) implies that $\delta S = 0$. But then $\delta X = 0$ in view of (6.5.53). Thus, the homogeneous version of (6.5.52) has the trivial solution only, so that (6.5.52) is solvable with a unique solution.

4*. Let δX, δS be the unique solution to (6.5.52), and let ΔX, ΔS be linked to δX, δS according to (6.5.51). Our local goal is to bound from above the Frobenius norms of δX and δS.

From (6.5.52)(c) it follows (cf. derivation of (6.5.54)) that

$$\begin{array}{lrcl} \text{(a)} & (\delta X)_{ij} & = & -\frac{\psi_{ij}}{\phi_{ij}} (\delta S)_{ij} + 2\frac{\mu_+ - x_i s_i}{\phi_{ii}} \delta_{ij}, \quad i, j = 1, \ldots, k, \\ \text{(b)} & (\delta S)_{ij} & = & -\frac{\phi_{ij}}{\psi_{ij}} (\delta X)_{ij} + 2\frac{\mu_+ - x_i s_i}{\psi_{ii}} \delta_{ij}, \quad i, j = 1, \ldots, k. \end{array} \tag{6.5.55}$$

Same as in the concluding part of 3*, relations (6.5.52)(a)–(b) imply that

$$\mathrm{Tr}(\Delta X \cdot \Delta S) = \mathrm{Tr}(\delta X \cdot \delta S) = \sum_{i,j} (\delta X)_{ij} (\delta S)_{ij} = 0. \tag{6.5.56}$$

Multiplying (6.5.55)(a) by $(\delta S)_{ij}$ and summing over i, j, we get, in view of (6.5.56), the relation

$$\sum_{i,j} \frac{\psi_{ij}}{\phi_{ij}} (\delta S)_{ij}^2 = 2 \sum_i \frac{\mu_+ - x_i s_i}{\phi_{ii}} (\delta S)_{ii}; \tag{6.5.57}$$

by symmetric reasoning, we get

$$\sum_{i,j} \frac{\phi_{ij}}{\psi_{ij}} (\delta X)_{ij}^2 = 2 \sum_i \frac{\mu_+ - x_i s_i}{\psi_{ii}} (\delta X)_{ii}. \tag{6.5.58}$$

Now let

$$\theta_i = x_i s_i, \tag{6.5.59}$$

so that in view of (6.5.49) and (6.5.50) one has

$$\begin{array}{ll} \text{(a)} & \sum_i \theta_i = k, \\ \text{(b)} & \sum_i (\theta_i - 1)^2 \le \kappa^2. \end{array} \tag{6.5.60}$$

Observe that

$$\phi_{ij} = \sqrt{x_i x_j}(s_i + s_j) = \sqrt{x_i x_j} \left(\frac{\theta_i}{x_i} + \frac{\theta_j}{x_j} \right) = \theta_j \sqrt{\frac{x_i}{x_j}} + \theta_i \sqrt{\frac{x_j}{x_i}}.$$

Thus,

$$\begin{array}{rcl} \phi_{ij} & = & \theta_j\sqrt{\frac{x_i}{x_j}} + \theta_i\sqrt{\frac{x_j}{x_i}}, \\ \psi_{ij} & = & \sqrt{\frac{x_i}{x_j}} + \sqrt{\frac{x_j}{x_i}}; \end{array} \tag{6.5.61}$$

since $1-\kappa \le \theta_i \le 1+\kappa$ by (6.5.60)(b), we get

$$1-\kappa \le \frac{\phi_{ij}}{\psi_{ij}} \le 1+\kappa. \tag{6.5.62}$$

By the geometric-arithmetic mean inequality we have $\psi_{ij} \ge 2$, whence in view of (6.5.62)

$$\phi_{ij} \ge (1-\kappa)\psi_{ij} \ge 2(1-\kappa) \quad \forall i,\, j. \tag{6.5.63}$$

We now have

$$\begin{array}{rcll} (1-\kappa)\sum_{i,j}(\delta X)_{ij}^2 & \le & \sum_{i,j}\frac{\phi_{ij}}{\psi_{ij}}(\delta X)_{ij}^2 & \text{[see (6.5.62)]} \\ & \le & 2\sum_i \frac{\mu_+ - x_i s_i}{\psi_{ii}}(\delta X)_{ii} & \text{[see (6.5.58)]} \\ & \le & 2\sqrt{\sum_i (\mu_+ - x_i s_i)^2}\sqrt{\sum_i \psi_{ii}^{-2}(\delta X)_{ii}^2} & \\ & \le & \sqrt{\sum_i ((1-\theta_i)^2 - 2\chi k^{-1/2}(1-\theta_i) + \chi^2 k^{-1})}\sqrt{\sum_{i,j}(\delta X)_{ij}^2} & \\ & & & \text{[see (6.5.63)]} \\ & \le & \sqrt{\chi^2 + \sum_i (1-\theta_i)^2}\sqrt{\sum_{i,j}(\delta X)_{ij}^2} & \\ & & & \text{[since } \textstyle\sum_i(1-\theta_i) = 0 \text{ by (6.5.60)(a)]} \\ & \le & \sqrt{\chi^2+\kappa^2}\sqrt{\sum_{i,j}(\delta X)_{ij}^2} & \text{[see (6.5.60)(b)]} \end{array}$$

and from the resulting inequality it follows that

$$\|\delta X\|_2 \le \rho \equiv \frac{\sqrt{\chi^2+\kappa^2}}{1-\kappa}. \tag{6.5.64}$$

Similarly,

$$\begin{aligned}
(1+\kappa)^{-1}\sum_{i,j}(\delta S)_{ij}^2 &\le \sum_{i,j}\tfrac{\psi_{ij}}{\phi_{ij}}(\delta S)_{ij}^2 && \text{[see (6.5.62)]}\\
&\le 2\sum_i \tfrac{\mu_+ - x_i s_i}{\phi_{ii}}(\delta S)_{ii} && \text{[see (6.5.57)]}\\
&\le 2\sqrt{\sum_i(\mu_+ - x_i s_i)^2}\sqrt{\sum_i \phi_{ii}^{-2}(\delta S)_{ii}^2}\\
&\le (1-\kappa)^{-1}\sqrt{\sum_i(\mu_+ - \theta_i)^2}\sqrt{\sum_{i,j}(\delta S)_{ij}^2} && \text{[see (6.5.63)]}\\
&\le (1-\kappa)^{-1}\sqrt{\chi^2+\kappa^2}\sqrt{\sum_{i,j}(\delta S)_{ij}^2} && \text{[same as above]},
\end{aligned}$$

and from the resulting inequality it follows that

$$\|\delta S\|_2 \le \frac{(1+\kappa)\sqrt{\chi^2+\kappa^2}}{1-\kappa} = (1+\kappa)\rho. \tag{6.5.65}$$

5*. We are ready to prove 2*(i)–(ii). We have

$$X_+ = X + \Delta X = X^{1/2}(I + \delta X)X^{1/2},$$

and the matrix $I + \delta X$ is positive definite due to (6.5.64). (Indeed, the right-hand side in (6.5.64) is $\rho \le 1$, whence the Frobenius norm (and therefore the maximum of modulae of eigenvalues) of δX is less than 1.) Note that by the just-indicated reasons $I + \delta X \preceq (1+\rho)I$, whence

$$X_+ \preceq (1+\rho)X. \tag{6.5.66}$$

Similarly, the matrix

$$S_+ = S + \Delta S = X^{-1/2}(X^{1/2}SX^{1/2} + \delta S)X^{-1/2}$$

is positive definite. Indeed, the eigenvalues of the matrix $X^{1/2}SX^{1/2}$ are $\ge \min_i \theta_i \ge 1-\kappa$, while the modulae of eigenvalues of δS, by (6.5.65), do not exceed $\frac{(1+\kappa)\sqrt{\chi^2+\kappa^2}}{1-\kappa} < 1-\kappa$. Thus, the matrix $X^{1/2}SX^{1/2} + \delta S$ is positive definite, whence $S_+ \succ 0$. We have proved 2*(i).

2*(ii) is easy to verify. First, by (6.5.52), we have $\Delta X \in \mathcal{L}$, $\Delta S \in \mathcal{L}^\perp$, and since $X \in \mathcal{L} - B$, $S \in \mathcal{L}^\perp + C$, we have $X_+ \in \mathcal{L} - B$, $S_+ \in \mathcal{L}^\perp + C$. Second, we have

$$\begin{aligned}
\mathrm{Tr}(X_+S_+) &= \mathrm{Tr}(XS + X\cdot\Delta S + \Delta X\cdot S + \Delta X\cdot\Delta S)\\
&= \mathrm{Tr}(XS + X\cdot\Delta S + \Delta X\cdot S)\\
&\quad [\text{since } \mathrm{Tr}(\Delta X\cdot\Delta S) = 0 \text{ due to } \Delta X \in \mathcal{L},\ \Delta S \in \mathcal{L}^\perp]\\
&= \mu_+ k\\
&\quad [\text{take the trace of both sides in (6.5.46)(c)}].
\end{aligned}$$

2*(ii) is proved.

6^*. It remains to verify 2^*(iii). We should bound from above the quantity

$$\Omega = \|\mu_+^{-1} X_+^{1/2} S_+ X_+^{1/2} - I\|_2 = \|X_+^{1/2}(\mu_+^{-1} S_+ - X_+^{-1})X_+^{1/2}\|_2,$$

and our plan is first to bound from above the closely related quantity

$$\begin{array}{rcl} \widehat{\Omega} = \|X^{1/2}(\mu_+^{-1} S_+ - X_+^{-1})X^{1/2}\|_2 & = & \mu_+^{-1}\|Z\|_2, \\ Z & = & X^{1/2}(S_+ - \mu_+ X_+^{-1})X^{1/2}, \end{array} \tag{6.5.67}$$

and then to bound Ω in terms of $\widehat{\Omega}$.

6^*1. Bounding $\widehat{\Omega}$. We have

$$\begin{array}{rcll} Z & = & X^{1/2}(S_+ - \mu_+ X_+^{-1})X^{1/2} & \\ & = & X^{1/2}(S + \Delta S)X^{1/2} - \mu_+ X^{1/2}[X + \Delta X]^{-1}X^{1/2} & \\ & = & XS + \delta S - \mu_+ X^{1/2}[X^{1/2}(I + \delta X)X^{1/2}]^{-1}X^{1/2} & \\ & & & \text{[see (6.5.51)]} \\ & = & XS + \delta S - \mu_+(I + \delta X)^{-1} & \\ & = & XS + \delta S - \mu_+(I - \delta X) - \mu_+[(I + \delta X)^{-1} - I + \delta X] & \\ & = & \underbrace{XS + \delta S + \delta X - \mu_+ I}_{Z^1} + \underbrace{(\mu_+ - 1)\delta X}_{Z^2} + \underbrace{\mu_+[I - \delta X - (I + \delta X)^{-1}]}_{Z^3}, & \end{array}$$

so that

$$\|Z\|_2 \le \|Z^1\|_2 + \|Z^2\|_2 + \|Z^3\|_2. \tag{6.5.68}$$

We intend to bound separately all three terms in the right-hand side of the latter inequality.

Bounding $\|Z^2\|_2$. We have

$$\|Z^2\|_2 = |\mu_+ - 1|\|\delta X\|_2 \le \chi k^{-1/2}\rho \tag{6.5.69}$$

(see (6.5.64) and take into account that $\mu_+ - 1 = -\chi k^{-1/2}$).

Bounding $\|Z^3\|_2$. Let λ_i be the eigenvalues of δX. We have

$$\begin{array}{rcll} \|Z^3\|_2 & = & \|\mu_+[(I + \delta X)^{-1} - I + \delta X]\|_2 & \\ & \le & \|(I + \delta X)^{-1} - I + \delta X\|_2 & \\ & & \text{[since } |\mu_+| \le 1\text{]} & \\ & = & \sqrt{\sum_i \left(\frac{1}{1+\lambda_i} - 1 + \lambda_i\right)^2} & \\ & & \text{[pass to the orthonormal eigenbasis of } \delta X\text{]} & \\ & = & \sqrt{\sum_i \frac{\lambda_i^4}{(1+\lambda_i)^2}} & (6.5.70) \\ & \le & \sqrt{\sum_i \frac{\rho^2\lambda_i^2}{(1-\rho)^2}} & \\ & & \text{[see (6.5.64) and note that } \sum_i \lambda_i^2 = \|\delta X\|_2^2 \le \rho^2\text{]} & \\ & \le & \frac{\rho^2}{1-\rho}. & \end{array}$$

Bounding $\|Z^1\|_2$. This is a bit more involved. We have

$$\begin{array}{rcl} Z^1_{ij} & = & (XS)_{ij} + (\delta S)_{ij} + (\delta X)_{ij} - \mu_+\delta_{ij} \\ & = & (\delta X)_{ij} + (\delta S)_{ij} + (x_i s_i - \mu_+)\delta_{ij} \\ & = & (\delta X)_{ij}\left[1 - \frac{\phi_{ij}}{\psi_{ij}}\right] + \left[2\frac{\mu_+ - x_i s_i}{\psi_{ii}} + x_i s_i - \mu_+\right]\delta_{ij} \\ & & \text{[we have used (6.5.55)(b)]} \\ & = & (\delta X)_{ij}\left[1 - \frac{\phi_{ij}}{\psi_{ij}}\right] \\ & & \text{[since } \psi_{ii} = 2\text{; see (6.5.61)],} \end{array}$$

whence, in view of (6.5.62),

$$|Z^1_{ij}| \le \left|1 - \frac{1}{1-\kappa}\right| |(\delta X)_{ij}| = \frac{\kappa}{1-\kappa}|(\delta X)_{ij}|,$$

so that

$$\|Z^1\|_2 \le \frac{\kappa}{1-\kappa}\|\delta X\|_2 \le \frac{\kappa}{1-\kappa}\rho \tag{6.5.71}$$

(the concluding inequality is given by (6.5.64)).

Assembling (6.5.69), (6.5.70), (6.5.71), and (6.5.68), we come to

$$\|Z\|_2 \le \rho\left[\frac{\chi}{\sqrt{k}} + \frac{\rho}{1-\rho} + \frac{\kappa}{1-\kappa}\right],$$

whence, by (6.5.67),

$$\widehat{\Omega} \le \frac{\rho}{1 - \chi k^{-1/2}}\left[\frac{\chi}{\sqrt{k}} + \frac{\rho}{1-\rho} + \frac{\kappa}{1-\kappa}\right]. \tag{6.5.72}$$

6*2. *Bounding* Ω. We have

$$\begin{array}{rcl} \Omega^2 & = & \|\mu_+^{-1} X_+^{1/2} S_+ X_+^{1/2} - I\|_2^2 \\ & = & \|X_+^{1/2} \underbrace{[\mu_+^{-1} S_+ - X_+^{-1}]}_{\Theta = \Theta^T} X_+^{1/2}\|_2^2 \\ & = & \mathrm{Tr}\left(X_+^{1/2}\Theta X_+ \Theta X_+^{1/2}\right) \\ & \le & (1+\rho)\mathrm{Tr}\left(X_+^{1/2}\Theta X \Theta X_+^{1/2}\right) \\ & & \text{[see (6.5.66)]} \\ & = & (1+\rho)\mathrm{Tr}\left(X_+^{1/2}\Theta X^{1/2} X^{1/2}\Theta X_+^{1/2}\right) \\ & = & (1+\rho)\mathrm{Tr}\left(X^{1/2}\Theta X_+^{1/2} X_+^{1/2}\Theta X^{1/2}\right) \\ & = & (1+\rho)\mathrm{Tr}\left(X^{1/2}\Theta X_+ \Theta X^{1/2}\right) \\ & \le & (1+\rho)^2\mathrm{Tr}\left(X^{1/2}\Theta X \Theta X^{1/2}\right) \\ & & \text{[again by (6.5.66)]} \\ & = & (1+\rho)^2\|X^{1/2}\Theta X^{1/2}\|_2^2 \\ & = & (1+\rho)^2\|X^{1/2}[\mu_+^{-1}S_+ - X_+^{-1}]X^{1/2}\|_2^2 \\ & = & (1+\rho)^2\widehat{\Omega}^2 \\ & & \text{[see (6.5.67)]} \end{array}$$

so that

$$\Omega \le (1+\rho)\widehat{\Omega} = \frac{\rho(1+\rho)}{1-\chi k^{-1/2}}\left[\frac{\chi}{\sqrt{k}} + \frac{\rho}{1-\rho} + \frac{\kappa}{1-\kappa}\right], \quad \rho = \frac{\sqrt{\chi^2+\kappa^2}}{1-\kappa} \tag{6.5.73}$$

(see (6.5.72) and (6.5.64)).

It is immediately seen that if $0 < \chi \le \kappa \le 0.1$, the right-hand side in the resulting bound for Ω is $\le \kappa$, as required in 2*(iii). □

Remark 6.5.2. We have carried out the complexity analysis for a large group of primal-dual path-following methods for SDP (i.e., for the case of $\mathbf{K} = \mathbf{S}^k_+$). In fact, the constructions and the analysis we have presented can be extended word by word to the case when $\mathbf{K}$ is a direct product of semidefinite cones—you should just bear in mind that all symmetric matrices we deal with, like the primal and the dual solutions X, S, the scaling matrices Q, the primal-dual search directions ΔX, ΔS, etc., are block-diagonal with common block-diagonal structure. In particular, our constructions and analysis work for the case of LP—this is the case when $\mathbf{K}$ is a direct product of 1D semidefinite cones. Note that in the case of LP, Zhang's family of primal-dual search directions reduces to a single direction: since now X, S, Q are diagonal matrices, the scaling (6.5.36) $\mapsto$ (6.5.37) does not vary the equations of augmented complementary slackness.

The recipe for translating all we have presented for the case of SDP to the case of LP is very simple: in the above, you should assume all matrices like X, S, ... to be diagonal, and look at what the operations with these matrices, required by the description of the method, do with their diagonals.[71]

6.6 Complexity bounds for linear programming, conic quadratic programming, and semidefinite programming

In what follows we list the best complexity bounds for LP, CQP, and SDP known so far. These bounds are yielded by IP methods and in essence say that the Newton complexity of finding an ϵ-solution to an instance, i.e., the total number of steps of a good interior point algorithm before an ϵ-solution is found, is $O(1)\sqrt{\theta(K)}\ln\frac{1}{\epsilon}$. This is what should be expected in view of the discussion in section 6.5.2. Note, however, that the complexity bounds to follow take into account the need to reach the highway—to come close to the central path before tracing it—while in section 6.5.2 we focused on how fast we could reduce the duality gap after the central path (the highway) has been reached.

Along with complexity bounds expressed in terms of the Newton complexity, we present the bounds on the number of real arithmetic operations required to build an ϵ-solution. Note that these latter bounds are typically conservative—when deriving them, we assume that the data of an instance are completely unstructured, which is usually not the case

[71] Incidentally, one of the first approaches to the design and the analysis of interior point methods for SDP was exactly opposite: you take an IP scheme for LP, replace the words "nonnegative vectors" in its description with "positive semidefinite diagonal matrices," and then erase the adjective "diagonal."

(cf. Warning in section 6.5.2). Exploiting the structure of the data, one can usually reduce significantly computational effort per step of an interior point method and consequently the arithmetic cost of an ϵ-solution.

6.6.1 Complexity of linear programming

Family of problems:

Problem instance: a program

$$\min_{x\in\mathbf{R}^n}\left\{c^Tx : a_i^Tx \le b_i,\ i=1,\dots,m;\ \|x\|_\infty \le R\right\}; \qquad (p)$$

Data:

$$\begin{aligned}&\mathrm{Data}(p) = [n; m; c; a_1, b_1; \dots; a_m, b_m; R],\\ &\mathrm{Size}(p) \equiv \dim \mathrm{Data}(p) = (m+1)(n+1)+2.\end{aligned}$$

ϵ-solution: an $x \in \mathbf{R}^n$ such that

$$\begin{aligned}\|x\|_\infty &\le R,\\ a_i^Tx &\le b_i+\epsilon,\ i=1,\dots,m,\\ c^Tx &\le \mathrm{Opt}(p)+\epsilon.\end{aligned}$$

(As always, the optimal value of an infeasible problem is $+\infty$.)

Newton complexity of ϵ-solution:[72]

$$\mathrm{Compl}^{\mathrm{Nwt}}(p,\epsilon) = O(1)\sqrt{m+n}\,\mathrm{Digits}(p,\epsilon),$$

where

$$\mathrm{Digits}(p,\epsilon) = \ln\left(\frac{\mathrm{Size}(p)+\|\mathrm{Data}(p)\|_1+\epsilon^2}{\epsilon}\right)$$

is the number of accuracy digits in ϵ-solution; see Lecture 5.

Arithmetic complexity of ϵ-solution:

$$\mathrm{Compl}(p,\epsilon) = O(1)(m+n)^{3/2}n^2\mathrm{Digits}(p,\epsilon).$$

[72] In what follows, the precise meaning of a statement "the Newton/arithmetic complexity of finding ϵ-solution of an instance (p) does not exceed N" is as follows: as applied to the input $(\mathrm{Data}(p),\epsilon)$, the method underlying our bound terminates in no more than N steps (respectively, N arithmetic operations) and outputs either a vector, which is an ϵ-solution to the instance, or the correct conclusion "(p) is infeasible."

6.6.2 Complexity of conic quadratic programming

Family of problems:

Problem instance: a program

$$\min_{x\in\mathbf{R}^n}\left\{c^Tx : \|A_ix+b_i\|_2 \le c_i^Tx+d_i,\ i=1,\dots,m;\ \|x\|_2\le R\right\}\quad \left[b_i\in\mathbf{R}^{k_i}\right] \qquad (p)$$

Data:

$$\mathrm{Data}(P)=[n;m;k_1,\dots,k_m;c;A_1,b_1,c_1,d_1;\dots;A_m,b_m,c_m,d_m;R],$$
$$\mathrm{Size}(p)\equiv\dim\mathrm{Data}(p)=(m+\textstyle\sum_{i=1}^m k_i)(n+1)+m+n+3.$$

ϵ-solution: an $x\in\mathbf{R}^n$ such that

$$\begin{aligned}\|x\|_2 &\le R,\\ \|A_ix+b_i\|_2 &\le c_i^Tx+d_i+\epsilon,\ i=1,\dots,m,\\ c^Tx &\le \mathrm{Opt}(p)+\epsilon.\end{aligned}$$

Newton complexity of ϵ-solution:

$$\mathrm{Compl}^{\mathrm{Nwt}}(p,\epsilon)=O(1)\sqrt{m+1}\mathrm{Digits}(p,\epsilon).$$

Arithmetic complexity of ϵ-solution:

$$\mathrm{Compl}(p,\epsilon)=O(1)(m+1)^{1/2}n\left(n^2+m+\sum_{i=0}^m k_i^2\right)\mathrm{Digits}(p,\epsilon).$$

6.6.3 Semidefinite programming

Family of problems:

Problem instance: a program

$$\min_{x\in\mathbf{R}^n}\left\{c^Tx : A_0+\sum_{j=1}^n x_jA_j\succeq 0,\ \|x\|_2\le R\right\}, \qquad (p)$$

where A_j, $j=0,1,\dots,n$, are symmetric block-diagonal matrices with m diagonal blocks $A_j^{(i)}$ of sizes $k_i\times k_i$, $i=1,\dots,m$.

Data:

$$\mathrm{Data}(p)=[n;m;k_1,\dots k_m;c;A_0^{(1)},\dots,A_0^{(m)};\dots;A_n^{(1)},\dots,A_n^{(m)};R],$$
$$\mathrm{Size}(p)\equiv\dim\mathrm{Data}(P)=\left(\textstyle\sum_{i=1}^m\frac{k_i(k_i+1)}{2}\right)(n+1)+m+n+3.$$

ϵ-solution: an x such that

$$\begin{aligned} \|x\|_2 &\leq R, \\ A_0 + \sum_{j=1}^{n} x_j A_j &\succeq -\epsilon I, \\ c^T x &\leq \mathrm{Opt}(p) + \epsilon. \end{aligned}$$

Newton complexity of ϵ-solution:

$$\mathrm{Compl}^{\mathrm{Nwt}}(p,\epsilon) = O(1)\left(1 + \sum_{i=1}^{m} k_i\right)^{1/2} \mathrm{Digits}(p,\epsilon).$$

Arithmetic complexity of ϵ-solution:

$$\mathrm{Compl}(p,\epsilon) = O(1)\left(1 + \sum_{i=1}^{m} k_i\right)^{1/2} n\left(n^2 + n\sum_{i=1}^{m} k_i^2 + \sum_{i=1}^{m} k_i^3\right) \mathrm{Digits}(p,\epsilon).$$

6.7 Concluding remarks

We have discussed interior point P methods for LP, CQP, and SDP as mathematical creatures and emphasizing the ideas underlying the algorithms and the complexity bounds ensured by the methods. Now it is time to address the issues of software implementations of the interior point algorithms and of practical performance of the resulting codes.

As far as the practical performance of recent interior point software is concerned, the situation depends on whether we are speaking about codes for LP or those for CQP and SDP.

• There exists extremely powerful commercial interior point software for LP, able to handle reliably truly large-scale LPs and quite competitive to the best simplex-type codes for LP. One of the best modern LP solvers—CPLEX—allows users to choose between a simplex-type and an interior point mode of execution, and in many cases the second option reduces the running time by orders of magnitude. With a state-of-the-art computer, CPLEX is able to solve routinely real-world LPs with tens and hundreds of thousands of variables and constraints. In the case of favorably structured constraint matrices, the numbers of variables and constraints can be increased to a few million.

• For the time being, interior point software for CQPs and SDPs is not as advanced, reliable, and powerful as the LP software. Roughly speaking, the codes available at the moment can solve SDPs with no more than 1000–2000 design variables. It is difficult to say something definite about interior point software for CQP: the first codes of this type are just becoming available.

There are two groups of factors causing the SDP software to be inferior to the interior point LP software: historical reasons and intrinsic ones. The historical aspect is simple: development of interior point software for LP started in the mid-eighties, while that for SDP and CQP started in the mid-nineties. For the time being, this is quite a difference. Unfortunately, there are intrinsic problems with interior point algorithms for large-scale

(many thousands of variables) SDPs and CQPs. Recall that for interior point methods the influence of the size of an SDP or CQP program on the complexity of its solving by an interior point method is twofold:

1. First, the size affects the Newton complexity of the process. Theoretically, the number of steps required to reduce the duality gap by a constant factor (say, 10) is proportional to $\sqrt{\theta(K)}$. ($\theta(K)$ is twice the total number of conic quadratic inequalities for CQP plus the total row size of the LMIs for SDP.) Thus, we could expect an unpleasant growth of the iteration count with $\theta(K)$. Fortunately, the iteration count for good interior point methods usually is much less than the one given by the worst-case complexity analysis. It is typically about few tens, independent of $\theta(K)$.

2. Second, the larger the instance, the larger the Newton system one should solve to generate a new primal (or primal-dual) search direction and, consequently, the larger the computational effort per step (which is dominated by the necessity to assemble and to solve the Newton system). Now, the system to be solved depends of course on the specific interior point method employed, but it is never simpler than the system (6.5.27) arising in the primal path-following method:

$$\underbrace{\mathcal{A}^*[\nabla^2 K(\bar{X})]\mathcal{A}}_{\mathcal{H}}\,\Delta x = \underbrace{-[t_+c + \mathcal{A}^*\nabla K(\bar{X})]}_{h}. \tag{Nwt}$$

The size n of this system is exactly the design dimension of problem (CP).

To process (Nwt), one should assemble the system (compute $\mathcal{H}$ and h) and then solve it. Whatever the cost of assembling, you should be able to store the resulting matrix $\mathcal{H}$ in memory and to factorize the matrix to get the solution. Both these operations—storing and factorizing $\mathcal{H}$—become prohibitively expensive when $\mathcal{H}$ is a large dense[73] matrix. (Imagine how miserable you might be with the need to store $\frac{5000\times 5001}{2} = 12,502,500$ reals representing a dense 5000×5000 symmetric matrix $\mathcal{H}$ and the need to perform $\approx \frac{5000^3}{6} \approx 2.08 \times 10^{10}$ arithmetic operations to find its Choleski factor.)

The need to assemble and solve large-scale systems of linear equations is intrinsic to interior point methods as applied to large-scale optimization programs, and in this respect there is no difference between LP, CQP, and SDP. The difference is in how difficult it is to handle these large-scale linear systems. In real life LPs, CQPs, and SDPs, the structure of the data allows one to assemble (Nwt) at a cost that is negligibly small compared to the cost of factorizing $\mathcal{H}$, which is a good news. Other good news is that in a typical real-world LP $\mathcal{H}$ is very well structured, a fact that reduces dramatically the effort of factorizing the matrix and storing its Choleski factor. All practical interior point solvers for LP utilize these favorable properties of real-life LPs, and this is where their capability to solve LPs with tens or hundreds of thousands of variables and constraints comes from. Spoil the structure of the problem and an interior point method will be unable to solve an LP with just a few thousands of variables. In contrast to real-life LPs, real-life SDPs typically result in dense matrices $\mathcal{H}$, and this is where severe limitations on the sizes of tractable in practice SDPs come from. In this respect, real life CQPs are somewhere between LPs and SDPs.

It should be mentioned that assembling matrices of the Newton systems and solving these systems by the standard linear algebra techniques is not the only possible way to implement an interior point method. Another option is to solve the Newton systems by

[73] That is, with $O(n^2)$ nonzero entries.

iterative linear algebra routines. With this approach, all we need to solve a system like (Nwt) is a possibility to multiply a given vector by the matrix of the system, and this does not require assembling and storing in memory the matrix itself. For example, to multiply a vector Δx by $\mathcal{H}$, we can use the multiplicative representation of $\mathcal{H}$ as presented in (Nwt). Theoretically, the outlined iterative schemes, as applied to real-life SDPs and CQPs, allow us to reduce the arithmetic cost of building search directions by orders of magnitude and to avoid the need to assemble and store huge dense matrices. These features look like an extremely attractive opportunity; the difficulty, however, is that iterative routines are much more affected by rounding errors than the usual linear algebra techniques. As a result, for the time being iterative linear algebra–based implementation of interior point methods is no more than a challenging goal.

Although the sizes of SDPs (and to some extent CQPs) that can be solved with the existing codes are not as impressive as those of LPs, the possibilities offered to a practitioner by SDP or CQP interior point methods could hardly be overestimated. Just 10 years ago we could not even dream of solving an SDP with more than a few tens of variables, while today we can routinely solve SDPs 20–25 times larger, and we have every reason to expect a significant progress in this direction.

6.8 Exercises to Lecture 6

6.8.1 Canonical barriers

Exercise 6.1. *Prove that the canonical barrier for the Lorentz cone is strongly convex.*

Hint. Rewrite the barrier equivalently as

$$L_k(x) = -\ln\left(t - \frac{x^T x}{t}\right) - \ln t$$

and use the fact that the function $t - \frac{x^T x}{t}$ is concave in (x, t).

Exercise 6.2. *Prove Proposition* 6.3.2.

Hint. Note that the property to be proved is stable with respect to taking direct products, so that it suffices to verify it in the cases of $\mathbf{K} = \mathbf{S}^k_+$ (which is done in Lecture 6) and $\mathbf{K} = \mathbf{L}^k$. Carry out the latter verification.

Exercise 6.3. *Let* $\mathbf{K}$ *be a direct product of Lorentz and semidefinite cones, and let* $K(\cdot)$ *be the canonical barrier for* K. *Prove that whenever* $X \in \mathrm{int}\mathbf{K}$ *and* $S = -\nabla K(X)$, *the matrices* $\nabla^2 K(X)$ *and* $\nabla^2 K(S)$ *are inverses of each other.*

Hint. Differentiate the identity

$$-\nabla K(-\nabla K(X)) = X$$

given by Proposition 6.3.2.

6.8.2 Scalings of canonical cones

We already know that the semidefinite cone $\mathbf{S}^k_+$ is highly symmetric: given two interior points X, X' of the cone, there exists a symmetry of $\mathbf{S}^k_+$—an affine transformation of the space where the cone lives—which maps the cone onto itself and maps X onto X'. The Lorentz cone possesses the same properties, and its symmetries are Lorentz transformations. Writing vectors from $\mathbf{R}^k$ as $x = \binom{u}{t}$ with $u \in \mathbf{R}^{k-1}$, $t \in \mathbf{R}$, we can write a Lorentz transformation as

$$\begin{pmatrix} u \\ t \end{pmatrix} \mapsto \alpha \begin{pmatrix} U\left[u - [\mu t - (\sqrt{1+\mu^2} - 1)e^T u]e\right] \\ \sqrt{1+\mu^2}t - \mu e^T u \end{pmatrix}. \tag{LT}$$

Here $\alpha > 0$, $\mu \in \mathbf{R}$, $e \in \mathbf{R}^{k-1}$, $e^T e = 1$, and an orthogonal $k \times k$ matrix U are the parameters of the transformation.

The first question is whether (LT) is indeed a symmetry of $\mathbf{L}^k$. Note that (LT) is the product of three linear mappings: we first act on vector $x = \binom{u}{t}$ by the special Lorentz transformation

$$L_{\mu,e}: \quad \begin{pmatrix} u \\ t \end{pmatrix} \mapsto \begin{pmatrix} u - [\mu t - (\sqrt{1+\mu^2} - 1)e^T u]e \\ \sqrt{1+\mu^2}t - \mu e^T u \end{pmatrix}, \tag{$*$}$$

then rotate the result by U around the t-axis, and finally multiply the result by $\alpha > 0$. The second and the third transformations clearly map the Lorentz cone onto itself. Thus, to verify that the transformation (LT) maps $\mathbf{L}^k$ onto itself, it suffices to establish the same property for the transformation $(*)$.

Exercise 6.4. *Prove the following.*

1. *Whenever $e \in \mathbf{R}^{k-1}$ is a unit vector and $\mu \in \mathbf{R}$, the linear transformation $(*)$ maps the cone $\mathbf{L}^k$ onto itself. Moreover, transformation $(*)$ preserves the space-time interval $x^T J_k x \equiv -x_1^2 - \cdots - x_{k-1}^2 + x_k^2$:*

$$[L_{\mu,e}x]^T J_k [L_{\mu,e}x] = x^T J_k x \quad \forall x \in \mathbf{R}^k \qquad [\Leftrightarrow L^T_{\mu,e} J_k L_{\mu,e} = J_k]$$

and $L^{-1}_{\mu,e} = L_{\mu,-e}$.

2. *Given a point $\bar{x} \equiv \binom{\bar{u}}{\bar{t}} \in \mathrm{int}\mathbf{L}^k$ and specifying a unit vector e and a real μ according to*

$$\bar{u} = \|\bar{u}\|_2 e,$$
$$\mu = \tfrac{\|\bar{u}\|_2}{\sqrt{\bar{t}^2 - \bar{u}^T\bar{u}}},$$

the resulting special Lorentz transformation $L_{\mu,e}$ maps $\bar{x}$ onto the point $\binom{0_{k-1}}{\sqrt{\bar{t}^2 - \bar{u}^T\bar{u}}}$ on the axis $\{x = \binom{0_{k-1}}{\tau} \mid \tau \geq 0\}$ of the cone $\mathbf{L}^k$. Consequently, the transformation $\sqrt{\frac{2}{\bar{t}^2 - \bar{u}^T\bar{u}}} L_{\mu,e}$ maps $\bar{x}$ onto the central point $e(\mathbf{L}^k) = \binom{0_{k-1}}{\sqrt{2}}$ of the axis—the point where $\nabla^2 L_k(\cdot)$ is the unit matrix.

By Exercise 6.4, given two points x', $x'' \in \mathrm{int}\mathbf{L}^k$, we can find two symmetries L', L'' of $\mathbf{L}^k$ such that $L'x' = e(\mathbf{L}^k)$, $L''x'' = e(\mathbf{L}^k)$, so that the linear mapping $(L'')^{-1}L'$ which

is a symmetry of $\mathbf{L}$ since both L', L'' are, maps x' onto x''. In fact the product $(L'')^{-1}L'$ is again a Lorentz transformation; these transformations form a subgroup in the group of all linear transformations of $\mathbf{R}^k$.

The importance of Lorentz transformations for us comes from the following fact.

Proposition 6.8.1. *The canonical barrier $L_k(x) = -\ln(x^T J_k x)$ of the Lorentz cone is semi-invariant with respect to Lorentz transformations: if L is such a transformation, then*

$$L_k(Lx) = L_k(x) + \text{const}(L).$$

Exercise 6.5. *Prove Proposition* 6.8.1.

As explained in Lecture 6, the semidefinite cone $\mathbf{S}^k_+$ also possesses a rich group of symmetries (which here are of the form $X \mapsto HXH^T$, $\text{Det}H \neq 0$); as in the case of the Lorentz cone, "richness" means that there are enough symmetries to map any interior point of $\mathbf{S}^k_+$ onto any other interior point of the cone. Recall also that the canonical barrier for the semidefinite cone is semi-invariant with respect to these symmetries.

Since our basic components $\mathbf{L}^k$ and $\mathbf{S}^k_+$ possess rich groups of symmetries, so do all canonical cones—those that are direct products of the Lorentz and the semidefinite ones. Given such a cone

$$\mathbf{K} = \mathbf{S}^{k_1}_+ \times \cdots \times \mathbf{S}^{k_p}_+ \times \mathbf{L}^{k_{p+1}} \times \cdots \times \mathbf{L}^{k_m} \subset E = \mathbf{S}^{k_1} \times \cdots \times \mathbf{S}^{k_p} \times \mathbf{R}^{k_{p+1}} \times \cdots \times \mathbf{R}^{k_m}, \quad \text{(Cone)}$$

let us call a scaling of $\mathbf{K}$ a linear transformation $\mathcal{Q}$ of E such that

$$\mathcal{Q}\begin{pmatrix} X_1 \\ \cdots \\ X_m \end{pmatrix} = \begin{pmatrix} \mathcal{Q}_1 X_1 \\ \cdots \\ \mathcal{Q}_m X_m \end{pmatrix}$$

and every $\mathcal{Q}_i$ is either a Lorentz transformation, if the corresponding direct factor of $\mathbf{K}$ is a Lorentz cone (in our notation this is the case when $i > p$), or a semidefinite scaling $\mathcal{Q}_i X_i = H_i X_i H_i^T$, $\text{Det}H_i \neq 0$, if the corresponding direct factor of $\mathbf{K}$ is the semidefinite cone (i.e., if $i \leq p$).

Exercise 6.6. *Prove the following.*

1. *If $\mathcal{Q}$ is a scaling of the cone $\mathbf{K}$, then $\mathcal{Q}$ is a symmetry of $\mathbf{K}$, i.e., it maps $\mathbf{K}$ onto itself, and the canonical barrier $K(\cdot)$ of $\mathbf{K}$ is semi-invariant with respect to $\mathcal{Q}$:*

$$K(\mathcal{Q}X) = K(X) + \text{const}(\mathcal{Q}).$$

2. *For every pair X', X'' of interior point of $\mathbf{K}$, there exists a scaling $\mathcal{Q}$ of $\mathbf{K}$ which maps X' onto X''. In particular, for every point $X \in \text{int}\mathbf{K}$ there exists a scaling $\mathcal{Q}$ which maps X onto the central point $e(\mathbf{K})$ of $\mathbf{K}$ defined as*

$$e(\mathbf{K}) = \begin{pmatrix} I_{k_1} \\ \cdots \\ I_{k_p} \\ \begin{pmatrix} 0_{k_{p+1}-1} \\ \sqrt{2} \end{pmatrix} \\ \cdots \\ \begin{pmatrix} 0_{k_m-1} \\ \sqrt{2} \end{pmatrix} \end{pmatrix},$$

where the Hessian of $K(\cdot)$ is the unit matrix:

$$\langle [\nabla^2 K(e(\mathbf{K}))]X, Y\rangle_E = \langle X, Y\rangle_E.$$

Those readers who passed through section 6.5.4 may guess that scalings play a key role in the LP-CQP-SDP interior point constructions and proofs. The reason is simple: in order to realize what happens with canonical barriers and related entities, like central paths, etc., at certain interior point X of the cone $\mathbf{K}$ in question, we apply an appropriate scaling to convert our point into a simple one, such as the central point $e(\mathbf{K})$ of the cone $\mathbf{K}$, and look what happens at this simple-to-analyze point. We then use the semi-invariance of canonical barriers with respect to scalings to transfer our conclusions to the original case of interest. Let us look at a couple of instructive examples.

6.8.3 Dikin ellipsoid

Let $\mathbf{K}$ be a canonical cone, i.e., a direct product of the Lorentz and the semidefinite cones, E be the space where $\mathbf{K}$ lives (see (Cone)), and $K(\cdot)$ be the canonical barrier for $\mathbf{K}$. Given $X \in \mathrm{int}\mathbf{K}$, we can define a local Euclidean norm

$$\|H\|_X = \sqrt{\langle [\nabla^2 K(X)]H, H\rangle_E}$$

on E.

Exercise 6.7. *Prove that $\|\cdot\|_X$ conforms with scalings: if X is an interior point of $\mathbf{K}$ and $\mathcal{Q}$ is a scaling of $\mathbf{K}$, then*

$$\|\mathcal{Q}H\|_{\mathcal{Q}X} = \|H\|_X \quad \forall H \in E \ \ \forall X \in \mathrm{int}\mathbf{K}.$$

In other words, if $X \in \mathrm{int}\mathbf{K}$ and $Y \in E$, then the $\|\cdot\|_X$-distance between X and Y equals to the $\|\cdot\|_{\mathcal{Q}X}$-distance between $\mathcal{Q}X$ and $\mathcal{Q}Y$.

Hint. Use the semi-invariance of $K(\cdot)$ with respect to $\mathcal{Q}$ to show that

$$D^k K(\mathcal{Q}X)[\mathcal{Q}H_1, \ldots, \mathcal{Q}H_k] = D^k K(X)[H_1, \ldots, H_k].$$

and then set $k = 2$, $H_1 = H_2 = H$.

Exercise 6.8. *For $X \in \mathrm{int}\mathbf{K}$, the Dikin ellipsoid of the point X is defined as the set*

$$W_X = \{Y \mid \|Y - X\|_X \le 1\};$$

see Fig. 6.2. *Prove that $W_X \subset \mathbf{K}$.*

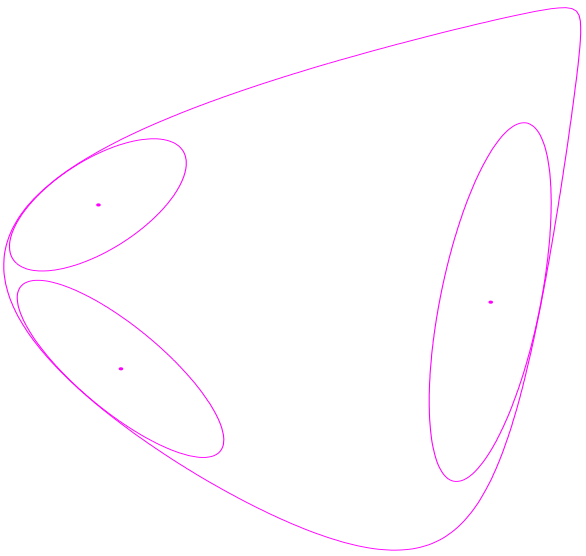


Figure 6.2. *A 2D cross section of* $\mathbf{S}^3_+$ *and cross sections of three Dikin ellipsoids.*

Hint. Note that the property to be proved is stable with respect to taking direct products, so that it suffices to verify it in the cases of $\mathbf{K} = \mathbf{L}^k$ and $\mathbf{K} = \mathbf{S}^k_+$. Further, use the result of Exercise 6.7 to verify that the Dikin ellipsoid conforms with scalings: the image of W_X under a scaling $\mathcal{Q}$ is exactly $W_{\mathcal{Q}X}$. Use this observation to reduce the general case to the one where $X = e(\mathbf{K})$ is the central point of the cone in question, and verify straightforwardly that $W_{e(\mathbf{K})} \subset \mathbf{K}$.

According to Exercise 6.8, the Dikin ellipsoid of a point $X \in \mathrm{int}\mathbf{K}$ is contained in $\mathbf{K}$; in other words, the distance from X to the boundary of $\mathbf{K}$, measured in the $\|\cdot\|_X$-norm, is not too small (it is at least 1). Can this distance be too large? The answer is no—the $\theta(K)$-enlargement of the Dikin ellipsoid contains a significant part of the boundary of $\mathbf{K}$. Specifically, given $X \in \mathrm{int}\mathbf{K}$, let us look at the vector $-\nabla K(X)$. By Proposition 6.3.2, this vector belongs to the interior of $\mathbf{K}$, and since $\mathbf{K}$ is self-dual, it means that the vector has positive inner products with all nonzero vectors from $\mathbf{K}$. It follows that the set (called a conic cap)

$$\mathbf{K}_X = \{Y \in \mathbf{K} \mid \langle -\nabla K(X), X - Y\rangle_E \geq 0\}$$

—the part of $\mathbf{K}$ below the affine hyperplane which is tangent to the level surface of $K(\cdot)$ passing through X—is a convex compact subset of $\mathbf{K}$ that contains an intersection of $\mathbf{K}$ and a small ball centered at the origin; see Fig. 6.3.

Exercise 6.9. *Let* $X \in \mathrm{int}\mathbf{K}$. *Prove the following.*

1. *The conic cap* $\mathbf{K}_X$ *conforms with scalings: if* $\mathcal{Q}$ *is a scaling of* $\mathbf{K}$, *then* $\mathcal{Q}(\mathbf{K}_X) = \mathbf{K}_{\mathcal{Q}X}$.

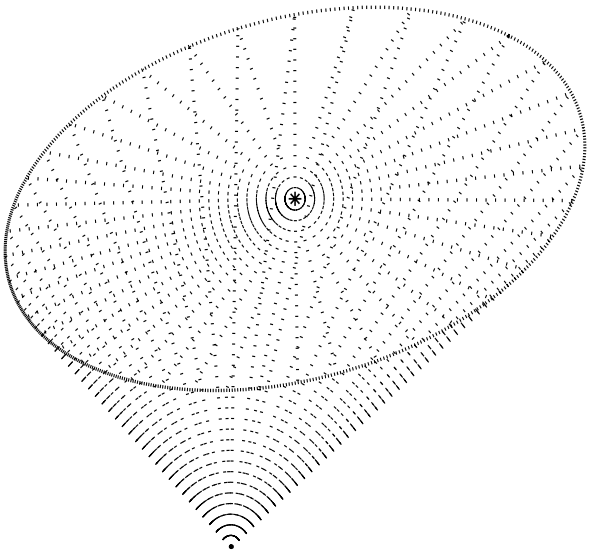


Figure 6.3. *Conic cap of* $\mathbf{L}^3$ *associated with* $X = \begin{pmatrix} 0.3 \\ 0 \\ 1 \end{pmatrix}$.

Hint. From the semi-invariance of the canonical barrier with respect to scalings it is clear that the image, under a scaling, of the hyperplane tangent to a level surface of K is again a hyperplane tangent to (perhaps, another) level surface of K.

2. *Whenever* $Y \in \mathbf{K}$, *one has*

$$\langle \nabla K(X), Y - X \rangle_E \leq \theta(K).$$

3. X *is orthogonal to the hyperplane* $\{H \mid \langle \nabla K(X), H \rangle_E = 0\}$ *in the local Euclidean structure associated with* X, *i.e.,*

$$\langle \nabla K(X), H \rangle_E = 0 \Leftrightarrow \langle [\nabla^2 K(X)]X, H \rangle_E = 0.$$

4. *The conic cap* $\mathbf{K}_X$ *is contained in the* $\|\cdot\|_X$*-ball, centered at* X, *with the radius* $\theta(K)$*:*

$$Y \in \mathbf{K}_X \Rightarrow \|Y - X\|_X \leq \theta(K).$$

Hint to 2–4. Use 1 to reduce the situation to the one where X is the central point of $\mathbf{K}$.

6.8.4 More on canonical barriers

Equipped with scalings, we can establish two additional useful properties of canonical barriers. Let $\mathbf{K}$ be a canonical cone and $K(\cdot)$ be the associated canonical barrier.

Exercise 6.10. *Prove that if* $X \in \mathrm{int}\mathbf{K}$*, then*

$$\max\{\langle \nabla K(X), H\rangle_E \mid \|H\|_X \le 1\} \le \sqrt{\theta(K)}.$$

Hint. Verify that the statement to be proved is scaling invariant, so that it suffices to prove it in the particular case when X is the central point $e(\mathbf{K})$ of the canonical cone $\mathbf{K}$. To verify the statement in this particular case, use Proposition 6.3.1.

Exercise 6.11. *Prove that if* $X \in \mathrm{int}\mathbf{K}$ *and* $H \in \mathbf{K}$*,* $H \neq 0$*, then*

$$\langle \nabla K(X), H\rangle_E < 0$$

and

$$\inf_{t\ge 0} K(X + tH) = -\infty.$$

Derive from the second statement the following proposition.

Proposition 6.8.2. *If* $\mathcal{N}$ *is an affine plane that intersects the interior of* $\mathbf{K}$*, then* K *is bounded below on the intersection* $\mathcal{N} \cap \mathbf{K}$ *if and only if the intersection is bounded.*

Hint. The first statement is an immediate corollary of Proposition 6.3.2. To prove the second fact, observe first that it is scaling invariant, so that it suffices to verify it in the case when X is the central point of $\mathbf{K}$ and then carry out the required verification.

6.8.5 Primal path-following method

We have seen that the primal path-following method (section 6.5.3) is a strange entity: it is a purely primal process for solving the conic problem

$$\min_x \left\{c^T x : X \equiv \mathcal{A}x - B \in \mathbf{K}\right\}, \tag{CP}$$

where $\mathbf{K}$ is a canonical cone, which iterates the updating

$$X \mapsto X_+ = X - \mathcal{A}\underbrace{[\mathcal{A}^*[\nabla^2 K(X)]\mathcal{A}]^{-1}[t_+ c + \mathcal{A}^*\nabla K(X)]}_{\delta x}. \qquad t \mapsto t_+ > t \tag{U}$$

Despite its primal nature, the method is able to produce dual approximate solutions; the corresponding formula is

$$S_+ = -t_+^{-1}[\nabla K(X) - [\nabla^2 K(X)]\mathcal{A}\delta x]. \tag{S}$$

What is the geometric meaning of (S)?

The answer is simple. Given a strictly feasible primal solution $Y = \mathcal{A}y - B$ and a value $\tau > 0$ of the penalty parameter, let us think how we could extend (τ, Y) by a strictly

feasible dual solution S to a triple (τ, Y, S) which is as close as possible, with respect to the distance $\mathrm{dist}(\cdot,\cdot)$, to the point $Z_*(\tau)$ of the primal-dual central path. Recall that the distance in question is

$$\mathrm{dist}((Y,S),Z_*(\tau)) = \sqrt{\langle [\nabla^2 K(Y)]^{-1}(\tau S + \nabla K(Y)), \tau S + \nabla K(Y)\rangle_E}. \qquad \text{(dist)}$$

The simplest way to resolve our question is to choose in the dual feasible plane

$$\mathcal{L}^\perp + C = \{S \mid \mathcal{A}^*(C-S) = 0\} \qquad [C : \mathcal{A}^*C = c]$$

the point S that minimizes the right-hand side of (dist). If we are lucky to get the resulting point in the interior of $\mathbf{K}$, we get the best possible completion of (τ, Y) to an admissible triple (τ, Y, S), the one where S is strictly dual feasible, and the triple is as close as possible to $Z_*(\tau)$.

Now, there is no difficulty in finding the above S—this is just a least squares problem

$$\min_S \left\{ \langle [\nabla^2 K(Y)]^{-1}(\tau S + \nabla K(Y)), \tau S + \nabla K(Y)\rangle_E : \mathcal{A}^* S = c \quad [\equiv \mathcal{A}^* C] \right\}.$$

Exercise 6.12. 1. *Let $\tau > 0$ and $Y = \mathcal{A}y - B \in \mathrm{int}\mathbf{K}$. Prove that the solution to the above least squares problem is*

$$\begin{array}{c} S_* = -\tau^{-1}\left[\nabla K(Y) - [\nabla^2 K(Y)]\mathcal{A}\delta\right], \\ \delta = [\mathcal{A}^*[\nabla^2 K(Y)]\mathcal{A}]^{-1}[\tau c + \mathcal{A}^*\nabla K(Y)], \end{array} \qquad (*)$$

and that the squared optimal value in the problem is

$$\begin{array}{rcl} \lambda^2(\tau, y) & \equiv & [\nabla K(Y)]^T \mathcal{A}\left[\mathcal{A}^*[\nabla^2 K(Y)]\mathcal{A}\right]^{-1}\mathcal{A}^*\nabla K(Y) \\ & = & \left(\|S_* + \tau^{-1}\nabla K(Y)\|_{-\tau^{-1}\nabla K(Y)}\right)^2 \\ & = & \left(\|\tau S_* + \nabla K(Y)\|_{-\nabla K(Y)}\right)^2. \end{array} \qquad (6.8.74)$$

2. *Derive from* 1 *and the result of Exercise* 6.8 *that if the Newton decrement $\lambda(\tau, y)$ of (τ, y) is < 1, then we are lucky—S_* is in the interior of $\mathbf{K}$.*

3. *Derive from* 1–2 *the following.*

Corollary 6.8.1. *Let* (CP) *and its dual be strictly feasible, let $Z_*(\cdot)$ be the primal-dual central path associated with* (CP)*, and let (τ, Y, S) be a triple comprised of $\tau > 0$, a strictly feasible primal solution $Y = \mathcal{A}y - B$, and a strictly feasible dual solution S. Assume that* $\mathrm{dist}((Y,S), Z_*(\tau)) < 1$. *Then*

$$\lambda(\tau, y) = \mathrm{dist}((Y, S_*), Z_*(\tau)) \le \mathrm{dist}((Y,S), Z_*(\tau)),$$

where S_* is the strictly feasible dual solution given by 1–2.

Hint. When proving the second equality in (6.8.74), use the result of Exercise 6.3.

The result stated in Exercise 6.12 is very instructive. First, we see that S_+ in the primal path-following method is exactly the best possible completion of (t_+, X) to an admissible triple (t_+, X, S). Second, we see the following.

Proposition 6.8.3. *If* (CP) *is strictly feasible and there exists* $\tau > 0$ *and a strictly feasible solution* y *of* (CP) *such that* $\lambda(\tau, y) < 1$*, then the problem is strictly dual feasible.*

Indeed, the above scheme, as applied to (τ, y), produces a strictly feasible dual solution!

In fact, given that (CP) *is strictly feasible, the existence of* $\tau > 0$ *and a strictly feasible solution* y *to* (CP) *such that* $\lambda(\tau, y) < 1$ *is a necessary and sufficient condition for* (CP) *to be strictly primal-dual feasible.*

Indeed, the sufficiency of the condition was already established. To prove its necessity, note that if (CP) is primal-dual strictly feasible, then the primal central path is well defined.[74] If $(\tau, Y_*(\tau) = \mathcal{A}y(\tau) - B)$ is on the primal central path, then, of course, $\lambda(\tau, y(\tau)) = 0$.

Note that the Newton decrement admits a nice and instructive geometric interpretation:

Let $\mathbf{K}$ *be a canonical cone,* K *be the associated canonical barrier,* $Y = \mathcal{A}y - B$ *be a strictly feasible solution to* (CP)*, and* $\tau > 0$*. Consider the barrier-generated family*

$$B_t(x) = tc^T x + B(x), \quad B(x) = K(\mathcal{A}x - B).$$

Then

$$\begin{aligned} \lambda(\tau, y) &= \max\{h^T \nabla B_\tau(y) \mid h^T [\nabla^2 B_\tau(y)] h \le 1\} \\ &= \max\{\langle \tau C + \nabla K(Y), H\rangle_E \mid \|H\|_Y \le 1,\ H \in \operatorname{Im} \mathcal{A}\} \quad [Y = \mathcal{A}y - B]. \end{aligned} \tag{6.8.75}$$

Exercise 6.13. *Prove* (6.8.75).

Exercise 6.14. *Let* $\mathbf{K}$ *be a canonical cone,* K *be the associated canonical barrier, and* $\mathcal{N}$ *be an affine plane intersecting* $\operatorname{int}\mathbf{K}$ *such that the intersection* $U = \mathcal{N} \cap \operatorname{int}\mathbf{K}$ *is unbounded. Prove that for every* $X \in U$ *one has*

$$\max\{\langle \nabla K(X), Y - X\rangle_E \mid \|Y - X\|_X \le 1, Y \in \mathcal{N}\} \ge 1.$$

Hint. Assume that the opposite inequality holds true for some $X \in U$ and use (6.8.75) and Proposition 6.8.3 to conclude that the problem with trivial objective

$$\min_X \{\langle 0, X\rangle_E : X \in \mathcal{N} \cap \mathbf{K}\}$$

and its conic dual are strictly feasible. Then use the result of Exercise 2.12 to get a contradiction.

The concluding exercise in this series deals with the toy example of application of the primal path-following method described at the end of section 6.5.3.

[74] In Lecture 6 we announced this fact but did not prove it. Interested readers can give a proof themselves.

Exercise 6.15. *Looking at the data in the table at the end of section* 6.5.3*, do you believe that the corresponding method is exactly the short-step primal path-following method from Theorem* 6.5.1 *with the stepsize policy* (6.5.31)*?*

In fact the data at the end of section 6.5.3 are given by a simple modification of the short-step path-following method: instead of the penalty updating policy (6.5.31), we increase at each step the value of the penalty in the largest ratio satisfying the requirement $\lambda(t_{i+1}, x_i) \leq 0.95$.

6.8.6 Infeasible start path-following method

In our presentation of the interior point path-following methods, we have ignored completely the initialization issue—how to come close to the path in order to start its tracing. There are several techniques for accomplishing this task. We are about to outline one of these techniques—the infeasible start path-following scheme (originating from Roos and Terlaky and from Nesterov). Among other attractive properties, a good pedagogical feature of this technique is that its analysis heavily relies on the results of exercises in sections 6.8.3, 6.8.4, and 6.8.5, thus illustrating the extreme importance of the facts, which at a first glance look a bit esoteric.

Situation and goals. Consider the following situation. We want to solve a conic problem

$$\min_x \left\{c^T x : X \equiv \mathcal{A}x - B \in \mathbf{K}\right\}, \tag{CP}$$

where K is a canonical cone. The corresponding primal-dual pair, in its geometric form, is

$$\min_X \{\langle C, X\rangle_E : X \in (\mathcal{L} - B) \cap \mathbf{K}\}, \tag{P}$$
$$\max_S \left\{\langle B, S\rangle_E : S \in (\mathcal{L}^\perp + C) \cap \mathbf{K}\right\} \tag{D}$$
$$\left[\mathcal{L} = \operatorname{Im}\mathcal{A}, \mathcal{L}^\perp = \operatorname{Ker}\mathcal{A}^*, \mathcal{A}^*C = c\right].$$

From now on we assume the primal-dual pair (P), (D) to be strictly primal-dual feasible.

To proceed, it is convenient to normalize the data as follows: when we shift B along the subspace $\mathcal{L}$, (P) remains unchanged, while (D) is replaced with an equivalent problem (since when shifting B along $\mathcal{L}$, the dual objective, restricted to the dual feasible set, gets a constant additive term). Similarly, when we shift C along $\mathcal{L}^\perp$, the dual problem (D) remains unchanged, and the primal (P) is replaced with an equivalent problem. Thus, we can shift B along $\mathcal{L}$ and C along $\mathcal{L}^\perp$, while not varying the primal-dual pair (P), (D) (or, better to say, converting it to an equivalent primal-dual pair). With appropriate shift of B along $\mathcal{L}$ we can enforce $B \in \mathcal{L}^\perp$, and with appropriate shift of C along $\mathcal{L}^\perp$ we can enforce $C \in \mathcal{L}$. Thus, we can assume that from the very beginning the data are normalized by the requirements

$$B \in \mathcal{L}^\perp, \quad C \in \mathcal{L}, \tag{Nrm}$$

which, in particular, implies that $\langle C, B\rangle_E = 0$, so that the duality gap at a pair (X, S) of primal-dual feasible solutions becomes

$$\text{DualityGap}(X, S) = \langle X, S\rangle_E = \langle C, X\rangle_E - \langle B, S\rangle_E \ \ [= \langle C, X\rangle_E - \langle B, S\rangle_E + \langle C, B\rangle_E].$$

Our goal is rather ambitious: to develop an interior point method for solving (P), (D) that requires neither a priori knowledge of a primal-dual strictly feasible pair of solutions, nor a specific initialization phase.

The scheme. The construction we are about to present achieves the announced goal as follows.

1. We write down the following system of conic constraints in variables X, S and additional scalar variables τ, σ:

$$\begin{array}{lrcl} \text{(a)} & X + \tau B - P & \in & \mathcal{L}, \\ \text{(b)} & S - \tau C - D & \in & \mathcal{L}^\perp, \\ \text{(c)} & \langle C, X\rangle_E - \langle B, S\rangle_E + \sigma - d & = & 0, \\ \text{(d)} & X & \in & \mathbf{K}, \\ \text{(e)} & S & \in & \mathbf{K}, \\ \text{(f)} & \sigma & \geq & 0, \\ \text{(g)} & \tau & \geq & 0. \end{array} \qquad \text{(C)}$$

   Here P, D, d are certain fixed entities which we choose in such a way that

   (i) we can easily point out a strictly feasible solution $\widehat{Y} = (\widehat{X}, \widehat{S}, \widehat{\sigma}, \widehat{\tau} = 1)$ to the system;

   (ii) the solution set $\mathcal{Y}$ of (C) is unbounded. Moreover, whenever $Y_i = (X_i, S_i, \sigma_i, \tau_i) \in \mathcal{Y}$ is an unbounded sequence, we have $\tau_i \to \infty$.

2. Imagine that we have a mechanism that allows us to run away to ∞ along $\mathcal{Y}$, i.e., to generate a sequence of points $Y_i = (X_i, S_i, \sigma_i, \tau_i) \in \mathcal{Y}$ such that $\|Y_i\| \equiv \sqrt{\|X_i\|_E^2 + \|S_i\|_E^2 + \sigma_i^2 + \tau_i^2} \to \infty$. In this case, by (ii) $\tau_i \to \infty$, $i \to \infty$. Let us define the normalizations

$$\widetilde{X}_i = \tau_i^{-1} X_i, \quad \widetilde{S}_i = \tau_i^{-1} S_i$$

   of X_i, S_i. Since $(X_i, S_i, \sigma_i, \tau_i)$ is a solution to (C), these normalizations satisfy the relations

$$\begin{array}{lrcl} \text{(a)} & \widetilde{X}_i & \in & (\mathcal{L} - B + \tau_i^{-1} P) \cap \mathbf{K}, \\ \text{(b)} & \widetilde{S}_i & \in & (\mathcal{L}^\perp + C + \tau_i^{-1} D) \cap \mathbf{K}, \\ \text{(c)} & \langle C, \widetilde{X}_i\rangle_E - \langle B, \widetilde{S}_i\rangle_E & \leq & \tau_i^{-1} d. \end{array} \qquad \text{(C}'\text{)}$$

   Since $\tau_i \to \infty$, relations (C′) say that as $i \to \infty$, the normalizations $\widetilde{X}_i, \widetilde{S}_i$ simultaneously approach primal-dual feasibility for (P), (D) (see (C′)(a), (b)) and primal-dual optimality (see (C′)(c) and recall that the duality gap, with our normalization $\langle C, B\rangle_E = 0$, is $\langle C, X\rangle_E - \langle B, S\rangle_E$).

3. *The* issue, of course, is how to build a mechanism that allows us to run away to ∞ along $\mathcal{Y}$. The mechanism we intend to use is as follows. (C) can be rewritten in the generic form

$$Y \equiv \begin{pmatrix} X \\ S \\ \sigma \\ \tau \end{pmatrix} \in (\mathcal{M} + R) \cap \widetilde{\mathbf{K}}, \tag{G}$$

where

$$\widetilde{\mathbf{K}} = \mathbf{K} \times \mathbf{K} \times \underbrace{\mathbf{S}^1_+}_{=\mathbf{R}_+} \times \underbrace{\mathbf{S}^1_+}_{=\mathbf{R}_+},$$

$$\mathcal{M} = \left\{ \begin{pmatrix} U \\ V \\ s \\ r \end{pmatrix} \;\middle|\; \begin{array}{c} U + rB \in \mathcal{L}, \\ V - rC \in \mathcal{L}^\perp, \\ \langle C, U\rangle_E - \langle B, V\rangle_E + s = 0 \end{array} \right\}$$

is a linear subspace in the space $\widetilde{E}$ where the cone $\widetilde{\mathbf{K}}$ lives, and

$$R = \begin{pmatrix} P \\ D \\ d - \langle C, P\rangle_E + \langle B, D\rangle_E \\ 0 \end{pmatrix} \in \widetilde{E}.$$

The cone $\widetilde{\mathbf{K}}$ is a canonical cone along with $\mathbf{K}$; as such, it is equipped with the corresponding canonical barrier $\widetilde{K}(\cdot)$. Let

$$\widehat{Y} = \begin{pmatrix} \widehat{X} \\ \widehat{S} \\ \widehat{\sigma} \\ \widehat{\tau} = 1 \end{pmatrix}$$

be the strictly feasible solution to (G) given by 1(i), and let

$$\widetilde{C} = -\nabla \widetilde{K}(\widehat{Y}).$$

Consider the auxiliary problem

$$\min_Y \left\{ \langle \widetilde{C}, Y\rangle_{\widetilde{E}} : Y \in (\mathcal{M} + R) \cap \widetilde{\mathbf{K}} \right\}. \tag{Aux}$$

By the origin of $\widetilde{C}$, the point $\widehat{Y}$ lies on the primal central path $\widetilde{Y}_*(t)$ of this auxiliary problem:

$$\widehat{Y} = \widetilde{Y}_*(1).$$

Let us trace the primal central path $\widetilde{Y}_*(\cdot)$, but decrease the value of the penalty instead of increasing it, thus enforcing the penalty to approach 0. What will happen in this

process? Recall that the point $\widetilde{Y}_*(t)$ of the primal central path of (Aux) minimizes the aggregate

$$t\langle \widetilde{C}, Y\rangle_{\widetilde{E}} + \widetilde{K}(Y)$$

over $Y \in \mathcal{Y}$. When t is small, we essentially are trying to minimize just $\widetilde{K}(Y)$. But the canonical barrier, restricted to an unbounded intersection of an affine plane and the associated canonical cone, is not bounded below on this intersection (see Proposition 6.8.2). Therefore, if we were minimizing the barrier $\widetilde{K}$ over $\mathcal{Y}$, the minimum would be achieved at infinity; it is natural to guess (and this is indeed true) that when minimizing a slightly perturbed barrier, the minimum will run away to infinity as the level of perturbations goes to 0. Thus, we may expect (and again it is indeed true) that $\|\widetilde{Y}_*(t)\| \to \infty$ as $t \to +0$, so that when tracing the path $\widetilde{Y}(t)$ as $t \to 0$, we are achieving our goal of running away to infinity along $\mathcal{Y}$.

Now let us implement the outlined approach.

Specifying *P, D, d*. Given the data of (CP), let us choose somehow $P >_{\mathbf{K}} B$, $D >_{\mathbf{K}} -C$, $\widehat{\sigma} > 0$ and set

$$d = \langle C, P - B\rangle_E - \langle B, D + C\rangle_E + \widehat{\sigma}.$$

Exercise 6.16. *Prove that with the above setup, the point*

$$\widehat{Y} = \begin{pmatrix} \widehat{X} = P - B \\ \widehat{S} = C + D \\ \widehat{\sigma} \\ \widehat{\tau} = 1 \end{pmatrix}$$

is a strictly feasible solution to (Aux). *Thus, our setup ensures* 1(i).

Verifying 1(ii). This step is crucial:

Exercise 6.17. *Let* (Aux′) *be the problem dual to* (Aux). *Prove that* (Aux), (Aux′) *is a strictly primal-dual feasible pair of problems.*

Hint. By construction, (Aux) is strictly feasible; to prove that (Aux′) is also strictly feasible, use Proposition 6.8.3.

Exercise 6.18. *Prove that with the outlined setup the feasible set $\mathcal{Y}$ of* (Aux) *is unbounded.*

Hint. Use the criterion of boundedness of the feasible set of a feasible conic problem (Exercise 2.11) which as applied to (Aux) reads as follows: the feasible set of (Aux) is bounded if and only if $\mathcal{M}^\perp$ intersects the interior of the cone dual to $\widetilde{\mathbf{K}}$. (Since $\widetilde{\mathbf{K}}$ is a canonical cone, the cone dual to it is $\widetilde{\mathbf{K}}$ itself.)

The result of Exercise 6.18 establishes the major part of 1(ii). The remaining part of the latter property is given by the next exercise.

Exercise 6.19. *Let $\bar{X}$, $\bar{S}$ be a strictly feasible pair of primal-dual solutions to* (P), (D) *(recall that the latter pair of problems was assumed to be strictly primal-dual feasible), so that there exists $\gamma \in (0, 1]$ such that*

$$\begin{array}{rcll} \gamma\|X\|_E & \le & \langle \bar{S}, X\rangle_E & \forall X \in \mathbf{K}, \\ \gamma\|S\|_E & \le & \langle \bar{X}, S\rangle_E & \forall S \in \mathbf{K}. \end{array}$$

Prove that if

$$Y = \begin{pmatrix} X \\ S \\ \sigma \\ \tau \end{pmatrix}$$

is feasible for (Aux)*, then*

$$\begin{array}{rcl} \|Y\|_{\widetilde{E}} & \le & \alpha\tau + \beta, \\ & & \alpha = \gamma^{-1}\left[\langle \bar{X}, C\rangle_E - \langle \bar{S}, B\rangle_E\right] + 1, \\ & & \beta = \gamma^{-1}\left[\langle \bar{X} + B, D\rangle_E + \langle \bar{S} - C, P\rangle_E + d\right]. \end{array} \tag{6.8.76}$$

Use this result to complete the verification of 1(ii).

Tracing the path $\widetilde{Y}_*(t)$ as $t \to 0$. The path $\widetilde{Y}_*(t)$ is the primal central path of a certain strictly primal-dual feasible primal-dual pair of conic problems associated with a canonical cone (see Exercise 6.17). The only difference with the situation discussed in Lecture 6 is that now we want to trace the path as $t \to +0$, starting the process from the point $\widehat{Y} = \widetilde{Y}_*(1)$ given by 1(i), rather than to trace the path as $t \to \infty$. It turns out that we have exactly the same possibilities to trace the path $\widetilde{Y}_*(t)$ as the penalty parameter approaches 0 as when tracing the path as $t \to \infty$; in particular, we can use short-step primal and primal-dual path-following methods with stepsize policies opposite those mentioned, respectively, in Theorem 6.5.1 and Theorem 6.5.2. (Opposite means that instead of increasing the penalty at each iteration in certain ratio, we decrease it in exactly the same ratio.) It can be verified (take it for granted!) that the results of Theorems 6.5.1 and 6.5.2 remain valid in this new situation as well. Thus, to generate a triple (t, Y, U) such that $t \in (0, 1)$, Y is strictly feasible for (Aux), U is strictly feasible for the problem (Aux$'$) dual to (Aux), and $\mathrm{dist}((Y, U), \widetilde{Z}_*(t)) \le \kappa \le 0.1$, it suffices to carry out

$$\mathcal{N}(t) = O(1)\sqrt{\theta(\widetilde{K})}\ln\frac{1}{t} = O(1)\sqrt{\theta(K)}\ln\frac{1}{t}$$

steps of the path-following method. Here $\widetilde{Z}_*(\cdot)$ is the primal-dual central path of the primal-dual pair of problems (Aux), (Aux$'$), and dist from now on is the distance to this path, as defined in section 6.4.2. Thus, we understand the cost of arriving at a close-to-the-path triple (t, Y, U) with a desired value $t \in (0, 1)$ of the penalty. Further, our original scheme explains how to convert the Y-component of such a triple into a pair X_t, S_t of approximate solutions to (P), (D),

$$X_t = \frac{1}{\tau[Y]}X[Y], \quad S_t = \frac{1}{\tau[Y]}S[Y],$$

where

$$Y = \begin{pmatrix} X[Y] \\ S[Y] \\ \sigma[Y] \\ \tau[Y] \end{pmatrix}.$$

What we do *not* know for the moment is

(?) *What is the quality of the resulting pair* (X_t, S_t) *of approximate solutions to* (P), (D) *as a function of* t*?*

Looking at (C′), we see that (?) is, essentially, the question of how rapidly the component $\tau[Y]$ of our close-to-the-path triple (t, Y, U) blows up when t approaches 0. In view of the bound (6.8.76), the latter question, in turn, becomes, How large is $\|Y\|_{\widetilde{E}}$ when t is small? The answers to all these questions are given in the following two exercises.

Exercise 6.20. *Let* (t, Y, U) *be a close-to-the-path triple, so that* $t > 0$*,* Y *is strictly feasible for* (Aux)*,* U *is strictly feasible for the dual to* (Aux) *problem* (Aux′)*, and*

$$\operatorname{dist}((Y, U), \widetilde{Z}_*(t)) \le \kappa \le 0.1.$$

Verify that

$$\begin{array}{lrcl} \text{(a)} & \max\{\langle -\nabla\widetilde{K}(Y), H\rangle_E : H \in \mathcal{M},\ \|H\|_Y \le 1\} & \ge & 1, \\ \text{(b)} & \max\{\left|\langle t\widetilde{C} + \nabla\widetilde{K}(Y), H\rangle_E\right| : H \in \mathcal{M},\ \|H\|_Y \le 1\} & \le & \kappa \le 0.1. \end{array} \tag{6.8.77}$$

Conclude from these relations that

$$\max\{\langle -t\widetilde{C}, H\rangle_E : H \in \mathcal{M},\ \|H\|_Y \le 1\} \ge 0.9. \tag{6.8.78}$$

Hint. To verify (6.8.77)(a), use the result of Exercise 6.14. To verify (6.8.77)(b), use Corollary 6.8.1 (with (Aux) playing the role of (CP), t playing the role of τ, and U playing the role of S) and the result of Exercise 6.13.

Now consider the following geometric construction. Given a triple (t, Y, U) satisfying the premise of Exercise 6.20, let us denote by W^1 the intersection of the Dikin ellipsoid of $\widehat{Y}$ with the feasible plane of (Aux) and by W^t the intersection of the Dikin ellipsoid of Y with the same feasible plane. Let us also extend the line segment $[\widehat{Y}, Y]$ to the left of $\widehat{Y}$ until it crosses the boundary of W^1 at a certain point Q. Further, let us choose $H \in \mathcal{M}$ such that $\|H\|_Y = 1$ and

$$\langle -t\widetilde{C}, H\rangle_{\widetilde{E}} \ge 0.9$$

(such an H exists in view of (6.8.78)) and set

$$M = Y + H; \quad N = \widehat{Y} + \omega H, \quad \omega = \frac{\|\widehat{Y} - P\|_2}{\|Y - P\|_2}.$$

The cross section of the entities involved by the 2D plane passing through Q, Y, M looks like Fig. 6.4.

Exercise 6.21. 1. *Prove that the points* Q, M, N *belong to the feasible set of* (Aux).

Hint. Use Exercise 6.8 to prove that Q, M are feasible for (Aux); note that N is a convex combination of Q, M.

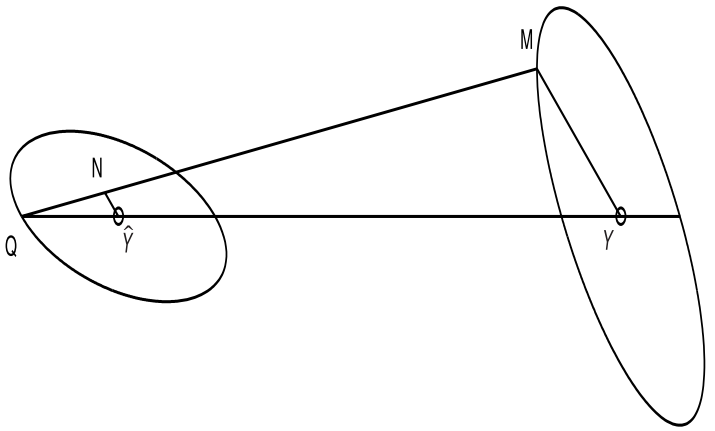


Figure 6.4.

2. *Prove that*

$$\langle \nabla \widetilde{K}(\widehat{Y}), N - \widehat{Y} \rangle_E = \frac{\omega}{t} \langle -t\widetilde{C}, H \rangle_{\widetilde{E}} \geq \frac{0.9\omega}{t}.$$

Hint. Recall that by definition $\widetilde{C} = -\nabla \widetilde{K}(\widehat{Y})$!

3. *Derive from* 1 and 2 *that*

$$\omega \leq \frac{\theta(\widetilde{K})t}{0.9}.$$

Conclude from the resulting bound on ω that

$$\begin{gathered} \|Y\|_{\widetilde{E}} \geq \tfrac{\Omega}{t} - \Omega', \\ \Omega = \frac{0.9 \min\limits_D [\|D\|_{\widetilde{E}} \,\big|\, \|D\|_{\widehat{Y}} = 1]}{\theta(\widetilde{K})}, \quad \Omega' = \max_D [\|D\|_{\widetilde{E}} \,\big|\, \|D\|_{\widehat{Y}} = 1]. \end{gathered} \tag{6.8.79}$$

Note that Ω and Ω' are positive quantities depending on our starting point $\widehat{Y}$ and completely independent of t!

Hint. To derive the bound on ω, use the result of Exercise 6.9.2.

Exercise 6.22. *Derive from the results of Exercises* 6.21 *and* 6.19 *that there exists a positive constant Θ (depending on the data of* (Aux)*) such that*

(#) *whenever a triple (t, Y, U) is close to the path (see Exercise* 6.20*) and*

$$Y = \begin{pmatrix} X \\ S \\ \sigma \\ \tau \end{pmatrix},$$

one has

$$\tau \geq \frac{1}{\Theta t} - \Theta.$$

Consequently, when $t \leq \frac{1}{2\Theta^2}$, *the pair* $(X_\tau = \tau^{-1}X, S_\tau = \tau^{-1}S)$ *satisfies the relations (cf.* (C′)*)*

$$\begin{array}{rl} X_\tau \in \mathbf{K} \cap (\mathcal{L} - B + 2t\Theta P) & [\text{primal } O(t)\text{-feasibility}], \\ S_\tau \in \mathbf{K} \cap (\mathcal{L}^\perp + C + 2t\Theta D) & [\text{dual } O(t)\text{-feasibility}], \\ \langle C, X_\tau\rangle_E - \langle B, S_\tau\rangle_E \leq 2t\Theta d & [O(t)\text{-duality gap}]. \end{array} \qquad (+)$$

Statement (#) says that in order to get an ϵ-primal-dual feasible ϵ-optimal solution to (P), (D), it suffices to trace the primal central path of (Aux), starting at the point $\widehat{Y}$ (penalty parameter equals 1) until a close-to-the-path point with penalty parameter $O(\epsilon)$ is reached, which requires $O(\sqrt{\theta(K)}\ln\frac{1}{O(\epsilon)})$ iterations. Thus, we arrive at a process with the same complexity characteristics as for the path-following methods discussed in Lecture 6. Note, however, that now we have absolutely no troubles with how to start tracing the path.

At this point, a careful reader should protest: relations (+) do say that when t is small, X_τ is nearly feasible for (P) and S_τ is nearly feasible for (D). But why do we know that X_τ, S_τ are nearly optimal for the respective problems? What pretends to ensure the latter property is the $O(t)$-duality gap relation in (+), and indeed, the left-hand side of this inequality looks like the duality gap, while the right-hand side is $O(t)$. But in fact the relation[75]

$$\text{DualityGap}(X, S) \equiv [\langle C, X\rangle_E - \text{Opt(P)}] + [\text{Opt(D)} - \langle B, S\rangle_E] = \langle C, X\rangle_E - \langle B, S\rangle_E$$

is valid only for primal-dual feasible pairs (X, S), while our X_τ, S_τ are only $O(t)$-feasible.

Here is the missing element:

Exercise 6.23. *Let the primal-dual pair of problems* (P), (D) *be strictly primal-dual feasible and be normalized by* $\langle C, B\rangle_E = 0$, *let* (X_*, S_*) *be a primal-dual optimal solution to the pair, and let* X, S ϵ*-satisfy the feasibility and optimality conditions for* (P), (D), *i.e.,*

$$\begin{array}{ll} \text{(a)} & X \in \mathbf{K} \cap (\mathcal{L} - B + \Delta X),\ \|\Delta X\|_E \leq \epsilon, \\ \text{(b)} & S \in \mathbf{K} \cap (\mathcal{L}^\perp + C + \Delta S),\ \|\Delta S\|_E \leq \epsilon, \\ \text{(c)} & \langle C, X\rangle_E - \langle B, S\rangle_E \leq \epsilon. \end{array}$$

Prove that

$$\begin{array}{rcl} \langle C, X\rangle_E - \text{Opt(P)} & \leq & \epsilon(1 + \|X_* + B\|_E), \\ \text{Opt(D)} - \langle B, S\rangle_E & \leq & \epsilon(1 + \|S_* - C\|_E). \end{array}$$

Exercise 6.24. *Implement the infeasible-start path-following method.*

[75] In fact, in the right-hand side there should also be the term $\langle C, B\rangle_E$. Recall, however, that with our setup this term is zero.

Solutions to Selected Exercises

Exercises to Lecture 1

Uniform approximation

Exercise 1.2. Let $\alpha < \infty$, and assume L is α-regular, i.e., the functions from L are continuously differentiable and

$$\|f'\|_\infty \le \alpha \|f\|_\infty \quad \forall f \in L.$$

Assume that $T \subset \Delta$ is such that the distance from a point in Δ to the closest point of T does not exceed $\beta < \alpha^{-1}$. Prove that under these assumptions

$$\kappa_L(T) \le \frac{1}{1-\alpha\beta}.$$

Solution. Let $f \in L$, $M = \|f\|_\infty$, and let $a \in \Delta$ be the point where $|f(a)| = M$. There exists a point $t \in T$ such that $|t-a| \le \beta$. Since L is regular, we have $|f(a) - f(t)| \le M\alpha\beta$, whence $|f(t)| \ge M(1-\alpha\beta)$, and consequently $\|f\|_{T,\infty} \ge |f(t)| \ge M(1-\alpha\beta)$. □

Exercise 1.5. Assume that $\Delta = [0, 2\pi]$, and let L be a linear space of functions on Δ comprising all trigonometric polynomials of degree $\le k$. Let T be an equidistant M-point grid on Δ:

$$T = \left\{ \frac{(2l+1)\pi}{M} \right\}_{l=0}^{M-1}.$$

1. Prove that if $M > k\pi$, then T is L-dense, with

$$\kappa_L(T) \le \frac{M}{M - k\pi}.$$

2. Prove that the above inequality remains valid if we replace T with its arbitrary shift modulo 2π, i.e., treat Δ as the unit circumference and rotate T by an angle.

3. Prove that if T is an arbitrary M-point subset of Δ with $M \le k$, then $\kappa_L(T) = \infty$.

Solution. 1. It suffices to apply the result of Exercise 1.2. In the case in question $\beta = \frac{\pi}{M}$, and, by the Bernshtein's Theorem, $\alpha = k$.

2. The solution follows from 1 because the space of trigonometric polynomials is invariant with respect to cyclic shift of the argument by any angle.

3. Let $T = \{t_i\}_{i=1}^{M}$. The function

$$f(t) = \prod_{i=1}^{M} \sin(t - t_i)$$

is a trigonometric polynomial of degree $M \le k$. This function vanishes on T (i.e., $\|f\|_{T,\infty} = 0$), although its uniform norm on Δ is positive. □

Exercise 1.6. Let $\Delta = [-1, 1]$ and let L be the space of all algebraic polynomials of degree $\le k$.

1. Assume that $2M > \pi k$ and T is the M-point set on Δ as follows:

$$T = \left\{ t_l = \cos\left(\frac{(2l+1)\pi}{2M} \right) \right\}_{l=0}^{M-1}.$$

Then T is L-dense, with

$$\kappa_L(T) \le \frac{2M}{2M - \pi k}.$$

2. Let T be an M-point set on Δ with $M \le k$. Then $\kappa_L(T) = \infty$.

Solution. 1. Let us pass from the functions $f \in L$ to the functions $f^+(\phi) = f(\cos(\phi))$, $\phi \in [0, 2\pi]$. Note that f^+ is a trigonometric polynomial of degree $\le k$. Let

$$T^+ = \left\{ \phi_l = \frac{(2l+1)\pi}{2M} \right\}_{l=0}^{2M-1}.$$

According to the result of Exercise 1.5.1, for every $f \in L$ we have

$$\begin{aligned} \|f\|_\infty &= \|f^+\|_\infty \\ &\le \tfrac{2M}{2M-\pi k} \max_{l=0,\dots,2M-1} |f^+(\phi_l)| \\ &= \tfrac{2M}{2M-\pi k} \max_{l=0,\dots,2M-1} |f(\cos(\phi_l))| \\ &= \tfrac{2M}{2M-\pi k} \max_{l=0,\dots,M-1} |f(t_l)|. \end{aligned}$$

(Note that when ϕ takes values in T^+, the quantity $\cos(\phi)$ takes values in T.)

2. Whenever the cardinality of T is $\le k$, L contains a nontrivial polynomial

$$f(t) = \prod_{t' \in T} (t - t')$$

which vanishes on T. □

Integration formulas and Gauss points

(∗) *Let Δ be a subset of $\mathbf{R}^k$, let L be an n-dimensional linear space comprising continuous real-valued functions on Δ, and let $I(f) : L \to \mathbf{R}$ be an integral—a linear functional such that $I(f) \geq 0$ for every $f \in L$ such that $f(t) \geq 0$ everywhere on Δ. Assume also that if a function $f \in L$ is nonnegative on Δ and is not identically 0, then $I(f) > 0$. Then there exists a precise n-point cubature formula for I, i.e., there are n points $t_1, \ldots, t_n \in \Delta$ and n nonnegative weights $\alpha_1, \ldots, \alpha_n$ such that*

$$I(f) = \sum_{i=1}^{n} \alpha_i f(t_i) \quad \forall f \in L.$$

Exercise 1.17. 1. Prove (*) for the case of finite Δ.

Solution. We may assume that I is not identically zero—otherwise there is nothing to prove. The sets $A_t = \{f \in L \mid f(t) \leq 0, I(f) = 1\}$ are convex sets belonging to a fixed hyperplane. We claim that there exist n sets of this type $A_{t_1}, \ldots, A_{t_n}$ with an empty intersection. Indeed, in the opposite case, by the Helley theorem, there would exist an element $f \in L$ belonging to all A_t, $t \in \Delta$, i.e., such that $f(t) \leq 0$ everywhere on Δ and $I(f) = 1$, which is impossible.

The fact that the intersection of $A_{t_1}, \ldots, A_{t_n}$ is empty means that the inequality $I(f) \leq 0$ is a corollary of the system of linear inequalities $f(t_i) = 0$, $i = 1, \ldots, n$. By the Farkas Lemma, it follows that the linear functional I is a combination, with nonnegative coefficients, of the n linear functionals $f \mapsto f(t_i)$. □

2. Prove (*) for the general case.

Solution. Introducing the same sets A_t as before, it suffices to prove that there exists a finite family of these sets with an empty intersection. Indeed, if we know that $A_{t_1}, \ldots, A_{t_N}$ have no point in common, then, by the Helley theorem, already a properly chosen n-element subfamily of this family of sets has an empty intersection, and we may complete the proof in the same way as in the case of finite Δ. To prove that there exists a finite family of the sets A_t with an empty intersection, assume that it is not the case, and let us lead this assumption to a contradiction. Let $t_1, t_2, \ldots$ be a sequence of points from Δ such that every point of this set belongs to the closure of the sequence, and let $f_1, \ldots, f_n$ be a basis in L. Under our assumption, for every N there exists a function $f^N \in L$ that is nonpositive at the points $t_1, \ldots, t_N$ and $I(f^N) = 1$. After an appropriate normalization, we can convert f^N to a function $g^N \in L$ such that g^N is nonpositive at the points $t_1, \ldots, t_N$, $I(g^N) \geq 0$, and the Euclidean norm of the vector λ^N of the coefficients of g^N in the basis $f_1, \ldots, f_n$ is 1. Passing to a subsequence $\{N_i\}$, we may assume that the vectors λ^{N_i} converge to a vector λ (which of course is nonzero). Then the functions g^{N_i}

pointwise converge to the function $g(t) = \sum_{j=1}^{n} \lambda_j f_j(t)$; this function clearly is nonpositive on the sequence $t_1, t_2, \ldots$, and since it is continuous on Δ, it is nonpositive everywhere. Since $0 \leq I(g^{N_i}) = \sum_{j=1}^{n} \lambda_j^{N_i} I(f_j)$, we conclude that $I(g) = \sum_{j=1}^{n} \lambda_j I(f_j) \geq 0$. Thus, g is a nonpositive on Δ function with nonnegative integral $I(g)$. At the same time g is not identically zero (since its vector of coefficients in the basis $f_1, \ldots, f_n$ is nonzero). This is the desired contradiction. □

Exercises to Lecture 2

Cones

Exercise 2.4.

2. Let $\mathbf{K}$ be a cone in $\mathbf{R}^n$ and $u \mapsto Au$ be a linear mapping from certain $\mathbf{R}^k$ to $\mathbf{R}^n$ with trivial null space and such that $\mathrm{Im}A \cap \mathrm{int}\mathbf{K} \neq \emptyset$. Prove that the inverse image of $\mathbf{K}$ under the mapping—the set

$$A^{-1}(\mathbf{K}) = \{u \mid Au \in \mathbf{K}\}$$

—is a cone in $\mathbf{R}^k$. Prove that the cone dual to $A^{-1}(\mathbf{K})$ is the image of $\mathbf{K}_*$ under the mapping $\lambda \mapsto A^T\lambda$:

$$(A^{-1}(\mathbf{K}))_* = \{A^T\lambda \mid \lambda \in \mathbf{K}_*\}.$$

Solution. The fact that $A^{-1}(K)$ is a cone is evident. Let us justify the announced description of the cone dual to $A^{-1}(\mathbf{K})$. If $\lambda \in \mathbf{K}_*$, then $A^T\lambda$ clearly belongs to $(A^{-1}(\mathbf{K}))_*$:

$$u \in A^{-1}(\mathbf{K}) \Rightarrow Au \in \mathbf{K} \Rightarrow \lambda^T(Au) = (A^T\lambda)^T u \geq 0.$$

Now let us prove the inverse implication: if $c \in (A^{-1}(\mathbf{K}))_*$, then $c = A^T\lambda$ for some $\lambda \in \mathbf{K}_*$. To this end consider the conic problem

$$\min_x \left\{c^T x \mid Ax \geq_{\mathbf{K}} 0\right\}.$$

The problem is strictly feasible and below bounded (why?), so that by the conic duality theorem the dual problem

$$\max_\lambda \left\{0^T\lambda \mid A^T\lambda = c, \lambda \geq_{\mathbf{K}_*} 0\right\}$$

is solvable. □

3. Let $\mathbf{K}$ be a cone in $\mathbf{R}^n$ and $y = Ax$ be a linear mapping from $\mathbf{R}^n$ onto $\mathbf{R}^N$ (i.e., the image of A is the entire $\mathbf{R}^N$). Assume $\mathrm{Null}(A) \bigcap \mathbf{K} = \{0\}$.

Prove then that the image of $\mathbf{K}$ under the mapping A—the set

$$A\mathbf{K} = \{Ax \mid x \in \mathbf{K}\}$$

—is a cone in $\mathbf{R}^N$.

Prove that the cone dual to $A(\mathbf{K})$ is

$$(A\mathbf{K})_* = \{\lambda \in \mathbf{R}^N \mid A^T\lambda \in \mathbf{K}_*\}.$$

Demonstrate by example that if in the above statement the assumption $\mathrm{Null}(A) \cap \mathbf{K} = \{0\}$ is weakened to $\mathrm{Null}(A) \cap \mathrm{int}\mathbf{K} = \emptyset$, then the image of $\mathbf{K}$ under the mapping A may happen to be nonclosed.

Solution. Let us temporarily set $B = A^T$ (note that B has trivial null space, since A is an onto mapping) and

$$\mathbf{L} = \{\lambda \in \mathbf{R}^N \mid B\lambda \in \mathbf{K}_*\}.$$

Let us prove that the image of B intersects the interior of $\mathbf{K}_*$. Indeed, otherwise we could separate the convex set $\mathrm{int}\mathbf{K}_*$ from the linear subspace $\mathrm{Im}\, B$: there would exist $x \neq 0$ such that

$$\inf_{\mu \in \mathrm{int}\mathbf{K}_*} x^T\mu \geq \sup_{\mu \in \mathrm{Im}\, B} x^T\mu,$$

whence $x \in (\mathbf{K}_*)_* = \mathbf{K}$ and $x \in (\mathrm{Im}\, B)^\perp = \mathrm{Null}(A)$, which is impossible.

It remains to apply to the cone $\mathbf{K}_*$ and the mapping B the rule on inverse image (rule 2). According to this rule, the set $\mathbf{L}$ is a cone, and its dual cone is the image of $(\mathbf{K}_*)_* = \mathbf{K}$ under the mapping $B^T = A$. Thus, $A(\mathbf{K})$ indeed is a cone, namely, the cone dual to $\mathbf{L}$, whence the cone dual to $A(\mathbf{K})$ is $\mathbf{L}$.

"Demonstrate by example…": When the 3D ice cream cone is projected onto its tangent plane, the projection is the open half-plane plus a single point on the boundary of this half-plane, which is not a closed set. □

Exercise 2.5. Let A be an $m \times n$ matrix of full column rank and $\mathbf{K}$ be a cone in $\mathbf{R}^m$.

1. Prove that at least one of the following facts always takes place:

   (i) There exists a nonzero $x \in \mathrm{Im}\, A$ which is $\geq_{\mathbf{K}} 0$.

   (ii) There exists a nonzero $\lambda \in \mathrm{Null}(A^T)$ which is $\geq_{\mathbf{K}_*} 0$.

   Geometrically: Given a primal-dual pair of cones $\mathbf{K}$, $\mathbf{K}_*$ and a pair L, $L^\perp$ of linear subspaces that are orthogonal complements of each other, we can find a nontrivial ray in the intersection $L \cap \mathbf{K}$ or in the intersection $L^\perp \cap \mathbf{K}_*$ or both.

2. Prove that the strict version of (ii) takes place (i.e., there exists $\lambda \in \mathrm{Null}(A^T)$ which is $>_{\mathbf{K}_*} 0$) if and only if (i) does not take place, and vice versa: the strict version of (i) takes place if and only if (ii) does not take place.

   Geometrically: If $\mathbf{K}$, $\mathbf{K}_*$ is a primal-dual pair of cones and L, $L^\perp$ are linear subspaces that are orthogonal complements of each other, then the intersection $L \cap \mathbf{K}$ is trivial (is the singleton $\{0\}$) if and only if the intersection $L^\perp \cap \mathrm{int}\mathbf{K}_*$ is nonempty.

Solution. 1. Assuming that (ii) does not take place, note that $A^T\mathbf{K}_*$ is a cone in $\mathbf{R}^n$ (Exercise 2.4), and the dual of this latter cone is the inverse image of $\mathbf{K}$ under the mapping A. Since the dual must contain nonzero vectors, (i) takes place.

2. Let $e >_{\mathbf{K}_*} 0$. Consider the conic problem

$$\max_{\lambda,t} \left\{ t \mid A^T\lambda = 0, \lambda - te \geq_{\mathbf{K}_*} 0 \right\}.$$

Note that this problem is strictly feasible, and the strict version of (ii) is equivalent to the fact that the optimal value in the problem is > 0. Thus, (ii) is not valid if and only if the optimal value in our strictly feasible maximization conic problem is ≤ 0. By the conic duality theorem this is the case if and only if the dual problem

$$\min_{z,\mu} \left\{ 0 \mid Az + \mu = 0, \mu^T e = 1, \mu \geq_{\mathbf{K}} 0 \right\}$$

is solvable with optimal value equal to 0, which clearly is the case if and only if the intersection of ImA and $\mathbf{K}$ is not the singleton $\{0\}$. □

Feasible and level sets of conic problems

Exercise 2.11. Let the problem

$$\min \left\{ c^T x \mid Ax - b \geq_{\mathbf{K}} 0 \right\} \tag{CP}$$

be feasible (A is of full column rank). Then the following properties are equivalent:

(i) The feasible set of the problem is bounded.

(ii) The set of primal slacks $K = \{y \geq_{\mathbf{K}} 0, y = Ax - b\}$ is bounded.[76]

(iii) Im $A \cap \mathbf{K} = \{0\}$.

(iv) The system of vector inequalities

$$A^T\lambda = 0, \lambda >_{\mathbf{K}_*} 0$$

is solvable.

Corollary. The property of (CP) to have bounded feasible set is independent of the particular value of b such that (CP) is feasible!

Solution. (i) $\Leftrightarrow$ (ii): This is an immediate consequence of **A**.

(ii) $\Rightarrow$ (iii). If there exists $0 \neq y = Ax \geq_{\mathbf{K}} 0$, then the set of primal slacks contains, along with any of its points $\bar{y}$, the entire ray $\{\bar{y} + ty \mid t > 0\}$, so that the set of primal slacks is unbounded. (Recall that it is nonempty—(CP) is feasible!) Thus, (iii) follows from (ii) (by contradiction).

(iii) $\Rightarrow$ (iv). See Exercise 2.5.2.

(iv) $\Rightarrow$ (ii). Let λ be given by (iv). For all primal slacks $y = Ax - b \in \mathbf{K}$ one has $\lambda^T y = \lambda^T[Ax - b] = (A^T\lambda)^T x - \lambda^T b = \lambda^T b$, and it remains to use the result of Exercise 2.3.2.

[76] Recall that we always assume that **A** holds!

Exercise 2.12. Let the problem

$$\min\left\{c^T x \mid Ax - b \geq_{\mathbf{K}} 0\right\} \tag{CP}$$

be feasible (A is of full column rank). Prove that the following two conditions are equivalent:

(i) (CP) has bounded level sets.

(ii) The dual problem

$$\max\left\{b^T\lambda \mid A^T\lambda = c, \lambda \geq_{\mathbf{K}_*} 0\right\}$$

is strictly feasible.

Corollary. The property of (CP) to have bounded level sets is independent of the particular value of b such that (CP) is feasible!

Solution. (i) $\Rightarrow$ (ii). Consider the linear vector inequality

$$\bar{A}x \equiv \begin{pmatrix} Ax \\ -c^T x \end{pmatrix} \geq_{\bar{\mathbf{K}}} 0,$$
$$\bar{\mathbf{K}} = \{(x,t) \mid x \geq_{\mathbf{K}} 0, t \geq 0\}.$$

If $\bar{x}$ is a solution of this inequality and x is a feasible solution to (CP), then the entire ray $\{x + t\bar{x} \mid t \geq 0\}$ is contained in the same level set of (CP) as x. Consequently, in the case of (i) the only solution to the inequality is the trivial solution $\bar{x} = 0$. In other words, the intersection of Im $\bar{A}$ with $\bar{\mathbf{K}}$ is trivial—{0}. Applying the result of Exercise 2.5.2, we conclude that the system

$$\bar{A}^T\begin{pmatrix}\lambda\\ \mu\end{pmatrix} \equiv A^T\lambda - \mu c = 0, \begin{pmatrix}\lambda\\ \mu\end{pmatrix} >_{\bar{\mathbf{K}}_*} 0$$

is solvable; if (λ, μ) solves the latter system, then $\lambda >_{\mathbf{K}_*} 0$ and $\mu > 0$, so that $\mu^{-1}\lambda$ is a strictly feasible solution to the dual problem.

(ii) $\Rightarrow$ (i). If $\lambda >_{\mathbf{K}_*} 0$ is feasible for the dual problem and x is feasible for (CP), then

$$\lambda^T(Ax - b) = (A^T\lambda)^T x - \lambda^T b = c^T x - \lambda^T b.$$

We conclude that if x runs through a given level set $\mathcal{L}$ of (CP), the corresponding slacks $y = Ax - b$ belong to a set of the form $\{y \geq_{\mathbf{K}} 0, \lambda^T y \leq \text{const}\}$. The sets of the latter type are bounded in view of the result of Exercise 2.3.2 (recall that $\lambda >_{\mathbf{K}_*} 0$). It remains to note that in view of **A** boundedness of the image of $\mathcal{L}$ under the mapping $x \mapsto Ax - b$ implies boundedness of $\mathcal{L}$. □

Exercises to Lecture 3

Optimal control in discrete time linear dynamic system. Consider a discrete time linear dynamic system

$$\begin{array}{rcl} x(t) & = & A(t)x(t-1) + B(t)u(t),\ t = 1, 2, \ldots, T, \\ x(0) & = & x_0. \end{array} \tag{S}$$

Here,

- t is the (discrete) time.
- $x(t) \in \mathbf{R}^l$ is the *state* vector: its value at instant t identifies the state of the controlled plant.
- $u(t) \in \mathbf{R}^k$ is the exogeneous input at time instant t; $\{u(t)\}_{t=1}^T$ is the *control*.
- For every $t = 1, \ldots, T$, $A(t)$ is a given $l \times l$, and $B(t)$ – a given $l \times k$ matrices.

A typical problem of optimal control associated with (S) is to minimize a given functional of the trajectory $x(\cdot)$ under given restrictions on the control. As a simple problem of this type, consider the optimization model

$$\min_x \left\{ c^T x(T) \mid \frac{1}{2} \sum_{t=1}^T u^T(t) Q(t) u(t) \le w \right\}, \tag{OC}$$

where $Q(t)$ are given positive definite symmetric matrices.

Exercise 3.1. 1. Use (S) to express $x(T)$ via the control and convert (OC) in a quadratically constrained problem with linear objective with respect to the u-variables.

Solution. From (S) it follows that

$$\begin{aligned}
x(1) &= A(1)x_0 + B(1)u(1); \\
x(2) &= A(2)x(1) + B(2)u(2) \\
&= A(2)A(1)x_0 + B(2)u(2) + A(2)B(1)u(1); \\
&\cdots \\
x(T) &= A(T)A(T-1)\ldots A(1)x_0 + \sum_{t=1}^T A(T)A(T-1)\ldots A(t+1)B(t)u(t) \\
&\equiv A(T)A(T-1)\ldots A(1)x_0 + \sum_{t=1}^T C(t)u(t), \\
&\quad C(t) = A(T)A(T-1)\ldots A(t+1)B(t).
\end{aligned}$$

Consequently, (OC) is equivalent to the problem

$$\min_{u(\cdot)} \left\{ \sum_{t=1}^T d_t^T u(t) \mid \frac{1}{2} \sum_{i=1}^T u^T(t) Q(t) u(t) \le w \right\} \quad [d_t = C^T(t)c]. \qquad \square \quad (*)$$

2. Convert the resulting problem to a conic quadratic program

Solution.

$$\begin{array}{ll}\text{minimize} & \displaystyle\sum_{t=1}^{T} d_t^T u(t)\\ \text{s.t.} & \\ & \begin{pmatrix} 2^{1/2}Q^{1/2}(t)u(t)\\ 1-s(t)\\ 1+s(t)\end{pmatrix} \ge_{\mathbf{L}^k} 0,\ t=1,\dots,T,\\ & \displaystyle\sum_{t=1}^{T} s(t) \le w;\end{array}$$

the design variables are $\{u(t)\in\mathbf{R}^k, s(t)\in\mathbf{R}\}_{t=1}^T$. □

3. Pass to the resulting problem to its dual and find the optimal solution to the latter problem.

Solution. The conic dual is

$$\begin{array}{ll}\text{maximize} & -w\omega - \displaystyle\sum_{t=1}^{T}[\mu(t)+\nu(t)]\\ \text{s.t.} & \\ & 2^{1/2}Q^{1/2}(t)\xi(t) = d_t,\ t=1,\dots,T\\ & \qquad [\Leftrightarrow \xi(t)=2^{-1/2}Q^{-1/2}(t)d_t],\\ & -\omega-\mu(t)+\nu(t) = 0,\ t=1,\dots,T\\ & \qquad [\Leftrightarrow \mu(t)=\nu(t)-\omega],\\ & \sqrt{\xi^T(t)\xi(t)+\mu^2(t)} \le \nu(t),\ t=1,\dots,T;\end{array}$$

the variables are $\{\xi(t)\in\mathbf{R}^k, \mu(t), \nu(t), \omega\in\mathbf{R}\}$. Equivalent reformulation of the dual problem is

$$\begin{array}{ll}\text{minimize} & w\omega + \displaystyle\sum_{t=1}^{T}[2\nu(t)-\omega]\\ \text{s.t.} & \\ & \sqrt{a_t^2+(\nu(t)-\omega)^2} \le \nu(t),\ t=1,\dots,T\\ & \qquad [a_t^2 = 2^{-1}d_t^T Q^{-1}(t)d_t],\end{array}$$

or, which is the same,

$$\begin{array}{ll}\text{minimize} & w\omega + \displaystyle\sum_{t=1}^{T}[2\nu(t)-\omega]\\ \text{s.t.} & \\ & \omega(2\nu(t)-\omega) \ge a_t^2,\ t=1,\dots,T,\end{array}$$

or, which is the same,

$$\min_{\omega}\left\{w\omega+\omega^{-1}\left(\sum_{t=1}^{T}a_t^2\right)\mid\ \omega>0\right\}.$$

It follows that the optimal solution to the dual problem is

$$\begin{array}{rcl}\omega_* &=& \sqrt{w^{-1}\left(\sum_{i=1}^{T}a_t^2\right)};\\ \nu_*(t) &=& \frac{1}{2}\left[a_t^2\omega_*^{-1}+\omega_*\right],\ t=1,\ldots,T;\\ \mu_*(t) &=& \nu_*(t)-\omega_*\\ &=& \frac{1}{2}\left[a_t^2\omega_*^{-1}-\omega_*\right],\ t=1,\ldots,T.\end{array}\qquad\square$$

Stable grasp. Recall that the stable grasp analysis problem is to check whether the system of constraints

$$\begin{array}{rcl}\|F^i\|_2 &\le& \mu(f^i)^Tv^i,\ i=1,\ldots,N,\\ (v^i)^TF^i &=& 0,\ i=1,\ldots,N,\\ \sum_{i=1}^{N}(f^i+F^i)+F^{\text{ext}} &=& 0,\\ \sum_{i=1}^{N}p^i\times(f^i+F^i)+T^{\text{ext}} &=& 0\end{array}\qquad\text{(SG)}$$

in the 3D vector variables F^i is or is not solvable. Here the data are given by a number of 3D vectors, namely,

- vectors v^i—unit inward normals to the surface of the body at the contact points;
- contact points p^i;
- vectors f^i—contact forces;
- vectors F^{ext} and T^{ext} of the external force and torque, respectively.

$\mu>0$ is a given friction coefficient; we assume that $f_i^Tv^i>0\ \forall i$.

Exercise 3.6. 1. Regarding (SG) as the system of constraints of a maximization program with trivial objective and applying the technique from section 2.5, build the dual problem.

Solution.

$$\begin{array}{ll}\text{minimize} & \displaystyle\sum_{i=1}^N \mu[(f^i)^T v^i]\phi_i - \left[\sum_{i=1}^N f^i + F^{\text{ext}}\right]^T \Phi - \left[\sum_{i=1}^N p^i \times f^i + T^{\text{ext}}\right]^T \Psi \\ \text{s.t.} & \Phi_i + \sigma_i v^i + \Phi - p^i \times \Psi = 0,\ i = 1, \ldots, N \\ & \|\Phi_i\|_2 \le \phi_i,\ i = 1, \ldots, N, \\ & \qquad [\Phi, \Phi_i, \Psi \in \mathbf{R}^3, \sigma_i, \phi_i \in \mathbf{R}]\end{array}$$

$$\Updownarrow$$

$$\min_{\sum_i \in \mathbf{R}, \Psi, \Phi \in \mathbf{R}^3} \left\{\sum_{i=1}^N \mu[(f^i)^T v^i]\|p^i \times \Psi - \Phi - \sigma_i v^i\|_2 - F^T\Phi - T^T\Psi\right\},$$

$$F = \sum_{i=1}^N f^i + F^{\text{ext}},\ \ T = \sum_{i=1}^N p^i \times f^i + T^{\text{ext}}. \qquad \square$$

Trusses. We are about to process the multiload TTD problem 3.4.1, which we write as (see (3.4.57))

$$\begin{array}{lrcl}\text{minimize} & \tau & & \\ \text{s.t.} & s_{ij}^2 & \le & 4t_i r_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, k, \\ & \displaystyle\sum_{i=1}^n r_{ij} & \le & \tfrac{1}{2}\tau,\ j = 1, \ldots, k, \\ & \displaystyle\sum_{i=1}^n t_i & \le & w, \\ & \displaystyle\sum_{i=1}^n s_{ij} b_i & = & f_j,\ j = 1, \ldots, k, \\ & t_i, r_{ij} & \ge & 0,\ i = 1, \ldots, n,\ j = 1, \ldots, k;\end{array} \qquad \text{(Pr)}$$

the design variables are $s_{ij}, r_{ij}, t_i, \tau$.

We assume that

- the ground structure $(n, m, b_1, \ldots, b_n)$ is such that the matrix $\sum_{i=1}^n b_i b_i^T$ is positive definite; and
- the loads of interest $f_1, \ldots, f_k$ are nonzero, and the material resource w is positive.

Exercise 3.7. 1. Applying the technique from section 2.5, build the problem (Dl) dual to the problem (Pr).

What is the design dimension of (Pr)? of (Dl)?

Solution.

$$
\begin{array}{lrcl}
\text{maximize} & -w\rho - \displaystyle\sum_{j=1}^{k} f_j^T v_j & & \\
\text{s.t.} & \alpha_{ij} + b_i^T v_j & = & 0,\ i = 1, \dots, n,\ j = 1, \dots, k; \\
& \gamma_{ij} - \beta_{ij} - \delta_j & = & 0,\ i = 1, \dots, n,\ j = 1, \dots, k; \\
& \displaystyle\sum_{j=1}^{k} [\beta_{ij} + \gamma_{ij}] - \rho & = & 0,\ i = 1, \dots, n; \\
& \frac{1}{2} \displaystyle\sum_{j=1}^{n} \delta_j & = & 1, \\
& \sqrt{\alpha_{ij}^2 + \beta_{ij}^2} & \le & \gamma_{ij},\ i = 1, \dots, n,\ j = 1, \dots, k; \\
& \delta_j & \ge & 0,\ j = 1, \dots, k; \\
& \rho & \ge & 0 \\
& & & [\alpha_{ij}, \beta_{ij}, \gamma_{ij}, \delta_j, \rho \in \mathbf{R}, v_j \in \mathbf{R}^m]
\end{array}
$$

$$\Updownarrow$$

$$
\begin{array}{lrcl}
\text{minimize} & w\rho + \displaystyle\sum_{j=1}^{k} f_j^T v_j & & \\
\text{s.t.} & \displaystyle\sum_{j=1}^{k} \lambda_j^{-1} (b_i^T v_j)^2 & \le & 2\rho,\ i = 1, \dots, n \\
& & & [\lambda_j = \delta_j/2]; \\
& \displaystyle\sum_{j=1}^{k} \lambda_j & = & 1; \\
& \lambda_j & \ge & 0,\ j = 1, \dots, k; \\
& \rho & \ge & 0 \\
& & & [\rho, \lambda_j \in \mathbf{R}, v_j \in \mathbf{R}^m].
\end{array}
$$

The design dimension of (Pr) is $2nk + n + 1$. The design dimension of (Dl) is $mk + k + 1$. □

Exercise 3.8. Let us fix a ground structure $(n, m, b_1, \dots, b_n)$ and a material resource w, and let $\mathcal{F}$ be a finite set of loads.

1. Assume that $\mathcal{F}_j \in \mathcal{F}$, $j = 1, \dots, k$, are subsets of $\mathcal{F}$ with $\cup_{j=1}^{k} \mathcal{F}_j = \mathcal{F}$. Let μ_j be the optimal value in the multiload TTD problem with the set of loads $\mathcal{F}_j$ and μ be the optimal value in the multiload TTD problem with the set of loads $\mathcal{F}$. Is it possible that $\mu > \sum_{j=1}^{k} \mu_j$?

Solution. The compliance $\mathrm{Compl}_f(t)$ clearly is nonincreasing in t and $\mathrm{Compl}_f(\theta t) = \theta^{-1}\mathrm{Compl}_f(t)$ for positive θ. Let t^j be an optimal solution to the multiload TTD problem with the set of loads $\mathcal{F}_j$, and let θ_j be positive reals with sum 1. The truss $t_\theta = \sum_{j=1}^k \theta_j t^j$ clearly satisfies the resource constraint, while for every $f \in \mathcal{F}_j$ one has

$$\mathrm{Compl}_f(t_\theta) \le \mathrm{Compl}_f(\theta_j t^j) = \theta_j^{-1}\mathrm{Compl}_f(t^j) \le \theta_j^{-1}\mu_j.$$

Setting $\theta_j = \mu_j(\sum_{l=1}^k \mu_l)^{-1}$, we get $\mathrm{Compl}_f(t_\theta) \le \sum_{j=1}^m \mu_j \; \forall f \in \cup_{j=1}^k \mathcal{F}_j = \mathcal{F}$. Thus, $\mu \le \sum_{j=1}^k \mu_j$. □

2. Assume that the ground structure includes $n = 1998$ tentative bars and that you are given a set $\mathcal{F}$ of $N = 1998$ loads. It is known that for every subset $\mathcal{F}'$ of $\mathcal{F}$ made up of no more than 999 loads, the optimal value in the multiload TTD problem, the set of loading scenarios being $\mathcal{F}'$, does not exceed 1. What can be said about the optimal value in the multiload TTD problem with the set of scenarios $\mathcal{F}$?

Solution. It does not exceed 2 (by the previous exercise). □

Answer a similar question in the case when $\mathcal{F}$ comprises $N' = 19980$ loads.

Solution. The optimal value still does not exceed 2. Indeed, let T_f, $t \in \mathcal{F}$, be the set of all trusses of given volume with compliances with respect to f not exceeding 2. Then T_f is a convex subset in 1997-dimensional affine plane (cut off the 1998-dimensional space of trusses by the linear equation "volume of the truss is given"). By the previous answer for every subset of $\mathcal{F}$ comprising 1998 loads, there exists a truss of given volume with compliance with respect to the subset not exceeding 2, i.e., every 1998 convex set from the family $\{T_f\}_{f\in\mathcal{F}}$ has a point in common. It follows, by the Helley theorem, that all sets from the family have a point in common, i.e., there exists a truss with compliance with respect to every load from $\mathcal{F}$ not exceeding 2. □

Does conic quadratic programming exist? Let $\epsilon > 0$ and a positive integer n be given. We intend to build a polyhedral ϵ-approximation of the Lorentz cone $\mathbf{L}^{n+1}$. Without loss of generality we may assume that n is an integer power of 2: $n = 2^\kappa$, $\kappa \in \mathbf{N}$.

Tower of variables. The first step of our construction is quite straightforward: we introduce extra variables to represent a conic quadratic constraint

$$\sqrt{y_1^2 + \cdots + y_n^2} \le t \tag{CQI}$$

of dimension $n + 1$ by a system of conic quadratic constraints of dimension three each. Namely, let us call our original y-variables variables of generation 0 and let us split them into pairs $(y_1, y_2), \ldots, (y_{n-1}, y_n)$. We associate with every one of these pairs its successor—an additional variable of generation 1. We split the resulting $2^{\kappa-1}$ variables of generation 1

into pairs and associate with every pair its successor—an additional variable of generation 2, and so on. After $\kappa-1$ steps we end up with two variables of the generation $\kappa-1$. Finally, the only variable of generation κ is the variable t from (CQI).

To introduce convenient notation, let us denote by y_i^ℓ the ith variable of generation ℓ, so that $y_1^0,\dots,y_n^0$ are our original y-variables $y_1,\dots,y_n$, $y_1^\kappa\equiv t$ is the original t-variable, and the parents of y_i^ℓ are the variables $y_{2i-1}^{\ell-1}$, $y_{2i}^{\ell-1}$.

Note that the total number of all variables in the tower of variables we end up with is $2n-1$.

It is clear that the system of constraints

$$\sqrt{[y_{2i-1}^{\ell-1}]^2+[y_{2i}^{\ell-1}]^2}\le y_i^\ell,\ \ i=1,\dots,2^{\kappa-\ell},\ \ \ell=1,\dots,\kappa, \tag{1}$$

is a representation of (CQI) in the sense that a collection $(y_1^0\equiv y_1,\dots,y_n^0\equiv y_n,y_1^\kappa\equiv t)$ can be extended to a solution of (1) if and only if (y,t) solves (CQI). Moreover, let $\Pi_\ell(x_1,x_2,x_3,u^\ell)$ be polyhedral ϵ_ℓ-approximations of the cone

$$\mathbf{L}^3=\left\{(x_1,x_2,x_3)\mid \sqrt{x_1^2+x_2^2}\le x_3\right\},$$

$\ell=1,\dots,\kappa$. Consider the system of linear constraints in variables y_i^ℓ, u_i^ℓ:

$$\Pi_\ell(y_{2i-1}^{\ell-1},y_{2i}^{\ell-1},y_i^\ell,u_i^\ell)\ge 0,\ \ i=1,\dots,2^{\kappa-\ell},\ \ \ell=1,\dots,\kappa. \tag{2}$$

Writing this system of linear constraints as $\Pi(y,t,u)\ge 0$, where Π is linear in its arguments, $y=(y_1^0,\dots,y_n^0)$, $t=y_1^\kappa$, and u is the collection of all u_i^ℓ, $\ell=1,\dots,\kappa$, and all y_i^ℓ, $\ell=1,\dots,\kappa-1$, we immediately conclude that Π is a polyhedral ϵ-approximation of $\mathbf{L}^{n+1}$ with

$$1+\epsilon=\prod_{\ell=1}^{\kappa}(1+\epsilon_\ell). \tag{3}$$

In view of this observation, we may focus on building polyhedral approximations of the Lorentz cone $\mathbf{L}^3$.

Polyhedral approximation of $\mathbf{L}^3$. The approximation we intend to use is given by the system of linear inequalities, as follows (positive integer ν is the parameter of the construction):

$$\begin{array}{lll}
\text{(a)} & \left\{\begin{array}{rcl}\xi^0 &\ge& |x_1|,\\ \eta^0 &\ge& |x_2|,\end{array}\right. & \\
\text{(b)} & \left\{\begin{array}{rcl}\xi^j &=& \cos\left(\frac{\pi}{2^{j+1}}\right)\xi^{j-1}+\sin\left(\frac{\pi}{2^{j+1}}\right)\eta^{j-1},\\ \eta^j &\ge& \left|-\sin\left(\frac{\pi}{2^{j+1}}\right)\xi^{j-1}+\cos\left(\frac{\pi}{2^{j+1}}\right)\eta^{j-1}\right|\end{array}\right. , & j=1,\dots,\nu,\\
\text{(c)} & \left\{\begin{array}{rcl}\xi^\nu &\le& x_3,\\ \eta^\nu &\le& \operatorname{tg}\left(\frac{\pi}{2^{\nu+1}}\right)\xi^\nu.\end{array}\right. &
\end{array} \tag{4}$$

Note that (4) can be straightforwardly written as a system of linear homogeneous inequalities $\Pi^{(\nu)}(x_1, x_2, x_3, u) \geq 0$, where u is the collection of $2(\nu+1)$ variables ξ^j, η^i, $j = 0, \ldots, \nu$.

Proposition 1. *$\Pi^{(\nu)}$ is a polyhedral $\delta(\nu)$-approximation of $\mathbf{L}^3 = \{(x_1, x_2, x_3) \mid \sqrt{x_1^2 + x_2^2} \leq x_3\}$ with*

$$\delta(\nu) = \frac{1}{\cos\left(\frac{\pi}{2^{\nu+1}}\right)} - 1. \tag{5}$$

Proof. We should prove that

(i) If $(x_1, x_2, x_3) \in \mathbf{L}^3$, then the triple (x_1, x_2, x_3) can be extended to a solution to (4).

(ii) If a triple (x_1, x_2, x_3) can be extended to a solution to (4), then $\|(x_1, x_2)\|_2 \leq (1 + \delta(\nu))x_3$.

(i): Given $(x_1, x_2, x_3) \in \mathbf{L}^3$, let us set $\xi^0 = |x_1|$, $\eta^0 = |x_2|$, thus ensuring (4)(a). Note that $\|(\xi^0, \eta^0)\|_2 = \|(x_1, x_2)\|_2$ and that the point $P^0 = (\xi^0, \eta^0)$ belongs to the first quadrant.

Now, for $j = 1, \ldots, \nu$ let us set

$$\begin{aligned} \xi^j &= \cos\left(\tfrac{\pi}{2^{j+1}}\right)\xi^{j-1} + \sin\left(\tfrac{\pi}{2^{j+1}}\right)\eta^{j-1}, \\ \eta^j &= \left|-\sin\left(\tfrac{\pi}{2^{j+1}}\right)\xi^{j-1} + \cos\left(\tfrac{\pi}{2^{j+1}}\right)\eta^{j-1}\right|, \end{aligned}$$

thus ensuring (4)(b), and let $P^j = (\xi^j, \eta^j)$. The point P^i is obtained from P^{j-1} by the following construction: We rotate clockwise P^{j-1} by the angle $\phi_j = \frac{\pi}{2^{j+1}}$, thus getting a point Q^{j-1}. If this point is in the upper half-plane, we set $P^j = Q^{j-1}$; otherwise, P^j is the reflection of Q^{j-1} with respect to the x-axis. From this description it is clear that

(I) $\|P^j\|_2 = \|P^{j-1}\|_2$, so that all vectors P^j are of the same Euclidean norm as P^0, i.e., of the norm $\|(x_1, x_2)\|_2$;

(II) since the point P^0 is in the first quadrant, the point Q^0 is in the angle $-\frac{\pi}{4} \leq \arg(P) \leq \frac{\pi}{4}$, so that P^1 is in the angle $0 \leq \arg(P) \leq \frac{\pi}{4}$. The latter relation, in turn, implies that Q^1 is in the angle $-\frac{\pi}{8} \leq \arg(P) \leq \frac{\pi}{8}$, whence P^2 is in the angle $0 \leq \arg(P) \leq \frac{\pi}{8}$. Similarly, P^3 is in the angle $0 \leq \arg(P) \leq \frac{\pi}{16}$, and so on: P^j is in the angle $0 \leq \arg(P) \leq \frac{\pi}{2^{j+1}}$.

By (I), $\xi^\nu \leq \|P^\nu\|_2 = \|(x_1, x_2)\|_2 \leq x_3$, so that the first inequality in (4)(c) is satisfied. By (II), P^ν is in the angle $0 \leq \arg(P) \leq \frac{\pi}{2^{\nu+1}}$, so that the second inequality in (4)(c) also is satisfied. We have extended a point from $\mathbf{L}^3$ to a solution to (4).

(ii): Let (x_1, x_2, x_3) be extended to a solution $(x_1, x_2, x_3, \{\xi^j, \eta^j\}_{j=0}^\nu)$ to (4). Let us set $P^j = (\xi^j, \eta^j)$. From (4)(a), (b) it follows that all vectors P^j are nonnegative. We have $\|P^0\|_2 \geq \|(x_1, x_2)\|_2$ by (4)(a). Now, (4)(b) says that the coordinates of P^j are $\geq$ absolute values of the coordinates of P^{j-1} taken in certain orthonormal system of coordinates, so that $\|P^j\|_2 \geq \|P^{j-1}\|_2$. Thus, $\|P^\nu\|_2 \geq \|(x_1, x_2)^T\|_2$. On the other hand, by (4)(c) one has $\|P^\nu\|_2 \leq \frac{1}{\cos(\frac{\pi}{2^{\nu+1}})}\xi^\nu \leq \frac{1}{\cos(\frac{\pi}{2^{\nu+1}})}x_3$, so that $\|(x_1, x_2)^T\|_2 \leq \delta(\nu)x_3$, as claimed. □

Specifying in (2) the mappings $\Pi_\ell(\cdot)$ as $\Pi^{(\nu_\ell)}(\cdot)$, we conclude that for every collection of positive integers $\nu_1, \dots, \nu_\kappa$ one can point out a polyhedral β-approximation $\Pi_{\nu_1,\dots,\nu_\kappa}(y,t,u)$ of $\mathbf{L}^n$, $n = 2^\kappa$:

$$
\begin{array}{ll}
(a_{\ell,i}) & \left\{ \begin{array}{rcl} \xi^0_{\ell,i} & \geq & |y^{\ell-1}_{2i-1}|, \\ \eta^0_{\ell,i} & \geq & |y^{\ell-1}_{2i}|, \end{array} \right. \\
(b_{\ell,i}) & \left\{ \begin{array}{rcl} \xi^j_{\ell,i} & = & \cos\left(\frac{\pi}{2^{j+1}}\right)\xi^{j-1}_{\ell,i} + \sin\left(\frac{\pi}{2^{j+1}}\right)\eta^{j-1}_{\ell,i}, \\ \eta^j_{\ell,i} & \geq & \left| -\sin\left(\frac{\pi}{2^{j+1}}\right)\xi^{j-1}_{\ell,i} + \cos\left(\frac{\pi}{2^{j+1}}\right)\eta^{j-1}_{\ell,i} \right|, \end{array} \right. \quad j = 1, \dots, \nu_\ell, \\
(c_{\ell,i}) & \left\{ \begin{array}{rcl} \xi^{\nu_\ell}_{\ell,i} & \leq & y^\ell_i, \\ \eta^{\nu_\ell}_{\ell,i} & \leq & \mathrm{tg}\left(\frac{\pi}{2^{\nu_\ell+1}}\right)\xi^{\nu_\ell}_{\ell,i}, \end{array} \right. \\
 & \qquad i = 1, \dots, 2^{\kappa-\ell}, \ \ell = 1, \dots, \kappa.
\end{array}
\tag{6}
$$

The approximation possesses the following properties:

1. The dimension of the u-vector (comprising all variables in (6) except $y_i = y^0_i$ and $t = y^\kappa_1$) is
$$p(n, \nu_1, \dots, \nu_\kappa) \leq n + O(1)\sum_{\ell=1}^{\kappa} 2^{\kappa-\ell}\nu_\ell.$$
2. The image dimension of $\Pi_{\nu_1,\dots,\nu_\kappa}(\cdot)$ (i.e., the number of linear inequalities plus twice the number of linear equations in (6)) is
$$q(n, \nu_1, \dots, \nu_\kappa) \leq O(1)\sum_{\ell=1}^{\kappa} 2^{\kappa-\ell}\nu_\ell.$$
3. The quality β of the approximation is
$$\beta = \beta(n; \nu_1, \dots, \nu_\kappa) = \prod_{\ell=1}^{\kappa} \frac{1}{\cos\left(\frac{\pi}{2^{\nu_\ell+1}}\right)} - 1.$$

Back to the general case. Given $\epsilon \in (0, 1]$ and setting

$$\nu_\ell = \left\lfloor O(1)\ell \ln\frac{2}{\epsilon} \right\rfloor, \ \ell = 1, \dots, \kappa,$$

with properly chosen absolute constant $O(1)$, we ensure that

$$
\begin{array}{rcl}
\beta(\nu_1, \dots, \nu_\kappa) & \leq & \epsilon, \\
p(n, \nu_1, \dots, \nu_\kappa) & \leq & O(1)n\ln\frac{2}{\epsilon}, \\
q(n, \nu_1, \dots, \nu_\kappa) & \leq & O(1)n\ln\frac{2}{\epsilon},
\end{array}
$$

as required.

Exercises to Lecture 4

Positive semidefiniteness, eigenvalues, and $\succeq$-ordering

Exercise 4.2. Diagonal-dominant matrices. Let $A = [a_{ij}]_{i,j=1}^m$ be a symmetric matrix satisfying the relation

$$a_{ii} \geq \sum_{j \neq i} |a_{ij}|, \ \ i = 1, \ldots, m.$$

Prove that A is positive semidefinite.

Solution. Let e be an eigenvector of A and λ be the corresponding eigenvalue. We may assume that the largest, in absolute value, of coordinates of e is equal to 1. Let i be the index of this coordinate; then

$$\lambda = a_{ii} + \sum_{j \neq i} a_{ij} e_j \geq a_{ii} - \sum_{j \neq i} |a_{ij}| \geq 0.$$

Thus, all eigenvalues of A are nonnegative, so that A is positive semidefinite. □

Variational description of eigenvalues

Exercise 4.8. Let f_* be a closed convex function with the domain $\mathrm{Dom} f_* \subset \mathbf{R}_+$, and let f be the Legendre transformation of f_*. Then for every pair of symmetric matrices X, Y of the same size with the spectrum of X belonging to $\mathrm{Dom} f$ and the spectrum of Y belonging to $\mathrm{Dom} f_*$ one has

$$\lambda(f(X)) \geq \lambda\left(Y^{1/2} X Y^{1/2} - f_*(Y)\right). \qquad (*)$$

Solution. By continuity reasons, it suffices to prove $(*)$ in the case of $Y \succ 0$ (why?). Let m be the size of X, let $k \in \{1, \ldots, m\}$, let $\mathcal{E}_k$ be the family of linear subspaces of $\mathbf{R}^m$ of codimension $k-1$, and let $E \in \mathcal{E}_k$ be such that

$$e \in E \Rightarrow e^T X e \leq \lambda_k(X) e^T e.$$

(Such an E exists by the variational characterization of eigenvalues as applied to X.) Let also $F = Y^{-1/2} E$; the codimension of F, same as the one of E, is $k-1$. Finally, let $g_1, .., g_m$ be an orthonormal system of eigenvectors of Y, so that $Y g_j = \lambda_j(Y) g_j$. We have

$$
\begin{aligned}
h \in F, h^T h = 1 \quad &\Rightarrow \\
h^T[Y^{1/2}XY^{1/2} - f_*(Y)]h \quad &= \quad (\underbrace{Y^{1/2}h}_{\in E})^T X (Y^{1/2}h) - \sum_{j=1}^m f_*(\lambda_j(Y))(g_j^T h)^2 \\
&\leq \quad \lambda_k(X)(Y^{1/2}h)^T(Y^{1/2}h) - \sum_{j=1}^m f_*(\lambda_j(Y))(g_j^T h)^2 \\
&\quad \quad [\text{since } Y^{1/2}h \in E] \\
&= \quad \lambda_k(X)(h^T Y h) - \sum_{j=1}^m f_*(\lambda_j(Y))(g_j^T h)^2 \\
&= \quad \lambda_k(X)\sum_{j=1}^m \lambda_j(Y)(g_j^T h)^2 - \sum_{j=1}^m f_*(\lambda_j(Y))(g_j^T h)^2 \\
&= \quad \sum_{j=1}^m [\lambda_k(X)\lambda_j(Y) - f_*(\lambda_j(Y))](g_j^T h)^2 \\
&\leq \quad \sum_{j=1}^m f(\lambda_k(X))(g_j^T h)^2 \\
&\quad \quad [\text{since } f = (f_*)_*] \\
&= \quad f(\lambda_k(X)) \\
&\quad \quad [\text{since } \sum_j (g_j^T h)^2 = h^T h = 1] \\
&= \quad \lambda_k(f(X)) \\
&\quad \quad [\text{since } f(\cdot) \text{ is nonincreasing due to } \mathrm{Dom} f_* \subset \mathbf{R}_+].
\end{aligned}
$$

We see that there exists $F \in \mathcal{E}_k$ such that

$$\max_{h \in F : h^T h = 1} h^T[Y^{1/2}XY^{1/2} - f_*(Y)]h \leq \lambda_k(f(X)).$$

From variational characterization of eigenvalues it follows that

$$\lambda_k(Y^{1/2}XY^{1/2} - f_*(Y)) \leq \lambda_k(f(X)). \qquad \square$$

Exercise 4.10. 4. (trace inequality) Prove that whenever $A, B \in \mathbf{S}^m$, one has

$$\lambda^T(A)\lambda(B) \geq \mathrm{Tr}(AB).$$

Solution. Denote $\lambda = \lambda(A)$, and let $A = V^T \mathrm{Diag}(\lambda) V$ be the spectral decomposition of A. Setting $\hat{B} = VBV^T$, note that $\lambda(\hat{B}) = \lambda(B)$ and $\mathrm{Tr}(AB) = \mathrm{Tr}(\mathrm{Diag}(\lambda)\hat{B})$. Thus, it suffices to prove the trace inequality in the particular case when A is a diagonal matrix with the diagonal $\lambda = \lambda(A)$. Denoting by μ the diagonal of B and setting

$$\sigma^0 = 0; \sigma^k = \sum_{i=1}^k \mu_i, \ k = 1, \ldots, m,$$

we have

$$\begin{aligned}
\mathrm{Tr}(AB) &= \sum_{i=1}^{m} \lambda_i \mu_i \\
&= \sum_{i=1}^{m} \lambda_i (\sigma^i - \sigma^{i-1}) \\
&= -\lambda_1 \sigma^0 + \sum_{i=1}^{m-1} (\lambda_i - \lambda_{i+1}) \sigma^i + \lambda_m \sigma^m \\
&= \sum_{i=1}^{m-1} (\lambda_i - \lambda_{i+1}) \sigma^i + \lambda_m \mathrm{Tr}(B) \\
&\leq \sum_{i=1}^{m-1} (\lambda_i - \lambda_{i+1}) \sum_{j=1}^{i} \lambda_j(B) + \lambda_m \sum_{j=1}^{m} \lambda_j(B) \\
&\quad \text{[since } \lambda_i \geq \lambda_{i+1} \text{ and in view of Exercise 4.10.3]} \\
&= \sum_{i=1}^{m} \lambda_i \lambda_i(B) \\
&= \lambda^T(A) \lambda(B). \qquad \square
\end{aligned}$$

Exercise 4.12. 3. Let X be a symmetric $n \times n$ matrix partitioned into blocks in a symmetric, with respect to the diagonal, fashion,

$$X = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{12}^T & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1m}^T & X_{2m}^T & \dots & X_{mm} \end{pmatrix},$$

so that the blocks X_{ii} are square. Let also $g : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ be convex function on the real line which is finite on the set of eigenvalues of X, and let $\mathcal{F}_n \subset \mathbf{S}^n$ be the set of all $n \times n$ symmetric matrices with all eigenvalues belonging to the domain of g. Assume that the mapping

$$Y \mapsto g(Y) : \mathcal{F}_n \to \mathbf{S}^n$$

is $\succeq$-convex:

$$g(\lambda' Y' + \lambda'' Y'') \preceq \lambda' g(Y') + \lambda'' g(Y'') \quad \forall (Y', Y'' \in \mathcal{F}_n, \lambda', \lambda'' \geq 0, \lambda' + \lambda'' = 1).$$

Prove that

$$(g(X))_{ii} \succeq g(X_{ii}), \ i = 1, \dots, m,$$

where the partition of $g(X)$ into the blocks $(g(X))_{ij}$ is identical to the partition of X into the blocks X_{ij}.

Solution. Let $\epsilon = (\epsilon_1, \dots, \epsilon_m)$ with $\epsilon_j = \pm 1$, and let

$$U_\epsilon = \begin{pmatrix} \epsilon_1 I_{n_1} & & & \\ & \epsilon_2 I_{n_2} & & \\ & & \ddots & \\ & & & \epsilon_m I_{n_m} \end{pmatrix},$$

where n_i is the row size of X_{ii}. Then U_ϵ are orthogonal matrices and one clearly has

$$D(X) \equiv \begin{pmatrix} X_{11} & & & \\ & X_{22} & & \\ & & \ddots & \\ & & & X_{mm} \end{pmatrix} = \frac{1}{2^m} \sum_{\epsilon:\epsilon_i=\pm 1,\, i=1,\dots,m} U_\epsilon^T X U_\epsilon.$$

We have

$$\begin{array}{rl} D(X) = \frac{1}{2^m} \displaystyle\sum_{\epsilon:\epsilon_i=\pm 1,\, i=1,\dots,m} U_\epsilon^T X U_\epsilon & \\ \Downarrow & \\ g(D(X)) \preceq \frac{1}{2^m} \displaystyle\sum_{\epsilon:\epsilon_i=\pm 1,\, i=1,\dots,m} g(U_\epsilon^T X U_\epsilon) & [\text{since } g \text{ is } \succeq\text{-convex}] \\ \Updownarrow & \\ g(D(X)) \preceq \frac{1}{2^m} \displaystyle\sum_{\epsilon:\epsilon_i=\pm 1,\, i=1,\dots,m} U_\epsilon^T g(X) U_\epsilon & \\ \Downarrow & \\ g(D(X)) \preceq D(g(X)). \quad \square & \end{array}$$

Cauchy's inequality for matrices

Exercise 4.19. 1. Denote $P = \left(\sum_i X_i^T X_i\right)^{1/2}$, $Q = \sum_i Y_i^T Y_i$ ($X_i, Y_i \in \mathbf{M}^{p,q}$). We should prove that

$$\sigma\left(\sum_i X_i^T Y_i\right) \le \lambda(P)\|\lambda(Q)\|_\infty^{1/2}$$

or, which is the same,

$$\sigma\left(\sum_i Y_i^T X_i\right) \le \lambda(P)\|\lambda(Q)\|_\infty^{1/2}.$$

By the variational description of singular values, it suffices to prove that for every $k = 1, 2, \dots, p$ there exists a subspace $L_k \subset \mathbf{R}^q$ of codimension $k-1$ such that

$$\forall \xi \in L_k: \quad \left\| \left(\sum_i Y_i^T X_i\right) \xi \right\|_2 \le \|\xi\|_2 \lambda_k(P)\|\lambda(Q)\|_\infty^{1/2}. \qquad (*)$$

Let $e_1, \ldots, e_q$ be the orthonormal eigenbasis of P: $Pe_i = \lambda_i(P)e_i$, and let L_k be the linear span of $e_k, e_{k+1}, \ldots, e_q$. For $\xi \in L_k$ one has

$$\begin{aligned}|\eta^T \left(\sum_i Y_i^T X_i\right) \xi| &\le \sum_i \|Y_i\eta\|_2 \|X_i\xi\|_2 \le \sqrt{\sum_i \|Y_i\eta\|_2^2}\sqrt{\sum_i \|X_i\xi\|_2^2} \\ &= \sqrt{\eta^T \left(\sum_i Y_i^T Y_i\right)\eta}\sqrt{\xi^T \left(\sum_i X_i^T X_i\right)\xi} \\ &\le \left(\|\lambda(Q)\|_\infty \|\eta\|_2^2\right)^{1/2} \left(\lambda_k^2(P)\|\xi\|_2^2\right)^{1/2} = \|\lambda(Q)\|_\infty^{1/2}\lambda_k(P)\|\eta\|_2\|\xi\|_2,\end{aligned}$$

whence

$$\left\|\left(\sum_i Y_i^T X_i\right)\xi\right\|_2 = \max_{\eta:\|\eta\|_2=1} \eta^T \left(\sum_i Y_i^T X_i\right)\xi \le \lambda_k(P)\|\lambda(Q)\|_\infty^{1/2}\|\xi\|_2,$$

as required in (*).

To make (*) equality, assume that $P \succ 0$ (the case of singular P is left to the reader), and let $Y_i = X_i P^{-1}$. Then

$$\sum_i Y_i^T Y_i = P^{-1}\left(\sum_i X_i^T X_i\right) P^{-1} = I$$

and

$$\sum_i X_i^T Y_i = \left(\sum_i X_i^T X_i\right) P^{-1} = P,$$

so that (*) becomes equality.

(1)⇒(2): it suffices to prove that if $A \in \mathbf{M}^{p,p}$, then

$$|\mathrm{Tr}(A)| \le \|\sigma(A)\|_1. \qquad (**)$$

Indeed, we have $A = U\Lambda V$, where U, V are orthogonal matrices and Λ is a diagonal matrix with the diagonal $\sigma(A)$. Denoting by e_i the standard basic orths in $\mathbf{R}^p$, we have

$$\begin{aligned}|\mathrm{Tr}(A)| &= |\mathrm{Tr}(U^T AU)| = |\mathrm{Tr}(\Lambda(VU))| \\ &= \left|\sum_i e_i^T \Lambda(VU)e_i\right| \le \sum_i |\sigma_i(A)e_i^T(VU)e_i| \le \sum_i \sigma_i(A),\end{aligned}$$

as required in (**).

Exercise 4.20. The true statements are 2 and 3.

Part 2 is an immediate consequence of the following.

Lemma 1. *Let $A_i \in \mathbf{S}_+^n$, $i = 1, \ldots, m$, and let $\alpha > 1$. Then*

$$\lambda_j\left(\left(\sum_{i=1}^m A_i^\alpha\right)^{1/\alpha}\right) \le \lambda_j^{\frac{1}{\alpha}}\left(\sum_{i=1}^m A_i\right)\lambda_1^{1-\frac{1}{\alpha}}\left(\sum_{i=1}^m A_i\right).$$

(Here, as always, $\lambda_j(B)$ are eigenvalues of a symmetric matrix B arranged in the nonascending order.)

Proof. Let $B = \sum_{i=1}^m A_i^\alpha$, $A = \sum_{i=1}^m A_i$. Since $\lambda_j(B^{1/\alpha}) = \left(\lambda_j(B)\right)^{1/\alpha}$, we should prove that

$$\lambda_j(B) \le \lambda_j(A)\lambda_1^{\alpha-1}(A). \tag{7}$$

By the variational description of eigenvalues, it suffices to verify that for every $j \le n$ there exists a linear subspace L_j in $\mathbf{R}^n$ of codimension $j-1$ such that

$$\xi^T B\xi \le \lambda_j(A)\lambda_1^{\alpha-1}(A) \quad \forall(\xi \in L_j, \|\xi\|_2 = 1). \tag{8}$$

Let $e_1, \dots, e_n$ be an orthonormal eigenbasis of A ($Ae_j = \lambda_j(A)e_j$), and let L_j be the linear span of the vectors $e_j, e_{j+1}, \dots, e_n$. Let $\xi \in L_j$ be a unit vector. We have

$$\begin{aligned}
\sum_i \xi^T A_i^\alpha \xi &= \sum_i \xi^T A_i \underbrace{[A_i^{\alpha-1}\xi]}_{\eta_i} \\
&\le \sum_i (\xi^T A_i \xi)^{1/2} (\eta_i^T A_i \eta_i)^{1/2} && \text{[since } A_i \succeq 0\text{]} \\
&\le \left(\sum_i \xi^T A_i \xi\right)^{1/2} \left(\sum_i \eta_i^T A_i \eta_i\right)^{1/2} && \text{[Cauchy's inequality]} \\
&= \left(\xi^T A \xi\right)^{1/2} \left(\sum_i \xi^T A_i^{2\alpha-1} \xi\right)^{1/2} \\
&\le \left(\xi^T A \xi\right)^{1/2} \left(\sum_i \lambda_1^{2\alpha-2}(A_i) \xi^T A_i \xi\right)^{1/2} \\
&\le \left(\max_i \lambda_1(A_i)\right)^{\alpha-1} \sum_i \xi^T A_i \xi \\
&\le \lambda_1^{\alpha-1}(A)\lambda_j(A) && \text{[since } \|\xi\|_2 = 1 \text{ and } \xi \in L_j\text{]}
\end{aligned}$$

as required in (8). □

Part 3 is less trivial than part 2. Let us look at the $(nm) \times (nm)$ square matrix

$$Q = \begin{pmatrix} A_1^{\alpha/2} & & \\ A_2^{\alpha/2} & & \\ \vdots & & \\ A_m^{\alpha/2} & & \end{pmatrix}.$$

(As always, blank spaces are filled with zeros.) Then

$$Q^T Q = \left(\begin{array}{c|c} B \equiv \sum_i A_i^\alpha & \\ \hline & \end{array}\right)$$

so that $\mathrm{Tr}([Q^TQ]^{1/\alpha}) = \mathrm{Tr}(B^{1/\alpha})$. Since the eigenvalues of Q^TQ are exactly the same as the eigenvalues of $X = QQ^T$, we conclude that

$$\mathrm{Tr}(B^{1/\alpha}) = \mathrm{Tr}([QQ^T]^{1/\alpha}) = \mathrm{Tr}(X^{1/\alpha}),$$

$$X = \begin{pmatrix} A_1^\alpha & A_1^{\alpha/2}A_2^{\alpha/2} & \cdots & A_1^{\alpha/2}A_m^{\alpha/2} \\ A_2^{\alpha/2}A_1^{\alpha/2} & A_2^\alpha & \cdots & A_2^{\alpha/2}A_m^{\alpha/2} \\ \vdots & \vdots & \ddots & \vdots \\ A_m^{\alpha/2}A_1^{\alpha/2} & A_m^{\alpha/2}A_2^{\alpha/2} & \cdots & A_m^\alpha \end{pmatrix}.$$

Applying the result of Exercise 4.12.1 with $F(Y) = -\mathrm{Tr}(Y^{1/\alpha})$, we get

$$[\mathrm{Tr}(B^{1/\alpha}) =] \quad \mathrm{Tr}(X^{1/\alpha}) = -F(X) \le -F\left(\begin{pmatrix} A_1^\alpha & & \\ & \ddots & \\ & & A_m^\alpha \end{pmatrix}\right) = \sum_{i=1}^m \mathrm{Tr}(A_i),$$

as required.

$\succeq$-convexity of some matrix-valued functions.

Exercise 4.22.

4. Prove that the function

$$F(x) = x^{1/2} : \mathbf{S}_+^m \to \mathbf{S}_+^m$$

is $\succeq$-concave and $\succeq$-monotone.

Solution. Since the function is continuous on its domain, it suffices to verify that it is $\succeq$-monotone and $\succeq$-concave on $\mathrm{int}\mathbf{S}_+^m$, where the function is smooth.

Differentiating the identity

$$F(x)F(x) = x \qquad (*)$$

in a direction h and setting $F(x) = y$, $DF(x)[h] = dy$, we get

$$y\,dy + dy\,y = h.$$

Since $y \succ 0$, this Lyapunov equation admits an explicit solution:

$$dy = \int_0^\infty \exp\{-ty\}h\exp\{-ty\}dt,$$

and we see that $dy \succeq 0$ whenever $h \succeq 0$. Applying Exercise 4.21.4, we conclude that F is $\succeq$-monotone.

Differentiating (*) twice in a direction h and denoting $d^2y = D^2F(x)[h,h]$, we get

$$y\,d^2y + d^2y\,y + 2(dy)^2 = 0,$$

whence, same as above,

$$d^2y = -\int_0^\infty \exp\{-ty\}(dy)^2\exp\{-ty\}dt \preceq 0.$$

Applying Exercise 4.21.3, we conclude that F is $\succeq$-concave. □

5. Prove that the function

$$F(x) = \ln x : \mathrm{int}\mathbf{S}^m_+ \to \mathbf{S}^m$$

is $\succeq$-monotone and $\succeq$-concave.

Solution. The function $x^{1/2} : \mathbf{S}^m_+ \to \mathbf{S}^m_+$ is $\succeq$-monotone and $\succeq$-concave by Exercise 4.22.4. Applying Exercise 4.21.6, we conclude that so are the functions $x^{1/2^k}$ for all positive integer k. It remains to note that

$$\ln x = \lim_{k\to\infty} 2^k\left[x^{1/2^k} - I\right]$$

and to use Exercise 4.21.7. □

6. Prove that the function

$$F(x) = \left(Ax^{-1}A^T\right)^{-1} : \mathrm{int}\mathbf{S}^n_+ \to \mathbf{S}^m$$

with matrix A of rank m is $\succeq$-concave and $\succeq$-monotone.

Solution. Since A is of rank m, the function $F(x)$ clearly is well defined and $\succ 0$ when $x \succ 0$. To prove that F is $\succeq$-concave, it suffices to verify that the set

$$\{(x,Y) \mid x, Y \succ 0, Y \preceq (Ax^{-1}A^T)^{-1}\}$$

is convex, which is nearly evident:

$$\begin{aligned}\{(x,Y) \mid x, Y \succ 0, Y \preceq (Ax^{-1}A^T)^{-1}\} &= \{(x,Y) \mid x, Y \succ 0, Y^{-1} \succeq Ax^{-1}A^T\}\\ &= \{(x,Y) \mid x, Y \succ 0, \begin{pmatrix} Y^{-1} & A\\ A^T & x\end{pmatrix} \succeq 0\}\\ &= \{(x,Y) \mid x, Y \succ 0, x \succeq A^TYA\}.\end{aligned}$$

To prove that F is $\succeq$-monotone, note that if $0 \preceq x \preceq x'$, then $0 \prec (x')^{-1} \preceq x^{-1}$, whence $0 \prec A(x')^{-1}A^T \preceq Ax^{-1}A^T$, whence, in turn, $F(x) = (Ax^{-1}A^T)^{-1} \preceq (A(x')^{-1}A^T)^{-1} = F(x')$. □

Lovasz capacity number

Exercise 4.33. Let Γ be an n-node graph and $\sigma(\Gamma)$ be the optimal value in the problem

$$\min_{\lambda,\mu,\nu}\left\{\lambda : \begin{pmatrix} \lambda, & -\frac{1}{2}(e+\mu)^T \\ -\frac{1}{2}(e+\mu), & A(\mu,\nu)\end{pmatrix} \succeq 0\right\}, \tag{Sh}$$

where $e = (1,\dots,1)^T \in \mathbf{R}^n$, $A(\mu,\nu) = \mathrm{Diag}(\mu) + Z(\nu)$, and $Z(\nu)$ is the matrix as follows:

- The dimension of ν is equal to the number of arcs in Γ, and the coordinates of ν are indexed by these arcs.
- The diagonal entries of Z, same as the off-diagonal entries of Z corresponding to empty cells ij (i.e., with i and j nonadjacent) are zeros.
- The off-diagonal entries of Z in a pair of symmetric nonempty cells ij, ji are equal to the coordinate of ν indexed by the corresponding arc.

Prove that $\sigma(\Gamma)$ is nothing but the Lovasz capacity $\Theta(\Gamma)$ of the graph.

Solution. In view of (4.10.122) all we need is to prove that $\sigma(\Gamma) \geq \Theta(\Gamma)$.

Let (λ,μ,ν) be a feasible solution to (Sh); we should prove that there exists x such that (λ, x) is a feasible solution to the Lovasz problem

$$\min_{\lambda,x}\{\lambda : \lambda I_n - \mathcal{L}(x) \succeq 0\}. \tag{L}$$

Setting $y = \frac{1}{2}(e+\mu)$, we see that

$$\begin{pmatrix} \lambda & -\frac{1}{2}(e+\mu)^T \\ -\frac{1}{2}(e+\mu) & Z(\nu)+\mathrm{Diag}(\mu)\end{pmatrix} = \begin{pmatrix} \lambda & -y^T \\ -y & Z(\nu)+2\mathrm{Diag}(y)-I_n\end{pmatrix} \succeq 0.$$

The diagonal entries of $Z = Z(\nu)$ are zero, while the diagonal entries of $Z + 2\mathrm{Diag}(y) - I_n$ must be nonnegative; we conclude that $y > 0$. Setting $Y = \mathrm{Diag}(y)$, we have

$$\begin{pmatrix} \lambda & -y^T \\ -y & Z+2Y-I_n\end{pmatrix} \succeq 0$$
$$\Rightarrow \begin{pmatrix} 1 & \\ & Y^{-1}\end{pmatrix}\begin{pmatrix} \lambda & -y^T \\ -y & Z+2Y-I_n\end{pmatrix}\begin{pmatrix} 1 & \\ & Y^{-1}\end{pmatrix} \succeq 0,$$

i.e.,

$$\begin{pmatrix} \lambda & -e^T \\ -e & Y^{-1}ZY^{-1}+2Y^{-1}-Y^{-2}\end{pmatrix} \succeq 0,$$

whence by the Schur complement lemma

$$\lambda\left[Y^{-1}ZY^{-1}+2Y^{-1}-Y^{-2}\right] - ee^T \succeq 0$$

or, which is the same,

$$\lambda I_n - \left[ee^T - \lambda Y^{-1}ZY^{-1}\right] \succeq \lambda(I_n - 2Y^{-1} + Y^{-2}) = \lambda(I_n - Y^{-1})^2.$$

We see that $\lambda I_n - \left[ee^T - \lambda Y^{-1}ZY^{-1}\right] \succeq 0$. It remains to note that the matrix in the brackets clearly is $\mathcal{L}(x)$ for certain x. □

$\mathcal{S}$-lemma. (For notation, see section 4.10.5.)

Exercise 4.42. We should prove that if $f(x) = x^T Ax + 2a^T x + \alpha$ and $g(x) = x^T Bx + 2b^T x + \beta$ are two quadratic forms ($A = A^T$, $B = B^T$) such that the premise in the implication

$$f(x) \le 0 \Rightarrow g(x) \le 0 \tag{9}$$

is strictly feasible, then the implication holds true if and only if

$$\exists \lambda \ge 0 : \quad g(x) \le \lambda f(x) \ \forall x.$$

Proof. The "if" part of the statement is evident. Let us prove the "only if" part. Thus, let us assume that the implication (9) is valid.

1. There clearly exist a sequence $\{\gamma_i > 0\}$ and $\delta > 0$ such that

   (i) $\gamma_i \to 0$, $i \to \infty$;

   (ii) all the matrices $A_i \equiv A + \gamma_i I$ are nonsingular; and

   (iii) $\bar{x}^T A_i \bar{x} + 2a^T \bar{x} + \alpha \le -\delta \ \forall i$.

2. Observe that since $\gamma_i > 0$, one has

$$x^T A_i x + 2a^T x + \alpha \le 0 \Rightarrow x^T Ax + 2a^T x + \alpha \le 0 \Rightarrow x^T Bx + 2b^T x + \beta \le 0,$$

i.e., one has

$$f_i(x) \equiv x^T A_i x + 2a^T x + \alpha \le 0 \Rightarrow g(x) = x^T Bx + 2b^T x + \beta \le 0.$$

Setting

$$\begin{aligned} \widehat{f_i}(y) &= f_i(y - A_i^{-1}a) = y^T A_i y + \underbrace{[\alpha - a^T A_i^{-1} a]}_{\alpha_i}, \\ \widehat{g_i}(y) &= g(y - A_i^{-1}a) = y^T By + 2b_i^T y + \beta_i, \end{aligned}$$

we have

$$\begin{array}{lll} \text{(a)} & \forall y : & y^T A_i y + \alpha_i \le 0 \Rightarrow y^T By + 2b_i^T y + \beta_i \le 0, \\ \text{(b)} & & \exists \tilde{y} : \ \tilde{y}^T A_i \tilde{y} + \alpha_i < 0. \end{array} \tag{10}$$

3. We claim that

$$\begin{array}{lll} \text{(a)} & \forall (y,t) : & y^T A_i y + \alpha_i t^2 \le 0 \Rightarrow y^T By + 2t b_i^T y + \beta_i t^2 \le 0, \\ \text{(b)} & & \exists (\bar{y}, \bar{t}) : \ \bar{y}^T A_i \bar{y} + \alpha_i \bar{t}^2 < 0. \end{array} \tag{11}$$

Indeed, (11)(b) is evident (set $\bar{y} = \tilde{y}$, $\bar{t} = 1$). Further, the implication (11)(a) with the premise strengthened by the assumption $t \ne 0$ is an immediate corollary of (10)(a). Thus, all we need in order to verify (11)(a) is to prove the implication

$$y^T A_i y \le 0 \Rightarrow y^T By \le 0. \tag{12}$$

Assume, on the contrary to what should be proved, that for some y it holds that

$$y^T A_i y \le 0, \quad y^T B y > 0.$$

Then $y \neq 0$, whence $y^T A y < y^T A_i y \le 0$, i.e., $y^T A y < 0$. It follows that

$$\begin{array}{rcl} f(sy - A_i^{-1}a) & \to & -\infty,\ s \to \infty \quad [\text{since } y^T A y < 0], \\ g(sy - A_i^{-1}a) & \to & +\infty,\ s \to \infty \quad [\text{since } y^T B y > 0], \end{array}$$

in contradiction to the assumption that $f(x) \le 0 \Rightarrow g(x) \le 0$.

4. By the usual $\mathcal{S}$-lemma, from (11) it follows that

$$\exists \lambda_i \ge 0: \quad y^T B y + 2t b_i^T y + \beta_i t^2 \le \lambda_i [y^T A_i y + \alpha_i t^2] \quad \forall (y, t),$$

whence (set $t = 1$) $\widehat{g}_i(y) \le \lambda_i \widehat{f}_i(y)\ \forall y$ or, which is the same,

$$x^T B x + 2b^T x + \beta \le \lambda_i [x^T A_i x + 2a^T x + \alpha] \quad \forall x.$$

Setting in the latter inequality $x = \bar{x}$ and taking into account (iii), we conclude that $\{\lambda_i \ge 0\}_{i=1}^{\infty}$ is a bounded sequence. Denoting by λ (any) limiting point of the sequence and taking into account (i), we come to the desired relations $\lambda \ge 0$, $g(x) \le \lambda f(x)\ \forall x$. □

Exercise 4.46.3. 3. Given data A, B satisfying the premise of (**SL**)(B), define the sets

$$Q_x = \{\lambda \ge 0 : x^T B x \ge \lambda x^T A x\}.$$

Prove that every two sets $Q_{x'}$, $Q_{x''}$ have a point in common.

Solution. The case when x', x'' are collinear is trivial. Assuming that x', x'' are linearly independent, consider the quadratic forms on the 2D plane:

$$\alpha(z) = (sx' + tx'')^T A (sx' + tx''),\ \beta(z) = (sx' + tx'')^T B (sx' + tx''),\ \ z = (s, t)^T.$$

By their origin, we have

$$\alpha(z) \ge 0, z \neq 0 \Rightarrow \beta(z) > 0. \qquad (!)$$

All we need is to prove that there exists $\lambda \ge 0$ such that $\beta(z) \ge \lambda \alpha(z)\ \forall z \in \mathbf{R}^2$. Such a λ clearly is a common point of $Q_{x'}$ and $Q_{x''}$.

As is well known from linear algebra, we can choose a coordinate system in $\mathbf{R}^2$ in such a way that the matrix α of the form $\alpha(\cdot)$ in these coordinates, let them be called u, v, is diagonal:

$$\alpha = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix};$$

let also

$$\beta = \begin{pmatrix} p & r \\ r & q \end{pmatrix}$$

be the matrix of the form $\beta(\cdot)$ in the coordinates u, v. Let us consider all possible combinations of signs of a, b:

- $a \geq 0, b \geq 0$. In this case, $\alpha(\cdot)$ is nonnegative everywhere, whence by (!) $\beta(\cdot) \geq 0$. Consequently, $\beta(\cdot) \geq \lambda\alpha(\cdot)$ with $\lambda = 0$.
- $a < 0, b < 0$. In this case the matrix of the quadratic form $\beta(\cdot) - \lambda\alpha(\cdot)$ is

$$\begin{pmatrix} p + \lambda|a| & r \\ r & q + \lambda|b| \end{pmatrix}.$$

  This matrix clearly is positive definite for all large-enough positive λ, so that here again $\beta(\cdot) \geq \lambda\alpha(\cdot)$ for properly chosen nonnegative λ.
- $a = 0, b < 0$. In this case $\alpha(1, 0) = 0$ (the coordinates in question are u, v), so that by (!) $p > 0$. The matrix of the form $\beta(\cdot) - \lambda\alpha(\cdot)$ is

$$\begin{pmatrix} p & r \\ r & q + \lambda|b| \end{pmatrix},$$

  and since $p > 0$ and $|b| > 0$, this matrix is positive definite for all large-enough positive λ. Thus, here again $\beta(\cdot) \geq \lambda\alpha(\cdot)$ for properly chosen $\lambda \geq 0$.
- $a < 0, b = 0$. This case is completely similar to the previous one.

Part 3 is proved. □

Exercise 4.47. Demonstrate by example that if $x^T Ax, x^T Bx, x^T Cx$ are three quadratic forms with symmetric matrices such that

$$\begin{array}{c} \exists \bar{x} : \bar{x}^T A\bar{x} > 0, \bar{x}^T B\bar{x} > 0, \\ x^T Ax \geq 0, x^T Bx \geq 0 \Rightarrow x^T Cx \geq 0, \end{array} \tag{13}$$

then not necessarily there exist $\lambda, \mu \geq 0$ such that $C \succeq \lambda A + \mu B$.

A solution.

$$A = \begin{pmatrix} \lambda^2 & 0 \\ 0 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} \mu\nu & 0.5(\mu - \nu) \\ 0.5(\mu - \nu) & -1 \end{pmatrix}, \quad C = \begin{pmatrix} \lambda\mu & 0.5(\mu - \lambda) \\ 0.5(\mu - \lambda) & -1 \end{pmatrix}.$$

With a proper setup, e.g.,

$$\lambda = 1.100, \quad \mu = 0.818, \quad \nu = 1.344,$$

the above matrices satisfy both (4.10.133) and (4.10.134). □

Exercise 4.52. Let $n \geq 3$,

$$\begin{aligned} f(x) &= \theta_1 x_1^2 - \theta_2 x_2^2 + \sum_{i=3}^{n} \theta_i x_i^2 : \mathbf{R}^n \to \mathbf{R}, \\ &\quad \theta_1 \geq \theta_2 \geq 0, \theta_1 + \theta_2 > 0, -\theta_2 \leq \theta_i \leq \theta_1 \quad \forall i \geq 3; \\ Y &= \{x : \|x\|_2 = 1, f(x) = 0\}. \end{aligned}$$

1. Let $x \in Y$. Prove that x can be linked in Y by a continuous curve with a point x' such that the coordinates of x' with indices $3, 4, \ldots, n$ vanish.

Proof. It suffices to build a continuous curve $\gamma(t) \in Y$, $0 \leq t \leq 1$, of the form

$$\gamma(t) = (x_1(t), x_2(t), tx_3, tx_4, \ldots, tx_n)^T, \quad 0 \leq t \leq 1,$$

which passes through x as $t = 1$. Setting $s = \sum_{i=3}^{n} \theta_i x_i^2 = \theta_2 x_2^2 - \theta_1 x_1^2$ and $g^2 = d^T d = 1 - x_1^2 - x_2^2$, we should verify that one can define continuous functions $x_1(t), x_2(t)$ of $t \in [0, 1]$ satisfying the system of equations

$$\left\{ \begin{array}{rcl} \theta_1 x_1^2(t) - \theta_2 x_2^2(t) + t^2 s &=& 0, \\ x_1^2(t) + x_2^2(t) + t^2 g^2 &=& 1 \end{array} \right.$$

along with the boundary conditions

$$\begin{aligned} x_1(1) &= x_1, \\ x_2(1) &= x_2. \end{aligned}$$

Substituting $v_1(t) = x_1^2(t)$, $v_2(t) = x_2^2(t)$ and taking into account that $\theta_1, \theta_2 \geq 0, \theta_1 + \theta_2 > 0$, we get

$$\begin{aligned} (\theta_1 + \theta_2) v_1(t) &= \theta_2(1 - t^2 g^2) - t^2 s \\ &= \theta_2(1 - t^2 g^2 - t^2(\theta_2 x_2^2 - \theta_1 x_1^2) \\ &= \theta_2(1 - t^2[g^2 + x_2^2]) + t^2 \theta_1 x_1^2 \\ &= \theta_2(1 - t^2[1 - x_1^2]) + t^2 \theta_1 x_1^2; \\ (\theta_1 + \theta_2) v_2(t) &= \theta_1(1 - t^2 g^2) + t^2 s \\ &= \theta_1(1 - t^2 g^2) + t^2(\theta_2 x_2^2 - \theta_1 x_1^2) \\ &= \theta_1(1 - t^2[g^2 + x_1^2]) + t^2 \theta_2 x_2^2 \\ &= \theta_1(1 - t^2[1 - x_2^2]) + t^2 \theta_2 x_2^2. \end{aligned}$$

We see that $v_1(t)$, $v_2(t)$ are continuous, nonnegative and equal x_1^2, x_2^2, respectively, as $t = 1$. Taking $x_1(t) = \kappa_1 v_1^{1/2}$, $x_2(t) = \kappa_2 v_2^{1/2}(t)$ with properly chosen $\kappa_i = \pm 1$, $i = 1, 2$, we get the required curve $\gamma(\cdot)$. □

2. Prove that there exists a point $z^+ = (z_1, z_2, z_3, 0, 0, \ldots, 0)^T \in Y$ such that
(i) $z_1 z_2 = 0$.
(ii) given a point $u = (u_1, u_2, 0, 0, \ldots, 0)^T \in Y$, you can either (ii)(a) link u by continuous curves in Y both to z^+ and to $\bar{z}^+ = (z_1, z_2, -z_3, 0, 0, \ldots, 0)^T \in Y$, or (ii)(b) link u both to $z^- = (-z_1, -z_2, z_3, 0, 0, \ldots, 0)^T$ and $\bar{z}^- = (-z_1, -z_2, -z_3, 0, 0, \ldots, 0)^T$. (Note that $z^+ = -\bar{z}^-$, $\bar{z}^+ = -z^-$.)

Proof. Recall that $\theta_1 \geq \theta_2 \geq 0$ and $\theta_1 + \theta_2 > 0$. Consider two possible cases: $\theta_2 = 0$ and $\theta_2 > 0$.

Case of $\theta_2 = 0$. In this case it suffices to set $z^+ = (0, 1, 0, 0, \ldots, 0)^T$. Indeed, the point clearly belongs to Y and satisfies (i). Further, if $u \in Y$ is such that $u_3 = \cdots = u_n = 0$, then from $\theta_2 = 0$ and the definition of Y it immediately follows that $u_1 = 0$, $u_2 = \pm 1$. Thus, either u coincides with $z^+ = \bar{z}^+$, or u coincides with $z^- = \bar{z}^-$. In both cases, (ii) takes place.

Case of $\theta_2 > 0$. Let us set

$$\begin{aligned}
\tau &= \min\left[\tfrac{\theta_1}{\theta_1-\theta_3}; \tfrac{\theta_2}{\theta_2+\theta_3}\right] \\
&\text{[note that } \tau > 0 \text{ due to } \theta_1, \theta_2 > 0 \text{ and } -\theta_2 \leq \theta_3 \leq \theta_1\text{]}, \\
z_1 &= \sqrt{(\theta_1+\theta_2)^{-1}\,(\theta_2 - \tau[\theta_2+\theta_3])}, \\
z_2 &= \sqrt{(\theta_1+\theta_2)^{-1}\,(\theta_1 - \tau[\theta_1-\theta_3])}, \\
z_3 &= \sqrt{\tau}, \\
z^+ &= (z_1, z_2, z_3, 0, 0, \ldots, 0)^T.
\end{aligned}$$

It is immediately seen that z^+ is well defined and satisfies (i). Now let us verify that z^+ satisfies (ii) as well. Let $u = (u_1, u_2, 0, 0, \ldots, 0)^T \in Y$, and let the vector-function $z(t)$, $0 \leq t \leq \sqrt{\tau}$, be defined by the relations

$$\begin{aligned}
z_1(t) &= \sqrt{(\theta_1+\theta_2)^{-1}\left(\theta_2 - t^2[\theta_2+\theta_3]\right)}, \\
z_2(t) &= \sqrt{(\theta_1+\theta_2)^{-1}\left(\theta_1 - t^2[\theta_1-\theta_3]\right)}, \\
z_3(t) &= t, \\
z_i(t) &= 0, \;\; i = 4, \ldots, n.
\end{aligned}$$

It is immediately seen that $z(\cdot)$ is well defined, is continuous, takes its values in Y, and $z(\sqrt{\tau}) = z^+$. Now, $z(0)$ is the vector

$$\bar{u} \equiv \left(\sqrt{\frac{\theta_2}{\theta_1+\theta_2}}, \sqrt{\frac{\theta_1}{\theta_1+\theta_2}}, 0, 0, \ldots, 0\right)^T.$$

From $u \in Y$, $u_3 = \cdots = u_n = 0$ it immediately follows that $|u_i| = |\bar{u}_i|$, $i = 1, 2$. Now consider four possible cases:

$(++)$ $u_1 = \bar{u}_1, u_2 = \bar{u}_2$,

$(--)$ $u_1 = -\bar{u}_1, u_2 = -\bar{u}_2$,

$(+-)$ $u_1 = \bar{u}_1, u_2 = -\bar{u}_2$,

$(-+)$ $u_1 = -\bar{u}_1, u_2 = \bar{u}_2$.

- In the case of $(++)$ the continuous curve $\gamma(t) \equiv z(t) \in Y$, $0 \leq t \leq \sqrt{\tau}$, links u with z^+, while the continuous curve $\bar{\gamma}(t) = (z_1(t), z_2(t), -z_3(t), 0, 0, \ldots, 0)^T \in Y$ links u with $\bar{z}^+$, so that (ii)(1) takes place.

- In the case of $(--)$ the continuous curve $\gamma(t) \equiv -z(t) \in Y$, $0 \le t \le \sqrt{\tau}$, links u with $\bar{z}^- = -z^+$, while the continuous curve $\bar{\gamma}(t) = (-z_1(t), -z_2(t), z_3(t), 0, 0, \dots, 0)^T \in Y$ links u with $z^- = -\bar{z}^+$, so that (ii)(2) takes place.
- In the case of $(+-)$ the continuous curves $\gamma(t) \equiv (z_1(t), -z_2(t), z_3(t), 0, 0, \dots, 0)^T \in Y$ and $\bar{\gamma}(t) = (z_1(t), -z_2(t), -z_3(t), 0, 0, \dots, 0)^T$, $0 \le t \le \sqrt{\tau}$, link u either with both points of the pair $(z^+, \bar{z}^+)$, or with both points of the pair $(z^-, \bar{z}^-)$, depending on whether $z_1 = 0$ or $z_1 \ne 0$, $z_2 = 0$. (Note that at least one of these possibilities does take place due to $z_1 z_2 = 0$; see (i).) Thus, in the case in question at least (ii)(1) or (ii)(2) does hold.
- In the case of $(-+)$ the continuous curves $\gamma(t) \equiv (-z_1(t), z_2(t), z_3(t), 0, 0, \dots, 0)^T \in Y$ and $\bar{\gamma}(t) = (z_1(t), -z_2(t), -z_3(t), 0, 0, \dots, 0)^T$, $0 \le t \le \sqrt{\tau}$, link u either with both points of the pair $(z^-, \bar{z}^-)$ or with both points of the pair $(z^+, \bar{z}^+)$, depending on whether $z_1 = 0$ or $z_1 \ne 0$, $z_2 = 0$, so that here again (ii)(1) or (ii)(2) does hold. □

3. Conclude from 1 and 2 that Y satisfies the premise of Proposition 4.10.2, and thus complete the proof of Proposition 4.10.4.

Proof. Let z^+ be given by part 2, and let $x, x' \in Y$. By part 1, we can link in Y the point x with a point $v = (v_1, v_2, 0, 0, \dots, 0)^T \in Y$, and the point x' with a point $v' = (v_1', v_2', 0, 0, \dots, 0)^T \in Y$. If for both $u = v$ and $u = v'$ (ii)(a) holds, then we can link both v, v' by continuous curves with z^+. Thus, both x, x' can be linked in Y with z^+ as well. We see that both x, x' can be linked in Y with the set $\{z^+; -z^+\}$, as required in the premise of Proposition 4.10.2. The same conclusion is valid if for both $u = v$, $u = v'$ (ii)(b) holds. Here both x and x' can be linked in Y with z^-, and the premise of Proposition 4.10.2 holds true.

Now consider the case when for one of the points $u = v$, $u = v'$, say, for $u = v$, (ii)(a) holds, while for the other one (ii)(b) takes place. Here we can link in Y the point v (and thus the point x) with the point z^+, and we can link in Y the point v' (and thus the point x') with the point $\bar{z}^- = -z^+$. Thus, we can link in Y both x and x' with the set $\{z^+; -z^+\}$, and the premise of Proposition 4.10.2 holds true. Thus, Y satisfies the premise of Proposition 4.10.2. □

Exercise 4.53. 2. Let A_i, $i = 1, 2, 3$, satisfy the premise of (**SL**)(D). Assuming $A_1 = I$, prove that the set

$$H_1 = \{(v_1, v_2)^T \in \mathbf{R}^2 \mid \exists x \in \mathbf{S}^{n-1} : v_1 = f_2(x), v_3 = f_3(x)\}$$

is convex.

Proof. Let $\ell = \{(v_1, v_2)^t \in \mathbf{R}^2 \mid pv_1 + qv_2 + r = 0\}$ be a line in the plane. We should prove that $W = X_1 \cap \ell$ is a convex or, which is the same, connected set

(Exercise 4.49). There is nothing to prove when $W = \emptyset$. Assuming $W \neq \emptyset$, let us set

$$f(x) = rf_1(x) + pf_2(x) + qf_3(x).$$

It is immediately seen that f is a homogeneous quadratic form on $\mathbf{R}^n$, and that

$$W = F(Y),$$

where

$$F(x) = \begin{pmatrix} f_2(x) \\ f_3(x) \end{pmatrix},$$
$$Y \equiv \{x \in S^{n-1} : f(y) = 0\} = -Y.$$

By Proposition 4.10.4, the image Z of the set Y under the canonical projection $S^{n-1} \to \mathbf{P}^{n-1}$ is connected. Since F is even, $W = F(Y)$ is the same as $G(Z)$ for certain continuous mapping $G : Z \to \mathbf{R}^2$ (Proposition 4.10.3). Thus, W is connected by (C.2).

We have proved that Z_1 is convex; the compactness of Z_1 is evident. □

Exercise 4.54. Demonstrate by example that (**SL**)(C) not necessarily remains valid when skipping the assumption $n \geq 3$ in the premise.

A solution. The linear combination

$$A + 0.005B - 1.15C$$

of the matrices built in the solution to Exercise 4.47 is positive definite. □

Exercise 4.55. Let A, B, C be three 2×2 symmetric matrices such that the system of inequalities $x^T Ax \geq 0$, $x^T Bx \geq 0$ is strictly feasible and the inequality $x^T Cx$ is a consequence of the system.

1. Assume that there exists a nonsingular matrix Q such that both the matrices QAQ^T and QBQ^T are diagonal. Prove that then there exist $\lambda, \mu \geq 0$ such that $C \succeq \lambda A + \mu B$.

Solution. Without loss of generality we may assume that the matrices A, B themselves are diagonal. The case when at least one of the matrices A, B is positive semidefinite is trivially reducible to the usual $\mathcal{S}$-lemma, so that we can assume that both matrices A and B are not positive semidefinite. Since the system of inequalities $x^T Ax > 0, x^T Bx > 0$ is feasible, we conclude that the determinants of the matrices A, B are negative. Applying appropriate dilatations of the coordinate axes, swapping, if necessary, the coordinates and multiplying A, B by appropriate positive constants we may reduce the situation to the one where $x^T Ax = x_1^2 - x_2^2$ and either (a) $x^T Bx = \theta^2 x_1^2 - x_2^2$ or (b) $x^T Bx = -\theta^2 x_1^2 + x_2^2$ with certain $\theta > 0$.

Case of (a). Here the situation is immediately reducible to the one considered in the $\mathcal{S}$-lemma. Indeed, in this case one of the inequalities in the system $x^TAx \geq 0$, $x^TBx \geq 0$ is a consequence of the other inequality of the system. Thus, a consequence x^TCx of the system is in fact a consequence of a properly chosen *single* inequality from the system. Thus, by the $\mathcal{S}$-lemma either $C \succeq \lambda A$ or $C \succeq \lambda B$ with certain $\lambda \geq 0$.

Case of (b). Observe, first, that $\theta < 1$, since otherwise the system $x^TAx \geq 0$, $x^TBx \geq 0$ is not strictly feasible. When $\theta < 1$, the solution set of our system is the union of the following four angles D^{++}, D^{+-}, D^{--}, D^{-+}:

$$\begin{array}{rcl} D^{++} & = & \{x \mid x_1 \geq 0, \theta x_1 \leq x_2 \leq x_1\}, \\ D^{+-} & = & \text{reflection of } D^{++} \text{ without respect to the } x_1\text{-axis}, \\ D^{--} & = & -D^{++}, \\ D^{-+} & = & \text{reflection of } D^{++} \text{ with respect to the } x_2\text{-axis}. \end{array}$$

Now, the case when C is positive semidefinite is trivial—here $C \succeq 0 \times A + 0 \times B$. Thus, we may assume that one eigenvalue of C is negative; the other should be nonnegative, since otherwise $x^TCx < 0$ whenever $x \neq 0$, while we know that $x^TCx \geq 0$ at the (nonempty!) solution set of the system $x^TAx > 0$, $x^TBx > 0$. Since one eigenvalue of C is negative, and the other one is nonnegative, the set $X_C = \{x \mid x^TCx \geq 0\}$ is the union of a certain angle D (which can reduce to a ray) and the angle $-D$. Since the inequality $x^TCx \geq 0$ is a consequence of the system $x^TAx \geq 0$, $x^TBx \geq 0$, we have $D \cup (-D) \supset D^{++} \cup D^{+-} \cup D^{--} \cup D^{-+}$. Geometry says that the latter inclusion can be valid only when D contains two neighboring, with respect to the cyclic order, of the angles D^{++}, D^{+-}, D^{--}, D^{-+}. But in this case the inequality x^TCx is a consequence of an appropriate single inequality from the pair $x^TAx \geq 0$, $x^TBx \geq 0$. Namely, when $D \supset D^{++} \cup D^{+-}$ or $D \supset D^{--} \cup D^{-+}$, X_C contains the solution set of the inequality $x^TBx \geq 0$, while in the cases of $D \supset D^{+-} \cup D^{--}$ and of $D \supset D^{-+} \cup D^{++}$, X_C contains the solution set of the inequality $x^TAx \geq 0$. Applying the usual $\mathcal{S}$-lemma, we conclude that in the case of (b) there exist $\lambda, \mu \geq 0$ (with one of these coefficients equal to 0) such that $C \succeq \lambda A + \mu B$. □

Exercises to Lecture 6

Canonical barriers

Exercise 6.2. Prove Proposition 6.3.2.

Solution. As explained in the Hint, it suffices to consider the case of $\mathbf{K} = \mathbf{L}^k$. Let $x = (u, t) \in \text{int}\mathbf{L}^k$, and let $s = (v, \tau) = -\nabla L_k(x)$. We should prove that $s \in \text{int}\mathbf{L}^k$ and that $\nabla L_k(s) = -x$. By (6.3.2), one has

$$v = -\frac{2}{t^2-u^Tu}u, \quad \tau = \frac{2}{t^2-u^Tu}t$$
$$\Downarrow$$
$$\tau \geq 0 \text{ and } \tau^2 - v^Tv = \frac{4}{t^2-x^Tx} > 0$$
$$\Downarrow$$
$$s \in \text{int}\mathbf{L}^k;$$
$$-\nabla L_k(s) = \begin{pmatrix} -\frac{2}{\tau^2-v^Tv}v \\ \frac{2}{\tau^2-v^Tv}\tau \end{pmatrix} = \begin{pmatrix} \frac{2}{\frac{4}{t^2-u^Tu}}\frac{2}{t^2-u^Tu}u \\ \frac{2}{\frac{4}{t^2-u^Tu}}\frac{2}{t^2-u^Tu}t \end{pmatrix} = \begin{pmatrix} u \\ t \end{pmatrix}. \qquad \square$$

Scalings of canonical cones

Exercise 6.4. Prove the following:

1. Whenever $e \in \mathbf{R}^{k-1}$ is a unit vector and $\mu \in \mathbf{R}$, the linear transformation

$$L_{\mu,e}: \quad \begin{pmatrix} u \\ t \end{pmatrix} \mapsto \begin{pmatrix} u - [\mu t - (\sqrt{1+\mu^2}-1)e^Tu]e \\ \sqrt{1+\mu^2}t - \mu e^Tu \end{pmatrix} \qquad (*)$$

maps the cone $\mathbf{L}^k$ onto itself. Besides this, transformation $(*)$ preserves the space-time interval $x^TJ_kx \equiv -x_1^2 - \cdots - x_{k-1}^2 + x_k^2$:

$$[L_{\mu,e}x]^TJ_k[L_{\mu,e}x] = x^TJ_kx \quad \forall x \in \mathbf{R}^k \qquad [\Leftrightarrow L_{\mu,e}^TJ_kL_{\mu,e} = J_k]$$

and $L_{\mu,e}^{-1} = L_{\mu,-e}$.

Solution. Let $x = \begin{pmatrix} u \\ t \end{pmatrix} \in \mathbf{L}^k$. Denoting $s \equiv \begin{pmatrix} v \\ \tau \end{pmatrix} = L_{\mu,e}x$, we have

$$v = u - [\mu t - (\sqrt{1+\mu^2}-1)e^Tu]e, \quad \tau = \sqrt{1+\mu^2}t - \mu e^Tu$$
$$\Downarrow$$
$$\begin{aligned} \tau &= \sqrt{1+\mu^2}t - \mu e^Tu \geq \sqrt{1+\mu^2}t - \mu\|u\|_2 \geq 0 \\ &\quad [\text{since } t \geq \|u\|_2,\ \|e\|_2 = 1] \\ \tau^2 - v^Tv &= (1+\mu^2)t^2 - 2\mu\sqrt{1+\mu^2}te^Tu + \mu^2(e^Tu)^2 \\ &\quad -u^Tu - [\mu t - (\sqrt{1+\mu^2}-1)e^Tu]^2 \\ &\quad +2(e^Tu)[\mu t - (\sqrt{1+\mu^2}-1)e^Tu] \\ &= t^2 - u^Tu \qquad [\text{just arithmetic!}]. \end{aligned}$$

Thus, $x \in \mathbf{L}^k \Rightarrow L_{\mu,e}x \in \mathbf{L}^k$. To replace here $\Rightarrow$ with $\Leftrightarrow$, it suffices to verify (a straightforward computation!) that

$$L_{\mu,e}^{-1} = L_{\mu,-e},$$

so that both $L_{\mu,e}$ and its inverse map $\mathbf{L}^k$ onto itself. $\square$

Dikin ellipsoid

Exercise 6.8. Prove that if $\mathbf{K}$ is a canonical cone, K is the corresponding canonical barrier, and $X \in \mathrm{int}\mathbf{K}$, then the Dikin ellipsoid

$$W_X = \{Y \mid \|Y - X\|_X \le 1\} \qquad [\|H\|_X = \sqrt{\langle [\nabla^2 K(X)]H, H\rangle_E}]$$

is contained in $\mathbf{K}$.

Solution. According to the Hint, it suffices to verify the inclusion $W_X \subset \mathbf{K}$ in the following two particular cases:

A. $\mathbf{K} = \mathbf{S}^k_+$, $X = I_k$.

B. $\mathbf{K} = \mathbf{L}^k$, $X = \begin{pmatrix} 0_{k-1} \\ \sqrt{2} \end{pmatrix}$.

Note that in both cases $\nabla^2 K(X)$ is the unit matrix, so that all we need to prove is that the unit ball, centered at our particular X, is contained in our particular $\mathbf{K}$.

A: We should prove that if $\|H\|_F \le 1$, then $I + H \succeq 0$, which is evident. The modulae of eigenvalues of H are $\le \|H\|_F \le 1$, so that all these eigenvalues are ≥ -1.

B: We should prove that if $du \in \mathbf{R}^{k-1}$, $dt \in \mathbf{R}$ satisfy $dt^2 + du^T du \le 1$, then the point $\begin{pmatrix} 0_{k-1} \\ \sqrt{2} \end{pmatrix} + \begin{pmatrix} du \\ dt \end{pmatrix} = \begin{pmatrix} du \\ \sqrt{2} + dt \end{pmatrix}$ belongs to $\mathbf{L}^k$. In other words, we should verify that $\sqrt{2} + dt \ge 0$ (which is evident) and that $(\sqrt{2} + dt)^2 - du^T du \ge 0$. Here is the verification of the latter statement:

$$\begin{array}{rcl} (\sqrt{2} + dt)^2 - du^T du & = & 2 + 2\sqrt{2}dt + dt^2 - du^T du \\ & = & 1 + 2\sqrt{2}dt + 2dt^2 + (1 - dt^2 - du^T du) \\ & \ge & 1 + 2\sqrt{2}dt + 2dt^2 \\ & & [\text{since } dt^2 + du^T du \le 1] \\ & = & (1 + \sqrt{2}dt)^2 \ge 0. \qquad \square \end{array}$$

Exercise 6.9. Let $\mathbf{K}$ be a canonical cone:

$$\mathbf{K} = \mathbf{S}^{k_1}_+ \times \cdots \times \mathbf{S}^{k_p}_+ \times \mathbf{L}^{k_{p+1}} \times \cdots \times \mathbf{L}^{k_m} \subset E = \mathbf{S}^{k_1} \times \cdots \times \mathbf{S}^{k_p} \times \mathbf{R}^{k_{p+1}} \times \cdots \times \mathbf{R}^{k_m} \quad \text{(Cone)}$$

and let $X \in \mathrm{int}\mathbf{K}$. Prove the following:

2. Whenever $Y \in \mathbf{K}$, one has

$$\langle \nabla K(X), Y - X\rangle_E \le \theta(K).$$

4. The conic cap $\mathbf{K}_X$ is contained in the $\|\cdot\|_X$-ball, centered at X, of the radius $\theta(K)$:

$$Y \in \mathbf{K}_X \Rightarrow \|Y - X\|_X \le \theta(K).$$

Solution to 2 *and* 4. According to the Hint, it suffices to verify the statement in the case when $X = e(\mathbf{K})$ is the central point of $\mathbf{K}$. In this case the Hessian $\nabla^2 K(X)$ is just the unit matrix, whence, by Proposition 6.3.1, $\nabla K(X) = -X$.

2. We should prove that if $Y \in \mathbf{K}$, then $\langle \nabla K(X), Y - X\rangle_E \le \theta(K)$. The statement clearly is stable with respect to taking direct products, so that it suffices to prove it in the cases of $\mathbf{K} = \mathbf{S}^k_+$ and $\mathbf{K} = \mathbf{L}^k$.

In the case of $\mathbf{K} = \mathbf{S}^k_+$ what should be proved is

$$\forall H \in \mathbf{S}^k_+ : \quad \mathrm{Tr}(I_k - H) \le k,$$

which is evident.

In the case of $\mathbf{K} = \mathbf{L}^k$ what should be proved is

$$\forall \left(\begin{pmatrix} u \\ t \end{pmatrix} : \|u\|_2 \le t \right) \quad \sqrt{2}(\sqrt{2} - t) \le 2,$$

which again is evident.

4. We should prove that if $X = e(\mathbf{K})$, $\langle -\nabla K(X), X - Y\rangle_E \ge 0$, and $Y \in \mathbf{K}$, then $\|Y - X\|_E \le \theta(K)$.

Let $Y \in \mathbf{K}$ be such that $\langle -\nabla K(X), X - Y\rangle_E \ge 0$, i.e., such that $\langle X, X - Y\rangle_E \ge 0$. We may think of Y as of a collection of a block-diagonal symmetric positive semidefinite matrix H with diagonal blocks of the sizes $k_1, \dots, k_p$ and $m - p$ vectors $\begin{pmatrix} u_i \\ t_i \end{pmatrix} \in \mathbf{L}^{k_i}$, $i = p+1, \dots, m$ (see (Cone)); the condition $\langle X, X - Y\rangle_E \ge 0$ now becomes

$$\mathrm{Tr}(\underbrace{I - H}_{D}) + \sum_{i=p+1}^{m} \sqrt{2}(\sqrt{2} - t_i) \ge 0. \qquad (*)$$

We now have, denoting by D_j the eigenvalues of D and by $n = \sum_{i=1}^{p} k_i$ the row size of D:

$$\begin{aligned}
\|X - Y\|_X^2 &= \|X - Y\|_E^2 = \mathrm{Tr}((I - H)^2) + \sum_{i=p+1}^{m} \left((\sqrt{2} - t_i)^2 + u_i^T u_i \right) \\
&= \sum_{j=1}^{n} D_j^2 + \sum_{i=p+1}^{m} \left(2 - 2\sqrt{2} t_i + t_i^2 + u_i^T u_i \right) \\
&\le \sum_{j=1}^{n} D_j^2 + \sum_{i=p+1}^{m} \left(2 - 2\sqrt{2} t_i + 2 t_i^2 \right) \\
&\qquad\qquad [\text{since } u_i^T u_i \le t_i^2] \\
&= \sum_{j=1}^{n} D_j^2 + \sum_{i=p+1}^{m} \left(1 + (1 - \sqrt{2} t_i)^2 \right).
\end{aligned}$$

Denoting $q = m - p$, $D_{n+i} = 1 - \sqrt{2}t_{p+i}$, $i = 1, \ldots, q$, we come to the relation

$$\|X - Y\|_X^2 = q + \sum_{j=1}^{n+q} D_j^2, \tag{14}$$

while $(*)$ and relations $H \succeq 0$, $t_i \geq 0$ imply that

$$\begin{aligned} D_j &\leq 1, \quad j = 1, \ldots, n+q, \\ \sum_{j=1}^{n+q} D_j &\geq -q, \end{aligned} \tag{15}$$

Let A be the maximum of the right-hand side in (14) over D_j's satisfying (15), and let $D^* = (D_1^*, \ldots, D_{n+q}^*)^T$ be the corresponding maximizer (which clearly exists—(15) defines a compact set!). In the case of $n + q = 1$ we clearly have $A = 1 + q$. Now let $n + q > 1$. We claim that among $n + q$ coordinates of D^*, $n + q - 1$ are equal 1, and the remaining coordinate equals to $-(n + 2q - 1)$. Indeed, if there were at least two entries in D^* which are less than 1, then subtracting from one of them a $\delta \neq 0$ small enough in absolute value and adding the same δ to the other coordinate, we preserve the feasibility of the perturbed point with respect to (15), and, with properly chosen sign of δ, increase $\sum_j D_j^2$, which is impossible. Thus, at least $n + q - 1$ coordinates of D^* are equal to 1. Among the points with this property that satisfy (15), the one with the largest $\sum_j D_j^2$ clearly has the remaining coordinate equal to $1 - n - 2q$, as claimed.

From our analysis it follows that

$$A = \begin{cases} q + 1, & n + q = 1, \\ q + (n + q - 1) + (n + 2q - 1)^2 = (n + 2q - 1)(n + 2q), & n + q > 1. \end{cases}$$

Recalling that $\theta(K) = n + 2q$ and taking into account (14) and the origin of A, we get

$$\|X - Y\|_X \leq \theta(K),$$

as claimed. □

More on canonical barriers

Exercise 6.11 Prove that if $\mathbf{K}$ is a canonical cone, K is the associated canonical barrier, $X \in \mathrm{int}\mathbf{K}$, and $H \in \mathbf{K}$, $H \neq 0$, then

$$\inf_{t \geq 0} K(X + tH) = -\infty. \tag{$*$}$$

Derive from this fact that

(!!) *Whenever $\mathcal{N}$ is an affine plane that intersects the interior of $\mathbf{K}$, K is below bounded on the intersection $\mathcal{N} \cap \mathbf{K}$ if and only if the intersection is bounded.*

Solution. As explained in the Hint, it suffices to verify $(*)$ in the particular case when X is the central point of $\mathbf{K}$. It is also clear that $(*)$ is stable with respect to taking direct products, so that we can restrict ourselves with the cases of $\mathbf{K} = \mathbf{S}^k_+$ and $\mathbf{K} = \mathbf{L}^k$.

Case of $\mathbf{K} = \mathbf{S}^k_+$, $X = e(\mathbf{S}^k_+) = I_k$. Denoting by $H_i \geq 0$ the eigenvalues of H and noting that at least one of H_i is > 0 due to $H \neq 0$, we have for $t > 0$

$$K(X + tH) = -\ln \mathrm{Det}(I_k + tH) = -\sum_{i=1}^{k} \ln(1 + tH_i) \to -\infty, \ t \to \infty.$$

Case of $\mathbf{K} = \mathbf{L}^k$, $X = e(\mathbf{L}^k) = \begin{pmatrix} 0_{k-1} \\ \sqrt{2} \end{pmatrix}$. Setting $H = \begin{pmatrix} u \\ s \end{pmatrix}$, we have $s \geq \|u\|_2$ and $s > 0$ due to $H \neq 0$. For $t > 0$ we have

$$\begin{array}{rcl} K(X + tH) & = & -\ln((\sqrt{2} + ts)^2 - t^2 u^T u) \\ & = & -\ln(2 + 2\sqrt{2}ts + t^2(s^2 - u^T u)) \\ & \leq & -\ln(2 + 2\sqrt{2}ts) \to -\infty, \ t \to \infty. \end{array}$$

To derive (!!), note that if $U = \mathcal{N} \cap \mathbf{K}$ is bounded, then K is below bounded on U just in view of convexity of K. (Moreover, from the fact that K is a barrier for $\mathbf{K}$ it follows that K attains its minimum on U.) It remains to prove that if U is unbounded, then K is not below bounded on U. If U is unbounded, there exists a nonzero direction $H \in \mathbf{K}$ that is parallel to $\mathcal{N}$. (Take as H a limiting point of the sequence $\|Y_i - X\|_2^{-1}(Y_i - X)$, where $Y_i \in U$, $\|Y_i\|_2 \to \infty$ as $i \to \infty$, and X is a once for ever fixed point from U.) By $(*)$, K is not below bounded on the ray $\{X + tH \mid t \geq 0\}$, and this ray clearly belongs to U. Thus, K is not below bounded on U. □

Primal path-following method

Exercise 6.15. Looking at the data in the table at the end of section 6.5.3, do you believe that the corresponding method is exactly the short-step primal path-following method from Theorem 6.5.1 with the stepsize policy (6.5.31)?

Solution. The table cannot correspond to the indicated method. Indeed, we see from the table that the duality gap along the 12-iteration trajectory is reduced by factor of about 10^6. Since the duality gap in a short-step method is nearly inverse proportional to the value of the penalty, the latter in the process of our 12 iterations should be increased by a factor of order of 10^5–10^6. In our case $\theta(K) = 3$, and the policy (6.5.31) increases the penalty at an iteration by the factor $(1 + 0.1/\sqrt{3}) \approx 1.0577$. With this policy, in 12 iterations the penalty would be increased by $1.0577^{12} < 2$, which is very far from 10^5! □

Infeasible start path-following method

Exercise 6.18. Consider the problem

$$\max_X \left\{ \langle \widetilde{C}, Y \rangle_{\widetilde{E}} \mid Y \in (\mathcal{M} + R) \cap \widetilde{\mathbf{K}} \right\}, \qquad \text{(Aux)}$$

where

$$\widetilde{\mathbf{K}} = \mathbf{K} \times \mathbf{K} \times \underbrace{\mathbf{S}^1_+}_{=\mathbf{R}_+} \times \underbrace{\mathbf{S}^1_+}_{=\mathbf{R}_+}$$

(**K** is a canonical cone),

$$\mathcal{M} = \left\{ \begin{pmatrix} U \\ V \\ s \\ r \end{pmatrix} \;\middle|\; \begin{array}{c} U + rB \in \mathcal{L}, \\ V - rC \in \mathcal{L}^\perp, \\ \langle C, U\rangle_E - \langle B, V\rangle_E + s = 0 \end{array} \right\}$$

is a linear subspace in the space $\widetilde{E}$ where the cone $\widetilde{\mathbf{K}}$ lives, and

$$C \in \mathcal{L}, \quad B \in \mathcal{L}^\perp.$$

It is given that the problem (Aux) is feasible. Prove that the feasible set of (Aux) is unbounded.

Solution. According to the Hint, we should prove that the linear space $\mathcal{M}^\perp$ does not intersect int$\widetilde{\mathbf{K}}$.

Let us compute $\mathcal{M}^\perp$. A collection

$$\begin{pmatrix} \xi \\ \eta \\ s \\ r \end{pmatrix}, \ \xi, \eta \in E, \ s, r \in \mathbf{S}^1 = \mathbf{R}$$

is in $\mathcal{M}^\perp$ if and only if the linear equation in variables X, S, σ, τ

$$\langle X, \xi\rangle_E + \langle S, \eta\rangle_E + \sigma s + \tau r = 0$$

is a corollary of the system of linear equations

$$X + \tau B \in \mathcal{L}, \quad S - \tau C \in \mathcal{L}^\perp, \quad \langle X, C\rangle_E - \langle S, B\rangle_E + \sigma = 0.$$

By linear algebra, this is the case if and only if there exist $U \in \mathcal{L}^\perp$, $V \in \mathcal{L}$, and a real λ such that

$$\begin{array}{lrcl} \text{(a)} & \xi & = & U + \lambda C, \\ \text{(b)} & \eta & = & V - \lambda B, \\ \text{(c)} & s & = & \lambda, \\ \text{(d)} & r & = & \langle U, B\rangle_E - \langle V, C\rangle_E. \end{array} \qquad \text{(Pr)}$$

We have obtained a parameterization of $\mathcal{M}^\perp$ via the parameters U, V, λ running through, respectively, $\mathcal{L}^\perp$, $\mathcal{L}$, and **R**. Now assume, contrary to what should be proved, that the intersection of $\mathcal{M}^\perp$ and int$\widetilde{\mathbf{K}}$ is nonempty. In other words, assume that there exist $U \in \mathcal{L}^\perp$, $V \in \mathcal{L}$, and $\lambda \in \mathbf{R}$ which, being substituted in (Pr), result in a collection (ξ, η, s, r) such that $\xi \in$ int**K**, $\eta \in$ int**K**, $s > 0$,

$r > 0$. From (Pr)(c) it follows that $\lambda > 0$. Since (Pr) is homogeneous, we may normalize our U, V, λ to make $\lambda = 1$, still keeping $\xi \in \mathrm{int}\mathbf{K}$, $\eta \in \mathrm{int}\mathbf{K}$, $s > 0$, $r > 0$. Assuming $\lambda = 1$ and taking into account that $U, B \in \mathcal{L}^\perp$, $V, C \in \mathcal{L}$, we get from (Pr)(a)–(b)

$$\langle \xi, \eta \rangle_E = \langle C, V \rangle_E - \langle B, U \rangle_E.$$

Adding this equality to (Pr)(d), we get

$$\langle \xi, \eta \rangle_E + r = 0,$$

which is impossible, since both $r > 0$ and $\langle \xi, \eta \rangle_E > 0$. (Recall that the cone $\mathbf{K}$ is self-dual and $\xi, \eta \in \mathrm{int}\mathbf{K}$.) We have come to the desired contradiction. □

Exercise 6.19. Let $\bar{X}, \bar{S}$ be a strictly feasible pair of primal-dual solutions to the primal-dual pair of problems

$$\min_X \{\langle C, X \rangle_E \mid X \in (\mathcal{L} - B) \cap \mathbf{K}\}, \tag{P}$$
$$\max_S \left\{\langle B, S \rangle_E \mid S \in (\mathcal{L}^\perp + C) \cap \mathbf{K}\right\} \tag{D}$$

so that there exists $\gamma \in (0, 1]$ such that

$$\begin{array}{rcll} \gamma \|X\|_E & \leq & \langle \bar{S}, X \rangle_E & \forall X \in \mathbf{K}, \\ \gamma \|S\|_E & \leq & \langle \bar{X}, S \rangle_E & \forall S \in \mathbf{K}. \end{array}$$

Prove that if

$$Y = \begin{pmatrix} X \\ S \\ \sigma \\ \tau \end{pmatrix}$$

is feasible for (Aux), then

$$\begin{array}{rcl} \|Y\|_{\widetilde{E}} & \leq & \alpha\tau + \beta, \\ & & \alpha = \gamma^{-1}\left[\langle \bar{X}, C \rangle_E - \langle \bar{S}, B \rangle_E\right] + 1, \\ & & \beta = \gamma^{-1}\left[\langle \bar{X} + B, D \rangle_E + \langle \bar{S} - C, P \rangle_E + d\right]. \end{array} \tag{16}$$

Solution. We have $\bar{X} = \bar{U} - B$, $\bar{U} \in \mathcal{L}$, $\bar{S} = \bar{V} + C$, $\bar{V} \in \mathcal{L}^\perp$. Taking into account the constraints of (Aux), we get

$$\begin{array}{rcl} \langle \bar{U}, S - \tau C - D \rangle_E & = & 0 \Rightarrow \\ \langle \bar{U}, S \rangle_E & = & \langle \bar{U}, \tau C + D \rangle_E \Rightarrow \\ \langle \bar{X}, S \rangle_E & = & -\langle B, S \rangle_E + \langle \bar{U}, \tau C + D \rangle_E, \\ \langle \bar{V}, X + \tau B - P \rangle_E & = & 0 \Rightarrow \\ \langle \bar{V}, X \rangle_E & = & \langle \bar{V}, -\tau B + P \rangle_E \Rightarrow \\ \langle \bar{S}, X \rangle_E & = & \langle C, X \rangle_E + \langle \bar{V}, -\tau B + P \rangle_E, \\ \Rightarrow & & \\ \langle \bar{X}, S \rangle_E + \langle \bar{S}, X \rangle_E & = & [\langle C, X \rangle_E - \langle B, S \rangle_E] + \tau\left[\langle \bar{U}, C \rangle_E - \langle \bar{V}, B \rangle_E\right] \\ & & + \left[\langle \bar{U}, D \rangle_E + \langle \bar{V}, P \rangle_E\right] \\ & = & d - \sigma + \tau\left[\langle \bar{U}, C \rangle_E - \langle \bar{V}, B \rangle_E\right] \\ & & + \left[\langle \bar{U}, D \rangle_E + \langle \bar{V}, P \rangle_E\right] \end{array}$$

whence

$$\gamma\left[\|X\|_E + \|S\|_E\right] + \sigma \quad \leq \quad \tau\left[\langle \bar{U}, C\rangle_E - \langle \bar{V}, B\rangle_E\right] + \left[\langle \bar{U}, D\rangle_E + \langle \bar{V}, P\rangle_E + d\right],$$

and (16) follows (recall that $\langle C, B\rangle_E = 0$). □

Exercise 6.23. Let $\mathbf{K}$ be a canonical cone, let the primal-dual pair of problems

$$\min_X \{\langle C, X\rangle_E \mid X \in (\mathcal{L} - B) \cap \mathbf{K}\}, \qquad \text{(P)}$$
$$\max_S \left\{\langle B, S\rangle_E \mid S \in (\mathcal{L}^\perp + C) \cap \mathbf{K}\right\} \qquad \text{(D)}$$

be strictly primal-dual feasible and be normalized by $\langle C, B\rangle_E = 0$, let (X_*, S_*) be a primal-dual optimal solution to the pair, and let X, S ϵ-satisfy the feasibility and optimality conditions for (P), (D), i.e.,

$$\begin{array}{ll} \text{(a)} & X \in \mathbf{K} \cap (\mathcal{L} - B + \Delta X),\ \|\Delta X\|_E \leq \epsilon, \\ \text{(b)} & S \in \mathbf{K} \cap (\mathcal{L}^\perp + C + \Delta S),\ \|\Delta S\|_E \leq \epsilon, \\ \text{(c)} & \langle C, X\rangle_E - \langle B, S\rangle_E \leq \epsilon. \end{array}$$

Prove that

$$\begin{array}{rcl} \langle C, X\rangle_E - \mathrm{Opt(P)} & \leq & \epsilon(1 + \|X_* + B\|_E), \\ \mathrm{Opt(D)} - \langle B, S\rangle_E & \leq & \epsilon(1 + \|S_* - C\|_E). \end{array}$$

Solution. We have $S - C - \Delta S \in \mathcal{L}^\perp$, $X_* + B \in \mathcal{L}$, whence

$$\begin{array}{rcl} 0 & = & \langle S - C - \Delta S, X_* + B\rangle_E \\ & = & \underbrace{\langle S, X_*\rangle_E}_{\geq 0:\, X_*, S \in \mathbf{K} = \mathbf{K}_*} - \mathrm{Opt(P)} + \langle S, B\rangle_E + \langle -\Delta S, X_* + B\rangle_E \\ & & \Rightarrow \\ -\mathrm{Opt(P)} & \leq & -\langle S, B\rangle_E + \langle \Delta S, X_* + B\rangle_E \leq -\langle S, B\rangle_E + \epsilon\|X_* + B\|_E. \end{array}$$

Combining the resulting inequality and (c), we get the first of the inequalities to be proved. The second is given by symmetric reasoning. □

Solutions to Selected Exercises

Exercises to Lecture 1

Uniform approximation

Exercise 1.2. Let $\alpha < \infty$, and assume L is α-regular, i.e., the functions from L are continuously differentiable and

$$\|f'\|_\infty \le \alpha \|f\|_\infty \quad \forall f \in L.$$

Assume that $T \subset \Delta$ is such that the distance from a point in Δ to the closest point of T does not exceed $\beta < \alpha^{-1}$. Prove that under these assumptions

$$\kappa_L(T) \le \frac{1}{1-\alpha\beta}.$$

Solution. Let $f \in L$, $M = \|f\|_\infty$, and let $a \in \Delta$ be the point where $|f(a)| = M$. There exists a point $t \in T$ such that $|t-a| \le \beta$. Since L is regular, we have $|f(a) - f(t)| \le M\alpha\beta$, whence $|f(t)| \ge M(1-\alpha\beta)$, and consequently $\|f\|_{T,\infty} \ge |f(t)| \ge M(1-\alpha\beta)$. □

Exercise 1.5. Assume that $\Delta = [0, 2\pi]$, and let L be a linear space of functions on Δ comprising all trigonometric polynomials of degree $\le k$. Let T be an equidistant M-point grid on Δ:

$$T = \left\{ \frac{(2l+1)\pi}{M} \right\}_{l=0}^{M-1}.$$

1. Prove that if $M > k\pi$, then T is L-dense, with

$$\kappa_L(T) \le \frac{M}{M - k\pi}.$$

2. Prove that the above inequality remains valid if we replace T with its arbitrary shift modulo 2π, i.e., treat Δ as the unit circumference and rotate T by an angle.

3. Prove that if T is an arbitrary M-point subset of Δ with $M \le k$, then $\kappa_L(T) = \infty$.

Solution. 1. It suffices to apply the result of Exercise 1.2. In the case in question $\beta = \frac{\pi}{M}$, and, by the Bernshtein's Theorem, $\alpha = k$.

2. The solution follows from 1 because the space of trigonometric polynomials is invariant with respect to cyclic shift of the argument by any angle.

3. Let $T = \{t_i\}_{i=1}^{M}$. The function

$$f(t) = \prod_{i=1}^{M} \sin(t - t_i)$$

is a trigonometric polynomial of degree $M \le k$. This function vanishes on T (i.e., $\|f\|_{T,\infty} = 0$), although its uniform norm on Δ is positive. □

Exercise 1.6. Let $\Delta = [-1, 1]$ and let L be the space of all algebraic polynomials of degree $\le k$.

1. Assume that $2M > \pi k$ and T is the M-point set on Δ as follows:

$$T = \left\{ t_l = \cos\left(\frac{(2l+1)\pi}{2M} \right) \right\}_{l=0}^{M-1}.$$

Then T is L-dense, with

$$\kappa_L(T) \le \frac{2M}{2M - \pi k}.$$

2. Let T be an M-point set on Δ with $M \le k$. Then $\kappa_L(T) = \infty$.

Solution. 1. Let us pass from the functions $f \in L$ to the functions $f^+(\phi) = f(\cos(\phi))$, $\phi \in [0, 2\pi]$. Note that f^+ is a trigonometric polynomial of degree $\le k$. Let

$$T^+ = \left\{ \phi_l = \frac{(2l+1)\pi}{2M} \right\}_{l=0}^{2M-1}.$$

According to the result of Exercise 1.5.1, for every $f \in L$ we have

$$\begin{aligned} \|f\|_\infty &= \|f^+\|_\infty \\ &\le \tfrac{2M}{2M-\pi k} \max_{l=0,\dots,2M-1} |f^+(\phi_l)| \\ &= \tfrac{2M}{2M-\pi k} \max_{l=0,\dots,2M-1} |f(\cos(\phi_l))| \\ &= \tfrac{2M}{2M-\pi k} \max_{l=0,\dots,M-1} |f(t_l)|. \end{aligned}$$

(Note that when ϕ takes values in T^+, the quantity $\cos(\phi)$ takes values in T.)

2. Whenever the cardinality of T is $\le k$, L contains a nontrivial polynomial

$$f(t) = \prod_{t' \in T} (t - t')$$

which vanishes on T. □

Integration formulas and Gauss points

(∗) *Let Δ be a subset of $\mathbf{R}^k$, let L be an n-dimensional linear space comprising continuous real-valued functions on Δ, and let $I(f) : L \to \mathbf{R}$ be an integral—a linear functional such that $I(f) \geq 0$ for every $f \in L$ such that $f(t) \geq 0$ everywhere on Δ. Assume also that if a function $f \in L$ is nonnegative on Δ and is not identically 0, then $I(f) > 0$. Then there exists a precise n-point cubature formula for I, i.e., there are n points $t_1, \ldots, t_n \in \Delta$ and n nonnegative weights $\alpha_1, \ldots, \alpha_n$ such that*

$$I(f) = \sum_{i=1}^{n} \alpha_i f(t_i) \quad \forall f \in L.$$

Exercise 1.17. 1. Prove (*) for the case of finite Δ.

Solution. We may assume that I is not identically zero—otherwise there is nothing to prove. The sets $A_t = \{f \in L \mid f(t) \leq 0, I(f) = 1\}$ are convex sets belonging to a fixed hyperplane. We claim that there exist n sets of this type $A_{t_1}, \ldots, A_{t_n}$ with an empty intersection. Indeed, in the opposite case, by the Helley theorem, there would exist an element $f \in L$ belonging to all A_t, $t \in \Delta$, i.e., such that $f(t) \leq 0$ everywhere on Δ and $I(f) = 1$, which is impossible.

The fact that the intersection of $A_{t_1}, \ldots, A_{t_n}$ is empty means that the inequality $I(f) \leq 0$ is a corollary of the system of linear inequalities $f(t_i) = 0$, $i = 1, \ldots, n$. By the Farkas Lemma, it follows that the linear functional I is a combination, with nonnegative coefficients, of the n linear functionals $f \mapsto f(t_i)$. □

2. Prove (*) for the general case.

Solution. Introducing the same sets A_t as before, it suffices to prove that there exists a finite family of these sets with an empty intersection. Indeed, if we know that $A_{t_1}, \ldots, A_{t_N}$ have no point in common, then, by the Helley theorem, already a properly chosen n-element subfamily of this family of sets has an empty intersection, and we may complete the proof in the same way as in the case of finite Δ. To prove that there exists a finite family of the sets A_t with an empty intersection, assume that it is not the case, and let us lead this assumption to a contradiction. Let $t_1, t_2, \ldots$ be a sequence of points from Δ such that every point of this set belongs to the closure of the sequence, and let $f_1, \ldots, f_n$ be a basis in L. Under our assumption, for every N there exists a function $f^N \in L$ that is nonpositive at the points $t_1, \ldots, t_N$ and $I(f^N) = 1$. After an appropriate normalization, we can convert f^N to a function $g^N \in L$ such that g^N is nonpositive at the points $t_1, \ldots, t_N$, $I(g^N) \geq 0$, and the Euclidean norm of the vector λ^N of the coefficients of g^N in the basis $f_1, \ldots, f_n$ is 1. Passing to a subsequence $\{N_i\}$, we may assume that the vectors λ^{N_i} converge to a vector λ (which of course is nonzero). Then the functions g^{N_i}

pointwise converge to the function $g(t) = \sum_{j=1}^{n} \lambda_j f_j(t)$; this function clearly is nonpositive on the sequence $t_1, t_2, \ldots$, and since it is continuous on Δ, it is nonpositive everywhere. Since $0 \leq I(g^{N_i}) = \sum_{j=1}^{n} \lambda_j^{N_i} I(f_j)$, we conclude that $I(g) = \sum_{j=1}^{n} \lambda_j I(f_j) \geq 0$. Thus, g is a nonpositive on Δ function with nonnegative integral $I(g)$. At the same time g is not identically zero (since its vector of coefficients in the basis $f_1, \ldots, f_n$ is nonzero). This is the desired contradiction. □

Exercises to Lecture 2

Cones

Exercise 2.4.

2. Let $\mathbf{K}$ be a cone in $\mathbf{R}^n$ and $u \mapsto Au$ be a linear mapping from certain $\mathbf{R}^k$ to $\mathbf{R}^n$ with trivial null space and such that $\mathrm{Im}A \cap \mathrm{int}\mathbf{K} \neq \emptyset$. Prove that the inverse image of $\mathbf{K}$ under the mapping—the set

$$A^{-1}(\mathbf{K}) = \{u \mid Au \in \mathbf{K}\}$$

—is a cone in $\mathbf{R}^k$. Prove that the cone dual to $A^{-1}(\mathbf{K})$ is the image of $\mathbf{K}_*$ under the mapping $\lambda \mapsto A^T\lambda$:

$$(A^{-1}(\mathbf{K}))_* = \{A^T\lambda \mid \lambda \in \mathbf{K}_*\}.$$

Solution. The fact that $A^{-1}(K)$ is a cone is evident. Let us justify the announced description of the cone dual to $A^{-1}(\mathbf{K})$. If $\lambda \in \mathbf{K}_*$, then $A^T\lambda$ clearly belongs to $(A^{-1}(\mathbf{K}))_*$:

$$u \in A^{-1}(\mathbf{K}) \Rightarrow Au \in \mathbf{K} \Rightarrow \lambda^T(Au) = (A^T\lambda)^T u \geq 0.$$

Now let us prove the inverse implication: if $c \in (A^{-1}(\mathbf{K}))_*$, then $c = A^T\lambda$ for some $\lambda \in \mathbf{K}_*$. To this end consider the conic problem

$$\min_x \left\{c^T x \mid Ax \geq_{\mathbf{K}} 0\right\}.$$

The problem is strictly feasible and below bounded (why?), so that by the conic duality theorem the dual problem

$$\max_\lambda \left\{0^T\lambda \mid A^T\lambda = c, \lambda \geq_{\mathbf{K}_*} 0\right\}$$

is solvable. □

3. Let $\mathbf{K}$ be a cone in $\mathbf{R}^n$ and $y = Ax$ be a linear mapping from $\mathbf{R}^n$ onto $\mathbf{R}^N$ (i.e., the image of A is the entire $\mathbf{R}^N$). Assume $\mathrm{Null}(A) \bigcap \mathbf{K} = \{0\}$.

Prove then that the image of $\mathbf{K}$ under the mapping A—the set

$$A\mathbf{K} = \{Ax \mid x \in \mathbf{K}\}$$

—is a cone in $\mathbf{R}^N$.

Prove that the cone dual to $A(\mathbf{K})$ is

$$(A\mathbf{K})_* = \{\lambda \in \mathbf{R}^N \mid A^T\lambda \in \mathbf{K}_*\}.$$

Demonstrate by example that if in the above statement the assumption $\mathrm{Null}(A) \cap \mathbf{K} = \{0\}$ is weakened to $\mathrm{Null}(A) \cap \mathrm{int}\mathbf{K} = \emptyset$, then the image of $\mathbf{K}$ under the mapping A may happen to be nonclosed.

Solution. Let us temporarily set $B = A^T$ (note that B has trivial null space, since A is an onto mapping) and

$$\mathbf{L} = \{\lambda \in \mathbf{R}^N \mid B\lambda \in \mathbf{K}_*\}.$$

Let us prove that the image of B intersects the interior of $\mathbf{K}_*$. Indeed, otherwise we could separate the convex set $\mathrm{int}\mathbf{K}_*$ from the linear subspace $\mathrm{Im}\, B$: there would exist $x \neq 0$ such that

$$\inf_{\mu \in \mathrm{int}\mathbf{K}_*} x^T\mu \geq \sup_{\mu \in \mathrm{Im}\, B} x^T\mu,$$

whence $x \in (\mathbf{K}_*)_* = \mathbf{K}$ and $x \in (\mathrm{Im}\, B)^\perp = \mathrm{Null}(A)$, which is impossible.

It remains to apply to the cone $\mathbf{K}_*$ and the mapping B the rule on inverse image (rule 2). According to this rule, the set $\mathbf{L}$ is a cone, and its dual cone is the image of $(\mathbf{K}_*)_* = \mathbf{K}$ under the mapping $B^T = A$. Thus, $A(\mathbf{K})$ indeed is a cone, namely, the cone dual to $\mathbf{L}$, whence the cone dual to $A(\mathbf{K})$ is $\mathbf{L}$.

"Demonstrate by example…": When the 3D ice cream cone is projected onto its tangent plane, the projection is the open half-plane plus a single point on the boundary of this half-plane, which is not a closed set. □

Exercise 2.5. Let A be an $m \times n$ matrix of full column rank and $\mathbf{K}$ be a cone in $\mathbf{R}^m$.

1. Prove that at least one of the following facts always takes place:

   (i) There exists a nonzero $x \in \mathrm{Im}\, A$ which is $\geq_{\mathbf{K}} 0$.

   (ii) There exists a nonzero $\lambda \in \mathrm{Null}(A^T)$ which is $\geq_{\mathbf{K}_*} 0$.

   Geometrically: Given a primal-dual pair of cones $\mathbf{K}$, $\mathbf{K}_*$ and a pair L, $L^\perp$ of linear subspaces that are orthogonal complements of each other, we can find a nontrivial ray in the intersection $L \cap \mathbf{K}$ or in the intersection $L^\perp \cap \mathbf{K}_*$ or both.

2. Prove that the strict version of (ii) takes place (i.e., there exists $\lambda \in \mathrm{Null}(A^T)$ which is $>_{\mathbf{K}_*} 0$) if and only if (i) does not take place, and vice versa: the strict version of (i) takes place if and only if (ii) does not take place.

   Geometrically: If $\mathbf{K}$, $\mathbf{K}_*$ is a primal-dual pair of cones and L, $L^\perp$ are linear subspaces that are orthogonal complements of each other, then the intersection $L \cap \mathbf{K}$ is trivial (is the singleton $\{0\}$) if and only if the intersection $L^\perp \cap \mathrm{int}\mathbf{K}_*$ is nonempty.

Solution. 1. Assuming that (ii) does not take place, note that $A^T\mathbf{K}_*$ is a cone in $\mathbf{R}^n$ (Exercise 2.4), and the dual of this latter cone is the inverse image of $\mathbf{K}$ under the mapping A. Since the dual must contain nonzero vectors, (i) takes place.

2. Let $e >_{\mathbf{K}_*} 0$. Consider the conic problem

$$\max_{\lambda,t} \left\{ t \mid A^T\lambda = 0, \lambda - te \geq_{\mathbf{K}_*} 0 \right\}.$$

Note that this problem is strictly feasible, and the strict version of (ii) is equivalent to the fact that the optimal value in the problem is > 0. Thus, (ii) is not valid if and only if the optimal value in our strictly feasible maximization conic problem is ≤ 0. By the conic duality theorem this is the case if and only if the dual problem

$$\min_{z,\mu} \left\{ 0 \mid Az + \mu = 0, \mu^T e = 1, \mu \geq_{\mathbf{K}} 0 \right\}$$

is solvable with optimal value equal to 0, which clearly is the case if and only if the intersection of ImA and $\mathbf{K}$ is not the singleton $\{0\}$. □

Feasible and level sets of conic problems

Exercise 2.11. Let the problem

$$\min \left\{ c^T x \mid Ax - b \geq_{\mathbf{K}} 0 \right\} \tag{CP}$$

be feasible (A is of full column rank). Then the following properties are equivalent:

(i) The feasible set of the problem is bounded.

(ii) The set of primal slacks $K = \{y \geq_{\mathbf{K}} 0, y = Ax - b\}$ is bounded.[76]

(iii) Im $A \cap \mathbf{K} = \{0\}$.

(iv) The system of vector inequalities

$$A^T\lambda = 0, \lambda >_{\mathbf{K}_*} 0$$

is solvable.

Corollary. The property of (CP) to have bounded feasible set is independent of the particular value of b such that (CP) is feasible!

Solution. (i) $\Leftrightarrow$ (ii): This is an immediate consequence of **A**.

(ii) $\Rightarrow$ (iii). If there exists $0 \neq y = Ax \geq_{\mathbf{K}} 0$, then the set of primal slacks contains, along with any of its points $\bar{y}$, the entire ray $\{\bar{y} + ty \mid t > 0\}$, so that the set of primal slacks is unbounded. (Recall that it is nonempty—(CP) is feasible!) Thus, (iii) follows from (ii) (by contradiction).

(iii) $\Rightarrow$ (iv). See Exercise 2.5.2.

(iv) $\Rightarrow$ (ii). Let λ be given by (iv). For all primal slacks $y = Ax - b \in \mathbf{K}$ one has $\lambda^T y = \lambda^T[Ax - b] = (A^T\lambda)^T x - \lambda^T b = \lambda^T b$, and it remains to use the result of Exercise 2.3.2.

[76] Recall that we always assume that **A** holds!

Exercise 2.12. Let the problem

$$\min\left\{c^T x \mid Ax - b \geq_{\mathbf{K}} 0\right\} \tag{CP}$$

be feasible (A is of full column rank). Prove that the following two conditions are equivalent:

(i) (CP) has bounded level sets.

(ii) The dual problem

$$\max\left\{b^T\lambda \mid A^T\lambda = c, \lambda \geq_{\mathbf{K}_*} 0\right\}$$

is strictly feasible.

Corollary. The property of (CP) to have bounded level sets is independent of the particular value of b such that (CP) is feasible!

Solution. (i) $\Rightarrow$ (ii). Consider the linear vector inequality

$$\bar{A}x \equiv \begin{pmatrix} Ax \\ -c^T x \end{pmatrix} \geq_{\bar{\mathbf{K}}} 0,$$
$$\bar{\mathbf{K}} = \{(x,t) \mid x \geq_{\mathbf{K}} 0, t \geq 0\}.$$

If $\bar{x}$ is a solution of this inequality and x is a feasible solution to (CP), then the entire ray $\{x + t\bar{x} \mid t \geq 0\}$ is contained in the same level set of (CP) as x. Consequently, in the case of (i) the only solution to the inequality is the trivial solution $\bar{x} = 0$. In other words, the intersection of Im $\bar{A}$ with $\bar{\mathbf{K}}$ is trivial—{0}. Applying the result of Exercise 2.5.2, we conclude that the system

$$\bar{A}^T\begin{pmatrix}\lambda\\ \mu\end{pmatrix} \equiv A^T\lambda - \mu c = 0, \begin{pmatrix}\lambda\\ \mu\end{pmatrix} >_{\bar{\mathbf{K}}_*} 0$$

is solvable; if (λ, μ) solves the latter system, then $\lambda >_{\mathbf{K}_*} 0$ and $\mu > 0$, so that $\mu^{-1}\lambda$ is a strictly feasible solution to the dual problem.

(ii) $\Rightarrow$ (i). If $\lambda >_{\mathbf{K}_*} 0$ is feasible for the dual problem and x is feasible for (CP), then

$$\lambda^T(Ax - b) = (A^T\lambda)^T x - \lambda^T b = c^T x - \lambda^T b.$$

We conclude that if x runs through a given level set $\mathcal{L}$ of (CP), the corresponding slacks $y = Ax - b$ belong to a set of the form $\{y \geq_{\mathbf{K}} 0, \lambda^T y \leq \text{const}\}$. The sets of the latter type are bounded in view of the result of Exercise 2.3.2 (recall that $\lambda >_{\mathbf{K}_*} 0$). It remains to note that in view of **A** boundedness of the image of $\mathcal{L}$ under the mapping $x \mapsto Ax - b$ implies boundedness of $\mathcal{L}$. □

Exercises to Lecture 3

Optimal control in discrete time linear dynamic system. Consider a discrete time linear dynamic system

$$\begin{array}{rcl} x(t) & = & A(t)x(t-1) + B(t)u(t),\ t = 1, 2, \ldots, T, \\ x(0) & = & x_0. \end{array} \tag{S}$$

Here,

- t is the (discrete) time.
- $x(t) \in \mathbf{R}^l$ is the *state* vector: its value at instant t identifies the state of the controlled plant.
- $u(t) \in \mathbf{R}^k$ is the exogeneous input at time instant t; $\{u(t)\}_{t=1}^T$ is the *control*.
- For every $t = 1, \ldots, T$, $A(t)$ is a given $l \times l$, and $B(t)$ – a given $l \times k$ matrices.

A typical problem of optimal control associated with (S) is to minimize a given functional of the trajectory $x(\cdot)$ under given restrictions on the control. As a simple problem of this type, consider the optimization model

$$\min_x \left\{ c^T x(T) \mid \frac{1}{2} \sum_{t=1}^T u^T(t) Q(t) u(t) \le w \right\}, \tag{OC}$$

where $Q(t)$ are given positive definite symmetric matrices.

Exercise 3.1. 1. Use (S) to express $x(T)$ via the control and convert (OC) in a quadratically constrained problem with linear objective with respect to the u-variables.

Solution. From (S) it follows that

$$\begin{aligned}
x(1) &= A(1)x_0 + B(1)u(1); \\
x(2) &= A(2)x(1) + B(2)u(2) \\
&= A(2)A(1)x_0 + B(2)u(2) + A(2)B(1)u(1); \\
&\cdots \\
x(T) &= A(T)A(T-1)\ldots A(1)x_0 + \sum_{t=1}^T A(T)A(T-1)\ldots A(t+1)B(t)u(t) \\
&\equiv A(T)A(T-1)\ldots A(1)x_0 + \sum_{t=1}^T C(t)u(t), \\
&\quad C(t) = A(T)A(T-1)\ldots A(t+1)B(t).
\end{aligned}$$

Consequently, (OC) is equivalent to the problem

$$\min_{u(\cdot)} \left\{ \sum_{t=1}^T d_t^T u(t) \mid \frac{1}{2} \sum_{i=1}^T u^T(t) Q(t) u(t) \le w \right\} \quad [d_t = C^T(t)c]. \qquad \square \quad (*)$$

2. Convert the resulting problem to a conic quadratic program

Solution.

$$\begin{array}{ll}\text{minimize} & \displaystyle\sum_{t=1}^{T} d_t^T u(t)\\ \text{s.t.} & \\ & \begin{pmatrix} 2^{1/2}Q^{1/2}(t)u(t)\\ 1-s(t)\\ 1+s(t)\end{pmatrix} \ge_{\mathbf{L}^k} 0,\ t=1,\dots,T,\\ & \displaystyle\sum_{t=1}^{T} s(t) \le w;\end{array}$$

the design variables are $\{u(t)\in\mathbf{R}^k, s(t)\in\mathbf{R}\}_{t=1}^T$. □

3. Pass to the resulting problem to its dual and find the optimal solution to the latter problem.

Solution. The conic dual is

$$\begin{array}{ll}\text{maximize} & -w\omega - \displaystyle\sum_{t=1}^{T}[\mu(t)+\nu(t)]\\ \text{s.t.} & \\ & 2^{1/2}Q^{1/2}(t)\xi(t) = d_t,\ t=1,\dots,T\\ & \qquad [\Leftrightarrow \xi(t)=2^{-1/2}Q^{-1/2}(t)d_t],\\ & -\omega-\mu(t)+\nu(t) = 0,\ t=1,\dots,T\\ & \qquad [\Leftrightarrow \mu(t)=\nu(t)-\omega],\\ & \sqrt{\xi^T(t)\xi(t)+\mu^2(t)} \le \nu(t),\ t=1,\dots,T;\end{array}$$

the variables are $\{\xi(t)\in\mathbf{R}^k, \mu(t), \nu(t), \omega\in\mathbf{R}\}$. Equivalent reformulation of the dual problem is

$$\begin{array}{ll}\text{minimize} & w\omega + \displaystyle\sum_{t=1}^{T}[2\nu(t)-\omega]\\ \text{s.t.} & \\ & \sqrt{a_t^2+(\nu(t)-\omega)^2} \le \nu(t),\ t=1,\dots,T\\ & \qquad [a_t^2 = 2^{-1}d_t^T Q^{-1}(t)d_t],\end{array}$$

or, which is the same,

$$\begin{array}{ll}\text{minimize} & w\omega + \displaystyle\sum_{t=1}^{T}[2\nu(t)-\omega]\\ \text{s.t.} & \\ & \omega(2\nu(t)-\omega) \ge a_t^2,\ t=1,\dots,T,\end{array}$$

or, which is the same,

$$\min_{\omega}\left\{w\omega+\omega^{-1}\left(\sum_{t=1}^{T}a_t^2\right)\mid\ \omega>0\right\}.$$

It follows that the optimal solution to the dual problem is

$$\begin{array}{rcl}\omega_* &=& \sqrt{w^{-1}\left(\sum_{i=1}^{T}a_t^2\right)};\\ \nu_*(t) &=& \frac{1}{2}\left[a_t^2\omega_*^{-1}+\omega_*\right],\ t=1,\ldots,T;\\ \mu_*(t) &=& \nu_*(t)-\omega_*\\ &=& \frac{1}{2}\left[a_t^2\omega_*^{-1}-\omega_*\right],\ t=1,\ldots,T.\end{array}\qquad\square$$

Stable grasp. Recall that the stable grasp analysis problem is to check whether the system of constraints

$$\begin{array}{rcl}\|F^i\|_2 &\le& \mu(f^i)^Tv^i,\ i=1,\ldots,N,\\ (v^i)^TF^i &=& 0,\ i=1,\ldots,N,\\ \sum_{i=1}^{N}(f^i+F^i)+F^{\text{ext}} &=& 0,\\ \sum_{i=1}^{N}p^i\times(f^i+F^i)+T^{\text{ext}} &=& 0\end{array}\qquad\text{(SG)}$$

in the 3D vector variables F^i is or is not solvable. Here the data are given by a number of 3D vectors, namely,

- vectors v^i—unit inward normals to the surface of the body at the contact points;
- contact points p^i;
- vectors f^i—contact forces;
- vectors F^{ext} and T^{ext} of the external force and torque, respectively.

$\mu>0$ is a given friction coefficient; we assume that $f_i^Tv^i>0\ \forall i$.

Exercise 3.6. 1. Regarding (SG) as the system of constraints of a maximization program with trivial objective and applying the technique from section 2.5, build the dual problem.

Solution.

$$\begin{array}{ll}\text{minimize} & \displaystyle\sum_{i=1}^N \mu[(f^i)^T v^i]\phi_i - \left[\sum_{i=1}^N f^i + F^{\text{ext}}\right]^T \Phi - \left[\sum_{i=1}^N p^i \times f^i + T^{\text{ext}}\right]^T \Psi \\ \text{s.t.} & \Phi_i + \sigma_i v^i + \Phi - p^i \times \Psi = 0,\ i = 1, \ldots, N \\ & \|\Phi_i\|_2 \le \phi_i,\ i = 1, \ldots, N, \\ & \qquad [\Phi, \Phi_i, \Psi \in \mathbf{R}^3, \sigma_i, \phi_i \in \mathbf{R}]\end{array}$$

$$\Updownarrow$$

$$\min_{\sum_i \in \mathbf{R}, \Psi, \Phi \in \mathbf{R}^3} \left\{\sum_{i=1}^N \mu[(f^i)^T v^i]\|p^i \times \Psi - \Phi - \sigma_i v^i\|_2 - F^T\Phi - T^T\Psi\right\},$$

$$F = \sum_{i=1}^N f^i + F^{\text{ext}},\ \ T = \sum_{i=1}^N p^i \times f^i + T^{\text{ext}}. \qquad \square$$

Trusses. We are about to process the multiload TTD problem 3.4.1, which we write as (see (3.4.57))

$$\begin{array}{lrcl}\text{minimize} & \tau & & \\ \text{s.t.} & s_{ij}^2 & \le & 4t_i r_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, k, \\ & \displaystyle\sum_{i=1}^n r_{ij} & \le & \tfrac{1}{2}\tau,\ j = 1, \ldots, k, \\ & \displaystyle\sum_{i=1}^n t_i & \le & w, \\ & \displaystyle\sum_{i=1}^n s_{ij} b_i & = & f_j,\ j = 1, \ldots, k, \\ & t_i, r_{ij} & \ge & 0,\ i = 1, \ldots, n,\ j = 1, \ldots, k;\end{array} \qquad \text{(Pr)}$$

the design variables are $s_{ij}, r_{ij}, t_i, \tau$.

We assume that

- the ground structure $(n, m, b_1, \ldots, b_n)$ is such that the matrix $\sum_{i=1}^n b_i b_i^T$ is positive definite; and
- the loads of interest $f_1, \ldots, f_k$ are nonzero, and the material resource w is positive.

Exercise 3.7. 1. Applying the technique from section 2.5, build the problem (Dl) dual to the problem (Pr).

What is the design dimension of (Pr)? of (Dl)?

Solution.

$$
\begin{array}{lrcl}
\text{maximize} & -w\rho - \displaystyle\sum_{j=1}^{k} f_j^T v_j & & \\
\text{s.t.} & \alpha_{ij} + b_i^T v_j & = & 0,\ i = 1, \dots, n,\ j = 1, \dots, k; \\
& \gamma_{ij} - \beta_{ij} - \delta_j & = & 0,\ i = 1, \dots, n,\ j = 1, \dots, k; \\
& \displaystyle\sum_{j=1}^{k} [\beta_{ij} + \gamma_{ij}] - \rho & = & 0,\ i = 1, \dots, n; \\
& \frac{1}{2} \displaystyle\sum_{j=1}^{n} \delta_j & = & 1, \\
& \sqrt{\alpha_{ij}^2 + \beta_{ij}^2} & \le & \gamma_{ij},\ i = 1, \dots, n,\ j = 1, \dots, k; \\
& \delta_j & \ge & 0,\ j = 1, \dots, k; \\
& \rho & \ge & 0 \\
& & & [\alpha_{ij}, \beta_{ij}, \gamma_{ij}, \delta_j, \rho \in \mathbf{R}, v_j \in \mathbf{R}^m]
\end{array}
$$

$$\Updownarrow$$

$$
\begin{array}{lrcl}
\text{minimize} & w\rho + \displaystyle\sum_{j=1}^{k} f_j^T v_j & & \\
\text{s.t.} & \displaystyle\sum_{j=1}^{k} \lambda_j^{-1} (b_i^T v_j)^2 & \le & 2\rho,\ i = 1, \dots, n \\
& & & [\lambda_j = \delta_j/2]; \\
& \displaystyle\sum_{j=1}^{k} \lambda_j & = & 1; \\
& \lambda_j & \ge & 0,\ j = 1, \dots, k; \\
& \rho & \ge & 0 \\
& & & [\rho, \lambda_j \in \mathbf{R}, v_j \in \mathbf{R}^m].
\end{array}
$$

The design dimension of (Pr) is $2nk + n + 1$. The design dimension of (Dl) is $mk + k + 1$. □

Exercise 3.8. Let us fix a ground structure $(n, m, b_1, \dots, b_n)$ and a material resource w, and let $\mathcal{F}$ be a finite set of loads.

1. Assume that $\mathcal{F}_j \in \mathcal{F}$, $j = 1, \dots, k$, are subsets of $\mathcal{F}$ with $\cup_{j=1}^{k} \mathcal{F}_j = \mathcal{F}$. Let μ_j be the optimal value in the multiload TTD problem with the set of loads $\mathcal{F}_j$ and μ be the optimal value in the multiload TTD problem with the set of loads $\mathcal{F}$. Is it possible that $\mu > \sum_{j=1}^{k} \mu_j$?

Solution. The compliance $\mathrm{Compl}_f(t)$ clearly is nonincreasing in t and $\mathrm{Compl}_f(\theta t) = \theta^{-1}\mathrm{Compl}_f(t)$ for positive θ. Let t^j be an optimal solution to the multiload TTD problem with the set of loads $\mathcal{F}_j$, and let θ_j be positive reals with sum 1. The truss $t_\theta = \sum_{j=1}^k \theta_j t^j$ clearly satisfies the resource constraint, while for every $f \in \mathcal{F}_j$ one has

$$\mathrm{Compl}_f(t_\theta) \le \mathrm{Compl}_f(\theta_j t^j) = \theta_j^{-1}\mathrm{Compl}_f(t^j) \le \theta_j^{-1}\mu_j.$$

Setting $\theta_j = \mu_j(\sum_{l=1}^k \mu_l)^{-1}$, we get $\mathrm{Compl}_f(t_\theta) \le \sum_{j=1}^m \mu_j \; \forall f \in \cup_{j=1}^k \mathcal{F}_j = \mathcal{F}$. Thus, $\mu \le \sum_{j=1}^k \mu_j$. □

2. Assume that the ground structure includes $n = 1998$ tentative bars and that you are given a set $\mathcal{F}$ of $N = 1998$ loads. It is known that for every subset $\mathcal{F}'$ of $\mathcal{F}$ made up of no more than 999 loads, the optimal value in the multiload TTD problem, the set of loading scenarios being $\mathcal{F}'$, does not exceed 1. What can be said about the optimal value in the multiload TTD problem with the set of scenarios $\mathcal{F}$?

Solution. It does not exceed 2 (by the previous exercise). □

Answer a similar question in the case when $\mathcal{F}$ comprises $N' = 19980$ loads.

Solution. The optimal value still does not exceed 2. Indeed, let T_f, $t \in \mathcal{F}$, be the set of all trusses of given volume with compliances with respect to f not exceeding 2. Then T_f is a convex subset in 1997-dimensional affine plane (cut off the 1998-dimensional space of trusses by the linear equation "volume of the truss is given"). By the previous answer for every subset of $\mathcal{F}$ comprising 1998 loads, there exists a truss of given volume with compliance with respect to the subset not exceeding 2, i.e., every 1998 convex set from the family $\{T_f\}_{f\in\mathcal{F}}$ has a point in common. It follows, by the Helley theorem, that all sets from the family have a point in common, i.e., there exists a truss with compliance with respect to every load from $\mathcal{F}$ not exceeding 2. □

Does conic quadratic programming exist? Let $\epsilon > 0$ and a positive integer n be given. We intend to build a polyhedral ϵ-approximation of the Lorentz cone $\mathbf{L}^{n+1}$. Without loss of generality we may assume that n is an integer power of 2: $n = 2^\kappa$, $\kappa \in \mathbf{N}$.

Tower of variables. The first step of our construction is quite straightforward: we introduce extra variables to represent a conic quadratic constraint

$$\sqrt{y_1^2 + \cdots + y_n^2} \le t \tag{CQI}$$

of dimension $n + 1$ by a system of conic quadratic constraints of dimension three each. Namely, let us call our original y-variables variables of generation 0 and let us split them into pairs $(y_1, y_2), \ldots, (y_{n-1}, y_n)$. We associate with every one of these pairs its successor—an additional variable of generation 1. We split the resulting $2^{\kappa-1}$ variables of generation 1

into pairs and associate with every pair its successor—an additional variable of generation 2, and so on. After $\kappa-1$ steps we end up with two variables of the generation $\kappa-1$. Finally, the only variable of generation κ is the variable t from (CQI).

To introduce convenient notation, let us denote by y_i^ℓ the ith variable of generation ℓ, so that $y_1^0,\dots,y_n^0$ are our original y-variables $y_1,\dots,y_n$, $y_1^\kappa\equiv t$ is the original t-variable, and the parents of y_i^ℓ are the variables $y_{2i-1}^{\ell-1}$, $y_{2i}^{\ell-1}$.

Note that the total number of all variables in the tower of variables we end up with is $2n-1$.

It is clear that the system of constraints

$$\sqrt{[y_{2i-1}^{\ell-1}]^2+[y_{2i}^{\ell-1}]^2}\le y_i^\ell,\ \ i=1,\dots,2^{\kappa-\ell},\ \ \ell=1,\dots,\kappa, \tag{1}$$

is a representation of (CQI) in the sense that a collection $(y_1^0\equiv y_1,\dots,y_n^0\equiv y_n,y_1^\kappa\equiv t)$ can be extended to a solution of (1) if and only if (y,t) solves (CQI). Moreover, let $\Pi_\ell(x_1,x_2,x_3,u^\ell)$ be polyhedral ϵ_ℓ-approximations of the cone

$$\mathbf{L}^3=\left\{(x_1,x_2,x_3)\mid \sqrt{x_1^2+x_2^2}\le x_3\right\},$$

$\ell=1,\dots,\kappa$. Consider the system of linear constraints in variables y_i^ℓ, u_i^ℓ:

$$\Pi_\ell(y_{2i-1}^{\ell-1},y_{2i}^{\ell-1},y_i^\ell,u_i^\ell)\ge 0,\ \ i=1,\dots,2^{\kappa-\ell},\ \ \ell=1,\dots,\kappa. \tag{2}$$

Writing this system of linear constraints as $\Pi(y,t,u)\ge 0$, where Π is linear in its arguments, $y=(y_1^0,\dots,y_n^0)$, $t=y_1^\kappa$, and u is the collection of all u_i^ℓ, $\ell=1,\dots,\kappa$, and all y_i^ℓ, $\ell=1,\dots,\kappa-1$, we immediately conclude that Π is a polyhedral ϵ-approximation of $\mathbf{L}^{n+1}$ with

$$1+\epsilon=\prod_{\ell=1}^{\kappa}(1+\epsilon_\ell). \tag{3}$$

In view of this observation, we may focus on building polyhedral approximations of the Lorentz cone $\mathbf{L}^3$.

Polyhedral approximation of $\mathbf{L}^3$. The approximation we intend to use is given by the system of linear inequalities, as follows (positive integer ν is the parameter of the construction):

$$\begin{array}{lll}
\text{(a)} & \left\{\begin{array}{rcl}\xi^0 &\ge& |x_1|,\\ \eta^0 &\ge& |x_2|,\end{array}\right. & \\
\text{(b)} & \left\{\begin{array}{rcl}\xi^j &=& \cos\left(\frac{\pi}{2^{j+1}}\right)\xi^{j-1}+\sin\left(\frac{\pi}{2^{j+1}}\right)\eta^{j-1},\\ \eta^j &\ge& \left|-\sin\left(\frac{\pi}{2^{j+1}}\right)\xi^{j-1}+\cos\left(\frac{\pi}{2^{j+1}}\right)\eta^{j-1}\right|\end{array}\right. , & j=1,\dots,\nu,\\
\text{(c)} & \left\{\begin{array}{rcl}\xi^\nu &\le& x_3,\\ \eta^\nu &\le& \operatorname{tg}\left(\frac{\pi}{2^{\nu+1}}\right)\xi^\nu.\end{array}\right. &
\end{array} \tag{4}$$

Note that (4) can be straightforwardly written as a system of linear homogeneous inequalities $\Pi^{(\nu)}(x_1, x_2, x_3, u) \geq 0$, where u is the collection of $2(\nu+1)$ variables ξ^j, η^i, $j = 0, \ldots, \nu$.

Proposition 1. *$\Pi^{(\nu)}$ is a polyhedral $\delta(\nu)$-approximation of $\mathbf{L}^3 = \{(x_1, x_2, x_3) \mid \sqrt{x_1^2 + x_2^2} \leq x_3\}$ with*

$$\delta(\nu) = \frac{1}{\cos\left(\frac{\pi}{2^{\nu+1}}\right)} - 1. \tag{5}$$

Proof. We should prove that

(i) If $(x_1, x_2, x_3) \in \mathbf{L}^3$, then the triple (x_1, x_2, x_3) can be extended to a solution to (4).

(ii) If a triple (x_1, x_2, x_3) can be extended to a solution to (4), then $\|(x_1, x_2)\|_2 \leq (1 + \delta(\nu))x_3$.

(i): Given $(x_1, x_2, x_3) \in \mathbf{L}^3$, let us set $\xi^0 = |x_1|$, $\eta^0 = |x_2|$, thus ensuring (4)(a). Note that $\|(\xi^0, \eta^0)\|_2 = \|(x_1, x_2)\|_2$ and that the point $P^0 = (\xi^0, \eta^0)$ belongs to the first quadrant.

Now, for $j = 1, \ldots, \nu$ let us set

$$\begin{aligned} \xi^j &= \cos\left(\tfrac{\pi}{2^{j+1}}\right)\xi^{j-1} + \sin\left(\tfrac{\pi}{2^{j+1}}\right)\eta^{j-1}, \\ \eta^j &= \left|-\sin\left(\tfrac{\pi}{2^{j+1}}\right)\xi^{j-1} + \cos\left(\tfrac{\pi}{2^{j+1}}\right)\eta^{j-1}\right|, \end{aligned}$$

thus ensuring (4)(b), and let $P^j = (\xi^j, \eta^j)$. The point P^i is obtained from P^{j-1} by the following construction: We rotate clockwise P^{j-1} by the angle $\phi_j = \frac{\pi}{2^{j+1}}$, thus getting a point Q^{j-1}. If this point is in the upper half-plane, we set $P^j = Q^{j-1}$; otherwise, P^j is the reflection of Q^{j-1} with respect to the x-axis. From this description it is clear that

(I) $\|P^j\|_2 = \|P^{j-1}\|_2$, so that all vectors P^j are of the same Euclidean norm as P^0, i.e., of the norm $\|(x_1, x_2)\|_2$;

(II) since the point P^0 is in the first quadrant, the point Q^0 is in the angle $-\frac{\pi}{4} \leq \arg(P) \leq \frac{\pi}{4}$, so that P^1 is in the angle $0 \leq \arg(P) \leq \frac{\pi}{4}$. The latter relation, in turn, implies that Q^1 is in the angle $-\frac{\pi}{8} \leq \arg(P) \leq \frac{\pi}{8}$, whence P^2 is in the angle $0 \leq \arg(P) \leq \frac{\pi}{8}$. Similarly, P^3 is in the angle $0 \leq \arg(P) \leq \frac{\pi}{16}$, and so on: P^j is in the angle $0 \leq \arg(P) \leq \frac{\pi}{2^{j+1}}$.

By (I), $\xi^\nu \leq \|P^\nu\|_2 = \|(x_1, x_2)\|_2 \leq x_3$, so that the first inequality in (4)(c) is satisfied. By (II), P^ν is in the angle $0 \leq \arg(P) \leq \frac{\pi}{2^{\nu+1}}$, so that the second inequality in (4)(c) also is satisfied. We have extended a point from $\mathbf{L}^3$ to a solution to (4).

(ii): Let (x_1, x_2, x_3) be extended to a solution $(x_1, x_2, x_3, \{\xi^j, \eta^j\}_{j=0}^\nu)$ to (4). Let us set $P^j = (\xi^j, \eta^j)$. From (4)(a), (b) it follows that all vectors P^j are nonnegative. We have $\|P^0\|_2 \geq \|(x_1, x_2)\|_2$ by (4)(a). Now, (4)(b) says that the coordinates of P^j are $\geq$ absolute values of the coordinates of P^{j-1} taken in certain orthonormal system of coordinates, so that $\|P^j\|_2 \geq \|P^{j-1}\|_2$. Thus, $\|P^\nu\|_2 \geq \|(x_1, x_2)^T\|_2$. On the other hand, by (4)(c) one has $\|P^\nu\|_2 \leq \frac{1}{\cos(\frac{\pi}{2^{\nu+1}})}\xi^\nu \leq \frac{1}{\cos(\frac{\pi}{2^{\nu+1}})}x_3$, so that $\|(x_1, x_2)^T\|_2 \leq \delta(\nu)x_3$, as claimed. □

Specifying in (2) the mappings $\Pi_\ell(\cdot)$ as $\Pi^{(\nu_\ell)}(\cdot)$, we conclude that for every collection of positive integers $\nu_1, \dots, \nu_\kappa$ one can point out a polyhedral β-approximation $\Pi_{\nu_1,\dots,\nu_\kappa}(y,t,u)$ of $\mathbf{L}^n$, $n = 2^\kappa$:

$$
\begin{array}{ll}
(a_{\ell,i}) & \left\{ \begin{array}{rcl} \xi^0_{\ell,i} & \geq & |y^{\ell-1}_{2i-1}|, \\ \eta^0_{\ell,i} & \geq & |y^{\ell-1}_{2i}|, \end{array} \right. \\
(b_{\ell,i}) & \left\{ \begin{array}{rcl} \xi^j_{\ell,i} & = & \cos\left(\frac{\pi}{2^{j+1}}\right)\xi^{j-1}_{\ell,i} + \sin\left(\frac{\pi}{2^{j+1}}\right)\eta^{j-1}_{\ell,i}, \\ \eta^j_{\ell,i} & \geq & \left| -\sin\left(\frac{\pi}{2^{j+1}}\right)\xi^{j-1}_{\ell,i} + \cos\left(\frac{\pi}{2^{j+1}}\right)\eta^{j-1}_{\ell,i} \right|, \end{array} \right. \quad j = 1, \dots, \nu_\ell, \\
(c_{\ell,i}) & \left\{ \begin{array}{rcl} \xi^{\nu_\ell}_{\ell,i} & \leq & y^\ell_i, \\ \eta^{\nu_\ell}_{\ell,i} & \leq & \mathrm{tg}\left(\frac{\pi}{2^{\nu_\ell+1}}\right)\xi^{\nu_\ell}_{\ell,i}, \end{array} \right. \\
 & \qquad i = 1, \dots, 2^{\kappa-\ell}, \ \ell = 1, \dots, \kappa.
\end{array}
\tag{6}
$$

The approximation possesses the following properties:

1. The dimension of the u-vector (comprising all variables in (6) except $y_i = y^0_i$ and $t = y^\kappa_1$) is
$$p(n, \nu_1, \dots, \nu_\kappa) \leq n + O(1)\sum_{\ell=1}^{\kappa} 2^{\kappa-\ell}\nu_\ell.$$
2. The image dimension of $\Pi_{\nu_1,\dots,\nu_\kappa}(\cdot)$ (i.e., the number of linear inequalities plus twice the number of linear equations in (6)) is
$$q(n, \nu_1, \dots, \nu_\kappa) \leq O(1)\sum_{\ell=1}^{\kappa} 2^{\kappa-\ell}\nu_\ell.$$
3. The quality β of the approximation is
$$\beta = \beta(n; \nu_1, \dots, \nu_\kappa) = \prod_{\ell=1}^{\kappa} \frac{1}{\cos\left(\frac{\pi}{2^{\nu_\ell+1}}\right)} - 1.$$

Back to the general case. Given $\epsilon \in (0, 1]$ and setting

$$\nu_\ell = \left\lfloor O(1)\ell \ln\frac{2}{\epsilon} \right\rfloor, \ \ell = 1, \dots, \kappa,$$

with properly chosen absolute constant $O(1)$, we ensure that

$$
\begin{array}{rcl}
\beta(\nu_1, \dots, \nu_\kappa) & \leq & \epsilon, \\
p(n, \nu_1, \dots, \nu_\kappa) & \leq & O(1)n\ln\frac{2}{\epsilon}, \\
q(n, \nu_1, \dots, \nu_\kappa) & \leq & O(1)n\ln\frac{2}{\epsilon},
\end{array}
$$

as required.

Exercises to Lecture 4

Positive semidefiniteness, eigenvalues, and $\succeq$-ordering

Exercise 4.2. Diagonal-dominant matrices. Let $A = [a_{ij}]_{i,j=1}^m$ be a symmetric matrix satisfying the relation

$$a_{ii} \geq \sum_{j \neq i} |a_{ij}|, \ \ i = 1, \ldots, m.$$

Prove that A is positive semidefinite.

Solution. Let e be an eigenvector of A and λ be the corresponding eigenvalue. We may assume that the largest, in absolute value, of coordinates of e is equal to 1. Let i be the index of this coordinate; then

$$\lambda = a_{ii} + \sum_{j \neq i} a_{ij} e_j \geq a_{ii} - \sum_{j \neq i} |a_{ij}| \geq 0.$$

Thus, all eigenvalues of A are nonnegative, so that A is positive semidefinite. □

Variational description of eigenvalues

Exercise 4.8. Let f_* be a closed convex function with the domain $\mathrm{Dom} f_* \subset \mathbf{R}_+$, and let f be the Legendre transformation of f_*. Then for every pair of symmetric matrices X, Y of the same size with the spectrum of X belonging to $\mathrm{Dom} f$ and the spectrum of Y belonging to $\mathrm{Dom} f_*$ one has

$$\lambda(f(X)) \geq \lambda\left(Y^{1/2} X Y^{1/2} - f_*(Y)\right). \qquad (*)$$

Solution. By continuity reasons, it suffices to prove $(*)$ in the case of $Y \succ 0$ (why?). Let m be the size of X, let $k \in \{1, \ldots, m\}$, let $\mathcal{E}_k$ be the family of linear subspaces of $\mathbf{R}^m$ of codimension $k-1$, and let $E \in \mathcal{E}_k$ be such that

$$e \in E \Rightarrow e^T X e \leq \lambda_k(X) e^T e.$$

(Such an E exists by the variational characterization of eigenvalues as applied to X.) Let also $F = Y^{-1/2} E$; the codimension of F, same as the one of E, is $k-1$. Finally, let $g_1, .., g_m$ be an orthonormal system of eigenvectors of Y, so that $Y g_j = \lambda_j(Y) g_j$. We have

$$
\begin{aligned}
h \in F, h^T h = 1 \quad &\Rightarrow \\
h^T[Y^{1/2}XY^{1/2} - f_*(Y)]h \quad &= \quad (\underbrace{Y^{1/2}h}_{\in E})^T X (Y^{1/2}h) - \sum_{j=1}^m f_*(\lambda_j(Y))(g_j^T h)^2 \\
&\leq \quad \lambda_k(X)(Y^{1/2}h)^T(Y^{1/2}h) - \sum_{j=1}^m f_*(\lambda_j(Y))(g_j^T h)^2 \\
&\quad \quad [\text{since } Y^{1/2}h \in E] \\
&= \quad \lambda_k(X)(h^T Y h) - \sum_{j=1}^m f_*(\lambda_j(Y))(g_j^T h)^2 \\
&= \quad \lambda_k(X)\sum_{j=1}^m \lambda_j(Y)(g_j^T h)^2 - \sum_{j=1}^m f_*(\lambda_j(Y))(g_j^T h)^2 \\
&= \quad \sum_{j=1}^m [\lambda_k(X)\lambda_j(Y) - f_*(\lambda_j(Y))](g_j^T h)^2 \\
&\leq \quad \sum_{j=1}^m f(\lambda_k(X))(g_j^T h)^2 \\
&\quad \quad [\text{since } f = (f_*)_*] \\
&= \quad f(\lambda_k(X)) \\
&\quad \quad [\text{since } \sum_j (g_j^T h)^2 = h^T h = 1] \\
&= \quad \lambda_k(f(X)) \\
&\quad \quad [\text{since } f(\cdot) \text{ is nonincreasing due to } \mathrm{Dom} f_* \subset \mathbf{R}_+].
\end{aligned}
$$

We see that there exists $F \in \mathcal{E}_k$ such that

$$\max_{h \in F : h^T h = 1} h^T[Y^{1/2}XY^{1/2} - f_*(Y)]h \leq \lambda_k(f(X)).$$

From variational characterization of eigenvalues it follows that

$$\lambda_k(Y^{1/2}XY^{1/2} - f_*(Y)) \leq \lambda_k(f(X)). \qquad \square$$

Exercise 4.10. 4. (trace inequality) Prove that whenever $A, B \in \mathbf{S}^m$, one has

$$\lambda^T(A)\lambda(B) \geq \mathrm{Tr}(AB).$$

Solution. Denote $\lambda = \lambda(A)$, and let $A = V^T \mathrm{Diag}(\lambda) V$ be the spectral decomposition of A. Setting $\hat{B} = VBV^T$, note that $\lambda(\hat{B}) = \lambda(B)$ and $\mathrm{Tr}(AB) = \mathrm{Tr}(\mathrm{Diag}(\lambda)\hat{B})$. Thus, it suffices to prove the trace inequality in the particular case when A is a diagonal matrix with the diagonal $\lambda = \lambda(A)$. Denoting by μ the diagonal of B and setting

$$\sigma^0 = 0; \sigma^k = \sum_{i=1}^k \mu_i, \ k = 1, \ldots, m,$$

we have

$$\begin{aligned}
\mathrm{Tr}(AB) &= \sum_{i=1}^{m} \lambda_i \mu_i \\
&= \sum_{i=1}^{m} \lambda_i (\sigma^i - \sigma^{i-1}) \\
&= -\lambda_1 \sigma^0 + \sum_{i=1}^{m-1} (\lambda_i - \lambda_{i+1}) \sigma^i + \lambda_m \sigma^m \\
&= \sum_{i=1}^{m-1} (\lambda_i - \lambda_{i+1}) \sigma^i + \lambda_m \mathrm{Tr}(B) \\
&\leq \sum_{i=1}^{m-1} (\lambda_i - \lambda_{i+1}) \sum_{j=1}^{i} \lambda_j(B) + \lambda_m \sum_{j=1}^{m} \lambda_j(B) \\
&\quad \text{[since } \lambda_i \geq \lambda_{i+1} \text{ and in view of Exercise 4.10.3]} \\
&= \sum_{i=1}^{m} \lambda_i \lambda_i(B) \\
&= \lambda^T(A) \lambda(B). \qquad \square
\end{aligned}$$

Exercise 4.12. 3. Let X be a symmetric $n \times n$ matrix partitioned into blocks in a symmetric, with respect to the diagonal, fashion,

$$X = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{12}^T & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1m}^T & X_{2m}^T & \dots & X_{mm} \end{pmatrix},$$

so that the blocks X_{ii} are square. Let also $g : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ be convex function on the real line which is finite on the set of eigenvalues of X, and let $\mathcal{F}_n \subset \mathbf{S}^n$ be the set of all $n \times n$ symmetric matrices with all eigenvalues belonging to the domain of g. Assume that the mapping

$$Y \mapsto g(Y) : \mathcal{F}_n \to \mathbf{S}^n$$

is $\succeq$-convex:

$$g(\lambda' Y' + \lambda'' Y'') \preceq \lambda' g(Y') + \lambda'' g(Y'') \quad \forall (Y', Y'' \in \mathcal{F}_n, \lambda', \lambda'' \geq 0, \lambda' + \lambda'' = 1).$$

Prove that

$$(g(X))_{ii} \succeq g(X_{ii}), \ i = 1, \dots, m,$$

where the partition of $g(X)$ into the blocks $(g(X))_{ij}$ is identical to the partition of X into the blocks X_{ij}.

Solution. Let $\epsilon = (\epsilon_1, \dots, \epsilon_m)$ with $\epsilon_j = \pm 1$, and let

$$U_\epsilon = \begin{pmatrix} \epsilon_1 I_{n_1} & & & \\ & \epsilon_2 I_{n_2} & & \\ & & \ddots & \\ & & & \epsilon_m I_{n_m} \end{pmatrix},$$

where n_i is the row size of X_{ii}. Then U_ϵ are orthogonal matrices and one clearly has

$$D(X) \equiv \begin{pmatrix} X_{11} & & & \\ & X_{22} & & \\ & & \ddots & \\ & & & X_{mm} \end{pmatrix} = \frac{1}{2^m} \sum_{\epsilon:\epsilon_i=\pm 1,\, i=1,\dots,m} U_\epsilon^T X U_\epsilon.$$

We have

$$\begin{array}{rl} D(X) = \frac{1}{2^m} \displaystyle\sum_{\epsilon:\epsilon_i=\pm 1,\, i=1,\dots,m} U_\epsilon^T X U_\epsilon & \\ \Downarrow & \\ g(D(X)) \preceq \frac{1}{2^m} \displaystyle\sum_{\epsilon:\epsilon_i=\pm 1,\, i=1,\dots,m} g(U_\epsilon^T X U_\epsilon) & [\text{since } g \text{ is } \succeq\text{-convex}] \\ \Updownarrow & \\ g(D(X)) \preceq \frac{1}{2^m} \displaystyle\sum_{\epsilon:\epsilon_i=\pm 1,\, i=1,\dots,m} U_\epsilon^T g(X) U_\epsilon & \\ \Downarrow & \\ g(D(X)) \preceq D(g(X)). \quad \square & \end{array}$$

Cauchy's inequality for matrices

Exercise 4.19. 1. Denote $P = \left(\sum_i X_i^T X_i\right)^{1/2}$, $Q = \sum_i Y_i^T Y_i$ ($X_i, Y_i \in \mathbf{M}^{p,q}$). We should prove that

$$\sigma\left(\sum_i X_i^T Y_i\right) \le \lambda(P)\|\lambda(Q)\|_\infty^{1/2}$$

or, which is the same,

$$\sigma\left(\sum_i Y_i^T X_i\right) \le \lambda(P)\|\lambda(Q)\|_\infty^{1/2}.$$

By the variational description of singular values, it suffices to prove that for every $k = 1, 2, \dots, p$ there exists a subspace $L_k \subset \mathbf{R}^q$ of codimension $k-1$ such that

$$\forall \xi \in L_k: \quad \left\| \left(\sum_i Y_i^T X_i\right) \xi \right\|_2 \le \|\xi\|_2 \lambda_k(P)\|\lambda(Q)\|_\infty^{1/2}. \qquad (*)$$

Let $e_1, \ldots, e_q$ be the orthonormal eigenbasis of P: $Pe_i = \lambda_i(P)e_i$, and let L_k be the linear span of $e_k, e_{k+1}, \ldots, e_q$. For $\xi \in L_k$ one has

$$\begin{aligned}|\eta^T \left(\sum_i Y_i^T X_i\right) \xi| &\le \sum_i \|Y_i\eta\|_2 \|X_i\xi\|_2 \le \sqrt{\sum_i \|Y_i\eta\|_2^2}\sqrt{\sum_i \|X_i\xi\|_2^2} \\ &= \sqrt{\eta^T \left(\sum_i Y_i^T Y_i\right)\eta}\sqrt{\xi^T \left(\sum_i X_i^T X_i\right)\xi} \\ &\le \left(\|\lambda(Q)\|_\infty \|\eta\|_2^2\right)^{1/2} \left(\lambda_k^2(P)\|\xi\|_2^2\right)^{1/2} = \|\lambda(Q)\|_\infty^{1/2}\lambda_k(P)\|\eta\|_2\|\xi\|_2,\end{aligned}$$

whence

$$\left\|\left(\sum_i Y_i^T X_i\right)\xi\right\|_2 = \max_{\eta:\|\eta\|_2=1} \eta^T \left(\sum_i Y_i^T X_i\right)\xi \le \lambda_k(P)\|\lambda(Q)\|_\infty^{1/2}\|\xi\|_2,$$

as required in (*).

To make (*) equality, assume that $P \succ 0$ (the case of singular P is left to the reader), and let $Y_i = X_i P^{-1}$. Then

$$\sum_i Y_i^T Y_i = P^{-1}\left(\sum_i X_i^T X_i\right) P^{-1} = I$$

and

$$\sum_i X_i^T Y_i = \left(\sum_i X_i^T X_i\right) P^{-1} = P,$$

so that (*) becomes equality.

(1)⇒(2): it suffices to prove that if $A \in \mathbf{M}^{p,p}$, then

$$|\mathrm{Tr}(A)| \le \|\sigma(A)\|_1. \qquad (**)$$

Indeed, we have $A = U\Lambda V$, where U, V are orthogonal matrices and Λ is a diagonal matrix with the diagonal $\sigma(A)$. Denoting by e_i the standard basic orths in $\mathbf{R}^p$, we have

$$\begin{aligned}|\mathrm{Tr}(A)| &= |\mathrm{Tr}(U^T AU)| = |\mathrm{Tr}(\Lambda(VU))| \\ &= \left|\sum_i e_i^T \Lambda(VU)e_i\right| \le \sum_i |\sigma_i(A)e_i^T(VU)e_i| \le \sum_i \sigma_i(A),\end{aligned}$$

as required in (**).

Exercise 4.20. The true statements are 2 and 3.

Part 2 is an immediate consequence of the following.

Lemma 1. *Let $A_i \in \mathbf{S}_+^n$, $i = 1, \ldots, m$, and let $\alpha > 1$. Then*

$$\lambda_j\left(\left(\sum_{i=1}^m A_i^\alpha\right)^{1/\alpha}\right) \le \lambda_j^{\frac{1}{\alpha}}\left(\sum_{i=1}^m A_i\right)\lambda_1^{1-\frac{1}{\alpha}}\left(\sum_{i=1}^m A_i\right).$$

(Here, as always, $\lambda_j(B)$ are eigenvalues of a symmetric matrix B arranged in the nonascending order.)

Proof. Let $B = \sum_{i=1}^m A_i^\alpha$, $A = \sum_{i=1}^m A_i$. Since $\lambda_j(B^{1/\alpha}) = \left(\lambda_j(B)\right)^{1/\alpha}$, we should prove that

$$\lambda_j(B) \le \lambda_j(A)\lambda_1^{\alpha-1}(A). \tag{7}$$

By the variational description of eigenvalues, it suffices to verify that for every $j \le n$ there exists a linear subspace L_j in $\mathbf{R}^n$ of codimension $j-1$ such that

$$\xi^T B\xi \le \lambda_j(A)\lambda_1^{\alpha-1}(A) \quad \forall(\xi \in L_j, \|\xi\|_2 = 1). \tag{8}$$

Let $e_1, \dots, e_n$ be an orthonormal eigenbasis of A ($Ae_j = \lambda_j(A)e_j$), and let L_j be the linear span of the vectors $e_j, e_{j+1}, \dots, e_n$. Let $\xi \in L_j$ be a unit vector. We have

$$\begin{aligned}
\sum_i \xi^T A_i^\alpha \xi &= \sum_i \xi^T A_i \underbrace{[A_i^{\alpha-1}\xi]}_{\eta_i} \\
&\le \sum_i (\xi^T A_i \xi)^{1/2} (\eta_i^T A_i \eta_i)^{1/2} && \text{[since } A_i \succeq 0\text{]} \\
&\le \left(\sum_i \xi^T A_i \xi\right)^{1/2} \left(\sum_i \eta_i^T A_i \eta_i\right)^{1/2} && \text{[Cauchy's inequality]} \\
&= \left(\xi^T A \xi\right)^{1/2} \left(\sum_i \xi^T A_i^{2\alpha-1} \xi\right)^{1/2} \\
&\le \left(\xi^T A \xi\right)^{1/2} \left(\sum_i \lambda_1^{2\alpha-2}(A_i) \xi^T A_i \xi\right)^{1/2} \\
&\le \left(\max_i \lambda_1(A_i)\right)^{\alpha-1} \sum_i \xi^T A_i \xi \\
&\le \lambda_1^{\alpha-1}(A)\lambda_j(A) && \text{[since } \|\xi\|_2 = 1 \text{ and } \xi \in L_j\text{]}
\end{aligned}$$

as required in (8). □

Part 3 is less trivial than part 2. Let us look at the $(nm) \times (nm)$ square matrix

$$Q = \begin{pmatrix} A_1^{\alpha/2} & & \\ A_2^{\alpha/2} & & \\ \vdots & & \\ A_m^{\alpha/2} & & \end{pmatrix}.$$

(As always, blank spaces are filled with zeros.) Then

$$Q^T Q = \left(\begin{array}{c|c} B \equiv \sum_i A_i^\alpha & \\ \hline & \end{array}\right)$$

so that $\mathrm{Tr}([Q^TQ]^{1/\alpha}) = \mathrm{Tr}(B^{1/\alpha})$. Since the eigenvalues of Q^TQ are exactly the same as the eigenvalues of $X = QQ^T$, we conclude that

$$\mathrm{Tr}(B^{1/\alpha}) = \mathrm{Tr}([QQ^T]^{1/\alpha}) = \mathrm{Tr}(X^{1/\alpha}),$$

$$X = \begin{pmatrix} A_1^\alpha & A_1^{\alpha/2}A_2^{\alpha/2} & \cdots & A_1^{\alpha/2}A_m^{\alpha/2} \\ A_2^{\alpha/2}A_1^{\alpha/2} & A_2^\alpha & \cdots & A_2^{\alpha/2}A_m^{\alpha/2} \\ \vdots & \vdots & \ddots & \vdots \\ A_m^{\alpha/2}A_1^{\alpha/2} & A_m^{\alpha/2}A_2^{\alpha/2} & \cdots & A_m^\alpha \end{pmatrix}.$$

Applying the result of Exercise 4.12.1 with $F(Y) = -\mathrm{Tr}(Y^{1/\alpha})$, we get

$$[\mathrm{Tr}(B^{1/\alpha}) =] \quad \mathrm{Tr}(X^{1/\alpha}) = -F(X) \le -F\left(\begin{pmatrix} A_1^\alpha & & \\ & \ddots & \\ & & A_m^\alpha \end{pmatrix}\right) = \sum_{i=1}^m \mathrm{Tr}(A_i),$$

as required.

$\succeq$-convexity of some matrix-valued functions.

Exercise 4.22.

4. Prove that the function

$$F(x) = x^{1/2} : \mathbf{S}_+^m \to \mathbf{S}_+^m$$

is $\succeq$-concave and $\succeq$-monotone.

Solution. Since the function is continuous on its domain, it suffices to verify that it is $\succeq$-monotone and $\succeq$-concave on $\mathrm{int}\mathbf{S}_+^m$, where the function is smooth.

Differentiating the identity

$$F(x)F(x) = x \qquad (*)$$

in a direction h and setting $F(x) = y$, $DF(x)[h] = dy$, we get

$$y\,dy + dy\,y = h.$$

Since $y \succ 0$, this Lyapunov equation admits an explicit solution:

$$dy = \int_0^\infty \exp\{-ty\}h\exp\{-ty\}dt,$$

and we see that $dy \succeq 0$ whenever $h \succeq 0$. Applying Exercise 4.21.4, we conclude that F is $\succeq$-monotone.

Differentiating (*) twice in a direction h and denoting $d^2y = D^2F(x)[h,h]$, we get

$$y\,d^2y + d^2y\,y + 2(dy)^2 = 0,$$

whence, same as above,

$$d^2y = -\int_0^\infty \exp\{-ty\}(dy)^2\exp\{-ty\}dt \preceq 0.$$

Applying Exercise 4.21.3, we conclude that F is $\succeq$-concave. □

5. Prove that the function

$$F(x) = \ln x : \mathrm{int}\mathbf{S}^m_+ \to \mathbf{S}^m$$

is $\succeq$-monotone and $\succeq$-concave.

Solution. The function $x^{1/2} : \mathbf{S}^m_+ \to \mathbf{S}^m_+$ is $\succeq$-monotone and $\succeq$-concave by Exercise 4.22.4. Applying Exercise 4.21.6, we conclude that so are the functions $x^{1/2^k}$ for all positive integer k. It remains to note that

$$\ln x = \lim_{k\to\infty} 2^k\left[x^{1/2^k} - I\right]$$

and to use Exercise 4.21.7. □

6. Prove that the function

$$F(x) = \left(Ax^{-1}A^T\right)^{-1} : \mathrm{int}\mathbf{S}^n_+ \to \mathbf{S}^m$$

with matrix A of rank m is $\succeq$-concave and $\succeq$-monotone.

Solution. Since A is of rank m, the function $F(x)$ clearly is well defined and $\succ 0$ when $x \succ 0$. To prove that F is $\succeq$-concave, it suffices to verify that the set

$$\{(x,Y) \mid x, Y \succ 0, Y \preceq (Ax^{-1}A^T)^{-1}\}$$

is convex, which is nearly evident:

$$\begin{aligned}\{(x,Y) \mid x, Y \succ 0, Y \preceq (Ax^{-1}A^T)^{-1}\} &= \{(x,Y) \mid x, Y \succ 0, Y^{-1} \succeq Ax^{-1}A^T\}\\ &= \{(x,Y) \mid x, Y \succ 0, \begin{pmatrix} Y^{-1} & A\\ A^T & x\end{pmatrix} \succeq 0\}\\ &= \{(x,Y) \mid x, Y \succ 0, x \succeq A^TYA\}.\end{aligned}$$

To prove that F is $\succeq$-monotone, note that if $0 \preceq x \preceq x'$, then $0 \prec (x')^{-1} \preceq x^{-1}$, whence $0 \prec A(x')^{-1}A^T \preceq Ax^{-1}A^T$, whence, in turn, $F(x) = (Ax^{-1}A^T)^{-1} \preceq (A(x')^{-1}A^T)^{-1} = F(x')$. □

Lovasz capacity number

Exercise 4.33. Let Γ be an n-node graph and $\sigma(\Gamma)$ be the optimal value in the problem

$$\min_{\lambda,\mu,\nu}\left\{\lambda : \begin{pmatrix} \lambda, & -\frac{1}{2}(e+\mu)^T \\ -\frac{1}{2}(e+\mu), & A(\mu,\nu)\end{pmatrix} \succeq 0\right\}, \tag{Sh}$$

where $e = (1,\dots,1)^T \in \mathbf{R}^n$, $A(\mu,\nu) = \mathrm{Diag}(\mu) + Z(\nu)$, and $Z(\nu)$ is the matrix as follows:

- The dimension of ν is equal to the number of arcs in Γ, and the coordinates of ν are indexed by these arcs.
- The diagonal entries of Z, same as the off-diagonal entries of Z corresponding to empty cells ij (i.e., with i and j nonadjacent) are zeros.
- The off-diagonal entries of Z in a pair of symmetric nonempty cells ij, ji are equal to the coordinate of ν indexed by the corresponding arc.

Prove that $\sigma(\Gamma)$ is nothing but the Lovasz capacity $\Theta(\Gamma)$ of the graph.

Solution. In view of (4.10.122) all we need is to prove that $\sigma(\Gamma) \geq \Theta(\Gamma)$.

Let (λ,μ,ν) be a feasible solution to (Sh); we should prove that there exists x such that (λ, x) is a feasible solution to the Lovasz problem

$$\min_{\lambda,x}\{\lambda : \lambda I_n - \mathcal{L}(x) \succeq 0\}. \tag{L}$$

Setting $y = \frac{1}{2}(e+\mu)$, we see that

$$\begin{pmatrix} \lambda & -\frac{1}{2}(e+\mu)^T \\ -\frac{1}{2}(e+\mu) & Z(\nu)+\mathrm{Diag}(\mu)\end{pmatrix} = \begin{pmatrix} \lambda & -y^T \\ -y & Z(\nu)+2\mathrm{Diag}(y)-I_n\end{pmatrix} \succeq 0.$$

The diagonal entries of $Z = Z(\nu)$ are zero, while the diagonal entries of $Z + 2\mathrm{Diag}(y) - I_n$ must be nonnegative; we conclude that $y > 0$. Setting $Y = \mathrm{Diag}(y)$, we have

$$\begin{pmatrix} \lambda & -y^T \\ -y & Z+2Y-I_n\end{pmatrix} \succeq 0$$
$$\Rightarrow \begin{pmatrix} 1 & \\ & Y^{-1}\end{pmatrix}\begin{pmatrix} \lambda & -y^T \\ -y & Z+2Y-I_n\end{pmatrix}\begin{pmatrix} 1 & \\ & Y^{-1}\end{pmatrix} \succeq 0,$$

i.e.,

$$\begin{pmatrix} \lambda & -e^T \\ -e & Y^{-1}ZY^{-1}+2Y^{-1}-Y^{-2}\end{pmatrix} \succeq 0,$$

whence by the Schur complement lemma

$$\lambda\left[Y^{-1}ZY^{-1}+2Y^{-1}-Y^{-2}\right] - ee^T \succeq 0$$

or, which is the same,

$$\lambda I_n - \left[ee^T - \lambda Y^{-1}ZY^{-1}\right] \succeq \lambda(I_n - 2Y^{-1} + Y^{-2}) = \lambda(I_n - Y^{-1})^2.$$

We see that $\lambda I_n - \left[ee^T - \lambda Y^{-1}ZY^{-1}\right] \succeq 0$. It remains to note that the matrix in the brackets clearly is $\mathcal{L}(x)$ for certain x. □

$\mathcal{S}$-lemma. (For notation, see section 4.10.5.)

Exercise 4.42. We should prove that if $f(x) = x^T Ax + 2a^T x + \alpha$ and $g(x) = x^T Bx + 2b^T x + \beta$ are two quadratic forms ($A = A^T$, $B = B^T$) such that the premise in the implication

$$f(x) \le 0 \Rightarrow g(x) \le 0 \tag{9}$$

is strictly feasible, then the implication holds true if and only if

$$\exists \lambda \ge 0 : \quad g(x) \le \lambda f(x) \ \forall x.$$

Proof. The "if" part of the statement is evident. Let us prove the "only if" part. Thus, let us assume that the implication (9) is valid.

1. There clearly exist a sequence $\{\gamma_i > 0\}$ and $\delta > 0$ such that

   (i) $\gamma_i \to 0$, $i \to \infty$;

   (ii) all the matrices $A_i \equiv A + \gamma_i I$ are nonsingular; and

   (iii) $\bar{x}^T A_i \bar{x} + 2a^T \bar{x} + \alpha \le -\delta \ \forall i$.

2. Observe that since $\gamma_i > 0$, one has

$$x^T A_i x + 2a^T x + \alpha \le 0 \Rightarrow x^T Ax + 2a^T x + \alpha \le 0 \Rightarrow x^T Bx + 2b^T x + \beta \le 0,$$

i.e., one has

$$f_i(x) \equiv x^T A_i x + 2a^T x + \alpha \le 0 \Rightarrow g(x) = x^T Bx + 2b^T x + \beta \le 0.$$

Setting

$$\begin{aligned} \widehat{f_i}(y) &= f_i(y - A_i^{-1}a) = y^T A_i y + \underbrace{[\alpha - a^T A_i^{-1} a]}_{\alpha_i}, \\ \widehat{g_i}(y) &= g(y - A_i^{-1}a) = y^T By + 2b_i^T y + \beta_i, \end{aligned}$$

we have

$$\begin{array}{lll} \text{(a)} & \forall y : & y^T A_i y + \alpha_i \le 0 \Rightarrow y^T By + 2b_i^T y + \beta_i \le 0, \\ \text{(b)} & & \exists \tilde{y} : \ \tilde{y}^T A_i \tilde{y} + \alpha_i < 0. \end{array} \tag{10}$$

3. We claim that

$$\begin{array}{lll} \text{(a)} & \forall (y,t) : & y^T A_i y + \alpha_i t^2 \le 0 \Rightarrow y^T By + 2t b_i^T y + \beta_i t^2 \le 0, \\ \text{(b)} & & \exists (\bar{y}, \bar{t}) : \ \bar{y}^T A_i \bar{y} + \alpha_i \bar{t}^2 < 0. \end{array} \tag{11}$$

Indeed, (11)(b) is evident (set $\bar{y} = \tilde{y}$, $\bar{t} = 1$). Further, the implication (11)(a) with the premise strengthened by the assumption $t \ne 0$ is an immediate corollary of (10)(a). Thus, all we need in order to verify (11)(a) is to prove the implication

$$y^T A_i y \le 0 \Rightarrow y^T By \le 0. \tag{12}$$

Assume, on the contrary to what should be proved, that for some y it holds that

$$y^T A_i y \le 0, \quad y^T B y > 0.$$

Then $y \neq 0$, whence $y^T A y < y^T A_i y \le 0$, i.e., $y^T A y < 0$. It follows that

$$\begin{array}{rcl} f(sy - A_i^{-1}a) & \to & -\infty,\ s \to \infty \quad [\text{since } y^T A y < 0], \\ g(sy - A_i^{-1}a) & \to & +\infty,\ s \to \infty \quad [\text{since } y^T B y > 0], \end{array}$$

in contradiction to the assumption that $f(x) \le 0 \Rightarrow g(x) \le 0$.

4. By the usual $\mathcal{S}$-lemma, from (11) it follows that

$$\exists \lambda_i \ge 0: \quad y^T B y + 2t b_i^T y + \beta_i t^2 \le \lambda_i [y^T A_i y + \alpha_i t^2] \quad \forall (y, t),$$

whence (set $t = 1$) $\widehat{g}_i(y) \le \lambda_i \widehat{f}_i(y)\ \forall y$ or, which is the same,

$$x^T B x + 2b^T x + \beta \le \lambda_i [x^T A_i x + 2a^T x + \alpha] \quad \forall x.$$

Setting in the latter inequality $x = \bar{x}$ and taking into account (iii), we conclude that $\{\lambda_i \ge 0\}_{i=1}^{\infty}$ is a bounded sequence. Denoting by λ (any) limiting point of the sequence and taking into account (i), we come to the desired relations $\lambda \ge 0$, $g(x) \le \lambda f(x)\ \forall x$. □

Exercise 4.46.3. 3. Given data A, B satisfying the premise of (**SL**)(B), define the sets

$$Q_x = \{\lambda \ge 0 : x^T B x \ge \lambda x^T A x\}.$$

Prove that every two sets $Q_{x'}$, $Q_{x''}$ have a point in common.

Solution. The case when x', x'' are collinear is trivial. Assuming that x', x'' are linearly independent, consider the quadratic forms on the 2D plane:

$$\alpha(z) = (sx' + tx'')^T A (sx' + tx''),\ \beta(z) = (sx' + tx'')^T B (sx' + tx''),\ \ z = (s, t)^T.$$

By their origin, we have

$$\alpha(z) \ge 0, z \neq 0 \Rightarrow \beta(z) > 0. \qquad (!)$$

All we need is to prove that there exists $\lambda \ge 0$ such that $\beta(z) \ge \lambda \alpha(z)\ \forall z \in \mathbf{R}^2$. Such a λ clearly is a common point of $Q_{x'}$ and $Q_{x''}$.

As is well known from linear algebra, we can choose a coordinate system in $\mathbf{R}^2$ in such a way that the matrix α of the form $\alpha(\cdot)$ in these coordinates, let them be called u, v, is diagonal:

$$\alpha = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix};$$

let also

$$\beta = \begin{pmatrix} p & r \\ r & q \end{pmatrix}$$

be the matrix of the form $\beta(\cdot)$ in the coordinates u, v. Let us consider all possible combinations of signs of a, b:

- $a \geq 0, b \geq 0$. In this case, $\alpha(\cdot)$ is nonnegative everywhere, whence by (!) $\beta(\cdot) \geq 0$. Consequently, $\beta(\cdot) \geq \lambda\alpha(\cdot)$ with $\lambda = 0$.
- $a < 0, b < 0$. In this case the matrix of the quadratic form $\beta(\cdot) - \lambda\alpha(\cdot)$ is

$$\begin{pmatrix} p + \lambda|a| & r \\ r & q + \lambda|b| \end{pmatrix}.$$

  This matrix clearly is positive definite for all large-enough positive λ, so that here again $\beta(\cdot) \geq \lambda\alpha(\cdot)$ for properly chosen nonnegative λ.
- $a = 0, b < 0$. In this case $\alpha(1, 0) = 0$ (the coordinates in question are u, v), so that by (!) $p > 0$. The matrix of the form $\beta(\cdot) - \lambda\alpha(\cdot)$ is

$$\begin{pmatrix} p & r \\ r & q + \lambda|b| \end{pmatrix},$$

  and since $p > 0$ and $|b| > 0$, this matrix is positive definite for all large-enough positive λ. Thus, here again $\beta(\cdot) \geq \lambda\alpha(\cdot)$ for properly chosen $\lambda \geq 0$.
- $a < 0, b = 0$. This case is completely similar to the previous one.

Part 3 is proved. □

Exercise 4.47. Demonstrate by example that if $x^T Ax, x^T Bx, x^T Cx$ are three quadratic forms with symmetric matrices such that

$$\begin{array}{c} \exists \bar{x} : \bar{x}^T A\bar{x} > 0, \bar{x}^T B\bar{x} > 0, \\ x^T Ax \geq 0, x^T Bx \geq 0 \Rightarrow x^T Cx \geq 0, \end{array} \tag{13}$$

then not necessarily there exist $\lambda, \mu \geq 0$ such that $C \succeq \lambda A + \mu B$.

A solution.

$$A = \begin{pmatrix} \lambda^2 & 0 \\ 0 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} \mu\nu & 0.5(\mu - \nu) \\ 0.5(\mu - \nu) & -1 \end{pmatrix}, \quad C = \begin{pmatrix} \lambda\mu & 0.5(\mu - \lambda) \\ 0.5(\mu - \lambda) & -1 \end{pmatrix}.$$

With a proper setup, e.g.,

$$\lambda = 1.100, \quad \mu = 0.818, \quad \nu = 1.344,$$

the above matrices satisfy both (4.10.133) and (4.10.134). □

Exercise 4.52. Let $n \geq 3$,

$$\begin{aligned} f(x) &= \theta_1 x_1^2 - \theta_2 x_2^2 + \sum_{i=3}^{n} \theta_i x_i^2 : \mathbf{R}^n \to \mathbf{R}, \\ &\quad \theta_1 \geq \theta_2 \geq 0, \theta_1 + \theta_2 > 0, -\theta_2 \leq \theta_i \leq \theta_1 \quad \forall i \geq 3; \\ Y &= \{x : \|x\|_2 = 1, f(x) = 0\}. \end{aligned}$$

1. Let $x \in Y$. Prove that x can be linked in Y by a continuous curve with a point x' such that the coordinates of x' with indices $3, 4, \ldots, n$ vanish.

Proof. It suffices to build a continuous curve $\gamma(t) \in Y$, $0 \leq t \leq 1$, of the form

$$\gamma(t) = (x_1(t), x_2(t), tx_3, tx_4, \ldots, tx_n)^T, \quad 0 \leq t \leq 1,$$

which passes through x as $t = 1$. Setting $s = \sum_{i=3}^{n} \theta_i x_i^2 = \theta_2 x_2^2 - \theta_1 x_1^2$ and $g^2 = d^T d = 1 - x_1^2 - x_2^2$, we should verify that one can define continuous functions $x_1(t), x_2(t)$ of $t \in [0, 1]$ satisfying the system of equations

$$\left\{ \begin{array}{rcl} \theta_1 x_1^2(t) - \theta_2 x_2^2(t) + t^2 s &=& 0, \\ x_1^2(t) + x_2^2(t) + t^2 g^2 &=& 1 \end{array} \right.$$

along with the boundary conditions

$$\begin{aligned} x_1(1) &= x_1, \\ x_2(1) &= x_2. \end{aligned}$$

Substituting $v_1(t) = x_1^2(t)$, $v_2(t) = x_2^2(t)$ and taking into account that $\theta_1, \theta_2 \geq 0, \theta_1 + \theta_2 > 0$, we get

$$\begin{aligned} (\theta_1 + \theta_2) v_1(t) &= \theta_2(1 - t^2 g^2) - t^2 s \\ &= \theta_2(1 - t^2 g^2 - t^2(\theta_2 x_2^2 - \theta_1 x_1^2) \\ &= \theta_2(1 - t^2[g^2 + x_2^2]) + t^2 \theta_1 x_1^2 \\ &= \theta_2(1 - t^2[1 - x_1^2]) + t^2 \theta_1 x_1^2; \\ (\theta_1 + \theta_2) v_2(t) &= \theta_1(1 - t^2 g^2) + t^2 s \\ &= \theta_1(1 - t^2 g^2) + t^2(\theta_2 x_2^2 - \theta_1 x_1^2) \\ &= \theta_1(1 - t^2[g^2 + x_1^2]) + t^2 \theta_2 x_2^2 \\ &= \theta_1(1 - t^2[1 - x_2^2]) + t^2 \theta_2 x_2^2. \end{aligned}$$

We see that $v_1(t)$, $v_2(t)$ are continuous, nonnegative and equal x_1^2, x_2^2, respectively, as $t = 1$. Taking $x_1(t) = \kappa_1 v_1^{1/2}$, $x_2(t) = \kappa_2 v_2^{1/2}(t)$ with properly chosen $\kappa_i = \pm 1$, $i = 1, 2$, we get the required curve $\gamma(\cdot)$. □

2. Prove that there exists a point $z^+ = (z_1, z_2, z_3, 0, 0, \ldots, 0)^T \in Y$ such that
(i) $z_1 z_2 = 0$.
(ii) given a point $u = (u_1, u_2, 0, 0, \ldots, 0)^T \in Y$, you can either (ii)(a) link u by continuous curves in Y both to z^+ and to $\bar{z}^+ = (z_1, z_2, -z_3, 0, 0, \ldots, 0)^T \in Y$, or (ii)(b) link u both to $z^- = (-z_1, -z_2, z_3, 0, 0, \ldots, 0)^T$ and $\bar{z}^- = (-z_1, -z_2, -z_3, 0, 0, \ldots, 0)^T$. (Note that $z^+ = -\bar{z}^-$, $\bar{z}^+ = -z^-$.)

Proof. Recall that $\theta_1 \geq \theta_2 \geq 0$ and $\theta_1 + \theta_2 > 0$. Consider two possible cases: $\theta_2 = 0$ and $\theta_2 > 0$.

Case of $\theta_2 = 0$. In this case it suffices to set $z^+ = (0, 1, 0, 0, \ldots, 0)^T$. Indeed, the point clearly belongs to Y and satisfies (i). Further, if $u \in Y$ is such that $u_3 = \cdots = u_n = 0$, then from $\theta_2 = 0$ and the definition of Y it immediately follows that $u_1 = 0$, $u_2 = \pm 1$. Thus, either u coincides with $z^+ = \bar{z}^+$, or u coincides with $z^- = \bar{z}^-$. In both cases, (ii) takes place.

Case of $\theta_2 > 0$. Let us set

$$\begin{aligned}
\tau &= \min\left[\tfrac{\theta_1}{\theta_1-\theta_3}; \tfrac{\theta_2}{\theta_2+\theta_3}\right] \\
&\text{[note that } \tau > 0 \text{ due to } \theta_1, \theta_2 > 0 \text{ and } -\theta_2 \leq \theta_3 \leq \theta_1\text{]}, \\
z_1 &= \sqrt{(\theta_1+\theta_2)^{-1}\,(\theta_2 - \tau[\theta_2+\theta_3])}, \\
z_2 &= \sqrt{(\theta_1+\theta_2)^{-1}\,(\theta_1 - \tau[\theta_1-\theta_3])}, \\
z_3 &= \sqrt{\tau}, \\
z^+ &= (z_1, z_2, z_3, 0, 0, \ldots, 0)^T.
\end{aligned}$$

It is immediately seen that z^+ is well defined and satisfies (i). Now let us verify that z^+ satisfies (ii) as well. Let $u = (u_1, u_2, 0, 0, \ldots, 0)^T \in Y$, and let the vector-function $z(t)$, $0 \leq t \leq \sqrt{\tau}$, be defined by the relations

$$\begin{aligned}
z_1(t) &= \sqrt{(\theta_1+\theta_2)^{-1}\left(\theta_2 - t^2[\theta_2+\theta_3]\right)}, \\
z_2(t) &= \sqrt{(\theta_1+\theta_2)^{-1}\left(\theta_1 - t^2[\theta_1-\theta_3]\right)}, \\
z_3(t) &= t, \\
z_i(t) &= 0, \;\; i = 4, \ldots, n.
\end{aligned}$$

It is immediately seen that $z(\cdot)$ is well defined, is continuous, takes its values in Y, and $z(\sqrt{\tau}) = z^+$. Now, $z(0)$ is the vector

$$\bar{u} \equiv \left(\sqrt{\frac{\theta_2}{\theta_1+\theta_2}}, \sqrt{\frac{\theta_1}{\theta_1+\theta_2}}, 0, 0, \ldots, 0\right)^T.$$

From $u \in Y$, $u_3 = \cdots = u_n = 0$ it immediately follows that $|u_i| = |\bar{u}_i|$, $i = 1, 2$. Now consider four possible cases:

$(++)$ $u_1 = \bar{u}_1, u_2 = \bar{u}_2$,

$(--)$ $u_1 = -\bar{u}_1, u_2 = -\bar{u}_2$,

$(+-)$ $u_1 = \bar{u}_1, u_2 = -\bar{u}_2$,

$(-+)$ $u_1 = -\bar{u}_1, u_2 = \bar{u}_2$.

- In the case of $(++)$ the continuous curve $\gamma(t) \equiv z(t) \in Y$, $0 \leq t \leq \sqrt{\tau}$, links u with z^+, while the continuous curve $\bar{\gamma}(t) = (z_1(t), z_2(t), -z_3(t), 0, 0, \ldots, 0)^T \in Y$ links u with $\bar{z}^+$, so that (ii)(1) takes place.

- In the case of $(--)$ the continuous curve $\gamma(t) \equiv -z(t) \in Y$, $0 \le t \le \sqrt{\tau}$, links u with $\bar{z}^- = -z^+$, while the continuous curve $\bar{\gamma}(t) = (-z_1(t), -z_2(t), z_3(t), 0, 0, \dots, 0)^T \in Y$ links u with $z^- = -\bar{z}^+$, so that (ii)(2) takes place.
- In the case of $(+-)$ the continuous curves $\gamma(t) \equiv (z_1(t), -z_2(t), z_3(t), 0, 0, \dots, 0)^T \in Y$ and $\bar{\gamma}(t) = (z_1(t), -z_2(t), -z_3(t), 0, 0, \dots, 0)^T$, $0 \le t \le \sqrt{\tau}$, link u either with both points of the pair $(z^+, \bar{z}^+)$, or with both points of the pair $(z^-, \bar{z}^-)$, depending on whether $z_1 = 0$ or $z_1 \ne 0$, $z_2 = 0$. (Note that at least one of these possibilities does take place due to $z_1 z_2 = 0$; see (i).) Thus, in the case in question at least (ii)(1) or (ii)(2) does hold.
- In the case of $(-+)$ the continuous curves $\gamma(t) \equiv (-z_1(t), z_2(t), z_3(t), 0, 0, \dots, 0)^T \in Y$ and $\bar{\gamma}(t) = (z_1(t), -z_2(t), -z_3(t), 0, 0, \dots, 0)^T$, $0 \le t \le \sqrt{\tau}$, link u either with both points of the pair $(z^-, \bar{z}^-)$ or with both points of the pair $(z^+, \bar{z}^+)$, depending on whether $z_1 = 0$ or $z_1 \ne 0$, $z_2 = 0$, so that here again (ii)(1) or (ii)(2) does hold. □

3. Conclude from 1 and 2 that Y satisfies the premise of Proposition 4.10.2, and thus complete the proof of Proposition 4.10.4.

Proof. Let z^+ be given by part 2, and let $x, x' \in Y$. By part 1, we can link in Y the point x with a point $v = (v_1, v_2, 0, 0, \dots, 0)^T \in Y$, and the point x' with a point $v' = (v_1', v_2', 0, 0, \dots, 0)^T \in Y$. If for both $u = v$ and $u = v'$ (ii)(a) holds, then we can link both v, v' by continuous curves with z^+. Thus, both x, x' can be linked in Y with z^+ as well. We see that both x, x' can be linked in Y with the set $\{z^+; -z^+\}$, as required in the premise of Proposition 4.10.2. The same conclusion is valid if for both $u = v$, $u = v'$ (ii)(b) holds. Here both x and x' can be linked in Y with z^-, and the premise of Proposition 4.10.2 holds true.

Now consider the case when for one of the points $u = v$, $u = v'$, say, for $u = v$, (ii)(a) holds, while for the other one (ii)(b) takes place. Here we can link in Y the point v (and thus the point x) with the point z^+, and we can link in Y the point v' (and thus the point x') with the point $\bar{z}^- = -z^+$. Thus, we can link in Y both x and x' with the set $\{z^+; -z^+\}$, and the premise of Proposition 4.10.2 holds true. Thus, Y satisfies the premise of Proposition 4.10.2. □

Exercise 4.53. 2. Let A_i, $i = 1, 2, 3$, satisfy the premise of (**SL**)(D). Assuming $A_1 = I$, prove that the set

$$H_1 = \{(v_1, v_2)^T \in \mathbf{R}^2 \mid \exists x \in \mathbf{S}^{n-1} : v_1 = f_2(x), v_3 = f_3(x)\}$$

is convex.

Proof. Let $\ell = \{(v_1, v_2)^t \in \mathbf{R}^2 \mid pv_1 + qv_2 + r = 0\}$ be a line in the plane. We should prove that $W = X_1 \cap \ell$ is a convex or, which is the same, connected set

(Exercise 4.49). There is nothing to prove when $W = \emptyset$. Assuming $W \neq \emptyset$, let us set

$$f(x) = rf_1(x) + pf_2(x) + qf_3(x).$$

It is immediately seen that f is a homogeneous quadratic form on $\mathbf{R}^n$, and that

$$W = F(Y),$$

where

$$F(x) = \begin{pmatrix} f_2(x) \\ f_3(x) \end{pmatrix},$$
$$Y \equiv \{x \in S^{n-1} : f(y) = 0\} = -Y.$$

By Proposition 4.10.4, the image Z of the set Y under the canonical projection $S^{n-1} \to \mathbf{P}^{n-1}$ is connected. Since F is even, $W = F(Y)$ is the same as $G(Z)$ for certain continuous mapping $G : Z \to \mathbf{R}^2$ (Proposition 4.10.3). Thus, W is connected by (C.2).

We have proved that Z_1 is convex; the compactness of Z_1 is evident. □

Exercise 4.54. Demonstrate by example that (**SL**)(C) not necessarily remains valid when skipping the assumption $n \geq 3$ in the premise.

A solution. The linear combination

$$A + 0.005B - 1.15C$$

of the matrices built in the solution to Exercise 4.47 is positive definite. □

Exercise 4.55. Let A, B, C be three 2×2 symmetric matrices such that the system of inequalities $x^T Ax \geq 0$, $x^T Bx \geq 0$ is strictly feasible and the inequality $x^T Cx$ is a consequence of the system.

1. Assume that there exists a nonsingular matrix Q such that both the matrices QAQ^T and QBQ^T are diagonal. Prove that then there exist $\lambda, \mu \geq 0$ such that $C \succeq \lambda A + \mu B$.

Solution. Without loss of generality we may assume that the matrices A, B themselves are diagonal. The case when at least one of the matrices A, B is positive semidefinite is trivially reducible to the usual $\mathcal{S}$-lemma, so that we can assume that both matrices A and B are not positive semidefinite. Since the system of inequalities $x^T Ax > 0, x^T Bx > 0$ is feasible, we conclude that the determinants of the matrices A, B are negative. Applying appropriate dilatations of the coordinate axes, swapping, if necessary, the coordinates and multiplying A, B by appropriate positive constants we may reduce the situation to the one where $x^T Ax = x_1^2 - x_2^2$ and either (a) $x^T Bx = \theta^2 x_1^2 - x_2^2$ or (b) $x^T Bx = -\theta^2 x_1^2 + x_2^2$ with certain $\theta > 0$.

Case of (a). Here the situation is immediately reducible to the one considered in the $\mathcal{S}$-lemma. Indeed, in this case one of the inequalities in the system $x^TAx \geq 0$, $x^TBx \geq 0$ is a consequence of the other inequality of the system. Thus, a consequence x^TCx of the system is in fact a consequence of a properly chosen *single* inequality from the system. Thus, by the $\mathcal{S}$-lemma either $C \succeq \lambda A$ or $C \succeq \lambda B$ with certain $\lambda \geq 0$.

Case of (b). Observe, first, that $\theta < 1$, since otherwise the system $x^TAx \geq 0$, $x^TBx \geq 0$ is not strictly feasible. When $\theta < 1$, the solution set of our system is the union of the following four angles D^{++}, D^{+-}, D^{--}, D^{-+}:

$$\begin{array}{rcl} D^{++} & = & \{x \mid x_1 \geq 0, \theta x_1 \leq x_2 \leq x_1\}, \\ D^{+-} & = & \text{reflection of } D^{++} \text{ without respect to the } x_1\text{-axis}, \\ D^{--} & = & -D^{++}, \\ D^{-+} & = & \text{reflection of } D^{++} \text{ with respect to the } x_2\text{-axis}. \end{array}$$

Now, the case when C is positive semidefinite is trivial—here $C \succeq 0 \times A + 0 \times B$. Thus, we may assume that one eigenvalue of C is negative; the other should be nonnegative, since otherwise $x^TCx < 0$ whenever $x \neq 0$, while we know that $x^TCx \geq 0$ at the (nonempty!) solution set of the system $x^TAx > 0$, $x^TBx > 0$. Since one eigenvalue of C is negative, and the other one is nonnegative, the set $X_C = \{x \mid x^TCx \geq 0\}$ is the union of a certain angle D (which can reduce to a ray) and the angle $-D$. Since the inequality $x^TCx \geq 0$ is a consequence of the system $x^TAx \geq 0$, $x^TBx \geq 0$, we have $D \cup (-D) \supset D^{++} \cup D^{+-} \cup D^{--} \cup D^{-+}$. Geometry says that the latter inclusion can be valid only when D contains two neighboring, with respect to the cyclic order, of the angles D^{++}, D^{+-}, D^{--}, D^{-+}. But in this case the inequality x^TCx is a consequence of an appropriate single inequality from the pair $x^TAx \geq 0$, $x^TBx \geq 0$. Namely, when $D \supset D^{++} \cup D^{+-}$ or $D \supset D^{--} \cup D^{-+}$, X_C contains the solution set of the inequality $x^TBx \geq 0$, while in the cases of $D \supset D^{+-} \cup D^{--}$ and of $D \supset D^{-+} \cup D^{++}$, X_C contains the solution set of the inequality $x^TAx \geq 0$. Applying the usual $\mathcal{S}$-lemma, we conclude that in the case of (b) there exist $\lambda, \mu \geq 0$ (with one of these coefficients equal to 0) such that $C \succeq \lambda A + \mu B$. □

Exercises to Lecture 6

Canonical barriers

Exercise 6.2. Prove Proposition 6.3.2.

Solution. As explained in the Hint, it suffices to consider the case of $\mathbf{K} = \mathbf{L}^k$. Let $x = (u, t) \in \text{int}\mathbf{L}^k$, and let $s = (v, \tau) = -\nabla L_k(x)$. We should prove that $s \in \text{int}\mathbf{L}^k$ and that $\nabla L_k(s) = -x$. By (6.3.2), one has

$$v = -\frac{2}{t^2-u^Tu}u, \quad \tau = \frac{2}{t^2-u^Tu}t$$
$$\Downarrow$$
$$\tau \geq 0 \text{ and } \tau^2 - v^Tv = \frac{4}{t^2-x^Tx} > 0$$
$$\Downarrow$$
$$s \in \text{int}\mathbf{L}^k;$$
$$-\nabla L_k(s) = \begin{pmatrix} -\frac{2}{\tau^2-v^Tv}v \\ \frac{2}{\tau^2-v^Tv}\tau \end{pmatrix} = \begin{pmatrix} \frac{2}{\frac{4}{t^2-u^Tu}}\frac{2}{t^2-u^Tu}u \\ \frac{2}{\frac{4}{t^2-u^Tu}}\frac{2}{t^2-u^Tu}t \end{pmatrix} = \begin{pmatrix} u \\ t \end{pmatrix}. \qquad \square$$

Scalings of canonical cones

Exercise 6.4. Prove the following:

1. Whenever $e \in \mathbf{R}^{k-1}$ is a unit vector and $\mu \in \mathbf{R}$, the linear transformation

$$L_{\mu,e}: \quad \begin{pmatrix} u \\ t \end{pmatrix} \mapsto \begin{pmatrix} u - [\mu t - (\sqrt{1+\mu^2}-1)e^Tu]e \\ \sqrt{1+\mu^2}t - \mu e^Tu \end{pmatrix} \qquad (*)$$

maps the cone $\mathbf{L}^k$ onto itself. Besides this, transformation $(*)$ preserves the space-time interval $x^TJ_kx \equiv -x_1^2 - \cdots - x_{k-1}^2 + x_k^2$:

$$[L_{\mu,e}x]^TJ_k[L_{\mu,e}x] = x^TJ_kx \quad \forall x \in \mathbf{R}^k \qquad [\Leftrightarrow L_{\mu,e}^TJ_kL_{\mu,e} = J_k]$$

and $L_{\mu,e}^{-1} = L_{\mu,-e}$.

Solution. Let $x = \begin{pmatrix} u \\ t \end{pmatrix} \in \mathbf{L}^k$. Denoting $s \equiv \begin{pmatrix} v \\ \tau \end{pmatrix} = L_{\mu,e}x$, we have

$$v = u - [\mu t - (\sqrt{1+\mu^2}-1)e^Tu]e, \quad \tau = \sqrt{1+\mu^2}t - \mu e^Tu$$
$$\Downarrow$$
$$\begin{aligned} \tau &= \sqrt{1+\mu^2}t - \mu e^Tu \geq \sqrt{1+\mu^2}t - \mu\|u\|_2 \geq 0 \\ &\quad [\text{since } t \geq \|u\|_2,\ \|e\|_2 = 1] \\ \tau^2 - v^Tv &= (1+\mu^2)t^2 - 2\mu\sqrt{1+\mu^2}te^Tu + \mu^2(e^Tu)^2 \\ &\quad -u^Tu - [\mu t - (\sqrt{1+\mu^2}-1)e^Tu]^2 \\ &\quad +2(e^Tu)[\mu t - (\sqrt{1+\mu^2}-1)e^Tu] \\ &= t^2 - u^Tu \qquad [\text{just arithmetic!}]. \end{aligned}$$

Thus, $x \in \mathbf{L}^k \Rightarrow L_{\mu,e}x \in \mathbf{L}^k$. To replace here $\Rightarrow$ with $\Leftrightarrow$, it suffices to verify (a straightforward computation!) that

$$L_{\mu,e}^{-1} = L_{\mu,-e},$$

so that both $L_{\mu,e}$ and its inverse map $\mathbf{L}^k$ onto itself. $\square$

Dikin ellipsoid

Exercise 6.8. Prove that if $\mathbf{K}$ is a canonical cone, K is the corresponding canonical barrier, and $X \in \mathrm{int}\mathbf{K}$, then the Dikin ellipsoid

$$W_X = \{Y \mid \|Y - X\|_X \le 1\} \qquad [\|H\|_X = \sqrt{\langle [\nabla^2 K(X)]H, H\rangle_E}]$$

is contained in $\mathbf{K}$.

Solution. According to the Hint, it suffices to verify the inclusion $W_X \subset \mathbf{K}$ in the following two particular cases:

A. $\mathbf{K} = \mathbf{S}^k_+$, $X = I_k$.

B. $\mathbf{K} = \mathbf{L}^k$, $X = \begin{pmatrix} 0_{k-1} \\ \sqrt{2} \end{pmatrix}$.

Note that in both cases $\nabla^2 K(X)$ is the unit matrix, so that all we need to prove is that the unit ball, centered at our particular X, is contained in our particular $\mathbf{K}$.

A: We should prove that if $\|H\|_F \le 1$, then $I + H \succeq 0$, which is evident. The modulae of eigenvalues of H are $\le \|H\|_F \le 1$, so that all these eigenvalues are ≥ -1.

B: We should prove that if $du \in \mathbf{R}^{k-1}$, $dt \in \mathbf{R}$ satisfy $dt^2 + du^T du \le 1$, then the point $\begin{pmatrix} 0_{k-1} \\ \sqrt{2} \end{pmatrix} + \begin{pmatrix} du \\ dt \end{pmatrix} = \begin{pmatrix} du \\ \sqrt{2} + dt \end{pmatrix}$ belongs to $\mathbf{L}^k$. In other words, we should verify that $\sqrt{2} + dt \ge 0$ (which is evident) and that $(\sqrt{2} + dt)^2 - du^T du \ge 0$. Here is the verification of the latter statement:

$$\begin{array}{rcl} (\sqrt{2} + dt)^2 - du^T du & = & 2 + 2\sqrt{2}dt + dt^2 - du^T du \\ & = & 1 + 2\sqrt{2}dt + 2dt^2 + (1 - dt^2 - du^T du) \\ & \ge & 1 + 2\sqrt{2}dt + 2dt^2 \\ & & [\text{since } dt^2 + du^T du \le 1] \\ & = & (1 + \sqrt{2}dt)^2 \ge 0. \qquad \square \end{array}$$

Exercise 6.9. Let $\mathbf{K}$ be a canonical cone:

$$\mathbf{K} = \mathbf{S}^{k_1}_+ \times \cdots \times \mathbf{S}^{k_p}_+ \times \mathbf{L}^{k_{p+1}} \times \cdots \times \mathbf{L}^{k_m} \subset E = \mathbf{S}^{k_1} \times \cdots \times \mathbf{S}^{k_p} \times \mathbf{R}^{k_{p+1}} \times \cdots \times \mathbf{R}^{k_m} \quad \text{(Cone)}$$

and let $X \in \mathrm{int}\mathbf{K}$. Prove the following:

2. Whenever $Y \in \mathbf{K}$, one has

$$\langle \nabla K(X), Y - X\rangle_E \le \theta(K).$$

4. The conic cap $\mathbf{K}_X$ is contained in the $\|\cdot\|_X$-ball, centered at X, of the radius $\theta(K)$:

$$Y \in \mathbf{K}_X \Rightarrow \|Y - X\|_X \le \theta(K).$$

Solution to 2 *and* 4. According to the Hint, it suffices to verify the statement in the case when $X = e(\mathbf{K})$ is the central point of $\mathbf{K}$. In this case the Hessian $\nabla^2 K(X)$ is just the unit matrix, whence, by Proposition 6.3.1, $\nabla K(X) = -X$.

2. We should prove that if $Y \in \mathbf{K}$, then $\langle \nabla K(X), Y - X\rangle_E \le \theta(K)$. The statement clearly is stable with respect to taking direct products, so that it suffices to prove it in the cases of $\mathbf{K} = \mathbf{S}^k_+$ and $\mathbf{K} = \mathbf{L}^k$.

In the case of $\mathbf{K} = \mathbf{S}^k_+$ what should be proved is

$$\forall H \in \mathbf{S}^k_+ : \quad \mathrm{Tr}(I_k - H) \le k,$$

which is evident.

In the case of $\mathbf{K} = \mathbf{L}^k$ what should be proved is

$$\forall \left(\begin{pmatrix} u \\ t \end{pmatrix} : \|u\|_2 \le t \right) \quad \sqrt{2}(\sqrt{2} - t) \le 2,$$

which again is evident.

4. We should prove that if $X = e(\mathbf{K})$, $\langle -\nabla K(X), X - Y\rangle_E \ge 0$, and $Y \in \mathbf{K}$, then $\|Y - X\|_E \le \theta(K)$.

Let $Y \in \mathbf{K}$ be such that $\langle -\nabla K(X), X - Y\rangle_E \ge 0$, i.e., such that $\langle X, X - Y\rangle_E \ge 0$. We may think of Y as of a collection of a block-diagonal symmetric positive semidefinite matrix H with diagonal blocks of the sizes $k_1, \dots, k_p$ and $m - p$ vectors $\begin{pmatrix} u_i \\ t_i \end{pmatrix} \in \mathbf{L}^{k_i}$, $i = p+1, \dots, m$ (see (Cone)); the condition $\langle X, X - Y\rangle_E \ge 0$ now becomes

$$\mathrm{Tr}(\underbrace{I - H}_{D}) + \sum_{i=p+1}^{m} \sqrt{2}(\sqrt{2} - t_i) \ge 0. \qquad (*)$$

We now have, denoting by D_j the eigenvalues of D and by $n = \sum_{i=1}^{p} k_i$ the row size of D:

$$\begin{aligned}
\|X - Y\|_X^2 &= \|X - Y\|_E^2 = \mathrm{Tr}((I - H)^2) + \sum_{i=p+1}^{m} \left((\sqrt{2} - t_i)^2 + u_i^T u_i \right) \\
&= \sum_{j=1}^{n} D_j^2 + \sum_{i=p+1}^{m} \left(2 - 2\sqrt{2} t_i + t_i^2 + u_i^T u_i \right) \\
&\le \sum_{j=1}^{n} D_j^2 + \sum_{i=p+1}^{m} \left(2 - 2\sqrt{2} t_i + 2 t_i^2 \right) \\
&\qquad\qquad [\text{since } u_i^T u_i \le t_i^2] \\
&= \sum_{j=1}^{n} D_j^2 + \sum_{i=p+1}^{m} \left(1 + (1 - \sqrt{2} t_i)^2 \right).
\end{aligned}$$

Denoting $q = m - p$, $D_{n+i} = 1 - \sqrt{2}t_{p+i}$, $i = 1, \ldots, q$, we come to the relation

$$\|X - Y\|_X^2 = q + \sum_{j=1}^{n+q} D_j^2, \tag{14}$$

while $(*)$ and relations $H \succeq 0$, $t_i \geq 0$ imply that

$$\begin{aligned} D_j &\leq 1, \quad j = 1, \ldots, n+q, \\ \sum_{j=1}^{n+q} D_j &\geq -q, \end{aligned} \tag{15}$$

Let A be the maximum of the right-hand side in (14) over D_j's satisfying (15), and let $D^* = (D_1^*, \ldots, D_{n+q}^*)^T$ be the corresponding maximizer (which clearly exists—(15) defines a compact set!). In the case of $n + q = 1$ we clearly have $A = 1 + q$. Now let $n + q > 1$. We claim that among $n + q$ coordinates of D^*, $n + q - 1$ are equal 1, and the remaining coordinate equals to $-(n + 2q - 1)$. Indeed, if there were at least two entries in D^* which are less than 1, then subtracting from one of them a $\delta \neq 0$ small enough in absolute value and adding the same δ to the other coordinate, we preserve the feasibility of the perturbed point with respect to (15), and, with properly chosen sign of δ, increase $\sum_j D_j^2$, which is impossible. Thus, at least $n + q - 1$ coordinates of D^* are equal to 1. Among the points with this property that satisfy (15), the one with the largest $\sum_j D_j^2$ clearly has the remaining coordinate equal to $1 - n - 2q$, as claimed.

From our analysis it follows that

$$A = \begin{cases} q + 1, & n + q = 1, \\ q + (n + q - 1) + (n + 2q - 1)^2 = (n + 2q - 1)(n + 2q), & n + q > 1. \end{cases}$$

Recalling that $\theta(K) = n + 2q$ and taking into account (14) and the origin of A, we get

$$\|X - Y\|_X \leq \theta(K),$$

as claimed. □

More on canonical barriers

Exercise 6.11 Prove that if $\mathbf{K}$ is a canonical cone, K is the associated canonical barrier, $X \in \mathrm{int}\mathbf{K}$, and $H \in \mathbf{K}$, $H \neq 0$, then

$$\inf_{t \geq 0} K(X + tH) = -\infty. \tag{$*$}$$

Derive from this fact that

(!!) *Whenever $\mathcal{N}$ is an affine plane that intersects the interior of $\mathbf{K}$, K is below bounded on the intersection $\mathcal{N} \cap \mathbf{K}$ if and only if the intersection is bounded.*

Solution. As explained in the Hint, it suffices to verify $(*)$ in the particular case when X is the central point of $\mathbf{K}$. It is also clear that $(*)$ is stable with respect to taking direct products, so that we can restrict ourselves with the cases of $\mathbf{K} = \mathbf{S}^k_+$ and $\mathbf{K} = \mathbf{L}^k$.

Case of $\mathbf{K} = \mathbf{S}^k_+$, $X = e(\mathbf{S}^k_+) = I_k$. Denoting by $H_i \geq 0$ the eigenvalues of H and noting that at least one of H_i is > 0 due to $H \neq 0$, we have for $t > 0$

$$K(X + tH) = -\ln \mathrm{Det}(I_k + tH) = -\sum_{i=1}^{k} \ln(1 + tH_i) \to -\infty, \ t \to \infty.$$

Case of $\mathbf{K} = \mathbf{L}^k$, $X = e(\mathbf{L}^k) = \begin{pmatrix} 0_{k-1} \\ \sqrt{2} \end{pmatrix}$. Setting $H = \begin{pmatrix} u \\ s \end{pmatrix}$, we have $s \geq \|u\|_2$ and $s > 0$ due to $H \neq 0$. For $t > 0$ we have

$$\begin{array}{rcl} K(X + tH) & = & -\ln((\sqrt{2} + ts)^2 - t^2 u^T u) \\ & = & -\ln(2 + 2\sqrt{2}ts + t^2(s^2 - u^T u)) \\ & \leq & -\ln(2 + 2\sqrt{2}ts) \to -\infty, \ t \to \infty. \end{array}$$

To derive (!!), note that if $U = \mathcal{N} \cap \mathbf{K}$ is bounded, then K is below bounded on U just in view of convexity of K. (Moreover, from the fact that K is a barrier for $\mathbf{K}$ it follows that K attains its minimum on U.) It remains to prove that if U is unbounded, then K is not below bounded on U. If U is unbounded, there exists a nonzero direction $H \in \mathbf{K}$ that is parallel to $\mathcal{N}$. (Take as H a limiting point of the sequence $\|Y_i - X\|_2^{-1}(Y_i - X)$, where $Y_i \in U$, $\|Y_i\|_2 \to \infty$ as $i \to \infty$, and X is a once for ever fixed point from U.) By $(*)$, K is not below bounded on the ray $\{X + tH \mid t \geq 0\}$, and this ray clearly belongs to U. Thus, K is not below bounded on U. □

Primal path-following method

Exercise 6.15. Looking at the data in the table at the end of section 6.5.3, do you believe that the corresponding method is exactly the short-step primal path-following method from Theorem 6.5.1 with the stepsize policy (6.5.31)?

Solution. The table cannot correspond to the indicated method. Indeed, we see from the table that the duality gap along the 12-iteration trajectory is reduced by factor of about 10^6. Since the duality gap in a short-step method is nearly inverse proportional to the value of the penalty, the latter in the process of our 12 iterations should be increased by a factor of order of 10^5–10^6. In our case $\theta(K) = 3$, and the policy (6.5.31) increases the penalty at an iteration by the factor $(1 + 0.1/\sqrt{3}) \approx 1.0577$. With this policy, in 12 iterations the penalty would be increased by $1.0577^{12} < 2$, which is very far from 10^5! □

Infeasible start path-following method

Exercise 6.18. Consider the problem

$$\max_X \left\{ \langle \widetilde{C}, Y \rangle_{\widetilde{E}} \mid Y \in (\mathcal{M} + R) \cap \widetilde{\mathbf{K}} \right\}, \qquad \text{(Aux)}$$

where

$$\widetilde{\mathbf{K}} = \mathbf{K} \times \mathbf{K} \times \underbrace{\mathbf{S}^1_+}_{=\mathbf{R}_+} \times \underbrace{\mathbf{S}^1_+}_{=\mathbf{R}_+}$$

(**K** is a canonical cone),

$$\mathcal{M} = \left\{ \begin{pmatrix} U \\ V \\ s \\ r \end{pmatrix} \;\middle|\; \begin{array}{c} U + rB \in \mathcal{L}, \\ V - rC \in \mathcal{L}^\perp, \\ \langle C, U\rangle_E - \langle B, V\rangle_E + s = 0 \end{array} \right\}$$

is a linear subspace in the space $\widetilde{E}$ where the cone $\widetilde{\mathbf{K}}$ lives, and

$$C \in \mathcal{L}, \quad B \in \mathcal{L}^\perp.$$

It is given that the problem (Aux) is feasible. Prove that the feasible set of (Aux) is unbounded.

Solution. According to the Hint, we should prove that the linear space $\mathcal{M}^\perp$ does not intersect int$\widetilde{\mathbf{K}}$.

Let us compute $\mathcal{M}^\perp$. A collection

$$\begin{pmatrix} \xi \\ \eta \\ s \\ r \end{pmatrix}, \ \xi, \eta \in E, \ s, r \in \mathbf{S}^1 = \mathbf{R}$$

is in $\mathcal{M}^\perp$ if and only if the linear equation in variables X, S, σ, τ

$$\langle X, \xi\rangle_E + \langle S, \eta\rangle_E + \sigma s + \tau r = 0$$

is a corollary of the system of linear equations

$$X + \tau B \in \mathcal{L}, \quad S - \tau C \in \mathcal{L}^\perp, \quad \langle X, C\rangle_E - \langle S, B\rangle_E + \sigma = 0.$$

By linear algebra, this is the case if and only if there exist $U \in \mathcal{L}^\perp$, $V \in \mathcal{L}$, and a real λ such that

$$\begin{array}{lrcl} \text{(a)} & \xi & = & U + \lambda C, \\ \text{(b)} & \eta & = & V - \lambda B, \\ \text{(c)} & s & = & \lambda, \\ \text{(d)} & r & = & \langle U, B\rangle_E - \langle V, C\rangle_E. \end{array} \qquad \text{(Pr)}$$

We have obtained a parameterization of $\mathcal{M}^\perp$ via the parameters U, V, λ running through, respectively, $\mathcal{L}^\perp$, $\mathcal{L}$, and **R**. Now assume, contrary to what should be proved, that the intersection of $\mathcal{M}^\perp$ and int$\widetilde{\mathbf{K}}$ is nonempty. In other words, assume that there exist $U \in \mathcal{L}^\perp$, $V \in \mathcal{L}$, and $\lambda \in \mathbf{R}$ which, being substituted in (Pr), result in a collection (ξ, η, s, r) such that $\xi \in$ int**K**, $\eta \in$ int**K**, $s > 0$,

$r > 0$. From (Pr)(c) it follows that $\lambda > 0$. Since (Pr) is homogeneous, we may normalize our U, V, λ to make $\lambda = 1$, still keeping $\xi \in \mathrm{int}\mathbf{K}$, $\eta \in \mathrm{int}\mathbf{K}$, $s > 0$, $r > 0$. Assuming $\lambda = 1$ and taking into account that $U, B \in \mathcal{L}^\perp$, $V, C \in \mathcal{L}$, we get from (Pr)(a)–(b)

$$\langle \xi, \eta \rangle_E = \langle C, V \rangle_E - \langle B, U \rangle_E.$$

Adding this equality to (Pr)(d), we get

$$\langle \xi, \eta \rangle_E + r = 0,$$

which is impossible, since both $r > 0$ and $\langle \xi, \eta \rangle_E > 0$. (Recall that the cone $\mathbf{K}$ is self-dual and $\xi, \eta \in \mathrm{int}\mathbf{K}$.) We have come to the desired contradiction. □

Exercise 6.19. Let $\bar{X}, \bar{S}$ be a strictly feasible pair of primal-dual solutions to the primal-dual pair of problems

$$\min_X \{\langle C, X \rangle_E \mid X \in (\mathcal{L} - B) \cap \mathbf{K}\}, \tag{P}$$
$$\max_S \left\{\langle B, S \rangle_E \mid S \in (\mathcal{L}^\perp + C) \cap \mathbf{K}\right\} \tag{D}$$

so that there exists $\gamma \in (0, 1]$ such that

$$\begin{array}{rcll} \gamma \|X\|_E & \leq & \langle \bar{S}, X \rangle_E & \forall X \in \mathbf{K}, \\ \gamma \|S\|_E & \leq & \langle \bar{X}, S \rangle_E & \forall S \in \mathbf{K}. \end{array}$$

Prove that if

$$Y = \begin{pmatrix} X \\ S \\ \sigma \\ \tau \end{pmatrix}$$

is feasible for (Aux), then

$$\begin{array}{rcl} \|Y\|_{\widetilde{E}} & \leq & \alpha\tau + \beta, \\ & & \alpha = \gamma^{-1}\left[\langle \bar{X}, C \rangle_E - \langle \bar{S}, B \rangle_E\right] + 1, \\ & & \beta = \gamma^{-1}\left[\langle \bar{X} + B, D \rangle_E + \langle \bar{S} - C, P \rangle_E + d\right]. \end{array} \tag{16}$$

Solution. We have $\bar{X} = \bar{U} - B$, $\bar{U} \in \mathcal{L}$, $\bar{S} = \bar{V} + C$, $\bar{V} \in \mathcal{L}^\perp$. Taking into account the constraints of (Aux), we get

$$\begin{array}{rcl} \langle \bar{U}, S - \tau C - D \rangle_E & = & 0 \Rightarrow \\ \langle \bar{U}, S \rangle_E & = & \langle \bar{U}, \tau C + D \rangle_E \Rightarrow \\ \langle \bar{X}, S \rangle_E & = & -\langle B, S \rangle_E + \langle \bar{U}, \tau C + D \rangle_E, \\ \langle \bar{V}, X + \tau B - P \rangle_E & = & 0 \Rightarrow \\ \langle \bar{V}, X \rangle_E & = & \langle \bar{V}, -\tau B + P \rangle_E \Rightarrow \\ \langle \bar{S}, X \rangle_E & = & \langle C, X \rangle_E + \langle \bar{V}, -\tau B + P \rangle_E, \\ \Rightarrow & & \\ \langle \bar{X}, S \rangle_E + \langle \bar{S}, X \rangle_E & = & [\langle C, X \rangle_E - \langle B, S \rangle_E] + \tau\left[\langle \bar{U}, C \rangle_E - \langle \bar{V}, B \rangle_E\right] \\ & & + \left[\langle \bar{U}, D \rangle_E + \langle \bar{V}, P \rangle_E\right] \\ & = & d - \sigma + \tau\left[\langle \bar{U}, C \rangle_E - \langle \bar{V}, B \rangle_E\right] \\ & & + \left[\langle \bar{U}, D \rangle_E + \langle \bar{V}, P \rangle_E\right] \end{array}$$

whence

$$\gamma\left[\|X\|_E + \|S\|_E\right] + \sigma \quad \leq \quad \tau\left[\langle \bar{U}, C\rangle_E - \langle \bar{V}, B\rangle_E\right] + \left[\langle \bar{U}, D\rangle_E + \langle \bar{V}, P\rangle_E + d\right],$$

and (16) follows (recall that $\langle C, B\rangle_E = 0$). □

Exercise 6.23. Let $\mathbf{K}$ be a canonical cone, let the primal-dual pair of problems

$$\min_X \{\langle C, X\rangle_E \mid X \in (\mathcal{L} - B) \cap \mathbf{K}\}, \qquad \text{(P)}$$
$$\max_S \left\{\langle B, S\rangle_E \mid S \in (\mathcal{L}^\perp + C) \cap \mathbf{K}\right\} \qquad \text{(D)}$$

be strictly primal-dual feasible and be normalized by $\langle C, B\rangle_E = 0$, let (X_*, S_*) be a primal-dual optimal solution to the pair, and let X, S ϵ-satisfy the feasibility and optimality conditions for (P), (D), i.e.,

$$\begin{array}{ll} \text{(a)} & X \in \mathbf{K} \cap (\mathcal{L} - B + \Delta X),\ \|\Delta X\|_E \leq \epsilon, \\ \text{(b)} & S \in \mathbf{K} \cap (\mathcal{L}^\perp + C + \Delta S),\ \|\Delta S\|_E \leq \epsilon, \\ \text{(c)} & \langle C, X\rangle_E - \langle B, S\rangle_E \leq \epsilon. \end{array}$$

Prove that

$$\begin{array}{rcl} \langle C, X\rangle_E - \mathrm{Opt(P)} & \leq & \epsilon(1 + \|X_* + B\|_E), \\ \mathrm{Opt(D)} - \langle B, S\rangle_E & \leq & \epsilon(1 + \|S_* - C\|_E). \end{array}$$

Solution. We have $S - C - \Delta S \in \mathcal{L}^\perp$, $X_* + B \in \mathcal{L}$, whence

$$\begin{array}{rcl} 0 & = & \langle S - C - \Delta S, X_* + B\rangle_E \\ & = & \underbrace{\langle S, X_*\rangle_E}_{\geq 0:\, X_*, S \in \mathbf{K} = \mathbf{K}_*} - \mathrm{Opt(P)} + \langle S, B\rangle_E + \langle -\Delta S, X_* + B\rangle_E \\ & & \Rightarrow \\ -\mathrm{Opt(P)} & \leq & -\langle S, B\rangle_E + \langle \Delta S, X_* + B\rangle_E \leq -\langle S, B\rangle_E + \epsilon\|X_* + B\|_E. \end{array}$$

Combining the resulting inequality and (c), we get the first of the inequalities to be proved. The second is given by symmetric reasoning. □