# MOSEK Modeling Cookbook

*Release 3.1*

MOSEK ApS

**03 July 2019**

# Contents

# Chapter 1

# Preface

This cookbook is about model building using convex optimization. It is intended as a modeling guide for the **MOSEK** optimization package. However, the style is intentionally quite generic without specific **MOSEK** commands or API descriptions.

There are several excellent books available on this topic, for example the books by Ben-Tal and Nemirovski *[BenTalN01]* and Boyd and Vandenberghe *[BV04]*, which have both been a great source of inspiration for this manual. The purpose of this manual is to collect the material which we consider most relevant to our users and to present it in a practical self-contained manner; however, we highly recommend the books as a supplement to this manual.

Some textbooks on building models using optimization (or mathematical programming) introduce various concepts through practical examples. In this manual we have chosen a different route, where we instead show the different sets and functions that can be modeled using convex optimization, which can subsequently be combined into realistic examples and applications. In other words, we present simple *convex building blocks*, which can then be combined into more elaborate convex models. We call this approach *extremely disciplined modeling*. With the advent of more expressive and sophisticated tools like conic optimization, we feel that this approach is better suited.

## Content

We begin with a comprehensive chapter on *linear optimization*, including modeling examples, duality theory and infeasibility certificates for linear problems. Linear problems are optimization problems of the form

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \\ & x \geq 0. \end{array}$$

*Conic optimization* is a generalization of linear optimization which handles problems of the form:

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \\ & x \in K, \end{array}$$

where $K$ is a *convex cone*. Various families of convex cones allow formulating different types of nonlinear constraints. The following chapters present modeling with four types of convex cones:

- *quadratic cones*,
- *power cone*,
- *exponential cone*,
- *semidefinite cone*.

It is "well-known" in the convex optimization community that this family of cones is sufficient to express almost all convex optimization problems appearing in practice.

Next we discuss issues arising in *practical optimization*, and we wholeheartedly recommend this short chapter to all readers before moving on to implementing mathematical models with real data.

Following that, we present a general *duality and infeasibility* theory for conic problems. Finally we diverge slightly from the topic of conic optimization and introduce the language of *mixed-integer optimization* and we discuss the relation between *convex quadratic optimization* and conic quadratic optimization.

# Chapter 2

# Linear optimization

In this chapter we discuss various aspects of linear optimization. We first introduce the basic concepts of linear optimization and discuss the underlying geometric interpretations. We then give examples of the most frequently used reformulations or modeling tricks used in linear optimization, and finally we discuss duality and infeasibility theory in some detail.

## 2.1 Introduction

### 2.1.1 Basic notions

The most basic type of optimization is *linear optimization*. In linear optimization we minimize a linear function given a set of linear constraints. For example, we may wish to minimize a linear function

$$x_1 + 2x_2 - x_3$$

under the constraints that

$$x_1 + x_2 + x_3 = 1, \quad x_1, x_2, x_3 \geq 0.$$

The function we minimize is often called the *objective function*; in this case we have a linear objective function. The constraints are also linear and consist of both linear *equalities* and *inequalities*. We typically use more compact notation

$$\begin{array}{ll} \text{minimize} & x_1 + 2x_2 - x_3 \\ \text{subject to} & x_1 + x_2 + x_3 = 1, \\ & x_1, x_2, x_3 \geq 0, \end{array} \tag{2.1}$$

and we call (2.1) a *linear optimization problem*. The domain where all constraints are satisfied is called the *feasible set*; the feasible set for (2.1) is shown in Fig. 2.1.

For this simple problem we see by inspection that the *optimal value* of the problem is $-1$ obtained by the *optimal solution*

$$(x_1^\star, x_2^\star, x_3^\star) = (0, 0, 1).$$

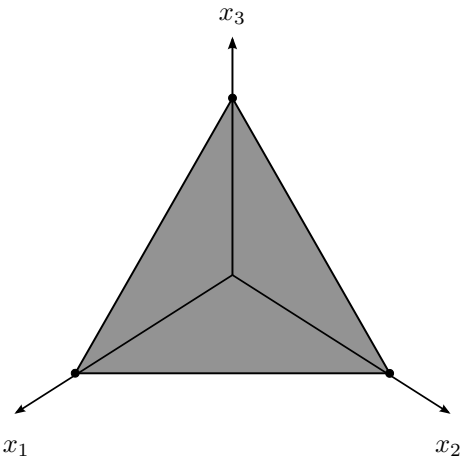



Fig. 2.1: Feasible set for $x_1 + x_2 + x_3 = 1$ and $x_1, x_2, x_3 \geq 0$.

Linear optimization problems are typically formulated using matrix notation. The standard form of a linear minimization problem is:

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \\ & x \geq 0. \end{array} \tag{2.2}$$

For example, we can pose (2.1) in this form with

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad c = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}.$$

There are many other formulations for linear optimization problems; we can have different types of constraints,

$$Ax = b, \quad Ax \geq b, \quad Ax \leq b, \quad l^c \leq Ax \leq u^c,$$

and different bounds on the variables

$$l^x \leq x \leq u^x$$

or we may have no bounds on some $x_i$, in which case we say that $x_i$ is a *free variable*. All these formulations are equivalent in the sense that by simple linear transformations and introduction of auxiliary variables they represent the same set of problems. The important feature is that the objective function and the constraints are all *linear* in $x$.

### 2.1.2 Geometry of linear optimization

A *hyperplane* is a subset of $\mathbb{R}^n$ defined as $\{x \mid a^T(x - x_0) = 0\}$ or equivalently $\{x \mid a^T x = \gamma\}$ with $a^T x_0 = \gamma$, see Fig. 2.2.

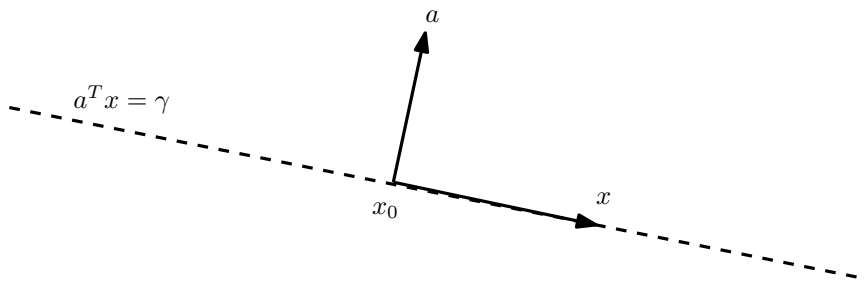



Fig. 2.2: The dashed line illustrates a hyperplane $\{x \mid a^T x = \gamma\}$.

Thus a linear constraint

$$Ax = b$$

with $A \in \mathbb{R}^{m \times n}$ represents an intersection of $m$ hyperplanes.

Next, consider a point $x$ above the hyperplane in Fig. 2.3. Since $x - x_0$ forms an acute angle with $a$ we have that $a^T(x - x_0) \geq 0$, or $a^T x \geq \gamma$. The set $\{x \mid a^T x \geq \gamma\}$ is called a *halfspace*. Similarly the set $\{x \mid a^T x \leq \gamma\}$ forms another halfspace; in Fig. 2.3 it corresponds to the area below the dashed line.

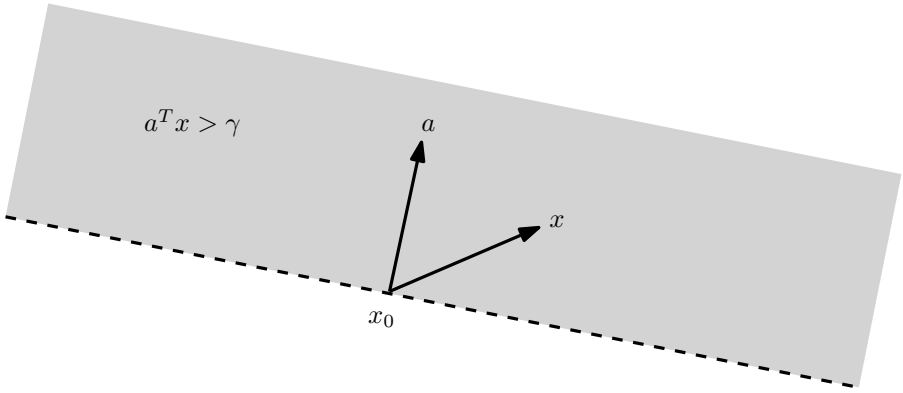



Fig. 2.3: The grey area is the halfspace $\{x \mid a^T x \geq \gamma\}$.

A set of linear inequalities

$$Ax \leq b$$

corresponds to an intersection of halfspaces and forms a *polyhedron*, see Fig. 2.4.

The polyhedral description of the feasible set gives us a very intuitive interpretation of linear optimization, which is illustrated in Fig. 2.5. The dashed lines are normal to the objective $c = (-1, 1)$, and to minimize $c^T x$ we move as far as possible in the opposite direction of $c$, to the furthest position where a dashed line intersect the polyhedron; an optimal solution is therefore always either a vertex of the polyhedron, or an entire facet of the polyhedron may be optimal.

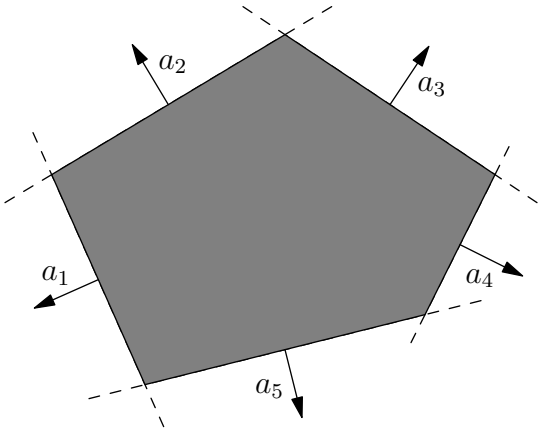



Fig. 2.4: A polyhedron formed as an intersection of halfspaces.

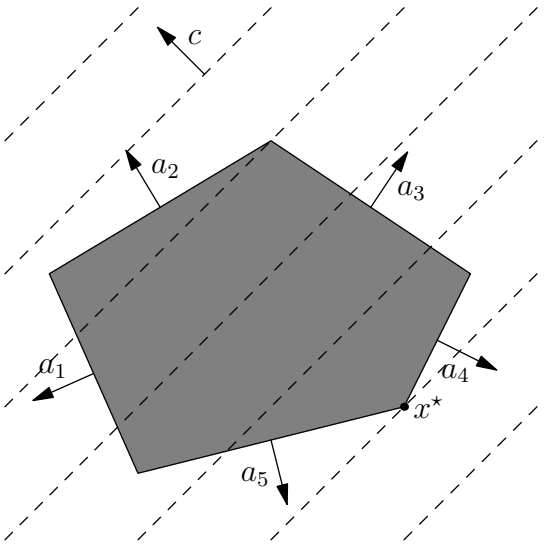



Fig. 2.5: Geometric interpretation of linear optimization. The optimal solution $x^\star$ is at a point where the normals to $c$ (the dashed lines) intersect the polyhedron.

The polyhedron shown in the figure is nonempty and bounded, but this is not always the case for polyhedra arising from linear inequalities in optimization problems. In such cases the optimization problem may be infeasible or unbounded, which we will discuss in detail in Sec. 2.3.

## 2.2 Linear modeling

In this section we present useful reformulation techniques and standard tricks which allow constructing more complicated models using linear optimization. It is also a guide through the types of constraints which can be expressed using linear (in)equalities.

### 2.2.1 Maximum

The inequality $t \geq \max\{x_1, \ldots, x_n\}$ is equivalent to a simultaneous sequence of $n$ inequalities

$$t \geq x_i, \ i = 1, \ldots, n$$

and similarly $t \leq \min\{x_1, \ldots, x_n\}$ is the same as

$$t \leq x_i, \ i = 1, \ldots, n.$$

Of course the same reformulation applies if each $x_i$ is not a single variable but a linear expression. In particular, we can consider convex piecewise-linear functions $f : \mathbb{R}^n \mapsto \mathbb{R}$ defined as the maximum of affine functions (see Fig. 2.6):

$$f(x) := \max_{i=1,\ldots,m} \{a_i^T x + b_i\}$$

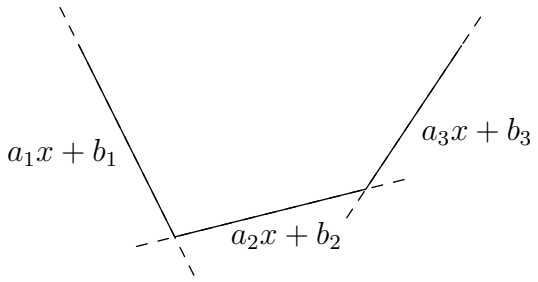



Fig. 2.6: A convex piecewise-linear function (solid lines) of a single variable $x$. The function is defined as the maximum of 3 affine functions.

The epigraph $f(x) \leq t$ (see Sec. 12) has an equivalent formulation with $m$ inequalities:

$$a_i^T x + b_i \leq t, \quad i = 1, \ldots, m.$$

Piecewise-linear functions have many uses linear in optimization; either we have a convex piecewise-linear formulation from the onset, or we may approximate a more complicated (nonlinear) problem using piecewise-linear approximations, although with modern nonlinear optimization software it is becoming both easier and more efficient to directly formulate and solve nonlinear problems without piecewise-linear approximations.

### 2.2.2 Absolute value

The absolute value of a scalar variable is a special case of maximum

$$|x| := \max\{x, -x\},$$

so we can model the epigraph $|x| \leq t$ using two inequalities

$$-t \leq x \leq t.$$

### 2.2.3 The $\ell_1$ norm

All norms are convex functions, but the $\ell_1$ and $\ell_\infty$ norms are of particular interest for linear optimization. The $\ell_1$ norm of vector $x \in \mathbb{R}^n$ is defined as

$$\|x\|_1 := |x_1| + |x_2| + \cdots + |x_n|.$$

To model the epigraph

$$\|x\|_1 \leq t, \tag{2.3}$$

we introduce the following system

$$|x_i| \leq z_i,\ i = 1, \ldots, n, \quad \sum_{i=1}^{n} z_i = t, \tag{2.4}$$

with additional (auxiliary) variable $z \in \mathbb{R}^n$. Clearly (2.3) and (2.4) are equivalent, in the sense that they have the same projection onto the space of $x$ and $t$ variables. Therefore, we can model (2.3) using linear (in)equalities

$$-z_i \leq x_i \leq z_i, \quad \sum_{i=1}^{n} z_i = t, \tag{2.5}$$

with auxiliary variables $z$. Similarly, we can describe the epigraph of the norm of an affine function of $x$,

$$\|Ax - b\|_1 \leq t$$

as

$$-z_i \leq a_i^T x - b_i \leq z_i, \quad \sum_{i=1}^{n} z_i = t,$$

where $a_i$ is the $i-$th row of $A$ (taken as a column-vector).

**Example 2.1** (Basis pursuit)**.** The $\ell_1$ norm is overwhelmingly popular as a convex approximation of the cardinality (i.e., number on nonzero elements) of a vector $x$. For example, suppose we are given an underdetermined linear system

$$Ax = b$$

where $A \in \mathbb{R}^{m\times n}$ and $m \ll n$. The *basis pursuit* problem

$$\begin{array}{ll} \text{minimize} & \|x\|_1 \\ \text{subject to} & Ax = b, \end{array} \tag{2.6}$$

uses the $\ell_1$ norm of $x$ as a heuristic for finding a sparse solution (one with many zero elements) to $Ax = b$, i.e., it aims to represent $b$ as a linear combination of few columns of $A$. Using (2.5) we can pose the problem as a linear optimization problem,

$$\begin{array}{lllllll} \text{minimize} & & & & e^T z & & \\ \text{subject to} & -z & \leq & x & \leq & z, \\ & & & Ax & = & b, \end{array} \tag{2.7}$$

where $e = (1, \ldots, 1)^T$.

### 2.2.4 The $\ell_\infty$ norm

The $\ell_\infty$ norm of a vector $x \in \mathbb{R}^n$ is defined as

$$\|x\|_\infty := \max_{i=1,\ldots,n} |x_i|,$$

which is another example of a simple piecewise-linear function. Using Sec. 2.2.2 we model

$$\|x\|_\infty \leq t$$

as

$$-t \leq x_i \leq t, \quad i = 1, \ldots, n.$$

Again, we can also consider affine functions of $x$, i.e.,

$$\|Ax - b\|_\infty \leq t,$$

which can be described as

$$-t \leq a_i^T x - b \leq t, \quad i = 1, \ldots, n.$$

**Example 2.2** (Dual norms)**.** It is interesting to note that the $\ell_1$ and $\ell_\infty$ norms are dual. For any norm $\|\cdot\|$ on $\mathbb{R}^n$, the *dual norm* $\|\cdot\|_*$ is defined as

$$\|x\|_* = \max\{x^T v \mid \|v\| \le 1\}.$$

Let us verify that the dual of the $\ell_\infty$ norm is the $\ell_1$ norm. Consider

$$\|x\|_{*,\infty} = \max\{x^T v \mid \|v\|_\infty \le 1\}.$$

Obviously the maximum is attained for

$$v_i = \begin{cases} +1, & x_i \ge 0, \\ -1, & x_i < 0, \end{cases}$$

i.e., $\|x\|_{*,\infty} = \sum_i |x_i| = \|x\|_1$. Similarly, consider the dual of the $\ell_1$ norm,

$$\|x\|_{*,1} = \max\{x^T v \mid \|v\|_1 \le 1\}.$$

To maximize $x^T v$ subject to $|v_1| + \cdots + |v_n| \le 1$ we simply pick the element of $x$ with largest absolute value, say $|x_k|$, and set $v_k = \pm 1$, so that $\|x\|_{*,1} = |x_k| = \|x\|_\infty$. This illustrates a more general property of dual norms, namely that $\|x\|_{**} = \|x\|$.

### 2.2.5 Homogenization

Consider the linear-fractional problem

$$\begin{array}{lrcl} \text{minimize} & \frac{a^T x + b}{c^T x + d} & & \\ \text{subject to} & c^T x + d & > & 0, \\ & Fx & = & g. \end{array} \tag{2.8}$$

Perhaps surprisingly, it can be turned into a linear problem if we homogenize the linear constraint, i.e. replace it with $Fy = gz$ for a single variable $z \in \mathbb{R}$. The full new optimization problem is

$$\begin{array}{lrcl} \text{minimize} & a^T y + bz & & \\ \text{subject to} & c^T y + dz & = & 1, \\ & Fy & = & gz, \\ & z & \ge & 0. \end{array} \tag{2.9}$$

If $x$ is a feasible point in (2.8) then $z = (c^T x + d)^{-1}$, $y = xz$ is feasible for (2.9) with the same objective value. Conversely, if $(y, z)$ is feasible for (2.9) then $x = y/z$ is feasible in (2.8) and has the same objective value, at least when $z \neq 0$. If $z = 0$ and $x$ is any feasible point for (2.8) then $x + ty, t \to +\infty$ is a sequence of solutions of (2.8) converging to the value of (2.9). We leave it for the reader to check those statements. In either case we showed an equivalence between the two problems.

Note that, as the sketch of proof above suggests, the optimal value in (2.8) may not be attained, even though the one in the linear problem (2.9) always is. For example, consider a pair of problems constructed as above:

$$\begin{array}{lrcl} \text{minimize} & x_1/x_2 & & \\ \text{subject to} & x_2 & > & 0, \\ & x_1 + x_2 & = & 1. \end{array} \qquad \begin{array}{lrcl} \text{minimize} & y_1 & & \\ \text{subject to} & y_1 + y_2 & = & z, \\ & y_2 & = & 1, \\ & z & \geq & 0. \end{array}$$

Both have an optimal value of $-1$, but on the left we can only approach it arbitrarily closely.

### 2.2.6 Sum of largest elements

Suppose $x \in \mathbb{R}^n$ and that $m$ is a positive integer. Consider the problem

$$\begin{array}{lrcll} \text{minimize} & mt + \sum_i u_i & & & \\ \text{subject to} & u_i + t & \geq & x_i, & i = 1, \ldots, n, \\ & u_i & \geq & 0, & i = 1, \ldots, n, \end{array} \tag{2.10}$$

with new variables $t \in \mathbb{R}$, $u_i \in \mathbb{R}^n$. It is easy to see that fixing a value for $t$ determines the rest of the solution. For the sake of simplifying notation let us assume for a moment that $x$ is sorted:

$$x_1 \geq x_2 \geq \cdots \geq x_n.$$

If $t \in [x_k, x_{k+1})$ then $u_l = 0$ for $l \geq k+1$ and $u_l = x_l - t$ for $l \leq k$ in the optimal solution. Therefore, the objective value under the assumption $t \in [x_k, x_{k+1})$ is

$$\text{obj}_t = x_1 + \cdots + x_k + t(m - k)$$

which is a linear function minimized at one of the endpoints of $[x_k, x_{k+1})$. Now we can compute

$$\text{obj}_{x_{k+1}} - \text{obj}_{x_k} = (k - m)(x_k - x_{k+1}).$$

It follows that $\text{obj}_{x_k}$ has a minimum for $k = m$, and therefore the optimum value of (2.10) is simply

$$x_1 + \cdots + x_m.$$

Since the assumption that $x$ is sorted was only a notational convenience, we conclude that in general the optimization model (2.10) computes the *sum of $m$ largest entries in $x$*. In Sec. 2.4 we will show a conceptual way of deriving this model.

## 2.3 Infeasibility in linear optimization

In this section we discuss the basic theory of primal infeasibility certificates for linear problems. These ideas will be developed further after we have introduced duality in the next section.

One of the first issues one faces when presented with an optimization problem is whether it has any solutions at all. As we discussed previously, for a linear optimization problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \\ & x \geq 0. \end{array} \tag{2.11}$$

the *feasible set*

$$\mathcal{F}_p = \{x \in \mathbb{R}^n \mid Ax = b,\ x \geq 0\}$$

is a convex polytope. We say the problem is *feasible* if $\mathcal{F}_p \neq \emptyset$ and *infeasible* otherwise.

**Example 2.3** (Linear infeasible problem)**.** Consider the optimization problem:

$$\begin{array}{lrcrcrcl} \text{minimize} & & 2x_1 & + & 3x_2 & - & x_3 & \\ \text{subject to} & & x_1 & + & x_2 & + & 2x_3 & = 1, \\ & - & 2x_1 & - & x_2 & + & x_3 & = -0.5, \\ & - & x_1 & & & + & 5x_3 & = -0.1, \\ & & & & & & x_i & \geq 0. \end{array}$$

This problem is infeasible. We see it by taking a linear combination of the constraints with coefficients $y = (1, 2, -1)^T$:

$$\begin{array}{rrcrcrcll} & x_1 & + & x_2 & + & 2x_3 & = 1, & / & \cdot 1 \\ - & 2x_1 & - & x_2 & + & x_3 & = -0.5, & / & \cdot 2 \\ - & x_1 & & & + & 5x_3 & = -0.1, & / & \cdot(-1) \\ \hline - & 2x_1 & - & x_2 & - & x_3 & = 0.1. & & \end{array}$$

This clearly proves infeasibility: the left-hand side is negative and the right-hand side is positive, which is impossible.

### 2.3.1 Farkas' lemma

In the last example we proved infeasibility of the linear system by exhibiting an explicit linear combination of the equations, such that the right-hand side (constant) is positive while on the left-hand side all coefficients are negative or zero. In matrix notation, such a linear combination is given by a vector $y$ such that $A^T y \leq 0$ and $b^T y > 0$. The next lemma shows that infeasibility of (2.11) is *equivalent* to the existence of such a vector.

**Lemma 2.1** (Farkas' lemma)**.** *Given $A$ and $b$ as in (2.11), exactly one of the two statements is true:*

1. *There exists $x \geq 0$ such that $Ax = b$.*
2. *There exists $y$ such that $A^T y \leq 0$ and $b^T y > 0$.*

*Proof.* Let $a_1, \ldots, a_n$ be the columns of $A$. The set $\{Ax \mid x \geq 0\}$ is a closed convex cone spanned by $a_1, \ldots, a_n$. If this cone contains $b$ then we have the first alternative. Otherwise the cone can be separated from the point $b$ by a hyperplane passing through 0, i.e. there exists $y$ such that $y^T b > 0$ and $y^T a_i \leq 0$ for all $i$. This is equivalent to the second alternative. Finally, 1. and 2. are mutually exclusive, since otherwise we would have

$$0 < y^T b = y^T A x = (A^T y)^T x \leq 0.$$

□

Farkas' lemma implies that either the problem (2.11) is feasible or there is a *certificate of infeasibility* $y$. In other words, every time we classify model as infeasible, we can certify this fact by providing an appropriate $y$, as in Example 2.3.

### 2.3.2 Locating infeasibility

As we already discussed, the infeasibility certificate $y$ gives coefficients of a linear combination of the constraints which is infeasible "in an obvious way", that is positive on one side and negative on the other. In some cases, $y$ may be very sparse, i.e. it may have very few nonzeros, which means that already a very small subset of the constraints is the root cause of infeasibility. This may be interesting if, for example, we are debugging a large model which we expected to be feasible and infeasibility is caused by an error in the problem formulation. Then we only have to consider the sub-problem formed by constraints with index set $\{j \mid y_j \neq 0\}$.

**Example 2.4** (All constraints involved in infeasibility)**.** As a cautionary note consider the constraints

$$0 \leq x_1 \leq x_2 \leq \cdots \leq x_n \leq -1.$$

Any problem with those constraints is infeasible, but dropping any one of the inequalities creates a feasible subproblem.

## 2.4 Duality in linear optimization

Duality is a rich and powerful theory, central to understanding infeasibility and sensitivity issues in linear optimization. In this section we only discuss duality in linear optimization at a descriptive level suited for practitioners; we refer to Sec. 8 for a more in-depth discussion of duality for general conic problems.

### 2.4.1 The dual problem

#### Primal problem

We consider as always a linear optimization problem in standard form:

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \\ & x \geq 0. \end{array} \tag{2.12}$$

We denote the optimal objective value in (2.12) by $p^\star$. There are three possibilities:

- The problem is infeasible. By convention $p^\star = +\infty$.
- $p^\star$ is finite, in which case the problem has an optimal solution.
- $p^\star = -\infty$, meaning that there are feasible solutions with $c^T x$ decreasing to $-\infty$, in which case we say the problem is *unbounded*.

#### Lagrange function

We associate with (2.12) a so-called *Lagrangian* function $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n_+ \to \mathbb{R}$ that augments the objective with a weighted combination of all the constraints,

$$L(x, y, s) = c^T x + y^T(b - Ax) - s^T x.$$

The variables $y \in \mathbb{R}^m$ and $s \in \mathbb{R}^n_+$ are called *Lagrange multipliers* or *dual variables*. For any feasible $x^* \in \mathcal{F}_p$ and any $(y^*, s^*) \in \mathbb{R}^m \times \mathbb{R}^n_+$ we have

$$L(x^*, y^*, s^*) = c^T x^* + (y^*)^T \cdot 0 - (s^*)^T x^* \leq c^T x^*.$$

Note the we used the nonnegativity of $s^*$, or in general of any Lagrange multiplier associated with an inequality constraint. The *dual function* is defined as the minimum of $L(x, y, s)$ over $x$. Thus the dual function of (2.12) is

$$g(y, s) = \min_x L(x, y, s) = \min_x x^T(c - A^T y - s) + b^T y = \begin{cases} b^T y, & c - A^T y - s = 0, \\ -\infty, & \text{otherwise.} \end{cases}$$

#### Dual problem

For every $(y, s)$ the value of $g(y, s)$ is a lower bound for $p^\star$. To get the best such bound we maximize $g(y, s)$ over all $(y, s)$ and get the *dual problem*:

$$\begin{array}{ll} \text{maximize} & b^T y \\ \text{subject to} & c - A^T y = s, \\ & s \geq 0. \end{array} \tag{2.13}$$

The optimal value of (2.13) will be denoted $d^\star$. As in the case of (2.12) (which from now on we call the *primal problem*), the dual problem can be infeasible ($d^\star = -\infty$), have an optimal solution ($-\infty < d^\star < +\infty$) or be unbounded ($d^\star = +\infty$). Note that the roles of $-\infty$ and $+\infty$ are now reversed because the dual is a maximization problem.

**Example 2.5** (Dual of basis pursuit)**.** As an example, let us derive the dual of the basis pursuit formulation (2.7). It would be possible to add auxiliary variables and constraints to force that problem into the standard form (2.12) and then just apply the dual transformation as a black box, but it is both easier and more instructive to directly write the Lagrangian:

$$L(x, z, y, u, v) = e^T z + u^T(x - z) - v^T(x + z) + y^T(b - Ax)$$

where $e = (1, \dots, 1)^T$, with Lagrange multipliers $y \in \mathbb{R}^m$ and $u, v \in \mathbb{R}^n_+$. The dual function

$$g(y, u, v) = \min_{x,z} L(x, z, y, u, v) = \min_{x,z} z^T(e - u - v) + x^T(u - v - A^T y) + y^T b$$

is only bounded below if $e = u + v$ and $A^T y = u - v$, hence the dual problem is

$$\begin{array}{ll} \text{maximize} & b^T y \\ \text{subject to} & e = u + v, \\ & A^T y = u - v, \\ & u, v \geq 0. \end{array} \tag{2.14}$$

It is not hard to observe that an equivalent formulation of (2.14) is simply

$$\begin{array}{ll} \text{maximize} & b^T y \\ \text{subject to} & \|A^T y\|_\infty \leq 1, \end{array} \tag{2.15}$$

which should be associated with duality between norms discussed in Example 2.2.

**Example 2.6** (Dual of a maximization problem)**.** We can similarly derive the dual of problem (2.13). If we write it simply as

$$\begin{array}{ll} \text{maximize} & b^T y \\ \text{subject to} & c - A^T y \geq 0, \end{array}$$

then the Lagrangian is

$$L(y, u) = b^T y + u^T(c - A^T y) = y^T(b - Au) + c^T u$$

with $u \in \mathbb{R}^n_+$, so that now $L(y, u) \geq b^T y$ for any feasible $y$. Calculating $\min_u \max_y L(y, u)$ is now equivalent to the problem

$$\begin{array}{ll} \text{minimize} & c^T u \\ \text{subject to} & Au = b, \\ & u \geq 0, \end{array}$$

so, as expected, the dual of the dual recovers the original primal problem.

### 2.4.2 Weak and strong duality

Suppose $x^*$ and $(y^*, s^*)$ are feasible points for the primal and dual problems (2.12) and (2.13), respectively. Then we have

$$b^T y^* = (Ax^*)^T y^* = (x^*)^T (A^T y^*) = (x^*)^T (c - s^*) = c^T x^* - (s^*)^T x^* \leq c^T x^*$$

so the dual objective value is a lower bound on the objective value of the primal. In particular, any dual feasible point $(y^*, s^*)$ gives a lower bound:

$$b^T y^* \leq p^\star$$

and we immediately get the next lemma.

**Lemma 2.2** (Weak duality)**.** $d^\star \leq p^\star$.

It follows that if $b^T y^* = c^T x^*$ then $x^*$ is optimal for the primal, $(y^*, s^*)$ is optimal for the dual and $b^T y^* = c^T x^*$ is the common optimal objective value. This way we can use the optimal dual solution to certify optimality of the primal solution and vice versa.

The remarkable property of linear optimization is that $d^\star = p^\star$ holds in the most interesting scenario when the primal problem is feasible and bounded. It means that the certificate of optimality mentioned in the previous paragraph *always exists*.

**Lemma 2.3** (Strong duality)**.** *If at least one of $d^\star, p^\star$ is finite then $d^\star = p^\star$.*

*Proof.* Suppose $-\infty < p^\star < \infty$; the proof in the dual case is analogous. For any $\varepsilon > 0$ consider the feasibility problem with variable $x \geq 0$ and constraints

$$\begin{bmatrix} -c^T \\ A \end{bmatrix} x = \begin{bmatrix} -p^\star + \varepsilon \\ b \end{bmatrix} \quad \text{that is} \quad \begin{array}{rcl} c^T x & = & p^\star - \varepsilon, \\ Ax & = & b. \end{array}$$

Optimality of $p^\star$ implies that the above problem is infeasible. By Lemma 2.1 there exists $\hat{y} = [y_0 \; y]^T$ such that

$$\begin{bmatrix} -c, & A^T \end{bmatrix} \hat{y} \leq 0 \text{ and } \begin{bmatrix} -p^\star + \varepsilon, & b^T \end{bmatrix} \hat{y} > 0.$$

If $y_0 = 0$ then $A^T y \leq 0$ and $b^T y > 0$, which by Lemma 2.1 again would mean that the original primal problem was infeasible, which is not the case. Hence we can rescale so that $y_0 = 1$ and then we get

$$c - A^T y \geq 0 \quad \text{and} \quad b^T y \geq p^\star - \varepsilon.$$

The first inequality above implies that $y$ is feasible for the dual problem. By letting $\varepsilon \to 0$ we obtain $d^\star \geq p^\star$. □

We can exploit strong duality to freely choose between solving the primal or dual version of any linear problem.

**Example 2.7** (Sum of largest elements)**.** Suppose that $x$ is now a constant vector. Consider the following problem with variable $z$:

$$\begin{array}{ll} \text{maximize} & x^T z \\ \text{subject to} & \sum_i z_i = m, \\ & 0 \leq z \leq 1. \end{array}$$

The maximum is attained when $z$ indicates the positions of $m$ largest entries in $x$, and the objective value is then their sum. This formulation, however, cannot be used when $x$ is another variable, since then the objective function is no longer linear. Let us derive the dual problem. The Lagrangian is

$$\begin{aligned} L(z, s, t, u) &= x^T z + t(m - e^T z) + s^T z + u^T(e - z) = \\ &= z^T(x - te + s - u) + tm + u^T e \end{aligned}$$

with $u, s \geq 0$. Since $s_i \geq 0$ is arbitrary and not otherwise constrained, the equality $x_i - t + s_i - u_i = 0$ is the same as $u_i + t \geq x_i$ and for the dual problem we get

$$\begin{array}{lrcll} \text{minimize} & mt + \sum_i u_i & & & \\ \text{subject to} & u_i + t & \geq & x_i, & i = 1, \ldots, n, \\ & u_i & \geq & 0, & i = 1, \ldots, n, \end{array}$$

which is exactly the problem (2.10) we studied in Sec. 2.2.6. Strong duality now implies that (2.10) computes the sum of $m$ biggest entries in $x$.

### 2.4.3 Duality and infeasibility: summary

We can now expand the discussion of infeasibility certificates in the context of duality. Farkas' lemma Lemma 2.1 can be dualized and the two versions can be summarized as follows:

**Lemma 2.4** (Primal and dual Farkas' lemma)**.** *For a primal-dual pair of linear problems we have the following equivalences:*

1. *The primal problem* (2.12) *is infeasible if and only if there is $y$ such that $A^T y \leq 0$ and $b^T y > 0$.*
2. *The dual problem* (2.13) *is infeasible if and only if there is $x \geq 0$ such that $Ax = 0$ and $c^T x < 0$.*

Weak and strong duality for linear optimization now lead to the following conclusions:

- If the problem is primal feasible and has finite objective value ($-\infty < p^\star < \infty$) then so is the dual and $d^\star = p^\star$. We sometimes refer to this case as *primal and dual feasible*. The dual solution certifies the optimality of the primal solution and vice versa.
- If the primal problem is feasible but unbounded ($p^\star = -\infty$) then the dual is infeasible ($d^\star = -\infty$). Part (ii) of Farkas' lemma provides a certificate of this fact, that is a

vector $x$ with $x \geq 0$, $Ax = 0$ and $c^T x < 0$. In fact it is easy to give this statement a geometric interpretation. If $x^0$ is any primal feasible point then the infinite ray

$$t \to x_0 + tx, \ t \in [0, \infty)$$

belongs to the feasible set $\mathcal{F}_p$ because $A(x_0 + tx) = b$ and $x_0 + tx \geq 0$. Along this ray the objective value is unbounded below:

$$c^T(x_0 + tx) = c^T x_0 + t(c^T x) \to -\infty.$$

- If the primal problem is infeasible ($p^\star = \infty$) then a certificate of this fact is provided by part (i). The dual problem may be unbounded ($d^\star = \infty$) or infeasible ($d^\star = -\infty$).

**Example 2.8** (Primal-dual infeasibility)**.** Weak and strong duality imply that the only case when $d^\star \neq p^\star$ is when both primal and dual problem are infeasible ($d^\star = -\infty$, $p^\star = \infty$), for example:

$$\begin{array}{ll} \text{minimize} & x \\ \text{subject to} & 0 \cdot x = 1. \end{array}$$

### 2.4.4 Dual values as shadow prices

Dual values are related to *shadow prices*, as they measure, under some nondegeneracy assumption, the sensitivity of the objective value to a change in the constraint. Consider again the primal and dual problem pair (2.12) and (2.13) with feasible sets $\mathcal{F}_p$ and $\mathcal{F}_d$ and with a primal-dual optimal solution $(x^*, y^*, s^*)$.

Suppose we change one of the values in $b$ from $b_i$ to $b_i'$. This corresponds to moving one of the hyperplanes defining $\mathcal{F}_p$, and in consequence the optimal solution (and the objective value) may change. On the other hand, the dual feasible set $\mathcal{F}_d$ is not affected. Assuming that the solution $(y^*, s^*)$ was a unique vertex of $\mathcal{F}_d$ this point remains optimal for the dual after a sufficiently small change of $b$. But then the change of the dual objective is

$$y_i^*(b_i' - b_i)$$

and by strong duality the primal objective changes by the same amount.

**Example 2.9** (Student diet)**.** An optimization student wants to save money on the diet while remaining healthy. A healthy diet requires at least $P = 6$ units of protein, $C = 15$ units of carbohydrates, $F = 5$ units of fats and $V = 7$ units of vitamins. The student can choose from the following products:

| | P | C | F | V | price |
|---|---|---|---|---|---|
| takeaway | 3 | 3 | 2 | 1 | 5 |
| vegetables | 1 | 2 | 0 | 4 | 1 |
| bread | 0.5 | 4 | 1 | 0 | 2 |

The problem of minimizing cost while meeting dietary requirements is

$$\begin{array}{ll} \text{minimize} & 5x_1 + x_2 + 2x_3 \\ \text{subject to} & 3x_1 + x_2 + 0.5x_3 \geq 6, \\ & 3x_1 + 2x_2 + 4x_3 \geq 15, \\ & 2x_1 + x_3 \geq 5, \\ & x_1 + 4x_2 \geq 7, \\ & x_1, x_2, x_3 \geq 0. \end{array}$$

If $y_1, y_2, y_3, y_4$ are the dual variables associated with the four inequality constraints then the (unique) primal-dual optimal solution to this problem is approximately:

$$(x, y) = ((1, 1.5, 3), (0.42, 0, 1.78, 0.14))$$

with optimal cost $p^\star = 12.5$. Note $y_2 = 0$ indicates that the second constraint is not binding. Indeed, we could increase $C$ to 18 without affecting the optimal solution. The remaining constraints are binding.

Improving the intake of protein by 1 unit (increasing $P$ to 7) will increase the cost by 0.42, while doing the same for fat will cost an extra 1.78 per unit. If the student had extra money to improve one of the parameters then the best choice would be to increase the intake of vitamins, with shadow price of just 0.14.

If one month the student only had 12 units of money and was willing to relax one of the requirements then the best choice is to save on fats: the necessary reduction of $F$ is smallest, namely $0.5 \cdot 1.78^{-1} = 0.28$. Indeed, with the new value of $F = 4.72$ the same problem solves to $p^\star = 12$ and $x = (1.08, 1.48, 2.56)$.

We stress that a truly balanced diet problem should also include upper bounds.

# Chapter 3

# Conic quadratic optimization

This chapter extends the notion of linear optimization with *quadratic cones*. Conic quadratic optimization, also known as second-order cone optimization, is a straightforward generalization of linear optimization, in the sense that we optimize a linear function under linear (in)equalities with some variables belonging to one or more (rotated) quadratic cones. We discuss the basic concept of quadratic cones, and demonstrate the surprisingly large flexibility of conic quadratic modeling.

## 3.1 Cones

Since this is the first place where we introduce a non-linear cone, it seems suitable to make our most important definition:

A set $K \subseteq \mathbb{R}^n$ is called a *convex cone* if

- for every $x, y \in K$ we have $x + y \in K$,
- for every $x \in K$ and $\alpha \geq 0$ we have $\alpha x \in K$.

For example a linear subspace of $\mathbb{R}^n$, the positive orthant $\mathbb{R}^n_{\geq 0}$ or any ray (half-line) starting at the origin are examples of convex cones. We leave it for the reader to check that the intersection of convex cones is a convex cone; this property enables us to assemble complicated optimization models from individual conic bricks.

### 3.1.1 Quadratic cones

We define the $n$-dimensional quadratic cone as

$$\mathcal{Q}^n = \left\{ x \in \mathbb{R}^n \mid x_1 \geq \sqrt{x_2^2 + x_3^2 + \cdots + x_n^2} \right\}. \tag{3.1}$$

The geometric interpretation of a quadratic (or second-order) cone is shown in Fig. 3.1 for a cone with three variables, and illustrates how the boundary of the cone resembles an ice-cream cone. The 1-dimensional quadratic cone simply states nonnegativity $x_1 \geq 0$.

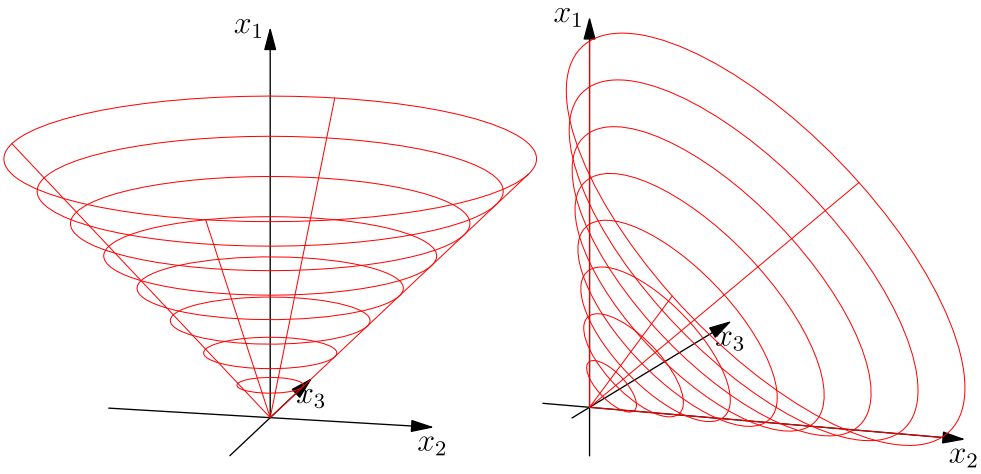



Fig. 3.1: Boundary of quadratic cone $x_1 \geq \sqrt{x_2^2 + x_3^2}$ and rotated quadratic cone $2x_1x_2 \geq x_3^2$, $x_1, x_2 \geq 0$.

### 3.1.2 Rotated quadratic cones

An $n-$dimensional *rotated quadratic cone* is defined as

$$\mathcal{Q}_r^n = \left\{ x \in \mathbb{R}^n \mid 2x_1x_2 \geq x_3^2 + \cdots + x_n^2,\ x_1, x_2 \geq 0 \right\}. \tag{3.2}$$

As the name indicates, there is a simple relationship between quadratic and rotated quadratic cones. Define an orthogonal transformation

$$T_n := \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 0 & 0 & I_{n-2} \end{bmatrix}. \tag{3.3}$$

Then it is easy to verify that

$$x \in \mathcal{Q}^n \quad \Longleftrightarrow \quad T_n x \in \mathcal{Q}_r^n,$$

and since $T$ is orthogonal we call $\mathcal{Q}_r^n$ a rotated cone; the transformation corresponds to a rotation of $\pi/4$ in the $(x_1, x_2)$ plane. For example if $x \in \mathcal{Q}^3$ and

$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}}(x_1 + x_2) \\ \frac{1}{\sqrt{2}}(x_1 - x_2) \\ x_3 \end{bmatrix}$$

then

$$2z_1z_2 \geq z_3^2,\ z_1, z_2 \geq 0 \quad \Longrightarrow \quad (x_1^2 - x_2^2) \geq x_3^2,\ x_1 \geq 0,$$

and similarly we see that

$$x_1^2 \geq x_2^2 + x_3^2,\ x_1 \geq 0 \quad \Longrightarrow \quad 2z_1z_2 \geq z_3^2,\ z_1, z_2 \geq 0.$$

Thus, one could argue that we only need quadratic cones $\mathcal{Q}^n$, but there are many examples where using an explicit rotated quadratic cone $\mathcal{Q}_r^n$ is more natural, as we will see next.

## 3.2 Conic quadratic modeling

In the following we describe several convex sets that can be modeled using conic quadratic formulations or, as we call them, are *conic quadratic representable*.

### 3.2.1 Absolute values

In Sec. 2.2.2 we saw how to model $|x| \leq t$ using two linear inequalities, but in fact the epigraph of the absolute value is just the definition of a two-dimensional quadratic cone, i.e.,

$$|x| \leq t \quad \Longleftrightarrow \quad (t, x) \in \mathcal{Q}^2.$$

### 3.2.2 Euclidean norms

The Euclidean norm of $x \in \mathbb{R}^n$,

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

essentially defines the quadratic cone, i.e.,

$$\|x\|_2 \leq t \quad \Longleftrightarrow \quad (t, x) \in \mathcal{Q}^{n+1}.$$

The epigraph of the squared Euclidean norm can be described as the intersection of a rotated quadratic cone with an affine hyperplane,

$$x_1^2 + \cdots + x_n^2 = \|x\|_2^2 \leq t \quad \Longleftrightarrow \quad (1/2, t, x) \in \mathcal{Q}_r^{n+2}.$$

### 3.2.3 Convex quadratic sets

Assume $Q \in \mathbb{R}^{n \times n}$ is a symmetric positive semidefinite matrix. The convex inequality

$$(1/2)x^T Q x + c^T x + r \leq 0$$

may be rewritten as

$$\begin{array}{rcl} t + c^T x + r & = & 0, \\ x^T Q x & \leq & 2t. \end{array} \tag{3.4}$$

Since $Q$ is symmetric positive semidefinite the epigraph

$$x^T Q x \leq 2t \tag{3.5}$$

is a convex set and there exists a matrix $F \in \mathbb{R}^{k\times n}$ such that

$$Q = F^T F \tag{3.6}$$

(see Sec. 6 for properties of semidefinite matrices). For instance $F$ could be the Cholesky factorization of $Q$. Then

$$x^T Q x = x^T F^T F x = \|Fx\|_2^2$$

and we have an equivalent characterization of (3.5) as

$$(1/2)x^T Q x \leq t \quad \Longleftrightarrow \quad (t, 1, Fx) \in \mathcal{Q}_r^{2+k}.$$

Frequently $Q$ has the structure

$$Q = I + F^T F$$

where $I$ is the identity matrix, so

$$x^T Q x = x^T x + x^T F^T F x = \|x\|_2^2 + \|Fx\|_2^2$$

and hence

$$(f, 1, x) \in \mathcal{Q}_r^{2+n}, \qquad (h, 1, Fx) \in \mathcal{Q}_r^{2+k}, \qquad f + h = t$$

is a conic quadratic representation of (3.5) in this case.

### 3.2.4 Second-order cones

A second-order cone is occasionally specified as

$$\|Ax + b\|_2 \leq c^T x + d \tag{3.7}$$

where $A \in \mathbb{R}^{m\times n}$ and $c \in \mathbb{R}^n$. The formulation (3.7) is simply

$$(c^T x + d, Ax + b) \in \mathcal{Q}^{m+1} \tag{3.8}$$

or equivalently

$$\begin{array}{rcl} s & = & Ax + b, \\ t & = & c^T x + d, \\ (t, s) & \in & \mathcal{Q}^{m+1}. \end{array} \tag{3.9}$$

As will be explained in Sec. 8, we refer to (3.8) as the dual form and (3.9) as the primal form. An alternative characterization of (3.7) is

$$\|Ax + b\|_2^2 - (c^T x + d)^2 \leq 0, \qquad c^T x + d \geq 0 \tag{3.10}$$

which shows that certain quadratic inequalities are conic quadratic representable.

### 3.2.5 Simple sets involving power functions

Some power-like inequalities are conic quadratic representable, even though it need not be obvious at first glance. For example, we have

$$|t| \leq \sqrt{x},\ x \geq 0 \iff (x, 1/2, t) \in \mathcal{Q}_r^3,$$

or in a similar fashion

$$t \geq \frac{1}{x},\ x \geq 0 \iff (x, t, \sqrt{2}) \in \mathcal{Q}_r^3.$$

For a more complicated example, consider the constraint

$$t \geq x^{3/2}, x \geq 0.$$

This is equivalent to a statement involving two cones and an extra variable

$$(s, t, x), (x, 1/8, s) \in \mathcal{Q}_r^3$$

because

$$2st \geq x^2,\ 2 \cdot \frac{1}{8}x \geq s^2,\ \implies 4s^2t^2 \cdot \frac{1}{4}x \geq x^4 \cdot s^2 \implies t \geq x^{3/2}.$$

In practice power-like inequalities representable with similar tricks can often be expressed much more naturally using the power cone (see Sec. 4), so we will not dwell on these examples much longer.

### 3.2.6 Harmonic mean

Consider next the hypograph of the harmonic mean,

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i^{-1}\right)^{-1} \geq t \geq 0, \quad x \geq 0.$$

It is not obvious either that the inequality defines a convex set, or whether it is conic quadratic representable. However, we can write it equivalently in the form

$$\sum_{i=1}^{n} \frac{t^2}{x_i} \leq nt,$$

which suggests the conic representation:

$$2x_i z_i \geq t^2, \quad x_i, z_i \geq 0, \quad 2\sum_{i=1}^{n} z_i = nt. \tag{3.11}$$

### 3.2.7 Quadratic forms with one negative eigenvalue

Assume that $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix with exactly one negative eigenvalue, i.e., $A$ has a spectral factorization (i.e., eigenvalue decomposition)

$$A = Q\Lambda Q^T = -\alpha_1 q_1 q_1^T + \sum_{i=2}^{n} \alpha_i q_i q_i^T,$$

where $Q^TQ = I$, $\Lambda = \mathbf{Diag}(-\alpha_1, \alpha_2, \ldots, \alpha_n)$, $\alpha_i \geq 0$. Then

$$x^T Ax \leq 0$$

is equivalent to

$$\sum_{j=2}^{n} \alpha_j (q_j^T x)^2 \leq \alpha_1 (q_1^T x)^2. \tag{3.12}$$

Suppose $q_1^T x \geq 0$. We can characterize (3.12) as

$$(\sqrt{\alpha_1} q_1^T x, \sqrt{\alpha_2} q_2^T x, \ldots, \sqrt{\alpha_n} q_n^T x) \in \mathcal{Q}^n. \tag{3.13}$$

### 3.2.8 Ellipsoidal sets

The set

$$\mathcal{E} = \{x \in \mathbb{R}^n \mid \|P(x-c)\|_2 \leq 1\}$$

describes an ellipsoid centred at $c$. It has a natural conic quadratic representation, i.e., $x \in \mathcal{E}$ if and only if

$$x \in \mathcal{E} \iff (1, P(x-c)) \in \mathcal{Q}^{n+1}.$$

## 3.3 Conic quadratic case studies

### 3.3.1 Quadratically constrained quadratic optimization

A general convex quadratically constrained quadratic optimization problem can be written as

$$\begin{array}{ll} \text{minimize} & (1/2)x^T Q_0 x + c_0^T x + r_0 \\ \text{subject to} & (1/2)x^T Q_i x + c_i^T x + r_i \leq 0, \quad i = 1, \ldots, p, \end{array} \tag{3.14}$$

where all $Q_i \in \mathbb{R}^{n\times n}$ are symmetric positive semidefinite. Let

$$Q_i = F_i^T F_i, \quad i = 0, \ldots, p,$$

where $F_i \in \mathbb{R}^{k_i \times n}$. Using the formulations in Sec. 3.2.3 we then get an equivalent conic quadratic problem

$$\begin{array}{lll} \text{minimize} & t_0 + c_0^T x + r_0 & \\ \text{subject to} & t_i + c_i^T x + r_i = 0, & i = 1, \ldots, p, \\ & (t_i, 1, F_i x) \in \mathcal{Q}_r^{k_i+2}, & i = 0, \ldots, p. \end{array} \tag{3.15}$$

Assume next that $k_i$, the number of rows in $F_i$, is small compared to $n$. Storing $Q_i$ requires about $n^2/2$ space whereas storing $F_i$ then only requires $nk_i$ space. Moreover, the amount of work required to evaluate $x^T Q_i x$ is proportional to $n^2$ whereas the work required to evaluate $x^T F_i^T F_i x = \|F_i x\|^2$ is proportional to $nk_i$ only. In other words, if $Q_i$ have low rank, then (3.15) will require much less space and time to solve than (3.14). We will study the reformulation (3.15) in much more detail in Sec. 10.

### 3.3.2 Robust optimization with ellipsoidal uncertainties

Often in robust optimization some of the parameters in the model are assumed to be unknown exactly, but there is a simple set describing the uncertainty. For example, for a standard linear optimization problem we may wish to find a robust solution for all objective vectors $c$ in an ellipsoid

$$\mathcal{E} = \{c \in \mathbb{R}^n \mid c = Fy + g,\ \|y\|_2 \leq 1\}.$$

A common approach is then to optimize for the *worst-case* scenario for $c$, so we get a robust version

$$\begin{array}{ll} \text{minimize} & \sup_{c \in \mathcal{E}} c^T x \\ \text{subject to} & Ax = b, \\ & x \geq 0. \end{array} \tag{3.16}$$

The worst-case objective can be evaluated as

$$\sup_{c \in \mathcal{E}} c^T x = g^T x + \sup_{\|y\|_2 \leq 1} y^T F^T x = g^T x + \|F^T x\|_2$$

where we used that $\sup_{\|u\|_2 \leq 1} v^T u = (v^T v)/\|v\|_2 = \|v\|_2$. Thus the robust problem (3.16) is equivalent to

$$\begin{array}{ll} \text{minimize} & g^T x + \|F^T x\|_2 \\ \text{subject to} & Ax = b, \\ & x \geq 0, \end{array}$$

which can be posed as a conic quadratic problem

$$\begin{array}{ll} \text{minimize} & g^T x + t \\ \text{subject to} & Ax = b, \\ & (t, F^T x) \in \mathcal{Q}^{n+1}, \\ & x \geq 0. \end{array} \tag{3.17}$$

### 3.3.3 Markowitz portfolio optimization

In classical Markowitz portfolio optimization we consider investment in $n$ stocks or assets held over a period of time. Let $x_i$ denote the amount we invest in asset $i$, and assume a stochastic model where the return of the assets is a random variable $r$ with known mean

$$\mu = \mathbf{E}r$$

and covariance

$$\Sigma = \mathbf{E}(r-\mu)(r-\mu)^T.$$

The return of our investment is also a random variable $y = r^T x$ with mean (or expected return)

$$\mathbf{E}y = \mu^T x$$

and variance (or risk)

$$(y - \mathbf{E}y)^2 = x^T \Sigma x.$$

We then wish to rebalance our portfolio to achieve a compromise between risk and expected return, e.g., we can maximize the expected return given an upper bound $\gamma$ on the tolerable risk and a constraint that our total investment is fixed,

$$\begin{array}{ll} \text{maximize} & \mu^T x \\ \text{subject to} & x^T \Sigma x \leq \gamma \\ & e^T x = 1 \\ & x \geq 0. \end{array} \tag{3.18}$$

Suppose we factor $\Sigma = GG^T$ (e.g., using a Cholesky or a eigenvalue decomposition). We then get a conic formulation

$$\begin{array}{ll} \text{maximize} & \mu^T x \\ \text{subject to} & (\sqrt{\gamma}, G^T x) \in \mathcal{Q}^{n+1} \\ & e^T x = 1 \\ & x \geq 0. \end{array} \tag{3.19}$$

In practice both the average return and covariance are estimated using historical data. A recent trend is then to formulate a robust version of the portfolio optimization problem to combat the inherent uncertainty in those estimates, e.g., we can constrain $\mu$ to an ellipsoidal uncertainty set as in Sec. 3.3.2.

It is also common that the data for a portfolio optimization problem is already given in the form of a factor model $\Sigma = F^T F$ of $\Sigma = I + F^T F$ and a conic quadratic formulation as in Sec. 3.3.1 is most natural. For more details see Sec. 10.3.

### 3.3.4 Maximizing the Sharpe ratio

Continuing the previous example, the Sharpe ratio defines an efficiency metric of a portfolio as the expected return per unit risk, i.e.,

$$S(x) = \frac{\mu^T x - r_f}{(x^T \Sigma x)^{1/2}},$$

where $r_f$ denotes the return of a *risk-free* asset. We assume that there is a portfolio with $\mu^T x > r_f$, so maximizing the Sharpe ratio is equivalent to minimizing $1/S(x)$. In other words, we have the following problem

$$\begin{array}{lccl} \text{minimize} & \frac{\|G^T x\|}{\mu^T x - r_f} & & \\ \text{subject to} & e^T x & = & 1, \\ & x & \geq & 0. \end{array}$$

The objective has the same nature as a quotient of two affine functions we studied in Sec. 2.2.5. We reformulate the problem in a similar way, introducing a scalar variable $z \geq 0$ and a variable transformation

$$y = zx.$$

Since a positive $z$ can be chosen arbitrarily and $(\mu - r_f e)^T x > 0$, we can without loss of generality assume that

$$(\mu - r_f e)^T y = 1.$$

Thus, we obtain the following conic problem for maximizing the Sharpe ratio,

$$\begin{array}{lccl} \text{minimize} & t & & \\ \text{subject to} & (t, G^T y) & \in & \mathcal{Q}^{k+1}, \\ & e^T y & = & z, \\ & (\mu - r_f e)^T y & = & 1, \\ & y, z & \geq & 0, \end{array}$$

and we recover $x = y/z$.

### 3.3.5 A resource constrained production and inventory problem

The resource constrained production and inventory problem *[Zie82]* can be formulated as follows:

$$\begin{array}{lll} \text{minimize} & \sum_{j=1}^n (d_j x_j + e_j / x_j) & \\ \text{subject to} & \sum_{j=1}^n r_j x_j \leq b, & \\ & x_j \geq 0, & j = 1, \ldots, n, \end{array} \tag{3.20}$$

where $n$ denotes the number of items to be produced, $b$ denotes the amount of common resource, and $r_j$ is the consumption of the limited resource to produce one unit of item $j$.

The objective function represents inventory and ordering costs. Let $c_j^p$ denote the holding cost per unit of product $j$ and $c_j^r$ denote the rate of holding costs, respectively. Further, let

$$d_j = \frac{c_j^p c_j^r}{2}$$

so that

$$d_j x_j$$

is the average holding costs for product $j$. If $D_j$ denotes the total demand for product $j$ and $c_j^o$ the ordering cost per order of product $j$ then let

$$e_j = c_j^o D_j$$

and hence

$$\frac{e_j}{x_j} = \frac{c_j^o D_j}{x_j}$$

is the average ordering costs for product $j$. In summary, the problem finds the optimal batch size such that the inventory and ordering cost are minimized while satisfying the constraints on the common resource. Given $d_j, e_j \geq 0$ problem (3.20) is equivalent to the conic quadratic problem

$$\begin{array}{ll} \text{minimize} & \sum_{j=1}^n (d_j x_j + e_j t_j) \\ \text{subject to} & \sum_{j=1}^n r_j x_j \leq b, \\ & (t_j, x_j, \sqrt{2}) \in \mathcal{Q}_r^3, \quad j = 1, \ldots, n. \end{array}$$

It is not always possible to produce a fractional number of items. In such case $x_j$ should be constrained to be integers. See Sec. 9.

# Chapter 4

# The power cone

So far we studied quadratic cones and their applications in modeling problems involving, directly or indirectly, quadratic terms. In this part we expand the quadratic and rotated quadratic cone family with *power cones*, which provide a convenient language to express models involving powers other than 2. We must stress that although the power cones include the quadratic cones as special cases, at the current state-of-the-art they require more advanced and less efficient algorithms.

## 4.1 The power cone(s)

$n$-dimensional power cones form a family of convex cones parametrized by a real number $0 < \alpha < 1$:

$$\mathcal{P}_n^{\alpha,1-\alpha} = \left\{ x \in \mathbb{R}^n \ : \ x_1^\alpha x_2^{1-\alpha} \geq \sqrt{x_3^2 + \cdots + x_n^2}, \ x_1, x_2 \geq 0 \right\}. \tag{4.1}$$

The constraint in the definition of $\mathcal{P}_n^{\alpha,1-\alpha}$ can be expressed as a composition of two constraints, one of which is a quadratic cone:

$$\begin{array}{rcl} x_1^\alpha x_2^{1-\alpha} & \geq & |z|, \\ z & \geq & \sqrt{x_3^2 + \cdots + x_n^2}, \end{array} \tag{4.2}$$

which means that the basic building block we need to consider is the three-dimensional power cone

$$\mathcal{P}_3^{\alpha,1-\alpha} = \left\{ x \in \mathbb{R}^3 \ : \ x_1^\alpha x_2^{1-\alpha} \geq |x_3|, \ x_1, x_2 \geq 0 \right\}. \tag{4.3}$$

More generally, we can also consider power cones with "long left-hand side". That is, for $m < n$ and a sequence of exponents $\alpha_1, \ldots, \alpha_m$ with $\alpha_1 + \cdots + \alpha_m = 1$, we have the most general power cone object defined as

$$\mathcal{P}_n^{\alpha_1,\cdots,\alpha_m} = \left\{ x \in \mathbb{R}^n \ : \ \prod_{i=1}^m x_i^{\alpha_i} \geq \sqrt{\sum_{i=m+1}^n x_i^2}, \ x_1, \ldots, x_m \geq 0 \right\}. \tag{4.4}$$

The left-hand side is nothing but the weighted geometric mean of the $x_i$, $i = 1, \ldots, m$ with weights $\alpha_i$. As we will see later, also this most general cone can be modeled as a composition of three-dimensional cones $\mathcal{P}_3^{\alpha,1-\alpha}$, so in a sense that is the basic object of interest.

There are some notable special cases we are familiar with. If we let $\alpha \to 0$ then in the limit we get $\mathcal{P}_n^{0,1} = \mathbb{R}_+ \times \mathcal{Q}^{n-1}$. If $\alpha = \frac{1}{2}$ then we have a rescaled version of the rotated quadratic cone, precisely:

$$(x_1, x_2, \ldots, x_n) \in \mathcal{P}_n^{\frac{1}{2},\frac{1}{2}} \iff (x_1/\sqrt{2}, x_2/\sqrt{2}, x_3, \ldots, x_n) \in \mathcal{Q}_r^n.$$

A gallery of three-dimensional power cones for varying $\alpha$ is shown in Fig. 4.1.

## 4.2 Sets representable using the power cone

In this section we give basic examples of constraints which can be expressed using power cones.

### 4.2.1 Powers

For all values of $p \neq 0, 1$ we can bound $x^p$ depending on the convexity of $f(x) = x^p$.

- For $p > 1$ the inequality $t \geq |x|^p$ is equivalent to $t^{1/p} \geq |x|$ and hence corresponds to

  $$t \geq |x|^p \iff (t, 1, x) \in \mathcal{P}_3^{1/p,1-1/p}.$$

  For instance $t \geq |x|^{1.5}$ is equivalent to $(t, 1, x) \in \mathcal{P}_3^{2/3,1/3}$.

- For $0 < p < 1$ the function $f(x) = x^p$ is concave for $x \geq 0$ and so we get

  $$|t| \leq x^p,\ x \geq 0 \iff (x, 1, t) \in \mathcal{P}_3^{p,1-p}.$$

- For $p < 0$ the function $f(x) = x^p$ is convex for $x > 0$ and in this range the inequality $t \geq x^p$ is equivalent to

  $$t \geq x^p \iff t^{1/(1-p)} x^{-p/(1-p)} \geq 1 \iff (t, x, 1) \in \mathcal{P}_3^{1/(1-p),-p/(1-p)}.$$

  For example $t \geq \frac{1}{\sqrt{x}}$ is the same as $(t, x, 1) \in \mathcal{P}_3^{2/3,1/3}$.

### 4.2.2 $p$-norm cones

Let $p \geq 1$. The $p$-norm of a vector $x \in \mathbb{R}^n$ is $\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$ and the $p$-norm ball of radius $t$ is defined by the inequality $\|x\|_p \leq t$. We take the $p$-norm cone (in dimension $n+1$) to be the convex set

$$\left\{(t, x) \in \mathbb{R}^{n+1} \ :\ t \geq \|x\|_p\right\}. \tag{4.5}$$

For $p = 2$ this is precisely the quadratic cone. We can model the $p$-norm cone by writing the inequality $t \geq \|x\|_p$ as:

$$t \geq \sum_i |x_i|^p / t^{p-1}$$

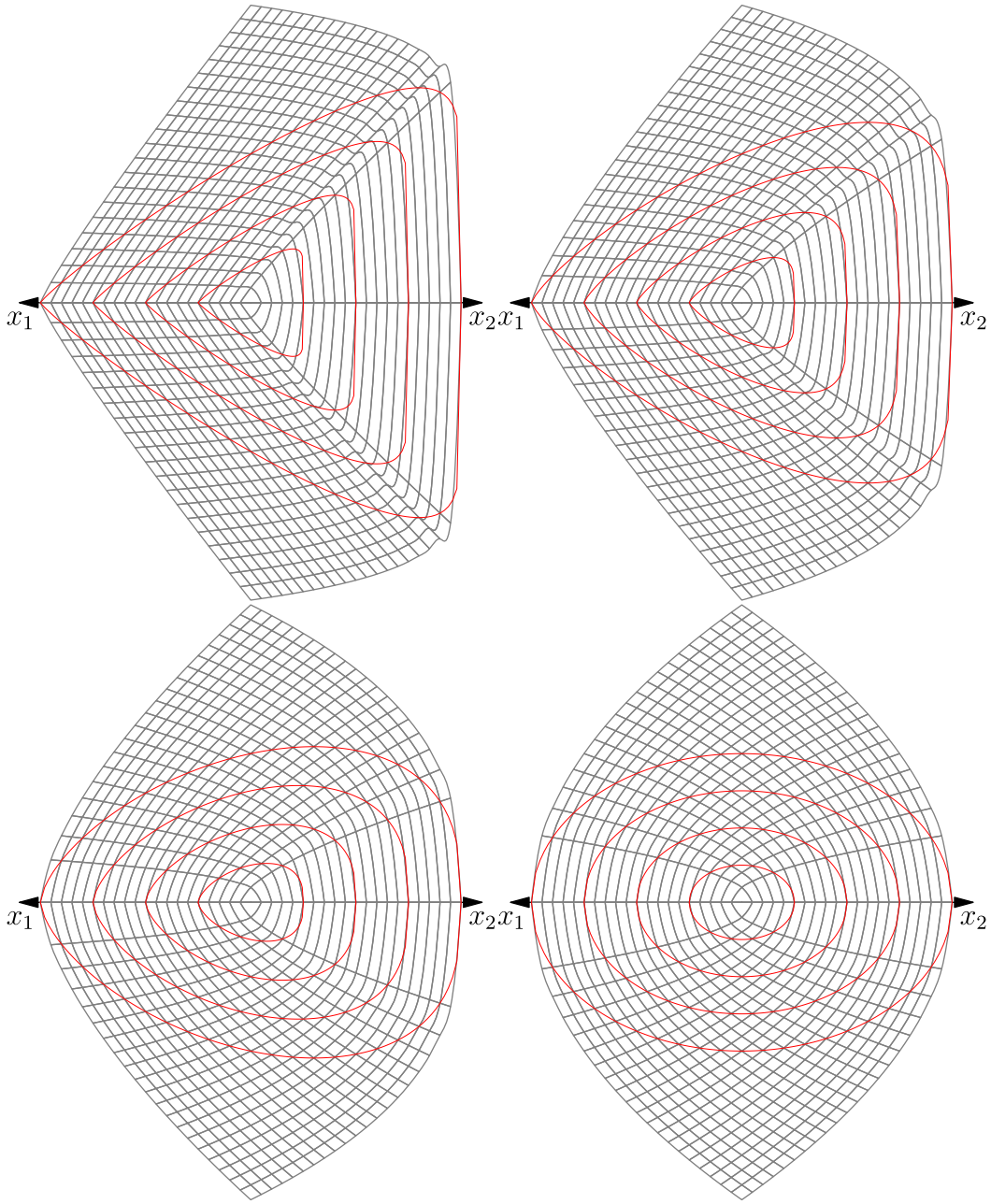



Fig. 4.1: The boundary of $\mathcal{P}_3^{\alpha,1-\alpha}$ seen from a point inside the cone for $\alpha = 0.1, 0.2, 0.35, 0.5$.

and bounding each summand with a power cone. This leads to the following model:

$$\begin{array}{rcl} r_i t^{p-1} & \geq & |x_i|^p \quad ((r_i, t, x_i) \in \mathcal{P}_3^{1/p,1-1/p}), \\ \sum r_i & = & t. \end{array} \tag{4.6}$$

When $0 < p < 1$ or $p < 0$ the formula for $\|x\|_p$ gives a concave, rather than convex function on $\mathbb{R}^n_+$ and in this case it is possible to model the set

$$\left\{ (t, x) \ : \ 0 \leq t \leq \left( \sum x_i^p \right)^{1/p}, x_i \geq 0 \right\}, \ p < 1, p \neq 0.$$

We leave it as an exercise (see previous subsection). The case $p = -1$ appears in Sec. 3.2.6.

### 4.2.3 The most general power cone

Consider the most general version of the power cone with "long left-hand side" defined in (4.4). We show that it can be expressed using the basic three-dimensional cones $\mathcal{P}_3^{\alpha,1-\alpha}$. Clearly it suffices to consider a short right-hand side, that is to model the cone

$$\mathcal{P}_{m+1}^{\alpha_1,\ldots,\alpha_m} = \left\{ x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_m^{\alpha_m} \geq |z|, \ x_1, \ldots, x_m \geq 0 \right\}, \tag{4.7}$$

where $\sum_i \alpha_i = 1$. Denote $s = \alpha_1 + \cdots + \alpha_{m-1}$. We split (4.7) into two constraints

$$\begin{array}{rcll} x_1^{\alpha_1/s} \cdots x_{m-1}^{\alpha_{m-1}/s} & \geq & |t|, & x_1, \ldots, x_{m-1} \geq 0, \\ t^s x_m^{\alpha_m} & \geq & |z|, & x_m \geq 0, \end{array} \tag{4.8}$$

and this way we expressed $\mathcal{P}_{m+1}^{\alpha_1,\ldots,\alpha_m}$ using two power cones $\mathcal{P}_m^{\alpha_1/s,\ldots,\alpha_{m-1}/s}$ and $\mathcal{P}_3^{s,\alpha_m}$. Proceeding by induction gives the desired splitting.

### 4.2.4 Geometric mean

The power cone $\mathcal{P}_{n+1}^{1/n,\ldots,1/n}$ is a direct way to model the inequality

$$|z| \leq (x_1 x_2 \cdots x_n)^{1/n}, \tag{4.9}$$

which corresponds to maximizing the geometric mean of the variables $x_i \geq 0$. In this special case tracing through the splitting (4.8) produces an equivalent representation of (4.9) using three-dimensional power cones as follows:

$$\begin{array}{rcl} x_1^{1/2} x_2^{1/2} & \geq & |t_3|, \\ t_3^{1-1/3} x_3^{1/3} & \geq & |t_4|, \\ & \cdots & \\ t_{n-1}^{1-1/(n-1)} x_{n-1}^{1/(n-1)} & \geq & |t_n|, \\ t_n^{1-1/n} x_n^{1/n} & \geq & |z|. \end{array}$$

### 4.2.5 Non-homogenous constraints

Every constraint of the form

$$x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_m^{\alpha_m} \geq |z|^\beta, \; x_i \geq 0,$$

where $\sum_i \alpha_i < \beta$ and $\alpha_i > 0$ is equivalent to

$$(x_1, x_2, \ldots, x_m, 1, z) \in \mathcal{P}_{m+2}^{\alpha_1/\beta, \ldots, \alpha_m/\beta, s}$$

with $s = 1 - \sum_i \alpha_i / \beta$.

## 4.3 Power cone case studies

### 4.3.1 Portfolio optimization with market impact

Let us go back to the Markowitz portfolio optimization problem introduced in Sec. 3.3.3, where we now ask to maximize expected profit subject to bounded risk in a long-only portfolio:

$$\begin{array}{lrcl} \text{maximize} & \mu^T x & & \\ \text{subject to} & \sqrt{x^T \Sigma x} & \leq & \gamma, \\ & x_i & \geq & 0, \; i = 1, \ldots, n, \end{array} \tag{4.10}$$

In a realistic model we would have to consider transaction costs which decrease the expected return. In particular if a really large volume is traded then the trade itself will affect the price of the asset, a phenomenon called *market impact*. It is typically modeled by decreasing the expected return of $i$-th asset by a slippage cost proportional to $x^\beta$ for some $\beta > 1$, so that the objective function changes to

$$\text{maximize} \left( \mu^T x - \sum_i \delta_i x_i^\beta \right).$$

A popular choice is $\beta = 3/2$. This objective can easily be modeled with a power cone as in Sec. 4.2:

$$\begin{array}{lrcl} \text{maximize} & \mu^T x - \delta^T t & & \\ \text{subject to} & (t_i, 1, x_i) & \in & \mathcal{P}_3^{1/\beta, 1-1/\beta} \quad (t_i \geq x_i^\beta), \\ & \cdots & & \end{array}$$

In particular if $\beta = 3/2$ the inequality $t_i \geq x_i^{3/2}$ has conic representation $(t_i, 1, x_i) \in \mathcal{P}_3^{2/3, 1/3}$.

### 4.3.2 Maximum volume cuboid

Suppose we have a convex, conic representable set $K \subseteq \mathbb{R}^n$ (for instance a polyhedron, a ball, intersections of those and so on). We would like to find a maximum volume axis-parallel

cuboid inscribed in $K$. If we denote by $p \in \mathbb{R}^n$ the leftmost corner and by $x_1, \ldots, x_n$ the edge lengths, then this problem is equivalent to

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & t \leq (x_1 \cdots x_n)^{1/n}, \\ & x_i \geq 0, \\ & (p_1 + e_1 x_1, \ldots, p_n + e_n x_n) \in K, \quad \forall e_1, \ldots, e_n \in \{0, 1\}, \end{array} \tag{4.11}$$

where the last constraint states that all vertices of the cuboid are in $K$. The optimal volume is then $v = t^n$. Modeling the geometric mean with power cones was discussed in Sec. 4.2.4.

Maximizing the volume of an arbitrary (not necessarily axis-parallel) cuboid inscribed in $K$ is no longer a convex problem. However, it can be solved by maximizing the solution to (4.11) over all sets $T(K)$ where $T$ is a rotation (orthogonal matrix) in $\mathbb{R}^n$. In practice one can approximate the global solution by sampling sufficiently many rotations $T$ or using more advanced methods of optimization over the orthogonal group.

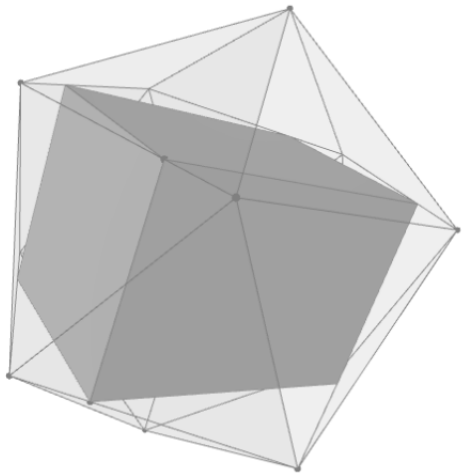


Fig. 4.2: The maximal volume cuboid inscribed in the regular icosahedron takes up approximately 0.388 of the volume of the icosahedron (the exact value is $3(1 + \sqrt{5})/25$).

### 4.3.3 $p$-norm geometric median

The *geometric median* of a sequence of points $x_1, \ldots, x_k \in \mathbb{R}^n$ is defined as

$$\operatorname{argmin}_{y \in \mathbb{R}^n} \sum_{i=1}^{k} \|y - x_i\|,$$

that is a point which minimizes the sum of distances to all the given points. Here $\|\cdot\|$ can be any norm on $\mathbb{R}^n$. The most classical case is the Euclidean norm, where the geometric median is the solution to the basic facility location problem minimizing total transportation cost from one depot to given destinations.

For a general $p$-norm $\|x\|_p$ with $1 \leq p < \infty$ (see Sec. 4.2.2) the geometric median is the solution to the obvious conic problem:

$$\begin{array}{ll} \text{minimize} & \sum_i t_i \\ \text{subject to} & t_i \geq \|y - x_i\|_p, \\ & y \in \mathbb{R}^n. \end{array} \tag{4.12}$$

In Sec. 4.2.2 we showed how to model the $p$-norm bound using $n$ power cones.

The *Fermat-Torricelli point* of a triangle is the Euclidean geometric mean of its vertices, and a classical theorem in planar geometry (due to Torricelli, posed by Fermat), states that it is the unique point inside the triangle from which each edge is visible at the angle of 120° (or a vertex if the triangle has an angle of 120° or more). Using (4.12) we can compute the $p$-norm analogues of the Fermat-Torricelli point. Some examples are shown in Fig. 4.3.

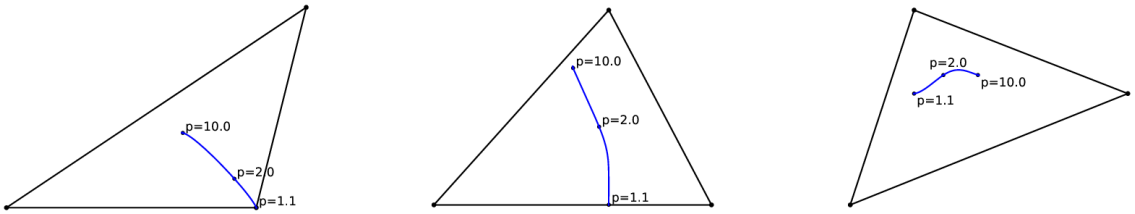



Fig. 4.3: The geometric median of three triangle vertices in various $p$-norms.

### 4.3.4 Maximum likelihood estimator of a convex density function

In *[TV98]* the problem of estimating a density function that is know in advance to be convex is considered. Here we will show that this problem can be posed as a conic optimization problem. Formally the problem is to estimate an unknown convex density function $g : \mathbb{R}_+ \to \mathbb{R}_+$ given an ordered sample $y_1 < y_2 < \ldots < y_n$ of $n$ outcomes of a distribution with density $g$.

The estimator $\tilde{g} \geq 0$ is a piecewise linear function

$$\tilde{g} : [y_1, y_n] \to \mathbb{R}_+$$

with break points at $(y_i, x_i)$, $i = 1, \ldots, n$, where the variables $x_i > 0$ are estimators for $g(y_i)$. The slope of the $i$-th linear segment of $\tilde{g}$ is

$$\frac{x_{i+1} - x_i}{y_{i+1} - y_i}.$$

Hence the convexity requirement leads to the constraints

$$\frac{x_{i+1} - x_i}{y_{i+1} - y_i} \leq \frac{x_{i+2} - x_{i+1}}{y_{i+2} - y_{i+1}}, \; i = 1, \ldots, n-2.$$

Recall the area under the density function must be 1. Hence,

$$\sum_{i=1}^{n-1}(y_{i+1}-y_i)\left(\frac{x_{i+1}+x_i}{2}\right)=1$$

must hold. Therefore, the problem to be solved is

$$\begin{array}{lrcll}\text{maximize} & \prod_{i=1}^{n} x_i & & & \\ \text{subject to} & \frac{x_{i+1}-x_i}{y_{i+1}-y_i}-\frac{x_{i+2}-x_{i+1}}{y_{i+2}-y_{i+1}} & \leq & 0, & i=1,\ldots,n-2, \\ & \sum_{i=1}^{n-1}(y_{i+1}-y_i)\left(\frac{x_{i+1}+x_i}{2}\right) & = & 1, & \\ & x & \geq & 0. & \end{array}$$

Maximizing $\prod_{i=1}^{n} x_i$ or the geometric mean $(\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$ will produce the same optimal solutions. Using that observation we can model the objective as shown in Sec. 4.2.4. Alternatively, one can use as objective $\sum_i \log x_i$ and model it using the exponential cone as in Sec. 5.2.

# Chapter 5

# Exponential cone optimization

So far we discussed optimization problems involving the major "polynomial" families of cones: linear, quadratic and power cones. In this chapter we introduce a single new object, namely the three-dimensional *exponential cone*, together with examples and applications. The exponential cone can be used to model a variety of constraints involving exponentials and logarithms.

## 5.1 Exponential cone

The exponential cone is a convex subset of $\mathbb{R}^3$ defined as

$$\begin{array}{rcl} K_{\text{exp}} & = & \left\{(x_1, x_2, x_3) \;:\; x_1 \geq x_2 e^{x_3/x_2},\; x_2 > 0\right\} \;\cup \\ & & \left\{(x_1, 0, x_3) \;:\; x_1 \geq 0,\; x_3 \leq 0\right\}. \end{array} \tag{5.1}$$

Thus the exponential cone is the closure in $\mathbb{R}^3$ of the set of points which satisfy

$$x_1 \geq x_2 e^{x_3/x_2},\; x_1, x_2 > 0. \tag{5.2}$$

When working with logarithms, a convenient reformulation of (5.2) is

$$x_3 \leq x_2 \log(x_1/x_2),\; x_1, x_2 > 0. \tag{5.3}$$

Alternatively, one can write the same condition as

$$x_1/x_2 \geq e^{x_3/x_2},\; x_1, x_2 > 0,$$

which immediately shows that $K_{\text{exp}}$ is in fact a cone, i.e. $\alpha x \in K_{\text{exp}}$ for $x \in K_{\text{exp}}$ and $\alpha \geq 0$. Convexity of $K_{\text{exp}}$ follows from the fact that the Hessian of $f(x, y) = y \exp(x/y)$, namely

$$D^2(f) = e^{x/y} \begin{bmatrix} y^{-1} & -xy^{-2} \\ -xy^{-2} & x^2 y^{-3} \end{bmatrix}$$

is positive semidefinite for $y > 0$.

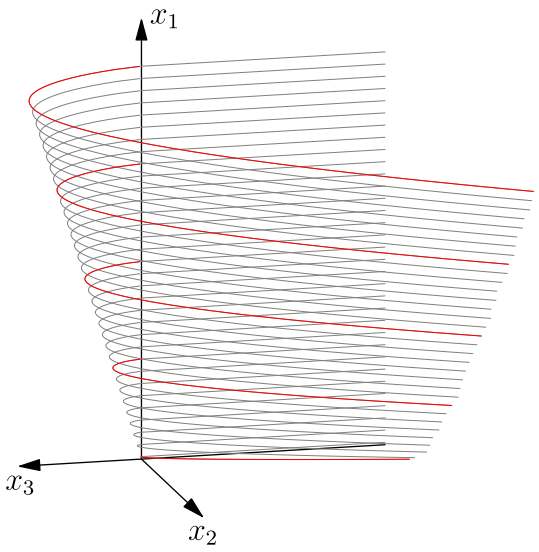



Fig. 5.1: The boundary of the exponential cone $K_{\text{exp}}$. The red isolines are graphs of $x_2 \to x_2 \log(x_1/x_2)$ for fixed $x_1$, see (5.3).

## 5.2 Modeling with the exponential cone

Extending the conic optimization toolbox with the exponential cone leads to new types of constraint building blocks and new types of representable sets. In this section we list the basic operations available using the exponential cone.

### 5.2.1 Exponential

The epigraph $t \geq e^x$ is a section of $K_{\text{exp}}$:

$$t \geq e^x \iff (t, 1, x) \in K_{\text{exp}}. \tag{5.4}$$

### 5.2.2 Logarithm

Similarly, we can express the hypograph $t \leq \log x,\ x \geq 0$:

$$t \leq \log x \iff (x, 1, t) \in K_{\text{exp}}. \tag{5.5}$$

### 5.2.3 Entropy

The entropy function $H(x) = -x \log x$ can be maximized using the following representation which follows directly from (5.3):

$$t \leq -x \log x \iff t \leq x \log(1/x) \iff (1, x, t) \in K_{\text{exp}}. \tag{5.6}$$

### 5.2.4 Relative entropy

The *relative entropy* or *Kullback-Leiber divergence* of two probability distributions is defined in terms of the function $D(x, y) = x \log(x/y)$. It is convex, and the minimization problem $t \geq D(x, y)$ is equivalent to

$$t \geq D(x, y) \iff -t \leq x \log(y/x) \iff (y, x, -t) \in K_{\exp}. \tag{5.7}$$

Because of this reparametrization the exponential cone is also referred to as the *relative entropy cone*, leading to a class of problems known as *REPs* (relative entropy problems). Having the relative entropy function available makes it possible to express epigraphs of other functions appearing in REPs, for instance:

$$x \log(1 + x/y) = D(x + y, y) + D(y, x + y).$$

### 5.2.5 Softplus function

In neural networks the function $f(x) = \log(1+e^x)$, known as the *softplus* function, is used as an analytic approximation to the rectifier activation function $r(x) = x^+ = \max(0, x)$. The softplus function is convex and we can express its epigraph $t \geq \log(1+e^x)$ by combining two exponential cones. Note that

$$t \geq \log(1 + e^x) \iff e^{x-t} + e^{-t} \leq 1$$

and therefore $t \geq \log(1 + e^x)$ is equivalent to the following set of conic constraints:

$$\begin{array}{rcl} u + v & \leq & 1, \\ (u, 1, x - t) & \in & K_{\exp}, \\ (v, 1, -t) & \in & K_{\exp}. \end{array} \tag{5.8}$$

### 5.2.6 Log-sum-exp

We can generalize the previous example to a *log-sum-exp* (logarithm of sum of exponentials) expression

$$t \geq \log(e^{x_1} + \cdots + e^{x_n}).$$

This is equivalent to the inequality

$$e^{x_1 - t} + \cdots + e^{x_n - t} \leq 1,$$

and so it can be modeled as follows:

$$\begin{array}{rcl} \sum u_i & \leq & 1, \\ (u_i, 1, x_i - t) & \in & K_{\exp},\ i = 1, \ldots, n. \end{array} \tag{5.9}$$

### 5.2.7 Log-sum-inv

The following type of bound has applications in capacity optimization for wireless network design:

$$t \geq \log\left(\frac{1}{x_1} + \cdots + \frac{1}{x_n}\right), \quad x_i > 0.$$

Since the logarithm is increasing, we can model this using a log-sum-exp and an exponential as:

$$\begin{array}{l} t \geq \log(e^{y_1} + \cdots + e^{y_n}), \\ x_i \geq e^{-y_i}, \quad i = 1, \ldots, n. \end{array}$$

### 5.2.8 Arbitrary exponential

The inequality

$$t \geq a_1^{x_1} a_2^{x_2} \cdots a_n^{x_n},$$

where $a_i$ are arbitrary positive constants, is of course equivalent to

$$t \geq \exp\left(\sum_i x_i \log a_i\right)$$

and therefore to $(t, 1, \sum_i x_i \log a_i) \in K_{\exp}$.

### 5.2.9 Lambert W-function

The *Lambert function* $W : \mathbb{R}_+ \to \mathbb{R}_+$ is the unique function satisfying the identity

$$W(x)e^{W(x)} = x.$$

It is the real branch of a more general function which appears in applications such as diode modeling. The $W$ function is concave. Although there is no explicit analytic formula for $W(x)$, the hypograph $\{(x,t) \ : \ 0 \leq x, \ 0 \leq t \leq W(x)\}$ has an equivalent description:

$$x \geq te^t = te^{t^2/t}$$

and so it can be modeled with a mix of exponential and quadratic cones (see Sec. 3.1.2):

$$\begin{array}{rcll} (x,t,u) & \in & K_{\exp}, & (x \geq t\exp(u/t)), \\ (1/2,u,t) & \in & \mathcal{Q}_r, & (u \geq t^2). \end{array} \tag{5.10}$$

### 5.2.10 Other simple sets

Here are a few more typical sets which can be expressed using the exponential and quadratic cones. The presentations should be self-explanatory; we leave the simple verifications to the reader.

Table 5.1: Sets representable with the exponential cone

| Set | Conic representation |
|---|---|
| $t \geq (\log x)^2,\ 0 < x \leq 1$ | $(\frac{1}{2}, t, u) \in \mathcal{Q}_r^3,\ (x, 1, u) \in K_{\exp},\ x \leq 1$ |
| $t \leq \log\log x,\ x > 1$ | $(u, 1, t) \in K_{\exp},\ (x, 1, u) \in K_{\exp}$ |
| $t \geq (\log x)^{-1},\ x > 1$ | $(u, t, \sqrt{2}) \in \mathcal{Q}_r^3,\ (x, 1, u) \in K_{\exp}$ |
| $t \leq \sqrt{\log x},\ x > 1$ | $(\frac{1}{2}, u, t) \in \mathcal{Q}_r^3,\ (x, 1, u) \in K_{\exp}$ |
| $t \leq \sqrt{x \log x},\ x > 1$ | $(x, u, \sqrt{2}t) \in \mathcal{Q}_r^3,\ (x, 1, u) \in K_{\exp}$ |
| $t \leq \log(1 - 1/x),\ x > 1$ | $(x, u, \sqrt{2}) \in \mathcal{Q}_r^3,\ (1 - u, 1, t) \in K_{\exp}$ |
| $t \geq \log(1 + 1/x),\ x > 0$ | $(x + 1, u, \sqrt{2}) \in \mathcal{Q}_r^3,\ (1 - u, 1, -t) \in K_{\exp}$ |

## 5.3 Geometric programming

Geometric optimization problems form a family of optimization problems with objective and constraints in special polynomial form. It is a rich class of problems solved by reformulating in logarithmic-exponential form, and thus a major area of applications for the exponential cone $K_{\exp}$. Geometric programming is used in circuit design, chemical engineering, mechanical engineering and finance, just to mention a few applications. We refer to *[BKVH07]* for a survey and extensive bibliography.

### 5.3.1 Definition and basic examples

A *monomial* is a real valued function of the form

$$f(x_1, \ldots, x_n) = c x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}, \tag{5.11}$$

where the exponents $a_i$ are arbitrary real numbers and $c > 0$. A *posynomial* (positive polynomial) is a sum of monomials. Thus the difference between a posynomial and a standard notion of a multi-variate polynomial known from algebra or calculus is that (i) posynomials can have arbitrary exponents, not just integers, but (ii) they can only have positive coefficients.

For example, the following functions are monomials (in variables $x, y, z$):

$$xy,\ 2x^{1.5}y^{-1}x^{0.3},\ 3\sqrt{xy/z},\ 1 \tag{5.12}$$

and the following are examples of posynomials:

$$2x + yz,\ 1.5x^3z + 5/y,\ (x^2y^2 + 3z^{-0.3})^4 + 1. \tag{5.13}$$

A *geometric program (GP)* is an optimization problem of the form

$$\begin{array}{lll} \text{minimize} & f_0(x) & \\ \text{subject to} & f_i(x) \leq 1, & i = 1, \ldots, m, \\ & x_j > 0, & j = 1, \ldots, n, \end{array} \tag{5.14}$$

where $f_0, \ldots, f_m$ are posynomials and $x = (x_1, \ldots, x_n)$ is the variable vector.

A geometric program (5.14) can be modeled in exponential conic form by making a substitution

$$x_j = e^{y_j}, \ j = 1, \ldots, n.$$

Under this substitution a monomial of the form (5.11) becomes

$$ce^{a_1 y_1} e^{a_2 y_2} \cdots e^{a_n y_n} = \exp(a_*^T y + \log c)$$

for $a_* = (a_1, \ldots, a_n)$. Consequently, the optimization problem (5.14) takes an equivalent form

$$\begin{array}{lll} \text{minimize } t & & \\ \text{subject to} & \log(\sum_k \exp(a_{0,k,*}^T y + \log c_{0,k})) & \leq t, \\ & \log(\sum_k \exp(a_{i,k,*}^T y + \log c_{i,k})) & \leq 0, \ i = 1, \ldots, m, \end{array} \tag{5.15}$$

where $a_{i,k} \in \mathbb{R}^n$ and $c_{i,k} \in \mathbb{R}$ for all $i, k$. These are now log-sum-exp constraints we already discussed in Sec. 5.2.6. In particular, the problem (5.15) is convex, as opposed to the posynomial formulation (5.14).

### Example

We demonstrate this reduction on a simple example. Take the geometric problem

$$\begin{array}{lrcl} \text{minimize} & x + y^2 z & & \\ \text{subject to} & 0.1\sqrt{x} + 2y^{-1} & \leq & 1, \\ & z^{-1} + yx^{-2} & \leq & 1. \end{array}$$

By substituting $x = e^u$, $y = e^v$, $z = e^w$ we get

$$\begin{array}{lrcl} \text{minimize } t & & & \\ \text{subject to} & \log(e^u + e^{2v+w}) & \leq & t, \\ & \log(e^{0.5u + \log 0.1} + e^{-v + \log 2}) & \leq & 0, \\ & \log(e^{-w} + e^{v-2u}) & \leq & 0. \end{array}$$

and using the log-sum-exp reduction from Sec. 5.2.6 we write an explicit conic problem:

$$\begin{array}{lll} \text{minimize } t & & \\ \text{subject to} & (p_1, 1, u - t), (q_1, 1, 2v + w - t) \in K_{\exp}, & p_1 + q_1 \leq 1, \\ & (p_2, 1, 0.5u + \log 0.1), (q_2, 1, -v + \log 2) \in K_{\exp}, & p_2 + q_2 \leq 1, \\ & (p_3, 1, -w), (q_3, 1, v - 2u) \in K_{\exp}, & p_3 + q_3 \leq 1. \end{array}$$

Solving this problem yields $(x, y, z) \approx (3.14, 2.43, 1.32)$.

### 5.3.2 Generalized geometric models

In this section we briefly discuss more general types of constraints which can be modeled with geometric programs.

#### Monomials

If $m(x)$ is a monomial then the constraint $m(x) = c$ is equivalent to two posynomial inequalities $m(x)c^{-1} \leq 1$ and $m(x)^{-1}c \leq 1$, so it can be expressed in the language of geometric programs. In practice it should be added to the model (5.15) as a linear constraint

$$\sum_k a_k y_k = \log c.$$

Monomial inequalities $m(x) \leq c$ and $m(x) \geq c$ should similarly be modeled as linear inequalities in the $y_i$ variables.

In similar vein, if $f(x)$ is a posynomial and $m(x)$ is a monomial then $f(x) \leq m(x)$ is still a posynomial inequality because it can be written as $f(x)m(x)^{-1} \leq 1$.

It also means that we can add lower and upper variable bounds: $0 < c_1 \leq x \leq c_2$ is equivalent to $x^{-1}c_1 \leq 1$ and $c_2^{-1}x \leq 1$.

#### Products and positive powers

Expressions involving products and positive powers (possibly iterated) of posynomials can again be modeled with posynomials. For example, a constraint such as

$$((xy^2 + z)^{0.3} + y)(1/x + 1/y)^{2.2} \leq 1$$

can be replaced with

$$xy^2 + z \leq t,\ t^{0.3} + y \leq u,\ x^{-1} + y^{-1} \leq v,\ uv^{2.2} \leq 1.$$

#### Other transformations and extensions

- If $f_1, f_2$ are already expressed by posynomials then $\max\{f_1(x), f_2(x)\} \leq t$ is clearly equivalent to $f_1(x) \leq t$ and $f_2(x) \leq t$. Hence we can add the maximum operator to the list of building blocks for geometric programs.

- If $f, g$ are posynomials, $m$ is a monomial and we know that $m(x) \geq g(x)$ then the constraint $\frac{f(x)}{m(x)-g(x)} \leq t$ is equivalent to $t^{-1}f(x)m(x)^{-1} + g(x)m(x)^{-1} \leq 1$.

- The objective function of a geometric program (5.14) can be extended to include other terms, for example:

$$\text{minimize } f_0(x) + \sum_k \log m_k(x)$$

where $m_k(x)$ are monomials. After the change of variables $x = e^y$ we get a slightly modified version of (5.15):

$$\begin{array}{lrcl} \text{minimize} & t + b^T y & & \\ \text{subject to} & \sum_k \exp(a_{0,k,*}^T y + \log c_{0,k}) & \leq & t, \\ & \log(\sum_k \exp(a_{i,k,*}^T y + \log c_{i,k})) & \leq & 0,\ i = 1, \ldots, m, \end{array}$$

(note the lack of one logarithm) which can still be expressed with exponential cones.

### 5.3.3 Geometric programming case studies

#### Frobenius norm diagonal scaling

Suppose we have a matrix $M \in \mathbb{R}^{n \times n}$ and we want to rescale the coordinate system using a diagonal matrix $D = \mathbf{Diag}(d_1, \ldots, d_n)$ with $d_i > 0$. In the new basis the linear transformation given by $M$ will now be described by the matrix $DMD^{-1} = (d_i M_{ij} d_j^{-1})_{i,j}$. To choose $D$ which leads to a "small" rescaling we can for example minimize the Frobenius norm

$$\|DMD^{-1}\|_F^2 = \sum_{ij} \left((DMD^{-1})_{ij}\right)^2 = \sum_{ij} M_{ij}^2 d_i^2 d_j^{-2}.$$

Minimizing the last sum is an example of a geometric program with variables $d_i$ (and without constraints).

#### Maximum likelihood estimation

Geometric programs appear naturally in connection with maximum likelihood estimation of parameters of random distributions. Consider a simple example. Suppose we have two biased coins, with head probabilities $p$ and $2p$, respectively. We toss both coins and count the total number of heads. Given that, in the long run, we observed $i$ heads $n_i$ times for $i = 0, 1, 2$, estimate the value of $p$.

The probability of obtaining the given outcome equals

$$\binom{n_0 + n_1 + n_2}{n_0} \binom{n_1 + n_2}{n_1} (p \cdot 2p)^{n_2} \left(p(1-2p) + 2p(1-p)\right)^{n_1} \left((1-p)(1-2p)\right)^{n_0},$$

and, up to constant factors, maximizing that expression is equivalent to solving the problem

$$\begin{array}{lrcl} \text{maximize} & p^{2n_2+n_1} s^{n_1} q^{n_0} r^{n_0} & & \\ \text{subject to} & q & \leq & 1-p, \\ & r & \leq & 1-2p, \\ & s & \leq & 3-4p, \end{array}$$

or, as a geometric problem:

$$\begin{array}{lrcl} \text{minimize} & p^{-2n_2-n_1} s^{-n_1} q^{-n_0} r^{-n_0} & & \\ \text{subject to} & q+p & \leq & 1, \\ & r+2p & \leq & 1, \\ & \frac{1}{3}s + \frac{4}{3}p & \leq & 1. \end{array}$$

For example, if $(n_0, n_1, n_2) = (30, 53, 16)$ then the above problem solves with $p = 0.29$.

### An Olympiad problem

The 26th Vojtěch Jarník International Mathematical Competition, Ostrava 2016. Let $a, b, c$ be positive real numbers with $a + b + c = 1$. Prove that

$$\left(\frac{1}{a} + \frac{1}{bc}\right)\left(\frac{1}{b} + \frac{1}{ac}\right)\left(\frac{1}{c} + \frac{1}{ab}\right) \geq 1728$$

Using the tricks introduced in Sec. 5.3.2 we formulate this problem as a geometric program:

$$\begin{array}{lrcl} \text{minimize} & pqr & & \\ \text{subject to} & p^{-1}a^{-1} + p^{-1}b^{-1}c^{-1} & \leq & 1, \\ & q^{-1}b^{-1} + q^{-1}a^{-1}c^{-1} & \leq & 1, \\ & r^{-1}c^{-1} + r^{-1}a^{-1}b^{-1} & \leq & 1, \\ & a + b + c & \leq & 1. \end{array}$$

Unsurprisingly, the optimal value of this program is 1728, achieved for $(a, b, c, p, q, r) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 12, 12, 12\right)$.

### Power control and rate allocation in wireless networks

We consider a basic wireless network power control problem. In a wireless network with $n$ logical transmitter-receiver pairs if the power output of transmitter $j$ is $p_j$ then the power received by receiver $i$ is $G_{ij}p_j$, where $G_{ij}$ models path gain and fading effects. If the $i$-th receiver's own noise is $\sigma_i$ then the *signal-to-interference-plus-noise (SINR) ratio* of receiver $i$ is given by

$$s_i = \frac{G_{ii}p_i}{\sigma_i + \sum_{j \neq i} G_{ij}p_j}. \tag{5.16}$$

Maximizing the minimal SINR over all receivers ($\max \min_i s_i$), subject to bounded power output of the transmitters, is equivalent to the geometric program with variables $p_1, \ldots, p_n, t$:

$$\begin{array}{lrcl} \text{minimize} & t^{-1} & & \\ \text{subject to} & p_{\min} \leq p_j \leq p_{\max}, & & j = 1, \ldots, n, \\ & t(\sigma_i + \sum_{j \neq i} G_{ij}p_j)G_{ii}^{-1}p_i^{-1} & \leq & 1, \ i = 1, \ldots, n. \end{array} \tag{5.17}$$

In the low-SNR regime the problem of system rate maximization is approximated by the problem of maximizing $\sum_i \log s_i$, or equivalently minimizing $\prod_i s_i^{-1}$. This is a geometric problem with variables $p_i, s_i$:

$$\begin{array}{lrcl} \text{minimize} & s_1^{-1} \cdot \cdots \cdot s_n^{-1} & & \\ \text{subject to} & p_{\min} \leq p_j \leq p_{\max}, & & j = 1, \ldots, n, \\ & s_i(\sigma_i + \sum_{j \neq i} G_{ij}p_j)G_{ii}^{-1}p_i^{-1} & \leq & 1, \ i = 1, \ldots, n. \end{array} \tag{5.18}$$

For more information and examples see *[BKVH07]*.

## 5.4 Exponential cone case studies

In this section we introduce some practical optimization problems where the exponential cone comes in handy.

### 5.4.1 Risk parity portfolio

Consider a simple version of the Markowitz portfolio optimization problem introduced in Sec. 3.3.3, where we simply ask to minimize the risk $r(x) = \sqrt{x^T \Sigma x}$ of a fully-invested long-only portfolio:

$$\begin{array}{lll} \text{minimize} & \sqrt{x^T \Sigma x} & \\ \text{subject to} & \sum_{i=1}^n x_i = 1, & \\ & x_i \geq 0, \ i = 1, \ldots, n, & \end{array} \tag{5.19}$$

where $\Sigma$ is a symmetric positive definite covariance matrix. We can derive from the first-order optimality conditions that the solution to (5.19) satisfies $\frac{\partial r}{\partial x_i} = \frac{\partial r}{\partial x_j}$ whenever $x_i, x_j > 0$, i.e. marginal risk contributions of positively invested assets are equal. In practice this often leads to concentrated portfolios, whereas it would benefit diversification to consider portfolios where all assets have the same total contribution to risk:

$$x_i \frac{\partial r}{\partial x_i} = x_j \frac{\partial r}{\partial x_j}, \ i, j = 1, \ldots, n. \tag{5.20}$$

We call (5.20) the *risk parity* condition. It indeed models equal risk contribution from all the assets, because as one can easily check $\frac{\partial r}{\partial x_i} = \frac{(\Sigma x)_i}{\sqrt{x^T \Sigma x}}$ and

$$r(x) = \sum_i x_i \frac{\partial r}{\partial x_i}.$$

Risk parity portfolios satisfying condition (5.20) can be found with an auxiliary optimization problem:

$$\begin{array}{ll} \text{minimize} & \sqrt{x^T \Sigma x} - c \sum_i \log x_i \\ \text{subject to} & x_i > 0, \quad i = 1, \ldots, n, \end{array} \tag{5.21}$$

for any $c > 0$. More precisely, the gradient of the objective function in (5.21) is zero when $\frac{\partial r}{\partial x_i} = c/x_i$ for all $i$, implying the parity condition (5.20) holds. Since (5.20) is scale-invariant, we can rescale any solution of (5.21) and get a fully-invested risk parity portfolio with $\sum_i x_i = 1$.

The conic form of problem (5.21) is:

$$\begin{array}{lllll} \text{minimize} & t - ce^T s & & & \\ \text{subject to} & (t, \Sigma^{1/2} x) & \in & \mathcal{Q}^{n+1}, & (t \geq \sqrt{x^T \Sigma x}), \\ & (x_i, 1, s_i) & \in & K_{\text{exp}}, & (s_i \leq \log x_i). \end{array} \tag{5.22}$$

### 5.4.2 Entropy maximization

A general entropy maximization problem has the form

$$\begin{array}{lrcl} \text{maximize} & -\sum_i p_i \log p_i & & \\ \text{subject to} & \sum_i p_i & = & 1, \\ & p_i & \geq & 0, \\ & p & \in & \mathcal{I}, \end{array} \tag{5.23}$$

where $\mathcal{I}$ defines additional constraints on the probability distribution $p$ (these are known as *prior information*). In the absence of complete information about $p$ the maximum entropy principle of Jaynes posits to choose the distribution which maximizes uncertainty, that is entropy, subject to what is known. Practitioners think of the solution to (5.23) as the most random or most conservative of distributions consistent with $\mathcal{I}$.

Maximization of the entropy function $H(x) = -x \log x$ was explained in Sec. 5.2.

Often one has an a priori distribution $q$, and one tries to minimize the distance between $p$ and $q$, while remaining consistent with $\mathcal{I}$. In this case it is standard to minimize the *Kullback-Leiber divergence*

$$\mathcal{D}_{KL}(p||q) = \sum_i p_i \log p_i/q_i$$

which leads to an optimization problem

$$\begin{array}{lrcl} \text{minimize} & \sum_i p_i \log p_i/q_i & & \\ \text{subject to} & \sum_i p_i & = & 1, \\ & p_i & \geq & 0, \\ & p & \in & \mathcal{I}. \end{array} \tag{5.24}$$

### 5.4.3 Hitting time of a linear system

Consider a linear dynamical system

$$\mathbf{x}'(t) = A\mathbf{x}(t) \tag{5.25}$$

where we assume for simplicity that $A = \mathbf{Diag}(a_1, \ldots, a_n)$ with $a_i < 0$ with initial condition $\mathbf{x}(0)_i = x_i$. The resulting dynamical system $\mathbf{x}(t) = \mathbf{x}(0)\exp(At)$ converges to 0 and one can ask, for instance, for the time it takes to approach the limit up to distance $\varepsilon$. The resulting optimization problem is

$$\begin{array}{lrcl} \text{minimize} & t & & \\ \text{subject to} & \sqrt{\sum_i \left(x_i \exp(a_i t)\right)^2} & \leq & \varepsilon, \end{array}$$

with the following conic form, where the variables are $t, q_1, \ldots, q_n$:

$$\begin{array}{lrcl} \text{minimize} & t & & \\ \text{subject to} & (\varepsilon, x_1 q_1, \ldots, x_n q_n) & \in & \mathcal{Q}^{n+1}, \\ & (q_i, 1, a_i t) & \in & K_{\exp},\ i = 1, \ldots, n. \end{array}$$

See Fig. 5.2 for an example. Other criteria for the target set of the trajectories are also possible. For example, polyhedral constraints

$$c^T \mathbf{x} \leq d,\ c \in \mathbb{R}^n_+, d \in \mathbb{R}_+$$

are also expressible in exponential conic form for starting points $\mathbf{x}(0) \in \mathbb{R}^n_+$, since they correspond to log-sum-exp constraints of the form

$$\log\left(\sum_i \exp(a_i \mathbf{x}_i + \log(c_i x_i))\right) \leq \log d.$$

For a robust version involving uncertainty on $A$ and $x(0)$ see *[CS14]*.

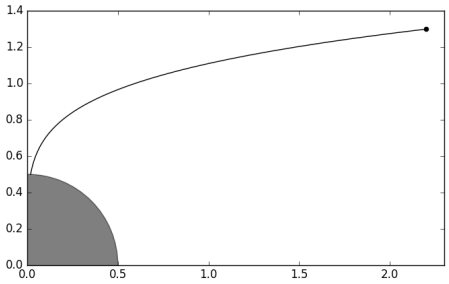



Fig. 5.2: With $A = \mathbf{Diag}(-0.3, -0.06)$ and starting point $x(0) = (2.2, 1.3)$ the trajectory reaches distance $\varepsilon = 0.5$ from origin at time $t \approx 15.936$.

### 5.4.4 Logistic regression

Logistic regression is a technique of training a binary classifier. We are given a training set of examples $x_1, \ldots, x_n \in \mathbb{R}^d$ together with labels $y_i \in \{0, 1\}$. The goal is to train a classifier capable of assigning new data points to either class 0 or 1. Specifically, the labels should be assigned by computing

$$h_\theta(x) = \frac{1}{1 + \exp(-\theta^T x)} \tag{5.26}$$

and choosing label $y = 0$ if $h_\theta(x) < \frac{1}{2}$ and $y = 1$ for $h_\theta(x) \geq \frac{1}{2}$. Here $h_\theta(x)$ is interpreted as the probability that $x$ belongs to class 1. The optimal parameter vector $\theta$ should be learned from the training set, so as to maximize the likelihood function:

$$\prod_i h_\theta(x_i)^{y_i}(1 - h_\theta(x_i))^{1-y_i}.$$

By taking logarithms and adding regularization with respect to $\theta$ we reach an unconstrained optimization problem

$$\text{minimize}_{\theta \in \mathbb{R}^d}\ \lambda\|\theta\|_2 + \sum_i -y_i \log(h_\theta(x_i)) - (1 - y_i)\log(1 - h_\theta(x_i)). \tag{5.27}$$

Problem (5.27) is convex, and can be more explicitly written as

$$\begin{array}{lrlll} \text{minimize} & \sum_i t_i + \lambda r & & & \\ \text{subject to} & t_i & \geq -\log(h_\theta(x)) & = \log(1+\exp(-\theta^T x_i)) & \text{if } y_i = 1, \\ & t_i & \geq -\log(1-h_\theta(x)) & = \log(1+\exp(\theta^T x_i)) & \text{if } y_i = 0, \\ & r & \geq \|\theta\|_2, & & \end{array}$$

involving softplus type constraints (see Sec. 5.2) and a quadratic cone. See Fig. 5.3 for an example.

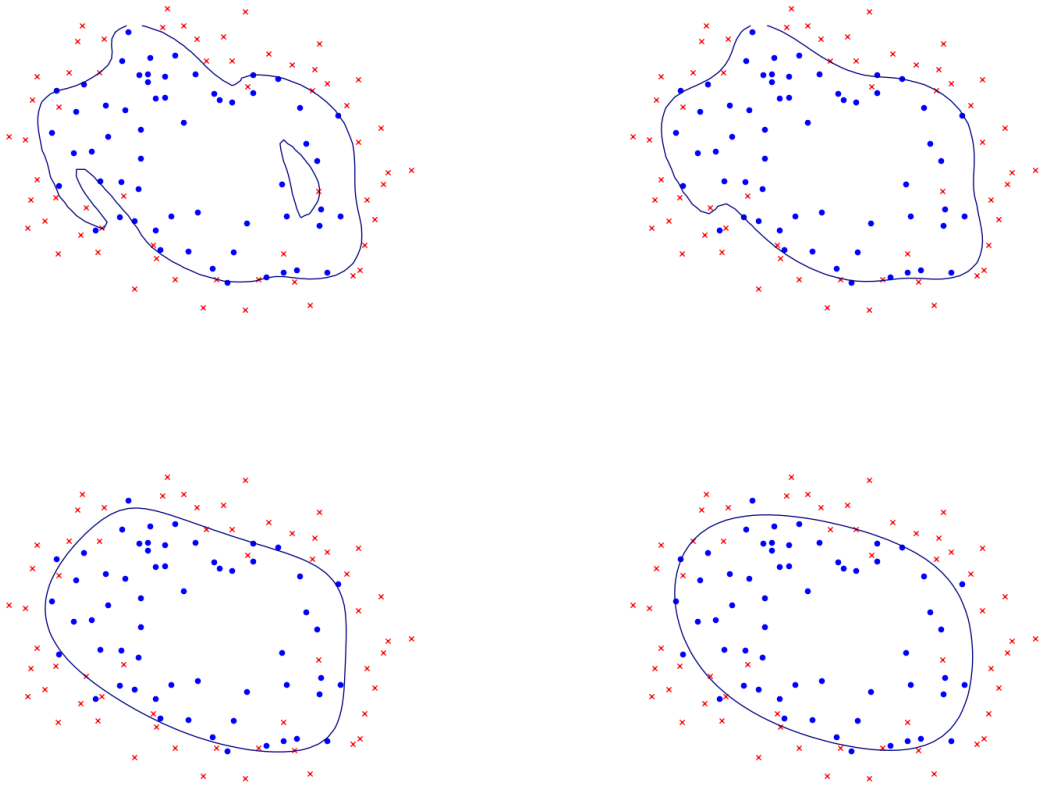


Fig. 5.3: Logistic regression example with none, medium and strong regularization (small, medium, large $\lambda$). The two-dimensional dataset was converted into a feature vector $x \in \mathbb{R}^{28}$ using monomial coordinates of degrees at most 6. Without regularization we get obvious overfitting.

# Chapter 6

# Semidefinite optimization

In this chapter we extend the conic optimization framework introduced before with symmetric positive semidefinite matrix variables.

## 6.1 Introduction to semidefinite matrices

### 6.1.1 Semidefinite matrices and cones

A symmetric matrix $X \in \mathcal{S}^n$ is called *symmetric positive semidefinite* if

$$z^T X z \geq 0, \quad \forall z \in \mathbb{R}^n.$$

We then define the cone of symmetric positive semidefinite matrices as

$$\mathcal{S}^n_+ = \{X \in \mathcal{S}^n \mid z^T X z \geq 0, \, \forall z \in \mathbb{R}^n\}. \tag{6.1}$$

For brevity we will often use the shorter notion *semidefinite* instead of *symmetric positive semidefinite*, and we will write $X \succeq Y$ ($X \preceq Y$) as shorthand notation for $(X - Y) \in \mathcal{S}^n_+$ ($(Y - X) \in \mathcal{S}^n_+$). As inner product for semidefinite matrices, we use the standard trace inner product for general matrices, i.e.,

$$\langle A, B \rangle := \mathrm{tr}(A^T B) = \sum_{ij} a_{ij} b_{ij}.$$

It is easy to see that (6.1) indeed specifies a convex cone; it is pointed (with origin $X = 0$), and $X, Y \in \mathcal{S}^n_+$ implies that $(\alpha X + \beta Y) \in \mathcal{S}^n_+$, $\alpha, \beta \geq 0$. Let us review a few equivalent definitions of $\mathcal{S}^n_+$. It is well-known that every symmetric matrix $A$ has a spectral factorization

$$A = \sum_{i=1}^n \lambda_i q_i q_i^T.$$

where $q_i \in \mathbb{R}^n$ are the (orthogonal) eigenvectors and $\lambda_i$ are eigenvalues of $A$. Using the spectral factorization of $A$ we have

$$x^T A x = \sum_{i=1}^n \lambda_i (x^T q_i)^2,$$

which shows that $x^T Ax \geq 0 \Leftrightarrow \lambda_i \geq 0,\ i = 1, \ldots, n$. In other words,

$$\mathcal{S}_+^n = \{X \in \mathcal{S}^n \mid \lambda_i(X) \geq 0,\ i = 1, \ldots, n\}. \tag{6.2}$$

Another useful characterization is that $A \in \mathcal{S}_+^n$ if and only if it is a *Grammian matrix* $A = V^T V$. Here $V$ is called the *Cholesky factor* of $A$. Using the Grammian representation we have

$$x^T Ax = x^T V^T Vx = \|Vx\|_2^2,$$

i.e., if $A = V^T V$ then $x^T Ax \geq 0$ for all $x$. On the other hand, from the positive spectral factorization $A = Q\Lambda Q^T$ we have $A = V^T V$ with $V = \Lambda^{1/2} Q^T$, where $\Lambda^{1/2} = \mathbf{Diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$. We thus have the equivalent characterization

$$\mathcal{S}_+^n = \{X \in \mathcal{S}_+^n \mid X = V^T V \text{ for some } V \in \mathbb{R}^{n \times n}\}. \tag{6.3}$$

In a completely analogous way we define the cone of *symmetric positive definite matrices* as

$$\begin{array}{rl} \mathcal{S}_{++}^n & = \{X \in \mathcal{S}^n \mid z^T Xz > 0,\ \forall z \in \mathbb{R}^n\} \\ & = \{X \in \mathcal{S}^n \mid \lambda_i(X) > 0,\ i = 1, \ldots, n\} \\ & = \{X \in \mathcal{S}_+^n \mid X = V^T V \text{ for some } V \in \mathbb{R}^{n \times n},\ \mathbf{rank}(V) = n\}, \end{array}$$

and we write $X \succ Y$ ($X \prec Y$) as shorthand notation for $(X - Y) \in \mathcal{S}_{++}^n$ ($(Y - X) \in \mathcal{S}_{++}^n$).

The one dimensional cone $\mathcal{S}_+^1$ simply corresponds to $\mathbb{R}_+$. Similarly consider

$$X = \begin{bmatrix} x_1 & x_3 \\ x_3 & x_2 \end{bmatrix}$$

with determinant $\det(X) = x_1 x_2 - x_3^2 = \lambda_1 \lambda_2$ and trace $\mathrm{tr}(X) = x_1 + x_2 = \lambda_1 + \lambda_2$. Therefore $X$ has positive eigenvalues if and only if

$$x_1 x_2 \geq x_3^2, \quad x_1, x_2 \geq 0,$$

which characterizes a three-dimensional scaled rotated cone, i.e.,

$$\begin{bmatrix} x_1 & x_3 \\ x_3 & x_2 \end{bmatrix} \in \mathcal{S}_+^2 \quad \Longleftrightarrow \quad (x_1, x_2, x_3\sqrt{2}) \in \mathcal{Q}_r^3.$$

More generally we have

$$x \in \mathbb{R}_+^n \quad \Longleftrightarrow \quad \mathbf{Diag}(x) \in \mathcal{S}_+^n$$

and

$$(t, x) \in \mathcal{Q}^{n+1} \quad \Longleftrightarrow \quad \begin{bmatrix} t & x^T \\ x & tI \end{bmatrix} \in \mathcal{S}_+^{n+1},$$

where the latter equivalence follows immediately from Lemma 6.1. Thus both the linear and quadratic cone are embedded in the semidefinite cone. In practice, however, linear and quadratic cones should never be described using semidefinite constraints, which would result in a large performance penalty by squaring the number of variables.

**Example 6.1.** As a more interesting example, consider the symmetric matrix

$$A(x,y,z) = \begin{bmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{bmatrix} \tag{6.4}$$

parametrized by $(x, y, z)$. The set

$$S = \{(x,y,z) \in \mathbb{R}^3 \mid A(x,y,z) \in \mathcal{S}^3_+\},$$

(shown in Fig. 6.1) is called a *spectrahedron* and is perhaps the simplest bounded semidefinite representable set, which cannot be represented using (finitely many) linear or quadratic cones. To gain a geometric intuition of $S$, we note that

$$\det(A(x,y,z)) = -(x^2 + y^2 + z^2 - 2xyz - 1),$$

so the boundary of $S$ can be characterized as

$$x^2 + y^2 + z^2 - 2xyz = 1,$$

or equivalently as

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

For $z = 0$ this describes a circle in the $(x, y)$-plane, and for $-1 \leq z \leq 1$ it characterizes an ellipse (for a fixed $z$).

### 6.1.2 Properties of semidefinite matrices

Many useful properties of (semi)definite matrices follow directly from the definitions (6.1)-(6.3) and their definite counterparts.

- The diagonal elements of $A \in \mathcal{S}^n_+$ are nonnegative. Let $e_i$ denote the $i$th standard basis vector (i.e., $[e_i]_j = 0$, $j \neq i$, $[e_i]_i = 1$). Then $A_{ii} = e_i^T A e_i$, so (6.1) implies that $A_{ii} \geq 0$.
- A block-diagonal matrix $A = \mathbf{Diag}(A_1, \ldots A_p)$ is (semi)definite if and only if each diagonal block $A_i$ is (semi)definite.
- Given a quadratic transformation $M := B^T A B$, $M \succ 0$ if and only if $A \succ 0$ and $B$ has full rank. This follows directly from the Grammian characterization $M = (VB)^T(VB)$. For $M \succeq 0$ we only require that $A \succeq 0$. As an example, if $A$ is (semi)definite then so is any permutation $P^T A P$.

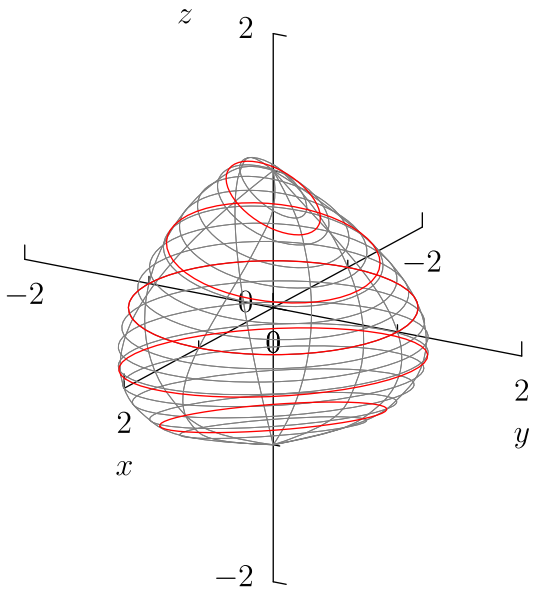



Fig. 6.1: Plot of spectrahedron $S = \{(x, y, z) \in \mathbb{R}^3 \mid A(x, y, z) \succeq 0\}$.

- Any principal submatrix of $A \in \mathcal{S}_+^n$ ($A$ restricted to the same set of rows as columns) is positive semidefinite; this follows by restricting the Grammian characterization $A = V^T V$ to a submatrix of $V$.

- The inner product of positive (semi)definite matrices is positive (nonnegative). For any $A, B \in \mathcal{S}_{++}^n$ let $A = U^T U$ and $B = V^T V$ where $U$ and $V$ have full rank. Then

$$\langle A, B \rangle = \operatorname{tr}(U^T U V^T V) = \|UV^T\|_F^2 > 0,$$

  where strict positivity follows from the assumption that $U$ has full column-rank, i.e., $UV^T \neq 0$.

- The inverse of a positive definite matrix is positive definite. This follows from the positive spectral factorization $A = Q\Lambda Q^T$, which gives us

$$A^{-1} = Q^T \Lambda^{-1} Q$$

  where $\Lambda_{ii} > 0$. If $A$ is semidefinite then the *pseudo-inverse* $A^\dagger$ of $A$ is semidefinite.

- Consider a matrix $X \in \mathcal{S}^n$ partitioned as

$$X = \begin{bmatrix} A & B^T \\ B & C \end{bmatrix}.$$

  Let us find necessary and sufficient conditions for $X \succ 0$. We know that $A \succ 0$ and $C \succ 0$ (since any principal submatrix must be positive definite). Furthermore, we can

simplify the analysis using a nonsingular transformation

$$L = \begin{bmatrix} I & 0 \\ F & I \end{bmatrix}$$

to diagonalize $X$ as $LXL^T = D$, where $D$ is block-diagonal. Note that $\det(L) = 1$, so $L$ is indeed nonsingular. Then $X \succ 0$ if and only if $D \succ 0$. Expanding $LXL^T = D$, we get

$$\begin{bmatrix} A & AF^T + B^T \\ FA + B & FAF^T + FB^T + BF^T + C \end{bmatrix} = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}.$$

Since $\det(A) \neq 0$ (by assuming that $A \succ 0$) we see that $F = -BA^{-1}$ and direct substitution gives us

$$\begin{bmatrix} A & 0 \\ 0 & C - BA^{-1}B^T \end{bmatrix} = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}.$$

In the last part we have thus established the following useful result.

**Lemma 6.1** (Schur complement lemma)**.** *A symmetric matrix*

$$X = \begin{bmatrix} A & B^T \\ B & C \end{bmatrix}.$$

*is positive definite if and only if*

$$A \succ 0, \quad C - BA^{-1}B^T \succ 0.$$

## 6.2 Semidefinite modeling

Having discussed different characterizations and properties of semidefinite matrices, we next turn to different functions and sets that can be modeled using semidefinite cones and variables. Most of those representations involve semidefinite matrix-valued affine functions, which we discuss next.

### 6.2.1 Linear matrix inequalities

Consider an affine matrix-valued mapping $A : \mathbb{R}^n \mapsto \mathcal{S}^m$:

$$A(x) = A_0 + x_1 A_1 + \cdots + x_n A_n. \tag{6.5}$$

A *linear matrix inequality (LMI)* is a constraint of the form

$$A_0 + x_1 A_1 + \cdots + x_n A_n \succeq 0 \tag{6.6}$$

in the variable $x \in \mathbb{R}^n$ with symmetric coefficients $A_i \in \mathcal{S}^m$, $i = 0, \ldots, n$. As a simple example consider the matrix in (6.4),

$$A(x, y, z) = A_0 + xA_1 + yA_2 + zA_3 \succeq 0$$

with

$$A_0 = I, \quad A_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Alternatively, we can describe the linear matrix inequality $A(x, y, z) \succeq 0$ as

$$X \in \mathcal{S}^3_+, \qquad x_{11} = x_{22} = x_{33} = 1,$$

i.e., as a semidefinite variable with fixed diagonal; these two alternative formulations illustrate the difference between primal and dual form of semidefinite problems, see Sec. 8.6.

### 6.2.2 Eigenvalue optimization

Consider a symmetric matrix $A \in \mathcal{S}^m$ and let its eigenvalues be denoted by

$$\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_m(A).$$

A number of different functions of $\lambda_i$ can then be described using a mix of linear and semidefinite constraints.

#### Sum of eigenvalues

The sum of the eigenvalues corresponds to

$$\sum_{i=1}^{m} \lambda_i(A) = \operatorname{tr}(A).$$

#### Largest eigenvalue

The largest eigenvalue can be characterized in epigraph form $\lambda_1(A) \leq t$ as

$$tI - A \succeq 0. \tag{6.7}$$

To verify this, suppose we have a spectral factorization $A = Q\Lambda Q^T$ where $Q$ is orthogonal and $\Lambda$ is diagonal. Then $t$ is an upper bound on the largest eigenvalue if and only if

$$Q^T(tI - A)Q = tI - \Lambda \succeq 0.$$

Thus we can minimize the largest eigenvalue of $A$.

### Smallest eigenvalue

The smallest eigenvalue can be described in hypograph form $\lambda_m(A) \geq t$ as

$$A \succeq tI, \tag{6.8}$$

i.e., we can maximize the smallest eigenvalue of $A$.

### Eigenvalue spread

The eigenvalue spread can be modeled in epigraph form

$$\lambda_1(A) - \lambda_m(A) \leq t$$

by combining the two linear matrix inequalities in (6.7) and (6.8), i.e.,

$$\begin{array}{c} zI \preceq A \preceq sI, \\ s - z \leq t. \end{array} \tag{6.9}$$

### Spectral radius

The spectral radius $\rho(A) := \max_i |\lambda_i(A)|$ can be modeled in epigraph form $\rho(A) \leq t$ using two linear matrix inequalities

$$-tI \preceq A \preceq tI.$$

### Condition number of a positive definite matrix

Suppose now that $A \in \mathcal{S}^m_+$. The condition number of a positive definite matrix can be minimized by noting that $\lambda_1(A)/\lambda_m(A) \leq t$ if and only if there exists a $\mu > 0$ such that

$$\mu I \preceq A \preceq \mu t I,$$

or equivalently if and only if $I \preceq \mu^{-1} A \preceq tI$. If $A = A(x)$ is represented as in (6.5) then a change of variables $z := x/\mu$, $\nu := 1/\mu$ leads to a problem of the form

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & I \preceq \nu A_0 + \sum_{i=1}^m z_i A_i \preceq tI, \end{array} \tag{6.10}$$

from which we recover the solution $x = z/\nu$. In essence, we first normalize the spectrum by the smallest eigenvalue, and then minimize the largest eigenvalue of the normalized linear matrix inequality. Compare Sec. 2.2.5.

### 6.2.3 Log-determinant

Consider again a symmetric positive-definite matrix $A \in \mathcal{S}^m_+$. The determinant

$$\det(A) = \prod_{i=1}^{m} \lambda_i(A)$$

is neither convex or concave, but $\log\det(A)$ is concave and we can write the inequality

$$t \leq \log\det(A)$$

in the form of the following problem:

$$\begin{array}{l} \left[\begin{array}{cc} A & Z \\ Z^T & \mathbf{diag}(Z) \end{array}\right] \succeq 0, \\ Z \text{ is lower triangular}, \\ t \leq \sum_i \log Z_{ii}. \end{array} \tag{6.11}$$

The equivalence of the two problems follows from Lemma 6.1 and subadditivity of determinant for semidefinite matrices. That is:

$$\begin{array}{rl} 0 \leq \det(A - Z^T\mathbf{diag}(Z)^{-1}Z) & \leq \det(A) - \det(Z^T\mathbf{diag}(Z)^{-1}Z) \\ & = \det(A) - \det(Z). \end{array}$$

On the other hand the optimal value $\det(A)$ is attained for $Z = LD$ if $A = LDL^T$ is the LDL factorization of $A$.

The last inequality in problem (6.11) can of course be modeled using the exponential cone as in Sec. 5.2. Note that we can replace that bound with $t \leq (\prod_i Z_{ii})^{1/m}$ to get instead the model of $t \leq \det(A)^{1/m}$ using Sec. 4.2.4.

### 6.2.4 Singular value optimization

We next consider a non-square matrix $A \in \mathbb{R}^{m\times p}$. Assume $p \leq m$ and denote the singular values of $A$ by

$$\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_p(A) \geq 0.$$

The singular values are connected to the eigenvalues of $A^TA$ via

$$\sigma_i(A) = \sqrt{\lambda_i(A^TA)}, \tag{6.12}$$

and if $A$ is square and symmetric then $\sigma_i(A) = |\lambda_i(A)|$. We show next how to optimize several functions of the singular values.

### Largest singular value

The epigraph $\sigma_1(A) \leq t$ can be characterized using (6.12) as

$$A^T A \preceq t^2 I,$$

which from Schur's lemma is equivalent to

$$\begin{bmatrix} tI & A \\ A^T & tI \end{bmatrix} \succeq 0. \tag{6.13}$$

The largest singular value $\sigma_1(A)$ is also called the *spectral norm* or the $\ell_2$-norm of $A$, $\|A\|_2 := \sigma_1(A)$.

### Sum of singular values

The *trace* norm or the *nuclear* norm of $X$ is the dual of the $\ell_2$-norm:

$$\|X\|_* = \sup_{\|Z\|_2 \leq 1} \operatorname{tr}(X^T Z). \tag{6.14}$$

It turns out that the nuclear norm corresponds to the sum of the singular values,

$$\|X\|_* = \sum_{i=1}^{m} \sigma_i(X) = \sum_{i=1}^{n} \sqrt{\lambda_i(X^T X)}, \tag{6.15}$$

which is easy to verify using singular value decomposition $X = U\Sigma V^T$. We have

$$\begin{aligned} \sup_{\|Z\|_2 \leq 1} \operatorname{tr}(X^T Z) &= \sup_{\|Z\|_2 \leq 1} \operatorname{tr}(\Sigma^T U^T Z V) \\ &= \sup_{\|Y\|_2 \leq 1} \operatorname{tr}(\Sigma^T Y) \\ &= \sup_{|y_i| \leq 1} \sum_{i=1}^{p} \sigma_i y_i = \sum_{i=1}^{p} \sigma_i. \end{aligned}$$

which shows (6.15). Alternatively, we can express (6.14) as the solution to

$$\begin{array}{ll} \text{maximize} & \operatorname{tr}(X^T Z) \\ \text{subject to} & \begin{bmatrix} I & Z^T \\ Z & I \end{bmatrix} \succeq 0, \end{array} \tag{6.16}$$

with the dual problem (see Example 8.8)

$$\begin{array}{ll} \text{minimize} & \operatorname{tr}(U + V)/2 \\ \text{subject to} & \begin{bmatrix} U & X^T \\ X & V \end{bmatrix} \succeq 0. \end{array} \tag{6.17}$$

In other words, using strong duality we can characterize the epigraph $\|A\|_* \leq t$ with

$$\begin{bmatrix} U & A^T \\ A & V \end{bmatrix} \succeq 0, \quad \operatorname{tr}(U + V)/2 \leq t. \tag{6.18}$$

For a symmetric matrix the nuclear norm corresponds to the sum of absolute values of eigenvalues, and for a semidefinite matrix it simply corresponds to the trace of the matrix.

### 6.2.5 Matrix inequalities from Schur's Lemma

Several quadratic or quadratic-over-linear matrix inequalities follow immediately from Schur's lemma. Suppose $A : \mathbb{R}^{m\times p}$ and $B : \mathbb{R}^{p\times p}$ are matrix variables. Then

$$A^T B^{-1} A \preceq C$$

if and only if

$$\begin{bmatrix} C & A^T \\ A & B \end{bmatrix} \succeq 0.$$

### 6.2.6 Nonnegative polynomials

We next consider characterizations of polynomials constrained to be nonnegative on the real axis. To that end, consider a polynomial *basis function*

$$v(t) = \left(1, t, \ldots, t^{2n}\right).$$

It is then well-known that a polynomial $f : \mathbb{R} \mapsto \mathbb{R}$ of even degree $2n$ is nonnegative on the entire real axis

$$f(t) := x^T v(t) = x_0 + x_1 t + \cdots + x_{2n} t^{2n} \geq 0, \ \forall t \tag{6.19}$$

if and only if it can be written as a sum of squared polynomials of degree $n$ (or less), i.e., for some $q_1, q_2 \in \mathbb{R}^{n+1}$

$$f(t) = (q_1^T u(t))^2 + (q_2^T u(t))^2, \qquad u(t) := (1, t, \ldots, t^n). \tag{6.20}$$

It turns out that an equivalent characterization of $\{x \mid x^T v(t) \geq 0, \forall t\}$ can be given in terms of a semidefinite variable $X$,

$$x_i = \langle X, H_i \rangle, \quad i = 0, \ldots, 2n, \qquad X \in \mathcal{S}_+^{n+1}. \tag{6.21}$$

where $H_i^{n+1} \in \mathbb{R}^{(n+1)\times(n+1)}$ are Hankel matrices

$$[H_i]_{kl} = \begin{cases} 1, & k+l = i, \\ 0, & \text{otherwise.} \end{cases}$$

When there is no ambiguity, we drop the superscript on $H_i$. For example, for $n = 2$ we have

$$H_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad H_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \ldots \quad H_4 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

To verify that (6.19) and (6.21) are equivalent, we first note that

$$u(t)u(t)^T = \sum_{i=0}^{2n} H_i v_i(t),$$

i.e.,

$$\begin{bmatrix} 1 \\ t \\ \vdots \\ t^n \end{bmatrix} \begin{bmatrix} 1 \\ t \\ \vdots \\ t^n \end{bmatrix}^T = \begin{bmatrix} 1 & t & \dots & t^n \\ t & t^2 & \dots & t^{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ t^n & t^{n+1} & \dots & t^{2n} \end{bmatrix}.$$

Assume next that $f(t) \geq 0$. Then from (6.20) we have

$$\begin{aligned} f(t) &= (q_1^T u(t))^2 + (q_2^T u(t))^2 \\ &= \langle q_1 q_1^T + q_2 q_2^T, u(t)u(t)^T \rangle \\ &= \textstyle\sum_{i=0}^{2n} \langle q_1 q_1^T + q_2 q_2^T, H_i \rangle v_i(t), \end{aligned}$$

i.e., we have $f(t) = x^T v(t)$ with $x_i = \langle X, H_i \rangle$, $X = (q_1 q_1^T + q_2 q_2^T) \succeq 0$. Conversely, assume that (6.21) holds. Then

$$f(t) = \sum_{i=0}^{2n} \langle H_i, X \rangle v_i(t) = \langle X, \sum_{i=0}^{2n} H_i v_i(t) \rangle = \langle X, u(t)u(t)^T \rangle \geq 0$$

since $X \succeq 0$. In summary, we can characterize the cone of nonnegative polynomials over the real axis as

$$K_\infty^n = \{x \in \mathbb{R}^{n+1} \mid x_i = \langle X, H_i \rangle,\ i = 0, \dots, 2n,\ X \in \mathcal{S}_+^{n+1}\}. \tag{6.22}$$

Checking nonnegativity of a univariate polynomial thus corresponds to a semidefinite feasibility problem.

### Nonnegativity on a finite interval

As an extension we consider a basis function of degree $n$,

$$v(t) = (1, t, \dots, t^n).$$

A polynomial $f(t) := x^T v(t)$ is then nonnegative on a subinterval $I = [a, b] \subset \mathbb{R}$ if and only $f(t)$ can be written as a *sum of weighted squares*,

$$f(t) = w_1(t)(q_1^T u_1(t))^2 + w_2(t)(q_2^T u_2(t))^2$$

where $w_i(t)$ are polynomials nonnegative on $[a, b]$. To describe the cone

$$K_{a,b}^n = \{x \in \mathbb{R}^{n+1} \mid f(t) = x^T v(t) \geq 0,\ \forall t \in [a, b]\}$$

we distinguish between polynomials of odd and even degree.

- *Even degree.* Let $n = 2m$ and denote

$$u_1(t) = (1, t, \dots, t^m), \qquad u_2(t) = (1, t, \dots, t^{m-1}).$$

We choose $w_1(t) = 1$ and $w_2(t) = (t-a)(b-t)$ and note that $w_2(t) \geq 0$ on $[a,b]$. Then $f(t) \geq 0, \forall t \in [a,b]$ if and only if

$$f(t) = (q_1^T u_1(t))^2 + w_2(t)(q_2^T u_2(t))^2$$

for some $q_1, q_2$, and an equivalent semidefinite characterization can be found as

$$\begin{aligned} K_{a,b}^n = \{x \in \mathbb{R}^{n+1} \mid x_i = \langle X_1, H_i^m \rangle + \langle X_2, (a+b)H_{i-1}^{m-1} - abH_i^{m-1} - H_{i-2}^{m-1} \rangle, \\ i = 0, \ldots, n,\ X_1 \in \mathcal{S}_+^m,\ X_2 \in \mathcal{S}_+^{m-1}\}. \end{aligned} \tag{6.23}$$

- *Odd degree.* Let $n = 2m+1$ and denote $u(t) = (1, t, \ldots, t^m)$. We choose $w_1(t) = (t-a)$ and $w_2(t) = (b-t)$. We then have that $f(t) = x^T v(t) \geq 0, \forall t \in [a,b]$ if and only if

  $$f(t) = (t-a)(q_1^T u(t))^2 + (b-t)(q_2^T u(t))^2$$

  for some $q_1, q_2$, and an equivalent semidefinite characterization can be found as

  $$\begin{aligned} K_{a,b}^n = \{x \in \mathbb{R}^{n+1} \mid x_i = \langle X_1, H_{i-1}^m - aH_i^m \rangle + \langle X_2, bH_i^m - H_{i-1}^m \rangle, \\ i = 0, \ldots, n,\ X_1, X_2 \in \mathcal{S}_+^m\}. \end{aligned} \tag{6.24}$$

### 6.2.7 Hermitian matrices

Semidefinite optimization can be extended to complex-valued matrices. To that end, let $\mathcal{H}^n$ denote the cone of Hermitian matrices of order $n$, i.e.,

$$X \in \mathcal{H}^n \qquad \Longleftrightarrow \qquad X \in \mathbb{C}^{n \times n}, \quad X^H = X, \tag{6.25}$$

where superscript '$H$' denotes Hermitian (or complex) transposition. Then $X \in \mathcal{H}_+^n$ if and only if

$$\begin{aligned} z^H X z &= (\Re z - i\Im z)^T (\Re X + i\Im X)(\Re z + i\Im z) \\ &= \begin{bmatrix} \Re z \\ \Im z \end{bmatrix}^T \begin{bmatrix} \Re X & -\Im X \\ \Im X & \Re X \end{bmatrix} \begin{bmatrix} \Re z \\ \Im z \end{bmatrix} \geq 0, \quad \forall z \in \mathbb{C}^n. \end{aligned}$$

In other words,

$$X \in \mathcal{H}_+^n \qquad \Longleftrightarrow \qquad \begin{bmatrix} \Re X & -\Im X \\ \Im X & \Re X \end{bmatrix} \in \mathcal{S}_+^{2n}. \tag{6.26}$$

Note that (6.25) implies skew-symmetry of $\Im X$, i.e., $\Im X = -\Im X^T$.

### 6.2.8 Nonnegative trigonometric polynomials

As a complex-valued variation of the sum-of-squares representation we consider trigonometric polynomials; optimization over cones of nonnegative trigonometric polynomials has several

important engineering applications. Consider a trigonometric polynomial evaluated on the complex unit-circle

$$f(z) = x_0 + 2\Re(\sum_{i=1}^{n} x_i z^{-i}), \qquad |z| = 1 \tag{6.27}$$

parametrized by $x \in \mathbb{R} \times \mathbb{C}^n$. We are interested in characterizing the cone of trigonometric polynomials that are nonnegative on the angular interval $[0, \pi]$,

$$K_{0,\pi}^n = \{x \in \mathbb{R} \times \mathbb{C}^n \mid x_0 + 2\Re(\sum_{i=1}^{n} x_i z^{-i}) \geq 0,\ \forall z = e^{jt}, t \in [0, \pi]\}.$$

Consider a complex-valued basis function

$$v(z) = (1, z, \ldots, z^n).$$

The Riesz-Fejer Theorem states that a trigonometric polynomial $f(z)$ in (6.27) is nonnegative (i.e., $x \in K_{0,\pi}^n$) if and only if for some $q \in \mathbb{C}^{n+1}$

$$f(z) = |q^H v(z)|^2. \tag{6.28}$$

Analogously to Sec. 6.2.6 we have a semidefinite characterization of the sum-of-squares representation, i.e., $f(z) \geq 0,\ \forall z = e^{jt},\ t \in [0, 2\pi]$ if and only if

$$x_i = \langle X, T_i \rangle, \quad i = 0, \ldots, n, \qquad X \in \mathcal{H}_+^{n+1} \tag{6.29}$$

where $T_i^{n+1} \in \mathbb{R}^{(n+1)\times(n+1)}$ are Toeplitz matrices

$$[T_i]_{kl} = \begin{cases} 1, & k - l = i \\ 0, & \text{otherwise} \end{cases}, \quad i = 0, \ldots, n.$$

When there is no ambiguity, we drop the superscript on $T_i$. For example, for $n = 2$ we have

$$T_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad T_1 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad T_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

To prove correctness of the semidefinite characterization, we first note that

$$v(z)v(z)^H = I + \sum_{i=1}^{n} (T_i v_i(z)) + \sum_{i=1}^{n} (T_i v_i(z))^H$$

i.e.,

$$\begin{bmatrix} 1 \\ z \\ \vdots \\ z^n \end{bmatrix} \begin{bmatrix} 1 \\ z \\ \vdots \\ z^n \end{bmatrix}^H = \begin{bmatrix} 1 & z^{-1} & \ldots & z^{-n} \\ z & 1 & \ldots & z^{1-n} \\ \vdots & \vdots & \ddots & \vdots \\ z^n & z^{n-1} & \ldots & 1 \end{bmatrix}.$$

Next assume that (6.28) is satisfied. Then

$$\begin{aligned} f(z) &= \langle qq^H, v(z)v(z)^H\rangle \\ &= \langle qq^H, I\rangle + \langle qq^H, \textstyle\sum_{i=1}^n T_i v_i(z)\rangle + \langle qq^H, \sum_{i=1}^n T_i^T \overline{v_i(z)}\rangle \\ &= \langle qq^H, I\rangle + \textstyle\sum_{i=1}^n \langle qq^H, T_i\rangle v_i(z) + \sum_{i=1}^n \langle qq^H, T_i^T\rangle \overline{v_i(z)} \\ &= x_0 + 2\Re(\textstyle\sum_{i=1}^n x_i v_i(z)) \end{aligned}$$

with $x_i = \langle qq^H, T_i\rangle$. Conversely, assume that (6.29) holds. Then

$$f(z) = \langle X, I\rangle + \sum_{i=1}^n \langle X, T_i\rangle v_i(z) + \sum_{i=1}^n \langle X, T_i^T\rangle \overline{v_i(z)} = \langle X, v(z)v(z)^H\rangle \geq 0.$$

In other words, we have shown that

$$K_{0,\pi}^n = \{x \in \mathbb{R}\times\mathbb{C}^n \mid x_i = \langle X, T_i\rangle,\ i = 0,\ldots,n,\ X \in \mathcal{H}_+^{n+1}\}. \tag{6.30}$$

### Nonnegativity on a subinterval

We next sketch a few useful extensions. An extension of the Riesz-Fejer Theorem states that a trigonometric polynomial $f(z)$ of degree $n$ is nonnegative on $I(a,b) = \{z \mid z = e^{jt},\ t \in [a,b] \subseteq [0,\pi]\}$ if and only if it can be written as a weighted sum of squared trigonometric polynomials

$$f(z) = |f_1(z)|^2 + g(z)|f_2(z)|^2$$

where $f_1, f_2, g$ are trigonemetric polynomials of degree $n, n-d$ and $d$, respectively, and $g(z) \geq 0$ on $I(a,b)$. For example $g(z) = z + z^{-1} - 2\cos\alpha$ is nonnegative on $I(0,\alpha)$, and it can be verified that $f(z) \geq 0,\ \forall z \in I(0,\alpha)$ if and only if

$$x_i = \langle X_1, T_i^{n+1}\rangle + \langle X_2, T_{i+1}^n\rangle + \langle X_2, T_{i-1}^n\rangle - 2\cos(\alpha)\langle X_2, T_i^n\rangle,$$

for $X_1 \in \mathcal{H}_+^{n+1}$, $X_2 \in \mathcal{H}_+^n$, i.e.,

$$\begin{aligned} K_{0,\alpha}^n = \{x \in \mathbb{R}\times\mathbb{C}^n \mid x_i = \langle X_1, T_i^{n+1}\rangle + \langle X_2, T_{i+1}^n\rangle + \langle X_2, T_{i-1}^n\rangle \\ -2\cos(\alpha)\langle X_2, T_i^n\rangle,\ X_1 \in \mathcal{H}_+^{n+1}, X_2 \in \mathcal{H}_+^n\}. \end{aligned} \tag{6.31}$$

Similarly $f(z) \geq 0,\ \forall z \in I(\alpha,\pi)$ if and only if

$$x_i = \langle X_1, T_i^{n+1}\rangle + \langle X_2, T_{i+1}^n\rangle + \langle X_2, T_{i-1}^n\rangle - 2\cos(\alpha)\langle X_2, T_i^n\rangle$$

i.e.,

$$\begin{aligned} K_{\alpha,\pi}^n = \{x \in \mathbb{R}\times\mathbb{C}^n \mid x_i = \langle X_1, T_i^{n+1}\rangle + \langle X_2, T_{i+1}^n\rangle + \langle X_2, T_{i-1}^n\rangle \\ +2\cos(\alpha)\langle X_2, T_i^n\rangle,\ X_1 \in \mathcal{H}_+^{n+1}, X_2 \in \mathcal{H}_+^n\}. \end{aligned} \tag{6.32}$$

## 6.3 Semidefinite optimization case studies

### 6.3.1 Nearest correlation matrix

We consider the set

$$S = \{X \in \mathcal{S}^n_+ \mid X_{ii} = 1,\ i = 1, \ldots, n\}$$

(shown in Fig. 6.1 for $n = 3$). For $A \in \mathcal{S}^n$ the nearest correlation matrix is

$$X^\star = \arg\min_{X \in S} \|A - X\|_F,$$

i.e., the projection of $A$ onto the set $S$. To pose this as a conic optimization we define the linear operator

$$\mathbf{svec}(U) = (U_{11}, \sqrt{2}U_{21}, \ldots, \sqrt{2}U_{n1}, U_{22}, \sqrt{2}U_{32}, \ldots, \sqrt{2}U_{n2}, \ldots, U_{nn}),$$

which extracts and scales the lower-triangular part of $U$. We then get a conic formulation of the nearest correlation problem exploiting symmetry of $A - X$,

$$\begin{array}{lrl} \text{minimize} & t & \\ \text{subject to} & \|z\|_2 & \leq t, \\ & \mathbf{svec}(A - X) & = z, \\ & \mathbf{diag}(X) & = e, \\ & X & \succeq 0. \end{array} \tag{6.33}$$

This is an example of a problem with both conic quadratic and semidefinite constraints in *primal form*. We can add different constraints to the problem, for example a bound $\gamma$ on the smallest eigenvalue by replacing $X \succeq 0$ with $X \succeq \gamma I$.

### 6.3.2 Extremal ellipsoids

Given a polytope we can find the largest ellipsoid contained in the polytope, or the smallest ellipsoid containing the polytope (for certain representations of the polytope).

#### Maximal inscribed ellipsoid

Consider a polytope

$$S = \{x \in \mathbb{R}^n \mid a_i^T x \leq b_i,\ i = 1, \ldots, m\}.$$

The ellipsoid

$$\mathcal{E} := \{x \mid x = Cu + d,\ \|u\| \leq 1\}$$

is contained in $S$ if and only if

$$\max_{\|u\|_2 \leq 1} a_i^T (Cu + d) \leq b_i, \quad i = 1, \ldots, m$$

or equivalently, if and only if

$$\|Ca_i\|_2 + a_i^T d \leq b_i, \quad i = 1, \ldots, m.$$

Since $\mathbf{Vol}(\mathcal{E}) \approx \det(C)^{1/n}$ the maximum-volume inscribed ellipsoid is the solution to

$$\begin{array}{ll} \text{maximize} & \det(C) \\ \text{subject to} & \|Ca_i\|_2 + a_i^T d \leq b_i, \quad i = 1, \ldots, m, \\ & C \succeq 0. \end{array}$$

In Sec. 6.2.2 we show how to maximize the determinant of a positive definite matrix.

### Minimal enclosing ellipsoid

Next consider a polytope given as the convex hull of a set of points,

$$S' = \mathbf{conv}\{x_1, x_2, \ldots, x_m\}, \quad x_i \in \mathbb{R}^n.$$

The ellipsoid

$$\mathcal{E}' := \{x \mid \|P(x - c)\|_2 \leq 1\}$$

has $\mathbf{Vol}(\mathcal{E}') \approx \det(P)^{-1/n}$, so the minimum-volume enclosing ellipsoid is the solution to

$$\begin{array}{ll} \text{maximize} & \det(P) \\ \text{subject to} & \|P(x_i - c)\|_2 \leq 1, \quad i = 1, \ldots, m, \\ & P \succeq 0. \end{array}$$

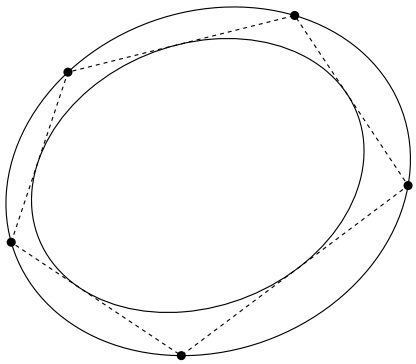


Fig. 6.2: Example of inner and outer ellipsoidal approximations of a pentagon in $\mathbb{R}^2$.

## 6.3.3 Polynomial curve-fitting

Consider a univariate polynomial of degree $n$,

$$f(t) = x_0 + x_1 t + x_2 t^2 + \cdots + x_n t^n.$$

Often we wish to fit such a polynomial to a given set of measurements or control points

$$\{(t_1, y_1), (t_2, y_2), \ldots, (t_m, y_m)\},$$

i.e., we wish to determine coefficients $x_i$, $i = 0, \ldots, n$ such that

$$f(t_j) \approx y_j, \quad j = 1, \ldots, m.$$

To that end, define the *Vandermonde* matrix

$$A = \begin{bmatrix} 1 & t_1 & t_1^2 & \ldots & t_1^n \\ 1 & t_2 & t_2^2 & \ldots & t_2^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_m & t_m^2 & \ldots & t_m^n \end{bmatrix}.$$

We can then express the desired curve-fit compactly as

$$Ax \approx y,$$

i.e., as a linear expression in the coefficients $x$. When the degree of the polynomial equals the number measurements, $n = m$, the matrix $A$ is square and non-singular (provided there are no duplicate rows), so we can can solve

$$Ax = y$$

to find a polynomial that passes through all the control points $(t_i, y_i)$. Similarly, if $n > m$ there are infinitely many solutions satisfying the *underdetermined* system $Ax = y$. A typical choice in that case is the *least-norm* solution

$$x_{\text{ln}} = \arg\min_{Ax=y} \|x\|_2$$

which (assuming again there are no duplicate rows) equals

$$x_{\text{ln}} = A^T(AA^T)^{-1}y.$$

On the other hand, if $n < m$ we generally cannot find a solution to the *overdetermined* system $Ax = y$, and we typically resort to a *least-squares* solution

$$x_{\text{ls}} = \arg\min \|Ax - y\|_2$$

which is given by the formula (see Sec. 8.5.1)

$$x_{\text{ls}} = (A^TA)^{-1}A^Ty.$$

In the following we discuss how the semidefinite characterizations of nonnegative polynomials (see Sec. 6.2.6) lead to more advanced and useful polynomial curve-fitting constraints.

- *Nonnegativity.* One possible constraint is nonnegativity on an interval,

$$f(t) := x_0 + x_1 t + \cdots + x_n t^n \geq 0, \, \forall t \in [a, b]$$

with a semidefinite characterization embedded in $x \in K^n_{a,b}$, see (6.23).

- *Monotonicity.* We can ensure monotonicity of $f(t)$ by requiring that $f'(t) \geq 0$ (or $f'(t) \leq 0$), i.e.,

$$(x_1, 2x_2, \ldots, nx_n) \in K^{n-1}_{a,b},$$

or

$$-(x_1, 2x_2, \ldots, nx_n) \in K^{n-1}_{a,b},$$

respectively.

- *Convexity or concavity.* Convexity (or concavity) of $f(t)$ corresponds to $f''(t) \geq 0$ (or $f''(t) \leq 0$), i.e.,

$$(2x_2, 6x_3, \ldots, (n-1)nx_n) \in K^{n-2}_{a,b},$$

or

$$-(2x_2, 6x_3, \ldots, (n-1)nx_n) \in K^{n-2}_{a,b},$$

respectively.

As an example, we consider fitting a smooth polynomial

$$f_n(t) = x_0 + x_1 t + \cdots + x_n t^n$$

to the points $\{(-1, 1), (0, 0), (1, 1)\}$, where smoothness is implied by bounding $|f_n'(t)|$. More specifically, we wish to solve the problem

$$\begin{array}{ll} \text{minimize} & z \\ \text{subject to} & |f_n'(t)| \leq z, \quad \forall t \in [-1, 1] \\ & f_n(-1) = 1, \quad f_n(0) = 0, \quad f_n(1) = 1, \end{array}$$

or equivalently

$$\begin{array}{ll} \text{minimize} & z \\ \text{subject to} & z - f_n'(t) \geq 0, \quad \forall t \in [-1, 1] \\ & f_n'(t) - z \geq 0, \quad \forall t \in [-1, 1] \\ & f_n(-1) = 1, \quad f_n(0) = 0, \quad f_n(1) = 1. \end{array}$$

Finally, we use the characterizations $K^n_{a,b}$ to get a conic problem

$$\begin{array}{ll} \text{minimize} & z \\ \text{subject to} & (z - x_1, -2x_2, \ldots, -nx_n) \in K^{n-1}_{-1,1} \\ & (x_1 - z, 2x_2, \ldots, nx_n) \in K^{n-1}_{-1,1} \\ & f_n(-1) = 1, \quad f_n(0) = 0, \quad f_n(1) = 1. \end{array}$$

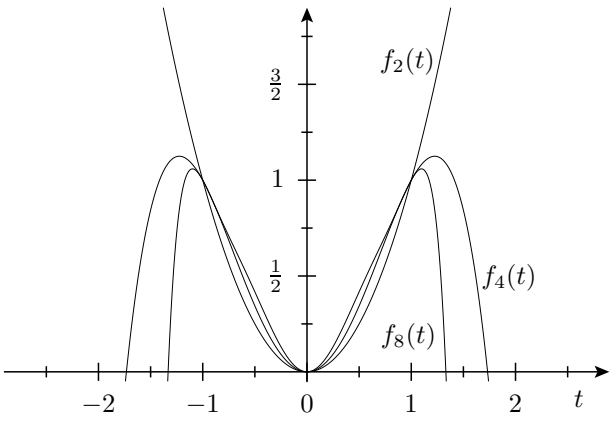



Fig. 6.3: Graph of univariate polynomials of degree 2, 4, and 8, respectively, passing through $\{(-1,1),(0,0),(1,1)\}$. The higher-degree polynomials are increasingly smoother on $[-1,1]$.

In Fig. 6.3 we show the graphs for the resulting polynomails of degree 2, 4 and 8, respectively. The second degree polynomial is uniquely determined by the three constraints $f_2(-1) = 1$, $f_2(0) = 0$, $f_2(1) = 1$, i.e., $f_2(t) = t^2$. Also, we obviously have a lower bound on the largest derivative $\max_{t\in[-1,1]} |f_n'(t)| \geq 1$. The computed fourth degree polynomial is given by

$$f_4(t) = \frac{3}{2}t^2 - \frac{1}{2}t^4$$

after rounding coefficients to rational numbers. Furthermore, the largest derivative is given by

$$f_4'(1/\sqrt{2}) = \sqrt{2},$$

and $f_4''(t) < 0$ on $(1/\sqrt{2}, 1]$ so, although not visibly clear, the polynomial is nonconvex on $[-1,1]$. In Fig. 6.4 we show the graphs of the corresponding polynomials where we added a convexity constraint $f_n''(t) \geq 0$, i.e.,

$$(2x_2, 6x_3, \ldots, (n-1)nx_n) \in K_{-1,1}^{n-2}.$$

In this case, we get

$$f_4(t) = \frac{6}{5}t^2 - \frac{1}{5}t^4$$

and the largest derivative increases to $\frac{8}{5}$.

### 6.3.4 Filter design problems

Filter design is an important application of optimization over trigonometric polynomials in signal processing. We consider a trigonometric polynomial

$$\begin{aligned} H(\omega) &= x_0 + 2\Re(\textstyle\sum_{k=1}^n x_k e^{-j\omega k}) \\ &= a_0 + 2\textstyle\sum_{k=1}^n (a_k \cos(\omega k) + b_k \sin(\omega k)) \end{aligned}$$

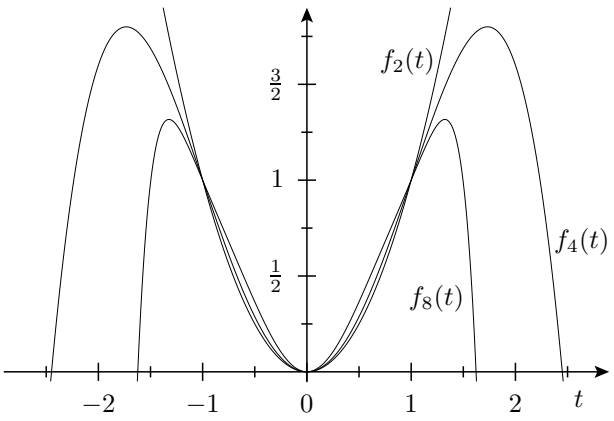



Fig. 6.4: Graph of univariate polynomials of degree $2, 4$, and $8$, respectively, passing through $\{(-1,1),(0,0),(1,1)\}$. The polynomials all have positive second derivative (i.e., they are convex) on $[-1,1]$ and the higher-degree polynomials are increasingly smoother on that interval.

where $a_k := \Re(x_k)$, $b_k := \Im(x_k)$. If the function $H(\omega)$ is nonnegative we call it a transfer function, and it describes how different harmonic components of a periodic discrete signal are attenuated when a filter with transfer function $H(\omega)$ is applied to the signal.

We often wish a transfer function where $H(\omega) \approx 1$ for $0 \leq \omega \leq \omega_p$ and $H(\omega) \approx 0$ for $\omega_s \leq \omega \leq \pi$ for given constants $\omega_p, \omega_s$. One possible formulation for achieving this is

$$\begin{array}{lrcccl} \text{minimize} & & & t & & \\ \text{subject to} & 0 & \leq & H(\omega) & & \forall \omega \in [0,\pi] \\ & 1-\delta & \leq & H(\omega) & \leq \quad 1+\delta & \forall \omega \in [0,\omega_p] \\ & & & H(\omega) & \leq \quad t & \forall \omega \in [\omega_s,\pi], \end{array}$$

which corresponds to minimizing $H(w)$ on the interval $[\omega_s, \pi]$ while allowing $H(w)$ to depart from unity by a small amount $\delta$ on the interval $[0, \omega_p]$. Using the results from Sec. 6.2.8 (in particular (6.30), (6.31) and (6.32)), we can pose this as a conic optimization problem

$$\begin{array}{lrcl} \text{minimize} & t & & \\ \text{subject to} & x & \in & K^n_{0,\pi}, \\ & (x_0 - (1-\delta), x_{1:n}) & \in & K^n_{0,\omega_p}, \\ & -(x_0 - (1+\delta), x_{1:n}) & \in & K^n_{0,\omega_p}, \\ & -(x_0 - t, x_{1:n}) & \in & K^n_{\omega_s,\pi}, \end{array} \tag{6.34}$$

which is a semidefinite optimization problem. In Fig. 6.5 we show $H(\omega)$ obtained by solving (6.34) for $n = 10$, $\delta = 0.05$, $\omega_p = \pi/4$ and $\omega_s = \omega_p + \pi/8$.

### 6.3.5 Relaxations of binary optimization

Semidefinite optimization is also useful for computing bounds on difficult non-convex or combinatorial optimization problems. For example consider the binary linear optimization

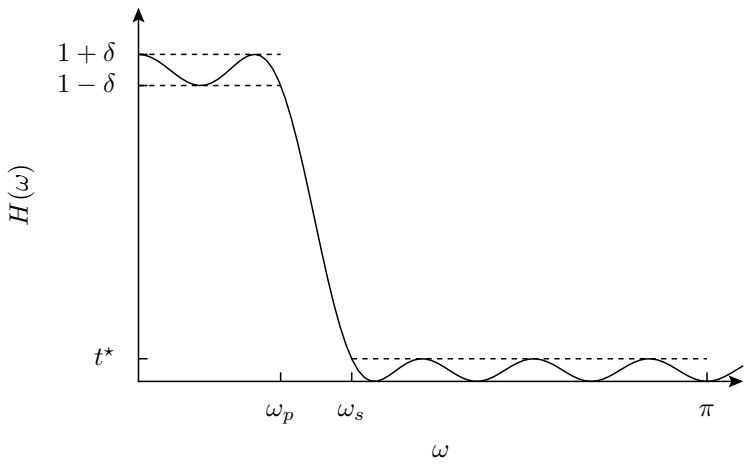



Fig. 6.5: Plot of lowpass filter transfer function.

problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \\ & x \in \{0,1\}^n. \end{array} \tag{6.35}$$

In general, problem (6.35) is a very difficult non-convex problem where we have to explore $2^n$ different objectives. Alternatively, we can use semidefinite optimization to get a lower bound on the optimal solution with polynomial complexity. We first note that

$$x_i \in \{0,1\} \quad \Longleftrightarrow \quad x_i^2 = x_i,$$

which is, in fact, equivalent to a rank constraint on a semidefinite variable,

$$X = xx^T, \qquad \mathbf{diag}(X) = x.$$

By relaxing the rank 1 constraint on $X$ we get a semidefinite relaxation of (6.35),

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \\ & \mathbf{diag}(X) = x, \\ & X \succeq xx^T, \end{array} \tag{6.36}$$

where we note that

$$X \succeq xx^T \quad \Longleftrightarrow \quad \begin{bmatrix} 1 & x^T \\ x & X \end{bmatrix} \succeq 0.$$

Since (6.36) is a semidefinite optimization problem, it can be solved very efficiently. Suppose $x^\star$ is an optimal solution for (6.35); then $(x^\star, x^\star(x^\star)^T)$ is also feasible for (6.36), but the feasible set for (6.36) is larger than the feasible set for (6.35), so in general the optimal solution of (6.36) serves as a lower bound. However, if the optimal solution $X^\star$ of (6.36) has

rank 1 we have found a solution to (6.35) also. The semidefinite relaxation can also be used in a branch-bound mixed-integer exact algorithm for (6.35).

We can tighten (or improve) the relaxation (6.36) by adding other constraints that cut away parts of the feasible set, without excluding rank 1 solutions. By tightening the relaxation, we reduce the gap between the optimal values of the original problem and the relaxation. For example, we can add the constraints

$$\begin{array}{l} 0 \leq X_{ij} \leq 1, \quad i = 1, \ldots, n, \; j = 1, \ldots, n, \\ X_{ii} \geq X_{ij}, \quad i = 1, \ldots, n, \; j = 1, \ldots, n, \end{array}$$

and so on. This will usually have a dramatic impact on solution times and memory requirements. Already constraining a semidefinite matrix to be *doubly nonnegative* ($X_{ij} \geq 0$) introduces additional $n^2$ linear inequality constraints.

### 6.3.6 Relaxations of boolean optimization

Similarly to Sec. 6.3.5 we can use semidefinite relaxations of boolean constraints $x \in \{-1, +1\}^n$. To that end, we note that

$$x \in \{-1, +1\}^n \quad \Longleftrightarrow \quad X = xx^T, \quad \mathbf{diag}(X) = e, \tag{6.37}$$

with a semidefinite relaxation $X \succeq xx^T$ of the rank-1 constraint.

As a (standard) example of a combinatorial problem with boolean constraints, we consider an undirected graph $G = (V, E)$ described by a set of vertices $V = \{v_1, \ldots, v_n\}$ and a set of edges $E = \{(v_i, v_j) \mid v_i, v_j \in V, \; i \neq j\}$, and we wish to find the cut of maximum capacity. A cut $C$ partitions the nodes $V$ into two disjoint sets $S$ and $T = V \setminus S$, and the cut-set $I$ is the set of edges with one node in $S$ and another in $T$, i.e.,

$$I = \{(u, v) \in E \mid u \in S, \; v \in T\}.$$

The capacity of a cut is then defined as the number of edges in the cut-set, $|I|$.

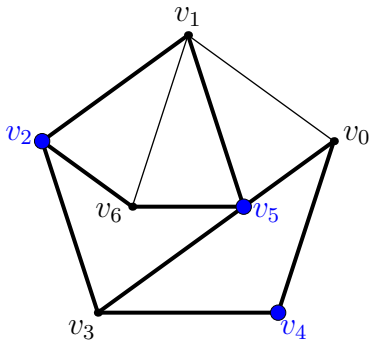



Fig. 6.6: Undirected graph. The cut $\{v_2, v_4, v_5\}$ has capacity 9 (thick edges).

To maximize the capacity of a cut we define the symmetric *adjacency* matrix $A \in \mathcal{S}^n$,

$$[A]_{ij} = \begin{cases} 1, & (v_i, v_j) \in E, \\ 0, & \text{otherwise}, \end{cases}$$

where $n = |V|$, and let

$$x_i = \begin{cases} +1, & v_i \in S, \\ -1, & v_i \notin S. \end{cases}$$

Suppose $v_i \in S$. Then $1 - x_i x_j = 0$ if $v_j \in S$ and $1 - x_i x_j = 2$ if $v_j \notin S$, so we get an expression for the capacity as

$$c(x) = \frac{1}{4} \sum_{ij} (1 - x_i x_j) A_{ij} = \frac{1}{4} e^T A e - \frac{1}{4} x^T A x,$$

and discarding the constant term $e^T A e$ gives us the MAX-CUT problem

$$\begin{array}{ll} \text{minimize} & x^T A x \\ \text{subject to} & x \in \{-1, +1\}^n, \end{array} \tag{6.38}$$

with a semidefinite relaxation

$$\begin{array}{ll} \text{minimize} & \langle A, X \rangle \\ \text{subject to} & \mathbf{diag}(X) = e, \\ & X \succeq 0. \end{array} \tag{6.39}$$

# Chapter 7

# Practical optimization

In this chapter we discuss various practical aspects of creating optimization models.

## 7.1 Conic reformulations

Many nonlinear constraint families were reformulated to conic form in previous chapters, and these examples are often sufficient to construct conic models for the optimization problems at hand. In case you do need to go beyond the cookbook examples, however, a few composition rules are worth knowing to guarantee correctness of your reformulation.

### 7.1.1 Nested nonlinearity

In representing nonlinear inequalities with composites, such as $\exp(\|x\|)$ or $x^4 = (x^2)^2$, it is often helpful to extract the nested nonlinearity into a separate constraint. That is, we would like to express $f(g(x))$ as $f(r)$ for a new artificial variable $r$ constrained as $r = g(x)$. The last constraint is usually not convex (unless $g$ is linear), so we need to derive an equivalent formulation using a relaxed inequality $r \geq g(x)$ or $r \leq g(x)$ instead.

Given an inequality $t \geq f(g(x))$, the nested function $g(x)$ can be replaced by a new artificial variable $r$ if either holds:

1. $g(x)$ is convex and $f(r)$ is convex and nondecreasing on the image of $g$. Then $t \geq f(g(x))$ is equivalent to $t \geq f(r),\ r \geq g(x)$.
2. $g(x)$ is concave and $f(r)$ is convex and nonincreasing on the image of $g$. Then $t \geq f(g(x))$ is equivalent to $t \geq f(r),\ r \leq g(x)$.

In case $g(x)$ is affine, that is $g(x) = a^T x + b$, it can always be extracted using $r = a^T x + b$.

**Example 7.1.** Here are some simple substitutions.

- The inequality $t \geq \exp(g(x))$ can be represented as $t \geq \exp(r)$ (using an exponential cone) with $r \geq g(x)$ for all convex functions $g(x)$. This is intuitive as we expect solvers to push $r \to -\infty$ in attempt to relax $t \geq \exp(r)$, why only the lower bound on $r$ is needed to prevent it.

- The inequality $2st \geq g(x)^2$ can be represented as $2st \geq r^2$ (a rotated quadratic cone) with $r \geq g(x)$ if $g(x)$ is nonnegative convex. This is intuitive as we expect solvers to push $r \to 0$ in attempt to relax $2st \geq r^2$, why only the lower bound on $r$ is needed if $r = g(x)$ is nonnegative.
- If we try to express $2st \geq \exp(x)$ as $2st \geq r \geq \exp(x)$ then we fail because the first constraint $2st \geq r$ is not convex. Nevertheless, we may restate it as $2st \geq \exp(x/2)^2$ and subsequently split into $2st \geq r^2$ and $r \geq \exp(x/2)$ following the previous example.

**Example 7.2.** As in the last example, it may require some work to find the correct reformulation of a nested nonlinear constraint. For instance, suppose that for $x, y, z \geq 0$ with $xyz > 1$ we want to write

$$t \geq \frac{1}{xyz - 1}.$$

A natural first attempt is:

$$t \geq \frac{1}{r}, \quad r \leq xyz - 1,$$

that is, we try to apply the decomposition (2) with $f(r) = 1/r$ and $g(x, y, z) = xyz - 1$. The function $f$ is indeed convex and nonincreasing on $r > 0$ and the inequality $tr \geq 1$ is representable with a rotated quadratic cone. Unfortunately $g$ is not concave. We know that a monomial like $xyz$ appears in connection with the power cone, but that requires a homogenous constraint such as $xyz \geq u^3$. This gives us an idea to try

$$t \geq \frac{1}{r^3}, \quad r^3 \leq xyz - 1,$$

that is $f(r) = 1/r^3$ and $g(x, y, z) = (xyz - 1)^{1/3}$. This provides the right balance: all conditions in (2) are satisfied. Introducing another variable $u$ we get the following model:

$$tr^3 \geq 1, \quad xyz \geq u^3, \quad u \geq (r^3 + 1^3)^{1/3},$$

We refer to Sec. 4 to verify that all the constraints above are representable using power cones. We leave it as an exercise to find other conic representations, based on other transformations of the original inequality.

### 7.1.2 Convex univariate piecewise-defined functions

Consider a univariate function with $k$ pieces:

$$f(x) = \begin{cases} f_1(x) & \text{if } x \leq \alpha_1, \\ f_i(x) & \text{if } \alpha_{i-1} \leq x \leq \alpha_i, \text{ for } i = 2, \ldots, k-1, \\ f_k(x) & \text{if } \alpha_{k-1} \leq x, \end{cases}$$

Suppose $f(x)$ is convex (in particular each piece is convex by itself) and $f_i(\alpha_i) = f_{i+1}(\alpha_i)$. In representing the epigraph $t \geq f(x)$ it is helpful to proceed via the equivalent representation:

$$\begin{array}{l} t = \sum_{i=1}^{k} t_i - \sum_{i=1}^{k-1} f_i(\alpha_i), \\ x = \sum_{i=1}^{k} x_i - \sum_{i=1}^{k-1} \alpha_i, \\ t_i \geq f_i(x_i) \text{ for } i = 1, \ldots, k, \\ x_1 \leq \alpha_1, \;\; \alpha_{i-1} \leq x_i \leq \alpha_i \text{ for } i = 2, \ldots, k-1, \;\; \alpha_{k-1} \leq x_k. \end{array}$$

*Proof.* In the special case when $k = 2$ and $\alpha_1 = f_1(\alpha_1) = f_2(\alpha_1) = 0$ the epigraph of $f$ is equal to the Minkowski sum $\{(t_1 + t_2, x_1 + x_2) \;:\; t_i \geq f_i(x_i)\}$. In general the epigraph over two consecutive pieces is obtained by shifting them to $(0, 0)$, computing the Minkowski sum and shifting back. Finally more than two pieces can be joined by continuing this argument by induction. □

As the reformulation grows in size with the number of pieces, it is preferable to keep this number low. Trivially, if $f_i(x) = f_{i+1}(x)$, these two pieces can be merged. Substituting $f_i(x)$ and $f_{i+1}(x)$ for $\max(f_i(x), f_{i+1}(x))$ is sometimes an invariant change to facilitate this merge. For instance, it always works for affine functions $f_i(x)$ and $f_{i+1}(x)$. Finally, if $f(x)$ is symmetric around some point $\alpha$, we can represent its epigraph via a piecewise function, with only half the number of pieces, defined in terms of a new variable $z = |x - \alpha| + \alpha$.

**Example 7.3** (Huber loss function)**.** The Huber loss function

$$f(x) = \begin{cases} -2x - 1 & \text{if } x \leq -1, \\ x^2 & \text{if } -1 \leq x \leq 1, \\ 2x - 1 & \text{if } 1 \leq x, \end{cases}$$

is convex and its epigraph $t \geq f(x)$ has an equivalent representation

$$\begin{array}{l} t \geq t_1 + t_2 + t_3 - 2, \\ x = x_1 + x_2 + x_3, \\ t_1 = -2x_1 - 1, \;\; t_2 \geq x_2^2, \;\; t_3 = 2x_3 - 1, \\ x_1 \leq -1, \;\; -1 \leq x_2 \leq 1, \;\; 1 \leq x_3, \end{array}$$

where $t_2 \geq x_2^2$ is representable by a rotated quadratic cone. No two pieces of $f(x)$ can be merged to reduce the size of this formulation, but the loss function does satisfy a simple

symmetry; namely $f(x) = f(-x)$. We can thus represent its epigraph by $t \geq \hat{f}(z)$ and $z \geq |x|$, where

$$\hat{f}(z) = \begin{cases} z^2 & \text{if } z \leq 1, \\ 2z - 1 & \text{if } 1 \leq z. \end{cases}$$

In this particular example, however, unless the absolute value $z$ from $z \geq |x|$ is used elsewhere, the cost of introducing it does not outweigh the savings achieved by going from three pieces in $x$ to two pieces in $z$.

## 7.2 Avoiding ill-posed problems

For a well-posed continuous problem a small change in the input data should induce a small change of the optimal solution. A problem is *ill-posed* if small perturbations of the problem data result in arbitrarily large perturbations of the solution, or even change the feasibility status of the problem. In such cases small rounding errors, or solving the problem on a different computer can result in different or wrong solutions.

In fact, from an algorithmic point of view, even computing a *wrong* solution is numerically difficult for ill-posed problems. A rigorous definition of the degree of ill-posedness is possible by defining a *condition number*. This is an attractive, but not very practical metric, as its evaluation requires solving several auxiliary optimization problems. Therefore we only make the modest recommendations to avoid problems

- that are nearly infeasible,
- with constraints that are linearly dependent, or nearly linearly dependent.

**Example 7.4** (Near linear dependencies)**.** To give some idea about the perils of near linear dependence, consider this pair of optimization problems:

$$\begin{array}{lrcl} \text{maximize} & x & & \\ \text{subject to} & 2x - y & \geq & -1, \\ & 2.0001x - y & \leq & 0, \end{array}$$

and

$$\begin{array}{lrcl} \text{maximize} & x & & \\ \text{subject to} & 2x - y & \geq & -1, \\ & 1.9999x - y & \leq & 0. \end{array}$$

The first problem has a unique optimal solution $(x, y) = (10^4, 2 \cdot 10^4 + 1)$, while the second problem is unbounded. This is caused by the fact that the hyperplanes (here: straight

lines) defining the constraints in each problem are almost parallel. Moreover, we can consider another modification:

$$\begin{array}{lrcl} \text{maximize} & x & & \\ \text{subject to} & 2x - y & \geq & -1, \\ & 2.001x - y & \leq & 0, \end{array}$$

with optimal solution $(x, y) = (10^3, 2 \cdot 10^3 + 1)$, which shows how a small perturbation of the coefficients induces a large change of the solution.

Typical examples of problems with nearly linearly dependent constraints are discretizations of continuous processes, where the constraints invariably become more correlated as we make the discretization finer; as such there may be nothing wrong with the discretization or problem formulation, but we should expect numerical difficulties for sufficiently fine discretizations.

One should also be careful not to specify problems whose optimal value is only achieved in the limit. A trivial example is

$$\text{minimize } e^x, \ x \in \mathbb{R}.$$

The infimum value 0 is not attained for any finite $x$, which can lead to unspecified behaviour of the solver. More examples of ill-posed conic problems are discussed in Fig. 7.1 and Sec. 8. In particular, Fig. 7.1 depicts some troublesome problems in two dimensions. In (a) minimizing $y \geq 1/x$ on the nonnegative orthant is unattained approaching zero for $x \to \infty$. In (b) the only feasible point is $(0, 0)$, but objectives maximizing $x$ are not blocked by the purely vertical normal vectors of active constraints at this point, falsely suggesting local progress could be made. In (c) the intersection of the two subsets is empty, but the distance between them is zero. Finally in (d) minimizing $x$ on $y \geq x^2$ is unbounded at minus infinity for $y \to \infty$, but there is no improving ray to follow. Although seemingly unrelated, the four cases are actually primal-dual pairs; (a) with (b) and (c) with (d). In fact, the missing normal vector property in (b)—desired to certify optimality—can be attributed to (a) not attaining the best among objective values at distance zero to feasibility, and the missing positive distance property in (c)—desired to certify infeasibility—is because (d) has no improving ray.

## 7.3 Scaling

Another difficulty encountered in practice involves models that are badly scaled. Loosely speaking, we consider a model to be badly scaled if

- variables are measured on very different scales,
- constraints or bounds are measured on very different scales.

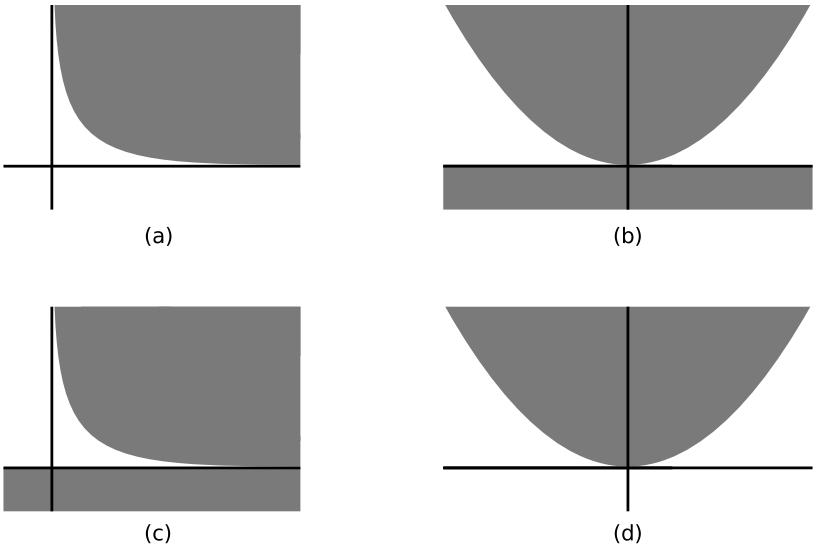



Fig. 7.1: Geometric representations of some ill-posed situations.

For example if one variable $x_1$ has units of molecules and another variable $x_2$ measures temperature, we might expect an objective function such as

$$x_1 + 10^{12} x_2$$

and in finite precision the second term dominates the objective and renders the contribution from $x_1$ insignificant and unreliable.

A similar situation (from a numerical point of view) is encountered when using *penalization* or *big-M* strategies. Assume we have a standard linear optimization problem (2.12) with the additional constraint that $x_1 = 0$. We may eliminate $x_1$ completely from the model, or we might add an additional constraint, but suppose we choose to formulate a penalized problem instead

$$\begin{array}{ll} \text{minimize} & c^T x + 10^{12} x_1 \\ \text{subject to} & Ax = b, \\ & x \geq 0, \end{array}$$

reasoning that the large penalty term will force $x_1 = 0$. However, if $\|c\|$ is small we have the exact same problem, namely that in finite precision the penalty term will completely dominate the objective and render the contribution $c^T x$ insignificant or unreliable. Therefore, the penalty term should be chosen carefully.

**Example 7.5** (Risk-return tradeoff)**.** Suppose we want to solve an efficient frontier variant of the optimal portfolio problem from Sec. 3.3.3 with an objective of the form

$$\text{maximize } \mu^T x - \frac{1}{2} \gamma x^T \Sigma x$$

where the parameter $\gamma$ controls the tradeoff between return and risk. The user might choose a very small $\gamma$ to discount the risk term. This, however, may lead to numerical

issues caused by a very large solution. To see why, consider a simple one-dimensional, unconstrained analogue:

$$\text{maximize } x - \frac{1}{2}\gamma x^2, \ x \in \mathbb{R}$$

and note that the maximum is attained for $x = 1/\gamma$, which may be very large.

**Example 7.6** (Explicit redundant bounds)**.** Consider again the problem (2.12), but with additional (redundant) constraints $x_i \leq \gamma$. This is a common approach for some optimization practitioners. The problem we solve is then

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \\ & x \geq 0, \\ & x \leq \gamma e, \end{array}$$

with a dual problem

$$\begin{array}{ll} \text{maximize} & b^T y - \gamma e^T z \\ \text{subject to} & A^T y + s - z = c, \\ & s, z \geq 0. \end{array}$$

Suppose we do not know *a-priori* an upper bound on $\|x\|_\infty$, so we choose a very large $\gamma = 10^{12}$ reasoning that this will not change the optimal solution. Note that the large variable bound becomes a penalty term in the dual problem; in finite precision such a large bound will effectively destroy accuracy of the solution.

**Example 7.7** (Big-M)**.** Suppose we have a mixed integer model which involves a big-M constraint (see Sec. 9), for instance

$$a^T x \leq b + M(1 - z)$$

as in Sec. 9.1.3. The constant $M$ must be big enough to guarantee that the constraint becomes redundant when $z = 0$, or we risk shrinking the feasible set, perhaps to the point of creating an infeasible model. On the other hand, the user might set $M$ to a huge, safe value, say $M = 10^{12}$. While this is mathematically correct, it will lead to a model with poorer linear relaxations and most likely increase solution time (sometimes dramatically). It is therefore very important to put some effort into estimating a reasonable value of $M$ which is safe, but not too big.

## 7.4 The huge and the tiny

Some types of problems and constraints are especially prone to behave badly with very large or very small numbers. We discuss such examples briefly in this section.

### Sum of squares

The sum of squares $x_1^2 + \cdots + x_n^2$ can be bounded above using the rotated quadratic cone:

$$t \geq x_1^2 + \cdots + x_n^2 \iff (\frac{1}{2}, t, x) \in \mathcal{Q}_r^n,$$

but in most cases it is better to bound the square root with a quadratic cone:

$$t' \geq \sqrt{x_1^2 + \cdots + x_n^2} \iff (t', x) \in \mathcal{Q}^n.$$

The latter has slightly better numerical properties: $t'$ is roughly comparable with $x$ and is measured on the same scale, while $t$ grows as the square of $x$ which will sooner lead to numerical instabilities.

### Exponential cone

Using the function $e^x$ for $x \leq -30$ or $x \geq 30$ would be rather questionable if $e^x$ is supposed to represent any realistic value. Note that $e^{30} \approx 10^{13}$ and that $e^{30.0000001} - e^{30} \approx 10^6$, so a tiny perturbation of the exponent produces a huge change of the result. Ideally, $x$ should have a single-digit absolute value. For similar reasons, it is even more important to have well-scaled data. For instance, in floating point arithmetic we have the "equality"

$$e^{20} + e^{-20} = e^{20}$$

since the smaller summand disappears in comparison to the other one, $10^{17}$ times bigger.

For the same reason it is advised to replace explicit inequalities involving $\exp(x)$ with log-sum-exp variants (see Sec. 5.2.6). For example, suppose we have a constraint such as

$$t \geq e^{x_1} + e^{x_2} + e^{x_3}.$$

Then after a substitution $t = e^{t'}$ we have

$$1 \geq e^{x_1 - t'} + e^{x_2 - t'} + e^{x_3 - t'}$$

which has the advantage that $t'$ is of the same order of magnitude as $x_i$, and that the exponents $x_i - t'$ are likely to have much smaller absolute values than simply $x_i$.

### Power cone

The power cone is not reliable when one of the exponents is very small. For example, consider the function $f(x) = x^{0.01}$, which behaves almost like the indicator function of $x > 0$ in the sense that

$$f(0) = 0, \text{ and } f(x) > 0.5 \text{ for } x > 10^{-30}.$$

Now suppose $x = 10^{-20}$. Is the constraint $f(x) > 0.5$ satisfied? In principle yes, but in practice $x$ could have been obtained as a solution to an optimization problem and it may in fact represent 0 up to numerical error. The function $f(x)$ is sensitive to changes in $x$ well below standard numerical accuracy. The point $x = 0$ does not satisfy $f(x) > 0.5$ but it is only $10^{-30}$ from doing so.

## 7.5 Semidefinite variables

Special care should be given to models involving semidefinite matrix variables (see Sec. 6), otherwise it is easy to produce an unnecessarily inefficient model. The general rule of thumb is:

- having many small matrix variables is more efficient than one big matrix variable,
- efficiency drops with growing number of semidefinite terms involved in linear constraints. This can have much bigger impact on the solution time than increasing the dimension of the semidefinite variable.

Let us consider a few examples.

### Block matrices

Given two matrix variables $X, Y \succeq 0$ *do not* assemble them in a block matrix and write the constraints as

$$\begin{bmatrix} X & 0 \\ 0 & Y \end{bmatrix} \succeq 0.$$

This increases the dimension of the problem and, even worse, introduces unnecessary constraints for a large portion of entries of the block matrix.

### Schur complement

Suppose we want to model a relaxation of a rank-one constraint:

$$\begin{bmatrix} X & x \\ x^T & 1 \end{bmatrix} \succeq 0.$$

where $x \in \mathbb{R}^n$ and $X \in \mathcal{S}^n_+$. The correct way to do this is to set up a matrix variable $Y \in \mathcal{S}^{n+1}_+$ with only a linear number of constraints:

$$\begin{array}{ll} Y_{i,n+1} = x_i, & i = 1, \ldots, n \\ Y_{n+1,n+1} = 1, & \\ Y \succeq 0, & \end{array}$$

and use the upper-left $n \times n$ part of $Y$ as the original $X$. Going the other way around, i.e. starting with a variable $X$ and aligning it with the corner of another, bigger semidefinite matrix $Y$ introduces $n(n+1)/2$ equality constraints and will quickly have formidable solution times.

### Sparse LMIs

Suppose we want to model a problem with a sparse linear matrix inequality (see Sec. 6.2.1) such as:

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & A_0 + \sum_{i=1}^k A_i x_i \succeq 0, \end{array}$$

where $x \in \mathbb{R}^k$ and $A_i$ are symmetric $n \times n$ matrices. The representation of this problem in primal form is:

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & A_0 + \sum_{i=1}^k A_i x_i = X, \\ & X \succeq 0, \end{array}$$

and the linear constraint requires a full set of $n(n+1)/2$ equalities, many of which are just $X_{k,l} = 0$, regardless of the sparsity of $A_i$. However the dual problem (see Sec. 8.6) is:

$$\begin{array}{ll} \text{maximize} & -\langle A_0, Z \rangle \\ \text{subject to} & \langle A_i, Z \rangle = c_i, \ i = 1, \ldots, k, \\ & Z \succeq 0, \end{array}$$

and the number of nonzeros in linear constraints is just joint number of nonzeros in $A_i$. It means that large, sparse LMIs should almost always be dualized and entered in that form for efficiency reasons.

## 7.6 The quality of a solution

In this section we will discuss how to validate an obtained solution. Assume we have a conic model with continuous variables only and that the optimization software has reported an optimal primal and dual solution. Given such a solution, we might ask how to verify that it is indeed feasible and optimal.

To that end, consider a simple model

$$\begin{array}{lrcl} \text{minimize} & -x_2 & & \\ \text{subject to} & x_1 + x_2 & \leq & 3, \\ & x_2^2 & \leq & 2, \\ & x_1, x_2 & \geq & 0, \end{array} \tag{7.1}$$

where a solver might approximate the solution as

$$x_1 = 0.0000000000000000 \text{ and } x_2 = 1.4142135623730951$$

and therefore the approximate optimal objective value is

$$-1.4142135623730951.$$

The true objective value is $-\sqrt{2}$, so the approximate objective value is wrong by the amount

$$1.4142135623730951 - \sqrt{2} \approx 10^{-16}.$$

Most likely this difference is irrelevant for all practical purposes. Nevertheless, in general a solution obtained using floating point arithmetic is only an approximation. Most (if not all) commercial optimization software uses double precision floating point arithmetic, implying that about 16 digits are correct in the computations performed internally by the software. This also means that irrational numbers such as $\sqrt{2}$ and $\pi$ can only be stored accurately within 16 digits.

### Verifying feasibility

A good practice after solving an optimization problem is to evaluate the reported solution. At the very least this process should be carried out during the initial phase of building a model, or if the reported solution is unexpected in some way. The first step in that process is to check that the solution is feasible; in case of the small example (7.1) this amounts to checking that:

$$\begin{array}{rcrcrcl} x_1 + x_2 & = & 0.0000000000000000 + 1.4142135623730951 & \leq & 3, & & \\ x_2^2 & = & 2.0000000000000004 & \leq & 2, & & (?) \\ x_1 & = & 0.0000000000000000 & \geq & 0, & & \\ x_2 & = & 1.4142135623730951 & \geq & 0, & & \end{array}$$

which demonstrates that one constraint is slightly violated due to computations in finite precision. It is up to the user to assess the significance of a specific violation; for example a violation of one unit in

$$x_1 + x_2 \leq 1$$

may or may not be more significant than a violation of one unit in

$$x_1 + x_2 \leq 10^9.$$

The right-hand side of $10^9$ may itself be the result of a computation in finite precision, and may only be known with, say 3 digits of accuracy. Therefore, a violation of 1 unit is not significant since the true right-hand side could just as well be $1.001 \cdot 10^9$. Certainly a violation of $4 \cdot 10^{-16}$ as in our example is within the numerical tolerance for zero and we would accept the solution as feasible. In practice it would be standard to see violations of order $10^{-8}$.

### Verifying optimality

Another question is how to verify that a feasible approximate solution is actually optimal, which is answered by duality theory, which is discussed in Sec. 2.4 for linear problems and in Sec. 8 in full generality. Following Sec. 8 we can derive an equivalent version of the dual problem as:

$$\begin{array}{lrcl} \text{maximize} & -3y_1 - y_2 - y_3 & & \\ \text{subject to} & 2y_2y_3 & \geq & (y_1+1)^2, \\ & y_1 & \geq & 0. \end{array}$$

This problem has a feasible point $(y_1, y_2, y_3) = (0, 1/\sqrt{2}, 1/\sqrt{2})$ with objective value $-\sqrt{2}$, matching the objective value of the primal problem. By duality theory this proves that both the primal and dual solutions are optimal. Again, in finite precision the dual objective value may be reported as, for example

$$-1.4142135623730950$$

leading to a duality gap of about $10^{-16}$ between the approximate primal and dual objective values. It is up to the user to assess the significance of any duality gap and accept or reject the solution as sufficiently good approximation of the optimal value.

## 7.7 Distance to a cone

The violation of a linear constraint $a^T x \leq b$ under a given solution $x^*$ is obviously

$$\max(0, a^T x^* - b).$$

It is less obvious how to assess violations of conic constraints. To this end suppose we have a convex cone $K$. The (Euclidean) projection of a point $x$ onto $K$ is defined as the solution of the problem

$$\begin{array}{ll} \text{minimize} & \|p - x\|_2 \\ \text{subject to} & p \in K. \end{array} \tag{7.2}$$

We denote the projection by $\text{proj}_K(x)$. The distance of $x$ to $K$

$$\text{dist}(x, K) = \|x - \text{proj}_K(x)\|_2,$$

i.e. the objective value of (7.2), measures the violation of a conic constraint involving $K$. Obviously

$$\operatorname{dist}(x, K) = 0 \iff x \in K.$$

This distance measure is attractive because it depends only on the set $K$ and not on any particular representation of $K$ using (in)equalities.

**Example 7.8** (Surprisingly small distance)**.** Let $K = \mathcal{P}_3^{0.1,0.9}$ be the power cone defined by the inequality $x^{0.1}y^{0.9} \geq |z|$. The point

$$x^* = (x, y, z) = (0, 10000, 500)$$

is clearly not in the cone, in fact $x^{0.1}y^{0.9} = 0$ while $|z| = 500$, so the violation of the conic inequality is seemingly large. However, the point

$$p = (10^{-8}, 10000, 500)$$

belongs to $\mathcal{P}_3^{0.1,0.9}$ (since $(10^{-8})^{0.1} \cdot 10000^{0.9} \approx 630$), hence

$$\operatorname{dist}(x^*, \mathcal{P}_3^{0.1,0.9}) \leq \|x^* - p\|_2 = 10^{-8}.$$

Therefore the distance of $x^*$ to the cone is actually very small.

### Distance to certain cones

For some basic cone types the projection problem (7.2) can be solved analytically. Denoting $[x]_+ = \max(0, x)$ we have that:

- For $x \in \mathbb{R}$ the projection onto the nonnegative half-line is $\operatorname{proj}_{\mathbb{R}_+}(x) = [x]_+$.
- For $(t, x) \in \mathbb{R} \times \mathbb{R}^{n-1}$ the projection onto the quadratic cone is

$$\operatorname{proj}_{\mathcal{Q}^n}(t, x) = \begin{cases} (t, x) & \text{if } t \geq \|x\|_2, \\ \frac{1}{2}\left(\frac{t}{\|x\|_2} + 1\right)(\|x\|_2, x) & \text{if } -\|x\|_2 < t < \|x\|_2, \\ 0 & \text{if } t \leq -\|x\|_2. \end{cases}$$

- For $X = \sum_{i=1}^n \lambda_i q_i q_i^T$ the projection onto the semidefinite cone is

$$\operatorname{proj}_{\mathcal{S}_+^n}(X) = \sum_{i=1}^n [\lambda_i]_+ q_i q_i^T.$$

# Chapter 8

# Duality in conic optimization

In Sec. 2 we introduced duality and related concepts for linear optimization. Here we present a more general version of this theory for conic optimization and we illustrate it with examples. Although this chapter is self-contained, we recommend familiarity with Sec. 2, which some of this material is a natural extension of.

Duality theory is a rich and powerful area of convex optimization, and central to understanding sensitivity analysis and infeasibility issues. Furthermore, it provides a simple and systematic way of obtaining non-trivial lower bounds on the optimal value for many difficult non-convex problems.

From now on we consider a conic problem in standard form

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \\ & x \in K, \end{array} \tag{8.1}$$

where $K$ is a convex cone. In practice $K$ would likely be defined as a product

$$K = K_1 \times \cdots \times K_m$$

of smaller cones corresponding to actual constraints in the problem formulation. The abstract formulation (8.1) encompasses such special cases as $K = \mathbb{R}^n_+$ (linear optimization), $K_i = \mathbb{R}$ (unconstrained variable) or $K_i = \mathcal{S}^n_+$ ($\frac{1}{2}n(n+1)$ variables representing the upper-triangular part of a matrix).

The *feasible set* for (8.1):

$$\mathcal{F}_p = \{x \in \mathbb{R}^n \mid Ax = b\} \cap K$$

is a section of $K$. We say the problem is *feasible* if $\mathcal{F}_p \neq \emptyset$ and *infeasible* otherwise. The *value* of (8.1) is defined as

$$p^\star = \inf\{c^T x \ : \ x \in \mathcal{F}_p\},$$

allowing the special cases of $p^\star = +\infty$ (the problem is infeasible) and $p^\star = -\infty$ (the problem is unbounded). Note that $p^\star$ may not be attained, i.e. the infimum is not necessarily a minimum, although this sort of problems are ill-posed from a practical point of view (see Sec. 7).

**Example 8.1** (Unattained problem value)**.** The conic quadratic problem

$$\begin{array}{ll} \text{minimize} & x \\ \text{subject to} & (x, y, 1) \in \mathcal{Q}_r^3, \end{array}$$

with constraint equivalent to $2xy \geq 1,\ x, y \geq 0$ has $p^\star = 0$ but this optimal value is not attained by any point in $\mathcal{F}_p$.

## 8.1 Dual cone

Suppose $K \subseteq \mathbb{R}^n$ is a closed convex cone. We define the *dual cone* $K^*$ as

$$K^* = \{y \in \mathbb{R}^n \ : \ y^T x \geq 0 \ \forall x \in K\}. \tag{8.2}$$

For simplicity we write $y^T x$, denoting the standard Euclidean inner product, but everything in this section applies verbatim to the inner product of matrices in the semidefinite context. In other words $K^*$ consists of vectors which form an acute (or right) angle with *every* vector in $K$. We see easily that $K^*$ is in fact a convex cone. The main example, associated with linear optimization, is the dual of the positive orthant:

**Lemma 8.1** (Dual of linear cone)**.**

$$(\mathbb{R}^n_+)^* = \mathbb{R}^n_+.$$

*Proof.* This is obvious for $n = 1$, and then we use the simple fact

$$(K_1 \times \cdots \times K_m)^* = K_1^* \times \cdots \times K_m^*,$$

which we invite the reader to prove. □

All cones studied in this cookbook can be explicitly dualized as we show next.

**Lemma 8.2** (Duals of nonlinear cones)**.** *We have the following:*

- *The quadratic, rotated quadratic and semidefinite cones are self-dual:*

$$(\mathcal{Q}^n)^* = \mathcal{Q}^n,\ (\mathcal{Q}_r^n)^* = \mathcal{Q}_r^n,\ (\mathcal{S}_+^n)^* = \mathcal{S}_+^n.$$

- *The dual of the power cone* (4.4) *is*

$$(\mathcal{P}_n^{\alpha_1,\cdots,\alpha_m})^* = \left\{ y \in \mathbb{R}^n \ : \ \left( \frac{y_1}{\alpha_1}, \ldots, \frac{y_m}{\alpha_m}, y_{m+1}, \ldots, y_n \right) \in \mathcal{P}_n^{\alpha_1,\cdots,\alpha_m} \right\}.$$

- *The dual of the exponential cone* (5.2) *is the closure*

$$(K_{\exp})^* = \operatorname{cl}\left\{ y \in \mathbb{R}^3 \ : \ y_1 \geq -y_3 e^{y_2/y_3 - 1},\ y_1 > 0, y_3 < 0 \right\}.$$

*Proof.* We only sketch some parts of the proof and indicate geometric intuitions. First, let us show $(\mathcal{Q}^n)^* = \mathcal{Q}^n$. For $(t, x), (s, y) \in \mathcal{Q}^n$ we have

$$st + y^T x \geq \|y\|_2 \|x\|_2 + y^T x \geq 0$$

by the Cauchy-Schwartz inequality $|u^T v| \leq \|u\|_2 \|v\|_2$, therefore $\mathcal{Q}^n \subseteq (\mathcal{Q}^n)^*$. Geometrically we are just saying that the quadratic cone $\mathcal{Q}^n$ has a right angle at the apex, so any two vectors inside the cone form at most a right angle. Now suppose $(s, y)$ is a point outside $\mathcal{Q}^n$. If $(-s, -y) \in \mathcal{Q}^n$ then $-s^2 - y^T y < 0$ and we showed $(s, y) \notin (\mathcal{Q}^n)^*$. Otherwise let $(t, x) = \text{proj}_{\mathcal{Q}^n}(s, y)$ be the projection (see Sec. 7.7); note that $(s, y)$ and $(t, -x) \in \mathcal{Q}^n$ form an obtuse angle.

For the semidefinite cone we use the property $\langle X, Y \rangle \geq 0$ for $X, Y \in \mathcal{S}^n_+$ (see Sec. 6.1.2). Conversely, assume that $Z \notin \mathcal{S}^n_+$. Then there exists a $w$ satisfying $\langle ww^T, Z \rangle = w^T Z w < 0$, so $Z \notin (\mathcal{S}^n_+)^*$.

As our last exercise let us check that vectors in $K_{\text{exp}}$ and $(K_{\text{exp}})^*$ form acute angles. By definition of $K_{\text{exp}}$ and $(K_{\text{exp}})^*$ we take $x, y$ such that:

$$x_1 \geq x_2 \exp(x_3/x_2), \quad y_1 \geq -y_3 \exp(y_2/y_3 - 1), \quad x_1, x_2, y_1 > 0, \; y_3 < 0,$$

and then we have

$$\begin{aligned} y^T x = x_1 y_1 + x_2 y_2 + x_3 y_3 &\geq -x_2 y_3 \exp(\tfrac{x_3}{x_2} + \tfrac{y_2}{y_3} - 1) + x_2 y_2 + x_3 y_3 \\ &\geq -x_2 y_3 (\tfrac{x_3}{x_2} + \tfrac{y_2}{y_3}) + x_2 y_2 + x_3 y_3 = 0, \end{aligned}$$

using the inequality $\exp(t) \geq 1 + t$. We refer to the literature for full proofs of all the statements. □

Finally it is nice to realize that $(K^*)^* = K$ and that the cones $\{0\}$ and $\mathbb{R}$ are each others' duals. We leave it to the reader to check these facts.

## 8.2 Infeasibility in conic optimization

We can now discuss infeasibility certificates for conic problems. Given an optimization problem, the first basic question we are interested in is its feasibility status. The theory of infeasibility certificates for linear problems, including the Farkas lemma (see Sec. 2.3) extends almost verbatim to the conic case.

**Lemma 8.3** (Farkas' lemma, conic version)**.** *Suppose we have the conic optimization problem* (8.1)*. Exactly one of the following statements is true:*

1. *Problem* (8.1) *is feasible.*
2. *Problem* (8.1) *is infeasible, but there is a sequence $x_n \in K$ such that $\|Ax_n - b\| \to 0$.*
3. *There exists $y$ such that $-A^T y \in K^*$ and $b^T y > 0$.*

Problems which fall under option 2. (*limit-feasible*) are ill-posed: an arbitrarily small perturbation of input data puts the problem in either category 1. or 3. This fringe case should therefore not appear in practice, and if it does, it signals issues with the optimization model which should be addressed.

**Example 8.2** (Limit-feasible model)**.** Here is an example of an ill-posed limit-feasible model created by fixing one of the root variables of a rotated quadratic cone to 0.

$$\begin{array}{ll} \text{minimize} & u \\ \text{subject to} & (u, v, w) \in \mathcal{Q}_r^3, \\ & v = 0, \\ & w \geq 1. \end{array}$$

The problem is clearly infeasible, but the sequence $(u_n, v_n, w_n) = (n, 1/n, 1) \in \mathcal{Q}_r^3$ with $(v_n, w_n) \to (0, 1)$ makes it limit-feasible as in alternative 2. of Lemma 8.3. There is no infeasibility certificate as in alternative 3.

Having cleared this detail we continue with the proof and example for the actually useful part of conic Farkas' lemma.

*Proof.* Consider the set

$$A(K) = \{Ax \ : \ x \in K\}.$$

It is a convex cone. Feasibility is equivalent to $b \in A(K)$. If $b \notin A(K)$ but $b \in \text{cl}(A(K))$ then we have the second alternative. Finally, if $b \notin \text{cl}(A(K))$ then the *closed* convex cone $\text{cl}(A(K))$ and the point $b$ can be strictly separated by a hyperplane passing through the origin, i.e. there exists $y$ such that

$$b^T y > 0, \quad (Ax)^T y \leq 0 \ \forall x \in K.$$

But then $0 \leq -(Ax)^T y = (-A^T y)^T x$ for all $x \in K$, so by definition of $K^*$ we have $-A^T y \in K^*$, and we showed $y$ satisfies the third alternative. Finally, 1. and 3. are mutually exclusive, since otherwise we would have

$$0 \leq (-A^T y)^T x = -y^T A x = -y^T b < 0$$

and the same argument works in the limit if $x_n$ is a sequence as in 2. That proves the lemma. □

Therefore Farkas' lemma implies that (up to certain ill-posed cases) either the problem (8.1) is feasible (first alternative) or there is a *certificate of infeasibility* $y$ (last alternative). In other words, every time we classify a model as infeasible, we can certify this fact by providing an appropriate $y$. Note that when $K = \mathbb{R}^n_+$ we recover precisely the linear version, Lemma 2.1.

**Example 8.3** (Infeasible conic problem)**.** Consider a minimization problem:

$$\begin{array}{ll} \text{minimize} & x_1 \\ \text{subject to} & -x_1 + x_2 - x_4 = 1, \\ & 2x_3 - 3x_4 = -1, \\ & x_1 \geq \sqrt{x_2^2 + x_3^2}, \\ & x_4 \geq 0. \end{array}$$

It can be expressed in the standard form (8.1) with

$$A = \begin{bmatrix} -1 & 1 & 0 & -1 \\ 0 & 0 & 2 & -3 \end{bmatrix}, \; b = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \; K = \mathcal{Q}^3 \times \mathbb{R}_+, \; c = [1, 0, 0, 0]^T.$$

A certificate of infeasibility is $y = [1, 0]^T$. Indeed, $b^T y = 1 > 0$ and $-A^T y = [1, -1, 0, 1] \in K^* = \mathcal{Q}^3 \times \mathbb{R}_+$. The certificate indicates that the first linear constraint alone causes infeasibility, which is indeed the case: the first equality together with the conic constraints yield a contradiction:

$$x_2 - 1 \geq x_2 - 1 - x_4 = x_1 \geq |x_2|.$$

## 8.3 Lagrangian and the dual problem

### Classical Lagrangian

In general constrained optimization we consider an optimization problem of the form

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f(x) \leq 0, \\ & h(x) = 0, \end{array} \tag{8.3}$$

where $f_0 : \mathbb{R}^n \mapsto \mathbb{R}$ is the objective function, $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ encodes inequality constraints, and $h : \mathbb{R}^n \mapsto \mathbb{R}^p$ encodes equality constraints. Readers familiar with the method of *Lagrange multipliers* or *penalization* will recognize the *Lagrangian* for (8.3), a function $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \mapsto \mathbb{R}$ that augments the objective with a weighted combination of all the constraints,

$$L(x, y, s) = f_0(x) + y^T h(x) + s^T f(x). \tag{8.4}$$

The variables $y \in \mathbb{R}^p$ and $s \in \mathbb{R}^m_+$ are called *Lagrange multipliers* or *dual variables*. The Lagrangian has the property that $L(x, y, s) \leq f_0(x)$ whenever $x$ is feasible for (8.1) and $s \in \mathbb{R}^m_+$. The optimal point satisfies the first-order optimality condition $\nabla_x L(x, y, s) = 0$. Moreover, the dual function $g(y, s) = \inf_x L(x, y, s)$ provides a lower bound for the optimal value of (8.3) for any $y \in \mathbb{R}^p$ and $s \in \mathbb{R}^m_+$, which leads to considering the dual problem of maximizing $g(y, s)$.

## Lagrangian for a conic problem

We next set up an analogue of the Lagrangian theory for the conic problem (8.1)

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \\ & x \in K, \end{array}$$

where $x, c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$ and $K \subseteq \mathbb{R}^n$ is a convex cone. We associate with (8.1) a Lagrangian of the form $L : \mathbb{R}^n \times \mathbb{R}^m \times K^* \to \mathbb{R}$

$$L(x, y, s) = c^T x + y^T(b - Ax) - s^T x.$$

For any feasible $x^* \in \mathcal{F}_p$ and any $(y^*, s^*) \in \mathbb{R}^m \times K^*$ we have

$$L(x^*, y^*, s^*) = c^T x^* + (y^*)^T \cdot 0 - (s^*)^T x^* \leq c^T x^*. \tag{8.5}$$

Note the we used the definition of the dual cone to conclude that $(s^*)^T x^* \geq 0$. The *dual function* is defined as the minimum of $L(x, y, s)$ over $x$. Thus the dual function of (8.1) is

$$g(y, s) = \min_x L(x, y, s) = \min_x x^T(c - A^T y - s) + b^T y = \begin{cases} b^T y, & c - A^T y - s = 0, \\ -\infty, & \text{otherwise.} \end{cases}$$

## Dual problem

From (8.5) every $(y, s) \in \mathbb{R}^m \times K^*$ satisfies $g(y, s) \leq p^\star$, i.e. $g(y, s)$ is a lower bound for $p^\star$. To get the best such bound we maximize $g(y, s)$ over all $(y, s) \in \mathbb{R}^m \times K^*$ and get the *dual problem*:

$$\begin{array}{ll} \text{maximize} & b^T y \\ \text{subject to} & c - A^T y = s, \\ & s \in K^*, \end{array} \tag{8.6}$$

or simply:

$$\begin{array}{ll} \text{maximize} & b^T y \\ \text{subject to} & c - A^T y \in K^*. \end{array} \tag{8.7}$$

The optimal value of (8.6) will be denoted $d^\star$. As in the case of (8.1) (which from now on we call the *primal problem*), the dual problem can be infeasible ($d^\star = -\infty$), have an optimal solution ($-\infty < d^\star < +\infty$) or be unbounded ($d^\star = +\infty$). As before, the value $d^\star$ is defined as a supremum of $b^T y$ and may not be attained. Note that the roles of $-\infty$ and $+\infty$ are now reversed because the dual is a maximization problem.

**Example 8.4** (More general constraints)**.** We can just as easily derive the dual of a problem with more general constraints, without necessarily having to transform the problem to

standard form beforehand. Imagine, for example, that some solver accepts conic problems of the form:

$$\begin{array}{ll} \text{minimize} & c^T x + c^f \\ \text{subject to} & l^c \leq Ax \leq u^c, \\ & l^x \leq x \leq u^x, \\ & x \in K. \end{array} \tag{8.8}$$

Then we define a Lagrangian with one set of dual variables for each constraint appearing in the problem:

$$\begin{array}{rcl} L(x, s_l^c, s_u^c, s_l^x, s_u^x, s_n^x) = & c^T x + c^f - (s_l^c)^T(Ax - l^c) - (s_u^c)^T(u^c - Ax) \\ & -(s_l^x)^T(x - l^x) - (s_u^x)^T(u^x - x) - (s_n^x)^T x \\ = & x^T(c - A^T s_l^c + A^T s_u^c - s_l^x + s_u^x - s_n^x) \\ & +(l^c)^T s_l^c - (u^c)^T s_u^c + (l^x)^T s_l^x - (u^c)^T s_u^x + c^f \end{array}$$

and that gives a dual problem

$$\begin{array}{ll} \text{maximize} & (l^c)^T s_l^c - (u^c)^T s_u^c + (l^x)^T s_l^x - (u^c)^T s_u^x + c^f \\ \text{subject to} & c + A^T(-s_l^c + s_u^c) - s_l^x + s_u^x - s_n^x = 0, \\ & s_l^c, s_u^c, s_l^x, s_u^x \geq 0, \\ & s_n^x \in K^*. \end{array}$$

**Example 8.5** (Dual of simple portfolio)**.** Consider a simplified version of the portfolio optimization problem, where we maximize expected return subject to an upper bound on the risk and no other constraints:

$$\begin{array}{ll} \text{maximize} & \mu^T x \\ \text{subject to} & \|Fx\|_2 \leq \gamma, \end{array}$$

where $x \in \mathbb{R}^n$ and $F \in \mathbb{R}^{m \times n}$. The conic formulation is

$$\begin{array}{ll} \text{maximize} & \mu^T x \\ \text{subject to} & (\gamma, Fx) \in \mathcal{Q}^{m+1}, \end{array} \tag{8.9}$$

and we can directly write the Lagrangian

$$L(x, v, w) = \mu^T x + v\gamma + w^T Fx = x^T(\mu + F^T w) + v\gamma$$

with $(v, w) \in (\mathcal{Q}^{m+1})^* = \mathcal{Q}^{m+1}$. Note that we chose signs to have $L(x, v, w) \geq \mu^T x$ since we are dealing with a maximization problem. The dual is now determined by the conditions $F^T w + \mu = 0$ and $v \geq \|w\|_2$, so it can be formulated as

$$\begin{array}{ll} \text{minimize} & \gamma\|w\|_2 \\ \text{subject to} & F^T w = -\mu. \end{array} \tag{8.10}$$

Note that it is actually more natural to view problem (8.9) as the dual form and problem (8.10) as the primal. Indeed we can write the constraint in (8.9) as

$$\begin{bmatrix} \gamma \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ -F \end{bmatrix} x \in (Q^{m+1})^*$$

which fits naturally into the form (8.7). Having done this we can recover the dual as a minimization problem in the standard form (8.1). We leave it to the reader to check that we get the same answer as above.

## 8.4 Weak and strong duality

### Weak duality

Suppose $x^*$ and $(y^*, s^*)$ are feasible points for the primal and dual problems (8.1) and (8.6), respectively. Then we have

$$b^T y^* = (Ax^*)^T y^* = (x^*)^T (A^T y^*) = (x^*)^T (c - s^*) = c^T x^* - (s^*)^T x^* \leq c^T x^* \tag{8.11}$$

so the dual objective value is a lower bound on the objective value of the primal. In particular, any dual feasible point $(y^*, s^*)$ gives a lower bound:

$$b^T y^* \leq p^\star$$

and we immediately get the next lemma.

**Lemma 8.4** (Weak duality)**.** $d^\star \leq p^\star$.

It follows that if $b^T y^* = c^T x^*$ then $x^*$ is optimal for the primal, $(y^*, s^*)$ is optimal for the dual and $b^T y^* = c^T x^*$ is the common optimal objective value. This way we can use the optimal dual solution to certify optimality of the primal solution and vice versa.

### Complementary slackness

Moreover, (8.11) asserts that $b^T y^* = c^T x^*$ is equivalent to orthogonality

$$(s^*)^T x^* = 0$$

i.e. *complementary slackness*. It is not hard to verify what complementary slackness means for particular types of cones, for example

- for $s, x \in \mathbb{R}_+$ we have $sx = 0$ iff $s = 0$ or $x = 0$,
- vectors $(s_1, \tilde{s}), (x_1, \tilde{x}) \in \mathcal{Q}^{n+1}$ are orthogonal iff $(s_1, \tilde{s})$ and $(x_1, -\tilde{x})$ are parallel,
- vectors $(s_1, s_2, \tilde{s}), (x_1, x_2, \tilde{x}) \in \mathcal{Q}_r^{n+2}$ are orthogonal iff $(s_1, s_2, \tilde{s})$ and $(x_2, x_1, -\tilde{x})$ are parallel.

One implicit case is worth special attention: complementary slackness for a linear inequality constraint $a^T x \leq b$ with a non-negative dual variable $y$ asserts that $(a^T x^* - b)y^* = 0$. This can be seen by directly writing down the appropriate Lagrangian for this type of constraint. Alternatively, we can introduce a slack variable $u = b - a^T x$ with a conic constraint $u \geq 0$ and let $y$ be the dual conic variable. In particular, if a constraint is non-binding in the optimal solution ($a^T x^* < b$) then the corresponding dual variable $y^* = 0$. If $y^* > 0$ then it can be related to a shadow price, see Sec. 2.4.4 and Sec. 8.5.2.

### Strong duality

The obvious question is now whether $d^\star = p^\star$, that is if optimality of a primal solution can always be certified by a dual solution with matching objective value, as for linear programming. This turns out not to be the case for general conic problems.

**Example 8.6** (Positive duality gap)**.** Consider the problem

$$\begin{array}{ll} \text{minimize} & x_3 \\ \text{subject to} & x_1 \geq \sqrt{x_2^2 + x_3^2}, \\ & x_2 \geq x_1,\ x_3 \geq -1. \end{array}$$

The only feasible points are $(x, x, 0)$, so $p^\star = 0$. The dual problem is

$$\begin{array}{ll} \text{maximize} & -y_2 \\ \text{subject to} & y_1 \geq \sqrt{y_1^2 + (1 - y_2)^2}, \end{array}$$

with feasible points $(y, 1)$, hence $d^\star = -1$.

Similarly, we consider a problem

$$\begin{array}{ll} \text{minimize} & x_1 \\ \text{subject to} & \begin{bmatrix} 0 & x_1 & 0 \\ x_1 & x_2 & 0 \\ 0 & 0 & 1 + x_1 \end{bmatrix} \in \mathcal{S}_+^3, \end{array}$$

with feasible set $\{x_1 = 0, x_2 \geq 0\}$ and optimal value $p^\star = 0$. The dual problem can be formulated as

$$\begin{array}{ll} \text{maximize} & -z_2 \\ \text{subject to} & \begin{bmatrix} z_1 & (1 - z_2)/2 & 0 \\ (1 - z_2)/2 & 0 & 0 \\ 0 & 0 & z_2 \end{bmatrix} \in \mathcal{S}_+^3, \end{array}$$

which has a feasible set $\{z_1 \geq 0, z_2 = 1\}$ and dual optimal value $d^\star = -1$.

To ensure strong duality for conic problems we need an additional regularity assumption. As with the conic version of Farkas' lemma Lemma 8.3, we stress that this is a technical condition to eliminate ill-posed problems which should not be formed in practice. In particular,

we invite the reader to think that strong duality holds for all well formed conic problems one is likely to come across in applications, and that having a duality gap signals issues with the model formulation.

We say problem (8.1) is *very nicely posed* if for all values of $c_0$ the feasibility problem

$$c^T x = c_0,\ Ax = b,\ x \in K$$

satisfies either the first or third alternative in Lemma 8.3.

**Lemma 8.5** (Strong duality)**.** *Suppose that* (8.1) *is very nicely posed and* $p^\star$ *is finite. Then* $d^\star = p^\star$.

*Proof.* For any $\varepsilon > 0$ consider the feasibility problem with variable $x$ and constraints

$$\begin{array}{rcl} \left[\begin{array}{c} -c^T \\ A \end{array}\right] x & = & \left[\begin{array}{c} -p^\star + \varepsilon \\ b \end{array}\right] \\ x & \in & K \end{array} \quad \text{that is} \quad \begin{array}{rcl} c^T x & = & p^\star - \varepsilon, \\ Ax & = & b, \\ x & \in & K. \end{array}$$

Optimality of $p^\star$ implies that the above problem is infeasible. By Lemma 8.3 and because we assumed very-nice-posedness there exists $\hat{y} = [y_0\ y]^T$ such that

$$\left[c,\ -A^T\right] \hat{y} \in K^* \text{ and } \left[-p^\star + \varepsilon,\ b^T\right] \hat{y} > 0.$$

If $y_0 = 0$ then $-A^T y \in K^*$ and $b^T y > 0$, which by Lemma 8.3 again would mean that the original problem was infeasible, which is not the case. Hence we can rescale so that $y_0 = 1$ and then we get

$$c - A^T y \in K^* \quad \text{and} \quad b^T y \geq p^\star - \varepsilon.$$

The first constraint means that $y$ is feasible for the dual problem. By letting $\varepsilon \to 0$ we obtain $d^\star \geq p^\star$. □

There are more direct conditions which guarantee strong duality, such as below.

**Lemma 8.6** (Slater constraint qualification)**.** *Suppose that* (8.1) *is* strictly feasible*: there exists a point* $x \in \operatorname{int}(K)$ *in the interior of* $K$ *such that* $Ax = b$*. Then strong duality holds if* $p^\star$ *is finite. Moreover, if both primal and dual problem are strictly feasible then* $p^\star$ *and* $d^\star$ *are attained.*

We omit the proof which can be found in standard texts. Note that the first problem from Example 8.6 does not satisfy Slater constraint qualification: the only feasible points lie on the boundary of the cone (we say the problem has *no interior*). That problem is not very nicely posed either: the point $(x_1, x_2, x_3) = (0.5c_0^2\varepsilon^{-1} + \varepsilon, 0.5c_0^2\varepsilon^{-1}, c_0) \in \mathcal{Q}^3$ violates the inequality $x_2 \geq x_1$ by an arbitrarily small $\varepsilon$, so the problem is infeasible but limit-feasible (second alternative in Lemma 8.3).

## 8.5 Applications of conic duality

### 8.5.1 Linear regression and the normal equation

Least-squares linear regression is the problem of minimizing $\|Ax - b\|_2^2$ over $x \in \mathbb{R}^n$, where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ are fixed. This problem can be posed in conic form as

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & (t, Ax - b) \in \mathcal{Q}^{m+1}, \end{array}$$

and we can write the Lagrangian

$$L(t, x, u, v) = t - tu - v^T(Ax - b) = t(1 - u) - x^T A^T v + b^T v$$

so the constraints in the dual problem are:

$$u = 1, \ A^T v = 0, \ (u, v) \in Q^{m+1}.$$

The problem exhibits strong duality with both the primal and dual values attained in the optimal solution. The primal solution clearly satisfies $t = \|Ax - b\|_2$, and so complementary slackness for the quadratic cone implies that the vectors $(u, -v) = (1, -v)$ and $(t, Ax - b)$ are parallel. As a consequence the constraint $A^T v = 0$ becomes $A^T(Ax - b) = 0$ or simply

$$A^T A x = A^T b$$

so if $A^T A$ is invertible then $x = (A^T A)^{-1} A^T b$. This is the so-called *normal equation* for least-squares regression, which we now obtained as a consequence of strong duality.

### 8.5.2 Constraint attribution

Consider again a portfolio optimization problem with mean-variance utility function, vector of expected returns $\alpha$ and covariance matrix $\Sigma = F^T F$:

$$\begin{array}{ll} \text{maximize} & \alpha^T x - \frac{1}{2} c x^T \Sigma x \\ \text{subject to} & Ax \leq b, \end{array} \tag{8.12}$$

where the linear part represents any set of additional constraints: total budget, sector constraints, diversification constraints, individual relations between positions etc. In the absence of additional constraints the solution to the unconstrained maximization problem is easy to derive using basic calculus and equals

$$\hat{x} = c^{-1} \Sigma^{-1} \alpha.$$

We would like to understand the difference $x^* - \hat{x}$, where $x^*$ is the solution of (8.12), and in particular to measure which of the linear constraints actually cause $x^*$ to deviate from $\hat{x}$ and to what degree. This can be quantified using the dual variables.

The conic version of problem (8.12) is

$$\begin{array}{ll} \text{maximize} & \alpha^T x - cr \\ \text{subject to} & Ax \leq b, \\ & (1, r, Fx) \in \mathcal{Q}_r \end{array}$$

with dual

$$\begin{array}{ll} \text{minimize} & b^T y + s \\ \text{subject to} & A^T y = \alpha + F^T u, \\ & (s, c, u) \in \mathcal{Q}_r, \\ & y \geq 0. \end{array}$$

Suppose we have a primal-dual optimal solution $(x^*, r^*, y^*, s^*, u^*)$. Complementary slackness for the rotated quadratic cone implies

$$(s^*, c, u^*) = \beta(r^*, 1, -Fx^*)$$

which leads to $\beta = c$ and

$$A^T y^* = \alpha - cF^T F x^*$$

or equivalently

$$x^* = \hat{x} - c^{-1}\Sigma^{-1} A^T y^*.$$

This equation splits the difference between the constrained and unconstrained solutions into contributions from individual constraints, where the weights are precisely the dual variables $y^*$. For example, if a constraint is not binding ($a_i^T x^* - b_i < 0$) then by complementary slackness $y_i^* = 0$ and, indeed, a non-binding constraint has no effect on the change in solution.

## 8.6 Semidefinite duality and LMIs

The general theory of conic duality applies in particular to problems with semidefinite variables so here we just state it in the language familiar to SDP practitioners. Consider for simplicity a primal semidefinite optimization problem with one matrix variable

$$\begin{array}{ll} \text{minimize} & \langle C, X \rangle \\ \text{subject to} & \langle A_i, X \rangle = b_i, \quad i = 1, \ldots, m, \\ & X \in \mathcal{S}_+^n. \end{array} \tag{8.13}$$

We can quickly repeat the derivation of the dual problem in this notation. The Lagrangian is

$$\begin{array}{rl} L(X, y, S) & = \langle C, X \rangle - \sum_i y_i(\langle A_i, X \rangle - b_i) - \langle S, X \rangle \\ & = \langle C - \sum_i y_i A_i - S, X \rangle + b^T y \end{array}$$

and we get the dual problem

$$\begin{array}{ll} \text{maximize} & b^T y \\ \text{subject to} & C - \sum_{i=1}^m y_i A_i \in \mathcal{S}_+^n. \end{array} \tag{8.14}$$

The dual contains an affine matrix-valued function with coefficients $C, A_i \in \mathcal{S}^n$ and variable $y \in \mathbb{R}^m$. Such a matrix-valued affine inequality is called a *linear matrix inequality (LMI)*. In Sec. 6 we formulated many problems as LMIs, that is in the form more naturally matching the dual.

From a modeling perspective it does not matter whether constraints are given as linear matrix inequalities or as an intersection of affine hyperplanes; one formulation is easily converted to other using auxiliary variables and constraints, and this transformation is often done transparently by optimization software. Nevertheless, it is instructive to study an explicit example of how to carry out this transformation. An linear matrix inequality

$$A_0 + x_1 A_1 + \cdots + x_n A_n \succeq 0$$

where $A_i \in \mathcal{S}_+^m$ is converted to a set of linear equality constraints using a slack variable

$$A_0 + x_1 A_1 + \cdots + x_n A_n = S, \qquad S \succeq 0.$$

Apart from introducing an explicit semidefinite variable $S \in \mathcal{S}_+^m$ we also added $m(m+1)/2$ equality constraints. On the other hand, a semidefinite variable $X \in \mathcal{S}_+^n$ can be rewritten as a linear matrix inequality with $n(n+1)/2$ scalar variables

$$X = \sum_{i=1}^n e_i e_i^T x_{ii} + \sum_{i=1}^n \sum_{j=i+1}^n (e_i e_j^T + e_j e_i^T) x_{ij} \succeq 0.$$

Obviously we should only use these transformations when necessary; if we have a problem that is more naturally interpreted in either primal or dual form, we should be careful to recognize that structure.

**Example 8.7** (Dualization and efficiency)**.** Consider the problem:

$$\begin{array}{ll} \text{minimize} & e^T z \\ \text{subject to} & A + \mathbf{Diag}(z) = X, \\ & X \succeq 0. \end{array}$$

with the variables $X \in \mathcal{S}_+^n$ and $z \in \mathbb{R}^n$. This is a problem in primal form with $n(n+1)/2$ equality constraints, but they are more naturally interpreted as a linear matrix inequality

$$A + \sum_i e_i e_i^T z_i \succeq 0.$$

The dual problem is

$$\begin{array}{ll} \text{maximize} & -\langle A, Z \rangle \\ \text{subject to} & \mathbf{diag}(Z) = e, \\ & Z \succeq 0, \end{array}$$

in the variable $Z \in \mathcal{S}_+^n$. The dual problem has only $n$ equality constraints, which is a vast improvement over the $n(n+1)/2$ constraints in the primal problem. See also Sec. 7.5.

**Example 8.8** (Sum of singular values revisited)**.** In Sec. 6.2.4, and specifically in (6.16), we expressed the problem of minimizing the sum of singular values of a nonsymmetric matrix $X$. Problem (6.16) can be written as an LMI:

$$\begin{array}{ll} \text{maximize} & \sum_{i,j} X_{ij} z_{ij} \\ \text{subject to} & I - \sum_{i,j} z_{ij} \begin{bmatrix} 0 & e_j e_i^T \\ e_i e_j^T & 0 \end{bmatrix} \succeq 0. \end{array}$$

Treating this as the dual and going back to the primal form we get:

$$\begin{array}{ll} \text{minimize} & \mathrm{Tr}(U) + \mathrm{Tr}(V) \\ \text{subject to} & S = -\frac{1}{2} X, \\ & \begin{bmatrix} U & S^T \\ S & V \end{bmatrix} \succeq 0, \end{array}$$

which is equivalent to the claimed (6.17). The dual formulation has the advantage of being linear in $X$.

# Chapter 9

# Mixed integer optimization

In other chapters of this cookbook we have considered different classes of convex problems with continuous variables. In this chapter we consider a much wider range of non-convex problems by allowing integer variables. This technique is extremely useful in practice, and already for linear programming it covers a vast range of problems. We introduce different building blocks for integer optimization, which make it possible to model useful non-convex dependencies between variables in conic problems. It should be noted that mixed integer optimization problems are very hard (technically NP-hard), and for many practical cases an exact solution may not be found in reasonable time.

## 9.1 Integer modeling

A general mixed integer conic optimization problem has the form

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \\ & x \in K, \\ & x_i \in \mathbb{Z}, \quad \forall i \in \mathcal{I}, \end{array} \tag{9.1}$$

where $K$ is a cone and $\mathcal{I} \subseteq \{1, \ldots, n\}$ denotes the set of variables that are constrained to be integers.

Two major techniques are typical for mixed integer optimization. The first one is the use of *binary variables*, also known as *indicator variables*, which only take values 0 and 1, and indicate the absence or presence of a particular event or choice. This restriction can of course be modeled in the form (9.1) by writing:

$$0 \leq x \leq 1 \text{ and } x \in \mathbb{Z}.$$

The other, known as *big-M*, refers to the fact that some relations can only be modeled linearly if one assumes some fixed bound $M$ on the quantities involved, and this constant enters the model formulation. The choice of $M$ can affect the performance of the model, see Example 7.7.

### 9.1.1 Implication of positivity

Often we have a real-valued variable $x \in \mathbb{R}$ satisfying $0 \leq x < M$ for a known upper bound $M$, and we wish to model the implication

$$x > 0 \implies z = 1. \tag{9.2}$$

Making $z$ a binary variable we can write (9.2) as a linear inequality

$$x \leq Mz,\ z \in \{0, 1\}. \tag{9.3}$$

Indeed $x > 0$ excludes the possibility of $z = 0$, hence forces $z = 1$. Since a priori $x \leq M$, there is no danger that the constraint accidentally makes the problem infeasible. A typical use of this trick is to model fixed setup costs.

**Example 9.1** (Fixed setup cost)**.** Assume that production of a specific item $i$ costs $u_i$ per unit, but there is an additional fixed charge of $w_i$ if we produce item $i$ at all. For instance, $w_i$ could be the cost of setting up a production plant, initial cost of equipment etc. Then the cost of producing $x_i$ units of product $i$ is given by the discontinuous function

$$c_i(x_i) = \begin{cases} w_i + u_i x_i, & x_i > 0 \\ 0, & x_i = 0. \end{cases}$$

If we let $M$ denote an upper bound on the quantities we can produce, we can then minimize the total production cost of $n$ products under some affine constraint $Ax = b$ with

$$\begin{array}{ll} \text{minimize} & u^T x + w^T z \\ \text{subject to} & Ax = b, \\ & x_i \leq M z_i, \quad i = 1, \ldots, n \\ & x \geq 0, \\ & z \in \{0, 1\}^n, \end{array}$$

which is a linear mixed-integer optimization problem. Note that by minimizing the production cost, we drive $z_i$ to 0 when $x_i = 0$, so setup costs are indeed included only for products with $x_i > 0$.

### 9.1.2 Semi-continuous variables

We can also model *semi-continuity* of a variable $x \in \mathbb{R}$,

$$x \in 0 \cup [a, b], \tag{9.4}$$

where $0 < a \leq b$ using a double inequality

$$az \leq x \leq bz,\ z \in \{0, 1\}.$$

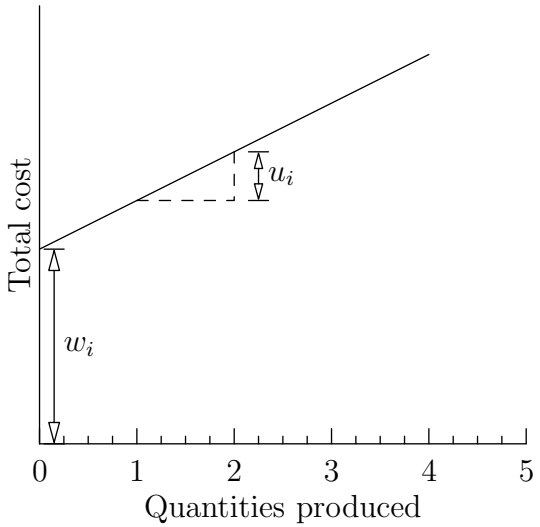



Fig. 9.1: Production cost with fixed setup cost $w_i$.

### 9.1.3 Indicator constraints

Suppose we want to model the fact that a certain linear inequality must be satisfied when some other event occurs. In other words, for a binary variable $z$ we want to model the implication

$$z = 1 \implies a^T x \leq b.$$

Suppose we know in advance an upper bound $a^T x - b \leq M$. Then we can write the above as a linear inequality

$$a^T x \leq b + M(1 - z).$$

Now if $z = 1$ then we forced $a^T x \leq b$, while for $z = 0$ the inequality is trivially satisfied and does not enforce any additional constraint on $x$.

### 9.1.4 Disjunctive constraints

With a disjunctive constraint we require that at least one of the given linear constraints is satisfied, that is

$$(a_1^T x \leq b_1) \vee (a_2^T x \leq b_2) \vee \cdots \vee (a_k^T x \leq b_k).$$

Introducing binary variables $z_1, \ldots, z_k$, we can use Sec. 9.1.3 to write a linear model

$$\begin{aligned} &z_1 + \cdots + z_k \geq 1, \\ &z_1, \ldots, z_k \in \{0, 1\}, \\ &a_i^T x \leq b_i + M(1 - z_i),\ i = 1, \ldots, k. \end{aligned}$$

Note that $z_j = 1$ implies that the $j$-th constraint is satisfied, but not vice-versa. Achieving that effect is described in the next section.

### 9.1.5 Constraint satisfaction

Say we want to define an optimization model that will behave differently depending on which of the inequalities

$$a^T x \leq b \qquad \text{or} \qquad a^T x \geq b$$

is satisfied. Suppose we have lower and upper bounds for $a^T x - b$ in the form of $m \leq a^T x - b \leq M$. Then we can write a model

$$b + mz \leq a^T x \leq b + M(1 - z),\ z \in \{0, 1\}. \tag{9.5}$$

Now observe that $z = 0$ implies $b \leq a^T x \leq b + M$, of which the right-hand inequality is redundant, i.e. always satisfied. Similarly, $z = 1$ implies $b + m \leq a^T x \leq b$. In other words $z$ is an indicator of whether $a^T x \leq b$.

In practice we would relax one inequality using a small amount of slack, i.e.,

$$b + (m - \epsilon)z + \epsilon \leq a^T x \tag{9.6}$$

to avoid issues with classifying the equality $a^T x = b$.

### 9.1.6 Exact absolute value

In Sec. 2.2.2 we showed how to model $|x| \leq t$ as two linear inequalities. Now suppose we need to model an exact equality

$$|x| = t. \tag{9.7}$$

It defines a non-convex set, hence it is not conic representable. If we split $x$ into positive and negative part $x = x^+ - x^-$, where $x^+, x^- \geq 0$, then $|x| = x^+ + x^-$ as long as either $x^+ = 0$ or $x^- = 0$. That last alternative can be modeled with a binary variable, and we get a model of (9.7):

$$\begin{array}{rcl} x & = & x^+ - x^-, \\ t & = & x^+ + x^-, \\ 0 & \leq & x^+, x^-, \\ x^+ & \leq & Mz, \\ x^- & \leq & M(1 - z), \\ z & \in & \{0, 1\}, \end{array} \tag{9.8}$$

where the constant $M$ is an a priori known upper bound on $|x|$ in the problem.

### 9.1.7 Exact 1-norm

We can use the technique above to model the exact $\ell_1$-norm equality constraint

$$\sum_{i=1}^{n} |x_i| = c, \tag{9.9}$$

where $x \in \mathbb{R}^n$ is a decision variable and $c$ is a constant. Such constraints arise for instance in *fully invested* portfolio optimizations scenarios (with short-selling). As before, we split $x$ into a positive and negative part, using a sequence of binary variables to guarantee that at most one of them is nonzero:

$$\begin{array}{rcl} x & = & x^+ - x^-, \\ 0 & \leq & x^+, x^-, \\ x^+ & \leq & cz, \\ x^- & \leq & c(e-z), \\ \sum_i x_i^+ + \sum_i x_i^- & = & c, \\ z & \in & \{0,1\}^n,\ x^+, x^- \in \mathbb{R}^n. \end{array} \tag{9.10}$$

### 9.1.8 Maximum

The exact equality $t = \max\{x_1, \ldots, x_n\}$ can be expressed by introducing a sequence of mutually exclusive indicator variables $z_1, \ldots, z_n$, with the intention that $z_i = 1$ picks the variable $x_i$ which actually achieves maximum. Choosing a safe bound $M$ we get a model:

$$\begin{array}{rcl} x_i & \leq & t \ \leq \ x_i + M(1-z_i),\ i = 1, \ldots, n, \\ z_1 + \cdots + z_n & = & 1, \\ z & \in & \{0,1\}^n. \end{array} \tag{9.11}$$

### 9.1.9 Boolean operators

Typically an indicator variable $z \in \{0,1\}$ represents a boolean value (true/false). In this case the standard boolean operators can be implemented as linear inequalities. In the table below we assume all variables are binary.

Table 9.1: Boolean operators

| Boolean | Linear |
|---|---|
| $z = x$ OR $y$ | $x \leq z,\ y \leq z,\ z \leq x + y$ |
| $z = x$ AND $y$ | $x \geq z,\ y \geq z,\ z + 1 \geq x + y$ |
| $z =$ NOT $x$ | $z = 1 - x$ |
| $x \implies y$ | $x \leq y$ |
| At most one of $z_1, \ldots, z_n$ holds (SOS1, set-packing) | $\sum_i z_i \leq 1$ |
| Exactly one of $z_1, \ldots, z_n$ holds (set-partitioning) | $\sum_i z_i = 1$ |
| At least one of $z_1, \ldots, z_n$ holds (set-covering) | $\sum_i z_i \geq 1$ |
| At most $k$ of $z_1, \ldots, z_n$ holds (cardinality) | $\sum_i z_i \leq k$ |

### 9.1.10 Fixed set of values

We can restrict a variable to take on only values from a specified finite set $\{a_1, \ldots, a_n\}$ by writing

$$\begin{array}{rcl} x & = & \sum_i z_i a_i \\ z & \in & \{0,1\}^n, \\ \sum_i z_i & = & 1. \end{array} \tag{9.12}$$

In (9.12) we essentially defined $z_i$ to be the indicator variable of whether $x = a_i$. In some circumstances there may be more efficient representations of a restricted set of values, for example:

- (sign) $x \in \{-1, 1\} \iff x = 2z - 1,\ z \in \{0, 1\}$,
- (modulo) $x \in \{1, 4, 7, 10\} \iff x = 3z + 1,\ 0 \le z \le 3,\ z \in \mathbb{Z}$,
- (fraction) $x \in \{0, 1/3, 2/3, 1\} \iff 3x = z,\ 0 \le z \le 3, z \in \mathbb{Z}$,
- (gap) $x \in (-\infty, a] \cup [b, \infty) \iff b - M(1-z) \le x \le a + Mz,\ z \in \{0, 1\}$ for sufficiently large $M$.

### 9.1.11 Continuous piecewise-linear functions

Consider a continuous, univariate, piecewise-linear, *non-convex* function $f : [\alpha_1, \alpha_5] \mapsto \mathbb{R}$ shown in Fig. 9.2. At the interval $[\alpha_j, \alpha_{j+1}]$, $j = 1, 2, 3, 4$ we can describe the function as

$$f(x) = \lambda_j f(\alpha_j) + \lambda_{j+1} f(\alpha_{j+1})$$

where $\lambda_j \alpha_j + \lambda_{j+1}\alpha_{j+1} = x$ and $\lambda_j + \lambda_{j+1} = 1$. If we add a constraint that only two (adjacent) variables $\lambda_j, \lambda_{j+1}$ can be nonzero, we can characterize every value $f(x)$ over the entire interval $[\alpha_1, \alpha_5]$ as some convex combination,

$$f(x) = \sum_{j=1}^{4} \lambda_j f(\alpha_j).$$

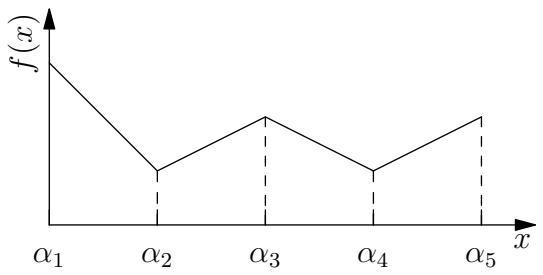



Fig. 9.2: A univariate piecewise-linear non-convex function.

The condition that only two adjacent variables can be nonzero is sometimes called an SOS2 constraint. If we introduce indicator variables $z_i$ for each pair of adjacent variables $(\lambda_i, \lambda_{i+1})$, we can model an SOS2 constraint as:

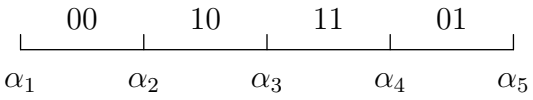
$$\lambda_1 \le z_1, \quad \lambda_2 \le z_1 + z_2, \quad \lambda_3 \le z_2 + z_3, \quad \lambda_4 \le z_4 + z_3, \quad \lambda_5 \le z_4$$
$$z_1 + z_2 + z_3 + z_4 = 1, \quad z \in \{0, 1\}^4,$$

so that we have $z_j = 1 \implies \lambda_i = 0, i \neq \{j, j+1\}$. Collectively, we can then model the epigraph $f(x) \leq t$ as

$$\begin{array}{ccccc} x = \sum_{j=1}^{n} \lambda_j \alpha_j, & & \sum_{j=1}^{n} \lambda_j f(\alpha_j) \leq t & & \\ \lambda_1 \leq z_1, & \lambda_j \leq z_j + z_{j-1}, \; j = 2, \ldots, n-1, & & \lambda_n \leq z_{n-1}, & \\ \lambda \geq 0, & \sum_{j=1}^{n} \lambda_j = 1, & \sum_{j=1}^{n-1} z_j = 1, & z \in \{0,1\}^{n-1}, & \end{array} \tag{9.13}$$

for a piecewise-linear function $f(x)$ with $n$ terms. This approach is often called the *lambda-method*.

For the function in Fig. 9.2 we can reduce the number of integer variables by using a *Gray encoding*




of the intervals $[\alpha_j, \alpha_{j+1}]$ and an indicator variable $y \in \{0,1\}^2$ to represent the four different values of Gray code. We can then describe the constraints on $\lambda$ using only two indicator variables,

$$\begin{array}{rcl} (y_1 = 0) & \rightarrow & \lambda_3 = 0, \\ (y_1 = 1) & \rightarrow & \lambda_1 = \lambda_5 = 0, \\ (y_2 = 0) & \rightarrow & \lambda_4 = \lambda_5 = 0, \\ (y_2 = 1) & \rightarrow & \lambda_1 = \lambda_2 = 0, \end{array}$$

which leads to a more efficient characterization of the epigraph $f(x) \leq t$,

$$\begin{array}{c} x = \sum_{j=1}^{5} \lambda_j \alpha_j, \quad \sum_{j=1}^{5} \lambda_j f(\alpha_j) \leq t, \\ \lambda_3 \leq y_1, \quad \lambda_1 + \lambda_5 \leq (1 - y_1), \quad \lambda_4 + \lambda_5 \leq y_2, \quad \lambda_1 + \lambda_2 \leq (1 - y_2), \\ \lambda \geq 0, \quad \sum_{j=1}^{5} \lambda_j = 1, \quad y \in \{0,1\}^2, \end{array}$$

The lambda-method can also be used to model multivariate continuous piecewise-linear non-convex functions, specified on a set of polyhedra $P_k$. For example, for the function shown in Fig. 9.3 we can model the epigraph $f(x) \leq t$ as

$$\begin{array}{c} x = \sum_{i=1}^{6} \lambda_i v_i, \qquad \sum_{i=1}^{6} \lambda_i f(v_i) \leq t, \\ \lambda_1 \leq z_1 + z_2, \qquad \lambda_2 \leq z_1, \qquad \lambda_3 \leq z_2 + z_3, \\ \lambda_4 \leq z_1 + z_2 + z_3 + z_4, \qquad \lambda_5 \leq z_3 + z_4, \qquad \lambda_6 \leq z_4, \\ \lambda \geq 0, \qquad \sum_{i=1}^{6} \lambda_i = 1, \qquad \sum_{i=1}^{4} z_i = 1, \\ z \in \{0,1\}^4. \end{array} \tag{9.14}$$

Note, for example, that $z_2 = 1$ implies that $\lambda_2 = \lambda_5 = \lambda_6 = 0$ and $x = \lambda_1 v_1 + \lambda_3 v_3 + \lambda_4 v_4$.

### 9.1.12 Lower semicontinuous piecewise-linear functions

The ideas in Sec. 9.1.11 can be applied to *lower semicontinuous* piecewise-linear functions as well. For example, consider the function shown in Fig. 9.4. If we denote the one-sided limits

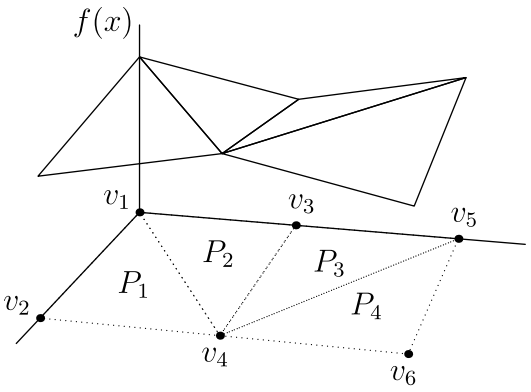



Fig. 9.3: A multivariate continuous piecewise-linear non-convex function.

by $f_-(c) := \lim_{x \uparrow c} f(x)$ and $f_+(c) := \lim_{x \downarrow c} f(x)$, respectively, the one-sided limits, then we can describe the epigraph $f(x) \leq t$ for the function in Fig. 9.4 as

$$\begin{array}{c} x = \lambda_1\alpha_1 + (\lambda_2 + \lambda_3 + \lambda_4)\alpha_2 + \lambda_5\alpha_3, \\ \lambda_1 f(\alpha_1) + \lambda_2 f_-(\alpha_2) + \lambda_3 f(\alpha_2) + \lambda_4 f_+(\alpha_2) + \lambda_5 f(\alpha_3) \leq t, \\ \lambda_1 + \lambda_2 \leq z_1, \quad \lambda_3 \leq z_2, \quad \lambda_4 + \lambda_5 \leq z_3, \\ \lambda \geq 0, \qquad \sum_{i=1}^5 \lambda_i = 1, \qquad \sum_{i=1}^3 z_i = 1, \qquad z \in \{0,1\}^3, \end{array} \tag{9.15}$$

where we have a different decision variable for the intervals $[\alpha_1, \alpha_2)$, $[\alpha_2, \alpha_2]$, and $(\alpha_2, \alpha_3]$. As a special case this gives us an alternative characterization of fixed charge models considered in Sec. 9.1.1.

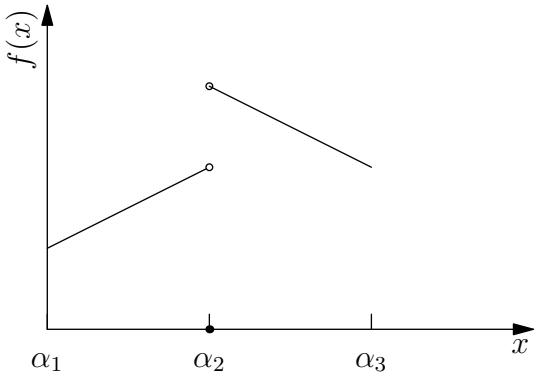



Fig. 9.4: A univariate lower semicontinuous piecewise-linear function.

## 9.2 Mixed integer conic case studies

### 9.2.1 Wireless network design

The following problem arises in wireless network design and some other applications. We want to serve $n$ clients with $k$ transmitters. A transmitter with range $r_j$ can serve all clients

within Euclidean distance $r_j$ from its position. The power consumption of a transmitter with range $r$ is proportional to $r^\alpha$ for some fixed $1 \leq \alpha < \infty$. The goal is to assign locations and ranges to transmitters minimizing the total power consumption while providing coverage of all $n$ clients.

Denoting by $x_1, \ldots, x_n \in \mathbb{R}^2$ locations of the clients, we can model this problem using binary decision variables $z_{ij}$ indicating if the $i$-th client is covered by the $j$-th transmitter. This leads to a mixed-integer conic problem:

$$\begin{array}{lll} \text{minimize} & \sum_j r_j^\alpha & \\ \text{subject to} & r_j \geq \|p_j - x_i\|_2 - M(1 - z_{ij}), & i = 1, \ldots, n,\ j = 1, \ldots, k, \\ & \sum_j z_{ij} \geq 1, & i = 1, \ldots, n, \\ & p_j \in \mathbb{R}^2,\ z_{ij} \in \{0, 1\}. & \end{array} \tag{9.16}$$

The objective can be realized by summing power bounds $t_j \geq r_j^\alpha$ or by directly bounding the $\alpha$-norm of $(r_1, \ldots, r_k)$. The latter approach would be recommended for large $\alpha$.

This is a type of clustering problem. For $\alpha = 1, 2$, respectively, we are minimizing the perimeter and area of the covered region. In practical applications the power exponent $\alpha$ can be as large as 6 depending on various factors (for instance terrain). In the linear cost model ($\alpha = 1$) typical solutions contain a small number of huge disks covering most of the clients. For increasing $\alpha$ large ranges are penalized more heavily and the disks tend to be more balanced.

### 9.2.2 Avoiding small trades

The standard portfolio optimization model admits a number of mixed-integer extensions aimed at avoiding solutions with very small trades. To fix attention consider the model

$$\begin{array}{ll} \text{maximize} & \mu^T x \\ \text{subject to} & (\gamma, G^T x) \in \mathcal{Q}^{n+1}, \\ & e^T x = w + e^T x^0, \\ & x \geq 0, \end{array} \tag{9.17}$$

with initial holdings $x^0$, initial cash amount $w$, expected returns $\mu$, risk bound $\gamma$ and decision variable $x$. Here $e$ is the all-ones vector. Let $\Delta x_j = x_j - x_j^0$ denote the change of position in asset $j$.

#### Transaction costs

A transaction cost involved with nonzero $\Delta x_j$ could be modeled as

$$T_j(x_j) = \begin{cases} 0, & \Delta x_j = 0, \\ \alpha_j \Delta x_j + \beta_j, & \Delta x_j \neq 0, \end{cases}$$

similarly to the problem from Example 9.1. Including transaction costs will now lead to the model:

$$\begin{array}{ll} \text{maximize} & \mu^T x \\ \text{subject to} & (\gamma, G^T x) \in \mathcal{Q}^{n+1}, \\ & e^T x + \alpha^T x + \beta^T z = w + e^T x^0, \\ & x - x^0 \leq Mz,\ x^0 - x \leq Mz, \\ & x \geq 0,\ z \in \{0,1\}^n, \end{array} \tag{9.18}$$

where the binary variable $z_j$ is an indicator of $\Delta x_j \neq 0$. Here $M$ is a sufficiently large constant, for instance $M = w + e^T x^0$ will do.

### Cardinality constraints

Another option is to fix an upper bound $k$ on the number of nonzero trades. The meaning of $z$ is the same as before:

$$\begin{array}{ll} \text{maximize} & \mu^T x \\ \text{subject to} & (\gamma, G^T x) \in \mathcal{Q}^{n+1}, \\ & e^T x = w + e^T x^0, \\ & x - x^0 \leq Mz,\ x^0 - x \leq Mz, \\ & e^T z \leq k, \\ & x \geq 0,\ z \in \{0,1\}^n. \end{array} \tag{9.19}$$

### Trading size constraints

We can also demand a lower bound on nonzero trades, that is $|\Delta x_j| \in \{0\} \cup [a, b]$ for all $j$. To this end we combine the techniques from Sec. 9.1.6 and Sec. 9.1.2 writing $p_j, q_j$ for the indicators of $\Delta x_j > 0$ and $\Delta x_j < 0$, respectively:

$$\begin{array}{ll} \text{maximize} & \mu^T x \\ \text{subject to} & (\gamma, G^T x) \in \mathcal{Q}^{n+1}, \\ & e^T x = w + e^T x^0, \\ & x - x^0 = x^+ - x^-, \\ & x^+ \leq Mp,\ x^- \leq Mq, \\ & a(p+q) \leq x^+ + x^- \leq b(p+q), \\ & p + q \leq e, \\ & x, x^+, x^- \geq 0,\ p, q \in \{0,1\}^n. \end{array} \tag{9.20}$$

## 9.2.3 Convex piecewise linear regression

Consider the problem of approximating the data $(x_i, y_i),\ i = 1, \ldots, n$ by a piecewise linear convex function of the form

$$f(x) = \max\{a_j x + b_j,\ j = 1, \ldots k\},$$

where $k$ is the number of segments we want to consider. The quality of the fit is measured with least squares as

$$\sum_i (f(x_i) - y_i)^2.$$

Note that we do not specify the locations of *nodes* (breakpoints), i.e. points where $f(x)$ changes slope. Finding them is part of the fitting problem.

As in Sec. 9.1.8 we introduce binary variables $z_{ij}$ indicating that $f(x_i) = a_j x_i + b_j$, i.e. it is the $j$-th linear function that achieves maximum at the point $x_i$. Following Sec. 9.1.8 we now have a mixed integer conic quadratic problem

$$\begin{array}{lrcll} \text{minimize} & \|y - f\|_2 & & & \\ \text{subject to} & a_j x_i + b_j & \leq & f_i, & i = 1, \ldots, n,\ j = 1, \ldots, k, \\ & f_i & \leq & a_j x_i + b_j + M(1 - z_{ij}), & i = 1, \ldots, n,\ j = 1, \ldots, k, \\ & \sum_j z_{ij} & = & 1, & i = 1 \ldots, n, \\ & z_{ij} & \in & \{0, 1\}, & i = 1, \ldots, n,\ j = 1, \ldots, k, \end{array} \tag{9.21}$$

with variables $a, b, f, z$, where $M$ is a sufficiently large constant.

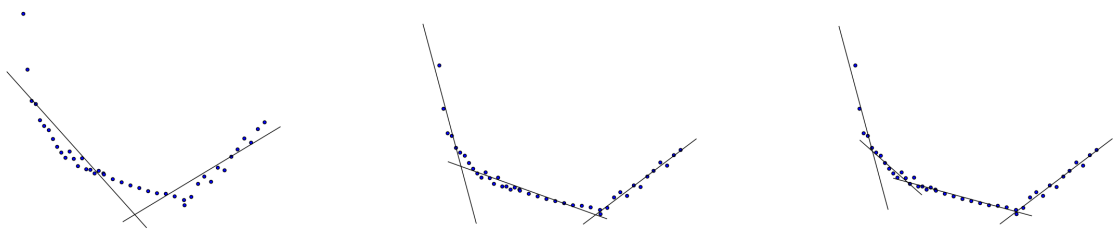


Fig. 9.5: Convex piecewise linear fit with $k = 2, 3, 4$ segments.

Frequently an integer model will have properties which formally follow from the problem's constraints, but may be very hard or impossible for a mixed-integer solver to automatically deduce. It may dramatically improve efficiency to explicitly add some of them to the model. For example, we can enhance (9.21) with all inequalities of the form

$$z_{i,j+1} + z_{i+i',j} \leq 1,$$

which indicate that each linear segment covers a contiguous subset of the sample and additionally force these segments to come in the order of increasing $j$ as $i$ increases from left to right. The last statement is an example of *symmetry breaking*.

# Chapter 10

# Quadratic optimization

In this chapter we discuss convex quadratic and quadratically constrained optimization. Our discussion is fairly brief compared to the previous chapters for three reasons; (*i*) convex quadratic optimization is a special case of conic quadratic optimization, (*ii*) for most convex problems it is actually more computationally efficient to pose the problem in conic form, and (*iii*) duality theory (including infeasibility certificates) is much simpler for conic quadratic optimization. Therefore, we generally recommend a conic quadratic formulation, see Sec. 3 and especially Sec. 3.2.3.

## 10.1 Quadratic objective

A standard (convex) quadratic optimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}x^T Q x + c^T x \\ \text{subject to} & Ax = b, \\ & x \geq 0, \end{array} \tag{10.1}$$

with $Q \in \mathcal{S}^n_+$ is conceptually a simple extension of a standard linear optimization problem with a quadratic term $x^T Q x$. Note the important requirement that $Q$ is symmetric positive semidefinite; otherwise the objective function would not be convex.

### 10.1.1 Geometry of quadratic optimization

Quadratic optimization has a simple geometric interpretation; we minimize a convex quadratic function over a polyhedron., see Fig. 10.1. It is intuitively clear that the following different cases can occur:

- The optimal solution $x^\star$ is at the boundary of the polyhedron (as shown in Fig. 10.1). At $x^\star$ one of the hyperplanes is tangential to an ellipsoidal level curve.
- The optimal solution is inside the polyhedron; this occurs if the unconstrained minimizer $\arg\min_x \frac{1}{2}x^T Q x + c^T x = -Q^\dagger c$ (i.e., the center of the ellipsoidal level curves) is inside the polyhedron. From now on $Q^\dagger$ denotes the pseudoinverse of $Q$; in particular $Q^\dagger = Q^{-1}$ if $Q$ is positive definite.

- If the polyhedron is unbounded in the opposite direction of $c$, and if the ellipsoid level curves are degenerate in that direction (i.e., $Qc = 0$), then the problem is unbounded. If $Q \in \mathcal{S}^n_{++}$, then the problem cannot be unbounded.
- The problem is infeasible, i.e., $\{x \mid Ax = b,\ x \geq 0\} = \emptyset$.

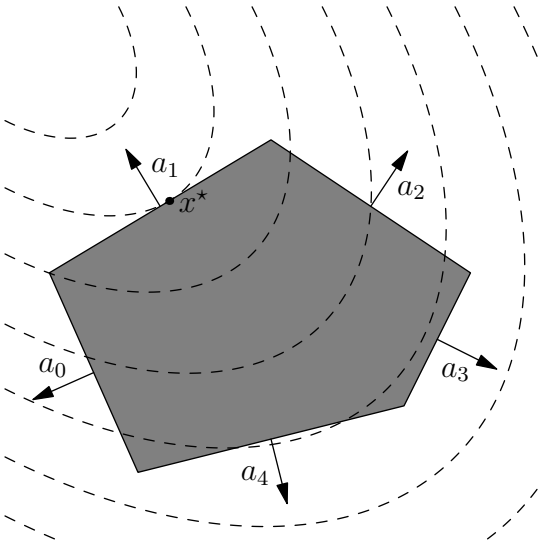



Fig. 10.1: Geometric interpretation of quadratic optimization. At the optimal point $x^\star$ the hyperplane $\{x \mid a_1^T x = b\}$ is tangential to an ellipsoidal level curve.

Possibly because of its simple geometric interpretation and similarity to linear optimization, quadratic optimization has been more widely adopted by optimization practitioners than conic quadratic optimization.

### 10.1.2 Duality in quadratic optimization

The Lagrangian function (Sec. 8.3) for (10.1) is

$$L(x, y, s) = \frac{1}{2} x^T Q x + x^T (c - A^T y - s) + b^T y \tag{10.2}$$

with Lagrange multipliers $s \in \mathbb{R}^n_+$, and from $\nabla_x L(x, y, s) = 0$ we get the necessary first-order optimality condition

$$Qx = A^T y + s - c,$$

i.e. $(A^T y + s - c) \in \mathcal{R}(Q)$. Then

$$\arg\min_x L(x, y, s) = Q^\dagger (A^T y + s - c),$$

which can be substituted into (10.2) to give a dual function

$$g(y, s) = \begin{cases} b^T y - \frac{1}{2}(A^T y + s - c) Q^\dagger (A^T y + s - c), & (A^T y + s - c) \in \mathcal{R}(Q), \\ -\infty & \text{otherwise.} \end{cases}$$

Thus we get a dual problem

$$\begin{array}{ll} \text{maximize} & b^T y - \frac{1}{2}(A^T y + s - c)Q^\dagger(A^T y + s - c) \\ \text{subject to} & (A^T y + s - c) \in \mathcal{R}(Q), \\ & s \geq 0, \end{array} \tag{10.3}$$

or alternatively, using the optimality condition $Qx = A^T y + s - c$ we can write

$$\begin{array}{ll} \text{maximize} & b^T y - \frac{1}{2}x^T Q x \\ \text{subject to} & Qx = A^T y - c + s, \\ & s \geq 0. \end{array} \tag{10.4}$$

Note that this is an unusual dual problem in the sense that it involves both primal and dual variables.

Weak duality, strong duality under Slater constraint qualification and Farkas infeasibility certificates work similarly as in Sec. 8. In particular, note that the constraints in both (10.1) and (10.4) are linear, so Lemma 2.4 applies and we have:

1. The primal problem (10.1) is infeasible if and only if there is $y$ such that $A^T y \leq 0$ and $b^T y > 0$.

2. The dual problem (10.4) is infeasible if and only if there is $x \geq 0$ such that $Ax = 0$, $Qx = 0$ and $c^T x < 0$.

### 10.1.3 Conic reformulation

Suppose we have a factorization $Q = F^T F$ where $F \in \mathbb{R}^{k \times n}$, which is most interesting when $k \ll n$. Then $x^T Q x = x^T F^T F x = \|Fx\|_2^2$ and the conic quadratic reformulation of (10.1) is

$$\begin{array}{ll} \text{minimize} & r + c^T x \\ \text{subject to} & Ax = b, \\ & x \geq 0, \\ & (1, r, Fx) \in \mathcal{Q}_r^{k+2}, \end{array} \tag{10.5}$$

with dual problem

$$\begin{array}{ll} \text{maximize} & b^T y - u \\ \text{subject to} & -F^T v = A^T y - c + s, \\ & s \geq 0, \\ & (u, 1, v) \in \mathcal{Q}_r^{k+2}. \end{array} \tag{10.6}$$

Note that in an optimal primal-dual solution we have $r = \frac{1}{2}\|Fx\|_2^2$, hence the complementary slackness for $\mathcal{Q}_r^{k+2}$ demands $v = -Fx$ and $-F^T F v = Qx$, as well as $u = \frac{1}{2}\|v\|_2^2 = \frac{1}{2}x^T Q x$. This justifies why some of the dual variables in (10.6) and (10.4) have the same names - they are in fact equal, and so both the primal and dual solution to the original quadratic problem can be recovered from the primal-dual solution to the conic reformulation.

## 10.2 Quadratically constrained optimization

A general convex quadratically constrained quadratic optimization problem is

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}x^T Q_0 x + c_0^T x + r_0 \\ \text{subject to} & \frac{1}{2}x^T Q_i x + c_i^T x + r_i \leq 0, \quad i = 1, \ldots, m, \end{array} \tag{10.7}$$

where $Q_i \in \mathcal{S}_+^n$. This corresponds to minimizing a convex quadratic function over an intersection of convex quadratic sets such as ellipsoids or affine halfspaces. Note the important requirement $Q_i \succeq 0$ for all $i = 0, \ldots, m$, so that the objective function is convex and the constraints characterize convex sets. For example, neither of the constraints

$$\frac{1}{2}x^T Q_i x + c_i^T x + r_i = 0, \quad \frac{1}{2}x^T Q_i x + c_i^T x + r_i \geq 0$$

characterize convex sets, and therefore cannot be included.

### 10.2.1 Duality in quadratically constrained optimization

The Lagrangian function for (10.7) is

$$\begin{array}{ll} L(x, \lambda) & = \frac{1}{2}x^T Q_0 x + c_0^T x + r_0 + \sum_{i=1}^m \lambda_i \left(\frac{1}{2}x^T Q_i x + c_i^T x + r_i\right) \\ & = \frac{1}{2}x^T Q(\lambda) x + c(\lambda)^T x + r(\lambda), \end{array}$$

where

$$Q(\lambda) = Q_0 + \sum_{i=1}^m \lambda_i Q_i, \quad c(\lambda) = c_0 + \sum_{i=1}^m \lambda_i c_i, \quad r(\lambda) = r_0 + \sum_{i=1}^m \lambda_i r_i.$$

From the Lagrangian we get the first-order optimality conditions

$$Q(\lambda)x = -c(\lambda), \tag{10.8}$$

and similar to the case of quadratic optimization we get a dual problem

$$\begin{array}{ll} \text{maximize} & -\frac{1}{2}c(\lambda)^T Q(\lambda)^\dagger c(\lambda) + r(\lambda) \\ \text{subject to} & c(\lambda) \in \mathcal{R}\left(Q(\lambda)\right), \\ & \lambda \geq 0, \end{array} \tag{10.9}$$

or equivalently

$$\begin{array}{ll} \text{maximize} & -\frac{1}{2}w^T Q(\lambda) w + r(\lambda) \\ \text{subject to} & Q(\lambda)w = -c(\lambda), \\ & \lambda \geq 0. \end{array} \tag{10.10}$$

Using a general version of the Schur Lemma for singular matrices, we can also write (10.9) as an equivalent semidefinite optimization problem,

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & \begin{bmatrix} 2(r(\lambda) - t) & c(\lambda)^T \\ c(\lambda) & Q(\lambda) \end{bmatrix} \succeq 0, \\ & \lambda \geq 0. \end{array} \tag{10.11}$$

Feasibility in quadratically constrained optimization is characterized by the following conditions (assuming Slater constraint qualification or other conditions to exclude ill-posed problems):

- Either the primal problem (10.7) is feasible or there exists $\lambda \geq 0,\ \lambda \neq 0$ satisfying

$$\sum_{i=1}^{m} \lambda_i \begin{bmatrix} 2r_i & c_i^T \\ c_i & Q_i \end{bmatrix} \succ 0. \tag{10.12}$$

- Either the dual problem (10.10) is feasible or there exists $x \in \mathbb{R}^n$ satisfying

$$Q_0 x = 0, \qquad c_0^T x < 0, \qquad Q_i x = 0, \qquad c_i^T x = 0, \quad i = 1, \ldots, m. \tag{10.13}$$

To see why the certificate proves infeasibility, suppose for instance that (10.12) and (10.7) are simultaneously satisfied. Then

$$0 < \sum_i \lambda_i \begin{bmatrix} 1 \\ x \end{bmatrix}^T \begin{bmatrix} 2r_i & c_i^T \\ c_i & Q_i \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = 2 \sum_i \lambda_i \left( \frac{1}{2} x^T Q_i x + c_i^T x + r_i \right) \leq 0$$

and we have a contradiction, so (10.12) certifies infeasibility.

### 10.2.2 Conic reformulation

If $Q_i = F_i^T F_i$ for $i = 0, \ldots, m$, where $F_i \in \mathbb{R}^{k_i \times n}$ then we get a conic quadratic reformulation of (10.7):

$$\begin{array}{ll} \text{minimize} & t_0 + c_0^T x + r_0 \\ \text{subject to} & (1, t_0, F_0 x) \in \mathcal{Q}_r^{k_0+2}, \\ & (1, -c_i^T x - r_i, F_i x) \in \mathcal{Q}_r^{k_i+2}, \quad i = 1, \ldots, m. \end{array} \tag{10.14}$$

The primal and dual solution of (10.14) recovers the primal and dual solution of (10.7) similarly as for quadratic optimization in Sec. 10.1.3. Let us see for example, how a (conic) primal infeasibility certificate for (10.14) implies an infeasibility certificate in the form (10.12). Infeasibility in (10.14) involves only the last set of conic constraints. We can derive the infeasibility certificate for those constraints from Lemma 8.3 or by directly writing the Lagrangian

$$\begin{array}{rl} L & = -\sum_i \left( u_i + v_i(-c_i^T x - r_i) + w_i^T F_i x \right) \\ & = x^T \left( \sum_i v_i c_i - \sum_i F_i^T w_i \right) + \left( \sum_i v_i r_i - \sum_i u_i \right). \end{array}$$

The dual maximization problem is unbounded (i.e. the primal problem is infeasible) if we have

$$\begin{array}{ll} \sum_i v_i c_i & = \sum_i F_i^T w_i, \\ \sum_i v_i r_i & > \sum_i u_i, \\ 2u_i v_i & \geq \|w_i\|^2, \ u_i, v_i \geq 0. \end{array}$$

We claim that $\lambda = v$ is an infeasibility certificate in the sense of (10.12). We can assume $v_i > 0$ for all $i$, as otherwise $w_i = 0$ and we can take $u_i = 0$ and skip the $i$-th coordinate. Let $M$ denote the matrix appearing in (10.12) with $\lambda = v$. We show that $M$ is positive definite:

$$\begin{aligned}[y,x]^T M[y,x] &= \textstyle\sum_i \left(v_i x^T Q_i x + 2v_i y c_i^T x + 2v_i r_i y^2\right) \\ &> \textstyle\sum_i \left(v_i \|F_i x\|^2 + 2y w_i^T F_i x + 2u_i y^2\right) \\ &\geq \textstyle\sum_i v_i^{-1} \|v_i F_i x + y w_i\|^2 \geq 0.\end{aligned}$$

## 10.3 Example: Factor model

Recall from Sec. 3.3.3 that the standard Markowitz portfolio optimization problem is

$$\begin{array}{ll}\text{maximize} & \mu^T x \\ \text{subject to} & x^T \Sigma x \leq \gamma, \\ & e^T x = 1, \\ & x \geq 0,\end{array} \tag{10.15}$$

where $\mu \in \mathbb{R}^n$ is a vector of expected returns for $n$ different assets and $\Sigma \in \mathcal{S}_+^n$ denotes the corresponding covariance matrix. Problem (10.15) maximizes the expected return of an investment given a budget constraint and an upper bound $\gamma$ on the allowed risk. Alternatively, we can minimize the risk given a lower bound $\delta$ on the expected return of investment, i.e., we can solve the problem

$$\begin{array}{ll}\text{minimize} & x^T \Sigma x \\ \text{subject to} & \mu^T x \geq \delta, \\ & e^T x = 1, \\ & x \geq 0.\end{array} \tag{10.16}$$

Both problems (10.15) and (10.16) are equivalent in the sense that they describe the same Pareto-optimal trade-off curve by varying $\delta$ and $\gamma$.

Next consider a factorization

$$\Sigma = V^T V \tag{10.17}$$

for some $V \in \mathbb{R}^{k \times n}$. We can then rewrite both problems (10.15) and (10.16) in conic quadratic form as

$$\begin{array}{ll}\text{maximize} & \mu^T x \\ \text{subject to} & (1/2, \gamma, Vx) \in \mathcal{Q}_r^{k+2}, \\ & e^T x = 1, \\ & x \geq 0,\end{array} \tag{10.18}$$

and

$$\begin{array}{ll}\text{minimize} & t \\ \text{subject to} & (1/2, t, Vx) \in \mathcal{Q}_r^{k+2}, \\ & \mu^T x \geq \delta, \\ & e^T x = 1, \\ & x \geq 0,\end{array} \tag{10.19}$$

respectively. Given $\Sigma \succ 0$, we may always compute a factorization (10.17) where $V$ is upper-triangular (Cholesky factor). In this case $k = n$, i.e., $V \in \mathbb{R}^{n\times n}$, so there is little difference in complexity between the conic and quadratic formulations. However, in practice, better choices of $V$ are either known or readily available. We mention two examples.

### Data matrix

$\Sigma$ might be specified directly in the form (10.17), where $V$ is a normalized data-matrix with $k$ observations of market data (for example daily returns) of the $n$ assets. When the observation horizon $k$ is shorter than $n$, which is typically the case, the conic representation is both more parsimonious and has better numerical properties.

### Factor model

For a factor model we have

$$\Sigma = D + U^T R U$$

where $D = \mathbf{Diag}(w)$ is a diagonal matrix, $U \in \mathbb{R}^{k\times n}$ represents the exposure of assets to risk factors and $R \in \mathbb{R}^{k\times k}$ is the covariance matrix of factors. Importantly, we normally have a small number of factors ($k \ll n$), so it is computationally much cheaper to find a Cholesky decomposition $R = F^T F$ with $F \in \mathbb{R}^{k\times k}$. This combined gives us $\Sigma = V^T V$ for

$$V = \begin{bmatrix} D^{1/2} \\ FU \end{bmatrix}$$

of dimensions $(n + k) \times n$. The dimensions of $V$ are larger than the dimensions of the Cholesky factors of $\Sigma$, but $V$ is very sparse, which usually results in a significant reduction of solution time. The resulting risk minimization conic problem can ultimately be written as:

$$\begin{array}{ll} \text{minimize} & d + t \\ \text{subject to} & (1/2, t, FUx) \in \mathcal{Q}_r^{k+2}, \\ & (1/2, d, w_1^{1/2} x_1, \cdots, w_n^{1/2} x_n) \in \mathcal{Q}_r^{n+2}, \\ & \mu^T x \geq \delta, \\ & e^T x = 1, \\ & x \geq 0. \end{array} \tag{10.20}$$

# Chapter 11

# Bibliographic notes

The material on linear optimization is very basic, and can be found in any textbook. For further details, we suggest a few standard references *[Chv83]*, *[BT97]* and *[PS98]*, which all cover much more that discussed here. *[NW06]* gives a more modern treatment of both theory and algorithmic aspects of linear optimization.

Material on conic quadratic optimization is based on the paper *[LVBL98]* and the books *[BenTalN01]*, *[BV04]*. The papers *[AG03]*, *[ART03]* contain additional theoretical and algorithmic aspects.

For more theory behind the power cone and the exponential cone we recommend the thesis *[Cha09]*.

Much of the material about semidefinite optimization is based on the paper *[VB96]* and the books *[BenTalN01]*, *[BKVH07]*. The section on optimization over nonnegative polynomials is based on *[Nes99]*, *[Hac03]*. We refer to *[LR05]* for a comprehensive survey on semidefinite optimization and relaxations in combinatorial optimization.

The chapter on conic duality follows the exposition in *[GartnerM12]*

Mixed-integer optimization is based on the books *[NW88]*, *[Wil93]*. Modeling of piecewise linear functions is described in the survey paper *[VAN10]*.

# Chapter 12

# Notation and definitions

$\mathbb{R}$ and $\mathbb{Z}$ denote the sets of real numbers and integers, respectively. $\mathbb{R}^n$ denotes the set of $n$-dimensional vectors of real numbers (and similarly for $\mathbb{Z}^n$ and $\{0,1\}^n$); in most cases we denote such vectors by lower case letters, e.g., $a \in \mathbb{R}^n$. A subscripted value $a_i$ then refers to the $i$-th entry in $a$, i.e.,

$$a = (a_1, a_2, \ldots, a_n).$$

The symbol $e$ denotes the all-ones vector $e = (1, \ldots, 1)^T$, whose length always follows from the context.

All vectors are interpreted as *column-vectors*. For $a, b \in \mathbb{R}^n$ we use the standard inner product,

$$\langle a, b \rangle := a_1 b_1 + a_2 b_2 + \cdots + a_n b_n,$$

which we also write as $a^T b := \langle a, b \rangle$. We let $\mathbb{R}^{m \times n}$ denote the set of $m \times n$ matrices, and we use upper case letters to represent them, e.g., $B \in \mathbb{R}^{m \times n}$ is organized as

$$B = \begin{bmatrix} b_{11} & b_{12} & \ldots & b_{1n} \\ b_{21} & b_{22} & \ldots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \ldots & b_{mn} \end{bmatrix}$$

For matrices $A, B \in \mathbb{R}^{m \times n}$ we use the inner product

$$\langle A, B \rangle := \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ij}.$$

For a vector $x \in \mathbb{R}^n$ we have

$$\mathbf{Diag}(x) := \begin{bmatrix} x_1 & 0 & \ldots & 0 \\ 0 & x_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & x_n \end{bmatrix},$$

i.e., a square matrix with $x$ on the diagonal and zero elsewhere. Similarly, for a square matrix $X \in \mathbb{R}^{n\times n}$ we have

$$\mathbf{diag}(X) := (x_{11}, x_{22}, \ldots, x_{nn}).$$

A set $S \subseteq \mathbb{R}^n$ is *convex* if and only if for any $x, y \in S$ and $\theta \in [0, 1]$ we have

$$\theta x + (1 - \theta)y \in S$$

A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex if and only if its domain $\mathbf{dom}(f)$ is convex and for all $\theta \in [0, 1]$ we have

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y).$$

A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is *concave* if and only if $-f$ is convex. The *epigraph* of a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is the set

$$\mathbf{epi}(f) := \{(x, t) \mid x \in \mathbf{dom}(f),\ f(x) \leq t\},$$

shown in Fig. 12.1.

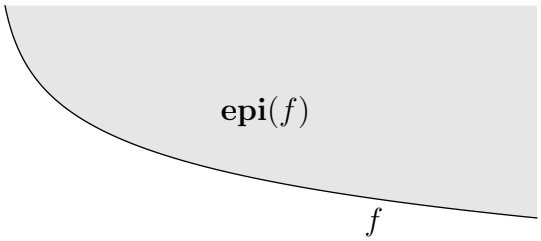



Fig. 12.1: The shaded region is the epigraph of the function $f(x) = -\log(x)$.

Thus, minimizing over the epigraph

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & f(x) \leq t \end{array}$$

is equivalent to minimizing $f(x)$. Furthermore, $f$ is convex if and only if $\mathbf{epi}(f)$ is a convex set. Similarly, the *hypograph* of a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is the set

$$\mathbf{hypo}(f) := \{(x, t) \mid x \in \mathbf{dom}(f),\ f(x) \geq t\}.$$

Maximizing $f$ is equivalent to maximizing over the hypograph

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & f(x) \geq t, \end{array}$$

and $f$ is concave if and only if $\mathbf{hypo}(f)$ is a convex set.

# Bibliography


[AG03] F. Alizadeh and D. Goldfarb. Second-order cone programming. *Math. Programming*, 95(1):3–51, 2003.

[ART03] E. D. Andersen, C. Roos, and T. Terlaky. On implementing a primal-dual interior-point method for conic quadratic optimization. *Math. Programming*, February 2003.

[BT97] D. Bertsimas and J. N. Tsitsiklis. *Introduction to linear optimization*. Athena Scientific, 1997.

[BKVH07] S. Boyd, S.J. Kim, L. Vandenberghe, and A. Hassibi. A Tutorial on Geometric Programming. *Optimization and Engineering*, 8(1):67–127, 2007. Available at http://www.stanford.edu/~boyd/gp_tutorial.html.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004. http://www.stanford.edu/~boyd/cvxbook/.

[CS14] Venkat Chandrasekaran and Parikshit Shah. Conic geometric programming. In *Information Sciences and Systems (CISS), 2014 48th Annual Conference on*, 1–4. IEEE, 2014.

[Cha09] Peter Robert Chares. *Cones and interior-point algorithms for structed convex optimization involving powers and exponentials*. PhD thesis, Ecole polytechnique de Louvain, Universitet catholique de Louvain, 2009.

[Chv83] V. Chvátal. *Linear programming*. W.H. Freeman and Company, 1983.

[GartnerM12] B. Gärtner and J. Matousek. *Approximation algorithms and semidefinite programming*. Springer Science & Business Media, 2012.

[Hac03] Y. Hachez. *Convex optimization over non-negative polynomials: structured algorithms and applications*. PhD thesis, Université Catholique De Lovain, 2003.

[LR05] M. Laurent and F. Rendl. Semidefinite programming and integer programming. *Handbooks in Operations Research and Management Science*, 12:393–514, 2005.

[LVBL98] M. S. Lobo, L. Vanderberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. *Linear Algebra Appl.*, 284:193–228, November 1998.

[NW88] G. L. Nemhauser and L. A. Wolsey. *Integer programming and combinatorial optimization*. John Wiley and Sons, New York, 1988.

[Nes99] Yu. Nesterov. Squared functional systems and optimization problems. In H. Frenk, K. Roos, T. Terlaky, and S. Zhang, editors, *High Performance Optimization*. Kluwer Academic Publishers, 1999.

[NW06] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science, 2nd edition, 2006.

[PS98] C. H. Papadimitriou and K. Steiglitz. *Combinatorial optimization: algorithms and complexity*. Dover publications, 1998.

[TV98] T. Terlaky and J.-Ph. Vial. Computing maximum likelihood estimators of convex density functions. *SIAM J. Sci. Statist. Comput.*, 19(2):675–694, 1998.

[VB96] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Rev.*, 38(1):49–95, March 1996.

[VAN10] J. P. Vielma, S. Ahmed, and G. Nemhauser. Mixed-integer models for nonseparable piecewise-linear optimization: unifying framework and extensions. *Operations research*, 58(2):303–315, 2010.

[Wil93] H. P. Williams. *Model building in mathematical programming*. John Wiley and Sons, 3rd edition, 1993.

[Zie82] H Ziegler. Solving certain singly constrained convex optimization problems in production planning. *Operations Research Letters*, 1982. URL: http://www.sciencedirect.com/science/article/pii/016763778290030X.

[BenTalN01] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. MPS/SIAM Series on Optimization. SIAM, 2001.

# Index

## N



## O



## P



## Q



## R

## S



## T



## U



## V