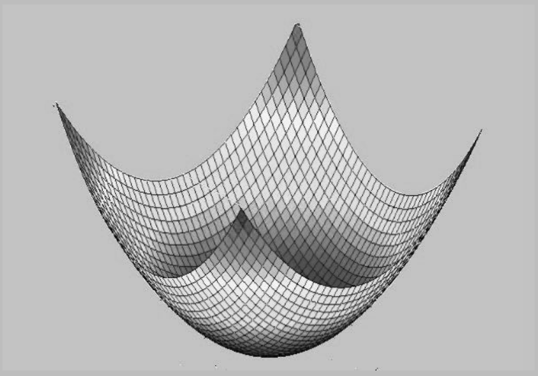


# MATHEMATICAL ESSENTIALS FOR CONVEX OPTIMIZATION

**Fatma Kilinç-Karzan, Tepper School of Business, Carnegie Mellon University**

**Arkadi Nemirovski, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology**

# Contents

# Preface

Convex optimization serves as a cornerstone in various fields of science, engineering, and mathematics, offering powerful tools for solving a wide range of practical problems. With the latest advancements in data sciences and engineering, convex optimization has flourished into a vibrant and rapidly evolving field.

This textbook aims to introduce fundamental theory necessary to establish a robust foundation for doing research on convex optimization. In particular, we have selected to cover both the indispensable basics suited for beginners – rooted in centuries-old research on convexity – as well as modern facets of convex optimization, e.g., cone-constrained conic programming, targeting more advanced readers.

Our emphasis is on foundations and mathematical prerequisites that underpin (primarily Convex) Optimization Theory, and not operational aspects like Modeling and Algorithms. This deliberate choice stems from our desire of emphasizing "timeless and essential classics" and providing an accessible, self-contained, concise, rigorous, yet practical mathematical toolkit. Our goal is to illuminate the entrance to the field of convex optimization, offering readers the background necessary to engage with and comprehend the state-of-the-art "operational" optimization literature, like excellent books [BV04, Nes18]. While applications and algorithms naturally evolve with changing trends and advancements, they will always rely on these timeless foundational blocks. Overall, we view the primary purpose of this book as learning and teaching as opposed to an extensive reference to be kept on shelf by experts.

To an expert in the field: the primary focus of this book is Convex Analysis and the basic theory of convex optimization. Convex Analysis boasts a rich and profound theoretical framework, chronicled in classical treatises such as [Roc70, IT79, HUL93]. However, our approach and presentation style here are tailored to meet the needs of those new to the subject. While we have condensed the scope compared to these classical resources, we have strived to maintain rigor and cover the fundamental descriptive mathematical groundwork that we believe is necessary for state-of-the-art research in Convex Optimization models and algorithms. With regards to convex optimization theory, once again our emphasis is on timeless building blocks such as duality and optimality conditions.

**Our intended audience,** are students, practitioners, researchers with backgrounds in mathematics, operations research, engineering, computer science, data sciences, statistics, and economics.

**As prerequisites** all we assume is basic knowledge of Linear Algebra and Calculus. In fact, we do not anticipate a deep, “ready-to-use” mastery of these subjects either; rather, we expect a basic mathematical culture. To clarify our expectations, consider the following: asserting that $2 \times 2 = 5$ does *not* necessarily indicate a deficiency in mathematical culture; it may simply be a miscalculation. In contrast, claiming that $2 \times 2$ is a triangle or a violin (occasionally encountered, albeit perhaps figuratively rather than literally, in our teaching experience) *does* signify a lack of mathematical culture: under any circumstance, the product of two real numbers should invariably yield another real number and cannot be a triangle or a violin.

Our choice of material is driven by years of experience teaching graduate-level courses on Nonlinear and Convex Optimization. We have organized this material into four main parts:

- basics on convex sets – instructive examples, “calculus” (convexity-preserving operations), main theorem on convex sets such as Caratheodory and Helly theorems, topology of convex sets, “descriptive basics” of Linear Programming (General Theorem on Alternative, Linear Programming duality),
- separation theorem and its applications – extreme points, extreme rays, recessive directions, and (finite-dimensional) Krein-Milman Theorem, structure of polyhedral sets,
- basics on convex functions – instructive examples, detecting convexity, “calculus,” subgradient inequality and preliminaries on subgradients, maxima and minima, Legendre transformation,
- basic theory of Convex Optimization – Lagrange Duality and Lagrange Duality Theorem for problems in standard form and in cone-constrained form, Conic Programming and Conic Duality Theorem, optimality conditions in Mathematical Programming, and convex-concave saddle points and Sion-Kakutani Theorem.

We envision that in a semester-long course on convex optimization, this material may cover about 40% or 60% of the course for building the foundational blocks before moving to other parts (modeling and/or algorithms) of convex optimization. For a course more geared towards linear and/or combinatorial optimization, an instructor may opt to use specific material from the first two parts.

For reader’s convenience, the elementary facts from Linear Algebra, Calculus, Real Analysis, and Matrix Analysis are summarized in the appendices of the textbook (reproducing, courtesy of World Scientific Publishing Co., appendices A – C in [Nem24]). In contrast to the main body of the textbook, these appendices usually do *not* feature accompanying proofs, which are readily available in standard undergraduate textbooks covering the respective subjects (see e.g., [Axl15, Edw12, Gel89, Pas22, Rud13, Str06]). A well-prepared reader may opt for a "fast-forward" approach by initially reviewing these appendices before delving into the main body of the book. Alternatively, they may commence their reading journey from Part 1, referring back to the appendices as necessary.

Certain sections in our text, with titles starting with ★, delve into more advanced and specialized topics, such as Conic, Perspective, and Legendre transforms, Majorization, Cone-convexity, Cone-monotonicity, among others. Although these starred sections hold significance in their respective domains, they are designed to be optional and can be skipped over depending on the goals of the reader.

**Our exposition** adheres to the usual standards of rigor needed to present mathematical subjects. Accordingly, we provide complete formal proofs for all of the theorems, propositions, lemmas, and the like. In addition to these, we include formal statements of similar nature labeled as "Facts" scattered throughout a chapter, but without accompanying proofs. The claims made in these "Facts" are also compulsory part of our exposition, and their knowledge is as mandatory for mastering the material as the knowledge of theorems, propositions, etc. Nonetheless, the statements within "Facts" are sufficiently elementary to be easily verified by a diligent reader. In essence, "Facts" serve as embedded exercises, and we firmly believe that engaging with these exercises as they appear in the text provides valuable hands-on practice for effective learning. This active participation is an indispensable facet of mastering the presented material and honing mathematical skills. At the same time, we recognize the importance of providing access to detailed self-contained proofs of "Facts." To this end, nearly all[1] "Facts" are repeated, this time with accompanying proofs, at the end of the respective parts.

**The exercises** are crafted to align with our educational objective of fostering hands-on learning and providing ample practice opportunities at various difficulty levels. In particular, they are categorized into three types. The traditional "Test Yourself" exercises enable readers to evaluate their grasp of the material presented in the main body of the textbook. In addition to these, we also offer "Try Yourself" type of exercises which typically require readers to prove something, aimed at fostering and assessing their creative skills. Finally, our "Educate Yourself" exercises address topics that we deem significant, extending beyond the core material covered in the textbook's main body. An illustrative example for this type of exercises is the investigation of conic representations of convex sets and functions, along with the calculus associated with these representations. This plays a crucial role in formulating and solving "well-structured" convex optimization problems, particularly those involving Second Order Conic and Semidefinite programs. We provide a separate solution manual for "Try Yourself" and "Educate Yourself" exercises.

**Acknowledgements.** The main body of this textbook existed for about 25 years,

[1] few exceptions are absolutely straightforward and presented as Facts solely for the sake of further references.

in a more restricted form, as appendices to the graduate course on Modern Convex Optimization taught by the second author first at TU Delft (1998) and then at Georgia Institute of Technology (since 2003). The first author was fortunate to have been a student at Georgia Tech thoroughly enjoying this material, and later on she adopted this material and has been teaching a similar course at Carnegie Mellon University (since 2012). These appendices originate from the descriptive part of graduate course "Optimization II" designed in 1980's by Prof. Aharon Ben-Tal and for over 20 years taught by him at the Technion – Israel Institute of Technology. This course was "inherited" and taught, in re-designed form, by the second author first at the Technion, and then - at Georgia Institute of Technology. It is our pleasure to acknowledge hereby, with the deepest gratitude, the instrumental role played by Prof. Ben-Tal in selecting and structuring significant part of the material to follow. Besides this, we are greatly indebted to Dr. Sergei Gelfand for the idea to convert these appendices into a "standalone" textbook.

Fatma Kilinç-Karzan, Tepper Business School, Carnegie Mellon University
Arkadi Nemirovski, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology
February 2024

## Main Notational Conventions

$\mathbf{N}, \mathbf{Z}, \mathbf{Q}, \mathbf{R}, \mathbf{C}$ stand for the set of all, respectively, nonnegative integers, integers, rational numbers, real numbers, and complex numbers.

**Vectors and matrices.** By default, *all vectors are column vectors.*

- The space of all $n$-dimensional vectors with real entries is denoted by $\mathbf{R}^n$, the set of all $m \times n$ matrices with real entries is denoted by $\mathbf{R}^{m\times n}$; notation $\mathbf{N}^n, \dots, \mathbf{C}^{m\times n}$ is interpreted similarly, where we restrict the entries to belong to the respective number domains. The set of symmetric $n \times n$ matrices is denoted by $\mathbf{S}^n$. By default, all vectors and matrices have real entries, and when speaking about $\mathbf{R}^n$ and $\mathbf{S}^n$ ($\mathbf{R}^{m\times n}$), $n$ ($m$ and $n$) are *positive* integers.

  By default, notation like $x_i$ (or $y_k$) refers to $i$-th entry of context-specified vector $x$ (or $k$-th entry of context-specified vector $y$). Similarly, notation like $x_{ij}$ (or $Y_{k\ell}$) refers to $(i,j)$-th entry of context-specified matrix $x$ (or $(k,\ell)$-th entry of context-specified matrix $Y$).
- Sometimes, "MATLAB notation" is used to save space: a vector with coordinates $x_1, \dots, x_n$ is written down as

$$x = [x_1; \dots; x_n]$$

  (pay attention to semicolon ";"). For example, $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ is written as $[1; 2; 3]$.

  More generally,
  — if $A_1, \dots, A_m$ are matrices with the same number of columns, we write $[A_1; \dots; A_m]$ to denote the matrix obtained by writing $A_2$ beneath $A_1$, $A_3$ beneath $A_2$, and so on.
  — if $A_1, \dots, A_m$ are matrices with the same number of rows, then $[A_1, \dots, A_m]$ stands for the matrix obtained by writing $A_2$ to the right of $A_1$, $A_3$ to the right of $A_2$, and so on.

  **Examples:**

  - $A_1 = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$, $A_2 = \begin{bmatrix} 7 & 8 & 9 \end{bmatrix} \implies [A_1; A_2] = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$
  - $A_1 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $A_2 = \begin{bmatrix} 7 \\ 8 \end{bmatrix} \implies [A_1, A_2] = \begin{bmatrix} 1 & 2 & 7 \\ 4 & 5 & 8 \end{bmatrix}$
  - $[1, 2, 3, 4] = [1; 2; 3; 4]^\top$
  - $$\begin{aligned}[[1,2;3,4],[5,6;7,8]] &= \left[\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}\right] = \begin{bmatrix} 1 & 2 & 5 & 6 \\ 3 & 4 & 7 & 8 \end{bmatrix} \\ &= [1,2,5,6;3,4,7,8]\end{aligned}$$
- We follow the standard convention that the sum of vectors over an empty set of indexes, i.e., $\sum_{i=1}^{0} x^i$, where $x^i \in \mathbf{R}^n$, has a value – it is the origin in $\mathbf{R}^n$.

**Intervals in $\mathbf{R}^n$.** Given two vectors $x, y \in \mathbf{R}^n$, we use the notation $[x, y]$ to denote the *segment* in $\mathbf{R}^n$ that connects $x$ and $y$, where both endpoints are included, i.e., $[x, y] := \{\lambda x + (1-\lambda)y : 0 \leq \lambda \leq 1\}$. Similarly, we define the open segment $(x, y) := \{\lambda x + (1-\lambda)y : 0 < \lambda < 1\}$ in $\mathbf{R}^n$ without the endpoints.

**Semidefinite order.** Relations $A \succeq B$, $B \preceq A$, $A - B \succeq 0$, $B - A \preceq 0$ all mean the same, namely, that $A, B$ are real symmetric matrices of the same size such that the difference $A - B$ is positive semidefinite. Positive definiteness of the difference of $A - B$ is expressed by every one of the relations $A \succ B$, $B \prec A$, $A - B \succ 0$, $B - A \prec 0$.

**Diag and Dg.** For $x \in \mathbf{R}^n$, $\mathrm{Diag}\{x\}$ stands for diagonal $n \times n$ matrix with the entries of $x$ on the diagonal. For a collection $X_1, \ldots, X_K$ of rectangular matrices, $\mathrm{Diag}\{X_1, \ldots, X_k\} = \left[\begin{array}{c|c|c} X_1 & & \\ \hline & \ddots & \\ \hline & & X_k \end{array}\right]$ stands for block-diagonal matrix with diagonal blocks $X_1, \ldots, X_K$. For an $n \times n$ matrix $X$, $\mathrm{Dg}\{X\}$ stands for $n$-dimensional vector composed of diagonal entries of $X$.

**Extended real axis.** We follow the standard conventions on operations of summation, multiplication, and comparison in the "extended real line" $\mathbf{R} \cup \{+\infty\} \cup \{-\infty\}$. These conventions are as follows:

- Operations with real numbers are understood in their usual sense.
- The sum of $+\infty$ and a real number, same as the sum of $+\infty$ and $+\infty$ is $+\infty$. Similarly, the sum of a real number and $-\infty$, same as the sum of $-\infty$ and $-\infty$ is $-\infty$. The sum of $+\infty$ and $-\infty$ is undefined.
- The product of a real number and $+\infty$ is $+\infty$, 0 or $-\infty$, depending on whether the real number is positive, zero or negative, and similarly for the product of a real number and $-\infty$. The product of two "infinities" is again infinity, with the usual rule for assigning the sign to the product.
- Finally, any real number is $< +\infty$ and $> -\infty$, and of course $-\infty < \infty$.

**Abbreviations.** From time to time we use the following abbreviations:
a.k.a. for "also known as"
iff for "if and only if"
w.l.o.g. for "without loss of generality"
w.r.t. for "with respect to"

**Symbols ▲ and ♦** mark respectively "try yourself" and "educate yourself" exercises.

# Part I

# Convex sets in $\mathbf{R}^n$ – From First Acquaintance to Linear Programming Duality

# 1

# First acquaintance with convex sets

### 1.1 Definition and examples

In the school geometry a figure is called *convex* if it contains, along with every pair of its points $x, y$, also the entire segment $[x, y]$ linking the points. This is exactly the definition of a convex set in the multidimensional case; all we need is to say what "the segment $[x, y]$ linking the points $x, y \in \mathbf{R}^n$" is. We state this formally in the following definition.

**Definition** I.1.1 [Convex set]
1) Let $x, y$ be two points in $\mathbf{R}^n$. The set

$$[x, y] := \{\lambda x + (1-\lambda)y : \ 0 \leq \lambda \leq 1\}$$

is called a *segment* with the endpoints $x, y$.
2) A subset $M$ of $\mathbf{R}^n$ is called *convex*, if it contains, along with every pair of its points $x, y$, also the entire segment $[x, y]$:

$$x, y \in M, \ 0 \leq \lambda \leq 1 \implies \lambda x + (1-\lambda)y \in M.$$

The definition of a segment $[x; y]$ is in full accordance with our "real life experience" in 2D or 3D: when $\lambda \in [0, 1]$, the point $x(\lambda) = \lambda x + (1-\lambda)y = x + (1-\lambda)(y - x)$ is the point where you arrive when traveling from $x$ directly towards $y$ after you have covered the fraction $(1-\lambda)$ of the entire distance from $x$ to $y$, and these points compose the "real world segment" with endpoints $x = x(1)$ and $y = x(0)$.

Note that an empty set is convex by the exact sense of the definition: for the empty set, you cannot present a counterexample to show that it is not convex.

A closed *ray* given by a direction $0 \neq d \in \mathbf{R}^n$ is also convex:

$$\mathbf{R}_+(d) := \{t\, d \in \mathbf{R}^n : \ t \geq 0\}\,.$$

Note also that the open ray given by $\{t\, d \in \mathbf{R}^n : \ t > 0\}$ is convex as well.

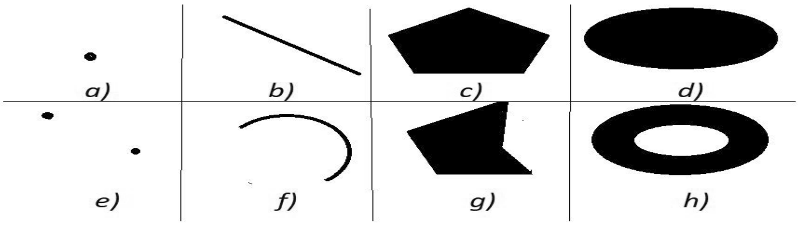


Figure I.1. *a – d)*: convex sets; *e – h)*: nonconvex sets

We next continue with a number of examples of convex sets.

### *1.1.1 Affine subspaces and polyhedral sets*

We start with a simple and important fact.

**Proposition** I.1.2 The solution set of an arbitrary (possibly, infinite) system

$$a_\alpha^\top x \leq b_\alpha,\ \alpha \in \mathcal{A} \tag{1.1}$$

of nonstrict linear inequalities with $n$ unknowns $x$, i.e., the set

$$S := \left\{x \in \mathbf{R}^n : a_\alpha^\top x \leq b_\alpha,\ \alpha \in \mathcal{A}\right\}$$

is convex.

**Proof.** Consider any $x', x'' \in S$ and any $\lambda \in [0,1]$. As $x', x'' \in S$, we have $a_\alpha^\top x' \leq b_\alpha$ and $a_\alpha^\top x'' \leq b_\alpha$ for any $\alpha \in \mathcal{A}$. Then, for every $\alpha \in \mathcal{A}$, multiplying the inequality $a_\alpha^\top x' \leq b_\alpha$ by $\lambda$, and the inequality $a_\alpha^\top x'' \leq b_\alpha$ by $1-\lambda$, respectively, and summing up the resulting inequalities, we get $a_\alpha^\top[\lambda x' + (1-\lambda)x''] \leq b_\alpha$. Thus, we deduce that $\lambda x' + (1-\lambda)x'' \in S$. □

Note that this verification of convexity of $S$ works also in the case when in the definition of $S$ some of nonstrict inequalities $a_\alpha^\top x \leq b_\alpha$ are replaced with their strict versions $a_\alpha^\top x < b_\alpha$.

Recall that linear and affine subspaces can be represented as the solution sets of systems of linear equations (Proposition A.47). Consequently, from Proposition I.1.2 we deduce that such sets are convex.

**Example** I.1.1 All linear subspaces and all affine subspaces of $\mathbf{R}^n$ are convex.

Another important special case of Proposition I.1.2 is the one when we have a finite system of nonstrict linear inequalities. Such sets have a special name as they are frequently encountered and studied.

**Definition** I.1.3 [Polyhedral set] A set in $\mathbf{R}^n$ is called *polyhedral* if it is the solution set of a finite system

$$Ax \le b$$

of $m$ nonstrict linear inequalities with $n$ variables (i.e., $A$ is an $m \times n$ matrix) for some nonnegative integer $m$.

Based on this definition and as an immediate consequence of Proposition I.1.2, we arrive at our second generic example of convex sets.

**Example** I.1.2 Any polyhedral set in $\mathbf{R}^n$ is convex.

**Remark** I.1.4 Note that every set given by Proposition I.1.2 is not only convex, but also closed (why?). In fact, Separation Theorem (see Theorem II.7.3) implies the following:

> *Every closed convex set in $\mathbf{R}^n$ is the solution set of an infinite system $a_i^\top x \le b_i$, $i = 1, 2, \ldots$, of nonstrict linear inequalities.*

**Remark** I.1.5 Replacing some of the nonstrict linear inequalities $a_\alpha^\top x \le b_\alpha$ in system (1.1) with their strict versions $a_\alpha^\top x < b_\alpha$ preserves, as we have already mentioned, convexity of the solution set, but can destroy its closedness.

### *1.1.2 Unit balls of norms*

Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$ i.e., a real-valued function on $\mathbf{R}^n$ satisfying the three characteristic properties of a norm (section B.1.1), specifically:

1. *Positivity:* $\|x\| \ge 0$ for all $x \in \mathbf{R}^n$, and $\|x\| = 0$ if and only if $x = 0$;
2. *Homogeneity:* For $x \in \mathbf{R}^n$ and $\lambda \in \mathbf{R}$, we have $\|\lambda x\| = |\lambda|\|x\|$;
3. *Triangle inequality:* For all $x, y \in \mathbf{R}^n$, we have $\|x + y\| \le \|x\| + \|y\|$.

**Fact** I.1.6 The *unit ball of a norm* $\|\cdot\|$, i.e., the set

$$\{x \in \mathbf{R}^n : \ \|x\| \le 1\},$$

same as every other $\|\cdot\|$-ball

$$B_r(a) := \{x \in \mathbf{R}^n : \|x - a\| \le r\},$$

(here $a \in \mathbf{R}^n$ and $r \ge 0$ are fixed) is convex.

In particular, Euclidean balls ($\|\cdot\|$-balls associated with the standard Euclidean norm $\|x\|_2 := \sqrt{x^\top x}$) are convex.

The standard examples of norms on $\mathbf{R}^n$ are the $\ell_p$-norms

$$\|x\|_p = \begin{cases} \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}, & \text{if } 1 \le p < \infty \\ \max\limits_{1 \le i \le n} |x_i|, & \text{if } p = \infty. \end{cases}$$

These indeed are norms (which is not clear in advance; for proof, see page 168, and for more details – page 214). When $p = 2$, we get the usual Euclidean norm. When $p = 1$, we get

$$\|x\|_1 = \sum_{i=1}^n |x_i|,$$

and its unit ball is the *hyperoctahedron*

$$V = \left\{ x \in \mathbf{R}^n : \ \sum_{i=1}^n |x_i| \le 1 \right\}.$$

When $p = \infty$, we get

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

and its unit ball is the *hypercube*

$$V = \{ x \in \mathbf{R}^n : \ -1 \le x_i \le 1, \, 1 \le i \le n \},$$

see Figure I.2.

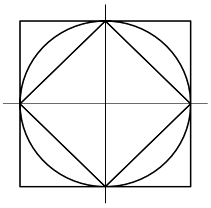


Figure I.2. $\|\cdot\|_p$-balls in 2D, $p = 1, 2, \infty$.

**Remark** I.1.7 As we have already mentioned, the fact that the $\ell_p$ norms, $1 \le p \le \infty$, indeed are norms, is not completely trivial and will be proved in full generality later. What is evident, is that $\|\cdot\|_p$ does possess properties of positivity and homogeneity; what requires effort, is the triangle inequality. There are, however, two special cases, i.e., $p = 1$ and $p = \infty$, where this inequality is easy. Indeed, from high school you know that for reals $a, b$ it always holds $|a + b| \le |a| + |b|$. It follows that

$$\|x + y\|_1 = \sum_i |x_i + y_i| \le \sum_i (|x_i| + |y_i|) = \sum_i |x_i| + \sum_i |y_i| = \|x\|_1 + \|y\|_1,$$

and

$$\begin{aligned} \|x + y\|_\infty = \max_i |x_i + y_i| \le \max_i \{|x_i| + |y_i|\} \le \max_{i,j} \{|x_i| + |y_j|\} \\ = \max_i |x_i| + \max_j |y_j| = \|x\|_\infty + \|y\|_\infty. \end{aligned}$$

Triangle inequality for Euclidean norm $\|\cdot\|_2$ should be already known to the reader; this is an immediate consequence of Cauchy-Schwarz inequality $|x^\top y| \le \|x\|_2 \|y\|_2$, see section B.1.1.

**Fact** I.1.8 Unit balls of norms on $\mathbf{R}^n$ are exactly the same as convex sets $V$ in $\mathbf{R}^n$ satisfying the following three properties:

(i) $V$ is symmetric with respect to the origin: $x \in V \implies -x \in V$;
(ii) $V$ is bounded and closed;
(iii) $V$ contains a neighborhood of the origin, i.e., there exists $r > 0$ such that the centered at the origin Euclidean ball of radius $r$ – the set $\{x \in \mathbf{R}^n : \|x\|_2 \leq r\}$ – is contained in $V$.

Any set $V$ satisfying the outlined properties is indeed the unit ball of a particular norm given by

$$\|x\|_V := \inf_t \left\{t : \ t^{-1}x \in V, \ t > 0\right\}. \tag{1.2}$$

### *1.1.3 Ellipsoids*

**Fact** I.1.9 Let $Q$ be an $n \times n$ matrix which is symmetric (i.e., $Q = Q^\top$) and positive definite (i.e., $x^\top Qx > 0$ for all $x \neq 0$). Then, for every nonnegative $r$, the *$Q$-ellipsoid of radius $r$ centered at $a$*, i.e., the set

$$\left\{x \in \mathbf{R}^n : \ (x-a)^\top Q(x-a) \leq r^2\right\}$$

is convex.

### *1.1.4 Neighborhood of a convex set*

**Example** I.1.3 Let $M$ be a nonempty convex set in $\mathbf{R}^n$, and let $\epsilon > 0$. Then, for every norm $\|\cdot\|$ on $\mathbf{R}^n$, the $\epsilon$-neighborhood of $M$, i.e., the set

$$M_\epsilon := \left\{y \in \mathbf{R}^n : \ \inf_{x \in M} \|y - x\| \ \leq \epsilon\right\}$$

is convex.

Justification of Example I.1.3 is left as an exercise at the end of this Part (see Exercise I.6).

## 1.2 Inner description of convex sets: convex combinations and convex hull

### *1.2.1 Convex combinations*

Recall the notion of *linear combination* $x$ of vectors $x^1, \ldots, x^m$; this is a vector represented as

$$x = \sum_{i=1}^m \lambda_i x^i,$$

where $\lambda_i \in \mathbf{R}$ are the coefficients. By including a specific restriction on which coefficients can be used in this definition, we arrive at important special types of linear combinations. For example, an *affine combination* is a linear combination where the sum of the coefficients is equal to 1. Given a nonempty set $X$, the smallest (w.r.t. inclusion) affine plane containing $X$ is composed of all affine combinations of the points of $X$, see section A.4.2. Another beast in this genre is *convex combination.*

**Definition** I.1.10 [Convex combination] A *convex combination* of vectors $x^1, \ldots, x^m$ is a linear combination

$$x := \sum_{i=1}^{m} \lambda_i x^i,$$

with nonnegative coefficients summing up to 1:

$$\lambda_i \geq 0, \, \forall i = 1, \ldots, m, \quad \sum_{i=1}^{m} \lambda_i = 1.$$

Equivalently, convex combination is an affine combination with nonnegative coefficients.

By Linear Algebra, a nonempty set $X \subseteq \mathbf{R}^n$ is a linear (or an affine) subspace if and only if $X$ is closed with respect to taking all linear, respectively, all affine combinations of its elements. Convex combinations play similar role when speaking about convex sets.

**Fact** I.1.11 A set $M \subseteq \mathbf{R}^n$ is convex if and only if it is closed with respect to taking all convex combinations of its elements. That is, $M$ is convex if and only if every convex combination of vectors from $M$ is again a vector from $M$.
*Hint:* Note that assuming $\lambda_1, \ldots, \lambda_m > 0$, one has

$$\sum_{i=1}^{m} \lambda_i x^i = \lambda_1 x^1 + (\lambda_2 + \lambda_3 + \ldots + \lambda_m) \sum_{i=2}^{m} \mu_i x^i, \quad \text{where } \mu_i := \frac{\lambda_i}{\lambda_2 + \lambda_3 + \ldots + \lambda_m}.$$

(cf. Corollary A.39).

### *1.2.2 Convex hull*

Recall that taking the intersection of linear subspaces results in another linear subspace. The same property holds true for convex sets as well (why?).

**Proposition** I.1.12 Let $\{M_\alpha\}_\alpha$ be an arbitrary family of convex subsets of $\mathbf{R}^n$. Then, their intersection

$$M = \bigcap_{\alpha} M_\alpha$$

is also convex.

As an immediate consequence of Proposition I.1.12, we come to the notion of *convex hull* $\mathrm{Conv}(M)$ of a subset $M \subseteq \mathbf{R}^n$ (cf. the notions of linear/affine span):

**Definition** I.1.13 [Convex hull] For any $M \subseteq \mathbf{R}^n$, the *convex hull* of $M$ [notation: $\mathrm{Conv}(M)$] is the intersection of all convex sets containing $M$ (and thus, by Proposition 1.2.2 $\mathrm{Conv}(M)$ is the smallest (w.r.t. inclusion) convex set containing $M$).

By Linear Algebra, the linear span of a set $M$ – the smallest (w.r.t. inclusion) linear subspace containing $M$ – can be described in terms of linear combinations: this is the set of all linear combinations of points from $M$. Analogous results hold for affine span of (nonempty) set and affine combinations of points from the set as well. We have an analogous description of convex hulls via convex combinations as well:

**Fact** I.1.14 [Convex hull via convex combinations] For a set $M \subseteq \mathbf{R}^n$,

$$\mathrm{Conv}(M) = \{\text{the set of all convex combinations of vectors from } M\}.$$

We will see in chapter 10 that when $M$ is a finite set in $\mathbf{R}^n$, $\mathrm{Conv}(M)$ is a bounded polyhedral set. Bounded polyhedral sets are also called *polytopes.*

We next continue with a number of important families of convex sets.

### *1.2.3 Simplex*

**Definition** I.1.15 [Simplex] The convex hull of $m+1$ affinely independent points $x^0, \ldots, x^m$ is called the *$m$-dimensional simplex with the vertices* $x^0, \ldots, x^m$. (See section A.4.3 for affine independence.)

Consider an $m$-dimensional simplex with vertices $x^0, \ldots, x^m$. Then, based on section A.4.3, every point $x$ from this simplex admits exactly one representation as a convex combination of these vertices. The coefficients $\lambda_i$, $i = 0, \ldots, m$, used in the convex combination representation of $x$ form the unique solution to the system of linear equations given by

$$\sum_{i=0}^{m} \lambda_i x^i = x, \quad \sum_{i=0}^{m} \lambda_i = 1.$$

This system in variables $\lambda_i$ is feasible if and only if $x \in M = \mathrm{Aff}(\{x^0, \ldots, x^m\})$, and the components of the solution (the barycentric coordinates of $x$) are affine functions of $x \in \mathrm{Aff}(M)$. The simplex itself is composed of points from $M$ with nonnegative barycentric coordinates.

### *1.2.4 Cones*

We next examine a very important class of convex sets.

A nonempty set $K \subseteq \mathbf{R}^n$ is called *conic* if it contains, along with every point $x \in K$, the entire ray $\mathbf{R}_+(x) = \{tx : t \geq 0\}$ spanned by the point:

$$x \in K \quad \Longrightarrow \quad tx \in K, \ \forall t \geq 0.$$

**Definition** I.1.16 [Cone] A *cone* is a nonempty, *convex*, and *conic* set.

**Fact** I.1.17 A set $K \subseteq \mathbf{R}^n$ is a cone if and only if it is nonempty and

- is conic, i.e., $x \in K, t \geq 0 \implies tx \in K$; and
- contains sums of its elements, i.e., $x, y \in K \implies x + y \in K$.

**Example** I.1.4 The solution set of an arbitrary (possibly, infinite) system of *homogeneous* linear inequalities with $n$ unknowns $x$, i.e., the set

$$K = \left\{x \in \mathbf{R}^n : \ a_\alpha^\top x \geq 0, \ \forall \alpha \in \mathcal{A}\right\},$$

is a cone.

In particular, the solution set of a *finite* system composed of $m$ homogeneous linear inequalities

$$Ax \geq 0$$

($A$ is $m \times n$ matrix) is a cone. A cone of this latter type is called *polyhedral*[1].

Specifically, the *nonnegative orthant* $\mathbf{R}^m_+ := \{x \in \mathbf{R}^m : \ x \geq 0\}$ is a polyhedral cone.

Note that the cones given by systems of linear homogeneous nonstrict inequalities are obviously closed. From Separation Theorem (see Theorem II.7.3) we will deduce the reverse as well, i.e., every closed cone is the solution set to such a system. Thus, Example I.1.4 is the generic example of a closed convex cone.

We already know that a norm $\|\cdot\|$ on $\mathbf{R}^n$ gives rise to specific convex sets in $\mathbf{R}^n$, namely, balls of this norm. In fact, a norm also gives rise to another important convex set.

**Proposition** I.1.18 For any norm $\|\cdot\|$ on $\mathbf{R}^n$, its epigraph, i.e., the set

$$K := \left\{[x; t] \in \mathbf{R}^{n+1} : \ t \geq \|x\|\right\}$$

is a closed cone in $\mathbf{R}^{n+1}$.

[1] The "literal" interpretation of the words "polyhedral cone" should be "a set of the form $\{x : Ax \leq b\}$ which is a cone;" this is not exactly the terminology just introduced. Luckily, there is no collision: a *polyhedral set $X = \{x : Ax \leq b\}$ is a cone if and only if $X = \{x : Ax \leq 0\}$*, see Exercise II.32.

**Proof.** Obviously, $K$ is nonempty as $[x;t]=[0;0]$ is in $K$. Also, $K$ is a conic set as any norm $\|\cdot\|$ is positively homogeneous. Moreover, the closedness of $K$ with respect to summation is readily given by the Triangle inequality: consider two points $[x;t]\in K$ and $[x';t']\in K$. Then, $t\geq\|x\|$ and $t'\geq\|x'\|$ which imply $t{+}t'\geq\|x\|{+}\|x'\|\geq\|x{+}x'\|$. Thus, $[x{+}x';t{+}t']\in K$. Invoking Fact I.1.17, we see that $K$ is a cone. In order to see that $K$ is closed recall that $\|\cdot\|$ is continuous (see Fact B.23). Thus, for any sequence of points $[x^i;t_i]\in K$ converging to a point $[x;t]$ as $i\to\infty$, we have $[\|x^i\|;t_i]\to[\|x\|;t]$ and therefore $t\geq\|x\|$. This establishes that the limit of any converging sequence from $K$ belongs to $K$, proving that $K$ is closed. □

A particular case of Proposition I.1.18 states that the epigraph of Euclidean norm, i.e.,

$$\mathbf{L}^m := \left\{[x;t]\in\mathbf{R}^{n+1}:\ t\geq\|x\|_2\right\},$$

is a closed cone. This is the *second-order*, (or *Lorentz* or *ice cream*) cone (see Figure I.3), and it plays a significant role in convex optimization.

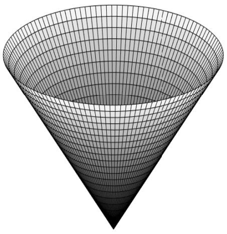


Figure I.3. [Boundary of] 3D Lorentz cone $\mathbf{L}^3$

To complete our first acquaintance with cones, we also mention the *semidefinite cone* $\mathbf{S}^m_+$ “living” in the space $\mathbf{S}^m$ of real symmetric $m\times m$ matrices and composed of positive semidefinite matrices from $\mathbf{S}^m$, i.e.,

$$\mathbf{S}^m_+ := \left\{X\in\mathbf{R}^{m\times m}:\ X=X^\top,\ a^\top Xa\geq 0,\ \forall a\in\mathbf{R}^m\right\};$$

see section D.2.2.

Cones form a very important family of convex sets, and one can develop theory of cones absolutely similar (and in a sense, equivalent) to that of all convex sets. For example, by introducing the notion of *conic combination* of vectors $x^1,\dots,x^k$ as a linear combination of these vectors with *nonnegative* coefficients, we can easily prove the following statements completely similar to those for general convex sets, with conic combination playing the role of convex ones:

- A set is a cone if and only if it is nonempty and is closed with respect to taking conic combinations of its elements;

- Intersection of a family of cones is again a cone; in particular, for every set $K\subseteq\mathbf{R}^n$ there exists the smallest (w.r.t. inclusion) cone containing $K$, called the *conic hull* of $K$:

**Definition** I.1.19 [Conic hull] For any $K \subseteq \mathbf{R}^n$, the *conic hull* of $K$ [notation: $\mathrm{Cone}(K)$] is the intersection of all cones containing $K$. Thus, $\mathrm{Cone}(K)$ is the smallest (w.r.t. inclusion) cone containing $K$.

• We can describe the conic hull of a set $K \subseteq \mathbf{R}^n$ in terms of its conic combinations:

**Fact** I.1.20 [Conic hull via conic combinations] The conic hull $\mathrm{Cone}(K)$ of a set $K \subseteq \mathbf{R}^n$ is the set of all conic combinations (i.e., linear combinations with nonnegative coefficients) of vectors from $K$:

$$\mathrm{Cone}(K) = \left\{ x \in \mathbf{R}^n : \exists N \geq 0, \lambda_i \geq 0, x^i \in K, i \leq N : x = \sum_{i=1}^{N} \lambda_i x^i \right\}.$$

Note that here we use the standard convention: the sum of vectors over an empty set of indexes, like $\sum_{i=1}^{0} z^i$, has a value – it is the origin of the space where vectors live. In particular, the set of conic combinations of vectors from empty set is $\{0\}$, in full accordance with Definition I.1.19.

## 1.3 Calculus of convex sets

Calculus of convex sets is, in a nutshell, the list of operations which preserve convexity.

**Proposition** I.1.21 The following operations preserve convexity of sets:

1. *Taking intersection*: if $M_\alpha$, $\alpha \in \mathcal{A}$, are convex sets, so is their intersection $\bigcap\limits_{\alpha} M_\alpha$.
2. *Taking direct product*: if $M_1 \subseteq \mathbf{R}^{n_1}$ and $M_2 \subseteq \mathbf{R}^{n_2}$ are convex sets, so is their direct product, i.e., the set
$$M_1 \times M_2 := \left\{ x = [x^1; x^2] \in \mathbf{R}^{n_1} \times \mathbf{R}^{n_2} = \mathbf{R}^{n_1+n_2} : \ x^1 \in M_1, \ x^2 \in M_2 \right\}.$$
3. *Arithmetic summation and multiplication by reals*: if $M_1, \ldots, M_k$ are nonempty convex sets in $\mathbf{R}^n$ and $\lambda_1, \ldots, \lambda_k$ are arbitrary reals, then the set
$$\lambda_1 M_1 + \ldots + \lambda_k M_k := \left\{ \sum_{i=1}^{k} \lambda_i x^i : \ x^i \in M_i, \ i = 1, \ldots, k \right\}$$
is convex.
*Warning:* "Linear combination $\lambda_1 M_1 + \ldots + \lambda_k M_k$ of sets" as defined above is just a notation. When operating with these "linear combinations of sets," one should be careful. For example. while is it true that $M_1 + M_2 = M_2 + M_1$ and that $M_1 + (M_2 + M_3) = (M_1 + M_2) + M_3$, and

even that $\lambda(M_1 + M_2) = \lambda M_1 + \lambda M_2$, it is, in general, *not* true that $(\lambda_1 + \lambda_2)M = \lambda_1 M + \lambda_2 M$.

4. *Taking image under an affine mapping*: if $M \subseteq \mathbf{R}^n$ is convex set and $x \mapsto \mathcal{A}(x) \equiv Ax + b$ is an affine mapping from $\mathbf{R}^n$ into $\mathbf{R}^m$ (where $A \in \mathbf{R}^{m\times n}$ and $b \in \mathbf{R}^m$), then the *image* of $M$ under the mapping $\mathcal{A}(\cdot)$, i.e., the set
$$\mathcal{A}(M) := \{\mathcal{A}(x) : \ x \in M\},$$
is convex.
5. *Taking inverse image under affine mapping*: if $M \subseteq \mathbf{R}^n$ is a convex set and $y \mapsto \mathcal{A}(y) = Ay + b$ is an affine mapping from $\mathbf{R}^m$ to $\mathbf{R}^n$ (where $A \in \mathbf{R}^{n\times m}$ and $b \in \mathbf{R}^n$), then the *inverse image* of $M$ under the mapping $\mathcal{A}(\cdot)$, i.e., the set
$$\mathcal{A}^{-1}(M) := \{y \in \mathbf{R}^m : \ \mathcal{A}(y) \in M\},$$
is convex.

The (completely straightforward) verification of this proposition is left to the reader.

### 1.3.1 Calculus of closed convex sets

Numerous important convexity-related results require not just convexity, but also closedness of the participating sets. Therefore, it makes sense to think to which extent the "calculus of convexity" as presented in Proposition I.1.21 is preserved when passing from general convex sets to *closed convex* sets. Here are the answers:

1. *Taking intersection*: if $M_\alpha$, $\alpha \in \mathcal{A}$, are closed convex sets, so is the set $\bigcap\limits_\alpha M_\alpha$.
2. *Taking direct product*: if $M_1 \subseteq \mathbf{R}^{n_1}$ and $M_2 \subseteq \mathbf{R}^{n_2}$ are closed convex sets, so is the set
$$M_1 \times M_2 = \left\{x = [x^1; x^2] \in \mathbf{R}^{n_1} \times \mathbf{R}^{n_2} = \mathbf{R}^{n_1+n_2} : \ x^1 \in M_1, \ x^2 \in M_2\right\}.$$
3. *Arithmetic summation* of nonempty closed convex sets $M_i$, $1 \le i \le k$, preserves convexity, but not necessarily preserves closedness. However, it does preserve closedness when at most one of the sets is unbounded.
An example of a pair of closed convex sets with non-closed sum is $M_1 = \{x \in \mathbf{R}^2 : x_1 > 0, x_2 \ge 1/x_1\}$, $M_2 = \{x \in \mathbf{R}^2 : x_2 = 0\}$. The sum of these two closed sets clearly is the open upper halfplane $\{x \in \mathbf{R}^2 : x_2 > 0\}$ (why?) and is not closed.
Let us verify that if at most one of nonempty closed convex sets is unbounded, then the sum of the sets is convex (this we already know from calculus of convexity) and closed. Closedness is given by the following observation:

> *the sum of nonempty closed sets, convex or not, with at most one of the sets unbounded, is closed.*

To justify this observation, it clearly suffices to verify its validity for a pair of sets, $M_1$ and $M_2$. Assuming both sets are nonempty and closed and $M_1$ is bounded, we should prove that if a sequence $\{x^i + y^i\}_i$ with $x^i \in M_1$ and $y^i \in M_2$ converges as $i \to \infty$, the limit $\lim\limits_{i\to\infty}(x^i + y^i)$ belongs to $M_1 + M_2$. Since $M_1$, and thus the sequence $\{x^i\}_i$, is bounded, passing to a subsequence we may assume that the sequence $\{x^i\}_i$ converges, as $i \to \infty$, to some $x$. Since the sequence $\{x^i + y^i\}_i$ converges as well, the sequence $\{y^i\}_i$ also converges to some $y$. As $M_1$ and $M_2$ are closed, we have $x \in M_1$, $y \in M_2$, and therefore $\lim\limits_{i\to\infty}(x^i + y^i) = x + y \in M_1 + M_2$, □

4. *Multiplication by a real:* For a nonempty closed convex set $M$ and a real $\lambda$, the set $\lambda M$ is closed and convex (why?).
5. *Image under an affine mapping* of a closed convex set $M$ is convex, but not necessarily closed; it is definitely closed when $M$ is bounded.

   As an example of closed convex set with a non-closed affine image consider the set $\{[x; y] \in \mathbf{R}^2 : x, y \geq 0, xy \geq 1\}$ (i.e., a branch of hyperbola) and its projection onto the $x$-axis. This set is convex and closed, but its projection onto the $x$-axis is the positive ray $\{x > 0\}$ which is not closed. Closedness of the affine image of a closed and bounded set is the special case of the general fact:

   > *the image of a closed and bounded set under a mapping that is continuous on this set is closed and bounded as well* (why?).

6. *Inverse image under affine mapping*: if $M \subseteq \mathbf{R}^n$ is convex and closed and $y \mapsto \mathcal{A}(y) = Ay + b$ is an affine mapping from $\mathbf{R}^m$ to $\mathbf{R}^n$, then the set

   $$\mathcal{A}^{-1}(M) := \{y \in \mathbf{R}^m : \mathcal{A}(y) \in M\}$$

   is a closed convex set in $\mathbf{R}^m$. Indeed, the convexity of $\mathcal{A}^{-1}(M)$ is given by the calculus of convexity, and its closedness is due to the following standard fact:

   > *the inverse image of a closed set in $\mathbf{R}^n$ under continuous mapping from $\mathbf{R}^m$ to $\mathbf{R}^n$ is closed* (why?).

We see that the "calculus of closed convex sets" is somehow weaker than the calculus of convexity *per se*. Nevertheless, we will see that these difficulties disappear when restricting the operands of our operations to be polyhedral, and not just closed and convex.

## 1.4 Topological properties of convex sets

Convex sets and closely related objects - convex functions - play the central role in Optimization. To play this role properly, the convexity alone is not sufficient; we need convexity *and* closedness.

### *1.4.1 The closure*

It is clear from definition of a closed set that the intersection of a family of closed sets in $\mathbf{R}^n$ is also closed (see Fact B.13). From this fact it, as always, follows that for every subset $M$ of $\mathbf{R}^n$ there exists the smallest (w.r.t. inclusion) closed set containing $M$. This leads us to the following definition.

**Definition** I.1.22 [Closure] Given a set $M \subseteq \mathbf{R}^n$, the *closure* of $M$ [notation: $\operatorname{cl} M$ or $\operatorname{cl}(M)$] is the smallest (w.r.t. inclusion) closed set (i.e., the intersections of all closed sets) containing $M$.

From Real Analysis, we have the following inner description of the closure of a set in a metric space (and, in particular, in $\mathbf{R}^n$).

**Fact** I.1.23 The closure of a set $M \subseteq \mathbf{R}^n$ is exactly the set composed of the limits of all converging sequences of elements from $M$.

**Example** I.1.5 Based on Fact I.1.23, it is easy to prove that, e.g., the closure of the open Euclidean ball

$$\{x \in \mathbf{R}^n : \ \|x-a\|_2 < r\} \quad [\text{where } r > 0]$$

is the closed Euclidean ball $\{x \in \mathbf{R}^n : \ \|x-a\|_2 \leq r\}$.

Another useful application example is the closure of a set defined by *strict* linear inequalities, i.e.,

$$M := \left\{x \in \mathbf{R}^n : \ a_\alpha^\top x < b_\alpha, \ \alpha \in \mathcal{A}\right\}.$$

*Whenever such a set $M$ is nonempty*, then its closure is given by the nonstrict versions of the same inequalities:

$$\operatorname{cl} M = \left\{x \in \mathbf{R}^n : \ a_\alpha^\top x \leq b_\alpha, \ \alpha \in \mathcal{A}\right\}.$$

Note here that nonemptiness of $M$ in this last example is essential. To see this, consider the set $M = \{x \in \mathbf{R} : \ x < 0, \ -x < 0\}$. Clearly, $M$ is empty, so that its closure also is the empty set. On the other hand, if we ignore the nonemptiness requirement on $M$ and apply formally the above rule, we would incorrectly claim that $\operatorname{cl} M = \{x \in \mathbf{R} : \ x \leq 0, \ -x \leq 0\} = \{0\}$.

### 1.4.2 The interior

Consider a set $M \subseteq \mathbf{R}^n$. Recall from Definition B.9 that a point $x \in M$ is an *interior* point of $M$, if some neighborhood of the point is contained in $M$, i.e., if there exists a ball of positive radius centered at $x$ which is contained in $M$:

$$\exists r > 0 : \quad B_r(x) := \{y \in \mathbf{R}^n : \ \|y - x\|_2 \leq r\} \subseteq M.$$

**Definition** I.1.24 [Interior] The set of all interior points of a given set $M \subseteq \mathbf{R}^n$ is called the *interior* of $M$ [notation: $\operatorname{int} M$ or $\operatorname{int}(M)$] (see Definition B.10).

**Example** I.1.6 We have the following sets and their corresponding interiors:

- The interior of an open set is the set itself.
- The interior of the closed ball $\{x \in \mathbf{R}^n : \ \|x-a\|_2 \leq r\}$ ($r > 0$ or $n \geq 1$) is the open ball $\{x \in \mathbf{R}^n : \ \|x-a\|_2 < r\}$ (why?).
- The interior of the *standard full-dimensional simplex*

$$\left\{\mu \in \mathbf{R}^n : \ \mu \geq 0, \ \sum_{i=1}^n \mu_i \leq 1\right\}$$

  is composed of all vectors $\mu_i > 0$ for all $i = 1, \ldots, n$ with $\sum_{i=1}^n \mu_i < 1$ (why?).
- The interior of a polyhedral set $\{x \in \mathbf{R}^n : \ Ax \leq b\}$ with matrix $A$ not containing zero rows is the set $\{x \in \mathbf{R}^n : \ Ax < b\}$ (why?).

  Note that here the requirement that the set is polyhedral, i.e., defined by a *finite* system of linear inequalities is critical. In particular, this statement is *not*, generally speaking, true for solution sets of infinite systems of linear inequalities. For example, the following set defined by an infinite system of linear inequalities

$$M := \left\{x \in \mathbf{R} : \ x \leq \frac{1}{n}, \quad n = 1, 2, \ldots\right\}$$

  is nothing but the nonpositive ray $\mathbf{R}_- = \{x \in \mathbf{R} : \ x \leq 0\}$, i.e., $M = \mathbf{R}_-$. Thus, $\operatorname{int} M = \{x \in \mathbf{R} : \ x < 0\}$, i.e., the negative ray. On the other hand, the following set defined by the strict versions of these inequalities

$$M' := \left\{x \in \mathbf{R} : \ x < \frac{1}{n}, \quad n = 1, 2, \ldots\right\}$$

  define the same nonpositive ray, i.e., $M' = \{x \in \mathbf{R} : \ x \leq 0\}$. Hence, $M' \neq \operatorname{int} M$ for this set M defined by an infinite system of inequalities.

  The following observation is evident:

**Fact** I.1.25 For any set $M$ in $\mathbf{R}^n$, its interior, $\operatorname{int} M$, is always open, and $\operatorname{int} M$ is the largest (with respect to the inclusion) open set contained in $M$.

The interior of a set is, of course, contained in the set, which, in turn, is contained in its closure:

$$\operatorname{int} M \subseteq M \subseteq \operatorname{cl} M. \tag{1.3}$$

**Definition** I.1.26 [Boundary] For any $M \subseteq \mathbf{R}^n$, the *boundary* of $M$ is the set

$$\operatorname{bd} M = \operatorname{cl} M \setminus \operatorname{int} M,$$

and the points on the boundary are called *boundary points* of $M$.

The boundary points of $M$ are exactly the points from $\mathbf{R}^n$ which can be approximated to whatever high accuracy both by points from $M$ and by points from outside of $M$ (check it!).

Given a set $M \subseteq \mathbf{R}^n$, it is important to note that the boundary points not necessarily belong to $M$, since $M = \operatorname{cl} M$ need not necessarily hold in general. In fact, all boundary points belong to $M$ if and only if $M = \operatorname{cl} M$, i.e., if and only if $M$ is closed.

The boundary of a set $M \subseteq \mathbf{R}^n$ is clearly closed as $\operatorname{bd} M = \operatorname{cl} M \cap (\mathbf{R}^n \setminus \operatorname{int} M)$ and both sets $\operatorname{cl} M$ and $\mathbf{R}^n \setminus \operatorname{int} M$ are closed (note that the set $\mathbf{R}^n \setminus \operatorname{int} M$ is closed since it is the complement of an open set). In addition, from the definition of the boundary, we have

$$M \subseteq (\operatorname{int} M \cup \operatorname{bd} M) = \operatorname{cl} M.$$

Therefore, any point from $M$ is either an interior or a boundary point of $M$.

### 1.4.3 The relative interior

Many of the constructions in Optimization possess nice properties in the interior of the set the construction is related to and may lose these nice properties at the boundary points of the set. This is why in many cases we are especially interested in interior points of sets and want the set of these interior points to be "sufficiently dense." What should we do if it is not the case, for example if there are no interior points at all (e.g., if we are looking at a segment in the plane)? It turns out that in these cases we can use a good surrogate of the "normal" interior, namely the *relative interior* defined as follows.

**Definition** I.1.27 [Relative interior] Let $M \subseteq \mathbf{R}^n$ be nonempty. We say that a point $x \in M$ is *relative interior* for $M$ if $M$ contains the intersection of a small enough ball centered at $x$ with $\operatorname{Aff}(M)$, i.e., if there exists $r > 0$ such that

$$(B_r(x) \cap \operatorname{Aff}(M)) := \{y \in \mathbf{R}^n : \ y \in \operatorname{Aff}(M), \ \|y - x\|_2 \leq r\} \subseteq M.$$

The *relative interior* of $M$ [notation: $\operatorname{rint} M$] refers to the set of all relative interior points of $M$.

By definition, the relative interior of empty set is empty.

**Example** I.1.7 We have the following sets and their corresponding relative interiors:

- The relative interior of a singleton is the singleton itself (since a point in the 0-dimensional space is the same as a ball of a positive radius).

- More generally, the relative interior of an affine subspace is the subspace itself.
- Given two distinct point $x \neq y$ in $\mathbf{R}^n$, the interior of a segment $[x, y]$ is empty whenever $n > 1$. In contrast to this, the relative interior of this set is always (independent of $n$) nonempty and it is precisely the interval $(x, y)$, i.e., the segment without the endpoints.

Geometrically speaking, the relative interior is the interior we get when we treat $M \subseteq \mathbf{R}^n$ as a subset of its affine hull (the latter, geometrically, is nothing but $\mathbf{R}^k$, $k$ being the affine dimension of $\mathrm{Aff}(M)$).

We can play with the notion of the relative interior in basically the same way as with the one of interior. Namely, for any $M \subseteq \mathbf{R}^n$, since $\mathrm{Aff}(M)$ is closed and contains $M$, it contains also the smallest closed set containing $M$, i.e., $\mathrm{cl}\, M$. Therefore, we have the following analogies of inclusions, cf. (1.3):

$$\mathrm{rint}\, M \subseteq M \subseteq \mathrm{cl}\, M \quad [\subseteq \mathrm{Aff}(M)]. \tag{1.4}$$

We can also define the *relative boundary.*

**Definition** I.1.28 [Relative boundary] For any $M \subseteq \mathbf{R}^n$, its *relative boundary* [notation: $\mathrm{rbd}\, M$] is defined as the set $\mathrm{rbd}\, M = \mathrm{cl}\, M \setminus \mathrm{rint}\, M$.

Note that for any $M \subseteq \mathbf{R}^n$, we naturally have $\mathrm{rbd}\, M$ is a closed set contained in $\mathrm{Aff}(M)$, and, as for the "actual" interior and boundary, we have

$$\mathrm{rint}\, M \subseteq M \subseteq \mathrm{cl}\, M = \mathrm{rint}\, M \cup \mathrm{rbd}\, M.$$

Of course, if $\mathrm{Aff}(M) = \mathbf{R}^n$, then the relative interior becomes the usual interior, and similarly for boundary. Note that $\mathrm{Aff}(M) = \mathbf{R}^n$ for sure is the case when $\mathrm{int}\, M \neq \varnothing$ (since then $M$ contains a ball $B$, and therefore the affine hull of $M$ is the entire $\mathbf{R}^n$, which is the affine hull of $B$).

### *1.4.4 Nice topological properties of convex sets*

An arbitrary set $M \subseteq \mathbf{R}^n$ may possess very pathological topology. In particular, both inclusions in the chain

$$\mathrm{rint}\, M \subseteq M \subseteq \mathrm{cl}\, M$$

can be very "loose." For example, let $M$ be the set of rational numbers in the segment $[0, 1] \subset \mathbf{R}$. Then, $\mathrm{rint}\, M = \mathrm{int}\, M = \varnothing$ since every neighborhood of every rational number contains irrational numbers. On the other hand, $\mathrm{cl}\, M = [0, 1]$. Thus, $\mathrm{rint}\, M$ is "incomparably smaller" than $M$, $\mathrm{cl}\, M$ is "incomparably larger" than $M$, and $M$ is contained in its relative boundary (by the way, what is this relative boundary?).

The following theorem demonstrates that the topology of a *convex* set $M$ is much better than what it might be for an arbitrary set.

**Theorem** I.1.29 Let $M$ be a convex set in $\mathbf{R}^n$. Then,

(i) The interior $\operatorname{int} M$, the closure $\operatorname{cl} M$ and the relative interior $\operatorname{rint} M$ are convex.

(ii) If $M$ is nonempty, then its relative interior $\operatorname{rint} M$ is nonempty.

(iii) The closure of $M$ is the same as the closure of its relative interior, i.e., $\operatorname{cl} M = \operatorname{cl}(\operatorname{rint} M)$. (In particular, every point of $\operatorname{cl} M$ is the limit of a sequence of points from $\operatorname{rint} M$.)

(iv) The relative interior remains unchanged when we replace $M$ with its closure, i.e., $\operatorname{rint} M = \operatorname{rint}(\operatorname{cl} M)$.

Moreover, (iii) and (iv) imply that

(v) The relative boundary remains unchanged when we replace $M$ with its closure.

We will use the following basic result to prove this theorem (we will present the proof of this lemma after the proof of the theorem).

**Lemma** I.1.30 Let $M$ be a convex set in $\mathbf{R}^n$. Then, for any $x \in \operatorname{rint} M$ and $y \in \operatorname{cl} M$, we have

$$[x, y) := \{(1-\lambda)x + \lambda y : \ 0 \leq \lambda < 1\} \subseteq \operatorname{rint} M.$$

**Proof of Theorem I.1.29.** (i): Prove yourself!

(ii): Let $M$ be a nonempty convex set, and let us prove that $\operatorname{rint} M \neq \varnothing$. By translation, we may assume that $0 \in M$. Furthermore, we may assume that the linear span of $M$, i.e., $\operatorname{Lin}(M)$, is the entire $\mathbf{R}^n$. Indeed, as far as linear operations and the Euclidean structure are concerned, $\operatorname{Lin}(M)$, as every other linear subspace in $\mathbf{R}^n$, is equivalent to $\mathbf{R}^k$ for a certain $k$. Since the notion of relative interior deals only with linear and Euclidean structures, we lose nothing thinking of $\operatorname{Lin}(M)$ as of $\mathbf{R}^k$ and taking it as our universe instead of the original universe $\mathbf{R}^n$. Thus, in the rest of the proof of (ii), we assume that $0 \in M$ and $\operatorname{Lin}(M) = \mathbf{R}^n$; what we need to prove is that the interior of $M$ (which in our case is the same as relative interior of $M$) is nonempty. Note that since $0 \in M$, we have $\operatorname{Aff}(M) = \operatorname{Lin}(M) = \mathbf{R}^n$.

As $\operatorname{Lin}(M) = \mathbf{R}^n$, we can find $n$ linearly independent vectors $a^1, \ldots, a^n$ in $M$. Let us also set $a^0 := 0$. The $n+1$ vectors $a^0, \ldots, a^n$ belong to $M$. *Since $M$ is convex*, the convex hull of these vectors, i.e.,

$$\Delta := \left\{ x = \sum_{i=0}^{n} \lambda_i a^i : \ \lambda \geq 0, \ \sum_{i=0}^{n} \lambda_i = 1 \right\} = \left\{ x = \sum_{i=1}^{n} \mu_i a^i : \ \mu \geq 0, \ \sum_{i=1}^{n} \mu_i \leq 1 \right\}$$

also belongs to $M$. Note that the set $\Delta$ is the image of the *standard full-dimensional simplex*

$$\left\{ \mu \in \mathbf{R}^n : \ \mu \geq 0, \ \sum_{i=1}^{n} \mu_i \leq 1 \right\}$$

under the linear transformation $\mu \mapsto A\mu$, where $A$ is the matrix with the columns

$a^1, \dots, a^n$. Recall from Example I.1.6 that the standard simplex has a nonempty interior. Since $A$ is nonsingular (due to the linear independence of $a^1, \dots, a^n$), multiplication by $A$ maps open sets onto open ones, so that $\Delta$ has a nonempty interior. Since $\Delta \subseteq M$, the interior of $M$ is nonempty.

(iii): The statement is evidently true when $M$ is empty, so we assume that $M \neq \varnothing$. We clearly have $\mathrm{cl}(\mathrm{rint}\, M) \subseteq \mathrm{cl}\, M$ due to $\mathrm{rint}\, M \subseteq M$. Thus, all we need to complete the proof of (iii) is to verify that every $y \in \mathrm{cl}\, M$ is the limit of a sequence of points $y^i \in \mathrm{rint}\, M$. Indeed, pick $x \in \mathrm{rint}\, M$ (recall that from part (ii) we have $\mathrm{rint}\, M \neq \varnothing$) and set $y^i := (1 - 1/i)y + (1/i)x$. By Lemma I.1.30, we have $y^i \in \mathrm{rint}\, M$, and clearly $y = \lim\limits_{i\to\infty} y^i$, completing the verification of (iii).

(iv): The statement is obviously true when $M$ is empty, so we assume that $M \neq \varnothing$. Since $M \subseteq \mathrm{cl}\, M$, we always have $\mathrm{rint}\, M \subseteq \mathrm{rint}\,(\mathrm{cl}\, M)$. To prove the reverse inclusion, consider any $z \in \mathrm{rint}\,(\mathrm{cl}\, M)$, and let us prove that $z \in \mathrm{rint}\, M$. Let $x \in \mathrm{rint}\, M$ (from part (ii), we already know that $\mathrm{rint}\, M \neq \varnothing$). As $x$ and $z$ are in $\mathrm{Aff}(M)$, for any $t \in \mathbf{R}$, the vectors $z^t := x + t(z - x)$ belong to $\mathrm{Aff}(M)$, and when $t$ approaches 1, $z^t$ approaches $z$. Since $z \in \mathrm{rint}\,(\mathrm{cl}\, M)$, it follows that there exists $\epsilon > 0$ such that $y := z^{1+\epsilon} \in \mathrm{cl}\, M$. It remains to note that $z = (1-\lambda)y + \lambda x$ with $\lambda = \frac{\epsilon}{1+\epsilon} \in (0,1)$, and therefore $z = (1-\lambda)y + \lambda x \in \mathrm{rint}\, M$ by Lemma I.1.30 (recall that $x \in \mathrm{rint}\, M$, $y \in \mathrm{cl}\, M$). □

**Remark** I.1.31 We see from the proof of Theorem I.1.29(iii) that to get the closure of a (nonempty) convex set, it suffices to take its "radial" closure, i.e., to take a point $x \in \mathrm{rint}\, M$, take all rays in $\mathrm{Aff}(M)$ starting at $x$ and look at the intersection of such a ray $\ell$ with $M$; such an intersection will be a convex set on the line which contains a one-sided neighborhood of $x$, i.e., is either a segment $[x, y^\ell]$, or the entire ray $\ell$, or a half-interval $[x, y^\ell)$. In the first two cases we do not need to do anything; in the third case, we need to add $y^\ell$ to $M$. After all rays are looked through and all "missed" endpoints $y^\ell$ are added to $M$, we obtain $\mathrm{cl}\, M$. To understand the role of convexity in this result, look at the *nonconvex* set of rational numbers from $[0,1]$. The interior ($\equiv$ relative interior) of this "highly percolated" set is empty, the closure is $[0,1]$, and there is no way to restore the closure in terms of the interior.

**Proof of Lemma I.1.30.** Given that $x \in M$, let us denote $\mathrm{Aff}(M) = x + L$, where $L$ is the linear subspace parallel to $\mathrm{Aff}(M)$. Then,

$$M \subseteq \mathrm{Aff}(M) = x + L.$$

Let $B$ be the unit Euclidean ball in $L$, i.e., $B = \{h \in L : \|h\|_2 \leq 1\}$. Since $x \in \mathrm{rint}\, M$, there exists a positive radius $r$ such that

$$x + rB \subseteq M. \tag{1.5}$$

Now consider any $\lambda \in [0,1)$, and let $z := (1-\lambda)x + \lambda y$. As $y \in \mathrm{cl}\, M$, we have $y = \lim\limits_{i\to\infty} y^i$ for certain sequence of points from $M$. By setting $z^i := (1-\lambda)x + \lambda y^i$, we get $z^i \to z$ as $i \to \infty$. Then, from (1.5) and the convexity of $M$, it follows that the sets $Z_i := \{(1-\lambda)x' + \lambda y^i : \ x' \in x + rB\}$ are contained in $M$. Clearly, $Z_i$ is exactly the set $z^i + r'B$, where $r' := (1-\lambda)r > 0$. Thus, $z$ is the limit of the sequence $z^i$, and $r'$-neighborhood (in $\mathrm{Aff}(M)$) of every one of the points $z^i$ belongs to $M$. For every $0 < r'' < r'$ and for all $i$ such that $z^i$ is close enough to $z$, the $r'$-neighborhood of $z^i$ contains the $r''$-neighborhood of $z$; thus, a neighborhood (in $\mathrm{Aff}(M)$) of $z$ belongs to $M$, hence $z \in \mathrm{rint}\, M$. □

A useful byproduct of Lemma I.1.30 is as follows:

**Corollary** I.1.32 Let $M \subseteq \mathbf{R}^n$ be convex. Then, every convex combination $\sum_i \lambda_i x^i$ of points $x^i \in \mathrm{cl}\, M$ such that at least one term with positive coefficient is associated with $x^i \in \mathrm{rint}\, M$ is in fact a point from $\mathrm{rint}\, M$.

Another useful byproduct of Lemma I.1.30 is as follows. Let $M_k$, $k \leq K$, be a finite collection of subsets of $\mathbf{R}^n$. The closure of the union of these sets is the union of their closures: $\mathrm{cl}(\cup_{k\leq K} M_k) = \cup_{k\leq K} \mathrm{cl}\, M_j$ (why?). Now let us ask ourselves similar question about intersection: *what is the relation between* $\mathrm{cl} \cap_{k\leq K} M_k$ *and* $\cap_{k\leq K} \mathrm{cl}\, M_k$*?* The set $\cap_{k\leq K} \mathrm{cl}\, M_k$ is closed and clearly contains $\cap_{k\leq K} M_k$ and thus always contains the closure of the latter set:

$$\mathrm{cl}\left(\bigcap_{k\leq K} M_k\right) \subseteq \bigcap_{k\leq K} \mathrm{cl}\, M_k. \tag{1.6}$$

In general, this inclusion can be "loose" – the right hand side set in (1.6) can be much larger than the left hand side one, even when all $M_k$ are convex. For example, when $K = 2$, $M_1 = \{x \in \mathbf{R}^2 : x_2 = 0\}$ is the $x_1$-axis, and $M_2 = \{x \in \mathbf{R}^2 : x_2 > 0\} \cup \{[0; 0]\}$, both sets are convex, their intersection is the singleton $\{0\}$, so that $\mathrm{cl}(M_1 \cap M_2) = \mathrm{cl}\{0\} = \{0\}$, while the intersection of $\mathrm{cl}\, M_1$ and $\mathrm{cl}\, M_2$ is the entire $x_1$-axis, which is simply $M_1$. In this example the right hand side in (1.6) is "incomparably larger" than the left hand side one. However, under suitable assumptions we can also achieve equality in (1.6).

**Proposition** I.1.33 Consider convex sets $M_k \subseteq \mathbf{R}^n$, $k \leq K$.

(i) If $\bigcap_{k\leq K} \mathrm{rint}\, M_k \neq \varnothing$, then $\mathrm{cl}(\cap_{k\leq K} M_k) = \cap_{k\leq K} \mathrm{cl}\, M_k$, i.e., (1.6) holds an equality.

(ii) Moreover, if $M_K \cap \mathrm{int}\, M_1 \cap \mathrm{int}\, M_2 \cap \ldots \cap \mathrm{int}\, M_{K-1} \neq \varnothing$, then we have $\bigcap_{k\leq K} \mathrm{rint}\, M_k \neq \varnothing$, i.e., the premise (and thus the conclusion) in (i) holds true, so that $\mathrm{cl}(\cap_{k\leq K} M_k) = \cap_{k\leq K} \mathrm{cl}\, M_k$.

**Proof.** (i): To prove that under the premise of (i) inclusion (1.6) is equality is the same as to verify that under the circumstances given $x \in \cap_k \mathrm{cl}\, M_k$, one has $x \in \mathrm{cl}\,(\cap_k M_k)$. Indeed, under the premise of (i) there exists $\bar{x} \in \cap_k \mathrm{rint}\, M_k$. Then, for every $k$ we have $\bar{x} \in \mathrm{rint}\, M_k$ and $x \in \mathrm{cl}\, M_k$, implying by Lemma I.1.30 that the set $\Delta := [\bar{x}, x) = \{(1-\lambda)\bar{x} + \lambda x : 0 \leq \lambda < 1\}$ is contained in $M_k$. Since $\Delta \subseteq M_k$ for all $k$, we have $\Delta \in \cap_k M_k$, and thus $\mathrm{cl}\,\Delta \subseteq \mathrm{cl}\,(\cap_k M_k)$. It remains to note that $x \in \mathrm{cl}\,\Delta$.

(ii): Let $\bar{x} \in M_K \cap \mathrm{int}\, M_1 \cap \ldots \cap \mathrm{int}\, M_{K-1}$. As $\bar{x} \in \mathrm{int}\, M_k$ for all $k < K$, there exists an open set $U \subset \cap_{k<K} M_k$ such that $\bar{x} \in U$. As $\bar{x} \in M_K \subseteq \mathrm{cl}\, M_K$, by Theorem I.1.29, $\bar{x}$ is the limit of a sequence of points from $\mathrm{rint}\, M_K$, so that there

exists $\widehat{x} \in U \cap \operatorname{rint} M_K$. Due to the origin of $U$, we have $\widehat{x} \in \operatorname{rint} M_k$ for all $k \leq K$, so that the premise of (i) indeed takes place. □

## 1.5 ★ Conic and perspective transforms of a convex set

Let $X \subseteq \mathbf{R}^n$ be a nonempty convex set. We can "lift" it to $\mathbf{R}^{n+1}$ by passing to the set

$$X^+ := \{[x;1] \in \mathbf{R}^n \times \mathbf{R} : \ x \in X\}.$$

Now let us look at the conic hull of $X^+$, given by

$$\begin{aligned} \operatorname{ConeT}(X) &:= \operatorname{Cone}(X^+) \\ &= \left\{ \begin{array}{r} [x;t] \in \mathbf{R}^n \times \mathbf{R}_+ : \ \exists (I, \ \lambda_i \geq 0, \ x^i \in X, \ \forall i \leq I) : \\ x = \sum_{i \leq I} \lambda_i x^i, \ t = \sum_{i \leq I} \lambda_i \end{array} \right\}. \end{aligned}$$

We will call this the *conic transform of* $X$, see Figure I.4. Note that this set is indeed a cone. Moreover, all vectors $[x;t]$ from this cone have $t \geq 0$, and, importantly, the only vector with $t = 0$ in the cone $\operatorname{ConeT}(X)$ is the origin in $\mathbf{R}^{n+1}$ (this is what you get when taking trivial – with all coefficients zero – conic combinations of vectors from $X^+$).

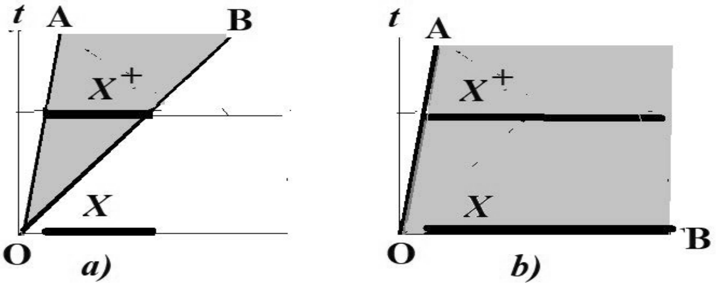


Figure I.4. Conic transform
*a)* conic transform of segment $X$ is the angle AOB
*b)* conic transform of ray $X$ is the angle AOB with relative interior of the ray OB excluded

All nonzero vectors $[x;t]$ from $\operatorname{ConeT}(X)$ have $t > 0$ and form a convex set which we call the *perspective transform* $\operatorname{Persp}(X)$ of $X$:

$$\operatorname{Persp}(X) := \{[x;t] \in \operatorname{ConeT}(X) : \ t > 0\} = \operatorname{ConeT}(X) \setminus \{0_{n+1}\}.$$

The name of this set is motivated by the following immediate observation:

**Proposition** I.1.34 [Perspective transform of a nonempty convex set] Let $X$ be a nonempty convex set in $\mathbf{R}^n$. Then, its perspective transform admits the representation

$$\mathrm{Persp}(X) = \{[x;t] \in \mathbf{R}^n \times \mathbf{R} : \ t > 0, \ x/t \in X\}. \tag{1.7}$$

In other words, to get $\mathrm{Persp}(X)$, we pass from $X$ to $X^+$ (i.e., lift $X$ to $\mathbf{R}^{n+1}$) and then take the union of all rays $\{[sx;s] \in \mathbf{R}^n \times \mathbf{R} : \ s > 0, \ x \in X\}$ emanating from the origin (with origin excluded) and passing through the points of $X^+$.

**Proof.** Let $\widehat{X} := \{[x;t] \in \mathbf{R}^n \times \mathbf{R} : \ t > 0, \ x/t \in X\}$, so that the claim in the proposition is $\mathrm{Persp}(X) = \widehat{X}$. Consider a point $[x;t] \in \widehat{X}$. Then, $t > 0$ and $y := x/t \in X$, and thus we have $[x,t] = t[y;1]$ so that the point $[x;t]$ from $\widehat{X}$ is a single-term conic combination – just positive multiple – of the point $[y;1] \in X^+$. As this holds for every point $[x;t] \in \widehat{X}$ we conclude $\widehat{X} \subseteq \mathrm{Persp}(X)$. To verify the opposite inclusion, recall that every point $[x;t] \in \mathrm{Persp}(X)$ is of the form $[\sum_i \lambda_i x^i; \sum_i \lambda_i]$ with $x^i \in X$, $\lambda_i \geq 0$, and $t = \sum_i \lambda_i > 0$. Then,

$$\left[\sum\nolimits_i \lambda_i x^i; \sum\nolimits_i \lambda_i\right] = t\left[\sum\nolimits_i (\lambda_i/t)x^i; 1\right] = t[y;1],$$

where $y := \sum_i (\lambda_i/t)x^i$. Note that $y \in X$ as it is a convex combination of points from $X$ and $X$ is convex. Thus, $[x;t]$ is such that $t > 0$ and $y = x/t \in X$, that is, $\widehat{X} \supseteq \mathrm{Persp}(X)$ as desired. □

As a byproduct of Proposition I.1.34, we conclude that the right hand side set in (1.7) is convex whenever $X$ is convex and nonempty – a fact not so evident "from scratch."

Note that $X^+$ is geometrically the same as $X$, and moreover we can view $X^+$ as simply the intersection of $\mathrm{ConeT}(X)$ (or $\mathrm{Persp}(X)$) with the hyperplane $t = 1$ in $\mathbf{R}^n \times \mathbf{R}$.

**Example** I.1.8

1. $\mathrm{ConeT}(\mathbf{R}^n) = \{[x;t] \in \mathbf{R}^{n+1} : t > 0\} \cup \{0_{n+1}\}$, and
   $\mathrm{Persp}(\mathbf{R}^n) = \{[x;t] \in \mathbf{R}^{n+1} : t > 0\}$.
2. $\mathrm{ConeT}(\mathbf{R}^n_+) = \left\{[x;t] \in \mathbf{R}^{n+1}_+ : t > 0\right\} \cup \{0_{n+1}\}$, and
   $\mathrm{Persp}(\mathbf{R}^n_+) = \left\{[x;t] \in \mathbf{R}^{n+1}_+ : t > 0\right\}$.
3. Given any norm $\|\cdot\|$ on $\mathbf{R}^n$, let $B$ be its unit ball. Then, we have $\mathrm{ConeT}(B) = \{[x;t] \in \mathbf{R}^{n+1} : \ t \geq \|x\|\}$, and $\mathrm{Persp}(B) = \{[t;x] \in \mathbf{R}^{n+1} : \ t \geq \|x\|, \ t > 0\}$.

Note that in all three examples in Example I.1.8, the set $X$ of which we are taking conic and perspective transforms is not just convex, but also closed. However, in the first two examples, the conic transform is a *non-closed* cone, while in the third example the conic transform is closed, albeit in all three cases the intersections of $\mathrm{ConeT}(X)$ with half-space $\{[x;t] \in \mathbf{R}^{n+1} : t \geq \alpha\}$ is closed, provided $\alpha > 0$. There is indeed a general fact underlying this phenomenon.

**Proposition** I.1.35 Let $X \subset \mathbf{R}^n$ be a nonempty convex set. Then, we have the following:
(i) For $\alpha > 0$, define $H_\alpha := \{[x;t] \in \mathbf{R}^{n+1} : t \geq \alpha\}$. When $X$ is closed, $\mathrm{ConeT}(X) \cap H_\alpha = \mathrm{Persp}(X) \cap H_\alpha$ and this intersection is closed for any $\alpha > 0$.
(ii) Moreover, the cone $\mathrm{ConeT}(X)$ is closed if and only if $X$ is closed and bounded. In fact, $\mathrm{ConeT}(X)$ is closed if and only if $\mathrm{cl}\,(\mathrm{Persp}(X)) = \mathrm{ConeT}(X)$.

**Proof.** (i): When $\alpha > 0$, we clearly have $\mathrm{ConeT}(X) \cap H_\alpha = \mathrm{Persp}(X) \cap H_\alpha$. To see that these intersections are closed whenever $X$ is closed, invoking (1.7) it suffices to prove that when $\{[x^i;t_i]\}_{i\geq 1}$ is a converging sequence such that $t_i \geq \alpha$ and $x^i/t_i \in X$, then the limit $[x;t]$ of this sequence satisfies $x/t \in X$ and $t \geq \alpha$. Since $t_i \to t$ as $i \to \infty$ and $t_i \geq \alpha$ holds for all $i$, we clearly have $t \geq \alpha$. Moreover, we have that the converging sequence $y^i := x^i/t_i$ is in $X$ thus $x^i/t_i \to x/t$ as $i \to \infty$ and the point $x/t$ is in $X$ since $X$ is closed.

(ii): First, we assume that nonempty convex set $X$ is closed and bounded, and we will prove that $\mathrm{ConeT}(X)$ is closed, that is, whenever a sequence $\{[x^i;t_i]\}_{i\geq 1}$ of points from $\mathrm{ConeT}(X)$ converges, the limit of the sequence belongs to $\mathrm{ConeT}(X)$. Indeed, consider such a sequence along with its limit $[x;t]$. When $t > 0$, all but finitely many terms of the sequence belong to the half-space $H_{t/2}$, and as by part (i) $\mathrm{ConeT}(X) \cap H_{t/2}$ is closed, we have $[x;t] \in \mathrm{ConeT}(X)$. When $t = 0$, then either (a) $t_i = 0$ for infinitely many values of $i$, or (b) $t_i > 0$ for all but finitely many values of $i$. In the case of (a) infinitely many terms in our sequence are of the form $[0_n;0]$ (since whenever $[y;0] \in \mathrm{ConeT}(X)$ we must have $y = 0$ as well), so that $[x;t] = 0_{n+1} \in \mathrm{ConeT}(X)$. In the case of (b) for all large enough $i$ we have $t_i > 0$ and $x^i/t_i \in X$, and since $t_i \to 0$ as $i \to \infty$ and $X$ is bounded we deduce $x^i \to 0$ as $i \to \infty$. Then, this together with $t_i \to 0$, $i \to \infty$, implies that $[x;t] = [0_{n+1};0]$, and we again have $[x;t] \in \mathrm{ConeT}(X)$. Thus, whenever $X$ is a nonempty closed and bounded set, $\mathrm{ConeT}(X)$ is closed.

Now assume that $\mathrm{ConeT}(X)$ is closed, and let us prove that $X$ is closed and bounded. Clearly, $X$ is closed if and only if $X^+$ is closed, and since $X^+$ is the intersection of the closed set $\mathrm{ConeT}(X)$ with the hyperplane $t = 1$ in $\mathbf{R}^n \times \mathbf{R}$, $X^+$ is indeed closed. it remains to prove that $X$ is bounded. Assume for contradiction that $X$ is unbounded. Then, we can find a sequence $x^i \in X$, $i \geq 1$, with $\|x^i\|_2 \to \infty$ as $i \to \infty$. Passing to a subsequence, we can assume that the $\|\cdot\|_2$-unit vectors $\xi^i := x^i/\|x^i\|_2$ converge to some unit vector $\xi$. Setting $t_i := 1/\|x^i\|_2$, we have $t_i > 0$, $t_i \to 0$ as $i \to \infty$, and $\xi^i/t_i = x^i \in X$, so that $[\xi^i;t_i] \in \mathrm{ConeT}(X)$. Since $\mathrm{ConeT}(X)$ is closed and by construction $[\xi^i;t_i] \to [\xi;0]$ as $i \to \infty$, we should have $[\xi;0] \in \mathrm{ConeT}(X)$, which is impossible as $\|\xi\|_2 = 1$.

The final claim, i.e., $\mathrm{ConeT}(X)$ is closed if and only if $\mathrm{cl}\,(\mathrm{Persp}(X)) = \mathrm{ConeT}(X)$, follows immediately as well. Indeed, whenever $X$ is nonempty and convex, we have $\mathrm{Persp}(X) = \mathrm{ConeT}(X) \setminus \{0_{n+1}\}$ and clearly $0_{n+1} \in \mathrm{cl}\,\mathrm{Persp}(X)$, implying that $\mathrm{cl}\,\mathrm{ConeT}(X) = \mathrm{cl}\,\mathrm{Persp}(X)$. As a result, whenever $X$ is nonempty and convex, $\mathrm{ConeT}(X)$ is closed iff $\mathrm{ConeT}(X) = \mathrm{cl}\,\mathrm{Persp}(X)$. $\square$

For a nonempty convex set $X$, let us also define the *closure* of $\mathrm{ConeT}(X)$, i.e., the set

$$\overline{\mathrm{ConeT}}(X) := \mathrm{cl}\left\{[x;t] \in \mathbf{R}^n \times \mathbf{R} : \; t > 0, \; x/t \in X\right\}.$$

Clearly, $\overline{\mathrm{ConeT}}(X)$ is a closed cone in $\mathbf{R}^{n+1}$ containing $X^+$. Moreover, it is immediately seen that $\overline{\mathrm{ConeT}}(X)$ is the smallest (w.r.t. inclusion) closed cone in $\mathbf{R}^{n+1}$ which contains $X^+$ and that this cone remains intact when extending $X$ to the closure of $X$. We will refer to $\overline{\mathrm{ConeT}}(X)$ as the *closed conic transform of $X$*. In some cases, $\overline{\mathrm{ConeT}}(X)$ admits a simple characterization. An immediate illustration of this is as follows:

**Fact** I.1.36 Let $K$ be a closed cone and let the set

$$X := \{x \in \mathbf{R}^n : \; Ax - b \in K\}$$

be nonempty. Then, $\overline{\mathrm{ConeT}}(X) = \{[x;t] \in \mathbf{R}^n \times \mathbf{R} : \; Ax - bt \in K, \; t \geq 0\}$.

For useful additional facts on closed conic transforms, see Exercise III.12.1-3.

# 2

# Theorems of Caratheodory, Radon, and Helly

We next examine three theorems from Convex Analysis that have important consequences in Optimization.

## 2.1 Caratheodory Theorem

Let us recall the notion of *dimension* from Linear Algebra. First of all, we define the *dimension* of a linear subspace. This is precisely the number of linearly independent vectors spanning the linear subspace. In the case of an affine subspace, we talk about its *affine dimension*, which is precisely the dimension of the linear subspace that is underlying (parallel) to the given affine subspace. Based on these notions, we are now ready to define dimension of a nonempty set $M$.

> **Definition** I.2.1 [Dimension of a nonempty set] Given a nonempty set $M \subseteq \mathbf{R}^n$, its *dimension* (also referred as its *affine dimension*) [notation: $\dim(M)$] is defined as the affine dimension of $\mathrm{Aff}(M)$, or, which is the same, linear dimension of the linear subspace parallel to $\mathrm{Aff}(M)$.

**Remark** I.2.2 Note that some subsets of $\mathbf{R}^n$ are in the scopes of several definitions of dimension. Specifically, linear subspace is also an affine subspace, and similarly, an affine subspace is a nonempty set as well. It is immediately seen that if a set is in the scope of more than one definition of dimension, all applicable definitions attribute the set with the same value of the dimension.

As an informal introduction to what follows, draw several points ("red points") on the 2D plane and a take a point ("blue point") in their convex hull. You will observe that whatever your selection of red points and the blue point in their convex hull, this point will belong to a properly selected triangle with red vertices. The general fact is as follows.

> **Theorem** I.2.3 [Caratheodory Theorem] Consider a nonempty set $M \subseteq \mathbf{R}^n$, and let $m := \dim(M)$. Then, every point $x \in \mathrm{Conv}(M)$ is a convex combination of at most $m + 1$ points from $M$.

**Proof.** Let $E := \mathrm{Aff}(M)$. Then, $\dim(E) = m$. Replacing, if necessary, the embedding space $\mathbf{R}^n$ of $M$ with $E$ (the latter is, geometrically, just $\mathbf{R}^m$), we can assume with loss of generality that $m = n$.

Let $x \in \operatorname{Conv}(M)$. By Fact I.1.14 on the structure of convex hull, there exists $x^1, \ldots, x^N$ from $M$ and convex combination weights $\lambda_1, \ldots, \lambda_N$ such that

$$x = \sum_{i=1}^N \lambda_i x^i, \quad \text{where} \quad \lambda_i \geq 0, \forall i = 1, \ldots, N, \ \sum_{i=1}^N \lambda_i = 1.$$

Among all such possible representations of $x$ as a convex combination of points from $M$, let us choose one with the smallest possible $N$, i.e., involving fewest number of points from $M$. Let this representation of $x$ be the above convex combination. We claim that $N \leq m + 1$; proving this claim is all we need to complete the proof of Caratheodory Theorem.

Let us assume for contradiction that $N > m + 1$. Now, consider the following system in $N$ variables $\mu_1, \ldots, \mu_N$:

$$\sum_{i=1}^N \mu_i x^i = 0, \quad \sum_{i=1}^N \mu_i = 0.$$

This is a system of $m+1$ scalar homogeneous linear equations (recall that we are in the case of $m = n$, that is, $x^i \in \mathbf{R}^m$). As $N > m + 1$, the number of variables in this system is *strictly greater* than the number of equations. Therefore, this system has a *nontrivial* solution, say $\delta_1, \ldots, \delta_N$, i.e.,

$$\sum_{i=1}^N \delta_i x^i = 0, \quad \sum_{i=1}^N \delta_i = 0, \quad \text{and } [\delta_1; \ldots; \delta_N] \neq 0.$$

Then, for every $t \in \mathbf{R}$, we have the following representation of $x$ as a linear combination of the points $x^1, \ldots, x^N$:

$$\sum_{i=1}^N (\lambda_i + t\delta_i) x^i = x.$$

For ease of reference, let us define $\lambda_i(t) := \lambda_i + t\delta_i$ for all $i$ and for all $t \in \mathbf{R}$. Note that for any $t \in \mathbf{R}$, by the definition of $\lambda_i$ and $\delta_i$, we always have

$$\sum_{i=1}^N \lambda_i(t) = \sum_{i=1}^N (\lambda_i + t\delta_i) = \sum_{i=1}^N \lambda_i + t \sum_{i=1}^N \delta_i = 1.$$

Moreover, when $t = 0$, $\lambda_i(0) = \lambda_i$ for all $i$, and thus this is a convex combination as all $\lambda_i(0) \geq 0$ for all $i$. On the other hand, from the selection of $\delta_i$, we know that $\sum_{i=1}^N \delta_i = 0$ and $[\delta_1; \ldots; \delta_N] \neq 0$, and thus at least one entry in $\delta$ must be negative. Therefore, when $t$ is large, some of the coefficients $\lambda_i(t)$ will be negative. There exists, of course, the largest $t= t^*$ for which $\lambda_i(t) \geq 0$ for all $i = 1, \ldots, N$ holds, and for $t = t^*$ at least some of $\lambda_i(t)$ are zero. Specifically, when setting $I^- := \{i : \ \delta_i < 0\}$,

$$i^* \in \operatorname*{argmin}_i \left\{ \frac{\lambda_i}{|\delta_i|} : \ i \in I^- \right\}, \quad \text{and} \quad t^* := \min_i \left\{ \frac{\lambda_i}{|\delta_i|} : \ i \in I^- \right\},$$

we have $\lambda_i(t^*) \geq 0$ for all $i$ and $\lambda_{i^*}(t^*) = 0$. This then implies that we have represented $x$ as a convex combination of less than $N$ points from $M$, which contradicts the definition of $N$ (being the smallest number of points $x^i$ needed in the convex combination representation of $x$). □

**Remark** I.2.4 Caratheodory Theorem is sharp: for every positive integer $n$, there is a set of $n+1$ affinely independent points in $\mathbf{R}^n$ (e.g., the origin and the $n$ standard basic orths) such that certain convex combination of these $n+1$ points (specifically, their average) cannot be represented as a convex combination using strictly less than $n+1$ points from the set.

Let us see an instructive corollary witnessing the power of Caratheodory Theorem.

**Corollary** I.2.5 Let $X \subseteq \mathbf{R}^n$ be a closed and bounded set. Then, $\mathrm{Conv}(X)$ is closed and bounded as well.

**Proof.** There is nothing to prove when $X$ is empty. Now let $X$ be nonempty, closed and bounded, and define $Y := \mathrm{Conv}(X)$. Boundedness of $Y$ is evident. In order to verify that $Y$ is closed, let $\{x^t\}_{t\geq 1}$ be a converging sequence of points from $Y$. By Caratheodory Theorem, every one of vectors $x^t$, being a convex combination of vectors from $X$, has a representation of the form $x^t = \sum_{i=1}^{n+1} \lambda_i^t x_i^t$ of at most $n+1$ vectors $x_i^t$ from $X$, $i \leq n+1$, where $\lambda_i^t$ are the corresponding convex combination weights. Since $X$ is bounded, passing to a subsequence $t_1 < t_2 < \ldots$, we can assume that the sequences $\{x_i^{t_s}\}_{s\geq 1}$ and $\{\lambda_i^{t_s}\}_{s\geq 1}$ converge as $s \to \infty$ for every $i \leq n+1$, the limits being, respectively, vectors $x_i$ and reals $\lambda_i$. We clearly have $\lim_{t\to\infty} x^t = \lim_{s\to\infty} \sum_{i=1}^{n+1} \lambda_i^{t_s} x_i^{t_s} = \sum_{i=1}^{n+1} \lambda_i x_i$. In addition, as $X$ is closed, $x_i = \lim_{s\to\infty} x_i^{t_s} \in X$. Moreover, clearly $\lambda \geq 0$ and $\sum_{i=1}^{n+1} \lambda_i = 1$. The bottom line is that $\lim_{t\to\infty} x^t = \sum_{i=1}^{n+1} \lambda_i x_i$ is a convex combination of points from $X$ and thus belongs to $Y$ as $Y = \mathrm{Conv}(X)$. □

**Remark** I.2.6 Note that the convex hull of a closed *unbounded* set is not always closed. For example, consider the set $X = \{[0;0]\} \cup \big\{[u;v] \in \mathbf{R}^2_+ : uv = 1\big\}$ which is closed and unbounded, and we have $\mathrm{Conv}(X) = \big\{[u;v] \in \mathbf{R}^2_+ : u > 0, v > 0\big\} \cup \{[0;0]\}$ which is not closed.

Let us see a concrete illustration, taken from [Nem24], of Caratheodory Theorem.

### *2.1.1 Caratheodory Theorem, Illustration*

Suppose that a supermarket sells 99 different market blend herbal teas, and every herbal tea is a certain blend of 26 herbs A,...,Z. In spite of such a variety of marketed blends, John is not satisfied with any one of them; the only herbal tea he likes is their mixture, in the proportion

$$1 : 2 : 3 : \ldots : 98 : 99.$$

Once it occurred to John that in order to prepare his favorite tea, there is no necessity to buy all 99 market blends; a smaller number of them will do. With

some arithmetics, John found a combination of 66 marketed blends which still allows to prepare his tea. Do you believe John's result can be improved?
**Answer:** In fact, just 26 properly selected market blends are enough.
*Explanation:* Let us represent a blend by its unit weight portion, say, 1g. Such a portion can be identified with 26-dimensional vector $x = [x_1; \ldots; x_{26}]$, where $x_i$ is the weight, in grams, of herb $\#i$ in the portion. Clearly, we have $x \in \mathbf{R}^{26}_+$ and $\sum_{i=1}^{26} x_i = 1$. When mixing market blends $x^1, x^2, \ldots, x^{99}$ to get unit weight portion $x$ of mixture, we take $\lambda_j \geq 0$ grams of market blend $x^j$, $j = 1, \ldots, 99$, and mix them together, that is, $x = \sum_{j=1}^{99} \lambda_j x^j$. Since the weight of the mixture represented by $x$ is 1 gram and $\lambda_j$s corresponds to the weight (in grams) of market blends $x^j$ used in $x$, we get $\sum_{j=1}^{99} \lambda_j = 1$. The bottom line is that blend $x$ can be obtained by mixing market blends $x^1, \ldots, x^{99}$ if and only if $x \in \mathrm{Conv}\{x^1, \ldots, x^{99}\}$.
Then, by Caratheodory Theorem, every blend which can be obtained by mixing market blends can be obtained by mixing $m+1$ of them, where $m$ is the affine dimension of the affine span of $x^1, \ldots, x^{99}$. In our case, this span belongs to the 25-dimensional affine plane $\left\{x \in \mathbf{R}^{26} : \sum_{i=1}^{26} x_i = 1\right\}$ that is, $m \leq 25$. □

Caratheodory Theorem admits a "conic analogy" as follows:

**Fact** I.2.7 [Caratheodory Theorem in conic form] Let $a \in \mathbf{R}^m$ be a conic combination (linear combination with nonnegative coefficients) of $N$ vectors $a^1, \ldots, a^N$. Then, $a$ is a conic combination of at most $m$ vectors from the collection $a^1, \ldots, a^N$.

## 2.2 Radon Theorem

As an informal introduction to what follows, draw 4 arbitrary yet distinct from each other points on the plane and try to color some of them in red, and remaining in blue in such a way that the convex hull of red points will intersect the convex hull of the blue ones. Experimentation will suggest that this is always possible. The general fact is as follows.

**Theorem** I.2.8 [Radon Theorem] Let $x^i \in \mathbf{R}^n$, $i \leq N$, where $N \geq n+2$. Then, there exists a partition $I \cup J = \{1, \ldots, N\}$ of the index set $\{1, \ldots, N\}$ into two nonempty disjoint: $I \cap J = \varnothing$ sets $I$ and $J$ such that

$$\mathrm{Conv}\left(\{x^i : i \in I\}\right) \cap \mathrm{Conv}\left(\{x^j : j \in J\}\right) \neq \varnothing.$$

**Proof.** Consider the following system of homogeneous equations in $N$ variables

$\mu_1, \ldots, \mu_N$

$$\sum_{i=1}^{N} \mu_i x^i = 0,$$
$$\sum_{i=1}^{N} \mu_i = 0.$$

Note that as $x^i \in \mathbf{R}^n$, this system has $n+1$ scalar linear equations. Moreover, as the premise of the theorem states that $N > n+1$, we deduce that this system of equations has a nontrivial solution $\lambda_1, \ldots, \lambda_N$:

$$\sum_{i=1}^{N} \lambda_i x^i = 0, \quad \sum_{i=1}^{N} \lambda_i = 0, \quad \text{and } [\lambda_1; \ldots; \lambda_N] \neq 0.$$

Let $I := \{i : \lambda_i \geq 0\}$ and $J := \{i : \lambda_i < 0\}$. Then, $I$ and $J$ are nonempty and form a partition of $\{1, \ldots, N\}$ (since the sum of all $\lambda_i$'s is zero and not all $\lambda_i$'s are zero). Moreover, we have

$$a := \sum_{i \in I} \lambda_i = \sum_{j \in J} (-\lambda_j) > 0.$$

Then, by setting

$$\alpha_i := \frac{\lambda_i}{a}, \text{ for } i \in I, \qquad \text{and} \qquad \beta_j := \frac{-\lambda_j}{a}, \text{ for } j \in J,$$

we get

$$\alpha_i \geq 0, \forall i, \quad \beta_j \geq 0, \forall j, \quad \sum_{i \in I} \alpha_i = 1, \quad \sum_{j \in J} \beta_j = 1.$$

In addition, we also have,

$$\left[\sum_{i \in I} \alpha_i x^i\right] - \left[\sum_{j \in J} \beta_j x^j\right] = \frac{1}{a}\left(\left[\sum_{i \in I} \lambda_i x^i\right] - \left[\sum_{j \in J} (-\lambda_j) x^j\right]\right) = \frac{1}{a} \sum_{i=1}^{N} \lambda_i x^i = 0.$$

We conclude that the vector $\sum_{i \in I} \alpha_i x^i = \sum_{j \in J} \beta_j x^j$ is the desired common point of $\mathrm{Conv}(\{x^i : i \in I\})$ and $\mathrm{Conv}(\{x^j : j \in J\})$. □

**Remark** I.2.9 Same as in Caratheodory Theorem, the bound in Radon Theorem is sharp: for every positive integer $n$, there exist $n+1$ points in $\mathbf{R}^n$ (e.g., the origin and the $n$ standard basic orths) which cannot be split into two disjoint subsets with intersecting convex hulls.

## 2.3 Helly Theorem

What follows is a multidimensional extension of the nearly evident fact:

> *Given finitely many segments $[a_i, b_i]$ on the line such that every two of the segments intersect, you can find a point common to all the segments.*

Multidimensional extension of this fact is as follows.

**Theorem** I.2.10 [Helly Theorem, I] Let $\mathcal{F} := \{S_1, \ldots, S_N\}$ be a finite family of convex sets in $\mathbf{R}^n$. Suppose that for every collection of at most $n+1$ sets from this family, the sets from the collection have a point in common. Then, all of the sets $S_i$, $i \leq N$, have a common point.

**Proof.** We will prove the theorem by induction on the number $N$ of sets in the family. The case of $N \leq n+1$ holds immediately due to the premise of the theorem. So, suppose that the statement holds for all families with certain number $N \geq n+1$ of sets, and let $S_1, \ldots, S_N, S_{N+1}$ be a family of $N+1$ convex sets which satisfies the premise of Helly Theorem. We need to prove that the intersection of the sets $S_1, \ldots, S_N, S_{N+1}$ is nonempty.

For each $i \leq N+1$, we construct the following $N$-set families

$$\mathcal{F}^i := \{S_1, S_2, \ldots, S_{i-1}, S_{i+1}, \ldots, S_{N+1}\}, \qquad \forall i \leq N+1,$$

where the $N$-set family $\mathcal{F}_i$ is obtained by deleting from our $N+1$-set family the set $S_i$. Note that each of these $N$-set families $\mathcal{F}^i$ satisfies the premise of the Helly Theorem, and thus, by the inductive hypothesis, the intersection of the members of $\mathcal{F}^i$ is nonempty:

$$T^i := S_1 \cap S_2 \cap \ldots \cap S_{i-1} \cap S_{i+1} \cap \ldots \cap S_{N+1} \neq \varnothing, \qquad \forall i \leq N+1.$$

For each $i \leq N+1$, choose a point $x^i \in T^i$ (recall that $T^i$ is nonempty!). Then, we have $N+1$ points $x^i \in \mathbf{R}^n$. As $N \geq n+1$, we have $N+1 \geq n+2$ and by Radon Theorem, we can partition the index set $\{1, \ldots, N+1\}$ into two nonempty disjoint subsets $I$ and $J$ in such a way that certain convex combination $x$ of the points $x^i$, $i \in I$, is a convex combination of the points $x^j$, $j \in J$, as well. Let us verify that $x$ belongs to all the sets $S_1, \ldots, S_{N+1}$, which will complete the inductive step. Indeed, select any index $i^* \leq N+1$ and let us prove that $x \in S_{i^*}$. We have either $i^* \in I$ or $i^* \in J$. Suppose first that $i^* \in I$. Then, all the sets $T^j$, $j \in J$, are contained in $S_{i^*}$ (since $S_{i^*}$ participates in all intersections which give $T^i$ with $i \neq i^*$). Consequently, all the points $x^j$, $j \in J$, belong to $S_{i^*}$, and therefore $x$, which is a convex combination of these points, also belongs to $S_{i^*}$ (recall that all $S_i$ are convex!), as required. Suppose now that $i^* \in J$. In this case, a similar reasoning shows that all the points $x^i$, $i \in I$, belong to $S_{i^*}$, and therefore $x$, which is a convex combination of these points $x^i$, $i \in I$, belongs to $S_{i^*}$. The induction and the proof are complete. □

For an alternative proof of Theorem I.2.10 which does not utilize Radon Theorem, see Exercise I.23.

**Remark** I.2.11 Helly Theorem admits a small and immediate refinement as follows:

> *Let $\mathcal{F} := \{S_1, \ldots, S_N\}$ be a family of $N$ convex sets in $\mathbf{R}^n$, and let $m$ be the dimension of $\bigcup_{i \leq N} S_i$. Assume that for every collection of at most $m+1$ sets from the family, the sets from the collection have a point in common. Then, all sets from the family have a point in common.*

The justification of this claim follows from viewing $S_i$ as subsets of $E$ rather than $\mathbf{R}^n$ and applying the standard Helly Theorem.

**Remark** I.2.12 Same as Caratheodory and Radon Theorems, Helly Theorem is sharp: for every positive integer $n$, there exists a finite family of convex sets in $\mathbf{R}^n$ such that every $n$ of them have a common point, but all of them have no common point. Indeed, take $n+1$ affinely independent points $x^1, \ldots, x^{n+1}$ in $\mathbf{R}^n$ (say, the origin and the $n$ basic orths in $\mathbf{R}^n$) and $n+1$ convex sets $S_1,\ldots,S_{n+1}$ with $S_i$ being the convex hull of points $x^1, \ldots, x^{i-1}, x^{i+1}, \ldots, x^{n+1}$. Every $n$ of these sets have a point in common (e.g., the common point of $S_1, S_3, S_4, \ldots, S_{n+1}$ is $x^2$), but there is no point common to all $n+1$ sets (why?).

### 2.3.1 Helly Theorem, Illustration A

[1] Suppose that we need to design a factory which, mathematically, is described by the following set of constraints in variables $x \in \mathbf{R}^n$:

$$\left.\begin{array}{lcll} Ax & \geq & d & [d_1, \ldots, d_{1000}\text{: demands}] \\ Bx & \leq & f & [f_1 \geq 0, \ldots, f_{10} \geq 0\text{: amounts of resources of various types}] \\ Cx & \leq & c, & [\text{other constraints}] \end{array}\right\} \qquad (F)$$

where $Ax \geq d$ with a vector $d \in \mathbf{R}^{1000}$ represents the demand constraints, $Bx \leq f$ with a vector $f \in \mathbf{R}^{10}_+$ corresponds to availability of various resources, and the constraint $Cx \leq c$ represents various additional restrictions. The data $A, B, C, c$ are given in advance, but $d$ is unknown and we are asked to determine $f$. In particular, we are asked to buy *in advance* the resources $f_i \geq 0$, $i = 1, \ldots, 10$, in such a way that the factory will be able to satisfy all demand scenarios $d$ from a given finite set $D$, that is, $(F)$ should be feasible for every $d \in D$. The unit cost of resource $i$ is given to us as $a_i > 0$. We are given that the amount $f_i$ of resource $i$ costs us $a_i f_i$ with known $a_i > 0$.

It is known that for every single demand scenario $d \in D$ proper (depending on the scenario $d$) investment of at most \$1 in resources suffices to meet the demand $d$.

How large should the investment in resources be in the cases when $D$ contains

1. just one scenario?
2. 3 scenarios?
3. 10 scenarios?
4. 100,000 scenarios?

**Answer:** We claim that in these scenarios the following investment amounts will suffice:

1. \$1 investment is enough.
2. \$3 investment is enough.
3. \$10 investment is enough.

[1] The illustration to follow is taken from [Nem24].

4. \$11 investment is enough.

*Explanation:*
Cases 1 — 3: In these cases, we know that every scenario $d$ from $D$ can be met by a vector of resources $f_d \geq 0$ incurring a cost of at most \$1. Thus, when we are given scenarios $d^1, \ldots, d^k$ from $D$, we can meet every one of them with the vector of resources $f := f_{d^1} + \ldots + f_{d^k}$, since $f \geq f_{d^i}$ for every $i = 1, \ldots, k$. Then, due to the structure of our model, $f$ meets demand scenario $d^i$ using the resources $f_{d^i}$.
Case 4: We claim that \$11 investment is always enough no matter what the cardinality of $D$ is. To see this, for every $d \in D$, we define

$$S[d, f] := \{x \in \mathbf{R}^n : \ Ax \geq d, \ Bx \leq f, \ Cx \leq c\},$$
$$F_d := \left\{f \in \mathbf{R}^{10} : a^\top f \leq 11, \ S[d, f] \neq \varnothing\right\}.$$

Based on these definitions, $F_d$ is precisely the set of vectors $f$ such that their cost $a^\top f$ is at most \$11 and the associated polyhedral set $S[f, d]$ is nonempty, that is, resource $f$ allows to meet demand $d$. Note that the set $F_d$ is convex as it is the linear image (in fact just the projection) of the convex set

$$\left\{f \in \mathbf{R}^{10}, x \in \mathbf{R}^n : \ a^\top f \leq 11, \ x \in S[d, f]\right\}.$$

The punchline in this illustration is that every 11 sets of the form $F_d$ have a common point. Suppose that we are given 11 scenarios $d^1, \ldots, d^{11}$ from $D$. Then, we can meet demand scenario $d^i$ by investing \$1 in properly selected vector of resources $f_{d^i} \geq 0$. As we proceeded in the cases 1—3, by investing \$11 in the single vector of resources $f = f_{d^1} + \ldots + f_{d^{11}}$, we can meet every one of 11 scenarios $d^1, \ldots, d^{11}$, whence $f \in F_{d^1} \cap \ldots \cap F_{d^{11}}$. Since every 11 of 100,000 convex sets $F_d \subseteq \mathbf{R}^{10}$, $d \in D$, have a point in common, all these sets have a common point, say $f_*$. That is, $f_* \in F_d$ for all $d \in D$, and thus by definition of $F_d$, we deduce that every one of the sets $S[d, f_*]$, $d \in D$, is nonempty, that is, vector of resources $f_*$ (which costs at most \$11) allows us to satisfy every demand scenario $d \in D$. □

### *2.3.2 Helly Theorem, Illustration B*

Consider an optimization problem

$$\mathrm{Opt}_* := \min_{x \in \mathbf{R}^{11}} \left\{c^\top x : \ g_i(x) \leq 0, \ i = 1, \ldots, 1000\right\}$$

with 11 variables $x_1, \ldots, x_{11}$ and convex constraints, i.e., every one of the sets

$$X_i := \left\{x \in \mathbf{R}^{11} : \ g_i(x) \leq 0\right\}, \quad i = 1, \ldots, 1000,$$

is convex. Suppose also that the problem is solvable with optimal value $\mathrm{Opt}_* = 0$. Clearly, when dropping one or more constraints, the optimal value can only decrease or remain the same.
**Question:** Is it possible to find a constraint such that even if we drop it, we preserve the optimal value? Two constraints which can be dropped simultaneously with no effect on the optimal value? Three of them?

**Answer:** We can in fact drop as many as $1000 - 11 = 989$ appropriately chosen constraints without changing the optimal value!
*Explanation:* The case of $c = 0$ is trivial - here one can drop all 1000 constraints without varying the optimal value! Therefore, from now on we assume $c \neq 0$.

Assume for contradiction that every 11-constraint relaxation of the original problem has negative optimal value. Since there are finitely many such relaxations, there exists $\epsilon < 0$ such that every problem of the form

$$\min_x \left\{c^\top x : \; g_{i_1}(x) \leq 0, \ldots, g_{i_{11}}(x) \leq 0\right\}$$

has a feasible solution with the objective value $< -\epsilon$. Besides this, such an 11-constraint relaxation of the original problem has also a feasible solution with the objective equal to 0 (namely, the optimal solution of the original problem), and since its feasible set is convex (as the intersection of the convex feasible sets of the participating constraints), the relaxation has a feasible solution $x$ with $c^\top x = -\epsilon$. In other words, every 11 of the 1000 convex sets

$$Y_i := \left\{x \in \mathbf{R}^{11} : \; c^\top x = -\epsilon, \; g_i(x) \leq 0\right\}, \quad i = 1, \ldots, 1000$$

have a point in common. Now, consider the hyperplane $H := \{x \in \mathbf{R}^{11} : \; c^\top x = -\epsilon\}$. Note that $Y_i \subseteq H$ for all $i$. Moreover, as $c \neq 0$, $\dim(H) = 10$ and thus $\dim(\bigcup_{i=1}^{1000} Y_i) \leq \dim(H) = 10$. Since every 11 of these sets $Y_i$ have a nonempty intersection and $\dim(\bigcup_{i=1}^{1000} Y_i) \leq 10$, all of them have a point in common. In other words, the original problem should have a feasible solution with negative objective value, which is not possible as $\mathrm{Opt}_* = 0$. □

### *2.3.3 ★ Helly Theorem for infinite families of convex sets*

In Helly Theorem as presented above we dealt with a family of finitely many convex sets. To extend the statement to the case of infinite families, we need to slightly strengthen the assumptions, essentially, to avoid complications stemming from the following two situations:

- lack of closedness: every two (and in fact – finitely many) of convex sets $A_i = \{x \in \mathbf{R} : 0 < x \leq 1/i\}$, $i = 1, 2, \ldots$, have a point in common, while all the sets have no common point;
- "intersection at $\infty$": every two (and in fact – finitely many) of the closed convex sets $A_i = \{x \in \mathbf{R} : x \geq i\}$, $i = 1, 2, \ldots$, have a point in common, while all the sets have no common point.

Resulting refined statement of Helly Theorem for a family of (possibly) infinitely many sets is as follows:

**Theorem** I.2.13 [Helly Theorem, II] Let $\mathcal{F}$ be an arbitrary family of convex sets in $\mathbf{R}^n$. Assume that

(i) for every collection of at most $n + 1$ sets from the family, the sets from the collection have a point in common;

and

(ii) every set in the family is closed, and the intersection of the sets from a certain finite subfamily of the family is bounded (e.g., one of the sets in the family is bounded).

Then, all the sets from the family have a point in common.

**Proof.** By (i), Theorem I.2.10 implies that all finite subfamilies of $\mathcal{F}$ have nonempty intersections, and also these intersections are convex (since intersection of a family of convex sets is convex by Proposition I.1.12); in view of (ii) these intersections are also closed. Adding to $\mathcal{F}$ intersections of sets from finite subfamilies of $\mathcal{F}$, we get a larger family $\mathcal{F}'$ composed of closed convex sets, and sets from a finite subfamily of this larger family again have a nonempty intersection. Moreover, from (ii) it follows that this new family contains a bounded set $Q$. Since all the sets are closed, the family of sets

$$\{Q \cap Q' : Q' \in \mathcal{F}\}$$

forms a *nested family of compact sets* (i.e., a family of compact sets with nonempty intersection of sets from every finite subfamily). Then, by a well-known theorem from Real Analysis such a family has a nonempty intersection[2]. □

[2] Here is the proof of this Real Analysis theorem: assume for contradiction that the intersection of the compact sets $Q_\alpha$, $\alpha \in \mathcal{A}$, is empty. Choose a set $Q_{\alpha^*}$ from the family; for every $x \in Q_{\alpha^*}$ there is a set $Q^x$ in the family which does not contain $x$ (otherwise $x$ would be a common point of all our sets). Since $Q^x$ is closed, there is an open ball $V_x$ centered at $x$ which does not intersect $Q^x$. The balls $V_x$, $x \in Q_{\alpha^*}$, form an open covering of the compact set $Q_{\alpha^*}$. Since $Q_{\alpha^*}$ is compact, there exists a finite subcovering $V_{x_1}, \dots, V_{x_N}$ of $Q_{\alpha^*}$ by the balls from the covering, see Theorem B.19. Since $Q^{x_i}$ does not intersect $V_{x_i}$, we conclude that the intersection of the finite subfamily $Q_{\alpha^*}, Q^{x_1}, \dots, Q^{x_N}$ is empty, which is a contradiction.

# 3

# Polyhedral representations and Fourier-Motzkin elimination

## 3.1 Polyhedral representations

Recall that by definition a polyhedral set $X$ in $\mathbf{R}^n$ is the solution set of a finite system of nonstrict linear inequalities in variables $x \in \mathbf{R}^n$:

$$X = \{x \in \mathbf{R}^n : \; Ax \le b\} = \left\{x \in \mathbf{R}^n : \; a_i^\top x \le b_i, \, 1 \le i \le m\right\}.$$

We call such a representation of $X$ its *polyhedral description.* A polyhedral set is always convex and closed (Proposition I.1.2). We next introduce the notion of *polyhedral representation* of a set $X \subseteq \mathbf{R}^n$.

**Definition** I.3.1 A set $X \subseteq \mathbf{R}^n$ is called *polyhedrally representable* if it admits a representation of the form

$$X = \left\{x \in \mathbf{R}^n : \; \exists u \in \mathbf{R}^k : Ax + Bu \le c\right\}, \tag{3.1}$$

where $A$, $B$ are $m \times n$ and $m \times k$ matrices and $c \in \mathbf{R}^m$. A representation of $X$ of the form of (3.1) is called a *polyhedral representation of* $X$, and the variables $u \in \mathbf{R}^k$ in such a representation are called *extra variables.*

Geometrically, a polyhedral representation of a set $X \subseteq \mathbf{R}^n$ is its representation as the *projection* $\left\{x \in \mathbf{R}^n : \; \exists u \in \mathbf{R}^k : [x; u] \in Y\right\}$ of a *polyhedral* set $Y = \left\{(x, u) \in \mathbf{R}^n \times \mathbf{R}^k : \; Ax + Bu \le c\right\}$. Here, $Y$ lives in the space of $n+k$ variables $x \in \mathbf{R}^n$ and $u \in \mathbf{R}^k$, and the polyhedral representation of $X$ is obtained by applying the linear mapping (the projection) $[x; u] \mapsto x : \mathbf{R}^{n+k} \to \mathbf{R}^n$ of the $(n + k)$-dimensional space of $(x, u)$-variables (the space where $Y$ lives) to the $n$-dimensional space of $x$-variables where $X$ lives.

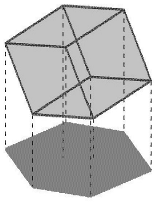


Figure I.5. Polyhedral representation of hexagon in $xy$-plane as the projection of rotated 3D cube onto the plane

Note that every polyhedrally representable set is the image under a linear mapping (even a projection) of a polyhedral, and thus convex, set. It follows that *a polyhedrally representable set is definitely convex* (Proposition I.1.21).

**Example** I.3.1 Every polyhedral set $X = \{x \in \mathbf{R}^n : Ax \leq b\}$ is polyhedrally representable: a polyhedral description of $X$ is nothing but a polyhedral representation with no extra variables ($k = 0$). Vice versa, a polyhedral representation of a set $X$ with no extra variables ($k = 0$) clearly is a polyhedral description of the set (which therefore is polyhedral).

**Example** I.3.2 Consider the set $X = \{x \in \mathbf{R}^n : \sum_{i=1}^n |x_i| \leq 1\}$. Note that this initial description of $X$ is *not* of the form $\{x \in \mathbf{R}^n : Ax \leq b\}$. Thus, from this description of $X$, we cannot immediately say whether it is polyhedral or not. However, $X$ admits a polyhedral representation, e.g., the following representation

$$X = \left\{ x \in \mathbf{R}^n : \exists u \in \mathbf{R}^n : \underbrace{-u_i \leq x_i \leq u_i}_{\Longleftrightarrow |x_i| \leq u_i}, 1 \leq i \leq n, \sum_{i=1}^n u_i \leq 1 \right\}. \tag{3.2}$$

Note that the set $X$ in question can be described by a system of linear inequalities *in $x$-variables only*, namely, as

$$X = \left\{ x \in \mathbf{R}^n : \sum_{i=1}^n \epsilon_i x_i \leq 1, \forall (\epsilon_i \in \{-1, +1\}), 1 \leq i \leq n) \right\},$$

thus, $X$ is polyhedral. However, the above polyhedral description of $X$ (which in fact is minimal in terms of the number of inequalities involved) requires $2^n$ inequalities — an astronomically large number when $n$ is just few tens. In contrast, the polyhedral representation of the same set given in (3.2) requires only $n$ extra variables $u$ and $2n + 1$ linear inequalities on $x, u$ and so the "complexity" of this representation is just linear in $n$.

**Example** I.3.3 Given a finite set of vectors $a^1, \ldots, a^m \in \mathbf{R}^n$, consider their conic hull $\mathrm{Cone}\,\{a^1, \ldots, a^m\} = \{\sum_{i=1}^m \lambda_i a^i : \lambda \geq 0\}$ (see section 1.2.4). From this definition, it is absolutely unclear whether this set is polyhedral. In contrast to this, its polyhedral representation is immediate:

$$\begin{aligned}\operatorname{Cone}\left\{a^1,\dots,a^m\right\} &= \left\{x\in\mathbf{R}^n:\ \exists\lambda\in\mathbf{R}^m_+:\ x=\sum_{i=1}^m\lambda_i a_i\right\}\\ &= \left\{x\in\mathbf{R}^n:\ \exists\lambda\in\mathbf{R}^m:\ \left\{\begin{array}{l}-\lambda\le 0\\ x-\sum_{i=1}^m\lambda_i a_i\le 0\\ -x+\sum_{i=1}^m\lambda_i a_i\le 0\end{array}\right.\right\}.\end{aligned}$$

In other words, the original description of $X$ is nothing but its polyhedral representation (in slight disguise), with $\lambda_i$'s in the role of extra variables.

## 3.2 Every polyhedrally representable set is polyhedral (Fourier-Motzkin elimination)

A surprising and deep fact is that the situation in Example I.3.2 above is quite general.

**Theorem** I.3.2 Every polyhedrally representable set is polyhedral.

**Proof: Fourier-Motzkin Elimination.** Recalling the definition of a polyhedrally representable set, our claim can be rephrased equivalently as follows:

> *The projection of a polyhedral set $Y$ in a space $\mathbf{R}^{n+k}_{x,u}$ of $(x,u)$-variables onto the subspace $\mathbf{R}^n_x$ of $x$-variables is a polyhedral set in $\mathbf{R}^n$.*

Note that it suffices to prove this claim in the case of exactly one extra variable since the projection which reduces the dimension by $k$ — "eliminates" $k$ extra variables — is the result of $k$ subsequent projections, every one reducing the dimension by 1, "eliminating" the extra variables one by one.

Thus, consider a polyhedral set with variables $x\in\mathbf{R}^n$ and $u\in\mathbf{R}$, i.e.,

$$Y:=\left\{(x,u)\in\mathbf{R}^{n+1}:\ a_i^\top x+b_i u\le c_i,\ 1\le i\le m\right\}.$$

We want to prove that the projection of $Y$ onto the space of $x$-variables, i.e.,

$$X:=\left\{x\in\mathbf{R}^n:\ \exists u\in\mathbf{R}:Ax+bu\le c\right\},$$

is polyhedral. To see this, let us split the indices of the inequalities defining $Y$ into three groups (some of these groups can be empty):

- inequalities with $b_i=0$: $I_0:=\{i:\ b_i=0\}$. These inequalities with index $i\in I_0$ do not involve $u$ at all;

- inequalities with $b_i>0$: $I_+:=\{i:\ b_i>0\}$. An inequality with index $i\in I_+$ can be rewritten equivalently as $u\le b_i^{-1}[c_i-a_i^\top x]$, and it imposes a (depending on $x$) *upper bound* on $u$;

- inequalities with $b_i < 0$: $I_- := \{i : b_i < 0\}$. An inequality with index $i \in I_-$ can be rewritten equivalently as $u \geq b_i^{-1}[c_i - a_i^\top x]$, and it imposes a (depending on $x$) *lower bound* on $u$.

We can now clearly answer the question of when $x$ can be in $X$, that is, when $x$ can be extended, by some $u$, to a point $(x,u)$ from $Y$: this is the case if and only if, first, $x$ satisfies all inequalities with $i \in I_0$, and, second, the inequalities with $i \in I_+$ giving the upper bounds on $u$ specified by $x$ are compatible with the inequalities with $i \in I_-$ giving the lower bounds on $u$ specified by $x$, meaning that every lower bound is less than or equal to every upper bound (the latter is necessary and sufficient to be able to find a value of $u$ which is greater than or equal to all lower bounds and less than or equal to all upper bounds). Thus,

$$X = \left\{ x \in \mathbf{R}^n : \begin{array}{ll} a_i^\top x \leq c_i, & \forall i \in I_0 \\ b_j^{-1}(c_j - a_j^\top x) \leq b_k^{-1}(c_k - a_k^\top x), & \forall j \in I_-,\ \forall k \in I_+. \end{array} \right\}.$$

We see that $X$ is given by finitely many nonstrict linear inequalities in $x$-variables only, as claimed. □

The outlined procedure for building polyhedral descriptions (i.e., polyhedral representations not involving extra variables) for projections of polyhedral sets is called *Fourier-Motzkin elimination.*

### *3.2.1 Some applications*

As an immediate application of Fourier-Motzkin elimination, let us take a linear program $\min\limits_x \{c^\top x : Ax \leq b\}$ and look at the set $T$ of possible objective values of all its feasible solutions (if any):

$$T := \left\{ t \in \mathbf{R} :\ \exists x \in \mathbf{R}^n : c^\top x = t,\ Ax \leq b \right\}.$$

Rewriting the linear equality $c^\top x = t$ as a pair of opposite inequalities, we see that $T$ is polyhedrally representable, and the above definition of $T$ is nothing but a polyhedral representation of this set, with $x$ in the role of the vector of extra variables. By Fourier-Motzkin elimination, $T$ is polyhedral − this set is given by a finite system of nonstrict linear inequalities in variable $t$ only. Thus, we immediately see that $T$ is

1. either empty (meaning that the LP in question is infeasible),
2. or is a below unbounded nonempty set of the form $\{t \in \mathbf{R} : -\infty \leq t \leq b\}$ with $b \in \mathbf{R} \cup \{+\infty\}$ (meaning that the LP is feasible and unbounded),
3. or is a below bounded nonempty set of the form $\{t \in \mathbf{R} : a \leq t \leq b\}$ with $a \in \mathbf{R}$ and $+\infty \geq b \geq a$. In this case, the LP is feasible and bounded, and its optimal value is $a$.

Note that given the list of linear inequalities defining $T$ (this list can be built algorithmically by Fourier-Motzkin elimination as applied to the original polyhedral representation of $T$), we can easily detect which one of the above cases indeed takes place, i.e., we can identify the feasibility and boundedness status of the LP and to find its optimal value. When it is finite (case 3. above), we can use the Fourier-Motzkin elimination backward, starting with $t = a \in T$ and extending this value to a pair $(t, x)$ with $t = a = c^\top x$ and $Ax \leq b$, that is, we can augment the optimal value by an optimal solution. Thus, we can say that Fourier-Motzkin elimination is a finite Real Arithmetics algorithm which allows to check whether an LP is feasible and bounded, and when it is the case, allows to find the optimal value and an optimal solution.

On the other hand, Fourier-Motzkin elimination is completely impractical since the elimination process can blow up exponentially the number of inequalities. Indeed, from the description of the process it is clear that if a polyhedral set is given by $m$ linear inequalities, then eliminating one variable, we can end up with as much as $m^2/4$ inequalities (this is what happens if there are $m/2$ indices in $I_+$, $m/2$ indices in $I_-$ and $I_0 = \varnothing$). Eliminating the next variable, we again can "nearly square" the number of inequalities, and so on. Thus, the number of inequalities in the description of $T$ can become astronomically large even when the dimension of $x$ is something like 10.

The actual importance of Fourier-Motzkin elimination is of theoretical nature. For example, the Linear Programming (LP)-related reasoning we have just carried out shows that

> *every feasible and bounded LP problem is solvable, i.e., it has an optimal solution.*

(We will revisit this result in more details in section 10.1.1). This is a fundamental fact for LP, and the above reasoning (even with the justification of the elimination "charged" to it) is, to the best of our knowledge, the shortest and most transparent way to prove this fundamental fact. Another application of the fact that polyhedrally representable sets are polyhedral is the Homogeneous Farkas Lemma to be stated and proved in section 4.1; this lemma will be instrumental in numerous subsequent theoretical developments.

## 3.3 Calculus of polyhedral representations

The fact that polyhedral sets are exactly the same as polyhedrally representable ones does not nullify the notion of a polyhedral representation. The point is that a set can admit "quite compact" polyhedral representation involving extra variables and require astronomically large, completely meaningless for any practical purpose, number of inequalities in its polyhedral description (think about Example I.3.2 and the associated set (3.2) when $n = 100$). Moreover, polyhedral representations admit a kind of "fully algorithmic calculus." Specifically, it turns out that all basic convexity-preserving operations (cf. Proposition I.1.21) as applied to polyhedral operands preserve polyhedrality. Moreover, polyhedral representations of the results are readily given by polyhedral representations of the operands. Here is the "algorithmic polyhedral analogy" of Proposition I.1.21:

1. *Taking finite intersection*: Let $M_i$, $1 \leq i \leq m$, be polyhedral sets in $\mathbf{R}^n$ given

by their polyhedral representations

$$M_i = \left\{x \in \mathbf{R}^n : \exists u^i \in \mathbf{R}^{k_i} : A_i x + B_i u^i \leq c^i\right\},\ 1 \leq i \leq m.$$

Then, the intersection of the sets $M_i$ is polyhedral with an explicit polyhedral representation, i.e.,

$$\begin{array}{l}\bigcap_{i=1}^m M_i \\ = \left\{x \in \mathbf{R}^n : \exists u = [u^1; \ldots; u^m] \in \mathbf{R}^{k_1+\ldots+k_m} : A_i x + B_i u^i \leq c^i,\ 1 \leq i \leq m\right\}.\end{array}$$

2. *Taking direct product*: Let $M_i \subseteq \mathbf{R}^{n_i}$, $1 \leq i \leq m$, be polyhedral sets given by polyhedral representations

$$M_i = \left\{x^i \in \mathbf{R}^{n_i} : \exists u^i \in \mathbf{R}^{k_i}]asntxtf{:}A_i x^i + B_i u^i \leq c^i\right\},\ 1 \leq i \leq m.$$

Then, the direct product

$$M_1 \times \ldots \times M_m := \left\{x = [x^1; \ldots; x^m] : x^i \in M_i,\ 1 \leq i \leq m\right\}$$

of the sets is a polyhedral set with explicit polyhedral representation, i.e.,

$$\begin{array}{l}M_1 \times \ldots \times M_m \\ = \left\{\begin{array}{r} x = [x^1; \ldots; x^m] \in \mathbf{R}^{n_1+\ldots+n_m} : \exists u = [u^1; \ldots; u^m] \in \mathbf{R}^{k_1+\ldots+k_m} : \\ A_i x^i + B_i u^i \leq c^i,\ 1 \leq i \leq m \end{array}\right\}.\end{array}$$

3. *Arithmetic summation and multiplication by reals*: Let $M_i \subseteq \mathbf{R}^n$, $1 \leq i \leq m$, be polyhedral sets given by polyhedral representations

$$M_i = \left\{x \in \mathbf{R}^n : \exists u^i \in \mathbf{R}^{k_i} : A_i x + B_i u^i \leq c^i\right\},\ 1 \leq i \leq m,$$

and let $\lambda_1, \ldots, \lambda_k$ be reals. Then, the set $\lambda_1 M_1 + \ldots + \lambda_m M_m := \{x = \lambda_1 x^1 + \ldots + \lambda_m x^m : x^i \in M_i,\ 1 \leq i \leq m\}$ is polyhedral with explicit polyhedral representation, specifically,

$$\begin{array}{l}\lambda_1 M_1 + \ldots + \lambda_m M_m \\ = \left\{\begin{array}{l} x \in \mathbf{R}^n : \exists (x^i \in \mathbf{R}^n, u^i \in \mathbf{R}^{k_i}, 1 \leq i \leq m) : \\ \qquad x \leq \sum_i \lambda_i x^i,\ x \geq \sum_i \lambda_i x^i,\ \ A_i x^i + B_i u^i \leq c^i,\ 1 \leq i \leq m \end{array}\right\}.\end{array}$$

4. *Taking the image under an affine mapping*: Let $M \subseteq \mathbf{R}^n$ be a polyhedral set given by polyhedral representation

$$M = \left\{x \in \mathbf{R}^n :\ \exists u \in \mathbf{R}^k : Ax + Bu \leq c\right\},$$

and let $\mathcal{P}(x) = Px + p : \mathbf{R}^n \to \mathbf{R}^m$ be an affine mapping. Then, the image of $M$ under this mapping, i.e., $\mathcal{P}(M) := \{Px + p :\ x \in M\}$, is a polyhedral set with explicit polyhedral representation given by

$$\mathcal{P}(M) = \left\{y \in \mathbf{R}^m :\ \exists (x \in \mathbf{R}^n, u \in \mathbf{R}^k) : \left\{\begin{array}{c} y \leq Px + p \\ y \geq Px + p \\ Ax + Bu \leq c \end{array}\right.\right\}.$$

5. *Taking the inverse image under affine mapping*: Let $M \subseteq \mathbf{R}^n$ be polyhedral set given by polyhedral representation
$$M = \left\{x \in \mathbf{R}^n : \ \exists u \in \mathbf{R}^k : Ax + Bu \leq c\right\},$$
and let $\mathcal{P}(y) = Py + p : \mathbf{R}^m \to \mathbf{R}^n$ be an affine mapping. Then, the inverse image of $M$ under this mapping, i.e., $\mathcal{P}^{-1}(M) := \{y \in \mathbf{R}^m : \ Py + p \in M\}$, is a polyhedral set with explicit polyhedral representation given by
$$\mathcal{P}^{-1}(M) = \left\{y \in \mathbf{R}^m : \ \exists u \in \mathbf{R}^k : A(Py + p) + Bu \leq c\right\}.$$

Note that the rules for intersection, taking direct products and taking inverse images, as applied to polyhedral *descriptions* of operands, lead to polyhedral descriptions of the results. In contrast to this, the rules for taking sums with coefficients and images under affine mappings heavily exploit the notion of polyhedral *representation*: even when the operands in these rules are given by polyhedral descriptions, there are no simple ways to point out polyhedral *descriptions* of the results.

Absolutely straightforward justification of the above calculus rules is the subject of Exercise I.27.

Finally, we note that the problem of minimizing a linear form $c^\top x$ over a set $M$ given by its polyhedral representation, i.e.,
$$M = \left\{x \in \mathbf{R}^n : \ \exists u \in \mathbf{R}^k : Ax + Bu \leq c\right\},$$
can be immediately reduced to an explicit LP program, namely,
$$\min_{x,u} \left\{c^\top x : \ Ax + Bu \leq c\right\}.$$

A reader with some experience in Linear Programming definitely used a lot of the above "calculus of polyhedral representations" when building LPs (perhaps without a clear understanding of what in fact is going on, same as Molière's Monsieur Jourdain all his life has been speaking prose without knowing it).

# 4

# General Theorem on Alternative and Linear Programming Duality

## 4.1 Homogeneous Farkas Lemma

Let $a_1, \ldots, a_N$ be vectors from $\mathbf{R}^n$, and let $a$ be another vector from $\mathbf{R}^n$. Here, we address the question: when does $a$ belong to the cone spanned by the vectors $a_1, \ldots, a_N$, i.e., when can $a$ be represented as a linear combination of $a_i$ with *nonnegative* coefficients? We immediately observe the following evident *necessary* condition:

$$\text{if } \ a = \sum_{i=1}^{N} \lambda_i a_i \qquad [\text{where } \lambda_i \geq 0,\ i = 1, \ldots, N],$$

then every vector $h$ that has nonnegative inner products with all $a_i$'s should also have nonnegative inner product with $a$:

$$\left\{ a = \sum_{i=1}^{N} \lambda_i a_i, \ \text{ with } \ \lambda_i \geq 0,\ \forall i, \ \text{ and } \ h^\top a_i \geq 0,\ \forall i \right\} \implies h^\top a \geq 0.$$

In fact, this evident necessary condition is also sufficient. This is given by the Homogeneous Farkas Lemma.

**Lemma** I.4.1 [Homogeneous Farkas Lemma (HFL)] Let $a, a_1, \ldots, a_N$ be vectors from $\mathbf{R}^n$. The vector $a$ is a conic combination of the vectors $a_i$ (linear combination with nonnegative coefficients), i.e., $a \in \mathrm{Cone}\,\{a_1, \ldots, a_N\}$, if and only if every vector $h$ satisfying $h^\top a_i \geq 0$, $i = 1, \ldots, N$, satisfies also $h^\top a \geq 0$. In other words, a homogeneous linear inequality

$$a^\top h \geq 0$$

in variable $h$ is a consequence of the system

$$a_i^\top h \geq 0, \qquad 1 \leq i \leq N$$

of homogeneous linear inequalities if and only if it can be obtained from the inequalities of the system by "admissible linear aggregation" – taking their weighted sum with nonnegative weights.

**Proof.** The necessity – the "only if" part of the statement – was proved before the Homogeneous Farkas Lemma was formulated. Let us prove the "if" part of the lemma. Thus, we assume that $h^\top a \geq 0$ is a *consequence* of the homogeneous system $h^\top a_i \geq 0\ \forall i$, i.e., every vector $h$ satisfying $h^\top a_i \geq 0\ \forall i$ satisfies also $h^\top a \geq 0$, and let us prove that $a$ is a conic combination of the vectors $a_i$.

An "intelligent" proof goes as follows. The set $\mathrm{Cone}\{a_1, \ldots, a_N\}$ of all conic combinations of $a_1, \ldots, a_N$ is polyhedrally representable (see Example I.3.3) and as such is polyhedral (Theorem I.3.2). Hence, we have

$$\mathrm{Cone}\{a_1, \ldots, a_N\} = \left\{x \in \mathbf{R}^n : \ p_j^\top x \geq b_j, 1 \leq j \leq J\right\}. \tag{4.1}$$

Now, observe that $0 \in \mathrm{Cone}\{a_1, \ldots, a_N\}$, and thus we conclude that $b_j \leq 0$ for all $j \in J$. Moreover, since $\lambda a_i \in \mathrm{Cone}\{a_1, \ldots, a_N\}$ for every $i$ and every $\lambda \geq 0$, we deduce $\lambda p_j^\top a_i \geq b_j$ for all $i, j$ and all $\lambda \geq 0$, whence $p_j^\top a_i \geq 0$ for all $i$ and $j$. For every $j$, the relation $p_j^\top a_i \geq 0$ for all $i$ implies, by the premise of the statement we want to prove, that $p_j^\top a \geq 0$. Then, as $0 \geq b_j$ for all $j$, we see that $p_j^\top a \geq b_j$ for all $j$, meaning that $a$ indeed belongs to $\mathrm{Cone}\{a_1, \ldots, a_N\}$ due to (4.1). □

This very short and elegant proof of Homogeneous Farkas Lemma is a nice illustration of the power of Fourier-Motzkin elimination.

## 4.2 Certificates for feasibility and infeasibility

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

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

where every $\Omega_i$ is either the relation "$>$" or the relation "$\geq$", and we assume $m \geq 1$, which is the only case of interest here.

The most basic question about $(\mathcal{S})$ is

($\mathcal{Q}$) *Does $(\mathcal{S})$ have a solution, i.e., is $(\mathcal{S})$ feasible?*

Knowing how to answer the question $(\mathcal{Q})$ enables us to answer many other questions. For example, verifying whether a given real number $a$ is a lower bound on the optimal value $\mathrm{Opt}^*$ of a linear program

$$\min_x \left\{c^\top x : \ Ax \geq b\right\} \tag{LP}$$

is the same as verifying whether the system

$$\begin{aligned} -c^\top x + a &> 0 \\ Ax - b &\geq 0 \end{aligned}$$

has no solutions.

The general question $(\mathcal{Q})$ above is too difficult, and it makes sense to pass from it to a seemingly simpler one:

*($\mathcal{Q}'$) How do we certify that $(\mathcal{S})$ has, or does not have, a solution?.*

Imagine that you are very smart and know the correct answer to $(\mathcal{Q})$ ; how can you convince everyone that your answer is correct? What can be an "evident for everybody" validity certificate for your answer?

If your claim is that $(\mathcal{S})$ is feasible, a *certificate* can be just to point out a solution $x^*$ to $(\mathcal{S})$. Given this certificate, one can substitute $x^*$ into the system and check whether $x^*$ is indeed a solution.

Suppose now that your claim is that $(\mathcal{S})$ has no solutions. What can be a "simple certificate" of this claim? How can one certify a *negative* statement? This is a highly nontrivial problem not just for mathematics; for example, in criminal law, how should someone accused in a murder prove his innocence? The "real life" answer to the question "how to certify a negative statement" is discouraging: such a statement normally *cannot* be certified[1]. In mathematics, the standard way to justify a negative statement **A**, like "there is no solution to such and such system of constraints" (e.g., "there is no solutions to the equation $x^5 + y^5 = z^5$ with positive integer variables $x, y, z$") is to lead the opposite to **A** statement, i.e., $\overline{\mathbf{A}}$ (in our example, "the solution exists"), to a contradiction. That is, assume that $\overline{\mathbf{A}}$ is true and derive consequences until a clearly false statement is obtained; when this happens, we know that $\overline{\mathbf{A}}$ is false (since legitimate consequences of a true statement must be true), and therefore **A** must be true. In general, there is no recipe for leading to contradiction something which in fact is false; this is why certifying negative statements usually is difficult.

Fortunately, finite systems of linear inequalities are simple enough to allow for a recipe for certifying their infeasibility: we start with the assumption that a solution exists and then demonstrate a contradiction in a very specific way – *by taking weighted sum of the inequalities in the system using nonnegative aggregation weights to produce a contradictory inequality.*

Let us start with a simple illustration: we would like to certify infeasibility of the following system of inequalities in variables $u, v, w$:

$$\begin{array}{rrrcr} 5u & -6v & -4w & > & 2 \\ & +4v & -2w & \geq & -1 \\ -5u & & +7w & \geq & 1 \end{array}$$

Let us assign these inequalities with "aggregation weights" $2, 3, 2$, multiply the inequalities by the respective weights and sum up the resulting inequalities:

$$\begin{array}{rr|rrrcr} & 2\times & 5u & -6v & -4w & > & 2 \\ + & & & & & & \\ & 3\times & & +4v & -2w & \geq & -1 \\ + & & & & & & \\ & 2\times & -5u & & +7w & \geq & 1 \\ \hline (*) & & 0\cdot u & +0\cdot v & +0\cdot w & > & 3 \end{array}$$

The resulting aggregated inequality $(*)$ is contradictory, it has no solutions at

[1] This is where the court rule "a person is presumed innocent until proven guilty" comes from – instead of requesting from the accused to certify the negative statement "I did not commit the crime," the court requests from the prosecution to certify the positive statement "The accused did commit the crime."

all. At the same time, $(*)$ is a consequence of our system – by construction of $(*)$, every solution to the original system of three inequalities is also feasible to $(*)$. Taken together, these two observations say that the system has no solutions, and the vector $[2;3;2]$ of our aggregation weights can be seen as an *infeasibility certificate* – taking weighted sum of inequalities from the system with the corresponding nonnegative weights, we lead the system to a contradiction.

As applied to a general system of inequalities $(\mathcal{S})$, a similar approach to certifying infeasibility would be to assign the inequalities with nonnegative aggregation weights, multiply them by these weights and sum up the resulting inequalities, arriving at an aggregated inequality, which, due to its origin, is a *consequence* of system $(\mathcal{S})$, meaning that every solution to the system solves the aggregated inequality as well. It follows that *when the aggregated inequality is contradictory*, i.e., it has no solutions at all, *the original system* $(\mathcal{S})$ *must be infeasible as well.* When this happens, the collection of weights used to generate the contradictory consequence inequality can be viewed as an infeasibility certificate for $(\mathcal{S})$.

Let us look what the outlined approach means when $(\mathcal{S})$ is composed of finitely many *linear* inequalities:

$$\left\{a_i^\top x\ \Omega_i\ b_i,\ i=1,\dots,m\right\} \qquad [\text{where } \Omega_i \text{ is either ``}>\text{'' or ``}\geq\text{''}]\,. \tag{$\mathcal{S}$}$$

In this case the "aggregated inequality" is linear as well:

$$\left(\sum_{i=1}^m \lambda_i a_i\right)^\top x \quad \Omega \quad \sum_{i=1}^m \lambda_i b_i, \tag{Comb($\lambda$)}$$

where $\Omega$ is "$>$" whenever $\lambda_i > 0$ for at least one $i$ with $\Omega_i =$ " $>$ ", and $\Omega$ is "$\geq$" otherwise. Now, when can a *linear* inequality

$$d^\top x\ \Omega\ e$$

be contradictory? Of course, it can happen only when $d = 0$. Furthermore, in this case, whether the inequality is contradictory depends on the relation $\Omega$ and the value of $e$: if $\Omega =$ " $>$ ", then the inequality is contradictory if and only if $e \geq 0$, and if $\Omega =$ " $\geq$ ", then it is contradictory if and only if $e > 0$. We have established the following simple result:

**Proposition** I.4.2 Consider a system of linear inequalities in unknowns $x \in \mathbf{R}^n$:

$$\begin{cases} a_i^\top x > b_i, & i = 1,\dots,m_{\mathrm{s}}, \\ a_i^\top x \geq b_i, & i = m_{\mathrm{s}}+1,\dots,m. \end{cases} \tag{$\mathcal{S}$}$$

Let us associate with $(\mathcal{S})$ two systems of linear inequalities and equations with unknowns $\lambda \in \mathbf{R}^m$:

$$\mathcal{T}_{\rm I}: \left\{\begin{array}{lrcl} (a) & \lambda & \geq & 0,\\ (b) & \sum_{i=1}^m \lambda_i a_i & = & 0,\\ (c_{\rm I}) & \sum_{i=1}^m \lambda_i b_i & \geq & 0,\\ \hline (d_{\rm I}) & \sum_{i=1}^{m_{\rm s}} \lambda_i & > & 0. \end{array}\right. \qquad \mathcal{T}_{\rm II}: \left\{\begin{array}{lrcl} (a) & \lambda & \geq & 0,\\ (b) & \sum_{i=1}^m \lambda_i a_i & = & 0,\\ (c_{\rm II}) & \sum_{i=1}^m \lambda_i b_i & > & 0. \end{array}\right.$$

If at least one of the systems $\mathcal{T}_{\rm I}$, $\mathcal{T}_{\rm II}$ is feasible, then the system $(\mathcal{S})$ is infeasible.

## 4.3 General Theorem on Alternative

Proposition I.4.2 states that in some cases it is easy to certify infeasibility of a linear system of inequalities: a "simple certificate" is a solution to another system of linear inequalities. Note, however, that the existence of a certificate of this latter type so far is only a *sufficient*, but not a *necessary*, condition for the infeasibility of $(\mathcal{S})$. A fundamental result in the theory of linear inequalities is that this sufficient condition is in fact also necessary:

**Theorem** I.4.3 [General Theorem on Alternative (GTA)] Consider the notation and setting of Proposition I.4.2. System $(\mathcal{S})$ has no solutions *if and only if* at least one of the systems $\mathcal{T}_{\rm I}$ or $\mathcal{T}_{\rm II}$ is feasible.

**Proof.** GTA is a more or less straightforward corollary of the Homogeneous Farkas Lemma. Indeed, in view of Proposition I.4.2, all we need to prove is that *if* $(\mathcal{S})$ has no solution, *then* at least one of the systems $\mathcal{T}_{\rm I}$, or $\mathcal{T}_{\rm II}$ is feasible. Thus, assume that $(\mathcal{S})$ has no solutions, and let us look at the consequences. Let us associate with $(\mathcal{S})$ the following system of *homogeneous* linear inequalities in variables $x$, $\tau$, $\epsilon$:

$$\begin{array}{lrrrcll} (a) & & \tau & -\epsilon & \geq & 0, & \\ (b) & a_i^\top x & -b_i\tau & -\epsilon & \geq & 0, & i=1,\ldots,m_{\rm s},\\ (c) & a_i^\top x & -b_i\tau & & \geq & 0, & i=m_{\rm s}+1,\ldots,m. \end{array} \tag{4.2}$$

First, we claim that *in every solution to* (4.2), *one has* $\epsilon \leq 0$. Indeed, assuming that (4.2) has a solution $x, \tau, \epsilon$ with $\epsilon > 0$, we conclude from (4.2.$a$) that $\tau > 0$. Then, from (4.2.$b-c$) it will follow that $\tau^{-1}x$ is a solution to $(\mathcal{S})$, while we assumed $(\mathcal{S})$ is infeasible. Therefore, we must have $\epsilon \leq 0$ in every solution to (4.2).

Now, we have that the homogeneous linear inequality

$$-\epsilon \geq 0 \tag{4.3}$$

is a consequence of the system of homogeneous linear inequalities (4.2). Then, by Homogeneous Farkas Lemma, there exist nonnegative weights $\nu$, $\lambda_i$, $i = 1, \ldots, m$, such that the aggregated inequality from (4.2) using these weights results in precisely the consequence inequality (4.3), i.e.,

$$\begin{array}{lrcl}(a) & \sum_{i=1}^{m}\lambda_i a_i & = & 0,\\ (b) & -\sum_{i=1}^{m}\lambda_i b_i+\nu & = & 0,\\ (c) & -\sum_{i=1}^{m_{\rm s}}\lambda_i-\nu & = & -1.\end{array} \tag{4.4}$$

Recall that by their origin, $\nu$ and all $\lambda_i$ are nonnegative. Now, it may happen that $\lambda_1,\ldots,\lambda_{m_{\rm s}}$ are zero. In this case $\nu=1$ by (4.4.$c$), and relations (4.4$a-b$) say that $\lambda_1,\ldots,\lambda_m$ is a solution for $\mathcal{T}_{\rm II}$. In the remaining case (that is, when not all $\lambda_1,\ldots,\lambda_{m_{\rm s}}$ are zero, or, which is the same, when $\sum_{i=1}^{m_{\rm s}}\lambda_i=1-\nu>0$), the same relations (4.4$a-b$) say that $\lambda_1,\ldots,\lambda_m$ is a solution for $\mathcal{T}_{\rm I}$. □

**Remark** I.4.4 We have derived GTA from Homogeneous Farkas Lemma (HFL). Note that HFL is nothing but a special case of GTA. Indeed, identifying when a linear inequality $a^\top x\leq b$ is a consequence of the system $a_i^\top x_i\leq b_i$, $1\leq i\leq m$ (this is the question answered by HFL in the case of $b=b_1=\ldots=b_m=0$) is exactly the same as identifying when the system of inequalities

$$a^\top x>b,\ \ a_i^\top x\leq b_i,\qquad i=1,\ldots,m \tag{$*$}$$

in variables $x$ is infeasible, and what in the latter case is said by GTA, is exactly what HFL states: *when $b=b_1=\ldots=b_m=0$, the system $(*)$ is infeasible if and only if the vector $a$ is a conic combination of the vectors $a_1,\ldots,a_m$*. Thus, it is completely sensible that GTA, in full generality, was derived from its independently justified special case, HFL.

## 4.4 Corollaries of GTA

Let us explicitly state two very useful principles derived from the General Theorem on Alternative:

**A.** A system of finitely many linear inequalities

$$a_i^\top x\ \Omega_i\ b_i,\ \ i=1,\ldots,m\qquad [\text{where } \Omega_i\in\{``\geq", ``>"\}]$$

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

**B.** A linear inequality

$$a_0^\top x\ \Omega_0\ b_0\qquad [\text{where } \Omega_0\in\{``\geq", ``>"\}]$$

in variables $x$ is a consequence of a *feasible* system of linear inequalities

$$a_i^\top x\ \Omega_i\ b_i,\ i=1,\ldots,m,\qquad [\text{where } \Omega_i\in\{``\geq", ``>"\}]$$

if and only if it can be obtained by *linear* aggregation with nonnegative weights from the inequalities of the system and the trivial identically true inequality $0^\top x>-1$.

In fact, when all $\Omega_i$ in **B** are non-strict, **B** can be reformulated equivalently as follows.

**Proposition** I.4.5 [Inhomogeneous Farkas Lemma] Linear inequality

$$a^\top x \leq b$$

is a consequence of the *feasible* system of linear inequalities

$$a_i^\top x \leq b_i, \quad 1 \leq i \leq m$$

if and only if there exist nonnegative aggregation weights $\lambda_i$, $i = 1, \ldots, m$, such that

$$a = \sum_{i=1}^m \lambda_i a_i \text{ and } b \geq \sum_{i=1}^m \lambda_i b_i.$$

We would like to emphasize that the preceding principles are highly nontrivial and very deep. Consider, e.g., the following system of 4 linear inequalities in two variables $u, v$:

$$\begin{aligned} -1 &\leq u \leq 1, \\ -1 &\leq v \leq 1. \end{aligned}$$

These inequalities clearly imply that

$$u^2 + v^2 \leq 2, \tag{!}$$

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

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

The concluding inequality $u + v \leq 2$ is linear and is a consequence of the original feasible system, and so we could have simply relied on Principle **B** to derive it. On the other hand, in the preceding demonstration of this linear consequence inequality both steps (!) and (!!) are "highly nonlinear." It is absolutely unclear a priori why the same consequence inequality can, as it is stated by Principle **B**, be derived from the system in a "linear" manner as well (of course it can – it suffices just to sum up two inequalities $u \leq 1$ and $v \leq 1$). In contrast, Inhomogeneous Farkas Lemma predicts that hundreds of pages of whatever complicated (but correct!) demonstration that such and such linear inequality is a consequence of such and such feasible finite system of linear inequalities can be replaced by simply demonstrating weights of prescribed signs such that the target inequality is the weighted sum, with these weights, of the inequalities from the system and the identically true linear inequality. One shall appreciate the elegance and depth of such a result!

Note that the General Theorem on Alternative and its corollaries **A** and **B** heavily exploit the fact that we are speaking about *linear* inequalities. For example, consider the following system of two quadratic and two linear inequalities in two variables:

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

along with the quadratic inequality

$$(e) \quad uv \geq 1.$$

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

## 4.5 Application: Linear Programming Duality

We are about to use the General Theorem on Alternative to obtain the basic results of the Linear Programming (LP) duality theory. To do so, we first introduce some basic terminology about mathematical programming problems.

### *4.5.1 Preliminaries: Mathematical and Linear Programming problems*

A (constrained) Mathematical Programming problem has the following form:

$$\text{(P)} \qquad \min_x \left\{ f(x) : \begin{array}{c} x \in X, \\ g(x) \equiv [g_1(x); \ldots; g_m(x)] \leq 0, \\ h(x) \equiv [h_1(x); \ldots; h_k(x)] = 0 \end{array} \right\}, \tag{4.5}$$

where

- [domain] $X$ is called the *domain* of the problem,
- [objective] $f$ is called the *objective* (function) of the problem,
- [constraints] $g_i$, $i = 1, \ldots, m$, are called the (functional) *inequality constraints*, and $h_j$, $j = 1, \ldots, k$, are called the *equality constraints*[2].

We always assume that $X \neq \varnothing$ and that the objective and the constraints are well-defined on $X$. Moreover, we typically skip indicating $X$ when $X = \mathbf{R}^n$.

We use the following standard terminology related to (4.5)

[2] Rigorously speaking, the constraints are not the *functions* $g_i$, $h_j$, but the *relations* $g_i(x) \leq 0$, $h_j(x) = 0$. We will use the word "constraints" in both of these senses, and it will always be clear what is meant. For example, we will say that "$x$ satisfies the constraints" to refer to the relations, and we will say that "the constraints are differentiable" to refer to the underlying functions.

- [feasible solution] a point $x \in \mathbf{R}^n$ is called a *feasible solution* to (4.5), if $x \in X$, $g_i(x) \le 0$, $i = 1, \dots, m$, and $h_j(x) = 0$, $j = 1, \dots, k$, i.e., if $x$ satisfies all restrictions imposed by the formulation of the problem.
  - [feasible set] the set of all feasible solutions is called the *feasible set* of the problem.
  - [feasible problem] a problem with a nonempty feasible set (i.e., the one which admits feasible solutions) is called *feasible* (or consistent).
- [optimal value] *the optimal value* of the problem refers to the quantity

$$\mathrm{Opt} := \begin{cases} \inf\limits_{x} \{f(x) : \, x \in X, \, g(x) \le 0, \, h(x) = 0\}, & \text{if the problem is feasible,} \\ +\infty, & \text{if the problem is infeasible.} \end{cases}$$

  - [below boundedness] the problem is called *below bounded*, if its optimal value is $> -\infty$, i.e., if the objective is bounded from below on the feasible set.
- [optimal solution] a point $x \in \mathbf{R}^n$ is called an *optimal solution* to (4.5), if $x$ is feasible and $f(x) \le f(x')$ for any other feasible solution $x'$, i.e., if

$$x \in \operatorname*{Argmin}_{x'} \{f(x') : \, x' \in X, \, g(x') \le 0, \, h(x') = 0\}.$$

  - [solvable problem] a problem is called *solvable*, if it admits optimal solutions.
  - [optimal set] the set of all optimal solutions to a problem is called its *optimal set.*

**Remark** I.4.6 In the above description of a Mathematical Programming problem and related basic notions, like feasibility, solvability, boundedness, etc., we "standardize" the situation by assuming that the objective should be minimized, and the inequality constraints are of the form $g_i(x) \le 0$. Needless to say, we can also speak about problems where the objective should be maximized and/or some of the inequality constraints are of the form $g_i(x) \ge 0$. There is no difficulty to reduce these "more general" forms of optimization problems to our standard form: maximizing $f(x)$ is the same as minimizing $-f(x)$, and the constraint of the form $g_i(x) \ge 0$ is the same as the constraint $-g_i(x) \le 0$. While this standardization is always possible, from time to time we take the liberty to speak about maximization problems and/or $\ge$-type constraints. With this in mind, it is worth to mention that when working with maximization problems, we should update the notions of optimal value, problem's boundedness, and optimal solution. For a maximization problem,

- the optimal value is the supremum of the values of the objective at feasible solutions, and is, by definition, $-\infty$ for infeasible problems, and
- boundedness means boundedness of the objective from above on the feasible set (or, which is the same, the fact that the optimal value is $< +\infty$),
- optimal solution is a feasible solution such that the objective value at this solution is greater than or equal to the objective value at every feasible solution.

Needless to say, when "standardizing" a maximization problem, i.e., replacing maximization of $f(x)$ with minimization of $-f(x)$, boundedness and optimal solutions remain intact while the optimal value "is negated," i.e., real number $a$ becomes $-a$, and $\pm\infty$ becomes $\mp\infty$.

**Linear Programming problems.** A Mathematical Programming problem (P) is called *Linear Programming* (LP) *problem*, if

- $X = \mathbf{R}^n$ is the entire space,
- $f, g_1, \dots, g_m$ are *real-valued affine* functions on $\mathbf{R}^m$, that is, functions of the form $a^\top x + b$, and
- there are no equality constraints at all.

Note that in principle we could allow for linear equality constraints $h_j(x) := a_j^\top x + b_j = 0$. However, a constraint of this type can be equivalently represented by a pair of opposite linear inequalities $a_j^\top x + b_j \leq 0, \ -a_j^\top x - b_j \leq 0$. To save space and words (and, as we have just explained, with no loss in generality), in the sequel we will focus on inequality constrained linear programming problems.

### *4.5.2 Dual to an LP problem: the origin*

Consider an LP problem

$$\mathrm{Opt} = \min_x \left\{c^\top x : \ Ax - b \geq 0\right\} \quad \left[\text{where } A = \begin{bmatrix} a_1^\top \\ a_2^\top \\ \cdots \\ a_m^\top \end{bmatrix} \in \mathbf{R}^{m\times n}\right]. \tag{LP}$$

The motivation for constructing the problem *dual* to an LP problem is the desire to generate, in a systematic way, lower bounds on the optimal value Opt of (LP).

Consider the problem

$$\min_x \left\{f(x) : \ g_i(x) \geq b_i, \ i = 1, \dots, m\right\}.$$

An evident way to bound from below a given function $f(x)$ in the domain given by a system of inequalities

$$g_i(x) \geq b_i, \quad i = 1, \dots, m, \tag{4.6}$$

is offered by what is called the *Lagrange duality*. We will discuss Lagrange Duality in full detail for general functions in Part IV. Here, let us do a brief precursor and examine the special case when we are dealing with linear functions only.

**Lagrange Duality:**

- *Let us look at all inequalities which can be obtained from* (4.6) *by linear aggregation, i.e., the inequalities of the form*

$$\sum_{i=1}^{m} y_i g_i(x) \geq \sum_{i=1}^{m} y_i b_i \tag{4.7}$$

*with the "aggregation weights"* $y_i \geq 0$ *for all* $i$. *Note that the inequality* (4.7), *due to its origin, is valid on the entire set* $\mathcal{X}$ *of feasible solutions of* (4.6).

- *Depending on the choice of aggregation weights, it may happen that the left hand side in* (4.7) *is* $\leq f(x)$ *for all* $x \in \mathbf{R}^n$. *Whenever this is the case, the right hand side* $\sum_{i=1}^{m} y_i b_i$ *of* (4.7) *is a lower bound on* $f(x)$ *for any* $x \in \mathcal{X}$. It follows that
  - *The optimal value of the problem*

$$\max_y \left\{ \sum_{i=1}^{m} y_i b_i : \begin{array}{rr} y \geq 0, & (a) \\ \sum_{i=1}^{m} y_i g_i(x) \leq f(x), \ \forall x \in \mathbf{R}^n & (b) \end{array} \right\} \tag{4.8}$$

*is a lower bound on the values of* $f$ *on the set of feasible solutions to the system* (4.6).

Let us now examine what happens with the Lagrange duality when $f$ and $g_i$ are homogeneous linear functions, i.e., $f(x) = c^\top x$ and $g_i(x) = a_i^\top x$ for all $i = 1, \ldots, m$. In this case, the requirement (4.8.$b$) merely says that $c = \sum_{i=1}^{m} y_i a_i$ (or, which is the same, $A^\top y = c$ due to the origin of the matrix $A$). Thus, problem (4.8) becomes the Linear Programming problem

$$\max_y \left\{ b^\top y : \ A^\top y = c, \ y \geq 0 \right\}, \tag{LP$^*$}$$

which is nothing but the LP dual of (LP).

By the construction of the dual problem (LP$^*$), we immediately have

> [Weak Duality] *The optimal value in* (LP$^*$) *is less than or equal to the optimal value in* (LP).

In fact, the "less than or equal to" in the latter statement is "equal to," provided that the optimal value Opt in (LP) is a number (i.e., (LP) is feasible and below bounded, in which case Fourier Motzkin elimination guarantees that Opt is a real number). To see that this indeed is the case, note that a real number $a$ is a lower bound on Opt if and only if $c^\top x \geq a$ holds for all $x$ satisfying $Ax \geq b$, or, which is the same, if and only if the system of linear inequalities

$$-c^\top x > -a, \quad Ax \geq b \tag{$\mathcal{S}_a$) :$$

has no solution. Then, by General Theorem on Alternative we deduce that at least one of a certain pair of systems of linear inequalities does have a solution. More precisely,

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

$$\mathcal{T}_{\rm I}: \quad \left\{\begin{array}{lrcl} (a) & \lambda=[\lambda_0;\lambda_1;\ldots;\lambda_m] & \geq & 0,\\ (b) & -\lambda_0 c+\sum_{i=1}^m \lambda_i a_i & = & 0,\\ \hline (c_{\rm I}) & -\lambda_0 a+\sum_{i=1}^m \lambda_i b_i & \geq & 0,\\ (d_{\rm I}) & \lambda_0 & > & 0; \end{array}\right.$$

*or*

$$\mathcal{T}_{\rm II}: \quad \left\{\begin{array}{lrcl} (a) & \lambda=[\lambda_0;\lambda_1;\ldots;\lambda_m] & \geq & 0,\\ (b) & -\lambda_0 c-\sum_{i=1}^m \lambda_i a_i & = & 0,\\ \hline (c_{\rm II}) & -\lambda_0 a-\sum_{i=1}^m \lambda_i b_i & > & 0. \end{array}\right.$$

Now assume that (LP) *is feasible*. We first claim that *under this assumption* $(\mathcal{S}_a)$ *has no solutions if and only if* $\mathcal{T}_{\rm I}$ *has a solution.* The implication "$\mathcal{T}_{\rm I}$ has a solution $\Longrightarrow$ $(\mathcal{S}_a)$ has no solution" is readily given by the preceding remarks. To verify the inverse implication, assume that $(\mathcal{S}_a)$ has no solution and the system $Ax \geq b$ has a solution, and let us prove that then $\mathcal{T}_{\rm I}$ has a solution. If $\mathcal{T}_{\rm I}$ has no solution, then by (*), $\mathcal{T}_{\rm II}$ must have a solution. Moreover, since any solution to $\mathcal{T}_{\rm II}$ where $\lambda_0>0$ is also a solution to $\mathcal{T}_{\rm I}$ as well, we must have $\lambda_0=0$ for (every) solution to $\mathcal{T}_{\rm II}$. But, the fact that $\mathcal{T}_{\rm II}$ has a solution $\lambda$ with $\lambda_0=0$ is independent of the values of $c$ and $a$; if this fact would take place, it would mean, by the same General Theorem on Alternative, that, e.g., the following instance of $(\mathcal{S}_a)$:

$$0^\top x \geq -1, \;\; Ax \geq b$$

has no solution as well. But, then we must have the system $Ax \geq b$ has no solution – a contradiction to the assumption that (LP) is feasible.

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

$$\begin{array}{rrcl} & y & \geq & 0,\\ \left[A^\top y \equiv\right] & \sum_{i=1}^m y_i a_i & = & c,\\ & b^\top y & \geq & a. \end{array} \tag{D}$$

We summarize these observations below.

**Proposition** I.4.7 If system (D) in unknowns $y, a$ associated with the LP program (LP) has a solution $(\bar{y}, \bar{a})$, then $\bar{a}$ is a lower bound on the optimal value in (LP). Vice versa, if (LP) is feasible and $\bar{a}$ is a lower bound on the optimal value of (LP), then $\bar{a}$ can be extended by a properly chosen $m$-dimensional vector $\bar{y}$ to a solution to (D).

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

$$\max_y \left\{ b^\top y : \; A^\top y = c, \; y \geq 0 \right\},$$

which is nothing but the dual to (LP) problem given by (LP$^*$). Note that (LP$^*$) is also a Linear Programming problem.

All we know about the dual problem so far is the following:

**Proposition** I.4.8 Whenever $y$ is a feasible solution to (LP$^*$), the corresponding value of the dual objective $b^\top y$ is a lower bound on the optimal value Opt in (LP). If (LP) is feasible, then for every $a \leq$ Opt there exists a feasible solution $y$ of (LP$^*$) with $b^\top y \geq a$.

### *4.5.3 Linear Programming Duality Theorem*

Proposition I.4.8 is in fact equivalent to the following complete statement of LP Duality Theorem.

**Theorem** I.4.9 [Duality Theorem in Linear Programming] Consider a linear programming problem

$$\min_x \left\{ c^\top x : \; Ax \geq b \right\}, \tag{LP}$$

along with its dual

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

Then,

1) [Primal-dual symmetry] The dual problem is an LP program, and its dual is equivalent to the primal problem;

2) [Weak duality] The value of the dual objective at every dual feasible solution is less than or equal to the value of the primal objective at every primal feasible solution, so that the dual optimal value is less than or equal to the primal one;

3) [Strong duality] The following 5 properties are equivalent to each other:

(i) The primal is feasible and bounded below.

(ii) The dual is feasible and bounded above.

(iii) The primal is solvable.

(iv) The dual is solvable.

(v) Both primal and dual are feasible.

Moreover, if any one of these properties (and then, by the equivalence just stated, every one of them) holds, then the optimal values of the primal and the dual problems are equal to each other.

Finally, if at least one of the problems in the primal-dual pair is feasible, then the optimal values in both problems are the same, i.e., either both are finite and equal to each other, or both are $+\infty$ (i.e., primal is infeasible and dual is not bounded above), or both are $-\infty$ (i.e., primal is unbounded below and dual is infeasible).

There is one last remark we should make to complete the story of primal and dual objective values given in Theorem I.4.9: in fact it is possible to have both primal and dual problems infeasible simultaneously (see Exercise I.38). This is the only case when the primal and the dual optimal values ($+\infty$ and $-\infty$, respectively) differ from each other.

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

$$\min_y \left\{ -b^\top y : \left[\begin{array}{c} I_m \\ A^\top \\ -A^\top \end{array}\right] y - \left[\begin{array}{r} 0 \\ c \\ -c \end{array}\right] \geq 0 \right\},$$

where $I_m$ is the $m \times m$ identity matrix. Applying the duality transformation to the latter problem, we come to the problem

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

which is clearly equivalent to (LP) (after we set $x = \zeta - \eta$ and eliminate $\xi$).

2) This part follows from the origin of the dual and is thus immediately given by Proposition I.4.8.

3) We prove the following implications.

(i) $\Longrightarrow$ (iv): If the primal is feasible and bounded below, its optimal value Opt (which of course is a lower bound on itself) can, by Proposition I.4.8, be (non-strictly) majorized by a quantity $b^\top y^*$, where $y^*$ is a feasible solution to (LP$^*$). Then, of course, $b^\top y^* = \mathrm{Opt}$ by already proven statement of item 2). On the other hand, by Proposition I.4.8, the optimal value in the dual is less than or equal to Opt. Thus, we conclude that the optimal value in the dual is attained and is equal to the optimal value in the primal.

(iv) $\Longrightarrow$ (ii): This is evident by the definition of solvability.

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

(iii) $\Longrightarrow$ (i): This is evident by the definition of solvability.

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

It remains to verify that if one problem in the primal-dual pair is feasible, then the primal and the dual optimal values are equal to each other. By primal-dual symmetry it suffices to consider the case when the primal problem is feasible. If also the primal is bounded from below, then by what has already been proved the dual problem is feasible and the primal and dual optimal values coincide with each other. If the primal problem is unbounded from below, then the primal optimal value is $-\infty$ and by Weak Duality the dual problem is infeasible, so that the dual optimal value is $-\infty$. □

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

**Theorem** I.4.10 [Necessary and sufficient optimality conditions in Linear Programming] Consider an LP program (LP) along with its dual (LP$^*$) as in Theorem I.4.9. A pair $(x, y)$ of primal and dual feasible solutions is composed of optimal solutions to the respective problems if and only if we have

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

or equivalently, if and only if

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

**Proof.** Indeed, the "zero duality gap" optimality condition is an immediate consequence of the fact that the value of primal objective at every primal feasible solution is greater than or equal to the value of the dual objective at every dual feasible solution, while the optimal values in the primal and the dual are equal to each other whenever one of the problem is feasible, see Theorem I.4.9. The equivalence between the "zero duality gap" and the "complementary slackness" optimality conditions is given by the following computation: whenever $x$ is primal feasible and $y$ is dual feasible, we have

$$y^\top(Ax - b) = (A^\top y)^\top x - b^\top y = c^\top x - b^\top y,$$

where the second equality follows from dual feasibility (i.e., $A^\top y = c$). Thus, for a primal-dual feasible pair $(x, y)$, the duality gap vanishes iff $y^\top(Ax - b) = 0$, and the latter, due to $y \geq 0$ and $Ax - b \geq 0$, happens iff $y_i[Ax - b]_i = 0$ for all $i$, that is, iff the complementary slackness takes place. □

**Geometry of primal-dual pair of LP problems.** Consider primal-dual pair of LP problems

$$\begin{array}{ll} \min_{x \in \mathbf{R}^n} \left\{c^\top x : Ax - b \geq 0\right\} & \text{(LP)} \\ \max_{y \in \mathbf{R}^m} \left\{b^\top y : A^\top y = c,\, y \geq 0\right\} & \text{(LP}^*\text{)} \end{array}$$

as presented in section 4.5.2 and assume that the system of equality constrains in

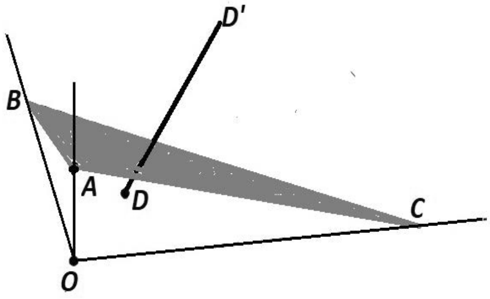


Figure I.6. Geometry of primal-dual LP pair, $m = 3$
$\mathcal{Q}$ : $\triangle ABC$; $\mathcal{Q}_*$: ray $DD'$; $\xi$: point $A$; $\xi_*$: point $D$
Pay attention to orthogonality of the ray and the plane of the triangle and orthogonality of vectors $\overrightarrow{OA}$ and $\overrightarrow{OD}$.

the dual problem is feasible, so that $A^\top\beta = c$ for some $\beta$. It turns out[3] that the pair (LP), (LP*) possesses nice and transparent geometry. Specifically, the data of the pair give rise to the following geometric entities:

- *a pair of linear subspaces $\mathcal{L}$, $\mathcal{L}_*$ in $\mathbf{R}^m$ which are orthogonal complements to each other*
  $[\mathcal{L} = \mathrm{Im}\,A, \mathcal{L}_* = \mathrm{Ker}A^\top]$
- *a pair of shift vectors $d, d_+ \in \mathbf{R}^m$*
  $[d = b, d_* = -\beta]$

which in turn give rise to

- *the pair of convex sets $\mathcal{Q} = [\mathcal{L} - d] \cap \mathbf{R}^m_+$, $\mathcal{Q}_* = [\mathcal{L}_* - d_*] \cap \mathbf{R}^m_+$*
  $[\mathbf{R}^m_+ = \{u \in \mathbf{R}^m : u \geq 0\}]$

*To solve* (LP), (LP*) *to optimality is exactly the same as to find orthogonal to each other vectors $\xi \in \mathcal{Q}$ and $\xi_* \in \mathcal{Q}_*$. Such a pair $\xi, \xi_*$ gives rise to primal-dual optimal pair(s) $(x^*, y^*)$* [one can take as $x^*$ any $x$ such that $Ax - b = \xi$ and set $y^* = \xi_*$], *and every primal-dual optimal pair $(x^*, y^*)$ can be obtained in this manner from a pair of orthogonal to each other vectors $\xi \in \mathcal{Q}$, $\xi_* \in \mathcal{Q}_*$. The required pairs $\xi, \xi_*$ exist iff both the sets $\mathcal{Q}$ and $\mathcal{Q}_*$ are nonempty.*

For illustration, see Figure I.6.

[3] for derivations, see Exercise IV.7 addressing Conic Duality, of which LP duality is a special case.

# 5

# Exercises for Part I

## 5.1 Elementaries

**Exercise** I.1 Mark in the following list the sets which are convex:

1. $\left\{x \in \mathbf{R}^2 : x_1 + i^2 x_2 \le 1,\ i = 1, \dots, 10\right\}$
2. $\left\{x \in \mathbf{R}^2 : x_1^2 + 2ix_1x_2 + i^2x_2^2 \le 1,\ i = 1, \dots, 10\right\}$
3. $\left\{x \in \mathbf{R}^2 : x_1^2 + ix_1x_2 + i^2x_2^2 \le 1,\ i = 1, \dots, 10\right\}$
4. $\left\{x \in \mathbf{R}^2 : x_1^2 + 5x_1x_2 + 4x_2^2 \le 1\right\}$
5. $\left\{x \in \mathbf{R}^{10} : x_1^2 + 2x_2^2 + 3x_3^2 + \ldots + 10x_{10}^2 \le 1000x_1 - 999x_2 + 998x_3 - \ldots + 992x_9 - 991x_{10}\right\}$
6. $\left\{x \in \mathbf{R}^2 : \exp\{x_1\} \le x_2\right\}$
7. $\left\{x \in \mathbf{R}^2 : \exp\{x_1\} \ge x_2\right\}$
8. $\left\{x \in \mathbf{R}^n : \sum_{i=1}^n x_i^2 = 1\right\}$
9. $\left\{x \in \mathbf{R}^n : \sum_{i=1}^n x_i^2 \le 1\right\}$
10. $\left\{x \in \mathbf{R}^n : \sum_{i=1}^n x_i^2 \ge 1\right\}$
11. $\left\{x \in \mathbf{R}^n : \max\limits_{i=1,\dots,n} x_i \le 1\right\}$
12. $\left\{x \in \mathbf{R}^n : \max\limits_{i=1,\dots,n} x_i \ge 1\right\}$
13. $\left\{x \in \mathbf{R}^n : \max\limits_{i=1,\dots,n} x_i = 1\right\}$
14. $\left\{x \in \mathbf{R}^n : \min\limits_{i=1,\dots,n} x_i \le 1\right\}$
15. $\left\{x \in \mathbf{R}^n : \min\limits_{i=1,\dots,n} x_i \ge 1\right\}$
16. $\left\{x \in \mathbf{R}^n : \min\limits_{i=1,\dots,n} x_i = 1\right\}$

**Exercise** I.2 Mark by **T** those of the following claims which always are true.

1. The linear image $Y = \{Ax : x \in X\}$ of a linear subspace $X$ is a linear subspace.
2. The linear image $Y = \{Ax : x \in X\}$ of an affine subspace $X$ is an affine subspace.
3. The linear image $Y = \{Ax : x \in X\}$ of a convex set $X$ is convex.
4. The affine image $Y = \{Ax + b : x \in X\}$ of a linear subspace $X$ is a linear subspace.
5. The affine image $Y = \{Ax + b : x \in X\}$ of an affine subspace $X$ is an affine subspace.
6. The affine image $Y = \{Ax + b : x \in X\}$ of a convex set $X$ is convex.
7. The intersection of two linear subspaces in $\mathbf{R}^n$ is always nonempty.
8. The intersection of two linear subspaces in $\mathbf{R}^n$ is a linear subspace.
9. The intersection of two affine subspaces in $\mathbf{R}^n$ is an affine subspace.
10. The intersection of two affine subspaces in $\mathbf{R}^n$, when nonempty, is an affine subspace.
11. The intersection of two convex sets in $\mathbf{R}^n$ is a convex set.
12. The intersection of two convex sets in $\mathbf{R}^n$, when nonempty, is a convex set.

**Exercise** I.3 ▲ Prove that the relative interior of a simplex with vertices $y^0, \dots, y^m$ is exactly the set

$$\left\{ \sum_{i=0}^{m} \lambda_i y_i : \ \lambda_i > 0, \ \sum_{i=0}^{m} \lambda_i = 1 \right\}.$$

**Exercise** I.4 Which of the following claims is true:

1. The set $X = \{x : Ax \le b\}$ is a cone if and only if $X = \{x : Ax \le 0\}$.
2. The set $X = \{x : Ax \le b\}$ is a cone if and only if $b = 0$.

**Exercise** I.5 Suppose $\mathbf{K}$ is a closed cone. Prove that the set $X = \{x : Ax - b \in \mathbf{K}\}$ is a cone if and only if $X = \{x : Ax \in \mathbf{K}\}$.

**Exercise** I.6 ▲ Prove that if $M$ is a nonempty convex set in $\mathbf{R}^n$ and $\epsilon > 0$, then for every norm $\|\cdot\|$ on $\mathbf{R}^n$, the $\epsilon$-neighborhood of $M$, i.e., the set

$$M_\epsilon = \left\{ y \in \mathbf{R}^n : \ \inf_{x \in M} \|y - x\| \le \epsilon \right\},$$

is convex.

**Exercise** I.7 Which of the following claims are always true? Explain why/why not.

1. The convex hull of a bounded set in $\mathbf{R}^n$ is bounded.
2. The convex hull of a closed set in $\mathbf{R}^n$ is closed.
3. The convex hull of a closed convex set in $\mathbf{R}^n$ is closed.
4. The convex hull of a closed and bounded set in $\mathbf{R}^n$ is closed and bounded.
5. The convex hull of an open set in $\mathbf{R}^n$ is open.

**Exercise** I.8 ▲ [This exercise together with its follow-up, i.e., Exercise II.9, and Exercise I.9 are the most boring exercises ever designed by the authors. Our excuse is that *There is no royal road to geometry* (Euclid of Alexandria, c. 300 BC)]
Let $A$, $B$ be nonempty subsets of $\mathbf{R}^n$. Consider the following claims. If the claim is always (i.e., for every data satisfying premise of the claim) true, give a proof; otherwise, give a counter example.

1. If $A \subseteq B$, then $\mathrm{Conv}(A) \subseteq \mathrm{Conv}(B)$.
2. If $\mathrm{Conv}(A) \subseteq \mathrm{Conv}(B)$, then $A \subseteq B$.
3. $\mathrm{Conv}(A \cap B) = \mathrm{Conv}(A) \cap \mathrm{Conv}(B)$.
4. $\mathrm{Conv}(A \cap B) \subseteq \mathrm{Conv}(A) \cap \mathrm{Conv}(B)$.
5. $\mathrm{Conv}(A \cup B) \subseteq \mathrm{Conv}(A) \cup \mathrm{Conv}(B)$.
6. $\mathrm{Conv}(A \cup B) \supseteq \mathrm{Conv}(A) \cup \mathrm{Conv}(B)$.
7. If $A$ is closed, so is $\mathrm{Conv}(A)$.
8. If $A$ is closed and bounded, so is $\mathrm{Conv}(A)$.
9. If $\mathrm{Conv}(A)$ is closed and bounded, so is $A$.

**Exercise** I.9 ▲ Let $A, B, C$ be nonempty subsets of $\mathbf{R}^n$ and $D$ be a nonempty subset of $\mathbf{R}^m$. Consider the following claims. If the claim is always (i.e., for every data satisfying premise of the claim) true, give a proof; otherwise, give a counter example.

1. $\mathrm{Conv}(A \cup B) = \mathrm{Conv}(\mathrm{Conv}(A) \cup B)$.
2. $\mathrm{Conv}(A \cup B) = \mathrm{Conv}(\mathrm{Conv}(A) \cup \mathrm{Conv}(B))$.
3. $\mathrm{Conv}(A \cup B \cup C) = \mathrm{Conv}(\mathrm{Conv}(A \cup B) \cup C)$.
4. $\mathrm{Conv}(A \times D) = \mathrm{Conv}(A) \times \mathrm{Conv}(D)$.
5. When $A$ is convex, the set $\mathrm{Conv}(A \cup B)$ (which is always the set of convex combinations of several points from $A$ and several points from $B$), can be obtained by taking convex combinations of points with *at most one of them* taken from $A$, and the rest taken from $B$. Similarly, if $A$ and $B$ are both convex, to get $\mathrm{Conv}(A \cup B)$, it suffices to add to $A \cup B$ all convex combinations of pairs of points, one from $A$ and one from $B$.

6. Suppose $A$ is a set in $\mathbf{R}^n$. Consider the affine mapping $x \mapsto Px + p : \mathbf{R}^n \to \mathbf{R}^m$, and the image of $A$ under this mapping, i.e., the set $PA + p := \{Px + p : x \in A\}$. Then, $\mathrm{Conv}(PA + p) = P\,\mathrm{Conv}(A) + p$.
7. Consider an affine mapping $y \mapsto P(y) : \mathbf{R}^m \to \mathbf{R}^n$ where $P(y) := Py + p$. Recall that given a set $X \in \mathbf{R}^n$, its inverse image under the mapping $P(\cdot)$ is given by $P^{-1}(X) := \{y \in \mathbf{R}^m : \; P(y) \in X\}$. Then, $\mathrm{Conv}(P^{-1}(A)) = P^{-1}(\mathrm{Conv}(A))$.
8. Consider an affine mapping $y \mapsto P(y) : \mathbf{R}^m \to \mathbf{R}^n$ where $P(y) := Py+p$. Then, $\mathrm{Conv}(P^{-1}(A)) \subseteq P^{-1}(\mathrm{Conv}(A))$.

**Exercise** I.10 ▲ Let $X_1, X_2 \in \mathbf{R}^n$ be two nonempty sets, and define $Y := X_1 \cup X_2$ and $Z := \mathrm{Conv}(Y)$. Consider the following claims. If the claim is always (i.e., for every data satisfying premise of the claim) true, give a proof; otherwise, give a counter example.

1. Whenever $X_1$ and $X_2$ are both convex, so is $Y$.
2. Whenever $X_1$ and $X_2$ are both convex, so is $Z$.
3. Whenever $X_1$ and $X_2$ are both bounded, so is $Y$.
4. Whenever $X_1$ and $X_2$ are both bounded, so is $Z$.
5. Whenever $X_1$ and $X_2$ are both closed, so is $Y$.
6. Whenever $X_1$ and $X_2$ are both closed, so is $Z$.
7. Whenever $X_1$ and $X_2$ are both compact, so is $Y$.
8. Whenever $X_1$ and $X_2$ are both compact, so is $Z$.
9. Whenever $X_1$ and $X_2$ are both polyhedral, so is $Y$.
10. Whenever $X_1$ and $X_2$ are both polyhedral, so is $Z$.
11. Whenever $X_1$ and $X_2$ are both polyhedral and bounded, so is $Y$.
12. Whenever $X_1$ and $X_2$ are both polyhedral and bounded, so is $Z$.

**Exercise** I.11 Consider two families of convex sets given by $\{F_i\}_{i\in I}$ and $\{G_j\}_{j\in J}$. Prove that the following relation holds:

$$\mathrm{Conv}\left(\bigcup_{i\in I,\, j\in J} (F_i \cap G_j)\right) \subseteq \mathrm{Conv}\left(\bigcup_{j\in J} [G_j \cap \mathrm{Conv}(\cup_{i\in I} F_i)]\right).$$

**Exercise** I.12 Let $C_1, C_2$ be two nonempty conic sets in $\mathbf{R}^n$, i.e., for each $i = 1, 2$, for any $x \in C_i$ and $t \geq 0$, we have $t \cdot x \in C_i$ as well. Note that $C_1, C_2$ are not necessarily convex. Prove that

1. $C_1 + C_2 \neq \mathrm{Conv}(C_1 \cup C_2)$ may happen if either $C_1$ or $C_2$ (or both) is nonconvex.
2. $C_1 + C_2 = \mathrm{Conv}(C_1 \cup C_2)$ always holds if $C_1, C_2$ are both convex.
3. $C_1 \cap C_2 = \bigcup_{\alpha\in[0,1]} (\alpha C_1 \cap (1-\alpha)C_2)$ always holds if $C_1, C_2$ are both convex.

**Exercise** I.13 ▲ Let $X \subseteq \mathbf{R}^n$ be a convex set with $\mathrm{int}\, X \neq \varnothing$, and consider the following set

$$K := \mathrm{cl}\,\{[x; t] : \; t > 0,\, x/t \in X\}.$$

Prove that the set $K$ is a closed cone with a nonempty interior.

## 5.2 Around ellipsoids

**Exercise** I.14 Verify each of the following statements:

1. Any ellipsoid $E \in \mathbf{R}^n$ is the images of the unit Euclidean ball $B_n = \{x \in \mathbf{R}^n : \|x\|_2 \leq 1\}$ under a one-to-one affine mapping. That is, $E \subset \mathbf{R}^n$ can be represented as $E = \{x : (x-c)^\top C(x-c) \leq 1\}$ with $C \succ 0$ and $c \in \mathbf{R}^n$ if and only if it can be represented as $E = \{c + Du : u \in B_n\}$ with nonsingular $D$, and in the latter representation $D$ can be selected to be symmetric positive definite.

2. Given $C \succ 0$, $D \succ 0$ and $c, d \in \mathbf{R}^n$, the ellipsoid $E_C := \{x : (x-c)^\top C(x-c) \leq 1\}$ is contained in the ellipsoid $E_D := \{x : (x-c)^\top D(x-c) \leq 1\}$ if and only if $C \succeq D$. If the ellipsoid $E_C$ is contained in the ellipsoid $E'_D = \{x : (x-d)^\top D(x-d) \leq 1\}$, then $C \succeq D$.
3. For a set $U \subset \mathbf{R}^n$, let $\mathrm{Vol}(U)$ be the ratio of the $n$-dimensional volume of $U$ and the $n$-dimensional volume of the unit ball $B_n$. Then, for an $n$-dimensional ellipsoid $E$ represented as $\{x = c + Du : \|u\|_2 \leq 1\}$ with nonsingular $D$ we have
$$\mathrm{Vol}(E) = |\mathrm{Det}(D)|,$$
and when $E$ is represented as $\{x : (x-c)^\top C(x-c) \leq 1\}$ with $C \succ 0$, we have
$$\mathrm{Vol}(E) = \mathrm{Det}^{-1/2}(C).$$

**Exercise** I.15 Given $C \succ 0$, an ellipsoid $\{x : (x-a)^\top C(x-a) \leq 1\}$ is the solution set of quadratic inequality $x^\top Cx - 2(Ca)^\top x + (a^\top Ca - 1) \leq 0$. Prove that the solution set $E$ of any quadratic inequality $f(x) := x^\top Cx - c^\top x + \sigma \leq 0$ with positive *semi*definite matrix $C$ is convex.

## 5.3 Truss Topology Design

**Exercise** I.16 ♦ [First acquaintance with Truss Topology Design]
**Preamble.** What follows is the first exercise in a "Truss Topology Design" (TTD) series ((other exercises in it are I.18, III.9, IV.11, IV.28). The underlying "real life" mechanical story is simple enough to be told and rich enough to illustrate numerous constructions and results presented in the main body of our textbook – ranging from Caratheodory Theorem to semidefinite duality, demonstrating on a real life example how the theory works.
**Trusses.** Truss is a mechanical construction, like railroad bridge, electric mast, of Eiffel Tower, composed of thin elastic *bars* linked with each other at *nodes* – points from physical space (3D space for spatial, and 2D space for planar trusses).

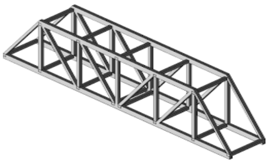


Figure I.7. Pratt Truss Bridge
source: `https://grabcad.com/library/pratt-truss-bridge-2`

When truss is subject to external load – collection of forces acting at the nodes – it starts to deform, so that the nodes move a little bit, leading to elongations/shortenings of bars, which, in turn, result in reaction forces. At the equilibrium, the reaction forces compensate the external ones, and the truss capacitates certain potential energy, called *compliance*. Mechanics models this story as follows.

- The nodes form a finite set $p_1, \ldots, p_K$ of distinct points in physical space $\mathbf{R}^d$ ($d = 2$ for planar, and $d = 3$ for spatial constructions). Virtual displacements of the nodes under the load are somehow restricted by "support conditions;" we will focus on the case when some of the nodes "are fixed" – cannot move at all (think about them as being in the wall), and the remaining "are free" – their virtual displacements form the entire $\mathbf{R}^d$. A virtual displacement $v$ of the nodal set can be identified with a vector of dimension $M = dm$, where $m$ is the number of free nodes; $v$ is block vector with $m$ $d$-dimensional blocks, indexed by the free nodes, representing physical displacements of these nodes.
- There are $N$ bars, $i$-th of them linking the nodes with indexes $\alpha_i$ and $\beta_i$ (with at least one of these nodes free) and with volume (3D or 2D, depending on whether the truss is spatial or planar) $t_i$.

- An external load is a collection of physical forces – vectors from $\mathbf{R}^d$ – acting at the free nodes (forces acting at the fixed nodes are of no interest – they are suppressed by the supports). Thus, an external load $f$ can be identified with block vector of the same structure as a virtual displacement – blocks are indexed by free nodes and represent the external forces acting at these nodes. Thus, displacements $v$ of the nodal set and external loads $f$ are vectors from the space $\mathcal{V}$ of *virtual displacements* – $M$-dimensional block vectors with $m$ $d$-dimensional blocks.
- The bars and the nodes together specify the symmetric positive semidefinite $M \times M$ *stiffness matrix* $A$ of the truss. The role of this matrix is as follows. A displacement $v \in \mathcal{V}$ of the nodal set results in reaction forces at free nodes (those at fixed nodes are of no interest – they are compensated by supports); assembling these forces into $M$-dimensional block-vector, we get a *reaction*, and this reaction is $-Av$. In other words, the potential energy capacitated in truss under displacement $v \in \mathcal{V}$ of nodes is $\frac{1}{2}v^\top Av$, and reaction, as it should be, is the minus gradient of the potential energy as a function of $v$ [1]. At the equilibrium under external load $f$, the total of the reaction and the load should be zero, that is, the equilibrium displacement satisfies

$$Av = f \tag{5.1}$$

Note that (5.1) may be unsolvable, meaning that the truss is crushed by the load in question. Assuming the equilibrium displacement $v$ exists, the truss at equilibrium capacitates potential energy $\frac{1}{2}v^\top Av$; this energy is called *compliance* of the truss w.r.t. the load. Compliance is convenient measure of rigidity of the truss with respect to the load, the less the compliance the better the truss withstands the load.

Let us build the stiffness matrix of a truss. As we have mentioned, the reaction forces originate from elongations/shortenings of bars under displacement of nodes. Consider $i$-th bar linking nodes with initial – prior to the external load being applied – positions $a_i = p_{\alpha_i}$ and $b_i = p_{\beta_i}$, and let us set

$$d_i = \|b_i - a_i\|_2,\ e_i = [b_i - a_i]/d_i.$$

Under displacement $v \in V$ of the nodal set,

- positions of the nodes linked by the bar become $a_i + \underbrace{v^{\alpha_i}}_{da}$, $b_i + \underbrace{v^{\beta_i}}_{db}$, where $v^\gamma$ is $\gamma$-th block in $v$ – the displacement of $\gamma$-th node
- as a result, elongation of the bar becomes, in the first-order in $v$ approximation, $e_i^\top[db - da]$, and the reaction forces caused by this elongation by Hooke's Law[2] are

$$\begin{array}{cl} d_i^{-1}S_ie_ie_i^\top[db-da] & \text{at node \# } \alpha_i \\ -d_i^{-1}S_ie_ie_i^\top[db-da] & \text{at node \# } \beta_i \\ 0 & \text{at all remaining nodes} \end{array}$$

where $S_i = t_i/d_i$ is the cross-sectional size of $i$-th bar. It follows that *when both nodes linked by i-th bar are free, the contribution of i-th bar to the reaction is*

$$-t_i\mathfrak{b}_i\mathfrak{b}_i^\top v,$$

[1] This is called *linearly elastic* model; it is the linearized in displacements approximation of the actual behavior of a loaded truss. This model works the better the smaller are the nodal displacements as compared to the inter-nodal distances, and is accurate enough to be used in typical real-life applications.

[2] Hooke's Law says that the magnitude of the reaction force caused by elongation/shortening of a bar is proportional to $Sd^{-1}\delta$, where $S$ is bar's cross-sectional size (area for spatial, and thickness for planar truss), $d$ is bar's (pre-deformation) length, and $\delta$ is the elongation. With units of length properly adjusted to bars' material, the proportionality coefficient becomes 1, and this is what we assume from now on.

*where $\mathfrak{b}_i \in \mathcal{V}$ is the vector with just two nonzero blocks:*
*— the block with index $\alpha_i$ – this block is $e_i/d_i = [b_i - a_i]/\|b_i - a_i\|_2^2$, and*
*— the block with index $\beta_i$ – this block is $-e_i/d_i = -[b_i - a_i]/\|b_i - a_i\|_2^2$.*
It is immediately seen that when just one of the nodes linked by $i$-th bar is free, the contribution of $i$-th bar to the reaction is given by similar relations, but with one, rather than 2, blocks in $\mathfrak{b}_i$ – the one corresponding to the free among the nodes linked by the bar.

The bottom line is that *The stiffness matrix of a truss composed of $N$ bars with volumes $t_i$, $1 \leq i \leq N$, is*

$$A = A(t) := \sum_i t_i \mathfrak{b}_i \mathfrak{b}_i^\top,$$

*where $\mathfrak{b}_i \in \mathcal{V} = \mathbf{R}^M$ are readily given by the geometry of nodal set and the indexes of nodes linked by bar $i$.*

**Truss Topology Design problem.** In the simplest Truss Topology Design (TTD) problem, one is given

- a finite *set of tentative nodes* in 2D or 3D along with support conditions indicating which of the nodes are fixed and which are free, and thus specifying the linear space $\mathcal{V} = \mathbf{R}^M$ of virtual displacements of the nodal set,
- the *set of $N$ tentative bars* – unordered pairs of (distinct from each other) nodes which are allowed to be linked by bars, and the total volume $W > 0$ of the truss,
- An external load $f \in \mathcal{V}$.

These data specify, as explained above, vectors $\mathfrak{b}_i \in \mathbf{R}^M$, $i = 1, \ldots, N$, and the stiffness matrix

$$A(t) = \sum_{i=1}^N t_i \mathfrak{b}_i \mathfrak{b}_i^\top = B \operatorname{Diag}\{t_1, \ldots, t_N\} B^\top \in \mathbf{S}^M \qquad [B = [\mathfrak{b}_1, \ldots, \mathfrak{b}_N]]$$

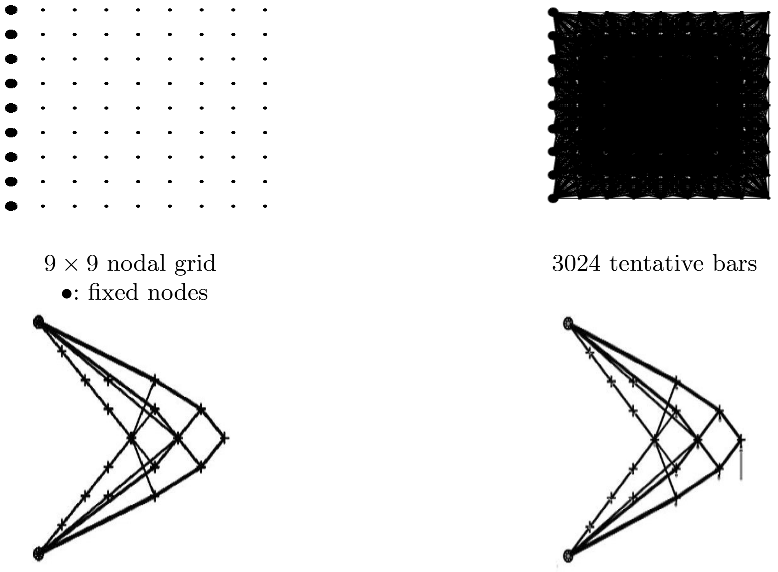
of truss, which under the circumstances can be identified with vector $t \in \mathbf{R}^N_+$ of bar volumes. What we want is to find the truss of given volume capable to “withstand best of all” the given load, that is, the one that minimizes the corresponding compliance.

When applying the TTD model, one starts with dense grid of tentative nodes and broad list of tentative bars (e.g., by allowing to link by a bar every pair of distinct from each other nodes, with at least one of the nodes in the pair free). At the optimal truss yielded by the optimal solution to the TTD problem, many tentative bars (usually vast majority of them) get zero volumes, and significant part of the tentative nodes become unused. Thus, TTD problem in fact is not about sizing – it allows to recover optimal structure of the construction, this is where “Topology Design” comes from.

To illustrate this point, here is a toy example (it will be our guinea pig in the entire series of TTD exercises):

**Console design:** We want to design a 2D truss as follows:

- The set of tentative nodes is the $9 \times 9$ grid $\{[p;q] \in \mathbf{R}^2 : p,q \in \{0,1,\dots,8\}\}$ with the 9 most-left nodes fixed and remaining 72 nodes free, resulting in $M = 144$-dimensional space $\mathcal{V}$ of virtual displacements
- The external load $f \in \mathcal{V} = \mathbf{R}^{144}$ is a single-force one, with the only nonzero force $[0;-1]$ applied at the 5-th node of the most-right column of nodes.
- We allow for all pairwise connections of pairs of distinct from each other nodes, with at least one of these nodes free, resulting in $N = 3204$ tentative bars
- The total volume of truss is $W = 1000$.



$9 \times 9$ nodal grid
●: fixed nodes

3024 tentative bars

optimal truss, 38 bars
compliance 0.1914

displacement under
load of interest

Figure I.8. Console. Bars and nodes' positions before (crosses) and after (gray dots) deformation. Gray segment starting at the most right node: external force

**Important:** *From now on, speaking about TTD problem, we always make the following assumption:*

$$\mathfrak{R}: \qquad \sum_{t=1}^N \mathfrak{b}_i \mathfrak{b}_i^\top \succ 0.$$

*Under this assumption, the stiffness matrix $A(t) = \sum_i t_i \mathfrak{b}_i \mathfrak{b}_i^\top$ associated with truss $t > 0$ is positive definite, so that such a truss can withstand whatever load $f$.* You can verify numerically that this is the case in Console design as stated above.

After this lengthy preamble (to justify its length, note that it is investment to a series of exercises, rather than just one of them), let us pass to the exercise per se. Consider a TTD problem.

1. Prove that truss $t \geq 0$ (recall that we identify truss with the corresponding vector of bar volumes) is capable to carry load $f$ if and only if the quadratic function

$$F(v) = f^\top v - \frac{1}{2} v^\top A(t) v$$

is bounded from above, and that whenever this takes place,

- the maximum of $F$ over $\mathcal{V}$ is achieved
- the maximizers of $F$ are exactly the equilibrium displacements $v$ – those with
$$A(t)v = f,$$
and for such a displacement, one has
$$[\max F =]\ F(v) = \frac{1}{2}v^\top A(t)v = \frac{1}{2}v^\top f$$
- the maximum value of $F$ is exactly the compliance of the truss w.r.t. the load $f$

2. Prove that a real $\tau$ is an upper bound on the compliance of truss $t \geq 0$ w.r.t. load $f$ if and only if the symmetric matrix
$$\mathcal{A} = \left[\begin{array}{c|c} B\,\mathrm{Diag}\{t\}B^\top & f \\ \hline f^\top & 2\tau \end{array}\right],\ B = [\mathfrak{b}_1, \ldots, \mathfrak{b}_N]$$
is positive semidefinite. As a result, pose the TTD problem as the optimization problem
$$\mathrm{Opt} = \min_{\tau,r}\left\{\tau : \left[\begin{array}{c|c} B\,\mathrm{Diag}\{t\}B^\top & f \\ \hline f^\top & 2\tau \end{array}\right] \succeq 0, t \geq 0, \sum_i t_i = W\right\} \tag{5.2}$$
Prove that the problem is solvable.
3. [computational study]

3.1. Solve the Console problem numerically and reproduce the numerical results presented above.

3.2. Resolve the problem with the set of all possible tentative bars reduced to the subset of "short" bars connecting neighboring nodes only:

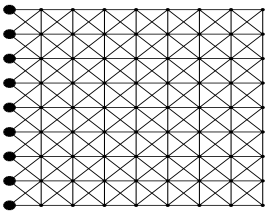


Figure I.9. 262 "short" tentative bars

and compare the resulting design and compliance to those in the previous item.

## 5.4 Around Caratheodory Theorem

**Exercise** I.17 ♦ Prove the following statement: Let $X \subset \mathbf{R}^n$ be nonempty. Then

1. if a point $x$ can be represented as a convex combination of a collection of vectors from $X$, then the collection can be selected to be affinely independent.
2. if a point $x$ can be represented as a conic combination of a collection of vectors from $X$, then the collection can be selected to be linearly independent,

Note that the claims above are refinements, albeit minor ones, of the Caratheodory Theorem (plain and conic, respectively). Indeed, when $M = \mathrm{Aff}(X)$ and $m$ is the dimension of $M$, every affine independent collection of points from $X$ contains at most $m + 1$ points (Proposition A.44), so that the first claim implies that if $x \in \mathrm{Conv}(X)$, then $x$ is a convex combination of at most $m+1$ points from $X$; however, the vectors participating in such a combination are not necessarily affinely independent, so that the first claim provides a bit more information than the plain Caratheodory's Theorem. Similarly, if $L = \mathrm{Lin}(X)$ and $m = \dim L$, then every linearly independent collection of vectors from $X$ contains at most $m \leq n$ points, that is, the second claim implies the Caratheodory's Theorem in conic form, and provides a bit more information than the latter theorem.

**Exercise** I.18 ♦ [3] Consider TTD problem, and let $N$ be the number of tentative bars, $M$ be the dimension of the corresponding space of virtual displacements $\mathcal{V}$, and $f$ be an external load. Prove that if truss $t \geq 0$ can withstand load $f$ with compliance $\leq \tau$ for some given real $\tau$, then there exists truss $\bar{t}$ of the same total volume as $t$ with compliance w.r.t. $f$ at most $\tau$ and at most $M+1$ bars of positive volume.

**Exercise** I.19 ♦ [Shapley-Folkman Theorem]

1. Prove that if a system of linear equations $Ax = b$ with $n$ variables and $m$ equations has a nonnegative solution, it has a nonnegative solution with at most $m$ positive entries.
2. Let $V_1, \ldots, V_n$ be $n$ nonempty sets in $\mathbf{R}^m$, and define

$$\overline{V} := \mathrm{Conv}(V_1 + V_2 + \ldots + V_n).$$

   1. Prove that

$$\overline{V} = \mathrm{Conv}(V_1) + \ldots + \mathrm{Conv}(V_n).$$

   2. Prove *Shapley-Folkman Theorem*:

> Let $x \in \overline{V}$. Then, there exists a representation of $x$ such that
>
> $$x = x_1 + \ldots + x_n, \quad x_i \in \mathrm{Conv}(V_i),$$
>
> in which at least $n - m$ of $x_i$'s belong to the respective sets $V_i$.

*Comment:* Shapley-Folkman Theorem says, informally, that when $n \gg m$, summing up $n$ nonempty sets in $\mathbf{R}^m$ possesses certain "convexification property" – every point from the convex hull $\overline{V}$ of the sum of our sets is the sum of points $x_i$ with all but $m$ of them belonging to $V_i$ rather than to $\mathrm{Conv}(V_i)$, and only $\leq m$ of the points belonging to $V_i$ "fractionally," that is, belonging to $\mathrm{Conv}(V_i)$, but not to $V_i$. This nice fact has numerous useful applications.

**Exercise** I.20 ♦ Caratheodory's Theorem in its plain and its conic forms are "existence" statements: if a point $x \in \mathbf{R}^m$ is a convex, respectively conic, combination of points $x^1, \ldots, x^N$, then *there exists* a representation of $x$ of the same type which involves at most $(m+1)$, respectively, $m$, terms. Extract from the proofs of the theorems *algorithms* for finding these "short" representations at the cost of solving at most $N$ solvable systems of linear equations with at most $N$ variables and $m$ equations each.

**Exercise** I.21 ♦ Prove *Kirchberger's Theorem*:

> Consider two sets of finitely many points $X = \{x^1, \ldots, x^k\}$ and $Y = \{y^1, \ldots, y^m\}$ in $\mathbf{R}^n$ such that $k + m \geq n + 2$ and all the points $x^1, \ldots, x^k, y^1, \ldots, y^m$ are distinct. Assume that for any subset $S \subseteq X \cup Y$ composed of $n+2$ points the convex hulls of the sets $X \cap S$ and $Y \cap S$ do not intersect: $\mathrm{Conv}(X \cap S) \cap \mathrm{Conv}(Y \cap S) = \varnothing$. Then, the convex hulls of $X$ and $Y$ also do not intersect: $\mathrm{Conv}(X) \cap \mathrm{Conv}(Y) = \varnothing$.

*Hint:* Assume, on contrary, that the convex hulls of $X$ and $Y$ intersect, so that

$$\sum_{i=1}^{k} \lambda_i x^i = \sum_{j=1}^{m} \mu_j y^j$$

[3] Preceding exercise in the TTD series is I.16.

for certain nonnegative $\lambda_i$, $\sum_{i=1}^k \lambda_i = 1$, and certain nonnegative $\mu_j$, $\sum_{j=1}^m \mu_j = 1$, and look at the expression of this type with the minimum possible total number of nonzero coefficients $\lambda_i$, $\mu_j$.

**Exercise** I.22 ♦ [Follow-up to Shapley-Folkman Theorem]

1. Let $X_1, \dots, X_K$ be nonempty convex sets in $\mathbf{R}^n$ and $X = \bigcup_{k \leq K} X_k$. Prove that

$$\mathrm{Conv}(X) = \left\{ x = \sum_{k=1}^K \lambda_k x^k : \ \lambda_k \geq 0, \, x^k \in X_k, \, \forall k \leq K, \ \sum_{k=1}^K \lambda_k = 1 \right\}.$$

2. Let $X_k$, $k \leq K$, be nonempty bounded polyhedral sets in $\mathbf{R}^n$ given by polyhedral representations:

$$X_k = \left\{ x \in \mathbf{R}^n : \ \exists u^k \in \mathbf{R}^{n_k} : P_k x + Q_k u^k \leq r_k \right\}.$$

Define $X := \bigcup_{k \leq K} X_k$. Prove that the set $\mathrm{Conv}(X)$ is a polyhedral set given by the polyhedral representation

$$\mathrm{Conv}(X) = \left\{ x \in \mathbf{R}^n : \begin{array}{ll} \exists x^k \in \mathbf{R}^n, \ u^k \in \mathbf{R}^{n_k}, \ \lambda_k \in \mathbf{R}, \ \forall k \leq K : & \\ P_k x^k + Q_k u^k - \lambda_k r_k \leq 0, \ k \leq K & (a) \\ \lambda_k \geq 0, \ \sum_{k=1}^K \lambda_k = 1 & (b) \\ x = \sum_{k=1}^K x^k & (c) \end{array} \right\}. \qquad (*)$$

Does the claim remain true when the assumption of boundedness of the sets $X_k$s is lifted?

After two preliminary items above, let us pass to the essence of the matter. Consider the situation as follows. We are given $n$ nonempty and bounded polyhedral sets $X_j \subset \mathbf{R}^r$, $j = 1, \dots, n$. We will think of $X_j$ as the "resource set" of the $j$-th production unit: entries in $x \in X_j$ are amounts of various resources, and $X_j$ describes the set of vectors of resources available, in principle, for $j$-th unit. Each production unit $j$ can possibly use any one of its $K_j < \infty$ different production plans. For each $j = 1, \dots, n$, the vector $y_j \in \mathbf{R}^p$ representing the production of the $j$-th unit depends on the vector $x_j$ of resources consumed by the unit and also on the production plan utilized in the unit. In particular, the production vector $y_j \in \mathbf{R}^p$ stemming from resources $x_j$ under $k$-th plan can be picked by us, at our will, from the set

$$Y_j^k[x_j] := \left\{ y_j \in \mathbf{R}^p : \ z_j := [x_j; -y_j] \in V_j^k \right\},$$

where $V_j^k$, $k \leq K_j$, are given bounded polyhedral "technological sets" of the units with projections onto the $x_j$-plane equal to $X_j$, so that for every $k \leq K_j$ it holds

$$x_j \in X_j \quad \Longleftrightarrow \quad \exists y_j : [x_j; -y_j] \in V_j^k. \qquad (5.3)$$

We assume that all the sets $V_j^k$ are given by polyhedral representations, and we define

$$V_j := \bigcup_{k \leq K_j} V_j^k.$$

Let $R \in \mathbf{R}^r$ be the vector of total resources available to all $n$ units and let $P \in \mathbf{R}^p$ be the vector of total demands for the products. For $j \leq n$, we want to select $x_j \in X_j$, $k_j \leq K_j$, and $y_j \in Y_j^{k_j}[x_j]$ in such a way that

$$\sum_j x_j \leq R \quad \text{and} \quad \sum_j y_j \geq P.$$

That is, we would like to find $z_j = [x_j; v_j] \in V_j$, $j \leq n$, in such a way that $\sum_j z_j \leq [R; -P]$. Note that the presence of "combinatorial part" in our decision – selection of production plans in finite sets – makes the problem difficult.

3. Apply Shapley-Folkman Theorem (Exercise I.19) to overcome, to some extent, the above difficulty and come up with a good and approximately feasible solution.

## 5.5 Around Helly Theorem

**Exercise** I.23 ▲ [Alternative proof of Helly Theorem] The goal of this exercise is to build an alternative proof of Helly's Theorem, without using Radon's Theorem.

1. Consider a system $a_i^\top x \leq b_i$, $i \leq N$, of $N$ linear inequalities in variables $x \in \mathbf{R}^n$. Helly's Theorem applied to the sets $A_i := \{x \in \mathbf{R}^n : \; a_i^\top x \leq b_i\}$ gives us that

   > (!) *If a system $a_i^\top x \leq b_i$, $i \leq N$, of linear inequalities in variables $x \in \mathbf{R}^n$ infeasible, so is a properly selected sub-system composed of at most $n+1$ inequalities from the system.*

   Find an alternative proof of (!) without relying on Helly's or Radon's Theorems.
2. Extract from item 1 Helly's Theorem for polyhedral sets: *If $A_1, \ldots, A_N$, $N \geq n+1$, are polyhedral sets in $\mathbf{R}^n$ and every $n+1$ of these sets have a point in common, then all the sets have a point in common.*
3. Extract from item 2 Helly's Theorem (Theorem I.2.10).

**Exercise** I.24 ▲ $A_0, A_1, \ldots, A_m$, $m = 2025$, are nonempty convex subsets of $\mathbf{R}^{2000}$, and $A_0$ is a triangle (convex hull of 3 affinely independent vectors). Which of the claims below are always (that is, for any $A_0, \ldots, A_m$ satisfying the above assumptions) true:

1. If every 3 among the sets $A_0, \ldots, A_m$ have a point in common, all $m+1$ sets have a point in common.
2. If every 4 among the sets $A_0, \ldots, A_m$ have a point in common, all $m+1$ sets have a point in common.
3. If every 2001 among the sets $A_0, \ldots, A_m$ have a point in common, all $m+1$ sets have a point in common.

**Exercise** I.25 ▲ Let $P_i := \{x \in \mathbf{R}^n : \; A_i x \leq b_i\}$ for $i \in \{1, \ldots, m\}$ and $C := \{x \in \mathbf{R}^n : \; Dx \geq d\}$ be nonempty polyhedral sets. Suppose that for any $n+1$ sets, $P_{i_1}, \ldots, P_{i_{n+1}}$, there is a translate of $C$, i.e., the set $C + u$ for some $u \in \mathbf{R}^n$, which is contained in all $P_{i_1}, \ldots, P_{i_{n+1}}$. Prove that there is a translate of $C$, which is contained in all of the sets $P_1, \ldots, P_m$.

**Exercise** I.26 ▲ A cake contains 300 g of raisins (you may think of every one of them as of a 3D ball of positive radius). John and Jill are about to divide the cake according to the following rules:

- first, Jill chooses a point $a$ in the cake;
- second, John makes a *cut* through $a$, that is, chooses a 2D plane $\Pi$ passing through $a$ and takes the part of the cake on one side of the plane (both $\Pi$ and the side are up to John, with the only restriction that the plane should pass through $a$); all the rest goes to Jill.

1. Prove that it may happen that Jill cannot guarantee herself 76 g of the raisins.
2. Prove that Jill always can choose $a$ in a way which guarantees her at least 74 g of the raisins.
3. Consider $n$-dimensional version of the problem, where the raisins are $n$-dimensional balls, the cake is a domain in $\mathbf{R}^n$, and "a cut" taken by John is defined as the part of the cake contained in the half-space

$$\left\{x \in \mathbf{R}^n : \; e^\top (x-a) \geq 0\right\},$$

   where $e \neq 0$ is the vector ("inner normal to the cutting hyperplane") chosen by John. Prove that for every $\epsilon > 0$, Jill can guarantee to herself at least $\frac{300}{n+1} - \epsilon$ g of raisins, but in general cannot guarantee to herself $\frac{300}{n+1} + \epsilon$ g.

   **Remarks:**

1. With some minor effort, you can prove that Jill can find a point which guarantees her $\frac{300}{n+1}$ g of raisins, and not $\frac{300}{n+1} - \epsilon$ g.

2. If, instead of dividing raisins, John and Jill would divide in the same fashion *uniform and convex* cake (that is, a closed and bounded convex body $X$ with a nonempty interior in $\mathbf{R}^n$, the reward being the $n$-dimensional volume of the part a person gets), the results would change dramatically: choosing as the point the center of masses of the cake
$$\bar{x} := \frac{\int\limits_X x dx}{\int\limits_X dx},$$
Jill would guarantee herself at least $\left(\frac{n}{n+1}\right)^n \approx \frac{1}{e}$ part of the cake. This is a not so easy corollary of the following extremely important and deep result:

> **Brunn-Minkowski Symmetrization Theorem:** *Let $X$ be as above, and let $[a, b]$ be the projection of $X$ on an axis $\ell$, say, on the last coordinate axis. Consider the " symmetrization" $Y$ of $X$, i.e., $Y$ is the set with the same projection $[a, b]$ on $\ell$ and for every hyperplane orthogonal to the axis $\ell$ and crossing $[a, b]$, the intersection of $Y$ with this hyperplane is an $(n-1)$-dimensional ball centered at the axis with precisely the same $(n-1)$-dimensional volume as the one of the intersection of $X$with the same hyperplane:*
> $$\begin{array}{l}\left\{z \in \mathbf{R}^{n-1} : [z; c] \in Y\right\} = \left\{z \in \mathbf{R}^{n-1} : \|z\|_2 \leq \rho(c)\right\}, \quad \forall c \in [a, b], \text{ and} \\ \mathrm{Vol}_{n-1}\left(\left\{z \in \mathbf{R}^{n-1} : [z; c] \in Y\right\}\right) = \mathrm{Vol}_{n-1}\left(\left\{z \in \mathbf{R}^{n-1} : [z; c] \in X\right\}\right), \quad \forall c \in [a, b].\end{array}$$
> *Then, $Y$ is a closed convex set.*

## 5.6 Around Polyhedral Representations

**Exercise** I.27 ▲ Justify calculus rules for polyhedral representations presented in section 3.3.

**Exercise** I.28 Given two sets $U, V \subseteq \mathbf{R}^m$, we define
$$U + V = \{x \in \mathbf{R}^m : \exists u \in U, \exists v \in V : x = u + v\}.$$
Let $D := \{x \in \mathbf{R}^n : Ax + b + Q_s \subseteq P, \forall s \in S\}$ where the set $\varnothing \neq P \subset \mathbf{R}^m$ admits polyhedral representation, the set $\varnothing \neq S \subset \mathbf{R}^k$ is given but arbitrary, and the sets $\varnothing \neq Q_s \subset \mathbf{R}^m$ are indexed by $s \in S$.

1. Suppose that $S$ is a finite set and for each $s \in S$ we have $Q_s = \{q_s\}$, i.e., is a single point. Then, will the set $D$ be polyhedrally representable?
2. State sufficient conditions on the structure of sets $Q_s$ and $S$ that will guarantee that the resulting set $D$ is polyhedral. Here, the goal is to have conditions as general as possible. Among your sufficient conditions, can you identify at least some of those that are necessary?

**Exercise** I.29 ♦ For $x \in \mathbf{R}^n$ and integer $k$, $1 \leq k \leq n$, let $s_k(x)$ be the sum of $k$ largest entries in $x$. For example, $s_1(x) = \max_i\{x_i\}$, $s_n(x) = \sum_{i=1}^n x_i$, $s_3([3; 1; 2; 2]) = 3 + 2 + 2 = 7$. Now let $1 \leq k \leq n$ be two integers. For any integer $k = 1, \ldots, n$, define
$$X_{k,n} := \{[x; t] \in \mathbf{R}^n \times \mathbf{R} : s_k(x) \leq t\}.$$
Observe that $X_{k,n}$ is a polyhedral set. Indeed, $s_k(x) \leq t$ holds if and only if for every $k$ indices $i_1 < i_2 < \ldots < i_k$ from $\{1, 2, \ldots, n\}$ we have $x_{i_1} + x_{i_2} + \ldots + x_{i_k} \leq t$, which is nothing but a linear inequality in variables $x, t$. Since there are $\binom{n}{k}$ possible ways of selecting $k$ indices from $\{1, 2, \ldots, n\}$, the number of linear inequalities describing $X_{k,n}$ is $\binom{n}{k}$, and these linear inequalities give the polyhedral description of $X_{k,n}$. The point of this exercise is to demonstrate that $X_{k,n}$ admits a "short" polyhedral representation, specifically,
$$X_{k,n} = \left\{[x; t] \in \mathbf{R}^n \times \mathbf{R} : \exists z \in \mathbf{R}^n, \exists s \in \mathbf{R} : x_i \leq z_i + s, \forall i, \ z \geq 0, \ \sum_{i=1}^n z_i + ks \leq t\right\}.$$

**Exercise** I.30 ▲ [Computational study: Fourier-Motzkin elimination as an LP algorithm] It was mentioned in section 3.2.1 that Fourier-Motzkin elimination provides us with an algorithm that terminates in finitely many steps for solving LP problems. This algorithm, however, is of no computational value due to the potential rapid growth of the number of inequalities one may need to handle when eliminating more and more variables. The goal of this exercise is to get an impression of this phenomenon.

Our "guinea pig" will be transportation problem with $n$ unit capacity suppliers and $n$ unit demand customers:

$$\min_{x,t}\left\{t:\ t\geq\sum_{i=1}^{n}\sum_{i=1}^{n}c_{ij}x_{ij},\ \sum_{i}x_{ij}\geq 1,\ \forall j,\ \sum_{j}x_{ij}\leq 1,\ \forall i,\ x_{ij}\geq 0,\forall i,j\right\}.$$

This problem has $n^2+1$ variables and $(n+1)^2$ linear inequality constraints, and let us solve it by applying the Fourier-Motzkin elimination to project the feasible set of the problem onto the axis of the $t$-variable, that is, to build a finite system $\mathcal{S}$ of univariate linear inequalities specifying this projection.

How many inequalities do you think there will be in $\mathcal{S}$ when $n=1,2,3,4$? Check your intuition by implementing and running the F-M elimination, assuming, for the sake of definiteness, that $c_{ij}=1$ for all $i,j$.

## 5.7 Around General Theorem on Alternative

**Exercise** I.31 1. Prove Gordan's Theorem on Alternative:

> A system of strict homogeneous linear inequalities $Ax<0$ in variables $x$ has a solution if and only if the system $A^\top\lambda=0,\ \lambda\geq 0$ in variables $\lambda$ has only the trivial solution $\lambda=0$.

2. Prove Motzkin's Theorem on Alternative:

> A system $Ax<0$, $Bx\leq 0$ of strict and nonstrict homogeneous linear inequalities has a solution if and only if the system $A^\top\lambda+B^\top\mu=0,\ \lambda\geq 0,\ \mu\geq 0$ in variables $\lambda,\mu$ has no solution with $\lambda\neq 0$.

**Exercise** I.32 For the systems of constraints to follow, write them down equivalently in the standard form $Ax<b, Cx\leq d$ and point out their feasibility status ("feasible – infeasible") along with the corresponding certificates (certificate for feasibility is a feasible solution to the system; certificate for infeasibility is a collection of weights of constraints which leads to a contradictory consequence inequality, as explained in GTA).

1. $x\leq 0$ $(x\in\mathbf{R}^n)$
2. $x\leq 0$, and $\sum_{i=1}^n x_i>0$ $(x\in\mathbf{R}^n)$
3. $-1\leq x_i\leq 1,\ 1\leq i\leq n,\ \sum_{i=1}^n x_i\geq n$ $(x\in\mathbf{R}^n)$
4. $-1\leq x_i\leq 1,\ 1\leq i\leq n,\ \sum_{i=1}^n x_i>n$ $(x\in\mathbf{R}^n)$
5. $-1\leq x_i\leq 1,\ 1\leq i\leq n,\ \sum_{i=1}^n ix_i\geq\frac{n(n+1)}{2}$ $(x\in\mathbf{R}^n)$
6. $-1\leq x_i\leq 1,\ 1\leq i\leq n,\ \sum_{i=1}^n ix_i>\frac{n(n+1)}{2}$ $(x\in\mathbf{R}^n)$
7. $x\in\mathbf{R}^2$, $|x_1|+x_2\leq 1,\ x_2\geq 0,\ x_1+x_2=1$
8. $x\in\mathbf{R}^2$, $|x_1|+x_2\leq 1,\ x_2\geq 0,\ x_1+x_2>1$
9. $x\in\mathbf{R}^4$, $x\geq 0$, the sum of two largest entries in $x$ does not exceed 2, and $x_1+x_2+x_3\geq 3$
10. $x\in\mathbf{R}^4$, $x\geq 0$, the sum of two largest entries in $x$ does not exceed 2, and $x_1+x_2+x_3>3$

**Exercise** I.33 Let $(\mathcal{S})$ be the following system of linear inequalities in variables $x\in\mathbf{R}^3$

$$x_1\leq 1,\ x_1+x_2\leq 1,\ x_1+x_2+x_3\leq 1 \qquad (\mathcal{S})$$

In the following list, point out which inequalities are/are not consequences of this system, and certify your claims. To certify that a given inequality is a consequence of the given system, you need to provide nonnegative aggregation weights $\lambda \in \mathbf{R}^3_+$ for the inequalities in $(\mathcal{S})$ such that the resulting consequence inequality implies the given inequality. To certify that a given inequality is not a consequence of the given system $(\mathcal{S})$, you need to find a point $x \in \mathbf{R}^3$ that satisfies the given system but violates the given inequality.

1. $3x_1 + 2x_2 + x_3 \leq 4$
2. $3x_1 + 2x_2 + x_3 \leq 2$
3. $3x_1 + 2x_2 \leq 3$
4. $3x_1 + 2x_2 \leq 2$
5. $3x_1 + 3x_2 + x_3 \leq 3$
6. $3x_1 + 3x_2 + x_3 \leq 2$

Make a generalization: prove that a linear inequality $px_1 + qx_2 + rx_3 \leq s$ is a consequence of $(\mathcal{S})$ if and only if $s \geq p \geq q \geq r \geq 0$.

**Exercise** I.34 Is the inequality $x_1 + x_2 \leq 1$ a consequence of the system $x_1 \leq 1,\ x_1 \geq 2$? If yes, can it be obtained by taking a legitimate weighted sum of inequalities from the system and the always true inequality $0^\top x \leq 1$, as it is suggested by the Inhomogeneous Farkas Lemma?

**Exercise** I.35 Certify the correct statements in the following list:

1. The polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 1\right\}$ is nonempty.
2. The polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 0.99\right\}$ is empty.
3. The linear inequality $x_1 + x_2 + x_3 \geq 2$ is violated somewhere on the polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 1\right\}$.
4. The linear inequality $x_1 + x_2 + x_3 \geq 2$ is violated somewhere on the polyhedral set $X = \left\{x \in \mathbf{R}^3 : x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 0.99\right\}$.
5. The linear inequality $x_1 + x_2 \leq 3/4$ is satisfied everywhere on the polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 1.05\right\}$.
6. The polyhedral set $Y = \left\{x \in \mathbf{R}^3 :\ x_1 \geq 1/3,\ x_2 \geq 1/3,\ x_3 \geq 1/3\right\}$ is not contained in the polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 1\right\}$.
7. The polyhedral set $Y = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 1\right\}$ is contained in the polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x_1 + x_2 \leq 2/3,\ x_2 + x_3 \leq 2/3,\ x_1 + x_3 \leq 2/3\right\}$.

## 5.8 Around Linear Programming Duality

**Exercise** I.36 ♦ Let the polyhedral set $P = \{x \in \mathbf{R}^n : Ax \leq b\}$, where $A = [a_1^\top; \ldots; a_m^\top]$, be nonempty. Prove that $P$ is bounded if and only if every vector from $\mathbf{R}^n$ can be represented as a linear combination of the vectors $a_i$ with nonnegative coefficients where at most $n$ coefficients are positive. As a result, given $A$, all nonempty sets of the form $\{x \in \mathbf{R}^n : Ax \leq b\}$ simultaneously are/are not bounded.

**Exercise** I.37 Consider the linear program

$$\mathrm{Opt} = \max_{x \in \mathbf{R}^2} \{x_1 :\ x_1 \geq 0,\ x_2 \geq 0,\ ax_1 + bx_2 \leq c\} \tag{$P$}$$

where $a, b, c$ are parameters. Answer the following questions:

1. Let $c = 1$. Is the problem feasible?
2. Let $a = b = 1,\ c = -1$. Is the problem feasible?
3. Let $a = b = 1,\ c = -1$. Is the problem bounded[4]?
4. Let $a = b = c = 1$. Is the problem bounded?
5. Let $a = 1,\ b = -1,\ c = 1$. Is the problem bounded?
6. Let $a = b = c = 1$. Is it true that $\mathrm{Opt} \geq 0.5$?

[4] Recall that a maximization problem is called *bounded*, if the objective is bounded from above on the feasible set, which is the same as its optimal value being $< \infty$

7. Let $a = b = 1,\ c = -1$. Is it true that $\mathrm{Opt} \le 1$?
8. Let $a = b = c = 1$. Is it true that $\mathrm{Opt} \le 1$?
9. Let $a = b = c = 1$. Is it true that $x_* = [1; 1]$ is an optimal solution of $(P)$?
10. Let $a = b = c = 1$. Is it true that $x_* = [1/2; 1/2]$ is an optimal solution of $(P)$?
11. Let $a = b = c = 1$. Is it true that $x_* = [1; 0]$ is an optimal solution of $(P)$?

**Exercise** I.38 Consider the LP program

$$\max_{x_1,x_2}\left\{-x_2 :\ \begin{array}{c} x_1 \le 0 \\ -x_1 \le -1 \\ x_2 \le 1 \end{array}\right\}$$

Write down the dual problem and check whether the optimal values are equal to each other.

**Exercise** I.39 Write down the problems dual to the following linear programs:

1. $\max\limits_{x\in\mathbf{R}^3}\left\{x_1 + 2x_2 + 3x_3 :\ \begin{array}{l} x_1 - x_2 + x_3 = 0, \\ x_1 + x_2 - x_3 \ge 100, \\ x_1 \le 0, \\ x_2 \ge 0, \\ x_3 \ge 0 \end{array}\right\}$
2. $\max\limits_{x\in\mathbf{R}^n}\left\{c^\top x :\ Ax = b,\ x \ge 0\right\}$
3. $\max\limits_{x\in\mathbf{R}^n}\left\{c^\top x :\ Ax = b,\ \underline{u} \le x \le \overline{u}\right\}$
4. $\max\limits_{x,y}\left\{c^\top x :\ Ax + By \le b,\ x \le 0,\ y \ge 0\right\}$

**Exercise** I.40 ▲ Consider a primal-dual pair of LO programs

$$\mathrm{Opt}(P) = \min_x\left\{c^\top x :\ Ax \ge b\right\} \qquad (P)$$
$$\mathrm{Opt}(D) = \max_y\left\{b^\top y :\ y \ge 0,\ A^\top y = c\right\} \qquad (D)$$

Prove that the feasible set of at least one of these problems is unbounded.

**Exercise** I.41 ▲ Consider the following linear program

$$\mathrm{Opt} = \min_{\{x_{ij}\}_{1\le i<j\le 4}}\left\{2\sum_{1\le i<j\le 4} x_{ij} :\ x_{ij} \ge 0,\ \forall 1 \le i < j \le 4,\ \sum_{j>i} x_{ij} + \sum_{j<i} x_{ji} \ge i,\ 1 \le i \le 4\right\}.$$

1. Show that the optimum objective value is at most 20.
2. Show that the optimum objective value is at least 10. $\mathrm{Opt} \ge 10$.

**Exercise** I.42 ♦ We say that an $n \times n$ matrix $P$ is *stochastic* if all of its entries are nonnegative and the sum of the entries of each row is equal to 1. Show that if $P$ is a stochastic matrix, then there is a nonzero vector $a \in \mathbf{R}^n$ such that $a^\top P = a^\top$ and $a \ge 0$.

**Exercise** I.43 ▲ Let $A \in \mathbf{R}^{n\times n}$ be a symmetric matrix. Consider the linear optimization problem

$$\min_x\left\{c^\top x :\ Ax \ge c,\ x \ge 0\right\}.$$

Prove that if $\bar{x}$ satisfies $A\bar{x} = c$ and $\bar{x} \ge 0$, then $\bar{x}$ is optimal.

**Exercise** I.44 ▲ Let $w \in \mathbf{R}^n$, and let $A \in \mathbf{R}^{n\times n}$ be a *skew-symmetric* matrix, i.e., $A^\top = -A$. Consider the following linear program

$$\mathrm{Opt}(P) = \min_{x\in\mathbf{R}^n}\left\{w^\top x :\ Ax \ge -w,\ x \ge 0\right\}.$$

Suppose that the problem is solvable. Provide a closed analytical form expression for $\mathrm{Opt}(P)$.

**Exercise** I.45 ▲ [Separation Theorem, polyhedral version] Let $P$ and $Q$ be two nonempty polyhedral sets in $\mathbf{R}^n$ such that $P \cap Q = \varnothing$. Suppose that the polyhedral descriptions of these sets are given as

$$P := \{x \in \mathbf{R}^n : \ Ax \leq b\} \text{ and } Q := \{x \in \mathbf{R}^n : \ Dx \geq d\}.$$

Using LP duality show that there exists a vector $c \in \mathbf{R}^n$ such that

$$c^\top x < c^\top y \text{ for all } x \in P \text{ and } y \in Q.$$

**Exercise** I.46 ▲ Suppose we are given the following linear program

$$\min_x \left\{ c^\top x : \ Ax = b, \ x \geq 0 \right\} \tag{$P$}$$

and its associated *Lagrangian* function given by

$$L(x, \lambda) := c^\top x + \lambda^\top (b - Ax).$$

The LP dual to $(P)$ is (replace $Ax = b$ with $Ax \geq b$, $-Ax \geq -b$)

$$\mathrm{Opt}(D) = \max_{\lambda_\pm, \mu} \left\{ b^\top [\lambda_+ - \lambda_-] : A^T [\lambda_+ - \lambda_-] + \mu = c, \lambda_\pm \geq 0, \mu \geq 0 \right\},$$

or, after eliminating $\mu$ and setting $\lambda = \lambda_+ - \lambda_-$,

$$\mathrm{Opt}(D) = \max_\lambda \left\{ b^\top \lambda : A^\top \lambda \leq b \right\}. \tag{$D$}$$

Now, let us consider the following "game": Player 1 chooses some $x \geq 0$, and player 2 chooses some $\lambda$ simultaneously; then, player 1 pays to player 2 the amount $L(x, \lambda)$. In this game, player 1 would like to minimize $L(x, \lambda)$ and player 2 would like to maximize $L(x, \lambda)$.

A pair $(x^*, \lambda^*)$ with $x^* \geq 0$, is called an *equilibrium* point (or *saddle point* or *Nash equilibrium* ) if

$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*), \quad \forall x \geq 0 \ \text{ and } \ \forall \lambda. \tag{$*$}$$

(That is, we have an equilibrium if no player is able to improve her performance by unilaterally modifying her choice.)

Show that a pair $(x^*, \lambda^*)$ is an equilibrium point if and only if $x^*$ and $\lambda^*$ are respectively optimal solutions to the problem $(P)$ and its dual respectively.

**Exercise** I.47 ▲ Given a polyhedral set $X = \left\{ x \in \mathbf{R}^n : \ a_i^\top x \leq b_i, \ \forall i = 1, \ldots, m \right\}$, consider the associated optimization problem

$$\max_{x,t} \{ t : \ B_1(x, t) \subseteq X \},$$

where $B_1(x, t) := \{ y \in \mathbf{R}^n : \ \|y - x\|_\infty \leq t \}$. Is it possible to pose this optimization problem as a linear program with polynomial in $m, n$ number of variables and constraints? If it is possible, give such a representation explicitly. If not, argue why.

**Exercise** I.48 ▲ Consider the following optimization problem

$$\min_{x \in \mathbf{R}^n} \left\{ c^\top x : \ \tilde{a}_i^\top x \leq b_i \text{ for some } \tilde{a}_i \in A_i, \ i = 1, \ldots, m, \ x \geq 0 \right\}, \tag{*}$$

where $A_i = \{ \bar{a}_i + \epsilon_i : \ \|\epsilon_i\|_\infty \leq \rho \}$ for $i = 1, \ldots, m$ and $\|u\|_\infty := \max_{j=1,\ldots,n} \{|u_j|\}$. In this problem, we basically mean that the constraint coefficient $\tilde{a}_{ij}$ ($j$-th component of the $i$-th constraint vector $\tilde{a}_i$) belongs to the interval uncertainty set $[\bar{a}_{ij} - \rho, \ \bar{a}_{ij} + \rho]$, where $\bar{a}_{ij}$ is its nominal value. That is, in $(*)$, we are seeking a solution $x$ such that each constraint is satisfied for *some* coefficient vector from the corresponding uncertainty set.

Note that in its current form $(*)$, this problem is not a linear program (LP). Prove that it can be written as an *explicit* linear program and give the corresponding LP formulation.

**Exercise** I.49 ⧫ Let $S = \{a_1, a_2, \ldots, a_n\}$ be a finite set composed of $n$ distinct from each other elements, and let $f$ be a real-valued function defined on the set of all subsets of $S$. We say that $f$ is *submodular* if for every $X, Y \subseteq S$, the following inequality holds

$$f(X) + f(Y) \geq f(X \cup Y) + f(X \cap Y).$$

1. Give an example of a submodular function $f$.
2. Let $f : 2^S \to \mathbf{Z}$ be an integer-valued submodular function such that $f(\varnothing) = 0$. Consider the polyhedron
$$P_f := \left\{ x \in \mathbf{R}^{|S|} : \sum_{t \in T} x_t \leq f(T), \ \forall T \subseteq S \right\},$$
Consider
$$\bar{x}_{a_k} := f(\{a_1, \ldots, a_k\}) - f(\{a_1, \ldots, a_{k-1}\}), \quad k = 1, \ldots, n.$$
Show that $\bar{x}$ is feasible to $P_f$.
3. Consider the following optimization problem associated with $P_f$
$$\max_x \left\{ c^\top x : \ x \in P_f \right\}.$$
Write down the dual of this LP.
4. Assume without loss of generality that $c_{a_1} \geq c_{a_2} \geq \ldots \geq c_{a_n}$. Identify a dual feasible solution and using LP Duality Theorem show that the solution $\bar{x}$ specified in part 2 is optimal to the primal maximization problem associated with $P_f$.

**Remark:** Note that when the submodular function $f$ is integer-valued, we immediately see from the characterization of the optimal primal solution $\bar{x}$ that for all integer vectors $c \in \mathbf{Z}^n$ such that there exists an optimum solution to the primal problem, there exists an optimum solution (e.g. $\bar{x}$) where all variables take integer values. A system of linear inequalities $Ax \leq b$ with $b \in \mathbf{Z}^m$ and $A \in \mathbf{Q}^{m \times n}$ satisfying such a property (i.e., whenever $c \in \mathbf{Z}^n$ is such that there is an optimal solution to $\max_x\{c^\top x : Ax \leq b\}$ then there is an integer optimum solution) is called *totally dual integral* (TDI). Thus, we conclude that the polyhedron $P_f$ associated with an integer-valued submodular function $f$ is TDI. TDI property is a well-known sufficient condition that guarantees that every extreme point (see section 8.2) of the associated polyhedron is integral. In particular, TDI property generalizes *total unimodularity* (TU), i.e., the other well-known sufficient condition for integrality of a polyhedron which plays a key role in network-flow based optimization.

# 6

# Proofs of Facts from Part I

**Fact I.1.6** The *unit ball of a norm* $\|\cdot\|$, i.e., the set

$$\{x \in \mathbf{R}^n : \ \|x\| \leq 1\},$$

same as every other $\|\cdot\|$-ball

$$B_r(a) := \{x \in \mathbf{R}^n : \|x-a\| \leq r\},$$

(here $a \in \mathbf{R}^n$ and $r \geq 0$ are fixed) is convex.

In particular, Euclidean balls ($\|\cdot\|$-balls associated with the standard Euclidean norm $\|x\|_2 := \sqrt{x^\top x}$) are convex.

Proof. Let us prove that the set $Q := \{x \in \mathbf{R}^n : \ \|x-a\| \leq r\}$ is convex. For any $x', x'' \in Q$ and $\lambda \in [0,1]$, we have

$$\begin{aligned} \|\lambda x' + (1-\lambda)x'' - a\| &= \|\lambda(x'-a) + (1-\lambda)(x''-a)\| \\ &\leq \|\lambda(x'-a)\| + \|(1-\lambda)(x''-a)\| \\ &= \lambda\|x'-a\| + (1-\lambda)\|x''-a\| \leq \lambda r + (1-\lambda)r = r. \end{aligned}$$

Here, the first inequality follows from Triangle inequality, and the second equality follows from homogeneity of norms, and the last inequality is due to $x', x'' \in Q$. Thus, from $\|\lambda x' + (1-\lambda)x'' - a\| \leq r$, we conclude that $\lambda x' + (1-\lambda)x'' \in Q$ as desired. □

**Fact I.1.8** Unit balls of norms on $\mathbf{R}^n$ are exactly the same as convex sets $V$ in $\mathbf{R}^n$ satisfying the following three properties:

(i) $V$ is symmetric with respect to the origin: $x \in V \implies -x \in V$;

(ii) $V$ is bounded and closed;

(iii) $V$ contains a neighborhood of the origin, i.e., there exists $r > 0$ such that the centered at the origin Euclidean ball of radius $r$ – the set $\{x \in \mathbf{R}^n : \|x\|_2 \leq r\}$ – is contained in $V$.

Any set $V$ satisfying the outlined properties is indeed the unit ball of a particular norm given by

$$\|x\|_V = \inf\left\{t \geq 0 : t^{-1}x \in V\right\}. \tag{1.2}$$

Proof. First, let $V$ be the unit ball of a norm $\|\cdot\|$, and let us verify the three stated properties. Note that $V = -V$ due to $\|x\| = \|-x\|$. $V$ is bounded and contains a neighborhood of the origin due to equivalence between $\|\cdot\|$ and $\|\cdot\|_2$ (Proposition B.3). Moreover, $V$ is closed. To see this note that $\|\cdot\|$ is Lipschitz continuous with constant 1 with respect to itself since by Triangle inequality and due to $\|x-y\| = \|y-x\|$ we have

$$|\|x\| - \|y\|| \le \|x-y\|, \quad \forall x, y \in \mathbf{R}^n,$$

which implies  by Proposition B.3 there exists $L_{\|\cdot\|} < \infty$ such that

$$|\|x\| - \|y\|| \le L_{\|\cdot\|}\|x-y\|_2, \quad \forall x, y \in \mathbf{R}^n,$$

that is, $\|\cdot\|$ is Lipschitz continuous (and thus continuous). And of course for any $a \in \mathbf{R}$, the sublevel set $\{x \in \mathbf{R}^n : f(x) \le a\}$ of a continuous function is closed.

For the reverse direction, consider any $V$ possessing properties (i –iii). Then, as $V$ is bounded and contains a neighborhood of the origin, the function $\|\cdot\|_V$ is well defined, is is positive outside of the origin and vanishes at the origin. Moreover, $\|\cdot\|_V$ is homogeneous – when the argument is multiplied by a real number $\lambda$, the value of the function is multiplied by $|\lambda|$ (by construction and since $V = -V$).

Now, let us show that the relation $V = \{y \in \mathbf{R}^n : \|y\|_V \le 1\}$ holds. Indeed, the inclusion $V \subseteq \{y : \|y\|_V \le 1\}$ is evident. So, we will verify that $\|y\|_V \le 1$ implies $y \in V$. Consider any $y$ such that $\|y\|_V \le 1$ and let $\bar{t} := \|y\|_V$ (note that $\bar{t} \in [0,1]$). There is nothing to prove when $\bar{t} = 0$, which due to the boundedness of $V$ implies that $y = 0$ and $V$ contains the origin. When $\bar{t} > 0$, then, by definition of $\|\cdot\|_V$, there exists a sequence of positive numbers $\{t_i\}$ that converges to $\bar{t}$ as $i \to \infty$ such that $y^i := t_i^{-1}y \in V$. Then, as $V$ is closed, $\bar{y} := \bar{t}^{-1}y \in V$. And since $0 < \bar{t} \le 1$, $y = \bar{t}\bar{y}$ is a convex combination of the origin and $\bar{y}$. As both $0 \in V$ and $\bar{y} \in V$ and $V$ is convex, we conclude $y \in V$.

Let us now check that $\|\cdot\|_V$ satisfies the Triangle inequality. As $\|\cdot\|_V$ is nonnegative,  all we have to check is that $\|x+y\|_V \le \|x\|_V + \|y\|_V$ when $x \neq 0$, $y \neq 0$. Setting $\bar{x} := x/\|x\|_V$, $\bar{y} := y/\|y\|_V$, we have by homogeneity $\|\bar{x}\|_V = \|\bar{y}\|_V = 1$. Then, from the relation $V = \{y \in \mathbf{R}^n : \|y\|_V \le 1\}$ we deduce $\bar{x} \in V$ and $\bar{y} \in V$. Now, as $V$ is convex and $\bar{x}, \bar{y} \in V$, we have

$$\frac{1}{\|x\|_V + \|y\|_V}(x+y) = \frac{\|x\|_V}{\|x\|_V + \|y\|_V}\bar{x} + \frac{\|y\|_V}{\|x\|_V + \|y\|_V}\bar{y} \in V.$$

That is, $\left\|\frac{1}{\|x\|_V + \|y\|_V}(x+y)\right\|_V \le 1$. Then, once again by homogeneity of $\|\cdot\|_V$ we conclude that $\|x+y\|_V \le \|x\|_V + \|y\|_V$. □

**Fact I.1.9** Let $Q$ be an $n \times n$ matrix which is symmetric (i.e., $Q = Q^\top$) and positive definite (i.e., $x^\top Q x > 0$ for all $x \neq 0$). Then, for every nonnegative $r$, the *$Q$-ellipsoid of radius $r$ centered at $a$*, i.e., the set

$$\left\{x \in \mathbf{R}^n : (x-a)^\top Q(x-a) \le r^2\right\}$$

is convex.

Proof. Note that since $Q$ is positive definite, the matrix $P := Q^{1/2}$ (see section 7n.symm.G for the definition of the matrix square root) is well-defined and positive definite. Then, we have $P$ is nonsingular and symmetric, and

$$\begin{aligned}\left\{x \in \mathbf{R}^n : (x-a)^\top Q(x-a) \le r^2\right\} &= \left\{x \in \mathbf{R}^n : (x-a)^\top P^\top P(x-a) \le r^2\right\} \\ &= \left\{x \in \mathbf{R}^n : \|P(x-a)\|_2 \le r\right\}.\end{aligned}$$

Now, note that whenever $\|\cdot\|$ is a norm on $\mathbf{R}^n$ and $P$ is a nonsingular $n \times n$ matrix, the function $x \mapsto \|Px\|$ is a norm itself (why?). Thus, the function $\|x\|_Q := \sqrt{x^\top Q x} = \|Q^{1/2}x\|_2$ is a norm, and the ellipsoid in question clearly is just the $\|\cdot\|_Q$-ball of radius $r$ centered at $a$. □

**Fact I.1.11** A set $M \subseteq \mathbf{R}^n$ is convex if and only if it is closed with respect to taking all convex combinations of its elements. That is, $M$ is convex if and only

if every convex combination of vectors from $M$ is again a vector from $M$.
*Hint:* Note that assuming $\lambda_1, \ldots, \lambda_m > 0$, one has

$$\sum_{i=1}^m \lambda_i x^i = \lambda_1 x^1 + (\lambda_2 + \lambda_3 + \ldots + \lambda_m) \sum_{i=2}^m \mu_i x^i, \quad \text{where } \mu_i := \frac{\lambda_i}{\lambda_2 + \lambda_3 + \ldots + \lambda_m}.$$

Proof. There is nothing to prove when $M$ is empty, so we assume $M \neq \varnothing$. If $M$ is closed with respect to taking arbitrary convex combinations of its points, it is closed with respect to taking 2-point combinations, which is exactly the same as to say that $M$ is convex. For the reverse direction, let $M$ be convex. We will prove that a point given as a convex combination of $N$ points from $M$ is itself in $M$ by induction on $N$. The claim is clearly true when $N = 1$ (independent of what $M$ is) and is true when $N = 2$ (since $M$ is convex). Suppose now that the claim is true for some $N \geq 2$. Consider $(N+1)$-term convex combination $x = \sum_{i=1}^{N+1} \lambda_i x^i$ of points $x^i$ from $M$. If $\lambda_1 = 1$, we have $x = x^1 \in M$. When $\lambda_1 < 1$, we have

$$x = \lambda_1 x^1 + (1 - \lambda_1) \left( \sum_{i=2}^N \frac{\lambda_i}{1 - \lambda_1} x^i \right).$$

Define $\bar{x} := \sum_{i=2}^N \frac{\lambda_i}{1-\lambda_1} x^i$. As $\sum_{i=2}^N \lambda_i = 1 - \lambda_1$, we see that $\bar{x}$ is an $N$-term convex combination of points from $M$ and thus belongs to $M$ by the inductive hypothesis. Hence, $x = \lambda_1 x^1 + (1-\lambda_1)\bar{x}$ is convex combination of $x^1, \bar{x} \in M$, and as $M$ is convex we conclude $x \in M$. This completes the inductive step. □

**Fact I.1.14** [Convex hull via convex combinations] For a set $M \subseteq \mathbf{R}^n$,

$$\mathrm{Conv}(M) = \{\text{the set of all convex combinations of vectors from } M\}.$$

Proof. Define $\widehat{M} := \{$the set of all convex combinations of vectors from $M\}$. Recall that a convex set is closed with respect to taking convex combinations of its members (Fact I.1.11); thus any convex set containing $M$ also contains $\widehat{M}$. As by definition $\mathrm{Conv}(M)$ is the intersection of all convex sets containing $M$, we have $\mathrm{Conv}(M) \supseteq \widehat{M}$. It remains to prove that $\mathrm{Conv}(M) \subseteq \widehat{M}$. We start with the claim that $\widehat{M}$ is convex. By Fact I.1.11 $\widehat{M}$ is convex if and only if every convex combination of points from $\widehat{M}$ is also in $\widehat{M}$. Indeed this criteria holds for $\widehat{M}$: let $\bar{x}^i \in \widehat{M}$ for $i = 1, \ldots, N$ and consider a convex combination of these points, i.e.,

$$\hat{x} := \sum_{i=1}^N \lambda_i \bar{x}^i,$$

where $\lambda_i \geq 0$ and $\sum_{i=1}^N \lambda_i = 1$. For each $i = 1, \ldots, N$, as $\bar{x}^i \in \widehat{M}$, by definition of $\widehat{M}$, we have $\bar{x}^i = \sum_{j=1}^{N_i} \mu_{i,j} x^{i,j}$, where $x^{i,j} \in M$, $\mu_{i,j} \geq 0$ and $\sum_{j=1}^{N_i} \mu_{i,j} = 1$. Then, we arrive at

$$\hat{x} := \sum_{i=1}^N \lambda_i \bar{x}^i = \sum_{i=1}^N \lambda_i \left( \sum_{j=1}^{N_i} \mu_{i,j} x^{i,j} \right) = \sum_{i=1}^N \sum_{j=1}^{N_i} (\lambda_i \mu_{i,j}) x^{i,j}.$$

Clearly, $\gamma_{i,j} := \lambda_i \mu_{i,j}$ is nonnegative for all $i, j$. Moreover,

$$\sum_{i=1}^N \sum_{j=1}^{N_i} \gamma_{i,j} = \sum_{i=1}^N \sum_{j=1}^{N_i} \lambda_i \mu_{i,j} = \sum_{i=1}^N \lambda_i \left( \sum_{j=1}^{N_i} \mu_{i,j} \right) = \sum_{i=1}^N \lambda_i = 1,$$

where the third and forth equalities follow from $\sum_{j=1}^{N_i} \mu_{i,j} = 1$ for all $i$ and $\sum_{i=1}^N \lambda_i = 1$, respectively. Therefore, $\hat{x} = \sum_{i=1}^N \sum_{j=1}^{N_i} \gamma_{i,j} x^{i,j}$ is nothing but a convex combination of points from $M$, and thus $\hat{x} \in \widehat{M}$ proving that $\widehat{M}$ is convex. Clearly, we also have $\widehat{M} \supseteq M$, and so by definition of $\mathrm{Conv}(M)$, we deduce $\widehat{M} \supseteq \mathrm{Conv}(M)$, as desired. □

**Fact I.1.17** A set $K \subseteq \mathbf{R}^n$ is a cone if and only if it is nonempty and

- is conic, i.e., $x \in K, t \geq 0 \implies tx \in K$; and
- contains sums of its elements, i.e., $x, y \in K \implies x + y \in K$.

Proof. Suppose $K$ is nonempty and possesses the above properties. Then, the first property already states that $K$ is conic, so we will show that $K$ is convex. For any $x, y \in K$ and $\lambda \in [0,1]$, we have

$$\lambda x + (1-\lambda)y = \bar{x} + \bar{y},$$

where $\bar{x} := \lambda x$ and $\bar{y} := (1-\lambda)y$. As $\lambda \in [0,1]$, $x, y \in K$ and $K$ is conic, we have $\bar{x}, \bar{y} \in K$. Moreover, since $K$ contains sum of its elements, we conclude $\lambda x + (1-\lambda)y = (\bar{x} + \bar{y}) \in K$, i.e., $K$ is convex.

For the reverse direction, if $K$ is a cone, then by definition $K$ is nonempty, conic and convex. Then, for any $x, y \in K$, as $K$ is convex we have $\frac{1}{2}x + \frac{1}{2}y \in K$, and as $K$ is conic we arrive at $x + y = 2\left(\frac{1}{2}x + \frac{1}{2}y\right) \in K$. □

**Fact I.1.20** [Conic hull via conic combinations] The conic hull Cone($K$) of a set $K \subseteq \mathbf{R}^n$ is the set of all conic combinations (i.e., linear combinations with nonnegative coefficients) of vectors from $K$:

$$\mathrm{Cone}(K) = \left\{ x \in \mathbf{R}^n : \exists N \geq 0, \lambda_i \geq 0, x^i \in K, i \leq N : x = \sum_{i=1}^{N} \lambda_i x^i \right\}.$$

Proof. The case of $K = \varnothing$ is trivial, see the comment on the value of an empty sum after the statement of Fact I.1.20. When $K \neq \varnothing$, this fact is an immediate corollary of Fact I.1.17. □

**Fact I.1.23** The closure of a set $M \subseteq \mathbf{R}^n$ is exactly the set composed of the limits of all converging sequences of elements from $M$.

Proof. Let $\overline{M}$ be the set of the limit points of all converging sequences of elements from $M$. We need to show that $\mathrm{cl}\, M = \overline{M}$. Let us first prove $\mathrm{cl}\, M \supseteq \overline{M}$. Suppose $x \in \overline{M}$. Then, $x$ is the limit of a converging sequence of points $\{x^i\} \in M \subseteq \mathrm{cl}\, M$. Since $\mathrm{cl}\, M$ is a closed set, we arrive at $x \in \mathrm{cl}\, M$.

For the reverse direction, note that by definition $\mathrm{cl}\, M$ is the smallest (w.r.t. inclusion) closed set that contains $M$, so it suffices to prove that $\overline{M}$ is a closed set satisfying $\overline{M} \supseteq M$. It is easy to see that $\overline{M} \supseteq M$ holds as for any $x \in M$, the sequence $\{x^i\}$ where $x^i = x$ is a converging sequence of points from $M$ with a limit point of $x$ and thus by definition of $\overline{M}$ we deduce $x \in \overline{M}$. Now, consider a converging sequence of points $\{x^i\} \subseteq \overline{M}$, and let us prove that the limit $\bar{x}$ of this sequence belongs to $\overline{M}$. For every $i$, since the point $x^i \in \overline{M}$ is the limit of a sequence of points from $M$, we can find a point $y^i \in M$ such that $\|x^i - y^i\|_2 \leq 1/i$. The sequence $\{y^i\}$ is composed of points from $M$ and clearly has the same limit as the sequence $\{x^i\}$, so that the latter limit is the limit of a sequence of points from $M$ and as such belongs to $\overline{M}$. □

**Fact I.1.36** Let $K$ be a closed cone, and let the set

$$X := \{x \in \mathbf{R}^n : Ax - b \in K\}$$

be nonempty. Then, $\overline{\mathrm{ConeT}}(X) = \{[x;t] \in \mathbf{R}^n \times \mathbf{R} : Ax - bt \in K, \ t \geq 0\}$.

Proof. Define $\widehat{X} := \{[x;t] \in \mathbf{R}^n \times \mathbf{R} : Ax - bt \in K, \ t \geq 0\}$; so, we should prove that $\widehat{X} = \overline{\mathrm{ConeT}}(X)$. Recall that $\overline{\mathrm{ConeT}}(M) := \mathrm{cl}\{[x;t] \in \mathbf{R}^n \times \mathbf{R} : t > 0, \ x/t \in M\}$ for any nonempty convex set $M$. Note that for the given set $X$, its perspective transform is

$$\begin{aligned} \mathrm{Persp}(X) &= \{[x;t] \in \mathbf{R}^n \times \mathbf{R} : t > 0, \ A(x/t) - b \in K\} \\ &= \{[x;t] \in \mathbf{R}^n \times \mathbf{R} : t > 0, \ Ax - bt \in K\}, \end{aligned}$$

where the last equality follows from $K$ being conic. So, $\mathrm{Persp}(X) \subseteq \widehat{X}$, and by taking the closures of both sides we arrive at $\overline{\mathrm{ConeT}}(X) = \mathrm{cl}(\mathrm{Persp}(X)) \subseteq \mathrm{cl}(\widehat{X}) = \widehat{X}$, where the last equality follows as $\widehat{X}$ clearly is closed. Hence, $\overline{\mathrm{ConeT}}(X) \subseteq \widehat{X}$. To verify the opposite inclusion, consider $[x;t] \in \widehat{X}$ and let us prove that $[x;t] \in \overline{\mathrm{ConeT}}(X)$. Let $\bar{x} \in X$ (recall that $X$ is nonempty). Then, $[\bar{x};1] \in \widehat{X}$. Moreover, as $\widehat{X}$ is a cone and the points $[x;t]$ and $[\bar{x};1]$ belong to $\widehat{X}$, we have $z_\epsilon := [x+\epsilon\bar{x}; t+\epsilon] \in \widehat{X}$ for all $\epsilon > 0$. Also, for $\epsilon > 0$, we have $t+\epsilon > 0$ and so $z_\epsilon \in \widehat{X}$ implies $\frac{1}{(t+\epsilon)}(x+\epsilon\bar{x}) \in X$, which is equivalent to $z_\epsilon = [x+\epsilon\bar{x}; t+\epsilon] \in \mathrm{Persp}(X)$. Finally, as $[x;t] = \lim_{\epsilon\to+0} z_\epsilon$, we have $[x;t] \in \mathrm{cl}(\mathrm{Persp}(X)) = \overline{\mathrm{ConeT}}(X)$, as desired. □

**Fact I.2.7** [Caratheodory Theorem in conic form] Let $a \in \mathbf{R}^m$ be a conic combination (linear combination with nonnegative coefficients) of $N$ vectors $a^1, \ldots, a^N$. Then, $a$ is a conic combination of at most $m$ vectors from the collection $a^1, \ldots, a^N$.

Proof. The proof follows the same lines as the proof of the "plain" Caratheodory Theorem. Consider the minimal, in terms of the number of positive coefficients, representation of $a$ as a conic combination of $a^1, \ldots, a^N$; w.l.o.g. we can assume that this is the representation $a = \sum_{i=1}^K \lambda_i a^i$, $\lambda_i > 0$, $i \le K$. We need to prove that $K \le m$. Assume for contradiction that $K > m$. Consider the system of $m$ scalar linear equations $\sum_{i=1}^K \delta_i a^i = 0$ in variables $\delta$. The number $K$ of unknowns in this system is larger than the number $m$ of equations. Thus, this system has a nontrivial solution $\bar{\delta}$. Passing, if necessary, from $\bar{\delta}$ to $-\bar{\delta}$, we may further assume that some of $\bar{\delta}_i$ are strictly negative. Define $\lambda_i(t) := \lambda_i + t\bar{\delta}_i$ for all $i$. Note that for all $t \ge 0$ we have $a = \sum_{i=1}^K \lambda_i(t) a^i$. Let $t^*$ be the largest $t \ge 0$ for which all $\lambda_i(t)$ are nonnegative ($t^*$ is well defined, since for large $t$ some of $\lambda_i(t)$, i.e., those corresponding to $\bar{\delta}_i < 0$, will become negative), we get $a = \sum_{i=1}^K \lambda_i(t^*) a^i$ with all coefficients $\lambda_i(t^*)$ nonnegative and at least one of them zero, contradicting the origin of $K$. □

# Part II

# Separation Theorem, Extreme Points, Recessive Directions, and Structure of Polyhedral Sets

# 7

# Separation Theorem

We next investigate *Separation Theorem* which is as indispensable when studying general convex sets as is General Theorem on Alternative when investigating properties of polyhedral sets.

## 7.1 Separation: definition

Recall that a *hyperplane* $M$ in $\mathbf{R}^n$ is, by definition, an affine subspace of dimension $n-1$. Then, by Proposition A.47, hyperplanes are precisely the level sets of nontrivial linear forms. That is,

$$\begin{array}{l} M \subset \mathbf{R}^n \text{ is a hyperplane} \\ \iff \exists a \in \mathbf{R}^n,\ a \neq 0,\ \exists b \in \mathbf{R} \text{ such that } M = \left\{x \in \mathbf{R}^n : a^\top x = b\right\}. \end{array}$$

We can associate with the hyperplane $M$ or, better to say, with the associated pair $a$, $b$ (defined by the hyperplane up to multiplication of $a$, $b$ by nonzero real number) the following sets:

- "upper" and "lower" open half-spaces

  $$M^{++} := \left\{x \in \mathbf{R}^n :\ a^\top x > b\right\}, \quad \text{and} \quad M^{--} := \left\{x \in \mathbf{R}^n :\ a^\top x < b\right\}.$$

  These sets clearly are convex, and since a linear form is continuous, and the sets are given by strict inequalities on the value of a continuous function, they indeed are open.

  These open half-spaces are uniquely defined by the hyperplane, up to swapping the "upper" and the "lower" ones (this is what happens when passing from a particular pair $a$, $b$ specifying M to a negative multiple of this pair).
- "upper" and "lower" closed half-spaces

  $$M^{+} := \left\{x \in \mathbf{R}^n :\ a^\top x \geq b\right\}, \quad \text{and} \quad M^{-} := \left\{x \in \mathbf{R}^n :\ a^\top x \leq b\right\}.$$

  These are also convex sets. Moreover, these two sets are polyhedral and thus closed. It is easily seen that the closed upper/lower half-space is the closure of the corresponding open half-space, and $M$ itself is the common boundary of all four half-spaces.

Also, note that our half-spaces and $M$ itself partition $\mathbf{R}^n$, i.e.,

$$\mathbf{R}^n = M^{--} \cup M \cup M^{++}$$

(partitioning by disjoint sets), and

$$\mathbf{R}^n = M^- \cup M^+$$

(where $M$ is the intersection of the sets $M^-$ and $M^+$).

We are now ready to define the basic notion of *separation* of two convex sets $T$ and $S$ by a hyperplane.

**Definition** II.7.1 [Separation] Let $S, T$ be two nonempty convex sets in $\mathbf{R}^n$.

- A hyperplane

$$M = \left\{x \in \mathbf{R}^n : \ a^\top x = b\right\} \quad \text{[where } a \neq 0]$$

is said to *separate* $S$ and $T$, if it satisfies both of the following properties:

– $S \subseteq \left\{x \in \mathbf{R}^n : \ a^\top x \leq b\right\}, \quad T \subseteq \left\{x \in \mathbf{R}^n : \ a^\top x \geq b\right\}$ (i.e., $S$ and $T$ belong to the opposite closed half-spaces into which $M$ splits $\mathbf{R}^n$), and,

– at least one of the sets $S, T$ is not contained in $M$ itself, i.e.,

$$S \cup T \not\subseteq M.$$

- The separation is called *strong*, if there exist $b', b'' \in \mathbf{R}$ satisfying $b' < b < b''$, such that

$$S \subseteq \left\{x \in \mathbf{R}^n : \ a^\top x \leq b'\right\}, \quad T \subseteq \left\{x \in \mathbf{R}^n : \ a^\top x \geq b''\right\}.$$

- A linear form $a \neq 0$ is said to *separate (strongly separate)* $S$ and $T$, if for properly chosen $b$ the hyperplane $\left\{x \in \mathbf{R}^n : \ a^\top x = b\right\}$ separates (strongly separates) $S$ and $T$.
- We say that $S$ and $T$ can be *(strongly) separated*, if there exists a hyperplane which (strongly) separates $S$ and $T$.

Let us examine the separation concept on a few simple examples.

**Example** II.7.1

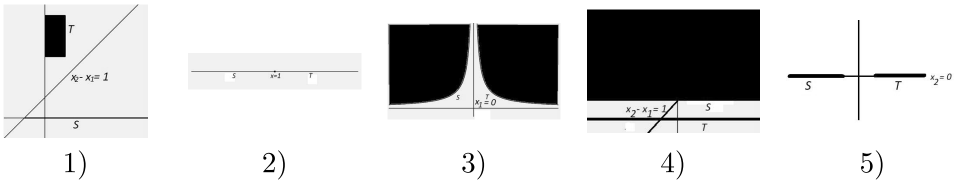


Figure II.1. Separation.

1) The hyperplane $\{x \in \mathbf{R}^2 : \ x_2 - x_1 = 1\}$ strongly separates the polyhedral sets $S = \{x \in \mathbf{R}^2 : x_2 = 0, \, x_1 \geq -1\}$ and $T = \{x \in \mathbf{R}^2 : 0 \leq x_1 \leq 1, \, 3 \leq x_2 \leq 5\}$.
2) The hyperplane $\{x \in \mathbf{R} : \ x = 1\}$ separates (but not strongly separates) the convex sets $S = \{x \in \mathbf{R} : \ x \leq 1\}$ and $T = \{x \in \mathbf{R} : \ x \geq 1\}$.
3) The hyperplane $\{x \in \mathbf{R}^2 : \ x_1 = 0\}$ separates (but not strongly separates) the convex sets $S = \{x \in \mathbf{R}^2 : x_1 < 0, \, x_2 \geq -1/x_1\}$ and $T = \{x \in \mathbf{R}^2 : x_1 > 0, \, x_2 > 1/x_1\}$.

4) The hyperplane $\{x \in \mathbf{R}^2 : \ x_2 - x_1 = 1\}$ does *not* separate the convex sets $S = \{x \in \mathbf{R}^2 : \ x_2 \geq 1\}$ and $T = \{x \in \mathbf{R}^2 : \ x_2 = 0\}$.
5) The hyperplane $\{x \in \mathbf{R}^2 : \ x_2 = 0\}$ does *not* separate the polyhedral sets $S = \{x \in \mathbf{R}^2 : \ x_2 = 0, \ x_1 \leq -1\}$ and $T = \{x \in \mathbf{R}^2 : \ x_2 = 0, \ x_1 \geq 1\}$.

The following equivalent description of separation is used often as well.

**Fact** II.7.2 Let $S, T$ be nonempty convex sets in $\mathbf{R}^n$. A linear form $a^\top x$ separates $S$ and $T$ if and only if

$$\begin{array}{ll} (a) & \sup\limits_{x \in S} a^\top x \leq \inf\limits_{y \in T} a^\top y, \quad \text{and} \\ (b) & \inf\limits_{x \in S} a^\top x < \sup\limits_{y \in T} a^\top y. \end{array}$$

This separation is strong if and only if $(a)$ holds as a strict inequality:

$$\sup_{x \in S} a^\top x < \inf_{y \in T} a^\top y.$$

## 7.2 Separation Theorem

One of the most fundamental results in convex analysis is the following separation theorem.

**Theorem** II.7.3 [Separation Theorem] Let $S$ and $T$ be nonempty convex sets in $\mathbf{R}^n$.

(i) $S$ and $T$ can be separated if and only if their relative interiors do not intersect, i.e., $\operatorname{rint} S \cap \operatorname{rint} T = \varnothing$.

(ii) $S$ and $T$ can be strongly separated if and only if the sets are at a positive distance from each other, i.e.,

$$\operatorname{dist}(S, T) := \inf\{\|x - y\|_2 : \ x \in S, \ y \in T\} > 0.$$

In particular, if $S$ and $T$ are nonempty non-intersecting closed convex sets and one of these sets is compact, then $S$ and $T$ can be strongly separated.

We will use the following simple and important lemma in the proof of the separation theorem.

**Lemma** II.7.4 A point $x \in \operatorname{rint} Q$ of a convex set $Q$ can be the minimizer (or maximizer) of a linear function $f(x) = a^\top x$ if and only if the function is constant on $Q$.

**Proof.** "If" part is evident. To prove the "only if" part, let $\bar{x} \in \operatorname{rint} Q$ be, say, a minimizer of $f$ over $Q$, then for any $y \in Q$ we need to prove that $f(\bar{x}) = f(y)$. There is nothing to prove if $y = \bar{x}$, so let us assume that $y \neq \bar{x}$. Since $Q$ is convex and $\bar{x}, y \in Q$, the segment $[\bar{x}, y]$ belongs to $Q$. Moreover, as $\bar{x} \in \operatorname{rint} Q$ we can extend this segment a little further away from $\bar{x}$ and still remain in $Q$. That is, there exists $z \in Q$ such that $\bar{x} = (1-\lambda)y + \lambda z$ with certain $\lambda \in [0,1)$. As $y \neq \bar{x}$, we have in fact $\lambda \in (0,1)$. Since $f$ is linear, we deduce

$$f(\bar{x}) = (1-\lambda)f(y) + \lambda f(z).$$

Because $\bar{x}$ is a minimizer of $f$ over $Q$ and $y, z \in Q$, we have $\min\{f(y),\, f(z)\} \geq f(\bar{x}) = (1-\lambda)f(y) + \lambda f(z)$. Then, from $\lambda \in (0,1)$ we conclude that this relation can be satisfied only when $f(\bar{x}) = f(y) = f(z)$. □

**Proof of Theorem II.7.3.** We will prove the separation theorem in several steps. We will first focus on the usual separation, i.e., case (i) of the theorem.

**(i) Necessity.** Assume that $S, T$ can be separated. Then, for certain $a \neq 0$ we have

$$\sup_{x\in S} a^\top x \leq \inf_{y\in T} a^\top y, \qquad \text{and} \qquad \inf_{x\in S} a^\top x < \sup_{y\in T} a^\top y. \tag{7.1}$$

Assume for contradiction that $\operatorname{rint} S$ and $\operatorname{rint} T$ have a common point $\bar{x}$. Then, from the first inequality in (7.1) and $\bar{x} \in S \cap T$, we deduce

$$a^\top \bar{x} \leq \sup_{x\in S} a^\top x \leq \inf_{y\in T} a^\top y \leq a^\top \bar{x}.$$

Thus, $\bar{x}$ maximizes the linear function $f(x) = a^\top x$ on $S$ and simultaneously minimizes this function on $T$. Then, as $\bar{x} \in \operatorname{rint} S$ and also $\bar{x} \in \operatorname{rint} T$ by Lemma II.7.4, we conclude $f(x) = f(\bar{x})$ on $S$ and on $T$, so that $f(\cdot)$ is constant on $S \cup T$. This then yields the desired contradiction to the second inequality in (7.1).

**(i) Sufficiency.** The proof of sufficiency part of the Separation Theorem is much more instructive. There are several ways to prove it. Below, we present a proof based on Theorem I.3.2.

**(i) Sufficiency, Step 1: Separation of a nonempty polytope and a point outside the polytope.** We start with seemingly a very particular case of the Separation Theorem – the one where $S = \operatorname{Conv}\{x^1, \ldots, x^N\}$ and $T$ is a singleton $T = \{x\}$ which does not belong to $S$. We will prove that in this case there exists a linear form which strongly separates $T = \{x\}$ and $S$.

The set $S = \operatorname{Conv}\{x^1, \ldots, x^N\}$ is given by the polyhedral representation

$$S = \left\{ z \in \mathbf{R}^n : \ \exists \lambda \text{ such that } \lambda \geq 0,\ \sum_{i=1}^N \lambda_i = 1,\ z = \sum_{i=1}^N \lambda_i a_i \right\},$$

and thus $S$ is polyhedral (Theorem I.3.2). Therefore, for a properly selected $k$, $a_1, \ldots, a_k$, and $b_1, \ldots, b_k$ we have:

$$S = \left\{ z \in \mathbf{R}^n : \ a_i^\top z \leq b_i,\ i \leq k \right\}.$$

Since $x \notin S$, there exists $i \leq k$ such that $a_i^\top x > b_i$, and thus the corresponding $a_i$ clearly strongly separates our $S$ and $T = \{x\}$.

**(i) Sufficiency, Step 2: Separation of a nonempty convex set and a point outside of the set.** Now consider the case when $S$ is an arbitrary nonempty convex set and $T = \{x\}$ is a singleton outside $S$ (here the difference with Step 1 is that now $S$ is not assumed to be a polytope).

Without loss of generality we may assume that $S$ contains 0 (if it is not the case, by taking any $p \in S$, we may translate $S$ and $T$ to the sets $S \mapsto -p + S$, $T \mapsto -p+T$; clearly, a linear form which separates the shifted sets, can be shifted to separate the original ones as well). Let $L$ be the linear span of $S$.

If $x \notin L$, the separation is easy: we can write $x = e + f$, where $e \in L$ and $f$ is from the subspace orthogonal to $L$, and thus

$$f^\top x = f^\top f > 0 = \max_{y \in S} f^\top y,$$

so that $f$ strongly separates $S$ and $T = \{x\}$.

Now, we consider the case when $x \in L$. Since $x \in L$, and $x \notin S$ as well as $\varnothing \neq S \subseteq L$, we deduce that $L$ contains at least two points and so $L \neq \{0\}$. Without loss of generality, we can assume that $L = \mathbf{R}^n$.

Let $\mathcal{B} := \{h \in \mathbf{R}^n : \ \|h\|_2 = 1\}$ be the unit sphere in $\mathbf{R}^n$. This is a closed and bounded set in $\mathbf{R}^n$ (boundedness is evident, and closedness follows from the fact that $\|\cdot\|_2$ is continuous). Thus, $\mathcal{B}$ is a compact set. Let us prove that there exists $f \in \mathcal{B}$ that separates $x$ and $S$ in the sense that

$$f^\top x \geq \sup_{y \in S} f^\top y. \tag{7.2}$$

Assume for contradiction that no such $f$ exists. Then, for every $h \in \mathcal{B}$ there exists $y_h \in S$ such that

$$h^\top y_h > h^\top x.$$

Since the inequality is strict, it immediately follows that there exists an open neighborhood $U_h$ of the vector $h$ such that

$$(h')^\top y_h > (h')^\top x, \quad \forall h' \in U_h. \tag{7.3}$$

Note that the family of open sets $\{U_h\}_{h \in \mathcal{B}}$ covers $\mathcal{B}$. As $\mathcal{B}$ is compact, we can find a finite subfamily $U_{h_1}, \ldots, U_{h_N}$ of this family which still covers $\mathcal{B}$. Let us take the corresponding points $y^1 := y_{h_1}$, $y^2 := y_{h_2}, \ldots, y^N := y_{h_N}$ and define the polytope $\widehat{S} := \operatorname{Conv}\{y^1, \ldots, y^N\}$. Due to the origin of $y^i$, all of these points are in $S$ and thus $S \supseteq \widehat{S}$ (recall that $S$ is convex). Since $x \notin S$, we deduce $x \notin \widehat{S}$. Then, by Step 1, $x$ can be strongly separated from $\widehat{S}$, i.e., there exists $a \neq 0$ such that

$$a^\top x > \sup_{y \in \widehat{S}} a^\top y = \max\left\{a^\top y^i\colon \ 1 \leq i \leq N\right\}. \tag{7.4}$$

By normalization, we may also assume that $\|a\|_2 = 1$, so that $a \in \mathcal{B}$. Recall that $U_{h_1}, \ldots, U_{h_N}$ form a covering of $\mathcal{B}$, and as $a \in \mathcal{B}$, we have that $a$ belongs to certain $U_{h_i}$. By construction of $U_{h_i}$ (see (7.3)), we have

$$a^\top y^i \equiv a^\top y_{h_i} > a^\top x,$$

which contradicts (7.4) as $y^i \in \widehat{S}$.

Thus, we conclude that there exists $f \in \mathcal{B}$ satisfying (7.2). We claim that

$f$ separates $S$ and $\{x\}$. Given that we already established(7.2), all we need to verify for establishing $f$ indeed separates $S$ and $\{x\}$ is to show that the linear form $f(y) = f^\top y$ is non-constant on $S \cup T$. This is evident as we are in the situation when $0 \in S$ and $L = \mathrm{Lin}(S) = \mathbf{R}^n$ and $f \neq 0$, so that $f(y)$ is non-constant already on $S$ (indeed, otherwise we would have $f^\top y = 0$ for $y \in S$ due to $0 \in S$, whence $f^\top y = 0$ for $y \in \mathrm{Lin}(S) = \mathbf{R}^n$, contradicting $f \neq 0$).

**(i) Sufficiency, Step 3: Separation of two nonempty and non-intersecting convex sets.** Now we are ready to prove that two nonempty and non-intersecting convex sets $S$ and $T$ can be separated. To this end consider the arithmetic difference of the sets $S$ and $T$, i.e.,

$$\Delta := S - T = \{x - y : \ x \in S, \ y \in T\}.$$

As $S$ and $T$ are nonempty and convex, $\Delta$ is nonempty and convex (by Proposition I.1.21.3). Also, as $S \cap T = \varnothing$, we have $0 \notin \Delta$. Then, by Step 2, we can separate $\Delta$ and $\{0\}$, i.e., there exists $f \neq 0$ such that

$$f^\top 0 = 0 \geq \sup_{z \in \Delta} f^\top z \quad \text{and} \quad f^\top 0 > \inf_{z \in \Delta} f^\top z.$$

In other words,

$$0 \geq \sup_{x \in S, y \in T} \left\{ f^\top x - f^\top y \right\} \quad \text{and} \quad 0 > \inf_{x \in S, y \in T} \left\{ f^\top x - f^\top y \right\},$$

which clearly means that $f$ separates $S$ and $T$.

**(i) Sufficiency, Step 4: Separation of nonempty convex sets with non-intersecting relative interiors.** Now we are ready to complete the proof of the "if" part of part (i) of the Separation Theorem. Let $S$ and $T$ be two nonempty convex sets such that $\mathrm{rint}\, S \cap \mathrm{rint}\, T = \varnothing$, then we will prove that $S$ and $T$ can be separated. Recall from Theorem I.1.29 that the sets $S' := \mathrm{rint}\, S$ and $T' := \mathrm{rint}\, T$ are nonempty and convex. Moreover, we are given that $S'$ and $T'$ do not intersect, thus they can be separated by Step 3. That is, there exists $f$ such that

$$\inf_{x \in T'} f^\top x \geq \sup_{y \in S'} f^\top x \quad \text{and} \quad \sup_{x \in T'} f^\top x > \inf_{y \in S'} f^\top x. \tag{7.5}$$

It is immediately seen that in fact $f$ separates $S$ and $T$. Indeed, the quantities in the left and the right hand sides of the first inequality in (7.5) clearly remain unchanged when we replace $S'$ with $\mathrm{cl}\, S'$ and $T'$ with $\mathrm{cl}\, T'$. Moreover, by Theorem I.1.29, $\mathrm{cl}\, S' = \mathrm{cl}\, S \supseteq S$ and $\mathrm{cl}\, T' = \mathrm{cl}\, T \supseteq T$, and we get $\inf\limits_{x \in T} f^\top x = \inf\limits_{x \in T'} f^\top x$, and similarly $\sup\limits_{y \in S} f^\top y = \sup\limits_{y \in S'} f^\top y$. Thus, we get from (7.5)

$$\inf_{x \in T} f^\top x \geq \sup_{y \in S} f^\top y.$$

It remains to note that $T' \subseteq T$, $S' \subseteq S$, so that the second inequality in (7.5) implies that

$$\sup_{x \in T} f^\top x > \inf_{y \in S} f^\top x.$$

**(ii) Necessity:** Prove yourself.

**(ii) Sufficiency:** Define $\rho := \operatorname{dist}(S,T) = \inf\{\|x-y\|_2 : \ x \in S, \ y \in T\}$. In case (ii), we are given that $\rho > 0$. Consider the set $\widehat{S} := \left\{x \in \mathbf{R}^n : \ \inf\limits_{y\in S} \|x-y\|_2 \le \rho/2\right\}$. As $S$ is convex, the set $\widehat{S}$ is convex (recall Example I.1.3). Moreover, $\widehat{S} \cap T = \varnothing$ (why?). Then, by part (i), $\widehat{S}$ and $T$ can be separated. Let $f$ be any linear form that separates $\widehat{S}$ and $T$. Then, the same form strongly separates $S$ and $T$ (why?). The last statement of (ii), i.e., "in particular" part, readily follows from the just proved statement due to the fact that if two closed nonempty sets in $\mathbf{R}^n$ do not intersect and one of them is compact, then the sets are at positive distance from each other (why?). □

**Remark** II.7.5 In Theorem II.7.3, a careful reader would notice that the considerations in the proof (i) Sufficiency, Step 1, i.e., separation of a polytope and a point not in the polytope, are based on solely Theorem I.3.2, and this is a "purely arithmetic" statement: when proving it, we never used things like convergence, compactness, square roots, etc., just rational arithmetics. Therefore, the result stated at Step 1 remain valid if we replace our universe $\mathbf{R}^n$ with the space $\mathbf{Q}^n$ of $n$-dimensional *rational* vectors (those with rational coordinates; of course, the multiplication by reals in this space should be restricted to multiplication by rationals). The possibility to separate a rational vector from a "rational" polytope by a *rational* linear form, which is the "rational" version of the result of Step 1, definitely are of interest (e.g., for Integer Programming). In fact, all results in Part 1 are derived from Fourier-Motzkin elimination and Theorem I.3.2, i.e., the existence of optimal solution(s) to feasible and bounded LP programs, Farkas Lemmas, General Theorem on Alternative, plain and conic Caratheodory and (finite family version of) Helly theorems, etc., remains valid when $\mathbf{R}^n$ is replaced with $\mathbf{Q}^n$ (provided, of course, that the related data are rational). In particular, any feasible and bounded LP program *with rational data* admits a rational optimal solution, which definitely is worthy of knowing.

In contrast to these "purely arithmetic" considerations at Step 1, at Step 2, i.e., for the separation of a closed convex set and a point outside of the set, we used compactness, which heavily exploits the fact that our universe is $\mathbf{R}^n$ and not, say, $\mathbf{Q}^n$ (in the latter space bounded and closed sets not necessarily are compact).

In fact, we could not avoid things like compactness arguments at Step 2, since the very fact we are proving is true in $\mathbf{R}^n$ but not in $\mathbf{Q}^n$. Indeed, consider the "rational plane," i.e., the universe composed of all 2-dimensional vectors with rational entries, and let $S$ be the half-plane in this rational plane given by the linear inequality

$$x_1 + \alpha x_2 \le 0,$$

where $\alpha$ is irrational. $S$ clearly is a "convex set" in $\mathbf{Q}^2$. But, it is immediately seen that a point outside this set cannot be separated from $S$ by a rational linear form.

# 8

# Consequences of Separation Theorem

Separation Theorem admits a number of important consequences. In this chapter, we will discuss these.

## 8.1 Supporting hyperplanes

By Separation Theorem, we immediately deduce that a nonempty closed convex set $M$ is precisely the intersection of all closed half-spaces containing $M$. Among these half-spaces, the most interesting are the "extreme" ones, i.e., those with boundary hyperplanes touching $M$. Such extreme hyperplanes are called supporting hyperplanes. While this notion of extreme makes sense for an arbitrary (not necessary closed) convex set, we will use it for closed convex sets only, and include the requirement of closedness in the definition:

**Definition** II.8.1 [Supporting hyperplane] Let $M$ be a convex closed set in $\mathbf{R}^n$, and let $x \in \operatorname{rbd} M$. A hyperplane

$$\Pi := \left\{y \in \mathbf{R}^n : \ a^\top y = a^\top x\right\} \qquad [\text{where } a \neq 0]$$

is called *supporting to $M$ at $x$*, if it separates $M$ and $\{x\}$, i.e., if

$$a^\top x \geq \sup_{y \in M} a^\top y \quad \text{and} \quad a^\top x > \inf_{y \in M} a^\top y. \tag{8.1}$$

Independent of whether $M$ is or is not closed, a point $x \in \operatorname{rbd} M$ is a limit of points from $M$, and thus the first inequality in (8.1) cannot be strict. As a result, we arrive at an equivalent definition of a supporting hyperplane for convex sets as follows.

*Given a closed convex set $M \in \mathbf{R}^n$ and a point $x \in \operatorname{rbd} M$, a hyperplane*

$$\{y \in \mathbf{R}^n : \ a^\top y = a^\top x\}$$

*is supporting to $M$ at $x$ if and only if the linear form $a(y) := a^\top y$ attains its maximum on $M$ at the point $x$ and is non-constant on $M$.*

**Example** II.8.1 The hyperplane $\{x \in \mathbf{R}^n : \ x_1 = 1\}$ clearly is supporting to the unit Euclidean ball $\{x \in \mathbf{R}^n : \ \|x\|_2 \leq 1\}$ at the point $x = e_1 = [1; 0; \ldots; 0]$.

The most important property of a supporting hyperplane is its existence:

**Proposition** II.8.2 [Existence of supporting hyperplanes] Let $M$ be a closed convex set in $\mathbf{R}^n$, and let $x \in \operatorname{rbd} M$. Then,

(i) there exists at least one hyperplane which is supporting to $M$ at $x$;

(ii) if a hyperplane $\Pi$ is supporting to $M$ at $x$, then $\Pi \cap M$ is a nonempty closed convex set, $x \in \Pi \cap M$, and $\dim(\Pi \cap M) < \dim(M)$.

**Proof.** To see (i) consider any $x \in \operatorname{rbd} M$. Then, $x \notin \operatorname{rint} M$, and therefore the point $\{x\}$ and $\operatorname{rint} M$ can be separated by the Separation Theorem. The associated separating hyperplane is exactly the desired hyperplane supporting to $M$ at $x$.

To prove (ii), note that if $\Pi = \left\{y \in \mathbf{R}^n : \; a^\top y = a^\top x\right\}$ is supporting to $M$ at $x \in \operatorname{rbd} M$, then the set $M' := M \cap \Pi$ is nonempty (as it contains $x$) and is closed and convex as both $\Pi$ and $M$ are. Moreover, the linear form $a^\top y$ is constant on $M'$ and therefore (why?) on $\operatorname{Aff}(M')$. At the same time, this form is non-constant on $M$ by definition of a supporting plane. Thus, $\operatorname{Aff}(M')$ is a proper (less than the entire $\operatorname{Aff}(M)$) subset of $\operatorname{Aff}(M)$, and therefore the affine dimension of $\operatorname{Aff}(M')$, which is by definition nothing but $\dim(M')$, is less than the affine dimension of $\operatorname{Aff}(M)$, which is precisely $\dim(M)$ [1]. □

## 8.2 Extreme points and Krein-Milman Theorem

Supporting hyperplanes are useful in proving the existence of *extreme points* of convex sets. Geometrically, an extreme point of a convex set is a point in the set which cannot be written as a convex combination of other points from the set. The importance of this notion originates from the following fact which we will soon prove: any "good enough" (in fact just nonempty compact) convex set $M$ is nothing but the convex hull of its extreme points, and the set of extreme points of such a set $M$ is the *smallest* set whose convex hull is equal to $M$. That is, every extreme point of a nonempty compact convex set $M$ is essential.

### *8.2.1 Extreme points: definition*

The exact definition of an extreme point is as follows:

**Definition** II.8.3 [Extreme points] Let $M$ be a nonempty convex set in $\mathbf{R}^n$. A point $x \in M$ is called an *extreme point* of $M$, if there is no nontrivial (of positive length) segment $[u, v] \in M$ for which $x$ is an interior point. That is, $x$ is an *extreme point* of $M$ if the relation

$$x = \lambda u + (1 - \lambda) v$$

[1] For dimension of a subset in $\mathbf{R}^n$, see Definition I.2.1 or/and section A.4.3. We have used the following immediate observation: *If $M \subset M'$ are two affine planes, then $\dim M \leq \dim M'$, with equality implying that $M = M'$. The readers are encouraged to prove this fact on their own.*

with $\lambda \in (0, 1)$ and $u, v \in M$ holds if and only if

$$u = v = x.$$

The set of all extreme points of $M$ is denoted by $\mathrm{Ext}(M)$.

In the case of polyhedral sets, extreme points are also referred to as *vertices.*

**Example** II.8.2

- The extreme points of a segment $[x, y] \in \mathbf{R}^n$ are exactly its endpoints $\{x, y\}$.
- The extreme points of a triangle are its vertices.
- The extreme points of a (closed) circle on the 2-dimensional plane are the points of the circumference.
- The convex set $M := \left\{x \in \mathbf{R}^2_+ : x_1 > 0, x_2 > 0\right\}$ does not have any extreme points.
- The only extreme point of the convex set $M := \{[0; 0]\} \cup \left\{x \in \mathbf{R}^2_+ : x_1 > 0, x_2 > 0\right\}$ is the point $[0; 0]$.
- The closed convex set $\{x \in \mathbf{R}^2 : x_1 = 0\}$ does not have any extreme points.

An equivalent definition of an extreme point is as follows:

**Fact** II.8.4 Let $M$ be a nonempty convex set and let $x \in M$. Then, $x$ is an extreme point of $M$ if and only if any (and then all) of the following holds:

(i) the only vector $h$ such that $x \pm h \in M$ is the zero vector;
(ii) in every representation $x = \sum_{i=1}^m \lambda_i x^i$ of $x$ as a convex combination, with positive coefficients, of points $x^i \in M$, $i \leq m$, one has $x^1 = \ldots = x^m = x$;
(iii) the set $M \setminus \{x\}$ is convex.

Fact II.8.4.(iii) also admits the following immediate corollary.

**Fact** II.8.5 All extreme points of the convex hull $\mathrm{Conv}(Q)$ of a set $Q$ belong to $Q$:

$$\mathrm{Ext}(\mathrm{Conv}(Q)) \subseteq Q.$$

### *8.2.2 Krein-Milman Theorem*

There are convex sets that do not necessarily possess extreme points; as an example you may take the open unit ball in $\mathbf{R}^n$. This example is not so interesting as the set in question is not closed, and when we replace it with its closure the resulting set is the closed unit ball with plenty of extreme points, i.e., all points of the boundary. There are, however, *closed* convex sets which do not possess extreme points. Consider for example, a line or an affine subspace of larger dimension as the convex set. Indeed, a nonempty *closed* convex set will have no extreme points *only when* it contains a line.

We will next prove that any nonempty closed convex set $M$ that does not contain lines for sure possesses extreme points. Furthermore, if $M$ is a nonempty convex compact set, it possesses a quite representative set of extreme points, i.e., their convex hull is the entire $M$.

**Theorem** II.8.6 Let $M$ be a nonempty closed convex set in $\mathbf{R}^n$. Then,
(i) the set of extreme points of $M$, i.e., $\mathrm{Ext}(M)$, is nonempty if and only if $M$ does not contain lines;
(ii) if $M$ is bounded, then $M = \mathrm{Conv}(\mathrm{Ext}(M))$, i.e., every point of $M$ is a convex combination of the points of $\mathrm{Ext}(M)$.

**Remark** II.8.7 Part (ii) of this theorem is the finite-dimensional version of the famous *Krein-Milman Theorem* (1940). In fact, Hermann Minkowski (1911) established Part (ii) of this theorem for the case $n = 3$, and Ernst Steinitz (1916) showed it for any (finite) $n$.

We will use a number of lemmas in the proof of Theorem II.8.6. The first one states that in the case of a closed convex set, we can add any "recessive" direction to any point in the set and still remain in the set.

**Lemma** II.8.8 Let $M \in \mathbf{R}^n$ be a nonempty closed convex set. Then, whenever $M$ contains a ray

$$\{\bar{x} + th : \ t \geq 0\}$$

starting at $\bar{x} \in M$ with the direction $h \in \mathbf{R}^n$, $M$ also contains all parallel rays starting at the points of $M$, i.e., for all $x \in M$

$$\{x + th : \ t \geq 0\} \subseteq M.$$

As a consequence, if $M$ contains a certain line, then it contains also all parallel lines passing through the points of $M$.

**Proof.** Suppose $\bar{x} + th \in M$ for all $t \geq 0$. Consider any point $x \in M$. Since $M$ is convex, for any fixed $\tau \geq 0$ we have

$$\epsilon\left(\bar{x} + \frac{\tau}{\epsilon}h\right) + (1-\epsilon)x \in M, \qquad \forall \epsilon \in (0,1).$$

By taking the limit as $\epsilon \to +0$ and noting that $M$ is closed, we deduce that $x + \tau h \in M$ for every $\tau \geq 0$. □

Note that Lemma II.8.8 admits a corollary as follows:

**Lemma** II.8.9 Let $M \in \mathbf{R}^n$ be a nonempty convex set, not necessarily closed. Suppose $\mathrm{cl}\, M$ contains a ray

$$\{\bar{x} + th : \ t \geq 0\}$$

starting at $\bar{x} \in \mathrm{cl}\, M$ with the direction $h \in \mathbf{R}^n$ and $\widehat{x} \in \mathrm{rint}\, M$. Then, $\mathrm{rint}\, M$ contains the ray

$$\{\widehat{x} + th : \ t \geq 0\}.$$

In particular, $\mathrm{cl}\, M$ contains a ray (a straight line) if and only if $M$ contains a ray (resp., a straight line) with the same direction.

**Proof.** With $\bar{x}$, $h$, and $\widehat{x}$ as above, for every $t > 0$ the point $x^t := \widehat{x} + 2th$ belongs to $\mathrm{cl}\, M$ by Lemma II.8.8. Taking into account that $\widehat{x} \in \mathrm{rint}\, M$ and invoking Lemma I.1.30, we conclude that $\widehat{x} + th = \frac{1}{2}[\widehat{x} + x^t] \in \mathrm{rint}\, M$. Thus, $\widehat{x} + th \in \mathrm{rint}\, M$ for all $t > 0$. Finally, this inclusion $\widehat{x} + th \in \mathrm{rint}\, M$ holds true for $t = 0$ as well due to $\widehat{x} \in \mathrm{rint}\, M$. □

Our last ingredient for the proof of Theorem II.8.6 is a lemma stating a nice transitive property of extreme points: that is, the extreme points of subsets of nonempty closed convex sets obtained from the intersection with a supporting hyperplane of the set are also extreme for the original set.

**Lemma** II.8.10 Let $M \subset \mathbf{R}^n$ be a nonempty closed convex set. Then, for any $\bar{x} \in \mathrm{rbd}\, M$ and any hyperplane $\Pi$ that is supporting to $M$ at $\bar{x}$, we have that the set $\Pi \cap M$ is nonempty closed and convex, and $\mathrm{Ext}(\Pi \cap M) \subseteq \mathrm{Ext}(M)$.

**Proof.** First statement, i.e., $\Pi \cap M$ is nonempty closed and convex, follows from Proposition II.8.2(ii). Moreover, by Proposition II.8.2(ii) we have $\bar{x} \in \Pi \cap M$.

Next, let $a \in \mathbf{R}^n$ be the linear form associated with $\Pi$, i.e.,

$$\Pi = \left\{ y \in \mathbf{R}^n : \ a^\top y = a^\top \bar{x} \right\},$$

so that

$$\inf_{x \in M} a^\top x < \sup_{x \in M} a^\top x = a^\top \bar{x} \tag{8.2}$$

(see Proposition II.8.2). Consider any extreme point $y$ of $\Pi \cap M$. Assume for contradiction that $y \notin \mathrm{Ext}(M)$. Then, there exists two distinct points $u, v \in M$ and $\lambda \in (0, 1)$ such that

$$y = \lambda u + (1 - \lambda) v.$$

As $y \in \Pi \cap M$ we have $a^\top y = a^\top \bar{x}$ and also as $u, v \in M$, from (8.2) we deduce that

$$a^\top y = a^\top \bar{x} \geq \max \left\{ a^\top u, \ a^\top v \right\}.$$

On the other hand, from the relation $y = \lambda u + (1 - \lambda) v$ we have

$$a^\top y = \lambda a^\top u + (1 - \lambda) a^\top v.$$

Combining these last two observations and taking into account that $\lambda \in (0, 1)$, we conclude that

$$a^\top y = a^\top u = a^\top v.$$

Then, by the definition of $\Pi$, these equalities imply that $u, v \in \Pi$. As $u, v \in M$ as well, this contradicts that $y \in \mathrm{Ext}(\Pi \cap M)$ as we have written $y = \lambda u + (1 - \lambda) v$ using distinct points $u, v \in \Pi \cap M$ and some $\lambda \in (0, 1)$. □

We are now ready to prove Theorem II.8.6.

**Proof of Theorem II.8.6.** Let us start with (i). The "only if" part for (i) follows from Lemma II.8.8. Indeed, for the "only if" part we need to prove that if $M$ possesses extreme points, then $M$ does not contain lines. That is, we need to prove that if $M$ contains lines, then it has no extreme points. But, this is indeed immediate: if $M$ contains a line, then, by Lemma II.8.8, there is a line in $M$ passing through every given point of $M$, so that no point can be extreme.

Now let us prove the "if" part of (i). Thus, from now on we assume that $M$ does not contain lines, and our goal is to prove that then $M$ possesses extreme points. Equipped with Lemma II.8.10 and Proposition II.8.2, we will prove this by induction on $\dim(M)$.

There is nothing to do if $\dim(M) = 0$, i.e., if $M$ is a single point – then, of course, $M = \mathrm{Ext}(M)$. Now, for the induction hypothesis, for some integer $k > 0$, we assume that all nonempty closed convex sets $T$ that do not contain lines and have $\dim(T) = k$ satisfy $\mathrm{Ext}(T) \neq \varnothing$. To complete the induction, we will show that this statement is valid for such sets of dimension $k+1$ as well. Let $M$ be a nonempty, closed, convex set that does not contain lines and has $\dim(M) = k+1$. Since $M$ does not contain lines and $\dim(M) > 0$, we have $M \neq \mathrm{Aff}(M)$. We claim that $M$ possesses a relative boundary point $\bar{x}$. To see this, note that there exists $z \in \mathrm{Aff}(M) \setminus M$, and thus for any fixed $x \in M$ the point

$$x_\lambda := x + \lambda(z - x)$$

does not belong to $M$ for some $\lambda > 0$ (and then, by convexity of $M$, for all larger values of $\lambda$), while $x_0 = x$ belongs to $M$. The set of those $\lambda \geq 0$ for which $x_\lambda \in M$ is therefore nonempty and bounded from above; this set clearly is closed (since $M$ is closed). Thus, there exists the largest $\lambda = \lambda^*$ for which $x_\lambda \in M$. We claim that $x_{\lambda^*} \in \mathrm{rbd}\, M$. Indeed, by construction $x_{\lambda^*} \in M$. If $x_{\lambda^*}$ were to be in $\mathrm{rint}\, M$, then all the points $x_\lambda$ with $\lambda$ values greater than $\lambda^*$ yet close to $\lambda^*$ would also belong to $M$, which contradicts the origin of $\lambda^*$.

Thus, there exists $\bar{x} \in \mathrm{rbd}\, M$. Then, by Proposition II.8.2(i), there exists a hyperplane $\Pi = \left\{x \in \mathbf{R}^n : a^\top x = a^\top \bar{x}\right\}$ which is supporting to $M$ at $\bar{x}$:

$$\inf_{x \in M} a^\top x < \max_{x \in M} a^\top x = a^\top \bar{x}.$$

Moreover, by Proposition II.8.2(ii), the set $T := \Pi \cap M$ is nonempty closed and convex and it satisfies $\dim(T) < \dim(M)$, i.e., $\dim(T) \leq k$. As $M$ does not contain lines, $T \subset M$ clearly does not contain lines either. Then, by the inductive hypothesis, $T$ possesses extreme points, i.e., $\mathrm{Ext}(T) \neq \varnothing$. Moreover, by Lemma II.8.10 $\mathrm{Ext}(M) \supseteq \mathrm{Ext}(\Pi \cap M) = \mathrm{Ext}(T) \neq \varnothing$. This completes the inductive step, and hence (i) is proved.

Now let us prove (ii). Let $M$ be nonempty, closed, convex, and bounded. We need to prove that

$$M = \mathrm{Conv}(\mathrm{Ext}(M)).$$

As $M$ is convex, we immediately observe that $M \supseteq \mathrm{Conv}(\mathrm{Ext}(M))$. Thus, all we need is to prove that every $x \in M$ is a convex combination of points from $\mathrm{Ext}(M)$. Here, we again use induction on $\dim(M)$. The case of $\dim(M) = 0$, i.e., when $M$ is a single point, is trivial. Assume that the statement holds for all $k$-dimensional closed convex and bounded sets. Let $M$ be a closed convex and bounded set with $\dim(M) = k + 1$. Consider any $x \in M$. To represent $x$ as a convex combination of points from $\mathrm{Ext}(M)$, let us pass through $x$ an arbitrary line $\ell = \{x + \lambda h : \lambda \in \mathbf{R}\}$ (where $h \neq 0$) in the affine span $\mathrm{Aff}(M)$ of $M$. Moving along this line from $x$ in each of the two possible directions, we eventually leave $M$ (since $M$ is bounded). Then, there exist nonnegative $\lambda^+$ and $\lambda^-$ such that the points

$$\bar{x}_+ := x + \lambda^+ h, \qquad \bar{x}_- := x - \lambda^- h$$

both belong to $\mathrm{rbd}\, M$. We claim that $\bar{x}_\pm$ admit convex combination representation using points from $\mathrm{Ext}(M)$ (this will complete the proof, since $x$ clearly is a convex combination of the two points $\bar{x}_\pm$). Indeed, by Proposition II.8.2(i) there exists a hyperplane $\Pi$ supporting to $M$ at $\bar{x}_+$, and by Proposition II.8.2(ii) the set $\Pi \cap M$ is nonempty, closed and convex with $\dim(\Pi \cap M) < \dim(M) = k + 1$. Moreover, as $M$ is bounded $\Pi \cap M$ is bounded as well. Then, by the inductive hypothesis, $\bar{x}_+ \in \mathrm{Conv}(\mathrm{Ext}(\Pi \cap M))$. Moreover, since by Lemma II.8.10 we have $\mathrm{Ext}(\Pi \cap M) \subseteq \mathrm{Ext}(M)$, we conclude $\bar{x}_+ \in \mathrm{Conv}(\mathrm{Ext}(M))$. Analogous reasoning is valid for $\bar{x}_-$ as well. □

## 8.3 Recessive directions and recessive cone

Lemma II.8.8 states that if $M$ is a nonempty closed convex set, then the set of all directions $h$ such that $x + th \in M$ for *some* $x$ and all $t \geq 0$ is exactly the same as the set of all directions $h$ such that $x + th \in M$ for *all* $x \in M$ and all $t \geq 0$. Directions of this type play an important role in the theory of convex sets, and consequently they have a name – they are called *recessive directions* of $M$.

**Definition** II.8.11 [Recessive directions and recessive cone] Given a nonempty closed convex set $M \subseteq \mathbf{R}^n$, a direction $h \in \mathbf{R}^n$ is called a *recessive direction* of $M$ if we have $x + th \in M$ for any $x \in M$ and any $t \geq 0$.
The set of all recessive directions is called the *recessive cone* of $M$ [notation: $\mathrm{Rec}(M)$].

**Remark** II.8.12 Given a closed convex set $M$, we immediately deduce that $\mathrm{Rec}(M)$ indeed is a closed cone (prove it!) and that

$$M + \mathrm{Rec}(M) = M. \tag{8.3}$$

Let us see some examples of recessive cones of sets.

**Example** II.8.3

- The recessive cone of $\mathbf{R}^n_+$ is itself. In fact, the recessive cone of any closed cone is itself.
- Consider the set $M := \{x \in \mathbf{R}^n : \sum_{i=1}^n x_i = 1\}$; then $\mathrm{Rec}(M) = \{h \in \mathbf{R}^n : \sum_{i=1}^n h_i = 0\}$.

- Consider the set $M := \{x \in \mathbf{R}^n_+ : \sum_{i=1}^n x_i = 1\}$; then $\mathrm{Rec}(M) = \{0\}$.
- Consider the set $M := \{x \in \mathbf{R}^n_+ : \sum_{i=1}^n x_i \geq 1\}$; then $\mathrm{Rec}(M) = \mathbf{R}^n_+$.
- Consider the set $M := \{x \in \mathbf{R}^2_+ : x_1 x_2 \geq 1\}$; then $\mathrm{Rec}(M) = \mathbf{R}^2_+$.
- Consider the set $M := \{x \in \mathbf{R}^n : x_n - a_n \geq \|(x_1, \ldots, x_{n-1}) - (a_1, \ldots, a_{n-1})\|_2\}$, where $a = (a_1, \ldots, a_n)$ is a given point. Then, $\mathrm{Rec}(M) = \mathbf{L}^n$.

**Fact** II.8.13 Let $M$ be a nonempty closed convex set in $\mathbf{R}^n$. Then,

(i) $\mathrm{Rec}(M) \neq \{0\}$ if and only if $M$ is unbounded.

(ii) If $M$ is unbounded, then all nonzero recessive directions of $M$ are positive multiples of recessive directions of unit Euclidean length, and the latter are *asymptotic directions* of $M$, i.e., a unit vector $h \in \mathbf{R}^n$ is a recessive direction of $M$ if and only if there exists a sequence $\{x^i \in M\}_{i \geq 1}$ such that $\|x^i\|_2 \to \infty$ as $i \to \infty$ and $h = \lim_{i\to\infty} x^i/\|x^i\|_2$.

(iii) $M$ does not contain lines if and only if the cone $\mathrm{Rec}(M)$ does not contain lines.

Here is how we can "visualize" (or compute) the recessive cone of a nonempty closed convex set:

**Fact** II.8.14 Let $M \subseteq \mathbf{R}^n$ be a nonempty closed convex set. Recall its closed conic transform is given by

$$\overline{\mathrm{ConeT}}(M) = \mathrm{cl}\left\{[x;t] \in \mathbf{R}^n \times \mathbf{R} : \ t > 0, \ x/t \in M\right\},$$

(see section 1.5). Then,

$$\mathrm{Rec}(M) = \left\{h \in \mathbf{R}^n : \ [h;0] \in \overline{\mathrm{ConeT}}(M)\right\}.$$

Finally, the recessive cones of nonempty polyhedral sets in fact admit a much simpler characterization.

**Fact** II.8.15 For any nonempty polyhedral set $M = \{x \in \mathbf{R}^n : Ax \leq b\}$, its recessive cone is given by

$$\mathrm{Rec}(M) = \{h \in \mathbf{R}^n : Ah \leq 0\},$$

i.e., $\mathrm{Rec}(M)$ is given by homogeneous version of linear constraints specifying $M$.

We have seen in Theorem II.8.6 that if $M$ is a nonempty convex compact set, it possesses a quite representative set of extreme points, i.e., their convex hull is the entire $M$. We close this section by extending this result as follows.

**Theorem** II.8.16 Let $M \subset \mathbf{R}^n$ be a nonempty closed convex set.

(i) If $M$ does not contain lines, then the set $\mathrm{Ext}(M)$ of extreme points of $M$ is nonempty, and

$$M = \mathrm{Conv}(\mathrm{Ext}(M)) + \mathrm{Rec}(M). \tag{8.4}$$

(ii) In every representation, if any, of $M$ as $M = V + K$ with a nonempty bounded set $V$ and a closed cone $K$ the cone $K$ is $\text{Rec}(M)$ and $V$ contains $\text{Ext}(M)$.

**Proof.**

(i): By Theorem II.8.6(i) we already know that any nonempty closed convex that does not contain lines must possess extreme points. We will prove the rest of Part (i) by induction on $\dim(M)$. There is nothing to prove when $\dim(M) = 0$, that is, $M$ is a singleton. So, suppose that the claim holds true for all sets of dimension $k$. Let $M$ be any nonempty closed convex that does not contain lines and has $\dim(M) = k + 1$. To complete the induction step, we will show that $M$ satisfies the relation (8.4). Consider $x \in M$ and let $e$ be a nonzero direction parallel to $\text{Aff}(M)$ (such a direction exists, since $\dim(M) = k + 1 \geq 1$). Recalling that $M$ does not contain lines and replacing, if necessary, $e$ with $-e$, we can assume that $-e$ is not a recessive direction of $M$. Same as in the proof of Theorem II.8.6, $x$ admits a representation $x = x^- + t_- e$ with $t_- \geq 0$ and $x^- \in \text{rbd}(M)$. Define $M_-$ to be the intersection of $M$ with the plane $\Pi_-$ supporting to $M$ at $x^-$. Then, $M_-$ is a nonempty closed convex subset of $M$ and $\dim(M_-) \leq k$. Also, $M_-$ does not contain lines as $M_- \subset M$ and $M$ does not contain lines. Thus, by inductive hypothesis, $x^-$ is the sum of a point from the nonempty set $\text{Conv}(\text{Ext}(M_-))$ and a recessive direction $h_-$ of $M_-$. As in the proof of Theorem II.8.6, $\text{Ext}(M_-) \subseteq \text{Ext}(M)$, and of course $h_- \in \text{Rec}(M)$ due to $\text{Rec}(M_-) \subseteq \text{Rec}(M)$ (why?). Thus, $x = v_- + h_- + t_- e$ with $v_- \in \text{Conv}(\text{Ext}(X))$ and $h_- \in \text{Rec}(M)$. Now, there are two possibilities: $e \in \text{Rec}(M)$ and $e \notin \text{Rec}(M)$. In the first case, $x = v_- + h$ with $h = h_- + t_- e \in \text{Rec}(M)$ (recall $h_- \in \text{Rec}(M)$ and in this case we also have $e \in \text{Rec}(M)$), that is, $x \in \text{Conv}(\text{Ext}(M)) + \text{Rec}(M)$. In the second case, we can apply the above construction to the vector $-e$ in the role of $e$, ending up with a representation of $x$ of the form $x = v_+ + h_+ - t_+ e$ where $v_+ \in \text{Conv}(\text{Ext}(M))$, $h_+ \in \text{Rec}(M)$ and $t_+ \geq 0$. Taking appropriate convex combination of the resulting pair of representations of $x$, we can cancel the terms with $e$ and arrive at $x = \lambda v_- + (1 - \lambda) v_+ + \lambda h_- + (1 - \lambda) h_+$, resulting in $x \in \text{Conv}(\text{Ext}(M)) + \text{Rec}(M)$. This reasoning holds true for every $x \in M$, hence we deduce $M \subseteq \text{Conv}(\text{Ext}(M)) + \text{Rec}(M)$. The opposite inclusion is given by (8.3) due to $\text{Conv}(\text{Ext}(M)) \subseteq M$. This then completes the proof of the inductive hypothesis, and thus Part (i) is proved.

(ii): Now assume that $M$, in addition to being nonempty, closed and convex, is represented as $M = V + K$, where $K$ is a closed cone and $V$ is a nonempty bounded set, and let us prove that $K = \text{Rec}(M)$ and $V \supseteq \text{Ext}(M)$. Indeed, every vector from $K$ clearly is a recessive direction of $V + K$, so that $K \subseteq \text{Rec}(M)$. To prove the opposite inclusion $K \supseteq \text{Rec}(M)$, consider any $h \in \text{Rec}(M)$, and let us prove that $h \in K$. Let us pick a point $v \in M$. The vectors $v + ih$, $i = 1, 2, \ldots$, belong to $M$ and therefore $v + ih = v^i + h^i$ for some $v^i \in V$ and $h^i \in K$ due to $M = V + K$. It follows that $h = i^{-1}[v^i - v] + i^{-1} h^i$ for $i = 1, 2, \ldots$, that is, $h = \lim_{i \to \infty} i^{-1} h^i$ (recall that $V$ is bounded). As $h^i \in K$ and $K$ is a cone, $i^{-1} h^i \in K$ and so $h$ is the limit of a sequence of points in $K$. Since $K$ is closed, we deduce $h \in K$, as claimed. Thus, $K = \mathrm{Rec}(M)$. It remains to prove that $\mathrm{Ext}(M) \subseteq V$. This is immediate: consider any $w \in \mathrm{Ext}(M)$, then as $M = V + K = V + \mathrm{Rec}(M)$ and $w \in M$, we have $w = v + e$ with some $v \in V \subseteq M$ and $e \in \mathrm{Rec}(M)$, implying that $w - e = v \in M$. Besides this, $w + e \in M$ as $w \in M$ and $e \in \mathrm{Rec}(M)$. Thus, $w \pm e \in M$. Since $w$ is an extreme point of $M$, we conclude that $e = 0$, that is, $w = v \in V$. □

Finally, let us consider what happens to the recessive directions after the projection operation.

**Proposition** II.8.17 Let $M^+ \in \mathbf{R}^n_x \times \mathbf{R}^k_u$ be a nonempty closed convex set such that its projection

$$M = \left\{x \in \mathbf{R}^n : \ \exists u : \ [x;u] \in M^+\right\}$$

is closed. Then,

$$[h_x;h_u] \in \mathrm{Rec}(M^+) \Longrightarrow h_x \in \mathrm{Rec}(M).$$

**Proof.** Consider any recessive direction $[h_x;h_u] \in \mathrm{Rec}(M^+)$. Then, for any $[\bar{x};\bar{u}] \in M^+$, the ray $\{[\bar{x};\bar{u}] + t[h_x;h_u] : \ t \geq 0\}$ is contained in $M^+$. The projection of this ray on the $x$-plane is given by the ray $\{\bar{x} + th_x : \ t \geq 0\}$, which is contained in $M$. Thus, $h_x \in \mathrm{Rec}(M)$. □

While Proposition II.8.17 states that $[h_x;h_u] \in \mathrm{Rec}(M^+) \implies h_x \in \mathrm{Rec}(M)$, in general, $\mathrm{Rec}(M)$ can be much larger than the projection of $\mathrm{Rec}(M^+)$ onto $x$-plane. Our next example illustrates this.

**Example** II.8.4 Consider the sets $M^+ = \{[x;u] \in \mathbf{R}^2 : \ u \geq x^2\}$ and $M = \{x \in \mathbf{R}^2 : \ \exists u \in \mathbf{R} : \ [x;u] \in M^+\}$. Then, $M$ is the entire $x$-axis and $\mathrm{Rec}(M) = M$ is the entire $x$-axis. On the other hand, $\mathrm{Rec}(M^+) = \{[0;h_u] : \ h_u \geq 0\}$ and the projection of $\mathrm{Rec}(M^+)$ onto the $x$-axis is just the origin.

In fact, the pathology highlighted in Example II.8.4 can be eliminated when we have that the set of extreme points of the convex representation $M^+$ of a convex set $M$ is bounded and the projection of $\mathrm{Rec}(M^+)$ is closed.

**Proposition** II.8.18 Let $M^+ \subset \mathbf{R}^n_x \times \mathbf{R}^k_u$ be a nonempty closed convex set such that $M^+ = V + \mathrm{Rec}(M^+)$ for some bounded and closed set $V$. Let $M$ be the projection of $M^+$ onto the $x$-plane, i.e.,

$$M = \left\{x \in \mathbf{R}^n_x : \ \exists u \in \mathbf{R}^k_u : \ [x;u] \in M^+\right\}.$$

Assume that the cone

$$K = \left\{h_x \in \mathbf{R}^n_x : \exists h_u \in \mathbf{R}^k_u : \ [h_x;h_u] \in \mathrm{Rec}(M^+)\right\}$$

is closed. Then, $M$ is closed and $K = \mathrm{Rec}(M)$.

**Proof.** Let $M^+$ satisfy the premise of the proposition. Define

$$W := \left\{x \in \mathbf{R}^n_x : \ \exists u \in \mathbf{R}^k_u : \ [x;u] \in V\right\},$$

that is, $W$ is the projection of $V$ onto the $x$-space. As $V$ is a closed and bounded (therefore compact) set, its projection $W$ is compact as well (recall that the image of a compact set under a continuous mapping is compact). Note that $M$ is nonempty and it satisfies $M = W + K$ (why?). Then, $M$ is the sum of a compact set $W$ and a closed set $K$, and thus $M$ is closed itself (why?). Besides this, $M$ is convex (recall that the projection of a convex set is convex). Thus, the nonempty closed convex set $M$ satisfies $M = W + K$ with nonempty bounded $W$ and closed cone $K$, implying by Theorem II.8.16 that $K = \mathrm{Rec}(M)$. □

Recall that we have investigated the relation between the recessive directions of a closed convex set $M \in \mathbf{R}^n_x$ and its closed convex representation $M^+ \in \mathbf{R}^n_x \times \mathbf{R}^k_u$ in Proposition II.8.17. In particular, we observed that while $[h_x; h_u] \in \mathrm{Rec}(M^+)$ implies $h_x \in \mathrm{Rec}(M)$, the recessive direction of $M$ "stemming" from those of $M^+$ can form a small part of $\mathrm{Rec}(M)$, as seen in Example II.8.4.

A surprising (and not completely trivial) fact is that *for polyhedral sets $M$, the projection of* $\mathrm{Rec}(M^+)$ *onto the $x$-plane is* $\mathrm{Rec}(M)$.

**Proposition** II.8.19 Let $M \in \mathbf{R}^n_x$ be a nonempty set admitting a polyhedral representation $M^+ \in \mathbf{R}^n_x \times \mathbf{R}^k_u$, i.e.,

$$\begin{aligned} M^+ &:= \left\{[x;u] \in \mathbf{R}^n_x \times \mathbf{R}^k_u :\ Ax + Bu \le c\right\}, \quad \text{and} \\ M &:= \left\{x \in \mathbf{R}^n_x :\ \exists u \in \mathbf{R}^k_u :\ [x;u] \in M^+\right\}. \end{aligned}$$

Then,

$$\begin{aligned} \mathrm{Rec}(M) &= \left\{h_x :\ \exists h_u :\ [h_x; h_u] \in \mathrm{Rec}(M^+)\right\} \\ &= \left\{h_x :\ \exists h_u :\ Ah_x + Bh_u \le 0\right\}. \end{aligned} \tag{8.5}$$

That is, polyhedral representation of $M$ naturally induces a polyhedral representation of $\mathrm{Rec}(M)$.

Proposition II.8.19 is an immediate consequence of Proposition II.8.18. To derive Proposition II.8.19 from Proposition II.8.18, it suffices to note that a nonempty polyhedral set is the sum of the convex hull of a finite set and a polyhedral cone, see section 10.1.

## 8.4 Dual cone

We start with the definition of dual cone.

**Definition** II.8.20 [Dual cone] Let $M \subseteq \mathbf{R}^n$ be a cone. The set of all vectors which have nonnegative inner products with all vectors from $M$, i.e., the set

$$M_* := \left\{a \in \mathbf{R}^n :\ a^\top x \ge 0,\ \forall x \in M\right\}, \tag{8.6}$$

is called the cone *dual* to $M$.

From its definition, it is clear that $M_*$ is a closed cone.

**Example** II.8.5 The cone dual to the nonnegative orthant $\mathbf{R}^n_+$ is composed of all $n$-dimensional vectors $y$ making nonnegative inner products with all entrywise nonnegative $n$-dimensional vectors $x$. As is immediately seen the vectors $y$ with this property are exactly entrywise nonnegative vectors: $[\mathbf{R}^n_+]_* = \mathbf{R}^n_+$.

Note that in the preceding example, $\mathbf{R}^n_+$ is given by finitely many homogeneous linear inequalities:

$$\mathbf{R}^n_+ = \left\{x \in \mathbf{R}^n : \; e_i^\top x \geq 0, \, i = 1, \ldots, n\right\},$$

where $e_i$ are the basic orths; and we observe that the dual cone is the conic hull of these basic orth. This is indeed a special case of the following general fact:

**Proposition** II.8.21 For any $F \subseteq \mathbf{R}^n$, the set

$$M := \left\{x \in \mathbf{R}^n : \; f^\top x \geq 0, \, \forall f \in F\right\}$$

is a closed cone, and its dual cone is

$$M_* = \operatorname{cl} \operatorname{Cone}(F),$$

where $\operatorname{Cone}(F)$, as always, is the conic hull of $F$, see Definition I.1.19. In addition, $M$ remains intact when $F$ is extended to its closed conic hull:

$$M := \left\{x \in \mathbf{R}^n : \; f^\top x \geq 0, \, \forall f \in F\right\} = \left\{x \in \mathbf{R}^n : \; f^\top x \geq 0, \, \forall f \in \operatorname{cl} \operatorname{Cone}(F)\right\}.$$

**Proof.** The inclusion $\operatorname{Cone}(F) \subseteq M_*$ is evident, and since $M_*$ is closed, we have also $\operatorname{cl} \operatorname{Cone}(F) \subseteq M_*$. Let us define $\overline{F} := \operatorname{cl} \operatorname{Cone}(F)$, so now we need to prove the inclusion $\overline{F} \supseteq M_*$. Consider $z \in M_*$, and assume for contradiction that $z \notin \overline{F}$. Note that $\overline{F}$ is convex, nonempty, and closed, so that by Separation Theorem (ii) there exists $g$ such that

$$g^\top z < \inf_{f \in \overline{F}} g^\top f.$$

Because $\overline{F}$ is a closed cone (and so $0 \in \overline{F}$), the right hand side infimum, being finite, must be 0. Then, $g^\top f \geq 0$ for all $f \in \overline{F}$ and $g^\top z < 0$. Since $f^\top g \geq 0$ for all $f \in \overline{F}$ and also $\overline{F} \supseteq F$, we deduce $f^\top g \geq 0$ for all $f \in F$, that is, $g \in M$ by the definition of $M$. But, then the inclusion $g \in M$ together with $z \in M_*$ contradicts the relation $z^\top g < 0$. Finally, we clearly have $f^\top x \geq 0$ for all $x \in F$ if and only if $f^\top x \geq 0$ for all $x \in \operatorname{cl} \operatorname{Cone}(F)$. □

**Remark** II.8.22 Note that, in contrast to Proposition II.8.21, in the concluding expression of the chain

$$[\mathbf{R}^n_+]_* = \left\{x \in \mathbf{R}^n : \; e_i^\top x \geq 0, \, i = 1, \ldots, n\right\}_* = \operatorname{Cone}(\{e_i : \, i = 1, \ldots, n\})$$

we did *not* need to take the closure. This is because the conic hull of a *finite* set $F$ is polyhedrally representable and is therefore a polyhedral cone (by Theorem I.3.2), and as such it is automatically closed.

This fact (i.e., no need to take the closure in Proposition II.8.21) holds true for the dual of *any* polyhedral cone: consider the set $\{x \in \mathbf{R}^n : \; a_i^\top x \geq 0, \; i =$

$1,\dots,I\}$ given by finitely many homogeneous nonstrict linear inequalities. This set is clearly a polyhedral cone, and its dual is the conic hull of $a_i$'s, i.e., $\mathrm{Cone}(\{a_i : i = 1,\dots,I\}) = \left\{\sum_{i=1}^I \lambda_i a_i : \lambda \geq 0\right\}$, and this dual cone clearly is also polyhedrally representable as

$$\mathrm{Cone}\,\{a_1,\dots,a_I\} = \left\{x \in \mathbf{R}^n : \exists \lambda \geq 0 : x = \sum_{i=1}^I \lambda_i a_i\right\},$$

and thus $\mathrm{Cone}\,\{a_1,\dots,a_I\}$ is polyhedral as well.

In the case of the cones of the form $\left\{x \in \mathbf{R}^n : f^\top x \geq 0, \forall f \in F\right\}$ stemming from infinite sets $F$ (in fact, *every* closed cone in $\mathbf{R}^n$ can be represented in this way using a properly selected countable set $F = \{f_i : i = 1, 2, \dots\}$) (why?), the closure operation in the computation of the dual cone, in general, cannot be omitted. This is so even when the set $F$ itself is closed convex and bounded (why? Hint: recall Proposition II.8.21).

Let us illustrate this in the next example.

**Example** II.8.6 Consider the cone given by $K := \left\{x \in \mathbf{R}^2 : f^\top x \geq 0, \forall f \in F\right\}$, where $F = \left\{f = [u; v] \in \mathbf{R}^2_+ : v \geq u^2, u \leq 1, v \leq 1\right\}$. Note that $F$ is a compact convex set contained in $\mathbf{R}^2_+$. Moreover, every vector $[u; v] \in \mathbf{R}^2$ with positive entries is a positive multiple of a vector from $F$ (draw a picture!). Thus, the set of vectors that have nonnegative inner products with all vectors from $F$, i.e., $K$, is exactly the same as the set of vectors that have nonnegative inner products with all vectors from $\mathbf{R}^2_+$. Hence, we arrive at $K = \left\{x \in \mathbf{R}^2 : f^\top x \geq 0, \forall f \in F\right\} = \mathbf{R}^2_+$, so $K_* = \mathbf{R}^2_+$ as well. Now observe that $K_* = \mathbf{R}^2_+$ is, as it should be by Proposition II.8.21, the *closure* of $\mathrm{Cone}(F)$, nevertheless $K_* = \mathrm{cl}\,\mathrm{Cone}(F)$ is larger than $\mathrm{Cone}(F)$ as $\mathrm{Cone}(F)$ is not closed! Note that $\mathrm{Cone}(F)$ is precisely the set obtained from $\mathbf{R}^2_+$ by eliminating all nonzero points on the boundary of $\mathbf{R}^2_+$.

**Fact** II.8.23 Let $M$ be a closed cone in $\mathbf{R}^n$, and let $M_*$ be the cone dual to $M$. Then

(i) Duality does not distinguish between a cone and its closure: whenever $M = \mathrm{cl}\,M'$ for a cone $M'$, we have $M_* = M'_*$.

(ii) Duality is symmetric: the cone dual to $M_*$ is $M$.

(iii) One has

$$\mathrm{int}\,M_* = \left\{y \in \mathbf{R}^n : y^\top x > 0, \forall x \in M \setminus \{0\}\right\},$$

and $\mathrm{int}\,M_*$ is nonempty if and only if $M$ is pointed (i.e., $M \cap [-M] = \{0\}$).

Moreover, when $M$, in addition to being closed, is pointed and nontrivial ($M \neq \{0\}$), one has

$$\mathrm{int}\,M_* = \left\{y \in \mathbf{R}^n : M_y := \{x \in M : x^\top y = 1\} \text{ is nonempty and compact}\right\}. \quad (8.7)$$

(iv) The cone dual to the direct product $M_1 \times \ldots \times M_m$ of cones $M_i$ is the direct product of their duals: $[M_1 \times \ldots \times M_m]_* = [M_1]_* \times \ldots \times [M_m]_*$.

Let us see some examples of dual cones.

**Example** II.8.7 Consider the epigraph of $\|\cdot\|_\infty$ on $\mathbf{R}^n$ given by

$$K_\infty := \left\{[x;t] \in \mathbf{R}^{n+1} : \; t \geq \|x\|_\infty\right\}.$$

Note that $K_\infty$ is a polyhedral cone (why?). The cone dual to $K_\infty$ is $K_1 = \{[x;t] \in \mathbf{R}^{n+1} : \; t \geq \|x\|_1\}$, which is nothing but the epigraph of $\|\cdot\|_1$:

$$\begin{aligned}
[K_\infty]_* &= \left\{[g;s] \in \mathbf{R}^{n+1} : \; st + g^\top x \geq 0, \; \forall([x;t] : \|x\|_\infty \leq t)\right\} \\
&= \left\{[g;s] \in \mathbf{R}^{n+1} : \; s + \min_{x:\|x\|_\infty \leq 1} g^\top x \geq 0\right\} \\
&= \left\{[g;s] \in \mathbf{R}^{n+1} : \; s \geq \|g\|_1\right\}.
\end{aligned}$$

**Definition** II.8.24 [Self-dual cone] A cone $\mathbf{K} \subset \mathbf{R}^n$ is called a *self-dual cone* if its dual cone is equal to itself, i.e., $\mathbf{K}_* = \mathbf{K}$.

We next examine a number of very important self-dual cones.

**Example** II.8.8 Let us compute the dual of the Lorentz cone

$$\mathbf{L}^n := \left\{[x;t] \in \mathbf{R}^{n-1} \times \mathbf{R} : \; t \geq \|x\|_2\right\}.$$

When $n = 1$, $\mathbf{L}^1$ is the nonnegative ray, and thus $\mathbf{L}^1 = \mathbf{R}_+$ and therefore $[\mathbf{L}^1]_* = \mathbf{R}_+ = \mathbf{L}^1$. When $n \geq 2$, we have

$$\begin{aligned}
[\mathbf{L}^n]_* &= \left\{[g;s] \in \mathbf{R}^{n-1} \times \mathbf{R} : \; g^\top x + ts \geq 0, \; \forall([x;t] : \|x\|_2 \leq t)\right\} \\
&= \left\{[g;s] \in \mathbf{R}^{n-1} \times \mathbf{R} : \; g^\top x + s \geq 0, \; \forall(x : \|x\|_2 \leq 1)\right\} \\
&= \left\{[g;s] \in \mathbf{R}^{n-1} \times \mathbf{R} : \; s + \min_{x:\|x\|_2 \leq 1} g^\top x \geq 0\right\} \\
&= \left\{[g;s] \in \mathbf{R}^{n-1} \times \mathbf{R} : \; s \geq \|g\|_2\right\},
\end{aligned}$$

where the concluding equality is due to Cauchy-Schwarz inequality, see Theorem B.1. Thus, $[\mathbf{L}^n]_* = \mathbf{L}^n$.

**Example** II.8.9 The cone dual to the semidefinite cone $\mathbf{S}^n_+$, by Theorem D.32, is itself:

$$[\mathbf{S}^n_+]_* := \left\{y \in \mathbf{S}^n : \; \langle y, x\rangle := \mathrm{Tr}(xy) \geq 0, \forall x \in \mathbf{S}^n_+\right\} = \mathbf{S}^n_+.$$

Based on these examples, we have arrived at the following conclusion.

*The cones $\mathbf{R}_+$, $\mathbf{L}^n$, and $\mathbf{S}^n_+$ are self-dual.*

By Fact II.8.23 the direct product of finitely many self-dual cones is self-dual, implying that *finite direct products of nonnegative orthants, Lorentz, and semidefinite cones are self-dual.*

## 8.5 ★ Dubovitski-Milutin Lemma

In this section, we deal with the following basic yet important question: Let $M^1, \ldots, M^k$ be cones (not necessarily closed) in $\mathbf{R}^n$, and $M$ be their intersection. Of course, $M$ also is a cone. But, how can we compute $M_*$ , i.e., the cone dual to $M$? To this end, we first examine the relationship between $M_*$ and the cone $\widetilde{M}$ that is defined as the sum of the dual cones $M^i_*$.

**Proposition** II.8.25 Let $M^1, \ldots, M^k$ be cones in $\mathbf{R}^n$. Define $M := \bigcap_{i=1}^k M^i$, and let $M_*$ be the dual cone of $M$. Let $M^i_*$ denote the dual cone of $M^i$, for $i = 1, \ldots, k$, and define $\widetilde{M} := M^1_* + \ldots + M^k_*$. Then, $M_* \supseteq \widetilde{M}$.

Moreover, if all the cones $M^1, \ldots, M^k$ are closed, then $M_* = \operatorname{cl} \widetilde{M}$. In particular, for closed cones $M^1, \ldots, M^k$, $M_* = \widetilde{M}$ holds if and only if $\widetilde{M}$ is closed.

**Proof.** For any $i = 1, \ldots, k$, any $a_i \in M^i_*$ and any $x \in M$, we have $a_i^\top x \geq 0$, and hence $(a_1 + \ldots + a_k)^\top x \geq 0$. Since the latter relation is valid for all $x \in M$, we conclude that $a_1 + \ldots + a_k \in M_*$. Thus, $\widetilde{M} \subseteq M_*$.

Now assume that the cones $M^1, \ldots, M^k$ are closed, and let us define $\widehat{M} := \operatorname{cl} \widetilde{M}$ so that we need to prove $M_* = \widehat{M}$. Recall that we have already seen $\widetilde{M} \subseteq M_*$, and as $M_*$ is closed we deduce $\widehat{M} = \operatorname{cl} \widetilde{M} \subseteq M_*$. Thus, all we need to prove is that if $a \in M_*$, then $a \in \widehat{M}$ as well. Assume for contradiction that there exists $a \in M_* \setminus \widehat{M}$. As $\widetilde{M}$ is clearly a cone, its closure $\widehat{M}$ is a closed cone. Then, as $a \notin \widehat{M}$, by Separation Theorem (ii), $a$ can be strongly separated from $\widehat{M}$ and thus also from $\widetilde{M} \subseteq \widehat{M}$. Therefore, for some $x \neq 0$ we have

$$a^\top x < \inf_{b \in \widetilde{M}} b^\top x = \inf_{a_i \in M^i_*, i=1,\ldots,k} (a_1 + \ldots + a_k)^\top x = \sum_{i=1}^k \inf_{a_i \in M^i_*} a_i^\top x. \tag{8.8}$$

As $a^\top x$ is a finite number, this inequality implies that $\inf\limits_{a_i \in M^i_*} a_i^\top x > -\infty$ holds for all $i = 1, \ldots, k$. Since $M^i_*$ is a cone, this is possible if and only if $\inf\limits_{a_i \in M^i_*} a_i^\top x = 0$. Then, we deduce that $x \in [M^i_*]_* = M^i$, where the last equality follows from the fact that each cone $M^i$ is closed and using Fact II.8.23.(ii). Thus, $x \in M^i$ for all $i$, and $\sum\limits_{i=1}^k \inf\limits_{a_i \in M^i_*} a_i^\top x = 0$. Therefore, $x \in M = \bigcap\limits_{i=1}^k M^i$, and (8.8) reduces to $a^\top x < 0$. But, this then contradicts the inclusion $a \in M_*$. □

**Remark** II.8.26 Note that in general $\widetilde{M}$ can be non-closed even when all the cones $M^1, \ldots, M^k$ are closed. Indeed, take $k = 2$, and let $M^1 = M^1_*$ be the second-order cone $\left\{(x, y, z) \in \mathbf{R}^3 : \ z \geq \sqrt{x^2 + y^2}\right\}$, and $M^2_*$ be the following ray in $\mathbf{R}^3$

$$\left\{(x, y, z) \in \mathbf{R}^3 : \ x = z, \ y = 0, \ x \leq 0\right\}.$$

Observe that the points from $\widetilde{M} \equiv M^1_* + M^2_*$ are exactly the points of the form $(x-t, y, z-t)$ with $t \geq 0$ and $z \geq \sqrt{x^2 + y^2}$. In particular, for any $\alpha > 0$, the points

$(0,1,\sqrt{\alpha^2+1}-\alpha) = (\alpha-\alpha, 1, \sqrt{\alpha^2+1}-\alpha)$ belong to $\widetilde{M}$. As $\alpha \to \infty$, these points converge to $\xi := (0,1,0)$, and thus $\xi \in \operatorname{cl} \widetilde{M}$. On the other hand, we clearly cannot find $x, y, z, t$ with $t \geq 0$ and $z \geq \sqrt{x^2+y^2}$ such that $(x-t, y, z-t) = (0,1,0)$, that is, $\xi \notin \widetilde{M}$.

Dubovitski-Milutin Lemma presents a simple sufficient condition for $\widetilde{M}$ to be closed and thus to coincide with $M_*$:

**Proposition** II.8.27 [Dubovitski-Milutin Lemma in finite dimensions] Let $M^1, \ldots, M^k$ be cones such that

$$M^k \cap \operatorname{int} M^1 \cap \operatorname{int} M^2 \cap \ldots \cap \operatorname{int} M^{k-1} \neq \varnothing.$$

Define $M := \bigcap\limits_{i=1}^{k} M^i$. Let also $M_*^i$ be the cones dual to $M^i$. Then,

(i) $\operatorname{cl} M = \bigcap\limits_{i=1}^{k} \operatorname{cl} M^i$; and

(ii) the cone $\widetilde{M} := M_*^1 + \ldots + M_*^k$ is closed, and thus by Proposition II.8.25 $M_* = M_*^1 + \ldots + M_*^k$. In other words, every linear form which is nonnegative on $M$ can be represented as a sum of $k$ linear forms which are nonnegative on the respective cones $M^1, \ldots, M^k$.

**Proof.** (i): This is given by Proposition I.1.33(ii).

(ii): First, we claim that under the premise of the proposition, without loss of generality we can assume that $M^1, \ldots, M^k$ are closed cones. This is because when replacing the cones $M^1, \ldots, M^k$ with their closures, we preserve the premise of the proposition, and also $\widetilde{M} = M_*^1 + \ldots + M_*^k = [\operatorname{cl} M^1]_* + \ldots + [\operatorname{cl} M^k]_*$ holds (recall that by definition of the dual cone $M_*^i = [\operatorname{cl} M^i]_*$), as well as $M_* = [\operatorname{cl} M]_* = \bigcap\limits_{i=1}^{k} \operatorname{cl} M^i$ where the last equality holds by Part (i).

To prove Part (ii) of the proposition all we need is to show that given closed cones $M^1, \ldots, M^k$ the cone $\widetilde{M} := M_*^1 + \ldots + M_*^k$ is closed. To this end, we will use induction on $k \geq 2$.

*Base case:* Suppose $k = 2$. Consider a sequence $\{f_t + g_t\}_{t=1}^{\infty}$ with $f_t \in M_*^1$, $g_t \in M_*^2$ and $(f_t + g_t) \to h$ as $t \to \infty$. We need to prove that $h = f + g$ for some appropriate $f \in M_*^1$ and $g \in M_*^2$. To this end, it suffices to verify that for an appropriate subsequence $t_j$ of indices there exists $f := \lim\limits_{j\to\infty} f_{t_j}$. Indeed, if this is the case, then $g = \lim\limits_{j\to\infty} g_{t_j}$ also exists since $f_t + g_t \to h$ as $t \to \infty$ and $f + g = h$, and also in this case we will have $f \in M_*^1$ and $g \in M_*^2$ (recall that $M_*^1$ and $M_*^2$ are closed cones). Let us verify the existence of the desired subsequence. Assume for contradiction that $\|f_t\|_2 \to \infty$ as $t \to \infty$. Passing to a subsequence, we may assume that the unit vectors $\phi_t := f_t/\|f_t\|_2$ have a limit $\phi$ as $t \to \infty$. Since $M_*^1$ is a closed cone, $\phi$ is a unit vector from $M_*^1$. Now, since $f_t + g_t \to h$ as $t \to \infty$, we have $\phi = \lim\limits_{t\to\infty} f_t/\|f_t\|_2 = -\lim\limits_{t\to\infty} g_t/\|f_t\|_2$ (recall that $\|f_t\|_2 \to \infty$ as $t \to \infty$, whence $h/\|f_t\|_2 \to 0$ as $t \to \infty$). Then, the vector $(-\phi) \in M_*^2$ as well (recall that

$M^2_*$ is a closed cone). Now, consider any $\bar{x} \in M^2 \cap \operatorname{int} M^1$ (by the premise of the proposition this set is non-empty). We have $\phi^\top \bar{x} \geq 0$ (since $\bar{x} \in M^1$ and $\phi \in M^1_*$) and $\phi^\top \bar{x} \leq 0$ (since $-\phi \in M^2_*$ and $\bar{x} \in M^2$). We conclude that $\phi^\top \bar{x} = 0$, which contradicts the facts that $0 \neq \phi$ (as $\|\phi\|_2 = 1$), $\phi \in M^1_*$ and $\bar{x} \in \operatorname{int} M^1$ (see Fact II.8.23.(iii)).

*Inductive step:* Assume that the statement is valid for $k-1 \geq 2$ cones, and let $M^1$,...,$M^k$ be $k$ cones satisfying the premise of the proposition. By this premise, the cone $M_1 := M^1 \cap \ldots \cap M^{k-1}$ has a nonempty interior, and $M^k$ intersects this interior. Applying to the pair of cones $M_1, M^k$ the already proved 2-cone version of the lemma, we see that the set $[M_1]_* + M^k_*$ is closed; here $[M_1]_*$ is the cone dual to $M_1$. Moreover, the cones $M^1, \ldots, M^{k-1}$ satisfy the premise of the $(k-1)$-cone version of the lemma. Then, by inductive hypothesis, the set $M^1_* + \ldots + M^{k-1}_*$ is closed. Then, as $M_1 := M^1 \cap \ldots \cap M^{k-1}$, Proposition II.8.25 implies that $[M_1]_* = M^1_* + \ldots + M^{k-1}_*$, and so $M^1_* + \ldots + M^k_* = [M_1]_* + M^k_*$. As $[M_1]_* + M^k_*$ is closed, we deduce that $M^1_* + \ldots + M^k_*$ is closed, as desired. This concludes the induction step. □

**Alternative to proof to Proposition II.8.27.** Here, we present an alternative proof of Proposition II.8.27 Part (ii) without relying on induction.

Let us start with the following fact that is important by its own right.

**Fact** II.8.28 Let $M \subseteq \mathbf{R}^n$ be a cone and $M_*$ be its dual cone. Then, for any $x \in \operatorname{int} M$, there exists a properly selected $C_x < \infty$ such that

$$\|f\|_2 \leq C_x f^\top x, \qquad \forall f \in M_*.$$

Now, as explained in the beginning of Part (ii) of the above proof of Proposition II.8.27, we can assume without loss of generality that the cones $M^1, \ldots, M^k$ satisfying the premise of the proposition are closed, and all we need to prove is that the cone $M^1_* + \ldots + M^k_*$ is closed. The latter is the same as to verify that whenever vectors $f^i_t \in M^i_*$, $i \leq k$, $t = 1, 2, \ldots$ are such that $f_t := \sum_{i=1}^k f^i_t \to h$ as $i \to \infty$, it holds $h \in M^1_* + \ldots + M^k_*$. Indeed, in the situation in question, selecting $\bar{x} \in M^k \cap \operatorname{int} M^1 \cap \ldots \cap \operatorname{int} M^{k-1}$ (by the promise this intersection is nonempty!) we have $\bar{x}^\top f^i_t \geq 0$ for all $i, t$ and $\sum_{i=1}^k \bar{x}^\top f^i_t \to \bar{x}^\top h$ as $t \to \infty$, implying that for all $i \leq k$ the sequences $\{\bar{x}^\top f^i_t\}_{t=1,2,\ldots}$ are bounded. Moreover, for any $i < k$ we have $\bar{x} \in \operatorname{int} M^i$ and $f^i_t \in M^i_*$, and so Fact II.8.28 guarantees that the sequence $\{f^i_t\}_{t=1,2,\ldots}$ is bounded. Thus, as the sequences $\{f^i_t\}_{t=1,2,\ldots}$ are bounded for any $i < k$ and the sequence $\sum_{i=1}^k f^i_t$ has a limit as $t \to \infty$, we conclude that the sequence $\{f^k_t\}_{t=1,2,\ldots}$ is bounded as well. Hence, all $k$ sequences $\{f^i_t\}_{t=1,2,\ldots}$ are bounded, so that passing to a subsequence $t_1 < t_2 < \ldots$ we can assume that $f^i := \lim_{j\to\infty} f^i_{t_j}$ is well defined for every $i \leq k$. Since $f^i_t \in M^i_*$ and the cones $M^i_*$ are closed, we have $f^i \in M^i_*$ for all $i \leq k$. Finally, as $h = \lim_{t\to\infty} \sum_i f^i_t = \lim_{j\to\infty} \sum_i f^i_{t_j} = \sum_i f^i$, we conclude that $h \in M^1_* + \ldots + M^k_*$, as claimed. □

## 8.6 Extreme rays and conic Krein-Milman Theorem

The story about extreme points of closed convex sets has a conic analogy, with nontrivial closed *pointed* cones playing the role of nonempty closed convex sets that do not contain straight lines and *extreme rays* of these cones in the role of extreme points.

Once again, in order to define extreme rays of nontrivial closed cones, we need to work with cones that do not contain straight lines. In the case of closed cones, this notion of not containing straight lines has a special name, i.e., pointed.

**Definition** II.8.29 [Pointed cone] A closed cone $M \subseteq \mathbf{R}^n$ is called *pointed* if $M \cap [-M] = \{0\}$, i.e., the zero vector is the only vector $x$ that satisfies $x \in M$ and $-x \in M$.

**Remark** II.8.30 Note that a *closed* cone $M \subseteq \mathbf{R}^n$ is pointed if and only if $M$ does not contain a straight line passing through the origin. Invoking Lemma II.8.8, we see that a closed cone is pointed if and only if it does not contain straight lines.

In our discussion, we will focus on *nontrivial* cones $M$, i.e., $M \neq \{0\}$. For nontrivial closed pointed cones, let us formally introduce the definition of extreme directions and extreme rays.

**Definition** II.8.31 [Extreme directions and extreme rays] Let $M \subset \mathbf{R}^n$ be a nontrivial closed pointed cone. A direction $d \in \mathbf{R}^n$ is called an *extreme direction* of $M$ if it possesses the following two properties:

- $d \in M \setminus \{0\}$, and
- in every representation of $d$ as the sum of two vectors from $M$, i.e., $d = d^1 + d^2$ with $d^1, d^2 \in M$, both $d^1$ and $d^2$ are nonnegative multiples of $d$.

It is clear that when $d$ is an extreme direction of $M$, so are all positive multiples of $d$, i.e., all nonzero vectors on the ray $\mathbf{R}_+(d)$ generated by $d$ are also extreme directions of $M$. A ray generated by an extreme direction of $M$ is called an *extreme ray* of $M$.
The set of all extreme directions and extreme rays of $M$ are denoted by $\mathrm{ExtD}(M)$ and $\mathrm{ExtR}(M)$, respectively.

**Example** II.8.10 The simplest example of nontrivial closed and pointed cone is the nonnegative orthant $\mathbf{R}^n_+$. Based on our extreme direction definition, the extreme directions of $\mathbf{R}^n_+$ should be the nonzero $n$-dimensional entrywise nonnegative vectors $d$ such that whenever $d = d^1 + d^2$ with $d^1 \geq 0$ and $d^2 \geq 0$, both $d^1$ and $d^2$ must be nonnegative multiples of $d$. Such a vector $d$ has at all entries nonnegative and at least one of them positive. In fact, the number of positive entries in $d$ is exactly one, since if there were at least two entries, say, $d_1$ and $d_2$, positive, we would have $d = [d_1; 0; \ldots; 0] + [0; d_2; d_3; \ldots; d_n]$ and both of the vectors in the right hand side would be nonzero and not proportional to $d$. Thus, any extreme direction of $\mathbf{R}^n_+$ must be a positive multiple of a basic orth. It is immediately seen that every vector of the latter type is an extreme direction of $\mathbf{R}^n_+$. Hence, the extreme directions of $\mathbf{R}^n_+$ are positive multiples of the basic orths, and the extreme rays of $\mathbf{R}^n_+$ are the nonnegative parts of the coordinate axes.

We next introduce the concept of a *base* which an important type of the cross section of a nontrivial closed pointed cone. Moreover, we will see that a base will be a compact convex set and we will establish a direct connection between the extreme rays of the underlying cone and extreme points of its base.

**Definition** II.8.32 [Base of a cone] Let $M \subseteq \mathbf{R}^n$ be a nontrivial closed pointed cone, and $M_*$ be its dual cone. A set of the form

$$B := \left\{x \in M : \ f^\top x = 1\right\} \tag{8.9}$$

is called a *base* of $M$ if this set intersects every emanating from the origin nontrivial ray in $M$. Equivalently, a set $B$ of the form (8.9) is a base of $M$ if for every $x \in M \setminus \{0\}$ there exists $t > 0$ such that $f^\top(tx) = 1$, or equivalently, if $f^\top x > 0$ whenever $x \in \ M \setminus \{0\}$.

**Fact** II.8.33 Let $M \subseteq \mathbf{R}^n$ be a nontrivial closed cone, and $M_*$ be its dual cone.
(i) $M$ is pointed
(i.1) if and only if $M$ does not contain straight lines,
(i.2) if and only if $M_*$ has a nonempty interior, and
(i.3) if and only if $M$ has a base.
(ii) Set (8.9) is a base of $M$
(ii.1) if and only if $f^\top x > 0$ for all $x \in M \setminus \{0\}$,
(ii.2) if and only if $f \in \operatorname{int} M_*$.
In particular, $f \in \operatorname{int} M_*$ if and only if $f^\top x > 0$ whenever $x \in M \setminus \{0\}$.
(iii) Every base of $M$ is nonempty, closed, and bounded. Moreover, whenever $M$ is pointed, for any $f \in M_*$ such that the set (8.9) is nonempty (note that this set is always closed for any $f$), this set is bounded if and only if $f \in \operatorname{int} M_*$, in which case (8.9) is a base of $M$.
(iv) $M$ has extreme rays if and only if $M$ is pointed. Furthermore, when $M$ is pointed, there is one-to-one correspondence between extreme rays of $M$ and extreme points of a base $B$ of $M$, specifically, the ray $R := \mathbf{R}_+(d)$, $d \in M \setminus \{0\}$ is extreme if and only if $R \cap B$ is an extreme point of $B$.

See Figure II.2 for an illustration of Fact II.8.33(iv).

In Example II.8.10 we have observed that every vector from $\mathbf{R}^n_+$ is the sum of finitely many extreme directions of $\mathbf{R}^n_+$. This observation is indeed the special case of the following result.

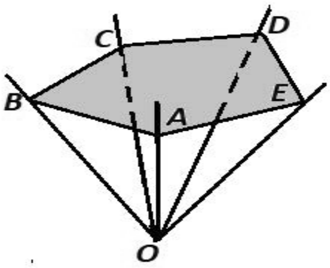


Figure II.2. Cone and its base (grey pentagon). Extreme rays of the cone are $OA$, $OB$,…,$OE$ intersecting the base at its extreme points $A, B, \ldots, E$.

**Theorem** II.8.34 [Krein-Milman Theorem, conic form] Let $M \subset \mathbf{R}^n$ be a nontrivial closed and pointed cone. Then, its set of extreme directions, $\mathrm{ExtD}(M)$, is nonempty, and every vector $d \in M$ is the sum of *finitely* many extreme directions of $M$.

**Proof.** Under the premise of the theorem, Fact II.8.33 implies that $M$ has a base $B$ which is a nonempty convex compact set. Then, by Krein-Milman Theorem (Theorem II.8.6, $B = \mathrm{Conv}(\mathrm{Ext}(B))$. Invoking Fact II.8.33(iv), we conclude that $M$ has extreme rays; each extreme ray of $M$ is generated by precisely one extreme point of $B$. Since vectors from $M$ are exactly the nonnegative multiples of vectors from $B$ and each point in $B$ is the convex combination of finitely many points from $\mathrm{Ext}(B)$ (by Caratheodory's Theorem, see Theorem I.2.3), we conclude that every point of $M$ belongs to the sum of finitely many properly selected extreme rays of $M$. □

Krein-Milman Theorem in conic form states that a nontrivial closed pointed cone has extreme rays – even enough of them to make their arithmetic sum the entire cone. Recall that when a cone $M$ is trivial, we have $M = \{0\}$ and thus $M$ does not contain nonzero vectors and therefore has no extreme directions (the latter, by definition, are nonzero vectors from the cone satisfying certain additional requirements). Then, by Fact II.8.33(iv) and Krein-Milman Theorem, we arrive at the following complete characterization of when a closed cone $M$ possesses extreme rays.

**Corollary** II.8.35 A closed cone $M \subseteq \mathbf{R}^n$ possesses extreme rays *if and only if* $M$ is nontrivial and pointed.

## 8.7 ★ Polar of a convex set

We next study the *polars* of convex sets, a concept closely related to the duals of cones.

**Definition** II.8.36 [Polar of a convex set] For any nonempty convex set $M \subseteq \mathbf{R}^n$, we define its *polar* [notation: Polar $(M)$] to be the set of all vectors $a \in \mathbf{R}^n$ such that $a^\top x \leq 1$ for all $x \in M$, i.e.,

$$\text{Polar}\,(M) := \left\{a \in \mathbf{R}^n :\ a^\top x \leq 1,\ \forall x \in M\right\}.$$

Let us see some basic examples.

**Example** II.8.11

1. Polar $(\mathbf{R}^n) = \{0\}$.
2. Polar $(\{0\}) = \mathbf{R}^n$.
3. Given a linear subspace in $L \subseteq \mathbf{R}^n$, we have Polar $(L) = L^\perp$ (why?).
4. Let $B$ be the unit Euclidean ball, i.e., $B := \{x \in \mathbf{R}^n : \|x\|_2 \leq 1\}$. Then, Polar $(B) = B$ (by Cauchy-Schwarz inequality).
5. Let $X \subset \mathbf{R}^n$ be nonempty and $D$ be a nonsingular $n \times n$ matrix. Then, Polar $(DX) = D^{-\top}$Polar $(X)$.
6. Let $E$ be an $n$-dimensional ellipsoid centered at the origin, i.e., $E := \{x : x^\top Cx \leq 1\}$ where $C \succ 0$. Then, Polar $(E) = \{x : x^\top C^{-1}x \leq 1\}$, i.e., its polar is another $n$-dimensional ellipsoid centered at the origin.
7. Finally, note that passing to polars reverts inclusions: when $\varnothing \neq X \subset Y \subset \mathbf{R}^n$, we have Polar $(Y) \subset$ Polar $(X)$.

For any nonempty convex set $M$, the following properties of its polar are evident:

1. $0 \in$ Polar $(M)$;
2. Polar $(M)$ is convex;
3. Polar $(M)$ is closed.

It turns out that these properties characterize polars completely:

**Proposition** II.8.37 Every closed convex set $M$ containing the origin is a polar set. Specifically, such a set is the polar of its polar:

$$M \text{ is closed, convex, and } 0 \in M \quad\Longleftrightarrow\quad M = \text{Polar}\,(\text{Polar}\,(M)).$$

**Proof.** Based on the evident properties of polars, all we need is to prove that if $M$ is closed and convex and $0 \in M$, then $M =$ Polar (Polar $(M)$). By definition, for all $x \in M$ and $y \in$ Polar $(M)$, we have

$$y^\top x \leq 1.$$

Thus, $M \subseteq$ Polar (Polar $(M)$).

To prove that this inclusion is in fact equality, we assume for contradiction that there exists $\bar{x} \in$ Polar (Polar $(M)) \setminus M$. Since $M$ is a nonempty closed convex set and $\bar{x} \notin M$, the point $\bar{x}$ can be strongly separated from $M$ (Separation Theorem (ii)). Thus, there exists $b \in \mathbf{R}^n$ such that

$$b^\top \bar{x} > \sup_{x \in M} b^\top x.$$

As $0 \in M$, we deduce $b^\top \bar{x} > 0$. Passing from $b$ to a proportional vector $a = \lambda b$ with appropriately chosen positive $\lambda$, we may ensure that

$$a^\top \bar{x} > 1 \geq \sup_{x \in M} a^\top x.$$

From the relation $1 \geq \sup\limits_{x \in M} a^\top x$ we conclude that $a \in \operatorname{Polar}(M)$. But, then the relation $a^\top \bar{x} > 1$ contradicts the assumption that $\bar{x} \in \operatorname{Polar}(\operatorname{Polar}(M))$. Hence, we conclude that indeed $M = \operatorname{Polar}(\operatorname{Polar}(M))$. □

We close this section with a few important properties of the polars.

**Fact** II.8.38 Let $M \subseteq \mathbf{R}^n$ be a convex set containing the origin. Then,

(i) $\operatorname{Polar}(M) = \operatorname{Polar}(\operatorname{cl} M)$;
(ii) $M$ is bounded if and only if $0 \in \operatorname{int}(\operatorname{Polar}(M))$;
(iii) $\operatorname{int}(\operatorname{Polar}(M)) \neq \varnothing$ if and only if $M$ does not contain straight lines;
(iv) Assume that $M$ is closed. Then $M$ is a closed cone if and only if $\operatorname{Polar}(M)$ is a closed cone;
(v) If $M$ is a cone (not necessarily closed), then

$$\operatorname{Polar}(M) = \left\{a \in \mathbf{R}^n : \; a^\top x \leq 0, \, \forall x \in M\right\} = -M_*. \tag{8.10}$$

For more information on polars, see Exercise II.38.

# 9

# Polyhedral sets

## 9.1 Extreme points of polyhedral sets

Consider a polyhedral set

$$M = \{x \in \mathbf{R}^n : \; Ax \leq b\},$$

where $A$ is an $m \times n$ matrix and $b \in \mathbf{R}^m$. We have seen a geometric characterization of extreme points for general convex sets in section 8.2.1. In the case of polyhedral sets $M$, we can also give an *algebraic* characterization of the extreme points as follows.

**Theorem** II.9.1 [Characterization of extreme points of polyhedral sets] Let $M = \{x \in \mathbf{R}^n : \; Ax \leq b\}$. A point $x \in M$ is an extreme point of $M$ if and only if *there are $n$ linearly independent* (i.e., with linearly independent vectors of coefficients) *inequalities* of the system $Ax \leq b$ that are *active* (i.e., hold as equalities) at $x$.

**Proof.** Let $a_i^\top$, $i = 1, \dots, m$, be the rows of $A$.

The "only if" part: let $x$ be an extreme point of $M$, and define the sets $I := \left\{i : \; a_i^\top x = b_i\right\}$ as the set of indices of active constraints and $F := \{a_i : \; i \in I\}$ as the set of vectors of active constraints. We will prove that the set $F$ contains $n$ linearly independent vectors, i.e., $\mathrm{Lin}(F) = \mathbf{R}^n$. Assume for contradiction that this is not the case. Then, as $\dim(F^\perp) = n - \dim(\mathrm{Lin}(F))$, we deduce $\dim(F^\perp) > 0$ and so there exists a nonzero vector $d \in F^\perp$. Consider the segment $\Delta_\epsilon := [x - \epsilon d, x + \epsilon d]$, where $\epsilon > 0$ will be the parameter of our construction. Since $d$ is orthogonal to the "active" vectors $a_i$ (those with $i \in I$), all points $y \in \Delta_\epsilon$ satisfy the relations $a_i^\top y = a_i^\top x = b_i$. Now, if $i$ is a "nonactive" index (one with $a_i^\top x < b_i$), then $a_i^\top y \leq b_i$ for all $y \in \Delta_\epsilon$, provided that $\epsilon$ is small enough. Since there are finitely many nonactive indices, we can choose $\epsilon > 0$ in such a way that all $y \in \Delta_\epsilon$ will satisfy all "nonactive" inequalities $a_i^\top x \leq b_i$, $i \notin I$, as well. So, we conclude that for the above choice of $\epsilon > 0$ we get $\Delta_\epsilon \subseteq M$. But, this is a contradiction to $x$ being an extreme point of $M$ as we have expressed $x$ as the midpoint of a nontrivial segment $\Delta_\epsilon$ (recall that $\epsilon > 0$ and $d \neq 0$).

To prove the "if" part, we assume that $x \in M$ is such that among the inequalities $a_i^\top x \leq b_i$ which are active at $x$ there are $n$ linearly independent ones. Without loss of generality, we assume that the indices of these linearly independent equations are $1, \dots, n$. Given this, we will prove that $x$ is an extreme point of $M$. Assume for contradiction that $x$ is not an extreme point. Then, there exists a vector $d \neq 0$ such that $x \pm d \in M$. In other words, for $i = 1, \ldots, n$ we would have $b_i \geq a_i^\top(x \pm d) \equiv b_i \pm a_i^\top d$ (where the last equivalence follows from $a_i^\top x = b_i$ for all $i \in I = \{1, \ldots, n\}$), which is possible only if $a_i^\top d = 0$, $i = 1, \ldots, n$. But the only vector which is orthogonal to $n$ linearly independent vectors in $\mathbf{R}^n$ is the zero vector (why?), and so we get $d = 0$, which contradicts to the assumption $d \neq 0$. □

Theorem II.9.1 states that at every extreme point of a polyhedral set $M = \{x \in \mathbf{R}^n : Ax \leq b\}$ we must have $n$ linearly independent constraints from $Ax \leq b$ holding as equalities. Since a system of $n$ linearly independent equality constraints in $n$ unknowns has a unique solution, such a system can specify *at most one extreme point* of $M$ (exactly one, when the (unique!) solution to the system satisfies the remaining constraints in the system $Ax \leq b$). Moreover, when $M$ is defined by $m$ inequality constraints, the number of such systems, and thus the number of extreme points of $M$, does not exceed the number $\mathrm{C}_m^n$ of $n \times n$ submatrices of the matrix $A \in \mathbf{R}^{m \times n}$. Hence, we arrive at the following corollary.

**Corollary** II.9.2 Every polyhedral set has finitely many extreme points.

Recall that there are nonempty polyhedral sets which do not have any extreme points; these are precisely the ones that contain lines.

Note that $\mathrm{C}_m^n$ is nothing but an upper (and typically very conservative) bound on the number of extreme points of a polyhedral set in $\mathbf{R}^n$ defined by $m$ inequality constraints. This is because some $n \times n$ submatrices of $A$ can be singular, and what is more important, the majority of the nonsingular ones typically produce "candidate" points which do not satisfy the remaining inequalities defining $M$.

**Remark** II.9.3 Historically, Theorem II.9.1 has been instrumental in developing an algorithm to solve linear programs, namely the *Simplex method.* Let us consider an LP in standard form

$$\min_{x \in \mathbf{R}^n} \left\{c^\top x : Px = p,\, x \geq 0\right\},$$

where $P \in \mathbf{R}^{k \times n}$. Note that we can convert any given LP to this form by adding a small number of new variables and constraints if needed. In the context of this LP, Theorem II.9.1 states that the extreme points of the feasible set are exactly the *basic feasible solutions* of the system $Px = p$, i.e., nonnegative vectors $x$ such that $Px = p$ and the set of columns of $P$ associated with positive entries of $x$ is linearly independent. As the feasible set of an LP in standard form clearly does not contain lines (note the constraints $x \geq 0$ which restricts the standard form LP domain to be subset of the pointed cone $\mathbf{R}_+^n$), among its optimal solutions (if they exist) at least one is an extreme point of the feasible set (Theorem II.10.3(ii)). This then suggests a simple algorithm to solve a solvable LP in standard form: go through the finite set of all extreme points of the feasible set (or equivalently all basic feasible solutions) and choose the one with the best objective value. This algorithm allows to find an optimal solution in finitely many arithmetic operations, provided that the LP is solvable, and underlies the basic idea for the Simplex method. As one will immediately recognize, the number of extreme points, although finite, may be quite large. The Simplex method operates in a smarter way and examines *only a subset* of the basic feasible solutions in an organized way and can handle other issues such as infeasibility and unboundedness.

Another useful consequence of Theorem II.9.1 is that if all the data in an LP are rational, then every one of its extreme points is a vector with rational entries. Thus, a solvable standard form LP with rational data has at least one rational optimal solution.

Theorem II.9.1 has further important consequences in terms of sizes of extreme points of polyhedral sets as well.

To this end, let us first recall a simple fact from Linear Algebra:

**Proposition** II.9.4 If a system of linear equations $Ax = b$ (with $A \in \mathbf{R}^{m\times n}$) is feasible, then it has a solution $x(b)$ of "magnitude of order of the magnitude of $b$;" that is, $\|x(b)\|_2 \leq C(A)\|b\|_2$ with parameter $C(A) < \infty$ which depends solely on $A$ and but not on $b$.

**Proof.** Let $r := \mathrm{rank}(A)$. There is nothing to prove when $r = 0$ as in this case $A$ is zero, and if the system $Ax = b$ has a solution, the vector zero is one of its solutions as well. When $r > 0$, we can assume without loss of generality that the first $r$ columns of $A$ are linearly independent, and the remaining columns are linear combinations of these $r$ columns. Then, if there is a solution to $Ax = b$, then there must be a solution where $x_i = 0$ for all $i > r$. We will take such a solution as $x(b)$. As the first $r$ columns of $A$ are linearly independent, $A$ has an $r \times r$ submatrix $\widehat{A}$ composed of the first $r$ columns of $A$ and $r$ properly selected rows of $A$. Let $\widehat{b}$ be the subvector of $b$ corresponding to the indices of rows selected for $\widehat{A}$. Define the vector $\widehat{x}(b) \in \mathbf{R}^r$ be the vector obtained from the first $r$ entries in $x(b)$. Since the entries in $x(b)$ with indexes greater than $r$ are zero, we have $x(b) = [\widehat{x}(b); 0]$ and so $\widehat{A}\widehat{x}(b) = \widehat{b}$. As $\widehat{A}$ is a nonsingular matrix, we deduce that

$$\|x(b)\|_2 = \|\widehat{x}(b)\|_2 = \|\widehat{A}^{-1}\widehat{b}\|_2 \leq \|\widehat{A}^{-1}\|\|\widehat{b}\|_2 \leq \|\widehat{A}^{-1}\|\|b\|_2,$$

where $\|\widehat{A}^{-1}\|$ is the spectral norm of $\widehat{A}^{-1}$. Then, setting $C(A) = \|\widehat{A}^{-1}\|$ concludes the proof. □

Surprisingly, a similar result holds for the solutions of systems of linear inequalities as well.

**Proposition** II.9.5 Consider a system of linear inequalities $Ax \leq b$ where $A \in \mathbf{R}^{m\times n}$. Whenever $b$ results in a feasible system $Ax \leq b$, then there exists $C(A) < \infty$ depending solely on $A$, but not on $b$, such that this system has a solution $x(b)$ with $\|x(b)\|_2 \leq C(A)\|b\|_2$.

**Proof.** This proof is quite similar to the one for Proposition II.9.4. Let $r := \mathrm{rank}(A)$. The case of $r = 0$ is trivial – in this case $A = 0$, and the system $Ax \leq b$ is feasible, it has the solution $x = 0$. When $r > 0$, we can assume without loss of generality that the first $r$ columns in $A$ are linearly independent. Let $\widehat{A} \in \mathbf{R}^{m\times r}$ be the submatrix of $A$ obtained from the first $r$ columns of $A$. As $\widehat{A}$ has all the linearly independent columns of $A$, the image spaces of $A$ and $\widehat{A}$ are the same. Thus, the system $Ax \leq b$ is feasible if and only if the system $\widehat{A}u \leq b$ is feasible. Moreover, given any feasible solution $u \in \mathbf{R}^r$ to $\widehat{A}u \leq b$, we can generate a feasible solution $x := [u; 0] \in \mathbf{R}^n$ by adding $n - r$ zeros at the end and still preserve the norm of the solution. Hence, without loss of generality we can assume that the columns of $A$ are linearly independent and $r = n$.

As $A \in \mathbf{R}^{m\times n}$ has $n$ linearly independent columns and each column is a vector in $\mathbf{R}^m$, we deduce that $m \geq n$ and $\{u : Au = 0\} = \{0\}$. Thus, we conclude that the polyhedral set $\{x : Ax \leq b\}$ does not contain lines. Therefore, when nonempty, by Krein-Milman Theorem this polyhedral set has an extreme point. Let us take this point as $x(b)$. Then, by Theorem II.9.1, at least $n$ of the inequality constraints from the system $Ax \leq b$ will be active at $x(b)$ and out of these active constraints there will be $n$ vectors $a_i$ (corresponding to rows of the matrix $A$) that are linearly independent. That is, $A_b x(b) = b$ holds for a certain nonsingular $n\times n$ submatrix $A_b$ of $A$. So, we conclude $\|x(b)\|_2 \leq \|A_b^{-1}\|\|b\|_2$. Since the number of $r\times r$ nonsingular submatrices in $A$ is finite, the maximum $C(A)$ of the spectral norms of the inverses of these submatrices is finite as well, and, as we have seen, for every $b$ for which the system $Ax \leq b$ is feasible, it has a solution $x(b)$ with $\|x(b)\|_2 \leq C(A)\|b\|_2$, as claimed. □

## 9.2 Extreme rays of polyhedral cones

Recall that for nontrivial closed pointed cones, we have defined the concepts of extreme directions and extreme rays in section 8.6. In the case of polyhedral cones, we can also give an algebraic characterization of its extreme directions analogous to Theorem II.9.1.

**Theorem** II.9.6 [Characterization of extreme directions of polyhedral cones] Consider a polyhedral cone $M = \left\{d \in \mathbf{R}^n : \; a_i^\top d \leq 0, \, i = 1, \ldots, m\right\}$. Suppose that $M$ is nontrivial and pointed. A direction $d \in M \setminus \{0\}$ is an extreme direction of $M$ if and only if there are $n - 1$ linearly independent (i.e., with linearly independent vectors $a_i$) constraints which are active at $d$ (i.e., such that $a_i^\top d = 0$).

**Proof.** As $M$ is a nonempty closed (recall that it is polyhedral!) pointed cone, from Fact II.8.23(iii), we deduce that its dual cone $M_*$ has a nonempty interior. Let $f \in \operatorname{int} M_*$. Consider the set

$$B := M \cap \left\{d \in \mathbf{R}^n : \; f^\top d = 1\right\}.$$

Then, by Fact II.8.33(ii), $B$ is a base of $M$. Thus, $B$ is nonempty and compact (see Fact II.8.33(iii)). As $M$ is polyhedral, by definition of $B$, we have $B$ is polyhedral as well. Since $B$ is a nonempty bounded polyhedral set, from Theorem II.8.6(ii) we have $B = \text{Conv}(\text{Ext}(B))$. Recall from Fact II.8.33(iv) that there is a one-to-one correspondence between extreme rays of $M$ and extreme points of a base $B$ of $M$; specifically, the ray $R := \mathbf{R}_+(d)$, $d \in M \setminus \{0\}$ is extreme if and only if $R \cap B$ is an extreme point of $B$.

Consider an extreme point $x_d$ of $B$. Applying Theorem II.9.1, we deduce that at every extreme point of $B$ we must have at least $n$ active constraints from the set of linear (in)equalities $f^\top x_d = 1$ and $a_i^\top x_d \leq 0$, $i = 1, \ldots, m$, where the corresponding vectors of active constraints must contain $n$ linearly independent vectors. Considering the definition of $B$, at every extreme point $x_d$ of $B$, we have the constraint $f^\top x_d = 1$ is active. Let $I$ be the set of indices $i \in \{1, \ldots, m\}$ such that $a_i^\top x_d = 0$. Then, the vectors generating the active constraints at $x_d$ are given by $\{f\} \cup \{a_i : i \in I\}$. Since among the active constraints at $x_d$, there are $n$ of them with linearly independent vectors, we conclude that there must be $n-1$ linearly independent vectors in $\{a_i : i \in I\}$. Then, as $x_d$ and $d$ generate the same extreme direction of $M$, we conclude that $d$ satisfies the desired algebraic characterization.

To establish the reverse direction, suppose that a direction $d \in M \setminus \{0\}$ satisfies $a_i^\top d = 0$ for $i \in I \subseteq \{1, \ldots, m\}$ where there are $n-1$ linearly independent vectors in $\{a_i : i \in I\}$. Let $x_d := \mathbf{R}_+(d) \cap B$. We claim that $x_d \in \text{Ext}(B)$. Note that $x_d$ satisfies $a_i^\top x_d = 0$ for $i \in I$ and $f^\top x_d = 1$. If $f \in \text{Lin}(\{a_i : i \in I\})$, then from the constraints $a_i^\top x_d = 0$ for all $i \in I$ we would have deduced $f^\top x_d = 0$, which is not the case. So, the set $\{f\} \cup \{a_i : i \in I\}$ must have $n$ linearly independent vectors. Then, by Theorem II.9.1 $x_d$ must be an extreme point of $B$. Once again, using Fact II.8.33(iv) we conclude that $d$ must be an extreme direction of $M$. □

Analogous to Corollary II.9.2, we have the following immediate corollary of this theorem.

**Corollary** II.9.7 Every nontrivial pointed polyhedral cone $M$ has finitely many extreme rays. Moreover, the sum of these extreme rays is the entire $M$.

**Proof.** Let $M$ be a nontrivial pointed polyhedral cone. As any polyhedral cone is defined by finitely many linear inequalities, using the algebraic characterization of the extreme directions given in Theorem II.9.6, we deduce that $M$ has finitely many extreme rays. The last claim of the corollary is justified by Theorem II.8.34, i.e., Krein-Milman Theorem in conic form, as this theorem states that $M$ has extreme rays, and their sum is the entire $M$. □

## 9.3 Important polyhedral sets and their extreme points

In this section, we will examine a number of important polyhedral sets and their extreme points. Throughout this section, we suppose that $k$ and $n$ are positive integers with $k \leq n$.

**Example** II.9.1 Suppose $k, n$ are integers satisfying $0 \leq k \leq n$. Consider the polytope

$$X := \left\{ x \in \mathbf{R}^n : \ 0 \leq x_i \leq 1, \, \forall i \leq n, \ \sum_{i=1}^n x_i = k \right\}.$$

The set of extreme points of $X$ is precisely the set of vectors with entries 0 and 1 which have exactly $k$ entries equal to 1. That is,

$$\operatorname{Ext}(X) = \left\{ x \in \{0,1\}^n : \ \sum_{i=1}^n x_i = k \right\}.$$

In particular, the extreme points of the "flat (a.k.a. *probabilistic*) simplex" $\{x \in \mathbf{R}^n_+ : \ \sum_{i=1}^n x_i = 1\}$ are the basic orths (see the Figure II.3 for an illustration of this set with $k = 1$).

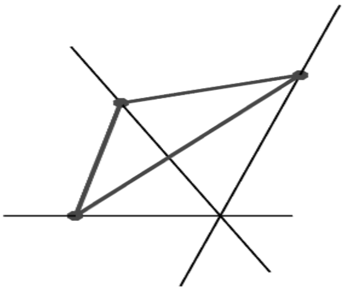


Figure II.3. Example II.9.1, $n = 3, k = 1$.

Let us justify the claim of this example. For convenience, we define $Y := \{x \in \{0,1\}^n : \ \sum_{i=1}^n x_i = k\}$; then we need to show that $\operatorname{Ext}(X) = Y$. Clearly, $Y \subseteq X$. Moreover, for any $y \in Y$, for each coordinate $i = 1, \ldots, n$ we have either $y_i = 0$ or $y_i = 1$, thus we have $n$ bound constraints active. Since the vectors of coefficients of these constraints provide us $n$ linearly independent vectors, we conclude by Theorem II.9.1 that $Y \subseteq \operatorname{Ext}(X)$. Now, consider any $w \in \operatorname{Ext}(X)$. Then, by Theorem II.9.1 among the constraints specifying $X$, $n$ constraints with linearly independent vectors of coefficients should be active at $w$. Thus, we must have at least $n - 1$ of the bound constraints $0 \leq x_i \leq 1$ active at $w$, i.e., at least $n - 1$ of the entries of $w$ must be $\{0, 1\}$. Let $i_*$ be the index of the remaining entry of $w$. As $w \in X$, it must also satisfy $\sum_{i=1}^n w_i = k$. As $k$ is an integer, we deduce $w_{i_*} = k - \sum_{i \neq i_*} w_i$ must be an integer as well. But, then as we also have $0 \leq w_{i_*} \leq 1$, we deduce that $w_{i_*} \in \{0, 1\}$. Thus, $w \in Y$ holds, as desired.

**Example** II.9.2 Suppose $k, n$ are integers satisfying $0 \leq k \leq n$. Consider the polytope

$$X = \left\{ x \in \mathbf{R}^n : \ 0 \leq x_i \leq 1, \, \forall i \leq n, \ \sum_{i=1}^n x_i \leq k \right\}.$$

The set of extreme points of $X$ is precisely the set of vectors with entries 0 and 1 which have at most $k$ entries equal to 1. That is,

$$\mathrm{Ext}(X) = \left\{ x \in \{0,1\}^n : \sum_{i=1}^n x_i \leq k \right\}.$$

In particular, the extreme points of the "full-dimensional simplex" $\{x \in \mathbf{R}^n_+ : \sum_{i=1}^n x_i \leq 1\}$ are the basic orths and the origin (see Figure II.4 for an illustration of this set with $k = 1$).

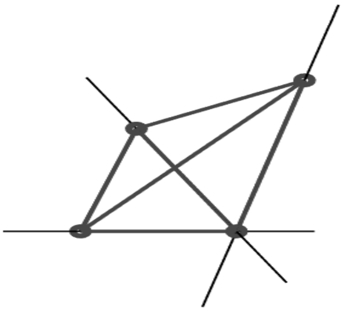


Figure II.4. Example II.9.2, $n = 3, k = 1$

Justification of this example follows the one of Example II.9.1 and is left as an exercise to the reader.

**Example** II.9.3 Suppose $k, n$ are integers satisfying $0 \leq k \leq n$. Consider the polytope

$$X = \left\{ x \in \mathbf{R}^n : |x_i| \leq 1, \forall i \leq n, \sum\nolimits_{i=1}^n |x_i| \leq k \right\}.$$

Extreme points of $X$ are exactly the vectors with entries $0, 1, -1$ which have exactly $k$ nonzero entries. That is,

$$\mathrm{Ext}(X) = \left\{ x \in \{-1,0,1\}^n : \sum_{i=1}^n |x_i| = k \right\}.$$

In particular, extreme points of the unit $\|\cdot\|_1$-ball $\{x \in \mathbf{R}^n : \|x\|_1 \leq 1\} = \{x \in \mathbf{R}^n : \sum_{i=1}^n |x_i| \leq 1\}$ are exactly the vectors $\{\pm e_i : i = 1, \ldots, n\}$ where $e_i$ is the $i$-th basic orth (see Figure II.5 for an illustration of this set with $k = 1$).

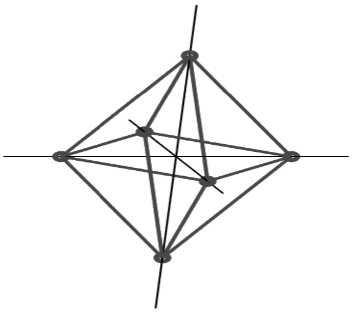


Figure II.5. Example II.9.3, $n = 3, k = 1$.

Similarly, extreme points of the unit $\|\cdot\|_\infty$-ball $\{x \in \mathbf{R}^n : \|x\|_\infty \le 1\} = \{x \in \mathbf{R}^n : |x_i| \le 1, \forall i = 1, \dots, n\}$ are the $2^n$ vectors with ±1 entries (see Figure II.6 for an illustration of this set).

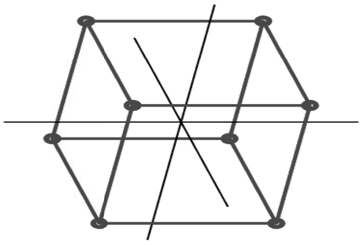


Figure II.6. Extreme points of the box $\{x \in \mathbf{R}^3 : \|x\|_\infty \le 1\}$.

Here is the justification of the claim of this example. For convenience, we define $Y := \{x \in \{-1, 0, 1\}^n : \sum_{i=1}^n |x_i| = k\}$; we need to show that $\mathrm{Ext}(X) = Y$. Clearly, $Y \subseteq X$. Consider any $y \in Y$. Without loss of generality, suppose that the nonzero entries of $y$ are the first $k$. Now, consider any $h$ such that $y \pm h \in X$. Since $y_i \in \{-1, +1\}$ for $i = 1, \dots, k$ and $y \pm h \in X$, we must have $h_i = 0$ for all $i = 1, \dots, k$. Also, $y + h \in X$ implies that $k \ge \sum_{i=1}^n |y_i + h_i| = \sum_{i=1}^k |y_i| + \sum_{i=k+1}^n |h_i| = k + \sum_{i=k+1}^n |h_i|$, and thus $|h_i| = 0$ for all $i = k+1, \dots, n$. This proves that $h = 0$ and thus $y$ must indeed be an extreme point of $X$. So, $Y \subseteq \mathrm{Ext}(X)$. Now, consider any $w \in \mathrm{Ext}(X)$, we will show that $w \in Y$. Note that $X$ has certain symmetry: for any $x \in X$ and any $d \in \{-1, +1\}^n$, we have $\mathrm{Diag}(d)x \in X$ as well. In particular, any $d \in \{-1, +1\}^n$ maps $X$ onto itself and therefore maps its extreme points $\mathrm{Ext}(X)$ onto $\mathrm{Ext}(X)$. As a result, we can assume without loss of generality that $w \ge 0$. Then, all we need to prove is that $w$ has $k$ entries equal to 1 and all remaining entries equal to 0. The set $X_+ := \{x \in X : x \ge 0\} = \{x \in \mathbf{R}^n : 0 \le x_i \le 1, \forall i \le n, \sum_{i=1}^k x_i \le k\}$ is contained in $X$ and $w \in X_+$. As $w \in \mathrm{Ext}(X)$ and $w \in X_+ \subset X$, we must have that $w$ is an extreme point of $X_+$ as well.

In the preceding reasoning, we have used the following evident fact: *if $P \subset Q$ are convex sets and $\bar{x} \in P$ is an extreme point of $Q$, then it is an extreme point of $P$ as well* (otherwise $\bar{x}$ would be the midpoint of a nontrivial segment contained in $P$ and therefore contained in $Q$).

Then, noting that $X_+$ is precisely the set from Example II.9.2 and $w \in \mathrm{Ext}(X_+)$, we conclude that $w$ has only 0 and 1 entries with at most $k$ entries equal to $k$. It remains to verify that the number of nonzero entries in $w$ is equal to $k$. Indeed, $w$ were to have fewer than $k$ nonzero entries, $w$ would have a zero entry, say, $w_1 = 0$, and $\sum_{i=1}^n |w_i| < k$, implying that there exists $\epsilon \in (0, 1)$ such that the vector $h = [\epsilon; 0; \dots; 0]$ will satisfy $(w \pm h) \in X$, which is impossible since $w \in \mathrm{Ext}(X)$.

As our last example we next discuss the so-called *Assignment polytope*, which is closely connected to the very important concept of *doubly stochastic matrices* and the Birkhoff Theorem.

**Definition** II.9.8 [Doubly stochastic matrix] A matrix $X = [x_{ij}]_{i,j=1}^n \in \mathbf{R}^{n\times n}$ is called *doubly stochastic*, if $x_{ij} \geq 0$ for all $i, j \in \{1,\dots,n\}$, $\sum_{i=1}^n x_{ij} = 1$ for all $j \in \{1,\dots,n\}$ (i.e., all column sums are equal to 1), and $\sum_{j=1}^n x_{ij} = 1$ for all $i \in \{1,\dots,n\}$ (i.e., all row sums are equal to 1).

The set of all doubly stochastic matrices (treated as elements of $\mathbf{R}^{n^2} = \mathbf{R}^{n\times n}$) form the following bounded polyhedral set:

$$\Pi_n := \left\{ X = [x_{ij}]_{i,j=1}^n : \begin{array}{c} x_{ij} \geq 0,\ \forall i,j \in \{1,\dots,n\}, \\ \sum_{i=1}^n x_{ij} = 1,\ \forall j \in \{1,\dots,n\}, \\ \sum_{j=1}^n x_{ij} = 1,\ \forall i \in \{1,\dots,n\} \end{array} \right\}.$$

The set $\Pi_n$ is also called the *Assignment (or balanced matching) polytope.* As $\Pi_n$ is a polytope, by Krein-Milman Theorem, $\Pi_n$ is the convex hull of its extreme points. What are these extreme points? The answer is given by the following fundamental result.

**Theorem** II.9.9 [Birkhoff–von Neumann Theorem] Extreme points of $\Pi_n$ are exactly the permutation matrices of order $n$, which are $n \times n$ Boolean (i.e., with 0/1 entries) matrices with exactly one nonzero element (equal to 1) in every row and every column.

**Proof.** It is indeed easy to see that every $n \times n$ permutation matrix is an extreme point of $\Pi_n$; we give this as Exercise II.42.

We now prove the difficult part, that is we will show that every extreme point of $\Pi_n$ is a permutation matrix. First, note that the $2n$ linear equations in the definition of $\Pi_n$, those saying that all row and column sums are equal to 1, are linearly dependent (observe that the sum of the first group of equalities is exactly the same as the sum of the second group of equalities). Thus, we lose nothing when assuming that there are just $2n - 1$ equality constraints in the description of $\Pi_n$. Now, let us prove the claim by induction on $n$. The base $n = 1$ is trivial. As the induction hypothesis suppose that the statement holds for $\Pi_{n-1}$. Let $X$ be an extreme point of $\Pi_n$. By Theorem II.9.1, among the constraints defining $\Pi_n$ (i.e., $2n - 1$ equalities and $n^2$ inequalities $x_{ij} \geq 0$) there should be $n^2$ linearly independent constraints which are satisfied at $X$ as equations. Thus, at least $n^2 - (2n-1) = (n-1)^2$ entries in $X$ should be zeros. It follows that at least one of the columns of $X$ contains $\leq 1$ nonzero entries (since otherwise the number of zero entries in $X$ would be at most $n(n-2) < (n-1)^2$). Thus, there exists at least one column with at most 1 nonzero entry; since the sum of entries in this column is 1, this nonzero entry, let it be $x_{\bar{i}\bar{j}}$, is equal to 1. As the entries in row $\bar{i}$ are nonnegative, sum up to 1 and $x_{\bar{i}\bar{j}} = 1$, $x_{\bar{i}\bar{j}} = 1$ is the only nonzero entry in its row and its column. Eliminating from $X$ the row $\bar{i}$ and the column $\bar{j}$, we get an $(n-1) \times (n-1)$ doubly stochastic matrix. By inductive hypothesis, this matrix is a convex combination of $(n-1)\times(n-1)$ permutation matrices. Augmenting every one of these matrices by the column and the row we have eliminated, we get a representation of $X$ as a convex combination of $n\times n$ permutation matrices: $X=\sum_\ell \lambda_\ell P^\ell$ with nonnegative $\lambda_\ell$ summing up to 1. Since $P^\ell\in\Pi_n$ and $X$ is an extreme point of $\Pi_n$, in this representation all terms with nonzero coefficients $\lambda_\ell$ must be equal to $\lambda_\ell X$, so that $X$ is one of the permutation matrices $P^\ell$ and as such is a permutation matrix. □

## 9.4 ★ Majorization

In this section we will introduce and study the *Majorization Principle*, which describes the convex hull of permutations of a given vector.

For any $x\in\mathbf{R}^n$, we define $X[x]$ to be the set of all convex combinations of $n!$ vectors obtained from $x$ by permuting its coordinates. That is,

$$\begin{aligned}X[x]&:=\operatorname{Conv}\left(\{Px:\ P \text{ is an } n\times n \text{ permutation matrix}\}\right)\\&=\{Dx:\ D\in\Pi_n\},\end{aligned}$$

where $\Pi_n$ is the set of all $n\times n$ doubly stochastic matrices. Here, the equality is due to the Birkhoff-von Neumann Theorem. Note that $X[x]$ is a *permutationally symmetric* set, that is given any vector from the set the vector obtained by permuting its entries is also in the set.

**Theorem** II.9.10 [Majorization Principle] Given two vectors $x,y\in\mathbf{R}^n$, we have $y\in X[x]$ if and only if $y$ satisfies

$$\begin{array}{l}s_j(y)\le s_j(x),\ j=1,\dots,n-1,\\ y_1+\ldots+y_n=x_1+\ldots+x_n,\end{array}\tag{9.1}$$

where $s_j(y)$ is the sum of the $j$ largest entries of the vector $y$.

**Proof:** To ease our notation, let us define the set

$$Y:=\left\{y\in\mathbf{R}^n:\ \begin{array}{l}s_j(y)\le s_j(x),\ j=1,\dots,n-1\\ s_n(y)=y_1+\ldots+y_n=x_1+\ldots+x_n=s_n(x)\end{array}\right\}.$$

Then, we need to show that $Y=X[x]$.

For any $k$, define $\mathcal{I}_k$ to be the family of all $k$-element subsets of the index set $\{1,2\dots,n\}$, and so

$$s_k(y)=\max_{I\in\mathcal{I}_k}\sum_{i\in I}y_i,.\tag{9.2}$$

We first prove that $Y\supseteq X[x]$. Consider any $y\in X[x]$. By the definition of $X[x]$, $y$ can be represented as convex combination

$$y=\sum_\sigma\lambda_\sigma x^\sigma,$$

where the sum is taken over all permutations $\sigma$ of indices $1, 2, \ldots, n$, and $x^\sigma$ is obtained from $x$ by permutation $\sigma$ of the entries, i.e.,

$$x^\sigma := [x_{\sigma(1)}; \ldots; x_{\sigma(n)}].$$

Consequently, for every $I \in \mathcal{I}_k$ we have

$$\sum_{i \in I} y_i = \sum_{i \in I} \sum_\sigma \lambda_\sigma x_{\sigma(i)} = \sum_\sigma \lambda_\sigma \sum_{i \in I} x_{\sigma(i)} \le \sum_{i \in I} \lambda_\sigma s_k(x) = s_k(x),$$

where the inequality is due to (9.2). Maximizing both sides of this inequality over $I \in \mathcal{I}_k$ and invoking (9.2) once again, we get $s_k(y) \le s_k(x)$ for all $k \le n$. In addition,

$$\sum_{i=1}^n y_i = \sum_{i=1}^n \sum_\sigma \lambda_\sigma x_{\sigma(i)} = \sum_\sigma \lambda_\sigma \sum_{i=1}^n x_{\sigma(i)} = \sum_\sigma \lambda_\sigma \sum_{i=1}^n x_i = \sum_{i=1}^n x_i,$$

that is, $s_n(y) = s_n(x)$. Thus, $y \in Y$, whence $Y \supseteq X[x]$.

We will now prove the difficult part of Majorization Principle which states that $Y \subseteq X[x]$. Consider any $y \in Y$ and let us prove that $y \in X[x]$. By symmetry, we may assume without loss of generality that the vectors $x$ and $y$ are ordered: $x_1 \ge x_2 \ge \ldots \ge x_n$ and $y_1 \ge y_2 \ge \ldots \ge y_n$. Assume for contradiction that $y \notin X[x]$. Since $X[x]$ clearly is a convex compact set and $y \notin X[x]$, there exists a linear form $c^\top z$ which strongly separates $y$ and $X[x]$, i.e.,

$$c^\top y > \max_{z \in X[x]} c^\top z.$$

As the set $X[x]$ is permutationally symmetric and the vector $y$ is ordered, without loss of generality we can select the vector $c$ to be ordered as well. This is because when permuting the entries of $c$, we preserve $\max\limits_{z \in X[x]} c^\top z$, and arranging the entries of $c$ in non-increasing order, we do not decrease $c^\top y$: assuming, say, that $c_1 < c_2$, swapping $c_1$ and $c_2$ we do not decrease $c^\top y$: $[c_2 y_1 + c_1 y_2] - [c_1 y_1 + c_2 y_2] = [c_2 - c_1][y_1 - y_2] \ge 0$. Next, by Abel's formula (discrete analogy of integration by parts) we have

$$\begin{aligned} c^\top y = \sum_{i=1}^n c_i y_i &= \sum_{i=1}^{n-1} (c_i - c_{i+1}) \sum_{j=1}^i y_j + c_n \sum_{j=1}^n y_j \\ &= \sum_{i=1}^{n-1} (c_i - c_{i+1}) s_i(y) + c_n s_n(y) \\ &\le \sum_{i=1}^{n-1} (c_i - c_{i+1}) s_i(x) + c_n s_n(x) = \sum_{i=1}^n c_i x_i = c^\top x. \end{aligned}$$

where the inequality follows from the "orderedness" of entries in $c$ and $y \in Y$. Thus, we conclude $c^\top y \le c^\top x$, which is the desired contradiction. □

# 10

# Structure of polyhedral sets

## 10.1 Main result

By definition, a polyhedral set $M$ is the set of all solutions to a finite system of nonstrict linear inequalities:

$$M := \{x \in \mathbf{R}^n : \ Ax \le b\}, \tag{10.1}$$

where $A \in \mathbf{R}^{m \times n}$ and $b \in \mathbf{R}^m$. This is an "outer" description of a polyhedral set, that is, it explains what we should delete from $\mathbf{R}^n$ to get $M$ (cf: "to create a sculpture, take a big stone and delete everything that is redundant"). We are about to establish an important result on the equivalent "inner" representation of a polyhedral set, that is, one explaining how to build the set starting with simple "building blocks."

Consider the following construction. Let us take two finite sets of vectors $V$ ("vertices;" this set should be nonempty) and $R$ ("rays;" this set can be empty) and build the set

$$\begin{aligned} M(V,R) &:= \mathrm{Conv}(V) + \mathrm{Cone}(R) \\ &= \left\{ \sum\nolimits_{v\in V} \lambda_v v + \sum\nolimits_{r \in R} \mu_r r : \begin{array}{l} \lambda_v \ge 0, \forall v \in V, \sum_{v \in V} \lambda_v = 1, \\ \mu_r \ge 0, \forall r \in R \end{array} \right\}. \end{aligned}$$

Thus, in the construction of $M(V,R)$ we take all vectors which can be represented as sums of convex combinations of the points from $V$ and conic combinations of the points from $R$. The set $M(V,R)$ clearly is convex as it is the arithmetic sum of two convex sets $\mathrm{Conv}(V)$ and $\mathrm{Cone}(R)$ (recall that our convention that $\mathrm{Cone}(\varnothing) = \{0\}$, see Fact I.1.20). We are now ready to present the promised inner description of a polyhedral set.

**Theorem** II.10.1 [Inner description of a polyhedral set] The sets of the form $M(V,R)$ are exactly the nonempty polyhedral sets: $M(V,R)$ is polyhedral, and every nonempty polyhedral set $M$ is $M(V,R)$ for properly chosen $V$ and $R$. The sets of the type $M(\{0\},R)$ are exactly the *polyhedral cones* (sets given by finitely many nonstrict homogeneous linear inequalities).

**Remark** II.10.2 We will see in section 10.2.2 that the inner characterization of the polyhedral sets given in Theorem II.10.1 can be made much more precise. Suppose that we are given a nonempty polyhedral set $M$. Then, we can select an inner characterization of it in the form of $M = \mathrm{Conv}(V) + \mathrm{Cone}(R)$ with finite $V$ and finite $R$, where the "conic" part $\mathrm{Cone}(R)$ (not the set $R$ itself!) is uniquely defined by $M$; in fact it will always hold that $\mathrm{Cone}(R) = \mathrm{Rec}(M)$, i.e., $R$ can be taken as the generators of the recessive cone of $M$ (see Comment to Lemma II.8.8). Moreover, if $M$ does not contain lines, then $V$ can be chosen as the set of all extreme points of $M$.

We will prove Theorem II.10.1 in section 10.2. Before proceeding with its proof, let us understand why this theorem is so important, i.e., why it is so nice to know both inner and outer descriptions of a polyhedral set.

Consider the following natural questions:

- A. Is it true that the inverse image of a polyhedral set $M \subset \mathbf{R}^n$ under an affine mapping $y \mapsto \mathcal{P}(y) = Py + p : \mathbf{R}^m \to \mathbf{R}^n$, i.e., the set
$$\mathcal{P}^{-1}(M) = \{y \in \mathbf{R}^m : Py + p \in M\}$$
is polyhedral?
- B. Is it true that the image of a polyhedral set $M \subset \mathbf{R}^n$ under an affine mapping $x \mapsto y = \mathcal{P}(x) = Px + p : \mathbf{R}^n \to \mathbf{R}^m$ – the set
$$\mathcal{P}(M) = \{Px + p : x \in M\}$$
is polyhedral?
- C. Is it true that the intersection of two polyhedral sets is again a polyhedral set?
- D. Is it true that the arithmetic sum of two polyhedral sets is again a polyhedral set?

The answers to all these questions are positive; one way to see it is to use calculus of polyhedral representations along with the fact that polyhedrally representable sets are exactly the same as polyhedral sets (see chapter 3). Another very instructive way is to use the just outlined results on the structure of polyhedral sets, which we will do now.

It is very easy to answer affirmatively to A, starting from the original "outer" definition of a polyhedral set: if $M = \{x : Ax \le b\}$, then, of course,
$$\mathcal{P}^{-1}(M) = \{y : A(Py + p) \le b\} = \{y : (AP)y \le b - Ap\}$$
and therefore $\mathcal{P}^{-1}(M)$ is a polyhedral set.

An attempt to answer affirmatively to B via the same "outer" definition fails – there is no easy way to convert the linear inequalities defining a polyhedral set into those defining its image, and it is absolutely unclear why the image in fact is given by finitely many linear inequalities. Note, however, that there is no difficulty to answer affirmatively to B with the inner description of a nonempty polyhedral set: if $M = M(V, R)$, then, evidently,
$$\mathcal{P}(M) = M(\mathcal{P}(V), PR),$$
where $PR := \{Pr : r \in R\}$ is the image of $R$ under the action of the homogeneous part of $\mathcal{P}$.

A positive answer to C becomes evident when we use the outer description of a polyhedral set: taking intersection of the solution sets to two systems of finitely many nonstrict linear inequalities, we, of course, again get the solution set to a system of this type – you simply should put together all inequalities from the both of the original systems.

On the other hand, it is very unclear how to give the affirmative answer to D using the outer description of a polyhedral set – what happens to the inequalities when we add the solution sets? In contrast to this, the inner description gives the answer immediately:

$$\begin{aligned} M(V,R)+M(V',R') &= \mathrm{Conv}(V)+\mathrm{Cone}(R)+\mathrm{Conv}(V')+\mathrm{Cone}(R') \\ &= [\mathrm{Conv}(V)+\mathrm{Conv}(V')]+[\mathrm{Cone}(R)+\mathrm{Cone}(R')] \\ &= \mathrm{Conv}(V+V')+\mathrm{Cone}(R\cup R') \\ &= M(V+V',R\cup R'). \end{aligned}$$

Note that in this computation we used two rules which should be justified: $\mathrm{Conv}(V)+\mathrm{Conv}(V')=\mathrm{Conv}(V+V')$ and $\mathrm{Cone}(R)+\mathrm{Cone}(R')=\mathrm{Cone}(R\cup R')$. The second is evident from the definition of the conic hull, and the first one follows from a very simple reasoning. To see it, note that $\mathrm{Conv}(V)+\mathrm{Conv}(V')$ is a convex set which by its definition contains $V+V'$ and thus also contains $\mathrm{Conv}(V+V')$. The inverse inclusion is proved as follows: if

$$x=\sum\nolimits_i \lambda_i v_i, \qquad y=\sum\nolimits_j \lambda'_j v'_j$$

are convex combinations of points from $V$, respectively, $V'$, then, as it is immediately seen (please check!),

$$x+y=\sum\nolimits_{i,j}\lambda_i\lambda'_j(v_i+v'_j)$$

and the right hand side expression is nothing but a convex combination of points from $V+V'$.

To conclude, *it is extremely useful to keep in mind both descriptions of polyhedral sets* – what is difficult to see with one of them, is absolutely clear with another.

As a seemingly "more important" application of the developed theory, let us look at Linear Programming.

### 10.1.1 Application: Descriptive theory of Linear Programming

A general linear program is the problem of minimizing a linear objective function over a polyhedral set:

$$\min_x \left\{c^\top x:\ x\in M\right\},\quad \text{where } M:=\{x\in\mathbf{R}^n: Ax\le b\}. \qquad (P)$$

Here, $c\in\mathbf{R}^n$ is the objective, $A\in\mathbf{R}^{m\times n}$ is the constraint matrix, and $b\in\mathbf{R}^m$ is the right hand side vector. Note that (P) is called a "Linear Programming problem in the canonical form;" there are other equivalent forms of this problem as well.

### *10.1.2 Application: Solvability of a Linear Programming problem*

According to the Linear Programming terminology discussed in section 4.5.1, $(P)$ is called

- *feasible*, if it admits a feasible solution, i.e., the system $Ax \leq b$ is feasible, and *infeasible* otherwise;
- *bounded*, if its objective is below bounded on the feasible set (e.g., due to the fact that the latter is empty), and *unbounded* otherwise;
- *solvable*, if it is feasible and the optimal solution exists, i.e., the objective function attains its minimum on the feasible set.

Whenever $(P)$ is feasible, the infimum of the values of the objective function at feasible solutions is called the *optimal value* $\mathrm{Opt}(P)$ of $(P)$. $\mathrm{Opt}(P)$ is finite when the problem $(P)$ is bounded from below and $-\infty$ when $(P)$ is unbounded. In the case of a minimization type problem, it is convenient to assign the optimal value of $+\infty$ whenever the problem is infeasible.

Note that our terminology is aimed to deal with minimization problems; if the problem is to maximize the objective, the terminology is updated in the natural way: when defining bounded/unbounded programs, we should speak about above boundedness rather than about the below boundedness of the objective on the feasible set, etc. As a result, the optimal value of an LP problem

- in the case of a minimization problem is the infimum of the objective over the feasible set, provided the latter is nonempty, and $+\infty$ when the problem is infeasible;
- in the case of a maximization problem is the supremum of the objective over the feasible set, provided the latter is nonempty, and $-\infty$ when the problem is infeasible.

This terminology is consistent with the usual way of converting a minimization problem into an equivalent maximization one by replacing the original objective $c$ with $-c$: the properties of feasibility – boundedness – solvability remain unchanged, and the optimal value in all cases changes its sign.

When talking about the possible outcomes of solving an LP problem, we talk about three possibilities: (i) infeasible LP problem, (ii) unbounded and feasible LP problem, and (iii) "solvable" LP problem. In particular, it seems that in the case of "bounded and feasible" LP problem, we are jumping straight to the conclusion that the corresponding optimization problem will be solvable. This, a bounded LP program is always solvable, indeed is true, although it is absolutely unclear in advance why (note that this statement absolutely does not hold for general convex programming problems without further assumptions). We have already established this fact, even twice — via Fourier-Motzkin elimination (section 3.2 and via the LP Duality Theorem). In fact yet another proof of this fundamental fact of Linear Programming follows immediately from the inner characterization of polyhedral sets as shown next.

**Theorem** II.10.3 Suppose we are given a feasible minimization type Linear Programming problem $(P)$ via an inner representation of its feasible set $M$:

$$M = \mathrm{Conv}(V) + \mathrm{Cone}(R),$$

where $V = \{v_i : i = 1, \ldots, I\}$ and $R = \{r_j : j = 1, \ldots, J\}$ are finite and nonempty sets (cf. Theorem II.10.1). Then,

(i) $(P)$ is solvable if and only if it is bounded, which is the case if and only if $c^\top r_j \geq 0$ for all $1 \leq j \leq J$.
In particular, the set $C$ of objectives $c$ for which $(P)$ is bounded is a polyhedral cone.

(ii) When $(P)$ is bounded, the optimal value of the problem is

$$\mathrm{Opt}(P) = \min_{v \in V} c^\top v,$$

whence $\mathrm{Opt}(P)$ is a concave function of the objective vector $c$, and there is an optimal solution which is the best, in terms of the objective, point of $V$. In addition, when the feasible set of $(P)$ does not contain lines and $(P)$ is bounded, there is at least one optimal solution of $(P)$ which is an extreme point of $M$.

**Proof.** (i): By Theorem II.10.1, we clearly have

$$\mathrm{Opt}(P) = \min_v \left\{c^\top v : v \in \mathrm{Conv}(V)\right\} + \inf_r \left\{c^\top r : r \in \mathrm{Cone}(R)\right\}.$$

We see that $\mathrm{Opt}(P)$ is finite if and only if $\inf_r \left\{c^\top r : r \in \mathrm{Cone}(R)\right\} > -\infty$, and the latter clearly is the case if and only if $c^\top r \geq 0$ for all $r \in R$. Then, in such a case $\inf_r \left\{c^\top r : r \in \mathrm{Cone}(R)\right\} = 0$, and clearly $\min_v \left\{c^\top v : v \in \mathrm{Conv}(V)\right\} = \min_v \left\{c^\top v : v \in V\right\}$.

(ii): The first claim in (ii) is an immediate byproduct of the proof of (i). The second claim follows from the fact that when $M$ does not contain lines, we can take $V = \mathrm{Ext}(M)$, see Remark II.10.2. □

## 10.2 Proof of Inner characterization of polyhedral sets

To simplify language let us call VR-sets ("V" from "vertex", "R" from rays) the sets of the form $M(V, R)$, and P-sets the nonempty polyhedral sets, i.e., defined by finitely many linear inequalities. We should prove that every P-set is a VR-set, and vice versa.

We start with proving that every VR-set is a P-set.

### *10.2.1 VR $\Longrightarrow$ P*

This is immediate: a VR-set is polyhedrally representable (why?) and thus is a P-set by Theorem I.3.2.

To complete the proof of Theorem II.10.1, we now need to show that every P-set is a VR-set.

### *10.2.2 P $\Longrightarrow$ VR*

Let $M$ be a P-set, so that $M$ is the set of all solutions to a feasible system of linear inequalities:

$$M = \{x \in \mathbf{R}^n : \ Ax \leq b\}, \tag{10.2}$$

where $A \in \mathbf{R}^{m \times n}$.

We will first study the case of P-sets that do not contain lines, and then reduce the general case to this one.

**Theorem** II.10.4 [Structure of a polyhedral set with no lines] A nonempty polyhedral set $M = \{x \in \mathbf{R}^n : \ Ax \leq b\}$ which does not contain lines admits a VR-representation given by $M = M(V, R) = \mathrm{Conv}(V) + \mathrm{Cone}(R)$, where $V$ is the set of extreme points of $M$ and $R = \{0\}$ if $M$ is bounded and $R$ is the set of extreme rays of $\mathrm{Rec}(M)$ if $M$ is unbounded.

**Proof.** As $M$ is a nonempty closed convex set that does not contain lines, by Theorem II.8.6(i) we know $\mathrm{Ext}(M) \neq \varnothing$, and by Theorem II.8.16 we have $M = \mathrm{Conv}(\mathrm{Ext}(M)) + \mathrm{Rec}(M)$. Moreover, by Corollary II.9.2, we have $\mathrm{Ext}(M)$ is a finite set.

If $M$ is bounded, then $\mathrm{Rec}(M) = \{0\}$, and thus the result follows. Suppose $M$ is unbounded. Then, $\mathrm{Rec}(M)$ is nontrivial. Also, $\mathrm{Rec}(M)$ is pointed as $M$ does not contain lines $\mathrm{Rec}(M)$ does not contain lines either. Moreover, from Fact II.8.15, we deduce that $\mathrm{Rec}(M) = \{x \in \mathbf{R}^n : \ Ax \leq 0\}$ and thus is a polyhedral cone. Then, by Corollary II.9.7 we have that $\mathrm{Rec}(M)$ has finitely many extreme rays and $\mathrm{Rec}(M)$ is the sum of its extreme rays. □

Next, we study the case when $M$ contains a line. We start with the following observation.

**Lemma** II.10.5 Nonempty polyhedral set $M = \{x \in \mathbf{R}^n : \ Ax \leq b\}$ contains lines if and only if $\mathrm{Ker}A \neq \{0\}$. Moreover, for a vector $h \neq 0$, the set $M$ contains a line with direction $h$ (i.e., $x + th \in M$ for all $x \in M$ and $t \in \mathbf{R}$) if and only if $h \in \mathrm{Ker}A$. That is, the nonzero vectors from $\mathrm{Ker}A$ are exactly the directions of lines contained in $M$.

**Proof.** If $h \neq 0$ is the direction of a line in $M$, then $A(x + th) \leq b$ for some $x \in M$ and all $t \in \mathbf{R}$, which is possible if and only if $Ah = 0$. Vice versa, if $h \neq 0$ is from the kernel of $A$, i.e., if $Ah = 0$, then the line $x + \mathbf{R}(h)$ with $x \in M$ clearly is contained in $M$. □

Given a nonempty set $M$ as in (10.2), define $L := \mathrm{Ker}A$, let $L^\perp$ be the orthogonal complement to $L$, and let $M'$ be the intersection of $M$ and $L^\perp$:

$$M' := \left\{x \in L^\perp : \ Ax \le b\right\}.$$

First, note that the set $M'$ clearly does not contain lines. This is because if $h \neq 0$ is the direction of a line satisfying $x + th \in M'$ for all $t \in \mathbf{R}$ and $x \in M'$, by definition of $M'$ we must have $x+th \in L^\perp$ for all $t$ and thus $h \in L^\perp$. On the other hand, by Lemma II.10.5, we must also have $h \in \mathrm{Ker}A = L$. Then, $h \in L \cap L^\perp$ implies $h = 0$, which is a contradiction.

Now, note that $M' \neq \varnothing$ ,and moreover, $M = M' + L$. Indeed, $M'$ contains the orthogonal projections of all points from $M$ onto $L^\perp$ (since to project a point onto $L^\perp$, you should move from this point along certain line with the direction in $L$, and all these movements, started in $M$, keep you in $M$ by Lemma II.10.5) and therefore is nonempty, first, and is such that $M' + L \supset M$, second. On the other hand, $M' \subset M$ and $M + L = M$ by Lemma II.10.5, hence $M' + L \subset M$. Thus, $M' + L = M$.

Finally, it is clear that $M'$ is a polyhedral set as the inclusion $x \in L^\perp$ can be represented by $\dim(L)$ linear equations (i.e., by $2\dim(L)$ nonstrict linear inequalities). To this end, all we need is a set of vectors $\xi_1, \ldots, \xi_{\dim(L)}$ forming a basis in $L$, and then $L^\perp := \{x \in \mathbf{R}^n : \xi_i^\top x = 0, \forall i = 1, \ldots, \dim(L)\}$.

Therefore, with these steps, given an arbitrary P-set $M$, we have represented is as the sum of a P-set $M'$ not containing lines and a linear subspace $L$. Then, as $M'$ does not contain lines, by Theorem II.10.4, we have $M' = M(V', R')$ where $V'$ is the set of extreme points of $M'$ and $R'$ is the set of extreme rays of $\mathrm{Rec}(M')$. Let us define $R''$ to be the finite set of generators for $L$, i.e., $L = \mathrm{Cone}(R'')$. Then, we arrive at

$$\begin{aligned} M &= M' + L \\ &= [\mathrm{Conv}(V') + \mathrm{Cone}(R')] + \mathrm{Cone}(R'') \\ &= \mathrm{Conv}(V') + [\mathrm{Cone}(R') + \mathrm{Cone}(R'')] \\ &= \mathrm{Conv}(V') + \mathrm{Cone}(R' \cup R'') \\ &= M(V', R' \cup R''). \end{aligned}$$

Thus, this proves that a P-set is indeed a VR-set as desired. □

Finally, let us justify Remark II.10.2. Suppose we are given $M = \mathrm{Conv}(V) + \mathrm{Cone}(R)$ with finite sets $V, R$. Justifying the first claim in this remark requires us to show that $\mathrm{Cone}(R) = \mathrm{Rec}(M)$. To see this, from $M = \mathrm{Conv}(V) + \mathrm{Cone}(R)$ it clearly follows $\mathrm{Cone}(R) \subseteq \mathrm{Rec}(M)$. To prove reverse inclusion, consider any $r \in \mathrm{Rec}(M)$. Then, by definition of the recessive cone, for any $v \in V \subseteq M$, we have $v + tr \in M$ for all $t > 0$. In addition, as $v + tr \in M$ for all $t > 0$, from $M = \mathrm{Conv}(V) + \mathrm{Cone}(R)$ we deduce that

$$\forall t > 0, \ \exists v_t \in \mathrm{Conv}(V) \text{ and } r_t \in \mathrm{Cone}(R) : \ v + tr = v_t + r_t.$$

Since $V$ is a finite set, $\mathrm{Conv}(V)$ is bounded, and so $r = \lim_{t\to\infty} t^{-1} r_t$. Moreover, this limit (and thus $r$) belongs to $\mathrm{Cone}(R)$ as $\mathrm{Cone}(R)$ is polyhedral and therefore closed. The last claim of Remark II.10.2 states that when a P-set $M$ does not contain lines, in every representation $M = M(V, R)$ with finite $V$ and $R$ one has $\mathrm{Ext}(M) \subseteq V$. To see this, suppose $x$ is an extreme point of $M$. We will first show that $x \in \mathrm{Conv}(V)$. Assume for contradiction that $x \in M \setminus \mathrm{Conv}(V)$. Then, from $x \in M = \mathrm{Conv}(V) + \mathrm{Cone}(R)$, we deduce that $x = \bar{x} + r$ with $\bar{x} \in \mathrm{Conv}(V)$ and $0 \neq r \in \mathrm{Cone}(R)$. But, this would imply $\bar{x} = x - r \in M$ and $x + r \in M$ as $x \in M$, $r \in \mathrm{Cone}(R) = \mathrm{Rec}(M)$ and $M$ is convex. Thus, we arrive at $x \pm r \in M$ with $r \neq 0$, contradicting the fact that $x \in \mathrm{Ext}(M)$. Therefore, $x \in \mathrm{Conv}(V)$, and hence $x \in V$ by Fact II.8.5.

# 11

# Exercises for Part II

## 11.1 Separation

**Exercise** II.1 Mark by "Y" those of the below listed cases where the linear form $f^\top x$ separates the sets $S$ and $T$:

- $S = \{0\} \subset \mathbf{R}$, $T = \{0\} \subset \mathbf{R}$, $f^\top x = x$
- $S = \{0\} \subset \mathbf{R}$, $T = [0,1] \subset \mathbf{R}$, $f^\top x = x$
- $S = \{0\} \subset \mathbf{R}$, $T = [-1,1] \subset \mathbf{R}$, $f^\top x = x$
- $S = \{x \in \mathbf{R}^3 : x_1 = x_2 = x_3\}$, $T = \{x \in \mathbf{R}^3 : x_3 \geq \sqrt{x_1^2 + x_2^2}\}$, $f^\top x = x_1 - x_2$
- $S = \{x \in \mathbf{R}^3 : x_1 = x_2 = x_3\}$, $T = \{x \in \mathbf{R}^3 : x_3 \geq \sqrt{x_1^2 + x_2^2}\}$, $f^\top x = x_3 - x_2$ $S = \{x \in \mathbf{R}^3 : -1 \leq x_1 \leq 1\}$, $T = \{x \in \mathbf{R}^3 : x_1^2 \geq 4\}$, $f^\top x = x_1$
- $S = \{x \in \mathbf{R}^2 : x_2 \geq x_1^2, x_1 \geq 0\}$, $T = \{x \in \mathbf{R}^2 : x_2 = 0\}$, $f^\top x = -x_2$

**Exercise** II.2 Consider the set

$$M = \left\{ x \in \mathbf{R}^{2004} : \begin{array}{rcl} x_1 + x_2 + \ldots + x_{2004} & \geq & 1 \\ x_1 + 2x_2 + 3x_3 \ldots + 2004x_{2004} & \geq & 10 \\ x_1 + 2^2x_2 + 3^2x_3 \ldots + 2004^2x_{2004} & \geq & 10^2 \\ \ldots\ldots\ldots\ldots & & \\ x_1 + 2^{2002}x_2 + 3^{2002}x_3 + \ldots + 2004^{2002}x_{2004} & \geq & 10^{2002} \end{array} \right\}$$

Is it possible to separate this set from the set $\{x_1 = x_2 = \ldots = x_{2004} \leq 0\}$? If yes, what could be a separating plane?

**Exercise** II.3 Can the sets $S = \{x \in \mathbf{R}^2 : x_1 > 0, x_2 \geq 1/x_1\}$ and $T = \{x \in \mathbf{R}^2 : x_1 < 0, x_2 \geq -1/x_1\}$ be separated? Can they be strongly separated?

**Exercise** II.4 ♦ Let $M \subset \mathbf{R}^n$ be a nonempty closed convex set. The *metric projection* $\mathrm{Proj}_M(x)$ of a point $x \in \mathbf{R}^n$ onto $M$ is the $\|\cdot\|_2$-closest to $x$ point of $M$, so that

$$\mathrm{Proj}_M(x) \in M \ \& \ \|x - \mathrm{Proj}_M(x)\|_2^2 = \min_{y \in M} \|x - y\|_2^2. \tag{$*$}$$

1. Prove that for every $x \in \mathbf{R}^n$ the minimum in the right hand side of $(*)$ is achieved, and $x_+$ is a minimizer if and only if
$$x_+ \in M \ \& \ \forall y \in M : [x - x_+]^\top [x_+ - y] \geq 0. \tag{11.1}$$
Derive from the latter fact that the minimum in $(*)$ is achieved at a unique point, the bottom line being that $\mathrm{Proj}_M(\cdot)$ is well defined.
2. Prove that when passing from a point $x \in \mathbf{R}^n$ to its metric projection $x_+ = \mathrm{Proj}_M(x)$, the distance to any point of $M$ does not increase, specifically,
$$\begin{array}{c} \forall y \in M : \|x_+ - y\|_2^2 \leq \|x - y\|_2^2 - \mathrm{dist}^2(x, M), \\ \mathrm{dist}(x, M) := \min_{u \in M} \|x - u\|_2 = \|x - x_+\|_2. \end{array} \tag{11.2}$$

3. Let $x \notin M$, so that, denoting $x_+ = \mathrm{Proj}_M(x)$, the vector $e = \frac{x-x_+}{\|x-x_+\|_2}$ is well defined. Prove that the linear form $e^\top z$ strongly separates $x$ and $M$, specifically,
$$\forall y \in M : e^\top y \le e^\top x - \mathrm{dist}(x, M).$$
*Note:* The fact just outlined underlies an alternative proof of Separation Theorem, where the first step is to prove that a point outside a nonempty closed convex set can be strongly separated from the set. In our proof, the first step was similar, but with $M$ restricted to be polyhedral, rather than merely convex and closed.
4. Prove that the mapping $x \mapsto \mathrm{Proj}_M(x) : \mathbf{R}^n \to M$ is *contraction* in $\|\cdot\|_2$:
$$\forall u, u' \in \mathbf{R}^n : \|\mathrm{Proj}_M(u) - \mathrm{Proj}_M(u')\|_2 \le \|u - u'\|_2.$$
5. Let $M$ be the probabilistic simplex: $M = \{x \in \mathbf{R}^n : x \ge 0, \sum_i x_i = 1\}$. Justify the following recipe for computing $\mathrm{Proj}_M(x)$:

   Let $\psi(t) = \sum_{i=1}^m [x_i - t]_+$. where $[s]_+ = \max[s, 0]$. $\psi$ is piecewise linear, with breakpoints $x_1, x_2, \ldots, x_n$, continuous function of $t \in \mathbf{R}$. $\psi(t) \to +\infty$ as $t \to -\infty$, and $\psi(t) \to 0$ as $t \to +\infty$. Consequently, there exists (and can be easily computed due to piecewise linearity of $\psi$) $t \in \mathbf{R}$ such that $\sum_i [x_i - t]_+ = 1$. The metric projection of $x$ onto $M$ is nothing but the vector $x_+$ with coordinates $[x_i - t]_+$, $1 \le i \le n$.

   What is the metric projection of the point $x = [1; 2; 2.5]$ onto the 3-dimensional probabilistic simplex?

**Exercise** II.5 ♦ [Follow-up to Exercise II.4] Let $p(z) = z^n + p_{n-1}z^{n-1} + ... + p_1 z + p_0$, $n \ge 1$ be a polynomial of complex variable $z$. By the Fundamental Theorem of Algebra, $p$ has $n$ roots $\lambda_1, ..., \lambda_n$. Treating complex numbers as 2D real vectors, prove that all roots of the derivative $p'(z) = nz^{n-1} + (n-1)p_{n-1}z^{n-2} + .. + p_1$ belong to the convex hull of $\lambda_1, ..., \lambda_n$.

**Exercise** II.6 ▲ Derive the statement in Remark I.1.4 from the Separation Theorem.

## 11.2 Extreme points

**Exercise** II.7 Find extreme points of the following sets:

1. $X = \{x \in \mathbf{R}^3 : x_1 + x_2 \le 1, x_2 + x_3 \le 1, x_3 + x_1 \le 1\}$
2. $X = \{x \in \mathbf{R}^4 : x_1 + x_2 \le 1, x_2 + x_3 \le 1, x_3 + x_4 \le 1, x_4 + x_1 \le 1\}$

**Exercise** II.8 ♦ Let $M \subset \mathbf{R}^n$ be a nonempty closed convex set not containing lines, and $f^\top x$ be a linear function of $x \in \mathbf{R}^n$ achieving its maximum over $X$. Prove that among maximizers of this function on $M$ there are extreme points of $M$.

**Exercise** II.9 [Follow-up to Exercise I.8] Let $A$, $B$ be subsets of $\mathbf{R}^n$. Mark by **T** those of the below claims which always (i.e., for every data satisfying premise of the claim) are true:

1. If $\mathrm{Conv}(A) = \mathrm{Conv}(B)$ , then $A = B$
2. If $\mathrm{Conv}(A) = \mathrm{Conv}(B)$ is nonempty and $A, B, \mathrm{Conv}(A)$ are closed, then $A \cap B \neq \varnothing$.
3. If $\mathrm{Conv}(A) = \mathrm{Conv}(B)$ is nonempty and bounded, $A \cap B \neq \varnothing$.
4. If $\mathrm{Conv}(A) = \mathrm{Conv}(B)$ is nonempty, closed and bounded, then $A \cap B \neq \varnothing$.

**Exercise** II.10 As is immediately seen, the only extreme point of the nonnegative orthant $\mathbf{R}^n_+ = \mathbf{R}_+ \times \mathbf{R}_+ \times \ldots \times \mathbf{R}_+$ is the origin, that is, the vector from $\{0\} \times \{0\} \times \ldots \times \{0\}$; as we know, the extreme points of $n$-dimensional unit box $\{x \in \mathbf{R}^n : 0 \le x_i \le 1, i \le n\} = [0,1] \times [0,1] \times \ldots \times [0,1]$ are zero/one vectors, that is, vectors from $\{0,1\} \times \{0,1\} \times \ldots \times \{0,1\}$. Prove the following generalization of these observations:

Let $X_i \subset \mathbf{R}^{n_i}$, $1 \le i \le K$, be closed convex sets. The set of extreme points of the direct product $X = X_1 \times \ldots \times X_K$ of these sets is the direct product of the sets of extreme points of $X_i$.

**Exercise** II.11 ♦ Looking at the sets of extreme points of closed convex sets like the unit Euclidean ball, a polytope, the paraboloid $\{[x;t] : t \geq x^\top x\}$, etc., we see that these sets are closed. Do you think this always is the case? Is it true that the set $\mathrm{Ext}(M)$ of extreme points of a closed convex set $M$ always is closed ?

**Exercise** II.12 ▲ Derive representation $(*)$ in Exercise I.29 from Example II.9.1 in section 9.3.

**Exercise** II.13 ♦ By Birkhoff Theorem, the extreme points of the polytope $\Pi_n = \{[x_{ij}] \in \mathbf{R}^{n\times n} : x_{ij} \geq 0, \sum_i x_{ij} = 1\,\forall j, \sum_j x_{ij} = 1\,\forall i\}$ are exactly the Boolean (i.e., with entries 0 and 1) matrices from this set. Prove that the same holds true for the "polytope of sub-doubly stochastic" matrices $\Pi_{m,n} = \{[x_{ij}] \in \mathbf{R}^{m\times n} : x_{ij} \geq 0, \sum_i x_{ij} \leq 1\,\forall j, \sum_j x_{ij} \leq 1\,\forall i\}$.

**Exercise** II.14 ♦ [Follow-up to Exercise II.13] Let $m, n$ be two positive integers with $m \leq n$, and $X_{m,n}$ be the set of $m \times n$ matrices $[x_{ij}]$ with $\sum_i |x_{ij}| \leq 1$ for all $j \leq n$ and $\sum_j |x_{ij}| \leq 1$ for all $i \leq m$. Describe the set $\mathrm{Ext}(X_{m,n})$. To get an educated guess, look at the matrices $\left[\begin{array}{ccc}1&0&0\\0&0&-1\end{array}\right], \left[\begin{array}{ccc}0&0&0\\0&0&-1\end{array}\right], \left[\begin{array}{ccc}0.5&-0.5&0\\-0.5&0.5&0\end{array}\right]$ from $X_{2,3}$.

**Exercise** II.15 ♦ [follow-up to Exercise II.13] Let $x$ be an $n \times n$ entrywise nonnegative matrix with all row and all column sums $\leq 1$. Is it true that for some doubly stochastic matrix $\overline{x}$, the matrix $\overline{x} - x$ is entrywise nonnegative?

**Exercise** II.16 ♦ [Assignment problem] Consider the problem as follows:

*There are $n$ jobs and $n$ workers. When worker $j$ is assigned with job $i$, we get profit $c_{ij}$. We want to assign every worker with a job in such a way that every worker is assigned with exactly one job and every job is assigned to exactly one worker. Under this restriction, we want to maximize the total profit.*

1. Pose the Assignment problem as a Boolean (i.e., with the decision variables restricted to be zeros and ones) Linear Programming problem.
2. Think how to solve the problem from item 1 via plain Linear Programming
3. [computational study] Consider the special case of Assignment problem where all profits $c_{ij}$ are zeros or ones; you can interpret $c_{ij} = 1/0$ as the fact that worker $j$ knows/does not know how to execute job $j$. In this situation Assignment problem requires from us to find an assignment which maximizes the total number of executed jobs. Assume now that the matrix $C = [c_{ij}]$ is generated at random, with entries taking, independently of each other, value 1 with probability $\epsilon \in (0,1)$ and value 0 with probability $1-\epsilon$. For $n \in \{4, 8, 16, 32, 64, 128, 256\}$ and $\epsilon \in \{1/2, 1/4, 1/8, 1/16\}$, run 100 simulations per pair $n, \epsilon$ to find the empirical mean of the ratio "number of executed jobs in optimal assignment"$/n$ and look at the results.

**Exercise** II.17 ▲ Let $\nu = (\nu_1, \ldots, \nu_K)$ with positive integer $\nu_i$, and let $\mathbf{S}^\nu = \mathbf{S}^{\nu_1} \times \ldots \times \mathbf{S}^{\nu_K}$ be the space of block-diagonal, with $K$ diagonal blocks of sizes $\nu_i \times \nu_i$, $i \leq K$, symmetric matrices, let $\mathbf{S}^\nu_+$ be the cone composed of positive semidefinite matrices from $\mathbf{S}^\nu$, and let $E$ be an $m$-dimensional affine plane in $\mathbf{S}^\nu$ which intersects $\mathbf{S}^\nu_+$. The intersection $X = E \cap \mathbf{S}^\nu_+$ is a closed nonempty convex set not containing lines and thus possessing extreme points. Let $W$ be such a point, $W^{ii}$ be the diagonal blocks of $W$, and $r_i$ be the ranks of $\nu_i \times \nu_i$ matrices $W^{ii}$. Prove that

$$\sum_{i=1}^{k} r_i(r_i+1) \leq \sum_{i=1}^{K} \nu_i(\nu_i+1) - 2m.$$

What happens in the diagonal case $\nu_1 = \ldots = \nu_K = 1$ ?

**Exercise** II.18 ♦ Let $M$ be a closed convex set in $\mathbf{R}^n$ and $\bar{x}$ be a point of $M$.

1. Prove that if there exists a linear form $a^\top x$ such that $\bar{x}$ is the *unique* maximizer of the form on $M$, then $\bar{x}$ is an extreme point of $M$.

2. Is the inverse of 1) true, i.e., is it true that every extreme point $\bar{x}$ of $M$ is the unique maximizer, over $x \in M$, of a properly selected linear form?

**Exercise** II.19 Identify and justify the correct claims in the following list:

1. Let $X \subset \mathbf{R}^n$ be a nonempty closed convex set, $P$ be an $m \times n$ matrix, $Y = PX := \{Px : x \in X\} \subset \mathbf{R}^n$, and $\overline{Y}$ be the closure of $Y$. Then
   - For every $x \in \mathrm{Ext}(X)$, $Px \in \mathrm{Ext}(\overline{Y})$
   - Every extreme point of $\overline{Y}$ which happens to belong to $Y$ is $Px$ for some $x \in \mathrm{Ext}(X)$
   - When $X$ does not contain lines, then every extreme point of $\overline{Y}$ which happens to belong to $Y$ is $Px$ for some $x \in \mathrm{Ext}(X)$
2. Let $X, Y$ be nonempty closed convex sets in $\mathbf{R}^n$, and let $Z = X + Y$, $\overline{Z} = \mathrm{cl}\, Z$. Then
   - If $w \in \mathrm{Ext}(\overline{Z}) \cap Z$, then $w = x + y$ for some $x \in \mathrm{Ext}(X)$ and $y \in \mathrm{Ext}(Y)$.
   - If $x \in \mathrm{Ext}(X)$, $y \in \mathrm{Ext}(Y)$, then $x + y \in \mathrm{Ext}(\overline{Z})$.

**Exercise** II.20 ♦ [faces of polyhedral set] Let $X = \{x \in \mathbf{R}^n : a_i^\top x \le b_i, i \le m\}$ be a nonempty polyhedral set and $f^\top x$ be a linear form of $x \in \mathbf{R}^n$ which is bounded above on $X$:

$$\mathrm{Opt}(f) = \sup_{x \in X} f^\top x < \infty$$

Prove that

1. $\mathrm{Opt}(f)$ is achieved – the set $\operatorname*{Argmax}_{x \in X} f^\top x := \{x \in X : f^\top x = \mathrm{Opt}(f)\}$ is nonempty.
2. The set $\operatorname*{Argmax}_{x \in X} f^\top x$ is as follows: there exists an index set $I \subset \{1, 2, \ldots, m\}$, perhaps empty, such that
$$\operatorname*{Argmax}_{x \in X} f^\top x = X_I := \{x : a_i^\top x \le b_i \,\forall i,\; a_i^\top x = b_i \,\forall i \in I\}$$
3. Vice versa, if $I \subset \{1, \ldots, m\}$ is such that the set $X_I = \{x : a_i^\top x \le b_i \,\forall i, a_i^\top x = b_i \,\forall i \in I\}$ is nonempty, then $X_I = X_* := \mathrm{Argmax}_{x \in X} f^\top x$ for properly selected $f$.

   *Note:* Nonempty sets of the form $X_I$, $I \subset \{1, \ldots, m\}$, are called *faces* of the polyhedral set $X$. This definition is not geometric – according to it, whether a given set $Y$ is or is not a face in $X$, may depend not on $X$ *per se*, but on its representation as the solution set of a finite system of linear inequalities. Facts 2—3, taken together, state that in fact being a face of a polyhedral set is a geometric property – faces are exactly the sets $\operatorname*{Argmax}_{x \in X} f^\top x$ of all maximizers of linear forms bounded from above on $X$.
4. Extreme points of a face of $X$ are extreme points of $X$.
5. Extreme points of $X$, if any, are exactly the faces of $X$ which are singletons.

   *Note:* As a corollary of 1—3, 5, we see that extreme points of polyhedral set $X$ are exactly the maximizers of those linear forms which achieve their maximum on $X$ at a unique point.

**Exercise** II.21 ♦ [Follow-up to Exercise II.20]

1. Let $X \subset Y$ be nonempty closed convex sets in $\mathbf{R}^n$. Is it true that $\mathrm{Ext}(Y) \cap X \subset \mathrm{Ext}(X)$ ?
2. Let $X$ be a nonempty closed convex set contained in the polyhedral set $\{x : Ax \le b\}$. Assuming that the set $\overline{X} = X \cap \{x : Ax = b\}$ is nonempty, is it true that $\mathrm{Ext}(\overline{X}) = \mathrm{Ext}(X) \cap \overline{X}$ ?
3. By the result of Exercise II.13, the extreme points of the polytope $\Pi_{m,n} = \{[x_{ij}] \in \mathbf{R}^{m \times n} : x_{ij} \ge 0, \sum_i x_{ij} \le 1 \,\forall j, \sum_j x_{ij} \le 1 \,\forall i\}$ are exactly the Boolean matrices from this polytope. Now let $\widehat{\Pi}_{m,n}$ be the part of $\Pi_{m,n}$ cut off $\Pi_{m,n}$ by imposing on prescribed row and columns of $m \times n$ matrix $x \in \Pi_{m,n}$ the requirement to be equal to 1, rather than to be $\le 1$. Assuming $\widehat{\Pi}_{m,n}$ nonempty, prove that the extreme points of this polytope are exactly the Boolean matrices contained in it.

**Exercise** II.22 Let $X \subset \mathbf{R}^m$ be a nonempty polyhedral set, $x \mapsto Px + p : \mathbf{R}^n \to \mathbf{R}^m$ be an affine mapping, and $Y$ be the image of $X$ under this mapping. Mark by **T** the statements in the below list which are always (i.e., for all $X, P, p$ compatible with the above assumptions) true:

1. $Y$ is a nonempty polyhedral set.
2. If $X$ does not contain lines, so is $Y$.
3. If $X$ does contain lines, so is $Y$.
4. If $v$ is an extreme point of $X$, then $Pv + p$ is an extreme point of $Y$.
5. If $z$ is an extreme point of $Y$, then $z = Pv + p$ for certain extreme point $z$ of $X$.
6. If $z$ is an extreme point of $Y$ and $X$ does not contain lines, then $z = Pv + p$ for certain extreme point $z$ of $X$.

**Exercise** II.23 ▲ Find extreme points of the following closed convex sets:

1. The set $\mathcal{S}_n = \{X \in \mathbf{S}^n : -I_n \preceq X \preceq I_n\}$
2. The set $\mathcal{S}_n^+ = \{X \in \mathbf{S}^n : 0 \preceq X \preceq I_n\}$
3. The set $\mathcal{D}_{k,n} = \{X \in \mathbf{S}^n : I_n \succeq X \succeq 0, \mathrm{Tr}(X) = k\}$, where $k$ is a positive integer $\leq n$.
4. The set $\mathcal{M}_n = \{X \in \mathbf{R}^{n \times n} : \|X\|_{2,2} \leq 1\}$ ($\|\cdot\|_{2,2}$ is the spectral norm)

**Exercise** II.24 ▲ Prove the following fact (which can be considered as a matrix extension of Birkhoff Theorem):

For positive integers $d, n$, let $\Pi_{d,n}$ be the set of all $n \times n$ block matrices with $d \times d$ symmetric blocks $X^{ij}$ satisfying

$$X^{ij} \succeq 0, \sum_j \mathrm{Tr}(X^{ij}) = 1 \forall i, \sum_i \mathrm{Tr}(X^{ij}) = 1 \forall j.$$

The extreme points of $\Pi_{d,n}$ are exactly the block matrices $[X^{ij}]_{i,j \leq n}$ as follows: for certain $n \times n$ permutation matrix $P$ and unit vectors $e_{ij} \in \mathbf{R}^d$, one has

$$X^{ij} = P_{ij} e_{ij} e_{ij}^\top \,\forall i, j.$$

**Exercise** II.25 ▲ Let $k, n$ be positive integers with $k \leq n$, and let $s_k(\lambda)$ for $\lambda \in \mathbf{R}^n$ be the sum of $k$ largest entries in $\lambda$. From the description of the extreme points of the polytope $X = \{x \in \mathbf{R}^n : 0 \leq x_i \leq 1, i \leq n, \sum_{i=1}^n x_i \leq k\}$, see Example II.9.2 in section 9.3, it follows that when $\lambda \in \mathbf{R}^n_+$, then

$$\max_{x \in X} \sum_{i=1}^n \lambda_i x_i = s_k(\lambda).$$

Prove the following matrix analogy of this fact:

For $k, n$ as above, let $\mathcal{X} = \{(X_1, \ldots, X_n) : X_i \in \mathbf{S}^d, 0 \preceq X_i \preceq I_d, i \leq n, \sum_{i=1}^n X_i \preceq kI_d\}$. Then for $\lambda \in \mathbf{R}^n_+$ one has

$$(X_1, \ldots, X_n) \in \mathcal{X} \implies \sum_{i=1}^n \lambda_i X_i \preceq s_k(\lambda) I_d,$$

with the concluding $\preceq$ being $=$ for properly selected $(X_1, \ldots, X_n) \in \mathcal{X}$.

## 11.3 Cones and extreme rays

**Exercise** II.26 ▲ Let $X$ be a nonempty closed and bounded set in $\mathbf{R}^n$. Which of the following statements are true?

1. $\mathrm{Conv}(X)$ is closed convex set.
2. $\mathrm{Cone}(X)$ is a closed cone.
3. When $X$ is convex, $\mathrm{Cone}(X)$ is closed cone.
4. When $0 \notin X$, $\mathrm{Cone}(X)$ is a closed cone.

5. When $0 \notin X$ and $X$ is convex, $\mathrm{Cone}(X)$ is closed cone.
6. When $X$ is polyhedral, $\mathrm{Cone}(X)$ is a closed cone.

**Exercise** II.27 ♦ Let $X \subset \mathbf{R}^n$ be a nonempty polyhedral set given by polyhedral representation:

$$X = \{x : \exists u : Ax + Bu \leq r\}$$

and let $K = \mathrm{Cone}(X)$ be the conic hull of $X$.

1. Is it true that $K$ is a closed cone?
2. Prove that $\overline{K} := \mathrm{cl}\, K$ is a polyhedral cone and find polyhedral representation of $\overline{K}$.
3. Assume that $X$ is given by plain – no extra variables – polyhedral representation: $X = \{x : Ax \leq b\}$. Build plain polyhedral representation of $\overline{K} := \mathrm{cl}\,\mathrm{Cone}(X)$.

**Exercise** II.28 As we know, the extreme directions of the nonnegative orthant $\mathbf{R}^n_+ = \mathbf{R}_+ \times \mathbf{R}_+ \times \ldots \times \mathbf{R}_+$ are the vectors with single positive entry and remaining entries equal to 0. Prove the following generalization of this observation:

Let $X_i \subset \mathbf{R}^{n_i}$, $1 \leq i \leq K$, be closed and pointed cones. The extreme directions of the direct product $X = X_1 \times \ldots \times X_K$ of these cones are the block-vectors $d = [d_1; \ldots; d_K]$ with $d_i \in \mathbf{R}^{n_i}$ of the following structure: all but one blocks in $d$ are zero, and the only nonzero block is an extreme direction of the corresponding factor $X_i$.

**Exercise** II.29 Describe all extreme rays of

1. positive semidefinite cone $\mathbf{S}^n_+$
2. Lorentz cone $\mathbf{L}^n$

## 11.4 Recessive cone

**Exercise** II.30 ▲ Let $M$ be a convex set, and let $\bar{x}$ and $h$ be such that $R_{\bar{x}} := \{\bar{x} + th : t \geq 0\} \subset M$.

1. Is it always true that whenever $x \in M$, the set $R_x = \{x + th, t \geq 0\}$ is contained in $M$ ?
2. Let $h$ be a recessive direction of $\overline{M} = \mathrm{cl}\, M$, and let $\bar{x}$ be a point from the relative interior of $M$. Is it always true that the set $R_{\bar{x}} = \{\bar{x} + th : t \geq 0\}$ is contained in $M$ ?

**Exercise** II.31 ▲ Let $M \subset \mathbf{R}^n$ be a cone, not necessary closed; recall that pointedness of a cone $M$ means that the only vector $x$ such that $x \in M$ and $-x \in M$ is the zero vector. Which of the following statements are always true:

1. $M$ is pointed if an only if the only representation of 0 as the sum of $k \geq 1$ vectors $x_i \in M$ is the representation with $x_i = 0$, $i \leq k$.
2. $M$ is pointed if and only if $M$ does not contain straight lines (one-dimensional affine planes) passing through the origin.
3. Assuming $M$ closed, $M$ is pointed if and only if $M$ does not contain straight lines.
4. $M$ is pointed cone if and only if the closure of $M$ is so.
5. The closure of $M$ is a pointed cone if and only if $M$ does not contain straight lines.

**Exercise** II.32 Literal interpretation of the words "polyhedral cone" is: a polyhedral set $\{x : Ax \leq b\}$ which is a cone. An immediate example is the solution set $\{x : Ax \leq 0\}$ of *homogeneous* system of linear inequalities. Prove that this example is generic: whenever a polyhedral set $K = \{x : Ax \leq b\}$ is a cone, one has $K = \{x : Ax \leq 0\}$.

**Exercise** II.33 ♦ Prove the following modification of Proposition II.8.18:

(!) *Let $X \subset \mathbf{R}^N$ be a nonempty closed convex set such that $X \subset V + \mathrm{Rec}(X)$ for some bounded and closed set $V$, let $x \mapsto \mathcal{A}(x) = Ax + b : \mathbf{R}^N \to \mathbf{R}^n$ be an affine mapping, and let $Y = \mathcal{A}(X) := \{y : \exists x \in X : y = \mathcal{A}(x)\}$ be the image of $X$ under this mapping. Let also*

$$K = \{h \in \mathbf{R}^n : \exists g \in \mathrm{Rec}(X) : h = Ag\}.$$

*Then the recessive cone of the closure $\overline{Y}$ of $Y$ is the closure $\overline{K}$ of $K$. In particular, when $K$ is closed (as definitely is the case when $\mathrm{Rec}(X)$ is polyhedral), it holds $\mathrm{Rec}(\overline{Y}) = K$.*

**Exercise** II.34 ♦ [follow-up to Exercise II.33]

1. Let $K_1 \subset \mathbf{R}^n, K_2 \subset \mathbf{R}^n$ be closed cones, and let $K = K_1 + K_2$.
   - Is it always true that $K$ is a cone?
   - Is it always true that $K$ is closed?
   - Let $K_2$ be polyhedral. Is it always true that $K$ is closed?
   - Let both $K_1$ and $K_2$ be polyhedral. Is it always true that $K$ is closed?
2. Let $X_i$, $i = 1, \dots, I$, be closed convex sets in $\mathbf{R}^n$ with nonempty intersection. Is it true that $\cap_i \mathrm{Rec}(X_i) = \mathrm{Rec}(\cap_i X_i)$?
3. Let $X_1$, $X_2$ be nonempty closed convex sets in $\mathbf{R}^n$, let $K_1 = \mathrm{Rec}(X_1)$, $K_2 = \mathrm{Rec}(X_2)$, $\overline{X} = \mathrm{cl}(X_1 + X_2)$, $\overline{K} = \mathrm{cl}(K_1 + K_2)$.
   - Is it always true that $\overline{K} \subset \mathrm{Rec}(\overline{X})$ ?
   - Is is always true that $\overline{K} = \mathrm{Rec}(\overline{X})$ ?
   - Assume that $X_i \subset V_i + K_i$ for properly selected closed and bounded set $V_i$, $i = 1, 2$, Is it true that $\overline{K} = \mathrm{Rec}(\overline{X})$ ?

**Exercise** II.35 ▲ Let $f(x) = x^\top Cx - c^\top x + \sigma$ be quadratic form with $C \succeq 0$. By Exercise I.15, the set $E = \{x : f(x) \le 0\}$ is convex (and of course closed). Assuming $E \neq \varnothing$, describe $\mathrm{Rec}(E)$.

## 11.5 Around majorization

**Exercise** II.36 ♦ Let $x \in \mathbf{R}^m$, let $X[x]$ be the convex hull of all permutations of $x$, and let $X_+[x]$ be the set of all vectors $x'$ dominated by a vector form $X[x]$:

$$X_+[x] = \{y : \exists z \in X[x] : y \le z\}.$$

1) Prove that $X_+[x]$ is a closed convex set.

2) Prove the following characterization of $X_+[x]$: $X_+[x]$ is exactly the set of solutions of the system of inequalities $s_j(y) \le s_j(x)$, $j = 1, \dots, m$, in variables $y$, where, as always $s_j(z)$ is the sum of the $j$ largest entries in vector $z$.

## 11.6 Around polars

**Exercise** II.37 Justify the last three claims in Example II.8.11.

**Exercise** II.38 ♦ [more on polars]

1. Recall that for $U \subset \mathbf{R}^n$, $\mathrm{Vol}(U)$ stands for the ratio of the $n$-dimensional volume of $U$ and the volume of the $n$-dimensional unit Euclidean ball. Check that for a centered at the origin ellipsoid $E = \{x : x^\top Cx \le 1\}$ $(C \succ 0)$ we have $\mathrm{Vol}(E)\mathrm{Vol}(\mathrm{Polar}\,(E)) = 1$.
2. Let $C \succ 0$ and let ellipsoid $E = \{x : (x - c)^\top C(x - c) \le 1\}$ contain the origin. Compute $\mathrm{Polar}\,(E)$.
3. Let $X_k$, $k \le K$, be closed convex sets in $\mathbf{R}^n$ containing the origin. Prove that

$$\begin{array}{rcll} \mathrm{Polar}\,(\mathrm{Conv}(\cup_k X_k)) & = & \cap_k \mathrm{Polar}\,(X_k) & (a) \\ \mathrm{Polar}\,(\cap_k X_k) & = & \mathrm{cl}\,\mathrm{Conv}(\cup_k \mathrm{Polar}\,(X_k)) & (b) \end{array}$$

**Exercise** II.39 ⧫ Let $X \subset \mathbf{R}^n$ be a cone given by polyhedral representation

$$X = \{x \in \mathbf{R}^n : \exists u : Ax + Bu \leq r\}$$

Is the dual to $X$ cone $X_*$ polyhedral? If yes, build a polyhedral representation of $X_*$.

**Exercise** II.40 ⧫

1. Let $X \subset \mathbf{R}^n$ be a nonempty polyhedral set given by polyhedral representation

$$X = \{x \in \mathbf{R}^n : \exists u : Ax + Bu \leq r\}$$

   Is the polar $\mathrm{Polar}\,(X)$ of $X$ polyhedral? If yes, point out a polyhedral representation of $\mathrm{Polar}\,(X)$. For non-polyhedral extension, see Exercise IV.36.
2. Compute the polars of
   1. probabilistic simplex $\Delta = \{x \in \mathbf{R}^n : x \geq 0, \sum_i x_i = 1\}$
   2. convex hull of nonempty finite set of points $a_1, \ldots, a_N$ from $\mathbf{R}^n$
   3. the set $\{x \in \mathbf{R}^n : x \leq b\}$
      *Solution:* $\mathrm{Polar}\,(\{x : x \leq b\}) = \{y : y \geq 0, y^\top b \leq 1\}$

## 11.7 Miscellaneous exercises

**Exercise** II.41 ▲ Let $X = \{x \in \mathbf{R}^n : Ax \leq b\}$ be a nonempty polyhedral set.

1. Prove that $X$ is bounded if and only if every one of the vectors $\pm e_i$, ($e_i$, $1 \leq i \leq n$, are the basic orth) can be represented as conic combination of columns of $A^\top$.
2. Certify the correct statements in the following list:
   - The polyhedral set $X = \{x \in \mathbf{R}^3 : x \geq [1/3; 1/3; 1/3], \sum_{i=1}^3 x_i \leq 1\}$ is bounded.
   - The polyhedral set $X = \{x \in \mathbf{R}^3 : x_1 \geq 1/3, x_2 \geq 1/3, \sum_{i=1}^3 x_i \leq 1\}$ is unbounded.

**Exercise** II.42 Prove the easy part of Theorem II.9.9, specifically, that every $n \times n$ permutation matrix is an extreme point of the polytope $\Pi_n$ of $n \times n$ doubly stochastic matrices.

**Exercise** II.43 ⧫ [robust LP] Consider *uncertain* Linear Programming problem – a family

$$\left\{ \min_{x \in \mathbf{R}^n} \{c^\top x : [A + \sum\nolimits_{\nu=1}^N \zeta_\nu \Delta_\nu] x \leq b + \sum\nolimits_{\nu=1}^N \zeta_\nu \delta_\nu\} : \zeta \in \mathcal{Z} \right\} \tag{11.3}$$

of LP instances of common sizes ($n$ variables, $m$ constraints). The associated story is as follows: we want to solve an LP program with the data not known exactly when the problem is being solved; what we know at this time, is that the "true problem" belongs to the parametric family given, according to (11.3), by the "nominal data" $c, A, b$, "basic perturbations $\Delta_\nu, \delta_\nu$" and the *perturbation set* $\mathcal{Z}$ through which run the data perturbations $\zeta$ specifying particular instances in the family. In this situation (quite typical for real life applications of LP, where partial data uncertainty is a rule rather than an exception), one way to "immunize" decisions against data uncertainty is to look for *robust solutions* – those remaining feasible for all perturbations of the data from the perturbation set – by solving the *Robust Counterpart* (RC) of our uncertain problem – the optimization problem

$$\min_x \left\{ c^\top x : [A + \sum\nolimits_{\nu=1}^N \zeta_\nu \Delta_\nu] x \leq b + \sum\nolimits_{\nu=1}^N \zeta_\nu \delta_\nu \,\forall (\zeta \in \mathcal{Z}) \right\} \tag{RC}$$

(RC) is *not* an LP program – it has finitely many decision variables and infinite (when $\mathcal{Z}$ is "massive") system of linear constraints on these variables. Optimization problems of this type are called *semi-infinite* and are, in general, difficult to solve. However, the RC of an uncertain LP is easy, provided that $\mathcal{Z}$ is a "computation-friendly" set, for example, nonempty set given by polyhedral representation:

$$\mathcal{Z} = \{\zeta : \exists u : P\zeta + Qu \leq r\} \tag{11.4}$$

Now goes the exercise *per se*:
Use LP duality to reformulate (RC), (11.4) as an explicit LP program.

**Exercise** II.44 ▲ Consider scalar linear constraint

$$a^\top x \le b \tag{1}$$

with uncertain data $a \in \mathbf{R}^n$ ($b$ is certain) varying in the set

$$\mathcal{U} = \{a : |a_i - a_i^*|/\delta_i \le 1, 1 \le i \le n, \sum\nolimits_{i=1}^{n} |a_i - a_i^*|/\delta_i \le k\} \tag{2}$$

where $a_i^*$ are given "nominal data," $\delta_i > 0$ are given quantities, and $k \le n$ is an integer (in literature, this is called "budgeted uncertainty"). Rewrite the Robust Counterpart

$$a^\top x \le b \, \forall a \in \mathcal{U} \tag{RC}$$

in a tractable LO form (that is, write down an explicit system ($S$) of linear inequalities in variables $x$ and additional variables such that $x$ satisfies (RC) if and only if $x$ can be extended to a feasible solution of ($S$)).

**Exercise** II.45 ▲ [computational study, follow-up to Exercise II.43]
*Preliminaries.* Consider oscillator transmitting harmonic wave with unit wavelength and placed at some point $P$ in 3D. Physics says that the electric field generated by the oscillator, when measured at a remote point $A$, is

$$e_A(t) \approx r^{-1} \underbrace{\alpha \cos\big(\omega t - 2\pi r + \theta + 2\pi d \cos(\phi) + \omega t\big)}_{E_A(t)} \tag{$*$}$$

where

- $t$ is time, $\omega$ is the frequency,
- $r$ is the distance from $A$ to the origin $O$, $d$ is the distance from $P$ to the origin, $\phi \in [0, \pi]$ is the angle between the directions $\overrightarrow{OP}$ and $\overrightarrow{OA}$,
- $\alpha$ and $\theta$ are responsible for how the oscillator is actuated.

The difference between the left and the right hand sides in ($*$) is of order of $r^{-2}$ and in all our subsequent considerations can be completely ignored.

It is convenient to assemble $\alpha$ and $\theta$ into the *actuation weight* – the complex number $w = \alpha e^{\imath\theta}$ ($\imath$ is the imaginary unit); with this convention, we have

$$E_A(t) = \Re\left[w D_P(\phi) e^{\imath\omega t - 2\pi r}\right], \, D_P(\phi) = e^{2\pi\imath d \cos(\phi)}.$$

where $\Re[\cdot]$ stands for the real part of a complex number. The complex-valued function $D_P(\phi) : [0, \pi] \to \mathbf{C}$, called *the diagram* of the oscillator, is responsible for the directional density of the energy emitted by the oscillator: when evaluated at certain 3D direction $\vec{e}$, this density is proportional to $|D_p(\phi)|^2$, where $\phi$ is the angle between the direction $\vec{e}$ and the direction $\overrightarrow{OP}$. Physics says that when our transmitting antenna is composed of $K$ harmonic oscillators located at points $P_1, \ldots, P_K$ and actuated with weights $w_1, \ldots, w_K$, the directional density of the energy transmitted by the resulting *antenna array*, as evaluated at a direction $\vec{e}$, is proportional to $|\sum_k w_k D_k(\phi_k(\vec{e}))|^2$, where $\phi_k(\vec{e})$ is the angle between the directions $\vec{e}$ and $\overrightarrow{OP_k}$.

Consider the design problem as follows. We are given linear array of $K$ oscillators placed at the points $P_k = (k-1)\delta\mathbf{e}$, $k \le K$, where $\mathbf{e}$ is the first basic orth (that is, the unit vector "looking" along the positive direction of the $x$-axis), and $\delta > 0$ is a given distance between consecutive oscillators. Our goal is to specify actuation weights $w_k$, $k \le K$, in order to send as much of total energy as possible along the directions which make at most a given angle $\gamma$ with $\mathbf{e}$. To this end, we intend to act as follows:

We want to select actuation weights $w_k$, $k \leq K$, in such a way that the magnitude $|D^w(\phi)|$ of the complex-valued function

$$D^w(\phi) = \sum_{k=1}^{K} w_k \mathrm{e}^{2\pi\imath(k-1)\delta\cos(\phi))}$$

of $\pi \in [0, \pi]$ is "concentrated" on the segment $0 \leq \phi \leq \gamma$. Let us normalize the weights by the requirement

$$D^w(0) = 1$$

and minimize under this restriction the "sidelobe level"

$$\max_{\gamma \leq \phi \leq \pi} |D^w(\phi)|$$

over $w$.

To get a computation-friendly version of this problem, we replace the full range $[0, \pi]$ of values of $\phi$ with $M$-point equidistant grid

$$\Gamma = \{\phi_\ell = \frac{\ell\pi}{M-1} : 0 \leq \ell \leq M-1\},$$

thus converting our design problem into the optimization problem

$$\mathrm{Opt} = \min_{t,w} \left\{ t : \begin{array}{l} |\sum_{k=1}^{K} w_k \mathrm{e}^{2\pi\imath(k-1)\delta\cos(\phi_\ell)}| \leq t \,\forall(\ell : \phi_\ell > \gamma) \\ \sum_{k=1}^{K} w_k \mathrm{e}^{2\pi\imath(k-1)\delta} = 1 \end{array}, w_k \in \mathbf{C}, k \leq K \right\} \qquad (P)$$

which is a convex problem in $2k$ real variables – real and imaginary parts of $w_1, \ldots, w_K$.

*Your tasks* are as follows:

1. Process problem $(P)$ numerically and find the optimal design $w^{\mathrm{n}} = \{w_k^{\mathrm{n}}, k \leq K\}$ along with the optimal value $\mathrm{Opt}^{\mathrm{n}}$. Here and in what follows, recommended setup is
   - number of oscillators $K = 24$, distance between consecutive oscillators $\delta = 0.125$
   - $\gamma = \pi/12$
   - cardinality $M$ of the equidistant grid $\Gamma$ is 512

   Draw the plot of the modulus of the resulting diagram

   $$D^{\mathrm{n}}(\phi) = \sum_{k=1}^{K} w_k^{\mathrm{n}} \mathrm{e}^{2\pi\imath(k-1)\delta\cos(\phi)}$$

   and compute the corresponding "energy concentration" $\mathcal{C}^{\mathrm{n}}$, with concentration of a diagram $D(\cdot)$ defined as

   $$\mathcal{C} = \frac{\sum_{\ell:\phi_\ell \leq \gamma} \sin(\phi_\ell)|D(\phi_\ell)|^2}{\sum_{\ell=1}^{M} \sin(\phi_\ell)|D(\phi_\ell)|^2}$$

   – up to discretization of $\phi$, this is the ratio of the energy emitted in the "cone of interest" (i.e., along the directions making angle at most $\gamma$ with $\mathbf{e}$) to the total emitted energy. Factors $\sin(\phi_\ell)$ reflect the fact that when computing the energy emitted in a spatial cone, we should integrate $|D(\cdot)|^2$ over the part of the unit sphere in $3D$ cut off the sphere by the cone.
2. Now note that "in reality" the optimal weights $w_k^{\mathrm{n}}$, $k \leq K$ are used to actuate physical devices and as such cannot be implemented with the same 16-digit accuracy with which they are computed; they definitely will be subject to small implementation errors. We can model these errors by assuming that the "real life" diagram is

   $$D(\phi) = \sum_{k=1}^{K} w_k^n (1 + \rho\xi_k) \mathrm{e}^{2\pi\imath(k-1)\delta\cos(\phi)}$$

where $\rho \geq 0$ is some (perhaps small) perturbation level and $\xi_k \in \mathbf{C}$ are "primitive" perturbations responsible for the implementation errors and running through the unit disk $\{\xi : |\xi| \leq 1\}$. It is not a great sin to assume that $\xi_k$ are independent across $k$ random variables uniformly distributed on the unit circumference in $\mathbf{C}$. Now the diagram becomes random and can violate the constraints of $(P)$ , unless $\rho = 0$; in the latter case, the diagram is the "nominal" one given by the optimal weights $w^{\mathrm{n}}$, so that it satisfies the constraints of $(P)$ with $t$ set to $\mathrm{Opt}^{\mathrm{n}}$.

Now, what happens when $\rho > 0$? In this case, the diagram $D(\cdot)$ and its deviation $v$ from the prescribed value 1 at the origin, its sidelobe level $\mathfrak{l} = \max_{\ell:\phi_\ell>\gamma} |D(\phi_\ell)|$, and energy concentration become random. A crucial "real life" question is how large are "typical values" of these quantities. To get impression of what happens, you are asked to carry out the numerical experiment as follows:

- select perturbation level $\rho \in \{10^{-\ell}, 1 \leq \ell \leq 6\}$
- for selected $\rho$, simulate and plot 100 realizations of the modulus of the actual diagram, and find empirical averages $\overline{v}$ of $v$, $\bar{\mathfrak{l}}$ of $\mathfrak{l}$, and $\overline{\mathcal{C}}$ of $\mathcal{C}$.

3. Apply Robust Optimization methodology from Exercise II.43 to build "immunized against implementation errors" solution to $(P)$, compute these solutions for perturbation levels $10^{-\ell}$, $1 \leq \ell \leq 6$, and subject the resulting designs to numerical study similar to the one outlined in the previous item.

   *Note:* $(P)$ is *not* a Linear Programming program, so that you cannot formally apply the results stated in Exercise II.43; what you can apply, is the Robust Optimization "philosophy."

**Exercise** II.46 ♦ Prove the statement "symmetric" the Dubovitski-Milutin Lemma:

The cone $M_*$ dual to the arithmetic sum of $k$ (close or not) cones $M^i \subset \mathbf{R}^n$, $i \leq k$, is the intersection of the $k$ cones $M^i_*$ dual to $M^i$.

**Exercise** II.47 ♦ Prove the following polyhedral version of the Dubovitski-Milutin Lemma:

Let $M^1, \ldots, M^k$ be polyhedral cones in $\mathbf{R}^n$, and let $M = \cap_i M^i$. The cone $M_*$ dual to $M$ is the sum of cones $M^i_*$, $i \leq k$, dual to $M^i$, so that a linear form $e^\top x$ is nonnegative on $M$ if and only it can be represented as the sum of linear forms $e_i^\top x$ nonnegative on the respective cones $M_i$.

**Exercise** II.48 ♦ [follow-up to Exercise II.47] Let $A \in \mathbf{R}^{m\times n}$ be a matrix with trivial kernel, $e \in \mathbf{R}^n$, and let the set

$$X = \{x : Ax \geq 0, e^\top x = 1\} \tag{$*$}$$

be nonempty and bounded. Prove that there exists $\lambda \in \mathbf{R}^m$ such that $\lambda > 0$ and $A^\top \lambda = e$.

Prove "partial inverse" of this statement: *if* $\mathrm{Ker} A = \{0\}$ *and* $e = A^\top \lambda$ *for some* $\lambda > 0$*, the set* $(*)$ *is bounded.*

**Exercise** II.49 ♦ Let $E$ be a linear subspace in $\mathbf{R}^n$, $K$ be a closed cone in $\mathbf{R}^n$, and $\ell(x) : E \to \mathbf{R}$ be a linear (linear, not affine!) function which is nonnegative on $K \cap E$. Which of the following claims are always true:

1. $\ell(\cdot)$ can be extended from $E$ onto the entire $\mathbf{R}^n$ to yield a linear function which is nonnegative on $K$
2. Assuming $\mathrm{int}\, K \cap E \neq \varnothing$, $\ell(\cdot)$ can be extended from $E$ onto the entire $\mathbf{R}^n$ to yield a linear function which is nonnegative on $K$.
3. Assuming, in addition to $\ell(x) \geq 0$ for $x \in K \cap E$, that $K = \{x : Px \leq 0\}$ is a polyhedral cone, $\ell(\cdot)$ can be extended from $E$ onto the entire $\mathbf{R}^n$ to yield a linear function which is nonnegative on $K$.

**Exercise** II.50 Let $n > 1$. Is the unit $\|\cdot\|_2$-ball $B_n = \{x \in \mathbf{R}^n : \|x\|_2 \leq 1\}$ a polyhedral set? Justify your answer.

**Exercise** II.51 ▲ The unit box $\{x \in \mathbf{R}^n : -1 \leq x_i \leq 1, i \leq n\}$ is cut off $\mathbf{R}^n$ by a system of $m = 2n$ linear inequalities and is a nonempty and bounded polyhedral set. However, when we eliminate any inequality from this system, the solution set of the resulting system becomes unbounded. To see that this situation is in a sense extreme, prove the following claim:

> *Consider the solution set of a system of $m$ linear inequalities in $n$ variables $x$, i.e., the set*
>
> $$X := \{x \in \mathbf{R}^n : \; Ax \leq b\},$$
>
> *where $A = [a_1^\top; a_2^\top; \ldots; a_m^\top]$. Suppose that $X$ is nonempty and bounded. Then, whenever $m > 2n$, one can drop from this system a properly selected inequality in such a way that the solution set of the resulting subsystem remains bounded.*

A provocative follow-up: Is it possible to cut off from $\mathbf{R}^{1000}$ a bounded set by using *only* a single linear inequality?

**Exercise** II.52 ▲ [computational study] Let $\omega^N = (\omega_1, \ldots, \omega_N)$ be an $N$-element i.i.d. sample drawn from the standard Gaussian distribution (zero mean, unit covariance) on $\mathbf{R}^d$. How many extreme points are there in the convex hull of the points from the sample?

1. Consider the planar case $d = 2$ and think how to list extreme points of $\mathrm{Conv}\{\omega_1, \ldots, \omega_N\}$. Fill the following table:

| $N$ | 2 | 4 | 8 | 16 | 32 | 64 | 128 |
|---|---|---|---|---|---|---|---|
| $U$ | | | | | | | |
| $M$ | | | | | | | |
| $L$ | | | | | | | |

   where $U$ is the maximal, $M$ is the mean, and $L$ is the minimal # of extreme points observed when processing 100 samples $\omega^N$ of a given cardinality

2. Think how to upper-bound the expected number of extreme points in $W = \mathrm{Conv}(\omega^N)$.

**Exercise** II.53 ▲ [computational study] Given positive integers $m, n$, with $n \geq 2$, consider randomly generated system $Ax \leq b$ of $m$ linear inequalities with $n$ variables. We assume that $A$, $b$ are generated by drawing the entries, independently of each other, from $\mathcal{N}(0, 1)$.

1. Consider the planar case $n = 2$. For $m = 2, 4, 8, 16$, generate 100 samples of $m \times 2$ systems and fill the following table:

| $m$ | 2 | 4 | 8 | 16 |
|---|---|---|---|---|
| $F$ | | | | |
| $B$ | | | | |

   where $F$ is the number of feasible systems, and $U$ is the number of feasible systems with bounded solution sets.

Intermezzo: related theoretical results originating from [Nem24, Exercise 2.23] are as follows. Given positive integers $m, n$ with $n \geq 2$, consider homogenous system $Ax \leq 0$ of $m$ inequalities with $n$ variables. We call this system *regular*, if its matrix $A$ is regular, regularity of a matrix $B$ meaning that all square submatrices of $B$ are nonsingular. Clearly, the entries of a regular matrix are nonzero, and when a $p \times q$ matrix $B$ is drawn at random from a probability distribution on $\mathbf{R}^{p \times q}$ which has a density w.r.t the Lebesgue measure, $B$ is regular with probability 1.

Given regular $m \times n$ homogeneous system of inequalities $Ax \leq 0$, let $g_i(x) = \sum_{j=1}^n A_{ij} x_j$, $i \leq m$, so that $g_j$ are nonconstant linear functions. Setting $\Pi_i = \{x : g_i(x) = 0\}$, we get a collection of $m$ hyperplanes in $\mathbf{R}^n$ passing through the origin. For a point $x \in \mathbf{R}^n$, the *signature* of $x$ is, by definition, the $m$-dimensional vector $\sigma(x)$ of signs of the reals $g_i(x)$, $1 \leq i \leq m$. Denoting by $\Sigma$ the set of all $m$-dimensional vectors with entries $\pm 1$, for $\sigma \in \Sigma$

the set $\mathcal{C}_\sigma = \{x : \sigma(x) = \sigma\}$ is either empty, or is a nonempty open convex set; when it is nonempty, let us call it a *cell* associated with $A$, and the corresponding $\sigma$ – an $A$-feasible signature. Clearly, for regular system, $\mathbf{R}^n$ is the union of all hyperplanes $\Pi_i$ and all cells associated with $A$. It turns out that

> *The number $N(m,n)$ of cells associated with a regular homogeneous $m \times n$ system $Ax \leq 0$ is independent of the system and is given by a simple recurrence:*

$$\begin{array}{rcl} N(1,2) & = & 2 \\ m \geq 2, n \geq 2 & \Longrightarrow & N(m,n) = N(m-1,n) + N(m-1,n-1) \quad [N(m,1) = 2,\, m \geq 1]. \end{array}$$

Next, when $A$ is drawn at random from probability distribution $P$ on $\mathbf{R}^{m\times n}$ which possesses *symmetric* density $p$, that is, such that $p([a_1^\top; a_2^\top; \ldots; a_m^\top]) = p([\epsilon_1 a_1^\top; \epsilon_2 a_2^\top; \ldots; \epsilon_m a_m^\top])$ for all $A = [a_1^\top; a_2^\top; \ldots; a_m^\top]$ and all $\epsilon_i = \pm 1$, then *the probability for a vector $\sigma \in \Sigma$ to be an $A$-feasible signature is*

$$\pi(m,n) = N(m,n)/2^m.$$

In particular, the probability for the system $Ax \leq 0$ to have a solution set with a nonempty interior (this is nothing but $A$-feasibility of the signature $[-1; \ldots; -1]$ is $\pi(m,n)$.
The inhomogeneous version of these results is as follows. An $m\times n$ system of linear inequalities $Ax \leq b$ is called regular, if the matrix $[A, -b]$ is regular. Setting $g_i(x) = \sum_{j=1}^n A_{ij}x_j - b_i$, $i \leq n$, the $[A,b]$-signature of $x$ is, as above, the vector of signs of the reals $g_i(x)$. For $\sigma \in \Sigma$, the set $\mathcal{C}_\sigma = \{x : \sigma(x)) = \sigma\}$ is either empty, or is a nonempty open convex set; in the latter case, we call $\mathcal{C}_\sigma$ an $[A,b]$-cell, and call $\sigma$ an $[A,b]$-feasible signature. Setting $\Pi_i = \{x : g_i(x) = 0\}$, we get $m$ hyperplanes in $\mathbf{R}^n$, and the entire $\mathbf{R}^n$ is the union of those hyperplanes and all $[A,b]$-cells. It turns out that

*The number $N(m,n)$ of cells associated with a regular $m \times n$ system $Ax \leq b$ is independent of the system and is equal to $\frac{1}{2}N(m+1,n+1)$.*

In addition, when $m\times(n+1)$ matrix $[A,b]$ is drawn at random from a probability distribution on $\mathbf{R}^{m\times(n+1)}$ possessing a symmetric density w.r.t. the Lebesgue distribution, *the probability for every $\sigma \in \Sigma$ to be $[A,b]$-feasible signature is*

$$\overline{\pi}(m,n) = N(m+1,n+1)/2^{m+1}.$$

In particular, the probability for the system $Ax \leq b$ to be strictly feasible is $\overline{\pi}(m,n)$.

2. Accompanying exercise: Prove that if $A$ is $m \times n$ regular matrix, then the system $Ax \leq 0$ has a nonzero solution if and only if the system $Ax < 0$ is feasible. Derive from this fact that if $[A,b]$ is regular, then the system $Ax \leq b$ is feasible if and only if it is strictly feasible, and that when the system $Ax \leq 0$ has a nonzero solution, the system $Ax \leq b$ is strictly feasible for every $b$.
3. Use the results from Intermezzo to compute the expected values of $F$ and $B$, see item 1.

**Exercise** II.54 ▲ [computational study]

1. For $\nu = 1, 2, \ldots, 6$, generate 100 systems of linear inequalities $Ax \leq b$ with $n = 2^\nu$ variables and $m = 2n$ inequalities, the entries in $A$, $b$ being drawn, independently of each other, from $\mathcal{N}(0,1)$. Fill the following table:

| $n$ | 2 | 4 | 8 | 16 | 32 | 64 |
|---|---|---|---|---|---|---|
| $F$ | | | | | | |
| $\mathbf{E}\{F\}$ | | | | | | |
| $B$ | | | | | | |

$F$: # of feasible systems in sample;
$B$: # of feasible systems with bounded soultion sets

To compute the expected value of $F$, use the results from [Nem24, Exercise 2.23] cited in item 2 of Exercise II.53.

2. Carry out experiment similar to the one in item 1, but with $m = n + 1$ rather than $m = 2n$.

| $n$ | 2 | 4 | 8 | 16 | 32 | 64 |
|---|---|---|---|---|---|---|
| $F$ | | | | | | |
| $\mathbf{E}\{F\}$ | | | | | | |
| $B$ | | | | | | |
| $\mathbf{E}\{B\}$ | | | | | | |

$F$: # of feasible systems in sample;
$B$: # of feasible systems with bounded soultion sets

# 12

# Proofs of Facts from Part II

**Fact II.7.2** Let $S, T$ be nonempty convex sets in $\mathbf{R}^n$. A linear form $a^\top x$ separates $S$ and $T$ if and only if

$$\begin{array}{ll}(a) & \sup\limits_{x\in S} a^\top x \leq \inf\limits_{y\in T} a^\top y, \quad \text{and} \\ (b) & \inf\limits_{x\in S} a^\top x < \sup\limits_{y\in T} a^\top y.\end{array}$$

This separation is strong if and only if $(a)$ holds as a strict inequality:

$$\sup_{x\in S} a^\top x < \inf_{y\in T} a^\top y.$$

Proof. When a linear form $a^\top x$ separates $S$ and $T$, $(a)$ holds true. Given $(a)$, $(b)$ could be wrong if and only if $\inf\limits_{x\in S} a^\top x = \sup\limits_{y\in T} a^\top y$. But, together with $(a)$, this can happen only if $a^\top x$ is constant on $S \cup T$, which is not the case. The above reasoning clearly can be reversed: given $(b)$, we have $a \neq 0$, and given $(a)$, both $\sup_{x\in S} a^\top x$ and $\inf_{y\in T} a^\top y$ are reals. Selecting $b$ in-between these reals, the hyperplane $a^\top x = b$ clearly separates $S$ and $T$. The "strong separation" claim is evident. □

**Fact II.8.4** Let $M$ be a nonempty convex set and let $x \in M$. Then, $x$ is an extreme point of $M$ if and only if any (and then all) of the following holds:

1. the only vector $h$ such that $x \pm h \in M$ is the zero vector;
2. in every representation $x = \sum_{i=1}^m \lambda_i x^i$ of $x$ as a convex combination, with positive coefficients, of points $x^i \in M$, $i \leq m$, one has $x^1 = \ldots = x^m = x$;
3. the set $M \setminus \{x\}$ is convex.

Proof.

1. If $x$ is extreme point and $x \pm h \in M$, then $h = 0$, since otherwise $x = \frac{1}{2}(x+h) + \frac{1}{2}(x-h)$ is an interior point of a nontrivial segment $[x-h, x+h]$, which is impossible. For the other direction, assume for contradiction that $x \pm h = 0$ implies $h = 0$ and that $x$ is not at extreme point of $M$. Then, as $x \notin \mathrm{Ext}(M)$, there exists $u, v \in M$ where both $u, v$ are not equal to $x$ and $\lambda \in (0,1)$ such that $x = \lambda u + (1-\lambda)v$. As $u \neq x$ and $v \neq x$ while $x = \lambda u + (1-\lambda)v$, we conclude that $u \neq v$. Now, consider any $\delta > 0$ such that $\delta < \min\{\lambda, 1-\lambda\}$ and define $h := \delta(u - v)$. Note that $h \neq 0$ and $x + h = (\lambda+\delta)u + (1-\lambda-\delta)v \in M$ and $x - h = (\lambda-\delta)u + (1-\lambda+\delta)v \in M$ due to $\lambda \pm \delta \in (0,1)$, $u, v \in M$ and convexity of $M$. This then leads to the desired contradiction with our assumption that $x \pm h \in M$ implies that $h = 0$.
   As a byproduct of our reasoning, we see that if $x \in M$ can be represented as $x = \lambda u + (1-\lambda)v$ with $u, v \in M$, $\lambda \in (0, 1]$, and $u \neq x$, then $x$ is *not* an extreme point of $M$.

2. In one direction, when $x$ is *not* an extreme point of $M$, there exists $h \neq 0$ such that $x \pm h \in M$ so that $x = \frac{1}{2}(x+h) + \frac{1}{2}(x-h)$ is a convex combination with positive coefficients and using two points $x \pm h$ that are both in $M$ and are distinct from $x$. To prove the opposite direction, let $x$ be an extreme point of $M$ and suppose $x = \sum_{i=1}^{m} \lambda_i x^i$ with $\lambda_i > 0$, $\sum_i \lambda_i = 1$, and let us prove that $x^1 = \ldots = x^m = x$. Indeed, assume for contradiction that at least one of $x^i$, say, $x^1$, differs from $x$, and $m > 1$. Since $\lambda_2 > 0$, we have $0 < \lambda_1 < 1$. Then, the point $v := (1-\lambda_1)^{-1} \sum_{i=2}^{m} \lambda_i x^i$ is well defined. Moreover, as $\sum_{i=2}^{m} \lambda_i = 1 - \lambda_1$, $v$ is a convex combination of $x^2, \ldots, x^m$ and therefore $v \in M$. Then, $x = \lambda_1 x^1 + (1-\lambda_1)v$ with $x, x^1, v \in M$, $\lambda_1 \in (0,1]$, and $x^1 \neq x$, which, by the concluding comment in item 1 of the proof, implies that $x \notin \mathrm{Ext}(M)$; this is the desired contradiction.
3. In one direction, let $x$ be an extreme point of $M$; let us prove that the set $M' := M \setminus \{x\}$ is convex. Were it not the case, there would exist $u, v \in M'$ and $\lambda \in [0,1]$ such that $\bar{x} := \lambda u + (1-\lambda)v \notin M'$, implying that $0 < \lambda < 1$ (since $u, v \in M'$). Since $M$ is convex, we have $\bar{x} \in M$, and since $\bar{x} \notin M'$ and $M \setminus M' = \{x\}$, we conclude that $\bar{x} = x$. Thus, $x$ is a convex combination, with positive coefficients, of two distinct from $x$ points from $M$, contradicting, by already proved item 2, the fact that $x$ is an extreme point of $M$. For the other direction, suppose that $M \setminus \{x\}$ is convex and we will prove that $x$ must be an extreme point of $M$. Assume for contradiction that $x \notin \mathrm{Ext}(M)$. Then, there exists $h \neq 0$ such that $x \pm h \in M$. As $h \neq 0$, both $x + h$ and $x - h$ are distinct from $x$, thus $x \pm h \in M \setminus \{x\}$. We see that $x \pm h \in M \setminus \{x\}$, $x = \frac{1}{2}(x+h) + \frac{1}{2}(x-h)$ and $x \notin M \setminus \{x\}$, contradicting the convexity of $M \setminus \{x\}$.

□

**Fact II.8.5** All extreme points of the convex hull $\mathrm{Conv}(Q)$ of a set $Q$ belong to $Q$:

$$\mathrm{Ext}(\mathrm{Conv}(Q)) \subseteq Q.$$

Proof. Assume for contradiction that $x \in \mathrm{Ext}(\mathrm{Conv}(Q))$ and $x \notin Q$. As $x \in \mathrm{Ext}(\mathrm{Conv}(Q))$, by Fact II.8.4.(iii) the set $\mathrm{Conv}(Q) \setminus \{x\}$ is convex and contains $Q$, contradicting the fact that $\mathrm{Conv}(Q)$ is the *smallest* convex set containing $Q$. □

**Fact II.8.13** Let $M$ be a nonempty closed convex set in $\mathbf{R}^n$. Then

(i) $\mathrm{Rec}(M) \neq \{0\}$ if and only if $M$ is unbounded.

(ii) If $M$ is unbounded, then all nonzero recessive directions of $M$ are positive multiples of recessive directions of unit Euclidean length, and the latter are *asymptotic directions* of $M$, i.e., a unit vector $h \in \mathbf{R}^n$ is a recessive direction of $M$ if and only if there exists a sequence $\{x^i \in M\}_{i \geq 1}$ such that $\|x^i\|_2 \to \infty$ as $i \to \infty$ and $h = \lim_{i\to\infty} x^i/\|x^i\|_2$.

(iii) $M$ does not contain lines if and only if the cone $\mathrm{Rec}(M)$ does not contain lines.

Proof.

(i) If $\mathrm{Rec}(M) \neq \{0\}$, then $M$ contains a ray and therefore $M$ is unbounded. For the reverse direction, suppose $M$ is unbounded and let us prove that $\mathrm{Rec}(M) \neq \{0\}$. As $M$ is unbounded, there exists a sequence of points $x^i \in M$ such that $\|x^i\|_2 > i$, for all $i = 1, 2, \ldots$. Then, the vectors $h^i := (x^i - x^1)/\|x^i - x^1\|_2$ are well defined unit vectors. Passing to a subsequence, we can assume that $h^i \to h$ as $i \to \infty$ (Theorem B.15), so that $h$ is a unit vector as well. For every $t \geq 0$, the points $x^1 + th^i$, for all $i$ with $\|x^i - x^1\|_2 > t$, are convex combinations of points $x^1$ and $x^i$ and both $x^1, x^i \in M$. Then, as $M$ is convex, $x^1 + th^i \in M$ for all large enough $i$. As $i \to \infty$, the points $x^1 + th^i$ converge to $x^1 + th$, and since $M$ is closed, we conclude $x^1 + th \in M$. Since this holds for every $t \geq 0$, the nonzero vector $h$ is a recessive direction of $M$.

(ii) Suppose $M$ is unbounded and consider any $h \in \mathrm{Rec}(M)$ such that $h$ is a unit vector. Pick any $x^0 \in M$ and define $x^i := x^0 + ih$, $i = 1, 2, \ldots$. Then, we get a sequence of points from $M$ diverging to infinity, i.e., $\|x^i\|_2 \to \infty$, $i \to \infty$, and also satisfying $h = \lim_{i\to\infty} \|x^i\|_2^{-1} x^i$. Thus, $h$ is an asymptotic direction of $M$, as claimed. To prove the reverse direction, if $x^i \in M$ are such that $\|x^i\|_2 \to \infty$ and $h := \lim_{i\to\infty} \|x^i\|_2^{-1} x^i$ exists, then $h$ is a recessive direction of $M$ by the same reasoning used in the proof of item (i), and of course $h$ is a unit vector.

(iii) Suppose $M$ contains a line with direction $h \neq 0$, i.e., for some $x \in M$ and for all $t \in \mathbf{R}$, we have $x + th \in M$. Then, by the definition of recessive direction, both $h$ and $-h$ are recessive directions of $M$, so $\pm h \in \mathrm{Rec}(M)$, and thus $\mathrm{Rec}(M)$ contains a line with the direction $h$. For the reverse direction suppose $h \neq 0$ and $\mathrm{Rec}(M)$ contains a line with direction $h$. Since $\mathrm{Rec}(M)$ is a closed convex cone, the line with the same direction passing through the origin is contained in $\mathrm{Rec}(M)$ (by Lemma II.8.8). Thus, $\pm h \in \mathrm{Rec}(M)$. Then, by the definition of $\mathrm{Rec}(M)$, for any $x \in M$ it holds that $x + th \in M$ for all $t \in \mathbf{R}$. Hence, $M$ contains a line with the direction $h$. □

**Fact II.8.14** Let $M \subseteq \mathbf{R}^n$ be a nonempty closed convex set. Recall its closed conic transform is given by

$$\overline{\mathrm{ConeT}}(M) = \mathrm{cl}\,\{[x;t] \in \mathbf{R}^n \times \mathbf{R} :\; t > 0,\; x/t \in M\},$$

(see section 1.5). Then,

$$\mathrm{Rec}(M) = \left\{h \in \mathbf{R}^n :\; [h;0] \in \overline{\mathrm{ConeT}}(M)\right\}.$$

Proof. Let $h$ be such that $[h;0] \in \overline{\mathrm{ConeT}}(M)$, and let us prove that $h \in \mathrm{Rec}(M)$. There is nothing to prove when $h = 0$, thus assume that $h \neq 0$. Since the vectors $g$ such that $[g;0] \in \overline{\mathrm{ConeT}}$ compose a closed cone and both $\overline{\mathrm{ConeT}}(M)$ and $\mathrm{Rec}(M)$ are cones as well, we lose nothing when assuming, in addition to $[h;0] \in \overline{\mathrm{ConeT}}(M)$ and $h \neq 0$, that $h$ is a unit vector. Since $[h;0] \in \overline{\mathrm{ConeT}}(M)$, by definition of the latter set there exists a sequence $[u^i;t_i] \to [h;0]$, $i \to \infty$, such that $t_i > 0$ and $x^i := u^i/t_i \in M$ for all $i$. Then, this together with $u^i \to h$ and $\|h\|_2 = 1$ imply that $\|x^i\|_2 \to \infty$ and $t_i\|x^i\|_2 \to 1$ as $i \to \infty$. As a result, $\lim_{i\to\infty} \|x^i\|_2^{-1} x^i = \lim_{i\to\infty} u^i = h$. By Fact II.8.13.ii, we see that $h \in \mathrm{Rec}(M)$.

For the reverse direction, consider any $h \in \mathrm{Rec}(M)$, and let us prove that $[h;0] \in \overline{\mathrm{ConeT}}(M)$. There is nothing to prove when $h = 0$, so we assume $h \neq 0$. Consider any $\bar{x} \in M$ and define $x^i := \bar{x} + ih$, $i = 1, 2, \ldots$. As $h \in \mathrm{Rec}(M)$, we have $x^i \in M$ for all $i$. Moreover, $\|x^i\|_2 \to \infty$ as $i \to \infty$ due to $h \neq 0$. We clearly have $\lim_{i\to\infty}[x^i/\|x^i\|_2; 1/\|x^i\|_2] = [h/\|h\|_2; 0]$, and the vectors $[y^i;t_i] := [x^i/\|x^i\|_2, 1/\|x^i\|_2]$ for all large enough $i$ satisfy the requirement $t_i > 0$, $y^i/t_i \in M$, so $[y^i;t_i] \in \overline{\mathrm{ConeT}}(M)$ for all large enough $i$. As $\overline{\mathrm{ConeT}}(M)$ is closed and $[y^i;t_i] \to [h/\|h\|_2;0]$ as $i \to \infty$, we deduce $[h/\|h\|_2;0] \in \overline{\mathrm{ConeT}}(M)$. Finally, $\overline{\mathrm{ConeT}}(M)$ is a cone, so $[h;0] \in \overline{\mathrm{ConeT}}(M)$ as well. □

**Fact II.8.15** For any nonempty polyhedral set $M = \{x \in \mathbf{R}^n : Ax \le b\}$, its recessive cone is given by

$$\mathrm{Rec}(M) = \{h \in \mathbf{R}^n : Ah \le 0\},$$

i.e., $\mathrm{Rec}(M)$ is given by homogeneous version of linear constraints specifying $M$.

Proof. Consider any $h$ such that $Ah \le 0$. Then, for any $\bar{x} \in M$, and $t \ge 0$, we have $A(\bar{x} + th) = A\bar{x} + tAh \le A\bar{x} \le b$, so $\bar{x} + th \in M$ for all $t \ge 0$. Hence, $h \in \mathrm{Rec}(M)$. For the reverse direction, suppose $h \in \mathrm{Rec}(M)$ and $\bar{x} \in M$. Then, for all $t \ge 0$ we have $A(\bar{x} + th) \le b$. This is equivalent to $Ah \le t^{-1}(b - A\bar{x})$ for all $t > 0$, which implies that $Ah \le 0$. □

**Fact II.8.23** Let $M$ be a closed cone in $\mathbf{R}^n$, and let $M_*$ be the cone dual to $M$. Then

1. Duality does not distinguish between a cone and its closure: whenever $M = \mathrm{cl}\, M'$ for a cone $M'$, we have $M_* = M'_*$.
2. Duality is symmetric: the cone dual to $M_*$ is $M$.
3. One has
$$\mathrm{int}\, M_* = \left\{y \in \mathbf{R}^n : \ y^\top x > 0, \forall x \in M \setminus \{0\}\right\},$$
and $\mathrm{int}\, M_*$ is nonempty if and only if $M$ is pointed (i.e., $M \cap [-M] = \{0\}$).
Moreover, when $M$, in addition to being closed, is pointed and nontrivial ($M \neq \{0\}$), one has
$$\mathrm{int}\, M_* = \left\{y \in \mathbf{R}^n : \ M_y := \{x \in M : x^\top y = 1\} \text{ is nonempty and compact}\right\}. \tag{8.7}$$
4. The cone dual to the direct product $M_1 \times \ldots \times M_m$ of cones $M_i$ is the direct product of their duals: $[M_1 \times \ldots \times M_m]_* = [M_1]_* \times \ldots \times [M_m]_*$.

Proof.

1. This is evident.
2. By definition, any $x \in M$ satisfies $x^\top y \geq 0$ for all $y \in M_*$, hence $M \subseteq [M_*]_*$. To prove $M = [M_*]_*$, assume for contradiction that there exists $\bar{x} \in [M_*]_* \backslash M$. By Separation Theorem, $\{\bar{x}\}$ can be strongly separated from $M$, i.e., there exists $y$ such that
$$y^\top \bar{x} < \inf_{x \in M} y^\top x.$$
As $M$ is a conic set and the right hand side infimum is finite, this infimum must be 0. Thus, $y^\top \bar{x} < 0$ while $y^\top x \geq 0$ for all $x \in M$ implying $y \in M_*$. But, then this contradicts to $\bar{x} \in [M_*]_*$.
3. Let us prove that $\mathrm{int}\, M_* \neq \varnothing$ if and only if $M$ is pointed. If $M$ is not pointed, then $\pm h \in M$ for some $h \neq 0$, implying that $y^\top [\pm h] \geq 0$ for all $y \in M_*$, that is, $y^\top h = 0$ for all $y \in M_*$. Thus, when $M$ is not pointed, $M_*$ belongs to a proper (smaller than the entire $\mathbf{R}^n$) linear subspace of $\mathbf{R}^n$ and thus $\mathrm{int}\, M_* = \varnothing$. This reasoning can be inverted: when $\mathrm{int}\, M_* = \varnothing$, the affine hull $\mathrm{Aff}(M_*)$ of $M_*$ cannot be the entire $\mathbf{R}^n$ (since $\mathrm{int}\, M_* = \varnothing$ and $\mathrm{rint}\, M_* \neq \varnothing$); taking into account that $0 \in M_*$, we have $\mathrm{Aff}(M_*) = \mathrm{Lin}(M_*)$, so that $\mathrm{Lin}(M_*) \subsetneq \mathbf{R}^n$, and therefore there exists nonzero $h$ orthogonal to $\mathrm{Lin}(M_*)$. We have $y^\top [\pm h] = 0$ for all $y \in M_*$, implying that $h$ and $-h$ belong to cone dual to $M_*$, that is, to $M$ (due to the already verified item 2). Thus, for some nonzero $h$ it holds $\pm h \in M$, that is, $M$ is not pointed.
Now let us prove that $y \in \mathrm{int}\, M_*$ if and only if $y^\top x > 0$ for every $x \in M \setminus \{0\}$. In one direction: assume that $y \in \mathrm{int}\, M_*$, so that for some $r > 0$ it holds $y + \delta \in M_*$ for all $\delta$ with $\|\delta_2\| \leq r$. If now $x \in M$, we have $0 \leq \min_{\delta:\|\delta\|_2 \leq r} [y + \delta]^\top x = y^\top x - r\|x\|_2$. Thus,
$$y \in \mathrm{int}\, M_* \implies \|x\|_2 \leq \frac{1}{r} y^\top x, \ \forall x \in M, \tag{12.1}$$
implying that $y^\top x > 0$ when $x \in M \setminus \{0\}$, as required. In the opposite direction: assume that $y^\top x > 0$ for all $x \in M \setminus \{0\}$, and let us prove that $y \in \mathrm{int}\, M_*$. There is nothing to prove when $M = \{0\}$ (and therefore $M_* = \mathbf{R}^n$). Assuming $M \neq \{0\}$, let $\bar{M} = \{x \in M : \|x\|_2 = 1\}$. This set is nonempty (since $M \neq \{0\}$), is closed (as $M$ is closed), and clearly is bounded, and thus is compact. We are in the situation when $y^\top x > 0$ for $x \in \bar{M}$, implying that $\min_{x \in \bar{M}} y^\top x$ (this minimum is achieved since $\bar{M}$ is a nonempty compact set) is strictly positive. Thus, $y^\top x \geq r > 0$ for all $x \in \bar{M}$, whence $[y + \delta]^\top x \geq 0$ for all $x \in \bar{M}$ and all $\delta$ with $\|\delta\|_2 \leq r$. Due to the origin of $\bar{M}$, inequality $[y + \delta]^\top x \geq 0$ for all $x \in \bar{M}$ implies that $[y + \delta]^\top x \geq 0$ for all $x \in M$. The bottom line is that the Euclidean ball of radius $r$ centered at $y$ belongs to $M_*$, and therefore $y \in \mathrm{int}\, M_*$, as claimed.
Now let us prove the "Moreover" part of item 3. Thus, let the cone $M$ be closed, pointed, and nontrivial. Consider any $y \in \mathrm{int}\, M_*$, then the set $M_y$, first, contains some positive multiple of every nonzero vector from $M$ and thus is nonempty (since $M \neq \{0\}$) and, second, is bounded (by (12.1)). Since $M_y$ is closed (as $M$ is closed), we conclude that $M_y$ is a nonempty compact

set. Thus, the left hand side set in (8.7) is contained in the right hand side one. To prove the opposite inclusion, let $y \in \mathbf{R}^n$ be such that $M_y$ is a nonempty compact set, and let us prove that $y \in \operatorname{int} M_*$. By the already proved part of item 3, all we need is to verify that if $x \neq 0$ and $x \in M$, then $y^\top x > 0$. Assume for contradiction that there exists $\bar{x} \in M \setminus \{0\}$ such that $\alpha := -y^\top \bar{x} \geq 0$. Then, by selecting any $\widehat{x} \in M_y$ ($M_y$ is nonempty!) and setting $e = \alpha\widehat{x} + \bar{x}$, we get $e \in M$ and $y^\top e = 0$. Note the $e \neq 0$; indeed, $e = 0$ means that the nonzero vector $\bar{x} \in M$ is such that $-\bar{x} = \alpha\widehat{x} \in M$, contradicting pointedness of $M$. The bottom line is that $e \in M \setminus \{0\}$ and $y^\top e = 0$, whence $e$ is a nonzero recessive direction of $M_y$. This is the desired contradiction as $M_y$ is compact!

4. This is evident.

□

**Fact II.8.28** Let $M \subseteq \mathbf{R}^n$ be a cone and $M_*$ be its dual cone. Then, for any $x \in \operatorname{int} M$, there exists a properly selected $C_x < \infty$ such that

$$\forall f \in M_* : \ \|f\|_2 \leq C_x f^\top x.$$

Proof. Since $x \in \operatorname{int} M$, there exists $\rho > 0$ such that $x - \delta \in M$ whenever $\|\delta\|_2 \leq \rho$. Then, as $f \in M^*$, we have $f^\top(x - \delta) \geq 0$ for any $\|\delta\|_2 \leq \rho$ , i.e., $f^\top x \geq \sup_\delta\{f^\top \delta : \|\delta\|_2 \leq \rho\} = \rho\|f\|_2$. Taking $C_x := 1/\rho$ (note that $C_x < \infty$ as $\rho > 0$) gives us the desired relation. □

**Fact II.8.33.** Let $M \subseteq \mathbf{R}^n$ be a nontrivial closed cone, and $M_*$ be its dual cone.
(i) $M$ is pointed
(i.1) if and only if $M$ does not contain straight lines,
(i.2) if and only if $M_*$ has a nonempty interior, and
(i.3) if and only if $M$ has a base.
(ii) Set (8.9) is a base of $M$
(ii.1) if and only if $f^\top x > 0$ for all $x \in M \setminus \{0\}$,
(ii.2) if and only if $f \in \operatorname{int} M_*$.
In particular, $f \in \operatorname{int} M_*$ if and only if $f^\top x > 0$ whenever $x \in M \setminus \{0\}$.
(iii) Every base of $M$ is nonempty, closed, and bounded. Moreover, whenever $M$ is pointed, for any $f \in M_*$ such that the set (8.9) is nonempty (note that this set is always closed for any $f$), this set is bounded if and only if $f \in \operatorname{int} M_*$, in which case (8.9) is a base of $M$.
(iv) $M$ has extreme rays if and only if $M$ is pointed. Furthermore, when $M$ is pointed, there is one-to-one correspondence between extreme rays of $M$ and extreme points of a base $B$ of $M$, specifically, the ray $R := \mathbf{R}_+(d)$, $d \in M \setminus \{0\}$ is extreme if and only if $R \cap B$ is an extreme point of $B$.

Proof. (i.1): Since $M$ is closed, convex, and contains the origin, $M$ contains a line if and only if $M$ contains a line passing through the origin, and since $M$ is conic, the latter happens if and only if $M$ is not pointed.

(i.2): This is precisely Fact II.8.23.3.

(i.3): As we have seen, (8.9) is a base of $M$ if and only if $f^\top x > 0$ for all $x \in M \setminus \{0\}$, which, by Fact II.8.23.3, holds if and only if $f \in \operatorname{int} M_*$.

(ii.1): This was explained when defining a base.

(ii.2): This is given by Fact II.8.23.3.

(iii): Suppose $B$ is a base of $M$. Then, $B$ is nonempty since $B$ intersects all nontrivial rays in $M$ emanating from the origin, and the set of these rays is nonempty since $M$ is nontrivial. Closedness of $B$ is evident. To prove that $B$ is bounded, note that by (ii.2) $f \in \operatorname{int} M_*$. Thus, there exists $r > 0$ such that $f - e \in M_*$, for all $\|e\|_2 \leq r$. Hence, $[f - e]^\top x \geq 0$ for all $x \in M$ and all $e$ with $\|e\|_2 \leq r$, implying $f^\top x \geq r\|x\|_2$ for all $x \in M$, and therefore $\|x\|_2 \leq r^{-1}$ for all $x \in B$.

Next, let $M$ be pointed and $f \in M_*$ be such that the set (8.9) is nonempty. Closedness of this set is evident. Let us show that this set is bounded if and only if $f \in \operatorname{int} M_*$. Indeed, when $f \in \operatorname{int} M_*$, $B$ is a base of $M$ by (ii.2) and therefore, as we have just seen, $B$ is bounded. For the other direction, suppose that $f \notin \operatorname{int} M_*$. Then, by Fact II.8.23.3, there exists $\bar{x} \in M \setminus \{0\}$ such that $f^\top \bar{x} = 0$. Also, as the set (8.9) is nonempty, there exists $\widehat{x} \in M$ such that $f^\top \widehat{x} = 1$. Now, observe that for any $\lambda \in [0,1)$ the vector $(1-\lambda)^{-1}[(1-\lambda)\widehat{x} + \lambda \bar{x}]$ belongs to $B$ and the norm of this vector goes to $+\infty$ as $\lambda \to 1$. But, then this implies that $B$ is unbounded, and so the proof of (iii) is completed.

(iv): Suppose $M$ is not pointed. Then, there exists a direction $e \neq 0$ such that $M$ contains the line generated by $e$; in particular $\pm e \in M$. Assume for contradiction that $d$ is an extreme direction of $M$. Then, as $M$ is a closed convex cone, $d \pm te \in M$ for all $t \in \mathbf{R}$. Thus, as $M$ is a cone, we have $d_\pm(t) := \frac{1}{2}[d \pm te] \in M$ for all $t$. Let us first suppose that $e$ is not collinear to $d$, then for any $t \neq 0$, the vector $d_\pm(t)$ is not a nonnegative multiple of $d$, but then this contradicts $d$ being an extreme direction of $M$. So, we now suppose that $e$ is collinear to $d$. But, in this case, for large enough $t$, one of the vectors $d_\pm(t)$, while being a multiple of $d$, is not a *nonnegative* multiple of $d$, which again is impossible. Thus, when $M$ is not pointed, $M$ does not have extreme rays.

Now let $M$ be pointed, and let the set $B$ given by (8.9) be a base of $M$ (a base does exist by (i.3)). $B$ is a nonempty closed and bounded convex set by (iii). Let us verify that the rays $\mathbf{R}_+(d)$ spanned by the extreme points of $B$ are exactly the extreme rays of $M$. First, suppose $d \in \operatorname{Ext}(B)$, and let us prove that $d$ is an extreme direction of $M$. Indeed, let $d = d_1 + d_2$ for some $d_1, d_2 \in M$; we should prove that $d_1, d_2$ are nonnegative multiples of $d$. There is nothing to prove when one of the vectors $d_1, d_2$ is zero, so we assume that both $d_1, d_2$ are nonzero. Then, since $B$ is a base, by (ii.1) we have $\alpha_i := f^\top d_i > 0$, $i = 1, 2$. Moreover, $\alpha_1 + \alpha_2 = f^\top d = 1$. Setting $\overline{d}_i := \alpha_i^{-1} d_i$, $i = 1, 2$, we have $\overline{d}_i \in B$, $i = 1, 2$, and $\alpha_1 \overline{d}_1 + \alpha_2 \overline{d_2} = d_1 + d_2 = d$. Recalling that $d$ is an extreme point of $B$ and $\alpha_i > 0$, $i = 1, 2$, we conclude that $\overline{d}_1 = \overline{d}_2 = d$, that is, $d_1$ and $d_2$ are positive multiples of $d$, as claimed. For the reverse direction, let $d$ be an extreme direction of $M$. We need to prove that the intersection of the ray $\mathbf{R}_+(d)$ and $B$ (this intersection is nonempty since $d \in M \setminus \{0\}$) is an extreme point of $B$. Passing from extreme direction $d$ to its positive multiple, we can assume that $d \in B$. To prove that $d \in \operatorname{Ext}(B)$, assume that there exists $h$ such that $d \pm h \in B$ and let us verify that $h = 0$. Indeed, as $d \in B$ we have $f^\top d = 1$, while from $d \pm h \in B$ we conclude that $f^\top h = 0$. Therefore, when $h \neq 0$, $h$ is not a multiple of $d$, whence the vectors $d \pm h$ are not multiples of $d$. On the other hand, both of the vectors $d \pm h$ belong to $M$ and $d$ is their average, which contradict the fact that $d$ is an extreme direction of $M$. Thus, $h = 0$, as claimed. □

**Fact II.8.38** Let $M$ be a convex set in $\mathbf{R}^n$ containing the origin. Then,

1. $\operatorname{Polar}(M) = \operatorname{Polar}(\operatorname{cl} M)$;
2. $M$ is bounded if and only if $0 \in \operatorname{int}(\operatorname{Polar}(M))$;
3. $\operatorname{int}(\operatorname{Polar}(M)) \neq \varnothing$ if and only if $M$ does not contain straight lines;
4. Assume that $M$ is closed. Then $M$ is a closed cone if and only if $\operatorname{Polar}(M)$ is a closed cone. If $M$ is a cone (not necessarily closed), then

$$\operatorname{Polar}(M) = \left\{a \in \mathbf{R}^n : \ a^\top x \leq 0, \forall x \in M\right\} = -M_*. \tag{8.10}$$

Proof.

1. This follows immediately from $\sup_{x \in M} a^\top x = \sup_{x \in \operatorname{cl} M} a^\top x$.
2. Suppose $M$ is bounded. Then, by Cauchy-Schwarz inequality all vectors $y$ with small enough norms satisfy $y \in \operatorname{Polar}(M)$ and so $0 \in \operatorname{int}(\operatorname{Polar}(M))$. To see the reverse direction, suppose $0 \in \operatorname{int}(\operatorname{Polar}(M))$. Note that $\operatorname{cl} M$ is a closed convex set and by item 1, we have

Polar $(M)$ = Polar $(\operatorname{cl} M)$, so $0 \in \operatorname{int}(\text{Polar}\,(\operatorname{cl} M))$, i.e., Polar $(\operatorname{cl} M)$ contains a ball centered at the origin with some radius $\rho > 0$. Then, by Proposition II.8.37, we have $\operatorname{cl} M =$ Polar (Polar $(\operatorname{cl} M)$) = Polar (Polar $(M)$), which implies that $x \in \operatorname{cl} M =$ Polar (Polar $(M)$) only if $1 \geq \sup_{y\in\mathbf{R}^n} \left\{y^\top x : \|y\|_2 \leq \rho\right\} = \rho\|x\|_2$. Thus, for all $x \in \operatorname{cl} M$ we have $\|x\|_2 \leq 1/\rho$.

3. By item 1, the polar remains intact when passing from $M$ to $\operatorname{cl} M$; by Lemma II.8.9, a nonempty convex set $M$ contains a straight line if and only if $\operatorname{cl} M$ is so. Thus, we lose nothing when assuming in the rest of the proof that $M$ is closed.
   Assume, first, that $M$ contains a straight line, and let us prove that int Polar $(M) = \varnothing$. Indeed, when the closed convex set $M$ contains a line, it contains a parallel line $\ell$ passing through $0 \in M$ (Lemma II.8.8), implying that Polar $(M) \subset$ Polar $(\ell)$. Since $\ell$ is a one-dimensional linear subspace of $\mathbf{R}^n$, Polar $(\ell)$ is the orthogonal complement $\ell^\perp$ of $\ell$, so that int Polar $(\ell) = \operatorname{int} \ell^\perp = \varnothing$, whence int Polar $(M) = \varnothing$ as well.
   Now let int Polar $(M) = \varnothing$, and let us prove that $M$ contains a straight line. Assume that it is not the case, and let us lead this assumption to a contradiction. Since $M$ does not contain lines, the closed cone $K = \operatorname{Rec}(M)$ is pointed, so that its dual cone $K_*$ has a nonempty interior (Fact II.8.33(i.2)). Thus, there exists $\overline{f}$ such that the ball of radius $2r > 0$ centered at $\overline{f}$ is contained in $K_*$, implying that $z^\top(f + e) \geq 0$ whenever $z \in K$, $\|e\|_2 \leq r$ and $\|f - \overline{f}\|_2 \leq r$. As a result,
   $$\forall(z \in K, f \in B := \{f : \|f - \overline{f}\|_2 \leq r\}) : f^\top z \geq r\|z\|_2. \tag{$*$}$$
   Now let
   $$C = \sup_{f\in -B, z\in M} f^\top z;$$
   we claim that $C < \infty$. Taking this claim for granted, observe that $C < \infty$ implies, by homogeneity, that $\sup_{f\in-\epsilon B, z\in M} f^\top z \leq \epsilon C$ for all $\epsilon > 0$, whence for properly selected small positive $\epsilon$ the ball $-\epsilon B$ is contained in Polar $(M)$, implying int Polar $(M) \neq \varnothing$, which is a desired contradiction.
   It remains to justify the above claim. To this end assume that $C = +\infty$, and let us lead this assumption to a contradiction. When $C = +\infty$, there exists a sequence $f_i \in -B$ and $z_i \in M$ such that $f_i^\top z_i \to +\infty$, $i \to \infty$, implying, due to $f_i \in -B$, that $\|z_i\|_2 \to \infty$ as $i \to \infty$. Passing to a subsequence, we can assume that $z_i/\|z_i\|_2 \to h$, $i \to \infty$; by its origin, $h$ is an asymptotic direction of $M$ and therefore is a unit vector from $K$ (Fact II.8.13(ii)). Assuming w.l.o.g. $z_i \neq 0$ for all $i$, we have
   $$f_i^\top z_i = \|z_i\|_2 \Bigg[\underbrace{f_i^\top h}_{\alpha_i} + \underbrace{f_i^\top [z_i/\|z_i\|_2 - h]}_{\beta_i}\Bigg]. \tag{!}$$
   As $i \to \infty$, $f_i \in -B$ remain bounded and $[z_i/\|z_i\|_2 - h] \to 0$, implying that $\beta_i \to 0$, $i \to \infty$, while $(*)$ combines with $h \in K$, $\|h\|_2 = 1$, and $f_i \in -B$ to imply that $\alpha_i \leq -r < 0$. Thus, $\alpha_i + \beta_i \leq -r/2$ for large enough values of $i$, so that (!) taken together with $\|z_i\|_2 \to \infty$, $i \to \infty$ says that $f_i^\top z_i \to -\infty$, $i \to \infty$, contradicting the origin of $f_i$ and $z_i$. Thus, $C < \infty$, as claimed. Verification of item 3 is complete.
4. Clearly, when $M$ is a nonempty conic set, the relation $y^\top x \leq 1$ for all $x \in M$ is exactly the same as $y^\top x \leq 0$ for all $x \in M$. Hence, when $M$ is a cone, its polar is a closed cone given by (8.10). On the other hand, when $M$ contains the origin and is convex and closed, it is the polar of its polar, so that when this polar is a cone, $M$ itself is a closed cone (by the just proved part of item 4 as applied to Polar $(M)$ in the role of $M$).

□

# Part III

# Convex Functions

# 13

# First acquaintance with convex functions

## 13.1 Definition and examples

**Definition** III.13.1 [Convex function] A function $f : Q \to \mathbf{R}$ defined on a subset $Q$ of $\mathbf{R}^n$ and taking real values is called *convex*, if

- the domain $Q$ of the function is convex, and
- for every $x, y \in Q$ and every $\lambda \in [0, 1]$ one has

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y). \tag{13.1}$$

The function $f$ is called *strictly convex* whenever the above inequality holds strictly for every $x \neq y$ and $0 < \lambda < 1$.

A function $f$ such that $-f$ is convex is called *concave*. In particular, the domain $Q$ of a concave function $f$ should be convex, and the function $f$ itself should satisfy the inequality opposite to (13.1), i.e.,

$$f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y), \quad \forall x, y \in Q \text{ and } \forall \lambda \in [0, 1].$$

**Example** III.13.1 The simplest example of a convex function is an *affine function*

$$f(x) = a^\top x + b,$$

which is simply the sum of a linear form and a constant. This function is clearly convex on the entire space, and in this case "convexity inequality" holds as an equality everywhere. In fact, an affine function is both convex and concave. Moreover, it is easily seen that a function which is both convex and concave on the entire space must be affine.

**Example** III.13.2 Here are several elementary examples of "nonlinear" convex functions of one variable:

- functions convex on the entire axis:
  $x^{2p}$, where $p$ is a positive integer;
  $\exp(x)$;
  $\exp(-x)$;
- functions convex on the nonnegative ray:
  $x^p$, where $p \geq 1$;
  $-x^p$, where $0 \leq p \leq 1$;

$x \ln x$;

- functions convex on the positive ray:

  $1/x^p$, where $p > 0$;

  $-\ln x$.

At the moment it is not clear why these functions are convex. We will soon derive a simple analytic criterion for detecting convexity which will immediately demonstrate that the above functions indeed are convex.

A very convenient equivalent definition of a convex function is in terms of its *epigraph*. Given a real-valued function $f$ defined on a subset $Q$ of $\mathbf{R}^n$, we define its epigraph as the set

$$\mathrm{epi}\{f\} := \left\{[x;t] \in \mathbf{R}^{n+1} : x \in Q,\ t \geq f(x)\right\}.$$

Geometrically, to define the epigraph, we plot the *graph* of the function, i.e., the surface $\{(x,t) \in \mathbf{R}^{n+1} : \ x \in Q,\ t = f(x)\}$ in $\mathbf{R}^{n+1}$, and add to this surface all points which are "above" it. Epigraph allows us to give an equivalent, more geometrical, definition of a convex function as follows.

**Proposition** III.13.2 [Epigraph based definition of convex functions] A function $f : Q \to \mathbf{R}$ defined on a subset $Q$ of $\mathbf{R}^n$ is convex if and only if its epigraph is a convex set in $\mathbf{R}^{n+1}$.

**Proof.** Let $f : Q \to \mathbf{R}$ be convex, we will show that $\mathrm{epi}(f)$ is convex. Consider any two points $[x';t']$, $[x'';t'']$ from $\mathrm{epi}\{f\}$ and any $\lambda \in [0,1]$. Then, $x', x'' \in Q$ and

$$\lambda[x';t'] + (1-\lambda)[x'';t''] = [\underbrace{\lambda x' + (1-\lambda)x''}_{:=x};\underbrace{\lambda t' + (1-\lambda)t''}_{:=t}],$$

and since $f$ is convex, we see that $Q$ is convex and $x \in Q$. Moreover, by definition $t = \lambda f(x') + (1-\lambda)f(x'') \geq f(x)$, where the inequality follows from convexity of $f$. Then, $[x;t] \in \mathrm{epi}\{f\}$. Thus, $\mathrm{epi}\{f\}$ is convex.

For the other direction, suppose that $\mathrm{epi}\{f\}$ is convex. Consider any $x', x'' \in Q$ and $\lambda \in [0,1]$. Then, we have $[x';f(x')] \in \mathrm{epi}\{f\}$ and $[x'';f(x'')] \in \mathrm{epi}\{f\}$ by definition of $\mathrm{epi}\{f\}$. Since $\mathrm{epi}\{f\}$ is convex, the point

$$\lambda[x';f(x')] + (1-\lambda)[x'';f(x'')] = [\lambda x' + (1-\lambda)x'';\ \lambda f(x') + (1-\lambda)f(x'')]$$

is in $\mathrm{epi}\{f\}$ as well. This, by definition of $\mathrm{epi}\{f\}$, implies the relations $\lambda x' + (1-\lambda)x'' \in Q$ and $f(\lambda x' + (1-\lambda)x'') \leq \lambda f(x') + (1-\lambda)f(x'')$, so that $f$ is convex. □

**More examples of convex functions: norms.** Equipped with Proposition III.13.2, we can extend our initial list of convex functions (affine functions and several one-dimensional functions) with more examples, namely *norms*. Let $\|x\|$ be a norm on $\mathbf{R}^n$ (see section 1.1.2). So far, we encountered three examples of norms: the Euclidean ($\ell_2$-) norm $\|x\|_2 = \sqrt{x^\top x}$, the $\ell_1$-norm $\|x\|_1 = \sum\limits_i |x_i|$ and the $\ell_\infty$-norm $\|x\|_\infty =$

$\max_i |x_i|$. It was also claimed (although not proved) that these are three members from an infinite family of norms

$$\|x\|_p := \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}, \quad \text{where } 1 \le p \le \infty$$

(the right hand side of the latter relation for $p = \infty$ is, *by definition*, $\max_i |x_i|$).

We say that a function $f : \mathbf{R}^n \to \mathbf{R}$ is *positively homogeneous of degree 1* if it satisfies

$$f(tx) = tf(x), \quad \forall x \in \mathbf{R}^n,\ t \ge 0.$$

Also, we say that the function $f : \mathbf{R}^n \to \mathbf{R}$ is *subadditive* if it satisfies

$$f(x + y) \le f(x) + f(y), \quad \forall x, y \in \mathbf{R}^n.$$

Note that every norm is positively homogeneous of degree 1 and subadditive. We are about to prove that all such functions (in particular, all norms) are convex:

**Proposition** III.13.3 Let $\pi(x)$ be a real-valued function on $\mathbf{R}^n$ which is positively homogeneous of degree 1. Then, $\pi$ is convex if and only if it is subadditive.

**Proof.** Note that the epigraph of a positively homogeneous of degree 1 function $\pi$ is a conic set since for any $\lambda \ge 0$ we have $[x; t] \in \mathrm{epi}\{\pi\}$ implies $\lambda[x; t] \in \mathrm{epi}\{\pi\}$. Moreover, by Proposition III.13.2 $\pi$ is convex if and only if $\mathrm{epi}\{\pi\}$ is convex. It is clear that a conic set is convex if and only if it contains the sum of every pair of its elements (why ?). This latter property is satisfied for the epigraph of a real-valued function if and only if the function is subadditive (evident). □

## 13.2 Jensen's inequality

The following basic observation is, we believe, one of the most useful observations ever made.

**Proposition** III.13.4 [Jensen's inequality] Let $f : Q \to \mathbf{R}$ be convex. Then, for every convex combination of points $x^i$ from $Q$, i.e.,

$$\sum_{i=1}^N \lambda_i x^i,$$

for some $\lambda \in \mathbf{R}^N_+$ satisfying $\sum_{i=1}^N \lambda_i = 1$, we have

$$f\left(\sum_{i=1}^N \lambda_i x^i\right) \le \sum_{i=1}^N \lambda_i f(x^i).$$

**Proof.** Note that the points $[x^i, f(x^i)]$ belong to the epigraph of $f$. As $f$ is convex, its epigraph is a convex set. Then, for any $\lambda \in \mathbf{R}^N_+$ satisfying $\sum_{i=1}^N \lambda_i = 1$, we have that the corresponding convex combination of the points given by

$$\sum_{i=1}^N \lambda_i [x^i; f(x^i)] = \left[\sum_{i=1}^N \lambda_i x^i; \sum_{i=1}^N \lambda_i f(x^i)\right]$$

also belongs to $\mathrm{epi}\{f\}$. By definition of the epigraph, this means exactly that $\sum\limits_{i=1}^N \lambda_i f(x^i) \geq f(\sum\limits_{i=1}^N \lambda_i x^i)$. □

Note that the definition of convexity of a function $f$ is exactly the requirement on $f$ to satisfy the Jensen inequality for the case of $N = 2$. We see that to satisfy this inequality for $N = 2$ is the same as to satisfy it for *all* $N \geq 2$.

**Remark** III.13.5 An instructive interpretation of Jensen's inequality is as follows: Given a convex function $f$, consider a discrete random variable $x$ taking values $x^i \in \mathrm{Dom}\, f$, $i \leq N$, with probabilities $\lambda_i$. Then,

$$f(\mathbf{E}[x]) \leq \mathbf{E}[f(x)],$$

where $\mathbf{E}[\cdot]$ stands for the expectation operator. The resulting inequality, under mild regularity conditions, holds true for general type random vectors $x$ taking values in $\mathrm{Dom}\, f$ with probability 1.

## 13.3 Convexity of sublevel sets

The *sublevel set* of a function $f : Q \to \mathbf{R}$ given by $\alpha \in \mathbf{R}$ is defined as

$$\mathrm{lev}_\alpha(f) := \{x \in Q : f(x) \leq \alpha\}.$$

We have the following simple yet useful observation on the sublevel sets of convex functions.

**Proposition** III.13.6 [Convexity of sublevel sets] Let $f : Q \to \mathbf{R}$ be convex. Then, for every $\alpha \in \mathbf{R}$, the sublevel set of $f$ given by $\alpha \in \mathbf{R}$, i.e., $\mathrm{lev}_\alpha(f)$, is convex.

**Proof.** Suppose $x, y \in \mathrm{lev}_\alpha(f)$ and $\lambda \in [0, 1]$. Then, from convexity of $f$, we have $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \leq \lambda\alpha + (1-\lambda)\alpha = \alpha$, so that $\lambda x + (1-\lambda)y \in \mathrm{lev}_\alpha(f)$. □

It is important to note that the convexity of sublevel sets does *not* characterize convex functions; there are nonconvex functions which possess this property (e.g., every monotone function on the axis has all of its sublevel sets convex). Thus, convexity of sublevel sets specifies a wider family of functions, the so called *quasiconvex* ones. The "proper" characterization of convex functions in terms of convex sets is given by Proposition III.13.2 – convex functions are exactly the functions with convex epigraphs.

## 13.4 Value of a convex function outside its domain

In its literal meaning, a function is not defined outside its domain and thus does not have any associated "value" outside its domain. Nevertheless, when speaking of *convex* functions, it is extremely convenient to think that the function outside its domain also has a value, namely, it takes the value of $+\infty$. With this convention, we revise our definition of convex functions as follows.

> *A* convex function $f$ *on* $\mathbf{R}^n$ *is a function taking values in the extended real axis* $\mathbf{R}\cup\{+\infty\}$ *and such that for all* $x, y \in \mathbf{R}^n$ *and all* $\lambda \in [0,1]$ *one has*
> $$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y). \tag{13.2}$$

When $f$ is taking values in the extended real axis, some terms in the inequality (13.2) may involve infinities. In such a case, the left- and right-hand side values of this inequality as well as its validity status is determined according to the standard conventions on operations of summation, multiplication, and comparison in the "extended real line" $\mathbf{R}\cup\{+\infty\}\cup\{-\infty\}$. These conventions are as follows:

- Operations with real numbers are understood in their usual sense.
- The sum of $+\infty$ and a real number, same as the sum of $+\infty$ and $+\infty$ is $+\infty$. Similarly, the sum of a real number and $-\infty$, same as the sum of $-\infty$ and $-\infty$ is $-\infty$. The sum of $+\infty$ and $-\infty$ is undefined.
- The product of a real number and $+\infty$ is $+\infty$, $0$ or $-\infty$, depending on whether the real number is positive, zero or negative, and similarly for the product of a real number and $-\infty$. The product of two "infinities" is again infinity, with the usual rule for assigning the sign to the product.
- Finally, any real number is $< +\infty$ and $> -\infty$, and of course $-\infty < \infty$.

The set where a function $f$ taking values in the extended real axis is finite is called the *domain* of $f$ and is denoted by $\mathrm{Dom}\, f$. Based on our revised definition of convex functions on the extended real axis, the function $f$ that is defined to be identically equal to $+\infty$ is a legitimate convex function. A convex function with nonempty domain (that is, a convex function which is not identically $+\infty$) is called *proper*.

When $f$ takes all its values in $\mathbf{R}\cup\{+\infty\}$, the inequality (13.2) is automatically valid when $\lambda = 0$ or $\lambda = 1$; when $0 < \lambda < 1$, it is automatically valid when $x = y$, same as when at least one of the points $x, y$ is not in $\mathrm{Dom}\, f$. Thus, we arrive at the following equivalent definition of convex functions:

> *Function* $f : \mathbf{R}^n \to \mathbf{R}\cup\{+\infty\}$ *is convex if and only if the inequality* (13.2) *holds for every* $x \ne y$, $x, y \in \mathrm{Dom}\, f$ *and for every* $\lambda \in (0,1)$.

Note that our initial definition of convex function included the requirement for the domain of the function to be convex; our new, equivalent, definition of convex function does not include such a requirement —after $f$ is extended outside of its domain by $+\infty$, inequality (13.2) automatically takes care of the convexity of $\mathrm{Dom}\, f$.

The simplest function with a given domain $Q$ is identically zero on $Q$ and identically

$+\infty$ outside of $Q$. This function, called the *characteristic* (a.k.a. *indicator*) *function of* $Q$ [1] is convex if and only if $Q$ is a convex set.

It is convenient to think of a convex function as of something which is defined everywhere, since it saves a lot of words. For example, with this convention we can write $f+g$ ($f$ and $g$ are convex functions on $\mathbf{R}^n$), and everybody will understand what is meant. Without this convention, we were supposed to add to this expression the following explanation as well: "$f+g$ is a function with the domain being the intersection of those of $f$ and $g$, and in this intersection it is defined as $(f+g)(x) = f(x)+g(x)$."

[1] This terminology is standard for Convex Analysis; in other areas of Math, characteristic, a.k.a. indicator, function of a set $Q \subset \mathbf{R}^n$ is defined as the function equal to 1 on the set and to 0 outside of it.

# 14

# How to detect convexity

In an optimization problem

$$\min_x \{f(x) : \ g_j(x) \leq 0, \, j = 1, \ldots, m\}$$

convexity of the objective function $f$ and the constraint functions $g_i$ is crucial. Indeed, convex problems — those with convex $f$ and $g_j$ — possess nice theoretical properties with important practical implications. For example, the local *necessary* optimality conditions for these problems are *sufficient for global optimality*. Moreover, much more importantly, convex problems can be efficiently (both in theoretical and, to some extent, in the practical meaning of the word) solved, which is not, unfortunately, the case for general nonconvex problems. This is why it is so important to know how to detect convexity of a given function.

The scheme of our investigation is typical for mathematics. Let us start with the example which you know from Analysis. How do you detect continuity of a function? Of course, there is a definition of continuity in terms of $\epsilon$ and $\delta$, but it would be an actual disaster if each time we need to prove continuity of a function, we were supposed to write down the proof that "for every positive $\epsilon$ there exists positive $\delta$ such that ...". In fact, we use another approach: we list once forever a number of standard operations which preserve continuity (like addition, multiplication, taking superpositions, etc.) and point out a number of standard examples of continuous functions (like the power function, the exponent, etc.). Note that both steps, proving that the operations in the list preserve continuity and proving that the standard functions are continuous, take certain effort and indeed is done in $\epsilon - \delta$ terms. But, after investing in this effort once, typically proving continuity of a given function becomes a much simpler task: it suffices to demonstrate that the function can be obtained, in finitely many steps, from our "raw materials" (the standard functions which are known to be continuous) by applying our "machinery" (the combination rules which preserve continuity). Normally, this demonstration is given by a single word "evident" or is even understood by default.

This is exactly the case with convexity. We will next point out the list of operations which preserve convexity and a number of standard convex functions.

## 14.1 Operations preserving convexity of functions

We start with the following basic operations preserving convexity of functions:

- *Stability under taking nonnegative weighted sums*: if $f, g$ are convex functions on $\mathbf{R}^n$, then their linear combination $\lambda f + \mu g$ with *nonnegative* coefficients $\lambda, \mu \in \mathbf{R}_+$ is also convex.

  [This can be verified straightforwardly using the convex function definition.]
- *Stability under affine substitutions of the argument*: given a convex function $f$ on $\mathbf{R}^n$ and an affine mapping $x \mapsto Ax + b$ from $\mathbf{R}^m$ into $\mathbf{R}^n$, the superposition $f(Ax + b)$ is convex.

  [This can be proved directly by verifying the convex function definition or by noting that the epigraph of the superposition is the inverse image of the epigraph of $f$ under an affine mapping.]
- *Stability under taking pointwise supremum*: given any nonempty (and possibly infinite!) family of convex functions $\{f_\alpha(\cdot)\}_{\alpha \in \mathcal{A}}$ on $\mathbf{R}^n$, their supremum $\sup\limits_{\alpha \in \mathcal{A}} f_\alpha(\cdot)$ is convex.

  [Note that the epigraph of the supremum is clearly the intersection of the epigraphs of the functions from the family. Then, recall that the intersection of every family of convex sets is convex.]
- *"Convex Monotone superposition"*: Let $f(x) := [f_1(x), \ldots, f_K(x)]$ be a vector-map with convex component functions $f_i : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ , and let $F$ be a convex function on $\mathbf{R}^K$. Suppose $F$ is *monotone nondecreasing*, i.e., for any $z, z' \in \mathbf{R}^K$ satisfying $z \leq z'$ we always have $F(z) \leq F(z')$. Then, the superposition function given by

$$\phi(x) := F(f(x)) = F(f_1(x), \ldots, f_K(x)) : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$$

is convex.

Here, note that the expression $F(f_1(x), \ldots, f_K(x))$ makes no evident sense at a point $x$ where some of $f_i$'s take the value of $+\infty$. At a such point, *by definition*, we assign the value of $+\infty$ to the superposition function.

Let us now justify the convex monotone composition rule. Consider any $x, x' \in \operatorname{Dom} \phi$. Then, $z := f(x)$ and $z' := f(x')$ are vectors from $\mathbf{R}^K$ which belong to $\operatorname{Dom} F$. Due to the convexity of the components of $f$, for any $\lambda \in (0, 1)$ we have the vector inequality

$$f(\lambda x + (1 - \lambda)x') \leq \lambda z + (1 - \lambda)z'.$$

In particular, the left hand side in this inequality is a vector from $\mathbf{R}^K$, i.e., it has no "infinite entries," and we may further use the monotonicity of $F$ to arrive at

$$\phi(\lambda x + (1 - \lambda)x') = F(f(\lambda x + (1 - \lambda)x')) \leq F(\lambda z + (1 - \lambda)z').$$

Moreover, using the convexity of $F$ we deduce

$$F(\lambda z + (1 - \lambda)z') \leq \lambda F(z) + (1 - \lambda)F(z').$$

Then, combining these two inequalities and noting that $F(z) = F(f(x)) = \phi(x)$ and $F(z') = F(f(x')) = \phi(x')$, we arrive at the desired convexity relation

$$\phi(\lambda x + (1 - \lambda)x') \leq \lambda \phi(x) + (1 - \lambda)\phi(x').$$

Imagine how many extra words would be necessary here if there were no convention on the value of a convex function outside its domain!

In the Convex Monotone superposition rule, monotone nondecreasing property of $F$ is crucial. (Look what happens when $n = K = 1$, $f_1(x) = x^2$, $F(z) = -z$). This rule, however, admits the following two useful variants where the monotonicity requirement is somehow relaxed (the justifications of these variants are left to the reader):

- *"Convex Affine superposition"*: Let $F(z)$ be a convex function on $\mathbf{R}^K$, and let the functions $f_i(x)$, $i \leq K$, be convex functions on $\mathbf{R}^n$. Suppose that for some $k \leq K$ the functions $f_1, \ldots, f_k$ are affine, and the function $F(z)$ is nondecreasing in the entries $z_s$ of $z$ with indices $s > k$. Then, the function $F(f_1(x), \ldots, f_K(x))$ is convex.
- Let $F(z)$ be a convex function on $\mathbf{R}^K$, and let the functions $f_i(x)$, $i \leq K$, be convex functions on $\mathbf{R}^n$. Define $f(x) := [f_1(x); \ldots; f_K(x)]$. Let $Y$ be a convex set in $\mathbf{R}^K$ such that $f(x) \in Y$ whenever all entries in $f(x)$ are finite. Suppose that for some $k \leq K$ the functions $f_1, \ldots, f_k$ are affine. Assume, next, that $F(z)$ is nondecreasing in only the entries $z_s$, $s > k$, of $z$ on $Y$, i.e., $F(z') \geq F(z)$ whenever $z', z$ are such that $z', z \in Y$ and $z'_s = z_s$ for all $s \leq k$ and $z'_s \geq z_s$ for all $s > k$. Then, $F(f(x))$ is convex on $\mathbf{R}^n$.
  *Example:* Let $f_i(x)$ be convex, $i \leq K$, and $F(z) := \|z\|_1$. Note that in general the function $F(f(x))$ is *not* necessarily convex (look what happens when $n = K = 1$ and $f_1(x) = x^2 - 1$). However, $F(f(x))$ is convex provided that $f_i(x)$ *are not just convex, but are nonnegative as well* (set $Y = \mathbf{R}^K_+$). More generally, the functions $\|f(x)\|_p$, $1 \leq p \leq \infty$, are convex provided that each component $f_i : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$, $i \leq K$ of the vector function $f(x) = [f_1(x); \ldots; f_K(x)]$ is convex *and nonnegative*.

We close this section with two more convexity preserving operations:

- *Stability under partial minimization*: if $f(x, y) : \mathbf{R}^n_x \times \mathbf{R}^m_y \to \mathbf{R} \cup \{+\infty\}$ is convex (as a function of $z = [x; y]$; this is called *joint convexity*) and the function

$$g(x) := \inf_y f(x, y)$$

is greater than $-\infty$ everywhere, then $g$ is convex.

The justification of this is as follows. First, the only values convex functions can take are real numbers and $+\infty$, and in the case of $g$ this is assumed. Now, consider any $x, x' \in \operatorname{Dom} g$ and any $\lambda \in [0, 1]$. Define $x'' := \lambda x + (1 - \lambda)x'$. We need to show that $g(x'') \leq \lambda g(x) + (1 - \lambda)g(x')$. There is clearly nothing to prove when $\lambda = 0$ or $\lambda = 1$, same as when $0 < \lambda < 1$ and either $x'$, or $x''$, or both do not belong to $\operatorname{Dom} g$. Thus, we assume that $x', x'' \in \operatorname{Dom} g$. For any positive $\epsilon$, we can find $y_\epsilon$ and $y'_\epsilon$ such that $[x; y_\epsilon] \in \operatorname{Dom} f$, $[x'; y'_\epsilon] \in \operatorname{Dom} f$ and $g(x) + \epsilon \geq f(x, y_\epsilon)$,

$g(x') + \epsilon \geq f(x', y'_\epsilon)$. Taking weighted sum of these two inequalities, we get

$$\begin{aligned}\lambda g(x) + (1-\lambda)g(x') + \epsilon &\geq \lambda f(x, y_\epsilon) + (1-\lambda) f(x', y'_\epsilon)\\ &\geq f(\lambda x + (1-\lambda)x', \lambda y_\epsilon + (1-\lambda)y'_\epsilon)\\ &= f(x'', \lambda y_\epsilon + (1-\lambda)y'_\epsilon),\end{aligned}$$

where the last inequality follows from the convexity of $f$. By definition of $g(x'')$ we have $f(x'', \lambda y_\epsilon + (1-\lambda)y'_\epsilon) \geq g(x'')$, and thus we get $\lambda g(x) + (1-\lambda)g(x') + \epsilon \geq g(x'')$. In particular, $x'' \in \operatorname{Dom} g$ (recall that $x, x' \in \operatorname{Dom}(g)$ and thus $g(x), g(x') \in \mathbf{R}$). Moreover, since the resulting inequality is valid for all $\epsilon > 0$, we come to $g(x'') \leq \lambda g(x) + (1-\lambda)g(x')$, as required.

- *Perspective transform of a convex function*: Given a convex function $f$ on $\mathbf{R}^n$, we define the function $g(x,y) := yf(x/y)$ with the domain $\{[x;y] \in \mathbf{R}^{n+1} : \ y > 0, x/y \in \operatorname{Dom} f\}$ to be its *perspective function*. The perspective function of a convex function is convex.

  Let us first examine a direct justification of this. Consider any $[x'; y']$ and $[x''; y'']$ from $\operatorname{Dom} g$ and any $\lambda \in [0,1]$. Define $x := \lambda x' + (1-\lambda)x''$, $y := \lambda y' + (1-\lambda)y''$. Then, $y > 0$. We also define $\lambda' := \lambda y'/y$ and $\lambda'' := (1-\lambda)y''/y$, so that $\lambda', \lambda'' \geq 0$ and $\lambda' + \lambda'' = 1$. As $f$ is convex, we deduce $x/y = \lambda x'/y + (1-\lambda)x''/y = \lambda' x'/y' + \lambda'' x''/y'' = \lambda' x'/y' + (1-\lambda')x''/y'' \in \operatorname{Dom} f$ and $f(x/y) \leq \lambda' f(x'/y') + (1-\lambda') f(x''/y'')$. Thus, as $y > 0$, we arrive at $yf(x/y) \leq y\lambda' f(x'/y') + y(1-\lambda')f(x''/y'') = \lambda[y' f(x'/y')] + (1-\lambda)[y'' f(x''/y'')]$, that is, $g(x,y) \leq \lambda g(x', y') + (1-\lambda) g(x'', y'')$.

  Here is an alternative smarter justification. There is nothing to prove when $\operatorname{Dom} f = \varnothing$. So, suppose that $\operatorname{Dom} f \neq \varnothing$. Consider the epigraph $\operatorname{epi}(f) = \{[x;s] : \ s \geq f(x)\}$ and the perspective transform of this nonempty convex set which is given by (see section 1.5)

$$\begin{aligned}\operatorname{Persp}(\operatorname{epi}\{f\}) &:= \left\{[[x;s];t] \in \mathbf{R}^{n+2} : \ t > 0, \ [x/t; s/t] \in \operatorname{epi}\{f\}\right\}\\ &= \left\{[x;s;t] \in \mathbf{R}^{n+2} : \ t > 0, \ s/t \geq f(x/t)\right\}\\ &= \left\{[x;s;t] \in \mathbf{R}^{n+2} : \ t > 0, \ s \geq tf(x/t)\right\}\\ &= \left\{[x;s;t] \in \mathbf{R}^{n+2} : \ t > 0, \ x/t \in \operatorname{Dom} f, \ s \geq tf(x/t)\right\}\\ &= \left\{[x;s;t] \in \mathbf{R}^{n+2} : \ t > 0, \ [x;t] \in \operatorname{Dom} g, \ s \geq g(x,t)\right\}\\ &= \left\{[x;s;t] \in \mathbf{R}^{n+2} : \ [x;t;s] \in \operatorname{epi}\{g\}\right\},\end{aligned}$$

  where the second from last equality follows from the fact that by definition of $g(x,t)$, whenever $t > 0$, the inclusion $x/t \in \operatorname{Dom} f$ takes place if and only if $[x;t] \in \operatorname{Dom} g$. Thus, we observe that $\operatorname{Persp}(\operatorname{epi}\{f\})$ is nothing but the image of $\operatorname{epi}\{g\}$ under the one-to-one linear transformation $[x;t;s] \mapsto [x;s;t]$. As $\operatorname{Persp}(\operatorname{epi}\{f\})$ is a convex set (recall from section 1.5 that the perspective transform of a nonempty convex set is convex), we conclude that $g$ is convex.

Now that we know what the basic operations preserving convexity of a function are, let us look at the standard convex functions these operations can be applied to. We have already seen several examples in Example III.13.2; but we still do not know why these functions are convex. The usual way to check convexity of a "simple" (i.e., given by a simple formula) function is based on *differential criteria of convexity*, which we will examine in the next section.

## 14.2 Criteria of convexity

The definition of convexity of a function immediately reveals that *convexity is a one-dimensional property*: a function $f$ on $\mathbf{R}^n$ taking values in $\mathbf{R} \cup \{+\infty\}$ is convex if and only if its restriction on every line, i.e., every function $g : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ of the type $g(t) := f(x + th)$ with $x, h \in \mathbf{R}^n$ is convex.

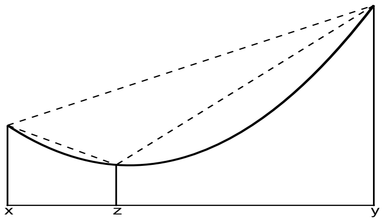


Figure III.1. Univariate convex function $f : [x, y] \to \mathbf{R}$. The average rate of change of $f$ on the entire segment $[x, y]$ is in-between the average rates of change "at the beginning," i.e., when passing from $x$ to $z$ and "at the end," i.e., when passing from $z$ to $y$.

We are about to show that any univariate function $f : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ is convex if and only if for every 3 real numbers $x < z < y$ such that $x, y \in \operatorname{Dom} f$ we have $z \in \operatorname{Dom} f$ and the average rate $\frac{f(y)-f(x)}{y-x}$ at which $f$ varies when moving from $x$ to $y$ is in-between the average rate $\frac{f(z)-f(x)}{z-x}$ at which $f$ changes "at the beginning," i.e., when moving from $x$ to $z$, and the average rate $\frac{f(y)-f(z)}{y-z}$ at which $f$ changes "at the end," i.e., when moving from $z$ to $y$, see Figure III.1:

$$\frac{f(z) - f(x)}{z - x} \leq \frac{f(y) - f(x)}{y - x} \leq \frac{f(y) - f(z)}{y - z},$$

so that

$$\begin{aligned} \frac{f(z) - f(x)}{z - x} &\leq \frac{f(y) - f(x)}{y - x}, \\ \frac{f(y) - f(x)}{y - x} &\leq \frac{f(y) - f(z)}{y - z}, \\ \frac{f(z) - f(x)}{z - x} &\leq \frac{f(y) - f(z)}{y - z}. \end{aligned} \tag{14.1}$$

As is immediately seen, every one of the three inequalities in (14.1) implies the other two.

Here is the justification of the above characterization of convexity of a univariate function $f$. Note that this convexity is nothing but the requirement that for any real numbers $x, y \in \operatorname{Dom} f$ with $x < y$ and every $\lambda \in (0, 1)$, for $z_\lambda := (1 - \lambda)x + \lambda y$ it holds $f(z_\lambda) \leq (1 - \lambda)f(x) + \lambda f(y)$, or, which is the same,

$$f(z_\lambda) - f(x) \leq \lambda(f(y) - f(x)). \tag{14.2}$$

When $\lambda \in (0, 1)$, the pair $(1, \lambda)$ is a positive multiple of the pair $(y - x, z_\lambda - x)$, thus (14.2) is equivalent to $(y - z)(f(z_\lambda) - f(x)) \leq (z_\lambda - x)(f(y) - f(x))$. Note that this inequality is the same as $\frac{f(z_\lambda)-f(x)}{z_\lambda - x} \leq \frac{f(y)-f(x)}{y-x}$. When $\lambda$ runs through the interval $(0, 1)$ the point $z_\lambda$ runs through the entire set $\{z : x < z < y\}$, and so we conclude that *$f$ is convex if and only if for every triple $x < z < y$ with $x, y \in \operatorname{Dom} f$ the first inequality in (14.1) holds true.* As every one of inequalities in (14.1) implies the other two, this justifies our "average rate of change" characterization of univariate convexity.

In the case of multivariate convex functions, we have the following immediate consequence of the preceding observations.

**Lemma** III.14.1 Let $x, x', x''$ be three distinct points in $\mathbf{R}^n$ with $x' \in [x, x'']$. Then, for any convex function $f$ that is finite on $[x, x'']$, we have

$$\frac{f(x') - f(x)}{\|x' - x\|_2} \leq \frac{f(x'') - f(x)}{\|x'' - x\|_2}. \tag{14.3}$$

**Proof.** Under the premise of the lemma, define $\phi(t) := f(x + t(x'' - x))$ and let $\lambda \in \mathbf{R}$ be such that $x' = x + \lambda(x'' - x)$. Note that $\lambda \in (0, 1)$ as $x' \in [x, x'']$ and the points $x, x', x''$ are all distinct from each other. As it was explained at the beginning of this section, the univariate function $\phi$ is convex along with $f$, and $0, 1, \lambda \in \operatorname{Dom} \phi$. Applying the first inequality in (14.1) to $\phi$ in the role of $f$ and the triple $(0, \lambda, 1)$ in the role of the triple $(x, z, y)$, we get $\frac{f(x')-f(x)}{\lambda} \leq f(x'') - f(x)$, which, due to $\lambda = \|x' - x\|_2 / \|x'' - x\|_2$, is nothing but (14.3). □

To sum up, to detect convexity of a function, in principle, it suffices to know how to detect convexity of functions of a single variable. Moreover, this latter question can be resolved by the standard Calculus tools.

### 14.2.1 Differential criteria of convexity

We have the following simple and complete characterization of convexity of smooth univariate functions from Calculus.

**Proposition** III.14.2 [Necessary and sufficient condition for convexity of smooth univariate functions] Let $(a, b)$ be an interval on the axis (where the cases of $a = -\infty$ and/or $b = +\infty$ are also possible). Then,

(i) A function $f$ that is differentiable everywhere on $(a, b)$ is convex on $(a, b)$ if and only if its derivative $f'$ is monotonically nondecreasing on $(a, b)$;

(ii) A function $f$ that is twice differentiable everywhere on $(a, b)$ is convex on $(a, b)$ if and only if its second derivative $f''$ is nonnegative everywhere on $(a, b)$.

**Proof.**

(i): We start by proving the necessity of the stated condition. Suppose that $f$ is differentiable and convex on $(a, b)$. We will prove that then $f'$ is monotonically nondecreasing. Let $x < y$ be two points from the interval $(a, b)$, and let us prove that $f'(x) \leq f'(y)$. Consider any $z \in (x, y)$. Invoking convexity of $f$ and applying (14.1), we have

$$\frac{f(z) - f(x)}{x - z} \leq \frac{f(y) - f(z)}{y - z}.$$

Passing to limit as $z \to x + 0$, we get

$$f'(x) \leq \frac{f(y) - f(x)}{y - x},$$

and passing to limit in the same inequality as $z \to y - 0$, we arrive at

$$\frac{f(y) - f(x)}{y - x} \leq f'(y),$$

and so $f'(x) \leq f'(y)$, as claimed.

Let us now prove the sufficiency of the condition in (i). Thus, we assume that $f'$ exists and is nondecreasing on $(a, b)$, and we will verify that $f$ is convex on $(a, b)$. By "average rate of change" description of the convexity of a univariate function, all we need is to verify that if $x < z < y$ and $x, y \in (a, b)$, then

$$\frac{f(z) - f(x)}{z - x} \leq \frac{f(y) - f(z)}{y - z}.$$

This is indeed evident: by the Lagrange Mean Value Theorem, the left hand side ratio is $f'(u)$ for some $u \in (x, z)$, and the right hand side one is $f'(v)$ for some $v \in (z, y)$. Since $v > u$ and $f'$ is nondecreasing on $(a, b)$, we conclude that the left hand side ratio is indeed less than or equal to the right hand side one.

(ii): This part is an immediate consequence of (i) as we know from Calculus that a differentiable function — in our case now this is the function $f'$— is monotonically nondecreasing on an interval if and only if its derivative is nonnegative on this interval.

□

Proposition III.14.2 immediately allows us to verify the convexity of functions listed in Example III.13.2. To this end, the only difficulty which we may encounter is that some of these functions (e.g., $x^p$ with $p \geq 1$, and $-x^p$ with $0 \leq p \leq 1$) are claimed to be convex on the half-interval $[0, +\infty)$, while Proposition III.14.2 talks about convexity of functions on open intervals. This difficulty can be addressed with the following simple result which allows us to extend the convexity of continuous functions beyond open sets.

**Proposition** III.14.3 Let $M$ be a convex set and let $f$ be a function with $\operatorname{Dom} f = M$. Suppose that $f$ is convex on $\operatorname{rint} M$ and is continuous on $M$, i.e.,

$$f(x_i) \to f(x), \quad \text{as } i \to \infty,$$

whenever $x_i, x \in M$ and $x_i \to x$ as $i \to \infty$. Then, $f$ is convex on $M$.

**Proof.** Consider any $x, y \in M$ and any $\lambda \in [0,1]$. Define $z := \lambda x + (1-\lambda)y$. We need to prove that

$$f(z) \leq \lambda f(x) + (1-\lambda) f(y).$$

As $x, y \in M$ and $M$ is convex, by Theorem I.1.29(iii), there exist sequences $\{x_i\}_{i\geq 1} \in \operatorname{rint} M$ and $\{y_i\}_{i\geq 1} \in \operatorname{rint} M$ converging to $x$ and to $y$, respectively. Then, the sequence $z_i := \lambda x_i + (1-\lambda) y_i$ is in $\operatorname{rint} M$ and it converges to $z$ as $i \to \infty$. Since $f$ is convex on $\operatorname{rint} M$, for all $i \geq 1$ we have

$$f(z_i) \leq \lambda f(x_i) + (1-\lambda) f(y_i).$$

By taking the limits of both sides of this inequality, noting that $f$ is continuous on $M$, and as $i \to \infty$ the sequences $x_i, y_i, z_i$ converge to $x, y, z \in M$, respectively, we obtain the desired inequality. □

**Illustration.** We are now able to justify the claim, made as early as in section 1.1.2, that the functions $\|\cdot\|_p$, $1 \leq p \leq \infty$, are norms. So far, we have verified it only for $p = 1, 2, \infty$ in Remark I.1.7. Now consider the case of $p \in (1, \infty)$. In order to prove that $\|\cdot\|_p$ is indeed a norm, using Fact I.1.8, all we need is to show that the set $V := \{x \in \mathbf{R}^n : \|x\|_p \leq 1\}$ is closed, bounded, symmetric with respect to origin, contains a neighborhood of the origin and is convex. All these except the convexity is evident. So, we will prove that $V$ is indeed convex. Define the function $f(x) := \|x\|_p^p = \sum_{i=1}^n |x_i|^p$. Then, $V = \{x \in \mathbf{R}^n : f(x) \leq 1\}$. By Proposition III.14.2, the univariate function $|x|^p$ is convex (recall that $p \in (1, \infty)$). Moreover, by "calculus of convexity" (section 14.1), $f$ is convex (as it is the sum of convex functions). Thus, $V$ is the sublevel set of a convex function, and by Proposition III.13.6, it is convex.

In the preceding illustration of convexity of norms we have started to examine multivariate functions. Let us continue our discussion of multivariate functions by examining a differential criteria for their convexity. Propositions III.14.2(ii) and III.14.3 give us the following convenient *necessary and sufficient condition* for convexity of *smooth multivariate* functions.

**Corollary** III.14.4 [Convexity criterion for smooth multivariate functions] Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a function where its domain $Q := \operatorname{Dom} f$ is a convex set with a nonempty interior. Suppose that $f$ is

- continuous on $Q$, and
- twice differentiable on $\operatorname{int} Q$.

Then, $f$ is convex on $Q$ if and only if for all $x \in \operatorname{int} Q$ its Hessian matrix $f''(x)$ is positive semidefinite, i.e.,

$$h^\top f''(x) h \geq 0, \quad \forall h \in \mathbf{R}^n.$$

That is, $f$ is convex on $Q$ if and only if the second order directional derivative of $f$ taken at any point $x \in \operatorname{int} Q$ along any direction $h \in \mathbf{R}^n$ is nonnegative, i.e., for all $x \in \operatorname{int} Q$, and $h \in \mathbf{R}^n$ we have

$$h^\top f''(x)h = \frac{d^2}{dt^2}\Big|_{t=0} f(x+th) \geq 0.$$

**Proof.** The "only if" part is evident: if $f$ is convex on $Q$ and $x \in \operatorname{int} Q$, then for any fixed direction $h \in \mathbf{R}^n$ the function $g : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ defined as

$$g(t) := f(x+th)$$

is convex in a certain neighborhood of the point $t = 0$ on the axis (recall that affine substitutions of argument preserves convexity). Since $f$ is twice differentiable in a neighborhood of $x$, the function $g$ is twice differentiable in a neighborhood of $t = 0$, as well. Thus, by Proposition III.14.2, we have $0 \leq g''(0) = h^\top f''(x)h$.

In order to prove the "if" part we need to show that every function $f : Q \to \mathbf{R} \cup \{+\infty\}$ that is continuous on $Q$ and satisfies $h^\top f''(x)h \geq 0$ for all $x \in \operatorname{int} Q$ and all $h \in \mathbf{R}^n$ is convex on $Q$

Let us first prove that $f$ is convex on $\operatorname{int} Q$. By Theorem I.1.29, $\operatorname{int} Q$ is a convex set. Since the convexity of a function on a convex set is a one-dimensional property, all we need to prove is that for any $x, y \in \operatorname{int} Q$ the univariate function $g : [0,1] \to \mathbf{R} \cup \{+\infty\}$ given by

$$g(t) := f(x + t(y-x))$$

is convex on the segment $[0,1]$. As $f$ is twice differentiable on $\operatorname{int} Q$, $g$ is continuous and twice differentiable on the segment $[0,1]$ and its second derivative is given by

$$g''(t) = (y-x)^\top f''(x + t(y-x))(y-x) \geq 0,$$

where the inequality follows from the premise on $f$. Then, by Propositions III.14.2(ii) and III.14.3, $g$ is convex on $[0,1]$. Thus, $f$ is convex on $\operatorname{int} Q$. As $f$ is convex on $\operatorname{int} Q$ and is continuous on $Q$, by Proposition III.14.3 we conclude that $f$ is convex on $Q$. □

**Corollary** III.14.5 [Sufficient condition for strict convexity of smooth functions] Consider the setting of Corollary III.14.4. Suppose, in addition, that $Q := \operatorname{Dom} f$ is open and the Hessian of $f$ is positive definite on $Q$, i.e.,

$$\frac{d^2}{dt^2} f(x+th) > 0, \quad \forall x \in Q \text{ and } \forall h \neq 0.$$

Then, $f$ is strictly convex.

**Proof.** Consider any $x, y \in Q$ with $x \neq y$ and any $\lambda \in (0,1)$. We need to show that $f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y)$. Consider the function $\phi : [0,1] \to \mathbf{R}$ given by $\phi(t) := f(tx + (1-t)y)$. Then, as $f$ is twice differentiable on $Q$, $\phi$ is twice differentiable on $[0,1]$. Moreover, based on the premise on $f$, we have $\phi''(t) > 0$ for all $t \in [0,1]$. Note that our target inequality is simply the relation $\phi(\lambda) < \lambda\phi(1) + (1-\lambda)\phi(0)$. Since $0 < \lambda < 1$, we can rewrite this target inequality as $\frac{\phi(\lambda)-\phi(0)}{\lambda} < \frac{\phi(1)-\phi(\lambda)}{1-\lambda}$. Finally, by the Mean Value Theorem and strict monotonicity of $\phi'$ we conclude that the desired target inequality holds. □

We conclude this section by highlighting that convexity of many "complicated" functions can be proved easily by the application of combination of "calculus of convexity" rules to simple functions which pass the "infinitesimal" convexity tests.

**Example** III.14.1 Consider the following *exponential posynomial* function $f : \mathbf{R}^n \to \mathbf{R}$, given by

$$f(x) = \sum_{i=1}^{N} c_i \exp(a_i^\top x),$$

where the coefficients $c_i$ are positive (this is why the function is called *posy*nomial). This function is in fact is convex on $\mathbf{R}^n$. How can we prove this?

An immediate proof is as follows:

1. The function $\exp(t)$ is convex on $\mathbf{R}$ as its second order derivative is positive as required by the infinitesimal convexity test for smooth univariate functions.
2. Thus, by stability of convexity under affine substitutions of argument, we deduce that all functions $\exp(a_i^\top x)$ are convex on $\mathbf{R}^n$.
3. Finally, by stability of convexity under taking linear combinations with nonnegative coefficients, we conclude that $f$ is convex on $\mathbf{R}^n$.

And if we were supposed to prove that the maximum of three exponential posynomials is convex? Then, all we need is to add to our three steps above the fourth one, which refers to the stability of convexity under taking pointwise supremum.

## 14.3 Important multivariate convex functions

Let us start with some simple yet important multivariate convex functions that can be detected solely based on "calculus of convexity" presented in section 14.1.

**Example** III.14.2 Let $k, n$ be two positive integers such that $k \leq n$. Consider the function $s_k : \mathbf{R}^n \to \mathbf{R}$ given by

$$s_k(x) := \sum_{i=1}^{k} x_{[i]},$$

where $x_{[i]}$ denotes the $i$-th largest entry in the vector $x$. That is, for every vector $x \in \mathbf{R}^n$, we have $x_{[1]} \geq x_{[2]} \geq \ldots \geq x_{[n]}$. By definition, $s_k(x)$ is simply the sum of $k$ largest elements in $x$. We claim that $s_k(x)$ is a convex function of $x$. Given any index set $I$, the function $\ell_I(x) := \sum_{i \in I} x_i$ is a linear function of $x$ and thus it is convex. Now, $s_k(x)$ is clearly the maximum of the linear functions $\ell_I(x)$ over all index sets $I$ with exactly $k$ elements from $\{1, \ldots, n\}$, and as such is convex.

Note also that $s_k(x)$ is a *permutation symmetric* function of $x$, that is, the value of the function $s_k(x)$ remains the same when permuting entries in its argument $x$. Taking together convexity and permutation symmetry of $s_k(x)$ will be very useful in our developments for functions of eigenvalues of symmetric matrices in chapter 18.

**Remark** III.14.6 An immediate implication of Example III.14.2 is that given a vector $x \in \mathbf{R}^n$, the function $\max_i\{x_i\}$, i.e., the value of its largest element, is convex in $x$. Thus, the function corresponding to the value of minimum element in $x$, i.e., $\min_i\{x_i\} = -\max_i\{-x_i\}$ is concave in $x$. That said, for $1 < k < n$, the "intermediate element" function, i.e., the function given by $x_{[k]}$, which stands for the $k$-largest element in $x$, is neither convex, nor concave function of $x$.

While "calculus of convexity" presented in section 14.1 is sufficient and quite practical in proving convexity of many functions, there are still several multivariate functions for which convexity seemingly cannot be extracted from this calculus and should be verified via Corollary III.14.4. Here are some important examples.

**Example** III.14.3 The function $f : \mathbf{R}^n \to \mathbf{R}$ given by

$$f(x) := \ln\left(\sum_{i=1}^n \exp(x_i)\right)$$

is convex.

Let us first verify the convexity of this function via direct computation using Corollary III.14.4. To this end, we define $p_i := \frac{\exp(x_i)}{\sum_j \exp(x_j)}$. Then, the second-order directional derivative of $f$ along the direction $h \in \mathbf{R}^n$ is given by

$$\omega := \left.\frac{d^2}{dt^2}\right|_{t=0} f(x+th) = \sum_i p_i h_i^2 - \left(\sum_i p_i h_i\right)^2.$$

Observing that $p_i > 0$ and $\sum_i p_i = 1$, we see that $\omega$ is the variance (the expectation of square minus the squared expectation) of discrete random variable taking values $h_i$ with probabilities $p_i$, and it is well known that the variance of any random variable is always nonnegative. Here is a direct verification of this fact:

$$\left(\sum_i p_i h_i\right)^2 = \left(\sum_i \sqrt{p_i}(\sqrt{p_i}h_i)\right)^2 \leq \left(\sum_i p_i\right)\left(\sum_i p_i h_i^2\right) = \sum_i p_i h_i^2,$$

where the inequality follows from Cauchy-Schwarz inequality and the last equality holds since $\sum_i p_i = 1$.

In fact, in this example we can prove convexity of $f$ via calculus of convexity as well. Recall that $\ln(z) = \min_{s\in\mathbf{R}}\{z\exp(s) - s - 1\}$. Thus, $\ln\left(\sum_i \exp(x_i)\right) = \min_{s\in\mathbf{R}}\left\{\left(\sum_i \exp(x_i+s)\right) - s + 1\right\}$. Then, by stability of convexity under partial minimization we conclude that $f$ is convex.

An even simpler proof is as follows:

$$\begin{aligned}\operatorname{epi}\left\{\ln\left(\sum_i \exp(x_i)\right)\right\} &= \left\{[x;t] : t \geq \ln\left(\sum_i \exp(x_i)\right)\right\}\\ &= \left\{[x;t] : \exp(t) \geq \sum_i \exp(x_i)\right\}\\ &= \left\{[x;t] : \sum_i \exp(x_i - t) \leq 1\right\},\end{aligned}$$

and the concluding set is convex (as a sublevel set of a convex function).

**Example** III.14.4 The function $f : \mathbf{R}^n_+ \to \mathbf{R}$ given by

$$f(x) = \prod_{i=1}^n x_i^{\alpha_i},$$

where $\alpha_i > 0$ for all $1 \leq i \leq n$ and satisfy $\sum_i \alpha_i \leq 1$, is convex.

To prove convexity of $f$ via Corollary III.14.4 all we need is to verify that for any $x \in \mathbf{R}^n$ satisfying $x > 0$ and for any $h \in \mathbf{R}^n$, we have $\frac{d^2}{dt^2}\Big|_{t=0} f(x + th) \geq 0$. Let $\eta_i := h_i/x_i$, then direct computation shows that

$$\frac{d^2}{dt^2}\bigg|_{t=0} f(x + th) = \left[(\sum_i \alpha_i\eta_i^2) - (\sum_i \alpha_i\eta_i)^2\right] f(x)$$

and as we have seen in Example III.14.3, the premise on $\alpha$ implies that

$$\left(\sum_i \alpha_i\eta_i\right)^2 = \left(\sum_i \sqrt{\alpha_i}(\sqrt{\alpha_i}\eta_i)\right)^2 \leq \left(\sum_i \alpha_i\right)\left(\sum_i \alpha_i\eta_i^2\right) \leq \sum_i \alpha_i\eta_i^2,$$

where the first inequality follows from Cauchy-Schwarz inequality and the last inequality holds since $\sum_i \alpha_i \leq 1$. Noting that $f(x) \geq 0$ whenever $x > 0$ completes the proof.

Example III.14.4 admits the following immediate extension:

**Example** III.14.5 The function $f : \operatorname{int} \mathbf{R}^n_+ \to \mathbf{R}$ given by

$$f(x) = \prod_{i=1}^n x_i^{-\alpha_i},$$

where $\alpha_i > 0$ for all $1 \leq i \leq n$, is convex.

Here "calculus of convexity" already works: the function

$$\ln(f(x)) = -\sum_{i=1}^n \alpha_i \ln(x_i)$$

is convex on $\operatorname{int} \mathbf{R}^n_+$ as the function $g(y) = -\ln y$ is convex on the positive ray. It remains to note that taking exponent preserves convexity by Convex Monotone superposition rule.

## 14.4 Gradient inequality

We next present an extremely important property of convex functions:

**Proposition** III.14.7 [Gradient inequality] Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$, $x \in \operatorname{int}(\operatorname{Dom} f)$, and let $Q$ be a convex set containing $x$. Suppose that

- $f$ is convex on $Q$, and
- $f$ is differentiable at $x$.

Let $\nabla f(x)$ be the gradient of the function $f$ at $x$. Then, the following inequality holds:

$$f(y) \geq f(x) + (y-x)^\top \nabla f(x), \quad \forall y \in Q. \tag{14.4}$$

Geometrically, this relation states that the graph of the function $f$ restricted onto the set $Q$, i.e.,

$$\left\{(y,t) \in \mathbf{R}^{n+1} : y \in \operatorname{Dom} f \cap Q,\ t = f(y)\right\}$$

is above the graph

$$\{(y,t) \in \mathbf{R}^{n+1} : t = f(x) + (y-x)^\top \nabla f(x)\}$$

of the linear form tangent to $f$ at $x$.

**Proof.** Let $y \in Q$. There is nothing to prove if $y \notin \operatorname{Dom} f$ (since then the left hand side of the gradient inequality is $+\infty$). Similarly, there is nothing to prove when $y = x$. Thus, we can assume that $y \neq x$ and $y \in \operatorname{Dom} f$. Let us set

$$y_\tau := x + \tau(y-x), \quad \text{where } 0 < \tau \leq 1,$$

so that $y_0 = x$, $y_1 = y$ and $y_\tau$ is an interior point of the segment $[x,y]$ for $0 < \tau < 1$. Applying Lemma III.14.1 to the triple $(x, x', x'')$ taken as $(x, y_\tau, y)$, we get

$$\frac{f(x+\tau(y-x)) - f(x)}{\tau\|y-x\|_2} \leq \frac{f(y)-f(x)}{\|y-x\|_2};$$

as $\tau \to +0$, the left hand side in this inequality, by the definition of the gradient, tends to $\frac{(y-x)^\top \nabla f(x)}{\|y-x\|_2}$, and so we get

$$\frac{(y-x)^\top \nabla f(x)}{\|y-x\|_2} \leq \frac{f(y)-f(x)}{\|y-x\|_2},$$

and as $\|y-x\|_2 > 0$ this is equivalent to

$$(y-x)^\top \nabla f(x) \leq f(y) - f(x).$$

Note that this inequality is exactly the same as (14.4). □

**Corollary** III.14.8 Suppose $Q$ is a convex set with a nonempty interior and $f$ is continuous on $Q$ and differentiable on $\operatorname{int} Q$. Then, $f$ is convex on $Q$ *if* *and only if* the gradient inequality (14.4) is valid for every pair $x \in \operatorname{int} Q$ and $y \in Q$.

**Proof.** Indeed, the "only if" part, i.e., the convexity of $f$ on $Q$ implying the gradient inequality for all $x \in \operatorname{int} Q$ and all $y \in Q$, is given by Proposition III.14.7. Let us prove the "if" part, i.e., establish the reverse implication. Suppose that $f$ satisfies the gradient inequality for all $x \in \operatorname{int} Q$ and all $y \in Q$, and let us verify that $f$ is convex on $Q$. As $f$ is continuous on $Q$ and $Q$ is convex, by Proposition III.14.3 it suffices to prove that $f$ is convex on $\operatorname{int} Q$. Recall also that by Theorem I.1.29 $\operatorname{int} Q$ is convex. Moreover, due to the gradient inequality, on $\operatorname{int} Q$ function $f$ is the supremum of the family of affine (and therefore convex) functions, i.e., for all $y \in \operatorname{int} Q$ we have

$$f(y) = \sup_{x \in \operatorname{int} Q} f_x(y), \quad \text{where } f_x(y) := f(x) + (y-x)^\top \nabla f(x).$$

As affine functions are convex and by stability of convexity under taking pointwise supremums, we conclude $f$ is convex on $\operatorname{int} Q$. □

### 14.5 Boundedness and Lipschitz continuity of a convex function

In this section we will discuss some local properties of convex functions such as boundedness and Lipschitz continuity. Recall that given a function $f$ and a set $Q \subseteq \operatorname{Dom} f$, we say that $f$ is *Lipschitz continuous* on $Q$ if there exists a constant $L > 0$ (referred to as the *Lipschitz constant* of $f$ on $Q$) such that

$$|f(x) - f(y)| \leq L\|x - y\|_2, \quad \forall x, y \in Q.$$

**Theorem** III.14.9 [Local boundedness and Lipschitz continuity of convex functions] Let $f$ be a convex function and let $K$ be a closed and bounded set contained in $\operatorname{rint}(\operatorname{Dom} f)$. Then, $f$ is Lipschitz continuous on $K$:

$$|f(x) - f(y)| \leq L\|x - y\|_2, \quad \forall x, y \in K. \tag{14.5}$$

for properly selected $L < \infty$. In particular, $f$ is bounded on $K$.

We shall prove this Theorem later in this section, after some preliminary effort.

**Remark** III.14.10 In Theorem III.14.9, all three assumptions on $K$, (1) closedness, (2) boundedness, and (3) $K \subseteq \operatorname{rint}(\operatorname{Dom} f)$, are essential. The following three examples illustrate their importance:

- Suppose $f(x) = 1/x$, then $\operatorname{Dom} f = (0, +\infty)$. Consider $K = (0, 1]$. We have assumptions (2) and (3) satisfied, but not (1). Note that $f$ is neither bounded nor Lipschitz continuous on $K$.
- Suppose $f(x) = x^2$, then $\operatorname{Dom} f = \mathbf{R}$. Consider $K = \mathbf{R}$. We have (1) and (3) satisfied, but not (2). Note that $f$ is neither bounded nor Lipschitz continuous on $K$.

- Suppose $f(x) = -\sqrt{x}$, then $\mathrm{Dom}\, f = [0, +\infty)$. Consider $K = [0, 1]$. We have (1) and (2) are satisfied, but not (3). Note that $f$ is not Lipschitz continuous on $K$ (indeed, we have $\lim\limits_{t\to+0} \frac{f(0)-f(t)}{t} = \lim\limits_{t\to+0} t^{-1/2} = +\infty$, while for a Lipschitz continuous $f$ the ratios $t^{-1}(f(0) - f(t))$ should be bounded). On the other hand, $f$ is bounded on $K$. With a properly chosen convex function $f$ of two variables and non-polyhedral compact domain (e.g., with $\mathrm{Dom}\, f$ being the unit circle), we can demonstrate also that lack of (3), even in presence of (1) and (2), may cause unboundedness of $f$ on $K$ as well.

Theorem III.14.9 says that a convex function $f$ is bounded on every compact (i.e., closed and bounded) subset of $\mathrm{rint}(\mathrm{Dom}\, f)$. In fact, in the case of convex functions we can make a much stronger statement on their *below* boundedness of $f$: any convex function $f$ is below bounded on *any* bounded subset of $\mathbf{R}^n$!

**Proposition** III.14.11 A convex function $f$ is bounded from below on every bounded subset of $\mathbf{R}^n$.

**Proof.** It is clear that $f$ is bounded from below at any point outside of the domain of $f$. Thus, without loss of generality we may assume that $\mathrm{Dom}\, f$ is full-dimensional and that $0 \in \mathrm{int}(\mathrm{Dom}\, f)$. By Theorem III.14.9, there exists a neighborhood $U$ of the origin – which can be thought of to be a centered at the origin ball of some radius $r > 0$ – where $f$ is bounded from above by some $C$. Now, consider an arbitrary $R > 0$ and an arbitrary point $x$ satisfying $\|x\|_2 \leq R$. Then, the point

$$y := -\frac{r}{R}x$$

is in $U$, and we have

$$0 = \frac{r}{r+R}x + \frac{R}{r+R}y.$$

As $f$ is convex and $\|y\|_2 \leq r$ implying $f(y) \leq C$, we conclude that

$$f(0) \leq \frac{r}{r+R}f(x) + \frac{R}{r+R}f(y) \leq \frac{r}{r+R}f(x) + \frac{R}{r+R}C.$$

By reorganizing the terms in this inequality, we get the lower bound

$$f(x) \geq \frac{r+R}{r}f(0) - \frac{R}{r}C.$$

Thus, $f$ is bounded below for any $x$ in the ball that is centered at 0 and has radius $R$. As $R > 0$ is arbitrary, for any bounded set, by selecting $R > 0$ large enough we can find such a ball with radius $R$ covering it. □

Our proof of Theorem III.14.9 relies on the following "local" version of the theorem.

**Proposition** III.14.12 Let $f$ be a convex function. Then, for any $\bar{x} \in \mathrm{rint}\,(\mathrm{Dom}\, f)$, we have

(i) $f$ is bounded at $\bar{x}$, i.e., there exists $r > 0$ such that $f$ is bounded in the

$r$-neighborhood $U_r(\bar{x})$ of $\bar{x}$ in the affine span of $\operatorname{Dom} f$:

$$\exists r > 0 \text{ and } C \text{ such that } |f(x)| \leq C, \quad \forall x \in U_r(\bar{x}),$$
$$\text{where } U_r(\bar{x}) := \{x \in \operatorname{Aff}(\operatorname{Dom} f) : \|x - \bar{x}\|_2 \leq r\};$$

(ii) $f$ is Lipschitz continuous at $\bar{x}$, i.e., there exist constants $\rho > 0$ and $L$ such that

$$|f(x) - f(x')| \leq L\|x - x'\|_2, \quad \forall x, x' \in U_\rho(\bar{x}).$$

**Proof.**

(i): $1^0$. We start with proving the *above boundedness* of $f$ in a neighborhood of $\bar{x}$. This is immediate: by the premise of the proposition we have $\bar{x} \in \operatorname{rint}(\operatorname{Dom} f)$, so there exists $\bar{r} > 0$ such that the neighborhood $U_{\bar{r}}(\bar{x})$ is contained in $\operatorname{Dom} f$. Now, we can find a small simplex $\Delta$ of the dimension $m := \dim(\operatorname{Aff}(\operatorname{Dom} f))$ with the vertices $x^0, \ldots, x^m$ in $U_{\bar{r}}(\bar{x})$ in such a way that $\bar{x}$ will be a convex combination of the vectors $x^i$ with *positive* coefficients, even with the coefficients $1/(m+1)$, i.e.,

$$\bar{x} = \sum_{i=0}^{m} \frac{1}{m+1} x^i.$$

Here is the justification of this claim that such a simplex $\Delta$ exists: First, when $\operatorname{Dom} f$ is a singleton, the claim is evident. So, we assume that $\dim(\operatorname{Dom} f) = m \geq 1$. Without loss of generality, we may assume that $\bar{x} = 0$, so that $0 \in \operatorname{Dom} f$ and therefore $\operatorname{Aff}(\operatorname{Dom} f) = \operatorname{Lin}(\operatorname{Dom} f)$. Then, by Linear Algebra, we can find $m$ vectors $y^1, \ldots, y^m$ in $\operatorname{Dom} f$ which form a basis of $\operatorname{Lin}(\operatorname{Dom} f) = \operatorname{Aff}(\operatorname{Dom} f)$. Setting $y^0 := -\sum\limits_{i=1}^{m} y^i$ and taking into account that $0 = \bar{x} \in \operatorname{rint}(\operatorname{Dom} f)$, we can find $\epsilon > 0$ such that the vectors $x^i := \epsilon y^i$, $i = 0, \ldots, m$, belong to $U_{\bar{r}}(\bar{x})$. By construction, $\bar{x} = 0 = \frac{1}{m+1} \sum\limits_{i=0}^{m} x^i$.

Note that $\bar{x} \in \operatorname{rint}(\Delta)$ (see Exercise I.3). Since $\Delta$ spans the same affine subspace as $\operatorname{Dom} f$, we can find a sufficiently small $r > 0$ such that $r \leq \bar{r}$ and $U_r(\bar{x}) \subseteq \Delta$. Now, by definition,

$$\Delta = \left\{ \sum_{i=0}^{m} \lambda_i x^i : \ \lambda_i \geq 0, \forall i, \ \sum_{i=0}^{m} \lambda_i = 1 \right\},$$

so that by Jensen's inequality, for all $x \in \Delta$ we have

$$f\left(\sum_{i=0}^{m} \lambda_i x^i\right) \leq \sum_{i=0}^{m} \lambda_i f(x^i) \leq \max_{i=0,\ldots,m} f(x^i) =: C_u.$$

Thus, as $\Delta \supseteq U_r(\bar{x})$, we conclude that $f(x) \leq C_u$ for all $x \in U_r(\bar{x})$ as well.

$2^0$. Now let us prove that if $f$ is above bounded, by some $C_\ell$, in $U_r := U_r(\bar{x})$, then in fact it is also below bounded in this neighborhood (and, consequently, $f$ is bounded in $U_r$). Consider any $x \in U_r$, so that $x \in \operatorname{Aff}(\operatorname{Dom} f)$ and $\|x - \bar{x}\|_2 \leq r$. Define $x' := \bar{x} - [x - \bar{x}] = 2\bar{x} - x$. Then, $x' \in \operatorname{Aff}(\operatorname{Dom} f)$ and $\|x' - \bar{x}\|_2 = \|x - \bar{x}\|_2 \leq r$, and so $x' \in U_r$. As $\bar{x} = \frac{1}{2}[x + x']$ and $f$ is convex, we have

$$2f(\bar{x}) \leq f(x) + f(x').$$

Note that this inequality holds for all $x \in U_r$, hence

$$f(x) \geq 2f(\bar{x}) - f(x') =: C_\ell, \quad \forall x \in U_r(\bar{x}).$$

Thus, $f$ is indeed bounded below by $C_\ell$ in $U_r$.

Setting $C := \max\{C_u, C_\ell\}$ then concludes the proof of part (i).

(ii): Part (ii) is indeed an immediate consequence of part (i) and Lemma III.14.1. From part (i) we already know that there exist positive constants $r, C$ such that $|f(x)| \leq C$ for all $x \in U_r(\bar{x})$. We will prove that $f$ is Lipschitz continuous in the neighborhood $U_{r/2}(\bar{x})$. Consider any $x, x' \in U_{r/2}(\bar{x})$ such that $x \neq x'$. Let us extend the segment $[x, x']$ through the point $x'$ until it reaches to a certain point $x''$ on the (relative) boundary of $U_r(\bar{x})$. Then,

$$x' \in (x, x'') \quad \text{and} \quad \|x'' - \bar{x}\|_2 = r.$$

Then, by $(14.3)$ we have

$$f(x') - f(x) \leq \|x' - x\|_2 \frac{f(x'') - f(x)}{\|x'' - x\|_2}.$$

As $|f(y)| \leq C$ for any $y \in U_r(\bar{x})$ and $x, x'' \in U_r(\bar{x})$, we observe that $|f(x'') - f(x)| \leq 2C$. Moreover, by the triangle inequality we have $\|x'' - x\|_2 \geq \|x'' - \bar{x}\|_2 - \|\bar{x} - x\|_2 \geq r - (r/2) = r/2$, where the last inequality holds from $\|x'' - \bar{x}\|_2 = r$ and $\|\bar{x} - x\|_2 \leq r/2$ (as $x \in U_{r/2}(\bar{x})$). Hence, the term $\frac{f(x'') - f(x)}{\|x'' - x\|_2}$ is bounded above by the quantity $(2C)/(r/2) = 4C/r$. Thus, we have

$$f(x') - f(x) \leq \frac{4C}{r}\|x' - x\|_2, \quad \forall x, x' \in U_{r/2}.$$

By swapping the roles of $x$ and $x'$, we arrive at

$$f(x) - f(x') \leq \frac{4C}{r}\|x' - x\|_2,$$

and hence

$$|f(x) - f(x')| \leq \frac{4C}{r}\|x - x'\|_2, \quad \forall x, x' \in U_{r/2},$$

as required in part (ii). □

We are now ready to prove Theorem III.14.9.

**Proof of Theorem III.14.9.** We can extract Theorem III.14.9 from Proposition III.14.12 by the standard Analysis reasoning. All we need is to prove that if $K$ is a bounded and closed (i.e., a compact) subset of $\operatorname{rint}(\operatorname{Dom} f)$, then $f$ is Lipschitz continuous on $K$ (the boundedness of $f$ on $K$ is an evident consequence of its Lipschitz continuity on $K$ and boundedness of $K$). Assume for contradiction that $f$ is not Lipschitz continuous on $K$. Then, for every integer $i$ there exists a pair of distinct points $x^i, y^i \in K$ such that

$$f(x^i) - f(y^i) \geq i\|x^i - y^i\|_2. \tag{14.6}$$

Since $K$ is compact, passing to a subsequence we can ensure that $x^i \to x \in K$ and $y^i \to y \in K$.

First, we claim that it is not possible to have $x = y$. To see this recall that by Proposition III.14.12 $f$ is Lipschitz continuous in a neighborhood $B$ of $x$. If $x = y$ were to hold, then since $x^i \to x,\ y^i \to y$, this neighborhood $B$ would contain all $x^i$ and $y^i$ with large enough indices $i$; but then, from the Lipschitz continuity of $f$ in $B$, the ratios $(f(x^i) - f(y^i))/\|x^i - y^i\|_2$ would form a bounded sequence, which we know is not the case. Thus, the case of $x = y$ is impossible.

Next, we claim that the case of $x \neq y$ is not possible, either. Proposition III.14.12 implies that $f$ is continuous on $\mathrm{Dom}\, f$ at both the points $x$ and $y$ (note that Lipschitz continuity at a point clearly implies the usual continuity at it), so that $f(x^i) \to f(x)$ and $f(y^i) \to f(y)$ as $i \to \infty$. Thus, the left hand side in $(14.6)$ remains bounded as $i \to \infty$. On the other hand, as $i \to \infty$ the right hand side in $(14.6)$ tends to $\infty$ since when $i$ tends to $\infty$ and $\|x^i - y^i\|_2$ tends to a nonzero limit $\|x - y\|_2$ (recall $x \neq y$ in this case).

This is the desired contradiction since precisely one of the cases $x = y$ and $x \neq y$ must hold. □

# 15

# Minima and maxima of convex functions

## 15.1 Minima of convex functions

### 15.1.1 Unimodality

As it was already mentioned, optimization problems involving convex functions possess nice theoretical properties. One of the *most important* of these properties is given by the following result which states that any local minimizer of a convex function is also its *global* minimizer.

**Theorem** III.15.1 [Unimodality] Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex function and let $Q \subseteq \mathbf{R}^n$ be a nonempty convex set. Suppose $x^* \in Q \cap \operatorname{Dom} f$ is a *local minimizer* of $f$ on $Q$, i.e.,

$$\begin{array}{c} \exists r > 0 \ \text{ such that} \\ f(y) \geq f(x^*) \ \forall y \in Q \text{ satisfying } \|y - x^*\|_2 < r. \end{array} \tag{15.1}$$

Then, $x^*$ is a *global minimizer* of $f$ on $Q$, i.e.,

$$f(x^*) \leq \min_y \{f(y) : y \in Q\}. \tag{15.2}$$

Moreover, the set $\operatorname*{Argmin}_Q f$ of all local ($\equiv$ global) minimizers of $f$ on $Q$ is convex.

If $f$ is strictly convex, $x \neq y$ and $\lambda \in (0,1)$, then the set of its global minimizers $\operatorname*{Argmin}_Q f$ is either empty or a singleton.

**Proof.** If $f$ is the convex function that is identical to $+\infty$, $\operatorname{Dom} f = \varnothing$ and there is nothing to prove. So, we assume that $\operatorname{Dom} f \neq \varnothing$.

We will first show that any local minimizer of $f$ is also a global minimizer of $f$. Let $x^*$ be a local minimizer of $f$ on $Q$. Consider any $y \in Q$ such that $y \neq x^*$. We need to prove that $f(y) \geq f(x^*)$. If $f(y) = +\infty$, this relation is automatically satisfied. So, we assume that $y \in \operatorname{Dom} f$. Note that by definition of a local minimizer, we also have $x^* \in \operatorname{Dom} f$ for sure. Now, for any $\tau \in (0,1)$, by Lemma III.14.1 we have

$$\frac{f(x^* + \tau(y - x^*)) - f(x^*)}{\tau\|y - x^*\|_2} \leq \frac{f(y) - f(x^*)}{\|y - x^*\|_2}.$$

Since $x^*$ is a local minimizer of $f$, the left hand side in this inequality is nonnegative for all small enough values of $\tau > 0$. Thus, we conclude that the right hand side is nonnegative as well, i.e., $f(y) \geq f(x^*)$.

Note that $\operatorname*{Argmin}_Q f$ is nothing but the sublevel set $\mathrm{lev}_\alpha(f)$ of $f$ associated with $\alpha$ taken as the minimal value $\min_Q f$ of $f$ on $Q$. Recall by Proposition III.13.6 any sublevel set of a convex function is convex, so this sublevel set $\operatorname*{Argmin}_Q f$ is convex.

Finally, let us prove that the set $\operatorname*{Argmin}_Q f$ associated with a strictly convex $f$ is, if nonempty, a singleton. Assume for contradiction that there are two distinct minimizers $x', x''$ in $\operatorname*{Argmin}_Q f$. Then, from strict convexity of $f$, we would have

$$f\left(\frac{1}{2}x' + \frac{1}{2}x''\right) < \frac{1}{2}(f(x') + f(x'')) = \min_Q f,$$

where the equality follows from $x', x'' \in \operatorname*{Argmin}_Q f$. But, this strict inequality is impossible since $\frac{1}{2}x' + \frac{1}{2}x'' \in Q$ as $Q$ is convex and by definition of $\min_Q f$ we cannot have a point in $Q$ with objective value strictly smaller than $\min_Q f$. □

### 15.1.2 Necessary and sufficient optimality conditions

Another pleasant fact for differentiable convex functions is that the Calculus-based necessary optimality condition (a.k.a., the Fermat rule) is *sufficient* for global optimality.

**Theorem** III.15.2 [Necessary and sufficient optimality condition for differentiable convex functions] Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex function and let $Q \subseteq \mathbf{R}^n$ be a nonempty convex set. Consider any $x^* \in \operatorname{int} Q$. Suppose that $f$ is differentiable at $x^*$. Then, $x^*$ is a minimizer of $f$ on $Q$ if and only if

$$\nabla f(x^*) = 0.$$

**Proof.** The *necessity* of the condition $\nabla f(x^*) = 0$ for local optimality is due to Calculus, and so it has nothing to do with convexity. The essence of the matter is, of course, the *sufficiency* of the condition $\nabla f(x^*) = 0$ for *global optimality* of $x^*$ in the case of *convex* function $f$. In fact, this sufficiency is readily given by the gradient inequality (14.4). In particular, when $\nabla f(x^*) = 0$ holds, (14.4) becomes

$$f(y) \geq f(x^*) + (y - x^*)\nabla f(x^*) = f(x^*), \quad \forall y \in Q.$$

□

A natural question is what happens if $x^*$ in Theorem III.15.2 is not necessarily an interior point of $Q \subseteq \mathbf{R}^n$. Let us now assume that $x^*$ is an *arbitrary* point of a convex set $Q$ and that $f$ is convex on $Q$ and differentiable at $x^*$ (the latter means exactly that $\operatorname{Dom} f$ contains a neighborhood of $x^*$ and $f$ possesses the first order derivative at $x^*$). Under these assumptions, our goal is to characterize when $x^*$ will be a minimizer of $f$ on $Q$.

In order to give such a characterization, given a convex set $Q$ and a point $x^* \in Q$, we will look at the *radial cone* of $Q$ at $x^*$ [notation: $T_Q(x^*)$] defined as the set

$$T_Q(x^*) := \{h \in \mathbf{R}^n : \ x^* + th \in Q, \quad \forall \text{ small enough } t > 0\} .$$

Geometrically, this is the set of all directions "looking" from $x^*$ towards $Q$, so that a small enough positive step from $x^*$ along the direction, i.e., adding to $x^*$ a small enough positive multiple of the direction, keeps the point in $Q$. That is, $T_Q(x^*)$ is the set of all "feasible" directions at $x^*$ that starting from $x^*$ we can go a positive distance along and remain in $Q$. From the convexity of $Q$ it immediately follows that the radial cone indeed is a cone (not necessary closed). For example, when $x^* \in \operatorname{int} Q$, we have $T_Q(x^*) = \mathbf{R}^n$. Let us examine a more interesting example, e.g., the polyhedral set

$$Q = \left\{x \in \mathbf{R}^n : \ a_i^\top x \le b_i, \, i = 1, \dots, m\right\} , \tag{15.3}$$

and its radial cone. For any $x^* \in Q$, we define $\mathcal{I}(x^*) := \{i : \ a_i^\top x^* = b_i\}$ as the set of indices of constraints that are *active* at $x^*$ (i.e., those satisfied at $x^*$ as equalities rather than as strict inequalities). Then,

$$T_Q(x^*) = \left\{h \in \mathbf{R}^n : \ a_i^\top h \le 0, \quad \forall i \in \mathcal{I}(x^*)\right\} . \tag{15.4}$$

The radial cone $T_Q(x^*)$ of a convex set $Q$ taken at $x^* \in Q$ gives rise to another cone, namely the *normal cone* [notation: $N_Q(x^*)$]. By definition, the normal cone is the negative of the dual of the radial cone $T_Q(x^*)$, i.e.,

$$N_Q(x^*) := \left\{h \in \mathbf{R}^n : \ h^\top (x' - x^*) {\le} 0, \, \forall x' \in Q\right\} . \tag{15.5}$$

Now, we are ready to present the necessary and sufficient condition for $x^*$ to be a minimizer of a convex function $f$ on $Q$.

**Proposition** III.15.3 Let $Q$ be a convex set and $x^* \in Q$, and suppose $f$ is a convex function on $Q$ which is differentiable at $x^*$. The *necessary and sufficient condition* for $x^*$ to be a minimizer of $f$ on $Q$ is that the derivative of $f$ taken at $x^*$ along every direction from $T_Q(x^*)$ should be nonnegative, i.e.,

$$h^\top \nabla f(x^*) \ge 0, \quad \forall h \in T_Q(x^*).$$

By the duality relation between $T_Q(x^*)$ and $N_Q(x^*)$, this is precisely the same condition as the inclusion

$$\nabla f(x^*) \in -N_Q(x^*).$$

Thus, with $Q$, $x^*$ and $f$ as above, $x^*$ minimizes $f$ over $Q$ if and only if $-\nabla f(x^*)$ belongs to the normal cone $N_Q(x^*)$ of $Q$ at $x^*$.

**Proof.** The necessity of this condition is an evident fact which has nothing to do with convexity. Suppose that $x^*$ is a local minimizer of $f$ on $Q$. Assume for contradiction that there exists $h \in T_Q(x^*)$ such that $h^\top \nabla f(x^*) < 0$. Then, by $h \in T_Q(x^*)$, we get $x^* + th \in Q$ for all small enough positive $t$. Moreover, as $h^\top \nabla f(x^*) < 0$, we would also get

$$f(x^* + th) < f(x^*),$$

for all small enough positive $t$. These two observations together thus imply that in every neighborhood of $x^*$ there are points $x$ from $Q$ with values $f(x)$ strictly smaller than $f(x^*)$. This clearly contradicts the assumption that $x^*$ is a local minimizer of $f$ on $Q$.

Once again, the sufficiency of this condition is given by the gradient inequality, exactly as in the case when $x^* \in \operatorname{int} Q$ discussed in the proof of Theorem III.15.2. □

Proposition III.15.3 states that under its premise the necessary and sufficient condition for $x^*$ to minimize $f$ on $Q$ is the inclusion $\nabla f(x^*) \in -N_Q(x^*)$. What does this condition actually mean? The answer depends on what the normal cone is: whenever we have an explicit description of it, we have an explicit form of the optimality condition. For example,

- Consider the case of $T_Q(x^*) = \mathbf{R}^n$, i.e., $x^* \in \operatorname{int} Q$. Then, the normal cone $N_Q(x^*)$ is the cone of all the vectors $h$ that have nonpositive inner products with every vector in $\mathbf{R}^n$, i.e., $N_Q(x^*) = \{0\}$. Consequently, in this case the necessary and sufficient optimality condition of Proposition III.15.3 becomes the Fermat rule $\nabla f(x^*) = 0$, which we already know.

- When $Q$ is an affine plane given by linear equalities $Ax = b$, $A \in \mathbf{R}^{m\times n}$, the radial cone at every point $x \in Q$ is the linear subspace $\{d : Ad = 0\}$, the normal cone is the orthogonal complement $\{u = A^\top v : v \in \mathbf{R}^m\}$ to this linear subspace, and the optimality condition reads

*Given $Q = \{x : Ax = b\}$, a point $x^* \in Q$, and a function $f$ that is convex on $Q$ and differentiable at $x^*$, $x^*$ is a minimizer of $f$ on $Q$ if and only if*

$$\exists v \in \mathbf{R}^m : \ \nabla f(x^*) + A^\top v = 0.$$

- When $Q$ is the polyhedral set $(15.3)$, the radial cone is the polyhedral cone $(15.4)$, i.e., it is the set of all directions which have nonpositive inner products with all $a_i$ for $i \in \mathcal{I}(x^*)$ (recall that these $a_i$ are coming from the constraints $a_i^\top x \le b_i$ specifying $Q$ that are satisfied as equalities at $x^*$). The corresponding normal cone is thus the set of all vectors which have nonpositive inner products with all these directions in $T_Q(x^*)$, i.e., of vectors $a$ such that the inequality $h^\top a \le 0$ is a consequence of the inequalities $h^\top a_i \le 0$, $i \in \mathcal{I}(x^*) = \{i : a_i^\top x^* = b_i\}$. From the Homogeneous Farkas Lemma we conclude that the normal cone is simply the conic hull of the vectors $a_i$, $i \in \mathcal{I}(x^*)$. Thus, in this case our necessary and sufficient optimality condition becomes:

*Given $Q = \{x \in \mathbf{R}^n : a_i^\top x \le b_i, i = 1, \dots, m\}$, a point $x^* \in Q$, and a function $f$ that is convex on $Q$ and differentiable at $x^*$, $x^*$ is a minimizer of $f$ on $Q$ if and only if there exist nonnegative reals $\lambda_i^*$ ("Lagrange multipliers") associated with "active" indices $i$ (those from $\mathcal{I}(x^*)$) such that*

$$\nabla f(x^*) + \sum_{i \in \mathcal{I}(x^*)} \lambda_i^* a_i = 0,$$

*or, equivalently, there exist nonnegative $\lambda_i^*$, $1 \le i \le m$, such that*

$$\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* a_i = 0, \qquad \text{(Karush-Kuhn-Tucker equation)}$$

$$\textit{and} \quad \lambda_i^*[a_i^\top x^* - b_i] = 0, \quad 1 \le i \le m. \qquad \text{(complementary slackness)}$$

This is our first acquaintance with the famous *Karush-Kuhn-Tucker* (KKT) optimality conditions. We will eventually extend them to the case where $Q$ is given by convex, rather than just linear, constraints (see Theorem IV.24.4). Indeed, KKT conditions are also *necessary* for *local* optimality in the case of nonconvex Mathematical Programming problems.

**Remark** III.15.4 Let us give an informal explanation of the preceding results on first-order optimality conditions through Physics (see Figure III.2 for a graphical illustration). Consider the optimization problem given by

$$\min_{x \in \mathbf{R}^n} \{f(x) : a_i(x) \le 0, \quad i = 1, \dots, m\}.$$

This problem can be interpreted as locating the equilibrium position of a particle that is moving through $\mathbf{R}^n$ while being affected by an external force (like gravity) with potential $f$, meaning that when the position of the particle is $x \in \mathbf{R}^n$, the force acting at the particle is $-\nabla f(x)$. The domain in which the particle can actually travel is $Q := \{x \in \mathbf{R}^n : a_i(x) \le 0, \ i \le m\}$; think about areas $a_i(x) > 0$ as rigid obstacles that the particle cannot penetrate into. When the particle touches $i$-th obstacle (i.e., is in position $x$ with $a_i(x) = 0$), the obstacle produces a reaction force directed along the inward normal $-\nabla a_i(x)$ to the boundary of the obstacle, so that the reaction force is $-\lambda_i \nabla a_i(x)$; here $\lambda_i \ge 0$ depends on the pressure on the obstacle exerted by the particle. At an equilibrium $x^*$ (which, by Physics, should minimize, at least locally, the potential over $Q$) the total of the forces acting at the particle should be zero, that is, for properly selected $\lambda_i \ge 0$ one should have $-\nabla f(x^*) - \sum_{i:a_i(x^*)=0} \lambda_i \nabla a_i(x^*) = 0$, which is exactly what is said by our Karush-Kuhn-Tucker (KKT) optimality condition as applied to the problem where the functions $a_i(x) = a_i^\top x - b_i$ are affine.

### 15.1.3 ★ Symmetry Principle

We close this section by discussing a simple characterization of minimizers of convex functions admitting certain symmetry. When applicable, this characterization is ex-

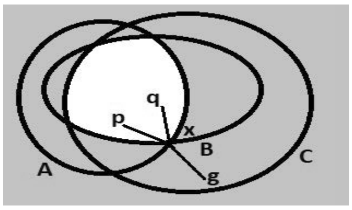


Figure III.2. Physical illustration of KKT optimality onditions for optimization problem $\min_{x\in\mathbf{R}^2}\{f(x) : a_i(x) \leq 0,\ i = 1, 2, 3\}$.
White area represents the feasible domain $Q$, while ellipses **A**, **B**, **C** represent the sets $a_1(x) \leq 0$, $a_2(x) \leq 0$, $a_3(x) \leq 0$. The point $x$ is a candidate feasible solution located at the intersection $\{u \in \mathbf{R}^2 : a_1(u) = a_2(u) = 0\}$ of boundaries of **A** and **B**. $g = -\nabla f(x)$ is external force acting at particle located at $x$, $p$ and $q$ are reaction forces created by obstacles **A** and **B**. The condition for $x$ to be at equilibrium reduces to $g + p + q = 0$, as on the picture. Equilibrium condition $g + p + q = 0$ translates to the KKT equation

$$\nabla f(x) + \lambda_1 a_1(x) + \lambda_2 \nabla a_2(x) = 0$$

holding for some nonnegative $\lambda_1, \lambda_2$.

tremely useful. This result is indeed an almost immediate consequence of Proposition III.13.6.

**Proposition** III.15.5 Let $f$ be a convex function with domain $Q := \mathrm{Dom}\, f \subseteq \mathbf{R}^n$. Suppose that $f$ and $Q$ admit a group of symmetries. That is, there exists a finite collection $\mathcal{G} = \{G_0, \ldots, G_m\}$ of distinct from each other $n \times n$ nonsingular matrices such that

(i) $\mathcal{G}$ is a finite group, i.e., $G \in \mathcal{G}$ implies $G^{-1} \in \mathcal{G}$ as well, and for any $G', G'' \in \mathcal{G}$, we also have $G'G'' \in \mathcal{G}$;
(ii) all matrices $G \in \mathcal{G}$ are symmetries of $Q$ and $f$, i.e., for any $G \in \mathcal{G}$, we have $Gx \in Q$ and $f(x) = f(Gx)$ for all $x \in Q$.

Suppose also that the set of minimizers of $f$ on $Q$, i.e., $Q_* := \mathrm{Argmin}_{x\in Q} f(x)$, is nonempty. Then, $f$ admits a *$\mathcal{G}$-symmetric minimizer*, that is, there exists some $x_* \in Q_*$ such that $Gx_* = x_*$ for all $G \in \mathcal{G}$.

**Proof.** Consider any fixed $x \in Q_*$. Since matrices $G \in \mathcal{G}$ are symmetries of $f$ and $Q$, they are symmetries of $Q_*$, i.e., $Gx \in Q_*$ for all $G \in \mathcal{G}$ (why?) as well. Now, because $f$ is convex, Proposition III.13.6 implies that the set $Q_* = \mathrm{Argmin}_{u\in Q} f(u) = \{u \in Q : f(u) \leq \min_Q f\}$ is convex and thus contains, along with the point $x$, the point $x_* := \frac{1}{m}\sum_{G\in\mathcal{G}} Gx$. It remains to show that $Hx_* = x_*$ for all $H \in \mathcal{G}$. Indeed, $Hx_* = \frac{1}{m}\sum_{G\in\mathcal{G}} HGx$, and the terms in the latter sum form a permutation of the terms in the sum specifying $x_*$ due to the fact that $\mathcal{G}$ is a group. Thus, $x_*$ is the desired $\mathcal{G}$-symmetric minimizer of $f$. □

**Illustration:** *Geometric-Arithmetic Mean inequality.* Let us use Symmetry Principle to justify the following classical inequality:

$$\text{For any } a_1, \ldots, a_m \in \mathbf{R}_+, \text{ we have } \sqrt[m]{a_1 \ldots a_m} \leq \frac{a_1 + a_2 + \ldots + a_m}{m}.$$

To prove this inequality, we define $a := a_1 + \ldots + a_m$, $Q := \{x \in \mathbf{R}^m_+ : \sum_{i=1}^m x_i = a\}$, and

$$f(x) := \begin{cases} \sqrt[m]{x_1 \ldots x_m}, & \text{if } x \in Q, \\ +\infty, & \text{otherwise.} \end{cases}$$

Note that $Q$ is nonempty and compact, and $f(x)$ is continuous and concave (see the end of section 14.2). Thus, by Theorem B.31 the set of maximizers of $f$ on $Q$, i.e., $Q_*$, is nonempty. Clearly, the $m!$ permutation matrices of size $m \times m$ form a group of symmetries of $f$ and $Q$, so that $Q_*$ contains a permutationally symmetric point $x_*$ (apply Proposition III.15.5 to minimize the convex function $-f$ over $Q$). Since the sum of all entries in a point from $Q$ is $a$, $Q$ contains exactly one permutationally symmetric point $\frac{1}{m}[a; a; \ldots; a]$. Then, as $\frac{1}{m}[a; a; \ldots; a]$ is a maximizer of $f$ over $Q$, we conclude that for every $[a_1; \ldots; a_m] \in Q$ we have

$$\sqrt[m]{a_1 \ldots a_m} = f([a_1; \ldots; a_m]) \leq f\left(\frac{1}{m}[a; \ldots; a]\right) = \frac{1}{m}a = \frac{a_1 + a_2 + \ldots + a_m}{m}.$$

□

## 15.2 Maxima of convex functions

So far we have seen that the fact that a point $x^* \in \operatorname{Dom} f$ is a global minimizer of a convex function $f$ depends only on the local behavior of $f$ at $x^*$. This is not the case with maximizers of a convex function. First of all, in all nontrivial cases, such a maximizer, if one exists at all, must belong to the relative boundary of the domain of the function.

**Theorem** III.15.6 Let $f$ be a convex function. If a point $x^* \in \operatorname{rint}(\operatorname{Dom} f)$ is a maximizer of $f$ over $\operatorname{Dom} f$, then $f$ is constant on $\operatorname{Dom} f$.

**Proof.** Define $Q := \operatorname{Dom} f$, and consider any $y \in Q$. We need to prove that $f(y) = f(x^*)$. There is nothing to prove if $y = x^*$, so we assume that $y \neq x^*$. Since, by assumption, $x^* \in \operatorname{rint} Q$, there exists a point $y' \in Q$ such that $x^*$ is an interior point of the segment $[y', y]$, i.e., there exists $\lambda \in (0, 1)$ such that

$$x^* = \lambda y' + (1 - \lambda)y.$$

Then, as $f$ is convex, we get

$$f(x^*) \leq \lambda f(y') + (1 - \lambda)f(y).$$

As $x^*$ is a maximizer of $f$ on $Q$, we have $f(x^*) \geq f(y)$ and $f(x^*) \geq f(y')$. Combining this with the preceding inequality leads to $\lambda f(y') + (1 - \lambda)f(y) = f(x^*)$. Since $\lambda \in (0, 1)$ and $\max\{f(y), f(y')\} \leq f(x^*)$, this equation can hold as equality only if $f(y') = f(y) = f(x^*)$. □

In particular, Theorem III.15.6 states that given a convex function $f$ the only way for a point $x^* \in \operatorname{rint}(\operatorname{Dom} f)$ to be a global maximizer of $f$ is if the function $f$ is constant over its domain.

Next, we provide further information on maxima of convex functions.

**Theorem** III.15.7 Let $f$ be a convex function on $\mathbf{R}^n$ and $E \subseteq \mathbf{R}^n$ be a nonempty set. Then,

$$\sup_{x \in \operatorname{Conv} E} f(x) = \sup_{x \in E} f(x). \tag{15.6}$$

In particular, if $S \subset \mathbf{R}^n$ is a nonempty convex and compact set, then the supremum of $f$ on $S$ is equal to the supremum of $f$ on the set of extreme points of $S$, i.e.,

$$\sup_{x \in S} f(x) = \sup_{x \in \operatorname{Ext}(S)} f(x). \tag{15.7}$$

**Proof.** We will first prove (15.6). As $\operatorname{Conv} E \supseteq E$, we have $\sup\limits_{x \in \operatorname{Conv} E} f(x) \geq \sup\limits_{x \in E} f(x)$. To prove the reverse direction, consider any $\bar{x} \in \operatorname{Conv} E$. Then, there exist $x^i \in E$ and convex combination weights $\lambda_i \geq 0$ satisfying $\sum\limits_i \lambda_i = 1$ and

$$\bar{x} = \sum_i \lambda_i x^i.$$

Applying Jensen's inequality (Proposition III.13.4), we get

$$f(\bar{x}) \leq \sum_i \lambda_i f(x^i) \leq \sum_i \lambda_i \sup_{x \in E} f(x) = \sup_{x \in E} f(x).$$

Since the preceding inequality holds for any $\bar{x} \in \operatorname{Conv} E$, we conclude $\sup\limits_{x \in \operatorname{Conv} E} f(x) \leq \sup\limits_{x \in E} f(x)$ holds as well, as desired.

To prove (15.7), note that when $S$ is a nonempty convex compact set, by Krein-Milman Theorem (Theorem II.8.6) we have $S = \operatorname{Conv}(\operatorname{Ext}(S))$. Then, (15.7) follows immediately from (15.6). □

Our last theorem on maxima of convex functions is as follows.

**Theorem** III.15.8 [Maxima of convex functions] Let $f$ be a proper convex function and let $Q := \operatorname{Dom} f$. Then,

(i) If $Q$ is closed and does not contain lines and the set of global maximizers of $f$, i.e.,

$$\operatorname*{Argmax}_Q f := \{x \in Q : f(x) \geq f(y), \forall y \in Q\},$$

is nonempty, then $\operatorname*{Argmax}\limits_Q f \cap \operatorname{Ext}(Q) \neq \varnothing$, i.e., at least one of the maximizers of $f$ is an extreme point of $Q$.

(ii) If the set $Q$ is polyhedral and $f$ is above bounded on $Q$, then the maximum of $f$ on $Q$ is achieved, i.e., $\operatorname*{Argmax}\limits_Q f \neq \varnothing$.

**Proof.**

Let us start with the following immediate observation:

> (!) *Let a convex function $f$ be bounded from above on a ray $\ell = \bar{x} + \mathbf{R}_+ e$. Then the function does not increase along the ray: when $x \in \ell$ and $s \geq 0$, we have $f(x+se) \leq f(x)$.*

Indeed, assuming for contradiction that there exists $x \in \ell$ and $x' = x + se$, $s \geq 0$, with $f(x') > f(x)$, we conclude that $s > 0$ and that the average rate of change $\kappa := \frac{f(x')-f(x)}{s}$ when moving from $x$ to $x'$ is positive: $\kappa > 0$. Since $f$ is convex, the average rate $\frac{f(x+te)-f(x)}{r}$ when moving from $x$ to $x+te$ with $t \geq s$ is at least $\kappa > 0$ (section 14.2), implying that $f(x+te) \to +\infty$ as $t \to \infty$, which is the desired contradiction, since $f$ is bounded from above on $\ell$.

As an immediate corollary of (!), we conclude that

> (!!) *If $Q$ is a convex set represented as $V+R$, where $V$ is nonempty, and $R$ is a cone, and $f$ is a convex bounded from above function on $Q$, then $f$ attains its maximum on $Q$ if and only if it attains its maximum on $V$, the maxima being equal to each other.*

Indeed, assume that $f$ attains its maximum on $Q$ at some point $\bar{x}$. As $Q = V + R$, we have $\bar{x} = \bar{v} + e$ for some $\bar{v} \in V$ and $e \in R$. By (!), we have $f(\bar{v}) \geq f(\bar{x})$ which combines with $V \subset Q$ to imply that $\bar{v}$ is a maximizer of $f$ on both $V$ and $Q$. Vice versa, assume that $f$ attains its maximum on $V$ at certain point $\bar{v}$. Every $x \in Q$ can be represented as $v + e$ with $v \in V$ and $e \in R$, so that by (!) we have $f(x) \leq f(v)$ and thus $f(x) \leq f(\bar{v})$, so that we again conclude that $\bar{v}$ maximizes $f$ both on $V$ and on $Q$.

Theorem III.15.8 is an immediate corollary of (!) and (!!). Indeed, in the case of (i), as $Q$ is a nonempty closed convex set that does not contain lines, by Theorem II.8.16 we deduce $\mathrm{Ext}(Q) \neq \varnothing$ and $Q$ admits a representation as

$$Q = \underbrace{\mathrm{Conv}(W)}_{V} + R, \tag{15.8}$$

$W = \mathrm{Ext}(Q)$, $R = \mathrm{Rec}(Q)$. In the situation of (i), $f$ attains its maximum on $Q$, and therefore, by (!!), has a maximizer $\bar{v}$ belonging to $V$. Due to the origin of $V$, $\bar{v} = \sum_{i \in I} \lambda_i v_i$ with nonempty finite set $I$, $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$, and $v_i \in \mathrm{Ext}(Q)$. By convexity of $f$, $f(\bar{v}) \leq \sum_{i \in I} \lambda_i f(v_i) \leq \max_{i \in I} f(v_i) =: f(v_{i_*})$, implying that the extreme point $v_{i_*}$f $Q$ maximizes $f$ on $Q$.

In the case of (ii), the polyhedral set $Q$ by Theorem II.10.1 admits representation (15.8) with nonempty finite $W$, implying that $f$, which is convex and real-valued on $V = \mathrm{Conv}(W) \subset Q = \mathrm{Dom}\, f$, attains its maximum on $V$, e.g., at its maximizer on the nonempty and finite set $W$. By (!!), this maximizer maximizes $f$ on $Q$ as well, so that $f$ achieves its maximum on $Q$, as claimed in (ii). □

# 16

# Subgradients

## 16.1 Proper lower semicontinuous convex functions and their representation

Recall that an equivalent definition of a convex function $f$ on $\mathbf{R}^n$ is that $f$ is a function taking values in $\mathbf{R} \cup \{+\infty\}$ such that its epigraph

$$\mathrm{epi}\{f\} = \left\{[x;t] \in \mathbf{R}^{n+1} : \ t \geq f(x)\right\}$$

is a convex set. Thus, there is no essential difference between convex functions and convex sets: a convex function generates a convex set, i.e., its epigraph, which of course remembers everything about the function. And the only specific property of the epigraph as a convex set is that it always possesses a very specific recessive direction, namely $h = [0;1]$. That is, the ray $\{z + th : t \geq 0\}$ directed by $h$ belongs to the epigraph set whenever the starting point $z$ of the ray is in the set. Whenever a convex set possesses a nonzero recessive direction $h$ such that $-h$ is not a recessive direction, the set in appropriate coordinates becomes the epigraph of a convex function. Thus, a convex function is, basically, nothing but a way to look, in the literal meaning of the latter verb, at a convex set.

Now, we know that the convex sets that are "actually nice" are the closed ones: they possess a lot of important properties (e.g., admit a good outer description) which are not shared by arbitrary convex sets. Therefore, among convex functions there also are "actually nice" ones, namely those with closed epigraphs. Closedness of the epigraph of a function can be "translated" to the functional language and there it becomes a special kind of continuity, namely *lower semicontinuity*.

Before formally defining lower semicontinuity, let us do a brief preamble on convergence of sequences on the extended real line. In the sequel, we will operate with limits of sequences $\{a_i\}_{i\geq 1}$ with terms $a_i$ from the extended real line $\overline{\mathbf{R}} := \mathbf{R} \cup \{+\infty\} \cup \{-\infty\}$. These limits are defined in the natural way: the relation

$$\lim_{i\to\infty} a_i = a \in \overline{\mathbf{R}}$$

means that for every $a' \in \overline{\mathbf{R}}$

- $a_i < a'$ for all but finitely many values of $i$, when $a' > a$, and
- $a_i > a'$ for all but finitely many values of $i$, when $a' < a$.

Equivalent way to treat convergence of sequences $\{a_i\}_i \subseteq \overline{\mathbf{R}}$ is as follows: let us fix somehow a strictly monotone continuous function $\theta$ on $\mathbf{R}$ which maps the axis onto the interval $(-1,1)$ (that is, $\lim_{s\to-\infty}\theta(s) = -1$, $\lim_{s\to\infty}\theta(s) = 1$), e.g., $\theta(s) = \frac{2}{\pi}\operatorname{atan}(s)$, and extend it from $\mathbf{R}$ to $\overline{\mathbf{R}}$ by setting

$$\overline{\theta}(a) = \begin{cases} -1, & \text{if } a = -\infty, \\ \theta(a), & \text{if } a \in \mathbf{R}, \\ 1, & \text{if } a = +\infty. \end{cases}$$

With this "encoding," $\overline{\mathbf{R}}$ becomes the segment $[-1,1]$, and the relation $a = \lim_{i\to\infty} a_i$ as defined above is the same as $\overline{\theta}(a) = \lim_{i\to\infty}\overline{\theta}(a_i)$, that is, this relation stands for the usual convergence, as $i\to\infty$, of reals $\overline{\theta}(a_i)$ to the real $\overline{\theta}(a)$. Note also that for $a,b\in\overline{\mathbf{R}}$ the relation $a\le b$ $(a<b)$ is exactly the same as the usual arithmetic inequality $\overline{\theta}(a)\le\overline{\theta}(b)$ $(\overline{\theta}(a)<\overline{\theta}(b)$, respectively).

With convergence and limits of sequences $\{a_i\}_i \subseteq \overline{\mathbf{R}}$ already defined, we can speak about upper (lower) limits of these sequences. For example, we can define $\liminf_{i\to\infty} a_i$ as $a\in\overline{\mathbf{R}}$ uniquely specified by the relation $\overline{\theta}(a) = \liminf_{i\to\infty}\overline{\theta}(a_i)$. Same as with lower limits of sequences of reals, $\liminf_{i\to\infty} a_i$ is the smallest (in terms of the relation $\le$ on $\overline{\mathbf{R}}$!) of the limits of converging (in $\overline{\mathbf{R}}$!) subsequences of the sequence $\{a_i\}_i$.

It is time to come back to lower semicontinuity.

**Definition** III.16.1 [Lower semicontinuity] Let $f:\mathbf{R}^n\to\mathbf{R}\cup\{+\infty\}$ be a function (not necessarily convex). $f$ is called *lower semicontinuous* (lsc for short) *at a point* $\bar{x}$, if for every sequence of points $\{x^i\}_i$ converging to $\bar{x}$ one has

$$f(\bar{x}) \le \liminf_{i\to\infty} f(x^i).$$

A function $f$ is called lower semicontinuous, if it is lower semicontinuous at every point.

A trivial example of an lsc function is a continuous one. Note, however, that an lsc function need not be continuous; what it is needed for lower semicontinuity is that the function can make only "jump downs." For example, the function

$$f(x) = \begin{cases} 0, & \text{if } x\neq 0, \\ a, & \text{if } x = 0, \end{cases}$$

is lsc if $a\le 0$ (the function can "jump down at $x=0$ or no jump at all"), and is *not* lsc if $a>0$ ("jump up"). For more illustrations, see Figure III.3.

Here is the connection between lower semicontinuity of a function and the geometry of its epigraph.

**Proposition** III.16.2 A function $f:\mathbf{R}^n\to\mathbf{R}\cup\{+\infty\}$ is lower semicontinuous if and only if its epigraph is closed.

**Proof.** First, suppose $\operatorname{epi}\{f\}$ is closed, and let us prove that $f$ is lsc. Consider a

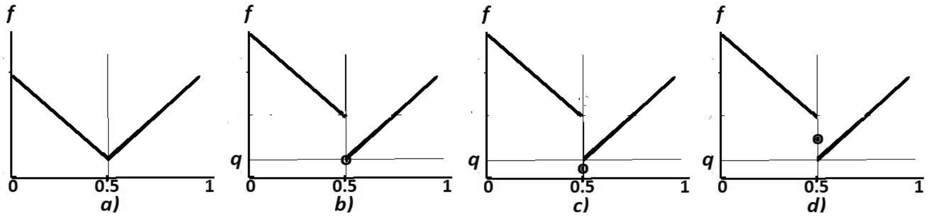


Figure III.3. Continuity and lower semicontinuity

a) continuous function $f(x)$, $\mathrm{Dom}\, f = [0,1]$

b-d) functions on $[0,1]$ discontinuous at $x = 0.5$. Given their common restriction on $\{0 \le x \le 1, x \neq 0.5\}$ and setting $q = \lim_{\epsilon\to+0} \inf_{\substack{|x-0.5|\le\epsilon \\ x\neq 0.5}} f(x)$, to be lsc, we should have $f(0.5) \le q$, as in b-c). Function d) is not lsc.

sequence $\{x^i\}_i$ such that $x^i \to x$ as $i \to \infty$, and let us prove that $f(x) \le a := \liminf_{i\to\infty} f(x^i)$. There is nothing to prove when $a = +\infty$. Assuming $a < +\infty$, by the definition of $\liminf$ there exists a sequence $i_1 < i_2 < \ldots$ such that $f(x^{i_j}) \to a$ as $j \to \infty$. Let us assume that $a > -\infty$ (we will verify later on that this is in fact the case). Then, as the points $[x^{i_j}; f(x^{i_j})] \in \mathrm{epi}\{f\}$ converge to $[x; a]$ and $\mathrm{epi}\{f\}$ is closed, we see that $[x; a] \in \mathrm{epi}\{f\}$, that is, $f(x) \le a$, as claimed. It remains to verify that $a > -\infty$. Indeed, assuming $a = -\infty$, we conclude that for every $t \in \mathbf{R}$ the points $[x^{i_j}; t]$ belong to $\mathrm{epi}\{f\}$ for all large enough values of $j$, which, as above, implies that $[x; t] \in \mathrm{epi}\{f\}$, that is, $t \ge f(x)$. The latter inequality cannot hold true for all real $t$, since $f$ does not take value $-\infty$; thus, $a = -\infty$ is impossible.

Now, for the opposite direction, let $f$ be lsc, and let us prove that $\mathrm{epi}\{f\}$ is closed. So, we should prove that if $[x^i; t_i] \to [x; t]$ as $i \to \infty$ and $[x^i; t_i] \in \mathrm{epi}\{f\}$, that is, $t_i \ge f(x^i)$ for all $i$, then $[x; t] \in \mathrm{epi}\{f\}$, that is, $t \ge f(x)$. Indeed, since $f$ is lsc and $f(x^i) \le t_i$, we have $f(x) \le \liminf_{i\to\infty} f(x^i) \le \liminf_{i\to\infty} t_i = \lim_{i\to\infty} t_i = t$. □

An immediate consequence of Proposition III.16.2 is as follows:

**Corollary** III.16.3 Given an arbitrary family of lsc functions $f_\alpha : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$, their supremum given by

$$f(x) := \sup_{\alpha\in\mathcal{A}} f_\alpha(x)$$

is lower semicontinuous.

**Proof.** The epigraph of the function $f$ is the intersection of the epigraphs of all functions $f_\alpha$, and the intersection of closed sets is always closed. □

Now let us look at *convex, proper, and lower semicontinuous* functions, that is, functions $\mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ with closed convex and nonempty epigraphs. To save words, let us call these functions *regular*.

What we are about to do is to translate to the functional language several constructions and results related to convex sets. In the usual life, a translation (e.g., of poetry) typically results in something less rich than the original. In contrast to this, in mathematics this is a powerful source of new ideas and constructions.

**"Outer description" of a proper lower semicontinuous convex function.** We know that any closed convex set is the intersection of closed half-spaces. What does this fact imply when the set is the epigraph of a regular function $f$? First of all, note that the epigraph is not a completely arbitrary convex set in $\mathbf{R}^{n+1}$: it has the recessive direction $e := [0_n; 1]$, i.e., the basic orth of the $t$-axis in the space of variables $x \in \mathbf{R}^n$, $t \in \mathbf{R}$ where the epigraph lives. This direction, of course, should be recessive for every closed half-space

$$\Pi = \left\{[x;t] \in \mathbf{R}^{n+1} : \alpha t \geq d^\top x - a\right\}, \text{ where } |\alpha| + \|d\|_2 > 0, \qquad (*)$$

containing $\operatorname{epi}\{f\}$. Note that in $(*)$ we are adopting a specific form of the nonstrict linear inequality describing the closed half-space $\Pi$ among many possible forms in the space where the epigraph lives; this form is the most convenient for us now. Thus, $e$ should be a recessive direction of $\Pi \supseteq \operatorname{epi}\{f\}$, and the recessiveness of $e$ for $\Pi$ means exactly that $\alpha \geq 0$. Thus, speaking about closed half-spaces containing $\operatorname{epi}\{f\}$, we in fact are considering some of the half-spaces $(*)$ with $\alpha \geq 0$.

Now, there are two essentially different possibilities for $\alpha$ to be nonnegative: (A) $\alpha > 0$, and (B) $\alpha = 0$. In the case of (B) the boundary hyperplane of $\Pi$ is "vertical," i.e., it is parallel to $e$, and in fact it "bounds" only $x$. And, in such cases, $\Pi$ is the set of all vectors $[x;t]$ with $x$ belonging to certain half-space in the $x$-subspace and $t$ being an arbitrary real number. These "vertical" half-spaces will be of no interest to us.

The half-spaces which indeed are of interest to us are the "nonvertical" ones: those given by the case (A), i.e., with $\alpha > 0$. For a non-vertical half-space $\Pi$, we can always divide the inequality defining $\Pi$ by $\alpha$ and make $\alpha = 1$. Thus, a "nonvertical" candidate eligible for the role of a closed half-space containing $\operatorname{epi}\{f\}$ can always be written as

$$\Pi = \left\{[x;t] \in \mathbf{R}^{n+1} : t \geq d^\top x - a\right\}. \qquad (**)$$

That is, a "nonvertical" closed half-space containing $\operatorname{epi}\{f\}$ *can be represented as the epigraph of an affine function of* $x$.

Now, when is such a candidate indeed a half-space containing $\operatorname{epi}\{f\}$? It is clear that the answer is yes *if and only if the affine function* $d^\top x - a$ *is less than or equal to* $f(x)$ *for all* $x \in \mathbf{R}^n$. This is indeed precisely what we call as "$d^\top x - a$ is an *affine minorant* of $f$." In fact, we have a very nice characterization of proper lsc convex functions through their affine minorants!

**Proposition** III.16.4 A proper lower semicontinuous convex function $f$ is the pointwise supremum of all its affine minorants. Moreover, at every point $\bar{x} \in \operatorname{rint}(\operatorname{Dom} f)$, the function $f$ is even not only the supremum, but simply the maximum of its affine minorants. That is, at every point $\bar{x} \in \operatorname{rint}(\operatorname{Dom} f)$, there exists an affine function $f_{\bar{x}}(x)$ such that $f_{\bar{x}}(x) \leq f(x)$ for all $x \in \mathbf{R}^n$ and $f_{\bar{x}}(\bar{x}) = f(\bar{x})$.

Note that Proposition III.16.4 essentially gives us an *outer description* of a proper lsc convex function. This outer description is in fact instrumental in developing algorithms for convex optimization. Before proceeding with the proof of Proposition III.16.4, we will first prove an intermediate result on proper convex (but not necessarily lower semicontinuous) functions. This result indeed forms the most important step in the proof of Proposition III.16.4.

**Proposition** III.16.5 Let $f$ be a proper convex function. By definition of affine minorant, we immediately have that the supremum of all affine minorants of $f$ is less than or equal to $f$ everywhere. Let $\bar{x} \in \mathrm{rint}\,(\mathrm{Dom}\, f)$. Then, there exists an affine minorant $d^\top x - a$ of $f$ which coincides with $f$ at $\bar{x}$, i.e.,

$$f(x) \geq d^\top x - a,\ \forall x \in \mathbf{R}^n, \quad \text{and} \quad d^\top \bar{x} - a = f(\bar{x}). \tag{16.1}$$

Moreover, this supremum is $+\infty$ outside of $\mathrm{cl}(\mathrm{Dom}\, f)$. Thus, the supremum of all affine minorants of $f$ is equal to $f$ everywhere except, perhaps, some subset of $\mathrm{rbd}(\mathrm{Dom}\, f)$.

**Proof.** I. We will first prove that at every $\bar{x} \in \mathrm{rint}\,(\mathrm{Dom}\, f)$ there exists an affine function $f_{\bar{x}}(x)$ such that $f_{\bar{x}}(x) \leq f(x)$ for all $x \in \mathbf{R}^n$ and $f_{\bar{x}}(\bar{x}) = f(\bar{x})$.

I.1$^0$ First of all, we can easily reduce the situation to the one when $\mathrm{Dom}\, f$ is full-dimensional. Indeed, by shifting $f$ we can make $\mathrm{Aff}(\mathrm{Dom}\, f)$ to be a linear subspace $L$ in $\mathbf{R}^n$; restricting $f$ onto this linear subspace, we clearly get a proper function on $L$. If we believe that our statement is true for the case when $\mathrm{Dom}\, f$ is full-dimensional, we can conclude that there exists an affine function *on* $L$, i.e.,

$$d^\top x - a \qquad \text{[where } x \in L]$$

(where $d \in L$) such that

$$f(x) \geq d^\top x - a,\ \forall x \in L, \quad \text{and} \quad f(\bar{x}) = d^\top \bar{x} - a.$$

This affine function $d^\top x - a$ on $L$ clearly can be extended, by the same formula, from $L$ on the entire $\mathbf{R}^n$ and is a minorant of $f$ on the entire $\mathbf{R}^n$ (note that outside of $L \supseteq \mathrm{Dom}\, f$, the function $f$ simply is $+\infty$!). This affine minorant on $\mathbf{R}^n$ is exactly what we need.

I.2$^0$. Now let us prove that our statement is valid when $\mathrm{Dom}\, f$ is full-dimensional. In such a case, $\bar{x} \in \mathrm{int}(\mathrm{Dom}\, f)$. Consider the point $\bar{y} := [\bar{x}; f(\bar{x})]$. Note that $\bar{y} \in \mathrm{epi}\{f\}$, and in fact we claim further that $\bar{y} \in \mathrm{rbd}(\mathrm{epi}\{f\})$. Assume for contradiction that $\bar{y} \in \mathrm{rint}\,(\mathrm{epi}\{f\})$. Recall that $e := [0_n; 1]$ is the special recessive direction of $\mathrm{epi}\{f\}$, thus $\bar{y}' := \bar{y} + e$ satisfies $\bar{y}' \in \mathrm{epi}\{f\}$ and so the segment $[\bar{y}', \bar{y}]$ is contained in $\mathrm{epi}\{f\}$. Since $\bar{y} \in \mathrm{rint}\,(\mathrm{epi}\{f\})$, we can extend this segment a little more through its endpoint $\bar{y}$, without leaving $\mathrm{epi}\{f\}$. But, this is clearly impossible, since in such a case the $t$-coordinate of the new endpoint would be $< f(\bar{x})$ while the $x$-component of it still would be $\bar{x}$. Thus, $\bar{y} \in \mathrm{rbd}(\mathrm{epi}\{f\})$.

Next, we claim that $\bar{y}'$ is an interior point of $\mathrm{epi}\{f\}$. This is immediate: we know from Theorem III.14.9 that $f$ is continuous at $\bar{x}$ (recall that $\bar{x} \in \mathrm{int}(\mathrm{Dom}\, f)$), so that there exists a neighborhood $U$ of $\bar{x}$ in $\mathrm{Aff}(\mathrm{Dom}\, f) = \mathbf{R}^n$ such that $f(x) \leq f(\bar{x}) + 0.5$

whenever $x \in U$, or, in other words, the set

$$V := \{[x;t] :\ x \in U,\ t > f(\bar{x}) + 0.5\}$$

is contained in $\mathrm{epi}\{f\}$; but this set clearly contains a neighborhood of $\bar{y}'$ in $\mathbf{R}^{n+1}$. We see that $\mathrm{epi}\{f\}$ is full-dimensional, so that $\mathrm{rint}(\mathrm{epi}\{f\}) = \mathrm{int}(\mathrm{epi}\, f)$ and $\mathrm{rbd}(\mathrm{epi}\{f\}) = \mathrm{bd}(\mathrm{epi}\, f)$.

Now let us look at a hyperplane $\Pi$ supporting $\mathrm{cl}(\mathrm{epi}\{f\})$ at the point $\bar{y} \in \mathrm{rbd}(\mathrm{epi}\{f\})$. W.l.o.g., we can represent this hyperplane via a nontrivial (i.e., with $|\alpha| + \|d\|_2 > 0$) linear inequality

$$\alpha t \geq d^\top x - a. \tag{16.2}$$

satisfied everywhere on $\mathrm{cl}(\mathrm{epi}\{f\})$, specifically, as the hyperplane where this inequality holds true as equality. Now, inequality (16.2) is satisfied everywhere on $\mathrm{epi}\{f\}$, and therefore at the point $\bar{y}' := [\bar{x}; f(\bar{x}) + 1] \in \mathrm{epi}\{f\}$ as well, and is satisfied as equality at $\bar{y} = [\bar{x}; f(\bar{x})]$ (since $\bar{y} \in \Pi$). These two observations clearly imply that $\alpha \geq 0$. We claim that $\alpha > 0$. Indeed, inequality (16.2) says that the linear form $h^\top[x;t] := \alpha t - d^\top x$ attains its minimum over $y \in \mathrm{cl}(\mathrm{epi}\{f\})$, equal to $-a$, at the point $\bar{y}$. Were $\alpha = 0$, we would have $h^\top \bar{y} = h^\top \bar{y}'$, implying that the set of minimizers of the linear form $h^\top y$ on the set $\mathrm{cl}(\mathrm{epi}\{f\})$ contains an interior point (namely, $\bar{y}'$) of the set. This is possible only when $h = 0$, that is, $\alpha = 0, d = 0$, which is not the case.

Now, as $\alpha > 0$, by dividing both sides of (16.2) by $\alpha$, we get a new inequality of the form

$$t \geq d^\top x - a \tag{16.3}$$

(here we keep the same notation for the right hand side coefficients as we will never come back to the old coefficients) which is valid on $\mathrm{epi}\{f\}$ and is equality at $\bar{y} = [\bar{x}; f(\bar{x})]$. Its validity on $\mathrm{epi}\{f\}$ implies that for all $[x;t]$ with $x \in \mathrm{Dom}\, f$ and $t = f(x)$, we have

$$f(x) \geq d^\top x - a \quad \forall x \in \mathrm{Dom}\, f. \tag{16.4}$$

Thus, we conclude that the function $d^\top x - a$ is an affine minorant of $f$ on $\mathrm{Dom}\, f$ and therefore on $\mathbf{R}^n$ ($f = +\infty$ outside $\mathrm{Dom}\, f$!). Finally, note that the inequality (16.4) becomes an equality at $\bar{x}$, since (16.3) holds as equality at $\bar{y}$. The affine minorant we have just built justifies the validity of the first claim of the proposition.

II. Let $\mathcal{F}$ be the set of all affine functions which are minorants of $f$, and define the function

$$\bar{f}(x) := \sup_{\phi \in \mathcal{F}} \phi(x).$$

We have proved that $\bar{f}(x)$ is equal to $f$ on $\mathrm{rint}\,(\mathrm{Dom}\, f)$ (and at any $x \in \mathrm{rint}\,(\mathrm{Dom}\, f)$ in fact sup in the right hand side can be replaced with max). To complete the proof of the proposition, we should prove that $\bar{f}$ is equal to $f$ outside of $\mathrm{cl}(\mathrm{Dom}\, f)$ as well. Note that this is the same as proving that $\bar{f}(x) = +\infty$ for all $x \in \mathbf{R}^n \setminus \mathrm{cl}(\mathrm{Dom}\, f)$. To see this, consider any $\bar{x} \in \mathbf{R}^n \setminus \mathrm{cl}(\mathrm{Dom}\, f)$. As $\mathrm{cl}(\mathrm{Dom}\, f)$ is a closed convex set, $\bar{x}$ can be strongly separated from $\operatorname{Dom} f$, see Separation Theorem (ii) (Theorem II.7.3). Thus, there exists $z \in \mathbf{R}^n$ such that

$$z^\top \bar{x} \geq z^\top x + \zeta, \quad \forall x \in \operatorname{Dom} f, \qquad \text{where } \zeta > 0. \tag{16.5}$$

In addition, we already know that there exists at least one affine minorant of $f$, i.e., there exist $a$ and $d$ such that

$$f(x) \geq d^\top x - a, \quad \forall x \in \operatorname{Dom} f. \tag{16.6}$$

Multiplying both sides of $(16.5)$ by a positive weight $\lambda$ and then adding it to $(16.6)$, we get

$$f(x) \geq \underbrace{(d + \lambda z)^\top x + [\lambda\zeta - a - \lambda z^\top \bar{x}]}_{=:\phi_\lambda(x)}, \quad \forall x \in \operatorname{Dom} f.$$

This inequality clearly says that $\phi_\lambda(\cdot)$ is an affine minorant of $f$ on $\mathbf{R}^n$ for every $\lambda > 0$. The value of this minorant at $x = \bar{x}$ is equal to $d^\top \bar{x} - a + \lambda\zeta$ and therefore it goes to $+\infty$ as $\lambda \to +\infty$. We see that the supremum of affine minorants of $f$ at $\bar{x}$ indeed is $+\infty$, as claimed. This concludes the proof of Proposition III.16.5. □

Let us now prove Proposition III.16.4.

**Proof of Proposition III.16.4.** Under the premise of the proposition, $f$ is a proper lsc convex function. Let $\mathcal{F}$ be the set of all affine functions which are minorants of $f$, and let

$$\bar{f}(x) := \sup_{\phi \in \mathcal{F}} \phi(x).$$

be the supremum of all affine minorants of $f$. Then, by Proposition III.16.5, we know that $\bar{f}$ is equal to $f$ everywhere on $\operatorname{rint}(\operatorname{Dom} f)$ and everywhere outside of $\operatorname{cl}(\operatorname{Dom} f)$. Thus, all we need to prove is that when $f$ is also lsc, $\bar{f}$ is equal to $f$ everywhere on $\operatorname{rbd}(\operatorname{Dom} f)$ as well.

Consider any $\bar{x} \in \operatorname{rbd}(\operatorname{Dom} f)$. Recall that by construction $\bar{f}$ is everywhere $\leq f$. So, there is nothing to prove if $\bar{f}(\bar{x}) = +\infty$. Thus, we assume that $\bar{f}(\bar{x}) = c < \infty$ holds, and we will prove that in this case $f(\bar{x}) = c$ holds as well. Since $\bar{f} \leq f$ everywhere, proving $f(\bar{x}) = c$ is the same as proving $f(\bar{x}) \leq c$. In fact $f(\bar{x}) \leq c$ holds due to the lower semicontinuity of $f$: pick any $x' \in \operatorname{rint}(\operatorname{Dom} f)$ and consider a sequence of points $x^i \in [x', \bar{x})$ converging to $\bar{x}$. For all $i$, by Lemma I.1.30, we have $x^i \in \operatorname{rint}(\operatorname{Dom} f)$ and thus by Proposition III.16.5 we conclude

$$f(x^i) = \bar{f}(x^i).$$

Also, since $x^i \in [x', \bar{x})$, there exists $\lambda_i \in (0, 1]$ such that $x^i = (1 - \lambda_i)\bar{x} + \lambda_i x'$. Note that as $i \to \infty$, we have $x^i \to \bar{x}$ and so $\lambda_i \to +0$. Since $\bar{f}$ is clearly convex (as it is the supremum of a family of affine and thus convex functions), we have

$$\bar{f}(x^i) \leq (1 - \lambda_i)\bar{f}(\bar{x}) + \lambda_i \bar{f}(x').$$

Noting that $\bar{f}(x') = f(x')$ (recall $x' \in \operatorname{rint}(\operatorname{Dom} f)$ and apply Proposition III.16.5) as well and putting things together, we get

$$f(x^i) \leq (1 - \lambda_i)\bar{f}(\bar{x}) + \lambda_i f(x').$$

Moreover, as $i \to \infty$, we have $\lambda_i \to +0$ and so the right hand side in our inequality converges to $\bar{f}(\bar{x}) = c$. In addition, as $i \to \infty$, we have $x^i \to \bar{x}$ and since $f$ is lower semicontinuous, we get $f(\bar{x}) \leq c$. □

We see why "translation of mathematical facts from one mathematical language to another" – in our case, from the language of convex sets to the language of convex functions – may be fruitful: because we invest a lot into the process rather than run it mechanically.

**Closure of a convex function.** Proposition III.16.4 presents a nice result on the outer description of a *proper lower semicontinuous convex* function: it is the supremum of a family of affine functions. Note that, the reverse is also true: the supremum of every family of affine functions is a proper lsc convex function, provided that this supremum is finite at least at one point. This is because we know from section 14.1 that supremum of every family of convex functions is convex and from Corollary III.16.3 that supremum of lsc functions, e.g., affine ones (these are in fact even continuous), is lower semicontinuous.

Now, what to do with a convex function which is not lower semicontinuous? There is a similar question about convex sets: what to do with a convex set which is not closed? We can resolve this question very simply by passing from the set to its closure and thus getting a "much easier to handle" object which is very "close" to the original one: the "main part" of the original set – its relative interior – remains unchanged, and the "correction" adds to the set something relatively small – (part of) its relative boundary. The same approach works for convex functions as well: if a proper convex function $f$ is not lower semicontinuous (i.e., its epigraph is convex and nonempty, but is not closed), we can "correct" the function by replacing it with a new function with the epigraph being the closure of $\mathrm{epi}\{f\}$. To justify this approach, we, of course, should be sure that the closure of the epigraph of a convex function is also an epigraph of such a function. This indeed is the case, and to see it, it suffices to note that a set $G$ in $\mathbf{R}^{n+1}$ is the epigraph of a function taking values in $\mathbf{R} \cup \{+\infty\}$ if and only if the intersection of $G$ with every vertical line $\{x = \text{const}, t \in \mathbf{R}\}$ is either empty, or is a closed ray of the form $\{x = \text{const}, t \geq \bar{t} > -\infty\}$. Now, it is absolutely evident that if $G = \mathrm{cl}(\mathrm{epi}\{f\})$, then the intersection of $G$ with a vertical line is either empty, or is a closed ray, or is the entire line (the last case indeed can take place – look at the closure of the epigraph of the function equal to $-\frac{1}{x}$ for $x > 0$ and $+\infty$ for $x \leq 0$). We see that in order to justify our idea of "proper correction" of a convex function we should prove that if $f$ is convex, then the last of the indicated three cases, i.e., the intersection of $\mathrm{cl}(\mathrm{epi}\{f\})$ with a vertical line is the entire line, never occurs. However, we know from Proposition III.14.11 that every convex function $f$ is bounded from below on every compact set. Thus, $\mathrm{cl}(\mathrm{epi}\{f\})$ indeed cannot contain an entire vertical line. Therefore, we conclude that the closure of the epigraph of a convex function $f$ is the epigraph of a certain function called *the closure* of $f$ [notation: $\mathrm{cl}\, f$] defined as:

$$\mathrm{cl}(\mathrm{epi}\{f\}) = \mathrm{epi}\{\mathrm{cl}\, f\}.$$

Of course, the function $\mathrm{cl}\, f$ is convex (its epigraph is convex as it is $\mathrm{cl}(\mathrm{epi}\{f\})$ and

$\mathrm{epi}\{f\}$ is convex). Moreover, since the epigraph of $\mathrm{cl}\, f$ is closed, $\mathrm{cl}\, f$ is lsc. And of course we have the following immediate observation.

**Observation** III.16.6 The closure of a lsc convex function $f$ is $f$ itself.

**Proof.** Indeed, when $f$ is convex, $\mathrm{epi}\{f\}$ is convex, and when $f$ is lsc, $\mathrm{epi}\{f\}$ is closed by Proposition III.16.2. Hence, under the premise of this observation, $\mathrm{epi}\{f\}$ is convex and closed and thus, by definition of $\mathrm{cl}\, f$, is the same as $\mathrm{epi}\{\mathrm{cl}\, f\}$, implying that $f = \mathrm{cl}\, f$. □

The following statement gives an instructive alternative description of $\mathrm{cl}\, f$ in terms of $f$.

**Proposition** III.16.7 Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex function and let $\mathrm{cl}\, f$ be its closure. Then,

(i) Affine minorants of $\mathrm{cl}\, f$ are exactly the same as affine minorants of $f$, and $\mathrm{cl}\, f$ is the supremum of these minorants, i.e., for all $x \in \mathbf{R}^n$ we have

$$\mathrm{cl}\, f(x) = \sup_{\phi} \left\{\phi(x) : \phi \text{ is an affine minorant of } f\right\}. \tag{16.7}$$

In addition for every $x \in \mathrm{rint}\,(\mathrm{Dom}(\mathrm{cl}\, f)) = \mathrm{rint}\,(\mathrm{Dom}\, f)$, we can replace sup in the right hand side of (16.7) with max.

Moreover,

$$\begin{array}{lll} (a) & f(x) \geq \mathrm{cl}\, f(x), & \forall x \in \mathbf{R}^n, \\ (b) & f(x) = \mathrm{cl}\, f(x), & \forall x \in \mathrm{rint}\,(\mathrm{Dom}\, f), \\ (c) & f(x) = \mathrm{cl}\, f(x), & \forall x \notin \mathrm{cl}(\mathrm{Dom}\, f). \end{array} \tag{16.8}$$

Thus, the "correction" $f \mapsto \mathrm{cl}\, f$ may vary $f$ only at the points from $\mathrm{rbd}(\mathrm{Dom}\, f)$, implying that

$$\mathrm{Dom}\, f \subseteq \mathrm{Dom}(\mathrm{cl}\, f) \subseteq \mathrm{cl}(\mathrm{Dom}\, f),$$

hence $\mathrm{rint}\,(\mathrm{Dom}\, f) = \mathrm{rint}\,(\mathrm{Dom}(\mathrm{cl}\, f))$.

Moreover, $\mathrm{cl}\, f$ is the supremum of all convex lower semicontinuous minorants of $f$.

(ii) For all $x \in \mathbf{R}^n$, we have

$$\mathrm{cl}\, f(x) = \lim_{r \to +0} \inf_{x' : \|x'-x\|_2 \leq r} f(x').$$

**Proof.** There is nothing to prove when $f \equiv +\infty$. In this case $\mathrm{cl}\, f \equiv +\infty$ as well, $\mathrm{Dom}\, f = \mathrm{Dom}\,\mathrm{cl}\, f = \varnothing$, and all claims are trivially satisfied. Thus, assume from now on that $f$ is proper.

$1^o$. By construction, $\mathrm{epi}\{\mathrm{cl}\, f\} = \mathrm{cl}(\mathrm{epi}\{f\}) \supseteq \mathrm{epi}\{f\}$, which implies (16.8.$a$). Also, as $\mathrm{cl}(\mathrm{epi}\{f\}) \subseteq [\mathrm{cl}(\mathrm{Dom}\, f)] \times \mathbf{R}$, we arrive at (16.8.$c$).

$2^o$. Note that from (16.8.$a$) we deduce that every affine minorant of $\mathrm{cl}\, f$ is an affine minorant of $f$ as well. Moreover, the reverse is also true.

> Indeed, let $g(x)$ be an affine minorant of $f$. Then, we clearly have $\mathrm{epi}\{g\} \supseteq \mathrm{epi}\{f\}$, and as $\mathrm{epi}\{g\}$ is closed, we also get $\mathrm{epi}\{g\} \supseteq \mathrm{cl}(\mathrm{epi}\{f\}) = \mathrm{epi}\{\mathrm{cl}\, f\}$. Note that $\mathrm{epi}\{g\} \supseteq \mathrm{epi}\{\mathrm{cl}\, f\}$ is simply the same as saying that $g$ is an affine minorant of $\mathrm{cl}\, f$.

Thus, affine minorants of $f$ and of $\mathrm{cl}\, f$ indeed are the same. Then, as $\mathrm{cl}\, f$ is lsc and proper (since $\mathrm{cl}\, f \leq f$ and $f$ is proper), by applying Proposition III.16.4 to $\mathrm{cl}\, f$ and also applying Proposition III.16.5 to $f$, we deduce (16.8.$b$) and (16.7).

Finally, if $g$ is a convex lsc minorant of $f$, then $g$ definitely is proper, and thus by Proposition III.16.4 it is the supremum of all its affine minorants. These minorants of $g$ are affine minorants of $f$ as well, and thus - affine minorants of $\mathrm{cl}\, f$. The bottom line is that $g$ is $\leq$ the supremum of all affine minorants of $f$, which, as we already know, is $\mathrm{cl}\, f$. Thus, convex lsc minorant of $f$ is a minorant of $\mathrm{cl}\, f$, implying that the supremum $\widetilde{f}$ of these lsc convex minorants of $f$ satisfies $\widetilde{f} \leq \mathrm{cl}\, f$. The latter inequality is equality, since $\mathrm{cl}\, f$ itself is a convex lsc minorant of $f$ by (16.8.$a$). This completes the proof of part (i).

$3^o$ To verify (ii) we need to prove the following two facts:

> (ii-1) For every sequence $\{x^i\}$ such that as $i \to \infty$ we have $x^i \to \bar{x}$ and $f(x^i) \to s$ (where $s$ may be a finite number or infinity), we have $s \geq \mathrm{cl}\, f(\bar{x})$.
> (ii-2) For every $\bar{x}$ there exists a sequence $x^i \to \bar{x}$ such that $\lim_{i\to\infty} f(x^i) = \mathrm{cl}\, f(\bar{x})$.

To prove (ii-1), note that under the premise of this claim we have $s \neq -\infty$, since $f$ is below bounded on bounded subsets of $\mathbf{R}^n$ (Proposition III.14.11). There is nothing to verify when $s = +\infty$. So, suppose $s \in \mathbf{R}$. Then, the point $[\bar{x}; s]$ is in $\mathrm{cl}(\mathrm{epi}\{f\}) = \mathrm{epi}\{\mathrm{cl}\, f\}$, and thus $\mathrm{cl}\, f(\bar{x}) \leq s$, as claimed.

To prove (ii-2), note that the claim is trivially true when $\mathrm{cl}\, f(\bar{x}) = +\infty$; indeed, in this case $f(\bar{x}) = +\infty$ as well due to (16.8.$a$), and for all $i = 1, 2, \ldots$ we can take $x^i = \bar{x}$. Now, consider a point $\bar{x}$ such that $\mathrm{cl}\, f(\bar{x}) < \infty$. Then, we have $[\bar{x}; \mathrm{cl}\, f(\bar{x})] \in \mathrm{epi}\{\mathrm{cl}\, f\} = \mathrm{cl}(\mathrm{epi}\{f\})$. Thus, there exists a sequence $[x^i; t_i] \in \mathrm{epi}\{f\}$ such that $[x^i; t_i] \to [\bar{x}; \mathrm{cl}\, f(\bar{x})]$ as $i \to \infty$. Passing to a subsequence, we can assume that $f(x^i)$ have a limit, finite or infinite, as $i \to \infty$. Hence, $\lim_{i\to\infty} x^i = \bar{x}$ and $\lim_{i\to\infty} f(x^i) = \liminf_{i\to\infty} f(x^i) \leq \lim_{i\to\infty} t_i = \mathrm{cl}\, f(\bar{x})$. Recall also that from (ii-1) we have $\lim_{i\to\infty} f(x^i) \geq \mathrm{cl}\, f(\bar{x})$, so we conclude $\lim_{i\to\infty} f(x^i) = \mathrm{cl}\, f(\bar{x})$ as desired. □

## 16.2 Subgradients

Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex function, and let $x \in \mathrm{Dom}\, f$. Recall from our discussion in the preceding section that $f$ may admit an affine minorant $d^\top x - a$ which coincides with $f$ at $x$, i.e.,

$$f(y) \geq d^\top y - a, \ \forall y \in \mathbf{R}^n, \quad \text{and} \quad f(x) = d^\top x - a.$$

The equality relation above is equivalent to $a = d^\top x - f(x)$, and substituting this representation of $a$ into the first inequality, we get

$$f(y) \geq f(x) + d^\top (y - x), \quad \forall y \in \mathbf{R}^n. \tag{16.9}$$

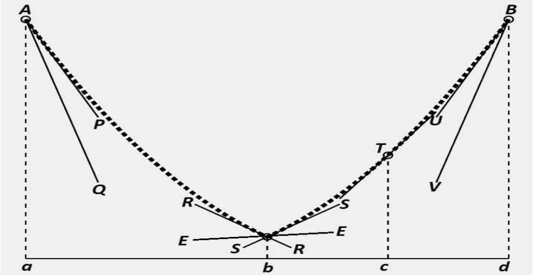


Figure III.4. Subdifferentials of univariate convex function
dotted: convex function with domain $[a,d]$
point $a$: at the boundary point $a$ of the domain, the subgradients of $f$ are the slopes of lines like $\overline{AP}$ and $\overline{AQ}$, and $\partial f(a)=\{g:g\leq \mathrm{slope}(\overline{AP})\}$
point $b$: at point $b$, subgradients are the slopes of lines like $\overline{RR}$, $\overline{EE}$, $\overline{SS}$, and $\partial f(b)=\{g:\mathrm{slope}(\overline{RR})\leq g\leq \mathrm{slope}(\overline{SS})\}$
point $c$: just one subgradient – the slope of the tangent line, taken at point $T$, to the graph of $f$
point $d$: similar to $a$, $\partial f(d)=\{g:g\geq \mathrm{slope}(\overline{UB})\}$

Thus, if $f$ admits an affine minorant which is exact at $x$, then there exists $d\in\mathbf{R}^n$ which gives rise to the inequality (16.9). In fact the reverse is also true: if $d$ is such that (16.9) holds, then the right hand side of (16.9), regarded as a function of $y$, is an affine minorant of $f$ which coincides with $f$ at $x$.

Now note that (16.9) expresses a specific property of a vector $d$ and leads to the following very important definition which generalizes the notion of gradient for smooth convex functions to nonsmooth convex function.

**Definition** III.16.8 [Subgradient of a convex function] Let $f:\mathbf{R}^n\to\mathbf{R}\cup\{+\infty\}$ be a convex function. Given $\bar{x}\in\mathbf{R}^n$, a vector $d\in\mathbf{R}^n$ is called a *subgradient* of $f$ at $\bar{x}$ if $d$ is the slope of an affine minorant of $f$ which is exact at $\bar{x}$. That is, $d$ is a subgradient of $f$ at $\bar{x}$ if $d$ satisfies

$$f(y)\geq f(\bar{x})+d^\top(y-\bar{x}),\quad \forall y\in\mathbf{R}^n.$$

The set of all subgradients of $f$ at a point $\bar{x}$ is called *subdifferential* of $f$ at $\bar{x}$ [notation: $\partial f(\bar{x})$].

Figure III.4 illustrates the notions just introduced..

Subgradients of convex functions play an important role in the theory and numerical methods for Convex Optimization – they are quite reasonable surrogates of gradients in the cases when the latter do not exist. Let us present a simple and instructive illustration of this. Recall that Theorem III.15.2 states that

> *A necessary and sufficient condition for a convex function* $f:\mathbf{R}^n\to\mathbf{R}\cup\{+\infty\}$ *to attain its minimum at a point* $x^*\in\mathrm{int}(\mathrm{Dom}\,f)$ *where* $f$ *is differentiable is that* $\nabla f(x^*)=0$.

The "nonsmooth" version of this statement is as follows.

**Proposition** III.16.9 [Necessary and sufficient optimality condition for nonsmooth convex functions] A necessary and sufficient condition for a convex function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ to attain its minimum at a point $x^* \in \operatorname{Dom} f$ is the inclusion $0 \in \partial f(x^*)$.

**Proof.** Given $x^* \in \operatorname{Dom} f$, by definition of the subdifferential, $0 \in \partial f(x^*)$ if and only if the vector $0$ is a subgradient of $f$ at $x^*$, which (by the definition of subgradient) holds if and only if

$$f(y) \geq f(x^*) + 0^\top (y - x^*), \quad \forall y \in \mathbf{R}^n,$$

which simply says that $x^*$ is a minimizer of $f$.

□

In fact, if a convex function $f$ is differentiable at $\bar{x} \in \operatorname{int}(\operatorname{Dom} f)$, then $\partial f(\bar{x})$ is the singleton $\{\nabla f(\bar{x})\}$ (see Proposition III.16.10 below). Thus, Proposition III.16.9 indeed is an extension of Theorem III.15.2 to the case when $f$ is possibly nondifferentiable.

Looking at the (absolutely trivial!) proof of Proposition III.16.9, one can ask: how such a trivial necessary and sufficient optimality condition can be useful? Why is it more informative than the tautological necessary and sufficient optimality condition "$f(x^*) \leq f(x)$ for all $x$"? Well, the definite usefulness of Fermat optimality condition stems not from the condition per se, but from our knowledge on gradients, including (but not reduced to) our ability to compute, to the extent given by the standard Calculus, the gradients. Similarly, the usefulness of Proposition III.16.9 stems from our knowledge on subgradients, including the ability, to the extent given by calculus of subgradients, to compute subdifferentials of convex functions. Let us become acquainted with the basics of this calculus.

Here is a summary of the most elementary properties of the subgradients.

**Proposition** III.16.10 Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex function. Then,

(i) $\partial f(x)$ is a closed convex set for any $x \in \operatorname{Dom} f$, $\partial f(x) \neq \varnothing$ whenever $x \in \operatorname{rint}(\operatorname{Dom} f)$ (this is the most important fact about subgradients), and $\partial f(x)$ is bounded whenever $x \in \operatorname{int}(\operatorname{Dom} f)$.

(ii) If $x \in \operatorname{int}(\operatorname{Dom} f)$ and $f$ is differentiable at $x$, then $\nabla f(x)$ is the only subgradient of $f$ at $x$: $\partial f(x) = \{\nabla f(x)\}$.

(iii) "closedness of the subdifferential mapping:" Let $g_i \in \partial f(x_i)$ and $[x_i; g_i] \to [x; g]$ as $i \to \infty$. Assume also that $x \in \operatorname{Dom} f$ and $f$ is continuous at $x$. Then, $g \in \partial f(x)$.

(iv) Let $Y$ be a nonempty convex compact subset of $\operatorname{int}(\operatorname{Dom} f)$. Then, the set $G = \{[x; g] \in \mathbf{R}^n \times \mathbf{R}^n : x \in Y, g \in \partial f(x)\}$ is compact.

**Proof.** (i) Closedness and convexity of $\partial f(x)$ are evident from their definition as (16.9) is an infinite system of nonstrict linear inequalities, indexed by $y \in \operatorname{Dom} f$, on variable $d$.

When $x \in \operatorname{rint}(\operatorname{Dom} f)$, Proposition III.16.5 provides us with an affine function which underestimates $f$ everywhere and coincides with $f$ at $x$. The slope of this affine function is clearly a subgradient of $f$ at $x$, and thus $\partial f(x) \neq \varnothing$.

Boundedness of $\partial f(x)$ when $x \in \operatorname{int}(\operatorname{Dom} f)$ is an immediate consequence of item (iv) to be proved in the mean time.

(ii) Suppose $x \in \operatorname{int}(\operatorname{Dom} f)$ and $f$ is differentiable at $x$. Then, by the gradient inequality we have $\nabla f(x) \in \partial f(x)$. Let us prove that in this case, $\nabla f(x)$ is the only subgradient of $f$ at $x$. Suppose that $d \in \partial f(x)$. Then, by definition of the subgradient, we have

$$f(y) - f(x) \geq d^\top (y - x) \quad \forall y \in \mathbf{R}^n.$$

Now, consider any fixed direction $h \in \mathbf{R}^n$, real number $t > 0$. By substituting $y = x + th$ in the preceding inequality and then dividing both sides of the resulting inequality by $t$, we obtain

$$\frac{f(x + th) - f(x)}{t} \geq d^\top h.$$

Taking the limit of both sides of this inequality as $t \to +0$, we get

$$h^\top \nabla f(x) \geq h^\top d.$$

Since $h$ was an arbitrary direction, this inequality is valid for all $h$, which is possible if and only if $d = \nabla f(x)$.

(iii) Under the premise of this part, for every $y \in \mathbf{R}^n$ and for all $i = 1, 2, \ldots$, we have

$$f(y) \geq f(x_i) + g_i^\top (y - x_i).$$

Passing to limit as $i \to \infty$, we get

$$f(y) \geq f(x) + g^\top (y - x), \quad \forall y \in \mathbf{R}^n,$$

and thus $g \in \partial f(x)$.

(iv) Let us start with the following observation:

> *Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be convex and consider $\bar{x} \in \operatorname{int}(\operatorname{Dom} f)$. Suppose $f$ is Lipschitz continuous, with constant $L$ with respect to $\|\cdot\|_2$, in a neighborhood $V$ of $\bar{x}$, that is, $|f(x) - f(y)| \leq L\|x - y\|_2$ for all $x, y \in V$. Then,*
>
> $$\|g\|_2 \leq L, \quad \forall g \in \partial f(\bar{x}).$$
>
> Indeed, when $g \in \partial f(x)$, by the subgradient inequality and Lipschitz continuity of $f$ we have $g^\top (y - x) \leq f(y) - f(x) \leq L\|x - y\|_2$ for all $y$ close to $x$, that is, $g^\top h \leq L\|h\|_2$ for all $h \in \mathbf{R}^n$, implying that $\|g\|_2 \leq L$.

Now, under the premise of (iv), we can find a compact set $Y' \subset \operatorname{int}(\operatorname{Dom} f)$ such that $Y \subset \operatorname{int} Y'$. By Theorem III.14.9, $f$ is Lipschitz continuous, with certain constant $L$, on $Y'$, implying by our observation that $\|g\|_2 \leq L$ for all $g \in \partial f(x)$ with $x \in Y$. Then, as $Y$ is bounded, the set $G = \{[x; g] \in \mathbf{R}^n \times \mathbf{R}^n : x \in Y, g \in \partial f(x)\}$ is bounded. Moreover, this set is closed by (iii) (recall that $Y$ is compact and $f$ is continuous on $Y$), so that $G$ is compact. □

Proposition III.16.10 sheds light onto why subgradients are good surrogates of gradients: at a point where gradient exists, the gradient is the unique subgradient; but, in contrast to the gradient, a subgradient exists basically everywhere (for sure in the relative interior of the domain of the function). Let us examine a simple function and its subgradients.

**Example** III.16.1 Consider the function $f : \mathbf{R} \to \mathbf{R}$ given by

$$f(x) = |x| = \max\{x, -x\}.$$

This function $f$ is, of course, convex (as maximum of two linear forms $x$ and $-x$). Whenever $x \neq 0$, $f$ is differentiable at $x$ with the derivative $+1$ for $x > 0$ and $-1$ for $x < 0$. At the point $x = 0$, $f$ is not differentiable; nevertheless, it must have subgradients at this point (since $0 \in \operatorname{int}(\operatorname{Dom} f)$). And indeed, it is immediately seen that the subgradients of $f$ at $x = 0$ are exactly the reals from the segment $[-1, 1]$. Thus,

$$\partial|x| = \begin{cases} \{-1\}, & \text{if } x < 0, \\ [-1, 1], & \text{if } x = 0, \\ \{+1\}, & \text{if } x > 0. \end{cases}$$

It is important to note that at the points from the relative boundary of the domain of a convex function, even a "good" one, we may not have any subgradients. That is, it is possible to have $\partial f(x) = \varnothing$ for a convex function $f$ at a point $x \in \operatorname{rbd}(\operatorname{Dom} f)$. We give an example of this next.

**Example** III.16.2 Consider the function

$$f(x) = \begin{cases} -\sqrt{x}, & \text{if } x \geq 0, \\ +\infty, & \text{if } x < 0. \end{cases}$$

Convexity of this function follows from convexity of its domain and Example III.13.2. Consider the point $[0; f(0)] \in \operatorname{rbd}(\operatorname{epi}\{f\})$. It is clear that at this point $[0; f(0)]$ there is no non-vertical supporting line to the set $\operatorname{epi}\{f\}$, and, consequently, there is no affine minorant of the function which is exact at $x = 0$.

A significant – and important – part of Convex Analysis deals with *subgradient calculus*, which is the set of rules for computing subgradients of "composite" functions, like sums, superpositions, maxima, etc., given subgradients of the operands. These rules extend the standard Calculus rules to nonsmooth convex functions, and they are very nice and useful. Here, we list several "self-evident" versions of these rules:

1. *Subdifferential of nonnegative weighted sum*: Let $f, g$ be convex functions on $\mathbf{R}^n$. Consider $x \in \operatorname{Dom} f \cap \operatorname{Dom} g$ and $\lambda, \mu \in \mathbf{R}_+$. If $d \in \partial f(x)$ and $e \in \partial g(x)$, then $\lambda d + \mu e \in \partial[\lambda f + \mu g](x)$.
2. *Subdifferential of pointwise supremum*: Let $\{f_\alpha(\cdot)\}_{\alpha \in \mathcal{A}}$ be a family of convex functions on $\mathbf{R}^n$. Consider the convex function $\overline{f}(x) := \sup_{\alpha \in \mathcal{A}} f_\alpha(x)$ and $\bar{x} \in \operatorname{Dom} \overline{f}$. Suppose that $\bar{\alpha}$ is such that $\overline{f}(\bar{x}) = f_{\bar{\alpha}}(\bar{x})$. Then, for any $\bar{d} \in \partial f_{\bar{\alpha}}(\bar{x})$, we

have $\bar{d} \in \partial \overline{f}(\bar{x})$.
Here is the justification: $\overline{f}(x) \geq f_{\bar{\alpha}}(x) \geq f_{\bar{\alpha}}(\bar{x}) + \bar{d}^\top (x - \bar{x})$ for all $x \in \mathbf{R}^n$, that is, $\overline{f}(x) \geq f_{\bar{\alpha}}(\bar{x}) + \bar{d}^\top (x - \bar{x})$ for all $x$, and this inequality becomes equality when $x = \bar{x}$ as $\overline{f}(\bar{x}) = f_{\bar{\alpha}}(\bar{x})$.

3. *Subdifferential of convex monotone/affine superposition [chain rule]*:
Let $f_1(x), \ldots, f_K(x)$ be convex functions on $\mathbf{R}^n$, and let $F$ be a convex function on $\mathbf{R}^K$. Suppose for some $0 \leq k \leq K$ the functions $f_1, \ldots, f_k$ are affine, and also $F(y)$ is nondecreasing in $y_{k+1}, \ldots, y_K$. Recall that the superposition

$$g(x) = \begin{cases} F(f_1(x), \ldots, f_K(x)), & \text{if } f_k(x) < +\infty \text{ for all } k \leq K, \\ +\infty, & \text{otherwise}, \end{cases}$$

is a convex function of $x$. Consider $\bar{x} \in \bigcap_{i=1}^K \mathrm{Dom}(f_i)$ such that

$$\bar{y} := [f_1(\bar{x}); \ldots; f_K(\bar{x})] \in \mathrm{Dom}\, F.$$

Let $g^i \in \partial f_i(\bar{x})$ for all $i \leq K$ and $e \in \partial F(\bar{y})$. Then, the vector $\overline{d} := \sum_{i=1}^K e_i g^i$ satisfies $\overline{d} \in \partial [F(f_1, \ldots, f_K)](\bar{x})$.

The justification of this is as follows. Let $h \in \mathbf{R}^n$ be an arbitrary direction, and consider $x = \bar{x} + h$ and $y = \bar{y} + [(g^1)^\top h; (g^2)^\top h; \ldots; (g^K)^\top h]$. Then, for all $i \leq K$, we have

$$y_i = \bar{y}_i + (g^i)^\top h = f_i(\bar{x}) + (g^i)^\top (x - \bar{x}).$$

For $i \leq k$ as $f_1, \ldots, f_k$ are affine we have $y_i = f_i(x)$; and for $i > k$, as $g^i \in \partial f_i(\bar{x})$, we have $f_i(x) \geq y_i$. Consequently, using the partial monotonicity of $F$, we conclude $F(f_1(x), \ldots, f_K(x)) \geq F(y)$. Now, $e \in \partial F(\bar{y})$, which implies the first inequality in the following chain:

$$\begin{aligned} F(y) \geq F(\bar{y}) + e^\top [y - \bar{y}] &= F(f_1(\bar{x}), \ldots, f_K(\bar{x})) + \left[\sum_i e_i g^i\right]^\top h \\ &= F(f_1(\bar{x}), \ldots, f_K(\bar{x})) + \overline{d}^\top (x - \bar{x}). \end{aligned}$$

Here, the first equality follows from $y_i = \bar{y}_i + (g^i)^\top h$ for all $i$, and the second equality is due to $x = \bar{x} + h$. Finally, this inequality combines with $F(f_1(x), \ldots, f_K(x)) \geq F(y)$ to imply that

$$F(f_1(x), \ldots, f_K(x)) \geq F(f_1(\overline{x}), \ldots, f_K(\overline{x})) + \overline{d}^\top (x - \overline{x}).$$

as desired.

Advanced versions of these rules, under appropriate assumptions, describe how the entire subdifferentials of the resulting functions are obtained from those of operands; the related considerations, however, are beyond our scope.

We close this section by providing an illustration of the second rule.

**Example** III.16.3 In this example, we will examine the subgradients of spectral norm of a symmetric matrix. For $X \in \mathbf{S}^n$, its spectral norm $\|X\|$ is given by

$$\|X\| := \max_{e \in \mathbf{R}^n} \left\{ |e^\top X e| : \ \|e\|_2 \leq 1 \right\}.$$

From this definition, it is clear that $\|X\|$ is a convex function of $X$.

Given $X$, we can compute a maximizer of the optimization problem defining $\|X\|$ by finding an eigenvalue-eigenvector pair $(\lambda_X, e_*)$ of $X$ such that $\lambda_X$ is the largest in magnitude of the eigenvalues of $X$. Let $e_X$ be the unit length normalization of $e_*$. Thus,

$$\begin{aligned}\|X\| = \max_{e:\|e\|_2 \leq 1} |e^\top X e| = \max_{e:\|e\|_2 \leq 1} |\mathrm{Tr}(X[ee^\top])| &= |\mathrm{Tr}(X[e_X e_X^\top])| \\ &= \mathrm{Tr}(X[\mathrm{sign}(\lambda_X) e_X e_X^\top]),\end{aligned}$$

where the third equality follows from the choice of $e_X$, and the last equality is due to $\mathrm{Tr}(X[e_X e_X^\top]) = \lambda_X$. Then, using item 2 in our subgradient calculus rules, we conclude that the symmetric matrix $E(X) := \mathrm{sign}(\lambda_X) e_X e_X^\top$ is a subgradient of the function $\|\cdot\|$ at $X$. That is,

$$\|Y\| \geq \mathrm{Tr}(Y E(X)) = \|X\| + \mathrm{Tr}(E(X)(Y - X)), \qquad \forall Y \in \mathbf{S}^n,$$

where the equality holds due to the choice of $E(X)$ guaranteeing $\|X\| = \mathrm{Tr}(X E(X))$. To see that the above is indeed a subgradient inequality, take into account that the inner product on $\mathbf{S}^n$ is the Frobenius inner product $\langle A, B \rangle = \mathrm{Tr}(AB)$.

Let us close this example by discussing smoothness properties of $\|X\|$. Recall that every norm $\|y\|$ in $\mathbf{R}^m$ is nonsmooth at the origin $y = 0$. However, $\|X\|$ is a nonsmooth function of $X$ even at points other than $X = 0$. In fact, $\|\cdot\|$ is continuously differentiable in a neighborhood of every point $X \in \mathbf{S}^n$ where the maximum magnitude eigenvalue is unique and is of multiplicity 1, and can lose smoothness at other points.

## 16.3 Subdifferentials and directional derivatives of convex functions

Let $f$ be a convex function on $\mathbf{R}^n$ and consider $x \in \mathrm{int}(\mathrm{Dom}\, f)$ (in fact our construction to follow admits an immediate generalization for the case when $x \in \mathrm{rint}\,(\mathrm{Dom}\, f)$ as well). Consider any direction $h \in \mathbf{R}^n$ and the univariate function

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

associated with $h$. Note that $\phi(t)$ is a real-valued convex function of $t$ in a neighborhood of 0, thus for all small enough positive $s$ and $t$ we have

$$\frac{\phi(0) - \phi(-s)}{s} \leq \frac{\phi(t) - \phi(0)}{t}.$$

Moreover, the right hand side of this inequality is nondecreasing in $t$, hence it implies the existence of *directional derivative of $f$ taken at $x$ along the direction $h$*, i.e., the quantity

$$Df(x)[h] = \lim_{t\to+0} \frac{f(x+th)-f(x)}{t}.$$

As a function of $h$, the function $Df(x)[h]$ is clearly positively homogeneous of degree 1:

$$\lambda \geq 0 \implies Df(x)[\lambda h] = \lambda Df(x)[h].$$

In addition, $Df(x)[h]$ is a convex function of $h$. To see this, let $r > 0$ be such that the Euclidean ball $B$ centered at $x$ and of radius $r$ is contained in $\mathrm{Dom}\, f$. Then, the functions

$$f_s(h) := \frac{f(x+sh)-f(x)}{s}$$

are convex over the domain $h \in B$ as long as $0 < s \leq 1$. Moreover, as $s \to +0$, on $B$ they pointwise converge to $Df(x)[h]$, which clearly implies the convexity of $Df(x)[h]$ as a function of $h \in B$. Finally, since as a function of $h$, $Df(x)[h]$ is positively homogeneous of degree 1, its convexity on $B$ clearly implies its convexity on the entire $\mathbf{R}^n$. Note that convexity of $f$ implies that

$$f(x+h) \geq f(x) + Df(x)[h], \quad \forall (x,h : x \in \mathrm{int}(\mathrm{Dom}\, f),\ x+h \in \mathrm{Dom}\, f). \quad (16.10)$$

We have arrived at the following result.

**Lemma** III.16.11 Let $f$ be a convex function on $\mathbf{R}^n$. For any $x \in \mathrm{int}(\mathrm{Dom}\, f)$, the subdifferential $\partial f(x)$ of $f$ at $x$ is exactly the same as the subdifferential of $Df(x)[\cdot]$ at the origin.

**Proof.** Suppose $d$ is a subgradient of $f$ at $x$, and thus $f(x+th)-f(x) \geq td^\top h$ holds for all $h \in \mathbf{R}^n$ and all small enough $t > 0$. This then implies that $Df(x)[h] \geq d^\top h$, that is, $d$ is a subgradient of $Df(x)[h]$ at $h = 0$. For the reverse direction, suppose that $d$ is a subgradient of $Df(x)[\cdot]$ at $h = 0$. Then, we have $Df(x)[h] \geq d^\top h$ for all $h$, implying, by (16.10), that $f(x+h) \geq f(x) + d^\top h$ whenever $x + h \in \mathrm{Dom}\, f$. □

Our goal is to demonstrate that when $x \in \mathrm{int}(\mathrm{Dom}\, f)$, the subdifferential $\partial f(x)$ is large enough to fully define $Df(x)[\cdot]$:

**Theorem** III.16.12 Let $f$ be a convex function on $\mathbf{R}^n$ and $x \in \mathrm{int}(\mathrm{Dom}\, f)$. Then,

$$Df(x)[h] = \max_d \left\{ d^\top h : \ d \in \partial f(x) \right\}. \quad (16.11)$$

Note that for a convex function $f$, at $x \in \mathrm{int}(\mathrm{Dom}\, f)$, we know $\partial f(x)$ is bounded and thus in (16.11) the use of $\max$ as opposed to $\sup$ is justified. We are about to obtain this theorem from the fundamental (finite-dimensional) *Hahn-Banach Theorem* given below.

**Theorem** III.16.13 [Hahn-Banach Theorem, finite-dimensional version] Let $D(\cdot)$ be a real-valued convex positively homogeneous, of degree 1, function on $\mathbf{R}^n$, $e \in \mathbf{R}^n$ and $E$ be a linear subspace of $\mathbf{R}^n$ such that

$$e^\top z \le D(z), \quad \forall z \in E.$$

Then, there exists $e' \in \mathbf{R}^n$ such that $[e']^\top z \equiv e^\top z$ for all $z \in E$ and $[e']^\top z \le D(z)$ for all $z \in \mathbf{R}^n$. In other words, a linear functional defined on a linear subspace of $\mathbf{R}^N$ and majorized on this subspace by $D(\cdot)$, can be extended from the subspace to a linear functional on the entire space in such a way that the extension is majorized by $D(\cdot)$ everywhere.

**Proof of Theorem III.16.12.** In this proof we will show that Theorem III.16.13 implies Theorem III.16.12.

Consider any $d \in \partial f(x)$. Then, by Lemma III.16.11 we have $d$ is in the subdifferential $Df(x)[\cdot]$ at the origin, i.e., $Df(x)[h] \ge d^\top h$. Thus, we conclude

$$Df(x)[h] \ge \max_d \left\{ d^\top h : \ d \in \partial f(x) \right\}.$$

To prove the opposite inequality, let us fix $h \in \mathbf{R}^n$, and let us verify that $g := Df(x)[h] \le \max_{d\in\partial f(x)} d^\top h$. There is nothing to prove when $h = 0$, so let $h \neq 0$. Setting $\phi(t) := Df(x)[th]$, we get a convex (since, as we already know, $Df(x)[\cdot]$ is convex) univariate function such that $\phi(t) = gt$ for $t \ge 0$. Then, this together with the convexity of $\phi$ implies that $\phi(t) \ge gt$ for all $t \in \mathbf{R}$. By applying Hahn-Banach Theorem to the function $D(z) := Df(x)[z]$ (we already know that this function satisfies the premise of Hahn-Banach Theorem), $E := \mathbf{R}(h)$ and the linear form $e^\top[th] = gt$, $t \in \mathbf{R}$, on $E$, we conclude that there exists $e \in \mathbf{R}^n$ such that $e^\top u \le Df(x)[u]$ for all $u \in \mathbf{R}^n$ and $e^\top h = g = Df(x)[h]$. By the first relation, $e$ is a subgradient of $Df(x)[\cdot]$ at the origin and thus, by Lemma III.16.11, $e \in \partial f(x)$, so that

$$\max_{d\in\partial f(x)} d^\top h \ge e^\top h = Df(x)[h].$$

Thus, the right hand side in (16.11) is $\ge$ the left hand side. The opposite inequality has already been proved, so that (16.11) is an equality. □

**Remark** III.16.14 The above reasoning implies the following fact:

> *Let $f$ be a convex function. Consider any $x \in \text{int}(\text{Dom}\, f)$ and any affine plane $M$ such that $x \in M$. Then, "every subgradient, taken at $x$, of the restriction $f\Big|_M$ of $f$ onto $M$ can be obtained from the subgradient of $f$." That is, if $e$ is such that $f(y) \ge f(x) + e^\top(y - x)$ for all $y \in M$, then there exists $e' \in \partial f(x)$ such that $e^\top(y-x) = (e')^\top(y-x)$ for all $y \in M$.*

For completeness we also present a proof of finite-dimensional version of Hahn-Banach Theorem (Theorem III.16.13).

**Proof of Theorem III.16.13.** We are given a linear functional defined on a linear subspace $E$ of $\mathbf{R}^n$ and this linear functional is majorized on this subspace by $D(\cdot)$; we want to prove that this linear functional can be extended from $E$ to a linear functional on the entire space majorized by $D$ everywhere. To this end, it clearly suffices to prove this fact when $E$ is of dimension $n-1$ (as given this fact for $E$ satisfying $\dim(E)=n-1$, we can build the desired extension in the general case by iterating extensions "increasing the dimension by 1"). Thus, suppose $\mathbf{R}^n=\mathbf{R}(g)+E$, where $g\notin E$. Note that in order to specify the desired vector $e'$ all we need is to determine what the value of $\alpha := (e')^\top g$ should be, since then $e'$ will be uniquely defined by the relation

$$(e')^\top(\lambda g+h)=\lambda\alpha+e^\top h,\quad \forall(\lambda\in\mathbf{R},\ h\in E).$$

Therefore, we wish to find $\alpha\in\mathbf{R}$ such that

$$\lambda\alpha+e^\top h\le D(\lambda g+h),\quad \forall(\lambda\in\mathbf{R}, h\in E).$$

Note that when $\lambda=0$, the preceding inequality is automatically satisfied due to the premise of the theorem on $e$. Moreover, as $D$ is positively homogeneous of degree 1 and $E$ is a linear subspace, all we need is to ensure that the preceding inequality is valid when $\lambda=\pm 1$, that is, to ensure that

$$\begin{aligned}&\alpha\le D(g+h)-e^\top h,\quad \forall h\in E &&(a)\\ &\alpha\ge -D(-g+h)+e^\top h,\quad \forall h\in E &&(b)\end{aligned}$$

Now, to justify that $\alpha\in\mathbf{R}$ satisfying the relations (a) and (b) above indeed exists, we need to show that every possible value of the right hand side in $(a)$ is $\ge$ than every possible value of the right hand side in $(b)$, that is, that $D(g+h)-e^\top h\ge -D(-g+h')+e^\top h'$ whenever $h,h'\in E$. By rearranging the terms, we thus need to show that

$$e^\top[h+h']\le D(h+g)+D(-g+h'),\quad \forall h,h'\in E. \tag{16.12}$$

Now, note that

$$\begin{aligned}D(h+g)+D(-g+h')&=2\left(\frac{1}{2}D(h+g)+\frac{1}{2}D(-g+h')\right)\\ &\ge 2D\left(\frac{1}{2}(h+g)+\frac{1}{2}(-g+h')\right)\\ &=D(h+h')\\ &\ge e^\top[h+h'],\end{aligned}$$

where the first inequality follows from convexity of the function $D$, the second equality is due to $D$ being positively homogeneous of degree 1, and the last inequality is due to the facts that $h+h'\in E$ and $e^\top z$ being majorized by $D(z)$ on $E$. Hence, (16.12) is proved. □

**Remark** III.16.15 The advantage of the preceding proof of Hahn-Banach Theorem in finite dimensional case is that it straightforwardly combines with what is called *transfinite induction* to yield Hahn-Banach Theorem in the case when $\mathbf{R}^n$

is replaced with arbitrary, perhaps infinite dimensional, linear space and extension of linear functional from linear subspace on the entire space which preserves majorization by a given convex and positively homogeneous, of degree 1, function on the space.

In the finite-dimensional case, alternatively, we can prove Hahn-Banach Theorem, via Separation Theorem: without loss of generality we can assume that $E \neq \mathbf{R}^n$. Define the sets $T := \{[x;t] : x \in \mathbf{R}^n,\, t \geq D(x)\}$ and $S := \{[h;t] : h \in E,\, t = e^\top h\}$. Then, we get two nonempty convex sets with non-intersecting relative interiors (as $E \neq \mathbf{R}^n$ and $D(h)$ majorizes $e^\top h$ on $E$). Then, by Separation Theorem there exists a nontrivial ($r \neq 0$) linear functional $r^\top[x;t] = d^\top x + at$, which separates $S$ and $T$, i.e., $\inf_{y\in T} r^\top y \geq \sup_{y\in S} r^\top y$. Moreover, since $S$ is a linear subspace, we deduce that $\sup_{y\in S} r^\top y$ is either 0 or $+\infty$. Also, as $T \neq \varnothing$, we conclude $+\infty > \inf_{y\in T} r^\top y \geq \sup_{y\in S} r^\top y$, and thus we must have $\sup_{y\in S} r^\top y = 0$. In addition, we claim that $a > 0$. Indeed,

$$0 = \sup_{y\in S} r^\top y \leq \inf_{y\in T} r^\top y = \inf_{x\in\mathbf{R}^n, t\in\mathbf{R}} \left\{d^\top x + at :\ t \geq D(x)\right\}.$$

As the right hand side value must be strictly greater than $-\infty$, we see $a \geq 0$. Also, if $a = 0$ were to hold, then from $r \neq 0$ we must have $d \neq 0$. Moreover, when $a = 0$ and $d \neq 0$, since $D(\cdot)$ is a finite valued function we have $\inf_{x\in\mathbf{R}^n, t\in\mathbf{R}} \left\{d^\top x + at :\ t \geq D(x)\right\} = -\infty$. But, this contradicts to the infimum being bounded below by 0. Now that $a > 0$, by multiplying $r$ by $a^{-1}$, we get a separator of the form $r = [-e'; 1]$. Then,

$$\begin{aligned} 0 = \sup_{y\in S} r^\top y = \sup_{[h;t]\in S} r^\top[h;t] = \sup_{[h;t]\in S} \{(-e')^\top h + t\} &= \sup_{h\in E, t = e^\top h} \{(-e')^\top h + t\} \\ &= \sup_{h\in E} \{(-e')^\top h + e^\top h\}, \end{aligned}$$

and so we conclude $(e')^\top h = e^\top h$ holds for all $h \in E$. Note also that the relation $0 = \sup_{[h;t]\in S} r^\top[h;t] \leq \inf_{[x;t]\in T} r^\top[x;t]$ is nothing but $(e')^\top x \leq D(x)$ for all $x \in \mathbf{R}^n$.

Hahn-Banach Theorem is extremely important on its own right, and our way of proving Theorem III.16.12 was motivated by the desire to acquaint the reader with Hahn-Banach Theorem. If justification of Theorem III.16.12 were to be our sole goal, we could achieve this goal in a much broader setting and at a cheaper cost, see solution to Exercise IV.29.D.4.

# 17

# ★ Legendre transform

## 17.1 Legendre transform : Definition and examples

Let $f$ be a proper convex function. We know that $f$ is "basically" the supremum of all its affine minorants. In fact, this is exactly the case when $f$ is lower semicontinuous in addition to being convex and proper; otherwise (i.e., if it is not lower semicontinuous) the corresponding equality takes place everywhere except, perhaps, some points from $\mathrm{rbd}(\mathrm{Dom}\, f)$. Now, let us look into the question of when an affine function $d^\top x - a$ is an affine minorant of $f$. This is the case if and only if for all $x$ we have

$$f(x) \geq d^\top x - a,$$

which holds if and only if for all $x$ we have

$$a \geq d^\top x - f(x).$$

Thus, we see that if the slope $d$ of an affine function $d^\top x - a$ is fixed, then in order for the function to be a minorant of $f$, it needs to satisfy

$$a \geq \sup_{x \in \mathbf{R}^n} \left\{ d^\top x - f(x) \right\}.$$

The supremum in the right hand side of this inequality is a certain function of $d$, and we arrive at the following important definition.

**Definition** III.17.1 [Legendre transform (Fenchel dual) of a convex function] Given a convex function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$, its *Legendre transform* (also called the *Fenchel conjugate* or *Fenchel dual*) [notation: $f^*$] is the function

$$f^*(d) := \sup_{x \in \mathbf{R}^n} \left\{ d^\top x - f(x) \right\} : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}.$$

Geometrically, the Legendre transform answers the following question: given a slope $d$ of an affine function, i.e., given the hyperplane $t = d^\top x$ in $\mathbf{R}^{n+1}$, what is the minimal "shift down" of this hyperplane so that it can be placed below the graph of $f$?

The definition of Legendre transform immediately leads to a simple yet useful observation.

**Fact** III.17.2 Given a proper convex function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$, its Legendre transform $f^*$ is a proper convex lower semicontinuous function.

Let us see some examples of simple functions and their Legendre transforms.

**Example** III.17.1

1. Given $a \in \mathbf{R}$, consider the constant function
$$f(x) \equiv a.$$
Its Legendre transform is given by
$$f^*(d) = \sup_{x \in \mathbf{R}^n} \left\{ d^\top x - f(x) \right\} = \sup_{x \in \mathbf{R}^n} \left\{ d^\top x - a \right\} = \begin{cases} -a, & \text{if } d = 0, \\ +\infty, & \text{otherwise.} \end{cases}$$
2. Consider the affine function
$$f(x) = c^\top x + a, \quad \forall x \in \mathbf{R}^n.$$
Its Legendre transform is given by
$$f^*(d) = \sup_{x \in \mathbf{R}^n} \left\{ d^\top x - f(x) \right\} = \sup_{x \in \mathbf{R}^n} \left\{ d^\top x - (c^\top x + a) \right\} = \begin{cases} -a, & \text{if } d = c, \\ +\infty, & \text{otherwise.} \end{cases}$$
3. Consider the strictly convex quadratic function
$$f(x) = \frac{1}{2} x^\top A x,$$
where $A \in \mathbf{S}^n$ is positive definite. Its Legendre transform is given by
$$f^*(d) = \sup_{x \in \mathbf{R}^n} \left\{ d^\top x - f(x) \right\} = \sup_{x \in \mathbf{R}^n} \left\{ d^\top x - \left( \frac{1}{2} x^\top A x \right) \right\} = \frac{1}{2} d^\top A^{-1} d,$$
where the final equality holds by examining the first-order necessary and sufficient optimality (maximum) condition for differentiable concave functions.
4. Consider the function $f : \mathbf{R} \to \mathbf{R}$ given by $f(x) = |x|^p/p$, where $p \in (1, \infty)$. Then, using the first-order optimality conditions we see that the Legendre transform of $f$ is given by
$$f^*(d) = \sup_{x \in \mathbf{R}} \left\{ dx - \frac{|x|^p}{p} \right\} = \frac{|d|^q}{q},$$
where $q$ satisfies $\frac{1}{p} + \frac{1}{q} = 1$.
5. Suppose $f$ is a proper convex function and the function $g$ is defined to be $g(x) = f(x - a)$. Then, the Legendre transform of $g$ satisfies
$$\begin{aligned} g^*(d) = \sup_{x \in \mathbf{R}^n} \left\{ d^\top x - g(x) \right\} &= \sup_{x \in \mathbf{R}^n} \left\{ d^\top x - f(x - a) \right\} \\ &= \sup_{x' \in \mathbf{R}^n} \left\{ d^\top (a + x') - f(x') \right\} = d^\top a + f^*(d). \end{aligned}$$

## 17.2 Legendre transform : Main properties

Given a function $f$, we define its *biconjugate* (denoted by $f^{**}$) as the Legendre transform of the function $f^*$, that is,

$$f^{**} := (f^*)^*.$$

The most elementary (and the most fundamental) fact about the Legendre transform is its involutive property ("symmetry") which we discuss next. In particular, this symmetry property of Legendre transform gives us an alternative representation of $f$ in terms of its affine minorants.

**Proposition** III.17.3 Let $f$ be a proper convex function. Then, its biconjugate $f^{**}$ is exactly the closure of $f$, i.e.,

$$f^{**} = \mathrm{cl}\, f.$$

In particular, when $f$ is proper convex and also *lower semicontinuous*, then it is precisely the Legendre transform of its Legendre transform, and thus

$$f(x) = f^{**}(x) = \sup_d \left\{x^\top d - f^*(d)\right\}.$$

**Proof.** First, by Fact III.17.2 $f^*$ is a proper lsc convex function, so that $f^{**}$, once again by Fact III.17.2, is a proper lsc convex function as well. Next, by definition,

$$f^{**}(x) = (f^*)^*(x) = \sup_{d\in\mathbf{R}^n} \left\{x^\top d - f^*(d)\right\} = \sup_{d\in\mathbf{R}^n, a\geq f^*(d)} \left\{d^\top x - a\right\}.$$

Now, recall from the origin of the Legendre transform that $a \geq f^*(d)$ if and only if the affine function $d^\top x - a$ is a minorant of $f$. Thus, $\sup\limits_{d\in\mathbf{R}^n, a\geq f^*(d)} \left\{d^\top x - a\right\}$ is exactly the supremum of all affine minorants of $f$, and this supremum, by Proposition III.16.7 is nothing but the closure of $f$. Finally, when $f$ is proper convex and lsc, $f = \mathrm{cl}\, f$ by Observation III.16.6, that is $f^{**} = \mathrm{cl}\, f$ is the same as $f^{**} = f$. □

The Legendre transform is a very powerful descriptive tool, i.e., it is a "global" transformation, so that *local* properties of $f^*$ correspond to *global* properties of $f$. Below we give a number important consequences of Legendre transform highlighting this.

Let $f$ be a proper convex lsc function.

**A.** By Proposition III.17.3, the Legendre transform $f^*(d) = \sup_x \left\{x^\top d - f(x)\right\}$ is a proper convex lsc function and $f(x) = \sup_d \left\{x^\top d - f^*(d)\right\}$. Since $f^*(d) \geq d^\top x - f(x)$ for all $x$, we have

$$x^\top d \leq f(x) + f^*(d), \quad \forall x, d \in \mathbf{R}^n. \tag{17.1}$$

Moreover, inequality in (17.1) becomes equality if and only if $x \in \mathrm{Dom}\, f$ and $d \in \partial f(x)$, same as if and only if $d \in \mathrm{Dom}\, f^*$ and $x \in \partial f^*(d)$.

**Proof.**

All we need is to justify the "moreover" part of the claim. Let $x, d \in \mathbf{R}^n$, and let us prove that $d^\top x = f(x) + f^*(d)$ if and only if $d \in \partial f(x)$. In one direction: when $d \in \partial f(x)$, we have $x \in \operatorname{Dom} f$ and $f(z) \geq f(x) + d^\top(z - x)$ for all $z \in \mathbf{R}^n$, so

$$f^*(d) = \sup_{z \in \mathbf{R}^n} \left\{ d^\top z - f(z) \right\} \leq \sup_{z \in \mathbf{R}^n} \left\{ d^\top z - f(x) - d^\top(z - x) \right\} = d^\top x - f(x).$$

Thus, $f^*(d) \leq f(x) - d^\top x$; since by $(17.1)$ strict inequality here is impossible, we conclude that when $d \in \partial f(x)$, the inequality in $(17.1)$ is equality. In the opposite direction: let $d, x$ be such that the inequality in $(17.1)$ is equality, and let us prove that $d \in \partial f(x)$. Indeed, in this case $x \in \operatorname{Dom} f$ and $d^\top x - f(x) = f^*(d) \geq d^\top z - f(z)$ for all $z \in \mathbf{R}^n$ (recall the definition of Legendre transform $f^*(d)$), that is, $f(z) \geq f(x) + d^\top(z - x)$ for all $z$, implying that $d \in \partial f(x)$.

We have seen that when $f$ is proper convex and lsc, then $d \in \partial f(x)$ if and only if $d^\top x = f(x) + f^*(d)$. By Fact III.17.2 and Proposition III.17.3, we can swap the roles of $f$ and $f^*$, so that the inequality in $(17.1)$ is equality if and only if $x \in \partial f^*(d)$. □

**B.** We always have $\inf_x f(x) = -f^*(0)$. Thus, $f$ is bounded from below if and only if $0 \in \operatorname{Dom} f^*$. Moreover, $f$ attains its minimum if and only if $\partial f^*(0) \neq \varnothing$, in which case $\operatorname{Argmin} f = \partial f^*(0)$.

**Proof.** By definition of the Legendre transform we have $f^*(0) = -\inf_x f(x)$. Thus, $f$ is below bounded if and only if $0 \in \operatorname{Dom} f^*$. To see the rest of the claim, suppose now that $0 \in \operatorname{Dom} f^*$ and so $f^*(0) \in \mathbf{R}$. As $\inf_x f(x) = -f^*(0)$ implies $f(x) + f^*(0) \geq 0$ for all $x \in \mathbf{R}^n$, we deduce that the equality $f(x) + f^*(0) = 0$ holds exactly for $x$'s that are the minimizers of $f$. Also, by part $\mathbf{A}$, we conclude that when $0 \in \operatorname{Dom} f^*$, inequality in $(17.1)$ with $d = 0$ holds as equality if and only if $x \in \partial f^*(0)$. Combining the last two statements, we conclude that when $f$ is below bounded (or, which is the same, when $0 \in \operatorname{Dom} f^*$), the set of minimizers of $f$ is exactly $\partial f^*(0)$. □

**C.** By part $\mathbf{A}$, we have (i) $\bar{d} \in \partial f(\bar{x})$ if and only if (ii) $\bar{x} \in \partial f^*(\bar{d})$, and moreover both (i) and (ii) hold simultaneously if and only if (iii) with $x = \bar{x}$, $d = \bar{d}$, inequality in $(17.1)$ holds as equality.

**C′.** Here is a nice special case of $\mathbf{C}$: *Let $f$ be a proper convex lsc function on $\mathbf{R}^n$. Assume that the domains of $f$ and $f^*$ are open, and that these functions are continuously differentiable and strictly convex on their domains. Then, the map $x \mapsto \nabla f(x) : \operatorname{Dom} f \to \mathbf{R}^n$ is a one-to-one mapping of $\operatorname{Dom} f$ onto $\operatorname{Dom} f^*$, and the inverse of this map is given by $y \mapsto \nabla f^*(y) : \operatorname{Dom} f^* \to \mathbf{R}^n$.*

**Proof.** First, we claim that under the given premise, $\nabla f(x)$ is an embedding of $\operatorname{Dom} f$ into $\mathbf{R}^n$, i.e., $\nabla f(x) = \nabla f(x')$ implies $x = x'$. Assume for contradiction that this is not the case. Consider the function $g(u) := f(u) - u^\top \nabla f(x)$. This function is strictly convex, since $f$ is so. However, when $\nabla f(x) = \nabla f(x')$, both $x$ and $x'$ are minimizers of $g(u)$, which is not possible as $g(u)$ is strictly convex and by Theorem III.15.1 its minimizer, if any exists, must be unique. Thus, $x \mapsto \nabla f(x)$ is an embedding of $\operatorname{Dom} f$ into $\mathbf{R}^n$. Next, when $x \in \operatorname{Dom} f$ and $d = \nabla f(x)$, we have $d \in \partial f(x)$, so that $d \in \operatorname{Dom} f^*$ and $x \in \partial f^*(d)$ by $\mathbf{C}$. As $\operatorname{Dom} f^*$ is open and $f^*$ is continuously differentiable at its domain, the relation $x \in \partial f^*(d)$ is the same as $x = \nabla f^*(d)$. Thus, $x \mapsto \nabla f(x)$ is an embedding of $\mathrm{Dom}\, f$ into $\mathrm{Dom}\, f^*$ and its left inverse is the mapping $d \mapsto \nabla f^*(d) : \mathrm{Dom}\, f^* \to \mathbf{R}^n$. Recalling that $f$ is proper lsc convex, so is $f^*$ (Fact III.17.2), and $(f^*)^* = f$ (Proposition III.17.3). Since the rest of our assumptions is symmetric with respect to $f$, $f^*$ as well, the previous reasoning as applied to $f^*$ in the role of $f$ demonstrates that the mapping $d \mapsto \nabla f^*(d)$ is an embedding of $\mathrm{Dom}\, f^*$ into $\mathrm{dom}\, f$ with left inverse $x \mapsto \nabla f(x)$. Taken together, our observations yield $\mathbf{C}'$. □

**D.** $\mathrm{Dom}\, f^* = \mathbf{R}^n$ if and only if $f(x)$ "grows at infinity faster than $\|x\|_2$", that is, if and only if the function $F(s) := \inf_{x:\|x\|_2 \leq s} f(x)/\|x\|_2$ blows up to $\infty$ as $s \to \infty$.

**Proof.** Suppose $F(s) \to \infty$ as $s \to \infty$. Consider any fixed $d \in \mathbf{R}^n$. Then, there exists $\bar{s}$ such that $f(x) \geq \|d\|_2\|x\|_2$ whenever $\|x\|_2 \geq \bar{s}$. Thus, whenever $\|x\|_2 \geq \bar{s}$ we have $x^\top d - f(x) \leq \|d\|_2\|x\|_2 - f(x) \leq 0$. Also, as a convex function $f$ is below bounded on the ball $\|x\|_2 \leq \bar{s}$, (in fact on any bounded set), the function $d^\top x - f(x)$ of $x$ is bounded from above and we conclude $d \in \mathrm{Dom}\, f^*$. As this conclusion holds for any $d$, we conclude that $\mathrm{Dom}\, f^* = \mathbf{R}^n$. Now, to see the reverse direction, suppose that $F(s)$ does not blow up to $\infty$ as $s \to \infty$. Then, we can find a sequence $\{x^i\}$ and $L \in \mathbf{R}$ such that $\|x^i\|_2 \to \infty$ as $i \to \infty$ and $f(x^i) \leq L\|x^i\|_2$ for all $i$. Passing to a subsequence, we can assume that as $i \to \infty$ we have $x^i/\|x^i\|_2 \to \xi$, where $\|\xi\|_2 = 1$. Now, select $d := 2L\xi$, then for all large enough $i$ we have $f(x^i) \leq L\|x^i\|_2 \leq \frac{2}{3}d^\top x^i$. In addition, $d^\top x^i \to \infty$ as $i \to \infty$. Consequently, $d^\top x^i - f(x^i) \geq d^\top x^i - L\|x^i\|_2 \geq \frac{1}{3}d^\top x^i$ for large enough $i$. Hence, $d^\top x^i - f(x^i) \to +\infty$ as $i \to \infty$, that is, $d \notin \mathrm{Dom}\, f^*$. □

The bottom line is that by investigating the Legendre transform of a convex function, we get a lot of "global" information on the function. This being said, detailed investigation of the properties of Legendre transform is beyond our scope. We close this section by listing several simple yet important consequences of Legendre transform.

## 17.3 Young, Hölder, and Moment inequalities

Legendre transform leads to several important inequalities. Recall from the definition of Legendre transformation, we have

$$f(x) + f^*(d) \geq x^\top d \quad \forall x, d \in \mathbf{R}^n.$$

We will see that specific choices of $f$ in this inequality leads to several well-known inequalities.

### 17.3.1 Young's inequality

Young's inequality reads: *Let $p$ and $q$ be positive reals such that $\frac{1}{p} + \frac{1}{q} = 1$. Then,*

$$xd \leq \frac{|x|^p}{p} + \frac{|d|^q}{q}, \quad \forall x, d \in \mathbf{R}.$$

**Proof.** Recall from Example III.17.1 that the Legendre transform of the function $|x|^p/p$ is $|d|^q/q$. □

### 17.3.2 Hölder's inequality

The admittedly simple-looking Young's inequality gives rise to the very nice and useful Hölder's inequality.

> *Let $1 \le p \le \infty$ and let $q = \frac{p}{p-1}$ (where $1/0 = +\infty$), so that $\frac{1}{p} + \frac{1}{q} = 1$. Then,*
>
> $$\sum_{i=1}^{n} |x_i y_i| \le \|x\|_p \|y\|_q, \qquad \forall x, y \in \mathbf{R}^n. \tag{17.2}$$

**Proof.** When $p = \infty$, we have $q = 1$ and $(17.2)$ becomes the obvious relation

$$\sum_{i=1}^{n} |x_i y_i| \le \left(\max_i\{|x_i|\}\right) \left(\sum_{i=1}^{n} |y_i|\right), \qquad \forall x, y \in \mathbf{R}^n.$$

By symmetry, we also see that when $p = 1$ we have $q = \infty$ and $(17.2)$ is evident. Now, let $1 < p < \infty$, so that also $1 < q < \infty$. In this case we should prove that

$$\sum_{i=1}^{n} |x_i y_i| \le \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^q\right)^{1/q}.$$

When $x = 0$ or $y = 0$, this inequality is evident. So, we assume that $x \ne 0$ and $y \ne 0$. As both sides of this inequality are positively homogeneous of degree 1 with respect to $x$, and similarly with respect to $y$, without loss of generality we can assume that $\|x\|_p = \|y\|_q = 1$. Now, under this normalization, we should prove that $\sum_{i=1}^{n} |x_i y_i| \le 1$. Recall that by Young's inequality we get $|x_i y_i| \le \frac{|x_i|^p}{p} + \frac{|y_i|^q}{q}$ for all $i$, and so

$$\sum_{i=1}^{n} |x_i y_i| \le \sum_{i=1}^{n} \left(\frac{|x_i|^p}{p} + \frac{|y_i|^q}{q}\right) = \frac{1}{p}\|x\|_p^p + \frac{1}{q}\|y\|_q^q = \frac{1}{p} + \frac{1}{q} = 1,$$

as desired. □

Note that for $p, q \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$, Hölder's inequality gives us

$$|x^\top y| \le \|x\|_p \|y\|_q. \tag{17.3}$$

When $p = q = 2$, this is the well-known Cauchy inequality. Moreover, for every $p \in [1, \infty]$ the inequality $(17.3)$ is *tight* in the sense that for every $x$ there exists $y$ with $\|y\|_q = 1$ such that

$$x^\top y = \|x\|_p \quad [= \|x\|_p \|y\|_q \text{ as } \|y\|_q = 1].$$

To justify this claim, note that when $x = 0$ we can select any $y$ with $\|y\|_q = 1$. When $x \ne 0$ and $p < \infty$ we, as is immediately seen, can specify $y$ by

$$y_i := \|x\|_p^{1-p} |x_i|^{p-1} \mathrm{sign}(x_i), \qquad \forall i = 1, \ldots, n,$$

where we set $0^{p-1} = 0$ when $p = 1$. Finally, when $p = \infty$, that is $q = 1$, we can specify index $i_*$ of the largest in magnitude entry of $x$ and set

$$y_i = \begin{cases} \mathrm{sign}(x_{i_*}), & \text{if } i = i_*, \\ 0, & \text{if } i \neq i_*. \end{cases}$$

These observations altogether lead us to an extremely important, although simple, fact:

$$\|x\|_p = \max_y \left\{ y^\top x : \ \|y\|_q \leq 1 \right\}, \qquad \text{where } \frac{1}{p} + \frac{1}{q} = 1. \tag{17.4}$$

Based on this, we, in particular, deduce that $\|x\|_p$ is convex (as an upper bound of a family of linear functions). Hence, by its convexity we deduce that for any $x', x''$ we have

$$\|x' + x''\|_p = 2 \left\| \frac{1}{2} x' + \frac{1}{2} x'' \right\|_p \leq 2 \left( \|x'\|_p/2 + \|x''\|_p/2 \right) = \|x'\|_p + \|x''\|_p,$$

which is nothing but the triangle inequality. Thus, $\|x\|_p$ satisfies the triangle inequality; it clearly possesses the other two characteristic properties of a norm, namely positivity and homogeneity, as well. Consequently, $\|\cdot\|_p$ is a norm—a fact that we announced twice and already proved (see Illustration after Proposition III.14.3).

A useful application of Hölder's inequality gives us another well-known inequality as follows.

### 17.3.3 Moment inequality

Moment inequality reads:

*For any $0 \neq a \in \mathbf{R}^n$, the function*

$$f(\pi) := \ln(\|a\|_{1/\pi}) : [0, 1] \to \mathbf{R}$$

*is a convex function of $\pi \in [0, 1]$. That is, by letting $p = 1/\pi$, and so $1 \leq p \leq \infty$, we have the following inequality*

$$\begin{array}{ll} 1 \leq r < s < \infty, \ p \in [r, s] \implies \|a\|_p \leq \|a\|_r^\lambda \|a\|_s^{1-\lambda} & \\ \lambda = \frac{r(s-p)}{p(s-r)} \quad \left[ \iff \lambda \in [0, 1] \ \& \ \lambda \frac{1}{r} + (1-\lambda) \frac{1}{s} = \frac{1}{p} \right] & . \end{array} \tag{17.5}$$

**Proof.** Let $\rho, \sigma \in (0, 1]$ be such that $\rho \neq \sigma$. Consider any $\lambda \in (0, 1)$ and set $\pi := \lambda\rho + (1-\lambda)\sigma$. By defining $\theta := \lambda\rho/\pi$, we get $0 < \theta < 1$ and $1 - \theta = (1-\lambda)\sigma/\pi$. Let $r := 1/\rho$, $s := 1/\sigma$, and $p := 1/\pi$, so we have $p = \theta r + (1-\theta)s$. Then,

$$\sum_{i=1}^n |a_i|^p = \sum_{i=1}^n |a_i|^{\theta r} |a_i|^{(1-\theta)s} \leq \left( \sum_i |a_i|^r \right)^\theta \left( \sum_{i=1}^n |a_i|^s \right)^{1-\theta},$$

where the inequality follows from Hölder's inequality. Raising both sides of the resulting inequality to the power $1/p$ we arrive at

$$\|a\|_p \leq \|a\|_r^{r\theta/p} \, \|a\|_s^{s(1-\theta)/p} = \|a\|_{1/\rho}^{\lambda} \, \|a\|_{1/\sigma}^{1-\lambda}.$$

By recalling that $p = \theta r + (1-\theta)s$ and $\ln(\cdot)$ is a monotone increasing function, we conclude that for any $\lambda \in (0,1)$ the function $f(\pi) = \ln(\|a\|_{1/\pi})$ satisfies the inequality

$$f(\lambda\rho + (1-\lambda)\sigma) \leq \lambda f(\rho) + (1-\lambda) f(\sigma), \quad \forall\, (\rho, \sigma : \rho, \sigma \in (0,1], \rho \neq \sigma)\,.$$

Since this function is continuous on $[0,1]$, it is convex, as claimed. □

Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$. We define its *dual* (a.k.a. *conjugate*) norm as the function

$$\|d\|_* := \sup_x \left\{ d^\top x : \ \|x\| \leq 1 \right\}.$$

As its name implies one can indeed show that this function $\|d\|_*$ is a norm.

**Fact** III.17.4 Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$. Then, its dual norm $\|\cdot\|_*$ indeed is a norm. Moreover, the norm dual to $\|\cdot\|_*$ is the original norm $\|\cdot\|$, and the unit balls of conjugate to each other norms are polars of each other.

For example, when $p \in [1,\infty]$, (17.4) says that the norm conjugate to $\|\cdot\|_p$ is $\|\cdot\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$.

We close this section by examining the Legendre transform of norms.

**Fact** III.17.5 Let $f(x) = \|x\|$ be a norm on $\mathbf{R}^n$. Then,

$$f^*(d) = \begin{cases} 0, & \text{if } \|d\|_* \leq 1, \\ +\infty, & \text{otherwise.} \end{cases}$$

That is, the Legendre transform of $\|\cdot\|$ is the characteristic function of the unit ball of the conjugate norm.

# 18

# ★ Functions of eigenvalues of symmetric matrices

One may think that the calculus of convexity-preserving operations presented in section 14.1 does not look really deep. On the other hand, these "simple" rules are extremely useful and allow us to detect convexity and offer nice characterizations for a particular class of functions of symmetric matrices, which we will examine in this chapter.

Let $X \in \mathbf{S}^n$ be an $n \times n$ symmetric matrix, and let $\lambda(X)$ denote the vector of eigenvalues of $X$ taken with their multiplicities and arranged in the non-ascending order, see section D.1.1.C. In this chapter, we present a really deep result stating that whenever $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is convex and permutation symmetric, then the function $F(X) := f(\lambda(X))$ is a convex function of $X$.

A function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is called *permutation symmetric* if

$$f(x) = f(Px) \text{ for every } n \times n \text{ permutation matrix } P.$$

That is, a function is permutation symmetric if and only if its value remains unchanged when we permute the coordinates in its argument.

We start with the following observation.

**Lemma** III.18.1 Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex permutation symmetric function. Then, for any $x \in \operatorname{Dom} f$ and $n \times n$ doubly stochastic matrix $\Pi$, we have $f(\Pi x) \le f(x)$.

**Proof.** Recall that by Birkhoff's Theorem (Theorem II.9.9), $\Pi$ is a convex combination of a permutation matrices $P^i$, i.e., there exists convex combination weights $\lambda_i \ge 0$ and permutation matrices $P^i$ such that $\Pi = \sum_i \lambda_i P^i$ and $\sum_i \lambda_i = 1$. Then, as $f$ is convex, we have

$$f(\Pi x) = f\left(\sum_i \lambda_i P^i x\right) \le \sum_i \lambda_i f(P^i x) = \sum_i \lambda_i f(x) = f(x),$$

where the inequality follows from convexity of $f$ and the second equality is due to the fact that $f$ is permutation-symmetric. □

Our developments will also rely on the following fundamental fact.

**Lemma** III.18.2 For any $A \in \mathbf{S}^n$, the diagonal $\mathrm{Dg}\{A\}$ of the matrix $A$ is the image of the vector $\lambda(A)$ of the eigenvalues of $A$ under multiplication by a doubly stochastic matrix. That is, there exists an $n \times n$ doubly stochastic matrix $\Pi$ such that

$$\mathrm{Dg}\{A\} = \Pi\lambda(A).$$

**Proof.** Consider the spectral decomposition of $A$, i.e.,

$$A = U^\top \operatorname{Diag}\{\lambda_1(A), \ldots, \lambda_n(A)\}\, U$$

where $U = [u_{ij}]_{i,j=1}^n$ is an orthogonal matrix. Define the matrix $\Pi := [u_{ji}^2]_{i,j=1}^n$. As $U$ is an orthogonal matrix, we have that $\Pi$ is indeed doubly stochastic. Moreover, by denoting the $i$-th basic orth with $e_i$, we get

$$\begin{aligned} A_{ii} = e_i^\top A e_i &= e_i^\top (U^\top \operatorname{Diag}\{\lambda_1(A), \ldots, \lambda_n(A)\}\, U) e_i \\ &= \mathrm{Tr}(e_i^\top (U^\top \operatorname{Diag}\{\lambda_1(A), \ldots, \lambda_n(A)\}\, U) e_i) \\ &= \mathrm{Tr}(U e_i e_i^\top U^\top \operatorname{Diag}\{\lambda_1(A), \ldots, \lambda_n(A)\}) \\ &= \sum_{j=1}^n u_{ji}^2 \lambda_j(A) = [\Pi\lambda(A)]_i. \end{aligned}$$

□

Let us denote with $\mathcal{O}_n$ the set of all $n\times n$ orthogonal matrices. Lemmas III.18.1 and III.18.2 together give us the following very useful relation.

**Proposition** III.18.3 Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex permutation symmetric function. Then, for every $n \times n$ symmetric matrix $A$, we have

$$f(\lambda(A)) \geq f(\mathrm{Dg}\{A\}).$$

Furthermore, we also have

$$f(\lambda(A)) = \max_{V \in \mathcal{O}_n} f(\mathrm{Dg}\{V^\top A V\}). \tag{18.1}$$

In particular, the function $F(A) := f(\lambda(A))$ is a convex function of $A$.

**Proof.** The first claim immediately follows from Lemmas III.18.1 and III.18.2.

To see the second claim, consider any $V \in \mathcal{O}_n$. Note that the matrix $V^\top A V$ has the same eigenvalues as $A$. Then, as $f$ is a convex permutation symmetric function, applying the first claim of this proposition to the matrix $V^\top A V$, we conclude

$$f(\mathrm{Dg}\{V^\top A V\}) \leq f(\lambda(V^\top A V)) = f(\lambda(A)).$$

Taking the supremum over $V \in \mathcal{O}_n$ of both sides of this relation gives us

$$f(\lambda(A)) \geq \sup_{V \in \mathcal{O}_n} f(\mathrm{Dg}\{V^\top A V\}).$$

Note that for a properly chosen $V \in \mathcal{O}_n$ we have $\mathrm{Dg}\{V^\top A V\} = \lambda(A)$. Thus, the preceding inequality holds as equality and the right hand side supremum is achieved.

For the final claim, note that for any $V \in \mathcal{O}_n$, we have the function $F_V(A) := f(\mathrm{Dg}\{V^\top A V\})$ is convex in $A$ (as it is the composition of a convex function $f$ and an affine map $A \mapsto \mathrm{Dg}\{V^\top A V\}$. Then, the final claim of the proposition follows from its

second claim as $F(A)$ is the pointwise supremum of convex functions $\{F_V(A)\}_{V\in\mathcal{O}_n}$.
□

Given a symmetric $n\times n$ matrix $X$, as a corollary of Proposition III.18.3, we arrive at the following immediate relations between eigenvalues and diagonal entries of $X$.

1. For all $p\geq 1$, we have $\sum\limits_{i=1}^n |X_{ii}|^p \leq \sum\limits_{i=1}^n |\lambda_i(X)|^p$.
   [Consider the function $f(x)=\sum_{i=1}^n |x_i|^p$.]
2. Whenever $X$ is positive semidefinite, we have $\prod\limits_{i=1}^n X_{ii}\geq \mathrm{Det}(X)$.
   [Consider $f(x)=-\sum_{i=1}^n \ln(x_i)$ over the domain where $x_i>0$ for all $i$.]
3. Define the function $s_k(x):\mathbf{R}^n\to\mathbf{R}$ to be the sum of $k$ largest entries of $x$ (i.e., the sum of the first $k$ entries in the vector obtained from $x$ by writing down the coordinates of $x$ in the non-ascending order). Then, the function $S_k(X):=s_k(\lambda(X))$ is convex, and
$$s_k(\mathrm{Dg}\{X\})\leq S_k(X). \tag{18.2}$$
   [Recall from Example III.14.2 that $s_k(x)$ is a convex permutation symmetric function.]

**Remark** III.18.4 Let us examine the convexity status of eigenvalues of symmetric matrices. Completely, analogous to our discussion in Remark III.14.6 for the vector case, we have by Proposition III.18.3, the largest eigenvalue $\lambda_1(A)$ of a symmetric $n\times n$ matrix $A$ is a convex function of $A$. Therefore, the smallest eigenvalue $\lambda_n(A)=-\lambda_1(-A)$ is concave in $A$. On the other hand, "intermediate" eigenvalues $\lambda_k(A)$, $1<k<n$, of $A$, are neither convex, nor concave functions of $A$. What is convex in $A$, is the sum $S_k(A)$ of $k\leq n$ largest eigenvalues of $A$, which we have just seen. This clearly implies that the sum of $k$ smallest eigenvalues of $A$ is concave in $A$. The sum of magnitudes (absolute values) of the $k$ largest in magnitude eigenvalues of $A$ is convex in $A$, since the function
$$\overline{s}_k(x)=s_k([|x_1|;\ldots;|x_n|])=\max_{\substack{i_1,\ldots,i_k\in\{1,2,\ldots,n\}:\\ i_1<i_2<\ldots<i_k}}[|x_{i_1}|+\ldots+|x_{i_k}|]$$
is permutation symmetric and convex on $\mathbf{R}^n$.

We say that a set $Q\in\mathbf{R}^n$ is *permutation symmetric* if for all $x\in Q$ and for all $n\times n$ permutation matrices $P$, we have $Px\in Q$ as well. Let us also mention the following useful corollary of Proposition III.18.3.

**Corollary** III.18.5 Let $Q$ be a nonempty closed convex and permutation symmetric set in $\mathbf{R}^n$. Then, the set
$$\mathcal{Q}:=\{A\in\mathbf{S}^n:\ \lambda(A)\in Q\}$$
is closed and convex.

**Proof.** Recall from Fact D.21 that $\lambda(A)$ is continuous in $A$, thus $\mathcal{Q}$ is closed.

To prove that $\mathcal{Q}$ is convex, consider the function

$$f(x) := \min_{y\in Q} \|x-y\|_2.$$

Note that $\|x-y\|_2$ is a convex function of $x$ and $y$ over the convex domain $\{[x;y] \in \mathbf{R}^n \times \mathbf{R}^n : y \in Q\}$, and as convexity is preserved by partial minimization, $f(x)$ is a convex real-valued function. Permutation symmetry of $Q$ and $\|\cdot\|_2$ clearly implies permutation symmetry of $f$. Then, by Proposition III.18.3 the function $F(A) := f(\lambda(A))$ is a convex function of $A$. Note that $z \in Q$ if and only if $f(z) = 0$ which holds if and only if $f(z) \leq 0$. Thus,

$$\mathcal{Q} = \{A \in \mathbf{S}^n : \ f(\lambda(A)) \leq 0\} = \{A \in \mathbf{S}^n : \ F(A) \leq 0\},$$

that is $\mathcal{Q}$ is a sublevel set of a convex function $F(A)$ of $A$. □

Consider a univariate real-valued function $g$ defined on some set $\operatorname{Dom} g \subseteq \mathbf{R}$. In section D.1.5 we have associated with the function $g(\cdot)$ the matrix-valued map $X \mapsto g(X) : \mathbf{S}^n \to \mathbf{S}^n$ as follows: the domain of this map is composed of all matrices $X \in \mathbf{S}^n$ with the spectrum $\sigma(X)$ (subset of $\mathbf{R}$ composed of all eigenvalues of $X$) contained in $\operatorname{Dom} g$, and for such an $X$, we set

$$g(X) := U \operatorname{Diag}\{g(\lambda_1), \ldots, g(\lambda_n)\} U^\top,$$

where $X = U \operatorname{Diag}\{\lambda_1, \ldots, \lambda_n\} U^\top$ is an eigenvalue decomposition of $X$.

The following fact is quite important:

**Fact** III.18.6 Let $g : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ be a convex function. Then, the function

$$F(X) = \begin{cases} \operatorname{Tr}(g(X)), & \text{if } \sigma(X) \subseteq \operatorname{Dom} g, \\ +\infty, & \text{otherwise,} \end{cases} \quad : \mathbf{S}^n \to \mathbf{R} \cup \{+\infty\}$$

is convex.

**Example** III.18.1 Proposition III.18.3 implies convexity of the following functions of $X \in \mathbf{S}^n$:

- $F(X) = -\operatorname{Det}^q(X)$ in the domain $X \succeq 0$, whenever $0 < q \leq \frac{1}{n}$.
  [Consider $f(x_1, \ldots, x_n) = -(x_1 \ldots x_n)^q : \mathbf{R}^n_+ \to \mathbf{R}$.]
- $F(X) = -\ln \operatorname{Det}(X)$ in the domain $X \succ 0$.
  [Consider $f(x_1, \ldots, x_n) = -\sum_{i=1}^n \ln x_i : \operatorname{int}(\mathbf{R}^n_+) \to \mathbf{R}$.]
- $F(X) = \operatorname{Det}^{-q}(X)$ in the domain $X \succ 0$, whenever $q \in \mathbf{R}$ is positive.
  [Consider $f(x_1, \ldots, x_n) = (x_1 \ldots x_n)^{-q} : \operatorname{int}(\mathbf{R}^n_+) \to \mathbf{R}$.]
- $\|X\|_p = \left(\sum\limits_{i=1}^n |\lambda_i(X)|^p\right)^{1/p}$, where $p \geq 1$.
  [Consider $f(x_1, \ldots, x_n) = \|x\|_p$.]
- $\|X_+\|_p = \left(\sum\limits_{i=1}^m (\max[\lambda_i(X), 0])^p\right)^{1/p}$, where $p \geq 1$.
  [Consider $f(x_1, \ldots, x_n) = \|x_+\|_p$, where $x_+ := [\max\{0, x_1\}; \ldots; \max\{0, x_n\}]$.]

**Fact** III.18.7 For $x \in \mathbf{R}^n$, let $x_\uparrow$ and $x_\downarrow$ be the vectors obtained by reordering the entries of $x$ in the non-decreasing and non-increasing orders, respectively. For example, $[1;3;2;1]_\uparrow = [1;1;2;3]$ and $[1;3;2;1]_\downarrow = [3;2;1;1]$.

(i) For every $x, y \in \mathbf{R}^n$ and every $n \times n$ doubly stochastic matrix $P$, we have

$$[x_\uparrow]^\top y_\uparrow \geq x^\top P y \geq [x_\uparrow]^\top y_\downarrow.$$

As a result,

(ii) [Trace inequality] For every $A, B \in \mathbf{S}^n$, we have

$$\lambda^\top(A)\lambda(B) \geq \mathrm{Tr}(AB) \geq \lambda^\top(A)[\lambda(B)]_\uparrow.$$

# 19

# Exercises for Part III

## 19.1 Around convex functions

**Exercise** III.1 Which of the following functions are convex on the indicated domains:

- $f(x) \equiv 1$ on $\mathbf{R}$
- $f(x) = x$ on $\mathbf{R}$
- $f(x) = |x|$ on $\mathbf{R}$
- $f(x) = -|x|$ on $\mathbf{R}$
- $f(x) = -|x|$ on $\mathbf{R}_+ = \{x \in \mathbf{R} : \; x \geq 0\}$
- $f(x) = |2x - 3|$ on $\mathbf{R}$
- $f(x) = |2x^2 - 3|$ on $\mathbf{R}$
- $\exp\{x\}$ on $\mathbf{R}$
- $\exp\{x^2\}$ on $\mathbf{R}$
- $\exp\{-x^2\}$ on $\mathbf{R}$
- $\exp\{-x^2\}$ on $\{x \in \mathbf{R} : \; x \geq 100\}$
- $\ln(x)$ on $\{x \in \mathbf{R} : \; x > 0\}$
- $-\ln(x)$ on $\{x \in \mathbf{R} : \; x > 0\}$

**Exercise** III.2 ▲

1. Prove the following fact:
   For every $C_i \in \mathbf{S}^m_+$, $i \leq I$, satisfying $\sum_{i \in I} C_i = I_m$ and for every $\lambda_i \in \mathbf{R}$, we have
   $$\mathrm{Tr}\left(\left(\sum\nolimits_{i\in I}\lambda_i C_i\right)^2\right) \leq \mathrm{Tr}\left(\sum\nolimits_{i\in I}\lambda_i^2 C_i\right).$$
2. Recall from Example III.14.3 in section 14.2 that for $a_i \geq 0$, $\sum_i a_i > 0$ the function $\ln(\sum_i a_i \exp(\lambda_i))$ is a convex function of $\lambda$. Prove the following matrix analogy of this fact:

   For every $A_i \in \mathbf{S}^m_+$, $1 \leq i \leq I$ such that $\sum_i A_i \succ 0$, the function
   $$f(\lambda) = \ln \mathrm{Det}\left(\sum\nolimits_i \exp(\lambda_i) A_i\right) : \mathbf{R}^I \to \mathbf{R}$$
   is convex.
3. Let $A_i$, $i \leq I$, be as in item 2. Is it true that the function
   $$g(x) = \ln \mathrm{Det}(\sum\nolimits_i x_i^{-1} A_i) : \{x \in \mathbf{R}^I : x > 0\} \to \mathbf{R}$$
   is convex?
4. Let $B_i$, $i \leq I$, be $m_i \times n$ matrices such that $\sum_i B_i^\top B_i \succ 0$, and let
   $$\Lambda = \{\lambda := (\lambda_1, \ldots, \lambda_I) : \lambda_i \in \mathbf{S}^{m_i}, \lambda_i \succ 0, i \leq I\}.$$
   Prove that the function
   $$h(\lambda) = \ln \mathrm{Det}\left(\sum\nolimits_i B_i^\top \lambda_i^{-1} B_i\right) : \Lambda \to \mathbf{R}$$
   is convex.

5. Let $B_i, i \leq I$, and $\Lambda$ be as in the previous item. Prove that the matrix-valued function
$$F(\lambda) = \left[\sum_i B_i^\top \lambda_i^{-1} B_i\right]^{-1} : \Lambda \to \operatorname{int} \mathbf{S}_+^n$$
is $\succeq$-concave, that is, the $\succeq$-hypograph
$$\{(\lambda, Y) : \lambda \in \Lambda, Y \preceq F(\lambda)\}$$
of the function is convex.

**Exercise** III.3 ♦ A function $f$ defined on a convex set $Q$ is called log-convex on $Q$, if it takes real positive values on $Q$ and the function $\ln f$ is convex on $Q$. Prove that

- a log-convex on $Q$ function is convex on $Q$
- the sum (more generally, linear combination with positive coefficients) of two log-convex functions on $Q$ also is log-convex on the set.

**Exercise** III.4 ♦ [Law of Diminishing Marginal Returns] Consider optimization problem
$$\operatorname{Opt}(r) = \max_x \{f(x) : G(x) \leq r \ \& \ x \in X\} \qquad (P[r])$$
where $X \subset \mathbf{R}^n$ is nonempty convex set, $f(\cdot) : X \to \mathbf{R}$ is concave, and $G(x) = [g_1(x); \ldots; g_m(x)] : X \to \mathbf{R}^m$ is vector-function with convex components, and let $\mathcal{R}$ be the set of those $r$ for which $(P[r])$ is feasible. Prove that

1. $\mathcal{R}$ is a convex set with nonempty interior and this set is monotone, meaning that when $r \in \mathcal{R}$ and $r' \geq r$, one has $r' \in \mathcal{R}$.
2. The function $\operatorname{Opt}(r) : \mathcal{R} \to \mathbf{R} \cup \{+\infty\}$ satisfies the concavity inequality:
$$\forall (r, r' \in \mathcal{R}, \lambda \in [0,1]) : \operatorname{Opt}(\lambda r + (1-\lambda) r') \geq \lambda \operatorname{Opt}(r) + (1-\lambda)\operatorname{Opt}(r'). \qquad (!)$$
3. If $\operatorname{Opt}(r)$ is finite at some point $\bar{r} \in \operatorname{int} \mathcal{R}$, then $\operatorname{Opt}(r)$ is real-valued everywhere on $\mathcal{R}$. Moreover, when $X = \mathbf{R}^n$, and $f$ and the components of $G$ are affine, so that $(P[r])$ is an LP program, we can replace in the above claim the inclusion $r \in \operatorname{int} \mathcal{R}$ with the inclusion $r \in \mathcal{R}$: in the LP case, the function $\operatorname{Opt}(r)$ is either identically $+\infty$ everywhere on $\mathcal{R}$, or is real-valued at every point of $\mathcal{R}$.

**Comment.** Think about problem $(P[r])$ as about problem where $r$ is the vector of resources you create, and $f(\cdot)$ is your profit, so that the problem is to maximize your profit given your resources and "technological constraints" $x \in X$. Now let $\bar{r} \in \mathcal{R}$ and $e$ be a nonnegative vector, and let us look what happens when you select your vector of resources on the ray $R = \bar{r} + \mathbf{R}_+ e$, assuming that $\operatorname{Opt}(r)$ on this ray is real-valued. Restricted on this ray, your best profit becomes a function $\phi(t)$ of nonnegative variable $t$:
$$\phi(t) = \operatorname{Opt}(\bar{r} + te).$$
Since $e \geq 0$, this function is nondecreasing, as it should be: the larger $t$, the more resources you have, and the larger is your profit. A not so nice news is that $\phi(t)$ is concave in $t$, meaning that the slope of this function does not increase as $t$ grows. In other words, if it costs you \$1 to pass from resources $\bar{x} + te$ to resources $\bar{x} + (t+1)e$, the return $\phi(t+1) - \phi(t)$ on one extra dollar of your investment goes down (or at least does not go up) as $t$ grows. This is called *The Law of Diminishing Marginal Returns.*

**Exercise** III.5 ▲ [follow-up to Exercise ref III.4] There are $n$ goods $j$ with per-unit prices $c_j > 0$, per-unit utilities $v_j > 0$, and the maximum available amounts $\overline{x}_j$, $j \leq n$. Given budget $R \geq 0$, you want to decide on amounts $x_j$ of goods to be purchased to maximize the total utility of the purchased goods, while respecting the budget and the availability constraints. Pose the problem as and verify that the optimal value $\operatorname{Opt}(R)$ is piecewise linear function of $R$. What are the breakpoints of this function? What are the slopes between breakpoints?

**Exercise** III.6 ▲ Let $\beta \in \mathbf{R}^n$ be such that $\beta_1 \geq \beta_2 \geq \ldots \geq \beta_n$. For $x \in \mathbf{R}^n$, let $x_{(k)}$ be the $k$-th largest entry in $x$. Consider the function

$$f(x) = \sum_k \beta_k x_{(k)} = [\beta_1 - \beta_2]s_1(x) + [\beta_2 - \beta_3]s_2(x) + \ldots + [\beta_{n-1} - \beta_n]s_{n-1}(x) + \beta_n s_n(x),$$

where, as always, $s_k(x) = \sum_{i=1}^k x_{(i)}$. As we know from Exercise I.29, the functions $s_k(x)$, $k < n$, are polyhedrally representable:

$$t \geq s_k(x) \iff \exists z \geq 0, s : x_i \leq z_i + s, i \leq n, \sum_i z_i + ks \leq t,$$

and $s_n(x)$ is just linear:

$$s_n(x) = \sum_i x_i$$

As a result, $f$ admits the polyhedral representation

$$t \geq f(x) \iff \exists Z = [z_{ik}] \in \mathbf{R}^{n \times n-1}, s_k, t_k, k < n : \left\{ \begin{array}{l} \forall (i \leq n, k < n) : z_{ik} \geq 0, x_i \leq z_{ik} + s_k, \\ \forall k < n : t_k \geq \sum_i z_{ik} + ks_k \\ t \geq \sum_{k=1}^{n-1} [\beta_k - \beta_{k+1}] t_k + \beta_n \sum_{i=1}^n x_i \end{array} \right.$$

This polyhedral representation has $2n^2 - n$ linear inequalities and $n^2 + n - 2$ extra variables. Now goes the exercise:

1. Find an alternative polyhedral representation of $f$ with $n^2 + 1$ linear inequalities and $2n$ extra variables.
2. [computational study] Generate at random orthogonal $n \times n$ matrix $U$ and vector $\beta$ with nonincreasing entries and solve numerically the problem

$$\min_x \left\{ f(x) := \sum_k \beta_k x_{(k)} : \|Ux\|_\infty \leq 1 \right\}$$

utilising the above polyhedral representations of $f$. For $n = 8, 16, 32, \ldots, 1024$, compare the running times corresponding to the 2 representations in question.

**Exercise** III.7 ⧫ Let $a \in \mathbf{R}^n$ be a nonzero vector, and let $f(\rho) = \ln(\|a\|_{1/\rho})$, $\rho \in [0, 1]$. Moment inequality, see section 17.3.3, states that $f$ is convex. Prove that the function is also nonincreasing and Lipschitz continuous, with Lipschitz constant $\ln n$, or, which is the same, that

$$1 \leq p \leq p' \leq \infty \implies \|a\|_p \geq \|a\|_{p'} \geq n^{\frac{1}{p'} - \frac{1}{p}} \|a\|_p.$$

**Exercise** III.8 ▲ This Exercise demonstrates power of Symmetry Principle. Consider the situation as follows: you are given noisy observations

$$\omega = Ax + \xi, \; A = \mathrm{Diag}\{\alpha_i, i \leq n\}$$

of unknown signal $x$ known to belong to the unit ball $\mathbf{B} = \{x \in \mathbf{R}^n : \|x\|_2 \leq 1\}$; here $\alpha_i > 0$ are given, and $\xi$ is the standard (zero mean, unit covariance) Gaussian observation noise. Your goal is to recover from this observation the vector $y = Bx$, $B = \mathrm{Diag}\{\beta_i, i \leq n\}$ being given. You intend to recover $y$ by *linear estimate*

$$\widehat{y}_H(\omega) = H\omega,$$

where $H$ is an $n \times n$ matrix you are allowed to choose. For example, selecting $H = BA^{-1} = \mathrm{Diag}\{\beta_i \alpha_i^{-1}\}$, you get an *unbiased* estimate:

$$\mathbf{E}\{\widehat{y}_H(Ax + \xi) - y\} = 0.$$

Let us quantify the quality of a candidate linear estimate $\widehat{y}_H$
— at a particular signal $x \in \mathbf{B}$ - by the quantity

$$\mathrm{Err}_x(H) = \sqrt{\mathbf{E}\{\|\widehat{y}_H(Ax+\xi) - Bx\|_2^2\}},$$

so that $\mathrm{Err}_x^2(H)$ is the expected squared $\|\cdot\|_2$-distance between the estimate and the estimated quantity,
— on the entire set $\mathbf{B}$ of possible signals – by *risk* $\mathrm{Risk}[H] = \max_{x\in\mathbf{B}} \mathrm{Err}_x(H)$.

1. Find closed form expressions for $\mathrm{Err}_x(H)$ and $\mathrm{Risk}(H)$.
2. Formulate the problem of finding the linear estimate with minimal risk as the problem of minimizing a convex function and prove that the problem is solvable, and admits an optimal solution $H^*$ which is diagonal: $H^* = \mathrm{Diag}\{\eta_i, i \le n\}$.
3. Reduce the problem yielding by item 2 to the problem of minimizing easy-to-compute convex univariate function. Consider the case when $\beta_i = i^{-1}$ and $\alpha_i = [\sigma i^2]^{-1}$, $1 \le i \le n$, set $n = 10000$ and fill the following table:

| $\sigma$ | 1.0 | 0.1 | 0.01 | 0.001 | 0.0001 | 0.00001 | 0.000001 |
|---|---|---|---|---|---|---|---|
| $\mathrm{Risk}[H^*]$ | | | | | | | |
| $\mathrm{Risk}[BA^{-1}]$ | | | | | | | |

where $H^*$ is the minimum risk linear estimate as yielded by the solution to univariate problem you end up with, and $\mathrm{Risk}[BA^{-1}]$ is the risk of unbiased linear estimate.

You should see from your numerical results that minimal risk of linear estimation is much smaller than the risk of the unbiased linear estimate. Explain on qualitative level why allowing for bias reduces the risk.

**Exercise** III.9 ♦ [1] Given the sets of $d$-dimensional tentative nodes ($d = 2$ or $d = 3$) and of tentative bars of a TTD problem satisfying assumption $\mathfrak{R}$, let $\mathcal{V} = \mathbf{R}^M$ be the space of virtual displacements of the nodes, $N$ be the number of tentative bars, and $W > 0$ be the allowed total bar volume, see Exercise I.16. Let, next, $\mathcal{C}(t,f) : \mathbf{R}_+^N \times \mathcal{V} \to \mathbf{R} \cup \{+\infty\}$ be the compliance of truss $t \ge 0$ w.r.t. load $f$ (we identify trusses with the corresponding vectors $t$ of bar volumes). Prove that

1. $\mathcal{C}(t,f)$ is a convex lsc function, positively homogeneous of homogeneity degree 1, of $[t;f]$ with $\mathbf{R}_{++}^N \times \mathcal{V} \subset \mathrm{Dom}\,\mathcal{C}$, where $\mathbf{R}_{++}^N = \mathrm{int}\,\mathbf{R}_+^N = \{t \in \mathbf{R}^n : t > 0\}$. This function is positively homogeneous, with degree -1, in $t$, when $f$ is fixed, and positively homogeneous, of degree 2, in $f$ when $t$ is fixed. Besides this, $\mathcal{C}(t,f)$ is nonincreasing in $t \ge 0$: if $0 \le t' \le t$, then $\mathcal{C}(t,f) \le \mathcal{C}(t',f)$ for every $f$.
2. The function $\mathrm{Opt}(W,f) = \inf_t \left\{\mathcal{C}(t,f) : t \ge 0, \sum_i t_i = W\right\}$ – the optimal value in the TTD problem (5.2) – with $W$ restricted to reside in $\mathbf{R}_{++} = \{W > 0\}$ is convex continuous function with the domain $\mathbf{R}_{++} \times \mathcal{V}$. This function is positively homogeneous, of degree -1, in $W > 0$ and homogeneous, of homogeneity degree 2, in $f$:
$$\forall(\lambda > 0, \mu \ge 0) : \mathrm{Opt}(\lambda W, \mu f) = \lambda^{-1}\mu^2 \mathrm{Opt}(W,f),\ \forall(W,f) \in \mathbf{R}_{++} \times \mathcal{V}.$$
Moreover, the infimum in $\inf_t \left\{\mathcal{C}(t,f) : t \ge 0, \sum_i t_i = W\right\}$ is achieved whenever $W > 0$.
3. When on certain bridge there is just one car, of unit weight, the compliance of the bridge does not exceed 1, whatever be the position of the car. How large could the compliance of the bridge when there are 100 cars of total weight 70 on it?

To formulate the next two tasks, let us associate with a free node $p$ the set $\mathcal{F}^p$ of all single-force loads stemming from forces $g$ of magnitude $\|g\|_2$ not exceeding 1 and acting at node $p$. For

[1] Preceding exercises in the TTD series are I.16, I.18.

a set $S$ of free nodes, $\mathcal{F}^S$ is the set of all loads with nonzero forces acting solely at the nodes from $S$ and with the sum of $\|\cdot\|_2$-magnitudes of the forces not exceeding 1, so that

$$\mathcal{F}^S = \mathrm{Conv}(\cup_{p\in S}\mathcal{F}^p)$$

(why?)

4. Let $S = \{p_1, ..., p_K\}$ be a $K$-element collection of free nodes from the nodal set. Assume that for every node $p$ from $S$ and every load $f \in \mathcal{F}^p$ there exists a truss of a given total weight $W$ such that its compliance w.r.t. $f$ does not exceed 1. Which, if any, of the following statements

   (i) For every load $f \in \mathcal{F}^S$, there exists a truss of total volume $W$ with compliance w.r.t. $f$ not exceeding 1

   (ii) There exists a truss of total volume $W$ with compliance w.r.t. every load from $\mathcal{F}^S$ not exceeding 1

   (iii) For properly selected $\gamma$ depending solely on $d$, there exists a truss of total volume $\gamma KW$ with compliance w.r.t. every load from $\mathcal{F}^S$ not exceeding 1

   is true?

★5. Prove the following statement:

> *In the situation of item 4 above, let $\gamma = 4$ when $d = 2$ and $\gamma = 7$ when $d = 3$. For every $k \leq K$ there exists a truss $\widehat{t}^k$ of total volume $\gamma W$ such that the compliance of $t$ w.r.t. every load from $\mathcal{F}^{p_k}$ does not exceed 1. As a result, there exists truss $\widetilde{t}$ of total volume $\gamma KW$ with compliance w.r.t. every load from $\mathcal{F}^S$ not exceeding 1.*

## 19.2 Around support, characteristic, and Minkowski functions

**Exercise** III.10 ♦ [characteristic and support functions of convex sets] Let $X \subset \mathbf{R}^n$ be a nonempty convex set. *Characteristic* (a.k.a. *indicator*) *function* of $X$ is, by definition, the function

$$\chi_X(x) = \left\{\begin{array}{cl} 0 & , x \in X \\ +\infty & , x \notin X \end{array}\right.$$

As is immediately seen, this function is convex and proper. The Legendre transform of this function is called the *support function* $\phi_X(x)$ of $X$:

$$\phi_X(x) = \sup_u [x^\top u - \chi_X(u)] = \sup_{u\in X} x^\top u.$$

1. Prove that $\chi_X$ is lower semicontinuous (lsc) if and only if $X$ is closed, and that the support functions of $X$ and $\mathrm{cl}\, X$ are the same.

In the remaining part of Exercise, we are interested in properties of support function,s and in view of item 1, it makes sense to assume from now on that $X$, on the top of being nonempty and convex, is also closed.

Prove the following facts:

2. $\phi_X(\cdot)$ is proper lsc convex function which is positively homogeneous of degree 1:

   $$\forall (x \in \mathrm{Dom}\, \phi_x, \lambda \geq 0) : \phi_X(\lambda x) = \lambda \phi_X(x).$$

   In particular, the domain of $\phi_X$ is a cone. Demonstrate by example that this cone not necessarily is closed (look at the support function of the closed convex set $\{[v; w] \in \mathbf{R}^2 : v > 0, w \leq \ln v\}$).

3. Vice versa, every proper convex lsc function $\phi$ which is positively homogeneous of degree 1,
$$(x \in \operatorname{Dom} f, \lambda \geq 0) \Longrightarrow \phi(\lambda x) = \lambda\phi(x)$$
is the support function of a nonempty closed convex set, specifically, its subdifferential $\partial\phi(0)$ taken at the origin. In particular, $\phi_X(\cdot)$ "remembers" $X$: if $X, Y$ are nonempty closed convex sets, then $\phi_X(\cdot) \equiv \phi_Y(\cdot)$ if and only if $X = Y$.
4. Let $X, Y$ be two nonempty closed convex sets. Then $\phi_X(\cdot) \geq \phi_Y(\cdot)$ if and only if $Y \subset X$.
5. $\operatorname{Dom} \phi_X = \mathbf{R}^n$ if and only if $X$ is bounded.
6. Let $X$ be the unit ball of some norm $\|\cdot\|$. Then $\phi_X$ is nothing but the norm $\|\cdot\|_*$ conjugate to $\|\cdot\|$. In particular, when $p \in [1, \infty]$ and $X = \{x \in \mathbf{R}^n : \|x\|_p \leq 1\}$, we have $\phi_X(x) \equiv \|x\|_q$, $\frac{1}{q} + \frac{1}{p} = 1$.
7. Let $x \mapsto Ax + b : \mathbf{R}^n \to \mathbf{R}^m$ be an affine mapping, and let $Y = AX + b = \{Ax + b : x \in X\}$. Then
$$\phi_Y(v) = \phi_X(A^\top v) + b^\top v.$$

**Exercise** III.11 ♦ [Minkowski functions of convex sets] The goal of this Exercise is to acquaint the reader with important special family of convex functions – Minkowski functions of convex sets.

Consider a proper *nonnegative* lower semicontinuous function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ which is *positively homogeneous of degree 1*, meaning that
$$x \in \operatorname{Dom} f, t \geq 0 \implies tx \in \operatorname{Dom} f \;\&\; f(tx) = tf(x).$$
Note that from the latter property of $f$ and its properness it follows that $0 \in \operatorname{Dom} f$ and $f(0) = 0$.

We can associate with $f$ its *basic sublevel set*
$$X = \{x \in \mathbf{R}^n : f(x) \leq 1\}.$$
Note that $X$ "remembers" $f$, specifically
$$\forall t > 0 : f(x) \leq t \iff f(t^{-1}x) \leq 1 \iff t^{-1}x \in X,$$
whence also
$$\begin{array}{c}\forall x \in \mathbf{R}^n : f(x) = \inf\left\{t : t > 0, t^{-1}x \in X\right\} \\ [\inf\{t : t > 0, t \in \varnothing\} = +\infty \text{ by definition}]\end{array} \tag{19.1}$$
Note that the basic sublevel set of our $f$ cannot be arbitrary: it is convex and closed (since $f$ is convex lsc) and contains the origin (since $f(0) = 0$).

Now, given a closed convex set $X \subset \mathbf{R}^n$ containing the origin, we can associate with it a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ by construction from (19.1), specifically, as
$$f(x) = \inf\left\{t : t > 0, t^{-1}x \in X\right\} \tag{19.2}$$
This function is called *the Minkowski function* (M.f.) of $X$.

Here goes your first task:

1. Prove that when $X \subset \mathbf{R}^n$ is convex, closed, bounded, and contains the origin, function $f$ given by (19.2) is proper, nonnegative, convex lsc function positively homogeneous of degree 1, and $X$ is the basic sublevel set of $f$. Moreover, $f$ is nothing but the support function $\phi_{X_*}$ of the polar $X_*$ of $X$.

Your next tasks are as follows:

2. What are the Minkowski functions of
   - the singleton $\{0\}$ ?
   - a linear subspace ?
   - a closed cone $\mathbf{K}$ ?

- the unit ball of a norm $\|\cdot\|$ ?

3. Prove that the Minkowski functions $f_X$, $f_Y$ of closed convex and containing the origin sets $X, Y$ are linked by the relation $f_X \geq f_Y$ if and only if $X \subset Y$
4. When the Minkowski function of a set $X$ (convex, closed, bounded, and containing the origin) does not take value $+\infty$?
5. What is the set of zeros of the Minkowski function of a set $X$ (convex, closed, bounded, and containing the origin)?
6. What is the M.f. of the intersection $\cap_{k\leq K} X_k$ of closed convex sets containing the origin?

**Exercise** III.12 ♦

1. Recall that the closed conic transform
$$\overline{\text{ConeT}}(X) = \text{cl}\left\{[x;t] \in \mathbf{R}^n \times \mathbf{R} :\ t > 0,\ x/t \in X\right\},$$
of a nonempty convex set $X \subset \mathbf{R}^n$ (see section 1.5) is a closed cone such that
$$\text{cl}(X) = \{x : [x;1] \in \overline{\text{ConeT}}(X).$$
What is the cone dual to $\overline{\text{ConeT}}(X)$ ?
2. Let $X \subset \mathbf{R}^n$ be a nonempty closed convex set and $X^+ = \overline{\text{ConeT}}(X)$. Prove that
$$X_t^+ := \{x : [x;t] \in X^+\} = \begin{cases} tX, & t > 0 \quad (a) \\ \text{Rec}(X), & t = 0 \quad (b) \\ \varnothing, & t < 0 \quad (c) \end{cases}$$
3. Let $X_1, \ldots, X_K$ be closed convex sets in $\mathbf{R}^n$ with nonempty intersection $X$. Prove that
$$\overline{\text{ConeT}}(X) = \cap_k \overline{\text{ConeT}}(X_k).$$
4. Let $X = \cap_{k\leq K} X_k$, where $X_1, \ldots, X_K$ are closed convex sets in $\mathbf{R}^n$ such that $X_K \cap \text{int}\, X_1 \cap \text{int}\, X_2 \ldots \cap \text{int}\, X_{K-1} \neq \varnothing$. Prove that $\phi_X(y) \leq a$ if and only if there exist $y_k$, $k \leq K$, such that
$$y = \sum_k y_k \ \&\ \sum_k \phi_{X_k}(y_k) \leq a. \qquad (*)$$
In words: *In the situation in question, the supremum of a linear form on $\cap_k X_k$ does not exceed some $a$ if and only if the form can be decomposed into the sum of $K$ forms with the sum of their suprema over the respective sets $X_k$ not exceeding $a$.*
5. Prove the following polyhedral version of the claim in item 4:
*Let $X_k = \{x \in \mathbf{R}^n : A_k x \leq b_k\}$, $k \leq K$, be polyhedral sets with nonempty intersection $X$. A linear form does not exceed some $a \in \mathbf{R}$ everywhere on $X$ if and only if the form can be decomposed into the sum of $K$ linear forms with the sum of their maxima on the respective sets $X_k$ not exceeding $a$.*

**Exercise** III.13 ▲ Let $X \subset \mathbf{R}^n$ be a nonempty polyhedral set given by polyhedral representation
$$X = \{x : \exists u : Ax + Bu \leq r\}.$$
Build polyhedral representation of the epigraph of the support function of $X$. For non-polyhedral extension, see Exercise IV.36.

**Exercise** III.14 ▲ Compute in closed analytic form the support functions of the following sets:

1. The ellipsoid $\{x \in \mathbf{R}^n : (x-c)^\top C(x-c) \leq 1\}$ with $C \succ 0$
2. The probabilistic simplex $\{x \in \mathbf{R}^n_+ : \sum_i x_i = 1\}$
3. The nonnegative part of the unit $\|\cdot\|_p$-ball: $X = \{x \in \mathbf{R}^n_+ : \|x\|_p \leq 1\}$, $p \in [1,\infty]$
4. The positive semidefinite part of the unit $\|\cdot\|_{p,\text{Sh}}$ norm: $X = \{x \in \mathbf{S}^n_+ : \|x\|_{p,\text{Sh}} \leq 1\}$
5. The paraboloid $\{x \in \mathbf{R}^{n+1} : x_{n+1} \geq \frac{1}{2}\sum_{i=1}^n x_i^2\}$ $(n \geq 1)$

## 19.3 Around subdifferentials

**Exercise** III.15 ⧫ Let $f$ be a convex function and $\bar{x} \in \operatorname{Dom} f \subset \mathbf{R}^n$. Prove that the property of $g \in \mathbf{R}^n$ to be a subgradient of $f$ at $\bar{x}$ is local: the inequality

$$f(x) \geq f(\bar{x}) + g^\top (x - \bar{x}) \tag{$*$}$$

hods true for all $x \in \mathbf{R}^n$ iff it holds true for all $x$ in a neighborhood of $\bar{x}$.

**Exercise** III.16 ⧫ [subdifferentials of norms] Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$, and $\|\cdot\|_*$ be its conjugate (see Fact III.17.4). Prove that

1. The subdifferential of $\|\cdot\|$ taken at the origin is the unit ball $B_*$ of $\|\cdot\|_*$, or, which is the same, the polar
$$\{u : u^\top x \leq 1 \, \forall (x : \|u\| \leq 1)\}$$
of the unit ball $B$ of the norm $\|\cdot\|$.
2. When $x \neq 0$, the subdifferential of $\|\cdot\|$ taken at $x$ is the set $\{u \in B_* : u^\top x = \|x\|\}$. In particular, the subdifferential of $\|\cdot\|$ remains intact when replacing $x$ with $tx$, $t > 0$, and is reflected with respect to the origin when $x$ is replaced with $tx$, $t < 0$.

**Exercise** III.17 ⧫ [Shatten norms] Let $p \in [1, \infty]$. The space $\mathbf{S}^n$ of symmetric $n \times n$ matrices can be equipped with *Shatten p-norms* – matrix analogies of the standard $\|\cdot\|_p$-norms on $\mathbf{R}^n$. Specifically, Shatten $p$-norm $\|\cdot\|_{p,\mathrm{Sh}}$ of symmetric matrix $X$ is defined as

$$\|X\|_{p,\mathrm{Sh}} = \|\lambda(X)\|_p,$$

where $\lambda(X)$, as always, is the vector of eigenvalues of $X$.

1. Prove that Shatten norms indeed are norms, and the norm conjugate to $\|\cdot\|_{p,\mathrm{Sh}}$ is $\|\cdot\|_{q,\mathrm{Sh}}$, $\frac{1}{p} + \frac{1}{q} = 1$:
$$\|X\|_{q,\mathrm{Sh}} = \max_Y \{\operatorname{Tr}(XY) : \|Y\|_{p,\mathrm{Sh}} \leq 1\} \tag{19.3}$$
2. Verify that $\|\cdot\|_{2,\mathrm{Sh}}$ is nothing but the Frobenius norm of $X$, and $\|\mathbb{X}\|_{\infty,\mathrm{Sh}}$ is the same as the spectral norm of $X$.

**Exercise** III.18 ⧫ [chain rule for subdifferentials] Let $Y \in \mathbf{R}^m$ and $X \in \mathbf{R}^n$ be nonempty convex sets, $\overline{y} \in Y$, $\overline{x} \in X$, $f(\cdot) : Y \to \mathbf{R}$ be a convex function, and $A(\cdot) : X \to Y$ with $A(\overline{x}) = \overline{y}$. Let, further, $\mathbf{K}$ be a closed cone in $\mathbf{R}^n$. Function $f$ is called $\mathbf{K}$-monotone on $Y$, if for $y, y' \in Y$ such that $y' - y \in \mathbf{K}$ it holds $f(y') \geq f(y)$, and $A$ is called $\mathbf{K}$-convex on $X$ if for all $x, x' \in X$ and $\lambda \in [0, 1]$ it holds $\lambda A(X) + (1 - \lambda) A(x') - A(\lambda x + (1 - \lambda) x') \in \mathbf{K}$. [2].

Prove that

1. $A$ is $\mathbf{K}$-convex on $X$ if and only if for every $\phi \in \mathbf{K}_*$ the real-valued function $\phi^\top A(x)$ is convex on $X$.
2. Let $A$ be $\mathbf{K}$-convex on $X$ and differentiable at $\overline{x}$. Prove that
$$\forall x \in X : A(x) - [A(\overline{x}) + A'(\overline{x})[x - \overline{x}]] \in \mathbf{K}. \tag{$*$}$$
3. Let $f$ be $\mathbf{K}$-monotone on $Y$ and $A$ be $\mathbf{K}$-convex on $X$. Prove that the real valued on $X$ function $f{\circ}A\,(x) = f(A(x))$ is convex.
4. Let $f$ be $\mathbf{K}$-monotone on $Y$. Prove that $\partial f(\overline{y}) \subset \mathbf{K}_*$ provided $\overline{y} \in \operatorname{int} Y$.
5. [chain rule] Let $\overline{y} \in \operatorname{int} Y$, $\overline{x} \in \operatorname{int} X$, let $f$ be $\mathbf{K}$-monotone on $Y$, $A$ be $\mathbf{K}$-convex on $X$ and differentiable at $\overline{x}$. Prove that
$$\partial f{\circ}A\,(\overline{x}) = [A'(\overline{x})]^\top \partial f(\overline{y}) = \{[A'(\overline{x})]^\top g : g \in \partial f(\overline{y})\} \tag{!}$$

[2] We shall study cone-monotonicity and cone-convexity in more details in Part IV.

**Exercise** III.19 ◆ Recall that the sum $S_k(X)$ of $k \leq n$ largest eigenvalues of $X \in \mathbf{S}^n$ is a convex function of $X$, see Remark III.18.4. Point out a subgradient of $S_k(\cdot)$ at a point $\overline{X} \in \mathbf{S}^n$. As a special case, find a subgradient of the maximal eigenvalue $\lambda_{\max}(X)$ of $X \in \mathbf{S}^n$ treated as a function of $X$.

## 19.4 Around Legendre transform

**Exercise** III.20 ▲ Compute Legendre transforms of the following univariate functions:

1. $f(x) = -\ln x$, $\mathrm{Dom}\, f = (0, \infty)$
2. $f(x) = \mathrm{e}^x$, $\mathrm{Dom}\, f = \mathbf{R}$.
3. $f(x) = x \ln x$, $\mathrm{Dom}\, f = [0, \infty)$ ($0 \ln 0 = 0$ by definition).
4. $f(x) = x^p/p$, $\mathrm{Dom}\, f = [0, \infty)$; here $p > 1$.

**Exercise** III.21 ▲ Compute Legendre transforms of the following functions:

- [log-barrier for nonnegative orthant $\mathbf{R}^n_+$] $f(x) = -\sum_{i=1}^n \ln x_i : \mathrm{int}\, \mathbf{R}^n_+ \to \mathbf{R}$
- [log-det barrier for semidefinite cone $\mathbf{S}^n_+$] $f(x) = -\ln \mathrm{Det}(x) : \mathrm{int}\, \mathbf{S}^n_+ \to \mathbf{R}$ (start with proving convexity of $f$).

**Exercise** III.22 ◆ [computing Legendre transform of the log-barrier $-\ln(x_n^2 - x_1^2 - ... - x_{n-1}^2)$ for Lorentz cone] Consider the optimization problem

$$\max_{x,t} \left\{ \xi^\top x + \tau t + \ln(t^2 - x^\top x) : (t, x) \in X = \{t > \sqrt{x^\top x}\} \right\}$$

where $\xi \in \mathbf{R}^n$, $\tau \in \mathbf{R}$ are parameters. Is the problem convex[3)]? What is the domain in the space of parameters where the problem is solvable? What is the optimal value? Is it convex in the parameters?

**Exercise** III.23 ◆ Consider the optimization problem

$$\max_{x,y} \left\{ f(x, y) = ax + by + \ln(\ln y - x) + \ln(y) : (x, y) \in X = \{y > \exp\{x\}\} \right\},$$

where $a, b \in \mathbf{R}$ are parameters. Is the problem convex? What is the domain in space of parameters where the problem is solvable? What is the optimal value? Is it convex in the parameters?

**Exercise** III.24 ▲ Compute Legendre transforms of the following functions:

- ["geometric mean"] $f(x) = -\prod_{i \leq n} x_i^{\pi_i} : \mathbf{R}^n_+ \to \mathbf{R}$, where $\pi_i > 0$ sum up to 1 and $n > 1$.
- ["inverse geometric mean"] $f(x) = \prod_{i \leq n} x_i^{-\pi_i} : \mathrm{int}\, \mathbf{R}^n_+ \to \mathbf{R}$, where $\pi_i > 0$.

## 19.5 Miscellaneous exercises

**Exercise** III.25 ◆ [multi-factor Hölder inequality] Given positive reals $q_1, ..., q_n$ and $p \in [1, \infty)$, we define the weighted $p$-norm of a vector $x \in \mathbf{R}^n$ as

$$|x|_p = \left( \sum_{j=1}^n q_j |x_j|^p \right)^{1/p}$$

This clearly is a norm which becomes the standard norm $\|\cdot\|_p$ when $q_j = 1$, $j \leq n$. Same as $\|x\|_p$, the quantity $|x|_p$ has limit, namely, $\|x\|_\infty$, as $p \to \infty$, and we define $|\cdot|_\infty$ as this limit.

[3] A *maximization* problem with objective $f(\cdot)$ and certain constraints and domain is called convex if the equivalent minimization problem with the objective $(-f)$ and the original constraints and domain is convex.

Now let $p_i$, $i \leq k$, be positive reals such that

$$\sum_{i=1}^{k} \frac{1}{p_i} = 1.$$

1. Prove that for nonnegative reals $a_1, ..., a_k$ one has

$$a_1 a_2 ... a_k \leq \frac{a_1^{p_1}}{p_1} + \ldots + \frac{a_k^{p_k}}{p_k}$$

or, equivalently (set $b_i = a_i^{p_i}$)

$$\forall b \geq 0 : b_1^{1/p_1} b_2^{1/p_2} ... b_k^{1/p_k} \leq \frac{b_1}{p_1} + \frac{b_2}{p_2} + ... + \frac{b_k}{p_k}.$$

Note: the special case $p_i = k$, $i \leq k$, of this inequality is the inequality between the geometric and the arithmetic means.

2. Let $x^1, ..., x^k \in \mathbf{R}^n$, and let $x^1 x^2 \ldots x^k$ be the entrywise product of $x^1, ..., x^k$:

$$[x^1 x^2 \ldots x^k]_j = x_j^1 x_j^2 \cdots x_j^k,\ 1 \leq j \leq n.$$

Prove that

$$|x^1 x^2 \ldots x^k|_1 \leq \sum_{i=1}^{k} \frac{|x_i^i|_{p_i}^{p_i}}{p_i}. \qquad (*)$$

3. Prove *multi-factor Hölder inequality*: for vectors $x^i \in \mathbf{R}^n$, $i \leq k$, one has

$$|x^1 x^2 \ldots x^k|_1 \leq |x^1|_{p_1} |x^2|_{p_2} \cdots |x^k|_{p_k} \qquad (\#)$$

Note: (#) was stated for positive *reals* $p_1, ..., p_k$ with $\sum_i 1/p_i = 1$. It is immediately seen that (#) remains true when $p_i = \infty$ for some $i$ (and, of course, $1/p_i$ is set to 0 for these $i$).

Note: (#) is the general form of Hölder inequality which in the main text was proved for $k = 2$ and $|\cdot|_{p_i} = \|\cdot\|_{p_i}$. Needless to say, this inequality extends to the case when $x_i$ are functions $x_i(\omega)$ on a space with measure $q(\cdot)$ , and the finite sums $\sum_{j=1}^{n} q_j f_j$ are replaced with integrals, resulting in

$$\int |\prod_{i=1}^{k} x_i(\omega)| q(d\omega) \leq \prod_{i=1}^{k} \left[\int |x_i(\omega)|^{p_i} q(d\omega)\right]^{1/p_i}$$

provided some measurability conditions are satisfied. In this textbook we, however, do not touch infinite-dimensional spaces of functions and related norms.

**Exercise** III.26 ♦ [Muirhead's inequality] For any $u \in \mathbf{R}^n$ and $z \in \mathbf{R}^n_{++} := \{z \in \mathbf{R}^n : z > 0\}$ define

$$f_z(u) = \frac{1}{n!} \sum_{\sigma} z_{\sigma(1)}^{u_1} \cdots z_{\sigma(n)}^{u_n},$$

where the sum is over all permutations $\sigma$ of $\{1, \ldots, n\}$. Show that if $P$ is a doubly stochastic $n \times n$ matrix, then

$$f_z(Pu) \leq f_z(u)\ \forall (u \in \mathbf{R}^n, z \in \mathbf{R}^n_{++}).$$

**Exercise** III.27 ♦ Prove that a convex lsc function $f$ with polyhedral domain is continuous on its domain. Does the conclusion remain true when lifting either one of the assumptions that (a) convex $f$ is lsc, and (b) $\mathrm{Dom}\, f$ is polyhedral?

**Exercise** III.28 ▲ Let $a_1, \dots, a_n > 0$, $\alpha, \beta > 0$. Solve the optimization problem

$$\min_x \left\{ \sum_{i=1}^n \frac{a_i}{x_i^\alpha} : x > 0, \sum_i x_i^\beta \le 1 \right\}$$

**Exercise** III.29 ▲ [computational study] Consider the following situation: there are $K$ "radars" with $k$-th of them capable to locate targets within ellipsoid $E_k = \{x \in \mathbf{R}^n : (x - c_k)^\top C_k (x - c_k) \le 1\}$ ($C_k \succ 0$); the measured position of target is

$$y_k = x + \sigma_k \zeta_k,$$

where $x$ is the actual position of the target, and $\zeta_k$ is the standard (zero mean, unit covariance) Gaussian observation noise; $\zeta_k$ are independent across $k$. Given measurements $y_1, \dots, y_K$ of target's location $x$ known to belong to the "common field of view" $E = \cap_k E_k$ of the radars, which we assume to possess a nonempty interior, we want to estimate a given linear form $e^\top x$ of $x$ by using linear estimate

$$\widehat{x} = \sum_k h_k^\top y_k + h.$$

We are interested in finding the estimate (e.g., the parameters $H_1, \dots, H_K$, $h$) minimizing the risk

$$\text{Risk2} = \max_{x \in E} \sqrt{\mathbf{E}\left\{ \left[ e^\top x - \sum\nolimits_k h_k^\top [x + \sigma_k \zeta_k] - h \right]^2 \right\}}$$

1. Pose the problem as convex optimization program
2. Process the problem numerically and look at the results.
   Recommended setup:

   - $K = 3$, $n = 2$, $[c_1, c_2, c_3] = \begin{bmatrix} 1.000 & -0.500 & -0.500 \\ 0 & 0.866 & -0.866 \end{bmatrix}$,

     $$C_1 = \begin{bmatrix} 0.2500 & 0 \\ 0 & 1.5000 \end{bmatrix},\ C_2 = \begin{bmatrix} 1.1875 & 0.5413 \\ 0.5413 & 0.5625 \end{bmatrix},\ C_3 = \begin{bmatrix} 1.1875 & -0.5413 \\ -0.5413 & 0.5625 \end{bmatrix}$$

   - $\sigma_1 = 0.1, \sigma_2 = 0.2, \sigma_3 = 0.3$
   - $e = [1; 1]/\sqrt{2}$.

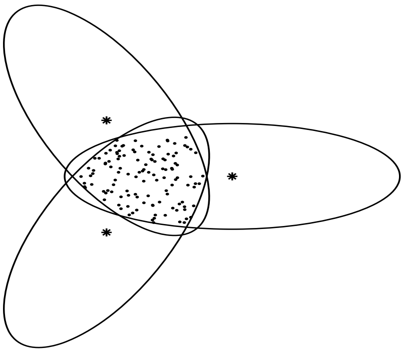


Figure III.5. 3 radars and their common filed of view (dotted)

**Exercise** III.30 ⧫ For any $k \leq m$ and $X \in \mathbf{S}^m$, recall that $S_k(X)$ denotes the sum of $k$ largest eigenvalues of the matrix $X$. Given $X \in \mathbf{S}^m$, define $R[X] := \left\{V^\top XV : V \in \mathcal{O}_m\right\}$ where $\mathcal{O}_m = \{V \in \mathbf{R}^{m\times m} : VV^\top = I_m\}$ is the set of all $m \times m$ orthogonal matrices. Prove that for any two symmetric matrices $X, Y \in \mathbf{S}^m$, we have

$$Y \in \mathrm{Conv}(R[X]) \text{ if and only if } S_k(Y) \leq S_k(X) \text{ for all } k < m \text{ and } \mathrm{Tr}(Y) = \mathrm{Tr}(X).$$

# 20

# Proofs of Facts from Part III

**Fact III.17.2** Given a proper convex function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$, its Legendre transform $f^*$ is a proper convex lower semicontinuous function.

Proof. This fact immediately follows from the definition of the Legendre transform $f^*$: indeed, we lose nothing when replacing $\sup\limits_{x\in\mathbf{R}^n} \{d^\top x - f(x)\}$ with $\sup\limits_{x\in\mathrm{Dom}\, f} \{d^\top x - f(x)\}$, so that the Legendre transform is the supremum of a nonempty (as $f$ is proper) family of affine functions and as such is convex and lower semicontinuous. Since this supremum is finite at least at one point (namely, at every $d$ which is the slope of an affine minorant of $f$ and we know that such a minorant exists), $f^*$ is a proper convex lsc function, as claimed. □

**Fact III.17.4** Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$. Then, its dual norm $\|\cdot\|_*$ indeed is a norm. Moreover, the norm dual to $\|\cdot\|_*$ is the original norm $\|\cdot\|$, and the unit balls of conjugate to each other norms are polars of each other.

Proof. From definition of $\|\cdot\|_*$ it immediately follows that $\|\cdot\|_*$ satisfies all three conditions specifying a norm. To justify that the norm dual to $\|\cdot\|_*$ is $\|\cdot\|$, note that the unit ball of the dual norm is, by the definition of this norm, the polar of the unit ball of $\|\cdot\|$, and the latter set, as the unit ball of any norm, is closed, convex, and contains the origin. As a result, the unit balls of $\|\cdot\|$, $\|\cdot\|_*$ are polars of each other (Proposition II.8.37), and the norm dual to dual is the original one – its unit ball is the polar of the unit ball of $\|\cdot\|_*$. □

**Fact III.17.5** Let $f(x) = \|x\|$ be a norm on $\mathbf{R}^n$. Then,

$$f^*(d) = \begin{cases} 0, & \text{if } \|d\|_* \le 1, \\ +\infty, & \text{otherwise.} \end{cases}$$

That is, the Legendre transform of $\|\cdot\|$ is the characteristic function of the unit ball of the conjugate norm.

Proof. Consider any fixed $d \in \mathbf{R}^n$. By the definition of Legendre transform we have

$$f^*(d) = \sup_{x\in\mathbf{R}^n} \left\{ d^\top x - f(x) \right\} = \sup_{x\in\mathbf{R}^n} \left\{ d^\top x - \|x\| \right\}.$$

Now, consider the function $g_d(x) := d^\top x - \|x\|$ so that $f_*(d) = \sup_{x\in\mathbf{R}^n} g_d(x)$. The function $g_d(x)$ is positively homogeneous, of degree 1, in $x$, so that its supremum over the entire space is either 0 (this happens when the function is nonpositive everywhere), or $+\infty$. By the same homogeneity, the function $g_d(x)$ is nonpositive everywhere if and only if it is nonpositive when $\|x\| = 1$, that is, if and only if $d^\top x \le 1$ whenever $\|x\| = 1$, or, which is the same, when $d^\top x \le 1$ whenever $\|x\| \le 1$. The bottom line is that $f^*(d) = \sup_x g_d(x)$ is either 0, or $+\infty$, with the first option taking place if and only if $d^\top x \le 1$ whenever $\|x\| \le 1$, that is, if and only if $\|d\|_* \le 1$. □

**Fact III.18.6** Let $g : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ be a convex function. Then, the function

$$F(X) = \begin{cases} \mathrm{Tr}(g(X)), & \text{if } \sigma(X) \subseteq \mathrm{Dom}\, g, \\ +\infty, & \text{otherwise,} \end{cases} \quad : \mathbf{S}^n \to \mathbf{R} \cup \{+\infty\}$$

is convex.

Proof. Define $\widehat{g}(x) := \sum_{i=1}^n g(x_i)$, so that $\widehat{g}$ is a permutation symmetric convex (since $g$ is convex) function on $\mathbf{R}^n$. Then, by Proposition III.18.3 the function $\widehat{F}(x) := \widehat{g}(\lambda(X))$ is a convex function of $X \in \mathbf{S}^n$. Now, $\mathrm{Dom}\, \widehat{F} = \{X \in \mathbf{S}^n : \sigma(X) \subseteq \mathrm{Dom}\, g\}$, so $\mathrm{Dom}\, \widehat{F} = \mathrm{Dom}\, F$. Given $X \in \mathrm{Dom}\, \widehat{F}$, consider the eigenvalue decomposition $X = U\, \mathrm{Diag}\{\lambda(X)\} U^\top$ of $X$. By definition of $g(X)$, we have $g(X) = U\, \mathrm{Diag}\{g(\lambda_1(X)), \ldots, g(\lambda_n(X))\} U^\top$ and $F(X) = \mathrm{Tr}(g(X)) = \sum_{i=1}^n g(\lambda_i(X)) = \widehat{g}(\lambda(X)) = \widehat{F}(X)$. We conclude that $F$ is nothing but the convex function $\widehat{F}$. □

**Fact III.18.7.** For $x \in \mathbf{R}^n$, let $x_\uparrow$ and $x_\downarrow$ be the vectors obtained by reordering the entries of $x$ in the non-decreasing and non-increasing orders, respectively. For example, $[1; 3; 2; 1]_\uparrow = [1; 1; 2; 3]$ and $[1; 3; 2; 1]_\downarrow = [3; 2; 1; 1]$.

(i) For every $x, y \in \mathbf{R}^n$ and every $n \times n$ doubly stochastic matrix $P$, we have

$$[x_\uparrow]^\top y_\uparrow \ge x^\top P y \ge [x_\uparrow]^\top y_\downarrow.$$

As a result,

(ii) [Trace inequality] For every $A, B \in \mathbf{S}^n$, we have

$$\lambda^\top(A)\lambda(B) \ge \mathrm{Tr}(AB) \ge \lambda^\top(A)[\lambda(B)]_\uparrow.$$

Proof.

(i) First, we claim that for all $x, y \in \mathbf{R}^n$ we have

$$x_\downarrow^\top y_\downarrow \ge x^\top y \ge x_\downarrow^\top y_\uparrow. \tag{*}$$

Indeed, by continuity, it suffices to verify this relation when all entries of $x$, same as all entries in $y$, are distinct from each other. In such a case, observe that the inequalities to be proved remain intact when we simultaneously reorder, in the same order, entries in $x$ and in $y$, so that we can assume without loss of generality that $x_1 \ge x_2 \ge \ldots \ge x_n$. Taking into account that $ac+bd-[ad+bc] = [a-b][c-d]$, we see that if $i < j$ and $\overline{y}$ is obtained from $y$ by swapping its $i$-th and $j$-th entries, we have $x^\top \overline{y} \le x^\top y$ when $y_i > y_j$ and $x^\top \overline{y} \ge x^\top y$ otherwise. Thus, the minimum (the maximum) of inner products $x^\top z$ over vectors $z$ obtained by reordering entries in $y$ is achieved when $z = y_\uparrow$ (respectively, $z = y_\downarrow$), as claimed.
In the situation of (i), by Birkhoff Theorem, $Py$ is a convex combination of vectors obtained from $y$ by reordering entries, and so the relation in (i) is immediately implied by $(*)$.

(ii) Let $A = U\, \mathrm{Diag}\{\lambda(A)\} U^\top$ be the eigenvalue decomposition of $A$. Then,

$$\mathrm{Tr}(AB) = \mathrm{Tr}(U\, \mathrm{Diag}\{\lambda(A)\} U^\top B) = \mathrm{Tr}(\mathrm{Diag}\{\lambda(A)\}(U^\top BU)) = (\lambda(A))^\top \mu,$$

where $\mu$ is the diagonal of the matrix $U^\top BU$, i.e., $\mu = \mathrm{Dg}\{U^\top BU\}$. Then, by Lemma III.18.2 we deduce that there exists a doubly stochastic matrix $P$ such that

$$\mu = \mathrm{Dg}\{U^\top BU\} = P\lambda(U^\top BU) = P\lambda(B).$$

Thus, $\mathrm{Tr}(AB) = (\lambda(A))^\top \mu = (\lambda(A))^\top P\lambda(B)$ for a doubly stochastic matrix $\Pi$. The desired inequality then follows from applying part (i) to the vectors $x := \lambda(A)$ and $y := \lambda(B)$.

□

# Part IV

# Convex Programming, Lagrange Duality, Saddle Points

# 21

# Convex Programming problems and Convex Theorem on Alternative

## 21.1 Mathematical Programming and Convex Programming problems

For reader's convenience, we start with reproducing the basic optimization terminology presented in section 4.5.1.

A (constrained) Mathematical Programming problem has the following form:

$$(\mathrm{P}) \qquad \min_x \left\{ f(x): \begin{array}{c} x \in X, \\ g(x) \equiv [g_1(x); \dots; g_m(x)] \le 0, \\ h(x) \equiv [h_1(x); \dots; h_k(x)] = 0 \end{array} \right\}, \tag{21.1}$$

where

- [domain] $X \subseteq \mathbf{R}^n$ is called the *domain* of the problem.
- [objective] $f$ is called the *objective* (function) of the problem,
- [constraints] $g_i$, $i = 1, \dots, m$, are called the (functional) *inequality constraints*, and $h_j$, $j = 1, \dots, k$, are called the *equality constraints*[1)].

We always assume that $X \neq \varnothing$ and that the objective and the constraints are well-defined on $X$. Moreover, we typically skip indicating $X$ when $X = \mathbf{R}^n$. Thus, *in the sequel, unless the domain is explicitly present in the formulation, it is the entire* $\mathbf{R}^n$.

We use the following standard terminology related to (21.1)

- [feasible solution] a point $x \in \mathbf{R}^n$ is called a *feasible solution* to (21.1), if $x \in X$, $g_i(x) \le 0$, $i = 1, \dots, m$, and $h_j(x) = 0$, $j = 1, \dots, k$, i.e., if $x$ satisfies all restrictions imposed by the formulation of the problem.
  - [feasible set] the set of all feasible solutions is called the *feasible set* of the problem.
  - [feasible problem] a problem with a nonempty feasible set (i.e., the one which admits feasible solutions) is called *feasible* (or consistent).
  - [active constraint] an inequality constraint $g_i(\cdot) \le 0$ is called *active at a given*

[1] Rigorously speaking, the constraints are not the *functions* $g_i$, $h_j$, but the *relations* $g_i(x) \le 0$, $h_j(x) = 0$. We will use the word "constraints" in both of these senses, and it will always be clear what is meant. For example, we will say that "$x$ satisfies the constraints" to refer to the relations, and we will say that "the constraints are differentiable" to refer to the underlying functions.

*feasible solution* $x$, if this constraint is satisfied at the point as an equality rather than strict inequality, i.e., if

$$g_i(x) = 0.$$

Each equality constraint $h_j(x) = 0$ by definition is active at every feasible solution $x$.

- [optimal value] *the optimal value* of the problem refers to the quantity

$$f^* := \begin{cases} \inf_x \{f(x) : \ x \in X,\ g(x) \le 0,\ h(x) = 0\}, & \text{if the problem is feasible} \\ +\infty, & \text{if the problem is infeasible} \end{cases} .$$

  - [below boundedness] the problem is called *below bounded*, if its optimal value is $> -\infty$, i.e., if the objective is bounded from below on the feasible set.

- [optimal solution] a point $x \in \mathbf{R}^n$ is called an *optimal solution* to (21.1), if $x$ is feasible and $f(x) \le f(x')$ for any other feasible solution $x'$, i.e., if

$$x \in \text{Argmin}\,\{f(x') : \ x' \in X,\ g(x') \le 0,\ h(x') = 0\}.$$

  - [solvable problem] a problem is called *solvable*, if it admits optimal solutions.
  - [optimal set] the set of all optimal solutions to a problem is called its *optimal set.*

The terminology above is for minimization problems; for its "maximization modifications," see section 4.5.1.

### *21.1.1 Convex Programming problem*

A Mathematical Programming problem (P) is called *convex* (or *Convex Programming* problem), if

- $X$ is a *convex* subset of $\mathbf{R}^n$,
- $f, g_1, \ldots, g_m$ are *real-valued convex* functions on $X$, and
- there are no equality constraints at all.

Note that instead of saying that there are no equality constraints, we could say that there are constraints of this type, but only *linear* (affine) ones; this latter case can be immediately reduced to the one without equality constraints by replacing $\mathbf{R}^n$ with the affine subspace given by the (linear) equality constraints.

## 21.2 Convex Theorem on Alternative

The simplest case of a convex problem is, of course, a Linear Programming problem – the one where $X = \mathbf{R}^n$ and the objective and all the constraints are linear. The main descriptive components of LP are LP duality and optimality conditions; our primary goal in this and forthcoming chapters is to extend duality and optimality conditions from Linear to Convex programming.

The origin of our developments is based on the following simple observation:

the fact that a point $x^*$ is an optimal solution can be expressed in terms of feasibility/infeasibility of certain systems of constraints. These systems in our current setup of convex optimization problems are given by

$$x \in X,\ f(x) \le c,\ g_j(x) \le 0,\ j = 1, \dots, m \tag{21.2}$$

and

$$x \in X,\ f(x) < c,\ g_j(x) \le 0,\ j = 1, \dots, m; \tag{21.3}$$

here $c$ is a parameter. Optimality of $x^*$ for the problem means precisely that for appropriately chosen $c$ (this choice, of course, is $c = f(x^*)$) the first of these systems is feasible and $x^*$ is its feasible solution, while the second system is infeasible. Next, in the case of LP, we converted the "negative" part of this simple observation –the claim that (21.3) is infeasible– into a positive statement, using the General Theorem on Alternative (Theorem I.4.3), and this gave us the LP Duality Theorem (Theorem I.4.9).

We will follow the same approach for convex optimization problems. To this end, we need a "convex analogy" to the Theorem on Alternative – something like the latter statement, but for the case when the inequalities in question are given by convex functions rather than the linear ones (and, besides, we now have to handle a "convex inclusion" $x \in X$).

Indeed, it is easy to *guess* the result we need. How did we come to the formulation of the Theorem on Alternative? The main question, basically, boiled down to how to express in an affirmative manner the fact that a system of linear inequalities has no solutions. To this end, we observed that if we can combine, in a linear fashion, the inequalities of the system and get an obviously false inequality like $0 \le -1$, then the system is infeasible. Note that this condition is nothing but a certain affirmative statement with respect to the weights with which we are combining the original inequalities.

Now, the scheme of the above reasoning has nothing tied to linearity (and even convexity) of the inequalities in question. Indeed, consider *an arbitrary* system of constraints of the type (21.3):

$$\begin{array}{rcl} f(x) & < & c \\ g_j(x) & \le & 0, \quad j = 1, \dots, m \\ x & \in & X. \end{array} \tag{I}$$

Here, all we assume is that $X$ is a nonempty subset in $\mathbf{R}^n$ and $f, g_1, \dots, g_m$ are real-valued functions on $X$. Then, it is absolutely evident that

*if there exist nonnegative weights $\lambda_1, \dots, \lambda_m$ such that the inequality*

$$f(x) + \sum\nolimits_{j=1}^{m} \lambda_j g_j(x) < c \tag{21.4}$$

*has no solutions in $X$, then* (I) *also has no solutions.*

Indeed, a solution to (I) is clearly a solution to (21.4) – the latter inequality is nothing but a combination of the inequalities from (I) with the weights 1 (for the first inequality) and $\lambda_j$ (for the remaining ones).

Now, what does it mean that (21.4) has no solutions in the domain $X$? A necessary and sufficient condition for this is that the infimum of the left hand side of (21.4) over the domain $x \in X$ is greater than or equal to $c$. Thus, we arrive at the following evident result.

**Proposition** IV.21.1 [Sufficient condition for infeasibility of (I)] Consider a system (I) with arbitrary data and assume that the system

$$\begin{array}{rcl} \inf\limits_{x \in X} \left[ f(x) + \sum_{j=1}^{m} \lambda_j g_j(x) \right] & \geq & c \\ \lambda_j & \geq & 0,\ j = 1, \ldots, m \end{array} \tag{II}$$

with unknowns $\lambda_1, \ldots, \lambda_m$ has a solution. Then, (I) is infeasible.

Let us stress that Proposition IV.21.1 is completely general; it does not require any assumptions (not even convexity) on the entities involved.

That said, Proposition IV.21.1, unfortunately, is not so helpful: the actual power of the Theorem on Alternative (and the key fact utilized in the proof of the Linear Programming Duality Theorem) is not the *sufficiency* of the condition of Proposition for infeasibility of (I), but the *necessity* of this condition. Justification of necessity of the condition in question has nothing to do with the evident reasoning that established its sufficiency. In the linear case ($X = \mathbf{R}^n$, $f, g_1, \ldots, g_m$ are linear), we established the necessity via the Homogeneous Farkas Lemma. We will next prove the necessity of the condition for the *convex* case. At this step, we already need some additional, although minor, assumptions; and in the general nonconvex case the sufficient condition stated in Proposition IV.21.1 simply is *not* necessary for the infeasibility of (I). This, of course, is very bad-yet-expected news – this is the reason why there are difficult optimization problems that we do not know how to solve efficiently.

The just presented "preface" outlines our action plan. Let us carry out our plan by formally defining the aforementioned "minor regularity assumptions."

**Definition** IV.21.2 [Slater Condition] Let $X \subseteq \mathbf{R}^n$ and let $g_1, \ldots, g_m$ be real-valued functions on $X$. We say that these functions satisfy the *Slater condition* on $X$, if there exists a *strictly feasible solution $x$, that is,* $x \in \operatorname{rint} X$ such that $g_j(x) < 0$, $j = 1, \ldots, m$.

We say that an inequality constrained problem

$$\begin{array}{l} \min_x \left\{ f(x) :\ g_j(x) \leq 0,\ j = 1, \ldots, m,\ x \in X \right\} \\ \text{[where } f, g_1, \ldots, g_m \text{ are real-valued functions on } X\text{]} \end{array} \tag{IC}$$

satisfies the Slater condition (synonym: is strictly feasible), if $g_1, \ldots, g_m$ satisfy this condition on $X$.

In the case where some of the constraints are linear, we rely on a slightly relaxed regularity condition.

**Definition** IV.21.3 [Relaxed Slater Condition] Let $X \subseteq \mathbf{R}^n$, and let $g_1, ..., g_m$ be real-valued functions on $X$. We say that $g_1, \ldots, g_m$ satisfy the *Relaxed Slater condition* on $X$, if there exists $x \in \operatorname{rint} X$ such that $g_j(x) \le 0$ for all $1 \le j \le m$, and $g_j(x) < 0$ for all $j$ *with non-affine* $g_j$.

An inequality constrained problem (IC) is said to satisfy the Relaxed Slater condition (synonym: is essentially strictly feasible), if $g_1, \ldots, g_m$ satisfy this condition on $X$.

A system of equality and inequality constraints

$$\begin{array}{l} g_j(x) \le 0,\ j = 1, \ldots, m, h_i(x) = 0,\ i = 1, \ldots, k,\ x \in X \\ \text{[where } g_1, \ldots, g_m, h_1, \ldots, h_k \text{ are real-valued functions on } X] \end{array} \tag{C}$$

is said to satisfy Relaxed Slater Condition (synonym: is essentially strictly feasible), if all $h_i$ are affine functions, and there exists an *essentially strictly feasible solution,* that is, a feasible solution $x \in \operatorname{rint} X$ where all inequality constraints $g_j(x) \le 0$ with non-affine $g_j$ are satisfied as strict inequalities. An optimization problem of minimizing a (real-valued on $X$) objective $f$ under constraints (C) is called *essentially strictly feasible*, if the system of constraints is so.

Note: (C) is essentially strictly feasible if and only if the equivalent inequality reformulation

$$g_j(x) \le 0, j \le m,\ \pm h_i(x) \le 0, i \le k,\ x \in X$$

of (C) is essentially strictly feasible.

Clearly, the validity of Slater condition implies the validity of the Relaxed Slater condition (why?). We are about to establish the following fundamental fact.

**Theorem** IV.21.4 [Convex Theorem on Alternative] Let $X \subseteq \mathbf{R}^n$ be convex, let $f, g_1, \ldots, g_m$ be real-valued convex functions on $X$, and let $g_1, \ldots, g_m$ satisfy the Relaxed Slater condition on $X$. Then, system (I) is feasible if and only if system (II) is infeasible.

Theorem IV.21.4 is a special case of Theorem IV.21.12 to be formulated and proved in the next section.

## 21.3 ★ Convex Theorem on Alternative – cone-constrained form

We will indeed present and prove a form of Theorem IV.21.4 that will be stronger. To this end, we need a few definitions and concepts related to cones.

**Definition** IV.21.5 [Regular cone] A cone $\mathbf{K} \subset \mathbf{R}^n$ is called a *regular cone* if $\mathbf{K}$ is closed, convex, full dimensional (i.e., possesses a nonempty interior), and is pointed (i.e., $\mathbf{K} \cap (-\mathbf{K}) = \{0\}$).

In our developments, we will frequently examine the dual cones as well. Therefore, we introduce the following elementary fact on the regularity of dual cones.

**Fact** IV.21.6 (i) A cone $\mathbf{K} \subseteq \mathbf{R}^n$ is regular if and only if its dual cone $\mathbf{K}_* = \{y \in \mathbf{R}^n : y^\top x \geq 0, \forall x \in \mathbf{K}\}$ is regular.
(ii) Given regular cones $\mathbf{K}_1, \ldots, \mathbf{K}_m$, their direct product $\mathbf{K}_1 \times \ldots \times \mathbf{K}_m$ is also regular.

There are a number of "magic cones" that are regular and play a crucial role in Convex Optimization. In particular, many convex optimization problems from practice can be posed as optimization problems involving domains expressed using these cones as the basic building blocks.

**Fact** IV.21.7 The following cones (see Examples discussed in section 1.2.4) are regular:

1. *Nonnegative ray* $\mathbf{R}_+$,
2. *Lorentz* (a.k.a., *second-order*, or *ice-cream*) cone, $\mathbf{L}^n = \{x \in \mathbf{R}^n : x_n \geq \sqrt{x_1^2 + \ldots + x_{n-1}^2}\}$ $(\mathbf{L}^1 := \mathbf{R}_+)$,
3. *Positive semidefinite cone*, $\mathbf{S}^n_+ = \{X \in \mathbf{S}^n : a^\top X a \geq 0, \forall a \in \mathbf{R}^n\}$.

Our developments will be based on an important concept that we introduce now.

**Definition** IV.21.8 [Cone-convexity] Let $\mathbf{K} \subset \mathbf{R}^\nu$ be a regular cone. A map $h(\cdot) : \operatorname{Dom} h \to \mathbf{R}^\nu$ is called $\mathbf{K}$-*convex* if $\operatorname{Dom} h$ is a convex set in some $\mathbf{R}^n$ and for every $x, y \in \operatorname{Dom} h$ and $\lambda \in [0,1]$ we have

$$\lambda h(x) + (1-\lambda)h(y) - h(\lambda x + (1-\lambda)y) \in \mathbf{K}. \tag{21.5}$$

Note that in the simplest case of $\mathbf{K} = \mathbf{R}^\nu_+$ (nonnegative orthant is a regular cone!) $\mathbf{K}$-convexity of a map $h$ means exactly that the components of $h$ are convex functions with common domain.

An instructive example of a "genuine cone-convex" function is as follows:

**Lemma** IV.21.9 Let $\mathbf{K} = \mathbf{S}^m_+$, and consider $h : \mathbf{R}^{m \times n} \to \mathbf{S}^m$ defined as $h(x) = xx^\top$. Then, $h$ is $\mathbf{K}$-convex.

**Proof.** Indeed, for any $x, y \in \mathbf{R}^{m \times n}$ and $\lambda \in [0,1]$, we have

$$(\lambda x + (1-\lambda)y)\,(\lambda x + (1-\lambda)y)^\top = \lambda xx^\top + (1-\lambda)yy^\top - \lambda(1-\lambda)(x-y)(x-y)^\top.$$

Therefore,

$$\lambda h(x) + (1-\lambda)h(y) - h(\lambda x + (1-\lambda)y) = \underbrace{\lambda(1-\lambda)}_{\geq 0}\underbrace{(x-y)(x-y)^\top}_{\succeq 0} \succeq 0.$$

□

See chapter 26 for other instructive examples of **K**-convex functions and their "calculus."

Indeed, **K**-convexity can be expressed in terms of the usual convexity due to the following immediate observation.

**Fact** IV.21.10 Let $\mathbf{K}$ be a regular cone in $\mathbf{R}^\nu$, $Z \subseteq \mathbf{R}^n$ be a nonempty convex set and $h : Z \to \mathbf{R}^\nu$ be a mapping with $\operatorname{dom} h = Z$. Then, $h$ is **K**-convex if and only if for all $\mu \in \mathbf{K}_*$ (where $\mathbf{K}_*$ is the cone dual to $\mathbf{K}$) we have that the real valued functions $\mu^\top h(\cdot)$ are convex on $Z$.

Given a regular cone $\mathbf{K} \subset \mathbf{R}^\nu$, we can associate with it **K**-*inequality between vectors of* $\mathbf{R}^\nu$: we say that $a \in \mathbf{R}^\nu$ is **K**-greater than or equal to $b \in \mathbf{R}^\nu$ (notation: $a \geq_\mathbf{K} b$, or, equivalently, $b \leq_\mathbf{K} a$) when $a - b \in \mathbf{K}$:

$$a{\geq_\mathbf{K}}b \quad\Longleftrightarrow\quad b{\leq_\mathbf{K}}a \quad\Longleftrightarrow\quad a - b \in \mathbf{K}.$$

For example, when $\mathbf{K} = \mathbf{R}^\nu_+$ is nonnegative orthant, $\geq_\mathbf{K}$ is the standard coordinate-wise vector inequality $\geq$. That is, $a \geq b$ means that every entry of $a$ is greater than or equal to, in the standard arithmetic sense, the corresponding entry in $b$. **K**-vector inequality possesses all algebraic properties of $\geq$.

**Fact** IV.21.11 Let $\mathbf{K} \subset \mathbf{R}^\nu$ be a regular cone. Then, any **K**-inequality $a \geq_\mathbf{K} b$ satisfies all of the following properties:

1. It is a partial order on $\mathbf{R}^\nu$, i.e., the relation $a \geq_\mathbf{K} b$ is
   - reflexive: $a \geq_\mathbf{K} a$ for all $a$;
   - anti-symmetric: $a \geq_\mathbf{K} b$ and $b \geq_\mathbf{K} a$ if and only if $a = b$;
   - transitive: if $a \geq_\mathbf{K} b$ and $b \geq_\mathbf{K} c$, then $a \geq_\mathbf{K} c$.
2. It is compatible with linear operations, i.e., $\geq_\mathbf{K}$-inequalities can be
   - summed up: if $a{\geq_\mathbf{K}}b$ and $c{\geq_\mathbf{K}}d$, then $a + c{\geq_\mathbf{K}}b + d$;
   - multiplied by nonnegative reals: if $a{\geq_\mathbf{K}}b$ and $\lambda$ is a nonnegative real, then $\lambda a{\geq_\mathbf{K}}\lambda b$.
3. It is compatible with convergence, i.e., one can pass to sidewise limits in $\geq_\mathbf{K}$-inequality:
   - if $a_t{\geq_\mathbf{K}}b_t$, $t = 1, 2, \ldots$, and $a_t \to a$ and $b_t \to b$ as $t \to \infty$, then $a{\geq_\mathbf{K}}b$.
4. It gives rise to strict version $>_\mathbf{K}$ of $\geq_\mathbf{K}$-inequality $a >_\mathbf{K}$b (equivalently: $b <_\mathbf{K} a$) meaning that $a - b \in \operatorname{int} \mathbf{K}$. The strict **K**-inequality possesses the basic properties of the coordinate-wise $>$, specifically,
   - $>_\mathbf{K}$ is stable: if $a >_\mathbf{K} b$ and $a'$, $b'$ are close enough to $a$, $b$ respectively, then $a' >_\mathbf{K} b'$;
   - if $a >_\mathbf{K} b$, $\lambda$ is a positive real, and $c{\geq_\mathbf{K}}d$, then $\lambda a >_\mathbf{K} \lambda b$ and $a + c >_\mathbf{K} b + d$.

In summary, the arithmetics of $\geq_\mathbf{K}$ and $>_\mathbf{K}$ inequalities is completely similar to the one of the usual $\geq$ and $>$ inequalities.

Verification of the claims made in Fact IV.21.11 is immediate and is left to the reader.

In the standard approach to nonlinear convex optimization, the Mathematical Programming problem that is convex has the following form

$$\min_{x\in X}\left\{f(x):\ g(x):=[g_1(x);\dots;g_m(x)]\leq 0,\ [h_1(x);\dots;h_k(x)]=0\right\},$$

where $X$ is a convex set, $f(x), g_i(x) : X \to \mathbf{R}, 1 \leq i \leq m$ are convex functions and $h_j(x), 1 \leq j \leq k$ are affine functions. In this form, the nonlinearity "sits" in the functions $g_i$ and/or non-polyhedrality of the set $X$. Since 1990s, it was realized that, along with this form, it is extremely convenient to consider the *conic form* where the nonlinearity "sits" in the inequality relations $\leq$. That is, the usual coordinate-wise $\leq$ is replaced with $\leq_{\mathbf{K}}$, where $\mathbf{K}$ is a regular cone. The resulting *convex program in cone-constrained form* reads

$$\min_{x\in X}\left\{f(x):\ \overline{g}(x):=Ax-b\leq 0,\ \underbrace{\widehat{g}(x)\leq_{\mathbf{K}} 0}_{\Longleftrightarrow\ \widehat{g}(x)\in-\mathbf{K}}\right\},\tag{21.6}$$

where $X \subset \mathbf{R}^n$ is a convex set, $f : X \to \mathbf{R}$ is a convex function, $\mathbf{K} \subset \mathbf{R}^\nu$ is a regular cone, $A \in \mathbf{R}^{k\times n}$, and $\widehat{g} : X \to \mathbf{R}^\nu$ is $\mathbf{K}$-convex. Note that $\mathbf{K}$-convexity of $\widehat{g}$ in our new notation is simply equivalent to the requirement

$$\widehat{g}(\lambda x+(1-\lambda)y))\leq_{\mathbf{K}}\lambda\widehat{g}(x)+(1-\lambda)\widehat{g}(y),\qquad \forall x,y\in X,\ \forall\lambda\in[0,1].$$

Indeed, when $\mathbf{K}$ is the nonnegative orthant, (21.6) recovers the Mathematical Programming form (21.1) of a convex problem.

It turns out that with "cone-constrained approach" to Convex Programming, we lose nothing when restricting ourselves with $X = \mathbf{R}^n$, linear $f(x)$, and affine $\widehat{g}(x)$; this specific version of (21.6) is called "conic problem" (to be considered in more details later). That said, it makes sense to speak about a "less extreme" form of a convex program, specifically, one presented in (21.6); we call problems of this form "convex problems in cone-constrained form," reserving the words "conic problems" for problems (21.6) with $X = \mathbf{R}^n$, linear $f$ and affine $\widehat{g}$, see section 23.4.

The developments from section 21.2 can be naturally extended to the cone-constrained case as follows. Let $X \subseteq \mathbf{R}^n$ be a nonempty convex set, $\mathbf{K} \subset \mathbf{R}^\nu$ be a regular cone, $f$ be a real-valued function on $X$, and $\widehat{g}(\cdot)$ be a mapping from $X$ into $\mathbf{R}^\nu$. Instead of feasibility/infeasibility of system (I) we can speak about feasibility/infeasibility of system of constraints

$$\begin{array}{rcl} f(x) & < & c\\ \overline{g}(x):=Ax-b & \leq & 0\\ \widehat{g}(x) & \leq_{\mathbf{K}} & 0\quad[\Longleftrightarrow\ \widehat{g}(x)\in-\mathbf{K}]\\ x & \in & X\end{array}\tag{ConI}$$

in variables $x$, i.e., this is the cone-constrained analogy of inequality-constrained system (I). We call system (ConI) *convex*, if, in addition to already assumed convexity of $X$, the function $f$ is convex on $X$, and the map $\widehat{g}$ is $\mathbf{K}$-convex on $X$.

Denoting by $\mathbf{K}_*$ the cone dual to $\mathbf{K}$, a *sufficient* condition for the infeasibility of (ConI) is the solvability of the following system of constraints

$$\begin{array}{rcl}\inf\limits_{x\in X}\left[f(x)+\overline{\lambda}^\top\overline{g}(x)+\widehat{\lambda}^\top\widehat{g}(x)\right] & \geq & c\\ \overline{\lambda} & \geq & 0\\ \widehat{\lambda} & \geq_{\mathbf{K}_*} & 0\ \ [\iff\ \widehat{\lambda}\in\mathbf{K}_*]\end{array}\tag{ConII}$$

in variables $\lambda=(\overline{\lambda},\widehat{\lambda})$.

> Indeed, given a feasible solution $\overline{\lambda},\widehat{\lambda}$ to (ConII) and "aggregating" the constraints in (ConI) with weights $1,\overline{\lambda},\widehat{\lambda}$ (i.e., taking sidewise inner products of constraints in (ConI) based on these aggregation weights and summing up the results), we arrive at the inequality
>
> $$f(x)+\overline{\lambda}^\top\overline{g}(x)+\widehat{\lambda}^\top\widehat{g}(x)<c$$
>
> which due to $\overline{\lambda}\geq 0$ and $\widehat{\lambda}\in\mathbf{K}_*$ is a consequence of (ConI) – it must be satisfied at every feasible solution to (ConI). On the other hand, the aggregated inequality contradicts the first constraint in (ConII), and we conclude that under the circumstances (ConI) is infeasible.

*"Cone-constrained" version of Slater/Relaxed Slater condition* is as follows: we say that the system of constraints

$$\overline{g}(x):=Ax-b\leq 0,\ \ \widehat{g}(x)\leq_{\mathbf{K}}0\tag{S}$$

in variables $x$ satisfies
— Slater condition on $X$, if there exists $\bar{x}\in\operatorname{rint}X$ such that $\overline{g}(\bar{x})<0$ and $\widehat{g}(\bar{x})<_{\mathbf{K}}0$ (i.e., $\widehat{g}(\bar{x})\in-\operatorname{int}\mathbf{K}$),
— Relaxed Slater condition on $X$, if there exists $\bar{x}\in\operatorname{rint}X$ such that $\overline{g}(\bar{x})\leq 0$ and $\widehat{g}(\bar{x})<_{\mathbf{K}}0$.
We say that optimization problem in cone-constrained form, i.e., problem of the form

$$\min_{x\in X}\{f(x):\ \overline{g}(x):=Ax-b\leq 0,\ \widehat{g}(x)\leq_{\mathbf{K}}0\}\tag{ConIC}$$

satisfies Slater/Relaxed Slater condition, if the system of its constraints satisfies this condition on $X$.

Note that in the case of $\mathbf{K}=\mathbf{R}^\nu_+$, (ConI) and (ConII) become, respectively, (I) and (II), and the cone-constrained versions of Slater/Relaxed Slater condition become the usual ones.

We are now ready to state *Convex Theorem on Alternative in cone-constrained form* dealing with convex cone-constrained system (ConI), and obtain Theorem IV.21.4 as a particular case of this result.

**Theorem** IV.21.12 [Convex Theorem on Alternative in cone-constrained form] Let $\mathbf{K} \subset \mathbf{R}^\nu$ be a regular cone, let $X \subseteq \mathbf{R}^n$ be nonempty and convex, let $f$ be real-valued convex function on $X$, $\overline{g}(x) = Ax - b$ be affine, and $\widehat{g}(x) : X \to \mathbf{R}^\nu$ be $\mathbf{K}$-convex. Suppose that system (S) satisfies the cone-constrained Relaxed Slater condition on $X$. Then, the system (ConI) is feasible if and only if the system (ConII) is infeasible.

**Note:** In some cases (ConI) may have no affine (i.e., polyhedral) part $\overline{g}(x) := Ax - b \leq 0$ and/or no "general part" $\widehat{g}(x) \leq_{\mathbf{K}} 0$; absence of one or both of these parts leads to self-evident modifications in (ConII). To unify our forthcoming considerations, it is convenient to assume that both of these parts are present. This assumption is for free: it is immediately seen that in our context, in absence of one or both of $g$-constraints in (ConI) we lose nothing when adding artificial affine part $\overline{g}(x) := 0^\top x - 1 \leq 0$ instead of missing affine part, and/or artificial general part $\widehat{g}(x) := 0^\top x - 1 \leq_{\mathbf{K}} 0$ with $\mathbf{K} = \mathbf{R}_+$ instead of missing general part. Thus, we lose nothing when assuming from the very beginning that both polyhedral and general parts are present.

It is immediately seen that Convex Theorem on Alternative (Theorem IV.21.4) is a special case of Theorem IV.21.12 corresponding to the case when $\mathbf{K}$ is a nonnegative orthant.

In the proof of Theorem IV.21.12, we will use Lemma IV.21.13, a generalization of the Inhomogeneous Farkas Lemma. We will present the proof of this result after the proof of Theorem IV.21.12.

**Lemma** IV.21.13 Let $X \subseteq \mathbf{R}^n$ be a convex set with $0 \in \operatorname{int} X$. Let $g(x) = Ax + a : \mathbf{R}^n \to \mathbf{R}^k$ and $d^\top x + \delta : \mathbf{R}^n \to \mathbf{R}$ be affine functions such that $g(0) \leq 0$ and

$$x \in X,\ g(x) \leq 0 \quad \Longrightarrow \quad d^\top x + \delta \leq 0.$$

Then, there exists a vector $\mu \geq 0$ such that

$$d^\top x + \delta \leq \mu^\top g(x),\ \forall x \in X.$$

**Proof of Theorem IV.21.12.** The first part of the statement – "if (ConII) has a solution, then (ConI) has no solutions" – has been already verified. What we need is to prove the reverse statement. Thus, let us assume that (ConI) has no solutions, and let us prove that then (ConII) has a solution.

$\mathbf{0^0}$. Without loss of generality we may assume that $X$ is full-dimensional: $\operatorname{rint} X = \operatorname{int} X$ (indeed, otherwise we can replace our "universe" $\mathbf{R}^n$ with the affine span of $X$). Besides this, if needed shifting $f$ by a constant, we can assume that $c = 0$. Thus, we are in the case where

$$\left.\begin{array}{rcl} f(x) & < & 0, \\ \overline{g}(x) := Ax - b & \leq & 0, \\ \widehat{g}(x) & \leq_{\mathbf{K}} & 0, \quad [\Longleftrightarrow \widehat{g}(x) \in -\mathbf{K}] \\ x & \in & X; \end{array}\right\} \qquad \text{(ConI)}$$

$$\left.\begin{array}{rcl}\inf\limits_{x\in X}\left[f(x)+\overline{\lambda}^\top\overline{g}(x)+\widehat{\lambda}^\top\widehat{g}(x)\right] & \geq & 0,\\ \overline{\lambda} & \geq & 0,\\ \widehat{\lambda} & \geq_{\mathbf{K}_*} & 0\ [\iff \widehat{\lambda}\in\mathbf{K}_*].\end{array}\right\}\qquad\text{(ConII)}$$

Moreover, because the system satisfies the cone-constrained version of Relaxed Slater condition on $X$, there exists some $\bar{x}\in\operatorname{rint}X=\operatorname{int}X$ such that $\overline{g}(\bar{x})\leq 0$ and $\widehat{g}(\bar{x})\in-\operatorname{int}\mathbf{K}$. If needed, by shifting $X$ (that is, passing from variables $x$ to variables $x-\bar{x}$; this clearly does not affect the statement we need to prove) we can assume that $\bar{x}$ is just the origin, so that we have

$$0\in\operatorname{int}X,\quad \overline{g}(0)\leq 0,\quad \widehat{g}(0)<_{\mathbf{K}}0.\tag{21.7}$$

Recall that we are in the situation when (ConI) is infeasible, that is, the optimization problem

$$\operatorname{Opt}(P):=\min_x\left\{f(x):\ x\in X,\ \overline{g}(x)\leq 0,\ \widehat{g}(x)\leq_{\mathbf{K}}0\right\}\tag{$P$}$$

satisfies $\operatorname{Opt}(P)\geq 0$, and our goal is to show that (ConII) is feasible.

**$1^0$.** Define

$$Y:=\{x\in X:\ \overline{g}(x)\leq 0\}=\{x\in X:\ Ax-b\leq 0\},$$

along with

$$\begin{array}{l}S:=\{t=[t_0;t_1]\in\mathbf{R}\times\mathbf{R}^\nu:\ t_0<0,\ t_1\leq_{\mathbf{K}}0\},\\ T:=\{t=[t_0;t_1]\in\mathbf{R}\times\mathbf{R}^\nu:\ \exists x\in Y\text{ s.t. }f(x)\leq t_0,\ \widehat{g}(x)\leq_{\mathbf{K}}t_1\},\end{array}$$

so that both sets $S$ and $T$ are nonempty (as $(P)$ is feasible by (21.7)) and convex (since $X$, $Y$, and $f$ are convex, and $\widehat{g}(x)$ is $\mathbf{K}$-convex on $X$). Moreover, $S$ and $T$ do not intersect since $\operatorname{Opt}(P)\geq 0$. Then, by Separation Theorem (Theorem II.7.3) $S$ and $T$ can be separated by an appropriately chosen linear form $\alpha$. Thus,

$$\sup_{t\in S}\alpha^\top t\leq\inf_{t\in T}\alpha^\top t\tag{21.8}$$

and the linear form $\alpha^\top t$ is non-constant on $S\cup T$, implying that $\alpha\neq 0$. Denote $\alpha=[\alpha_0;\alpha_1]$ with $\alpha_0\in\mathbf{R}$ and $\alpha_1\in\mathbf{R}^\nu$. We claim that $\alpha_0\geq 0$ and $\alpha_1\in\mathbf{K}_*$. Suppose not, then either $\alpha_0<0$ or $\alpha_1\notin\mathbf{K}_*$ (i.e., there exists $\bar{\tau}\in\mathbf{K}$ such that $\alpha_1^\top\bar{\tau}<0$) or both. But, in any of these cases we would have $\sup_{t\in S}\alpha^\top t=+\infty$ (look at what happens with $\alpha^\top t$ on the ray $\{t=[t_0;t_1]:t_0\leq-1,t_1=0\}\subset S$ when $\alpha_0<0$, and on the ray $\{[t_0=-1,t_1=-s\bar{\tau}],s\geq 0\}\subset S$ when $\alpha_1\notin\mathbf{K}_*$ and $\bar{\tau}\in\mathbf{K}$ is such that $\alpha_1^\top\bar{\tau}<0$) and this would contradict (21.8) combined with nonemptiness of $T$. Then, taking into account that $\alpha_0\geq 0$ and $\alpha_1\in\mathbf{K}_*$, we have $\sup_{t\in S}\alpha^\top t=0$, and (21.8) reads

$$0\leq\inf_{t\in T}\alpha^\top t.\tag{21.9}$$

**$2^0$.** We now claim that $\alpha_0>0$. Note that the point $\bar{t}:=[\bar{t}_0;\bar{t}_1]$ with the components $\bar{t}_0:=f(0)$ and $\bar{t}_1:=\widehat{g}(0)$ belongs to $T$ (since $0\in Y$ by (21.7)), thus by (21.9) it holds that $\alpha_0f(0)+\alpha_1^\top\widehat{g}(0)\geq 0$. Assume for contradiction that $\alpha_0=0$. Then, we deduce $\alpha_1^\top\widehat{g}(0)\geq 0$, which due to $-\widehat{g}(0)\in\operatorname{int}\mathbf{K}$ (see (21.7)) and $\alpha_1 \in \mathbf{K}_*$ is possible only when $\alpha_1 = 0$, see Fact II.8.23.3. Thus, we conclude that $\alpha_1 = 0$ on the top of $\alpha_0 = 0$, which is impossible, since $\alpha \neq 0$.

Given that we have proved $\alpha_0 > 0$, in the sequel, we set $\bar{\alpha}_1 := \alpha_1/\alpha_0$, and

$$h(x) := f(x) + \bar{\alpha}_1^\top \widehat{g}(x).$$

Note that $h(\cdot)$ is convex on $X$ due to convexity of $X$ and $f$, $\mathbf{K}$-convexity of $\widehat{g}$, and the inclusion $\bar{\alpha}_1 \in \mathbf{K}_*$, see Fact IV.21.10.

Observe that (21.9) remains valid when we replace $\alpha$ with $\bar{\alpha} := \alpha/\alpha_0$. Moreover, when $x \in Y$, we have $[f(x); \widehat{g}(x)] \in T$ and thus from (21.9) and the definition of $h(x)$ we deduce that

$$x \in Y \quad \Longrightarrow \quad h(x) \geq 0. \tag{21.10}$$

**$3^0$.i.** Consider the convex sets

$$\begin{aligned} Q &:= \left\{[x;\tau] \in \mathbf{R}^n \times \mathbf{R} : \ x \in X, \ \overline{g}(x) := Ax - b \leq 0, \ \tau < 0\right\}, \\ W &:= \left\{[x;\tau] \in \mathbf{R}^n \times \mathbf{R} : \ x \in X, \ \tau \geq h(x)\right\}. \end{aligned}$$

These sets clearly are nonempty and do not intersect (since the $x$-component $x$ of a point from $Q \cap W$ would satisfy the premise but violate the conclusion in (21.10)). By Separation Theorem, there exists $[e;\beta] \neq 0$ such that

$$\sup_{[x;\tau]\in Q} \left\{e^\top x + \beta\tau\right\} \leq \inf_{[x;\tau]\in W} \left\{e^\top x + \beta\tau\right\}.$$

As $W$ is nonempty, we have $\inf_{[x;\tau]\in W}[e^\top x + \beta\tau] < +\infty$. Then, by taking into account the definition of $Q$, we deduce $\beta \geq 0$ (since otherwise the left hand side in the preceding inequality would be $+\infty$). With $\beta \geq 0$ in mind and considering the definitions of $Q$ and $W$, the preceding inequality reads

$$\sup_x \left\{e^\top x : \ x \in X, \ \overline{g}(x) \leq 0\right\} \leq \inf_x \left\{e^\top x + \beta h(x) : \ x \in X\right\}. \tag{21.11}$$

Let us define $a := \sup_x \left\{e^\top x : \ x \in X, \ \overline{g}(x) \leq 0\right\}$. Note that in (21.11), $\sup_x$ and $\inf_x$ are taken over nonempty sets, implying that $a \in \mathbf{R}$.

**$3^0$.ii.** Recall that we have seen that $\beta \geq 0$ in (21.11). We claim that in fact $\beta > 0$. Assume for contradiction that $\beta = 0$. Then, using the definition of $a$ and $\beta = 0$, (21.11) implies

$$[e^\top x - a \leq 0, \ \forall (x \in X : \overline{g}(x) \leq 0)] \quad \& \quad [e^\top x \geq a, \ \forall x \in X]. \tag{21.12}$$

Taking into account that $0 \in X$ and $\overline{g}(0) \leq 0$ by (21.7), the first relation in (21.12) says that $a \geq 0$. Then, as $a \geq 0$, from the second relation in (21.12) we deduce that $e^\top x \geq 0$ for $x \in X$. As $0 \in \operatorname{int} X$, there exists a small enough $\varepsilon > 0$ such that $(0 - \varepsilon e) \in X$. Thus, from $e^\top(-\varepsilon e) \geq 0$, we deduce that $e = 0$. But, $e = 0$ is impossible, since we are in the case when $[e;\beta] \neq 0$ and $\beta = 0$.

**$3^0$.iii.** Thus, in (21.11) we must have $\beta > 0$. Then, by replacing $e$ with $\beta^{-1}e$, we can assume that (21.11) holds true with $\beta = 1$. Once again recalling $a = \sup_x \left\{e^\top x : \ x \in X, \ \overline{g}(x) \leq 0\right\}$, the inequality (21.11) becomes

$$a \leq h(x) + e^\top x, \qquad \forall x \in X. \tag{21.13}$$

By the definition of $a$, we have also

$$d^\top x + \delta := e^\top x - a \leq 0, \qquad \forall (x \in X : \overline{g}(x) \leq 0)$$

so that the data $X$, $g(x) := \overline{g}(x)$ augmented with the just defined affine function $d^\top x + \delta$ satisfy the premise in Lemma IV.21.13 (recall that (21.7) holds true). Applying Lemma IV.21.13, we conclude that there exists $\mu \geq 0$ such that

$$e^\top x - a \leq \mu^\top \overline{g}(x), \qquad \forall x \in X.$$

Combining this relation and (21.13), we conclude that

$$h(x) + \mu^\top \overline{g}(x) \geq 0, \qquad \forall x \in X. \tag{21.14}$$

Recalling that $h(x) = f(x) + \bar{\alpha}_1^\top \widehat{g}(x)$ with $\bar{\alpha}_1 \in \mathbf{K}_*$ and setting $\overline{\lambda} = \mu$, $\widehat{\lambda} = \bar{\alpha}_1$, we get $\overline{\lambda} \geq 0$, $\widehat{\lambda} \in \mathbf{K}_*$ while by (21.14) it holds

$$f(x) + \overline{\lambda}^\top \overline{g}(x) + \widehat{\lambda}^\top \widehat{g}(x) \geq 0 \; \forall x \in X,$$

meaning that $\overline{\lambda}, \widehat{\lambda}$ solve (ConII) (recall that we are in the case of $c = 0$). □

**Proof of Lemma IV.21.13.** Recall $g(x) = Ax + a$. Let us define the following cones

$$\begin{aligned} M_1 &:= \operatorname{cl}\{[x;t] \in \mathbf{R}^n \times \mathbf{R} : \; t > 0, \, x/t \in X\}, \\ M_2 &:= \{y = [x;t] \in \mathbf{R}^n \times \mathbf{R} : \; Ax + ta \leq 0\}. \end{aligned}$$

$M_2$ is a polyhedral cone, and $M_1$ is a closed cone (in fact it is the closed conic transform $\overline{\text{ConeT}}(X)$ of $X$, see section 1.5) with a nonempty interior (since the point $[0;\ldots;0;1] \in \mathbf{R}^{n+1}$ belongs to $\operatorname{int} M_1$ due to $0 \in \operatorname{int} X$). Moreover, $(\operatorname{int} M_1) \cap M_2 \neq \varnothing$ as the point $[0;\ldots;0;1] \in \operatorname{int} M_1$ belongs to $M_2$ as $g(0) \leq 0$.

Let $f := [d;\delta] \in \mathbf{R}^n \times \mathbf{R}$. We claim that the linear form $f^\top [x;t] = d^\top x + t\delta$ is nonpositive on $M_1 \cap M_2$. Indeed, consider any $y := [z;t] \in M_1 \cap M_2$, and define $y_s = [z_s;t_s] = (1-s)y + s[0;\ldots;0;1]$. Then, for $s = 0$ we have $y_s = y$ and it is in $M_1 \cap M_2$, and when $s = 1$ we have $y_s = [0;\ldots;0;1] \in M_1 \cap M_2$. As $M_1 \cap M_2$ is convex, we observe that $y_s \in M_1 \cap M_2$ for $0 \leq s \leq 1$. Now, consider the case when $0 < s \leq 1$. Then, we have $t_s > 0$ (since $t \geq 0$ due to $y \in M_1$), and $y_s \in M_1$ implies that $w_s := z_s/t_s \in \operatorname{cl} X$ (why?), while $y_s \in M_2$ along with $t_s > 0$ implies that $g(w_s) \leq 0$. Since $0 \in \operatorname{int} X$, $w_s \in \operatorname{cl} X$ implies that $\theta w_s \in X$ for all $\theta \in [0,1)$, while $g(0) \leq 0$ along with $g(w_s) \leq 0$ implies that $g(\theta w_s) \leq 0$ for $\theta \in [0,1)$. Then, by invoking the premise of the lemma, we conclude that

$$d^\top (\theta w_s) + \delta \leq 0, \; \forall \theta \in [0,1).$$

Hence, whenever $0 < s \leq 1$, we have $d^\top w_s + \delta \leq 0$, or, equivalently, $f^\top y_s \leq 0$. As $s \to +0$, we have $f^\top y_s \to f^\top y = f^\top [z;t]$, implying that $f^\top [z;t] \leq 0$, as claimed.

Therefore, we have shown that $f^\top [x;t] = d^\top x + t\delta \leq 0$ for every $[x;t] \in (M_1 \cap M_2)$. That is, $(-f) \in (M_1 \cap M_2)_*$, where, as usual, $M_*$ is the cone dual to the cone $M$. To summarize, we have $M_1$, $M_2$ are closed cones such that $(\operatorname{int} M_1) \cap M_2 \neq \varnothing$ and $(-f) \in (M_1 \cap M_2)_*$. Applying to $M_1, M_2$ the Dubovitski-Milutin Lemma (Proposition II.8.27), we conclude that $(M_1 \cap M_2)_* = (M_1)_* + (M_2)_*$.

Since $(-f) \in (M_1 \cap M_2)_*$, there exist $\psi \in (M_1)_*$ and $\phi \in (M_2)_*$ such that $f = [d; \delta] = -\phi - \psi$. The inclusion $\phi \in (M_2)_*$ means that the homogeneous linear inequality $\phi^\top y \geq 0$ in variables $y \in \mathbf{R}^{n+1}$ is a consequence of the system of homogeneous linear inequalities given by $[A, a]y \leq 0$. Hence, by Homogeneous Farkas Lemma (Lemma I.4.1) $-\phi$ is a conic combination of the transposes of the rows of the matrix $[A, a]$, so that $\phi^\top [x; 1] = -\mu^\top g(x)$ for some nonnegative $\mu$ and all $x \in \mathbf{R}^n$. Thus, for all $x \in \mathbf{R}^n$, we deduce

$$d^\top x + \delta = [d; \delta]^\top [x; 1] = f^\top [x; 1] = -\phi^\top [x; 1] - \psi^\top [x; 1] = \mu^\top g(x) - \psi^\top [x; 1].$$

Finally, note that $[x; 1] \in M_1$ whenever $x \in X$. Then, as $\psi \in (M_1)_*$, we have $\psi^\top [x; 1] \geq 0$ for all $x \in X$. Thus, for all $x \in X$, we have $0 \leq \psi^\top [x; 1] = \mu^\top g(x) - \left(d^\top x + \delta\right)$ and so $\mu$ satisfies precisely the requirements stated in the lemma. □

To complete the story about Convex Theorem on Alternative, let us present an example which demonstrates that the relaxed Slater condition is crucial for the validity of Theorem IV.21.12.

**Example** IV.21.1 Consider the following special case of (ConI):

$$f(x) \equiv x < 0, \;\; \overline{g}(x) \equiv 0 \leq 0, \;\; \widehat{g}(x) \equiv x^2 \leq 0 \qquad \text{(ConI)}$$

(here the embedding space is $\mathbf{R}$, $X = \mathbf{R}$, $c = 0$, and $\mathbf{K} = \mathbf{R}_+$, that is, this is just a system of scalar convex constraints). System (ConII) here is the system of constraints

$$\inf_{x \in \mathbf{R}} \left[x + \overline{\lambda} \cdot 0 + \widehat{\lambda} x^2\right] \geq 0, \quad \overline{\lambda} \geq 0, \quad \widehat{\lambda} \geq 0 \qquad \text{(ConII)}$$

on variables $\overline{\lambda}, \widehat{\lambda} \in \mathbf{R}$.

System (ConI) clearly is infeasible. System (ConII) is infeasible as well – it is immediately seen that whenever $\overline{\lambda}$ and $\widehat{\lambda}$ are nonnegative, the quantity $x + \overline{\lambda} \cdot 0 + \widehat{\lambda} x^2$ is negative for all small in magnitude $x < 0$, that is, the first inequality in (ConII) is incompatible with the remaining two inequalities of the system.

Note that in this example the only missing component of the premise in Theorem IV.21.12 is the relaxed Slater condition. Let us now examine what happens when we replace the constraint $\widehat{g}(x) \equiv x^2 \leq 0$ with $\widehat{g}(x) \equiv x^2 - 2\epsilon x \leq 0$ with $\epsilon > 0$. In this case, we keep (ConI) infeasible, and gain the validity of the relaxed (relaxed, not plain!) Slater condition. Then, as all the conditions of Convex Theorem on Alternative are now met, we deduce that (ConII) which now reads

$$\inf_{x} \left[x + \overline{\lambda} \cdot 0 + \widehat{\lambda}(x^2 - 2\epsilon x)\right] \geq 0, \quad \overline{\lambda} \geq 0, \quad \widehat{\lambda} \geq 0,$$

must be feasible. In fact, $\overline{\lambda} = 0$, $\widehat{\lambda} = \frac{1}{2\epsilon}$ is a feasible solution to (ConII).

# 22

# Lagrange Function and Lagrange Duality

## 22.1 Lagrange function

Convex Theorem on Alternative brings to our attention the function

$$\underline{L}(\lambda) := \inf_{x\in X}\left[f(x)+\sum\nolimits_{j=1}^{m}\lambda_j g_j(x)\right], \tag{22.1}$$

and the related aggregate function

$$L(x,\lambda) := f(x)+\sum\nolimits_{j=1}^{m}\lambda_j g_j(x) \tag{22.2}$$

from which $\underline{L}(\lambda)$ originates. The aggregate function in (22.2) is called the *Lagrange* (or *Lagrangian*) *function* of the inequality constrained optimization program

$$\begin{array}{l}\min\{f(x):\ g_j(x)\leq 0,\ j=1,\ldots,m,\ x\in X\}\\ \text{[where } f,g_1,\ldots,g_m \text{ are real-valued functions on } X]\end{array} \tag{IC}$$

The Lagrange function $L(x,\lambda)$ of an optimization problem is a very important entity as most of the optimality conditions are expressed in terms of this function. Let us start with translating our developments from section 21.2 to the language of the Lagrange function.

## 22.2 Convex Programming Duality Theorem

We start by developing the duality theorem for convex optimization problems.

**Theorem** IV.22.1 Consider an arbitrary inequality constrained optimization program

$$\begin{array}{l}\min\{f(x):\ g_j(x)\leq 0,\ j=1,\ldots,m,\ x\in X\}\\ \text{[where } f,g_1,\ldots,g_m \text{ are real-valued functions on } X]\end{array} \tag{IC}$$

along with its Lagrange function

$$L(x,\lambda) := f(x)+\sum\nolimits_{i=1}^{m}\lambda_i g_i(x):\ X\times\mathbf{R}^m_+\to\mathbf{R}.$$

Then,

(i) [Weak Duality] For every $\lambda\geq 0$, the infimum of the Lagrange function in $x \in X$, that is,

$$\underline{L}(\lambda) := \inf_{x \in X} L(x, \lambda)$$

is a lower bound on the optimal value of (IC), so that the optimal value of the optimization problem

$$\sup_{\lambda \geq 0} \underline{L}(\lambda) \tag{IC$^*$}$$

also is a lower bound for the optimal value in (IC);

(ii) [Convex Duality Theorem] If the problem (IC)

- is convex,
- is below bounded, and
- satisfies the Relaxed Slater condition,

then the optimal value of (IC$^*$) is attained and is equal to the optimal value in (IC).

**Proof.** Let $c^*$ be the optimal value of (IC).

(i) This part is nothing but Proposition IV.21.1 (why?). It makes sense, however, to repeat here the corresponding one-line reasoning:

Consider any $\lambda \in \mathbf{R}^m_+$. Then, by definition of the Lagrange function, for any $x$ that is feasible for (IC) we have

$$L(x, \lambda) = f(x) + \sum\nolimits_{j=1}^{m} \lambda_j g_j(x) \leq f(x),$$

where the inequality follows from the facts that $\lambda \in \mathbf{R}^m_+$ and the feasibility of $x$ for (IC) implies that $g_j(x) \leq 0$ for all $j = 1, \ldots, m$. Then, we immediately arrive at

$$\underline{L}(\lambda) = \inf_{x \in X} L(x, \lambda) \leq \inf_{x \in X: g(x) \leq 0} L(x, \lambda) \leq \inf_{x \in X: g(x) \leq 0} f(x) = c^*,$$

as desired.

(ii) This part is an immediate consequence of the Convex Theorem on Alternative. Note that the system

$$f(x) < c^*, \quad g_j(x) \leq 0, \, j = 1, \ldots, m$$

has no solutions in $X$, and by Theorem IV.21.4, the system (II) associated with $c = c^*$ has a solution, i.e., there exists $\lambda^* \geq 0$ such that $\underline{L}(\lambda^*) \geq c^*$. But, we know from part (i) that the strict inequality here is impossible and, moreover, that $\underline{L}(\lambda) \leq c^*$ for every $\lambda \geq 0$. Thus, $\underline{L}(\lambda^*) = c^*$ and $\lambda^*$ is a maximizer of $\underline{L}$ over $\lambda \geq 0$. □

## 22.3 Lagrange duality and saddle points

Theorem IV.22.1 establishes a certain connection between two optimization problems, i.e., the "primal" problem

$$\begin{array}{l}\min_x \{f(x) : \ g_j(x) \leq 0, \, j = 1, \dots, m, \ x \in X\} \\ \text{[where } f, g_1, \dots, g_m \text{ are real-valued functions on } X] \, ,\end{array} \tag{IC}$$

and its *Lagrange dual problem*

$$\begin{array}{l}\max_\lambda \{\underline{L}(\lambda) : \ \lambda \in \mathbf{R}^m_+\} \\ \left[\text{where } \underline{L}(\lambda) = \inf\limits_{x \in X} L(x, \lambda) \text{ and } L(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i g_i(x)\right] .\end{array} \tag{IC$^*$}$$

Here, the variables $\lambda$ of the dual problem are called the *Lagrange multipliers* of the primal problem. Theorem IV.22.1 states that the optimal value of the dual problem is at most that of the primal, and under some favorable circumstances (i.e., when the primal problem is convex, below bounded, and satisfies the Relaxed Slater condition) the optimal values in this pair of problems are equal to each other.

In our formulation there may seem to be some asymmetry between the primal and the dual problems. In fact, both of these problems are related to the Lagrange function in a quite symmetric way. Indeed, consider the problem

$$\min_{x \in X} \overline{L}(x), \qquad \text{where } \overline{L}(x) := \sup_{\lambda \geq 0} L(x, \lambda).$$

By definition of the Lagrange function $L(x, \lambda)$, the function $\overline{L}(x)$ is clearly $+\infty$ at every point $x \in X$ which is not feasible for (IC) and is $f(x)$ on the feasible set of (IC), so that this problem is equivalent to (IC). We see that both the primal and the dual problems originate from the Lagrange function: in the primal problem, we *minimize* over $X$ the result of *maximization* of $L(x, \lambda)$ in $\lambda \geq 0$, i.e., the primal problem is

$$\min_{x \in X} \sup_{\lambda \in \mathbf{R}^m_+} L(x, \lambda),$$

and in the dual program we *maximize* over $\lambda \geq 0$ the result of *minimization* of $L(x, \lambda)$ in $x \in X$, i.e., the dual problem is

$$\max_{\lambda \in \mathbf{R}^m_+} \inf_{x \in X} L(x, \lambda).$$

This is a particular (and the most important) example of a *two-person zero-sum game* which we will explore later.

We have seen that under certain convexity and regularity assumptions the optimal values in (IC) and (IC$^*$) are equal to each. There is also another way to say when these optimal values are equal – this is always the case when the Lagrange function possesses a *saddle point*, i.e., there exists a pair $x^* \in X, \lambda^* \geq 0$ such that at the pair the function $L(x, \lambda)$ attains its minimum as a function of $x \in X$ and attains its maximum as a function of $\lambda \geq 0$:

$$L(x, \lambda^*) \geq L(x^*, \lambda^*) \geq L(x^*, \lambda) \quad \forall x \in X, \ \forall \lambda \geq 0.$$

This then leads to the following easily demonstrable fact (do it by yourself or look at Theorem IV.28.2).

**Proposition** IV.22.2 The primal-dual pair of solutions $(x^*, \lambda^*) \in X \times \mathbf{R}^k_+$ is a saddle point of the Lagrange function $L$ of (IC) if and only if $x^*$ is an optimal solution to (IC), $\lambda^*$ is an optimal solution to (IC$^*$) and the optimal values in the indicated problems are equal to each other.

# 23

# ★ Convex Programming in cone-constrained form

The results from sections 22.1 and 22.2 related to convex optimization problems in the standard MP format admit instructive extensions to the case of convex problems in cone-constrained form. We next present these extensions.

## 23.1 Convex problem in cone-constrained form

Convex problem in cone-constrained form is an optimization problem of the form

$$\mathrm{Opt}(P) = \min_{x\in X} \left\{ f(x) : \ \overline{g}(x) \leq 0, \ \widehat{g}(x) \leq_K 0 \right\}, \tag{$P$}$$

where $X \subseteq \mathbf{R}^n$ is a nonempty convex set, $f : X \to \mathbf{R}$ is a convex function, $\overline{g}(x) := Ax - b$ is an affine function from $\mathbf{R}^n$ to $\mathbf{R}^k$, $\mathbf{K} \subset \mathbf{R}^\nu$ is a regular cone, and $\widehat{g}(\cdot) : X \to \mathbf{R}^\nu$ is $\mathbf{K}$-convex.

**Example** IV.23.1 Recall that the positive semidefinite cone $\mathbf{S}^n_+$ and the notation $A \succeq B$, $B \preceq A$, $A \succ B$, $B \prec A$ for the associated non-strict and strict conic inequalities were introduced in section D.2.2. As we know from Fact IV.21.7 and Example II.8.9, the cone $\mathbf{S}^n_+$ is regular and self-dual. Recall from Lemma IV.21.9 that the function from $\mathbf{S}^n$ to $\mathbf{S}^n$ given by $\widehat{g}(x) = xx^\top = xx = x^2$ is $\succeq$-convex. As a result, the problem

$$\begin{aligned} \mathrm{Opt}(P) &= \min_{x=(t,y)\in\mathbf{R}\times\mathbf{S}^n} \left\{ t : \ \mathrm{Tr}(y) \leq t, \ y^2 \preceq B \right\} \qquad (23.1) \\ &= \min_{x=(t,y)\in\mathbf{R}\times\mathbf{S}^n} \left\{ t : \ \langle y, I_n\rangle - t \leq 0, \ y^2 \preceq B \right\} \end{aligned}$$

where $B$ is a positive definite matrix, and $\langle\cdot,\cdot\rangle$ is the Frobenius inner product, is a convex program in cone-constrained form.

## 23.2 Cone-constrained Lagrange function

Consider $(P)$ with a convex objective $f$, a convex domain $X$, an affine map $\overline{g}(\cdot) : \mathbf{R}^n \to \mathbf{R}^k$, a regular cone $\mathbf{K} \subset \mathbf{R}^\nu$, and a $\mathbf{K}$-convex function $\widehat{g}(\cdot) : X \to \mathbf{R}^\nu$. Let $\Lambda := \mathbf{R}^k_+ \times \mathbf{K}_*$, where $\mathbf{K}_*$ is the cone dual to $\mathbf{K}$, and consider $\lambda := [\overline{\lambda}; \widehat{\lambda}] \in \Lambda$. Then, the *cone-constrained Lagrange function* of $(P)$ is defined as

$$L(x;\lambda) := f(x) + \overline{\lambda}^\top \overline{g}(x) + \widehat{\lambda}^\top \widehat{g}(x) : X \times \Lambda \to \mathbf{R}.$$

By construction, for any $\lambda \in \Lambda$, we have that $L(x;\lambda)$ as a function of $x$ underestimates $f(x)$ everywhere on the feasible domain of $(P)$.

**Example** IV.23.2 (continued from Example IV.23.1) We see that the cone-constrained Lagrange function of (23.1) is given by

$$L(t, y; \overline{\lambda}, \widehat{\lambda}) = t + \overline{\lambda}[\mathrm{Tr}(y) - t] + \mathrm{Tr}(\widehat{\lambda}(y^2 - B)) : [\mathbf{R} \times \mathbf{S}^n] \times [\mathbf{R}_+ \times \mathbf{S}^n_+] \to \mathbf{R}.$$

*Cone-constrained Lagrange dual* of $(P)$ is the optimization problem

$$\mathrm{Opt}(D) := \max\left\{\underline{L}(\lambda) : \ \lambda \in \Lambda\right\}, \qquad [\text{where } \underline{L}(\lambda) := \inf_{x \in X} L(x;\lambda)], \qquad (D)$$

where $\Lambda := \mathbf{R}^k_+ \times \mathbf{K}_*$. From $L(x;\lambda) \le f(x)$ for all $x \in X$ and $\lambda \in \Lambda$, we clearly extract that

$$\mathrm{Opt}(D) \le \mathrm{Opt}(P). \qquad \text{[Weak Duality]}$$

Note that Weak Duality is independent of any assumptions of convexity on $f$, $X$ and on $\mathbf{K}$-convexity of $\widehat{g}$.

**Example** IV.23.3 (continued from Example IV.23.1) It is immediately seen (check it, or look at the solution to Exercise IV.22) that for the cone-constrained Lagrange function of (23.1) we have $\mathrm{Dom}(\underline{L}) = \{[\overline{\lambda}; \widehat{\lambda}] : \ \overline{\lambda} = 1, \ \widehat{\lambda} \succ 0\}$ and

$$\underline{L}(\overline{\lambda}, \widehat{\lambda}) = \begin{cases} -\frac{1}{4}\mathrm{Tr}(\widehat{\lambda}^{-1}) - \mathrm{Tr}(\widehat{\lambda}B), & \text{if } [\overline{\lambda}, \widehat{\lambda}] \in \mathrm{Dom}(\underline{L}) \\ -\infty, & \text{otherwise} \end{cases}. \qquad (23.2)$$

Then, the cone-constrained Lagrange dual of (23.1) is the problem

$$\begin{aligned} \mathrm{Opt}(D) &= \max_{\widehat{\lambda} \succ 0, \overline{\lambda} \ge 0} \left\{-\frac{1}{4}\mathrm{Tr}(\widehat{\lambda}^{-1}) - \mathrm{Tr}(\widehat{\lambda}B) : \ \overline{\lambda} = 1\right\} \\ &= \max_{\widehat{\lambda} \succ 0} \left\{-\frac{1}{4}\mathrm{Tr}(\widehat{\lambda}^{-1}) - \mathrm{Tr}(\widehat{\lambda}B)\right\}. \qquad (23.3) \end{aligned}$$

## 23.3 Convex Programming Duality Theorem in cone-constrained form

For cone-constrained problems, we have the following strong duality theorem.

**Theorem** IV.23.1 [Convex Programming Duality Theorem in cone-constrained form] Consider convex cone-constrained problem $(P)$, that is, $X$ is convex, $f$ is real-valued and convex on $X$, and $\widehat{g}(\cdot)$ is well defined and $\mathbf{K}$-convex on $X$. Assume that the problem is below bounded and satisfies the Relaxed Slater condition. Then, $(D)$ is solvable and $\mathrm{Opt}(P) = \mathrm{Opt}(D)$.

Note that the only nontrivial part (ii) of Theorem IV.22.1 is nothing but the special case of Theorem IV.23.1 where $\mathbf{K}$ is a nonnegative orthant.

**Proof of Theorem IV.23.1.** This proof is immediate. Under the premise of the theorem, $c := \mathrm{Opt}(P)$ is a real, and the system of constraints (ConI) associated with this $c$ has no solutions. Relaxed Slater Condition along with Convex Theorem on Alternative in cone-constrained form (Theorem IV.21.12) imply the solvability of (ConII), i.e., the existence of $\lambda_* = [\overline{\lambda}_*; \widehat{\lambda}_*] \in \Lambda$ such that

$$\underline{L}(\lambda_*) = \inf_{x\in X} \left\{ f(x) + \overline{\lambda}_*^\top \overline{g}(x) + \widehat{\lambda}_*^\top \widehat{g}(x) \right\} \geq c = \mathrm{Opt}(P).$$

Thus, we deduce that $(D)$ has a feasible solution with objective value $\geq \mathrm{Opt}(P)$, By Weak Duality, this value is exactly $\mathrm{Opt}(P)$, the solution in question is optimal for $(D)$, and $\mathrm{Opt}(P) = \mathrm{Opt}(D)$. □

**Example** IV.23.4 (continued from Example IV.23.1) Problem (23.1) is clearly below bounded and satisfies Slater condition (since $B \succ 0$). By Theorem IV.23.1 the dual problem (23.3) is solvable and has the same optimal value as (23.1). The solution for the (convex!) dual problem (23.3) can be found by applying the Fermat rule. To this end, note also that for a positive definite $n \times n$ matrix $y$ and $h \in \mathbf{S}^n$ it holds that

$$\frac{d}{dt}\Big|_{t=0}(y + th)^{-1} = -y^{-1}hy^{-1}$$

(why?). Then, the Fermat rule says that the optimal solution to (23.3) is

$$\overline{\lambda}_* = 1, \quad \widehat{\lambda}_* = \frac{1}{2}B^{-1/2}$$

and $\mathrm{Opt}(P) = \mathrm{Opt}(D) = -\mathrm{Tr}(B^{1/2})$.

**23.3.A “Subgradient interpretation” of Lagrange multipliers.** Consider any $\Delta := [\overline{\delta}; \widehat{\delta}] \in \mathbf{R}^k \times \mathbf{R}^\nu$, and define the maps $\overline{g}_\Delta(x) := Ax - b - \overline{\delta} = \overline{g}(x) - \overline{\delta}$ and $\widehat{g}_\Delta(x) := \widehat{g}(x) - \widehat{\delta}$ along with the parametric family of convex cone-constrained problems $(P_\Delta)$ given by

$$\begin{aligned}\mathrm{Opt}(P_\Delta) &:= \min_{x\in X} \{ f(x) : \ \overline{g}_\Delta(x) \leq 0, \ \widehat{g}_\Delta(x) \leq_K 0 \} \\ &= \min_{x\in X} \left\{ f(x) : \ \overline{g}(x) - \overline{\delta} \leq 0, \ \widehat{g}(x) - \widehat{\delta} \leq_K 0 \right\}.\end{aligned}$$

Notice that $(P)$ is part of this family, as it is precisely $(P_0)$.

The cone-constrained Lagrange duals of these problems $(P_\Delta)$ also form a parametric family. Specifically, by setting $\lambda = [\overline{\lambda}; \widehat{\lambda}]$ and $\Lambda = \mathbf{R}^k_+ \times \mathbf{K}_*$, we arrive at the cone-constrained Lagrange function of $(P_\Delta)$ as

$$L_\Delta(x, \lambda) := f(x) + \overline{\lambda}^\top (\overline{g}(x) - \overline{\delta}) + \widehat{\lambda}^\top (\widehat{g}(x) - \widehat{\delta}),$$

the resulting dual family of problems $(D_\Delta)$ given by

$$\begin{aligned}&\mathrm{Opt}(D_\Delta) := \max_{\lambda\in\Lambda} \{ \underline{L}_\Delta(\lambda) \}, \\ &\text{where } \underline{L}_\Delta(\lambda) := \inf_{x\in X} \{ L_\Delta(x, \lambda) \} = \inf_{x\in X} \left\{ f(x) + \overline{\lambda}^\top (\overline{g}(x) - \overline{\delta}) + \widehat{\lambda}^\top (\widehat{g}(x) - \widehat{\delta}) \right\}.\end{aligned}$$

Since $L_\Delta(x, \lambda) = L_0(x, \lambda) - \overline{\lambda}^\top \overline{\delta} - \widehat{\lambda}^\top \widehat{\delta}$, we deduce $\underline{L}_\Delta(\lambda) = \underline{L}_0(\lambda) - \overline{\lambda}^\top \overline{\delta} - \widehat{\lambda}^\top \widehat{\delta}$, where by definition $\underline{L}_0(\lambda) = \inf_{x\in X}\left\{f(x) + \overline{\lambda}^\top \overline{g}(x) + \widehat{\lambda}^\top \widehat{g}(x)\right\}$. Thus,

$$\mathrm{Opt}(D_\Delta) = \max_{\lambda\in\Lambda}\{\underline{L}_\Delta(\lambda)\} = \max_{\lambda\in\Lambda}\left\{\underline{L}_0(\lambda) - \overline{\lambda}^\top\overline{\delta} - \widehat{\lambda}^\top\widehat{\delta}\right\}.$$

We have the following nice and instructive fact that provides further insights to the optimum value sensitivity of these parametric families of problems.

**Fact** IV.23.2 Consider parametric family $(P_\Delta)$ of convex cone-constrained problems along with the family $(D_\Delta)$ of their cone-constrained Lagrange duals. Then,

(i) If $\underline{L}_0(\mu) > -\infty$ for some $\mu = [\overline{\mu};\widehat{\mu}] \in \Lambda$, then the primal optimal value $\mathrm{Opt}(P_\Delta)$ takes values in $\mathbf{R}\cup\{+\infty\}$ and is a convex function of $\Delta$.

(ii) If $(D_0)$ is solvable with optimal solution $\lambda_* = [\overline{\lambda}_*;\widehat{\lambda}_*]$ and $\mathrm{Opt}(D_0) = \mathrm{Opt}(P_0)$, then $-\lambda_*$ is a subgradient of $\mathrm{Opt}(P_\Delta)$ at the point $\Delta = 0$, i.e.,

$$\mathrm{Opt}(P_\Delta) \geq \mathrm{Opt}(P_0) - \overline{\lambda}_*^\top\overline{\delta} - \widehat{\lambda}_*^\top\widehat{\delta}, \qquad \forall(\Delta = [\overline{\delta};\widehat{\delta}]).$$

The premises in (i) and (ii) definitely take place when $(P_0)$ satisfies the Relaxed Slater condition and is below bounded.

**Example** IV.23.5 (continued from Example IV.23.1) Problem (23.1) can be embedded into the parametric family of problems

$$\mathrm{Opt}(P_R) := \min_{x=(t,y)\in\mathbf{R}\times\mathbf{S}^n}\left\{t:\ \mathrm{Tr}(y) \leq t,\ y^2 \preceq B + R\right\} \qquad (P[R])$$

with $R$ varying through $\mathbf{S}^n$. Taking into account all we have established so far for this problem and also considering Fact IV.23.2, we arrive at

$$\mathrm{Tr}((B+R)^{1/2}) = -\mathrm{Opt}(P_R) \leq \mathrm{Tr}(B^{1/2}) + \frac{1}{2}\mathrm{Tr}(B^{-1/2}R), \quad \forall(R \succ -B). \quad (23.4)$$

Note that this is nothing but the Gradient inequality for the concave (see Fact III.18.6) function $\mathrm{Tr}(X^{1/2}) : \mathbf{S}^n_+ \to \mathbf{R}$, see Fact D.24.

## 23.4 Conic Programming and Conic Duality Theorem

In this section, we will consider a special case of convex problem in cone-constrained form, namely *Conic programs* in today's optimization terminology. In conic programs, $f(x)$ is linear, $X$ is the entire space, and $\widehat{g}(x) = Px - p$ is affine. Note that affine mapping is $\mathbf{K}$-convex whatever be the cone $\mathbf{K}$. Thus, conic problem automatically satisfies convexity restrictions from cone-constrained Convex Programming Duality Theorem (Theorem IV.23.1) and is an optimization problem of the form

$$\mathrm{Opt}(P) = \min_{x\in\mathbf{R}^n}\left\{c^\top x:\ Ax - b \leq 0,\ Px - p \leq_{\mathbf{K}} 0\right\}, \qquad (23.5)$$

where $\mathbf{K}$ is a regular cone in certain $\mathbf{R}^\nu$.

The simplest example of a conic problem is an LP problem, where $\mathbf{K}$ is a nonnegative orthant. Another instructive example is the conic reformulation of convex quadratic quadratically constrained problem. This example relies on the following useful observation.

**Fact** IV.23.3 Consider the convex quadratic constraint $x^\top A^\top Ax \leq b^\top x + c$, where $A \in \mathbf{R}^{d\times n}$, $b \in \mathbf{R}^n$, and $c \in \mathbf{R}$. This constraints can be equivalently rewritten as a conic constraint involving Lorentz cone, i.e.,

$$\begin{aligned} & x^\top A^\top Ax \leq b^\top x + c \\ \iff & [2Ax;\ b^\top x + c - 1;\ b^\top x + c + 1] \in \mathbf{L}^{d+2} \\ \iff & 4x^\top Ax + (b^\top x + c - 1)^2 \leq (b^\top x + c + 1)^2 \ \text{ and } \ b^\top x + c + 1 \geq 0. \end{aligned}$$

Fact IV.23.3 immediately leads to the following useful result.

**Fact** IV.23.4 Given $A_j \in \mathbf{R}^{d_j\times n}$, $b_j \in \mathbf{R}^n$, and $c_j \in \mathbf{R}$, for $0 \leq j \leq m$, the convex quadratic quadratically constrained optimization problem

$$\min_x \left\{x^\top A_0^\top A_0 x + b_0^\top x + c_0 :\ x^\top A_j^\top A_j x + b_j^\top x + c_j \leq 0,\ 1 \leq j \leq m\right\}$$

is equivalent to the conic problem on the product of $m+1$ Lorentz cones, specifically, it admits the conic formulation given by

$$\min_{x,t} \left\{ t : \begin{array}{ll} A[x;t] + b & := [\alpha_0(x,t); \alpha_1(x); \ldots; \alpha_m(x)] \\ & \in \mathbf{L}^{d_0+2} \times \mathbf{L}^{d_1+2} \times \ldots \times \mathbf{L}^{d_m+2} \end{array} \right\},$$

where $\alpha_0(x,t) := [2A_0x;\ t - b_0^\top x - c_0 - 1;\ t - b_0^\top x - c_0 + 1]$ and $\alpha_j(x) := [2A_jx;\ -b_j^\top x - c_j - 1;\ -b_j^\top x - c_j + 1]$ for all $1 \leq j \leq m$.

The cone-constrained Lagrange dual problem of problem (23.5) reads

$$\max_{\overline{\lambda},\widehat{\lambda}} \left[\inf_x \left\{c^\top x + \overline{\lambda}^\top (Ax - b) + \widehat{\lambda}^\top (Px - p) :\ \overline{\lambda} \geq 0,\ \widehat{\lambda} \in \mathbf{K}_*\right\}\right],$$

which, as it is immediately seen, is nothing but the problem

$$\mathrm{Opt}(D) = \max_{\overline{\lambda},\widehat{\lambda}} \left\{-b^\top \overline{\lambda} - p^\top \widehat{\lambda} :\ A^\top \overline{\lambda} + P^\top \widehat{\lambda} + c = 0,\ \overline{\lambda} \geq 0,\ \widehat{\lambda} \in \mathbf{K}_*\right\}. \qquad (D)$$

Recall that by Fact IV.21.6.i the cone dual to a regular cone also is a regular cone. As a result, the problem $(D)$, called the *conic dual* of the conic problem (23.5), also is a conic problem. An immediate computation (utilizing the fact that $(\mathbf{K}_*)_* = \mathbf{K}$ for every regular cone $\mathbf{K}$) shows that *conic duality is symmetric*.

**Fact** IV.23.5 Conic duality is symmetric, i.e., the conic dual to conic problem $(D)$ is (equivalent to) conic problem (23.5).

In view of primal-dual symmetry, Convex Duality Theorem in cone-constrained form (Theorem IV.23.1) in the Conic Programming case takes the following nice form.

**Theorem** IV.23.6 [Conic Duality Theorem] Consider a primal-dual pair of conic problems

$$\mathrm{Opt}(P) := \min_{x\in\mathbf{R}^n} \left\{c^\top x : Ax - b \leq 0,\, Px - x {\leq_{\mathbf{K}}} 0\right\}, \tag{$P$}$$

$$\mathrm{Opt}(D) := \max_{\overline{\lambda},\widehat{\lambda}} \left\{-b^\top\overline{\lambda} - p^\top\widehat{\lambda} :\ A^\top\overline{\lambda} + P^\top\widehat{\lambda} + c = 0,\ \overline{\lambda} \geq 0,\, \widehat{\lambda} \in \mathbf{K}_*\right\}. \tag{$D$}$$

Then, we always have $\mathrm{Opt}(D) \leq \mathrm{Opt}(P)$. Moreover, if one of the problems in the pair is bounded and satisfies the Relaxed Slater condition, then the other problem in the pair is *solvable*, and $\mathrm{Opt}(P) = \mathrm{Opt}(D)$. Finally, if both of the problems satisfy Relaxed Slater condition, then both are solvable with equal optimal values.

**Proof.** This proof is immediate. Weak duality has already been verified. To verify the second claim, note that by primal-dual symmetry we can assume that the bounded problem satisfying Relaxed Slater condition is $(P)$. But, then the claim in question is given by Theorem IV.23.1. Finally, if both problems satisfy Relaxed Slater condition (and in particular are feasible), by Weak Duality, both are bounded, and therefore solvable with equal optimal values by the preceding claim. □

**Application example: $\mathcal{S}$-Lemma** $\mathcal{S}$-Lemma is an extremely useful fact that has applications in optimization, engineering, and control.

**Lemma** IV.23.7 ($\mathcal{S}$-Lemma) Let $A, B \in \mathbf{S}^n$ be such that

$$\exists \bar{x} :\quad \bar{x}^\top A\bar{x} > 0. \tag{23.6}$$

Then, the implication

$$x^\top Ax \geq 0 \quad\Longrightarrow\quad x^\top Bx \geq 0 \tag{23.7}$$

holds if and only if

$$\exists \lambda \geq 0 :\quad B \succeq \lambda A. \tag{23.8}$$

Note that $\mathcal{S}$-Lemma is the statement of the same flavor as Homogeneous Farkas Lemma: the latter states that a homogeneous linear inequality $b^\top x \geq 0$ is a consequence of a system of homogeneous linear inequalities $a_i^\top x \geq 0$, $1 \leq i \leq k$, if and only if the target inequality can be obtained from the inequalities of the system by taking weighted sum with nonnegative weights; we could add to “taking weighted sum” also “and adding identically true homogeneous linear inequality” – by the simple reason that there exists only one inequality of the latter type, $0^\top x \geq 0$. Similarly, $\mathcal{S}$-Lemma says that (whenever (23.6) holds) homogeneous quadratic inequality $x^\top Bx \geq 0$ is a consequence of (single-inequality) system of homogeneous quadratic inequalities $x^\top Ax \geq 0$ if and only if the target inequality can be obtained by taking weighted sum, with nonnegative weights, of inequalities of the system (that is, by taking a nonnegative multiple of the inequality $x^\top Ax \geq$

0) and adding an identically true homogeneous quadratic inequality (there are plenty of them, these are inequalities $x^\top Cx \geq 0$ with $C \succeq 0$).

Note that the possibility for the target inequality to be obtained by summing up, with nonnegative weights, inequalities from certain system and adding an identically true inequality is clearly a *sufficient* condition for the target inequality to be a consequence of the system. The actual power of Homogeneous Farkas Lemma and $\mathcal{S}$-Lemma is in the fact that this evident sufficient condition is also necessary for the conclusion in question to be valid (in the case of linear inequalities – whenever the system is finite, in the case of $\mathcal{S}$-Lemma – when the system is a single-inequality one and (23.6) takes place). The fact that in the quadratic case to guarantee the necessity, the system should be a single-inequality one, whatever unpleasant, is a must. In fact, a straightforward "quadratic version" of Homogeneous Farkas Lemma fails, in general, to be true already when there are just two quadratic inequalities in the system. This being said, even that poor, as compared to its linear inequalities analogy, $\mathcal{S}$-Lemma is extremely useful...

In preparation to $\mathcal{S}$-Lemma, we will first prove the following weaker statement.

**Lemma** IV.23.8 Given $A, B \in \mathbf{S}^n$ such that $\exists \bar{x}$ satisfying $\bar{x}^\top A\bar{x} > 0$, the implication

$$\{X \succeq 0,\ \mathrm{Tr}(AX) \geq 0\} \quad \Longrightarrow \quad \mathrm{Tr}(BX) \geq 0 \tag{23.9}$$

holds if and only if $B \succeq \lambda A$ for some $\lambda \geq 0$.

**Proof.** The "if" part of this lemma is evident. To prove the "only if" part, consider the conic problem

$$\mathrm{Opt}(P) = \min_X \{\mathrm{Tr}(BX) :\ \mathrm{Tr}(AX) \geq 0,\ X \succeq 0\} \tag{$P$}$$

along with its conic dual, which is given by

$$\mathrm{Opt}(D) = \max_{\lambda, Y} \{0 \cdot \lambda + \mathrm{Tr}(0_{n\times n}Y) :\ \lambda \geq 0,\ Y \succeq 0,\ Y + \lambda A = B\} \tag{$D$}$$

(derive the conic dual of $(P)$ yourself by utilizing the fact that $\mathbf{S}^n_+$ is self-dual). Note that from the premise of (23.6), we deduce that for large enough nonnegative $t$, the solution $\bar{X} := I_n + t\bar{x}\bar{x}^\top$ will ensure that the Slater condition holds true. Moreover, under the premise of (23.9) $\mathrm{Opt}(P)$ is bounded from below by 0 as well. Then, by Conic Duality Theorem, the dual $(D)$ is solvable, implying that $B \succeq \lambda A$ for some $\lambda \geq 0$, as required in this lemma and completing the proof. □

Note that (23.7) is nothing but (23.9) with $X$ restricted to be of rank $\leq 1$. Indeed, $X \succeq 0$ is of rank $\leq 1$ if and only if $X = xx^\top$ for some vector $x$, and in this case $\mathrm{Tr}(PX) = x^\top Px$ for every symmetric $P$ of appropriate size. We are now ready to complete the proof of $\mathcal{S}$-Lemma.

**Proof of Lemma IV.23.7 ($\mathcal{S}$-Lemma).** The "if" part is evident. To prove the "only if" part, assume that implication (23.7) holds true, and let us verify that $B \succeq \lambda A$ for some $\lambda \geq 0$. By Lemma IV.23.8, all we need to this end is to show that the validity implication (23.7) implies the validity of implication (23.9). Thus, assume that $x^\top Ax \geq 0$ does imply that $x^\top Bx \geq 0$, and let $X \succeq 0$ be such that $\mathrm{Tr}(AX) \geq 0$; all we need is to prove that in this case $\mathrm{Tr}(BX) \geq 0$ holds as well. To this end, let $X^{1/2}AX^{1/2} = U\,\mathrm{Diag}\{\mu\}U^\top$ be the eigenvalue decomposition of $X^{1/2}AX^{1/2}$. By defining $\overline{\mu} := \mathrm{Tr}(\mathrm{Diag}\{\mu\})$ and using the relation $X^{1/2}AX^{1/2} = U\,\mathrm{Diag}\{\mu\}U^\top$, we arrive at

$$\overline{\mu} = \mathrm{Tr}(\mathrm{Diag}\{\mu\}) = \mathrm{Tr}(U^\top X^{1/2}AX^{1/2}U) = \mathrm{Tr}(X^{1/2}AX^{1/2}) = \mathrm{Tr}(AX) \geq 0.$$

Now, consider an $n$-dimensional Rademacher random vector $\zeta$, i.e., $n$-dimensional vector with entries which, independently of each other, take values $\pm 1$ with probabilities 1/2. By setting $\xi := X^{1/2}U\zeta$, we get

$$\begin{aligned}\xi^\top A\xi = \zeta^\top(U^\top X^{1/2}AX^{1/2}U)\zeta &= \zeta^\top\,\mathrm{Diag}\{\mu\}\zeta \\ &= \mathrm{Tr}(\mathrm{Diag}\{\mu\}\zeta\zeta^\top) = \mathrm{Tr}(\mathrm{Diag}\{\mu\}) = \overline{\mu} \geq 0.\end{aligned}$$

Thus, $\xi^\top A\xi \geq 0$ for all realizations of $\xi$. Recalling that we are in the case when (23.7) holds, we conclude that $\xi^\top B\xi \geq 0$ for all realizations of $\xi$, or, which is the same, $\zeta^\top(U^\top X^{1/2}BX^{1/2}U)\zeta \geq 0$ for all realizations of $\zeta$. Passing to expectations and recalling that $\zeta$ is Rademacher random vector, we get

$$\begin{aligned}0 \leq \mathbf{E}_\zeta\left[\zeta^\top(U^\top X^{1/2}BX^{1/2}U)\zeta\right] &= \mathrm{Tr}(U^\top X^{1/2}BX^{1/2}U) \\ &= \mathrm{Tr}(X^{1/2}BX^{1/2}) = \mathrm{Tr}(BX),\end{aligned}$$

that is, $\mathrm{Tr}(BX) \geq 0$, so that (23.9) does hold true. □

**Inhomogeneous $\mathcal{S}$-Lemma.** $\mathcal{S}$-Lemma provides a necessary and sufficient condition for a homogeneous quadratic inequality $x^\top Bx \geq 0$ to be a consequence of strictly feasible homogeneous quadratic inequality $x^\top Ax \geq 0$. What about the inhomogeneous case? When an inhomogeneous quadratic inequality

$$x^\top Bx + 2b^\top x + \beta \geq 0 \tag{$B$}$$

is a consequence of a strictly feasible inhomogeneous quadratic inequality

$$x^\top Ax + 2a^\top x + \alpha \geq 0 \tag{$A$}$$

? The answer is easy to guess. The implication $(A) \implies (B)$ is nothing but the implication

$$\forall(t \neq 0, x):\quad x^\top Ax + 2ta^\top x + \alpha t^2 \geq 0 \implies x^\top Bx + 2tb^\top x + \beta t^2 \geq 0 \tag{$*$}$$

(plug $x/t$ instead of $x$ into $(A)$ and look what happens with $(B)$). We understand when a bit stronger implication

$$\forall(t, x):\quad x^\top Ax + 2ta^\top x + \alpha t^2 \geq 0 \implies x^\top Bx + 2tb^\top x + \beta t^2 \geq 0 \tag{$**$}$$

holds true: the homogeneous inequality in the premise of $(**)$ is strictly feasible along with $(A)$, so that by the homogeneous $\mathcal{S}$-Lemma $(**)$ holds true if and only if

$$\exists\lambda \geq 0:\quad \begin{bmatrix} B & b \\ b^\top & \beta \end{bmatrix} \succeq \lambda \begin{bmatrix} A & a \\ a^\top & \alpha \end{bmatrix}. \tag{23.10}$$

Thus, *if* we knew that in the case of strictly feasible $(A)$ the validity of implication $(*)$ is the same as the validity of implication $(**)$, we could be sure that the first of these implications takes place if and only if (23.10) takes place. The above "if" indeed is true.

**Lemma** IV.23.9 [Inhomogeneous $\mathcal{S}$-Lemma] Let $A, B \in \mathbf{S}^n$. Suppose there exists $\bar{x}$ such that $\bar{x}^\top A\bar{x} + 2a^\top \bar{x} + \alpha > 0$. Then, the implication $(A) \implies (B)$ takes place if and only if (23.10) takes place.

**Proof.** Suppose the premise holds, i.e., $\bar{x}^\top A\bar{x} + 2a^\top \bar{x} + \alpha > 0$ for some $\bar{x}$. Based on the discussion preceding this lemma all we need to verify is that the validity of $(*)$ is exactly the same as the validity of $(**)$. Clearly, the validity of $(**)$ implies the validity of $(*)$, so our task boils down to demonstrating that under the premise of the lemma, the validity of $(*)$ implies the validity of $(**)$. Thus, assume that $(*)$ is valid, and let us prove that $(**)$ is valid as well. All we need to prove is that $y^\top Ay \geq 0$ implies $y^\top By \geq 0$. Thus, assume that $y$ is such that $y^\top Ay \geq 0$, and let us prove that $y^\top By \geq 0$ as well. Define $x_t := t\bar{x} + (1-t)y$, and consider the univariate quadratic functions $q_a(t) := x_t^\top Ax_t + 2ta^\top x_t + \alpha t^2$, $q_b(t) := x_t^\top Bx_t + 2tb^\top x_t + \beta t^2$. We have so far seen that

(a) for all $t \neq 0$, $q_a(t) \geq 0 \implies q_b(t) \geq 0$,

(b) $q_a(1) > 0$ and $q_a(0) \geq 0$,

and we would like to show that $q_b(0) \geq 0$. Note that $q_a$ and $q_b$ are linear or quadratic functions of $t$ and thus they are continuous in $t$. Now, consider the following cases (draw $q_a$ in these cases!):

- If $q_a(0) > 0$, by continuity of $q_a$ we have $q_a(t) > 0$ for all small enough nonzero $t$, and so in such a case, by (a) we also get $q_b(t) \geq 0$ for all small enough in magnitude nonzero $t$, implying, by continuity, that $q_b(0) \geq 0$.
- If $q_a(0) = 0$, the reasoning goes as follows. When $t$ varies from 0 to 1, the linear or quadratic function $q_a(t)$ varies from 0 to something positive. It follows that
  - either $q_a(t) \geq 0$, $0 \leq t \leq 1$, implying by (a) that $q_b(t) \geq 0$ for $t \in (0,1]$, and so $q_b(0) \geq 0$ holds by continuity of $q_b(t)$ at $t = 0$,
  - or $q_a(\bar{t}) < 0$ holds for some $\bar{t} \in (0,1)$. Assuming that this is the case, the linear or quadratic function $q_a(t)$ is zero at $t = 0$, negative somewhere on $(0,1)$, and positive at $t = 1$. Therefore, $q_a$ is quadratic, and not linear, function of $t$ which has exactly one root in the interval $(0,1)$. Let this root in $(0,1)$ be $t_1$. Recall that the other root of the quadratic function $q_a$ is $t = 0$, thus we must have $q_a(t) = c(t-0)(t-t_1)$ for some $c \in \mathbf{R}$. From $t_1 < 1$ and $q_a(1) > 0$ it follows that $c > 0$; this, in turn, combines with $t_1 > 0$ to imply that $q_a(t) > 0$ when $t < 0$. By (a) it follows that $q_b(t) \geq 0$ for $t < 0$, whence by continuity $q_b(0) \geq 0$.

□

As an important consequence of Inhomogeneous $\mathcal{S}$-Lemma we arrive at the following observation that states that the semidefinite programming relaxation of the quadratically constraint quadratic program with a single strictly feasible quadratic constraint is exact.

**Corollary** IV.23.10 Let $A, B \in \mathbf{S}^n$. Suppose there exists $\bar{x}$ such that $\bar{x}^\top A\bar{x} + 2a^\top \bar{x} + \alpha > 0$. Then,

$$\begin{aligned}\beta^* &:= \inf_x \left\{x^\top Bx + 2b^\top x : \; x^\top Ax + 2a^\top x + \alpha \geq 0\right\} \\ &= \max_{\beta,\lambda} \left\{\beta : \; \lambda \geq 0, \; \begin{bmatrix} B - \lambda A & b - \lambda a \\ b^\top - \lambda a^\top & -\lambda\alpha - \beta \end{bmatrix} \succeq 0\right\}.\end{aligned}$$

(Here, as always, the optimal value of an infeasible maximization problem is $-\infty$.)

**Proof.** Define the quadratic functions $q(x) := x^\top Ax + 2a^\top x + \alpha$ and $q_\beta)(x) := x^\top Bx + 2b^\top x - \beta$. Note that $\beta^*$ is the supremum of all $\beta$'s satisfying $\beta \leq \inf_x\{x^\top Bx + 2b^\top x : \; x^\top Ax + 2a^\top x + \alpha \geq 0\}$, that is, those $\beta$ for which the implication

$$q(x) \geq 0 \quad \Longrightarrow \quad q_\beta(x) \geq 0$$

holds true. By the premise of the corollary, there exists $\bar{x}$ satisfying $q(\bar{x}) > 0$, so that by Inhomogeneous $\mathcal{S}$-lemma (see Lemma IV.23.9) $\beta^*$ is the supremum of those $\beta$ which can be augmented by appropriate $\lambda$ to yield feasible solutions to the maximization problem in the corollary's formulation, or, which is the same, $\beta^*$ is the optimal value in the latter problem. □

**Example** IV.23.6 (continued from Example IV.23.1) Invoking Schur Complement Lemma (Proposition D.33), we can rewrite (23.1) equivalently as the conic problem

$$\mathrm{Opt}(P) = \min_{t \in \mathbf{R}, y \in \mathbf{S}^n} \left\{t : \; \mathrm{Tr}(y) \leq t, \; \begin{bmatrix} B & y \\ y & I_n \end{bmatrix} \succeq 0\right\}. \tag{23.11}$$

The conic dual of (23.11) can be obtained as follows: we equip the scalar inequality $t - \mathrm{Tr}(y) \geq 0$ with Lagrange multiplier $\overline{\lambda} \geq 0$, the inequality $\begin{bmatrix} B & y \\ y & I_n \end{bmatrix} \succeq 0$ with Lagrange multiplier $\begin{bmatrix} U & V \\ V^\top & W \end{bmatrix} \succeq 0$ (recall that the semidefinite cone is self-dual, so that legitimate Lagrange multipliers for $\succeq$-constraints are $\succeq$-nonnegative), and sum up the termwise inner products of the constraints of (23.11), thus arriving at the aggregated inequality

$$\overline{\lambda}(t - \mathrm{Tr}(y)) + 2\mathrm{Tr}(yV^\top) \geq -\mathrm{Tr}(BU) - \mathrm{Tr}(W).$$

which by construction is a consequence of the constraints in (23.11). We then impose on the Lagrange multipliers, in addition to the above conic constraints, the restriction that the left hand side in the aggregated inequality is identically in $t \in \mathbf{R}, y \in \mathbf{S}^n$ equal to the objective in (23.11), that is, the restrictions

$$\overline{\lambda} = 1, \quad V + V^\top = I_n.$$

Then, the dual problem is given by

$$\mathrm{Opt}(D) = \max_{\substack{\overline{\lambda}\in\mathbf{R},\\ U,W\in\mathbf{S}^n,V\in\mathbf{R}^{n\times n}}} \left\{ -(\mathrm{Tr}(BU)+\mathrm{Tr}(W)) : \begin{array}{l} \overline{\lambda}=1, \\ V+V^\top = I_n \\ \begin{bmatrix} U & V \\ V^\top & W \end{bmatrix} \succeq 0 \end{array} \right\}. \tag{23.12}$$

Thus, the dual is precisely the problem of maximizing under the outlined restriction the right hand side of the aggregated inequality.

Since problem (23.11) satisfies the Slater condition (as $B \succ 0$) and is below bounded (why?), the dual problem is solvable and $\mathrm{Opt}(P) = \mathrm{Opt}(D)$. Moreover, the dual problem also satisfies the Relaxed Slater condition (why?), so that both the primal and the dual problems are solvable.

# 24

# Optimality Conditions in Convex Programming

Using our results on convex optimization duality, we next derive optimality conditions for convex programs.

## 24.1 Saddle point form of optimality conditions

**Theorem** IV.24.1 [Saddle Point formulation of Optimality Conditions in Convex Programming]
Consider optimization problem

$$\begin{array}{l}\min_x \{f(x) : \ g_j(x) \leq 0,\, j = 1, \ldots, m,\ x \in X\} \\ \text{[where } f, g_1, \ldots, g_m \text{ are real-valued functions on } X]\,,\end{array} \tag{IC}$$

along with its Lagrange dual problem

$$\begin{array}{l}\max_\lambda \{\underline{L}(\lambda) : \ \lambda \in \mathbf{R}^m_+\} \\ \left[\text{where } \underline{L}(\lambda) = \inf\limits_{x \in X} L(x, \lambda) \text{ and } L(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i g_i(x)\right].\end{array} \tag{IC$^*$}$$

Let $x^* \in X$. Then,

(i) A *sufficient* condition for $x^*$ to be an optimal solution to (IC) is the existence of the vector of Lagrange multipliers $\lambda^* \geq 0$ such that $(x^*, \lambda^*)$ is a *saddle point* of the Lagrange function $L(x, \lambda)$, i.e., a point where $L(x, \lambda)$ attains its minimum as a function of $x \in X$ and attains its maximum as a function of $\lambda \geq 0$:

$$L(x, \lambda^*) \geq L(x^*, \lambda^*) \geq L(x^*, \lambda) \quad \forall x \in X,\ \forall \lambda \geq 0. \tag{24.1}$$

(ii) Furthermore, if the problem (IC) *is convex* and *satisfies the Relaxed Slater condition*, then the above condition is *necessary* for optimality of $x^*$: if $x^*$ is optimal for (IC), then there exists $\lambda^* \geq 0$ such that $(x^*, \lambda^*)$ is a saddle point of the Lagrange function.

**Proof.** (i): Assume that for a given $x^* \in X$ there exists $\lambda^* \geq 0$ such that (24.1) is satisfied, and let us prove that then $x^*$ is optimal for (IC). First, we claim that $x^*$ is feasible. Assume for contradiction that $g_j(x^*) > 0$ for some $j$. Then, of course, $\sup\limits_{\lambda \geq 0} L(x^*, \lambda) = +\infty$ (look what happens when all $\lambda$'s, except $\lambda_j$, are fixed, and $\lambda_j \to +\infty$). But, $\sup\limits_{\lambda\geq 0} L(x^*,\lambda) = +\infty$ is forbidden by the second inequality in (24.1). Thus, $x^*$ must be feasible to (IC).

Since $x^*$ is feasible, $\sup\limits_{\lambda\geq 0} L(x^*,\lambda) = f(x^*)$, and we conclude from the second inequality in (24.1) that $L(x^*,\lambda^*) = \sup\limits_{\lambda\geq 0} L(x^*,\lambda) = f(x^*)$. Finally, let us examine the first inequality in (24.1). This relation now reads

$$f(x) + \sum_{j=1}^{m} \lambda_j^* g_j(x) \geq f(x^*), \quad \forall x \in X.$$

Recall that for any $x$ feasible for (IC), we have $g_j(x) \leq 0$ for all $j$. Together with $\lambda^* \geq 0$, we then deduce that for any $x$ feasible to (IC), we have $f(x) \geq f(x) + \sum_{j=1}^{m} \lambda_j^* g_j(x)$. But, then the above relation immediately implies that $x^*$ is optimal. □

(ii): Assume that (IC) is a convex program, $x^*$ is its optimal solution and the problem satisfies the Relaxed Slater condition. We will prove that then there exists $\lambda^* \geq 0$ such that $(x^*,\lambda^*)$ is a saddle point of the Lagrange function, i.e., that (24.1) is satisfied. As we know from the Convex Programming Duality Theorem (Theorem IV.22.1.ii), the dual problem (IC$^*$) has a solution $\lambda^* \geq 0$ and the optimal value of the dual problem is equal to the optimal value of the primal one, i.e., to $f(x^*)$:

$$f(x^*) = \underline{L}(\lambda^*) \equiv \inf_{x\in X} L(x,\lambda^*). \tag{24.2}$$

Then, we immediately conclude that

$$\lambda_j^* > 0 \quad \Longrightarrow \quad g_j(x^*) = 0$$

(this is called *complementary slackness*: positive Lagrange multipliers can be associated only with active (satisfied at $x^*$ as equalities) constraints). Indeed, from (24.2) it for sure follows that

$$f(x^*) \leq L(x^*,\lambda^*) = f(x^*) + \sum_{j=1}^{m} \lambda_j^* g_j(x^*);$$

the terms in the summation expression in the right hand side are nonpositive (since $x^*$ is feasible for (IC) and $\lambda^* \geq 0$), and the sum itself is nonnegative due to our inequality. Note that this is possible if and only if all the terms in the summation expression are zero, and this is precisely the complementary slackness.

From the complementary slackness we immediately conclude that $f(x^*) = L(x^*,\lambda^*)$, so that (24.2) results in

$$L(x^*,\lambda^*) = f(x^*) = \inf_{x\in X} L(x,\lambda^*).$$

On the other hand, since $x^*$ is feasible for (IC), from the definition of the Lagrangian function we deduce that $L(x^*,\lambda) \leq f(x^*)$ whenever $\lambda \geq 0$. Combining our observations, we conclude that

$$L(x^*,\lambda) \leq L(x^*,\lambda^*) \leq L(x,\lambda^*)$$

for all $x \in X$ and all $\lambda \geq 0$. □

Note that Theorem IV.24.1.i is valid for an arbitrary inequality constrained optimization problem, not necessarily a convex one. However, in the nonconvex case the *sufficient* condition for optimality given by Theorem IV.24.1.i is extremely far from being necessary and is "almost never" satisfied. In contrast to this, in the convex case the condition in question is not only sufficient, but also "nearly necessary" – it for sure is necessary when (IC) is a convex program satisfying the Relaxed Slater condition.

Theorem IV.24.1 presents Saddle Point form of optimality conditions for convex problems in the standard Mathematical Programming form (that is, with constraints represented by scalar convex inequalities). Similar results can be obtained for convex cone-constrained problems as follows.

**Theorem** IV.24.2 [Saddle Point formulation of Optimality Conditions in Convex Cone-constrained Programming] Consider a convex cone-constrained problem

$$\mathrm{Opt}(P) = \min_{x\in X} \left\{ f(x) : \ \overline{g}(x) := Ax - b \leq 0, \ \widehat{g}(x) \leq_{\mathbf{K}} 0 \right\}, \qquad (P)$$

($X$ is convex, $f : X \to \mathbf{R}$ is convex, $\overline{g}(\cdot) : \mathbf{R}^n \to \mathbf{R}^k$ is affine, $\mathbf{K}$ is a regular cone, and $\widehat{g}$ is $\mathbf{K}$-convex on $X$) along with its Cone-constrained Lagrange dual problem

$$\mathrm{Opt}(D) = \max_{\lambda:=[\overline{\lambda};\widehat{\lambda}]} \left\{ \underline{L}(\lambda) := \inf_{x\in X} \left[ f(x) + \overline{\lambda}^\top \overline{g}(x) + \widehat{\lambda}^\top \widehat{g}(x) \right] : \ \overline{\lambda} \geq 0, \ \widehat{\lambda} \in \mathbf{K}_* \right\}. \qquad (D)$$

Suppose that $(P)$ is bounded and satisfies Relaxed Slater condition. Then, a point $x^* \in X$ is an optimal solution to $(P)$ if and only if $x^*$ can be augmented by properly selected $\lambda^* \in \Lambda := \mathbf{R}^k_+ \times [\mathbf{K}_*]$ to be a saddle point of the cone-constrained Lagrange function

$$L(x; [\overline{\lambda}; \widehat{\lambda}]) := f(x) + \overline{\lambda}^\top \overline{g}(x) + \widehat{\lambda}^\top \widehat{g}(x)$$

on $X \times \Lambda$.

**Proof.** The proof basically repeats the one of Theorem IV.24.1. In one direction: assume that $x^* \in X$ can be augmented by $\lambda^* = [\overline{\lambda}^*; \widehat{\lambda}^*] \in \Lambda$ to form a saddle point of $L(x;\lambda)$ on $X \times \Lambda$, and let us prove that $x^*$ is an optimal solution to $(P)$. Observe, first, that from

$$L(x^*; \lambda^*) = \sup_{\lambda\in\Lambda} L(x^*; [\overline{\lambda}; \widehat{\lambda}]) = f(x^*) + \sup_{\lambda\in\Lambda} \left[ \overline{\lambda}^\top \overline{g}(x^*) + \widehat{\lambda}^\top \widehat{g}(x^*) \right] \qquad (24.3)$$

it follows that the linear form $\widehat{\lambda}^\top \widehat{g}(x^*)$ of $\widehat{\lambda}$ is bounded from above on the cone $\mathbf{K}_*$, implying that $-\widehat{g}(x^*) \in [\mathbf{K}_*]_* = \mathbf{K}$. Similarly, (24.3) says that the linear form $\overline{\lambda}^\top \overline{g}(x^*)$ of $\overline{\lambda}$ is bounded from above on the cone $\mathbf{R}^k_+$, implying that $-\overline{g}(x^*)$ belongs to the dual of this cone, that is, to $\mathbf{R}^k_+$. Thus, $x^*$ is feasible for $(P)$. As $x^*$ is feasible for $(P)$, the right hand side of the second equalty in (24.3) is $f(x^*)$, and thus (24.3) says that $L(x^*;\lambda^*) = f(x_*)$. With this in mind, the relation $L(x;\lambda^*) \geq L(x^*;\lambda^*)$ (which is satisfied for all $x \in X$, since $(x^*,\lambda^*)$ is a saddle point of $L$) reads $L(x;\lambda^*) \geq f(x^*)$. This combines with the relation $f(x) \geq L(x;\lambda^*)$ (which, due to $\lambda^* \in \Lambda$, holds true for all $x$ feasible for $(P)$) to imply that $\mathrm{Opt}(P) \geq f(x^*)$. The bottom line is that $x^*$ is a feasible solution to $(P)$ satisfying $\mathrm{Opt}(P) \geq f(x^*)$, thus, $x^*$ is an optimal solution to $(P)$, as claimed.

In the opposite direction: let $x^*$ be an optimal solution to $(P)$, and let us verify that $x^*$ is the first component of a saddle point of $L(x;\lambda)$ on $X \times \Lambda$. Indeed, $(P)$ is convex essentially strictly feasible cone-constrained problem; being solvable, it is below bounded. Applying Convex Programming Duality Theorem in cone-constrained form (Theorem IV.23.1), the dual problem $(D)$ is solvable with optimal value $\mathrm{Opt}(D) = \mathrm{Opt}(P)$. Denoting by $\lambda^* = [\overline{\lambda}^*; \widehat{\lambda}^*]$ an optimal solution to $(D)$, we have

$$f(x^*) = \mathrm{Opt}(P) = \mathrm{Opt}(D) = \underline{L}(\lambda^*) = \inf_{x\in X} L(x;\lambda^*), \tag{24.4}$$

whence $f(x^*) \leq L(x^*;\lambda^*) = f(x^*) + [\overline{\lambda}^*]^\top \overline{g}(x^*) + [\widehat{\lambda}^*]^\top \widehat{g}(x^*)$, that is, $[\overline{\lambda}^*]^\top \overline{g}(x^*) + [\widehat{\lambda}^*]^\top \widehat{g}(x^*) \geq 0$. Both terms in the latter sum are nonpositive (as $x^*$ is feasible for $(P)$ and $\lambda^* \in \Lambda$), while their sum is nonnegative, so that $[\overline{\lambda}^*]^\top \overline{g}(x^*) = 0$ and $[\widehat{\lambda}^*]^\top \widehat{g}(x^*) = 0$. We conclude that the inequality $f(x^*) \leq L(x^*;\lambda^*)$ is in fact equality, so that (24.4) reads $\inf_{x\in X} L(x;\lambda^*) = L(x^*;\lambda^*)$. Next, $L(x^*;\lambda) \leq f(x^*)$ for $\lambda \in \Lambda$ due to feasibility of $x^*$ for $(P)$, which combines with the already proved equality $L(x^*;\lambda^*) = f(x^*)$ to imply that $\sup_{\lambda\in\Lambda} L(x^*;\lambda) = L(x^*;\lambda^*)$. Thus, $(x^*,\lambda^*)$ is the desired saddle point of $L$. □

## 24.2 Karush-Kuhn-Tucker form of optimality conditions

Theorem IV.24.1 provides, basically, the strongest known optimality conditions for a Convex Programming problem. These conditions, however, are "implicit" – they are expressed in terms of saddle point of the Lagrange function, and it is unclear how to verify whether a given solution is a saddle point of the Lagrange function. Fortunately, the proof of Theorem IV.24.1 yields more or less explicit optimality conditions.

Recall that the normal cone $N_X(x)$ of a set $X \subseteq \mathbf{R}^n$ at a point $x \in X$ as defined by (15.5) is

$$N_X(x) = \left\{h \in \mathbf{R}^n : \; h^\top(x'-x) \leq 0, \, \forall x' \in X\right\}.$$

Now let us define the notion of *Karush-Kuhn-Tucker point* of inequality constrained optimization problem.

**Definition** IV.24.3 [Karush-Kuhn-Tucker point of inequality constrained

Mathematical Programming problem] Consider optimization problem

$$\begin{array}{l}\min_x \{f(x) :\ g_j(x) \leq 0,\, j = 1,\dots,m,\ x \in X\}\\ \text{[where } f, g_1,\dots,g_m \text{ are real-valued functions on } X\text{]}\,,\end{array} \qquad \text{(IC)}$$

A point $x^* \in \mathbf{R}^n$ is called *Karush-Kuhn-Tucker (KKT) point* of (IC), if $x^*$ is feasible, $f$, $g_1$,…,$g_m$ are differentiable at $x^*$, and there exist *nonnegative* Lagrange multipliers $\lambda_j^*$, $j = 1,\dots,m$, such that

$$\lambda_j^* g_j(x^*) = 0,\ j = 1,\dots,m, \qquad \text{[complementary slackness]} \tag{24.5}$$

$$\text{and}\quad \nabla f(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla g_j(x^*) \in -N_X(x^*), \qquad \text{[KKT equation]} \tag{24.6}$$

where $N_X(x^*)$ is the normal cone of $X$ at $x^*$.

We are now ready to state "more explicit" optimality conditions for convex programs based on KKT points.

**Theorem** IV.24.4 [Karush-Kuhn-Tucker Optimality Conditions in Convex Programming] Let (IC) be a convex program (i.e., $X$ is nonempty and convex, and $f, g_1,\dots,g_m$ are convex on $X$). Let $x^* \in X$, and let the functions $f$, $g_1$,…,$g_m$ be differentiable at $x^*$. Then,

(i) [Sufficiency] If $x^*$ is a KKT point of (IC), then $x^*$ is an optimal solution to the problem.

(ii) [Necessity and sufficiency] If, in addition to the premise, the Relaxed Slater condition holds, $x^*$ is an optimal solution to (IC) if and only if $x^*$ is a KKT point of the problem.

**Proof.** (i): Suppose $x^*$ is a KKT point of problem (IC), and let us prove that $x^*$ is an optimal solution to (IC). By Theorem IV.24.1, it suffices to demonstrate that augmenting $x^*$ by properly selected $\lambda \geq 0$, we get a saddle point $(x^*, \lambda)$ of the Lagrange function on $X \times \mathbf{R}^m_+$. Let $\lambda^*$ be the Lagrange multipliers associated with $x^*$ according to the definition of a KKT point. We claim that $(x^*, \lambda^*)$ is a saddle point of the Lagrange function. Indeed, complementary slackness says that $L(x^*, \lambda^*) = f(x^*)$, while due to feasibility of $x^*$ we have $\sup_{\lambda \geq 0} \sum_{j=1}^{n} \lambda_j g_j(x^*) = 0$. Taken together, these observations say that $L(x^*, \lambda^*) = \sup_{\lambda \geq 0} L(x^*, \lambda)$. Moreover, the function $\phi(x) := L(x, \lambda^*) : X \to \mathbf{R}$ is convex and differentiable at $x^*$, and by the KKT equation we have $\nabla \phi(x^*) \in -N_X(x^*)$. Invoking Proposition III.15.3, we conclude that $x^*$ minimizes $\phi$ on $X$, that is, $L(x, \lambda^*) \geq L(x^*, \lambda^*)$ for all $x \in X$. Thus, $(x^*, \lambda^*)$ is a saddle point of the Lagrange function.

(ii): In view of (i), all we need to prove (ii) is to demonstrate that if $x^*$ is an optimal solution to (IC), (IC) is a convex program that satisfies the Relaxed Slater condition, and the objective and constraints of (IC) are differentiable at

$x^*$, then $x^*$ is a KKT point. Indeed, let $x^*$ and (IC) satisfy the above "if." By Theorem IV.24.1.ii, $x^*$ can be augmented by some $\lambda^* \geq 0$ to yield a saddle point $(x^*, \lambda^*)$ of $L(x, \lambda)$ on $X \times \mathbf{R}^m_+$. Then, the saddle point inequalities (24.1) give us

$$f(x^*) + \sum_{j=1}^m \lambda_j^* g_j(x^*) = L(x^*, \lambda^*) \geq \sup_{\lambda \geq 0} L(x^*, \lambda) = f(x^*) + \sup_{\lambda \geq 0} \sum_{j=1}^m \lambda_j g_j(x^*). \tag{24.7}$$

Moreover, as $x^*$ is feasible to (IC), we have $g_j(x^*) \leq 0$ for all $j$, whence

$$\sup\nolimits_{\lambda \geq 0} \sum\nolimits_j \lambda_i g_j(x^*) = 0.$$

Therefore (24.7) implies that $\sum_j \lambda_j^* g_j(x^*) \geq 0$. This inequality, in view of $\lambda_j^* \geq 0$ and $g_j(x^*) \leq 0$ for all $j$, implies that $\lambda_j^* g_j(x^*) = 0$ for all $j$, i.e., complementary slackness (24.5) condition holds. The relation $L(x, \lambda^*) \geq L(x^*, \lambda^*)$ for all $x \in X$, implies that the function $\phi(x) := L(x, \lambda^*)$ attains its minimum on $X$ at $x = x^*$. Note also that $\phi(x)$ is convex on $X$ and differentiable at $x^*$, thus, by Proposition III.15.3, we deduce that the KKT equation (24.6) also holds. □

Note that the optimality conditions stated in Theorem III.15.2 and Proposition III.15.3 are particular cases of Theorem IV.24.4 corresponding to $m = 0$.

**Remark** IV.24.5 A standard special case of Theorem IV.24.4 that is worth discussing explicitly is when $x^*$ is in the (relative) interior of $X$.

When $x^* \in \operatorname{int} X$, we have $N_X(x^*) = \{0\}$, so that (24.6) reads

$$\nabla f(x^*) + \sum_{j=1}^m \lambda_j^* \nabla g_j(x^*) = 0.$$

When $x^* \in \operatorname{rint} X$, $N_X(x^*)$ is the orthogonal complement to the linear subspace $\mathcal{L}$ to which $\operatorname{Aff}(X)$ is parallel, so that (24.6) reads

$$\nabla f(x^*) + \sum_{j=1}^m \lambda_j^* \nabla g_j(x^*) \text{ is orthogonal to } \mathcal{L} := \operatorname{Lin}(X - x_*).$$

## 24.3 Cone-constrained KKT optimality conditions

The cone-constrained version of the notion of a KKT point is defined as follows:

**Definition** IV.24.6 [Karush-Kuhn-Tucker point of a cone-constrained problem] Consider cone-constrained optimization problem

$$\begin{array}{c} \min \{f(x) : \overline{g}(x) := Ax - b \leq 0,\ \widehat{g}(x) \leq_{\mathbf{K}} 0,\ x \in X\} \\ \left[\begin{array}{l} \text{where } \ X \subseteq \mathbf{R}^n,\ f : X \to \mathbf{R},\ \overline{g} : \mathbf{R}^n \to \mathbf{R}^k, \\ \qquad\quad \widehat{g} : X \to \mathbf{R}^\nu,\ \mathbf{K} \subset \mathbf{R}^\nu \text{ is a regular cone.} \end{array}\right] \end{array} \tag{ConeC}$$

A point $x^* \in \mathbf{R}^n$ is called *Karush-Kuhn-Tucker (KKT) point* of (ConeC), if $x^*$ is feasible, $f$ and $\widehat{g}$ are differentiable at $x^*$, and there exist Lagrange multipliers $\overline{\lambda} = [\overline{\lambda}_1; \dots; \overline{\lambda}_k] \geq 0$ and $\widehat{\lambda} \in \mathbf{K}_*$ such that

$$\overline{\lambda}_j[\overline{g}(x^*)]_j = 0, \forall j \leq k, \ \& \ \widehat{\lambda}^\top \widehat{g}(x^*) = 0, \quad \text{[complementary slackness]} \quad (24.8)$$

$$\text{and } \nabla_x \left[f(x) + \overline{\lambda}^\top \overline{g}(x) + \widehat{\lambda}^\top \widehat{g}(x)\right]\Big|_{x=x^*} \in -N_X(x^*), \quad \text{[KKT equation]} \quad (24.9)$$

where $N_X(x^*)$ is the normal cone of $X$ at $x^*$, see (15.5).

Based on this definition, cone-constrained version of Theorem IV.24.4 is as follows.

**Theorem** IV.24.7 [Karush-Kuhn-Tucker Optimality Conditions in Cone-constrained Convex Programming] Consider a convex cone-constrained problem

$$\text{Opt}(P) = \min_{x \in X} \{f(x) : \ \overline{g}(x) := Ax - b \leq 0, \ \widehat{g}(x) \leq_{\mathbf{K}} 0\} \qquad (P)$$

($X$ is convex, $f : X \to \mathbf{R}$ is convex, $A \in \mathbf{R}^{k \times n}$, $\mathbf{K}$ is a regular cone, and $\widehat{g}$ is $\mathbf{K}$-convex on $X$). Suppose $x^* \in X$ is a feasible solution to the problem, and let $f$ and $\widehat{g}$ be differentiable at $x^*$.

(i) If $x^*$ is a KKT point (as defined by Definition IV.24.6) of $(P)$, then $x^*$ is an optimal solution to $(P)$.

(ii) If $x^*$ is optimal solution to $(P)$ and, if addition to the above premise, $(P)$ satisfies the cone-constrained Relaxed Slater condition, then $x^*$ is a KKT point, as defined by Definition IV.24.6, of $(P)$

The proof of this theorem follows verbatim by the proof of Theorem IV.24.4, with Theorem IV.24.2 in the role of Theorem IV.24.1.

**Application: Optimal value in parametric convex cone-constrained problem.** What follows is a far-reaching extension of subgradient interpretation of Lagrange multipliers presented in section 23.3.A. Consider a parametric family of convex cone-constrained problems defined by a parameter $p \in P$

$$\text{Opt}(p) := \min_{x \in X} \{f(x, p) : \ g(x, p) \leq_{\mathbf{M}} 0\}, \qquad (\text{P}[p])$$

where

(a) $X \subseteq \mathbf{R}^n$ and $P \subseteq \mathbf{R}^\mu$ are nonempty convex sets,
(b) $\mathbf{M}$ is a regular cone in some $\mathbf{R}^\nu$.
(c) $f : X \times P \to \mathbf{R}$ is convex, and $g : X \times P \to \mathbf{R}^\nu$ is $\mathbf{M}$-convex. [1]

Next, we make the following assumption:

[1] In what follows, splitting the constraint in a cone-constrained problem into a system of scalar linear inequalities and a conic inequality does not play any role, and in order to save notation, (P[$p$]) uses "single-constraint" format of cone-constrained problem. Of course, the two-constraint format $\overline{g}(x) := Ax - b \leq 0$, $\widehat{g}(x) \leq_{\mathbf{K}} 0$ reduces to the single-constraint one by setting $g(x) = [\overline{g}(x); \widehat{g}(x)]$ and $\mathbf{M} = \mathbf{R}^k_+ \times \mathbf{K}$.

(d) Suppose $\overline{x} \in X$ and $\overline{p} \in P$ are such that

- $\overline{x}$ is a KKT point, as defined by Definition IV.24.6, of convex cone-constrained problem (P[$\overline{p}$]) (and therefore, by Theorem IV.24.7, is an optimal solution to the problem), and
- $f(x,p)$ and $g(x,p)$ are differentiable at the point $[\overline{x};\overline{p}]$, the derivatives being

$$Df([\overline{x};\overline{p}])[[dx;dp]] = F_x^\top dx + F_p^\top dp,\ Dg([\overline{x};\overline{p}])[[dx;dp]] = G_x dx + G_p dp.$$

and let $\overline{\mu} \in \mathbf{M}_*$ be the Lagrange multiplier associated with $\overline{x}$ and (P[$\overline{p}$]), so that

$$\overline{\mu}^\top g(\overline{x},\overline{p}) = 0 \ \text{ and } \ [x-\overline{x}]^\top [F_x + G_x^\top \overline{\mu}] \geq 0,\ \forall x \in X. \tag{24.10}$$

(cf. Definition IV.24.6).

**Proposition** IV.24.8 Under Assumptions (a–d), Opt$(\cdot)$ is a convex function on $P$ taking values in $\mathbf{R} \cup \{+\infty\}$ and finite at $\overline{p}$, and the vector

$$F_p + G_p^\top \overline{\mu}$$

is a subgradient of Opt$(\cdot)$ at $\overline{p}$:

$$\mathrm{Opt}(p) \geq \mathrm{Opt}(\overline{p}) + [p-\overline{p}]^\top [F_p + G_p^\top \overline{\mu}],\ \forall p \in P. \tag{24.11}$$

**Proof.** Observe, first, that as $f$ is convex by Gradient inequality we have

$$f(x,p) \geq f(\overline{x},\overline{p}) + F_x^\top [x-\overline{x}] + F_p^\top [p-\overline{p}], \quad \forall (x \in X, p \in P). \tag{24.12}$$

Also, from $\overline{\mu} \in \mathbf{M}_*$ and $\mathbf{M}$-convexity of $g$ by Fact IV.21.10 it follows that the function $\overline{\mu}^\top g(x,p)$ is convex on $X \times P$, so that by Gradient inequality we have

$$\overline{\mu}^\top g(x,p) \geq \underbrace{\overline{\mu}^\top g(\overline{x},\overline{p})}_{=0 \text{ by } (24.10)} + \overline{\mu}^\top G_x[x-\overline{x}] + \overline{\mu}^\top G_p[p-\overline{p}], \quad \forall (x \in X, p \in P). \tag{24.13}$$

Now let $p \in P$ and $x$ be feasible for (P[$p$]). Then,

$$\begin{aligned} f(x,p) &\geq f(x,p) + \overline{\mu}^\top g(x,p) \\ &\geq f(\overline{x},\overline{p}) + F_x^\top [x-\overline{x}] + F_p^\top [p-\overline{p}] + \overline{\mu}^\top G_x[x-\overline{x}] + \overline{\mu}^\top G_p[p-\overline{p}] \\ &= \mathrm{Opt}(\overline{p}) + (F_x + G_x^\top \overline{\mu})^\top [x-\overline{x}] + (F_p + G_p^\top \overline{\mu})^\top [p-\overline{p}] \\ &\geq \mathrm{Opt}(\overline{p}) + (F_p + G_p^\top \overline{\mu})^\top [p-\overline{p}], \end{aligned}$$

where the first inequality follows from $\overline{\mu} \in \mathbf{M}_*$ and $g(x,p) \leq_{\mathbf{M}} 0$, the second is due to (24.12) and (24.13), the equality holds by recalling that $\overline{x}$ is optimal for $(P[\overline{p}])$, and the last inequality is due to (24.10). As the resulting inequality holds true for all $x$ feasible for (P[$p$]), it justifies (24.11).

To complete the proof, we need to verify the convexity of Opt$(\cdot)$. By the relation in (24.11), Opt$(p)$ for $p \in P$ is either a real, or $+\infty$, as is required for a convex function. For any $p', p'' \in P \cap \mathrm{Dom}(\mathrm{Opt}(\cdot))$ and $\lambda \in [0,1]$, we need to check

$$\mathrm{Opt}(\lambda p' + (1-\lambda)p'') \leq \lambda \mathrm{Opt}(p') + (1-\lambda)\mathrm{Opt}(p''). \tag{24.14}$$

This is immediate (cf. proof of Fact IV.23.2): given $\epsilon > 0$, we can find $x', x'' \in X$ such that

$$g(x',p') \leq_{\mathbf{M}} 0,\ g(x'',p'') \leq_{\mathbf{M}} 0,\ f(x',p') \leq \mathrm{Opt}(p') + \epsilon,\ f(x'',p'') \leq \mathrm{Opt}(p'') + \epsilon.$$

Setting $p := \lambda p' + (1-\lambda)p''$, $x := \lambda x' + (1-\lambda)x''$ and invoking convexity of $f$ and $\mathbf{M}$-convexity of $g$, we get

$$g(x,p) \leq_{\mathbf{M}} 0,\ f(x) \leq [\lambda\mathrm{Opt}(p') + (1-\lambda)\mathrm{Opt}(p'')] + \epsilon.$$

Finally, since $\epsilon > 0$ is arbitrary, we arrive at (24.14). □

Note that the result of section 23.3.A is nothing but what Proposition IV.24.8 states in the case of $f$ independent of $p$, $\mathbf{M} = \mathbf{R}^k_+ \times \mathbf{K}$, $p = [\overline{\delta}, \widehat{\delta}] \in \mathbf{R}^k \times \mathbf{R}^\nu$, and $g(x,p) = [\overline{g}(x) - \overline{\delta}; \widehat{g}(x) - \widehat{\delta}]$.

## 24.4 Optimality conditions in Conic Programming

We continue by discussing the case of conic programming.

**Theorem** IV.24.9 [Optimality Conditions in Conic Programming] Consider a primal-dual pair of conic problems (cf. section 23.4)

$$\mathrm{Opt}(P) = \min_{x\in\mathbf{R}^n} \left\{c^\top x:\ Ax - b \leq 0,\ Px - p \leq_{\mathbf{K}} 0\right\} \tag{P}$$

$$\mathrm{Opt}(D) = \max_{\overline{\lambda},\widehat{\lambda}} \left\{-b^\top\overline{\lambda} - p^\top\widehat{\lambda} : A^\top\overline{\lambda} + P^\top\widehat{\lambda} + c = 0,\ \overline{\lambda} \geq 0,\ \widehat{\lambda} \in \mathbf{K}_*\right\}. \tag{D}$$

Suppose that both problems satisfy Relaxed Slater condition. Then, a pair of feasible solutions $x_*$ to $(P)$ and $\lambda_* := [\overline{\lambda}_*; \widehat{\lambda}_*]$ to $(D)$ is optimal to the respective problems if and only if

$$\mathrm{DualityGap}(x_*;\lambda_*) := c^\top x_* - [-b^\top\overline{\lambda}_* - p^\top\widehat{\lambda}_*] = 0, \qquad \text{[Zero Duality Gap]}$$

which holds if and only if

$$\overline{\lambda}_*^\top[b - Ax_*] + \widehat{\lambda}_*^\top[p - Px_*] = 0. \qquad \text{[Complementary Slackness]}$$

**Remark** IV.24.10 Under the premise of Theorem IV.24.9, from the feasibility of $x_*$ and $\lambda_*$ for their respective problems it follows that $b - Ax_* \geq 0$ and $p - Px_* \in \mathbf{K}$ and $\overline{\lambda}_* \geq 0$ and $\widehat{\lambda}_* \in \mathbf{K}_*$. Therefore, Complementary slackness (which says that the sum of two inner products, every one of a vector from a regular cone and a vector from the dual of this cone, and as such automatically nonnegative) is zero is a really strong restriction. This comment is applicable to relation $[\widehat{\lambda}_*]^\top\widehat{g}(x^*) = 0$ in (24.8 ).

**Proof of Theorem IV.24.9.** By Conic Duality Theorem (Theorem IV.23.6) we are in the case when $\mathrm{Opt}(P) = \mathrm{Opt}(D) \in \mathbf{R}$, and therefore for any $x$ and $\lambda := [\overline{\lambda}; \widehat{\lambda}]$, we have

$$\mathrm{DualityGap}(x;\lambda) = [c^\top x - \mathrm{Opt}(P)] + [\mathrm{Opt}(D) - (-b^\top\overline{\lambda} - p^\top\widehat{\lambda})].$$

Now, when $x$ is feasible for $(P)$, the *primal optimality gap* $c^\top x - \mathrm{Opt}(P)$ is nonnegative and is zero iff $x$ is optimal for $(P)$. Similarly, when $\lambda = [\overline{\lambda}; \widehat{\lambda}]$ is feasible for $(D)$, the *dual optimality gap* $\mathrm{Opt}(D) - (-b^\top \overline{\lambda} - p^\top \widehat{\lambda})$ is nonnegative and is zero iff $\lambda$ is optimal for $(D)$. We conclude that whenever $x$ is feasible for $(P)$, and $\lambda$ is feasible for $(D)$. the duality gap $\mathrm{DualityGap}(x; \lambda)$ (which, as we have seen, is the sum of the corresponding optimality gaps) is nonnegative and is zero iff both these optimality gaps are zero, that is, iff $x$ is optimal for $(P)$, and $\lambda$ is optimal for $(D)$.

It remains to note that Complementary Slackness condition is equivalent to Zero Duality Gap one. To this end, note that since $x_*$ and $\lambda_*$ are feasible for their respective problems, we have

$$\begin{aligned}\mathrm{DualityGap}(x_*; \lambda_*) &= c^\top x_* + b^\top \overline{\lambda}_* + p^\top \widehat{\lambda}_* \\ &= -[A^\top \overline{\lambda}_* + P^\top \widehat{\lambda^*}]^\top x_* + b^\top \overline{\lambda}_* + p^\top \widehat{\lambda}_* \\ &= \overline{\lambda}_*^\top [b - Ax_*] + \widehat{\lambda}_*^\top [p - Px_*].\end{aligned}$$

Therefore, Complementary Slackness, for the solutions $x_*$ and $\lambda_*$ that are feasible for the respective problems, is exactly the same as Zero Duality Gap. □

**Example** IV.24.1 (continued from Example IV.23.1) Consider the primal-dual pair of conic problems (23.11) and (23.12). We claim that the primal solution $y = -B^{1/2}$, $t = -\mathrm{Tr}(B^{1/2})$ and the dual solution $\overline{\lambda} = 1, U = \frac{1}{2}B^{-1/2}, V = \frac{1}{2}I_n$, $W = \frac{1}{2}B^{1/2}$ are optimal for the respective problems. Indeed, it is immediately seen that these solutions are feasible for the respective problems (to check feasibility of the dual solution, use Schur Complement Lemma). Moreover, the objective value of the primal solution equals to the objective value of the dual solution, and both these quantities are equal to $-\mathrm{Tr}(B^{1/2})$. Thus, the zero duality gap indeed holds true.

# 25

# Duality in Linear and Convex Quadratic Programming

The fundamental role of the Lagrange function and Lagrange Duality in Optimization is clear already from the Optimality Conditions given by Theorem IV.24.1, but this role is not restricted by this theorem only. There are several cases when we can explicitly write down the Lagrange dual, and whenever it is the case, we get a pair of explicitly formulated and closely related to each other optimization programs – the *primal-dual pair*; analyzing the problems simultaneously, we get more information about their properties (and get a possibility to solve the problems numerically in a more efficient way) than it is possible when we restrict ourselves with only one problem of the pair. The detailed investigation of Duality in "well-structured" Convex Programming deals with cone-constrained Lagrange duality and conic problems. This being said, there are cases where already "plain" Lagrange duality is quite appropriate. Let us look at two of these particular cases.

## 25.1 Linear Programming Duality

Let us start with some general observation. Note that the Karush-Kuhn-Tucker condition under the assumption of Theorem IV.24.4 (i.e., problem

$$\min_x \{f(x) : g_(x) \leq 0, j = 1, ..., m, x \in X\} \tag{IC}$$

is convex, $x^*$ is a feasible solutio to the problem, and $f, g_1, \ldots, g_m$ are differentiable at $x^*$) is exactly the condition that $(x^*, \lambda^* := (\lambda_1^*, \ldots, \lambda_m^*))$ is a saddle point of the Lagrange function

$$L(x, \lambda) = f(x) + \sum_{j=1}^m \lambda_j g_j(x) \tag{25.1}$$

on $X \times \mathbf{R}_+^m$: equalities (24.5) taken together with feasibility of $x^*$ state that $L(x^*, \lambda)$ attains its maximum in $\lambda \geq 0$ at $\lambda^*$, and (24.6) states that when $\lambda$ is fixed at $\lambda^*$ the function $L(x, \lambda^*)$ attains its minimum in $x \in X$ at $x = x^*$.

Now consider the particular case of (IC) where $X = \mathbf{R}^n$ is the entire space, the objective $f$ is convex and everywhere differentiable and the constraints $g_1, \ldots, g_m$ are *linear*. In this case the Relaxed Slater Condition holds whenever there is a feasible solution to (IC), and when that is the case, Theorem IV.24.4 states that the KKT (Karush-Kuhn-Tucker) condition is necessary and sufficient for optimality of $x^*$; as we just have explained, this is the same as to say that the necessary and sufficient condition of optimality for $x^*$ is that $x^*$ along with certain $\lambda^* \geq 0$ form a saddle point of the Lagrange function. Combining these observations with Proposition IV.22.2, we get the following simple result.

**Proposition** IV.25.1 Let (IC) be a convex program with $X = \mathbf{R}^n$, everywhere differentiable objective $f$ and linear constraints $g_1, \ldots, g_m$.

Then, $x^*$ is an optimal solution to (IC) if and only if there exists $\lambda^* \geq 0$ such that $(x^*, \lambda^*)$ is a saddle point of the Lagrange function (25.1) (regarded as a function of $x \in \mathbf{R}^n$ and $\lambda \geq 0$). In particular, (IC) is solvable if and only if $L$ has saddle points, and if it is the case, then both (IC) and its Lagrange dual

$$\max_{\lambda \geq 0} \left\{ \underline{L}(\lambda) := \inf_x L(x, \lambda) \right\} \tag{IC$^*$}$$

are solvable with equal optimal objective values.

Let us look what Proposition IV.25.1 says in the Linear Programming case, i.e., when (IC) is the problem given by

$$\min_x \left\{ f(x) := c^T x : g_j(x) := b_j - a_j^T x \leq 0, \, j = 1, \ldots, m \right\} \tag{$P$}$$

In order to get to the Lagrange dual, we should form the Lagrange function of (IC) given by

$$L(x, \lambda) = f(x) + \sum_{j=1}^m \lambda_j g_j(x) = \left( c - \sum\nolimits_{j=1}^m \lambda_j a_j \right)^\top x + \sum_{j=1}^m \lambda_j b_j,$$

and minimize it in $x \in \mathbf{R}^n$; this will give us the dual objective. In our case the minimization in $x$ is immediate: the minimal value is equal to $-\infty$, if $c - \sum_{j=1}^m \lambda_j a_j \neq 0$, and it is $\sum_{j=1}^m \lambda_j b_j$, otherwise. Hence, we see that the Lagrange dual is given by

$$\max_\lambda \left\{ b^\top \lambda : \sum_{j=1}^m \lambda_j a_j = c, \, \lambda \geq 0 \right\}. \tag{$D$}$$

Therefore, the Lagrange dual problem is precisely the usual LP dual to $(P)$, and Proposition IV.25.1 is one of the equivalent forms of the Linear Programming Duality Theorem (Theorem I.4.9) which we already know.

## 25.2 Quadratic Programming Duality

Now consider the case when the original problem is linearly constrained convex quadratic program given by

$$\min_x \left\{ f(x) := \frac{1}{2} x^\top Q x + c^\top x : g_j(x) := b_j - a_j^\top x \leq 0, \, j = 1, \ldots, m \right\} \tag{$P$}$$

where the objective is a strictly convex quadratic form, so that the matrix $Q = Q^\top$ is positive definite, i.e., $x^\top Qx > 0$ whenever $x \neq 0$. It is convenient to rewrite the constraints in the vector-matrix form using the notation

$$g(x) = b - Ax \leq 0, \text{ where } b := [b_1; \ldots ; b_m], \ A := \begin{bmatrix} a_1^\top \\ \vdots \\ a_m^\top \end{bmatrix}.$$

In order to form the Lagrange dual to $(P)$ program, we write down the Lagrange function

$$\begin{aligned} L(x, \lambda) &= f(x) + \sum_{j=1}^{m} \lambda_j g_j(x) \\ &= \frac{1}{2}x^\top Qx + c^\top x + \lambda^\top (b - Ax) = \frac{1}{2}x^\top Qx - (A^\top \lambda - c)^\top x + b^\top \lambda \end{aligned}$$

and minimize it in $x$. Since the function is convex and differentiable in $x$, the minimum, if exists, is given by the Fermat rule

$$\nabla_x L(x, \lambda) = 0,$$

which in our situation becomes

$$Qx = A^\top \lambda - c.$$

Since $Q$ is positive definite, it is nonsingular, so that the Fermat equation has a unique solution which is the minimizer of $L(\cdot, \lambda)$. This solution is

$$x(\lambda) := Q^{-1}(A^\top \lambda - c).$$

Substituting the expression of $x(\lambda)$ into the expression for the Lagrange function, we get the dual objective

$$\underline{L}(\lambda) = -\frac{1}{2}(A^\top \lambda - c)^\top Q^{-1}(A^\top \lambda - c) + b^\top \lambda.$$

Thus, the dual problem is to maximize this objective over the nonnegative orthant. Let us rewrite this dual problem equivalently by introducing additional variables

$$t := -Q^{-1}(A^\top \lambda - c) \quad \Longrightarrow \quad (A^\top \lambda - c)^\top Q^{-1}(A^\top \lambda - c) = t^\top Qt.$$

With this substitution, the dual problem becomes

$$\max_{\lambda, t} \left\{ -\frac{1}{2}t^\top Qt + b^\top \lambda : \ A^\top \lambda + Qt = c, \ \lambda \geq 0 \right\}. \tag{D}$$

We see that the dual problem also turns out to be linearly constrained convex quadratic program.

**Remark** IV.25.2 Note also a feasible quadratic program in the form of $(P)$ with a positive definite matrix $Q$ automatically is solvable. This relies on the following simple general fact:

> Let (IC) be a feasible program with closed domain $X$, continuous on $X$ objective and constraints and such that $f(x) \to \infty$ as $x \in X$ "goes to infinity" (i.e., $\|x\|_2 \to \infty$). Then (IC) is solvable.

In the case of our quadratic program $(P)$, as $Q$ is positive definite, we have $f(x) \to \infty$ as $\|x\|_2 \to \infty$. Then, solvability of $(P)$ follows from the simple fact stated above. You are encouraged to prove this simple fact on your own.

Based on this remark, Proposition IV.25.1 leads to the following result.

**Theorem** IV.25.3 [Duality Theorem in Quadratic Programming] Let $(P)$ be a feasible quadratic program with positive definite symmetric matrix $Q$ in the objective. Then, both $(P)$ and $(D)$ are solvable, and the optimal values in these problems are equal to each other.

A pair of primal and dual *feasible* solutions, say $(x; (\lambda, t))$, to these problems are optimum

(i) if and only if *"zero duality gap"* optimality condition holds, i.e., the primal objective value at $x$ is equal to the dual objective value at $(\lambda, t)$
or equivalently,

(ii) if and only if the following holds

$$\lambda_i (Ax - b)_i = 0, \; i = 1, \ldots, m, \quad \text{and} \quad t = -x. \tag{25.2}$$

**Proof.** (i) Proposition IV.25.1 implies that the optimal value in minimization problem $(P)$ is equal to the optimal value in the maximization problem $(D)$. It follows that the value of the primal objective at any primal feasible solution is $\geq$ the value of the dual objective at any dual feasible solution, and equality is possible if and only if these values coincide with the optimal values in the problems, as claimed in (i).

(ii) Let $\Delta(x, (\lambda, t))$ be the difference between the primal objective value of the primal feasible solution $x$ and the dual objective value of the dual feasible solution $(\lambda, t)$

$$\begin{aligned} \Delta(x, (\lambda, t)) &:= (c^\top x + \frac{1}{2} x^\top Q x) - (b^\top \lambda - \frac{1}{2} t^\top Q t) \\ &= (A^\top \lambda + Qt)^\top x + \frac{1}{2} x^\top Q x + \frac{1}{2} t^\top Q t - b^\top \lambda \\ &= \lambda^\top (Ax - b) + \frac{1}{2} (x + t)^\top Q (x + t), \end{aligned}$$

where the second equation follows since $A^\top \lambda + Qt = c$. Whenever $x$ is primal feasible, we have $Ax - b \geq 0$, and similarly dual feasibility of $(\lambda, t)$ implies that $\lambda \geq 0$. Since $Q$ is positive definite as well, we then deduce that the first and the second terms in the above representation of $\Delta(x, (\lambda, t))$ are nonnegative for every pair $(x; (\lambda, t))$ of primal and dual feasible solutions. Thus, for such a pair $\Delta(x, (\lambda, t)) = 0$ holds if and only if $\lambda^\top (Ax - b) = 0$ and $(x + t)^\top Q (x + t) = 0$. The first of these equalities, due to $\lambda \geq 0$ and $Ax \geq b$, is equivalent to $\lambda_j (Ax - b)_j = 0$

for all $j = 1, \ldots, m$; the second equality, due to positive definiteness of $Q$, is equivalent to $x + t = 0$. □

# 26

# ★ Cone-convex functions: elementary calculus and examples

So far, speaking about **K**-convex functions, we assumed the cone $K$ to be regular, that is, pointed, closed, and with a nonempty interior. In the considerations of this chapter, nonemptiness of $\operatorname{int}\mathbf{K}$ is of no importance, so that within this chapter we specify **K**-convex mapping $f$, **K** being a closed pointed cone in some embedding Euclidean space $F$, as a mapping $f$ defined on a convex domain $\operatorname{Dom} f \subset \mathbf{R}^n$ and taking values in $F$ such that

$$\forall (x, y \in \operatorname{Dom} f, \lambda \in [0,1]) : f(\lambda x + (1-\lambda)y) \leq_{\mathbf{K}} \lambda f(x) + (1-\lambda) f(y),$$

where, as always, $a \leq_{\mathbf{K}} b$ means that $b - a \in \mathbf{K}$. In particular, we allow for $\mathbf{K} = \{0\}$, that is $a \leq_{\mathbf{K}} b$ equivalent to $a = b$; in this extreme case, **K**-convexity of $f$ means that $f$ is affine on $\operatorname{Dom} f$.

Note that Fact IV.21.10, as is clear from its proof, remains true when the regularity of cone **K** in its premise is relaxed to closedness and pointedness.

The calculus presented below resembles the usual calculus of real-valued monotone and convex functions, and almost all claims to follow are nearly self-evident, and therefore not all of them are accompanied by proofs. Absence of a proof in the text means that providing the proof is an exercise for the reader[1].

### 26.1 Epigraph characterization of cone-convexity

Let $f(x) : \operatorname{Dom} f \to \mathbf{R}^\nu$ be a mapping with convex domain $\operatorname{Dom} f \subseteq \mathbf{R}^n$, and let $\mathbf{K} \subset \mathbf{R}^\nu$ be a closed pointed cone. The following claim is evident.

**Fact** IV.26.1 A function $f$ is **K**-convex if and only if its **K**-epigraph

$$\operatorname{epi}_{\mathbf{K}}(f) := \{(x, y) \in \operatorname{Dom} f \times \mathbf{R}^\nu : \ y \geq_{\mathbf{K}} f(x)\}$$

is convex.

Let us examine some **K**-convex functions. In what follows, we use $\succeq$-convexity as synonym of $\mathbf{S}^m_+$-convexity, with context-specified $m$.

**Example** IV.26.1 Consider the "convex matrix-valued quadratic function" of $x \in \mathbf{R}^{p\times q}$ given by

$$f(x) := [AxB][AxB]^\top + [CxD + D^\top x^\top C^\top] + H,$$

[1] This being said, all missing proofs can be found in solutions to Exercise IV.29.

where $A, C \in \mathbf{R}^{m\times p}, B \in \mathbf{R}^{q\times n}, D \in \mathbf{R}^{q\times m}, H \in \mathbf{S}^m$. Note that $f : \mathbf{R}^{p\times q} \mapsto \mathbf{S}^m$. We claim that $f$ is $\succeq$-convex. Indeed, by the Schur Complement Lemma we have

$$\mathrm{epi}_{\mathbf{S}^m_+}\{f\} = \left\{(x,y) : \ \left[\begin{array}{c|c} y - (CxD + D^\top x^\top C^\top) - H & AxB \\ \hline B^\top x^\top A^\top & I_n \end{array}\right] \succeq 0\right\}$$

and the right hand side set is clearly convex (as it is the inverse image of $\mathbf{S}^{m+n}_+$ under an affine mapping).

**Example** IV.26.2 Consider the matrix-valued fractional-quadratic function

$$f(u,v) := u^\top v^{-1} u$$

with the domain $\mathrm{Dom}\, f = \{(u,v) : \ u \in \mathbf{R}^{p\times q}, \ v \in \mathrm{int}\,\mathbf{S}^p_+\}$. We claim that $f$ is $\succeq$-convex. Indeed, by the Schur Complement Lemma we have

$$\mathrm{epi}_{\mathbf{S}^q_+}\{f\} = \left\{[(u,v,y) : \ \left[\begin{array}{c|c} y & u^\top \\ \hline u & v \end{array}\right] \succeq 0, v \succ 0\right\}$$

and the right hand side set is convex (by the same reason as in Example IV.26.1).

## 26.2 Testing cone-convexity and cone-monotonicity

### *26.2.1 Cone-monotonicity*

Let us start with a new (for us) notion which will play an important role in "calculus of cone-convexity."

**Definition** IV.26.2 [$(\mathbf{U},\mathbf{K})$-monotonicity] Let $E$ and $F$ be Euclidean spaces equipped with closed pointed cones $\mathbf{U}$ and $\mathbf{K}$, $Q$ be a nonempty convex subset of $E$, and $f(x) : Q \to F$ be a mapping. We say that this mapping is $(\mathbf{U},\mathbf{K})$-*monotone on* $Q$, if $f(x) \leq_{\mathbf{K}} f(x')$ whenever $x, x' \in Q$ are such that $x \leq_{\mathbf{U}} x'$.

For example, when $\mathbf{U}$ and $\mathbf{K}$ are nonnegative orthants in $E = \mathbf{R}^n$ and $F = \mathbf{R}^m$, $(\mathbf{U},\mathbf{K})$-monotonicity of $F$ on $Q$ means that whenever $x \leq x'$ and $x, x' \in Q$, one has $F(x) \leq F(x')$. An instructive extreme example is the one where $\mathbf{U} = \{0\}$; in this case, every mapping $f : Q \to F$ is $(\mathbf{U},\mathbf{K})$-monotone.

### *26.2.2 Differential criteria for cone-convexity and cone-monotonicity*

We next present a differential characterization of cone-convexity and cone-monotonicity. The following claim is nearly evident.

**Proposition** IV.26.3 Let $E, F$ be Euclidean spaces equipped with closed pointed cones $\mathbf{U}$ and $\mathbf{K}$, let $\mathrm{Dom}\, f \subseteq E$ be a convex set with a nonempty interior, and let $f : \mathrm{Dom}\, f \to F$ be a mapping. Then,

(i) $f$ is $\mathbf{K}$-convex if and only if the scalar function $\langle g, f(x)\rangle : \mathrm{Dom}\, f \to \mathbf{R}$ is convex for every $g \in \mathbf{K}_*$. In particular, assuming that $f$ is continuous on $\mathrm{Dom}\, f$ and twice differentiable on $\mathrm{int}(\mathrm{Dom}\, f)$, we deduce that $f$ is $\mathbf{K}$-convex if and only if

$$\frac{d^2}{dt^2}\bigg|_{t=0} f(x+th) \geq_{\mathbf{K}} 0, \qquad \forall (x \in \mathrm{int}(\mathrm{Dom}\, f),\ h \in E).$$

(ii) Assuming $f$ is continuous on $\mathrm{Dom}\, f$ and differentiable on $\mathrm{int}(\mathrm{Dom}\, f)$, $f$ is $(\mathbf{U}, \mathbf{K})$-monotone on $\mathrm{Dom}\, f$ if and only if

$$\frac{d}{dt}\bigg|_{t=0} f(x+th) \geq_{\mathbf{K}} 0, \qquad \forall (h \in \mathbf{U},\ x \in \mathrm{int}(\mathrm{Dom}\, f)).$$

**Example** IV.26.3 The function $f(x) = xx^\top : \mathbf{R}^{m\times n} \to \mathbf{S}^m$ is $\mathbf{S}^m_+$-convex. Indeed,

$$\frac{d}{dt}\bigg|_{t=0} f(x+th)[h] = xh^\top + hx^\top, \quad \frac{d^2}{dt^2}\bigg|_{t=0} f(x+th) = 2hh^\top \succeq 0.$$

We can arrive at the same conclusion by specifying appropriately the data in Example IV.26.1, and compare our now tools with the "bare hands" verification of $\succeq$-convexity of similar matrix-valued function in the proof of Lemma IV.21.9.

**Example** IV.26.4 The function $x \mapsto f(x) := x^{-1} : \mathrm{int}\, \mathbf{S}^m_+ \to \mathrm{int}\, \mathbf{S}^m_+$ is $(-\mathbf{S}^m_+, \mathbf{S}^m_+)$-monotone and $\mathbf{S}^m_+$-convex.
Indeed, as we know from Example C.8 in section C.1.6, for $x \in \mathrm{Dom}\, f$ and $h \in \mathbf{S}^m$ it holds

$$Df(x)[h] := \frac{d}{dt}\bigg|_{t=0} f(x+th) = -x^{-1}hx^{-1}.$$

Thus, $Df(x)[h] \succeq 0$ whenever $h \in -\mathbf{S}^m_+$, which, by Proposition IV.26.3.ii, implies the desired monotonicity. From the above expression for $Df(x)[h]$ it follows that

$$\begin{aligned} D^2 f(x)[h,h] &:= \frac{d^2}{dt^2}\bigg|_{t=0} f(x+th) \\ &= \frac{d}{dt}\bigg|_{t=0} \left(-(x+th)^{-1}h(x+th)^{-1}\right) \\ &= (x^{-1}hx^{-1})hx^{-1} + x^{-1}h(x^{-1}hx^{-1}) \\ &= 2x^{-1}hx^{-1}hx^{-1} \\ &= 2x^{-1/2}\,(x^{-1/2}hx^{-1/2})^2\,x^{-1/2} \succeq 0. \end{aligned}$$

Then, by Proposition IV.26.3.i we deduce the $\mathbf{S}^m_+$-convexity of $f(x)$ as well.

One can easily construct numerical examples showing that the function $f(x) = x^{-2} : \mathrm{int}\, \mathbf{S}^m_+ \to \mathbf{S}^m_+$ is neither $(-\mathbf{S}^m_+, \mathbf{S}^m_+)$-monotone, nor $\mathbf{S}^m_+$ convex.

Our next example is less trivial and relies on the following well-known result that is important on its own.

**Theorem** IV.26.4 Let $f(s) : (a,b) \to \mathbf{R}$ be an analytic function on an interval of real axis. Then, the function $F(x) = f(x) : \operatorname{Dom} F \to \mathbf{S}^m$ where $\operatorname{Dom} F := \{x \in \mathbf{S}^m : \lambda_i(x) \in (a,b), 1 \le i \le m\}$ (see section D.1.5 for the precise definition of the map $F$) is infinitely many times differentiable on the open set $\Delta$.

Justification of this well known fact requires tools (integral Cauchy formula) which go beyond prerequisites in Calculus we take for granted in this book.
We are now ready for our next example.

**Example** IV.26.5 The function $f(x) = x^{1/2} : \mathbf{S}^m_+ \to \mathbf{S}^m$ is $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-monotone and $\mathbf{S}^m_+$-concave (the latter, of course, means that $-f(x)$ is $\mathbf{S}^m_+$-convex).

Here is the justification. By Proposition D.26, $f(x)$ is continuous on $\mathbf{S}^m_+$, so that it suffices to prove that the function possesses the announced monotonicity and concavity properties in the interior of $\mathbf{S}^m_+$. By Theorem IV.26.4, $f(x)$ is infinitely many times differentiable on $\operatorname{int} \mathbf{S}^m_+$. Let us compute the derivative of $f(x)$. Given $x \succ 0$ and $h \in \mathbf{S}^m$ and setting $d := Df(x)[h]$, we have, by differentiating the identity $f^2(x) \equiv x$,

$$x^{1/2} \cdot d + d \cdot x^{1/2} = h. \tag{26.1}$$

Rewriting this linear equation in variable $d \in \mathbf{S}^m$ in an orthonormal eigenbasis of $x$ we see that this equation has a unique solution. $A$ (and therefore *the*) solution to this equation is given by

$$d = \int\limits_0^\infty \exp\{-x^{1/2}t\} h \exp\{-x^{1/2}t\} dt. \tag{26.2}$$

Indeed, consider $u \in \mathbf{R}^{n \times n}$. Then, one has

$$\frac{d}{dt}\exp\{tu\} = \frac{d}{dt}\left[\sum_{i=0}^\infty \frac{1}{i!} t^i u^i\right] = u \sum_{i=0}^\infty \frac{1}{i!} t^i u^i = u \exp\{tu\} = \exp\{tu\} u$$

(for the definition of the exponent of a square matrix, see Remark D.25). Moreover, by integrating in $t$, over the ray $\mathbf{R}_+$, the identity

$$\begin{aligned}&\frac{d}{dt}\left[\exp\{-x^{1/2}t\} h \exp\{-x^{1/2}t\}\right]\\ &= -x^{1/2}\exp\{-x^{1/2}t\} h \exp\{-x^{1/2}t\} - \exp\{-x^{1/2}t\} h \exp\{-x^{1/2}t\} x^{1/2},\end{aligned}$$

we arrive at the result.[2]

[2] Note that when $x \succ 0$ and $h$ commute (as is the case when $x, h \in \mathbf{S}^1$ and $x > 0$), (26.2) becomes

$$\begin{array}{rcl} d & = & \int_0^\infty \exp\{-x^{1/2}t\} h \exp\{-x^{1/2}t\} dt = h \int_0^\infty \exp\{-2x^{1/2}t\} dt \\ & = & h \int_0^\infty \frac{d}{dt}\left[-[2x^{1/2}]^{-1}\exp\{-2x^{1/2}t\}\right] dt = [2x^{1/2}]^{-1} h, \end{array}$$

in full accordance with the formula $D[\sqrt{s}][ds] = \frac{1}{2\sqrt{s}} ds$ for the derivative of the univariate $\sqrt{s}$. Thus, (26.2) is (not that predictable in advance) matrix version of the formula for the derivative of the usual square root.

From (26.2) we deduce that $d \succeq 0$ whenever $h \succeq 0$, and thus by Proposition IV.26.3.ii, we conclude that $f$ is $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-monotone. Now, differentiating the identity (26.1) with $x$ replaced by $x + th$, that is, the identity

$$f(x+th)Df(x+th)[h] + Df(x+th)[h]f(x+th) \equiv h$$

in $t$ and setting $t = 0$, we get

$$0 = 2d^2 + x^{1/2}D^2f(x)[h,h] + D^2f(x)[h,h]x^{1/2},$$

whence, as above,

$$D^2f(x)[h,h] = -2\int_0^\infty \exp\{-tx^{1/2}\}d^2\exp\{-tx^{1/2}\}dt,$$

so that $D^2f(x)[h,h] \preceq 0$, thus $f(x) = x^{1/2}$ is $\mathbf{S}^m_+$-concave.

## 26.3 Elementary calculus of cone-convexity

We start with the elementary calculus of cone-convexity.

**A.** Let $F$ be a Euclidean space and $\mathbf{K} \subset F$ be a closed pointed cone. Then,

**A.1.** An affine mapping $f(x) = Ax + b : \mathbf{R}^n \to F$ is $\mathbf{K}$-convex.

**A.2.** When $f_i(x)$, $i = 1, \ldots, m$, are $\mathbf{K}$-convex functions with common domain $D \subseteq E$, $E$ being a finite-dimensional space, and $\lambda_i \geq 0$, the function $\sum_i \lambda_i f_i(x)$ with the domain $D$ is $\mathbf{K}$-convex.

**A.3.** When $f(x) : \mathrm{Dom}\, f \to F$ is $\mathbf{K}$-convex, and $x = Ay + b : G \to E$ is an affine mapping, $E$ being the embedding space of $\mathrm{Dom}\, f$, the function $g(y) = f(Ay + b)$ with the domain $\mathrm{Dom}\, g = \{y : Ay + b \in \mathrm{Dom}\, f\}$ is $\mathbf{K}$-convex.

**A.4.** When $f(x) : \mathrm{Dom}\, f \to \mathbf{R}^\nu$ is $\mathbf{U}$-convex, $\mathbf{U}$ being a closed pointed cone in $\mathbf{R}^\nu$, and $y = Az + b : \mathbf{R}^\nu \to F$ is an affine mapping such that $Au \in \mathbf{K}$ whenever $u \in \mathbf{U}$, the function $g(x) := Af(x) + b$, with $\mathrm{Dom}\, g := \mathrm{Dom}\, f$, is $\mathbf{K}$-convex.

We next present the "conic version" of Convex Monotone superposition rules from section 14.1.

**B.** Let

- $\mathbf{K} \subset \mathbf{R}^\nu$ be a closed pointed cone,
- $F(y) : \mathrm{Dom}\, F \to \mathbf{R}^\nu$ be a mapping with convex domain $\mathrm{Dom}\, F \subseteq \mathbf{R}^{n_1} \times \ldots \times \mathbf{R}^{n_K}$, so that an argument $y = [y_1; \ldots; y_K]$ of $F$ is a block vector with blocks $y_k$ of dimension $n_k$, $1 \leq k \leq K$,
- $\mathbf{U}_k \subset \mathbf{R}^{n_k}$, $k \leq K$, be closed pointed cones,
- $f_k(x) : D \to \mathbf{R}^{n_k}$, $1 \leq k \leq K$, be mappings with common convex domain $D \subseteq \mathbf{R}^n$.

Assume that $F$ is $\mathbf{K}$-convex and $(\mathbf{U}_1 \times \ldots \times \mathbf{U}_K, \mathbf{K})$-monotone, $f_k$ are $\mathbf{U}_k$-convex, $k \leq K$, and

$$f(x) := [f_1(x); \ldots; f_K(x)] \in \mathrm{Dom}\, F, \quad \forall x \in D.$$

Then, the function

$$G(x) := F(f(x)) : D \to \mathbf{R}^\nu$$

is **K**-convex.

**Remark** IV.26.5 Note that some of $\mathbf{U}_k$ may be trivial: $\mathbf{U}_k = \{0\}$. For these $k$, $\mathbf{U}_k$-convexity of $f_k$ is the same as $f_k$ being an affine function, and $(\mathbf{U}_k, \mathbf{K})$-monotonicity of $F$ in $y_k$ holds true automatically. Thus, the above rule covers the usual Convex Monotone superposition rule from section 14.1, where affinity of some of the inner functions $f_k$ allowed to lift the requirement for the outer function $F$ to be monotone in the respective $y_k$.

Let us illustrate these calculus rules on some examples.

**Example** IV.26.6 Suppose $f(x) : \mathbf{S}^m_+ \to \mathbf{S}^m_+$ and $g(x) : \mathbf{S}^m_+ \to \mathbf{S}^m_+$ are $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-monotone and $\mathbf{S}^m_+$-concave, then, as an immediate corollary of **B**, so is the mapping $h(x) := f(g(x))$.

Therefore, recalling also Example IV.26.5, we conclude that for any positive integer $k$ the function $x^{1/2^k} : \mathbf{S}^m_+ \to \mathbf{S}^m_+$ is $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-monotone and $\mathbf{S}^m_+$-concave. Taking into account that

$$\lim_{\alpha \to +0} \frac{s^\alpha - 1}{\alpha} = \ln s,\ s > 0.$$

we conclude that the matrix logarithm, i.e., the function

$$f(x) := \ln(x) : \operatorname{int} \mathbf{S}^m_+ \to \mathbf{S}^m$$

is $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-monotone and $\mathbf{S}^m_+$-concave.

Finally, in contrast to these, the matrix exponent $\exp\{x\} : \mathbf{S}^m \to \mathbf{S}^m$ possesses no monotonicity/convexity properties unless $m = 1$.

**Example** IV.26.7 Let the function $f(x) : \operatorname{Dom} f \to \operatorname{int} \mathbf{S}^m_+$ be continuous on the convex domain $\operatorname{Dom} f \subseteq \mathbf{R}^n$ with a nonempty interior, and assume that $f$ is twice differentiable on $\operatorname{int} \operatorname{Dom} f$. When $f$ is $\mathbf{S}^m_+$-concave, the function $g(x) := f^{-1}(x)$ is $\mathbf{S}^m_+$-convex.

Indeed, for $x \in \operatorname{Dom} f$ and $h \in \mathbf{R}^n$ we have

$$\begin{aligned}
Dg(x)[h] &= -g(x)Df(x)[h]g(x),\\
D^2g(x) &= 2g(x)Df(x)[h]g(x)Df(x)[h]g(x) - g(x)D^2f(x)[h,h]g(x)\\
&= 2g^{1/2}(x)\left(g^{1/2}(x)Df(x)[h]g^{1/2}(x)\right)^2 g^{1/2}(x) - g(x)D^2f(x)[h,h]g(x) \succeq 0
\end{aligned}$$

(recall that $D^2f(x)[h,h] \preceq 0$ by Proposition IV.26.3.i as $f$ is $\mathbf{S}^m_+$-concave).

**Example** IV.26.8 All of the following functions

$$
\begin{array}{l}
f(u,v,w,z) = \left[\begin{array}{c|c} u & v^\top \\ \hline v & w - z^{1/2} \end{array}\right], \text{ where} \\
\qquad \operatorname{Dom} f := \{(u,v,w,z) : u \in \mathbf{S}^m, v \in \mathbf{R}^{n\times m}, w \in \mathbf{S}^n, z \in \mathbf{S}^n, z \succeq 0\} \\
g(u,v,w,z) = \left[\begin{array}{c|c} u & v^\top \\ \hline v & w + z^{-1/2} \end{array}\right], \text{ where} \\
\qquad \operatorname{Dom} g := \{(u,v,w,z) : u \in \mathbf{S}^m, v \in \mathbf{R}^{n\times m}, w \in \mathbf{S}^n, z \in \mathbf{S}^n, z \succ 0\} \\
h(u,v,w,z) = \left[\begin{array}{c|c} u & v^\top \\ \hline v & w + z^{-1} \end{array}\right], \text{ where} \\
\qquad \operatorname{Dom} h := \{(u,v,w,z) : u \in \mathbf{S}^m, v \in \mathbf{R}^{n\times m}, w \in \mathbf{S}^n, z \in \mathbf{S}^n, z \succ 0\} \\
e(u,v,w,z) = \left[\begin{array}{c|c} u & v^\top \\ \hline v & w + z^2 \end{array}\right], \text{ where} \\
\qquad \operatorname{Dom} e := \{(u,v,w,z) : u \in \mathbf{S}^m, v \in \mathbf{R}^{n\times m}, w \in \mathbf{S}^n, z \in \mathbf{S}^n\}
\end{array}
$$

are $\mathbf{S}^{m+n}_+$-convex.

To justify our claim, note that the function

$$F(y_1,y_2,y_3) := \left[\begin{array}{c|c} y_1 & y_2^\top \\ \hline y_2 & y_3 \end{array}\right] : \mathbf{S}^m \times \mathbf{R}^{n\times m} \times \mathbf{S}^n \to \mathbf{S}^{m+n}$$

is affine and therefore $\mathbf{S}^{m+n}_+$-convex. By Proposition IV.26.3.ii, this function is $(\mathbf{S}^m_+, \mathbf{S}^{m+n}_+)$-monotone in $y_3$. Applying **B** with $\mathbf{U}_1 = \{0\} \subset \mathbf{S}^m$, $\mathbf{U}_2 = \{0\} \subset \mathbf{R}^{m\times n}$, $\mathbf{U}_3 = \mathbf{S}^n$ and

$$f_1(u,v,w,z) := u,\ \ f_2(u,v,w,z) := v,\ \ f_3(u,v,w,z) := w - z^{1/2}$$

(the latter function is $\mathbf{S}^n_+$-convex in its domain $\{(u,v,w,z) : z \succeq 0\}$ by Example IV.26.5, we conclude that $f$ is $\mathbf{S}^{m+n}_+$-convex. Similar reasoning, with $f_3(u,v,w,z)$ replaced with

- $w + z^{-1/2}$ (this function is $\mathbf{S}^n_+$-convex in the domain $z \succ 0$ by $\succeq$-concavity of $z^{1/2}$, $z \succeq 0$ (Example IV.26.5) and Example IV.26.7)

- $w + z^{-1}$ (this function is $\mathbf{S}^n_+$-convex in the domain $z \succ 0$ by Example IV.26.7),

- $w + z^2$ (this function is $\mathbf{S}^n_+$-convex by Example IV.26.3),

justifies the announced $\mathbf{S}^{m+n}_+$-convexity of the functions $g, h, e$ as well.

# 27

# ★ Mathematical Programming Optimality Conditions

The goal of this chapter is to develop optimality conditions for a general-type Mathematical Programming problem

$$\min_x \left\{ f(x) : \begin{array}{l} \overbrace{g_1(x) \le 0, \ldots, g_m(x) \le 0,}^{\text{inequality constraints}} \\ \underbrace{h_1(x) = 0, \ldots, h_k(x) = 0}_{\text{equality constraints}} \end{array} \right\} \tag{27.1}$$

where $m, k$ are nonnegative integers, and the objective $f$ and the constraints $g_j$, $h_i$ are real-valued functions well-defined each on its own subset of $\mathbf{R}^n$. This topic seemingly "goes beyond convexity;" however, related developments utilize Convex Analysis tools, and it would be unwise to skip it completely.

Same as in the case of convex programs, optimality conditions are aimed at answering the following question:

> *Given a feasible solution $x_*$ to* (27.1), *what are necessary/sufficient conditions for $x_*$ to be an optimal solution to the problem?*

The conditions we are looking for should be verifiable: given $x_*$ and local information on the objective and the constraints (i.e., their values and derivatives taken at $x_*$) we should be able to check whether the conditions are/are not satisfied.

Beyond the convex case, there are no verifiable sufficient conditions for $x_*$ to be *globally optimal* (i.e., $f(x_*) \le f(x)$ for every feasible $x$) are known. The existing optimality conditions focus on *local optimality* of $x_*$ defined as follows:

**Definition** IV.27.1 [Local optimality] A feasible solution to (27.1) is called *locally optimal*, if the objective and the constraints are well defined in a neighborhood of $x_*$ and $x_*$ has the best –the smallest– objective value among all feasible solutions that are close enough to it, that is, there exists $r > 0$ such that

$$x \text{ is feasible for (27.1)} \quad \text{and} \quad \|x - x_*\|_2 \le r \implies f(x_*) \le f(x).$$

The classical MP optimality conditions we are about to present are applicable only when $x_*$ is a *regular* feasible solution to (27.1), with regularity defined as follows:

**Definition** IV.27.2 [Regular solution] A vector $x_* \in \mathbf{R}^n$ is called a *regular solution* to problem (27.1), if
- $x_*$ is a feasible solution to the problem,
- the objective and the constraints are well defined and continuously differentiable in a neighborhood of $x_*$, and
- the taken at $x_*$ gradients of active at $x_*$ (i.e., satisfied at $x_*$ as equalities) constraints are linearly independent.

Geometrically and informally, regularity of a solution $x_*$ means that the set $S$ cut off a small enough neighborhood of $x_*$ by the equality versions of the constraints active at $x_*$ is a smooth surface.

## 27.1 Formulating Optimality conditions

**27.1.1 Default Assumption.** From now on, unless the opposite is explicitly stated,

> *we assume that $x_*$ is a regular solution to (27.1) and that the objective and the constraints are twice continuously differentiable in a neighborhood of $x_*$.*

We will express our optimality conditions in terms of the Lagrange function of (27.1) given by

$$L(x;\lambda,\mu) := f(x) + \sum_{j=1}^{m} \lambda_j g_j(x) + \sum_{i=1}^{k} \mu_i h_i(x).$$

This function is well defined on the direct product of a neighborhood $X$ of $x_*$ and the entire space $\mathbf{R}^m_\lambda \times \mathbf{R}^k_\mu$ of *Lagrange multipliers* $\lambda, \mu$ and is twice continuously differentiable in $x \in X$ and linear in $[\lambda; \mu]$.

We are now ready to state our optimality conditions (first the necessary one, then the sufficient one) for general MPs under our *Default Assumption*.

**Proposition** IV.27.3 [Necessary Optimality condition for general MP] Suppose *Default Assumption* takes place, and let $x_*$ be a locally optimal solution to problem (27.1). Then,
medskip

(i) [first order part] $x_*$ is a KKT (Karush-Kuhn-Tucker) point of the problem, i.e., there exist $\lambda^* \in \mathbf{R}^m_+$ and $\mu^* \in \mathbf{R}^k$ satisfying
[complementary slackness]: $\lambda^*_j g_j(x_*) = 0$, for all $1 \le j \le m$
[KKT equation]: $\nabla_x\big|_{x=x_*} L(x;\lambda^*,\mu^*) = 0$.

(ii) [second order part] The second order directional derivatives of $L(\cdot;\lambda^*,\mu^*)$

taken at $x_*$ along the directions from the linear subspace

$$T_{\rm n} := \left\{ d \in \mathbf{R}^n : \begin{array}{l} d \text{ is orthogonal to the taken at } x_* \text{ gradients} \\ \text{of all constraints that are active at } x_* \end{array} \right\}$$

are nonnegative, i.e.,

$$d^\top \nabla_x^2\big|_{x=x_*} L(x;\lambda^*,\mu^*)d \geq 0, \quad \forall d \in T_{\rm n}.$$

Note that under our Default Assumption if $x_*$ is a KKT point of (27.1), then the corresponding Lagrange multipliers $\lambda^*, \mu^*$ are uniquely defined by $x_*$. Indeed, the KKT equation says that $-\nabla f(x_*) = \sum_j \lambda_j^* \nabla g_j(x_*) + \sum_i \mu_i^* \nabla h_i(x_*)$. By complementary slackness, the Lagrange multipliers $\lambda_j^*$ for non-active at $x_*$ inequality constraints are zero, and the rest of $\lambda_j^*$'s taken together with $\mu_i^*$'s form the coefficients in a representation of $-\nabla f(x_*)$ as a linear combination of *linearly independent* (since $x_*$ is regular) gradients, taken at $x_*$, of the active at $x_*$ constraints and thus are uniquely defined.

**Proposition** IV.27.4 [Sufficient Optimality condition for general MP] Suppose *Default Assumption* takes place, and let $x_*$ be such that

(i) [first order part] $x_*$ is a KKT point of (27.1) as defined in item (i) of Proposition IV.27.3, and

(ii) [second order part] For the Lagrange multipliers $\lambda^*, \mu^*$ associated with the KKT point $x_*$, we have that the second order directional derivatives of $L(\cdot;\lambda^*,\mu^*)$ taken at $x_*$ along the nonzero directions from the linear subspace

$$T_{\rm s} := \left\{ d \in \mathbf{R}^n : \begin{array}{l} d \text{ is orthogonal to the taken at } x_* \text{ gradients} \\ \text{of all equality constraints } h_i \text{ and all inequality} \\ \text{constraints } g_j \textit{ corresponding to } \lambda_j^* > 0 \end{array} \right\}$$

are positive, i.e.,

$$d^\top \nabla_x^2\big|_{x=x_*} L(x;\lambda^*,\mu^*)d > 0, \quad \forall d \in T_{\rm s}\backslash\{0\}.$$

Then, $x_*$ is a locally optimal solution to (27.1).

Pay attention to the fact that by complementary slackness $\lambda_j^* > 0$ only when the constraint $g_j(x) \leq 0$ is active at $x_*$. As a result, the linear subspaces participating in the second order parts of these two Optimality conditions are embedded one into another: $T_{\rm n} \subseteq T_{\rm s}$, and in general this inclusion is strict. In fact, $T_{\rm n} = T_{\rm s}$ holds if and only if all active at $x_*$ inequality constraints $g_j$ are associated with positive, and not just nonnegative, Lagrange multipliers $\lambda_j^*$.

## 27.2 Justifying Optimality conditions

**Justifying Optimality conditions: preliminary step** It may happen that some of the inequality constraints are non-active at $x_*$. It is immediately seen

that in our context, eliminating these constraints from (27.1) changes nothing: on one hand, $x_*$ remains a regular solution to the resulting problem and is a *locally* optimal solution to the latter problem if and only if it is locally optimal to the former one. On the other hand, satisfiability statuses of optimality conditions in both problems are exactly the same, since these conditions are expressed in terms of two entities:
(a) the functions of $x$ obtained from the respective Lagrange functions by setting $\lambda$, $\mu$ to $\lambda^*, \mu^*$; note that complementary slackness enforces these functions for both problems in question to be the same;
(b) linear subspaces $T_{\rm n}$ and $T_{\rm s}$; these linear subspaces are defined in terms of the taken at $x_*$ gradients of the *active* at $x_*$ constraints and for both problems in question are the same as well.
The bottom line is that *it suffices to verify validity of our optimality conditions in the special case when all inequality constraints in (27.1) are active at* $x_*$, and this is what we assume from now on.

### *27.2.1 Main tool: Implicit Function Theorem*

We are about to justify Optimality conditions by utilizing the following fundamental fact (which is one of the forms of Implicit Function Theorem available in any graduate textbook on multivariate calculus):

**Theorem** IV.27.5 Let $x_* \in \mathbf{R}^n$ and $\phi_1, \dots, \phi_p$ be real-valued functions well defined and $\kappa \geq 1$ times continuously differentiable in a neighborhood $U$ of $x_*$ and normalized by the condition $\phi_\ell(x_*) = 0$, $\ell \leq p$. Assume that the taken at $x_*$ gradients of $\phi_\ell(x)$, $\ell \leq p$, are linearly independent. Then, there exists a neighborhood $X \subseteq U$ of $x_*$, a neighborhood $Y$ of $y_* := 0 \in \mathbf{R}^n$ and a one-to one mapping $y(x)$ of $X$ onto $Y$ which, along with its inverse $x(y)$ (this is a one-to-one mapping of $Y$ onto $X$), possesses the following properties

1. $y_* := y(x_*) = 0$ ( $\Longleftrightarrow$ $x(0) = x_*$);
2. $y(x)$ and $x(y)$ are $\kappa$ times continuously differentiable on $X$, resp. on $Y$;
3. In $y$-variables, the functions $\phi_\ell(\cdot)$, $\ell \leq p$, become just the first $p$ coordinates $y_\ell$, $\ell \leq p$, of $y$:

$$\phi_\ell(x(y)) \equiv y_\ell,\ \forall (y \in Y, \ell \leq p) \quad [\Longleftrightarrow\ \phi_\ell(x) = y_\ell(x),\ \forall (x \in X, \ell \leq p)] \tag{27.2}$$

### *27.2.2 Strategy*

Let us apply the Implicit Function Theorem to the $m + k$ functions given by

$$\phi_j(x) := g_j(x),\ j \leq m, \quad \phi_{m+i}(x) := h_i(x),\ i \leq k. \tag{27.3}$$

Since we are in the case when $x_*$ is a regular solution and *all* constraints of the problem of interest are active at $x_*$, the resulting $p = m + k$ functions $\phi_\ell$

satisfy the premise in the Implicit Function Theorem, with $\kappa$ set to 2. Let $X$, $Y$, $x(\cdot)$, $y(\cdot)$ be the entities given by the theorem as applied to our $\phi_\ell$ and $x_*$. Taking into account (27.2), substituting $x = x(y)$ and defining $\phi(y) := f(x(y))$, we convert the original problem (27.1) into the linearly constrained Mathematical Programming problem

$$\min_y \{\phi(y) : \ y_j \leq 0, j \leq m, \ y_{m+i} = 0, i \leq k\}. \tag{27.4}$$

Taking into account that $y(x)$ and $x(y)$ are twice continuously differentiable inverse to each other mappings of neighborhoods $X$ of $x_*$ and $Y$ of $y_* = y(x_*) = 0$ onto each other, the function $\phi(y)$ is twice continuously differentiable in a neighborhood of $y_* = 0$, and $x_*$ is a locally optimal solution to (27.1) if and only if $y_* = 0$ is a locally optimal solution to (27.4). Moreover, by looking at the problem (27.4), we conclude that $y_* = 0$ is its regular solution.

Our course of actions will be as follows:

**A.** We start with verifying that our optimality conditions as applied to problem (27.4) and its regular solution $y_* = 0$ are indeed valid.

**B.** We "transfer" these *valid* optimality conditions from problem (27.4) to problem of interest (27.1). Taking into account that $x_*$ is locally optimal for the latter problem if and only if $y_* = 0$ is a locally optimal solution to the former problem, we end up with *valid* necessary/sufficient optimality conditions for the problem of interest. As we shall see, these valid conditions are *exactly* the conditions stated in Propositions IV.27.3, IV.27.4; this will complete the justification of these two propositions.

Note that when implementing our strategy we should not bother about complementary slackness, since in both problems (27.1), (27.4) all constraints are active at $x_*$, resp. $y_*$, making complementary slackness trivially true.

### *27.2.3 Justifying optimality conditions for (27.4)*

We are about to verify the validity of Propositions IV.27.3 and IV.27.4 as applied to problem (27.4), which is nearly immediate. Let

$$\widehat{L}(y; \lambda, \mu) := \phi(y) + \sum_{j=1}^{m} \lambda_j y_j + \sum_{i=1}^{k} \mu_i y_{m+i}$$

be the Lagrange function of (27.4). Here are the straightforward specifications of the optimality conditions stated in Propositions IV.27.3 and IV.27.4 as applied to the solution $y_* = 0$ of problem (27.4):

- **Claim N**: *The condition* **N***:*

**N**.(i) [first order part] *$y_* = 0$ is a KKT point of* (27.4)*, i.e., there exist $\lambda^* \in \mathbf{R}^m_+$ and $\mu^* \in \mathbf{R}^k$ such that $\nabla_y\big|_{y=0}\widehat{L}(y;\lambda^*,\mu^*) = 0$.*

**N**.(ii) [second order part] *The second order directional derivatives of the function $\widehat{L}(\cdot;\lambda^*,\mu^*)$, or, which is the same under the circumstances, of the function $\phi(y)$, taken at the point $y_* = 0$ along any direction from the linear space*

$$\overline{T}_{\text{n}} := \left\{d \in \mathbf{R}^n :\ e_\ell^\top d = 0, \forall \ell \le m+k\right\}$$

*($e_1,\dots,e_n$, as always, are the standard basic orth in $\mathbf{R}^n$) are nonnegative, i.e.,*

$$e_\ell^\top d = 0,\ \forall \ell \le m+k \quad \Longrightarrow \quad d^\top \nabla^2_y\big|_{y=0}\widehat{L}(y;\lambda^*,\mu^*)d = d^\top\nabla^2\phi(0)d \ge 0.$$

*is necessary for $y_* = 0$ to be a locally optimal solution to* (27.4)*.*

- **Claim S:** *The condition* **S***:*

**S**.(i) [first order part] *$y_* = 0$ is a KKT point of* (27.4) *(see item* **N***.(i) above),*

**S**.(ii) [second order part] *The second order directional derivative of the function $\widehat{L}(\cdot;\lambda^*,\mu^*)$ (where $\lambda^*,\mu^*$ are the Lagrange multipliers associated with the KKT point $y_* = 0$), or, which is the same under the circumstances, of the function $\phi(y)$, taken at the point $y_* = 0$ along every nonzero direction from the linear subspace*

$$\overline{T}_{\text{s}} := \left\{d \in \mathbf{R}^n :\ e_\ell^\top d = 0,\ \forall(\ell \le m : \lambda^*_\ell > 0),\ e_{m+\ell}^\top d = 0, \forall \ell \le k\right\}$$

*is positive, i.e.,*

$$\begin{array}{l} d \neq 0\ \&\ e_\ell^\top d = 0,\ \forall(\ell \le m : \lambda^*_j > 0)\ \&\ e_{m+\ell}^\top d = 0,\ \forall \ell \le k \\ \Longrightarrow\ d^\top\nabla^2_y\big|_{y=0}\widehat{L}(y;\lambda^*,\mu^*)d = d^\top\nabla^2\phi(0)d > 0. \end{array}$$

*is sufficient for $y_* = 0$ to be a locally optimal solution to* (27.4)*.*

We are about to justify Claims N, S.

**1)** Observe that the feasible set of (27.4) is the polyhedral cone

$$\mathbf{F} := \left\{d \in \mathbf{R}^n :\ e_\ell^\top d \le 0,\ \forall \ell \le m,\ e_{m+\ell}^\top d = 0,\ \forall \ell \le k\right\},$$

so that a necessary condition for $y_* = 0$ to be locally optimal solution to (27.4) is that for every $d \in \mathbf{F}$, 0 is the local minimizer of the restriction of $\phi(\cdot)$ on the ray $\{td : t \ge 0\}$. Next, given a univariate function $\psi$ well defined and twice continuously differentiable in a neighborhood of the origin, the elementary calculus says that a necessary condition for 0 to be a local minimizer of the restriction of $\psi$ onto the nonnegative ray is

$$\psi'(0) \ge 0\ \&\ \psi''(0) \ge 0 \text{ when } \psi'(0) = 0.$$

Thus, we conclude that

**C:** *A necessary condition for $y_* = 0$ to be a locally optimal solution to* (27.4) *is that for every direction $d \in \mathbf{F}$*

(a) *the first-order directional derivative $d^\top \nabla\phi(0)$ of $\phi$ taken at the origin along the direction $d$ is nonnegative, and*

(b) *if the first-order directional derivative from* (a) *is zero, then the second-order directional derivative $d^\top \nabla^2\phi(0)d$ of $\phi$ taken at the origin along the direction $d$ is nonnegative.*

Now, **C**.(a) is nothing but the fact that the homogeneous linear inequality $d^\top \nabla\phi(0) \geq 0$ in variables $d \in \mathbf{R}^n$ is a consequence of the following system of homogeneous linear inequalities in variables $d$:

$$-e_j^\top d \geq 0, \, \forall j \leq m, \qquad \pm\, e_{m+i}^\top d \geq 0, \, \forall i \leq k.$$

By Homogeneous Farkas Lemma, this is the same as to say that $-\nabla\phi(0)$ is a linear combination, with nonnegative coefficients, of the vectors $e_j, j \leq m$, $\pm e_{m+i}, i \leq k$, or, which again is the same, that $-\nabla\phi(0)$ is a linear combination of the basic orth $e_\ell$, $\ell \leq m+k$, with the first $m$ coefficients being nonnegative. Thus, **C**.(a) is nothing but the fact that

$$\nabla\phi(0) + \sum_{j=1}^m \lambda_j^* e_j + \sum_{i=1}^k \mu_i^* e_{m+i} = 0$$

for some nonnegative $\lambda_j^*$. The bottom line is that **C**.*(a) is nothing but the fact that $y_* = 0$ is a KKT point of problem* (27.4).

Now, let us define and examine the set

$$\mathbf{K} := \left\{ d \in \mathbf{F} : \; d^\top \nabla\phi(0) = 0 \right\}.$$

When **C**.(a) holds true, by the representation of $\nabla\phi(0)$ in terms of $\lambda^* \in \mathbf{R}^m_+$, $\mu^* \in \mathbf{R}^k$ and also by definition of **F**, we have

$$\begin{aligned}
\mathbf{K} &= \left\{ d \in \mathbf{F} : d^\top \nabla\phi(0) = 0 \right\} \\
&= \left\{ d \in \mathbf{F} : d^\top \left( \sum_{j=1}^m \lambda_j^* e_j + \sum_{i=1}^k \mu_i^* e_{m+i} \right) = 0 \right\} \\
&= \left\{ d \in \mathbf{R}^n : \begin{array}{l} e_j^\top d \leq 0, \, \forall j \leq m, \; e_{m+i}^\top d = 0, \, \forall i \leq k, \\ \sum_{j=1}^m \lambda_j^* d_j + \sum_{i=1}^k \mu_i^* d_{m+i} = 0 \end{array} \right\} \\
&= \left\{ d \in \mathbf{R}^n : \begin{array}{l} e_j^\top d \leq 0, \, \forall j \leq m, \; e_{m+i}^\top d = 0, \, \forall i \leq k, \\ \sum_{j=1}^m \lambda_j^* d_j = 0 \end{array} \right\} \\
&= \left\{ d \in \mathbf{R}^n : \begin{array}{l} e_j^\top d \leq 0, \, \forall (j \leq m : \lambda_j^* = 0) \\ e_j^\top d = 0, \, \forall (j \leq m : \lambda_j^* > 0) \\ e_{m+i}^\top d = 0, \, \forall i \leq k \end{array} \right\}.
\end{aligned}$$

Note that **C**.(b) wants from the quadratic form $d^\top \nabla^2\phi(0)d$ to be nonnegative on the cone **K**. The bottom line is that we have justified the following:

**Fact N:** *The condition* $\mathbf{N}^*$*:*

> $y_* = 0$ *is a KKT point of* (27.4) *such that second order directional derivatives* $d^\top \nabla^2 \phi(0) d$ *of* $\phi$ *(or, which is the same, of the Lagrange function* $\widehat{L}(\cdot; \lambda^*, \mu^*)$*, with* $\lambda^*, \mu^*$ *given by the KKT property of* $y_*$*) taken at* $y_* = 0$ *along all directions from cone* $\mathbf{K}$ *are nonnegative*

*is necessary for* $y_* = 0$ *to be a locally optimal solution to* (27.4).

**2)** We have already outlined that a necessary condition for 0 to be local minimizer of the restriction onto $\mathbf{R}_+$ of a well defined and twice continuously differentiable in a neighborhood of 0 univariate function $\psi$ is "$\psi'(0) \geq 0$ and $\psi''(0) \geq 0$ when $\psi'(0) = 0$." In fact, a slightly strengthened version of this condition is a sufficient condition for local optimality of 0 for $\psi$ over $\mathbf{R}_+$ as well. Specifically, if for $\psi$ it holds that $\psi'(0) \geq 0$ and $\psi''(0) > 0$ when $\psi'(0) = 0$, then 0 is a local minimizer of the restriction of $\psi$ onto $\mathbf{R}_+$. This suggests the following *educated guess:*

> *Let all first-order directional derivatives* $d^\top \nabla \phi(0)$ *of the function* $\phi$ *taken at the origin along directions from the cone* $\mathbf{F}$ *be nonnegative, and the second-order directional derivatives* $d^\top \nabla^2 \phi(0) d$ *taken at the origin along all nonzero directions from* $\mathbf{F}$ *which are orthogonal to* $\nabla \phi(0)$ *be positive. Then* $y_* = 0$ *is a locally optimal solution to* (27.4).

An absolutely straightforward verification demonstrates that our educated guess is correct. Modulo this verification (it requires nothing but the fact that a continuous real-valued function on a nonempty closed and bounded set attains its minimum on the set and is left to the reader), we have arrived at the following:

**Fact S:** *The condition* $\mathbf{S}^*$*:*

> $y_* = 0$ *is a KKT point of problem* (27.4) *such that second-order directional derivatives* $d^\top \nabla^2 \phi(0) d$ *of* $\phi$ *(or, which is the same, of the Lagrange function* $\widehat{L}(\cdot; \lambda^*, \mu^*)$*, with* $\lambda^*, \mu^*$ *given by the KKT property of* $y_*$*) taken at* $y_* = 0$ *along all nonzero directions from cone* $\mathbf{K}$ *are positive*

*is sufficient for* $y_* = 0$ *to be a locally optimal solution to* (27.4).

Note that the sufficient optimality condition $\mathbf{S}^*$ is obtained from $\mathbf{N}^*$ by strengthening the nonnegativity of the quadratic form $d^\top \nabla^2 \phi(0) d = d^\top \nabla^2_y\big|_{y=0} \widehat{L}(y; \lambda^*, \mu^*) d$ on the cone $\mathbf{K}$ to positivity of the form on the "nonzero part" $\mathbf{K} \backslash \{0\}$ of that cone.

**3)** So far, we have established pretty close to each other conditions $\mathbf{N}^*$, $\mathbf{S}^*$ which are necessary, resp. sufficient, for $y_* = 0$ to be a locally optimal solution to (27.4). Unfortunately, these conditions are difficult to verify, since checking nonnegativity/positivity outside of the origin of a quadratic form on a polyhedral cone, in contrast to checking the form's nonnegativity/positivity outside of the origin on a linear subspace, is a computationally intractable task even when the cone is a simple as the nonnegative orthant. To overcome, to some extent, this difficulty, we

— modify the necessary optimality condition $\mathbf{N}^*$ by replacing nonnegativity of the quadratic form $d^\top \nabla^2 \phi(0) d$ on the cone $\mathbf{K}$ with nonnegativity of the form on

the largest linear subspace contained in $\mathbf{K}$. This linear subspace, as is immediately seen, is $\overline{T}_{\rm n}$, and the resulting "spoiled" necessary optimality condition is nothing but condition $\mathbf{N}$;
— modify the sufficient optimality condition $\mathbf{S}^*$ by replacing positivity of the quadratic form $d^\top\nabla^2\phi(0)d$ on the "nonzero part" $\mathbf{K}\backslash\{0\}$ of the cone $\mathbf{K}$ with positivity of the form on the nonzero part of the smallest linear subspace containing $\mathbf{K}$. This linear subspace, as is immediately seen, is $\overline{T}_{\rm s}$, and the resulting "spoiled" sufficient optimality condition is nothing but condition $\mathbf{S}$.

Claims N and S are justified.

### *27.2.4 Justifying Propositions IV.27.3, IV.27.4*

Let us start with summarizing some of our observations.

**O.1.** $x_*$ is a locally optimal solution to (27.1) if and only if $y_* = y(x_*) = 0$ is a locally optimal solution to (27.4).

**O.2.** Setting

$$\phi_\ell(x) := e_\ell^\top y(x),\ \forall \ell = 1,\dots,n,$$

where $e_\ell$ are the standard basic orth in $\mathbf{R}^n$, we get twice continuously differentiable in a neighborhood $X$ of $x_*$ functions such that

$$\phi_\ell(x) = \begin{cases} g_\ell(x), & \text{if } \ell \le m,\\ h_{\ell-m}(x), & \text{if } m < \ell \le m+k,\end{cases} \qquad \forall x \in X$$

(see (27.2), (27.3)).

**O.3.** The Lagrange functions $L(x;\lambda,\mu)$, $\widehat{L}(y;\lambda,\mu)$ of problems (27.1) and (27.4) are linked by the relation

$$L(x;\lambda,\mu) = \widehat{L}(y(x);\lambda,\mu),\ \forall (x\in X, \lambda\in\mathbf{R}^m, \mu\in\mathbf{R}^k).$$

Let

$$J := [\nabla y_1(x_*),\dots,\nabla y_n(x_*)]^\top$$

be the taken at $x = x_*$ Jacobian of the mapping $x \mapsto y(x)$; this $n\times n$ matrix is nonsingular, the inverse being the taken at $y_* := y(x_*) = 0$ Jacobian of the mapping $y \mapsto x(y)$. By Chain Rule we have

$$\nabla_x\big|_{x=x_*} L(x;\lambda,\mu) = J^\top \nabla_y\big|_{y=y_*=0}\widehat{L}(y;\lambda,\mu),\ \forall(\lambda\in\mathbf{R}^m,\mu\in\mathbf{R}^k).$$

Taking into account nonsingularity of $J$ and recalling that all constraints in (27.1) and (27.4) are active at $x_*$, $y_*$, respectively, we conclude that *$x_*$ is a KKT point of* (27.1) *if and only if $y_* = 0$ is a KKT point of* (27.4), *and in this case the Lagrange multipliers $\lambda^*, \mu^*$ certifying the KKT properties of the points are the same for both points.*

**O.4.** By the same Chain Rule we have

$$\nabla\phi_\ell(x_*) = J^\top e_\ell, \quad \forall \ell \leq n,$$

implying that *a direction $d$ is orthogonal to $\nabla\phi_\ell(x_*)$ if and only if the direction $Jd$ is orthogonal to $e_\ell$.*

**O.5.** Finally, by elementary calculus we have

$$\begin{array}{rcl}\forall(d \in \mathbf{R}^n, \lambda, \mu): & & \\ d^\top \nabla_x^2\big|_{x=x_*} L(x;\lambda,\mu)d &=& (Jd)^\top \nabla_y^2\big|_{y=y_*=0} \widehat{L}(y;\lambda,\mu)(Jd) \\ & & + \left(\nabla_y\big|_{y=y_*} \widehat{L}(y;\lambda,\mu)\right)^\top \frac{d^2}{dt^2}\Big|_{t=0} y(x_*+td). \end{array} \tag{27.5}$$

### *Justifying Proposition IV.27.3*

Let $x_*$ be a locally optimal solution to (27.1), so that by **O.1** the point $y_* = 0$ is a locally optimal solution to (27.4). Due to the latter fact and (already verified) Claim N, condition **N** is satisfied, whence, in particular, $y_*$ is a KKT point of (27.4), the associated Lagrange multipliers being some $\lambda^* \geq 0, \mu^*$. Consequently, $x_*$ is a KKT point of (27.1), see **O.3**, the associated Lagrange multipliers being the same $\lambda^*, \mu^*$; we have justified the first order part of the conclusion in Proposition IV.27.3. Next, by **O.4** we have $\overline{T}_{\rm n} = JT_{\rm n}$. Besides this, by **O.5** with $\lambda, \mu$ set to $\lambda^*, \mu^*$, we have

$$d^\top \nabla_x^2\big|_{x=x_*} L(x;\lambda^*,\mu^*)d = (Jd)^\top \nabla_y^2\big|_{y=0} \widehat{L}(y;\lambda^*,\mu^*)(Jd) \tag{27.6}$$

for all $d$, since the last term in the right hand side of (27.5) vanishes when $\lambda = \lambda^*, \mu = \mu^*$ – we already know that $y_* = 0$ is a KKT point of (27.4), the Lagrange multipliers being $\lambda^*, \mu^*$. As we have just seen, when $d \in T_{\rm n}$, one has $Jd \in \overline{T}_{\rm n}$, so that the right hand side in (27.6) is nonnegative by the second order part of condition **N**. Thus, the second order part of the conclusion in Proposition IV.27.3 takes place as well. □

### *Justifying Proposition IV.27.4*

Let the premise of Proposition IV.27.4 take place. Then, by the first order part of this premise, $x_*$ is a KKT point of (27.1), the associated Lagrange multipliers being some $\lambda^* \geq 0$ and $\mu^*$. By **O.3**, $y_* = 0$ is a KKT point of problem (27.4), the Lagrange multipliers being the same $\lambda^*, \mu^*$. Thus, the first order part of condition **S** is satisfied. Next, the linear subspace $T_{\rm s}$ is cut off $\mathbf{R}^n$ by the requirement that a direction from $T_{\rm s}$ should be orthogonal to the gradients, taken at $x_*$, of some of the functions $\phi_\ell$, and the linear subspace $\overline{T}_{\rm s}$ is cut off $\mathbf{R}^n$ by the requirement that a direction from $\overline{T}_{\rm s}$ should be orthogonal to some of the vectors $e_\ell$. Indexes $\ell$ participating in these requirements are fully specified, via the same for both subspaces in question rules, by the vectors of Lagrange multipliers associated with the inequality constraints of respective problems (27.1), (27.4). As it was already explained, these two vectors of Lagrange multipliers coincide with each other, which combines with **O.4** to imply that $\overline{T}_{\rm s} = JT_{\rm s}$. Invoking (27.5) with $\lambda, \mu$ set to $\lambda^*, \mu^*$ and taking into account that $y_* = 0$ is a KKT point of (27.4), the Lagrange multipliers being $\lambda^*, \mu^*$, we conclude that (27.6) still holds true, so that second order part of the premise in Proposition IV.27.4 implies that the second order part of condition **S** is satisfied. Thus, condition **S** is satisfied, so that $y_* = 0$ is a locally optimal solution to (27.4) due to the already verified Claim **S**. It remains to note that local optimality of $y_*$ as a solution to (27.4) implies, by **O.1**, local optimality of $x_*$ as a solution to (27.1). □

## 27.3 Concluding remarks

**A.** We have obtained the claims in Propositions IV.27.3, IV.27.4, by "translating" to the problem of interest (27.1) Claims N, S stating optimality conditions established for problem (27.4). In turn, Claims N, S were "spoiled" versions of "tight" necessary optimality condition $\mathbf{N}^*$, and "tight" sufficient optimality condition $\mathbf{S}^*$ for (27.4). We could also translate to problem (27.1) the conditions $\mathbf{N}^*$, $\mathbf{S}^*$ as they are, thus arriving at "tight" necessary and sufficient conditions $\mathbf{N}^+$, $\mathbf{S}^+$ for local optimality of $x_*$ in the problem of interest (27.1). As is immediately seen, $\mathbf{N}^+$, $\mathbf{S}^+$ are obtained from Propositions IV.27.3, IV.27.4 as follows. Define

$$H := \nabla_x^2\big|_{x=x_*} L(x; \lambda^*, \mu^*).$$

Then

— in the second order part of the claim of Proposition IV.27.3, the requirement $d^\top Hd \geq 0$ for all directions $d$ from the linear subspace $T_{\rm n}$ is strengthened to $d^\top Hd \geq 0$ for all directions $d$ from the following cone (which contains $T_{\rm n}$)

$$\overline{\mathbf{K}} := \left\{ d \in \mathbf{R}^n : \begin{array}{ll} d^\top \nabla h_i(x_*) = 0, & \forall i, \\ d^\top \nabla g_j(x_*) = 0, & \text{for all } j \text{ with } \lambda_j^* > 0, \\ d^\top \nabla g_j(x_*) \leq 0, & \text{for all } j \text{ with } g_j(x^*) = 0 \text{ and } \lambda_j^* = 0 \end{array} \right\}.$$

— in the second order part of the claim of Proposition IV.27.4, the requirement $d^\top Hd > 0$ for all nonzero directions $d$ from the linear subspace

$$T_{\rm s} = \{d : d^\top \nabla h_i(x)*) = 0,\, i \leq k,\, d^\top \nabla g)i(x_*) = 0\, \forall (j \leq m : \lambda_j^* > 0)\}$$

is relaxed to $d^\top Hd > 0$ for all nonzero directions $d$ from the cone $\overline{\mathbf{K}}$ which is contained in $T_{\rm s}$.

As we have explained, the rationale to spoil $\mathbf{N}^+$, $\mathbf{S}^+$ is the desire to end up with *efficiently verifiable* optimality conditions. Our current goal is to indicate special cases where the tight conditions are verifiable "as is." Here are these cases (below, $x_*$ is a regular feasible solution to (27.1) which happens to be a KKT point of the problem, and $\lambda^*$ is the associated vector of Lagrange multipliers for the inequality constraints):

- [nondegeneracy] all active at $x_*$ inequality constraints $g_j$ are associated with

positive Lagrange multipliers $\lambda^*_j$.
In this case, $\overline{\mathbf{K}} = T_{\mathrm{n}} = T_{\mathrm{s}}$.

- there is exactly one active at $x_*$ inequality constraint with zero Lagrange multiplier $\lambda^*_j$.
  In this case, $\overline{\mathbf{K}}$ is cut off $T_{\mathrm{s}}$ by a single homogeneous linear inequality, so that for a homogeneous quadratic form to be nonnegative/positive outside of the origin on the cone $\overline{\mathbf{K}}$ and on the subspace $T_{\mathrm{s}}$ is the same. Consequently, in the case in question $\mathbf{N}^+$ is obtained from Proposition IV.27.3 by replacing in the formulation of the second order part the subspace $T_{\mathrm{n}}$ with larger subspace $T_{\mathrm{s}}$, and Proposition IV.27.4 states exactly the same as condition $\mathbf{S}^+$;
- There are exactly two active at $x_*$ inequality constraints with zero Lagrange multipliers $\lambda^*_\cdot$, say, $j$-th and $j'$-th.
  In this case, the cone $\overline{\mathbf{K}}$ is cut off the linear subspace $T_{\mathrm{s}}$ by two homogeneous linear constraints:

$$\overline{\mathbf{K}} = \left\{d \in T_{\mathrm{s}} : \ a^\top d \leq 0, \ b^\top d \leq 0\right\},$$

  where $a$ and $b$ are the orthoprojections of $\nabla g_j(x_*)$ and $\nabla g_{j'}(x_*)$ onto $T_{\mathrm{s}}$. Note that $a$ and $b$ are linearly independent due to the regularity of $x_*$ and the origin of $T_{\mathrm{s}}$. As a result, a homogeneous quadratic form $d^\top Hd$ is nonnegative/positive outside of the origin on the cone $\overline{\mathbf{K}}$ if and only if it is so everywhere on the set $\{d \in T_{\mathrm{s}} : \ d^\top \underbrace{[ab^\top + ba^\top]}_{:=E} d \geq 0\}$, or, invoking $\mathcal{S}$-lemma (Lemma IV.23.7), if and only if there exists $\theta \geq 0$ such that the quadratic form $d^\top[H - \theta E]d$ is nonnegative/positive outside of the origin on the linear subspace $T_{\mathrm{s}}$. The bottom line is that in the case in question optimality conditions $\mathbf{N}^+$ and $\mathbf{S}^+$ are efficiently verifiable and may "outperform" the standard optimality conditions stated in propositions IV.27.3, IV.27.4. In order to see an example of when they may "outperform," look what the conditions say in the case of the problem $\min_{x_1,x_2}\{f(x) : x_1 \leq 0, x_2 \leq 0\}$ when $x_* = 0$ and $\nabla f(0) = 0$.

**B.** Finally, we remark that on the closest inspection, the first order part of Proposition IV.27.3 remains a necessary condition for local optimality in the case when we relax our Default Assumption by replacing twice continuous differentiability of the objective and the constraints in a neighborhood of a regular solution $x_*$ with plain continuous differentiability; the validity of this modification is readily given by inspecting the relevant parts of the above derivations and the Implicit function Theorem applied with $\kappa = 1$ rather than with $\kappa = 2$. The resulting "truncated" version of proposition IV.27.3 is called the necessary First Order Optimality condition.

Aside from rare cases when the optimality conditions allow to find an optimal solution in closed analytical form, their role in traditional Mathematical Programming is in "guiding" algorithms and justifying their convergence properties. Specifically, when the current iterate of an iterative algorithm does not satisfy the necessary optimality condition (e.g., the gradient of smooth objective which we want to minimize over the entire space is nonzero), the condition usually suggests a way to "improve" the iterate (say, in the example above moving along the antigradient direction reduces the value of the objective, provided the step is chosen properly), and the algorithm utilizes this improvement and in this sense is guided by the optimality condition. However, in this textbook, we do not touch algorithms at all. Therefore, to illustrate the role of optimality conditions, we consider the rare situation where these conditions allow to establish an important theoretical result, namely, $\mathcal{S}$-Lemma (Lemma IV.23.7).

### *27.3.1 Illustration: $\mathcal{S}$-Lemma revisited*

The proof to follow is incomparably less elegant that the original one; the advantage of the new proof is its "bare hands" nature – it is what you can develop if you do not want to think much.

The only nontrivial part of $\mathcal{S}$-Lemma is the claim that if a homogeneous quadratic inequality

$$x^\top Bx \geq 0, \tag{$B$}$$

is a consequence of strictly feasible quadratic inequality

$$x^\top Ax \geq 0, \tag{$A$}$$

then $B-\theta A \succeq 0$ for properly selected $\theta \geq 0$. The proof of this claim via optimality conditions goes as follows. Define $f(x) := x^\top Bx$ and $h(x) := x^\top Ax - 1$ and consider the following equality constrained optimization problem

$$\text{Opt} := \min_x \{f(x) : \; h(x) = 0\} \tag{$P$}$$

Since $(A)$ is strictly feasible, this problem $(P)$ is feasible, and since $(B)$ is a consequence of $(A)$, we have $\text{Opt} \geq 0$. *Assume for a moment that the problem is solvable*, and let $x_*$ be its optimal solution. Note that $x_*$ is a regular solution to our problem, since $\nabla h(x_*) = 2Ax_* \neq 0$ due to $x_*^\top Ax_* = 1$. Applying Necessary optimality condition, we get a $\mu^*$ such that

$$\begin{aligned}
&\nabla f(x_*) + \mu^* \nabla h(x_*) = 0, \quad \text{and} \\
&d^\top(\nabla^2 f(x_*) + \mu^* \nabla^2 h(x_*))d \geq 0, \quad \forall (d : d^\top \nabla h(x_*) = 0).
\end{aligned}$$

That is, by defining $H := B + \mu^* A$, we arrive at

$$(a) : Hx_* = 0, \quad \text{and} \quad (b) : d^\top Hd \geq 0, \; \forall (d : d^\top Ax_* = 0).$$

Thus, we conclude that $x_*^\top Hx_* = x_*^\top(B + \mu^* A)x_* = 0$, that is, $\text{Opt} + \mu^* = 0$, implying that $\mu^* \leq 0$ due to $\text{Opt} \geq 0$. Next, for any $g \in \mathbf{R}^n$ and setting $\gamma := g^\top Ax_*$ and $d := g - (g^\top Ax_*)x_* = g - \gamma x_*$, we get

$$(c) : d^\top Ax_* = 0$$

due to $x_*^\top A x_* = 1$. Consequently, for all $g \in \mathbf{R}^n$ we have

$$\begin{aligned} g^\top H g &= (d + \gamma x_*)^\top H (g + \gamma x_*) \\ &= \underbrace{d^\top H d}_{\geq 0 \text{ by (c,b)}} + 2\gamma d^\top \underbrace{H x_*}_{\substack{=0 \\ \text{by (a)}}} + \gamma^2 x_*^\top \underbrace{H x_*}_{=0} \geq 0, \end{aligned}$$

that is, $H \succeq 0$. Thus, setting $\theta = -\mu^* \geq 0$, we get $B - \theta A \succeq 0$, as claimed.

It remains to get rid of the assumption that $(P)$ is solvable. To this end let $\epsilon > 0$ and define $B_\epsilon := B + \epsilon I$. When $x$ satisfies $(A)$, we have $x^\top B_\epsilon x = x^\top B x + \epsilon x^\top x \geq \epsilon x^\top x$, so that replacing in $(P)$ matrix $B$ with $B_\epsilon$, we get a feasible optimization problem with nonnegative optimal value and objective tending to $\infty$ along every sequence of feasible solutions with norms tending to $\infty$, implying that the problem is solvable. By the above reasoning applied to $B_\epsilon$ in the role of $B$, there exists $\theta_\epsilon \geq 0$ such that $B_\epsilon \succeq \theta_\epsilon A$. Let $\overline{x}$ be a once forever fixed feasible solution to $(P)$. Then, the feasibility of $\overline{x}$ (i.e., $\overline{x}^\top A \overline{x} = 1$) and $B_\epsilon \succeq \theta_\epsilon A$ together imply $\overline{x}^\top B_\epsilon \overline{x} \geq \theta_\epsilon \overline{x}^\top A \overline{x} = \theta_\epsilon$. By plugging in the definition of $B_\epsilon$, we see that $\overline{x}^\top B \overline{x} + \epsilon \|\overline{x}\|_2^2 \geq \theta_\epsilon$, and so as $\epsilon \to +0$ the sequence $\theta_\epsilon$ remains bounded. Thus, defining $\theta$ to be a limiting point of $\theta_\epsilon$ as $\epsilon \to +0$, we get $\theta \geq 0$, and $B_\epsilon \succeq \theta_\epsilon A$ combines with $B_\epsilon \to B$ as $\epsilon \to +0$ to imply that $B \succeq \theta A$. □

# 28

# Saddle points

### 28.1 Definition and Game Theory interpretation

When speaking about the "saddle point" formulation of optimality conditions in Convex Programming, we touched the topic of Saddle Points, which is very interesting in its own right. Let us recall our situation and the definition of saddle points.

**Definition** IV.28.1 [Saddle points] Let $X \subseteq \mathbf{R}^n$ and $\Lambda \subseteq \mathbf{R}^m$ be two nonempty sets, and let

$$L(x, \lambda) : X \times \Lambda \to \mathbf{R}$$

be a real-valued function of $x \in X$ and $\lambda \in \Lambda$. We say that a point $(x^*, \lambda^*) \in X \times \Lambda$ is a *saddle point* of $L$ on $X \times \Lambda$, if $L$ attains at this point its maximum in $\lambda \in \Lambda$ and attains at the point its minimum in $x \in X$, i.e.,

$$L(x, \lambda^*) \geq L(x^*, \lambda^*) \geq L(x^*, \lambda), \quad \forall (x, \lambda) \in X \times \Lambda. \tag{28.1}$$

The notion of a saddle point admits natural interpretation in *game terms*. Consider what is called a *two-person zero-sum game* where player I chooses $x \in X$ and player II chooses $\lambda \in \Lambda$; after the players have chosen their decisions, player I pays to player II the sum $L(x, \lambda)$. Of course, I is interested to minimize his payment, while II is interested to maximize his income. What is the natural notion of the *equilibrium* in such a game, i.e., what are the choices $(x, \lambda)$ of the players I and II such that every one of the players is not interested to vary his choice independently on whether he knows the choice of his opponent? It is immediately seen that the equilibria are exactly the saddle points of the cost function $L$. Indeed, if the players decisions $(x, y)$ are chosen to be a saddle point $(x^*, \lambda^*)$ satisfying (28.1), then the player I is not interested to pass from $x^*$ to another choice, given that II keeps his choice $\lambda = \lambda^*$ fixed: the first inequality in (28.1) shows that such a choice cannot decrease the payment of I. Similarly, player II is not interested to choose something different from $\lambda^*$, given that I keeps his choice $x = x^*$ as such an action cannot increase the income of II. On the other hand, if the players decisions $(x, \lambda)$ is not a saddle point, then either the player I can decrease his payment passing from $x$ to another choice, given that II keeps his choice at $\lambda$ (this is the case when the first inequality in (28.1) is violated), or similarly for the player II. Thus, we conclude that equilibria are exactly the saddle points.

The game interpretation of the notion of a saddle point motivates deep insight into the structure of the set of saddle points. Consider the following two situations:

(A) player I makes his choice first, and player II makes his choice already knowing the choice of player I;

(B) vice versa, player II chooses first, and player I makes his choice already knowing the choice of player II.

In the case (A) the reasoning of player I is as follows: If I choose some $x$, then player II of course will choose $\lambda$ which maximizes, for my $x$, my payment $L(x,\lambda)$, so that I will pay the sum

$$\overline{L}(x) := \sup_{\lambda\in\Lambda} L(x,\lambda);$$

Consequently, my policy should be to choose $x$ which minimizes my *loss function* $\overline{L}$, i.e., the one which solves the optimization problem

$$\mathrm{Opt}(\mathrm{P}) = \inf_{x\in X} \overline{L}(x); \tag{P}$$

with this policy my anticipated payment will be

$$\mathrm{Opt}(\mathrm{P}) = \inf_{x\in X} \overline{L}(x) = \inf_{x\in X} \sup_{\lambda\in\Lambda} L(x,\lambda).$$

In the case (B), similar reasoning of player II results in his *profit function* to be given by

$$\underline{L}(\lambda) := \inf_{x\in X} L(x,\lambda),$$

and thus player II's objective becomes to maximize (in $\lambda$) $\underline{L}(\lambda)$ resulting in the following optimization problem

$$\mathrm{Opt}(\mathrm{D}) = \sup_{\lambda\in\Lambda} \underline{L}(\lambda). \tag{D}$$

Based on this policy, the anticipated profit of II is given by

$$\mathrm{Opt}(\mathrm{D}) = \sup_{\lambda\in\Lambda} \underline{L}(\lambda) = \sup_{\lambda\in\Lambda} \inf_{x\in X} L(x,\lambda).$$

Note that these two reasonings relate to two *different* games: the one with priority of player II (when making his decision, player II already knows the choice of player I) in the case of (A), and the one with similar priority of player I in the case of (B). Therefore, we should not, generally speaking, expect that the anticipated loss of player I in (A) is equal to the anticipated profit of player II in (B). What can be guessed is that the anticipated loss of player I in (B) is *less than or equal to* the anticipated profit of player II in (A), since the conditions of the game (B) are better for player I than those of (A). Thus, we may guess that *independent of any structural property of the function* $L(x,\lambda)$, the following inequality holds:

$$\mathrm{Opt}(\mathrm{D}) = \sup_{\lambda\in\Lambda} \inf_{x\in X} L(x,\lambda) \le \mathrm{Opt}(\mathrm{P}) = \inf_{x\in X} \sup_{\lambda\in\Lambda} L(x,\lambda). \tag{28.2}$$

This inequality indeed is true, which is seen from the following reasoning:

$$\begin{array}{lll} & \inf_{x\in X} L(x,\lambda) \le L(y,\lambda), & \forall y \in X \\ \Longrightarrow & \sup_{\lambda\in\Lambda} \inf_{x\in X} L(x,\lambda) \le \sup_{\lambda\in\Lambda} L(y,\lambda) =: \overline{L}(y), & \forall y \in X. \end{array}$$

Consequently, the quantity $\sup_{\lambda\in\Lambda} \inf_{x\in X} L(x,\lambda)$ is a lower bound for the function $\overline{L}(y)$ for all $y \in X$, and therefore it is a lower bound for the infimum of the latter function over $y \in X$, i.e., it is a lower bound for $\inf_{y\in X} \sup_{\lambda\in\Lambda} L(y,\lambda)$.

Now let us look at what happens when the game in question has a saddle point $(x^*,\lambda^*)$, so that the following relations hold:

$$L(x,\lambda^*) \ge L(x^*,\lambda^*) \ge L(x^*,\lambda), \quad \forall (x,\lambda) \in X \times \Lambda. \tag{28.3}$$

We claim that if it is the case, then we have the following property:

> (*) $x^*$ *is an optimal solution to* (P)*,* $\lambda^*$ *is an optimal solution to* (D) *and the optimal values in these two optimization problems are equal to each other* (and are equal to the quantity $L(x^*,\lambda^*)$).

Indeed, from (28.3) it follows that

$$\underline{L}(\lambda^*) \ge L(x^*,\lambda^*) \ge \overline{L}(x^*),$$

hence, of course,

$$\sup_{\lambda\in\Lambda} \underline{L}(\lambda) \ge \underline{L}(\lambda^*) \ge L(x^*,\lambda^*) \ge \overline{L}(x^*) \ge \inf_{x\in X} \overline{L}(x).$$

On the other hand, from (28.2) we deduce that $\sup_{\lambda\in\Lambda} \underline{L}(\lambda) \le \inf_{x\in X} \overline{L}(x)$, which is possible if and only if all the inequalities in the chain are equalities, which is exactly what is said in property (*)..

Thus, if $(x^*,\lambda^*)$ is a saddle point of $L$, then (*) takes place. We are about to demonstrate that the inverse is also true.

**Theorem** IV.28.2 [Structure of the set of saddle points] Let $L : X \times Y \to \mathbf{R}$ be a function. The function $L$ has a saddle point if and only if the related optimization problems (P) and (D) are solvable and $\mathrm{Opt}(P) = \mathrm{Opt}(D)$. In such a case, the saddle points of $L$ are exactly all pairs $(x^*,\lambda^*)$ with $x^*$ being an optimal solution to (P) and $\lambda^*$ being an optimal solution to (D), and the value of the cost function $L(\cdot,\cdot)$ at every one of these points is equal to the common optimal value in (P) and (D).

**Proof.** We have already established "half" of the theorem: if there are saddle points of $L$, then their components are optimal solutions to (P), respectively, (D), and the optimal objective values, Opt(P) and Opt(D), of these two problems are equal to each other and to the value of $L$ at the saddle point in question.

To complete the proof, we should demonstrate that if $x^*$ is an optimal solution to (P), $\lambda^*$ is an optimal solution to (D) and the optimal objective values of these two problems are equal to each other, then $(x^*, \lambda^*)$ is a saddle point of $L$. But, this is also immediate as for all $x \in X,\ \lambda \in \Lambda$ we have

$$L(x, \lambda^*) \geq \underline{L}(\lambda^*) = \overline{L}(x^*) \geq L(x^*, \lambda).$$

Here, the first inequality holds by definition of $\underline{L}$, the equality holds by the assumption that the optimum objective values of the problems (P) and (D) are the same, and the last inequality follows from the definition of $\overline{L}$. Hence,

$$L(x, \lambda^*) \geq L(x^*, \lambda) \quad \forall x \in X, \lambda \in \Lambda.$$

By substituting $\lambda = \lambda^*$ in the right hand side of this inequality, we get $L(x, \lambda^*) \geq L(x^*, \lambda^*)$, and substituting $x = x^*$ in the right hand side of our inequality, we get $L(x^*, \lambda^*) \geq L(x^*, \lambda)$. Thus, $(x^*, \lambda^*)$ is indeed a saddle point of $L$. □

## 28.2 Existence of Saddle Points: Sion-Kakutani Theorem

It is easily seen that a "quite respectable" function $L$ on a direct product–type convex domain may have no saddle points.

**Example** IV.28.1 Consider the function $L(x, \lambda) := (x - \lambda)^2$ on the unit square $[0, 1] \times [0, 1]$. Then, we have

$$\begin{aligned} \overline{L}(x) &= \sup_{\lambda \in [0,1]} (x - \lambda)^2 = \max\{x^2, (1 - x)^2\}, \\ \underline{L}(\lambda) &= \inf_{x \in [0,1]} (x - \lambda)^2 = 0,\ \lambda \in [0, 1], \end{aligned}$$

so that the optimal objective value of (P) is $\frac{1}{4}$, and the optimal objective value of (D) is 0. Hence, Theorem IV.28.2 implies that $L$ has no saddle points.

On the other hand, there are generic cases when $L$ has a saddle point. For example, when

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) : X \times \mathbf{R}^m_+ \to \mathbf{R}$$

is the Lagrange function of a solvable convex program satisfying the Slater condition. Setting $\Lambda = \mathbf{R}^m_+$, we get

$$\overline{L}(x) = \left\{ \begin{array}{ll} f(x) & , g_j(x) \leq 0,\ j \leq m \\ +\infty & , \text{otherwise} \end{array} \right. : X \to \mathbf{R} \cup \{+\infty\},$$

so that, in the notation from Theorem IV.28.2, problem (P) stemming from $L, \Lambda, X$ is (equivalent to) the problem

$$\min_x \{f(x) : g_j(x) \leq 0, j \leq m, x \in X\} \qquad (*)$$

underlying the Lagrange function in question, and problem $(D)$ is the standard Lagrange dual of $(*)$, cf. section 22.3. Theorem IV.24.1 states that in the case in question saddle points exist, which, in particular, combines with Theorem IV.28.2 to imply that the value of $L$ at a saddle point is equal to the optimal value of

$(P)$, that is, of $(*)$ – a fact that was not explicitly articulated in Theorem IV.24.1 and which we will use later.

Note that in the case in question $\Lambda = \mathbf{R}^m_+$ and $X$ are convex, and $L$ is convex in $x$ for every fixed $\lambda \in \Lambda$ and is linear (and therefore concave) in $\lambda$ for every fixed $x \in X$. As we shall see in a while, these structural properties of $L$ –convexity in $x$ for every fixed $\lambda$ and concavity in $\lambda$ for every fixed $x$– take upon themselves the "main responsibility" for the fact that in the case in question the saddle points exist. We state this formally as the following result.

**Theorem** IV.28.3 [Existence of saddle points of a convex-concave function (Sion-Kakutani)] Let $X$ and $\Lambda$ be nonempty convex compact sets in $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively, and let

$$L(x, \lambda) : X \times \Lambda \to \mathbf{R}$$

be a continuous function which is convex in $x \in X$ for every fixed $\lambda \in \Lambda$ and is concave in $\lambda \in \Lambda$ for every fixed $x \in X$. Then, $L$ has saddle points on $X \times \Lambda$.

In the proof of Theorem IV.28.3 we will use a basic fact from Analysis.

**Fact** IV.28.4 Let $X$ and $\Lambda$ be nonempty compact sets in $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively, and let $L(x, \lambda) : X \times \Lambda \to \mathbf{R}$ be a continuous function. Then, the problems (P) and (D) are solvable.

In order to prove Theorem IV.28.3, we need one more critical ingredient that is quite significant on its own right.

**Lemma** IV.28.5 [Minimax Lemma] Let $X$ be a nonempty convex compact set, and let $f_0, \ldots, f_N$ be a collection of $N+1$ convex and continuous functions on $X$. Then, there exist convex combination weights $\mu_i \in \mathbf{R}_+$, $i = 0, \ldots, N$ such that $\sum_{i=0}^N \mu_i = 1$ and

$$\min_{x \in X} \max_{i=0,\ldots,N} f_i(x) = \min_{x \in X} \sum_{i=0}^N \mu_i f_i(x).$$

**Remark** IV.28.6 Minimum of *every* convex combination of a collection of *arbitrary* functions is less than or equal to the minimax of the collection. This evident fact can be also obtained from applying (28.2) to the function

$$M(x, \mu) := \sum_{i=0}^N \mu_i f_i(x)$$

on the direct product of $X$ and the standard simplex

$$\Delta = \left\{ \mu \in \mathbf{R}^{N+1} : \ \mu_i \geq 0, \forall 0 \leq i \leq N, \ \sum_{i=0}^N \mu_i = 1 \right\}.$$

The interesting part of the Minimax Lemma states that if $f_i$ are convex and continuous on a convex compact set $X$, then the indicated inequality is in fact equality. You can easily verify that this is nothing but the claim that the function $M$ possesses a saddle point. Thus, the Minimax Lemma is in fact a particular case of the Sion-Kakutani Theorem (i.e., Theorem IV.28.3).

We will first give a direct proof of this particular case of the Theorem IV.28.3 stated in Lemma IV.28.5 and then use it to prove the general case given in Theorem IV.28.3.

## 28.3 Proof of Sion-Kakutani Theorem

### *28.3.1 Proof of Minimax Lemma*

Consider the optimization program

$$\min_{t,x} \{t : \ x \in X, \ f_i(x) - t \le 0, \forall i = 0, \ldots, N\}. \tag{S}$$

This is clearly a convex problem with the optimal objective value equal to

$$t^* := \min_{x\in X} \max_{i=0,\ldots,N} f_i(x).$$

Note that $(t, x)$ is feasible solution for $(S)$ if and only if $x \in X$ and $t \ge \max\limits_{i=0,\ldots,N} f_i(x)$. Problem $(S)$ clearly satisfies the Slater condition and is solvable (since $X$ is a compact set and $f_i$, $i = 0, \ldots, N$, are continuous on $X$; therefore their maximum is also continuous on $X$ and thus attains its minimum on the compact set $X$). Let $(t^*, x^*)$ be an optimal solution to Problem $(S)$. According to Theorem IV.24.1, there exists $\lambda^* \ge 0$ such that $((t^*, x^*), \lambda^*)$ is a saddle point of the corresponding Lagrange function

$$L(t, x; \lambda) = t + \sum_{i=0}^{N} \lambda_i (f_i(x) - t) = t\left(1 - \sum_{i=0}^{N} \lambda_i\right) + \sum_{i=0}^{N} \lambda_i f_i(x),$$

on $(t, x; \lambda) \in [\mathbf{R}\times X]\times\mathbf{R}^{N+1}_+$ and the value of this function at $((t^*, x^*), \lambda^*)$ is equal to $t^*$, i.e., the optimal objective value of $(S)$, see discussion preceding Theorem IV.28.3.

Now, since $L(t, x; \lambda^*)$ attains its minimum in $(t, x)$ over the set $\mathbf{R}\times X$ at $(t^*, x^*)$, we should have

$$\sum_{i=0}^{N} \lambda_i^* = 1$$

(otherwise the minimum of $L$ in $(t, x)$ would be $-\infty$ (look what happens when $\sum_i \lambda_i < 1$ and $t \to -\infty$, and what happens when $\sum_i \lambda_i > 1$ and $t \to +\infty$). Thus,

$$\left(\min_{x\in X} \max_{i=0,\ldots,N} f_i(x)\right) = \quad t^* = \min_{t\in\mathbf{R},x\in X} \left(t \cdot 0 + \sum_{i=0}^{N} \lambda_i^* f_i(x)\right),$$

so that

$$\min_{x\in X}\max_{i=0,\ldots,N} f_i(x) = \min_{x\in X}\sum_{i=0}^{N}\lambda_i^* f_i(x)$$

for some $\lambda_i^* \geq 0,\ \forall 0 \leq i \leq N$ satisfying $\sum_{i=0}^{N}\lambda_i^* = 1$, as desired. □

We are now ready to prove Theorem IV.28.3, a.k.a., Sion-Kakutani Theorem.

### *28.3.2 From Minimax Lemma to the proof of Sion-Kakutani Theorem*

Based on Theorem IV.28.2, all we need to prove is that

(i) the optimization problems (P) and (D) are solvable, and
(ii) the optimal values in (P) and (D) are equal to each other.

In fact, (i) is valid independent of convexity-concavity of $L$ and the convexity of the sets $X$ and $\Lambda$, and it is simply given by Fact IV.28.4.

The essence of the matter in the proof of Theorem IV.28.3 lies in (ii). In order to prove (ii), of course, we will heavily exploit convexity-concavity of $L$ and the convexity of the sets $X$ and $\Lambda$. We will prove this in several steps and use the Minimax Lemma (i.e., Lemma IV.28.5) in the last step.

$0^0$. We need to prove that the optimal objective values of (P) and (D) (which, by (i), are well defined real numbers) are equal to each other, i.e., the following relation holds:

$$\inf_{x\in X}\sup_{\lambda\in\Lambda} L(x,\lambda) = \sup_{\lambda\in\Lambda}\inf_{x\in X} L(x,\lambda).$$

We already know from (28.2) (which by the way makes no structural assumptions on $L,\ X$, or $\Lambda$) that

$$\inf_{x\in X}\sup_{\lambda\in\Lambda} L(x,\lambda) \geq \sup_{\lambda\in\Lambda}\inf_{x\in X} L(x,\lambda).$$

So, all we need is to prove the reverse inequality. Without loss of generality we can assume that $\mathrm{Opt}(D)$ is zero, i.e., $\mathrm{Opt}(D) = \sup\limits_{\lambda\in\Lambda}\inf\limits_{x\in X} L(x,\lambda) = 0$ (we can achieve this by shifting the function $L$ by a constant quantity if needed). Then, all we need to prove is that $\mathrm{Opt}(P)$ cannot be positive, i.e., $\mathrm{Opt}(P) = \inf\limits_{x\in X}\sup\limits_{\lambda\in\Lambda} L(x,\lambda) \leq 0$.

$1^0$. What does it mean that $\mathrm{Opt}(D)$ is zero? When $\mathrm{Opt}(D) = 0$, then the function $\underline{L}(\lambda) = \inf\limits_{x\in X} L(x,\lambda)$ is nonpositive for every $\lambda \in \Lambda$, or, equivalently, for every $\lambda \in \Lambda$, the function $L(x,\lambda)$, which is a convex and continuous function of $x \in X$, has nonpositive minimal value over $x \in X$. Since $X$ is compact, this minimal value is achieved, so that for any $\lambda \in \Lambda$ the set

$$X(\lambda) := \{x \in X :\ L(x,\lambda) \leq 0\}$$

is nonempty. Moreover, as $X$ is convex and for every $\lambda \in \Lambda$ the function $L$ is convex in $x \in X$, the set $X(\lambda)$ is convex (it is nothing but the sublevel set of a convex function, so Proposition III.13.6 applies here). Note also that the set $X(\lambda)$ is closed since $X$ is closed and $L(x, \lambda)$ is continuous in $x \in X$. Thus, if $\mathrm{Opt}(D) = 0$, then the set $X(\lambda)$ is a nonempty convex compact set for every $\lambda \in \Lambda$.

$2^0$. What does it mean that $\mathrm{Opt}(P) \leq 0$? It means exactly that there is a point $x \in X$ where the function $\overline{L}$ is nonpositive, i.e., there exists a point $x \in X$ where $L(x, \lambda) \leq 0$ for all $\lambda \in \Lambda$. In other words, to prove that $\mathrm{Opt}(P) \leq 0$ is the same as to prove that *the sets $X(\lambda)$, $\lambda \in \Lambda$, have a point in common.*

$3^0$. With the above observations we see that the situation is as follows: we are given a family of nonempty closed convex subsets $X(\lambda)$, $\lambda \in \Lambda$, of a compact set $X$, and we need to prove that these sets have a point in common. By Helly Theorem II, in order to prove that all $X(\lambda)$ have a point in common, it suffices to prove that every $(N+1)$ sets of this family, where $N := \dim(X)$, have a point in common. Let $X(\lambda_0), \ldots, X(\lambda_N)$ be $N+1$ sets from our family; we should prove that the sets have a point in common. To this end, by defining

$$f_i(x) := L(x, \lambda_i),\ i = 0, \ldots, N;$$

all we need to prove is that there exists a point $x$ where all our functions are nonpositive, or, equivalently prove that the minimax of our collection of functions $f_i$ for $i = 0, \ldots, N$, i.e., the quantity

$$\alpha := \min_{x \in X} \max_{i=0,\ldots,N} f_i(x),$$

is nonpositive.

Note that as $L$ is convex and continuous in $x$, all of the functions $f_i$ are convex and continuous. Since $X$ is compact and all $f_i$ are convex and continuous, we can then apply the Minimax Lemma (i.e., Lemma IV.28.5) and deduce that $\alpha$ is the minimum in $x \in X$ of certain convex combination $\phi(x) := \sum_{i=0}^N \nu_i f_i(x)$ (where $\nu_i \geq 0, \sum_i \nu_i = 1$) of the functions $f_i(x)$. Thus, we arrive at

$$\phi(x) = \sum_{i=0}^N \nu_i f_i(x) = \sum_{i=0}^N \nu_i L(x, \lambda_i) \leq L\left(x, \sum_{i=0}^N \nu_i \lambda_i\right),$$

where the last inequality follows from the concavity of $L$ in $\lambda$ (this is the only –and crucial– point where we use this assumption). Then, we see that $\phi(\cdot)$ is majorized by $L(\cdot, \bar{\lambda})$ where $\bar{\lambda} := \sum_{i=0}^N \nu_i \lambda_i$ for come convex combination weights $\nu_i$, so that $\bar{\lambda} \in \Lambda$ by convexity of $\Lambda$. Thus, the minimum of $\phi$ in $x \in X$ –and we already know that this minimum is exactly $\alpha$– is nonpositive (recall that the minimum of $L$ in $x$ is nonpositive for every $\lambda \in \Lambda$). □

## 28.4 Sion-Kakutani Theorem: A refinement

The next theorem lifts the assumption of boundedness of $X$ and $\Lambda$ in Theorem IV.28.3 – now only one of these sets need to be bounded – at the price of some weakening of the conclusion. In particular, when only one of the sets is bounded, we can no longer claim to be able to find the associated saddle points (as they may not exist), yet we can switch the order in which we take the minimum over $x$ and the maximum over $\lambda$, i.e., the optimum objective values of the associated problems are still the same.

**Theorem** IV.28.7 [Swapping min and max in convex-concave saddle point problem (Sion-Kakutani)] Let $X$ and $\Lambda$ be nonempty convex sets in $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively, and suppose that $X$ is compact. Let

$$L(x,\lambda) : X \times \Lambda \to \mathbf{R}$$

be a continuous function which is convex in $x \in X$ for every fixed $\lambda \in \Lambda$ and is concave in $\lambda \in \Lambda$ for every fixed $x \in X$. Then,

$$\inf_{x\in X} \sup_{\lambda\in\Lambda} L(x,\lambda) = \sup_{\lambda\in\Lambda} \inf_{x\in X} L(x,\lambda). \tag{28.4}$$

**Proof.** Once again, we already know from (28.2) (which by the way makes no structural assumptions on $L,\ X$, or $\Lambda$) that

$$\inf_{x\in X} \sup_{\lambda\in\Lambda} L(x,\lambda) \geq \sup_{\lambda\in\Lambda} \inf_{x\in X} L(x,\lambda).$$

Hence, there is nothing to prove when the right had side in (28.4) is $+\infty$. So, we assume that this is not the case. Since $X$ is compact and $L$ is continuous in $x \in X$, $\inf_{x\in X} L(x,\lambda) > -\infty$ for every $\lambda \in \Lambda$, so that the right hand side in (28.4) cannot be $-\infty$, either. Then, $\sup\limits_{\lambda\in\Lambda} \inf\limits_{x\in X} L(x,\lambda) \in (-\infty,\infty)$, and thus it must be a real number. By shifting the function $L$ by a constant (if needed), without loss of generality we can assume that this real number is 0, i.e.,

$$\sup_{\lambda\in\Lambda} \inf_{x\in X} L(x,\lambda) = 0.$$

All we need to prove now is that the left hand side in (28.4) is nonpositive. Assume, on the contrary, that it is positive and thus it is strictly greater than some number $c > 0$. Then, for every $x \in X$ there exists $\lambda_x \in \Lambda$ such that $L(x,\lambda_x) > c$. By continuity, there exists a neighborhood $V_x$ of $x$ in $X$ (i.e., the intersection with $X$ of an open set containing $x$) such that $L(x',\lambda_x) \geq c$ for all $x' \in V_x$. Since $X$ is compact, we can find finitely many points $x_1,\ldots,x_p$ in $X$ such that the union set given by $\bigcup_{i=1}^p V_{x_i}$ is exactly $X$. Then, we deduce $\max_{1\leq i\leq p} L(x,\lambda_{x_i}) \geq c$ for every $x \in X$.

Now, define

$$\bar{\Lambda} := \operatorname{Conv}\{\lambda_{x_1},\ldots,\lambda_{x_n}\}.$$

Then, $\bar{\Lambda}$ is compact by design and $\max_{\lambda\in\bar{\Lambda}} L(x,\lambda) \geq c$ for every $x \in X$, so that $c \leq \min_{x\in X} \max_{\lambda\in\bar{\Lambda}} L(x,\lambda)$. We can now apply Theorem IV.28.3 to $L$ and the convex *compact* sets $X$ and $\bar{\Lambda}$ to get the equality in the following chain:

$$c \leq \min_{x\in X} \max_{\lambda\in\bar{\Lambda}} L(x,\lambda) = \max_{\lambda\in\bar{\Lambda}} \min_{x\in X} L(x,\lambda) \leq \sup_{\lambda\in\Lambda} \min_{x\in X} L(x,\lambda) = 0.$$

This implies $c \leq 0$ and gives the desired contradiction (recall that $c > 0$). □

In the case when one of the sets, say $\Lambda$ is unbounded, a slightly stronger premise of Theorem IV.28.7 still allows us to replace (28.4) with the existence of a saddle point.

**Theorem** IV.28.8 [Existence of saddle point in convex-concave saddle point problem (Sion-Kakutani, Semi-Bounded case)] Let $X$ and $\Lambda$ be nonempty closed convex sets in $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively, and suppose that $X$ is compact. Let

$$L(x,\lambda) : X \times \Lambda \to \mathbf{R}$$

be a continuous function which is convex in $x \in X$ for every fixed $\lambda \in \Lambda$ and is concave in $\lambda \in \Lambda$ for every fixed $x \in X$. Furthermore, assume that for every $a \in \mathbf{R}$, there exist a collection of points $x_1^a, \ldots, x_{n_a}^a \in X$ such that the set

$$\{\lambda \in \Lambda : \; L(x_i^a, \lambda) \geq a, \, 1 \leq i \leq n_a\}$$

is bounded[†]. Then, $L$ has saddle points on $X \times \Lambda$.

[†] This is definitely the case when $L(\bar{x}, \lambda)$ is *coercive* in $\lambda$ for some $\bar{x} \in X$, meaning that the sets $\{\lambda \in \Lambda : L(\bar{x}, \lambda) \geq a\}$ are bounded for every $a \in \mathbf{R}$, or, equivalently, whenever $\lambda_i \in \Lambda$ and $\|\lambda_i\|_2 \to \infty$ as $i \to \infty$, we have $L(\bar{x}, \lambda_i) \to -\infty$ as $i \to \infty$.

**Proof.** Since $X$ is compact and $L$ is continuous, the function

$$\underline{L}(\lambda) = \min_{x \in X} L(x, \lambda)$$

is real-valued and continuous on $\Lambda$. Furthermore, for every $a \in \mathbf{R}$, the set $\{\lambda \in \Lambda : \; \underline{L}(\lambda) \geq a\}$ is clearly contained in the set $\{\lambda : \; L(x_i^a, \lambda) \geq a, \, 1 \leq i \leq n_a\}$ and thus is bounded. Thus, $\underline{L}(\lambda)$ is a continuous function on a closed set $\Lambda$, and its superlevel sets $\{\lambda \in \Lambda : \; \underline{L}(\lambda) \geq a\}$ are bounded. Therefore, $\underline{L}$ attains its maximum on $\Lambda$. Then, by Theorem IV.28.7 we deduce that $\inf_{x \in X} [\overline{L}(x) := \sup_{\lambda \in \Lambda} L(x, \lambda)]$ is finite. Hence, the function $\overline{L}(\cdot)$ is not $+\infty$ identically in $x \in X$. Also, as $L$ is continuous, $\overline{L}$ is lower semicontinuous. Thus, $\overline{L} : X \to \mathbf{R} \cup \{+\infty\}$ is a lower semicontinuous proper (i.e., not identically $+\infty$) function on $X$, and since we additionally have that $X$ is compact, we conclude that $\overline{L}$ attains its minimum on $X$. Thus, both problems $\max_{\lambda \in \Lambda} \underline{L}(\lambda)$ and $\min_{x \in X} \overline{L}(x)$ are solvable, and their optimal objective values are equal by Theorem IV.28.7. Then, by Theorem IV.28.2, we conclude that $L$ has a saddle point. □

# 29

# Exercises for Part IV

## 29.1 Around Conic Duality

**Exercise** IV.1 Given Linear Dynamical System

$$\begin{array}{rcl} x_0 & = & 0 \\ x_{t+1} & = & Ax_t + Bu_t,\ t = 0, 1, \ldots, N-1 \end{array} \tag{LDS}$$

$(A : n \times n, B : n \times m)$ with controls $u_t$ subject to the "energy constraints"

$$\|u_t\|_2 \le 1,\ 0 \le t < N, \tag{EN}$$

pose the problem of maximizing $f^\top x_N$ ($f$ is a given vector) as a conic problem on the product of Lorentz cones, write down the conic dual of this problem, and answer the following questions:

1. Is the problem essentially strictly feasible?
2. Is the problem bounded?
3. Is the problem solvable?
4. Is the dual problem feasible?
5. Is the dual problem solvable?
6. Are the optimal values equal to each other?
7. What do optimality conditions say?

**Exercise** IV.2 ▲ Consider conic constraint $Ax - b \in K$ where $K \subset \mathbf{R}^m$ is a regular cone and matrix $A$ is of full column rank (i.e., has linearly independent columns, or, which is the same, has trivial kernel). Suppose that the constraint is feasible. Show that the following properties are all equivalent to each other:

(i) the feasible region $\{x \in \mathbf{R}^n : Ax - b \in K\}$ is bounded;

(ii) $\mathrm{Im}(A) \cap K = \{\ 0\ \}$, where $\mathrm{Im}(A) := \{Ax : x \in \mathbf{R}^n\}$;

(iii) the following system of vector inequalities is solvable

$$A^\top \lambda = 0, \quad \lambda \in \operatorname{int} K_*.$$

Using these conclude that the property of whether a given conic optimization problem has a bounded feasible region or not is independent of the choice of $b$, provided that the problem is feasible.

**Exercise** IV.3 ♦ Given a cone $K$ in a Euclidean space $E$ with an inner product $\langle \cdot, \cdot \rangle$, we call a pair of elements $x \in K$ and $y \in K_*$ *complementary* if they satisfy $\langle x, y \rangle = 0$.

In this question, we will examine complementarity relations for the second-order cones $\mathbf{L}^n$ and the positive semidefinite cone $\mathbf{S}^n_+$.

1. Consider $\mathbf{L}^n := \left\{x = [\tilde{x}; x_n] \in \mathbf{R}^{n-1} \times \mathbf{R} : x_n \ge \|\tilde{x}\|_2\right\}$; as we know, this cone is self-dual (Example II.8.8). Prove that $x, s \in \mathbf{L}^n$ satisfy $\langle x, s \rangle = 0$ iff $x_n \tilde{s} + s_n \tilde{x} = 0$ holds.

2. Consider the space of $n \times n$ symmetric matrices, i.e., $E = \mathbf{S}^n$ equipped with the Frobenius inner product $\langle X, Y \rangle = \mathrm{Tr}(XY) = \sum_{i=1}^n \sum_{j=1}^n X_{ij} Y_{ij}$. Let $K = \mathbf{S}^n_+ := \{X \in \mathbf{S}^n : x^\top X x \geq 0 \, \forall x \in \mathbf{R}^n\}$ be the positive semidefinite cone; recall that this cone is self-dual (Example II.8.9). Prove that $X, Y \in \mathbf{S}^n_+$ are complementary, i.e., $\langle X, Y \rangle = 0$, iff their matrix product is zero, i.e., $XY = YX = 0$. In particular, matrices from a complementary pair commute and therefore share a common orthonormal eigenbasis.

**Exercise** IV.4 ⧫ By General Theorem on Alternative, a system of $m$ scalar linear constraints $Ax \geq b$ in variables $x \in \mathbf{R}^n$ (or, which is the same, the conic inequality $Ax \geq_{\mathbf{R}^m_+} b$) has no solutions if and only if it can be led to contradiction by aggregation: there exist nonnegative weights $\lambda_1, \ldots, \lambda_m$ such that the associated weighted sum $\lambda^\top Ax \geq \lambda^\top b$ of inequalities from the system is a contradictory inequality, that is, $A^\top \lambda = 0$ and $b^\top \lambda > 0$. For a general conic constraint of the form

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

where $\mathbf{K} \subset \mathbf{R}^m$ is a regular cone, similar recipe for certifying infeasibility would read

$$\exists \lambda \in \mathbf{K}_* : A^\top \lambda = 0 \ \& \ b^\top \lambda > 0. \tag{II}$$

The goal of Exercise is to investigate relation between feasibility statuses of (I) and of (II).

Your first task is easy:

1. Prove that if (II) is feasible, then (I) is infeasible.

The rest of your effort is aimed at investigating to which extent item 1 can be inverted: if and when it is true that when (II) has no solutions, then (I) is feasible? General Theorem on Alternative says that this indeed is the case when $\mathbf{K}$ is the nonnegative orthant $\mathbf{R}^m_+$. In the general case, the situation is different.

2. Let (I) be the univariate conic inequality

$$Ax := [1; 0; 1]x \geq_{\mathbf{L}^3} b := [0; 1; 0] \tag{i}$$

where $\mathbf{L}^3$ is the 3D Lorentz cone. Write down the associated system (II) and check that both this system and (i) are infeasible. Conclude from this example that in general, solvability of (II) is only sufficient, but not necessary, condition for infeasibility of (I).

3. Prove that (II) is infeasible if and only if (I) is *nearly feasible*, meaning that for every $\epsilon > 0$ there exists $b'$ such that $\|b' - b\|_2 \leq \epsilon$ and the conic constraint $Ax \geq_{\mathbf{K}} b'$ is feasible. Equivalently: (II) is infeasible if and only if $b$ belongs to the closure $\overline{B}$ of the set $B = A\mathbf{R}^n - \mathbf{K}$ of those right hand side vectors in (I) for which (I) is feasible.

Conclusion: *Solvability of* (II) *is necessary and sufficient for infeasibility of* (I) *if and only if the set* $B = A\mathbf{R}^n - \mathbf{K}$ *of the right hand sides in the conic constraint* (I) *resulting in constraint's solvability is closed; in fact, solvability of* (II) *is necessary and sufficient condition for* $b$ *to belong to the closure* $\overline{B}$ *of* $B$. Now, when $\mathbf{K} = \{y : Py \geq 0\}$ is a polyhedral cone, e.g., $\mathbf{R}^m_+$, $B$ is polyhedral (since its definition in the case under consideration is its polyhedral representation as well) and therefore closed, which explains why when the cone $\mathbf{K}$ is polyhedral infeasibility of (II) is equivalent to solvability of (I). At the same time, when $\mathbf{K}$ is not polyhedral, $B$ can be non-closed, as is the case in example from item 2. Let us look at the geometry of this example. (i) wants of us to find a point in the intersection of the cone $\mathbf{L}^3 = \{x \in \mathbf{R}^3 : x_3 \geq \sqrt{x_1^2 + x_2^2}\}$ with the line $\ell = \{[t; -1; t] \in \mathbf{R}^3 : t \in \mathbf{R}\}$. $\ell$ belongs to the 2D plane $L = \{x \in \mathbf{R}^3 : x_2 = -1\}$, and the intersection of $\mathbf{L}^3$ with this plane is the set $\{[x_1; -1; x_3] : x_3^2 - x_1^2 \geq 1, x_3 \geq 0\}$, or, which is the same, the set $\{[x_1; -1; x_3] : (x_3 - x_1)(x_3 + x_1) \geq 1, x_3 - x_1 \geq 0\}$; introducing the coordinates $u = x_1 + x_3$, $v = x_1 - x_3$ on the 2D plane $L$, the intersection of $L$ and $\mathbf{L}^3$ in these coordinates becomes the inner part $H = \{[u; v] : u \geq 1/v, v > 0\}$ of the branch $\Gamma = \{[u; v] : uv = 1, v > 0\}$ of hyperbola. In $u, v$-coordinates the line $\ell$ is just the line $v = 0$. Thus, geometrically the situation is as follows: to intersect $\ell$ and $\mathbf{L}^3$ is the same as to intersect $H$ with the $v$-axis of the $[u; v]$-plane; the intersection is clearly empty, so that (i) is infeasible. At the same time, our line is an asymptote of $\Gamma$, so that the shift $v = \epsilon$ of the line $v = 0$ makes the intersection of the shifted line with $H$ nonempty, whatever small $\epsilon > 0$ be. The outlined shift of $\ell$ in our original $x$-coordinates reduces to passing from $b = [0; 1; 0]$ to $b_\epsilon = [0; 1; -\epsilon]$. The bottom line is that $b \notin B$ and $b \in \overline{B}$, since $b = \lim_{\epsilon \to +0} b_\epsilon$ and $b_\epsilon \in B$.

The result of item 3 attracts our attention to the following question: *What are natural sufficient conditions which guarantee the closedness of the set $A\mathbf{R}^n - \mathbf{K}$ ?* Here is a simple answer:

4. Prove that when the only common point of the image space $L := \{y \in \mathbf{R}^m : \exists x : y = Ax\}$ of $A$ and of $\mathbf{K}$ is the origin, the set $B := A\mathbf{R}^n - \mathbf{K} = L - \mathbf{K}$ is closed. Prove that the same holds true when the condition $L \cap \mathbf{K} = \{0\}$ is "heavily violated," meaning that $L \cap \operatorname{int} \mathbf{K} \neq \varnothing$.

**Exercise** IV.5 ♦ [follow-up to Exercise IV.4] Let $\mathbf{K} \subset \mathbf{R}^m$ be a regular cone, $P \in \mathbf{R}^{m \times n}$, $Q \in \mathbf{R}^{m \times k}$, and $p \in \mathbf{R}^m$. Consider the set

$$\overline{K} \quad = \quad \{x \in \mathbf{R}^n : \exists u \in \mathbf{R}^k : Px + Qu + p \in \mathbf{K}\}$$

This set clearly is convex. When the cone $\mathbf{K}$ is polyhedral, the above description of $\overline{K}$ is its polyhedral representation, so that the set $\overline{K}$ is polyhedral and as such is closed.

The goal of this exercise is to understand what happens with closedness of $\overline{K}$ when $\mathbf{K}$ is a general-type regular cone.

1. Is it true that $\overline{K}$ is closed whenever $\mathbf{K}$ is a regular cone?
   *Hint:* Look what happens when $\mathbf{K} = \mathbf{L}^3$, $P = I_3$, $Q = [0; 1; 1] \in \mathbf{R}^{3 \times 1}$, and $p = [0; 0; 0]$
2. Prove when $\mathbf{K}$ is a regular cone and $\operatorname{Im} Q \cap \mathbf{K} = \{0\}$, $\overline{K}$ is closed.

**Exercise** IV.6 ♦ Let $\mathfrak{n}(x)$ be a norm on $\mathbf{R}^n$ such that $\mathfrak{n}$ is continuously differentiable outside of the origin, and let

$$\mathfrak{n}_*(y) = \max_x \{y^\top x : \mathfrak{n}(x) \leq 1\}.$$

be the norm conjugate to $\mathfrak{n}$ (see Fact III.17.4), so that $\mathfrak{n}_*(\cdot)$ is a norm such that

$$x^\top y \leq \mathfrak{n}(x)\mathfrak{n}_*(y)\, \forall x, y \in \mathbf{R}^n$$

and $(\mathfrak{n}_*)_* = \mathfrak{n}$, implying that for every $x \neq 0$ there exists $y \neq 0$ such that

$$x^\top y = \mathfrak{n}(x)\mathfrak{n}_*(y).$$

Here are your tasks:

1. Let $M$ be a $d \times d$ matrix, $d \geq 2$, with diagonal entries equal to 1. Assume that $M\lambda \leq 0$ for some nonzero vector $\lambda \geq 0$. How large could be $\min_{i,j} M_{ij}$ ?
2. For $d \geq 2$, let $p_1, \ldots, p_d$ be $\mathfrak{n}_*(\cdot)$-unit vectors, $w_1, \ldots, w_d$ be $\mathfrak{n}(\cdot)$-unit vectors, and let $p_i^\top w_i = 1$, $1 \leq i \leq d$. Assume that $0 \in \operatorname{Conv}\{p_1, \ldots, p_d\}$. How small could be $\max_{i \neq j} \mathfrak{n}(w_i - w_j)$ ?
3. Let $x \in \mathbf{R}^n$ be nonzero.
   1. Let $g = \nabla \mathfrak{n}(x)$.
      1. What is $\mathfrak{n}_*(g)$ ?
      2. What is $g^\top x$ ?
      3. Let $e$ be such that $\mathfrak{n}_*(e) \leq \mathfrak{n}_*(g)$ and $e^\top x = g^\top x$. Is it true that $e = g$ ?
   2. Given $N$ points $y_i \in \mathbf{R}^n$, consider the problem of finding the smallest $\mathfrak{n}(\cdot)$-ball containing $y_1, \ldots, y_N$.
      1. Write down the problem as a conic one, and write down the conic dual of this problem. Are both the problems solvable with equal optimal values?
      2. Assume that the data are such that the optimal value in $(P)$ is equal to 1. How small can be $\max_{i,j} \mathfrak{n}(y_i - y_j)$ ?
         *Hint:* write down and analyze optimality conditions.
      3. In the situation of item 3.2.2, assume that $\mathfrak{n}(x) = \|x\|_2$ is the standard Euclidean norm. How small can be $\max_{i,j} \mathfrak{n}(y_i - y_j)$ now?

### 29.1.1 ★ Geometry of primal-dual pair of conic problems

**Exercise** IV.7 ⧫ [geometry of primal-dual pair of conic problem] The goal of the Exercise is to reveal notable geometry of primal-dual pair of conic problem.

It is convenient to work with the primal problem in the form

$$\mathrm{Opt}(P)=\min_x\left\{c^\top x: Ax-b\geq_{\mathbf{K}}0,\, Px=p\right\} \tag{$P$}$$

where $\mathbf{K}$ is a regular cone in certain $\mathbf{R}^N$. As is immediately seen, the conic dual of $(P)$ reduces to the problem

$$\mathrm{Opt}(D)=\max_{y,z}\left\{b^\top y+p^\top z: y\in\mathbf{K}_*,\, A^\top y+P^\top z=c\right\}\ ^{1} \tag{$D$}$$

From now on we make the following, in fact, rather weak,

**Assumption:** *The systems of linear equality constraints in* $(P)$ *and* $(D)$ *are solvable.*

Let us fix $\overline{x}$ and $(\overline{y},\overline{z})$ such that

$$P\overline{x}=p\ \&\ A^\top\overline{y}+P^\top\overline{z}=c. \tag{\#}$$

Your first task is as follows:

1. Pass in $(P)$ from variables $x$ to *primal slack* $\xi=Ax-b$. Specifically, prove that in terms of primal slack $(P)$ becomes the problem

$$\begin{array}{c}\mathrm{Opt}(\mathcal{P})=\min_\xi\left\{\overline{y}^\top\xi:\xi\in\mathbf{K}\cap[\mathcal{L}-\overline{\xi}]\right\}\\ \left[\mathcal{L}=\{\xi:\exists x:\xi=Ax, Px=0\},\ \overline{\xi}=b-A\overline{x}\right]\end{array} \tag{$\mathcal{P}$}$$

namely, prove that

(i) Every feasible solution $x$ to $(P)$ induces feasible solution $\xi=Ax-b$ to $(\mathcal{P})$, and the value of the objective of $(P)$ at $x$ differs from the value of the objective of $(\mathcal{P})$ at $\xi=Ax-b$ by the independent of $x$ constant:

$$\overline{y}^\top\xi=c^\top x-\left[\overline{y}^\top b+\overline{z}^\top p\right]. \tag{$A$}$$

(ii) Vice versa, every feasible solution $\xi$ to $(\mathcal{P})$ is of the form $Ax-b$ for some feasible solution $x$ to $(P)$.

The bottom line is that $(P)$ can be reformulated equivalently as $(\mathcal{P})$, and the optimal values of these two problems are linked by the relation

$$\mathrm{Opt}(\mathcal{P})=\mathrm{Opt}(P)-\left[\overline{y}^\top b+\overline{z}^\top p\right].$$

Next task is as follows:

[1] building conic dual to a conic problem is a purely mechanical process; however, this process as presented in section 23.4 operates with conic problem in a form slightly different from the one of $(P)$, namely, with linear inequality constraints instead of linear equalities. To apply this process to $(P)$, it suffices to represent the linear equalities $Px=p$ by a pair of opposite linear inequalities $Px-p\geq 0, -Px+p\geq 0$. Applying the recipe from section 23.4 to the resulting problem, the dual reads

$$\max_{y,z',z''}\left\{b^\top y+[z'-z'']^\top A^\top y+P^\top[z'-z'']=c,\, y\in\mathbf{K}_*, z'\geq 0, z''\geq 0\right\}.$$

Passing from $z',z''$ to $z=z'-z''$, we reduce the latter problem to $(D)$.

2. Pass from problem $(D)$ in variables $y$, $z$ to problem

$$\begin{array}{c}\max\limits_{y}\left\{\overline{\xi}^{\top} y : y \in \mathbf{K}_* \cap [\mathcal{L}^{\perp} + \overline{y}]\right\} \\ \left[\mathcal{L}^{\perp} := \{y : y^{\top}\xi = 0\,\forall \xi \in \mathcal{L}\} = \{y : \exists z : A^{\top}y + P^{\top}z = 0\}\right]\end{array} \tag{$\mathcal{D}$}$$

in variable $y$ only, specifically, prove that
(i) The orthogonal complement $\mathcal{L}^{\perp}$ of $\mathcal{L}$ indeed is the linear subspace $\{y : \exists z : A^{\top}y + P^{\top}z = 0\}$.
(ii) $y$-component of feasible solution $(y, z)$ to $(D)$ is a feasible solution to $(\mathcal{D})$, and vice versa – every feasible solution $y$ to $(\mathcal{D})$ can be augmented by $z$ to yield a feasible solution $(y, z)$ to $(D)$. Besides this, whenever $(y, z)$ is feasible for $(D)$, we have

$$b^{\top}y + p^{\top}z = \overline{\xi}^{\top}y + c^{\top}\overline{x}. \tag{$B$}$$

The bottom line is that $(D)$ can be reformulated equivalently as $(\mathcal{D})$, and the optimal values of these two problems are linked by the relation

$$\mathrm{Opt}(\mathcal{D}) = \mathrm{Opt}(D) - c^{\top}\overline{x}.$$

The summary of items 1 and 2 is as follows:

- Primal-dual pair $(P)$, $(D)$ of conic problems reduces to pair of problems $(\mathcal{P})$, $(\mathcal{D})$, "reduces" meaning that feasible solutions $x$ and $(y, z)$ to $(P)$, $(D)$ induce feasible solutions $\xi = Ax - b$ and $y$ to $(\mathcal{P})$, $(\mathcal{D})$, and every pair of feasible solutions to the latter problems can be obtained, in the fashion just described, from a pair of feasible solutions to $(P)$, $(D)$;
- Geometrically, $(\mathcal{P})$, $(\mathcal{D})$ are as follows:

- Problems' data are (a) primal-dual pair of regular cones $\mathbf{K}$, $\mathbf{K}_*$ in some $\mathbf{R}^N$, (b) pair of linear subspaces $\mathcal{L}_{\mathcal{P}}$, $\mathcal{L}_{\mathcal{D}}$ in $\mathbf{R}^N$ which are orthogonal complements to each other, and (c) pair of vectors $\overline{y}, \overline{\xi}$ in $\mathbf{R}^N$.
- $(\mathcal{P})$ is the problem of minimizing linear objective $\overline{y}^{\top}\xi$ over the intersection of the *primal feasible plane* $\mathcal{M}_{\mathcal{P}} := \mathcal{L}_{\mathcal{P}} - \overline{\xi}$ with the cone $\mathbf{K}$, while $(\mathcal{D})$ is the problem of maximizing the linear objective $\overline{\xi}^{\top}y$ over the intersection of the *dual feasible plane* $\mathcal{M}_{\mathcal{D}} := \mathcal{L}_{\mathcal{D}} + \overline{y}$ and the dual cone $\mathbf{K}_*$.

Pay attention to the "nearly perfect" primal-dual symmetry; the only asymmetry is that in the primal feasible plane the shift vector is $-\overline{\xi}$ – minus the vector of coefficients of the objective in $(\mathcal{D})$, while in the dual feasible plane the shift vector is $\overline{y}$ – the vector of coefficients of the objective in $(\mathcal{P})$. This minor asymmetry stems from the fact that by tradition one of the problems (in our presentation, $(\mathcal{P})$) is written as a minimization program, and the other problem from the pair as a maximization one.

In fact, the symmetry can be made perfect, and the objectives – eliminated at all.

3. Consider pairs of problems $(P)$, $(D)$ along with problems $(\mathcal{P})$, $(\mathcal{D})$, and let $x$, $(y, z)$ be feasible solutions to $(P)$, $(D)$, and $\xi$, $y$ – the feasible solutions to $(\mathcal{P})$, $(\mathcal{D})$ induced by $x$ and $(y, z)$, respectively. Prove that the *duality gap*

$$\mathrm{DualityGap}(x; y, z) := c^{\top}x - [b^{\top}y + p^{\top}z]$$

– the difference between the objective of primal problem $(P)$ evaluated at primal feasible solution $x$ and the objective of the dual problem $(D)$ evaluated at the dual feasible solution $(y, z)$ – is nothing but the inner product $\xi^{\top}y$ of $\xi$ and $y$.

Since solving the primal-dual pair $(P)$, $(D)$ is the same as minimizing the duality gap over pairs $(x, (y, z))$ of their feasible solutions, we conclude that

> *Solving $(P), (D)$ is the same as finding in the feasible set $\mathcal{M}_{\mathcal{P}} \cap \mathbf{K}$ of $(\mathcal{P})$ and the feasible set $\mathcal{M}_{\mathcal{D}} \cap \mathbf{K}_*$ of $(\mathcal{D})$ pair of vectors $\xi$, $y$ as close to orthogonality as possible.*

Note that for every pair $\xi \in \mathcal{M}_{\mathcal{P}} \cap \mathbf{K}$, $y \in \mathcal{M}_{\mathcal{D}} \cap \mathbf{K}_*$ we have $\xi \in \mathbf{K}$, $y \in \mathbf{K}_*$, that is, $\xi^\top y \geq 0$; the zero value of the latter inner product means zero duality gap for the associated with $\xi$, $y$ feasible solutions $x$, $(y, z)$ to $(P)$, $(D)$. The latter, by Weak Duality, implies optimality of $x$ in $(P)$ and of $(y, z)$ in $(D)$. By Conic Duality Theorem, the desired orthogonal to each other $\xi \in \mathcal{M}_{\mathcal{P}} \cap \mathbf{K}$ and $y \in \mathcal{M}_{\mathcal{D}} \cap \mathbf{K}_*$ definitely exist, provided that the primal and the dual feasible planes intersect the interiors of the respective cones.

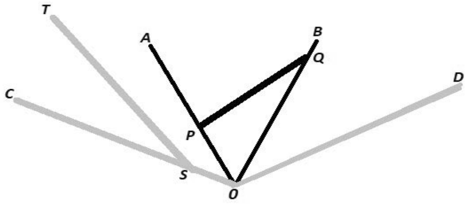


Figure IV.1. Geometry of primal-dual conic pair
$\angle AOB$ – cone $\mathbf{K}$; $\angle COD$ – cone $\mathbf{K}_*$; segment $[P, Q]$ – feasible set of $(\mathcal{P})$;
ray $[ST)$ – feasible set of $(\mathcal{D})$; $Q$ is the primal, and $S$ is the dual optimal solution.
Pay attention to the orthogonality of $\overrightarrow{PQ}$ to $\overrightarrow{ST}$ and of $\overrightarrow{OQ}$ to $\overrightarrow{OS}$.

We comment that the geometric formulation of $(\mathcal{P})$, $(\mathcal{D})$ – "find orthogonal to each other vectors in the intersections of given affine planes with given cones[2]" for the authors, who in their high-school years were taught traditional geometry, sounds as a problem from their old geometry textbooks (modulo the fact that the dimensionality now is arbitrary, and not 2 or 3). It is extremely surprising that in spite of its quite old-fashioned appearance, this geometric problem happens to be responsible for an extremely wide spectrum of applications, ranging from feeding poultry and cattle to decision making, signal and image processing, engineering design, etc., etc.

## 29.2 Around $\mathcal{S}$-Lemma

**Exercise** IV.8 Recall that $\mathcal{S}$-Lemma guarantees that the validity of the implication

$$x^T A x \geq 0 \implies x^T B x \geq 0 \qquad [A, B \in \mathbf{S}^n]$$

is the same as the existence of $\lambda \geq 0$ such that $B \succeq \lambda A$ only under the assumption that the inequality $x^T A x \geq 0$ is strictly feasible. Does the lemma remain true when this assumption is lifted?

**Exercise** IV.9 ♦ Given $A \in \mathbf{S}^n$, consider the set $Q_A = \{x \in \mathbf{R}^n : \ x^\top A x \leq 0\}$.

1 Let $B \in \mathbf{S}^n$ be such that $B \neq A$ and $Q_B = Q_A$. Then, is it always true that there exists $\rho > 0$ such that $B = \rho A$?

2 Suppose that $A \in \mathbf{S}^n$ satisfies $A_{ij} \geq 0$ for all $i, j$. Under this condition, does your answer to item 1 change?

3 Suppose that $A \in \mathbf{S}^n$ satisfies $\lambda_{\min}(A) < 0 < \lambda_{\max}(A)$. Under this condition, does your answer to item 1 change?

[2] not *arbitrary* planes and *arbitrary* cones: the planes should be shifts of linear subspaces which are orthogonal complements of each other, the cones should be duals of each other.

**Exercise** IV.10 ♦ For two nonzero reals $a, b$, one has $2|ab| = \min_{\lambda>0}[\lambda^{-1}a^2 + \lambda b^2]$, implying by the Schur Complement Lemma that $2|ab| \leq c$ if and only if there exists $\lambda > 0$ such that $\left[\begin{array}{c|c} c-\lambda b^2 & a \\ \hline a & \lambda \end{array}\right] \succeq 0$. Assuming $b \neq 0$, we have also $2|ab| \leq c$ if and only if there exists $\lambda \geq 0$ such that $\left[\begin{array}{c|c} c-\lambda b^2 & a \\ \hline a & \lambda \end{array}\right] \succeq 0$. Note also that $c \geq 2|ab|$ is the same as $c \geq 2a\delta b$ for all $\delta \in [-1, 1]$.

Prove the following matrix analogy of the above observation[3]:

Let $A \in \mathbf{R}^{p\times r}$, $B \in \mathbf{R}^{p\times s}$, let $B \neq 0$, and let $\mathcal{D} = \{\Delta \in \mathbf{R}^{r\times s} : \|\Delta\| \leq 1\}$, where $\|\cdot\|$ is the spectral norm. Then $C \succeq [A\Delta B^\top + B\Delta^\top A^\top]$ for all $\Delta \in \mathcal{D}$ if and only if there exists $\lambda \geq 0$ such that $\left[\begin{array}{c|c} C-\lambda BB^\top & A \\ \hline A^\top & \lambda I_r \end{array}\right] \succeq 0$. In particular, when $a, b \in \mathbf{R}^p$ and $b \neq 0$, one has $C \succeq \pm[ab^\top + ba^\top]$ if and only if there exists $\lambda \geq 0$ such that $\left[\begin{array}{c|c} C-\lambda bb^\top & a \\ \hline a^\top & \lambda \end{array}\right] \succeq 0$.

**Exercise** IV.11 ♦ [Robust TTD] [4] Let us come back to TTD problem (5.2). Assume we have solved this problem and have at our disposal the resulting *nominal truss* withstanding best of all, the total truss volume being a given $W > 0$, the load of interest $f$. Now, we cannot ignore the possibility that "in real life" the truss can be affected, aside of the load of interest $f$, by perhaps small, but still nonzero, occasional load composed of forces acting at the free nodes utilized by the nominal truss (think of railroad bridge and wind). In order for our truss to be useful, it should withstand well all small enough occasional loads of this type. Note that our design gives no guarantees of this type – when building the nominal truss, we took into account just one loading scenario $f$.

1. To get impression of potential dangers of "small occasional loads," run numerical study as follows:
   - Compute the optimal console $t^*$ (see "Console design" Exercise I.16)
   - Looking one by one at the free nodes $p^1, ..., p^\mu$ actually used by the nominal console, associate with every one of them single-force occasional load, the corresponding force acting at node under consideration, generate this force as random 2D vector of Euclidean length 0.01 (that is, 1% of the magnitude of the single nonzero force in the load of interest), and compute the compliance of the nominal truss w.r.t. to the resulting occasional load. Conclude that the nominal console can be crushed by small occasional load and is therefore completely impractical.
2. Proposed cure is, of course, to use Robust Optimization methodology – to immunize the truss against small occasional loads, that is, to control its compliance w.r.t. the load of interest *and* all small occasional loads. An immediate question is where the occasional loads should be applied. There is no sense to allow them to act at all free nodes from the original set of tentative nodes – we have all reasons to believe that some, if not most, of these nodes will not be used in the optimal truss, so that we should not bother about forces acting at these nodes. On the other hand, we should take into account occasional loads acting at the nodes actually used by the optimal robust truss, and we do *not* know in advance what these nodes are. A reasonable compromise here as follows. After the nominal optimal truss is built, we can reduce the nodal set to the nodes actually used in this truss, allow for all pair connection of these nodes and resolve the TTD problem on this reduced sets of tentative nodes and tentative bars, now taking into account not only the load of interest, but all small occasional loads distributed along the nodes of our new nodal set. This approach can be implemented as follows.

[3] S. Boyd, L. El Ghaoui, E. Feron, V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory* – SIAM 1994.

[4] Preceding exercises in the TTD series are I.16, I.18, III.9.

- We specify $\overline{\mathcal{V}}$ as the set of virtual displacements of nodes of our reduced nodal set, preserving the original status ("fixed" – "free") of these nodes, and denote by $\overline{f}$ the natural projection of the load of interest on $\overline{\mathcal{V}}$; note that all nonzero blocks in $f$ – those representing nonzero physical forces from the collection specifying $f$ – are inherited by $\overline{f}$, since the free nodes where these nonzero forces are applied should clearly be used by the nominal truss.
- We specify $\mathcal{F}$ as the "ellipsoidal envelope" of $\overline{f}$ and all small in magnitude (measured in $\|\cdot\|_2$-norm) loads from $\overline{\mathcal{V}}$. Specifically, we use $\overline{f}$ as one of the half-axes of $\mathcal{F}$; the other $\overline{M}-1$ half-axes of $\mathcal{F}$ ($\overline{M} = \dim \overline{\mathcal{V}}$) are orthogonal to each other and to $\overline{f}$ vectors from $\overline{\mathcal{V}}$ of $\|\cdot\|_2$-norm $\rho\|\overline{f}\|_2$, where the "uncertainty level" $\rho \in [0,1]$ is a parameter of our construction. Note that
$$\mathcal{F} = \{g = Ph : h^\top h \leq 1\}$$
for properly selected $\overline{M} \times \overline{M}$ matrix $P$.
- We define the *robust compliance* $\overline{\mathcal{C}}(\bar{t})$ of a truss $\bar{t} \in \mathbf{R}^{\overline{N}}_+$ ($\overline{N}$ is the number of bars in our new – reduced – set of tentative bars), as the supremum, over $g \in \mathcal{F}$, of the usual compliances (computed for the new nodal set) of $\bar{t}$ w.r.t. load $g$, and pose the Robust Counterpart of the TTD problem as the problem of minimizing this robust compliance over trusses $\bar{t} \geq 0$ of total volume $W$. Solving this problem, we arrive at the *robust truss*.

An immediate question is how to solve the Robust Counterpart. Those who solved Exercise I.16.3 know that as stated right now, the Robust Counterpart is the *semiinfinite* – with infinitely many convex constraints – optimization program

$$\overline{\mathrm{Opt}} = \min_{\bar{t},\tau} \left\{ \tau : \bar{t} \in \mathbf{R}^{\overline{N}}_+, \sum_{i=1}^{\overline{N}} \bar{t}_i = W, \left[\begin{array}{c|c} \overline{B}\,\mathrm{Diag}\{\bar{t}\}\overline{B}^\top & g \\ \hline g^\top & 2\tau \end{array}\right] \succeq 0, \forall g \in \mathcal{F} \right\} \qquad (\#)$$

where $\overline{B}$ is the matrix built for the new TTD data in the same fashion as the matrix $B$ was built for the original data.
Here go your tasks:

1. Reformulate (#) as a "normal" convex optimization problem – one with efficiently computable convex objective and finitely many explicitly verifiable convex constraints.
2. Solve the Console design version of the latter problem and subject the resulting robust truss to the same tests as those proposed above for quantifying the "real-life" quality of the nominal truss.

## 29.3 Miscellaneous exercises

**Exercise** IV.12 ▲ Find the minimizer of a linear function
$$f(x) = c^\top x$$
on the set
$$V_p = \{x \in \mathbf{R}^n \mid \sum_{i=1}^n |x_i|^p \leq 1\};$$
here $p$, $1 < p < \infty$, is a parameter. What happens with the solution when the parameter becomes 0.5?

**Exercise** IV.13 ▲ Every one of 3 random variables $\xi_1$, $\xi_2$, $\xi_3$ takes values 0 and 1 with probabilities 0.5, and every two of these 3 variables are independent. Is it true that all 3 variables are mutually independent? If not, how large could be probability of the event $\xi_1 = \xi_2 = \xi_3 = 1$?

**Exercise** IV.14 ▲ [computational study] Consider situation as follows: at discrete time instants $t = 1, 2, \dots, T$ we observe the states $y_t \in \mathbf{R}^\kappa$ of dynamical system; our observations are

$$y_t + \sigma\xi_t, t = 1, 2, \dots, T,$$

where $\sigma > 0$ is a given noise intensity and $\xi_t$ are independent across $t$ zero mean Gaussian noises with unit covariance matrix. All we know about the trajectory of the system is that

$$\|y_{t+1} - 2y_t + y_{t-1}\|_2 \leq dt^2\alpha,$$

where $dt > 0$ is the continuous time interval between consecutive discrete time instants; in other words, the Euclidean norm of the (finite-difference approximation of the) acceleration of the system is $\leq \alpha$. Given time delay $d$, we want to estimate the linear form $f^\top y_{T+d}$ of the system's state at time $T + d \geq 1$, and we intend to use a linear estimate

$$\widehat{y} = \sum_{t=1}^{T} h_t^\top \omega_t.$$

1. Write down optimization problem specifying the minimum risk linear estimate, with the risk of an estimate defined as

$$\mathrm{Risk}[\widehat{y}] = \sqrt{\sup_{y \in \mathcal{Y}} \mathbf{E}\{|\widehat{y} - f^\top y_{T+d}|^2\}},$$

where $\mathcal{Y}$ is the set of all trajectories $y = \{y_t, -\infty < t < \infty\}$ satisfying all constraints (!).
2. Use Conic Duality to convert the problem from the previous item into a Conic Quadratic problem.
3. Carry out numerical experimentation with minimum risk linear estimate.

**Exercise** IV.15 ▲ [computational study] Consider the following problem:

> A particle is moving through $\mathbf{R}^d$. Given positions and velocities of the particle at times $t = 0$ and $t = 1$, find the trajectory of the particle on $[0, 1]$ with minimum possible (upper bound) on acceleration.

1. Formulate the (discretized in time version of the) problem as a Conic Quadratic problem and write down its conic dual. Are the problems solvable? Are the optimal values equal to each other? What is said by optimality conditions?
2. Run numerical experiments in 2D and 3D and look at the results.

**Exercise** IV.16 ▲ [computational study] The study offered to you in this Exercise is aimed at answering the following question::

> *A steel rod is heated at time $t = 0$, the magnitude of the temperature being $\leq R$, and is left to cool, the temperature at the endpoints being all the time kept 0. We measure the temperature of the rod at locations $s_i$ and times $t_i > 0$, $1 \leq i \leq m$; the measurements are affected by Gaussian noise with zero mean and covariance matrix $\sigma^2 I_m$. Given the measurements, we want to recover the distribution of temperature of the rod at time $\bar{t} > 0$.*

**Building the model.** With properly selected units of temperature and length (so that the rod becomes the segment $[0, 1]$), evolution of the temperature $u(t, s)$ ($t \geq 0$ is time, $s \in [0, 1]$ is location) is governed by the *Heat equation*

$$\frac{\partial}{\partial t}u(t, s) = \frac{\partial^2}{\partial s^2}u(t, s) \qquad [u(t, 0) = u(t, 1) \equiv 0]$$

It is convenient to represent functions on $[0,1]$ as

$$f(s) = \sum_{k=1}^{\infty} f_k \phi_k(s),\ \phi_k(s) = \sqrt{2}\sin(\pi k s).$$

Functions $\phi_k$ form an orthonormal basis in the space $L_2 = L_2[0,1]$ of square summable real-valued functions on $[0,1]$ equipped with the inner product

$$\langle f, g\rangle = \int_0^1 f(s)g(s)ds,$$

the corresponding norm being $\|f\|_2 = \sqrt{\int_0^1 f^2(s)ds}$.

The claim that functions $\phi_k$ form an orthonormal basis in $L_2$ means that a series

$$\sum_{k=1}^{\infty} f_k \phi_k(s),$$

converges in $\|\cdot\|_2$ to some $f \in L_2$ if and only if $\sum_k f_k^2 < \infty$, and in this case $f_k = \langle f, \phi_k\rangle$; moreover,

$$\langle \sum_k f_k \phi_k(\cdot), \sum_k g_k \phi_k(\cdot)\rangle = \sum_k f_k g_k$$

for all square-summable sequences $\{f_k\}$ and $\{g_\cdot\}$. In particular,

$$u(t,s) = \sum_{k=1}^{\infty} u_k(t)\phi_k(s),\ u_k(t) = \int_0^1 u(t,s)\phi_k(s)ds.$$

Assuming $|u(0,\cdot)| \leq R$, we have

$$\sum_k u_k^2(0) \leq R^2, \tag{29.1}$$

and in terms of the coefficients $u_k(t)$ of the rod's temperature, the Heat equation becomes very simple:

$$\frac{d}{dt}u_k(t) = -\pi^2 k^2 u_k(t) \implies u_k(t) = \exp\{-\pi^2 k^2 t\}u_k(0).$$

As a result, when $t > 0$, the coefficients $u_k(t)$ go to 0 exponentially fast as $k \to \infty$, so that the series

$$\sum_k u_k(t)\phi_k(s)$$

converges to the solution $(t,s)$ of the heat equation not only in $\|\cdot\|_2$, but uniformly on $[0,1]$ as well, implying, due to $\phi_k(0) = \phi_k(1) = 0$, that the series does satisfy the boundary conditions $u(t,0) = u(t,1) = 0$, $t > 0$.

Now our problem can be posed as follows:

*The sequence of coefficients* $\{u_k^t\}_{k=1}^\infty$ *of* $u(t,\cdot)$ *in the orthonormal basis* $\{\phi_k(\cdot)\}_{k\geq 1}$ *of* $L_2$ *evolves according to*

$$u_k^t = \exp\{-\pi^2 k^2 t\} u_k^0,$$

*with*

$$u^0 := \{u_k^0\}_{k\geq 1} \in \mathbf{B} := \{\{c_k\}_{k\geq 1} : \sum_k c_k^2 \leq R^2\}.$$

*Given* $m$ *noisy observations*

$$\omega_i = \Omega_i[u^0] + \sigma\xi_i,\ \Omega_i[u^0] = \sum_{k=1}^\infty \exp\{-\pi^2 k^2 t_i\} u_k^0 \phi_k(s_i),$$

*where* $\xi_1,\dots,\xi_m$ *are independent of each other* $\mathcal{N}(0,1)$ *observation noises, and* $t_i > 0$, $s_i \in [0,1]$ *are given, we want to recover the sequence* $\{u_k^{\bar{t}}\}_{k\geq 1}$.
*We quantify the performance of a candidate estimate* $\omega := (\omega_1,\dots,\omega_m) \mapsto \widehat{u} := \{\widehat{u}_k(\omega)\}_{k\geq 1}$ *by the risk*

$$\mathrm{Risk}[\widehat{u}] = \sqrt{\max_{u^0\in\mathbf{B}} \mathbf{E}_\xi \left\{ \sum_{k\geq 1} [\widehat{u}_k(\Omega_1[u^0]+\xi_1,\dots,\Omega_m[u^0]+\xi_m) - \exp\{-\pi^2 k^2 \bar{t}\} u_k^0]^2 \right\}}$$

*that is,* Risk$^2$ *is the worst, with respect to the distribution of temperature at time* $t = 0$ *of* $\|\cdot\|_2$*-norm not exceeding* $R$*, expected squared norm* $\|\cdot\|_2^2$ *of the recovery error.*

Our last modeling step is to replace infinite sequences $\{u_k^0\}_{k\geq 1}$ with their finite initial segments $\{u_k^0\}_{1\leq k\leq K}$, that is, to approximate the situation by the one where $u_0^k = 0$ when $k > K$. The simplest way to do it is as follows. Let $\ t = \min[\min_i t_i, \bar{t}]$, so that $\ t > 0$. For $u_0 \in \mathbf{B}$ and $K \geq 1$, the magnitude of the total contribution of the coefficients $u_0^k$, $k > K$, to $u(t,s)$ with $t \geq \ t$ does not exceed

$$\sum_{k=K+1}^\infty \max_s |\phi_k(s)| \exp\{-\pi^2 k^2\ t\} |u_0^k| \leq \delta := \sqrt{2}R \sum_{k=K+1}^\infty \exp\{-\pi^2 k^2\ t\}.$$

Given a "really small" tolerance $\bar{\delta} > 0$, say, $\bar{\delta} = 10^{-10}$, we can easily find $K = K(\bar{\delta})$ such that $\delta \leq \bar{\delta}$. Thus, as far as the temperatures we measure and the temperatures we want to recover are concerned, zeroing out coefficients $u_0^k$ with $k > K(\bar{\delta})$ changes these temperatures by at most $\bar{\delta}$. Common sense (which can be easily justified by formal analysis) says, that with $\bar{\delta}$ as small as $10^{-10}$, these changes have no effect on the quality of our recovery, at least when $\sigma \gg \bar{\delta}$.

Now goes your task:

1. Assuming $u_0^k = 0$ for $k > K$, model the problem of interest as the following estimation problem:

   > *"In the nature" there exists* $K$*-dimensional signal* $u$ *known to belong to the centered at the origin Euclidean ball* $B^R = \{u \in \mathbf{R}^K : u^\top u \leq R^2\}$ *of a given radius* $R$*. Given noisy observations*
   >
   > $$\omega = Au + \sigma\xi, \qquad [A : m\times K, \xi \sim \mathcal{N}(0, I_m)]$$
   >
   > *we want to recover* $Bu$*, quantifying the recovery error of a candidate estimate* $\omega \mapsto \widehat{u}(\omega)$ *by its risk*
   >
   > $$\mathrm{Risk2}[\widehat{u}] = \sqrt{\sup_{u\in B^R} \mathbf{E}_{\xi\sim\mathcal{N}(0,I_m)} \{[\widehat{u}(Au+\sigma\xi) - Bu]^\top [\widehat{u}(Au+\sigma\xi) - Bu]\}}$$
   >
   > *where* $B$ *is a given* $K\times K$ *matrix.*

   Write down the expressions for the matrices $A$ and $B$.

2. Build convex optimization problem responsible for the minimum risk *linear estimate* – estimate of the form $\widehat{u}(\omega) = H^\top \omega$.
3. Compute the minimum risk linear estimate and run simulations to test its performance. Recommended setup:
   - $\bar{t} \in \{0.01, 0.001, 0.0001, 0.00001\}$
   - $m = 100$, $t_i$ are drawn at random from the uniform distribution on $[\bar{t}, 2\bar{t}]$, $s_i$ are drawn at random from the uniform distribution on $[0, 1]$;
   - $R = 10^4$, $\sigma = 10$, $\bar{\delta} = 10^{-10}$;
   - To accelerate computations, truncate $K(\bar{\delta})$ at the level 100.

**Exercise** IV.17 ▲ Given positive definite $A \in \mathbf{S}^n$, let us set

$$P[A] = \{X \in \mathbf{S}^n : X \succeq 0, X^2 \preceq A\},\ Q[A] = \{X \in \mathbf{S}^n : X \succeq 0, X \preceq A^{1/2}.$$

From $\succeq$-monotonicity of the matrix square root on $\mathbf{S}^n_+$ (Example IV.26.5 in section 26.2) it follows that $P[A] \subseteq Q[A]$. Your task is to answer the following question:

> Are $P[A]$ and $Q[A]$ "comparable," meaning that for some $c$ independent of $A$ (but perhaps depending on $n$) one has
> $$Q[A] \subset c \cdot P[A] \qquad ?$$

**Exercise** IV.18 Find the optimal value in the convex optimization problem

$$\mathrm{Opt}(a) = \min_x \left\{ \sum\nolimits_{i=1}^n [-(1 + a_i)x_i + x_i \ln x_i] : x \geq 0, \sum\nolimits_i x_i \leq 1 \right\}$$

where $0 \ln 0 = 0$ by definition, so that the function $x \ln x$ is well defined and continuous on the nonnegative ray $x \geq 0$.

**Exercise** IV.19 ♦ Given $m \times n$ matrix $A$ with trivial kernel, consider the matrix-valued function $F(X) = [A^\top X^{-1} A]^{-1} : \mathrm{Dom}\, F := \{X \in \mathbf{S}^m, X \succ 0\} \to \mathbf{S}^n_+$. Prove that $F$ is $\succeq$-concave on its domain.

## 29.4 Around convex cone-constrained and conic problems

**Exercise** IV.20 ♦ [cone-constrained semidefinite problems]

1. Let $X, Y \in \mathbf{S}^m_+$, Prove that $\mathrm{Tr}(XY) = 0$ if and only if $XY = YX = 0$.
2. Given an ordered collection $\nu = \{n_1, \ldots, n_k\}$ of positive integers, let $\mathbf{S}^\nu$ be the space of block-diagonal symmetric matrices with $k$ diagonal blocks of sizes $n_1 \times n_1, \ldots, n_k \times n_k$, and let $\mathbf{S}^\nu_+$ be the cone of positive semidefinite matrices from $\mathbf{S}^\nu$. Equipping $\mathbf{S}^\nu$ with the Frobenius inner product, $\mathbf{S}^\nu_+$ is clearly a self-dual regular cone in the resulting Euclidean space.
   Convex cone-constrained problem on the cone $\mathbf{S}^\nu_+$ is of the generic form
   $$\mathrm{Opt(SDP)} = \min_{x \in X} \left\{ f(x) : \overline{g}(x) := Ax - b \leq 0,\ \widehat{g}(x) := \mathrm{Diag}\{g_1(x), \ldots, g_k(x)\} \leq_{\mathbf{S}^\nu_+} 0 \right\} \tag{SDP}$$
   where $X$ is a nonempty convex set in some $\mathbf{R}^n$, the function $f : X \to \mathbf{R}$ is convex, and the mapping $\widehat{g} : X \to \mathbf{S}^\nu$ is $\mathbf{S}^\nu_+$-convex.
   Prove that in the case of convex cone-constrained semidefinite problem (SDP) Theorem IV.24.7 reads

**Theorem IV.24.7.SDP** *Consider a convex cone-constrained semidefinite problem* (SDP)*, let $x^* \in X$ be a feasible solution to the problem, and let $f$ and $\widehat{g}$ be differentiable at $x^*$.*
(i) *If $x^*$ is a KKT point of* (SDP)*, the Lagrange multipliers being $\overline{\lambda}^* \geq 0$ and $\widehat{\lambda}^* \in \mathbf{S}^\nu_+$,*

*meaning that*

$$\begin{array}{ll}\overline{\lambda}^*_i[\overline{g}(x^*)]_i = 0\,\forall i \;\&\; \widehat{\lambda}^*\widehat{g}(x^*) = 0 & \text{[sdp complementary slackness]}\\ \nabla_x\left[f(x) + [\overline{\lambda}^*]^\top\overline{g}(x) + \mathrm{Tr}(\widehat{\lambda}^*\widehat{g}(x))\right]\big|_{x=x^*} \in -N_X(x^*) & \text{[ KKT equation]}\end{array}$$

*(here, as always, $N_X(x)$ is the normal cone of $X$, see (15.5)), then $x^*$ is an optimal solution to* (SDP).
(ii) *If $x^*$ is optimal solution to* (SDP) *and, if addition to the above premise,* (SDP) *satisfies the cone-constrained Relaxed Slater condition, then $x^*$ is an sdp KKT point, as defined in* (i),

**Exercise** IV.21 ▲ [follow-up to Exercise IV.20] In the sequel, we fix the dimension $n$ of the embedding space and denote by $E_C = \{x \in \mathbf{R}^n : x^\top Cx \leq 1\}$ the centered at the origin ellipsoid associated with positive definite $n \times n$ matrix $C$. Given positive $K$ and $K$ ellipsoids $E_{A_k}$, $k \leq K$, consider two optimization problems:
— $\mathcal{O}$: find the smallest volume centered at the origin ellipsoid containing $\cup_{k\leq K}E_{A_k}$
— $\mathcal{I}$: find the largest volume centered at the origin ellipsoid contained in $\cap_{k\leq K}E_{A_k}$.

1. Pose $\mathcal{O}$ and $\mathcal{I}$ as solvable convex cone-constrained semidefinite programs
2. Prove that problems $\mathcal{O}$ and $\mathcal{I}$ reduce to each other at the cost of appropriate modification of the data
3. Prove that there exist matrices $\Lambda_k \succeq 0$ such that $\Lambda := \sum_k \Lambda_k \succ 0$ and
$$\Lambda_k = \Lambda_k A_k \Lambda,\; k \leq K.$$

**Exercise** IV.22 ▲ Recall convex cone-constrained problem in Example IV.23.1, section 23.1

$$\mathrm{Opt}(\mathrm{P}) = \min_{x=(t,y)\in\mathbf{R}\times\mathbf{S}^n}\{t : \underbrace{t \geq \mathrm{Tr}(y)}_{\Longleftrightarrow\,\langle y, I_n\rangle - t \leq 0}, y^2 \preceq B\} \tag{23.1}$$

where $B$ is a positive definite matrix.

1. Verify (23.2)
2. Find Lagrange multipliers certifying that $t_* = -\mathrm{Tr}(B^{1/2})$, $y_* = -B^{1/2}$ is a cone-constrained KKT point of problem (23.1) (and thus, by Theorem IV.24.7, is an optimal solution to the problem).
3. Consider the parametric family
$$\mathrm{Opt}(p := (v,w)) = \min_{t\in\mathbf{R},y\in\mathbf{S}^n}\{t : t \geq \mathrm{Tr}(y), yv^{-1}y \preceq w\} \tag{P[$p$]}$$
of convex cone-constrained problems, with $p \in P = \{p = (v,w) : v \in \mathrm{int}\,\mathbf{S}^n_+, w \in \mathrm{int}\,\mathbf{S}^n_+\}$, so that (23.1) is problem (P[$\overline{p}$]) corresponding to
$$\overline{p} = (I_n, B).$$
Prove that $\mathrm{Opt}(p)$ is convex function of $p \in P$ and find a subgradient of this function at the point $\overline{p}$.

**Exercise** IV.23 ▲ [follow-up to Exercise IV.4] Given positive integers $m, n$, consider two parametric families of convex sets:

- $S_1[P] = \{(X,Y) \in \mathcal{R}_1 := \mathbf{S}^m \times \mathbf{S}^n : \left[\begin{array}{c|c} X & P \\ \hline P^\top & Y\end{array}\right] \succeq 0\}$, where the "parameter" $P$ runs through the space $\mathbf{R}^{m\times n}$ of $m \times n$ matrices, let it be temporarily denoted $\mathcal{P}_1$;
- $S_2[P] = \{(X,Y) \in \mathcal{R}_2 := \mathbf{S}^m \times \mathbf{R}^{m\times n} : \left[\begin{array}{c|c} X & Y \\ \hline Y^\top & P\end{array}\right] \succeq 0\}$, where the "parameter" $P$ runs through the positive semidefinite cone $\mathbf{S}^n_+$, let it be temporarily denoted $\mathcal{P}_2$.

Prove that for $\chi = 1, 2$ the set-valued mappings $P \to S_\chi[P]$ are super-additive on their domains:

$$P, Q \in \mathcal{P}_\chi \implies P + Q \in \mathcal{P}_\chi \ \& \ \underbrace{S_\chi[P] + S_\chi[Q] \subset S_\chi[P + Q]}_{(*)}.$$

and that the concluding inclusion
— not necessarily is equality for $\chi = 1$, and
— is equality for $\chi = 2$.

**Exercise** IV.24 In the simplest Steiner problem, one is given $m$ distinct points $a_1, \ldots, a_m$ in $\mathbf{R}^n$ and is looking for a point $x_*$ such that the sum of Euclidean distances between the points and $x_*$ is as small as possible (think, e.g., about $m$ oil wells on 2D plane and the problem of locating collector to be linked to the wells by pipes in a way minimizing the total length of the pipes).

1. Pose the problem as conic problem, the cone being direct product of $m$ Lorentz cones.
2. Build the dual problem. Are the primal and the dual problems solvable? Are the primal and the dual optimal values equal to each other?
3. Write down optimality conditions and see what they say
   *Hint:* You are advised to consider separately the cases where optimal solution differs from all of the points $a_1, \ldots, a_m$, and the case when it is one of the points.
4. Solve the problem in the case when $n = 2$, $m = 3$ and $a_1, a_2, a_3$ are vertices of triangle on 2D plane.
   Note: The point on the plane minimizing the sum of distances to the vertices of a given triangle is called *Fermat point* of the triangle. Quoting “Fermat point” in Wikipedia, ”This question [to minimize the sum of distances from a point to the vertices of triangle] was proposed by Fermat, as a challenge to Evangelista Torricelli. He solved the problem in a similar way to Fermat’s [...] His pupil, Viviani, published the solution in 1659.”

**Exercise** IV.25 ▲ Consider a primal-dual pair of conic problems

$$\mathrm{Opt}(P) = \min_x \left\{ c^\top x : \ Ax \geq_{\mathbf{K}} b \right\} \qquad (P)$$

$$\mathrm{Opt}(D) = \max_y \left\{ b^\top y : \ y \geq_{\mathbf{K}_*} 0, \ A^\top y = c \right\} \qquad (D)$$

($\mathbf{K} \subset \mathbf{R}^n$ is a regular cone) and assume that both problems are feasible.

1. Find the recessive cones $\mathrm{Rec}(P)$ and $\mathrm{Rec}(D)$ of the primal and the dual feasible sets.
2. Prove that the feasible set of at least one of the problems is unbounded.

**Exercise** IV.26 (semidefinite duality) A *semidefinite program* is a conic program involving the positive semidefinite cone. As a matter of fact, *Semidefinite programming* – the family of semidefinite programs – possesses extremely powerful “expressive abilities.” It is prudent to say that *for all practical purposes*, whatever it means, Semidefinite programming is “the same” as the entire Convex programming. In this exercise we would like to acquaint the reader with the specific form taken by Conic duality when the cone involved is the positive semidefinite cone.

Formally, a semidefinite program is of the form

$$\begin{array}{r} \mathrm{Opt}(P) = \min_{x \in \mathbf{R}^n} \left\{ c^\top x : Ax - b := \sum_j a_j x_j - b \geq 0, \right. \\ \left. \mathcal{A}x - B := x_1 A_1 + \ldots + x_n A_n - B \succeq 0 \right\}, \end{array} \qquad (P)$$

where $a_j, b$ are vectors from some $\mathbf{R}^p$, and $A_j, B$ are matrices from some $\mathbf{S}^q$. “Real life” form of a semidefinite program usually is a bit different, namely,

$$\begin{array}{r} \mathrm{Opt}(\mathcal{P}) = \min_{x \in \mathbf{R}^n} \left\{ c^\top x : \ Ax - b := \sum_j a_j x_j - b \geq 0, \right. \\ \left. \mathcal{A}_i x - B^i := x_1 A_1^i + \ldots + x_n A_n^i - B^i \succeq 0, \ \forall i \leq m \right\}, \end{array} \qquad (\mathcal{P})$$

where $A_j^i, B^i \in \mathbf{S}^{q_i}$. In the formulation $(\mathcal{P})$ as opposed to the formulation $(P)$ we have a bunch of positive semidefinite cone constraints, i.e., $\mathcal{A}_i x - B^i \succeq 0$, $i \leq m$, instead of a single constraint

$\mathcal{A}x - B \succeq 0$. We can always rewrite $(\mathcal{P})$ in the form of $(P)$ by assembling $A_j^i$, $B^i$ into block-diagonal matrices $A_j = \mathrm{Diag}\{A_j^1, \ldots, A_j^m\}$, $B = \mathrm{Diag}\{B^1, \ldots, B^m\}$. Taking into account that a block-diagonal symmetric matrix is positive semidefinite if and only if all the diagonal blocks are positive semidefinite, we deduce that $(\mathcal{P})$ is equivalent to the problem

$$\min_{x\in\mathbf{R}^n}\left\{c^\top x:\ Ax - b := \sum_j x_j a_j - b \geq 0,\ \mathcal{A}x - B := \sum_j x_j A_j - B \succeq 0\right\}$$

of the form $(P)$. When proving theorems, it is usually better to work with program in the form of $(P)$ – it saves notation; in contrast, when working with "real life" semidefinite programs, it is usually better to operate with problems in more detailed form $(\mathcal{P})$.

Your task is as follows:

1. Verify that the conic dual of $(\mathcal{P})$ is the semidefinite program

$$\max_{\lambda,\{\Lambda_i,i\le m\}}\left\{b^\top\lambda + \sum_{i=1}^m \mathrm{Tr}(\Lambda_i B^i):\ \begin{array}{c}\lambda\in\mathbf{R}^p_+, \Lambda_i\in\mathbf{S}^{q_i}_+, i\le m\\ A^\top\lambda + \sum_{i=1}^m \mathcal{A}_i^*\Lambda_i = c,\end{array}\right\}, \qquad (\mathcal{D})$$

where for the linear mapping $x \mapsto \sum_j x_j A_j : \mathbf{R}^n \to \mathbf{S}^q$ its conjugate linear mapping $X \mapsto \mathcal{A}^*X : \mathbf{S}^q \to \mathbf{R}^n$ is given by the identity

$$\mathrm{Tr}(X[\mathcal{A}x]) \equiv [\mathcal{A}^*X]^\top x \quad \forall (x\in\mathbf{R}^n, X\in\mathbf{S}^q),$$

or, which is the same,

$$\mathcal{A}^*X = [\mathrm{Tr}(A_1X);\ldots;\mathrm{Tr}(A_nX)].$$

**Exercise** IV.27 ♦ [example of semidefinite relaxation] Let $T_k \succeq 0, k \le K$, be positive semidefinite $m\times m$ matrices such that $\sum_k T_k \succ 0$, $\mathcal{T}\subset\mathbf{R}^K_+$ be a convex compact set intersecting the interior of $\mathbf{R}^K_+$, and $A$ be a symmetric $m\times m$ matrix. Let also $\phi_{\mathcal{T}}(z) = \max_{t\in\mathcal{T}} z^\top t$ be the support function of $\mathcal{T}$. Prove that

$$\begin{array}{rll}\mathrm{Opt} & := & \min_z\left\{\phi_{\mathcal{T}}(z): z\ge 0, A\preceq\sum_k z_kT_k\right\} \qquad (a)\\ & = & \max_{\Lambda,t}\left\{\mathrm{Tr}(A\Lambda):\Lambda\succeq 0, t\in\mathcal{T}, \mathrm{Tr}(T_k\Lambda)\le t_k, k\le K\right\} \quad (b)\end{array}$$

and that both minimization and maximization problems above are solvable.

**Comment.** Exercise IV.27 tells us an interesting story. Consider the problem of maximizing the homogeneous quadratic form $x^\top Ax$ over the set

$$\mathcal{X} = \{x\in\mathbf{R}^m : \exists t\in\mathcal{T} : x^\top T_kx\le t_k,\ k\le K\}$$

with the above restrictions on $T_k$ and $\mathcal{T}$. Such a set is called *ellitope*, and this notion covers many interesting sets, e.g.,

- finite and bounded intersections of centered at the origin ellipsoids/elliptic cylinders; this is what you get when take $\mathcal{T} = \{[1;\ldots;1]\}$. In particular, the intersection of $K$ symmetric with respect to the origin stripes — sets of the form $\{x : |a_k^\top x|\le 1\}$, like the unit box $\{x\in\mathbf{R}^m : \|x\|_\infty\le 1\}$ is an ellitope, provided that the intersection is bounded – set $\mathcal{T} = \{[1;\ldots;1]\}$ and $T_k = a_ka_k^\top$;
- $\|\cdot\|_p$-balls with $2\le p\le\infty$: $\{x\in\mathbf{R}^m : \|x\|_p\le 1\} = \{x\in\mathbf{R}^m : \exists t\in\mathcal{T} : x_k^2\le t_k, k\le K = m\}$ with $\mathcal{T} = \{t\in\mathbf{R}^m_+ : \|t\|_{p/2}\le 1\}$.

It is known that computing the maximum $\mathrm{Opt}_* = \max_{x\in\mathcal{X}} x^\top Ax$ of quadratic form over an ellitope, even as simple as the unit box, is a computationally intractable problem, even when $A$ is restricted to be positive semidefinite. However, the difficult to compute quantity $\mathrm{Opt}_*$ admits the *semidefinite relaxation bound* built as follows: whenever $z\in\mathbf{R}^K_+$ is such that $A\preceq\sum_k z_kT_k$ and $x\in\mathcal{X}$, there exists $t\in\mathcal{T}$ such that $x^\top T_kx\le t_k$, $k\le K$, and we have

$$x^\top Ax \le x^\top\left[\sum_k z_kT_k\right]x = \sum_k z_kx^\top T_kx \le \sum_k z_kt_k \le \phi_{\mathcal{T}}(z),$$

implying that the optimal value Opt of $(a)$ is an upper bound on $\mathrm{Opt}_*$. For reasonable $\mathcal{T}$, in particular, those in the above examples, Opt, in contrast to $\mathrm{Opt}_*$, is efficiently computable.

Now, the fact that Opt is the optimal value not only of $(a)$, but of $(b)$ as well, also admits a useful interpretation. Namely, instead of maximizing $x^\top Ax$ over $x \in \mathcal{X}$, let us maximize the *expectation* of $\xi^\top A\xi$ over *randomized* solutions $\xi$, that is, random vectors $\xi$ which satisfy the constraints specifying $\mathcal{X}$ *at average* — random vectors $\xi$ with distributions $P$ satisfying the condition

$$\exists t = t[P] \in \mathcal{T} : \mathbf{E}_{\xi\sim P}\{\xi^\top T_k\xi\} \leq t_k,\, k \leq K. \qquad (\#)$$

Setting $\Lambda = \mathbf{E}_{\xi\in P}\{\xi\xi^\top\}$, $(\#)$ implies that $\Lambda \succeq 0$ and $\mathrm{Tr}(\Lambda T_k) \leq t_k$, $k \leq K$, for some $t \in \mathcal{T}$, that is, $(\Lambda, t)$ is feasible for $(b)$. Vice versa, if $(\Lambda, t)$ is feasible for $(b)$, then, representing $\Lambda$ as the covariance matrix of random vector $\xi$ (which always can be done, in many ways, since $\Lambda \succeq 0$), we get randomized solution $\xi$ such that $(\xi, t)$ satisfies $(\#)$. The bottom line is that the equality in $(a)$, $(b)$ allows for alternative interpretation of the semidefinite relaxation upper bound Opt on $\mathrm{Opt}_*$ — as the maximal expected value of $\xi^\top A\xi$ over all random vectors $\xi$ belonging to $\mathcal{X}$ "at average." This interpretation underlies all known results on *tightness* of semidefinite relaxation, like the relation

$$\mathrm{Opt}_* \leq \mathrm{Opt} \leq 3\ln(\sqrt{3}K)\mathrm{Opt}_*.$$

Derivation of this (in fact, unimprovable, up to absolute constants involved, without additional restrictions on $A$ and $\mathcal{X}$), bound on the quality of semidefinite relaxations, while not being too difficult, goes beyond the scope of this textbook. It should be added that in special cases better approximation bounds can be found. For example, when $A \succeq 0$ and $\mathcal{X} = \{x : \|x\|_\infty \leq 1\}$ is the unit box (or, more generally, matrices $T_1, ..., T_K$ commute with each other), the "tightness factor" $3\ln(\sqrt{3}K)$ can be improved to $\frac{\pi}{2} \approx 1.571$ (Nesterov's $\frac{\pi}{2}$ Theorem), and when $A$ is further restricted to have nonpositive off-diagonal entries and zero row sums – to 1.138 (MAXCUT Theorem of Goemans and Williamson).

**Exercise** IV.28 ♦ [5] What follows is the concluding exercise in the "Truss Topology Design" series. We have already used TTD problem to present instructive "real life" illustrations of the power of several results of Convex Analysis, specifically, Caratheodory Theorem (Exercise I.18), epigraph description of convexity and Helly Theorem (Exercise III.9) and $\mathcal{S}$-Lemma (Exercise IV.11), not speaking about the Schur Complement Lemma which was instrumental in all these exercises. Now it is time to illustrate the power of conic duality.

In the sequel, we assume that the reader is reasonably well acquainted with Truss Topology Design story as told in Exercise I.16 and use without additional comments the notions, notation, and the results presented in this Exercise, including the default assumption $\mathfrak{R}$ which remains in force below. In addition, we assume from now on that the load of interest $f$ is nonzero – this is the only nontrivial case in TTD.

Recall that the TTD problem as posed in Exercise I.16.2 reads

$$\mathrm{Opt} = \min_{\tau,r}\left\{\tau : \left[\begin{array}{c|c} B\,\mathrm{Diag}\{t\}B^\top & f \\ \hline f^\top & 2\tau \end{array}\right] \succeq 0, t \geq 0, \sum_i t_i = W\right\} \qquad (P)$$

In our present language, this is a semidefinite program, and we know from Exercise I.16 that this problem is solvable.

Your first task is easy:

1. Build the semidefinite dual of $(P)$ and prove that the dual problem is solvable with the same optimal value Opt as the primal problem $(P)$.

Since passing from a semidefinite problem to its dual is a purely mechanical process, on one hand, and the subsequent tasks will be formulated in terms of the dual problem, here is the dual as given by Conic Duality:

$$\max_{V,g,\theta,\lambda,\mu}\left\{-2f^\top g - W\mu : 2\theta = 1, \mathfrak{b}_i^\top V\mathfrak{b}_i + \lambda_i - \mu = 0\,\forall i, \lambda \geq 0, \left[\begin{array}{c|c} V & g \\ \hline g^\top & \theta \end{array}\right] \succeq 0\right\}$$

[5] Preceding exercises in the TTD series are I.16, I.18, III.9, IV.11.

Eliminating variable $\tau$ (which is fixed by the corresponding constraint), we rewrite the dual as

$$\max_{V,g,\lambda,\mu}\left\{-2f^\top g - W\mu : \mathfrak{b}_i^\top V\mathfrak{b}_i + \lambda_i - \mu = 0\,\forall i, \lambda \geq 0, \left[\begin{array}{c|c} V & g \\ \hline g^\top & \frac{1}{2}\end{array}\right] \succeq 0\right\} \qquad (D)$$

What is left to you, is to verify the derivation and to prove that $(D)$ is solvable with the same optimal value Opt as $(P)$.

Your next task still is easy:

2. Verify that eliminating, by partial optimization, variables $V$ and $\lambda$, problem $(D)$ reduces to the problem

$$\max_{g,\mu}\left\{-2f^\top g - W\mu : \left[\begin{array}{c|c} \mu & \mathfrak{b}_i^\top g \\ \hline \mathfrak{b}_i^\top g & \frac{1}{2}\end{array}\right] \succeq 0\,\forall i\right\} \qquad (\overline{D})$$

   and the latter problem is solvable with the same optimal value Opt as $(P)$ and $(D)$.

Pay attention to the first surprising fact: semidefinite constraints in $(\overline{D})$ involve the cone $\mathbf{S}^2_+$ of $2\times 2$ positive semidefinite matrices, and this cone, as we know, is, up to one-to-one linear transformation, just the Lorentz cone $\mathbf{L}^3$. Thus, $(\overline{D})$ is a conic quadratic problem.

Your next task is

3. Pass from problem $(\overline{D})$ to its semidefinite dual $(\overline{P})$ and prove that the latter problem is solvable with optimal value Opt.

At the first glance, the task seems crazy: the dual of the dual is the primal! Note, however, that $(\overline{D})$ is *not* the plain conic dual to $(P)$ problem $(D)$ – it is obtained from $(D)$ by eliminating part of variables, and nobody told us that this elimination keeps the dual to $(\overline{D})$ equivalent to the dual of $(D)$, that is, to $(P)$.

By the same reasons as in item 1, we take upon ourselves writing down $(\overline{P})$:

$$\min_{s,t,q}\left\{\frac{1}{2}\sum_i s_i : \sum_i t_i = W, \sum_i q_i\mathfrak{b}_i = f, \left[\begin{array}{c|c} t_i & q_i \\ \hline q_i & s_i\end{array}\right] \succeq 0\,\forall i\right\} \qquad (\overline{P})$$

What is left to you is to prove that $(\overline{P})$ is solvable with optimal value Opt.

Now – the main surprise:

4. Verify that $(\overline{P})$ allows eliminating, by partial minimization, variables $t_i$ and $s_i$, which reduces $(\overline{P})$ to solvable optimization problem

$$\min_q\left\{\frac{1}{2W}\left(\sum_i |q_i|\right)^2 : \sum_i q_i\mathfrak{b}_i = f\right\} \qquad (\#.1)$$

   with the same optimal value Opt as all preceding problems, $(P)$ included.

This indeed is a great surprise – $(\#.1)$ is equivalent to *Linear Programming* problem

$$G = \min_q\left\{\|q\|_1 : \sum_i q_i\mathfrak{b}_i = f\right\}, \qquad (\#.2)$$

the optimal value in this problem being

$$G = \sqrt{2W\text{Opt}}.$$

The challenge is, of course, to extract from optimal solution to $(\#.2)$ an optimal truss $t^*$ – one with total bar volume $W$ and compliance, w.r.t. load $f$, equal to Opt, and this is your final task:

5. Extract from optimal solution to $(\#.2)$ an optimal truss.

## 29.5 ★ Cone-convexity

**Exercise** IV.29 ♦ [elementary properties of cone-convex functions] The goal of this Exercise is to extend elementary properties of convex functions onto cone-convex mappings.

**A.** Let $\mathcal{X},\mathcal{Y}$ be Euclidean spaces equipped with norms $\|\cdot\|_{\mathcal{X}}$, $\|\cdot\|_{\mathcal{Y}}$. Let, next, $\mathbf{X}$ be a closed pointed cone in $\mathcal{X}$, $\mathbf{Y}$ be a closed *and pointed* cone in $\mathcal{Y}$, and $f: X\to\mathcal{Y}$ be a mapping defined on a nonempty convex set $X\subset\mathcal{X}$. Recall that for a closed and pointed cone $\mathbf{K}$ in Euclidean space $\mathcal{K}$ and $x,x'\in\mathcal{K}$, relation $x\leq_{\mathbf{K}}x'$, same as $x'\geq_{\mathbf{K}}x$, means that $x'-x\in\mathbf{K}$.

Recall that $f$ is called

- $(\mathbf{X},\mathbf{Y})$-monotone on $X$, if

$$\left\{x,x'\in X \text{ and } x\leq_{\mathbf{X}}x'\right\} \implies f(x)\leq_{\mathbf{Y}}f(x');$$

- $\mathbf{Y}$-convex on $X$, if

$$f(\lambda x+(1-\lambda)x')\leq_{\mathbf{Y}}\lambda f(x)+(1-\lambda)f(x')$$

for every $x,x'\in X$ and $\lambda\in[0,1]$.

For example,

— an affine mapping $f(x)=Ax+a:\mathcal{X}\to\mathcal{Y}$ is $\mathbf{Y}$-convex, whatever be pointed closed cone $\mathbf{Y}$;

— when $\mathcal{Y}=\mathbf{R}$ and $\mathbf{Y}=\mathbf{R}_+$, $\mathbf{Y}$-convex on $X$ functions are the convex, in the standard definition, real-valued functions on $X$.

A.1. In the situation of **A**, let $\mathbf{Y}^*$ be the cone dual to $\mathbf{Y}$. For $e\in\mathcal{Y}$, let $f_e(x)=\langle e,f(x)\rangle_{\mathcal{Y}}:X\to\mathbf{R}$. Prove that $f$ is
— $\mathbf{Y}$-convex on $X$ if and only if the function $f_e$ is convex on $X$ whenever $e\in\mathbf{Y}^*$
— $(\mathbf{X},\mathbf{Y})$-monotone on $X$ if and only if the function $f_e$ is $\mathbf{X}$-monotone on $X$ (i.e., $x,x'\in X, x\leq_{\mathbf{X}}x' \implies f_e(x)\leq f_e(x')$) for every $e\in\mathbf{Y}^*$.

A.2. In the situation of **A**, let $f$ be $\mathbf{Y}$-convex. Prove that $f$ is locally bounded and locally Lipschitz continuous on the interior of $X$, meaning that if $\bar{X}\subset\operatorname{int}X$ is a closed and bounded set, then there exists $M<\infty$ such that $\|f(x)\|_{\mathcal{Y}}\leq M$ holds for all $x\in\bar{X}$ (this is local boundedness) and there exists $L<\infty$ such that $\|f(x)-f(z')\|_{\mathcal{Y}}\leq L\|x-x'\|_{\mathcal{X}}$ holds for all $x,x'\in\bar{X}$ (this is local Lipschitz continuity).

**B.** Now let us look at elementary operations preserving cone convexity. From now on, $\mathrm{Lin}(\mathcal{X},\mathcal{Y})$ denotes the linear space of linear mappings acting from Euclidean space $\mathcal{X}$ to Euclidean space $\mathcal{Y}$. Prove the following statements:

B.1. ["nonnegative linear combinations"] Let $X$ be a nonempty convex subset of Euclidean space $\mathcal{X}$, $\mathcal{Y}_j$, $j\leq J$, and $\mathcal{Y}$ be Euclidean spaces equipped with pointed closed cones $\mathbf{Y}_j$, $\mathbf{Y}$, and $\alpha_j\in\mathrm{Lin}(\mathcal{Y}_j,\mathcal{Y})$ be "nonnegative coefficients", meaning that $\alpha_jy_j\in\mathbf{Y}$ whenever $y_j\in\mathbf{Y}_j$. When mappings $f_j(x):X\to\mathcal{Y}_j$. are $\mathbf{Y}_j$-convex, $j\leq J$, their "linear combination with coefficients $\alpha_j$" – the mapping

$$f(x)=\sum_j\alpha_jf_j(x):X\to\mathcal{Y}$$

– is $\mathbf{Y}$-convex.

B.2. [affine substitution of variables] In the situation of **A**, let $z\mapsto Az+a:\mathcal{Z}\to\mathcal{X}$ be an affine mapping, and let $f$ be $\mathbf{Y}$-convex on $X$. Then, the function $g(z):=f(Az+a)$ is $\mathbf{Y}$-convex on the set $Z=\{z:Az+a\in X\}$.

B.3. [monotone composition] Let $\mathcal{U}_j$, $j\leq J$, be Euclidean spaces equipped with closed pointed cones $\mathbf{U}_j$, let $\mathcal{U}=\mathcal{U}_1\times\ldots\times\mathcal{U}_J$, $\mathbf{U}=\mathbf{U}_1\times\ldots\times\mathbf{U}_J$, and let $\mathcal{Y}$ be an Euclidean space equipped with closed pointed cone $\mathbf{Y}$. Next, let $X$ be nonempty convex set in Euclidean space $\mathcal{X}$, $U$ be a nonempty convex set in $\mathcal{U}$, let $f_j(x):X\to\mathcal{U}_j$ be $\mathbf{U}_j$-convex functions, $j\leq J$, such that $f(x) = [f_1(x); \ldots; f_J(x)] \in U$ whenever $x \in X$. Finally, let mapping $F : U \to \mathcal{Y}$ be $(\mathbf{U}, \mathbf{Y})$-monotone and $\mathbf{Y}$-convex on $U$. Then the composition

$$G(x) = F(f(x)) : X \to \mathcal{Y}$$

is $\mathbf{Y}$-convex on $X$.

**Note:** we do not forbid some of $\mathbf{U}_j$ to be the trivial cones $\{0\}$. When $\mathbf{U}_j$ is trivial, $(\mathbf{U}_j, \mathbf{Y})$-monotonicity of $F(u_1, \ldots, u_J)$ with respect to $u_j$ automatically holds true, while $\mathbf{U}_j$-convexity of $f_j$ holds true when $f_j$ is affine, cf. Convex Monotone Superposition rule in section 14.1.

**C.** The gradient inequality and existence of directional derivative can be extended from the usual convex functions (i.e., $\mathbf{R}_+$-convex functions taking values in $\mathbf{R}$) to the cone-convex ones. Prove the following statements:

C.1. ["gradient inequality"] In the situation of **A**, let $\bar{x} \in X$ and $f$ be $\mathbf{Y}$-convex on $X$ and differentiable at $\bar{x}$. Then

$$\forall y \in X : f(y) \geq_{\mathbf{Y}} f(\bar{x}) + f'(\bar{x})(y - \bar{x}),$$

where $f'(\bar{x})$ is the Jacobian of $f$ at $\bar{x}$.

C.2. [existence of directional derivative] In the situation of **A**, let $f$ be $\mathbf{Y}$-convex on $X$, let $\bar{x} \in \operatorname{int} X$ and $d \in \mathcal{X}$. Then

$$\exists Df(\bar{x})[d] := \lim_{t \to +0} \frac{f(\bar{x} + td) - f(\bar{x})}{t}$$

and

$$(t \geq 0 \ \& \ \bar{x} + td \in X) \implies f(\bar{x} + td) \geq_{\mathbf{Y}} f(\bar{x}) + tDf(\bar{x})[d]. \qquad (\#)$$

Besides this, as a function of $d \in \mathcal{X}$, $Df(\bar{x})[d]$ is positively homogeneous of degree 1 (i.e., $Df(\bar{x})[td] = tDf(\bar{x})[d]$ when $t \geq 0$) and $\mathbf{Y}$-convex.

**D.** Subdifferentials of the usual convex functions admit natural extensions to the cone-convex mappings. Specifically, in the situation of **A**, let $\bar{x} \in X$. Let us say that $g \in \operatorname{Lin}(\mathcal{X}, \mathcal{Y})$ is a *sub-Jacobian* of $f$ at $\bar{x}$, if

$$\forall y \in X : f(y) \geq_{\mathbf{Y}} f(\bar{x}) + g[y - x].$$

For example, C.1 says that if $f$ is $\mathbf{Y}$-convex on $X$ and differentiable at $\bar{x} \in X$, then the taken at $x$ Jacobian $f'(\bar{x})$ of $f$ is a sub-Jacobian of $f$ at $\bar{x}$. Clearly, for a usual convex function its sub-Jacobians at a point are exactly the linear forms on $\mathcal{X}$ given by subgradients $f'(x)$ of $f$ at $x$ according to

$$gh = \langle f'(x), h \rangle_{\mathcal{X}}, \; h \in \mathcal{X}.$$

Let $\mathcal{J}f(x)$ be the set of all sub-Jacobians of $f$ at $x \in X$. Prove the following statements:

D.1. In the situation of **A**, for $x \in X$ one has $g \in \mathcal{J}f(x)$ if and only if for every $e \in \mathbf{Y}^*$ the vector $g^*e \in \mathcal{X}$ is a subgradient of $f_e$ at $x$; here for $g \in \operatorname{Lin}(\mathcal{X}, \mathcal{Y})$, $g^* \in \operatorname{Lin}(\mathcal{Y}, \mathcal{X})$ is the conjugate of $g$: $\langle gu, v \rangle_{\mathcal{Y}} = \langle u, g^*v \rangle_{\mathcal{X}}$ for all $u \in \mathcal{X}$, $v \in \mathcal{Y}$.

D.2. In the situation of **A**, let $f$ be $\mathbf{Y}$-convex on $X$. Then

— D.2.1. For every $x \in X$, the set $\mathcal{J}f(x)$ is a closed convex subset of $\operatorname{Lin}(\mathcal{X}, \mathcal{Y})$;

— D.2.2. The mapping $x \mapsto \mathcal{J}f(x)$ is locally bounded on the interior of $X$, that is, for every closed and bounded set $\bar{X} \subset \operatorname{int} X$, the induced norms $\|g\|_{\mathcal{X},\mathcal{Y}} = \max_z\{\|gz\|_{\mathcal{Y}} : \|z\|_{\mathcal{X}} \leq 1\}$ of linear mappings $g \in \mathcal{J}f(x)$, $x \in \bar{X}$ are bounded away from $+\infty$;

— D.2.3. The multivalued mapping $x \mapsto \mathcal{J}f(x)$ is closed on $\operatorname{int} X$: if $x_i \in \operatorname{int} X$ converge as $i \to \infty$ to $\bar{x} \in \operatorname{int} X$ and linear mappings $g_i \in \mathcal{J}f(x_i)$ converge as $i \to \infty$ to some $\bar{g} \in \operatorname{Lin}(\mathcal{X}, \mathcal{Y})$, then $\bar{g} \in \mathcal{J}f(\bar{x})$.

The most attractive property of subgradients of the usual convex function is their existence, at least at interior points of the function's domain. This fact extends to the cone-convex mappings. Prove the following statements:

D.3. [existence of sub-Jacobians] In the situation of **A**, let $\bar{x} \in \operatorname{int} X$ and $f$ be **Y**-convex on $X$. Then $\mathcal{J}f(\bar{x})$ is nonempty.

For a real-valued convex function $f$ and $x \in \operatorname{int}\operatorname{Dom} f$, $d \in \mathcal{X}$, one has

$$Df(x)[d] = \max_{y \in \partial f(x)} \langle y, d\rangle_{\mathcal{X}}.$$

A similar fact holds true for cone-convex functions:

D.4. In the situation of **A**, let $f$ be **Y**-convex on $X$. Let also $\bar{x} \in \operatorname{int} X$ and $d \in \mathcal{X}$. Then for properly selected $g \in \mathcal{J}f(\bar{x})$ one has

$$Df(\bar{x})[d] = gd,$$

while for every $g' \in \mathcal{J}f(\bar{x})$ one has

$$Df(\bar{x})[d] \geq_{\mathbf{Y}} g'd.$$

There is a natural relation between sub-Jacobians of **Y**-convex function $f$ and subgradients of functions $f_e = \langle e, f\rangle_{\mathcal{Y}}$, $e \in \mathbf{Y}^*$:

D.5. In the situation of **A**, let $f$ be **Y**-convex on $X$ and $\bar{x} \in \operatorname{int} X$. For $e \in \mathbf{Y}^*$, $h \in \partial f_e(\bar{x})$ (that is, $f_e(y) \geq f_e(\bar{x}) + \langle h, y - \bar{x}\rangle_{\mathcal{X}}$ for all $y \in X$) if and only if $h = g^*e$ for some $g \in \mathcal{J}f(\bar{x})$.

Finally, the chain rule:

D.6. [chain rule] Let $\mathcal{U}_j$, $j \leq J$, be Euclidean spaces equipped with closed pointed cones $\mathbf{U}_j$, let $\mathcal{U} = \mathcal{U}_1 \times \ldots \times \mathcal{U}_J$, $\mathbf{U} = \mathbf{U}_1 \times \ldots \times \mathbf{U}_J$, and let $\mathcal{Y}$ be an Euclidean space equipped with closed pointed cone **Y**. Next, let $X$ be nonempty convex set in Euclidean space $\mathcal{X}$, $U$ be a nonempty convex set in $\mathcal{U}$, let $f_j(x) : X \to \mathcal{U}_j$ be $\mathbf{U}_j$-convex on $X$ functions, $j \leq J$, such that $f(x) = [f_1(x); \ldots; f_J(x)] \in U$ whenever $x \in X$. Finally, let mapping $F : U \to \mathcal{Y}$ be $(\mathbf{U}, \mathbf{Y})$-monotone and **Y**-convex on $U$. As we know from B.3, the composition

$$G(x) = F(f(x)) : X \to \mathcal{Y}$$

is $\mathcal{Y}$-convex on $X$. Now let $\bar{x} \in \operatorname{int} X$, $\bar{u}_j = f_j(\bar{x})$ be such that $\bar{u} = [\bar{u}_1; \ldots; \bar{u}_J] \in \operatorname{int} U$. Finally, let $g_j \in \mathcal{J}f_j(\bar{x})$, $j \leq J$, and $g \in \mathcal{J}F(\bar{u})$. Then the linear mapping $[u_1; \ldots; u_J] \mapsto g[u_1; \ldots; u_J]$ is $(\mathbf{U}, \mathbf{Y})$-monotone, and the linear mapping

$$h \mapsto \widehat{g}h := g[g_1h; \ldots; g_Jh] : \mathcal{X} \to \mathcal{Y}$$

is sub-Jacobian of $G$ at $\bar{x}$.

**Exercise** IV.30 Univariate function $f(x) = x^{-1/2} : \{x > 0\} \to \mathbf{R}$ is nonincreasing and convex, and $\nabla f(x) = -x^{-3/2}/2$, $x > 0$. Now let $P$ be $m \times n$ matrix with trivial kernel.

1. Prove that the mapping $F(X) = [PXP^\top]^{-1/2} : \mathbf{S}^n_{++} \to \mathbf{S}^m$, where $\mathbf{S}^n_{++} = \operatorname{int}\mathbf{S}^n_+ = \{X \in \mathbf{S}^n : X \succ 0\}$, is $(\mathbf{S}^n_+, \mathbf{S}^m_+)$-antimonotone and $\mathbf{S}^m_+$-convex
2. Assuming $P = I_2$, compute numerically $F(X)$ and $dF(X)[dX]$ for $X = \left[\begin{array}{cc} 2 & -1 \\ -1 & 1 \end{array}\right]$ and $dX = \left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array}\right]$. For the above $X$, compute also the Jacobian $J$ of $F$ at $X$ – the matrix of the linear mapping $dX \mapsto DF(X)[dX] : \mathbf{S}^2 \to \mathbf{S}^2$ – in the basis $[1, 0; 0, 0]$, $[0, 0; 0, 1]$, $[0, 1/\sqrt{2}; 1/\sqrt{2}, 0]$ of $\mathbf{S}^2$.
3. How the "Gradient inequality" (Exercise IV.29.C.1) for the $\mathbf{S}^n_+$-convex mapping $F$ looks like?

## 29.6 ★ Around conic representations of sets and functions

### *29.6.1 Conic representations: definitions*

Let $\mathfrak{K}$ be a family of regular cones in Euclidean spaces which contains the nonnegative ray $\mathbf{R}_+$ and is closed with respect to taking finite direct products and passing from a cone to its dual. Instructive examples are the families $\mathfrak{N}$ of nonnegative orthants, $\mathfrak{L}$ of finite direct products of Lorentz cones, and $\mathfrak{S}$ of finite direct products of semidefinite cones.

- A $\mathfrak{K}$-*representation* ($\mathfrak{K}$-r.) of a set $X \subset \mathbf{R}^n$ is its representation of the form

$$X = \{x \in \mathbf{R}^n : \exists u : Px + Qu - r \in \mathbf{K}\} \tag{29.2}$$

with $\mathbf{K} \in \mathfrak{K}$ – representation of $X$ as the projection of the solution set of conic inequality $Px + Qu \geq_{\mathbf{K}} r$ in variables $x, u$ onto the plane of $x$-variables where $X$ lives. A set $X$ admitting conic representation with cone from $\mathfrak{K}$ is called $\mathfrak{K}$-*representable* ($\mathfrak{K}$-r for short).

- A $\mathfrak{K}$-*representation of a function* $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is, by definition, $\mathfrak{K}$-representation of the epigraph of $f$:

$$[t; x] \in \text{epi}\{f\} := \{[x; t] : t \geq f(x)\} \iff \exists u : Px + tp + Qu - r \in \mathbf{K} \text{ with } \mathbf{K} \in \mathfrak{K}.$$

Functions admitting $\mathfrak{K}$-representation are called $\mathfrak{K}$-*representable* ($\mathfrak{K}$-r for short)

We are already acquainted with $\mathfrak{N}$-representability – it is that was called polyhedral representability. By Fourier-Motzkin elimination, polyhedral representable sets $X \subset \mathbf{R}^n$ admit polyhedral representations not involving additional variables $u$, and similarly for $\mathfrak{N}$-representable functions; this is not the case for more general families $\mathfrak{K}$, like families $\mathfrak{L}$ of Lorentz- and $\mathfrak{S}$ of semidefinite-representable sets.

The following exercise explains what is the rationale underlying the above restrictions on $\mathfrak{K}$ and why we are interested in $\mathfrak{K}$-representations.

**Exercise** IV.31 ♦ Check that

1. Every finite system $P_0 y \geq r_0$, $P_i y - r_i \in \mathbf{K}_i$, $i \leq I$, of scalar linear inequalities and conic inequalities, involving cones from $\mathfrak{K}$, in variables $y$ is equivalent to a single conic inequality, with cone from $\mathfrak{K}$, in these variables:

$$\begin{array}{l} \{P_0 y - r_0 \geq 0, P_i y - r_i \in \mathbf{K}_i, 1 \leq i \leq I\} \\ \iff \left\{[P_0; P_1; \ldots; P_I]y - [r_0; r_1; \ldots; r_I] \in \mathbf{K} := \underbrace{\mathbf{R}_+ \times \ldots \times \mathbf{R}_+}_{\dim r_0 \text{ times}} \times \mathbf{K}_1 \times \mathbf{K}_2 \times \ldots \times \mathbf{K}_I\right\} \end{array}$$

and $\mathbf{K} \in \mathfrak{K}$ (since $\mathbf{R}_+ \in \mathfrak{K}$ and $\mathfrak{K}$ is closed with respect to taking finite direct products).
As a result, representation of a set $X$ as

$$X = \{x : \exists u : P_0 x + Q_0 u - r^0 \geq 0, P_i x + Q_i u - r_i \in \mathbf{K}_i, 1 \leq i \leq I\} \qquad [\mathbf{K}_i \in \mathfrak{K}] \quad (!)$$

– as the projection of the solution set of a finite system of linear and $\mathfrak{K}$-conic inequalities in variables $x, u$ onto the plane of $x$-variables where $X$ lives, can be straightforwardly converted into a $\mathfrak{K}$-r. of $X$.

**Important:** Item 1 allows us from now on to refer to representations of the form (!) as to $\mathfrak{K}$-representations of $X$, skipping (always straightforward and purely mechanical) conversion of such a representation into the "canonical" representation (29.2).

2. $\mathfrak{K}$-r. of a function straightforwardly induces $\mathfrak{K}$-r.'s of its sublevel sets:

$$\begin{array}{l} \left\{\{t \geq f(x)\} \iff \{\exists u : Px + tp + Qu - r \in \mathbf{K}\}\right\} \\ \Longrightarrow X_a := \{x : f(x) \leq a\} = \{x : \exists u : Px + Qu - [r - ap] \in \mathbf{K}\} \end{array} \qquad [a \in \mathbf{R}, \mathbf{K} \in \mathfrak{K}]$$

3. Given $\mathfrak{K}$-representations of a set $X \subset \mathbf{R}^n$ and a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$:

$$\begin{array}{l} X = \{x \in \mathbf{R}^n : \exists u : P_X x + Q_X u - r_X \in \mathbf{K}_X\}, \\ \mathrm{epi}\{f\} = \{[x;t] : \exists v : P_f x + t p_f + Q_f v - r_f \in \mathbf{K}_f\} \end{array} \qquad [\mathbf{K}_X \in \mathfrak{K}, \mathbf{K}_f \in \mathfrak{K}]$$

   we can straightforwardly convert the optimization problem

$$\min_{x \in X} f(x) \tag{$*$}$$

   into conic problem on a cone from $\mathfrak{K}$, namely, the problem

$$\min_{x,t,u,v} \left\{ t : A[x;t;u;v] - b := [P_X x + Q_X u; P_f x + t p_f + Q_f v] - [r_X; r_f] \in \underbrace{\mathbf{K} := \mathbf{K}_X \times \mathbf{K}_f}_{\in \mathfrak{K}} \right\}$$

   As a result, a solver $\mathcal{S}$ capable to solve conic problems on cones from $\mathfrak{K}$ can be straightforwardly utilized when solving problems ($*$) with $X$ and $f$ given by $\mathfrak{K}$-r.'s.

4. Given a conic problem

$$\min_x \left\{ c^\top x : Ax - b \in \mathbf{K}, Rx \geq r \right\} \tag{$P$}$$

   on a cone from $\mathfrak{K}$, its conic dual – the conic problem

$$\begin{array}{c} \max_{y,z} \left\{ \langle b, y \rangle + r^\top z : A^* y + R^\top z = c, y \in \mathbf{K}_*, z \geq 0 \right\} \\ \left[ \begin{array}{c} \langle \cdot, \cdot \rangle \text{ is the inner product in the Euclidean space where } \mathbf{K} \text{ lives, } \mathbf{K}_* \text{ is the cone dual to } \mathbf{K}, \\ A^* \text{ is the conjugate of } A\text{: } \langle Ax, y \rangle \equiv x^\top A^* y \ \forall x, y \end{array} \right] \end{array} \tag{$D$}$$

   also is a conic problem on a cone from $\mathfrak{K}$ (since $\mathfrak{K}$ is closed with respect to passing from a cone to its dual and contains nonnegative orthants).

Note that the option mentioned in the last item of Exercise IV.31 is implemented in "CVX: MATLAB software for disciplined convex programming" due to M. Grant and S. Boyd `http://cvxr.com/cvx` – second to none in its scope and user-friendliness tool for numerical processing of well-structured convex problems, the underlying family $\mathfrak{K}$ being the semidefinite family $\mathfrak{S}$. We conclude that it makes sense to develop a kind of calculus allowing to recognize $\mathfrak{K}$-representability of sets/functions and to build, when possible, their $\mathfrak{K}$-representations. The desired calculus exists and is pretty simple, general and fully algorithmic. The goal of subsequent exercises is to make you acquainted with the most frequently used elements of this calculus; for more on this subject, see [BTN].

### 29.6.2 Conic representability: elementary calculus

Elementary calculus of conic representability is completely similar to calculus of polyhedral representations from section 3.3.

**Exercise** IV.32 [elementary calculus of $\mathfrak{K}$-representable sets] Check that basic convexity-preserving[6] operations with sets preserve $\mathfrak{K}$-representability. Specifically,

1. Finite intersection of $\mathfrak{K}$-r sets $X_i = \{x \in \mathbf{R}^n : \exists u^i : P_i x + Q_i u^i - r_i \in \mathbf{K}_i\}$, $i \leq I$ (here and in what follows all cones involved are from $\mathfrak{K}$) is $\mathfrak{K}$-r:

$$\begin{array}{ll} \bigcap\limits_{i \leq I} X_i = & \{x \in \mathbf{R}^n : \exists u = [u^1; \ldots; u^I] : \\ & Px + Qu - r := [P_1 x + Q_1 u^1; \ldots; P_I x + Q_I u^I] - [r_1; \ldots; r_I] \in \underbrace{\mathbf{K} := \mathbf{K}_1 \times \ldots \times \mathbf{K}_I}_{\in \mathfrak{K}} \} \end{array}$$

[6] "convexity-preserving" is crucial – clearly, $\mathfrak{K}$-r sets and functions must be convex!

2. Direct product of finitely many $\mathfrak{K}$-r sets $X_i = \{x \in \mathbf{R}^n : \exists u^i : P_i x + Q_i u^i - r_i \in \mathbf{K}_i\}$, $i \leq I$ is $\mathfrak{K}$-r:

$$\begin{array}{l} X_1 \times \ldots \times X_I = \{x = [x^1; \ldots ; x^I] : \exists u = [u^1; \ldots ; u^I] : \\ Px + Qu - r := [P_1 x^1 + Q_1 u^1; \ldots ; P_I x^I + Q_I u^I] - [r_1; \ldots ; r_I] \in \underbrace{\mathbf{K} := \mathbf{K}_1 \times \ldots \times \mathbf{K}_I}_{\in \mathfrak{K}}\} \end{array}$$

3. Affine image $Y = \{y = Ax + b : x \in X\}$ of $\mathfrak{K}$-r set $X = \{x \in \mathbf{R}^n : \exists u : Px + Qu - r \in \mathbf{K}\}$ is $\mathfrak{K}$-r:

$$Y = \{y : \exists [x; u] : Ax + b = y, Px + Qu - r \in \mathbf{K}\}$$

   is the projection onto the $y$-plane of a set given by explicit finite system of linear and $\mathfrak{K}$-conic inequalities and as such admits an explicit $\mathfrak{K}$-r. by item 1 of Exercise IV.31.

4. Inverse affine image $Y = \{y : Ay + b \in X\}$ of $\mathfrak{K}$-r set $X = \{x \in \mathbf{R}^n : \exists u : Px + Qu - r \in \mathbf{K}\}$ is $\mathfrak{K}$-r:

$$Y = \{y : \exists u : PAy + Qy - [r - Pb] \in \mathbf{K}\}.$$

5. The arithmetic sum $X = X_1 + \ldots + X_I$ of $\mathfrak{K}$-r sets $X_i = \{x \in \mathbf{R}^n : \exists u^i : P_i x + Q_i u^i - r_i \in \mathbf{K}_i\}$, $i \leq I$, is $\mathfrak{K}$-r:

$$X = \{x : \exists [x^1; \ldots ; x^I; u^1; \ldots ; u^I] : x - \sum_i x^i = 0, P_i x^i + Q_i u^i - r_i \in \mathbf{K}_i, i \leq I\}$$

   and it remains to apply item 1 of Exercise IV.31.

**Exercise** IV.33 ♦ [elementary calculus of $\mathfrak{K}$-representable functions] Check that the following convexity-preserving operations with functions preserve $\mathfrak{K}$-representability:

0. Restricting onto $\mathfrak{K}$-r set: $\mathfrak{K}$-r. $t \geq f(x) \iff \exists u : P_f x + t_f p + Q_f u - r_f \in \mathbf{K}_f$ of a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ taken together with $\mathfrak{K}$-r. $X = \{x \in \mathbf{R}^n : \exists v : P_X x + Q_X v - r_X \in \mathbf{K}_X\}$ of a set $X \subset \mathbf{R}^n$ induce $\mathfrak{K}$-r.

$$t \geq f\big|_X(x) \iff \exists u, v : P_f x + t p_f + Q_f u - r_f \in \mathbf{K}_f,\ P_X x + Q_X v - r_X \in \mathbf{K}_X$$

   of the restriction $f\big|_X(x) = \left\{ \begin{array}{ll} f(x) & , x \in X \\ +\infty & , x \notin X \end{array} \right.$ of $f$ onto $X$

1. Taking linear combination $\sum_{i=1}^I \lambda_i f_i$ with positive coefficients:

$$\begin{array}{c} t \geq f_i(x) \iff \exists u^i : P_i x + t p_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \\ \Downarrow \\ t \geq f(x) := \sum_{i=1}^I \lambda_i f_i(x) \iff \exists [t_1; \ldots ; t_I; u^1; \ldots ; u^i] : t \geq \sum_i \lambda_i t_i, P_i x + t_i p_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \end{array}$$

2. Direct summation:

$$\begin{array}{c} t \geq f_i(x^i) \iff \exists u^i : P_i x^i + t p_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \\ \Downarrow \\ t \geq f(x^1, \ldots , x^I) := \sum_{i=1}^I f_i(x^i) \iff \exists [t_1; \ldots ; t_I; u^1; \ldots ; u^i] : \\ t \geq \sum_i t_i, P_i x^i + t_i p_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \end{array}$$

3. Taking finite maxima:

$$\begin{array}{c} t \geq f_i(x) \iff \exists u^i : P_i x + t p_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \\ \Downarrow \\ t \geq f(x) := \max_{i \leq I} f_i(x) \iff \exists [u^1; \ldots ; u^i] : P_i x + t p_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \end{array}$$

4. Affine substitution of variables:

$$\begin{array}{c} t \geq f(x) \iff \exists u : Px + tp + Qu - r \in \mathbf{K} \\ \Downarrow \\ t \geq g(y) := f(Ay + b) \iff \exists u : PAu + tp + Qu - [r - Pb] \in \mathbf{K} \end{array}$$

   In fact, claims in items 1–4 are special cases of the following observation:

5. Monotone superposition: let functions $f_i(x)$, $i \leq I$, be $\mathfrak{K}$-r with the first $K$ of the functions being affine, and let $F(y) : \mathbf{R}^I \to \mathbf{R} \cup \{+\infty\}$ be $\mathfrak{K}$-r and monotonically nondecreasing in $y_{K+1}, \dots, y_I$.
$$y, y' \in \mathbf{R}^I, y \geq y', y_i = y'_i, i \leq K \implies F(y) \geq F(y').$$
Then the functions
$$g(x) = \left\{ \begin{array}{ll} F(f_1(x), \dots, f_I(x)) & , f_i(x) < \infty \, \forall i \\ +\infty & , \text{otherwise.} \end{array} \right.$$
is $\mathfrak{K}$-r, specifically,
$$\left\{ \begin{array}{c} f_i \text{ are affine}, i \leq K, \ \& \ t \geq f_i(x) \iff \exists u^i : P_i x + t p_i + Q_i u^i - r_i \in \mathbf{K}_i, \ K < i \leq I \\ t \geq F(y) \iff \exists u : Py + tp + Qu - r \in \mathbf{K} \end{array} \right\}$$
$$\Downarrow$$
$$t \geq g(x) \iff \exists t_i, 1 \leq i \leq I, u^i, K < i \leq I, u : \left\{ \begin{array}{ll} \underbrace{t_i - f_i(x) = 0}_{\text{linear equations}} & , i \leq K \\ P_i x + t_i p_i + Q_i u^i - r_i \in \mathbf{K}_i & , K < i \leq I \\ P[t_1; \dots; t_k] + tp + Qu - r \in \mathbf{K} & \end{array} \right.$$

### *29.6.3 $\mathfrak{R}/\mathfrak{L}/\mathfrak{S}$ hierarchy*

**Exercise** IV.34 ♦

1. Let $\mathfrak{K}$ and $\mathfrak{M}$ be two families of regular cones, each containing nonnegative rays and closed with respect to taking finite direct products and passing from a cone to its dual cone. Assume that every cone $\mathbf{M} \in \mathfrak{M}$ admits $\mathfrak{K}$-representation:
$$\mathbf{M} = \{y : \exists v : P_{\mathbf{M}} y + Q_{\mathbf{M}} v - r_{\mathbf{M}} \in \underbrace{\mathbf{K}_{\mathbf{M}}}_{\in \mathfrak{K}}\}.$$
Show that a $\mathfrak{M}$-r. $X = \{x \in \mathbf{R}^n : \exists u : Px + Qu - r \in \underbrace{\mathbf{M}}_{\in \mathfrak{M}}\}$ of a set $X$ can be straightforwardly converted into $\mathfrak{K}$-r. of $X$.
2. [Cf. Exercise IV.35] Note that $\mathbf{R}^n_+$ belongs to $\mathfrak{L}$ (same as to every other family of cones we are considering here – all these families contain nonnegative rays and are closed with respect to taking finite direct products), thus, every polyhedral representable set/function is Lorentz-representable as well by item 1. Check that the Lorentz cone $\mathbf{L}^m$ is semidefinite-representable as well, specifically,
$$\begin{array}{rcl} \mathbf{L}^m & := & \{x \in \mathbf{R}^m : x_m \geq \sqrt{\sum_{i=1}^{m-1} x_i^2}\} \\ & = & \left\{ x \in \mathbf{R}^m : \text{Arrow}(x) := \left[ \begin{array}{c|c|c|c} x_m & x_1 & \dots & x_{m-1} \\ \hline x_1 & x_m & & \\ \hline \vdots & & \ddots & \\ \hline x_{m-1} & & & x_m \end{array} \right] \succeq 0 \right\} \end{array}$$
implying by item 1 that cones from $\mathfrak{L}$ admit explicit $\mathfrak{S}$-representations and thus that Lorentz-representable sets and functions are semidefinite representable as well, with $\mathfrak{S}$-r.'s readily given by $\mathfrak{L}$-r.'s.

**Exercise** IV.35 ♦ It is easy "to see" the nonnegative orthant $\mathbf{R}^n_+$ in the semidefinite cone $\mathbf{S}^n_+$ – $\mathbf{R}^n_+$ is nothing but the intersection of $\mathbf{S}^n_+$ with the linear subspace $L$ of diagonal matrices from $\mathbf{S}^n$. Formally: Let $A$ be the embedding of $\mathbf{R}^n$ into $\mathbf{S}^n$ which maps any vector $\xi$ into the diagonal matrix $\text{Diag}\{\xi\}$; then $z \in \mathbf{R}^n_+$ if and only if $Az \in \mathbf{S}^n_+$. Alternatively, you can get $\mathbf{R}^n_+$ as the linear image of the positive semidefinite cone. In particular, $\mathbf{R}^n_+$ is the image of $\mathbf{S}^n_+$ under the linear mapping which maps a symmetric $n \times n$ matrix $Z$ into the vector $\text{Dg}(Z)$ composed of diagonal entries of $Z$. As a result, a Linear Programming problem $\min_{x \in \mathbf{R}^n} \{c^\top z : \ Ax \leq b\}$ can be converted into the equivalent semidefinite problem $\min_{X \in \mathbf{S}^n} \{\sum_i c_i X_{ii} : \ X \succeq 0, \ A\text{Dg}(X) \leq b\}$.

Indeed, similar possibilities exist for the Lorentz cone $\mathbf{L}^n$, including the possibility to reformulate a conic problem involving direct products of Lorentz cones as a semidefinite program. Specifically,

1. Prove that $x \in \mathbf{L}^n$ if and only if the following "arrow" matrix

$$\mathrm{Arrow}(x) := \left[\begin{array}{c|c|c|c|c} x_n & x_1 & x_2 & \dots & x_{n-1} \\ \hline x_2 & x_n & & & \\ \hline \vdots & & \ddots & & \\ \hline x_{n-1} & & & & x_n \end{array}\right]$$

is positive semidefinite.

2. Represent $\mathbf{L}^n$ as the image of $\mathbf{S}^n_+$ under a linear mapping.

### 29.6.4 More calculus

The calculus rules to follow are less trivial:

**Exercise** IV.36 ♦ [passing from a set to its support function and polar] Let $X \subset \mathbf{R}^n$ be a nonempty closed convex set given by essentially strictly feasible $\mathfrak{K}$-representation:

$$\begin{array}{rcl} X & = & \{x \in \mathbf{R}^n : \exists u : Ax + Bu - c \geq 0, Px + Qu - r \in \mathbf{K}\} \\ & & \& \ \exists \bar{x}, \bar{u} : A\bar{x} + B\bar{u} - c \geq 0, P\bar{x} + Q\bar{u} - r \in \mathrm{int}\,\mathbf{K}. \end{array} \qquad (*)$$

This representation induces $\mathfrak{K}$-r. of the support function $\phi_X(y) = \sup_{x \in X} y^\top x$, specifically,

$$t \geq \phi_X(y) \iff \exists (\lambda, \xi) : \begin{array}{l} A^\top \lambda + P^* \xi + y = 0, B^\top \lambda + Q^* \xi = 0 \\ c^\top \lambda + \langle r, \xi \rangle + t \geq 0, \lambda \geq 0, \xi \in \mathbf{K}_* \end{array}.$$

where $\langle \cdot, \cdot \rangle$ is the inner product in the Euclidean space where $\mathbf{K}$ lives and, as always, $\mathbf{K}_*$ is the cone dual to $\mathbf{K}$. In addition. $(*)$ induces $\mathfrak{K}$-r. of the polar $\mathrm{Polar}\,(X)$ of $X$:

$$\begin{array}{rcl} \mathrm{Polar}\,(X) & := & \{y : y^\top x \leq 1 \, \forall x \in X\} \\ & = & \left\{ y : \exists (\lambda, \xi) : \begin{array}{l} A^\top \lambda + P^* \xi + y = 0, B^\top \lambda + Q^* \xi = 0 \\ c^\top \lambda + \langle r, \xi \rangle + 1 \geq 0, \lambda \geq 0, \xi \in \mathbf{K}_* \end{array} \right\} \end{array}$$

**Exercise** IV.37 ♦ Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a proper convex lower semiconscious function given by essentially strictly feasible $\mathfrak{K}$-representation:

$$\begin{array}{l} t \geq f(x) \iff \exists u : Ax + tq + Bu \geq c, Px + tp + Qu - r \in \mathbf{K} \\ \qquad \& \ \exists \overline{x}, \overline{t}, \overline{u} : A\overline{x} + \overline{t}q + B\overline{u} \geq c, P\overline{x} + \overline{t}p + Q\overline{u} - r \in \mathrm{int}\,\mathbf{K} \end{array}$$

Build $\mathfrak{K}$-r. of the Legendre transform

$$f^*(y) = \sup_x \left[ y^\top x - f(x) \right]$$

of $f$.

### 29.6.5 Raw materials

Rules of grammar become useful only after we have at our disposal words in "dictionary form" which we can combine using these rules. Similarly, calculus of conic representations becomes useful only after a rich enough dictionary of "raw materials," "atoms" – specific $\mathfrak{K}$-representable sets and functions – is built. In contrast to calculus rules which are, basically, independent of what is the family $\mathfrak{K}$ of cones in question, raw materials do depend on $\mathfrak{K}$. Here we restrict ourselves with few instructive examples of Lorentz- and Semidefinite-representable sets and functions; for in-depth acquaintance with this topic, we refer the reader to [BTN].

We understand well what are the "atomic" $\mathfrak{N}$-representable functions and sets – these are half-spaces and affine functions. Other polyhedrally representable sets are intersections of finite families of half-spaces, and other polyhedrally representable functions – maxima of finitely many affine functions restricted on a polyhedral domain. In other words, all $\mathfrak{N}$-representable functions and sets are obtained from the above atoms via the calculus we have just outlined.

In the next two exercises we present instructive examples of $\mathfrak{L}$-r functions and sets.

**Exercise** IV.38 ▲ [$\mathfrak{L}$-representability of $\|\cdot\|_2$ and $\|\cdot\|_2^2$] Check that the functions $\|x\|_2$ and $x^\top x$ on $\mathbf{R}^n$ admits $\mathfrak{L}$-r.'s as follows:

$$\begin{array}{l}\{[x;t]\in\mathbf{R}^n_x\times\mathbf{R}_t : t\geq\|x\|_2\}=\left\{[x;t]\in\mathbf{R}^n\times\mathbf{R}:[x;t]\in\mathbf{L}^{n+1}\right\}\\ \{[x;t]\in\mathbf{R}^n_x\times\mathbf{R}_t : t\geq x^\top x\}=\left\{[x;t]\in\mathbf{R}^n\times\mathbf{R}:[2x;t-1;t+1]\in\mathbf{L}^{n+2}\right\}\end{array}$$

**Exercise** IV.39 ♦ [$\mathfrak{L}$-representability of power functions] Justify the following claims

1. Let $k$ be a positive integer. Then the set

$$\mathfrak{S}_k=\left\{[t;x_1;x_2;\dots;x_{2^k}]\geq 0 : t\leq\left[\prod_{i=1}^{2^k}x_i\right]^{1/2^k}\right\}$$

– the intersection of the hypograph of the geometric mean of $2^k$ nonnegative variables $x_1,...,x_{2^k}$ with the half-space $\{[t;x]\in\mathbf{R}^{2^k}_x\times\mathbf{R}_t : t\geq 0\}$ – admits $\mathfrak{L}$-representation, specifically,

$$\mathfrak{S}_k=\left\{\begin{array}{l}[t;x_1;x_2;\dots;x_{2^k}]\geq 0:\exists\{u_{i,\ell}\geq 0, 1\leq\ell\leq k, 1\leq i\leq 2^\ell\}:\\ \qquad u_{ik}=x_i, 1\leq i\leq 2^k\\ \qquad \left.\begin{array}{r}[2u_{i\ell};u_{2i-1,\ell+1}-u_{2i,\ell+1};u_{2i-1,\ell+1}+u_{2i,\ell+1}]\in\mathbf{L}^3,\\ 1\leq i\leq 2^\ell, 1\leq\ell<k\end{array}\right\}\\ \qquad [2t;u_{1,1}-u_{2,1};u_{1,1}+u_{2,1}]\in\mathbf{L}^3.\end{array}\right. \qquad (*)$$

Surprisingly, item 1 paves road to $\mathfrak{L}$-representations of power functions.

2. Build explicit $\mathfrak{L}$-r's of the univariate functions as follows:

2.1. $f(x)=\max[0,x]^\theta$ with rational $\theta=p/q\geq 1$.

2.2. $f(x)=\left\{\begin{array}{ll}x^{p_+/q_+} & ,x\geq 0\\ |x|^{p_-/q_-} & ,x\leq 0\end{array}\right.$ , where $p_\pm$, $q_\pm$ are positive integers with $p_+/q_+\geq 1$, $p_-/q_-\geq 1$

2.3. $f(x)=\left\{\begin{array}{ll}-x^{p/q} & ,x\geq 0\\ +\infty & ,x<0\end{array}\right.$ with positive integers $p,q$ such that $p/q\leq 1$

2.4. $f(x)=\left\{\begin{array}{ll}x^{-p/q} & ,x>0\\ +\infty & ,x\leq 0\end{array}\right.$ with positive integers $p,q$

3. Build $\mathfrak{L}$-r's of the following sets:

3.1. The hypograph

$$\{[x;t]\in\mathbf{R}^n_+\times\mathbf{R}_t : t\leq f(x):=x_1^{\pi_1}x_2^{\pi_2}\dots x_n^{\pi_n}\}$$

of algebraic monomial of $n$ nonnegative variables, where $\pi_i$ are positive rationals such that $\sum_i\pi_i\leq 1$ (the latter inequality for nonnegative $\pi_i$'s is a necessary and sufficient for $f$ to be concave on $\mathbf{R}^n_+$).

3.2. The epigraph of algebraic monomial $f(x)=x_1^{-\pi_1}x_2^{-\pi_2}\dots x_n^{-\pi_n}$ of $n$ positive variables, where $\pi_i$ are positive rationals.

3.3. The epigraph of $\|\cdot\|_\pi$ on $\mathbf{R}^n$ with rational $\pi\geq 1$.

By Exercise IV.34, expressive abilities of semidefinite representations are at least as strong as those of Lorentz representability. In fact, $\mathfrak{S}$-representability is strong enough to bring, "for all practical purposes," the entire Convex Optimization within the grasp of Semidefinite Optimization. In our next exercise we are just touching the tip of the "semidefinite iceberg."

**Exercise** IV.40 ♦

1. For starters, build $\mathfrak{S}$-r's of the maximum eigenvalue of a symmetric matrix and of the spectral norm $\|\cdot\|_{2,2}$ (the maximum singular value) of a rectangular matrix.
   *Hint:* Note that for a $p \times q$ matrix $A$, the eigenvalues of the symmetric $(p+q)\times(p+q)$ matrix $\left[\begin{array}{c|c} & A \\ \hline A^\top & \end{array}\right]$ are the singular values of $A$, minus these singular values, and perhaps a number of zeros.

As a matter of fact, the single most valuable $\mathfrak{S}$-representation is the one for the sums $S_k(X)$ of $k$ largest eigenvalues of a symmetric matrix $X$; convexity of these sums in $X$ was established in chapter 18.

2. Build $\mathfrak{S}$-r. of the sum $S_k(X)$ of $k \le m$ largest eigenvalues of $m\times m$ symmetric matrix $X$.
   *Hint:* Recall the polyhedral representation, built in Exercise I.29, of the "vector analogy" of $S_k(X)$ – the sum $s_k(x)$ of $k$ largest entries in $m$-dimensional vector $x$:

$$t \ge s_k(x) \iff \exists z \ge 0, s : x \le z + s\mathbf{1}, \sum_i z_i + ks \le t,$$

   where $\mathbf{1}$ is the all-ones vector.

The importance of $\mathfrak{S}$-representability of $S_k(\cdot)$ becomes clear from the following

3. Let $f(x) : \mathbf{R}^m \to \mathbf{R}\cup\{+\infty\}$ be a convex function symmetric with respect to permutations of entries in the argument, and let

$$F(X) = f(\lambda(X)) : \mathbf{S}^m \to \mathbf{R}\cup\{+\infty\};$$

   recall that $F$ is convex by Proposition III.18.3. Show that $F(X)$ admits the following representation:

$$t \ge F(x) \iff \exists u \in \mathbf{R}^m : \begin{array}{ll} f(u) \le t & (a) \\ u_1 \ge u_2 \ge \ldots \ge u_m & (b) \\ S_k(X) \le u_1 + \ldots + u_k,\ 1 \le k < m & (c_k) \\ \mathrm{Tr}(X) = u_1 + \ldots + u_m & (c_m) \end{array} \qquad (29.3)$$

   Combine this fact with $\mathfrak{S}$-representability of $S_k(\cdot)$ to arrive at the following

**Corollary** IV.29.1 In the situation of item 3, assume that $f$ is not just symmetric, but is $\mathfrak{S}$-representable as well. A $\mathfrak{S}$-r. of $f$ gives rise to explicit $\mathfrak{S}$-r. of $F(X)$.

Corollary underlies $\mathfrak{S}$-representations of numerous highly important functions and sets, e.g., *Shatten norms* of rectangular matrices – $p$-norms of the vector of matrix's singular values, or the hypograph $t \le \mathrm{Det}^{1/m}(X)$ of the (appropriate power of the) determinant of $X \in \mathbf{S}^m_+$, or the epigraph of the function $\mathrm{Det}^{-1}(X)$ of $X \succ 0$.

**Exercise** IV.41 ♦ An interesting example of $\mathfrak{S}$-representable sets deals with matrix square and matrix square root:

1. [$\succeq$-epigraph of the matrix square] Prove that the function $F(X) = X^\top X : \mathbf{R}^{m\times n} \to \mathbf{S}^n$ is $\succeq$-convex and find a $\mathfrak{S}$-r. of its $\succeq$-epigraph $\{(X,Y) \in \mathbf{R}^{m\times n}\times\mathbf{S}^n : Y \succeq X^\top X\}$.
2. [$\succeq$-hypograph of the matrix square root] Prove that the set $\{(X,Y) \in \mathbf{S}^n\times\mathbf{S}^n : X \succeq 0, Y \preceq X^{1/2}\}$ is convex and find its $\mathfrak{S}$-r.

Note: Solutions to items 1–2 provide us with $\mathfrak{S}$-r's of the sets $\{(X,Y) \in \mathbf{S}^n\times\mathbf{S}^n : X \succeq 0, 0 \preceq X \preceq Y^{1/2}\}$ and $\{(X,Y) \in \mathbf{S}^n\times\mathbf{S}^n : X \succeq 0, X^2 \preceq Y\}$. These sets are different, and the second is "essentially smaller" than the first one, see Exercise IV.17.

**Exercise** IV.42 ♦ [important example of $\mathfrak{S}$-representation] Consider the situation as follows. Given a *basic set* $\mathcal{B} \subset \mathbf{R}^n$ which is the solution set of a strictly feasible quadratic inequality:

$$\mathcal{B} = \{u \in \mathbf{R}^n : u^\top Qu + 2q^\top u + \kappa \leq 0\},$$

we consider *target set*

$$\mathcal{Q} = \{x \in \mathbf{R}^m : x^\top Sx + 2s^\top x + \sigma \leq 0\} \qquad [S \in \mathbf{S}^m, s \in \mathbf{R}^m, \sigma \in \mathbf{R}]$$

and affine mapping

$$u \mapsto P(x) := Pu + p : \mathbf{R}^n \to \mathbf{R}^m.$$

We are interested in the situation when the image of the basic set under the mapping $P(\cdot)$ is contained in the target set, and want to describe this situation in terms of the parameters $S, s, \sigma, P, p$.

Your task is as follows. Let us set

$$\mathcal{M}(S, s, \sigma; P, p; \lambda) = [P, p]^\top S[P, p] + \left[\begin{array}{c|c} -\lambda Q & P^\top s - \lambda q \\ \hline s^\top P - \lambda q^\top & 2s^\top p + \sigma - \lambda\kappa \end{array}\right].$$

Prove that the inclusion $P(\mathcal{B}) \subset \mathcal{Q}$ is equivalent to the existence of $\lambda \geq 0$ such that

$$\mathcal{M}(S, s, \sigma; P, p; \lambda) \preceq 0. \tag{!}$$

*Why this is important:* Pay attention to the fact that

1. When $(S, s, \sigma)$ is fixed and $S \succeq 0$, $\mathcal{M}(S, s, \sigma; P, p; \lambda)$ is $\succeq$-convex in $(P, p, \lambda)$; moreover, introducing $\succeq$-upper bound $V$ on $[P, p]^\top S[P, p]$, we have
$$\begin{array}{c} \mathcal{M}(S, s, \sigma; P, p; \lambda) \preceq 0 \\ \Updownarrow \\ \exists V \in \mathbf{S}^{m+1} : \left[\begin{array}{c|c} V & [P, p]^\top S^{1/2} \\ \hline S^{1/2}[P, p] & I_n \end{array}\right] \succeq 0 \ \& \ V + \left[\begin{array}{c|c} -\lambda Q & P^\top s - \lambda q \\ \hline s^\top P - \lambda q^\top & 2s^\top p + \sigma - \lambda\kappa \end{array}\right] \preceq 0, \end{array}$$
and we arrive at $\mathfrak{S}$-r. of the set of parameters $P, p$ of affine mappings mapping $\mathcal{B}$ into $\mathcal{Q}$.
2. When $P, p$ are fixed, (!) is a linear matrix inequality in variables $S, s, \sigma$; as a result, we get an $\mathfrak{S}$-r. of the set of parameters $S, s, \sigma$ of quadratic forms resulting in $P(\mathcal{B}) \subset \mathcal{Q}$ for fixed $P(\cdot)$.

Item 1 here allows to pose as an explicit semidefinite program the arising in numerous applications problem of finding the largest volume ellipsoid contained in the intersection $\mathcal{S}$ of finitely many ellipsoids (or, more generally, sublevel sets of convex quadratic functions, e.g., linear functions). To this end it suffices to specify the basic set as the unit Euclidean ball in $\mathbf{R}^m$ and to note that an ellipsoid ("flat" of full-dimensional) in $\mathbf{R}^m$ is the image of this ball under affine transformation $u \mapsto Pu + p$ with symmetric positive semidefinite $P$. Item 1 says that the set of parameter $(P \succeq 0, p)$ of ellipsoids contained in $\mathcal{S}$ admits explicit $\mathfrak{S}$-r. Taking into account that the volume of the ellipsoid $P\mathcal{B} + p$ with $P \succeq 0$ is proportional to $\mathrm{Det}(P)$ and that the function $-\mathrm{Det}^{1/m}(P)$ of $P \in \mathbf{S}^m_+$ admits explicit $\mathfrak{S}$-r., see Exercise IV.40.3, we arrive at the semidefinite reformulation of the problem of interest.

Similarly, item 2 allows to handle another problem with a wide spectrum of applications – the problem of finding the smallest volume ellipsoid containing the union of finitely many given ellipsoids. Indeed, representing these ellipsoids as the images of the unit ball in $\mathbf{R}^m$ under given affine mappings, representing the ellipsoid of interest as

$$E = \{x \in \mathbf{R}^n : (x - c)^\top S(x - c) \leq 1\}$$

with $S \succ 0$, and passing from variables $S, c$ to variables $S, s = -Sc$, we get

$$\begin{array}{c} E = \{x \in \mathbf{R}^n : x^\top Sx + 2s^\top x + \sigma(S, s) \leq 0\}, \\ [S \succ 0, \sigma(S, s) = s^\top S^{-1}s - 1] \end{array} \tag{!!}$$

so that the fact that $E$ contains a given ellipsoid $P\mathcal{B} + p$ is equivalent to the existence of $\lambda \geq 0$ such that

$$\mathcal{M}(S, s, \sigma(S, s); P, p; \lambda) = [P, p]^\top S[P; p] + \left[\begin{array}{c|c} -\lambda Q & P^\top s - \lambda q \\ \hline s^\top P - \lambda q^\top & 2s^\top p + s^\top S^{-1}s - 1 - \lambda\kappa \end{array}\right] \preceq 0,$$

or, which is the same, to the existence of $\lambda \geq 0$ and $\mu$ such that

$$[P.p]^\top S[P; p] + \left[\begin{array}{c|c} -\lambda Q & P^\top s - \lambda q \\ \hline s^\top P - \lambda q^\top & 2s^\top p + \mu - 1 - \lambda\kappa \end{array}\right] \preceq 0, \ \left[\begin{array}{c|c} S & s \\ \hline s^\top & \mu \end{array}\right] \succeq 0.$$

Thus, we can point out an explicit $\mathfrak{S}$-r. of the set of parameters $(S \succ 0, s)$ of ellipsoids containing all ellipsoids from a given finite collection. The volume of ellipsoid (!!) is proportional to $\mathrm{Det}^{-1/2}S$, so that the problem of interest is to maximize $\mathrm{Det}(S)$ over a subset of $\mathbf{S}^m_+$ given by an $\mathfrak{S}$-r. Recalling that the function $-\mathrm{Det}^{1/m}(S)$ of $S \succeq 0$ admits explicit $\mathfrak{S}$-r., we again reduce the problem of interest to an explicit semidefinite program.

As a simple illustration, consider the *inscribed ellipsoid* algorithm[7]. This algorithm is aimed at solving optimization problem

$$\mathrm{Opt} = \min_{x\in\mathbf{R}^n} \{f(x) : \|x\|_\infty \leq 1\}$$

where $f$ is convex continuous function on the unit box $B_n = \{x \in \mathbf{R}^n : \|x\|_\infty \leq 1\}$. The algorithm is applicable when one can compute the value and a subgradient of $f$ at any desired point $x \in \mathrm{int}\, B_n$ and works as follows: at the beginning of iteration $i = 1, 2, \ldots$ we have at our disposal a *localizer* $G_i$ - a polytope known to belong to $B_n$ and to contain the optimal set $X_*$ of the problem, with $G_1 = B_n$. At iteration $i$ we

- compute $i$-th *search point* $x_i$ – the center of the largest volume ellipsoid $E_i$ contained in $G_i$; we already know that this requires solving auxiliary semidefinite problem and can be done efficiently;
- compute $f(x_i)$ and a subgradient $g_i$ of $f$ at $x_i$. We define the *approximate solution* $x^i$ generated at the first $i$ iterations as the best – with the smallest value of the objective – among the search points $x_1, \ldots, x_i$, and set $f^i = f(x^i)$;
- update the localizer by setting

$$G_{i+1} = \{x \in G_i : f(x_i) + g_i^T(x - x_i) \leq f^i\}.$$

Note that the polytope $G_{i+1}$ indeed is a localizer. Indeed, $X_* \subset G_i$, since $G_i$ is a localizer; the only possibility for the inclusion $X_* \subset G_{i+1}$ to be violated is for some $x_* \in X_*$ to satisfy $f(x_i)+g_i^T(x_*-x_i) > f^i$, which is impossible – the left hand side in this inequality is $\leq f(x_*)$ due to $g_i \in \partial f(x_i)$, and $f(x_*) \leq f^i = f(x^i)$.

The above recurrence is terminated when the average linear size $\mathrm{Vol}^{1/n}(E_i)$ of $E_i$ becomes less than $\epsilon$, where $\epsilon \in (0, 1/2)$ is a prescribed tolerance. Theory says that

- the number $I$ of iterations before termination is bounded by $O(1)n\ln(1/\epsilon)$, and
- the resulting approximate solution $x^I \in B_n$ satisfies the error bound

$$f(x^I) - \mathrm{Opt} \leq \epsilon[\max_{B_n} f - \min_{B_n} f].$$

Moreover, the outlined algorithm possesses provably optimal (in certain precise sense not to be discussed here) complexity.

An interested reader is highly recommended to implement and run this algorithm, preferably on a low-dimensional problem (say, with $n = 5$ or $n = 10$) in order to make solving auxiliary problems not too time-consuming.

Recommended setup:

- $n = 5$ or $n = 10$
- $f(x) = \max_{i\leq 1000n} |a_i^T x + b_i|^3$ with i.i.d. $\mathcal{N}(0, 1)$ entries in $a_i$ and $b$
- $\epsilon = 1.\mathrm{e}{-6}$

[7] S.P Tarasov, L.G. Khachiyan, I.I. Erlikh, ”The method of inscribed ellipsoids”, *Dokl. Akad. Nauk SSSR*, 298:5 (1988), 1081–1085; English translation: *Dokl. Math.*, 37:1 (1988), 226–230.

# 30

# Proofs of Facts from Part IV

**Fact IV.21.6**

(i) A cone $\mathbf{K} \subseteq \mathbf{R}^n$ is regular if and only if its dual cone $\mathbf{K}_* = \{y \in \mathbf{R}^n : y^\top x \geq 0, \forall x \in \mathbf{K}\}$ is regular.

(ii) Given regular cones $\mathbf{K}_1, \ldots, \mathbf{K}_m$, their direct product $\mathbf{K}_1 \times \ldots \times \mathbf{K}_m$ is also regular.

Proof. (i) This is an immediate consequence of Fact II.8.23.

(ii) This is evident. □

**Fact IV.21.7**

The following cones (see Examples discussed in section 1.2.4) are regular:

1. *Nonnegative ray* $\mathbf{R}_+$,
2. *Lorentz* (a.k.a., *second-order*, or *ice-cream*) cone, $\mathbf{L}^n = \{x \in \mathbf{R}^n : x_n \geq \sqrt{x_1^2 + \ldots + x_{n-1}^2}\}$ ($\mathbf{L}^1 := \mathbf{R}_+$)
3. *Positive semidefinite cone*, $\mathbf{S}^n_+ = \{X \in \mathbf{S}^n : a^\top X a \geq 0, \forall a \in \mathbf{R}^n\}$.

Proof. Regularity of the above cones is self-evident. □

**Fact IV.21.10** Let $\mathbf{K}$ be a regular cone in $\mathbf{R}^\nu$, $Z \subseteq \mathbf{R}^n$ be a nonempty convex set and $h : Z \to \mathbf{R}^\nu$ be a mapping with $\operatorname{dom} h = Z$. Then, $h$ is $\mathbf{K}$-convex if and only if for all $\mu \in \mathbf{K}_*$ (where $\mathbf{K}_*$ is the cone dual to $\mathbf{K}$) the real valued functions $\mu^\top h(\cdot)$ are convex on $Z$.

Proof. Since $\mathbf{K}$ is a closed cone, it is the cone dual to $\mathbf{K}_*$, that is, (21.5) takes place for all $x, y \in Z$ and $\lambda \in [0,1]$ if and only if

$$\mu^\top h(\lambda x + (1-\lambda)y) \leq \lambda \mu^\top h(x) + (1-\lambda)\mu^\top h(y), \quad \forall (x, y \in Z, \lambda \in [0,1], \mu \in \mathbf{K}_*).$$

□

**Fact IV.23.2** Consider parametric family $(P_\Delta)$ of convex cone-constrained problems along with the family $(D_\Delta)$ of their cone-constrained Lagrange duals. Then,

(i) If $\underline{L}_0(\mu) > -\infty$ for some $\mu = [\overline{\mu}; \widehat{\mu}] \in \Lambda$, then the primal optimal value $\mathrm{Opt}(P_\Delta)$ takes values in $\mathbf{R} \cup \{+\infty\}$ and is a convex function of $\Delta$.

(ii) If $(D_0)$ is solvable with optimal solution $\lambda_* = [\overline{\lambda}_*; \widehat{\lambda}_*]$ and $\mathrm{Opt}(D_0) = \mathrm{Opt}(P_0)$, then $-\lambda_*$ is a subgradient of $\mathrm{Opt}(P_\Delta)$ at the point $\Delta = 0$, i.e.,

$$\mathrm{Opt}(P_\Delta) \geq \mathrm{Opt}(P_0) - \overline{\lambda}_*^\top \overline{\delta} - \widehat{\lambda}_*^\top \widehat{\delta}, \qquad \forall (\Delta = [\overline{\delta}; \widehat{\delta}]).$$

The premises in (i) and (ii) definitely take place when $(P_0)$ satisfies the Relaxed

Slater condition and is below bounded.

Proof. Under the premise of (i), by Weak duality we have for all $\Delta$

$$\mathrm{Opt}(P_\Delta) \geq \underline{L}_\Delta(\mu) = \underline{L}_0(\mu) - \overline{\mu}^\top \overline{\delta} - \widehat{\mu}^\top \widehat{\delta}. \tag{30.1}$$

That is, for all $\Delta$, $\mathrm{Opt}(P_\Delta) > -\infty$ (as $\underline{L}_0(\mu) > -\infty$). To prove that in this case $\mathrm{Opt}(P_\Delta)$ is convex in $\Delta$, it suffices to verify that when $\Delta', \Delta'' \in \mathrm{Dom}\,\mathrm{Opt}(P.)$ and $\theta \in (0,1)$, we have $\mathrm{Opt}(P_{\theta\Delta'+(1-\theta)\Delta''}) \leq \theta\mathrm{Opt}(P_{\Delta'}) + (1-\theta)\mathrm{Opt}(P_{\Delta''})$. Let us denote $\Delta' = [\overline{\delta}'; \widehat{\delta}']$ and $\Delta'' = [\overline{\delta}''; \widehat{\delta}'']$. Then, by the definitions of $\mathrm{Opt}(P_{\Delta'})$ and $\mathrm{Opt}(P_{\Delta''})$, given $\epsilon > 0$, we can find $x'_\epsilon, x''_\epsilon \in X$ such that

$$\begin{aligned} f(x'_\epsilon) &\leq \mathrm{Opt}(P_{\Delta'}) + \epsilon, \qquad \overline{g}(x'_\epsilon) \leq \overline{\delta}', \qquad \widehat{g}(x'_\epsilon) \leq_{\mathbf{K}} \widehat{\delta}', \\ f(x''_\epsilon) &\leq \mathrm{Opt}(P_{\Delta''}) + \epsilon, \qquad \overline{g}(x''_\epsilon) \leq \overline{\delta}'', \qquad \widehat{g}(x''_\epsilon) \leq_{\mathbf{K}} \widehat{\delta}''. \end{aligned}$$

Hence, by convexity of $f$ and $\overline{g}$ and $\mathbf{K}$-convexity of $\widehat{g}$, the point $x_\epsilon := \theta x'_\epsilon + (1-\theta)x''_\epsilon$ satisfies

$$f(x_\epsilon) \leq (\theta\mathrm{Opt}(P_{\Delta'}) + (1-\theta)\mathrm{Opt}(P_{\Delta''}))+\epsilon, \quad \overline{g}(x_\epsilon) \leq \theta\overline{\delta}'+(1-\theta)\overline{\delta}'', \quad \widehat{g}(x_\epsilon) \leq_{\mathbf{K}} \theta\widehat{\delta}'+(1-\theta)\widehat{\delta}'',$$

implying that $\mathrm{Opt}(P_{\theta\Delta'+(1-\theta)\Delta''}) \leq (\theta\mathrm{Opt}(P_{\Delta'}) + (1-\theta)\mathrm{Opt}(P_{\Delta''}))+\epsilon$. Since $\epsilon > 0$ is arbitrary, we arrive at the desired relation $\mathrm{Opt}(P_{\theta\Delta'+(1-\theta)\Delta''}) \leq \theta\mathrm{Opt}(P_{\Delta'}) + (1-\theta)\mathrm{Opt}(P_{\Delta''})$. This justifies part (i).

To justify (ii), note that under the premise of this claim, the premise of (i) is satisfied by taking $\mu := \lambda_*$. Then, (30.1) holds true for this $\mu$ as well. Moreover, as $\underline{L}_\Delta(\lambda_*) = \mathrm{Opt}(P_0)$, we see that (30.1) with $\mu = \lambda_*$ says that $-\lambda_*$ is a subgradient of $\mathrm{Opt}(P.)$ at the origin.

Finally, when $(P_0)$ is below bounded and satisfies the relaxed Slater condition, the premise of (ii) (and therefore of (i) as well) holds true by Theorem IV.23.1. □

**Fact IV.23.3** Consider the convex quadratic constraint $x^\top A^\top Ax \leq b^\top x + c$, where $A \in \mathbf{R}^{d\times n}$, $b \in \mathbf{R}^n$, and $c \in \mathbf{R}$. This constraints can be equivalently rewritten as a conic constraint involving Lorentz cone, i.e.,

$$\begin{aligned} & x^\top A^\top Ax \leq b^\top x + c \\ \Longleftrightarrow\ & [2Ax;\ b^\top x + c - 1;\ b^\top x + c + 1] \in \mathbf{L}^{d+2} \\ \Longleftrightarrow\ & 4x^\top Ax + (b^\top x + c - 1)^2 \leq (b^\top x + c + 1)^2 \ \text{ and } \ b^\top x + c + 1 \geq 0. \end{aligned}$$

Proof. Note that $b^\top x + c = \frac{(b^\top x+c+1)^2-(b^\top x+c-1)^2}{4}$. Thus, if $x$ is feasible to the quadratic constraint (and, in particular, $b^\top x + c \geq 0$, implying that $b^\top x + c + 1 > 0$), $x$ is feasible for the conic constraint, and clearly vice versa. □

**Fact IV.23.5** Consider the conic problem given in (23.5)

$$\mathrm{Opt}(P) = \min_{x\in\mathbf{R}^n} \left\{c^\top x :\ Ax - b \leq 0,\ Px - p \leq_{\mathbf{K}} 0\right\}, \tag{$P$}$$

along with its conic dual problem

$$\mathrm{Opt}(D) = \max_{\overline{\lambda},\widehat{\lambda}} \left\{-b^\top\overline{\lambda} - p^\top\widehat{\lambda} :\ A^\top\overline{\lambda} + P^\top\widehat{\lambda} + c = 0,\ \overline{\lambda} \geq 0,\ \widehat{\lambda} \in \mathbf{K}_*\right\}. \tag{$D$}$$

Conic duality is symmetric, i.e., the conic dual to conic problem $(D)$ is (equivalent to) conic problem $(P)$, i.e., (23.5).

Proof. In order to apply our recipe for building the conic dual to a conic problem, let us rewrite $(D)$ in the minimization form with $\leq 0$-type affine constraints and a single conic constraint, i.e.,

$$-\mathrm{Opt}(D) = \min_{\overline{\lambda},\widehat{\lambda}} \left\{ b^\top\overline{\lambda} + p^\top\widehat{\lambda} :\ \begin{array}{rl} A^\top\overline{\lambda} + P^\top\widehat{\lambda} + c & \leq 0 \\ -A^\top\overline{\lambda} - P^\top\widehat{\lambda} - c & \leq 0 \\ -\overline{\lambda} & \leq 0 \end{array},\ -\widehat{\lambda} \in -\mathbf{K}_* \right\} \tag{$D$}$$

The dual to this problem, in view of $[\mathbf{K}_*]_* = \mathbf{K}$, reads

$$\max_{u,v,w,y} \left\{c^\top [u-v] : \; b + A[u-v] - w = 0, \; P[u-v] + p - y = 0, \; u \geq 0, \; v \geq 0, \; w \geq 0, \; y \in \mathbf{K}\right\}.$$

By setting $x := v - u$ and eliminating $y$ and $w$, the latter problem becomes

$$\max_x \left\{-c^\top x : \; Ax - b \leq 0, \; Px - p \leq_{\mathbf{K}} 0\right\},$$

which is noting but $(P)$. □

**Fact IV.28.4** Let $X$ and $\Lambda$ be compact sets in $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively, and let $L(x,\lambda) : X \times \Lambda \to \mathbf{R}$ be a continuous function. Then, the problems (P) and (D) are solvable.

Proof. Since $X$ and $\Lambda$ are compact sets and $L$ is continuous on $X \times \Lambda$, due to the well-known Analysis theorem (Theorem B.29 and Remark B.30) $L$ is uniformly continuous on $X \times \Lambda$. That is, for every $\epsilon > 0$ there exists $\delta(\epsilon) > 0$ such that

$$\|x - x'\|_2 + \|\lambda - \lambda'\|_2 \leq \delta(\epsilon) \quad \Longrightarrow \quad |L(x,\lambda) - L(x',\lambda')| \leq \epsilon. \tag{30.2}$$

In particular, for every $\lambda = \lambda' \in \Lambda$ we have

$$\|x - x'\|_2 \leq \delta(\epsilon) \quad \Longrightarrow \quad |L(x,\lambda) - L(x',\lambda)| \leq \epsilon,$$

which immediately implies

$$\|x - x'\|_2 \leq \delta(\epsilon) \quad \Longrightarrow \quad |\overline{L}(x) - \overline{L}(x')| \leq \epsilon,$$

so that the function $\overline{L}$ is continuous on $X$. This together with the fact that $X$ is compact imply that the problem (P) is solvable. The same reasoning applied to the variables $\lambda$ leads to the conclusion that problem (D) is solvable. □

# Appendix A

# Prerequisites from Linear Algebra

> Note: Appendices A – D reproduce , courtesy of World Scientific Publishing Co., appendices A – C in the textbook A. Nemirovski, *Introduction to Linear Optimization*, World Scientific, 2024.

Regarded as mathematical entities, the objective and the constraints in a Mathematical Programming problem are functions of several real variables; therefore before entering the Optimization Theory, we need to recall several basic notions and facts about the spaces $\mathbf{R}^n$ where these functions live, same as about the functions themselves.

The following material is considered basic, and we expect the reader to be already familiar with it, and thus we use a "cook book" style here.

## A.1 Space $\mathbf{R}^n$: algebraic structure

Basically all events and constructions to be considered will take place in the *space* $\mathbf{R}^n$ *of* $n$*-dimensional real vectors*. This space can be described as follows.

### A.1.1 A point in $\mathbf{R}^n$

*A point* in $\mathbf{R}^n$ (called also an $n$*-dimensional vector*) is an ordered collection $x = [x_1; \dots; x_n]$ of $n$ reals, called the *coordinates*, or *components*, or *entries* of vector $x$; the space $\mathbf{R}^n$ itself is the set of all collections of this type. Note that we follow MATLAB notation for representing vectors and matrices. That is, the notation $x = [x_1; \dots; x_n]$ represents a column vector, while $x = [x_1, \dots, x_n]$ represents a row vector.

### A.1.2 Linear operations

$\mathbf{R}^n$ is equipped with two *basic operations*:

- *Addition of vectors*. This operation takes on input two vectors $x = [x_1; \dots; x_n]$ and $y = [y_1; \dots; y_n]$ and produces from them a new vector with entries which are sums of the corresponding entries in $x$ and in $y$, i.e.,

$$x + y = [x_1 + y_1; \dots; x_n + y_n].$$

- *Multiplication of vectors by reals.* This operation takes on input a real $\lambda$ and an $n$-dimensional vector $x = [x_1; \dots; x_n]$ and produces from them a new vector with entries which are $\lambda$ times the entries of $x$, i.e.,

$$\lambda x = [\lambda x_1; \dots; \lambda x_n].$$

As far as addition and multiplication by reals are concerned, the arithmetic of $\mathbf{R}^n$ inherits most of the common rules of Real Arithmetic, like $x+y = y+x$, $(x+y)+z = x+(y+z)$, $(\lambda+\mu)(x+y) = \lambda x + \mu x + \lambda y + \mu y$, $\lambda(\mu x) = (\lambda\mu)x$, etc.

In the above presentation, it was tacitly assumed that $n$ is a positive integer. To avoid unnecessary trivial comments in the sequel, it makes sense to define $\mathbf{R}^0$ as the linear space with exactly one element, denoted by 0, and the only possible in this case linear operations: $0+0=0$, $\lambda \cdot 0 = 0$, $\lambda \in \mathbf{R}$.

## A.2 Linear subspaces

**Definition** A.1 [Linear subspace] A *linear subspace* in $\mathbf{R}^n$ is, by definition, a nonempty subset of $\mathbf{R}^n$ which is closed with respect to addition of vectors and multiplication of vectors by reals. That is,

$$L \subseteq \mathbf{R}^n \text{ is a linear subspace} \quad \Longleftrightarrow \quad \left\{ \begin{array}{c} L \neq \varnothing; \\ x, y \in L \implies x + y \in L; \\ x \in L, \lambda \in \mathbf{R} \implies \lambda x \in L. \end{array} \right.$$

### A.2.1 Examples of linear subspaces

We have some immediate examples of linear subspaces.

**Example** A.2 Each one of the following is a linear subspace:

1. The entire $\mathbf{R}^n$.
2. The *trivial* subspace containing the single zero vector $0 = [0; \dots; 0]$ (this vector/point is called also *the origin*).
   Here, pay attention to the notation: we use the same symbol 0 to denote the real zero and the $n$-dimensional vector with all coordinates equal to zero; these two zeros are not the same, and one should understand from the context (it is always very easy) which zero is meant.
3. The set $\{x \in \mathbf{R}^n : x_1 = 0\}$ of all vectors $x$ with the first coordinate equal to zero.

In fact, the last example in the above list admits a natural extension:

**Example** A.3 The set of all solutions to a *homogeneous* (i.e., with zero right hand side) system of linear equations

$$\left\{ x \in \mathbf{R}^n : \begin{array}{rcl} a_{11}x_1 + \ldots + a_{1n}x_n & = & 0 \\ a_{21}x_1 + \ldots + a_{2n}x_n & = & 0 \\ \ldots & & \\ a_{m1}x_1 + \ldots + a_{mn}x_n & = & 0 \end{array} \right\} \tag{A.1}$$

is always a linear subspace in $\mathbf{R}^n$.

In fact, we will see in Proposition A.46 that this example is "generic," that is, *every* linear subspace in $\mathbf{R}^n$ is the solution set of a (finite) system of homogeneous linear equations.

**Definition** A.4 [Linear combination] Given a set of vectors $x^1, \ldots, x^N \in \mathbf{R}^n$ and a set of reals $\lambda_1, \ldots, \lambda_N$, the vector $\sum_{i=1}^N \lambda_i x^i$ is called a *linear combination* of the vectors $x^1, \ldots, x^N$, and the reals $\lambda_1, ..., \lambda_N$ are called the *coefficients* of the combination.

**Definition** A.5 [Linear span] Given a nonempty set $X \subseteq \mathbf{R}^n$ of vectors, the *linear span* of $X$ [notation: $\mathrm{Lin}(X)$] is the linear subspace that consists of all vectors $x$ which can be represented as linear combinations of vectors from $X$. That is,

$$\mathrm{Lin}(X) := \left\{ x \in \mathbf{R}^n : \exists N, x^i \in X, i = 1, \ldots, N, \lambda \in \mathbf{R}^N \text{ s.t. } x = \sum_{i=1}^N \lambda_i x^i \right\}.$$

By definition, the linear span of empty set is the trivial linear subspace, the origin, i.e., $\mathrm{Lin}(\varnothing) = \{0\}$.

**Example** A.6 Given a nonempty set $X \subseteq \mathbf{R}^n$ of vectors, their linear span $\mathrm{Lin}(X)$ is a linear subspace.

The definition of linear span also immediately points out to an "outer" characterization as follows.

**Fact** A.7 For $X \subset \mathbf{R}^n$, $\mathrm{Lin}(X)$ is the smallest linear subspace which contains $X$: if $L$ is a linear subspace such that $L \supseteq X$, then $L \supseteq \mathrm{Lin}(X)$.

We will see in Theorem A.16 that the "linear span" example is generic as well. That is, *every linear subspace in* $\mathbf{R}^n$ *is the linear span of an appropriately chosen finite set of vectors from* $\mathbf{R}^n$. Note that this latter characterization of the linear span is of "inner" description type.

### A.2.2 Sums and intersections of linear subspaces

Let $\{L_\alpha\}_{\alpha\in I}$ be a family (finite or infinite) of linear subspaces of $\mathbf{R}^n$. From this family, one can build two sets:

1. *The sum* $\sum\limits_\alpha L_\alpha$ of the subspaces $L_\alpha$ which consists of all vectors which can be represented as finite sums of vectors taken each from its own subspace of the family;
2. *The intersection* $\bigcap\limits_\alpha L_\alpha$ of the subspaces from the family.

**Theorem** A.8 Let $\{L_\alpha\}_{\alpha\in I}$ be a family of linear subspaces of $\mathbf{R}^n$. Then,

(i) The sum $\sum\limits_\alpha L_\alpha$ of the subspaces from the family is itself a linear subspace of $\mathbf{R}^n$; it is the smallest of those subspaces of $\mathbf{R}^n$ which contain every subspace $L_\alpha$ from the family;

(ii) The intersection $\bigcap\limits_\alpha L_\alpha$ of the subspaces from the family is itself a linear subspace of $\mathbf{R}^n$; it is the largest of those subspaces of $\mathbf{R}^n$ which are contained in every subspace $L_\alpha$ from the family.

### A.2.3 Linear independence, bases, dimensions

**Definition** A.9 [Linear independence] A collection $X = \{x^1, \ldots, x^N\}$ of vectors from $\mathbf{R}^n$ is called *linearly independent*, if no nontrivial (i.e., with at least one nonzero coefficient) linear combination of vectors from $X$ is zero. That is,

$$\sum_{i=1}^N \lambda_i x^i = 0 \quad\Longrightarrow\quad \lambda_1 = \ldots = \lambda_N = 0.$$

**Example** A.10 (Example of a linearly independent set) The collection of $n$ *standard basis vectors*, (a.k.a. *standard basic orths*) *in* $\mathbf{R}^n$, i.e., the vectors $e_1 := [1;0;\ldots;0]$, $e_2 := [0;1;0;\ldots;0]$, …, $e_n := [0;\ldots;0;1]$, is linearly independent.

**Example** A.11 (Examples of linearly dependent sets) The following sets are all linearly dependent:

1. $X = \{0\}$;
2. $X = \{e_1, e_1\}$;
3. $X = \{e_1, e_2, e_1 + e_2\}$.

**Definition** A.12 [Basis] A collection of vectors $f^1, \ldots, f^m$ is called a *basis* in $\mathbf{R}^n$, if

- the collection is linearly independent, and

• every vector from $\mathbf{R}^n$ is a linear combination of vectors from the collection (i.e., $\mathrm{Lin}\{f^1, \dots, f^m\} = \mathbf{R}^n$).

**Example** A.13 (Basis) The collection of standard basis vectors $e_1, \dots, e_n$ is a basis in $\mathbf{R}^n$.

**Example** A.14 (Non-basis) Every one of the following examples does not form a basis of $\mathbf{R}^n$.

1. The collection $\{e_2, \dots, e_n\}$: this collection is linearly independent, but not every vector $\mathbf{R}^n$ is a linear combination of the vectors from the collection.
2. The collection $\{e_1, e_1, e_2, \dots, e_n\}$: every vector in $\mathbf{R}^n$ is a linear combination of vectors from this collection, but the collection is not linearly independent.

Besides the bases of the entire $\mathbf{R}^n$, we can also speak about the bases of linear subspaces.

**Definition** A.15 [Basis of a linear subspace] A collection $\{f^1, \dots, f^m\}$ of vectors is called a *basis of a linear subspace* $L$, if
• the collection is linearly independent, and
• $L = \mathrm{Lin}\{f^1, \dots, f^m\}$, i.e., all vectors $f^i$ belong to $L$, and every vector from $L$ is a linear combination of the vectors $f^1, \dots, f^m$.

In order to avoid trivial remarks, we follow the standard convention:

> *An empty set of vectors is linearly independent, and an empty linear combination of vectors* $\sum\limits_{i \in \varnothing} \lambda_i x^i$ *has a value, specifically, equals to zero.*

With this convention, the trivial linear subspace $L = \{0\}$ also has a basis, specifically, an empty set of vectors. This convention also is fully compatible with our convention $\mathrm{Lin}(\varnothing) = \{0\}$.

**Theorem** A.16 Let $L$ be a linear subspace of $\mathbf{R}^n$. Then, $L$ admits a (finite) basis, and all bases of $L$ are composed of the same number of vectors.

**Definition** A.17 [Dimension of a linear subspace] Given a linear subspace $L$, the number of elements in its basis is called the *dimension* of $L$ [notation: $\dim(L)$].

We have seen that $\mathbf{R}^n$ admits a basis composed of $n$ elements, i.e., the standard basis vectors $e_i$. From Theorem A.16, it follows that *every* basis of $\mathbf{R}^n$ contains exactly $n$ vectors, and the dimension of $\mathbf{R}^n$ is $n$.

**Theorem** A.18 The larger is a linear subspace of $\mathbf{R}^n$, the larger is its dimension: Suppose $L \subseteq L'$ are linear subspaces of $\mathbf{R}^n$. Then, $\dim(L) \leq \dim(L')$. Moreover, $\dim(L) = \dim(L')$ holds if and only if $L = L'$.

We have seen that $\dim(\mathbf{R}^n) = n$; according to the above convention, the trivial linear subspace $\{0\}$ of $\mathbf{R}^n$ admits an empty basis, so that its dimension is 0. Since every linear subspace $L$ of $\mathbf{R}^n$ satisfies $\{0\} \subseteq L \subseteq \mathbf{R}^n$, it follows from Theorem A.18 that the dimension of a linear subspace in $\mathbf{R}^n$ is an integer between 0 and $n$.

**Theorem** A.19 Let $L$ be a linear subspace in $\mathbf{R}^n$. Then,

(i) Every linearly independent subset of vectors from $L$ can be extended to a basis for $L$;
(ii) From every spanning subset $X$ for $L$, i.e., a set $X$ such that $\mathrm{Lin}(X) = L$, one can extract a basis for $L$.

It follows from Theorem A.19 that

- every linearly independent subset of $L$ contains at most $\dim(L)$ vectors, and if it contains exactly $\dim(L)$ vectors, it is a basis of $L$;
- every spanning set for $L$ contains at least $\dim(L)$ vectors, and if it contains exactly $\dim(L)$ vectors, it is a basis of $L$.

**Theorem** A.20 Let $L$ be a linear subspace in $\mathbf{R}^n$. Suppose $L$ is nontrivial (i.e., $L \neq \{0\}$) and $\{f^1, \ldots, f^m\}$ is a basis of $L$. Then, every vector $x \in L$ admits a *unique representation*

$$x = \sum_{i=1}^{m} \lambda_i(x) f^i$$

as a linear combination of vectors from the basis. Moreover, the mapping

$$x \mapsto (\lambda_1(x), \ldots, \lambda_m(x)) : L \to \mathbf{R}^m$$

is a one-to-one mapping of $L$ onto $\mathbf{R}^m$ which is linear, i.e., for every $i = 1, \ldots, m$ one has

$$\begin{array}{rcl} \lambda_i(x+y) & = & \lambda_i(x) + \lambda_i(y) \quad \forall (x, y \in L); \\ \lambda_i(\nu x) & = & \nu \lambda_i(x) \quad \forall (x \in L, \nu \in \mathbf{R}). \end{array} \tag{A.2}$$

**Definition** A.21 [Coordinates] Given a linear subspace $L$, a basis $f^1, \dots, f^m$ of $L$, and a point $x \in L$, the reals $\lambda_i(x)$, $i = 1, \dots, m$, in the unique representation of $x$ given by $x = \sum\limits_{i=1}^{m} \lambda_i(x) f^i$ are called the *coordinates* of $x \in L$ in the basis $f^1, \dots, f^m$.

For example, the coordinates of a vector $x \in \mathbf{R}^n$ in the *standard basis* $e_1, \dots, e_n$ of $\mathbf{R}^n$ – the one composed of the standard basis vectors – are exactly the entries of $x$.

**Theorem** A.22 [Dimension formula] Let $L_1, L_2$ be linear subspaces of $\mathbf{R}^n$. Then,

$$\dim(L_1 \cap L_2) + \dim(L_1 + L_2) = \dim(L_1) + \dim(L_2).$$

### A.2.4 Linear mappings and matrices

A function $\mathcal{A}(x)$, also called as *mapping*, defined on $\mathbf{R}^n$ and taking values in $\mathbf{R}^m$ is called *linear*, if it preserves linear operations:

$$\begin{array}{l} \mathcal{A}(x + y) = \mathcal{A}(x) + \mathcal{A}(y) \quad \forall (x, y \in \mathbf{R}^n), \text{ and} \\ \mathcal{A}(\lambda x) = \lambda \mathcal{A}(x) \quad \forall (x \in \mathbf{R}^n, \lambda \in \mathbf{R}). \end{array}$$

It is immediately seen that a linear mapping from $\mathbf{R}^n$ to $\mathbf{R}^m$ can be represented as a multiplication by an $m \times n$ matrix. That is, given a linear mapping $\mathcal{A} : \mathbf{R}^n \to \mathbf{R}^m$, there exists a matrix $A \in \mathbf{R}^{m \times n}$ such that

$$\mathcal{A}(x) = Ax.$$

Moreover, this matrix $A$ is uniquely defined by the mapping: the columns $A_j$ of $A$ are just the images of the standard basis vectors $e_j$ under the mapping $\mathcal{A}$, i.e.,

$$A_j = \mathcal{A}(e_j).$$

Linear mappings from $\mathbf{R}^n$ into $\mathbf{R}^m$ can be added to each other, i.e.,

$$(\mathcal{A} + \mathcal{B})(x) = \mathcal{A}(x) + \mathcal{B}(x),$$

and multiplied by reals, i.e.,

$$(\lambda \mathcal{A})(x) = \lambda \mathcal{A}(x) \ \forall \lambda \in \mathbf{R}.$$

Moreover, the results of these operations again are linear mappings from $\mathbf{R}^n$ to $\mathbf{R}^m$. The addition of linear mappings and multiplication of these mappings by reals correspond to the same operations with the matrices representing the mappings: adding/multiplying by reals mappings, we add, respectively, multiply by reals the corresponding matrices.

Given two linear mappings $\mathcal{A}(x) : \mathbf{R}^n \to \mathbf{R}^m$ and $\mathcal{B}(y) : \mathbf{R}^m \to \mathbf{R}^k$, we can build their superposition

$$\mathcal{C}(x) \equiv \mathcal{B}(\mathcal{A}(x)) : \mathbf{R}^n \to \mathbf{R}^k,$$

which is again a linear mapping, now from $\mathbf{R}^n$ to $\mathbf{R}^k$. In the language of matrices representing the mappings, the superposition corresponds to matrix multiplication: the $k \times n$ matrix $C$ representing the mapping $\mathcal{C}$ is the product of the matrices representing $\mathcal{A}$ and $\mathcal{B}$:

$$\mathcal{A}(x) = Ax, \quad \mathcal{B}(y) = By \quad \Longrightarrow \quad \mathcal{C}(x) \equiv \mathcal{B}(\mathcal{A}(x)) = B \cdot (Ax) = (\underbrace{BA}_{=C})x.$$

### A.2.5 Determinant and rank

Let us recall some basic facts about ranks and determinants of matrices.

### A.2.6 Determinant

Let $A = [a_{ij}]_{\substack{1\leq i\leq n,\\ 1\leq j\leq n}}$ be a square matrix. A *diagonal* of the matrix $A$ is the collection of $n$ cells with indices $(1, j_1), (2, j_2), \ldots, (n, j_n)$, where $j_1, j_2, \ldots, j_n$ are distinct from each other, so that the mapping $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ given by $\sigma(i) = j_i$, $1 \leq i \leq n$, is a *permutation*, i.e., a one-to-one mapping of the set $\{1, \ldots, n\}$ onto itself. There are $n!$ different permutations of $\{1, \ldots, n\}$, and these permutations form the group $\Sigma_n$ with the group operation –the product– $(\sigma^1\sigma^2)(i) = \sigma^1(\sigma^2(i))$, $1 \leq i \leq n$. Any permutation $\sigma \in \Sigma_n$ can be assigned a *sign* $\mathrm{sign}(\sigma) \in \{\pm 1\}$ in such a way that the sign of the identity permutation $\sigma(i) \equiv i$ is $+1$, the sign of the product of two permutations is the product of their signs: $\mathrm{sign}(\sigma^1\sigma^2) = \mathrm{sign}(\sigma^1)\mathrm{sign}(\sigma^2)$ for all $\sigma^1, \sigma^2$, and the sign of a *transposition* –permutation with swaps two distinct indices and keeps all other indices intact– is $(-1)$; these properties specify the sign of a permutation in a unique fashion.

Then, for any $n \times n$ real or complex matrix $A = [a_{ij}]_{1\leq i,j\leq n}$, the quantity

$$\mathrm{Det}(A) := \sum_{\sigma \in \Sigma_n} \mathrm{sign}(\sigma) \prod_{i=1}^{n} a_{i\sigma(i)}$$

is called the *determinant* of $A$. We have the following main properties of the determinant:

1. $\mathrm{Det}(A) = \mathrm{Det}(A^\top)$.
2. [polylinearity] $\mathrm{Det}(A)$ is linear in rows of $A$: when all rows but one are fixed, $\mathrm{Det}(A)$ is a linear function of the remaining row. An analogous property holds for the columns as well.
3. [antisymmetry] When swapping rows with two distinct indices, the determinant is multiplied by $(-1)$. An analogous property holds for the columns as well.
4. $\mathrm{Det}(I_n) = 1$.

   *Note:* The last three properties uniquely define $\mathrm{Det}(\cdot)$.
5. [multiplicativity] For two $n \times n$ matrices $A, B$ one has $\mathrm{Det}(AB) = \mathrm{Det}(A)\mathrm{Det}(B)$.
6. An $n \times n$ matrix $A$ is *nonsingular*, that is, $AB = I_n$ for properly selected $B$, if and only if $\mathrm{Det}(A) \neq 0$.

7. [Cramer's rule] For a nonsingular $n \times n$ matrix $A$, the linear system $Ax = b$ in variables $x$ has a unique solution, and the entries of this solution are given by

$$x_i := \frac{\mathrm{Det}(A,i,b)}{\mathrm{Det}(A)},\ 1 \leq i \leq n,$$

where $\mathrm{Det}(A,i,b)$ is the determinant of the matrix obtained from $A$ by replacing its $i$-th column with the right hand side vector $b$.

8. [decomposition] Let $A$ be an $n \times n$ matrix with $n > 1$. Then, for every $i \in \{1,\ldots,n\}$ we have

$$\mathrm{Det}(A) = \sum_{j=1}^{n} a_{ij} C^{ij},$$

where $C^{ij}$ is the *algebraic complement* of the $(i,j)$-th entry in $A$, i.e., $C^{ij} := (-1)^{i+j}\mathrm{Det}(A^{ij})$, and here $A^{ij}$ is the $(n-1)\times(n-1)$ matrix obtained from $A$ by eliminating $i$-th row and $j$-th column.

9. An affine mapping $x \mapsto Ax + b : \mathbf{R}^n \to \mathbf{R}^n$ multiplies $n$-dimensional volumes by $|\mathrm{Det}(A)|$. In particular, for a real $n \times n$ matrix $A$ the quantity $|\mathrm{Det}(A)|$ is the $n$-dimensional volume of the parallelotope

$$X = \left\{ x = \sum_j s_j A_j :\ 0 \leq s_j \leq 1,\ \forall j = 1,\ldots,n \right\}$$

spanned by the columns $A_1,\ldots,A_n$ of $A$.

### A.2.7 Rank

Let $A$ be an $m \times n$ matrix. Every matrix is associated with two linear subspaces, namely its image space and kernel. The *image space* of $A$ [notation: $\mathrm{Im}(A)$] is defined as the linear span of columns of $A$, i.e., the subspace of the destination space $\mathbf{R}^m$ formed by the vectors that admit a representation of the form $Ax$ for some $x \in \mathbf{R}^n$. The *kernel* (a.k.a. *nullspace*) of $A$ [notation: $\mathrm{Ker}(A)$] is given by $\mathrm{Ker}(A) := \{x :\ Ax = 0\}$. Note that $\mathrm{Ker}(A)$ is the orthogonal complement of $\mathrm{Im}(A^\top)$.

Let $R_p := \{i_1 < i_2 < \ldots < i_p\}$ be a collection of $p \geq 1$ distinct from each other indices of rows of $A$, and $C_q := \{j_1 < j_2 < \ldots < j_q\}$ be a collection of $q \geq 1$ distinct from each other indices of columns of $A$. The matrix $B \in \mathbf{R}^{p\times q}$ with entries $B_{k\ell} = A_{i_k,j_\ell}$, $1 \leq k \leq p, 1 \leq \ell \leq q$ is called the submatrix of $A$ with row indices $R_p$ and column indices $C_q$; this is precisely what we get from the matrix $A$ in the intersection of rows with indices from $R_p$ and columns with indices from $C_q$.

The *rank* of $A$ (which is denoted by $\mathrm{rank}(A)$) is, by definition, the largest of the row, or, which is the same, the column sizes of *nonsingular* square submatrices of $A$. When no such submatrix exist, that is, when $A$ is the zero matrix, the rank of $A$ by definition is 0.

The main properties of rank are as follows:

1. $\mathrm{rank}(A)$ is the dimension of the image space $\mathrm{Im}\,A$, i.e., $\mathrm{rank}(A) = \dim(\mathrm{Im}\,A)$. Equivalently, $\mathrm{rank}(A)$ is the maximum of cardinalities of linearly independent collections of columns of $A$. Moreover, a collection of columns of $A$ with $\mathrm{rank}(A)$ distinct indices is linearly independent if and only if its intersection with the collection of $\mathrm{rank}(A)$ properly selected rows is a nonsingular submatrix of $A$.
2. $\mathrm{rank}(A) = \mathrm{rank}(A^\top)$, so that $\mathrm{rank}(A)$ is the maximum of cardinalities of linearly independent collections of rows of $A$. Thus, $\mathrm{rank}(A)$ is the codimension of the kernel of $A$. That is, for a given $A \in \mathbf{R}^{m\times n}$
$$\dim(\mathrm{Ker}(A)) = n - \mathrm{rank}(A).$$
A collection of rows of $A$ with $\mathrm{rank}(A)$ distinct indices is linearly independent if and only if its intersection with the collection of $\mathrm{rank}(A)$ properly selected columns is a nonsingular submatrix of $A$.
3. Whenever the product $AB$ of two matrices makes sense, we have
$$\mathrm{rank}(AB) \le \min\left\{\mathrm{rank}(A), \mathrm{rank}(B)\right\}.$$
For two matrices $A$, $B$ of the same size, we have
$$|\,\mathrm{rank}(A) - \mathrm{rank}(B)| \le \mathrm{rank}(A+B) \le \mathrm{rank}(A) + \mathrm{rank}(B).$$
4. An $n \times n$ matrix $B$ is nonsingular if and only if $\mathrm{rank}(B) = n$.

## A.3 Space $\mathbf{R}^n$: Euclidean structure

So far, we were interested solely in the *algebraic structure* of $\mathbf{R}^n$, or, equivalently, in the properties of the *linear* operations (addition of vectors and multiplication of vectors by scalars) the space is endowed with. Now let us consider another structure on $\mathbf{R}^n$ –the *standard Euclidean structure*– which allows to speak about distances, angles, convergence, etc., and thus makes the space $\mathbf{R}^n$ a much richer mathematical entity.

### A.3.1 Euclidean structure

The standard Euclidean structure on $\mathbf{R}^n$ is given by the *standard inner product* – an operation which takes on input two vectors $x, y$ and produces from them a real number, specifically, the real number given by
$$\langle x, y\rangle := x^\top y = \sum_{i=1}^n x_i y_i.$$
The basic properties of the inner product are as follows:

1. [bi-linearity]: The real-valued function $\langle x, y\rangle$ of two vector arguments $x, y \in \mathbf{R}^n$ is linear with respect to every one of the arguments when the other argument is being fixed:
$$\begin{aligned}\langle \lambda u + \mu v, y\rangle &= \lambda\langle u, y\rangle + \mu\langle v, y\rangle, \quad \forall (u, v, y \in \mathbf{R}^n,\ \lambda, \mu \in \mathbf{R}),\\ \langle x, \lambda u + \mu v\rangle &= \lambda\langle x, u\rangle + \mu\langle x, v\rangle, \quad \forall (x, u, v \in \mathbf{R}^n,\ \lambda, \mu \in \mathbf{R}).\end{aligned}$$

2. [symmetry]: The function $\langle x, y\rangle$ is symmetric:

$$\langle x, y\rangle = \langle y, x\rangle, \quad \forall (x, y \in \mathbf{R}^n).$$

3. [positive definiteness]: The quantity $\langle x, x\rangle$ is always nonnegative, and it is zero if and only if $x$ is zero.

**Remark** A.23 The outlined three properties (i.e., bi-linearity, symmetry and positive definiteness) form a definition of an *Euclidean inner product.* In fact, there are infinitely many different ways to satisfy these properties. In other words, there are infinitely many different Euclidean inner products on $\mathbf{R}^n$. The standard inner product $\langle x, y\rangle = x^\top y$ is just a particular case of this general notion. Although in this book we normally work with the standard inner product, the reader should remember that the facts we are about to recall are valid for all Euclidean inner products, and not only for the standard one.

The notion of an inner product underlies a number of purely algebraic constructions, in particular, those of *inner product representation of linear forms* and of *orthogonal complement.*

### A.3.2 Inner product representation of linear forms on $\mathbf{R}^n$

A *linear form* on $\mathbf{R}^n$ is a real-valued function $f(x)$ on $\mathbf{R}^n$ which is *additive*, i.e., $f(x+y) = f(x) + f(y)$ for all $x, y \in \mathbf{R}^n$, and *homogeneous*, i.e., $f(\lambda x) = \lambda f(x)$ for all $x \in \mathbf{R}^n$ and $\lambda \in \mathbf{R}$.

**Example** A.24 (Example of a linear function) The function $f(x) := \sum\limits_{j=1}^{n} j x_j$ is linear.

**Example** A.25 (Examples of non-linear functions) Each one of the following functions is not linear:

1. $f(x) := x_1 + 1$,
2. $f(x) := x_1^2 - x_2^2$,
3. $f(x) := \sin(x_1)$.

When we add two linear forms or when we multiply a linear form by a real number, we again get a linear form (scientifically speaking: "linear forms on $\mathbf{R}^n$ form a linear space"). *Euclidean structure allows us to identify linear forms on* $\mathbf{R}^n$ *with vectors from* $\mathbf{R}^n$ as follows.

**Theorem** A.26 Let $\langle \cdot, \cdot \rangle$ be a Euclidean inner product on $\mathbf{R}^n$.

(i) Let $f(x)$ be a linear form on $\mathbf{R}^n$. Then, there exists a uniquely defined vector $f \in \mathbf{R}^n$ such that the linear form $f(x)$ is just the inner product with $f$ and $x$:

$$f(x) = \langle f, x\rangle, \quad \forall x \in \mathbf{R}^n.$$

(ii) Vice versa, every vector $f \in \mathbf{R}^n$ defines, via the formula

$$f(x) := \langle f, x \rangle,$$

a linear form on $\mathbf{R}^n$.

(iii) The above one-to-one correspondence between the linear forms and vectors on $\mathbf{R}^n$ is linear: adding linear forms (or multiplying a linear form by a real number), we add (respectively, multiply by the real number) the vector(s) representing the form(s).

### A.3.3 Orthogonal complement

The Euclidean structure allows us to associate with a linear subspace $L \subseteq \mathbf{R}^n$ another linear subspace, namely the *orthogonal complement* (or the *annulator*) of $L$ [notation: $L^\perp$]. By definition, $L^\perp$ consists of all vectors which are orthogonal to every vector from $L$. That is,

$$L^\perp := \{f \in \mathbf{R}^n : \langle f, x \rangle = 0, \quad \forall x \in L\}.$$

**Theorem** A.27 (i) Whenever $L$ is a linear subspace of $\mathbf{R}^n$, one has

$$(L^\perp)^\perp = L.$$

(ii) The larger is a linear subspace $L$, the smaller is its orthogonal complement: if $L_1 \subset L_2$ are linear subspaces of $\mathbf{R}^n$, then $L_1^\perp \supset L_2^\perp$.

(iii) The intersection of a subspace and its orthogonal complement is trivial, and the sum of these subspaces is the entire $\mathbf{R}^n$:

$$L \cap L^\perp = \{0\}, \qquad L + L^\perp = \mathbf{R}^n.$$

**Remark** A.28 From Theorem A.27.iii and the Dimension formula (Theorem A.22) it follows that for every subspace $L$ in $\mathbf{R}^n$ we have

$$\dim(L) + \dim(L^\perp) = n.$$

Moreover, every vector $x \in \mathbf{R}^n$ admits a unique decomposition as a sum of two vectors

$$x = x_L + x_{L^\perp},$$

where $x_L$ belongs to $L$ and $x_{L^\perp}$ belongs to $L^\perp$. This decomposition is called the *orthogonal decomposition* of $x$ *taken with respect to* $L, L^\perp$. In this decomposition, $x_L$ is called the *orthogonal projection* of $x$ onto $L$, and $x_{L^\perp}$ is called the orthogonal projection of $x$ onto the orthogonal complement of $L$. Both projections depend on $x$ linearly. That is, we have

$$(x+y)_L = x_L + y_L, \ \forall x, y \in \mathbf{R}^n, \quad \text{and} \quad (\lambda x)_L = \lambda x_L, \ \forall x \in \mathbf{R}^n, \forall \lambda \in \mathbf{R}.$$

The mapping $x \mapsto x_L$ is called the *orthogonal projector* onto $L$.

### A.3.4 Orthonormal bases

A collection of vectors $f^1, \dots, f^m$ is called *orthonormal* w.r.t. Euclidean inner product $\langle \cdot, \cdot \rangle$ if each vector from the collection is orthogonal to every other vector from it, i.e.,

$$i \neq j \quad \Longrightarrow \quad \langle f^i, f^j \rangle = 0,$$

and the inner product of every vector $f^i$ with itself is unit, i.e.,

$$\langle f^i, f^i \rangle = 1, \; i = 1, \dots, m.$$

**Theorem** A.29 An orthonormal collection of vectors $f^1, \dots, f^m$ is always linearly independent and is thus a basis of its linear span $L = \mathrm{Lin}(f^1, \dots, f^m)$ (such a basis in a linear subspace is called *orthonormal*). The coordinates of a vector $x \in L$ w.r.t. an orthonormal basis $f^1, \dots, f^m$ of $L$ are given by explicit formulas:

$$x = \sum_{i=1}^{m} \lambda_i(x) f^i \iff \lambda_i(x) = \langle x, f^i \rangle \; \forall i = 1, \dots, m.$$

**Proof.** Consider any $x \in \mathbf{R}^n$ and any $i = 1, \dots, m$. Starting from the representation $x = \sum\limits_{j=1}^{m} \lambda_j(x) f^j$, and by taking inner product of both sides of this equality with $f^i$, we get

$$\begin{aligned} \langle x, f_i \rangle &= \langle \textstyle\sum_{j=1}^m \lambda_j(x) f^j, f^i \rangle \\ &= \sum_{j=1}^{m} \lambda_j(x) \langle f^j, f^i \rangle && \text{[by bilinearity of inner product]} \\ &= \lambda_i(x). && \text{[by orthonormality of } \{f^i\}] \end{aligned}$$

Plugging $x = 0$ in this representation results in the corresponding coefficients $\lambda_i(0) = 0$ for all $i$, i.e., all the coefficients are zero. Hence, an orthonormal system is linearly independent. □

**Example** A.30 (An orthonormal basis in $\mathbf{R}^n$) The standard basis $\{e_1, \dots, e_n\}$ is orthonormal *with respect to the standard inner product* $\langle x, y \rangle = x^\top y$ on $\mathbf{R}^n$ (but is not orthonormal w.r.t. other Euclidean inner products on $\mathbf{R}^n$).

**Theorem** A.31 (i) If $f^1, \dots, f^m$ is an orthonormal basis in a linear subspace $L$, then the inner product of two vectors $x, y \in L$ in the coordinates $\lambda_i(\cdot)$ w.r.t. this basis is given by the standard formula

$$\langle x, y \rangle = \sum_{i=1}^{m} \lambda_i(x) \lambda_i(y).$$

(ii) Every linear subspace $L$ of $\mathbf{R}^n$ admits an orthonormal basis. Moreover, every orthonormal system $f^1, \ldots, f^m$ of vectors from $L$ can be extended to an orthonormal basis in $L$.

**Proof.** To prove (i), consider any two vectors $x, y \in L$ along with their basis representations, i.e., $x = \sum\limits_{i=1}^{m} \lambda_i(x) f^i$ and $y = \sum\limits_{i=1}^{m} \lambda_i(y) f^i$. Then,

$$\begin{aligned}\langle x, y\rangle &= \langle \textstyle\sum_{i=1}^{m} \lambda_i(x) f^i, \sum_{i=1}^{m} \lambda_i(y) f^i \rangle \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i(x)\lambda_j(y)\langle f^i, f^j\rangle && \text{[by bilinearity of the inner product]} \\ &= \sum_{i=1}^{m} \lambda_i(x)\lambda_i(y). && \text{[by orthonormality of the vectors } \{f^i\}]\end{aligned}$$

The proof of (ii) is given by the *Gram-Schmidt orthogonalization process* (which is important by its own right) as follows. We start with an arbitrary basis $h^1, \ldots, h^m$ in $L$ and step by step convert it into an orthonormal basis $f^1, \ldots, f^m$. At the beginning of a step $t$ of the construction, we already have an orthonormal collection $f^1, \ldots, f^{t-1}$ such that $\mathrm{Lin}\{f^1, \ldots, f^{t-1}\} = \mathrm{Lin}\{h^1, \ldots, h^{t-1}\}$. For any $t = 1, \ldots, m$, at the step $t$, we proceed as follows:

1. Build the vector

$$g^t := h^t - \sum_{j=1}^{t-1} \langle h^t, f^j\rangle f^j.$$

It is easily seen (check it!) that

- we have

$$\mathrm{Lin}\{f^1, \ldots, f^{t-1}, g^t\} = \mathrm{Lin}\{h^1, \ldots, h^t\}; \tag{A.3}$$

- moreover, $g^t \neq 0$ (derive this fact from (A.3) and the linear independence of the collection $h^1, \ldots, h^m$);
- also, $g^t$ is orthogonal to $f^1, \ldots, f^{t-1}$.

2. Since $g^t \neq 0$, the quantity $\langle g^t, g^t\rangle$ is positive (by the positive definiteness of the inner product), so that the vector

$$f^t := \frac{1}{\sqrt{\langle g^t, g^t\rangle}} g^t$$

is well defined. It is immediately seen (check it!) that the collection $f^1, \ldots, f^t$ is orthonormal and

$$\mathrm{Lin}\{f^1, \ldots, f^t\} = \mathrm{Lin}\{f^1, \ldots, f^{t-1}, g^t\} = \mathrm{Lin}\{h^1, \ldots, h^t\},$$

which completes the step $t$ of the orthogonalization process.

After $m$ steps of the orthogonalization process, we end up with an orthonormal system $f^1,\dots,f^m$ of vectors from $L$ such that

$$\mathrm{Lin}\{f^1,\dots,f^m\} = \mathrm{Lin}\{h^1,\dots,h^m\} = L,$$

so that $f^1,\dots,f^m$ is an orthonormal basis in $L$.

The construction can be easily modified (do it!) to extend a given orthonormal system of vectors from $L$ to an orthonormal basis of $L$. □

Theorem A.31 admits an important corollary which states that *All Euclidean spaces of the same dimension are "the same."*

**Corollary** A.32 Suppose $L$ is an $m$-dimensional subspace in a space $\mathbf{R}^n$ equipped with an Euclidean inner product $\langle\cdot,\cdot\rangle$. Then, there exists a one-to-one mapping $x \mapsto A(x)$ of $L$ onto $\mathbf{R}^m$ such that

- the mapping preserves linear operations:

$$\begin{aligned}&A(x+y) = A(x) + A(y), \quad \forall (x, y \in L) \quad \text{and}\\ &A(\lambda x) = \lambda A(x), \quad \forall (x \in L, \lambda \in \mathbf{R});\end{aligned}$$

- the mapping converts the $\langle\cdot,\cdot\rangle$ inner product on $L$ into the standard inner product on $\mathbf{R}^m$:

$$\langle x, y\rangle = (A(x))^\top A(y), \quad \forall x, y \in L.$$

**Proof.** Indeed, by Theorem A.31.ii $L$ admits an orthonormal basis $f^1,\dots,f^m$; using Theorem A.31.i, one can immediately check that the mapping

$$x \mapsto A(x) = [\lambda_1(x);\dots;\lambda_m(x)]$$

which maps $x \in L$ into the $m$-dimensional vector composed of the coordinates of $x$ in the basis $f^1,\dots,f^m$, meets all the requirements. □

**Fact** A.33 (i) Let $L$ be a linear subspace of $\mathbf{R}^n$, and $f^1,\dots,f^m$ be an orthonormal basis in $L$. Then, for every $x \in \mathbf{R}^n$, the *orthoprojection* $x_L$ of $x$ onto $L$ is given by the formula

$$x_L := \sum_{i=1}^m (x^\top f^i) f^i.$$

(ii) Let $L_1, L_2$ be linear subspaces in $\mathbf{R}^n$. Then,

$$(L_1+L_2)^\perp = L_1^\perp \cap L_2^\perp \quad \text{and} \quad (L_1 \cap L_2)^\perp = L_1^\perp + L_2^\perp.$$

## A.4 Affine subspaces in $\mathbf{R}^n$

In this book, many events we consider take place not in the entire $\mathbf{R}^n$, but in its *affine subspaces*. Geometrically, affine subspaces are planes of different dimensions in $\mathbf{R}^n$. Let us become acquainted with these subspaces.

### A.4.1 Affine subspaces and affine hulls

In geometry, a linear subspace $L$ of $\mathbf{R}^n$ is a special plane – the one passing through the origin of the space (i.e., containing the zero vector). To get an arbitrary plane $M$, it suffices to *translate* an appropriate special plane $L$, i.e., add a fixed *shifting vector* $\bar{a}$ to all points from $L$. This geometric intuition leads to the following definition.

**Definition** A.34 [Affine subspace] An *affine subspace* (a.k.a. *affine plane*) in $\mathbf{R}^n$ is a set of the form

$$M = \bar{a} + L = \{\bar{a} + x : \ x \in L\}, \tag{A.4}$$

where $L$ is a linear subspace in $\mathbf{R}^n$ and $\bar{a}$ is a vector from $\mathbf{R}^n$.

According to our convention on arithmetic of sets, we were supposed to write in (A.4) $\{\bar{a}\} + L$ instead of $\bar{a} + L$ as we did not define arithmetic sum of a vector and a set. It is usual to ignore this difference and omit the brackets when writing down singleton sets in similar expressions: we shall write $\bar{a} + L$ instead of $\{\bar{a}\} + L$, $\mathbf{R}d$ instead of $\mathbf{R}\{d\}$, etc.

**Example** A.35 Let $L$ be the linear subspace composed of vectors with first entries equal to zero. Suppose that we shift $L$ by a vector $\bar{a} = [\bar{a}_1; \ldots; \bar{a}_n]$. Then, we obtain the set $M := \bar{a} + L$ of all vectors $x$ with $x_1 = \bar{a}_1$. According to our terminology, this set $M$ is an affine subspace.

An immediate question about the notion of an affine subspace is what "degrees of freedom" we may have in decomposition (A.4). For example, would $M$ uniquely determine the linear subspace $L$ and the shifting vector $\bar{a}$? The next proposition provides an answer to this question.

**Proposition** A.36 Given an affine subspace $M$ in $\mathbf{R}^n$, the linear subspace $L$ in its decomposition (A.4) is uniquely determined by $M$. Specifically, $L$ is the set of all differences of the vectors from $M$, i.e.,

$$L = M - M = \{x - y : \ x, y \in M\}. \tag{A.5}$$

In contrast, the shifting vector $\bar{a}$ is not uniquely defined by $M$ and can be chosen as an arbitrary vector from $M$.

In this book, given an affine subspace $M$, we refer to the linear subspace $L := M - M$ as the subspace *parallel* to the affine subspace $M$.

### A.4.2 Intersections of affine subspaces, affine combinations, and affine hulls

An immediate conclusion of Proposition A.36 is as follows:

**Corollary** A.37 Let $\{M_\alpha\}$ be an arbitrary family of affine subspaces in $\mathbf{R}^n$. Whenever the set $M := \bigcap\limits_{\alpha} M_\alpha$ is nonempty, then $M$ is an affine subspace.

From Corollary A.37 it immediately follows that for every nonempty subset $Y$ of $\mathbf{R}^n$ there exists the smallest affine subspace containing $Y$; this is the intersection of all affine subspaces containing $Y$. This smallest affine subspace containing $Y$ is called the *affine hull* of $Y$ [notation: $\mathrm{Aff}(Y)$].

All this resembles a lot the story about linear spans. In fact, we can further extend this analogy and get an "inner" description of the affine hull $\mathrm{Aff}(Y)$ in terms of elements of $Y$ similar to the one of the linear span (recall that the linear span of $X$ is also characterized as the set of all linear combinations of vectors from $X$).

Given a nonempty set $Y$, let us choose an arbitrary point $y^0 \in Y$, and consider the set

$$X := Y - y^0.$$

All affine subspaces containing $Y$ should also contain $y^0$. Therefore, by Proposition A.36, $\mathrm{Aff}(Y)$ can be represented as $M = y^0 + L$, where $L$ is a linear subspace. It is absolutely evident that an affine subspace $M = y^0 + L$ contains $Y$ if and only if the subspace $L$ contains $X$, and that the larger is $L$, the larger is $M$:

$$L \subset L' \implies M = y^0 + L \subset M' = y^0 + L'.$$

Thus, in order to find the smallest among the *affine subspaces containing* $Y$, it suffices to find the smallest among the *linear subspaces containing* $X$ and then translate the latter space by $y^0$:

$$\mathrm{Aff}(Y) = y^0 + \mathrm{Lin}(X) = y^0 + \mathrm{Lin}(Y - y^0). \tag{A.6}$$

Now, recall that by definition $\mathrm{Lin}(Y - y^0)$ is the set of all linear combinations of vectors from $Y - y^0$, so that a generic element of $\mathrm{Lin}(Y - y^0)$ is

$$x = \sum_{i=1}^{k} \mu_i (y^i - y^0) \qquad \text{[here } k \text{ may depend on } x\text{]}$$

with $y^i \in Y$ and coefficients $\mu_i \in \mathbf{R}$ for $i = 1, \ldots, k$. Then, a generic element of $\mathrm{Aff}(Y)$ is given by

$$y = y^0 + \sum_{i=1}^{k} \mu_i (y^i - y^0) = \sum_{i=0}^{k} \lambda_i y^i,$$

where

$$\lambda_0 := 1 - \sum_i \mu_i, \quad \text{and} \quad \lambda_i := \mu_i,\ i = 1, \ldots, k.$$

Hence, we deduce that a generic element of $\mathrm{Aff}(Y)$ is a linear combination of vectors from $Y$. Note, however, that in this combination the coefficients $\lambda_i$ are not completely arbitrary: their sum is equal to 1. Linear combinations of this type –with the unit sum of coefficients– have a special name; they are called *affine combinations*.

We have seen that every vector from $\mathrm{Aff}(Y)$ is an affine combination of vectors from $Y$. In fact, the reverse is also true, i.e., $\mathrm{Aff}(Y)$ contains all affine combinations of vectors from $Y$. Indeed, if

$$y = \sum_{i=1}^{k} \lambda_i y^i$$

is an affine combination of vectors from $Y$, then, using the equality $\sum_i \lambda_i = 1$, we can write it also as

$$y = y^0 + \sum_{i=1}^{k} \lambda_i (y^i - y^0),$$

where $y^0$ is the "marked" vector we used in our previous reasoning. And any vector $y$ of this form, as we already know, belongs to $\mathrm{Aff}(Y)$. Thus, we arrive at the following result.

**Proposition** A.38 [Structure of affine hull] For any nonempty set $Y$, we have

$$\mathrm{Aff}(Y) = \{\text{the set of all affine combinations of vectors from } Y\}.$$

When $Y$ itself is an affine subspace, it, of course, coincides with its affine hull, and the previous proposition leads to the following consequence.

**Corollary** A.39 An affine subspace $M$ is closed with respect to taking affine combinations of its members, i.e., every combination of this type is a vector from $M$. Vice versa, a nonempty set which is closed with respect to taking affine combinations of its members is an affine subspace.

### A.4.3 Affinely spanning sets, affinely independent sets, affine dimension

Affine subspaces are closely related to linear subspaces, and the basic notions associated with linear subspaces have natural and useful affine analogies. Here, we introduce these notions and discuss their basic properties.

**Affinely spanning sets.** Consider an affine subspace $M = \bar{a} + L$. We say that a subset $Y$ of $M$ is *affinely spanning* for $M$ (we also say that $Y$ *spans* $M$ *affinely*, or that $M$ *is affinely spanned by* $Y$), if $M = \mathrm{Aff}(Y)$, or, due to Proposition A.38 equivalently, if every point of $M$ is an affine combination of points from $Y$. Hence, we arrive at the following immediate consequence of section A.4.2.

**Proposition** A.40 Let $M = \bar{a} + L$ be an affine subspace, let $Y$ be a subset of $M$, and consider any $y^0 \in Y$. Then, the set $Y$ affinely spans $M$, i.e., $M = \mathrm{Aff}(Y)$, if and only if the set

$$X := Y - y^0$$

spans the linear subspace $L$, i.e., $L = \mathrm{Lin}(X)$.

**Affinely independent sets.** A linearly independent set of vectors $x^1, \dots, x^k$ is a set such that no nontrivial linear combination of $x^1, \dots, x^k$ equals to zero (see Definition A.9). An equivalent definition is given by Theorem A.16.iv. That is, vectors $x^1, \dots, x^k$ are linearly independent if for any linear combination

$$x = \sum_{i=1}^{k} \lambda_i x^i$$

the coefficients $\lambda_i$ are *uniquely* determined by the vector $x$. This equivalent form reflects the essence of the matter – what we indeed need, is the uniqueness of the coefficients in expansions. Accordingly, this equivalent form is the prototype for the notion of an affinely independent set: we want to introduce this notion in such a way that the coefficients $\lambda_i$ in an *affine* combination

$$y = \sum_{i=0}^{k} \lambda_i y^i$$

of "affinely independent" set of vectors $y^0, \dots, y^k$ would be uniquely defined by $y$. *Non*-uniqueness would mean that we can write $y$ as an affine combination of the vectors $y^0, \dots, y^k$ using two different sets of coefficients $\lambda_i$ and $\lambda_i'$ such that $\sum_{i=0}^{k} \lambda_i = \sum_{i=0}^{k} \lambda_i' = 1$. That is,

$$y = \sum_{i=0}^{k} \lambda_i y^i = \sum_{i=0}^{k} \lambda_i' y^i.$$

In such a case, we arrive at

$$\sum_{i=0}^{k} (\lambda_i - \lambda_i') y^i = 0,$$

which implies that the vectors $y^0, \dots, y^k$ are linearly dependent. Moreover, there exists a nontrivial combination of these vectors which represents the zero vector and the sum of the coefficients in this representation satisfy $\sum_i (\lambda_i - \lambda_i') = \sum_i \lambda_i - \sum_i \lambda_i' = 1 - 1 = 0$. Our reasoning can be reversed: if there exists a nontrivial linear combination of $y^i$'s with zero sum of coefficients which results in the zero vector, then the coefficients in the representation of any vector as an affine combination of $y^i$'s are not uniquely defined. Thus, in order to get uniqueness we should for sure forbid relations

$$\sum_{i=0}^{k} \mu_i y^i = 0$$

with nontrivial coefficients $\mu_i$ satisfying $\sum_{i=0}^{k} \mu_i = 0$. Thus, this discussion motivates the following definition.

**Definition** A.41 [Affine independence] A collection $y^0, \ldots, y^k$ of vectors in $\mathbf{R}^n$ is called *affinely independent* if no nontrivial linear combination of the vectors with zero sum of coefficients is the zero vector, i.e.,

$$\sum_{i=0}^{k} \lambda_i y^i = 0 \text{ and } \sum_{i=0}^{k} \lambda_i = 0 \quad \Longrightarrow \quad \lambda_0 = \lambda_1 = \ldots = \lambda_k = 0.$$

(To compare against the definition of linear independence, see Definition A.9.)

With this definition of affinely independent set of vectors, we arrive at the following result analogous to Theorem A.20.

**Corollary** A.42 Let $y^0, \ldots, y^k$ be affinely independent. Then, the coefficients $\lambda_i$ in any affine combination

$$y = \sum_{i=0}^{k} \lambda_i y^i \qquad \left[\text{where } \sum_{i=0}^{k} \lambda_i = 1\right]$$

of the vectors $y^0, \ldots, y^k$ are uniquely defined by the value $y$ of the combination.

Verification of affine independence of a collection can be immediately reduced to the verification of linear independence of a closely related collection.

**Proposition** A.43 $k+1$ vectors $y^0, \ldots, y^k$ are affinely independent if and only if the set of $k$ vectors given by $(y^1 - y^0), (y^2 - y^0), \ldots, (y^k - y^0)$ are linearly independent.

Based on this last proposition we deduce for example that the set of vectors $0, e_1, \ldots, e_n$ composed of the origin and the standard basis vectors is affinely independent. Note that this collection is linearly dependent (as every set of vectors containing zero is linearly dependent). The difference between the two notions of independence we deal with is important to keep in mind: linear independence means that no nontrivial linear combination of the vectors can be zero, while affine independence means that no nontrivial linear combination *from certain restricted class of them* (with zero sum of coefficients) can be zero. Therefore, there are more affinely independent sets than the linearly independent ones: a linearly independent set is for sure affinely independent, but not vice versa.

**Affine bases and affine dimension.** Propositions A.38 and A.40 reduce the notions of affine spanning/affinely independent sets to the notions of spanning/linearly independent ones. Combined with Theorem A.16, they result in the following analogies of the latter two statements:

**Proposition** A.44 [Affine dimension] Let $M = \bar{a} + L$ be an affine subspace in $\mathbf{R}^n$. Then, the following two quantities are finite integers which are equal to each other:

(i) the minimal number of elements in the subsets of $M$ which affinely span $M$;

(ii) the maximal number of elements in affinely independent subsets of $M$. The common value of these two integers is exactly equal to $\dim(L) + 1$.

By definition, the *affine dimension* of an affine subspace $M = \bar{a} + L$ is the dimension $\dim(L)$ of $L$, i.e., $\dim(M) := \dim(L)$. Thus, if $M$ is of affine dimension $k$, then the minimal cardinality of sets affinely spanning $M$, same as the maximal cardinality of affinely independent subsets of $M$, is $k + 1$.

**Theorem** A.45 [Affine bases] Let $M = \bar{a} + L$ be an affine subspace in $\mathbf{R}^n$.

**A.** For any $Y \subseteq M$, the following three properties of $Y$ are equivalent:

(i) $Y$ is an affinely independent set which affinely spans $M$;

(ii) $Y$ is affinely independent and contains $\dim(L) + 1$ elements;

(iii) $Y$ affinely spans $M$ and contains $\dim(L) + 1$ elements.

A subset $Y$ of $M$ possessing these preceding equivalent to each other properties is called an *affine basis* of $M$. Affine bases of $M$ are exactly the collections $y^0, \dots, y^{\dim(L)}$ such that $y^0 \in M$ and $(y^1 - y^0), \dots, (y^{\dim(L)} - y^0)$ is a basis of $L$.

**B.** Every affinely independent collection of vectors of $M$ either itself is an affine basis of $M$, or can be extended to such a basis by adding new vectors. In particular, every affine subspace $M$ admits an affine basis.

**C.** If $Y$ affinely spans $M$, then we can always extract from $Y$ an affine basis of $M$.

We already know that the standard basis vectors $e_1, \dots, e_n$ form a basis of the entire space $\mathbf{R}^n$. And what about affine bases in $\mathbf{R}^n$? According to Theorem A.45.**A**, you can choose such an affine basis for an affine subspace $M$ as any collection of vectors $e^0, e^0 + e_1, \dots, e^0 + e_n$, where $e^0$ is an arbitrary vector from $M$.

**Barycentric coordinates.** Let $M$ be an affine subspace, and let $y^0, \dots, y^k$ be an affine basis of $M$. Since the basis, by definition, affinely spans $M$, every vector $y$ from $M$ is an affine combination of the vectors of the basis, i.e.,

$$y = \sum_{i=0}^{k} \lambda_i y^i \qquad \left[\text{where } \sum_{i=0}^{k} \lambda_i = 1\right].$$

Moreover, since the vectors of the affine basis are affinely independent, the coefficients of this combination are uniquely defined by $y$ (Corollary A.42). These coefficients are called *barycentric coordinates* of $y$ with respect to the affine basis in question. In contrast to the usual coordinates with respect to a (linear) basis, the barycentric coordinates could not be quite arbitrary: their sum should be equal to 1.

### A.4.4 Dual description of linear subspaces and affine subspaces

So far, we have introduced the notions of linear subspace and affine subspace and have presented a scheme of generating these entities. For example, to get, a linear subspace, we start from an arbitrary nonempty set $X \subset \mathbf{R}^n$ and add to it all linear combinations of the vectors from $X$. By replacing linear combinations with the affine ones, we get a way to generate affine subspaces.

The just indicated way of generating linear or affine subspaces resembles the approach of a worker building a house: he starts with the base and then adds to it new elements until the house is ready. There is yet another way to generate such subspaces that resembles the approach of an artist creating a sculpture: he takes something large and then deletes extra parts of it. The "artist's way" to represent linear subspaces and affine subspaces is quite instructive as well and we will examine it next.

### A.4.5 Affine subspaces and systems of linear equations

Let $L$ be a linear subspace in $\mathbf{R}^n$. According to Theorem A.27.i it is an orthogonal complement itself, namely, $L$ is the orthogonal complement to the linear subspace $L^\perp$. Now let the vectors $a_1, \dots, a_m$ be a finite spanning set in $L^\perp$. A vector $x$ which is orthogonal to $a_1, \dots, a_m$ is orthogonal to the entire $L^\perp$ (since every vector from $L^\perp$ is a linear combination of $a_1, \dots, a_m$ and the inner product is bilinear). Of course, vice versa is also true: a vector orthogonal to the entire $L^\perp$ is orthogonal to every vector in $a_1, \dots, a_m$. Therefore, we arrive at

$$L = (L^\perp)^\perp = \left\{x \in \mathbf{R}^n : a_i^\top x = 0,\ i = 1, \dots, m\right\}. \tag{A.7}$$

Thus, we get the following very important, although simple, result.

**Proposition** A.46 ["Outer" description of a linear subspace] Every linear subspace $L$ in $\mathbf{R}^n$ is a set of all solutions to a homogeneous system of linear equations

$$a_i^\top x = 0, \quad i = 1, \dots, m, \tag{A.8}$$

given by properly chosen $m$ and vectors $a_1, \dots, a_m \in \mathbf{R}^n$.

Recall from Example A.3 that the solution set to a homogeneous system of linear equations with $n$ variables is always a linear subspace in $\mathbf{R}^n$. Thus, Proposition A.46 is indeed an "if and only if" statement.

From Proposition A.46 and the facts about the dimension of linear subspaces we can easily derive several important consequences:

- The systems of equations (A.8) which define a given linear subspace $L$ are exactly the systems given by the vectors $a_1, \dots, a_m$ which span $L^\perp$ [1].

[1] The reasoning which led us to Proposition A.46 states that $a_1, \dots, a_m$ span $L^\perp$ implies (A.8) defines $L$. Here, we claim that the reverse is also true.

- If $m$ is the smallest possible number of equations in (A.8), then $m$ is also the dimension of $L^\perp$. That is, by Remark A.28, we have $\mathrm{codim}(L) := n - \dim(L)$.[2]

Now, an affine subspace $M$ is, by definition, a translation of a linear subspace, i.e., $M = \bar{a} + L$. Recall that the vectors $x$ from $L$ are exactly the solutions of certain *homogeneous* system of linear equations

$$a_i^\top x = 0, \quad i = 1, \ldots, m.$$

It is absolutely clear that adding to these vectors a fixed vector $\bar{a}$, we get exactly the set of solutions to the *inhomogeneous* feasible system of linear equations

$$a_i^\top x = b_i, \quad i = 1, \ldots, m,$$

where $b_i := a_i^\top \bar{a}$ for all $i$. The reverse is also true, i.e., the set of solutions to a *feasible* system of linear equations

$$a_i^\top x = b_i, \quad i = 1, \ldots, m,$$

with $n$ variables is the sum of a particular solution to the system and the solution set to the underlying homogeneous system of linear equations (the latter set, as we already know, is a linear subspace in $\mathbf{R}^n$). That is, the set of solutions to a *feasible* system of linear equations is an affine subspace. Thus, we arrive at the following result.

**Proposition** A.47 [“Outer” description of an affine subspace]
Every affine subspace $M = a + L$ in $\mathbf{R}^n$ is a set of all solutions to a *feasible* system of linear equations

$$a_i^\top x = b_i, \quad i = 1, \ldots, m, \tag{A.9}$$

given by properly chosen $m$ and vectors $a_1, \ldots, a_m$.

Vice versa, the set of all solutions to a feasible system of linear equations with $n$ variables is an affine subspace in $\mathbf{R}^n$.

The linear subspace $L$ associated with $M$ is exactly the set of solutions of the homogeneous (with the right hand side set to 0) version of system (A.9).

We see, in particular, that an affine subspace is always closed.

**Remark** A.48 The “outer” description of a linear or affine subspace (the artist’s one) is in many cases much more useful than the “inner” description via linear/affine combinations (the worker’s one). For example, using the outer description, it is very easy to check whether or not a given vector belongs to a given linear (or affine) subspace. In contrast, this task is not that easy with the inner

[2] Note that this statement holds true also in the extreme case when $L = \mathbf{R}^n$ (i.e., when $\mathrm{codim}(L) = 0$) due to the fact that the solution set of an *empty* set of equations or inequalities in variables $x \in \mathbf{R}^n$ is the entire space; indeed, were it not the case, there would exist a vector in $\mathbf{R}^n$ violating one or more equalities/inequalities composing the system, which clearly is not the case when the system is empty.

description[3]. In fact both descriptions are "complementary" to each other and work perfectly well in parallel: what is difficult to see with one of them, is clear with another. The idea of using "inner" and "outer" descriptions of the entities we meet with (e.g., linear subspaces, affine subspaces, convex sets, optimization problems) – the general idea of *duality* – is, in our humble opinion, the main driving force of Convex Analysis and Optimization, and so not surprisingly in this book we will all the time meet with different implementations of this fundamental idea.

### A.4.6 Structure of the simplest affine subspaces

Here, we mainly introduce some terminology. According to their dimension, affine subspaces in $\mathbf{R}^n$ are named as follows:

- Subspaces of dimension 0 are translations of the only 0-dimensional linear subspace $\{0\}$, i.e., they are singleton sets – vectors from $\mathbf{R}^n$. These subspaces are called *points*; a point is a solution to a system of $n$ linear equations in $n$ unknowns with a nonsingular matrix.
- Subspaces of dimension 1 are *lines*. These subspaces are translations of one-dimensional linear subspaces of $\mathbf{R}^n$. A one-dimensional linear subspace has a single-element basis given by a nonzero vector $d$ and is simply the set of all possible multiples of this vector. Consequently, a line $l$ is a set of the form

  $$l := \{\bar{a} + td : t \in \mathbf{R}\}$$

  given by a pair of vectors $\bar{a}$ (the origin of the line) and $d$ (the direction of the line), $d \neq 0$. Naturally, this is the inner description of the line $l$. Note that in this description the origin of the line and its direction are not uniquely defined by the line; you can choose as origin any point on the line and also you can multiply a particular direction by any nonzero real number.

  In the barycentric coordinates a line $l$ is described as follows:

  $$l = \{\lambda_0 y^0 + \lambda_1 y^1 : \lambda_0 + \lambda_1 = 1\} = \{\lambda y^0 + (1-\lambda) y^1 : \lambda \in \mathbf{R}\},$$

  where $y^0, y^1$ is an affine basis of $l$ (such a basis can be chosen as any pair of distinct points on the line).

  The outer description of a line is as follows: it is the set of solutions to a system of $n-1$ linearly independent linear equations in $n$ variables (unknowns).
- Subspaces of dimension $k$ where $2 \leq k < n-1$ have no special names; sometimes they are called affine planes of dimension $k$.
- Affine subspaces of dimension $n-1$, due to the important role they play in Convex

[3] In principle it is not difficult to certify that a given point belongs to, say, a linear subspace given as the linear span of some set – it suffices to point out a representation of the point as a linear combination of vectors from the set. But how could you certify that the point does *not* belong to the subspace?

Analysis, have a special name; they are called *hyperplanes*. The outer description of a hyperplane is that a hyperplane is the solution set of a *single* linear equation

$$a^\top x = b$$

with nontrivial left hand side ($a \neq 0$). In other words, a hyperplane is the level set $a(x) =$ const of a nonconstant linear form $a(x) = a^\top x$.

- The "largest possible" affine subspace –the one of dimension $n$– is unique and it is the entire $\mathbf{R}^n$. This subspace is given by an empty system of linear equations.

## A.5 Exercises

Exercise 1 1. Mark in the list below those subsets of $\mathbf{R}^n$ which are linear subspaces. For the ones that are linear subspaces, find out their dimensions and point out bases. For the ones that are not linear subspaces provide counterexamples.

1. $\mathbf{R}^n$
2. $\{0\}$
3. $\varnothing$
4. $\left\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n ix_i = 0\right\}$
5. $\left\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n ix_i^2 = 0\right\}$
6. $\left\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n ix_i = 1\right\}$
7. $\left\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n ix_i^2 = 1\right\}$

2. Suppose that we know $L$ is a subspace of $\mathbf{R}^n$ with exactly one basis. What is $L$?

Exercise 2 Consider the sets given in Exercise 1 and identify the ones that are affine subspaces. For the ones that are affine subspaces, find their affine dimensions and point out their linear subspaces parallel to them. For the ones that are not affine subspaces, provide counterexamples.

Exercise 3 1. What is the orthogonal complement (w.r.t. the standard inner product) of the subspace $\left\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n x_i = 0\right\}$ in $\mathbf{R}^n$?

2. Find an orthonormal basis (w.r.t. the standard inner product) in the linear subspace $\{x \in \mathbf{R}^n : x_1 = 0\}$ of $\mathbf{R}^n$

Exercise 4 Suppose $a \in \mathbf{R}^n$ where $a_i > 0$ for all $i = 1, \dots, n$, and consider the affine subspace

$$M = \left\{x \in \mathbf{R}^n : \sum_{i=1}^n a_i x_i = 1\right\}$$

Point out the linear subspace parallel to $M$ and find an affine basis in $M$.

Exercise 5 Let $\varnothing \neq C \subseteq \mathbf{R}^n$ and $x \in \mathbf{R}^n$ be given.

1. Is it always true that $\mathrm{Aff}(C - \{x\}) = \mathrm{Aff}(C) - \{x\}$?
2. Is it always true that $\mathrm{Lin}(C - \{x\}) = \mathrm{Aff}(C) - \{x\}$?
3. Do your answers to the previous questions change if you further assume $x \in \mathrm{Aff}(C)$?

Exercise 6 Suppose that we are given $n$ sets $E_1, E_2, \dots, E_n$ in $\mathbf{R}^{100}$ that are distinct from each other and they satisfy

$$E_1 \subset E_2 \subset \ldots \subset E_n.$$

How large can $n$ be, if

1. every one of $E_i$ is a linear subspace?
2. every one of $E_i$ is an affine subspace?
3. every one of $E_i$ is a convex set?

Exercise 7 Prove that the *Triangle inequality in Euclidean norm*, i.e., $\|x+y\|_2 \leq \|x\|_2 + \|y\|_2$, holds true as *equality* if and only if $x$ and $y$ are nonnegative multiples of some vector (which always can be taken to be $x+y$).

## A.6 Proofs of Facts

**Fact A.33** (i) Let $L$ be a linear subspace of $\mathbf{R}^n$, and $f^1, \ldots, f^m$ be an orthonormal basis in $L$. Then, for every $x \in \mathbf{R}^n$, the *orthoprojection* $x_L$ of $x$ onto $L$ is given by the formula

$$x_L := \sum_{i=1}^{m} (x^\top f^i) f^i.$$

(ii) Let $L_1, L_2$ be linear subspaces in $\mathbf{R}^n$. Then,

$$(L_1 + L_2)^\perp = L_1^\perp \cap L_2^\perp \quad \text{and} \quad (L_1 \cap L_2)^\perp = L_1^\perp + L_2^\perp.$$

Proof. To see part (i), note that $x_L$ belongs to $L$ and therefore $x_L = \sum_i \lambda_i f^i$ for some $\lambda_i$'s, and $x - x_L$ is orthogonal to $L$. Hence, we have

$$x = (x - x_L) + \sum_{i=1}^{m} \lambda_i f^i.$$

Then, for any $j$, by taking inner products of both sides of the above equality with $f^j$, we arrive at

$$(f^j)^\top x = (f^j)^\top (x - x_L) + \sum_{i=1}^{m} \lambda_i (f^j)^\top f^i = \sum_{i=1}^{m} \lambda_i (f^j)^\top f^i = \lambda_j.$$

The result follows by plugging in the expressions for $\lambda_j$ in the equation $x_L = \sum_i \lambda_i f^i$.

Let us now prove part (ii). Since $L_1 \subseteq L_1 + L_2$ and $L_2 \subseteq L_1 + L_2$, a vector orthogonal to $L_1 + L_2$ is orthogonal to both $L_1$ and $L_2$. Vice versa, a vector with the latter property is orthogonal to sums of vectors from $L_1$ and from $L_2$, that is, it is orthogonal to $L_1 + L_2$. Thus, the vectors orthogonal to $L_1 + L_2$ are exactly the vectors orthogonal to both $L_1$ and $L_2$. This proves the first equality. Applying this equality to orthogonal complements of $L_1$ and $L_2$ in the role of $L_1$, $L_2$, we get the second equality. □

# Appendix B

# Prerequisites from Real Analysis

## B.1 Space $\mathbf{R}^n$: metric structure and topology

Euclidean structure on the space $\mathbf{R}^n$ gives rise to a number of extremely important *metric* notions – distances, convergence, etc. For the sake of definiteness, we associate these notions with the standard inner product $\langle x, y\rangle = x^\top y$.

### *B.1.1 Euclidean norm and distances*

By positive definiteness, the quantity $x^\top x$ is always nonnegative, so that the quantity

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

is well-defined. This quantity is called the (standard) *Euclidean norm* of vector $x$ (or simply the norm of $x$) and is treated as the distance from the origin to $x$. The distance between two arbitrary points $x, y \in \mathbf{R}^n$ is, by definition, the norm $d_2(x, y) = \|x - y\|_2$ of the difference $x - y$. These notions we have just introduced indeed satisfy all the basic requirements on the general notions of a norm $\|\cdot\| : \mathbf{R}^n \to \mathbf{R}$ and a distance $d(x, y) : \mathbf{R}^m \times \mathbf{R}^n \to \mathbf{R}$. Specifically:

1. *Positivity of norms:* The norm of a vector is always nonnegative. Furthermore, it is zero if and only if the vector is zero:
$$\|x\| \geq 0 \quad \forall x; \qquad \|x\| = 0 \iff x = 0.$$
2. *Homogeneity of norms:* When a vector is multiplied by a real, its norm is multiplied by the absolute value of the real:
$$\|\lambda x\| = |\lambda| \cdot \|x\| \quad \forall (x \in \mathbf{R}^n, \lambda \in \mathbf{R}).$$
3. *Triangle inequality:* Norm of the sum of two vectors is less than or equal to the sum of their norms:
$$\|x + y\| \leq \|x\| + \|y\| \quad \forall (x, y \in \mathbf{R}^n).$$

In contrast to the properties of positivity and homogeneity, which are absolutely evident, the Triangle inequality is not trivial and definitely requires a proof. Its proof goes through a fact which is extremely important by its own right – the *Cauchy Inequality*, which perhaps is the most frequently used inequality in Mathematics.

**Theorem** B.1 [Cauchy Inequality] For any $x, y \in \mathbf{R}^n$, we have

$$|x^\top y| \le \|x\|_2 \|y\|_2.$$

Moreover, the above relation holds as equality if and only if one of the vectors is proportional to the other one:

$$\begin{aligned} &|x^\top y| = \|x\|_2 \|y\|_2 \\ &\iff \quad \exists \alpha \in \mathbf{R} \text{ such that } x = \alpha y \text{ or } \exists \beta \in \mathbf{R} \text{ such that } y = \beta x. \end{aligned}$$

**Proof.** Without loss of generality we may assume that both $x$ and $y$ are nonzero (otherwise the Cauchy inequality is clearly equality, and one of the vectors is constant times (specifically, zero times) the other one, as desired). Assuming $x, y \neq 0$, consider the function

$$f(\lambda) = (x - \lambda y)^\top (x - \lambda y) = x^\top x - 2\lambda x^\top y + \lambda^2 y^\top y.$$

By positive definiteness of the inner product, this function – which is a second order polynomial – is nonnegative on the entire axis, hence the discriminant $(x^\top y)^2 - (x^\top x)(y^\top y)$ of $f$ is nonpositive:

$$(x^\top y)^2 \le (x^\top x)(y^\top y).$$

By taking square roots of both sides, we arrive at the Cauchy Inequality. We also see that the inequality holds as equality if and only if the discriminant of the second order polynomial $f(\lambda)$ is zero, i.e., if and only if the polynomial has a (multiple) real root. But, due to the positive definiteness of the inner product, $f(\cdot)$ has a root $\lambda$ if and only if $x = \lambda y$, which proves the second part of the theorem. □

*From Cauchy's Inequality to the Triangle Inequality:* Let $x, y \in \mathbf{R}^n$. Then,

$$\begin{aligned} \|x + y\|_2^2 &= (x + y)^\top (x + y) && \text{[by the definition of the Euclidean norm]} \\ &= x^\top x + y^\top y + 2x^\top y && \text{[by opening the parentheses]} \\ &\le \underbrace{x^\top x}_{=\|x\|_2^2} + \underbrace{y^\top y}_{=\|y\|_2^2} + 2\|x\|_2\|y\|_2 && \text{[by Cauchy's Inequality]} \\ &= (\|x\|_2 + \|y\|_2)^2 \\ \Longrightarrow \ \|x + y\|_2 &\le \|x\|_2 + \|y\|_2. \end{aligned}$$

□

We also have the following simple and useful fact.

**Fact** B.2 For *every norm* $\|\cdot\|$ *on* $\mathbf{R}^n$ and any $x, y \in \mathbf{R}^n$, we have

$$|\|x\| - \|y\|| \le \|x - y\|.$$

**Proof.** Indeed, $x = (x - y) + y$, and so by the Triangle inequality $\|x\| \le \|x - y\| + \|y\|$. That is, $\|x\| - \|y\| \le \|x - y\|$. Swapping $x$ and $y$, we get $\|y\| - \|x\| \le \|y - x\| = \|x - y\|$, as well. Thus, the result follows. □

The properties of a norm (i.e., of the distance to the origin) we have established induce properties of the distances between pairs of arbitrary points in $\mathbf{R}^n$. Specifically, $d_2(x,y) = \|x-y\|_2$, same as any other norm-induced distance $d_{\|\cdot\|}(x,y) = \|x-y\|$, possesses the following standard properties of a general distance $d(x,y) : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$:

1. *Positivity of distances:* The distance $d(x,y)$ between two points $x,y \in \mathbf{R}^n$ is positive, except for the case when the points coincide ($x=y$) in which case the distance between $x$ and $y$ is zero:
$$d(x,y) \geq 0,\ \forall (x,y \in \mathbf{R}^n); \qquad d(x,y) = 0 \iff x = y.$$
2. *Symmetry of distances:* For any $x,y \in \mathbf{R}^n$, the distance from $x$ to $y$ is the same as the distance from $y$ to $x$:
$$d(x,y) = d(y,x),\ \forall (x,y \in \mathbf{R}^n).$$
3. *Triangle inequality for distances:* For every $x,y,z \in \mathbf{R}^n$, the distance from $x$ to $z$ does not exceed the sum of distances between $x$ and $y$ and between $y$ and $z$:
$$d(x,z) \leq d(x,y) + d(y,z), \quad \forall (x,y,z \in \mathbf{R}^n).$$

A norm-induced distance $d(x,y) = \|x-y\|$ possesses the following additional properties as well:

1. *Shift invariance:* $d(x+h, y+h) = d(x,y)$ for all $x,y,h \in \mathbf{R}^n$.
2. *Homogeneity:* $d(\lambda x, \lambda y) = |\lambda|\, d(x,y)$ for all $x,y \in \mathbf{R}^n$ and $\lambda \in \mathbf{R}$.

**Equivalence of norms on $\mathbf{R}^n$.** As is immediately seen, any distance $d(x,y) = \|x-y\|$ induced by an arbitrary norm $\|\cdot\|$ on $\mathbf{R}^n$ possesses the above properties 1 – 5. In fact (check it!), every distance $d(x,y)$ on $\mathbf{R}^n$ satisfying properties 1 – 5 is indeed induced by a norm, specifically, the norm $d(0,x)$ (in the case of properties 1 – 5, $d(0,x)$ indeed is a norm).

We next discuss a very important fact (characteristic for *finite-dimensional* linear spaces) which states that all these distances are within constant factors from each other.

**Proposition** B.3 Any two norms $\|\cdot\|$, $\|\cdot\|'$ on $\mathbf{R}^n$ are equivalent, i.e., there exists a positive constant $c$ (which depends on the given pair of norms) such that
$$c \leq \frac{\|x\|'}{\|x\|} \leq \frac{1}{c}, \qquad \forall (x \in \mathbf{R}^n, x \neq 0). \tag{B.1}$$

**Proof.** Clearly it suffices to justify equivalence of any norm to a fixed one, say, the Euclidean norm $\|\cdot\|_2$. Thus, let us take $\|\cdot\|' \equiv \|\cdot\|_2$. It suffices to prove that there exist positive constants $c$ and $C$ such that both of the following conditions hold:
$$\begin{array}{lll} (a) & \|x\| \leq C\|x\|_2, & \forall x \neq 0; \\ (b) & \|x\|_2 \leq c\|x\|, & \forall x \neq 0. \end{array} \tag{B.2}$$

To prove $(a)$, all we need is to set $C := \sum_{i=1}^n \|e_i\|$, where $e_1, \dots, e_n$ are the standard basis vectors. This is because

$$\|x\| = \left\| \sum_{i=1}^n x_i e_i \right\| \le \sum_{i=1}^n |x_i| \|e_i\| \le \left( \max_{i=1,\dots,n} |x_i| \right) \sum_{i=1}^n \|e_i\| \le C\|x\|_2.$$

In order to prove $(b)$, we will show that $0 < \inf_x \{\|x\|/\|x\|_2 : x \neq 0\}$. Assume, on the contrary, that the infimum in question is 0. Then, there exists a sequence of nonzero vectors $x^t$, $t = 1, 2, \dots$ such that $\|x^t\|/\|x^t\|_2 \to 0$ as $t \to \infty$. Since both $\|\cdot\|$ and $\|\cdot\|_2$ are homogeneous, we lose nothing by scaling $x^t$ to have $\|x^t\|_2 = 1$ for all $t$, whence we have $\|x^t\| \to 0$ as $t \to \infty$. As $\|x^t\|_2 = 1$ for all $t$, for any $1 \le i \le n$, the sequences $x_i^t$ of the $i$-th entries in $x^t$ are bounded. Passing to a subsequence, we can assume w.l.o.g. that all these sequences have limits as $t \to \infty$, i.e., $\lim_{t\to\infty} x_i^t = z_i$ for some $z \in \mathbf{R}^n$. Passing to limit in the equality $\sum_{i=1}^n (x_i^t)^2 = 1$, we get $\|z\|_2 = 1$. Clearly, $\|x^t - z\|_2 \to 0$ as $t \to \infty$. Moreover, from $(a)$ we have $\|x^t - z\| \le C\|x^t - z\|_2$ which implies that $\|x^t - z\| \to 0$ as $t \to \infty$. Then, by Triangle inequality, we deduce $\|z\| \le \|z - x^t\| + \|x^t\| \to 0$ as $t \to \infty$, and thus $\|z\| = 0$. This is the desired contradiction since $z \neq 0$ and therefore $\|z\| > 0$. □

**Note:** In the preceding proof, we used for granted the following fundamental property of real line (stemming from centuries-long development of rigorous theory of real numbers):

> **Theorem 0.** *Every bounded sequence of reals $\{x_t\}_{t\ge1}$ has a converging subsequence, that is, for properly selected $t_1 < t_2 < \dots$ and $\bar{x} \in \mathbf{R}$ the sequence $x_{t_i}$ converges to $\bar{x}$ as $i \to \infty$: for every $\epsilon > 0$ one has $|\bar{x} - x_{t_i}| < \epsilon$ for all but finitely many indices $i$.*

As an immediate consequence, from every bounded sequence $\{x^t\}_{t\ge1}$ of vectors from $\mathbf{R}^n$ (boundedness meaning that the sequences of $i$-th entries in $x^t$ are bounded for every $i \le n$) one can extract a subsequence $\{x^{t_s} : s = 1, 2, \dots, t_1 < t_2 < \dots\}$ such that for every $i \le n$ the sequences of $i$-th coordinates $x_i^{t_s}$ of vectors $x^{t_s}$ converge as $s \to \infty$ to some reals – the fact we used in the proof of Proposition B.3. Indeed, the sequences $\{x_i^t\}_{t\ge1}$ are bounded for every $i$. By Theorem 0 we can extract from $\{x^t\}_{t\ge1}$ a subsequence $\{x^{t_s}, s = 1, 2, \dots, t_1 < t_2 < \dots\}$ with converging first entries, from this subsequence – a subsequence with converging second entries, and so on. In $n$ steps of this process we get a subsequence of the original sequence of vectors such that $i$-th entries in the vectors from our subsequence converge to some reals $\bar{x}_i$, $i \le n$. This is formalized as Theorem B.15 in Section B.1.4.

### B.1.2 Convergence

Equipped with distances, we can define the fundamental notion of *convergence of a sequence of vectors*. Specifically, we say that *a sequence $x^1, x^2, \dots$ of vectors from $\mathbf{R}^n$ converges to a vector $\bar{x}$*, or, equivalently, that *$\bar{x}$ is the limit of the* *sequence* $\{x^i\}$ [notation: $\bar{x} = \lim\limits_{i\to\infty} x^i$], if the distances from $\bar{x}$ to $x^i$ go to 0 as $i \to \infty$:

$$\bar{x} = \lim_{i\to\infty} x^i \iff \|\bar{x} - x^i\|_2 \to 0 \ \text{ as } \ i \to \infty,$$

or, equivalently, for every $\epsilon > 0$ there exists $i = i(\epsilon)$ such that the distance between every point $x^i$, $i \geq i(\epsilon)$, and $\bar{x}$ does not exceed $\epsilon$:

$$\left\{\|\bar{x} - x^i\|_2 \to 0 \text{ as } i \to \infty\right\} \iff \left\{\forall \epsilon > 0, \exists i(\epsilon) : i \geq i(\epsilon) \implies \|\bar{x} - x^i\|_2 \leq \epsilon\right\}.$$

Note that by Proposition B.3 every two norm-induced distances on $\mathbf{R}^n$ are within multiplicative factors from each other, meaning that *replacing in the definition of convergence the Euclidean distance with any other norm-induced distance, we do not affect convergence per se – the fact that a sequence converges, and the limit of such a sequence.*

**Fact** B.4 All of the following statements are correct:

(i) For any $\bar{x} \in \mathbf{R}^n$, we have $\bar{x} = \lim\limits_{t\to\infty} x^t$ if and only if for every index $i = 1, \ldots, n$ the $i$-th coordinate of the vectors $x^t$ converge to the $i$-th coordinate of the vector $\bar{x}$ as $t \to \infty$.

(ii) If a sequence converges, its limit is uniquely defined.

(iii) Convergence is compatible with linear operations:

- if $x^t \to x$ and $y^t \to y$ as $t \to \infty$, then $x^t + y^t \to x + y$ as $t \to \infty$;
- if $x^t \to x$ and $\lambda_t \to \lambda$ as $t \to \infty$, then $\lambda_t x^t \to \lambda x$ as $t \to \infty$.

We also introduce two other notions related to sequences in $\mathbf{R}$ to analyze their "extremes." Given a sequence $\{x_t\} \subset \mathbf{R}$, its *limit inferior* (lower limit) [notation: $\liminf_{t\to\infty} x_t$] is defined by

$$\liminf_{t\to\infty} x_t := \lim_{t\to\infty} \left(\inf_{m\geq t} x_m\right).$$

That is,

$$\liminf_{t\to\infty} x_t = \sup_{t\geq 0} \inf_{m\geq t} x_m = \sup\left\{\inf\{x_m : m \geq t\} : t \geq 0\right\}.$$

In other words, if $\{x_t\}$ is a sequence of reals, $a = \liminf_{t\to\infty} x_t$ is

- either $-\infty$, if there is a subsequence diverging to $-\infty$,
- or $+\infty$, if $x_t \to \infty$, $t \to \infty$,
- or the smallest real $a$ which can be represented as the limit of a subsequence of $\{x_t\}$.

In contrast to the usual limit, lim inf is well defined for every sequence of reals, e.g., $\liminf_{t\to\infty}(-1)^t = -1$.

Similarly, the *limit superior* (upper limit) [notation: $\limsup_{t\to\infty} a_t$] is defined by

$$\limsup_{t\to\infty} x_t := \lim_{t\to\infty} \left(\sup_{m\geq t} x_m\right),$$

which is nothing but

$$\limsup_{t\to\infty} x_t = \inf_{t\geq 0}\sup_{m\geq t} x_m = \inf\left\{\sup\{x_m : m \geq t\} : t \geq 0\right\}.$$

### *B.1.3 Closed and open sets*

Now that we have at our disposal the notions of distance and convergence, we can speak about *closed* and *open* sets.

**Definition** B.5 [Closed set] A set $X \subseteq \mathbf{R}^n$ is called *closed*, if it contains limits of all converging sequences of elements from $X$:

$$\left\{x^i \in X,\ x = \lim_{i\to\infty} x^i\right\} \implies x \in X.$$

**Example** B.6 (Examples of closed sets) Each of the following sets is closed:

1. $\mathbf{R}^n$.
2. $\varnothing$.
3. The set composed of the points $x^i = (i, 0, \ldots, 0)$, $i = 1, 2, 3, \ldots$
4. Any finite subset of $\mathbf{R}^n$.
5. Any linear subspace in $\mathbf{R}^n$, i.e., any set of the form $\left\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n a_{ij}x_j = 0,\ i = 1, \ldots, m\right\}$ (see Proposition A.46).
6. Any affine subspace of $\mathbf{R}^n$, i.e., any nonempty set of the form $\left\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n a_{ij}x_j = b_i,\ i = 1, \ldots, m\right\}$ (see Proposition A.47).

**Example** B.7 (Examples of non-closed sets) Each of the following sets is *not* closed, provided $n > 0$:

1. $\mathbf{R}^n \setminus \{0\}$.
2. The set composed of points $x^i = (1/i, 0, \ldots, 0)$, for $i = 1, 2, 3, \ldots$
3. The set $\{x \in \mathbf{R}^n : x_j > 0,\ \forall j = 1, \ldots, n\}$.
4. The set $\left\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n x_j > 5\right\}$.

**Definition** B.8 [Open set] A set $X \subseteq \mathbf{R}^n$ is called *open*, if whenever $x$ belongs to $X$, all points close enough to $x$ also belong to $X$:

$$\forall (x \in X)\ \exists (\delta > 0) : \|x' - x\|_2 < \delta \implies x' \in X.$$

**Definition** B.9 [Neighborhood, interior point] An open set containing a point $x \in \mathbf{R}^n$ is called a *neighborhood* of $x$. A point $x \in \mathbf{R}^n$ from a set $X$ is called an *interior point* of $X$ if $X$ contains a neighborhood of $x$.

Note that based on these definitions, open sets are exactly the sets such that every point of the set is its interior point.

**Definition** B.10 [Interior of a set] Given a set $X \subseteq \mathbf{R}^n$, the *interior* of $X$ [notation: $\mathrm{int}(X)$ or $\mathrm{int}\, X$] is defined as the set of all interior points of $X$:

$$\mathrm{int}(X) := \{x \in X : \exists \delta > 0 \text{ such that } x' \in X\, \forall (x' : \|x - x'\|_2 < \delta)\}.$$

Note that a set $X \subseteq \mathbf{R}^n$ is open if and only if $X = \mathrm{int}(X)$.

**Example** B.11 (Examples of open sets) Each of the following sets is open:

1. $\mathbf{R}^n$.
2. $\varnothing$.
3. The set $\left\{x \in \mathbf{R}^n : \sum\limits_{j=1}^n a_{ij} x_j > b_i,\ \forall i = 1, \ldots, m\right\}$.
4. The complement of any finite set.

**Example** B.12 (Examples of non-open sets) Each of the following sets is *not* open:

1. A nonempty finite set.
2. The sequence $x^i = (1/i, 0, \ldots, 0)$, $i = 1, 2, 3, \ldots$.
3. The composed of points from a sequence $x^i \in ]bR^n$, $i = 1, 2, 3, \ldots$
4. The set $\left\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n x_j \geq 5\right\}$.

Based on these examples, we see that $\mathbf{R}^n$ and $\varnothing$ are both open and closed. Also, note that there are sets that are neither closed nor open, e.g.,

$$\left\{x \in \mathbf{R}^2 : x_1 > 0,\ x_2 \geq 0\right\}.$$

**Fact** B.13 (i) A set $X \subseteq \mathbf{R}^n$ is closed if and only if its complement $\overline{X} := \mathbf{R}^n \setminus X$ is open.

(ii) Intersection of every (finite or infinite) family of closed sets is closed. Union of every (finite or infinite) family of open sets is open.

(iii) Union of finitely many closed sets is closed. Intersection of finitely many open sets is open.

We close this section with another important definition.

**Definition** B.14 [Closure of a set] The *closure* of a set $X \subset \mathbf{R}^n$ [notation: $\mathrm{cl}\, X$ or $\mathrm{cl}(X)$] is the smallest closed set which contains $X$, i.e., it is the intersection of all closed sets containing $X$ (this intersection indeed is closed by Fact B.13.ii).

Equivalently (the equivalence is established in Fact I.1.23), $\mathrm{cl}\, X$ is the set composed of the limits of all converging sequences of points from $X$:

$$\mathrm{cl}\, X = \left\{ x \in \mathbf{R}^n : \ \exists \left\{x^i\right\} \in X \text{ such that } x = \lim_{i\to\infty} x^i \right\}.$$

Note that a set $X \subseteq \mathbf{R}^n$ is closed if and only if $X = \mathrm{cl}(X)$.

### *B.1.4 Local compactness of* $\mathbf{R}^n$

The following is a fundamental fact about convergence in $\mathbf{R}^n$, which in certain sense is characteristic for this series of spaces.

**Theorem** B.15 From every bounded sequence $\{x^i\}_{i=1}^{\infty}$ of points from $\mathbf{R}^n$ one can extract a converging subsequence $\{x^{i_j}\}_{j=1}^{\infty}$.

**Proof.** Let $\{x^i \in \mathbf{R}^n\}_{i\geq 1}$ be a bounded sequence. Since every bounded sequence of reals has a converging subsequence, we can extract from $\{x^i\}$ a subsequence with the first entries of its members converging to a limit. Note that this subsequence itself will be bounded. Then, from this subsequence, we can extract a subsequence with the second entries of the members converging to a limit. Extracting in this fashion subsequences from already generated ones, after $n$ steps we get a subsequence of the original sequence with every entry of the selected members converging to a limit. That is, the selected final subsequence converges entrywise (and then - in the Euclidean norm as well) to the vector composed of the above limits. □

We next introduce an important concept related to the closed sets.

**Definition** B.16 [Compact set] A set $X$ is called *compact* if every sequence in $X$ has a subsequence that converges to an element again contained in $X$.

**Theorem** B.17 Any set $X \subseteq \mathbf{R}^n$ that is closed and bounded is compact. Moreover, any compact set $X \subseteq \mathbf{R}^n$ is closed and bounded.

**Proof.** Let $X$ be a closed and bounded subset of $\mathbf{R}^n$. Consider any sequence $\{x^i\}$ of points from $X$. Note that this sequence is bounded as $X$ is bounded. Then, by Theorem B.15 there exists a subsequence of our sequence which has a limit. Since $X$ is closed, this limit belongs to $X$, as required in the definition of compact sets.

To prove the last statement in this theorem, let $X$ be a compact set in $\mathbf{R}^n$. Note that if the set $X$ were unbounded, it would be possible to select a sequence of points from $X$ with norms diverging to $+\infty$, and such a sequence has no converging subsequences, which contradicts to $X$ being compact. Furthermore, if $X$ were to be not closed, we could find a sequence of points from $X$ converging to a point *not* from $X$. For such a sequence, no subsequence converges to a point from

$X$ (since limits of all these subsequences are the same as the limit of the entire sequence, and the latter does not belong to $X$), which once again contradicts to $X$ being compact. □.

Given a set $X \subseteq \mathbf{R}^n$, we refer to a family $\mathcal{C}$ of open sets such that $X \subseteq \bigcup\limits_{U\in\mathcal{C}} U$ as an *open covering of* $X$. We first establish a general statement regarding open coverings, and then show that a much stronger version of this property gives an alternative characterization of compact sets.

**Proposition** B.18 Given a set $X \subseteq \mathbf{R}^n$ and its open covering $\mathcal{C}$, one can extract a countable subcovering – there exists a sequence $U_1, U_2, \ldots$ of members of $\mathcal{C}$ such that $X \subseteq \bigcup\limits_{i\geq 1} U_i$.

**Proof.** Observe, first, that the family $\mathcal{V}$ of all open balls $\{x \in \mathbf{R}^n : \|x - \bar{x}\|_2 < r\}$ where the radius $r$ and all coordinates of the centers $\bar{x}$ are rational is countable, i.e., all balls from $\mathcal{V}$ can be arranged into sequence $V_1, V_2, \ldots$. This is because the rational data – radius and coordinates of center – of a ball from $\mathcal{V}$ form a collection of $n+1$ rational numbers, or, which is the same, a collection of $2(n+1)$ integers, and all collections of the latter type can be arranged into a sequence: we first list all collections of $2(n+1)$ integers with the total of their magnitudes not exceeding 1 (there are finitely many collections of this type), then list the yet unlisted collections with the total of magnitudes of members not exceeding 2, and so on. This construction clearly arranges into a sequence all collections of $2(n+1)$ integers, that is, all balls with rational centers and positive rational radii.

Now let us build a countable subcovering $U_1, U_2, \ldots$ of a given open covering $\mathcal{C}$ of $X \subseteq \mathbf{R}^n$. We will process the balls $V_1$, $V_2$,... one by one. When processing $V_i$, we check whether this set is covered by a member of $\mathcal{C}$. If $V_i$ is covered by a member of $\mathcal{C}$, we add this member to the sequence $U_1, U_2, \ldots$ we are building. Otherwise, we add to the $U$-sequence nothing and pass to processing $V_{i+1}$.

Now, every point $x$ of $X$ belongs to certain member $W_x$ of covering $\mathcal{C}$ as $X \subseteq \bigcup\limits_{C\in\mathcal{C}} C$. Since $W_x$ is open, among the balls $V_i$ we can find a ball, let us call its index $i(x)$, such that the center of $V_{i(x)}$ is close to $x$ and the radius of $V_{i(x)}$ is small enough to ensure that $x \in V_{i(x)} \subset W_x$. Then, at the step $i(x)$ of processing $V_1, V_2, \ldots$ the family $\mathcal{C}$ contains a member, namely, $W_x$, which covers $V_{i(x)}$. Therefore, at this step we will add to the $U$-sequence a set which contains $V_{i(x)}$ and hence contains $x$. We conclude that every $x \in X$ is contained in some set of the sequence $U_1, U_2, \ldots$ of members of $\mathcal{C}$, and thus this sequence is a countable subcovering of $X$ as desired. □

We close by providing a characterization of compact sets through their open coverings.

**Theorem** B.19 A set $X \subseteq \mathbf{R}^n$ is compact if and only if it possesses the following property:

From every open covering $\mathcal{C}$ of $X$ one can extract a *finite* sub-covering, i.e., a finite subfamily $\mathcal{C}' \subseteq \mathcal{C}$ such that $\mathcal{C}'$ still is a covering of $X$.

**Proof.** First, suppose that $X$ satisfies the stated property. We will prove that $X$ is closed and bounded, which by Theorem B.17 will imply that $X$ is compact. Suppose $X$ is unbounded, and consider the covering of $X$ given by all open balls centered at points from $X$. Then, we would be unable to extract a finite subcovering of $X$ from this open covering. Now, assume for contradiction that $X$ is not closed, and consider the covering of $X$ given by the open sets $\{x : \|x - \bar{x}\|_2 > 1/i\}$, $i = 1, 2, \ldots$ associated with a point $\bar{x} \in \mathrm{cl}(X) \setminus X$. Then, we would be unable to extract a finite subcovering of $X$ from this open covering of $X$ which is a contradiction.

Now, let $X$ be closed and bounded (and thus compact by Theorem B.17). Consider an open covering $\mathcal{C}$ of $X$. By Proposition B.18, we can extract from $\mathcal{C}$ a countable subcovering: $X \subset \bigcup_{i \geq 1} U_i$, $U_i \in \mathcal{C}$. We claim that in fact just finitely many of the sets $U_i$ suffices to cover $X$. This, of course, is enough to show that $X$ obeys the desired property stated in the theorem. Now, assume for contradiction that for every $i = 1, 2, \ldots$ there is a point $x^i \in X \setminus \bigcup_{j \leq i} U_j$. Since $X$ is compact, by definition we can extract from $\{x^i\}$ a subsequence converging to some $\bar{x} \in X$. Since $U_i$ are open and cover $X$, there exists $i_*$ such that $\bar{x} \in U_{i_*}$. Also, as $U_{i_*}$ is open and the subsequence in question converges to $\bar{x}$, the subsequence (and therefore the entire sequence $\{x^i\}$) visits $U_{i_*}$ infinitely many times. This gives us the desired contradiction, since by construction all points $x^i$ with $i > i_*$ do not belong to $U_{i_*}$. □

## B.2 Continuous functions on $\mathbf{R}^n$

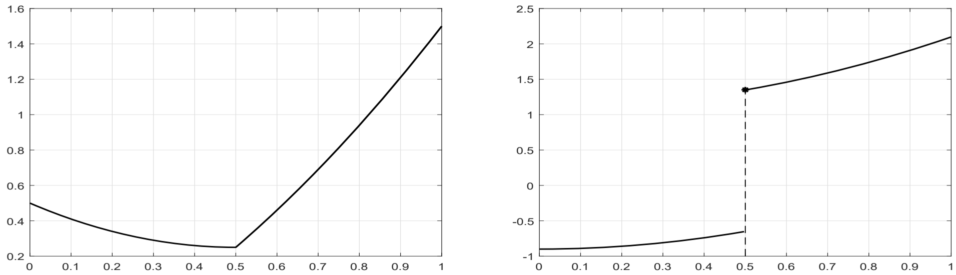
### B.2.1 Continuity of a function

In this section, we consider $X \subseteq \mathbf{R}^n$ and examine a function (also called *mapping*) $f(x) : X \to \mathbf{R}^m$ defined on $X$ and taking values in $\mathbf{R}^m$. We start with basic definitions.

**Definition** B.20 [Continuity at a point] A function $f(x) : X \to \mathbf{R}^m$ is called *continuous at a point* $\bar{x} \in X$, if for every sequence $x^i$ of points of $X$ converging to $\bar{x}$ the sequence $f(x^i)$ converges to $f(\bar{x})$. Equivalently, $f : X \to \mathbf{R}^m$ is continuous at $\bar{x} \in X$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that

$$x \in X, \ \|x - \bar{x}\|_2 < \delta \implies |f(x) - f(\bar{x})| < \epsilon.$$

**Definition** B.21 [Continuous function] A function $f(x) : X \to \mathbf{R}^m$ is called *continuous on* $X$, if $f$ is continuous at every point from $X$. Equivalently, $f$ is continuous if $f$ preserves convergence: whenever a sequence of points $x^i \in X$ converges to a point $x \in X$, the sequence $f(x^i)$ converges to $f(x)$.



continuous function $f : [0,1] \to \mathbf{R}$ discontinuous function $f : [0,1] \to \mathbf{R}$

Figure V.1. Continuity vs. discontinuity

**Example** B.22 (Continuous mappings) An *affine* mapping

$$f(x) := \left[\sum_{j=1}^{m} A_{1j}x_j + b_1; \ldots; \sum_{j=1}^{m} A_{mj}x_j + b_m\right] \equiv Ax + b : \mathbf{R}^n \to \mathbf{R}^m$$

is continuous on the entire $\mathbf{R}^n$ (and thus – on every subset of $\mathbf{R}^n$) (check it!).

**Fact** B.23 The norm $\|x\|_2$ is a real-valued function and it is continuous on $\mathbf{R}^n$ (and thus – on every subset of $\mathbf{R}^n$) (check it!). In fact, *every* norm $\|\cdot\|$ on $\mathbf{R}^n$ is continuous.

### *B.2.2 Elementary continuity-preserving operations*

All “elementary” operations with mappings preserve continuity.

**Theorem** B.24 Let $X \subseteq \mathbf{R}^n$.

(i) [stability of continuity w.r.t. linear operations] If $f_1(x), f_2(x)$ are continuous functions on $X$ taking values in $\mathbf{R}^m$ and $\lambda_1(x)$, $\lambda_2(x)$ are continuous real-valued functions on $X$, then the function

$$f(x) = \lambda_1(x)f_1(x) + \lambda_2(x)f_2(x) : X \to \mathbf{R}^m$$

is continuous on $X$.

(ii) [stability of continuity w.r.t. superposition] Let

- $X \subseteq \mathbf{R}^n$, $Y \subseteq \mathbf{R}^m$;
- $f : X \to \mathbf{R}^m$ be a continuous mapping such that $f(x) \in Y$ for every $x \in X$;
- $g : Y \to \mathbf{R}^k$ be a continuous mapping.

Then, the composite mapping

$$h(x) := g(f(x)) : X \to \mathbf{R}^k$$

is continuous on $X$.

All these claims are self-evident.

### B.2.3 Basic properties of continuous functions on $\mathbf{R}^n$

The basic properties of continuous functions on $\mathbf{R}^n$ can be summarized as the next three theorems.

**Theorem** B.25 Let $X$ be a nonempty compact subset of $\mathbf{R}^n$. If a mapping $f : X \to \mathbf{R}^m$ is continuous on $X$, it is bounded on $X$, i.e., there exists $C \in \mathbf{R}$ such that $\|f(x)\|_2 \le C$ for all $x \in X$. Moreover, its image $f(X) := \{f(x) : \ x \in X\}$ is closed. Hence, the image $f(X)$ of a compact set $X \subset \mathbf{R}^n$ under a continuous mapping $f$ is compact.

**Proof.** Assume for contradiction that $f$ is unbounded. Then, for every $i \in \mathbf{N}$ there exists a point $x^i \in X$ such that $\|f(x^i)\|_2 > i$. Since $X$ is bounded, by Theorem B.15, we can extract from the sequence $\{x^i\}$ a subsequence $\{x^{i_j}\}_{j=1}^{\infty}$ which converges to a point $\bar{x} \in X$. The real-valued function $g(x) = \|f(x)\|_2$ is continuous (as the superposition of two continuous mappings, see Theorem B.24.ii) and therefore its values at the points $x^{i_j}$ should converge, as $j \to \infty$, to its value at $\bar{x}$. On the other hand, $g(x^{i_j}) \ge i_j \to \infty$ as $j \to \infty$, and we get the desired contradiction.

In order to prove that $f(X)$ is closed, we should prove that if a sequence $y^i := f(x^i)$, $x^i \in X$, of points from $f(X)$ converges to $\bar{y}$, as $i \to \infty$, then $\bar{y} \in f(X)$. Indeed, since $X$ is compact, we can extract from $\{x^i\}$ a subsequence converging to some $\bar{x} \in X$. By continuity of $f$, the corresponding subsequence of the sequence $\{y^i = f(x^i)\}$ converges to $f(\bar{x})$. Thus, $\bar{y} = f(\bar{x}) \in f(X)$, as claimed. □

A stronger notion of continuity is defined as follows:

**Definition** B.26 [Uniformly continuous function] A function $f(x) : X \to \mathbf{R}^m$ is called *uniformly continuous on* $X$, if for every $\epsilon > 0$ there exists $\delta > 0$ such that

$$x, y \in X, \ \|x - y\|_2 < \delta \implies \|f(x) - f(y)\|_2 < \epsilon.$$

**Remark** B.27 It is important to note the difference between the usual continuity and the uniform continuity. The usual continuity of a function $f$ means that given $\epsilon > 0$ *and a point* $x \in X$, it is possible to choose $\delta(\epsilon, x) > 0$ such that for all $y \in X$, $\|x - y\|_2 < \delta(\epsilon, x) \implies \|f(x) - f(y)\|_2 < \epsilon$; here the appropriate value of $\delta$ can depend on $\epsilon$ *and on* $x$. The uniform continuity means that this positive $\delta$ may be chosen as a function of *only* $\epsilon$.

**Example** B.28 (Uniformly continuous functions) All of the following mappings are uniformly continuous on $X := [1, 100]$:

- $f(x) := ax + b$
- $f(x) := |x|$
- $f(x) = x^2$
- $f(x) = 1/x$

**Theorem** B.29 Let $X \subset \mathbf{R}^n$ be compact. If a mapping $f : X \to \mathbf{R}^m$ is continuous on $X$, then it is *uniformly continuous.*

**Proof.** Assume for contradiction that there exists $\epsilon > 0$ such that for every $\delta > 0$ one can find a pair of points $x, y \in X$ such that $\|x - y\|_2 < \delta$ and $\|f(x) - f(y)\|_2 \geq \epsilon$. In particular, for every $i = 1, 2, \ldots$, we can find two points $x^i, y^i$ in $X$ such that $\|x^i - y^i\|_2 \leq 1/i$ and $\|f(x^i) - f(y^i)\|_2 \geq \epsilon$. As $X$ is compact, by Theorem B.15, we can extract from the sequence $\{x^i\}$ a subsequence $\{x^{i_j}\}_{j=1}^{\infty}$ which converges to certain point $\bar{x} \in X$. Since $\|y^{i_j} - x^{i_j}\|_2 \leq 1/i_j \to 0$ as $j \to \infty$, the sequence $\{y^{i_j}\}_{j=1}^{\infty}$ converges to the same point $\bar{x}$ as the sequence $\{x^{i_j}\}_{j=1}^{\infty}$ (why?) Since $f$ is continuous, we have

$$\lim_{j\to\infty} f(y^{i_j}) = f(\bar{x}) = \lim_{j\to\infty} f(x^{i_j}),$$

whence $\lim\limits_{j\to\infty}(f(x^{i_j}) - f(y^{i_j})) = 0$. But, this contradicts the assumption that $\|f(x^{i_j}) - f(y^{i_j})\|_2 \geq \epsilon > 0$ for all $j$. □

**Remark** B.30 The fact that a function that is continuous on a compact set is automatically uniformly continuous on the set is one of the most useful features of compact sets. Note also that compactness – closedness and boundednes – of the domain is crucial here: every one of the functions $f(x) = 1/x : (0, 1] \to \mathbf{R}$, $f(x) = x^2 : \mathbf{R} \to \mathbf{R}$ is continuous on the respective domain, and neither one is uniformly continuous on it.

**Theorem** B.31 Let $X \subset \mathbf{R}^n$ be nonempty and compact and let $f$ be a real-valued continuous function on $X$. Then, $f$ attains both its minimum and its maximum on $X$, i.e.,

$$\begin{aligned}\operatorname*{Argmin}_X f &:= \left\{x \in X : \ f(x) = \inf_{y\in X} f(y)\right\} \neq \varnothing,\\ \operatorname*{Argmax}_X f &:= \left\{x \in X : \ f(x) = \sup_{y\in X} f(y)\right\} \neq \varnothing.\end{aligned}$$

**Proof.** We will prove that $f$ attains its maximum on $X$; the proof for minimum is completely analogous. Since $f$ is bounded on $X$ by Theorem B.25, the quantity

$$f^* := \sup_{x\in X} f(x)$$

is finite. Thus, we can find a sequence $\{x^i\}$ of points from $X$ such that $f^* =$

$\lim\limits_{i\to\infty} f(x^i)$. As $X$ is compact, we can extract from the sequence $\{x^i\}$ a subsequence $\{x^{i_j}\}_{j=1}^{\infty}$ which converges to a certain point $\bar{x} \in X$. Since $f$ is continuous on $X$, we have

$$f(\bar{x}) = \lim_{j\to\infty} f(x^{i_j}) = \lim_{i\to\infty} f(x^i) = f^*,$$

so that the maximum of $f$ on $X$ indeed is achieved (e.g., at the point $\bar{x}$). □

Theorem B.31 admits the following useful generalization for unbounded sets:

> **Theorem** B.32 Let $X \subseteq \mathbf{R}^n$ be a nonempty and closed set, and let $f$ be a real-valued continuous function on $X$ which "goes to $+\infty$ at $\infty$," i.e., $f(x^t) \to \infty$, $t \to \infty$, along every sequence $\{x^t \in X, t \geq 1\}$ with $\|x^t\|_2 \to \infty$, $t \to \infty$, or, which is the same, such that the sublevel sets $X_a := \{x \in X : f(x) \leq a\}$ of $f$ on $X$ are bounded for all $a \in \mathbf{R}$. Then, $f$ attains its minimum on $X$:
>
> $$\operatorname*{Argmin}_X f \neq \varnothing.$$

Needless to say, Theorem B.32 admits a maximization version; the reader is strongly advised to formulate this version on their own.

**Proof.** Since $X \neq \varnothing$, there exists a real number $a \in \mathbf{R}$ such that the set $X_a$ is nonempty. As $X$ is closed and $f$ is continuous on $X$, $X_a$ is closed. Moreover, under the premise of the theorem, this set is also bounded, so that $X_a$ is compact by Theorem B.17. By Theorem B.31, $f$ attains its minimum on $X_a$, and due to the origin of $X_a$, the minimizers of $f$ on $X_a$ are nothing but the minimizers of $f$ on $X$. □

## B.3 Semicontinuity

Theorem B.32 admits a useful extension in which we can relax the requirement for $f$ to be continuous. The related framework is as follows.

### *B.3.1 Functions with values in the extended real axis*

So far, when speaking about scalar functions, we dealt with real-valued functions defined on subsets of $\mathbf{R}^n$. However, in various minimization-related situations it is more convenient to speak about functions on $\mathbf{R}^n$ taking values in *extended real axis* $\mathbf{R} \cup \{+\infty\}$. To this end, we will

- augment $\mathbf{R}$ with "fictitious" point, denoted $+\infty$, and extend the usual orders $<$, $\leq$ onto the resulting *extended real line* $\mathbf{R}\cup\{+\infty\}$ by the natural convention that a real number is $< +\infty$ (and therefore is $\leq +\infty$), and, of course, $+\infty \leq +\infty$;
- consider functions $f(x) : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$; for such a function, its *domain* $\operatorname{Dom} f$ is defined as the set where the values of $f$ are reals:

$$\operatorname{Dom} f := \{x \in \mathbf{R}^n : \ f(x) \in \mathbf{R}\}.$$

Note that our "old" partially defined real-valued functions on $\mathbf{R}^n$, i.e., the functions $f : X \to \mathbf{R}$ with $X \subseteq \mathbf{R}^n$, can be extended by value $+\infty$ outside of $X$ to become functions on the *entire* $\mathbf{R}^n$ taking values in $\mathbf{R} \cup \{+\infty\}$. Such an extended function clearly "remembers its origin" – its domain is $X$, and its restriction onto $X$ coincides with the original $f$. Thus, two entities – the set $X$ on which the original $f$ was defined, and the original function $f$ itself – are now encoded in a single entity, i.e., the function on $\mathbf{R}^n$ taking values in $\mathbf{R} \cup \{+\infty\}$. While there is nothing deep in this encoding (this is just a convention), it saves a lot of words. In the sequel, if otherwise is not explicitly stated, "function" means a function on $\mathbf{R}^n$ taking values in $\mathbf{R} \cup \{+\infty\}$.

Note that the function with *empty* domain – one which is identically equal to $+\infty$ – is a fully legitimate inhabitant of our new world. In this world, functions with nonempty domains have a special name; they are called *proper*.

### *B.3.2 Epigraph of a function*

Given a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$, we associate with it its *epigraph* – the set

$$\mathrm{epi}\{f\} := \{[x;t] \in \mathbf{R}^n \times \mathbf{R} : \ t \geq f(x)\}.$$

Clearly, a set $E$ in $\mathbf{R}^n \times \mathbf{R}$ is an epigraph if and only if for every $x \in \mathbf{R}^n$ the intersection of $E$ with the "vertical" line $\ell_x := \{[x;t] : \ t \in \mathbf{R}\}$ is either empty, or is a ray of the form $\{t \in \mathbf{R} : \ t \geq t_x\}$ with some real number $t_x$. Moreover, a set with these properties remembers the underlying function of which it is epigraph, i.e., the value of this function at a point $x \in \mathbf{R}^n$ is $+\infty$ when $\ell_x$ does not intersect $E$, and is $t_x$ otherwise. The domain of a function is just the projection of its epigraph onto the $x$-plane:

$$\mathrm{Dom}\, f := \{x \in \mathbf{R}^n : \ \exists t \text{ such that } [x;t] \in \mathrm{epi}\{f\}\}.$$

### *B.3.3 Lower semicontinuity*

Functions with *closed* epigraphs have a name – they are called *lower semicontinuous* (*lsc* for short). Here is their characterization:

**Theorem** B.33 Let $f(x) : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$. The following properties of $f$ are equivalent to each other

(i) [lower semicontinuity] $\mathrm{epi}\{f\}$ is closed;

(ii) [closedness of sublevel sets] For every $\alpha \in \mathbf{R}$, the sublevel set of $f$ defined as

$$\mathrm{lev}_\alpha(f) := \{x \in \mathbf{R}^n : \ f(x) \leq \alpha\}$$

is closed;

(iii) Whenever a sequence $\{x^i \in \mathbf{R}^n, i \geq 1\}$ has a limit, say $x$, one has

$$f(x) \leq \liminf_{i\to\infty} f(x^i). \tag{B.3}$$

**Proof.** (i) $\Longrightarrow$ (ii). We want to prove that if epi$\{f\}$ is closed and $\alpha \in \mathbf{R}$, then $\mathrm{lev}_\alpha(f)$ is closed. This is immediate: if $x^i \in \mathrm{lev}_\alpha(f)$ and $x^i \to x$, $i \to \infty$, then $f(x^i) \leq \alpha$ for all $i$, so that the sequence $\overline{x}^i := [x^i; \alpha] \in \mathbf{R}^n \times \mathbf{R}$ belongs to epi$\{f\}$; as $i \to \infty$, this sequence converges to $[x; \alpha]$, and since epi$\{f\}$ is closed, $[x; \alpha] \in \mathrm{epi}\{f\}$, implying that $f(x) \leq \alpha$, that is, $x \in \mathrm{lev}_\alpha(f)$.

(ii) $\Longrightarrow$ (iii) Assume that the sublevel sets of $f$ are closed, and let us prove that whenever $x^i \to x$ as $i \to \infty$, (B.3) holds true. Passing to a subsequence of $\{x^i\}$, it suffices to consider the situation when $\liminf_{i\to\infty} f(x^i) = \lim_{i\to\infty} f(x^i) =: \beta$. Note that $\beta$ is either $+\infty$, or a real number, or $-\infty$. In the first case we clearly have $f(x) \leq \beta$. Suppose the second or the third case holds, and assume for the contradiction that $\beta < f(x)$. Let $\gamma$ be a real number such that $\beta < \gamma < f(x)$. Since $\gamma > \beta$, for all large enough $i$ we have $f(x^i) \leq \gamma$, that is, $x^i \in \mathrm{lev}_\gamma(f)$, implying that $x = \lim_{i\to\infty} x^i$ is the limit of a converging sequence of points from $\mathrm{lev}_\gamma(f)$. As the sublevel sets of $f$ are closed, we conclude that $x \in \mathrm{lev}_\gamma(f)$, that is, $f(x) \leq \gamma$, which contradicts the origin of $\gamma$.

(iii) $\Longrightarrow$ (i) Assume that $f$ obeys (iii), and let us prove that epi$\{f\}$ is closed. Suppose the points $[x^i; t_i] \in \mathrm{epi}\{f\}$ converge as $i \to \infty$ to $[x; t]$; we need to prove that $[x; t] \in \mathrm{epi}\{f\}$. As $[x^i; t_i] \in \mathrm{epi}\{f\}$, we have $f(x^i) \leq t_i$, and as $t_i \to t$, $i \to \infty$, we deduce $\liminf_{i\to\infty} f(x^i) \leq t$. Since $f$ obeys (ii), we conclude that $f(x) \leq t$, that is, $[x; t] \in \mathrm{epi}\{f\}$. □

Clearly, all finite-valued continuous functions are also lower semicontinuous. Moreover, there are many finite-valued discontinuous functions that are lower semicontinuous, i.e., precisely the ones with closed epigraphs. Based on the definition of lower semicontinuity, we also see that there are finite-valued discontinuous function that are not lsc. For example, the discontinuous function in Figure V.1 is not lsc.

The following corollary shows that a large class of continuous functions with extended values are also lsc.

**Corollary** B.34 Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a function that is continuous on its domain $\mathrm{Dom}\, f$ and let its domain be closed. Then, $f$ is lower semicontinuous.

**Proof.** For any $\alpha \in \mathbf{R}$, consider the sublevel set of $f$ given by $\mathrm{lev}_\alpha(f) = \{x \in \mathrm{Dom}\, f : f(x) \leq \alpha\}$. Under the promise of the corollary we will show that these sublevel sets are all closed, and then the result will follow from Theorem B.33.ii. Consider a converging sequence $\{x^i\}$ contained in $\mathrm{lev}_\alpha(f)$ and let $x := \lim_{i\to\infty} x^i$. As $\mathrm{Dom}\, f$ is closed and $x^i \in \mathrm{Dom}\, f$, we have $x \in \mathrm{Dom}\, f$. Since $f$ is continuous on $\mathrm{Dom}\, f$, we deduce $f(x) = \lim_{i\to\infty} f(x^i)$. Finally, as $f(x^i) \leq \alpha$ for all $i$, we get $f(x) \leq \alpha$, that is, $x \in \mathrm{lev}_\alpha(f)$. This proves that $\mathrm{lev}_\alpha(f)$ is closed as desired. □

The closedness of the domain in Corollary B.33 is indeed crucial – look what happens with the univariate function which is zero on the positive ray and $+\infty$ on the nonpositive ray.

Another useful example of lsc function is the pointwise supremum

$$f(x) = \sup_{\alpha \in \mathcal{A}} f_\alpha(x)$$

of a whatever family of lsc functions. Indeed, we clearly have $\mathrm{epi}\{f\} = \bigcap\limits_{\alpha \in \mathcal{A}} \mathrm{epi}\{f_\alpha\}$, and the intersection of a family of closed sets is closed.

We are now ready to state and prove the promised extension of Theorem B.32.

**Theorem** B.35 Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a proper lower semicontinuous function with bounded sublevel sets. Then, $f$ is below bounded and attains its minimum.

**Proof.** Since $f$ is proper, there exists $\alpha \in \mathbf{R}$ such that the set $\mathrm{lev}_\alpha(f)$ is nonempty. Since $f$ is lsc, $\mathrm{lev}_\alpha(f)$ is closed. Moreover, under the premise of the theorem $\mathrm{lev}_\alpha(f)$ is also bounded, and thus it is compact. Now, let $x^i \in \mathrm{lev}_\alpha(f)$ be a sequence such that $f(x^i) \to \inf_{x \in \mathrm{lev}_\alpha(f)} f(x)$ (note also that $\inf_{x \in \mathrm{lev}_\alpha(f)} f(x) = \inf_{x \in \mathbf{R}^n} f(x)$). Since $\mathrm{lev}_\alpha(f)$ is compact, passing to a subsequence, we can assume that $x^i \to \bar{x}$ as $i \to \infty$, whence, by Theorem B.33.iii, $f(\bar{x}) \le \liminf_{i \to \infty} f(x^i) = \inf_{x \in \mathbf{R}^n} f(x)$, implying that $\inf_{x \in \mathbf{R}^n} f(x)$ is $f(\bar{x}) \in \mathbf{R}$ and therefore $\inf_{x \in \mathbf{R}^n} f(x)$ is achieved and finite. □

### *B.3.4 Hypograph and upper semicontinuity*

While in the preceding sections we focused on minimizers of the functions and stated conditions under which such minimizers exist, there are of course natural "maximization" analogies of these developments. In the "maximization" analogies, the functions are allowed to take values in $\mathbf{R} \cup \{-\infty\}$, and the role of the epigraph is played by the *hypograph* of a function, i.e., the set

$$\mathrm{h}\mathit{ypo}\{f\} = \{[x;t] : t \le f(x)\},$$

and lower semicontinuity is replaced with *upper semicontinuity*, i.e., closedness of the hypograph. For example, the discontinuous function in Figure V.1 is upper semicontinuous.

In fact, a function $f : \mathbf{R}^n \to (\mathbf{R} \cup \{\pm\infty\})$ is upper semicontinuous if and only if $-f$ is lower semicontinuous; and such an $f$ is continuous if and only if it is both upper and lower semicontinuous.

## B.4 Exercises

Exercise 8 Mark in the list below those sets which are closed and those which are open (sets are in $\mathbf{R}^n$, $\|\cdot\|$ is a norm on $\mathbf{R}^n$, n¿0):

1. All vectors with integer coordinates.
2. All vectors with rational coordinates.

3. All vectors with positive coordinates.
4. All vectors with nonnegative coordinates.
5. $\{x \in \mathbf{R}^n : \ \|x\| < 1\}$.
6. $\{x \in \mathbf{R}^n : \ \|x\| = 1\}$.
7. $\{x \in \mathbf{R}^n : \ \|x\| \leq 1\}$
8. $\{x \in \mathbf{R}^n : \ \|x\| \geq 1\}$
9. $\{x \in \mathbf{R}^n : \ \|x\| > 1\}$
10. $\{x \in \mathbf{R}^n : \ 1 < \|x\| \leq 2\}$

Exercise 9 Consider the function $f(x_1, x_2) : \mathbf{R}^2 \to \mathbf{R}$ defined as

$$f(x_1, x_2) = \begin{cases} \frac{x_1^2 - x_2^2}{x_1^2 + x_2^2}, & \text{if } (x_1, x_2) \neq 0 \\ 0, & \text{if } x_1 = x_2 = 0. \end{cases}$$

Check whether this function is continuous on the following sets:

1. $\mathbf{R}^2$
2. $\mathbf{R}^2 \setminus \{0\}$
3. $\left\{x \in \mathbf{R}^2 : x_1 = 0\right\}$
4. $\left\{x \in \mathbf{R}^2 : x_2 = 0\right\}$
5. $\left\{x \in \mathbf{R}^2 : x_1 + x_2 = 0\right\}$
6. $\left\{x \in \mathbf{R}^2 : x_1 - x_2 = 0\right\}$
7. $\left\{x \in \mathbf{R}^2 : |x_1 - x_2| \leq x_1^4 + x_2^4\right\}$

Exercise 10 Let $f : \mathbf{R}^n \to \mathbf{R}^m$ be a continuous mapping. Among the following statements, mark those which are always true:

1. If $U$ is an open set in $\mathbf{R}^m$, then so is the set $f^{-1}(U) := \{x \in \mathbf{R}^n : \ f(x) \in U\}$.
2. If $U$ is an open set in $\mathbf{R}^n$, then so is the set $f(U) = \{f(x) : \ x \in U\}$.
3. If $F$ is a closed set in $\mathbf{R}^m$, then so is the set $f^{-1}(F) = \{x \in \mathbf{R}^n : \ f(x) \in F\}$.
4. If $F$ is a closed set in $\mathbf{R}^n$, then so is the set $f(F) = \{f(x) : \ x \in F\}$.

Exercise 11 Prove that in general *neither one* of Theorems B.25, B.29, and B.31 remains valid when $X$ is closed, but not bounded, same as when $X$ is bounded, but not closed.

## B.5 Proofs of Facts

**Fact B.4** All of the following statements are correct:

(i) For any $\bar{x} \in \mathbf{R}^n$, we have $\bar{x} = \lim\limits_{t\to\infty} x^t$ if and only if for every index $i = 1, \ldots, n$ the $i$-th coordinate of the vectors $x^t$ converge to the $i$-th coordinate of the vector $\bar{x}$ as $t \to \infty$.

(ii) If a sequence converges, its limit is uniquely defined.

(iii) Convergence is compatible with linear operations:

- if $x^t \to x$ and $y^t \to y$ as $t \to \infty$, then $x^t + y^t \to x + y$ as $t \to \infty$;
- if $x^t \to x$ and $\lambda_t \to \lambda$ as $t \to \infty$, then $\lambda_t x^t \to \lambda x$ as $t \to \infty$.

Proof.

(i) Indeed, $\|\bar{x} - x^i\|_2 = \sqrt{\sum_{j=1}^n (\bar{x}_j - (x^i)_j)^2}$ converges to 0 as $i \to \infty$ if and only if $\bar{x}_j - (x^i)_j$ converges to 0 as $i \to \infty$ for every $j \leq n$.

(ii) Suppose $x'$ and $x''$ both are limits of a converging sequence $\{x^i\}$. Then, for every $\epsilon > 0$ and all large enough $i$ it holds that $\|x' - x^i\|_2 \leq \epsilon$, $\|x'' - x^i\|_2 \leq \epsilon$, and so by Triangle inequality $\|x' - x''\|_2 \leq 2\epsilon$. Since $\epsilon > 0$ is arbitrary, we conclude that $\|x' - x''\|_2 \leq 0$, and thus $x' = x''$.

(iii) By elementary Calculus, both claims are valid for sequences of real numbers; this combines with item (i) to imply the validity of the claims for vectors. □

**Fact B.13** (i) A set $X \subseteq \mathbf{R}^n$ is closed if and only if its complement $\overline{X} := \mathbf{R}^n \setminus X$ is open.

(ii) Intersection of every (finite or infinite) family of closed sets is closed. Union of every (finite or infinite) family of open sets is open.

(iii) Union of finitely many closed sets is closed. Intersection of finitely many open sets is open.

Proof.

(i) Clearly, a point $\bar{x} \in \mathbf{R}^n$ is a limit of a sequence of points from a set $X \subseteq \mathbf{R}^n$ if and only if every centered at $\bar{x}$ ball of positive radius contains points from $X$. Consequently,

- when $X$ is closed and $\bar{x} \notin X$, there is a ball of positive radius centered at $\bar{x}$ and not intersecting $X$, implying that the complement $\overline{X}$ of a closed set $X$ is open;
- when $X$ is open, none of the points from $X$ is the limit of a sequence of the points from the complement $\overline{X}$ of $X$, implying that the limit of every converging sequence of points from $\overline{X}$ belongs to $\overline{X}$, that is, $\overline{X}$ is closed.

(ii) Evident.

(iii) When $X_1, \dots, X_N$ are closed and $\{x^i\}$ is a converging sequence of points from $X = \cup_{i \le N} X_i$, the sequence visits certain $X_j$ infinitely many times, that is, the sequence has a subsequence with all terms from certain $X_j$. The limit of this subsequence (which is the limit of $\{x^i\}$ as well) belongs to $X_j$ since this set is closed, and therefore the limit of $\{x^i\}$ belongs to $X$. Thus, $X$ is closed. Passing from the sets to their complements and invoking (i), we extract from what just have been proved that the intersection of finitely many open sets is open. □

**Fact B.23** The norm $\|x\|_2$ is a real-valued function and it is continuous on $\mathbf{R}^n$ (and thus – on every subset of $\mathbf{R}^n$) (check it!). In fact, *every* norm $\|\cdot\|$ on $\mathbf{R}^n$ is continuous.

Proof. By Fact B.2, we have $|\|x\|_2 - \|y\|_2| \le \|x - y\|_2$, so that $\lim_{t\to\infty} x^t = x$ implies that $0 \le |\|x\|_2 - \|x^t\|_2| \le \|x - x^t\|_2 \to 0$, $t \to \infty$, hence $\lim_{t\to\infty} \|x^t\|_2 = \|x\|_2$, implying the continuity of $\|\cdot\|_2$. This reasoning can be word by word repeated for an arbitrary norm $\|\cdot\|$ and induced by this norm convergence in the role of those coming from $\|\cdot\|_2$. This combines with the the fact that convergence of a sequence from $\mathbf{R}^n$ and its limit are independent of the norm-induced distance on $\mathbf{R}^n$ we use, see section B.1.2, to imply the continuity of every norm on $\mathbf{R}^n$. □

# Appendix C

# Prerequisites from Calculus

## C.1 Differentiable functions on $\mathbf{R}^n$

We next turn to prerequisites from calculus, specifically differentiable functions.

### *C.1.1 The derivative*

The reader definitely is familiar with the notion of *derivative* of a real-valued function $f : \mathbf{R} \to \mathbf{R}$ of real variable $x$:

$$f'(x) := \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}. \quad ^{1)}$$

This definition does not work when we pass from functions of single real variable to functions of several real variables, or, which is the same, to functions with vector arguments. Indeed, in this case the shift in the argument $\Delta x$ should be a vector, and we do not know what does it mean to *divide* by a vector...

A proper way to extend the notion of the derivative to real- and vector-valued functions of vector argument is to realize what in fact the meaning of the derivative is in the univariate case: $f'(x)$ gives us the precise description of *how to approximate $f$ in a neighborhood of $x$ by a linear function*. Specifically, *if $f'(x)$ exists, then the linear function $f'(x)\Delta x$ of $\Delta x$ approximates the change $f(x + \Delta x) - f(x)$ in $f$ up to a remainder which is of higher order as compared to $\Delta x$ as $\Delta x \to 0$:*

$$|f(x + \Delta x) - f(x) - f'(x)\Delta x| \leq \bar{o}(|\Delta x|) \text{ as } \Delta x \to 0.$$

In the above formula, we meet with the notation $\bar{o}(|\Delta x|)$, and here is the explanation of this notation:

[1] Here in what follows, when speaking about limits like the one in the right hand side of the definition of $f'$, we exclude the value $\Delta x = 0$ (at which the quantity we are looking at is undefined). In other words, we are speaking about limit taken w.r.t. "percolated neighborhood" of the origin: by definition, relation $a = \lim_{\Delta x \to 0} g(\Delta x)/\Delta x$ means that for every $\epsilon > 0$ there exists $\delta > 0$ such that $|g(\Delta x)/\Delta x - a| \leq \epsilon$ for all *nonzero* $\Delta x$ satisfying $|\Delta x| \leq \delta$.

$\bar{o}(|\Delta x|)$ *is a common name of all functions* $\phi(\Delta x)$ *of* $\Delta x$ *which are well-defined in a neighborhood of the point* $\Delta x = 0$ *on the axis, vanish at the point* $\Delta x = 0$ *and are such that*

$$\frac{\phi(\Delta x)}{|\Delta x|} \to 0 \text{ as } \Delta x \to 0.$$

For example,

1. $(\Delta x)^2 = \bar{o}(|\Delta x|)$, $\Delta x \to 0$,
2. $|\Delta x|^{1.01} = \bar{o}(|\Delta x|)$, $\Delta x \to 0$,
3. $\sin^2(\Delta x) = \bar{o}(|\Delta x|)$, $\Delta x \to 0$,
4. $\Delta x \neq \bar{o}(|\Delta x|)$, $\Delta x \to 0$.

Later on we shall meet with the notation "$\bar{o}(|\Delta x|^k)$ as $\Delta x \to 0$", where $k$ is a positive integer. The definition of '$\bar{o}(|\Delta x|^k)$ is completely similar to the one for the case of $k = 1$:
$\bar{o}(|\Delta x|^k)$ *is a common name of all functions* $\phi(\Delta x)$ *of* $\Delta x$ *which are well-defined in a neighborhood of the point* $\Delta x = 0$ *on the axis, vanish at the point* $\Delta x = 0$ *and are such that*

$$\frac{\phi(\Delta x)}{|\Delta x|^k} \to 0 \text{ as } \Delta x \to 0.$$

Note that if $f(\cdot)$ is a function defined in a neighborhood of a point $x$ on the axis, then there perhaps are many linear functions $a\Delta x$ of $\Delta x$ which well approximate $f(x + \Delta x) - f(x)$, in the sense that the remainder in the approximation

$$f(x + \Delta x) - f(x) - a\Delta x$$

tends to 0 as $\Delta x \to 0$. Among these approximations, however, there exists *at most one* which approximates $f(x+\Delta x) - f(x)$ "very well" – so that the remainder is $\bar{o}(|\Delta x|)$, and not merely tends to 0 as $\Delta x \to 0$. Indeed, if

$$f(x + \Delta x) - f(x) - a\Delta x = \bar{o}(|\Delta x|),$$

then, dividing both sides by $\Delta x$, we get

$$\frac{f(x + \Delta x) - f(x)}{\Delta x} - a = \frac{\bar{o}(|\Delta x|)}{\Delta x}.$$

By definition of $\bar{o}(\cdot)$, the right hand side in this equality tends to 0 as $\Delta x \to 0$, whence

$$a = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} = f'(x).$$

Thus, *if* a linear function $a\Delta x$ of $\Delta x$ approximates the change $f(x + \Delta x) - f(x)$ in $f$ up to the remainder which is $\bar{o}(|\Delta x|)$ as $\Delta x \to 0$, *then* $a$ is the derivative of $f$ at $x$. You can easily verify that the inverse statement is also true: *if* the derivative of $f$ at $x$ exists, *then* the linear function $f'(x)\Delta x$ of $\Delta x$ approximates the change $f(x + \Delta x) - f(x)$ in $f$ up to the remainder which is $\bar{o}(|\Delta x|)$ as $\Delta x \to 0$.

The advantage of the "$\bar{o}(|\Delta x|)$"-definition of derivative is that it can be naturally extended onto vector-valued functions of vector arguments (by just replacing "axis" with $\mathbf{R}^n$ in the definition of $\bar{o}$) and enlightens the *essence* of the notion of derivative: when it exists, this is exactly *the linear function of* $\Delta x$ *which approximates the change* $f(x+\Delta x)-f(x)$ *in* $f$ *up to a remainder which is* $\bar{o}(|\Delta x|)$. The precise definition is as follows:

**Definition** C.1 [Frechet differentiability] Let $f$ be a function which is well-defined in a neighborhood of a point $x \in \mathbf{R}^n$ and takes values in $\mathbf{R}^m$. We say that $f$ is *differentiable* at $x$, if there exists a linear function $Df(x)[\Delta x]$ of $\Delta x \in \mathbf{R}^n$ taking values in $\mathbf{R}^m$ which approximates the change $f(x+\Delta x) - f(x)$ in $f$ up to a remainder which is $\bar{o}(\|\Delta x\|_2)$, i.e.,

$$\|f(x+\Delta x) - f(x) - Df(x)[\Delta x]\|_2 \leq \bar{o}(\|\Delta x\|_2). \tag{C.1}$$

Equivalently, a function $f$ which is well-defined in a neighborhood of a point $x \in \mathbf{R}^n$ and takes values in $\mathbf{R}^m$ is called *differentiable* at $x$, if there exists a linear function $Df(x)[\Delta x]$ of $\Delta x \in \mathbf{R}^n$ taking values in $\mathbf{R}^m$ such that for every $\epsilon > 0$ there exists $\delta > 0$ satisfying the relation

$$\|\Delta x\| \leq \delta \implies \|f(x+\Delta x) - f(x) - Df(x)[\Delta x]\|_2 \leq \epsilon \|\Delta x\|_2.$$

Note that due to equivalence of norms on finite-dimensional spaces, in this definition the standard Euclidean norms in which we measure change in $f$ and magnitude of $\Delta x$ can be replaced with any other pair of norms on $\mathbf{R}^m$ and $\mathbf{R}^n$ without affecting differentiability and $Df(x)[\cdot]$.

### *C.1.2 Derivative and directional derivatives*

We have defined what it means for a function $f : \mathbf{R}^n \to \mathbf{R}^m$ to be differentiable at a point $x$, but have not stated yet what the *derivative* is. The reader may guess that the derivative is exactly the "linear function $Df(x)[\Delta x]$ of $\Delta x \in \mathbf{R}^n$ taking values in $\mathbf{R}^m$ which approximates the change $f(x+\Delta x) - f(x)$ in $f$ up to a remainder which is less than or equal to $\bar{o}(\|\Delta x\|_2)$" participating in the definition of differentiability. While this guess is correct, we cannot merely call the entity participating in the definition the derivative – why do we know that this entity is unique? Perhaps there are many different linear functions of $\Delta x$ approximating the change in $f$ up to a remainder which is $\bar{o}(\|\Delta x\|_2)$. In fact there is no more than a single linear function with this property due to the following observation.

**Proposition** C.2 *Let* $f$ *be differentiable at* $x$, *and* $Df(x)[\Delta x]$ *be a linear function participating in the definition of differentiability. Then,*

$$Df(x)[\Delta x] = \lim_{t \to +0} \frac{f(x+t\Delta x) - f(x)}{t}, \qquad \forall \Delta x \in \mathbf{R}^n. \tag{C.2}$$

*In particular,* the derivative $Df(x)[\cdot]$ *is uniquely defined by* $f$ *and* $x$.

**Proof.** For all $\Delta x \in \mathbf{R}^n$ and for any $t > 0$, we have

$$\begin{aligned}
&\|f(x+t\Delta x) - f(x) - Df(x)[t\Delta x]\|_2 \leq \bar{o}(\|t\Delta x\|_2)\\
\Longrightarrow\ &\|\frac{f(x+t\Delta x)-f(x)}{t} - \frac{Df(x)[t\Delta x]}{t}\|_2 \leq \frac{\bar{o}(\|t\Delta x\|_2)}{t}\\
\Longleftrightarrow\ &\left\|\frac{f(x+t\Delta x)-f(x)}{t} - Df(x)[\Delta x]\right\|_2 \leq \frac{\bar{o}(\|t\Delta x\|_2)}{t}\\
\Longrightarrow\ &Df(x)[\Delta x] = \lim_{t\to+0} \frac{f(x+t\Delta x)-f(x)}{t},
\end{aligned}$$

where the second line is obtained by dividing with $t > 0$, the third line follows since $Df(x)[\cdot]$ is linear, and the last line is obtained by passing to limit as $t \to +0$ and noting that $\frac{\bar{o}(\|t\Delta x\|_2)}{t} \to 0$ as $t \to +0$. □

We conclude this section with two important remarks as follows:

1. The right hand side limit in (C.2) is an important entity called the *directional derivative of $f$ taken at $x$ along (a direction)* $\Delta x$; note that this quantity is defined in the "purely univariate" fashion – by dividing the change in $f$ by the magnitude of a shift in a direction $\Delta x$ and passing to limit as the magnitude of the shift approaches 0. Relation (C.2) states that the derivative, if it exists, is, at every $\Delta x$, nothing but the directional derivative of $f$ taken at $x$ along $\Delta x$. Note, however, that differentiability is much more than the existence of directional derivatives along all directions $\Delta x$. In particular, differentiability requires also *the directional derivatives to be "well-organized" – to depend* linearly *on the direction* $\Delta x$. It is easily seen that just mere existence of directional derivatives does not imply their "good organization." For example, the Euclidean norm

$$f(x) = \|x\|_2$$

at $x = 0$ possesses directional derivatives along all directions:

$$\lim_{t\to+0} \frac{f(0+t\Delta x)-f(0)}{t} = \|\Delta x\|_2.$$

These derivatives, however, depend *non-linearly* on $\Delta x$, so that the Euclidean norm is *not* differentiable at the origin (although is differentiable everywhere outside the origin, but this is another story).

2. It should be stressed that the derivative, if it exists, is what it is: *a linear function of* $\Delta x \in \mathbf{R}^n$ *taking values in* $\mathbf{R}^m$. As we shall see in a while, we can *represent* this function by something "tractable," like a vector or a matrix, and we can understand how to compute such a representation. However, a careful reader should bear in mind that *a* representation is not exactly the same as *the* represented entity. Sometimes the difference between derivatives and the entities which represent them is reflected in the terminology: what we call the *derivative*, is also called the *differential*, while the word "derivative" is reserved for the vector/matrix representing the differential.

### C.1.3 Representations of the derivative

By definition, the derivative of a mapping $f : \mathbf{R}^n \to \mathbf{R}^m$ at a point $x$ is a linear function $Df(x)[\Delta x]$ taking values in $\mathbf{R}^m$. How could we represent such a function?

**Case of $m = 1$: The gradient.** Let us start with real-valued functions (i.e., with the case of $m = 1$); in this case the derivative is a *linear* real-valued function on $\mathbf{R}^n$. As we remember, the standard Euclidean structure on $\mathbf{R}^n$ allows to represent every linear function on $\mathbf{R}^n$ as the inner product of the argument with certain fixed vector. In particular, the derivative $Df(x)[\Delta x]$ of a scalar function can be represented as

$$Df(x)[\Delta x] = [\text{vector}]^\top \Delta x;$$

what is denoted with "vector" in this relation, is called the *gradient* of $f$ at $x$ and is denoted by $\nabla f(x)$:

$$Df(x)[\Delta x] = (\nabla f(x))^\top \Delta x. \tag{C.3}$$

How to compute the gradient? The answer is given by (C.2). Indeed, let us look what (C.3) and (C.2) say when $\Delta x$ is the $i$-th standard basis vector. According to (C.3), $Df(x)[e_i]$ is the $i$-th coordinate of the vector $\nabla f(x)$. Then, using (C.2), we arrive at

$$\left.\begin{array}{rcl} Df(x)[e_i] & = & \lim\limits_{t\to+0} \frac{f(x+te_i)-f(x)}{t}, \\ Df(x)[e_i] & = & -Df(x)[-e_i] = -\lim\limits_{t\to+0} \frac{f(x-te_i)-f(x)}{t} = \lim\limits_{t\to-0} \frac{f(x+te_i)-f(x)}{t} \end{array}\right\}$$
$$\Longrightarrow\ Df(x)[e_i] = \tfrac{\partial f(x)}{\partial x_i}.$$

Thus,

> *If a real-valued function $f$ is differentiable at $x$, then the first-order partial derivatives of $f$ at $x$ exist, and the gradient of $f$ at $x$ is just the vector with the coordinates which are the first-order partial derivatives of $f$ taken at $x$:*
>
> $$\nabla f(x) := \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$
>
> *The derivative of $f$, taken at $x$, is the linear function of $\Delta x$ given by*
>
> $$Df(x)[\Delta x] = (\nabla f(x))^\top \Delta x = \sum_{i=1}^n \frac{\partial f(x)}{\partial x_i} (\Delta x)_i.$$

**General case: The Jacobian.** Now consider $f : \mathbf{R}^n \to \mathbf{R}^m$ with $m \geq 1$. In this case, $Df(x)[\Delta x]$, regarded as a function of $\Delta x$, is a linear mapping from $\mathbf{R}^n$ to $\mathbf{R}^m$. Recall that the standard way to represent a linear mapping from $\mathbf{R}^n$ to $\mathbf{R}^m$ is to represent it as the multiplication by an $m \times n$ matrix:

$$Df(x)[\Delta x] = [m \times n \text{ matrix}] \cdot \Delta x. \tag{C.4}$$

What is denoted by "matrix" in (C.4), is called *the Jacobian* of $f$ at $x$ and is denoted by $f'(x)$. How to compute the entries of the Jacobian? Once again the answer is readily given by (C.2). Indeed, on one hand, we have

$$Df(x)[\Delta x] = f'(x)\Delta x, \tag{C.5}$$

where

$$[Df(x)[e_j]]_i = [f'(x)]_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n.$$

On the other hand, by denoting

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix},$$

the same computation as in the case of gradient demonstrates that

$$[Df(x)[e_j]]_i = \frac{\partial f_i(x)}{\partial x_j}.$$

Thus, we arrive at the following conclusion:

> *If a vector-valued function $f(x) = [f_1(x); \ldots; f_m(x)]$ is differentiable at $x$, then the first-order partial derivatives of all $f_i$ at $x$ exist, and the Jacobian of $f$ at $x$ is just the $m \times n$ matrix with the entries $\left[\frac{\partial f_i(x)}{\partial x_j}\right]_{i,j}$ (so that the rows in the Jacobian are $[\nabla f_1(x)]^\top$,..., $[\nabla f_m(x)]^\top$). The derivative of $f$, taken at $x$, is the linear vector-valued function of $\Delta x$ given by*
>
> $$Df(x)[\Delta x] = f'(x)\Delta x = \begin{bmatrix} [\nabla f_1(x)]^\top \Delta x \\ \vdots \\ [\nabla f_m(x)]^\top \Delta x \end{bmatrix}.$$

**Remark** C.3 Note that for a real-valued function $f : \mathbf{R}^n \to \mathbf{R}$ we have defined both the gradient $\nabla f(x)$ and the Jacobian $f'(x)$. These two entities are "nearly the same," but not exactly the same. The Jacobian is a vector-row and the gradient is a vector-column; these two are linked by the relation

$$f'(x) = (\nabla f(x))^\top.$$

Of course, both these representations of the derivative of $f$ yield the same linear approximation of the change in $f$:

$$Df(x)[\Delta x] = (\nabla f(x))^\top \Delta x = f'(x)\Delta x.$$

### *C.1.4 Existence of the derivative*

We have seen that the existence of the derivative of $f$ at a point implies the existence of the first-order partial derivatives of the components $(f_1, \ldots, f_m)$ of

$f$. The reverse statement is not exactly true. In particular, the existence of all first-order partial derivatives $\frac{\partial f_i(x)}{\partial x_j}$ for all $i, j$ not necessarily implies the existence of the derivative. In fact, we need a bit more, which is described next.

**Theorem** C.4 [Sufficient condition for differentiability] Given a mapping $f = (f_1, \ldots, f_m) : \mathbf{R}^n \to \mathbf{R}^m$, $f$ is differentiable at the point $\bar{x} \in \mathbf{R}^n$ if all of the following conditions holds:

(i) the mapping $f$ is well-defined in a neighborhood $U$ of the point $\bar{x} \in \mathbf{R}^n$;

(ii) the first-order partial derivatives of the components $f_i$ of $f$ exist everywhere in $U$; and

(iii) the first-order partial derivatives of the components $f_i$ of $f$ are continuous at the point $\bar{x}$.

### *C.1.5 Calculus of derivatives*

The following elementary rules for the calculus of derivatives are useful to know.

**Theorem** C.5

(i) [Differentiability and linear operations] Let $f_1(x)$, $f_2(x)$ be mappings defined in a neighborhood of a point $\bar{x} \in \mathbf{R}^n$ and taking values in $\mathbf{R}^m$, and $\lambda_1(x), \lambda_2(x)$ be real-valued functions defined in a neighborhood of $\bar{x}$. Whenever $f_1, f_2, \lambda_1, \lambda_2$ are differentiable at $\bar{x}$, so is the function $f(x) := \lambda_1(x)f_1(x) + \lambda_2(x)f_2(x)$, and its derivative at $\bar{x}$ is given by

$$\begin{aligned} Df(\bar{x})[\Delta x] &= [D\lambda_1(\bar{x})[\Delta x]]f_1(\bar{x}) + \lambda_1(\bar{x})Df_1(\bar{x})[\Delta x] \\ &\quad + [D\lambda_2(\bar{x})[\Delta x]]f_2(\bar{x}) + \lambda_2(\bar{x})Df_2(\bar{x})[\Delta x], \\ \Longrightarrow \ f'(\bar{x}) &= f_1(\bar{x})[\nabla\lambda_1(\bar{x})]^\top + \lambda_1(\bar{x})f_1'(\bar{x}) \\ &\quad + f_2(\bar{x})[\nabla\lambda_2(\bar{x})]^\top + \lambda_2(\bar{x})f_2'(\bar{x}). \end{aligned}$$

(ii) [Chain rule] Let a mapping $f : \mathbf{R}^n \to \mathbf{R}^m$ be differentiable at $\bar{x}$, and a mapping $g : \mathbf{R}^m \to \mathbf{R}^k$ be differentiable at $\bar{y} := f(\bar{x})$. Then, the superposition function given by $h(x) = g(f(x))$ is differentiable at $\bar{x}$, and its derivative at $\bar{x}$ is given by

$$\begin{aligned} Dh(\bar{x})[\Delta x] &= Dg(\bar{y})[Df(\bar{x})[\Delta x]], \\ \Longrightarrow \ h'(\bar{x}) &= g'(\bar{y})f'(\bar{x}). \end{aligned}$$

If the outer function $g$ is real-valued, then the latter formula implies that

$$\nabla h(\bar{x}) = [\nabla g(\bar{y})]^\top f'(\bar{x})$$

(recall that for a real-valued function $\phi$, $\phi' = (\nabla\phi)^\top$).

### C.1.6 Computing the derivative

Representations of the derivative via first-order partial derivatives normally allow us to compute it by the standard Calculus rules, in a completely mechanical fashion, not thinking at all of *what* we are computing. The examples to follow (especially Example C.8) demonstrate that it often makes sense to bear in mind *what* the derivative is; this sometimes yields the result much faster than blindly implementing Calculus rules.

**Example** C.6 (Gradient of an affine function) An *affine* function

$$f(x) = a + \sum_{i=1}^{n} g_i x_i \equiv a + g^\top x : \mathbf{R}^n \to \mathbf{R}$$

is differentiable at every point (Theorem C.4) and its gradient, of course, equals to $g$:

$$\begin{aligned}(\nabla f(x))^\top \Delta x &= \lim_{t\to+0} t^{-1}\left(f(x+t\Delta x) - f(x)\right) && \text{[by (C.2)]}\\ &= \lim_{t\to+0} t^{-1}(tg^\top \Delta x) && \text{[by plugging the definition of } f\text{].}\end{aligned}$$

Hence, we arrive at

$$\boxed{\nabla(a + g^\top x) = g.}$$

**Example** C.7 (Gradient of a quadratic form) For now, let us define a homogeneous quadratic form on $\mathbf{R}^n$ as a function

$$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_i x_j = x^\top Ax,$$

where $A$ is an $n \times n$ matrix. Note that the matrices $A$ and $A^\top$ define the same quadratic form, and therefore the *symmetric* matrix $B := \frac{1}{2}(A+A^\top)$ also produces the same quadratic form as $A$ and $A^\top$. Thus, we always may assume (and do assume from now on) that the matrix $A$ producing the quadratic form in question is symmetric.

A quadratic form is a simple polynomial and as such is differentiable at every point (Theorem C.4). What is the gradient of $f$ at a point $x$? Here is the computation:

$$\begin{aligned}&(\nabla f(x))^\top \Delta x\\ &= Df(x)[\Delta x]\\ &= \lim_{t\to+0} t^{-1}\left((x+t\Delta x)^\top A(x+t\Delta x) - x^\top Ax\right)\\ &= \lim_{t\to+0} t^{-1}\Big(x^\top Ax + t(\Delta x)^\top Ax + tx^\top A\Delta x + t^2(\Delta x)^\top A\Delta x - x^\top Ax\Big)\\ &= \lim_{t\to+0} t^{-1}\Big(2t(Ax)^\top\Delta x + t^2(\Delta x)^\top A\Delta x\Big)\\ &= 2(Ax)^\top\Delta x,\end{aligned}$$

where the second equality follows from (C.2), the third equality is obtained by opening the parenthesis, and in the last line we used that $A$ is symmetric.

Hence, we conclude that for a symmetric matrix $A$, we have

$$\boxed{\nabla(x^\top Ax) = 2Ax.}$$

**Example** C.8 (Derivative of $X^{-1}$ on the domain of nonsingular $n \times n$ matrices) Define the mapping $F(X) := X^{-1}$ on the open set of nonsingular $n \times n$ matrices. Suppose $X \in \mathbf{R}^{n\times n}$ is nonsingular. Then, for any $\Delta X \in \mathbf{R}^{n\times n}$ we have

$$\begin{aligned} DF(X)[\Delta X] &= \lim_{t\to+0} t^{-1}\left((X + t\Delta X)^{-1} - X^{-1}\right) \\ &= \lim_{t\to+0} t^{-1}\left((X(I + tX^{-1}\Delta X))^{-1} - X^{-1}\right) \\ &= \lim_{t\to+0} t^{-1}\big((I + tX^{-1}\Delta X)^{-1}X^{-1} - X^{-1}\big) \end{aligned}$$

Thus, by defining $Y := X^{-1}\Delta X$, for all $\Delta X \in \mathbf{R}^{n\times n}$ we arrive at

$$\begin{aligned} DF(X)[\Delta X] &= \left(\lim_{t\to+0} t^{-1}\left((I + tY)^{-1} - I\right)\right) X^{-1} \\ &= \left(\lim_{t\to+0} t^{-1}\left(I - (I + tY)\right)(I + tY)^{-1}\right) X^{-1} \\ &= \left(\lim_{t\to+0} (-Y(I + tY)^{-1})\right) X^{-1} \\ &= -YX^{-1} \\ &= -X^{-1}\Delta XX^{-1}. \end{aligned}$$

Therefore, we arrive at the important relation

$$\boxed{D(X^{-1})[\Delta X] = -X^{-1}\Delta XX^{-1}.}$$

(cf., the derivative of the univariate function $x^{-1}$ at $x \neq 0$ is $-x^{-2}$).

**Example** C.9 (Derivative of the log-det barrier) The *log-det barrier* is given by

$$F(X) = \ln \operatorname{Det}(X),$$

where $X$ is an $n \times n$ matrix (or, if you prefer, $n^2$-dimensional vector). The log-det barrier plays an extremely important role in modern optimization. In this example, we will compute its derivative.

Note that $F(X)$ is well-defined and differentiable in a neighborhood of every point $\bar{X}$ with positive determinant. (Indeed, $\operatorname{Det}(X)$ is a polynomial of the entries of $X$ and thus it is everywhere continuous and differentiable with continuous partial derivatives, while the function $\ln(t)$ is continuous and differentiable on the positive ray. Then, by Theorems B.24.ii and C.5.ii, $F$ is differentiable at every $X$ such that $\operatorname{Det}(X) > 0$). While the computation of the derivative of $F$ by the standard techniques will not be very pleasant, we next illustrate that this can be done easily by resorting to the fundamental definition of the derivative.

Let us consider a point $\bar{X}$ such that $\mathrm{Det}(\bar{X}) > 0$, and define $G(X) := \mathrm{Det}(X)$. Then, we have

$$\begin{aligned} DF(\bar{X})[\Delta X] &= D\ln(G(\bar{X}))\,\big(DG(\bar{X})[\Delta X]\big) \\ &= (G(\bar{X}))^{-1}DG(\bar{X})[\Delta X] \\ &= (\mathrm{Det}(\bar{X}))^{-1}\lim_{t\to+0} t^{-1}\big(\mathrm{Det}(\bar{X}+t\Delta X)-\mathrm{Det}(\bar{X})\big) \\ &= (\mathrm{Det}(\bar{X}))^{-1}\lim_{t\to+0} t^{-1}\big(\mathrm{Det}\big(\bar{X}(I+t\bar{X}^{-1}\Delta X)\big)-\mathrm{Det}(\bar{X})\big) \\ &= (\mathrm{Det}(\bar{X}))^{-1}\lim_{t\to+0} t^{-1}\big(\mathrm{Det}(\bar{X})\big(\mathrm{Det}(I+t\bar{X}^{-1}\Delta X\big)-1)\big) \\ &= \lim_{t\to+0} t^{-1}\big(\mathrm{Det}(I+t\bar{X}^{-1}\Delta X)-1\big) \\ &= \mathrm{Tr}(\bar{X}^{-1}\Delta X) = \sum_{i=1}^{n}\sum_{j=1}^{n}(\bar{X}^{-1})_{ji}\,(\Delta X)_{ij}, \end{aligned}$$

where the first equality follows from the chain rule, the second one from the fact that $\ln'(x) = x^{-1}$ for any $x > 0$, the third one from the definition of $G$ and (C.2), the fifth one follows from the relation $\mathrm{Det}(AB) = \mathrm{Det}(A)\mathrm{Det}(B)$ for every $A, B$ of appropriate size, and the second to last equality, i.e.,

$$\lim_{t\to+0} t^{-1}(\mathrm{Det}(I+tA)-1) = \mathrm{Tr}(A) \equiv \sum_{i=1}^{n} A_{ii}, \tag{C.6}$$

is immediately given by recalling what $\mathrm{Det}(I + tA)$ is. Note that $\mathrm{Det}(I + tA)$ is a polynomial of $t$ which is the sum of products, taken along all diagonals of an $n \times n$ matrix and assigned certain signs, of the entries of $I + tA$. At every one of these diagonals, except for the main one, there are at least two cells with the entries proportional to $t$, so that the corresponding products do not contribute to the constant and the linear in $t$ terms in $\mathrm{Det}(I + tA)$ and thus do not affect the limit in (C.6). The only product which does contribute to the linear and the constant terms in $\mathrm{Det}(I + tA)$ is the product $(1 + tA_{11})(1 + tA_{22}) \dots (1 + tA_{nn})$ coming from the main diagonal. Moreover, it is clear that in this product the constant term is 1, and the linear in $t$ term is $t(A_{11} + \dots + A_{nn})$, and thus (C.6) follows.

## C.2 Higher order derivatives

Let $f : \mathbf{R}^n \to \mathbf{R}^m$ be a mapping which is well-defined and differentiable at every point $x$ from an open set $U$. The Jacobian of this mapping $J(x)$ is a mapping from $\mathbf{R}^n$ to the space $\mathbf{R}^{m\times n}$ matrices, i.e., it is a mapping taking values in certain $\mathbf{R}^M$ ($M = mn$). The derivative of this mapping $J(x)$, if it exists, is called the *second derivative* of $f$, which again is a mapping from $\mathbf{R}^n$ to certain $\mathbf{R}^M$ and as such can be differentiable, and so on, so that we can speak about the second, the third, ... derivatives of a vector-valued function of vector argument. A *sufficient* condition for the existence of $k$ derivatives of $f$ in $U$ is that $f$ is $\mathrm{C}^k$ in $U$, i.e., that all partial derivatives of $f$ of orders $\leq k$ exist and are continuous everywhere in $U$ (cf. Theorem C.4).

The preceding description explains what it means that $f$ has $k$ derivatives in $U$. Note, however, that according to this description, highest order derivatives at a point $x$ are just long vectors; say, the second-order derivative of a scalar function $f$ of 2 variables is the Jacobian of the mapping $x \mapsto f'(x) : \mathbf{R}^2 \to \mathbf{R}^2$, i.e., a mapping from $\mathbf{R}^2$ to $\mathbf{R}^{2\times 2} = \mathbf{R}^4$; the third-order derivative of $f$ is therefore the Jacobian of a mapping from $\mathbf{R}^2$ to $\mathbf{R}^4$, i.e., a mapping from $\mathbf{R}^2$ to $\mathbf{R}^{4\times 2} = \mathbf{R}^8$, and so on. The question which should be addressed now is: *What is a natural and transparent way to represent the highest order derivatives?*

The answer is as follows:

> $(*)$ *Let $f : \mathbf{R}^n \to \mathbf{R}^m$ be $\mathrm{C}^k$ on an open set $U \subseteq \mathbf{R}^n$. The derivative of order $\ell \leq k$ of $f$, taken at a point $x \in U$, can be naturally identified with a function*
>
> $$D^\ell f(x)[\Delta x^1, \Delta x^2, \ldots, \Delta x^\ell]$$
>
> *of $\ell$ vector arguments $\Delta x^i \in \mathbf{R}^n$, $i = 1, \ldots, \ell$, and taking values in $\mathbf{R}^m$. This function is linear in every one of the arguments $\Delta x^i$, the other arguments being fixed, and is symmetric with respect to permutation of arguments $\Delta x^1, \ldots, \Delta x^\ell$.*
> *In terms of $f$, the quantity $D^\ell f(x)[\Delta x^1, \Delta x^2, \ldots, \Delta x^\ell]$ (full name: "the $\ell$-th derivative (or differential) of $f$ taken at a point $x$ along the directions $\Delta x^1, \ldots, \Delta x^\ell$") is given by*
>
> $$\begin{array}{l} D^\ell f(x)[\Delta x^1, \Delta x^2, \ldots, \Delta x^\ell] \\ = \frac{\partial^\ell}{\partial t_\ell \partial t_{\ell-1} \ldots \partial t_1}\Big|_{t_1=\ldots=t_\ell=0} f(x + t_1\Delta x^1 + t_2\Delta x^2 + \ldots + t_\ell \Delta x^\ell). \end{array} \quad \text{(C.7)}$$

The explanation to our claims is as follows. Let $f : \mathbf{R}^n \to \mathbf{R}^m$ be $\mathrm{C}^k$ on an open set $U \subseteq \mathbf{R}^n$.

1. When $\ell = 1$, $(*)$ states that the first-order derivative of $f$, taken at $x$, is a linear function $Df(x)[\Delta x^1]$ of $\Delta x^1 \in \mathbf{R}^n$, taking values in $\mathbf{R}^m$, and that the value of this function at every $\Delta x^1$ is given by the relation

   $$Df(x)[\Delta x^1] = \frac{\partial}{\partial t_1}\Big|_{t_1=0} f(x + t_1 \Delta x^1) \qquad \text{(C.8)}$$

   (cf. (C.2)), which is in complete accordance with what we already know about the derivative.
2. To understand what the second derivative is, let us take the first derivative $Df(x)[\Delta x^1]$, *let us temporarily fix somehow the argument* $\Delta x^1$ and treat the derivative as a function of $x$. As a function of $x$, $\Delta x^1$ being fixed, the quantity $Df(x)[\Delta x^1]$ is again a mapping which maps $U$ into $\mathbf{R}^m$ and is differentiable by Theorem C.4 (provided, of course, that $k \geq 2$). The derivative of this mapping will be a certain linear function of $\Delta x \equiv \Delta x^2 \in \mathbf{R}^n$, depending on $x$ as on a parameter; and of course it depends on $\Delta x^1$ as on a parameter as well. Thus,

the derivative of $Df(x)[\Delta x^1]$ in $x$ is a certain function

$$D^2 f(x)[\Delta x^1, \Delta x^2]$$

of $x \in U$ and $\Delta x^1, \Delta x^2 \in \mathbf{R}^n$ and taking values in $\mathbf{R}^m$. What we know about this function is that it is linear in $\Delta x^2$. In fact, it is also linear in $\Delta x^1$, since it is the derivative in $x$ of certain function (namely, of $Df(x)[\Delta x^1]$) *linearly depending on the parameter* $\Delta x^1$, so that the derivative of the function *in* $x$ is linear in the parameter $\Delta x^1$ as well (differentiation is a linear operation with respect to a function we are differentiating: summing up functions and multiplying them by real constants, we sum up, respectively, multiply by the same constants, the derivatives). Thus, $D^2 f(x)[\Delta x^1, \Delta x^2]$ is linear in $\Delta x^1$ when $x$ and $\Delta x^2$ are fixed, and is linear in $\Delta x^2$ when $x$ and $\Delta x^1$ are fixed. Moreover, we have

$$\begin{array}{rcll} D^2 f(x)[\Delta x^1, \Delta x^2] & = & \frac{\partial}{\partial t_2}\Big|_{t_2=0} Df(x + t_2\Delta x^2)[\Delta x^1] & \text{[cf. (C.8)]} \\ & = & \frac{\partial}{\partial t_2}\Big|_{t_2=0} \frac{\partial}{\partial t_1}\Big|_{t_1=0} f(x + t_2\Delta x^2 + t_1\Delta x^1) & \text{[by (C.8)]} \\ & = & \frac{\partial^2}{\partial t_2 \partial t_1}\Bigg|_{t_1=t_2=0} f(x + t_1\Delta x^1 + t_2\Delta x^2) & \end{array} \tag{C.9}$$

as claimed in (C.7) for $\ell = 2$. The only piece of information about the second derivative which is contained in (∗) and is not justified yet is that $D^2 f(x)[\Delta x^1, \Delta x^2]$ is symmetric in $\Delta x^1, \Delta x^2$. This fact is readily given by the representation (C.7), since, as it was proven in Calculus, if a function $\phi$ possesses *continuous* partial derivatives of orders $\leq \ell$ in a neighborhood of a point, then these derivatives in this neighborhood are independent of the order in which they are taken. Then, it follows that

$$\begin{aligned} D^2 f(x)[\Delta x^1, \Delta x^2] &= \frac{\partial^2}{\partial t_2 \partial t_1}\Bigg|_{t_1=t_2=0} \underbrace{f(x + t_1\Delta x^1 + t_2\Delta x^2)}_{:=\phi(t_1,t_2)} && \text{[by (C.9)]} \\ &= \frac{\partial^2}{\partial t_1 \partial t_2}\Bigg|_{t_1=t_2=0} \phi(t_1, t_2) \\ &= \frac{\partial^2}{\partial t_1 \partial t_2}\Bigg|_{t_1=t_2=0} f(x + t_2\Delta x^2 + t_1\Delta x^1) \\ &= D^2 f(x)[\Delta x^2, \Delta x^1] && \text{[once again by (C.9)]} \end{aligned}$$

3. Now it is clear how to proceed: to define $D^3 f(x)[\Delta x^1, \Delta x^2, \Delta x^3]$, we fix in the second-order derivative $D^2 f(x)[\Delta x^1, \Delta x^2]$ the arguments $\Delta x^1, \Delta x^2$ and treat it as a function of $x$ only, thus arriving at a mapping which maps $U$ into $\mathbf{R}^m$ and depends on $\Delta x^1, \Delta x^2$ as on parameters (linearly in every one of them). Differentiating the resulting mapping in $x$, we arrive at a function $D^3 f(x)[\Delta x^1, \Delta x^2, \Delta x^3]$ which by construction is linear in every one of the arguments $\Delta x^1$, $\Delta x^2$, $\Delta x^3$ and satisfies (C.7); the latter relation, due to the Calculus result on the symmetry of partial derivatives, implies that $D^3 f(x)[\Delta x^1, \Delta x^2, \Delta x^3]$ is symmetric in $\Delta x^1, \Delta x^2, \Delta x^3$. After we have at our

disposal the third derivative $D^3 f$, we can build from it in the already explained fashion the fourth derivative, and so on, until $k$-th derivative is defined.

**Remark** C.10 Since $D^\ell f(x)[\Delta x^1, \dots, \Delta x^\ell]$ is linear in every one of $\Delta x^i$, we can expand the derivative in a multiple sum:

$$\Delta x^i = \sum_{j=1}^{n} \Delta x^i_j \, e_j,$$

$$\Longrightarrow\ D^\ell f(x)[\Delta x^1, \dots, \Delta x^\ell] = D^\ell f(x) \left[ \sum_{j_1=1}^{n} \Delta x^1_{j_1} e_{j_1}, \dots, \sum_{j_\ell=1}^{n} \Delta x^\ell_{j_\ell} e_{j_\ell} \right] \qquad \text{(C.10)}$$
$$= \sum_{1 \le j_1, \dots, j_\ell \le n} D^\ell f(x) \, [e_{j_1}, \dots, e_{j_\ell}] \, \Delta x^1_{j_1} \dots \Delta x^\ell_{j_\ell}.$$

What is the origin of the coefficients $D^\ell f(x)[e_{j_1}, \dots, e_{j_\ell}]$? According to (C.7), one has

$$D^\ell f(x)[e_{j_1}, \dots, e_{j_\ell}] = \left.\frac{\partial^\ell}{\partial t_\ell \partial t_{\ell-1} \dots \partial t_1}\right|_{t_1 = \dots = t_\ell = 0} f(x + t_1 e_{j_1} + t_2 e_{j_2} + \dots + t_\ell e_{j_\ell})$$
$$= \frac{\partial^\ell}{\partial x_{j_\ell} \partial x_{j_{\ell-1}} \dots \partial x_{j_1}} f(x).$$

so that the coefficients in (C.10) are nothing but the partial derivatives, of order $\ell$, of $f$.

**Remark** C.11 An important particular case of the relation (C.7) is the one when the vectors $\Delta x^1, \Delta x^2, \dots, \Delta x^\ell$ are all the same. Let $d := \Delta x^1 = \dots = \Delta x^\ell$. According to (C.7), we have

$$D^\ell f(x)[d, d, \dots, d] = \left.\frac{\partial^\ell}{\partial t_\ell \partial t_{\ell-1} \dots \partial t_1}\right|_{t_1 = \dots = t_\ell = 0} f(x + t_1 d + t_2 d + \dots + t_\ell d).$$

This relation can be interpreted as follows: consider the function

$$\phi(t) := f(x + td)$$

of a real variable $t$. Then, (check it!)

$$\phi^{(\ell)}(0) = \left.\frac{\partial^\ell}{\partial t_\ell \partial t_{\ell-1} \dots \partial t_1}\right|_{t_1 = \dots = t_\ell = 0} f(x + t_1 d + t_2 d + \dots + t_\ell d) = D^\ell f(x)[d, \dots, d].$$

In fact, $D^\ell f(x)[d, \dots, d]$ has a special name. It is called *$\ell$-th directional derivative of $f$ taken at $x$ along the direction $d$.* To define this quantity, we pass from function $f$ of several variables to the univariate function $\phi(t) := f(x + td)$ , which restricts $f$ onto the line passing through $x$ and directed by $d$, and then take the "usual" derivative of order $\ell$ of the resulting function of single real variable $t$ at the point $t = 0$ (which corresponds to the point $x$ of our line).

**Representation of higher order derivatives.** $k$-th order derivative $D^k f(x)[\cdot, \dots, \cdot]$ of a C$^k$ function $f : \mathbf{R}^n \to \mathbf{R}^m$ is what it is – it is a symmetric $k$-linear mapping on $\mathbf{R}^n$ taking values in $\mathbf{R}^m$ and depending on $x$ as a parameter. Choosing somehow coordinates in $\mathbf{R}^n$, we can represent such a mapping in the form

$$D^k f(x)[\Delta x^1, \ldots, \Delta x^k] = \sum_{1 \le i_1, \ldots, i_k \le n} \frac{\partial^k f(x)}{\partial x_{i_k} \partial x_{i_{k-1}} \ldots \partial x_{i_1}} (\Delta x^1)_{i_1} \ldots (\Delta x^k)_{i_k}.$$

We may say that the derivative can be represented by $k$-index collection of $m$-dimensional vectors $\frac{\partial^k f(x)}{\partial x_{i_k} \partial x_{i_{k-1}} \ldots \partial x_{i_1}}$. This collection, however, is a difficult-to-handle entity, so that such a representation does not help. There is, however, a case when the collection becomes an entity we know to handle; this is the case of the second-order derivative of a scalar function ($k = 2, m = 1$). In this case, the collection in question is just a symmetric matrix

$$H(x) := \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right]_{1 \le i,j \le n}.$$

This matrix is called the *Hessian* of $f$ at $x$. Note that

$$D^2 f(x)[\Delta x^1, \Delta x^2] = [\Delta x^1]^\top H(x) \Delta x^2.$$

### *C.2.1 Calculus of C$^k$ mappings*

The calculus of C$^k$ mappings can be summarized as follows:

**Theorem** C.12 1 Let $U$ be an open set in $\mathbf{R}^n$, $f_1(\cdot), f_2(\cdot) : \mathbf{R}^n \to \mathbf{R}^m$ be C$^k$ in $U$, and let real-valued functions $\lambda_1(\cdot), \lambda_2(\cdot)$ be C$^k$ in $U$. Then, the function

$$f(x) = \lambda_1(x) f_1(x) + \lambda_2(x) f_2(x)$$

is C$^k$ in $U$.

2 Let $U$ be an open set in $\mathbf{R}^n$, $V$ be an open set in $\mathbf{R}^m$, let a mapping $f : \mathbf{R}^n \to \mathbf{R}^m$ be such that it is C$^k$ in $U$ and $f(x) \in V$ for $x \in U$, and, finally, let a mapping $g : \mathbf{R}^m \to \mathbf{R}^p$ be C$^k$ in $V$. Then, the superposition

$$h(x) = g(f(x))$$

is C$^k$ in $U$.

**Remark** C.13 For higher order derivatives, in contrast to the first-order ones, there is no simple "chain rule" for computing the derivative of superposition. For example, the second-order derivative of the superposition $h(x) = g(f(x))$ of two C$^2$-mappings is given by the formula

$$\begin{aligned} &Dh(x)[\Delta x^1, \Delta x^2] \\ &= Dg(f(x))[D^2 f(x)[\Delta x^1, \Delta x^2]] + D^2 g(x)[Df(x)[\Delta x^1], Df(x)[\Delta x^2]] \end{aligned}$$

(check it!). We see that both the first- and the second-order derivatives of $f$ and $g$ contribute to the second-order derivative of the superposition $h$.

The only case when there does exist a simple formula for high order derivatives of a superposition is the case *when the inner function is affine*: if $f(x) = Ax + b$ and $h(x) = g(f(x)) = g(Ax + b)$ with a $\mathrm{C}^\ell$ mapping $g$, then

$$D^\ell h(x)[\Delta x^1, \ldots, \Delta x^\ell] = D^\ell g(Ax + b)[A\Delta x^1, \ldots, A\Delta x^\ell]. \tag{C.11}$$

### *C.2.2 Examples of higher-order derivatives*

We next go over some examples of computing higher-order derivatives.

**Example** C.14 (Second-order derivative of an affine function) Consider $f(x) = a + b^\top x$. Then, of course, its second-order derivative is identically zero. Indeed, as we have seen in Example C.6,

$$Df(x)[\Delta x^1] = b^\top \Delta x^1$$

is independent of $x$, and therefore the derivative of $Df(x)[\Delta x^1]$ in $x$, which should give us the second derivative $D^2 f(x)[\Delta x^1, \Delta x^2]$, is zero. Clearly, the third, the fourth, etc., derivatives of an affine function are zero as well.

**Example** C.15 (Second-order derivative of a homogeneous quadratic form) Consider $f(x) = x^\top Ax$, where $A$ is a symmetric $n \times n$ matrix. As we have seen in Example C.7,

$$Df(x)[\Delta x^1] = 2x^\top A\,\Delta x^1.$$

Differentiating this mapping in $x$, we get

$$D^2 f(x)[\Delta x^1, \Delta x^2] = \lim_{t \to +0} t^{-1}\left(2(x + t\Delta x^2)^\top A\,\Delta x^1 - 2x^\top A\,\Delta x^1\right) = 2(\Delta x^2)^\top A\,\Delta x^1.$$

Hence,

$$\boxed{D^2 f(x)[\Delta x^1, \Delta x^2] = 2(\Delta x^2)^\top A\,\Delta x^1.}$$

Note that the second derivative of a quadratic form is independent of $x$. Consequently, the third, the fourth, etc., derivatives of a quadratic form are identically zero.

**Example** C.16 (Second-order derivative of the log-det barrier) Consider $F(X) = \ln \mathrm{Det}(X)$. As we have seen, this function of an $n \times n$ matrix is well-defined and differentiable on the set $U$ of matrices with positive determinant (which is an open set in the space $\mathbf{R}^{n \times n}$ of $n \times n$ matrices). In fact, this function is $\mathrm{C}^\infty$ in $U$.

Let us compute its second-order derivative. Recall from Example C.9 that

$$DF(X)[\Delta X^1] = \mathrm{Tr}(X^{-1}\Delta X^1). \tag{C.12}$$

To differentiate the right hand side in $X$, we will need the derivative of the mapping $G(X) = X^{-1}$ which is defined on the open set of nonsingular $n \times n$ matrices. Recall from Example C.8 that we have the relation

$$D(X^{-1})[\Delta X] = -X^{-1}\Delta X X^{-1}, \quad \left[X \in \mathbf{R}^{n \times n}, \mathrm{Det}(X) \neq 0\right]$$

which is the "matrix extension" of the standard relation $(x^{-1})' = -x^{-2}$, $x \in \mathbf{R}$.

Now we are ready to compute the second derivative of the log-det barrier

$F(X) = \ln \mathrm{Det}(X)$. Starting from its first derivative $DF(X)[\Delta X^1] = \mathrm{Tr}(X^{-1}\Delta X^1)$, and differentiating we obtain

$$\begin{aligned}
D^2F(X)[\Delta X^1, \Delta X^2] &= \lim_{t\to+0} t^{-1}\left(\mathrm{Tr}\left((X+t\Delta X^2)^{-1}\Delta X^1\right) - \mathrm{Tr}(X^{-1}\Delta X^1)\right)\\
&= \lim_{t\to+0} \mathrm{Tr}\left(t^{-1}\left((X+t\Delta X^2)^{-1}\Delta X^1 - X^{-1}\Delta X^1\right)\right)\\
&= \lim_{t\to+0} \mathrm{Tr}\left(t^{-1}\left((X+t\Delta X^2)^{-1} - X^{-1}\right)\Delta X^1\right)\\
&= \mathrm{Tr}\left(\left(\lim_{t\to+0} t^{-1}\left((X+t\Delta X^2)^{-1} - X^{-1}\right)\right)\Delta X^1\right)\\
&= \mathrm{Tr}\left(\left(-X^{-1}\Delta X^2 X^{-1}\right)\Delta X^1\right),
\end{aligned}$$

where the last equation follows from $D(X^{-1})[\Delta X^2] = -X^{-1}\Delta X^2 X^{-1}$ for all nonsingular $X \in \mathbf{R}^{n\times n}$. Hence, we arrive at the formula

$$\boxed{D^2F(X)[\Delta X^1, \Delta X^2] = -\mathrm{Tr}(X^{-1}\Delta X^2 X^{-1}\Delta X^1) \quad [X \in \mathbf{R}^{n\times n}, \mathrm{Det}(X) > 0].}$$

Since $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ (check it!) for all matrices $A, B$ such that the product $AB$ makes sense and is square, the right hand side in the above formula is symmetric in $\Delta X^1$, $\Delta X^2$, as it should be for the second derivative of a $\mathrm{C}^2$ function.

### *C.2.3 Taylor expansion*

Assume that $f : \mathbf{R}^n \to \mathbf{R}^m$ is $\mathrm{C}^k$ in an open neighborhood $U$ of a point $\bar{x}$. The *Taylor expansion of order* $k$ of $f$, built at the point $\bar{x}$, is the function defined as

$$\begin{aligned}
F_k(x) := f(\bar{x}) &+ \frac{1}{1!}Df(\bar{x})[x-\bar{x}] + \frac{1}{2!}D^2f(\bar{x})[x-\bar{x}, x-\bar{x}] \\
&+ \frac{1}{3!}D^3f(\bar{x})[x-\bar{x}, x-\bar{x}, x-\bar{x}] + \ldots + \frac{1}{k!}D^kf(\bar{x})[\underbrace{x-\bar{x}, \ldots, x-\bar{x}}_{k \text{ times}}].
\end{aligned} \tag{C.13}$$

We are already acquainted with the Taylor expansion of order 1

$$F_1(x) = f(\bar{x}) + Df(\bar{x})[x-\bar{x}],$$

this is the affine function of $x$ which approximates "very well" $f(x)$ in a neighborhood of $\bar{x}$, namely, within approximation error $\bar{o}(|x-\bar{x}|)$. We next state that a similar fact is true for Taylor expansions of higher order.

**Theorem** C.17 Let $f : \mathbf{R}^n \to \mathbf{R}^m$ be $\mathrm{C}^k$ in a neighborhood of $\bar{x}$, and let $F_k(x)$ be the Taylor expansion of $f$ at $\bar{x}$ of degree $k$. Then,

(i) $F_k(x)$ is a vector-valued polynomial of full degree less than or equal to $k$, i.e., every one of the coordinates of the vector $F_k(x)$ is a polynomial of $x_1, \ldots, x_n$, and the sum of powers of $x_i$'s in every term of this polynomial does not exceed $k$.

(ii) $F_k(x)$ approximates $f(x)$ in a neighborhood of $\bar{x}$ up to a remainder which is $\bar{o}(\|x-\bar{x}\|_2^k)$ as $x \to \bar{x}$. That is, for every $\epsilon > 0$, there exists $\delta > 0$

such that

$$\|x - \bar{x}\|_2 \le \delta \implies \|F_k(x) - f(x)\|_2 \le \epsilon \|x - \bar{x}\|_2^k.$$

$F_k(\cdot)$ is the unique polynomial with components of full degree less than or equal to $k$ which approximates $f$ up to a remainder which is $\bar{o}(\|x - \bar{x}\|^k)$.

(iii) The value and the derivatives of $F_k$ of orders $1, 2, \dots, k$, taken at $\bar{x}$, are the same as the value and the corresponding derivatives of $f$ taken at the same point.

As stated in Theorem C.17, $F_k(x)$ approximates $f(x)$ for $x$ close to $\bar{x}$ up to a remainder which is $\bar{o}(|x - \bar{x}|^k)$. In many cases, it is not enough to know that the reminder is "$\bar{o}(\|x - \bar{x}\|_2^k)$" — we need an explicit bound on this remainder. The standard bound of this type is as follows:

**Theorem** C.18 Let $k$ be a positive integer, and let $f : \mathbf{R}^n \to \mathbf{R}^m$ be C$^{k+1}$ in a ball $B_r := B_r(\bar{x}) = \{x \in \mathbf{R}^n : \|x - \bar{x}\|_2 < r\}$ of a radius $r > 0$ centered at a point $\bar{x}$. Assume that the directional derivatives of order $k + 1$, taken at every point of $B_r$ along every $\|\cdot\|_2$-unit direction, do not exceed certain $L < \infty$:

$$\|D^{k+1} f(x)[d, \dots, d]\|_2 \le L \quad \forall (x \in B_r),\ \forall (d \in \mathbf{R}^n : \|d\|_2 = 1).$$

Then, the following holds for the Taylor expansion $F_k$ of order $k$ of $f$ taken at $\bar{x}$

$$\|f(x) - F_k(x)\|_2 \le \frac{L\|x - \bar{x}\|_2^{k+1}}{(k+1)!}, \quad \forall x \in B_r.$$

Thus, in a neighborhood of $\bar{x}$ the remainder of the *k-th order* Taylor expansion, taken at $\bar{x}$, is of order of $L\|x - \bar{x}\|_2^{k+1}$, where $L$ is the maximal (over all unit directions and all points from the neighborhood) $\|\cdot\|_2$-magnitudes of the directional derivatives *of order* $k + 1$ of $f$.

# Appendix D

# Prerequisites: Symmetric Matrices and Positive Semidefinite Cone

## D.1 Symmetric matrices

Let $\mathbf{S}^m$ be the space of symmetric $m \times m$ matrices, and $\mathbf{R}^{m\times n}$ be the space of rectangular $m \times n$ matrices with real entries. From the viewpoint of their linear structure (i.e., the operations of addition and multiplication by real numbers) $\mathbf{S}^m$ is nothing but the arithmetic linear space $\mathbf{R}^{m(m+1)/2}$ of dimension $\frac{m(m+1)}{2}$: by arranging the elements of a symmetric $m \times m$ matrix $X$ in a single column, say, in the row-by-row order, you get a usual $m^2$-dimensional column vector; multiplication of a matrix by a real number and addition of matrices correspond to the same operations with the "representing vector(s)." When $X$ runs through $\mathbf{S}^m$, the vector representing $X$ runs through $m(m+1)/2$-dimensional subspace of $\mathbf{R}^{m^2}$ consisting of vectors satisfying the "symmetry condition" – the coordinates coming from symmetric to each other pairs of entries in $X$ are equal to each other. Similarly, $\mathbf{R}^{m,n}$ as a linear space is just $\mathbf{R}^{mn}$, and it is natural to equip $\mathbf{R}^{m,n}$ with the inner product defined as the usual inner product of the vectors representing the matrices:

$$\langle X, Y\rangle := \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}Y_{ij} = \mathrm{Tr}(XY^\top) \quad \left[= \sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij}X_{ij} = \mathrm{Tr}(YX^\top)\right]$$

Here Tr stands for the *trace* – the sum of diagonal elements of a (square) matrix. With this inner product (called the *Frobenius inner product*), $\mathbf{R}^{m\times n}$ becomes a legitimate Euclidean space, and we may use in connection with this space all notions based upon the Euclidean structure, e.g., the *(Frobenius) norm* of a matrix

$$\|X\|_2 := \sqrt{\langle X, X\rangle} = \sqrt{\sum\nolimits_{i=1}^{m}\sum\nolimits_{j=1}^{n} X_{ij}^2} = \sqrt{\mathrm{Tr}(XX^\top)},$$

and likewise the notions of orthogonality, orthogonal complement of a linear subspace, etc. The same applies to the space $\mathbf{S}^m$ equipped with the Frobenius inner product; of course, the Frobenius inner product of symmetric matrices can be written without the transposition sign:

$$\langle X, Y\rangle = \mathrm{Tr}(XY), \ X, Y \in \mathbf{S}^m.$$

The following simple fact is very useful:

**Fact** D.1 Let $X, Y$ be rectangular matrices such that $XY$ makes sense and is a square matrix. Then, $\mathrm{Tr}(YX)$ also makes sense and

$$\mathrm{Tr}(XY) = \mathrm{Tr}(YX).$$

Here is another equally simple and useful fact.

**Fact** D.2 If $X, Y \in \mathbf{R}^{m\times n}$, the Frobenius inner product of $X$ and $Y$ is equal to the Frobenius inner product of $X^\top$ and $Y^\top$:

$$\mathrm{Tr}(XY^\top) = \mathrm{Tr}\left((X^\top)(Y^\top)^\top\right).$$

Moreover, when $U$ is an orthogonal $m \times m$ matrix (i.e., $UU^\top = U^\top U = I_m$, which is the same as $U^{-1} = U^\top$), and $V$ is an orthogonal $n \times n$ matrix (i.e., $VV^\top = V^\top V = I_n$), the Frobenius inner product of $UXV$ and $UYV$ is the same as the Frobenius inner product of $X$ and $Y$:

$$\mathrm{Tr}(XY^\top) = \mathrm{Tr}\left((UXV)(UYV)^\top\right).$$

### *D.1.1 Main facts on symmetric matrices*

Here, we will focus on the space $\mathbf{S}^m$ of symmetric matrices. We start with the following most important property of these matrices.

**Theorem** D.3 [Eigenvalue decomposition] An $n\times n$ matrix $A$ is symmetric if and only if it admits an orthonormal system of eigenvectors, i.e., there exist orthonormal basis $\{e_1, \ldots, e_n\}$ such that

$$Ae_i = \lambda_i e_i, \quad i = 1, \ldots, n, \tag{D.1}$$

holds for some reals $\lambda_i$.

In connection with Theorem D.3, let us recall the following notions and facts:

**D.1.1.A. Eigenvectors and eigenvalues.**

**Definition** D.4 [Eigenvector and eigenvalue] An *eigenvector* of an $n \times n$ matrix $A$ is a *nonzero* vector $e$ (real or complex) such that $Ae = \lambda e$ holds for some (real or complex) scalar $\lambda$. This scalar is called the *eigenvalue* of $A$ *corresponding to the eigenvector* $e$.

Eigenvalues of $A$ are exactly the roots of the *characteristic polynomial* of $A$ given by

$$\pi(z) = \mathrm{Det}(zI - A) = z^n + b_1 z^{n-1} + b_2 z^{n-2} + \ldots + b_n.$$

In the sequel, we denote by $\sigma(A)$ the *spectrum* of $A \in \mathbf{S}^n$, i.e., the set of all real numbers which are eigenvalues of the matrix. To avoid misunderstanding, let us stress that $\sigma(A)$ has as many points as many distinct eigenvalues $A$ has and “pays no attention” to the multiplicities of these eigenvalues; for example, $\sigma(I_n) = \{1\}$.

Theorem D.3 states, in particular, that for a symmetric matrix $A$, all eigenvalues are real, and the corresponding eigenvectors can be chosen to be real and to form an orthonormal basis in $\mathbf{R}^n$.

**Remark** D.5 When $Q$ is a square and nonsingular $n \times n$ matrix, the *similarity transformation* $A \mapsto QAQ^{-1}$ preserves the characteristic polynomial (and therefore the eigenvalues). Indeed,

$$\begin{aligned}\mathrm{Det}(zI - QAQ^{-1}) &= \mathrm{Det}(Q[zI - A]Q^{-1}) = \mathrm{Det}(Q)\mathrm{Det}(zI - A)\mathrm{Det}(Q^{-1})\\ &= \mathrm{Det}(zI - A).\end{aligned}$$

**D.1.1.B. Eigenvalue decomposition of a symmetric matrix.** We will give an alternative characterization of symmetric matrices. To this end, we start with a definition and a fact that are important on their own.

**Definition** D.6 [Orthogonal matrix] An $n \times n$ matrix $U$ is called *orthogonal* if $U^{-1} = U^\top$.

There are several equivalent characterizations of orthogonal matrices.

**Fact** D.7 An $n \times n$ matrix $U$ is orthogonal if and only if any (and all) of the following holds:

- $U^{-1} = U^\top$,
- $U^\top U = I$,
- $UU^\top = I$,
- the columns of $U$ form an orthonormal basis in $\mathbf{R}^n$,
- the rows of $U$ form an orthonormal basis in $\mathbf{R}^n$.

Theorem D.3 admits an equivalent reformulation as follows (check the equivalence!).

**Theorem** D.8 An $n \times n$ matrix $A$ is symmetric if and only if it can be represented in the form

$$A = U\Lambda U^\top, \tag{D.2}$$

where

- $U$ is an orthogonal matrix,
- $\Lambda$ is the diagonal matrix with the diagonal entries $\lambda_1, \ldots, \lambda_n$.

Representation (D.2) with orthogonal $U$ and diagonal $\Lambda$ is called the *eigenvalue decomposition* of $A$. In such a representation,

- the columns of $U$ form an orthonormal system of eigenvectors of $A$, and
- the diagonal entries in $\Lambda$ are the eigenvalues of $A$ corresponding to these eigenvectors.

**D.1.1.C. Vector of eigenvalues.** When speaking about eigenvalues $\lambda_i(A)$ of a symmetric $n \times n$ matrix $A$, we always arrange them in the non-increasing order:

$$\lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_n(A).$$

Moreover, we use $\lambda(A) \in \mathbf{R}^n$ to denote the vector of eigenvalues of $A$ taken in the above order. We let $\lambda_{\max}(A)$ (and respectively $\lambda_{\min}(A)$) correspond to the largest (smallest) eigenvalue of $A$.

**D.1.1.D. Freedom in eigenvalue decomposition.** It is important to note that in the eigenvalue decomposition (D.2), a part of the data $\Lambda$, $U$ is uniquely defined by $A$, while the other data admit certain "freedom." Specifically, the sequence $\lambda_1, \ldots, \lambda_n$ of eigenvalues of $A$ (i.e., diagonal entries of $\Lambda$) is exactly the sequence of roots of the characteristic polynomial of $A$ (every root is repeated according to its multiplicity) and thus is uniquely defined by $A$ (provided that we arrange the entries of the sequence in the non-increasing order). On the other hand, it is possible that the columns of $U$ are not uniquely defined by $A$. What is uniquely defined, are the *linear spans* $E(\lambda)$ of the columns of $U$ corresponding to all eigenvalues equal to certain $\lambda$, and such a linear span is nothing but the *spectral subspace* $\{x : Ax = \lambda x\}$ of $A$ corresponding to the eigenvalue $\lambda$. Given $A \in \mathbf{S}^n$, there are as many spectral subspaces of $A$ as the number of different eigenvalues of $A$. Moreover, the spectral subspaces corresponding to different eigenvalues of a given symmetric matrix are orthogonal to each other, and their sum is the entire space. When building an orthogonal matrix $U$ for the eigenvalue decomposition (D.2), one chooses an orthonormal eigenbasis in the spectral subspace corresponding to the largest element in $\sigma(A)$ and makes the vectors of this basis the first columns in $U$, then chooses an orthonormal basis in the spectral subspace corresponding to the second largest element of $\sigma(A)$ and makes the vectors from this basis the next columns of $U$, and so on.

**D.1.1.E. "Simultaneous" decomposition of commuting symmetric matrices.** Given a number of symmetric matrices $A_1, \ldots, A_k \in \mathbf{S}^n$, it is useful to know when they commute with each other, i.e., $A_iA_j = A_jA_i$ for all $i, j$. It turns out that there is a complete characterization of this property given as follows.

**Fact** D.9 The matrices $A_1, \ldots, A_k \in \mathbf{S}^n$ commute with each other ($A_iA_j = A_jA_i$ for all $i, j$) if and only if they can be "simultaneously diagonalized," i.e., there exist a single orthogonal matrix $U$ and diagonal matrices $\Lambda_1$,...,$\Lambda_k$ such that

$$A_i = U\Lambda_iU^\top,\ i = 1, \ldots, k.$$

The proof of Fact D.9 relies on two simple facts that are important by their own rights. We state these facts next.

**Fact** D.10 Let $A, B \in \mathbf{R}^{n\times n}$ be commuting and $\lambda$ be a real eigenvalue of $A$. Then, the spectral subspace $E = \{x \in \mathbf{R}^n : Ax = \lambda x\}$ of $A$ corresponding to $\lambda$ is invariant for $B$ (i.e., $Be \in E$ for every $e \in E$).

**Fact** D.11 If $A$ is an $n \times n$ matrix and $L$ is an invariant subspace of $A$ (i.e., $L$ is a linear subspace such that $Ae \in L$ whenever $e \in L$), then the orthogonal complement $L^\perp$ of $L$ is invariant for the matrix $A^\top$. In particular, if $A$ is symmetric and $L$ is invariant subspace of $A$, then $L^\perp$ is an invariant subspace of $A^\top$ as well.

### *D.1.2 Variational characterization of eigenvalues*

To put what follows into a proper perspective, let us ask ourselves the following simple question: *for a vector $x$ with coordinates $x_1, \dots, x_n$, let $x^{(k)}$ be the $k$-th largest of the coordinates – the one which will be at $k$-th place when the entries in $x$ are rearranged in the non-ascending order. For example, for $x = [2; 1; 2]$ we have $x^{(1)} = 2$, $x^{(2)} = 2$, $x^{(3)} = 1$. Now the question is how to characterize $x^{(k)}$ without referring to rearranging the entries in $x$.* The answer is as follows. Let us think about $n$ stones with the weights $x_1, x_2, \dots, x_n$. Then $x^{(k)}$ can be described as follows: let us somehow throw away from the collection of our $n$ stones $k-1$ of them and look at the largest weight, $w$, of the remaining stones; this weight depends on which $k-1$ of the stones were thrown away. Now let us take the minimum of the weights $w$ we can get in this fashion over all possible ways to select $k-1$ stones to throw away; this minimum is exactly $x^{(k)}$. By the way, it is immediately self-evident that when $x, y \in \mathbf{R}^n$ and $y \geq x$, then $y^{(k)} \geq x^{(k)}$ for every $k \leq n$. However, even self-evident facts need proof, and here is the proof: when increasing entries $x_i$ in $x$ to $y_i$'s (which precisely corresponds to increasing the weights of every one of our stones) we clearly cannot decrease the above $w$'s (that is, the largest of the weights in groups obtained by throwing away a particular collection of $k-1$ stones). And since none of the $w$'s decrease, their minimum does not decrease as well.

Variational characterization of eigenvalues of a symmetric matrix is the extremely useful matrix version of the elementary considerations above.

**Theorem** D.12 [Variational Characterization of Eigenvalues (VCE)] For any $A \in \mathbf{S}^n$, we have

$$\lambda_\ell(A) = \min_{E \in \mathcal{E}_\ell} \max_{x \in E, x^\top x = 1} x^\top A x, \quad \ell = 1, \dots, n, \tag{D.3}$$

where $\mathcal{E}_\ell$ is the family of all linear subspaces in $\mathbf{R}^n$ of the dimension $n - \ell + 1$.

Based on VCE, in order to get the largest eigenvalue $\lambda_1(A)$, we maximize the quadratic form $x^\top A x$ over the unit sphere $S = \left\{x \in \mathbf{R}^n : x^\top x = 1\right\}$ and the maximum is exactly $\lambda_1(A)$. To get the second largest eigenvalue $\lambda_2(A)$, we act as follows: we choose a linear subspace $E$ of dimension $n-1$ and maximize the quadratic form $x^\top A x$ over the cross-section of $S$ by this subspace; the maximum value of the form depends on $E$, and we minimize this maximum over linear subspaces $E$ of the dimension $n-1$; the resulting value is exactly $\lambda_2(A)$. To get $\lambda_3(A)$, we replace in the latter construction subspaces of the dimension $n-1$ by those of the dimension $n-2$, and so on. In particular, the smallest eigenvalue $\lambda_n(A)$ is just the minimum, over all linear subspaces $E$ of the dimension $n-n+1 = 1$, i.e., over all lines passing through the origin, of the quantities $x^\top Ax$, where $x \in E$ has unit norm ($x^\top x = 1$). In other words, $\lambda_n(A)$ is just the minimum of the quadratic form $x^\top Ax$ over the unit sphere $S$.

**Proof of the VCE.** Let $e_1, \dots, e_n$ be an orthonormal eigenbasis of $A$, i.e., $Ae_\ell = \lambda_\ell(A)e_\ell$ and $e_\ell$ are unit vectors for all $1 \le \ell \le n$. For $1 \le \ell \le n$, let

$$F_\ell := \mathrm{Lin}\{e_1, \dots, e_\ell\}, \quad \text{and} \quad G_\ell := \mathrm{Lin}\{e_\ell, e_{\ell+1}, \dots, e_n\}.$$

Finally, for $x \in \mathbf{R}^n$, let $\xi(x)$ be the vector of coordinates of $x$ in the orthonormal basis $e_1, \dots, e_n$. Note that

$$x^\top x = \xi(x)^\top \xi(x),$$

since $\{e_1, \dots, e_n\}$ is an orthonormal basis, and that

$$\begin{array}{rl} x^\top Ax & = x^\top A \sum_{i=1}^n \xi_i(x)e_i = x^\top \sum_{i=1}^n \lambda_i(A)\xi_i(x)e_i = \sum_{i=1}^n \lambda_i(A)\xi_i(x) \underbrace{(x^\top e_i)}_{=\xi_i(x)} \\ & = \sum_{i=1}^n \lambda_i(A)\xi_i^2(x). \end{array} \tag{D.4}$$

Now, consider a fixed $\ell$ such that $1 \le \ell \le n$, and set $E = G_\ell$. Note that $E$ is a linear subspace of the dimension $n - \ell + 1$. Moreover, from (D.4), we deduce that

$$x^\top Ax = \sum_{i=\ell}^n \lambda_i(A)\xi_i(x)^2$$

holds for any $x \in E$. Therefore,

$$\begin{array}{rcl} \max\limits_x \left\{x^\top Ax : \ x \in E, \, x^\top x = 1\right\} & = & \max_\xi \left\{\sum_{i=\ell}^n \lambda_i(A)\xi_i^2 : \ \sum_{i=\ell}^n \xi_i^2 = 1\right\} \\ & = & \max\limits_{\ell \le i \le n} \lambda_i(A) = \lambda_\ell(A). \end{array}$$

Thus, for appropriately chosen $E \in \mathcal{E}_\ell$, the inner maximum in the right hand side of (D.3) equals to $\lambda_\ell(A)$. Hence, we have

$$\min_{E \in \mathcal{E}_\ell} \max_{x \in E, x^\top x = 1} x^\top Ax \le \lambda_\ell(A).$$

It remains to prove the reverse inequality. To this end, consider a linear subspace $E$ of the dimension $n - \ell + 1$ and observe that its intersection with the linear subspace $F_\ell$ of the dimension $\ell$ is nontrivial (indeed, $\dim E + \dim F_\ell = (n - \ell + 1) + \ell > n$, so that $\dim(E \cap F) > 0$ by the Dimension formula). Consider any unit vector $y \in E \cap F_\ell$. Since $y$ is a unit vector from $F_\ell$, it admits a representation of the form,

$$y = \sum_{i=1}^\ell \eta_i e_i \quad \text{with} \quad \sum_{i=1}^\ell \eta_i^2 = 1,$$

whence, by (D.4),

$$y^\top Ay = \sum_{i=1}^{\ell} \lambda_i(A)\eta_i^2 \geq \min_{1\leq i\leq \ell} \lambda_i(A) = \lambda_\ell(A).$$

Since $y \in E$, we conclude that

$$\max_{x\in E, x^\top x=1} x^\top Ax \geq y^\top Ay \geq \lambda_\ell(A).$$

Since $E$ is an arbitrary subspace form $\mathcal{E}_\ell$, we conclude that the right hand side in (D.3) is $\geq \lambda_\ell(A)$. □

Our reasoning in the proof of the VCE uses the relation (D.4) which is a simple and useful byproduct. Therefore, we state it formally as a corollary.

**Corollary** D.13 For any $A \in \mathbf{S}^n$, the quadratic form $x^\top Ax$ is exactly equal to the weighted sum of squares of the coordinates $\xi_i(x)$ of $x$ taken with respect to an orthonormal eigenbasis of $A$. Moreover, the weights in this sum are exactly the eigenvalues of $A$. That is,

$$x^\top Ax = \sum\nolimits_i \lambda_i(A)\xi_i^2(x).$$

### *D.1.3 Corollaries of the VCE*

VCE admits a number of extremely important corollaries as follows:

**D.1.2.A. Eigenvalue characterization of positive (semi)definite matrices.**

**Definition** D.14 [Positive definite matrix] A matrix $A$ is called *positive definite* [notation: $A \succ 0$], if it is symmetric and $x^\top Ax > 0$ holds for all $x \neq 0$.

**Definition** D.15 [Positive semidefinite matrix] A matrix $A$ is called *positive semidefinite* [notation: $A \succeq 0$], if it is symmetric and $x^\top Ax \geq 0$ holds for all $x$.

Based on VCE we arrive at the following eigenvalue characterization of positive (semi)definite matrices.

**Proposition** D.16 A symmetric matrix $A$ is positive semidefinite if and only if its eigenvalues are nonnegative. Moreover, $A$ is positive definite if and only if all eigenvalues of $A$ are positive.

**Proof.** Indeed, $A$ is positive definite (positive semidefinite) if and only if the minimum value of $x^\top Ax$ over the unit sphere is positive (nonnegative). Recall also that by VCE, the minimum value of $x^\top Ax$ over the unit sphere is exactly the minimum eigenvalue of $A$. □

**D.1.2.B. $\succeq$-Monotonicity of the vector of eigenvalues.**

We write $A \succeq B$ ($A \succ B$) to express that $A, B$ are symmetric matrices of the same size such that $A - B$ is positive semidefinite (respectively, positive definite).

**Proposition** D.17 Consider $A, B \in \mathbf{S}^n$. If $A \succeq B$, then $\lambda(A) \geq \lambda(B)$. Also, if $A \succ B$, then $\lambda(A) > \lambda(B)$.

**Proof.** Indeed, when $A \succeq B$, then, of course,

$$\max_{x \in E : x^\top x = 1} x^\top Ax \geq \max_{x \in E : x^\top x = 1} x^\top Bx$$

for every linear subspace $E$, whence

$$\lambda_\ell(A) = \min_{E \in \mathcal{E}_\ell} \max_{x \in E : x^\top x = 1} x^\top Ax \geq \min_{E \in \mathcal{E}_\ell} \max_{x \in E : x^\top x = 1} x^\top Bx = \lambda_\ell(B), \ \ell = 1, \ldots, n,$$

i.e., $\lambda(A) \geq \lambda(B)$. The case of $A \succ B$ can be considered similarly. □

**D.1.2.C. Eigenvalue Interlacement Theorem.**

This is an extremely important result, and we formulate it as follows.

**Theorem** D.18 [Eigenvalue Interlacement Theorem] For $A \in \mathbf{S}^n$, let $\bar{A}$ be the angular $(n-k) \times (n-k)$ submatrix of $A$ where $k \leq n$. Then, for every $\ell \leq n-k$, the $\ell$-th eigenvalue of $\bar{A}$ separates the $\ell$-th and the $(\ell+k)$-th eigenvalues of $A$:

$$\lambda_\ell(A) \succeq \lambda_\ell(\bar{A}) \succeq \lambda_{\ell+k}(A). \tag{D.5}$$

**Proof.** Recall that by VCE,

$$\lambda_\ell(\bar{A}) = \min_{E \in \bar{\mathcal{E}}_\ell} \max_{x \in E : x^\top x = 1} x^\top Ax,$$

where $\bar{\mathcal{E}}_\ell$ is the family of all linear subspaces of the dimension $n-k-\ell+1$ contained in the linear subspace $\{x \in \mathbf{R}^n : \ x_{n-k+1} = x_{n-k+2} = \ldots = x_n = 0\}$. Since $\bar{\mathcal{E}}_\ell \subset \mathcal{E}_{\ell+k}$, we have

$$\lambda_\ell(\bar{A}) = \min_{E \in \bar{\mathcal{E}}_\ell} \max_{x \in E : x^\top x = 1} x^\top Ax \geq \min_{E \in \mathcal{E}_{\ell+k}} \max_{x \in E : x^\top x = 1} x^\top Ax = \lambda_{\ell+k}(A).$$

We have proved the left inequality in (D.5). Applying this inequality to the matrix $-A$, we get

$$-\lambda_\ell(\bar{A}) = \lambda_{n-k-\ell}(-\bar{A}) \geq \lambda_{n-\ell}(-A) = -\lambda_\ell(A),$$

or, which is the same, $\lambda_\ell(\bar{A}) \leq \lambda_\ell(A)$. This then proves the first inequality in (D.5). □

### *D.1.4 Spectral norm and Lipschitz continuity of vector of eigenvalues*

**Spectral and induced norms of matrices.** A useful example of a norm (to refresh your memory on what a norm is, see section B.1) is the *spectral norm* of an $m \times n$ matrix $A$:

$$\|A\| = \max_{x:\|x\|_2 \le 1} \|Ax\|_2,$$

The spectral norm is indeed a special case of the *induced norm of a linear mapping.* $x \mapsto y = Ax : E \to F$, where $E$ and $F$ are (finite-dimensional) linear spaces equipped with norms $\|\cdot\|_E$ and $\|\cdot\|_F$ respectively. The norm $\|A\|_{F,E}$ of $A$ induced by these two norms is defined as

$$\|A\|_{F,E} := \max_x \{\|Ax\|_F : \ \|x\|_E \le 1\}.$$

The definition of induced norm immediately imply (check it!) that, first, for all $x \in E$, $y \in F$ one has

$$\|Ax\|_F \le \|A\|_{F,E}\,\|x\|_E,$$

and, second, the latter inequality is equality for properly selected nonzero $x \in E$. In other words, the induced norm is the largest "amplification factor" – the largest factor by which the norm $\|\cdot\|_F$ of $Ax$ can be larger than the norm $\|\cdot\|_E$ of a nonzero vector $x$.

Clearly, when $E = \mathbf{R}^n$, $F = \mathbf{R}^m$, and $\|\cdot\|_E$, $\|\cdot\|_F$ are the standard Euclidean norms on the respective spaces, the induced norm of $A$ is exactly the spectral norm of $A$.

**Fact** D.19 (i) Let $E$ and $F$ be finite-dimensional linear spaces. The induced norm $\|\cdot\|_{F,E}$ is indeed a norm on the space $\mathrm{Lin}(E,F)$ of linear mappings from $E$ to $F$.
(ii) Let $E, F, G$ be finite-dimensional linear spaces equipped with the norms $\|\cdot\|_E$, $\|\cdot\|_F$, $\|\cdot\|_G$ respectively. Let $y = Ax : E \to F$ and $z = By : F \to G$ be linear mappings. Then,

$$\|BA\|_{G,E} \le \|B\|_{G,F}\,\|A\|_{F,E}.$$

We also have the following useful properties of the spectral norm $\|\cdot\|$ on $\mathbf{R}^{m\times n}$.

**Fact** D.20 Let $\|\cdot\|$ be the spectral norm on $\mathbf{R}^{m\times n}$.
(i) For any $A \in \mathbf{R}^{m\times n}$, we have

$$\|A\| = \max_{x,y}\left\{y^\top Ax : \ \|x\|_2 \le 1, \ \|y\|_2 \le 1\right\},$$

hence also $\|A\| = \|A^\top\|$.
(ii) For any $A \in \mathbf{S}^n$, we have $\|A\| = \max\{|\lambda_{\max}(A)|, |\lambda_{\min}(A)|\}$. Moreover, for any $A \in \mathbf{R}^{m\times n}$ we have

$$\|A\|^2 = \|A^\top A\| = \lambda_{\max}(A^\top A) = \lambda_{\max}(AA^\top) = \|AA^\top\| = \|A^\top\|^2.$$

**Lipschitz continuity of the vector of eigenvalues.** Another useful consequence of VCE is the Lipschitz continuity of the vector of eigenvalues of a matrix as a function of the matrix.

**Fact** D.21 The vector-valued function $A \mapsto \lambda(A) : \mathbf{S}^n \to \mathbf{R}^n$ is Lipschitz continuous, specifically, denoting by $\|\cdot\|$ the spectral norm, for all $k \leq n$, we have

$$|\lambda_k(A) - \lambda_k(A')| \leq \|A - A'\| \quad \forall (A, A' \in \mathbf{S}^n).$$

We also have the following immediate consequence of Fact D.21.

**Corollary** D.22 Let $S$ be a subset of $\mathbf{R}^n$, and $\mathcal{S}$ be the set of all matrices $A \in \mathbf{S}^n$ with $\lambda(A) \in S$. Whenever $S$ is open, or closed, or bounded, so is $\mathcal{S}$.

### *D.1.5 Functions of symmetric matrices*

Let $A$ be a symmetric $m \times m$ matrix, and $f$ be a real-valued function defined on a subset of real line containing the spectrum $\sigma(A)$ of $A$. We define the $m \times m$ symmetric *matrix* $f(A)$ as follows:

> Let $\sigma(A) = \{\mu_1, \ldots, \mu_s\}$, and let $\mathbf{R}^m = E_1 + \ldots + E_s$ be the decomposition of $\mathbf{R}^m$ into the sum of mutually orthogonal spectral subspaces of $A$, i.e., $E_i := \{x \in \mathbf{R}^m : Ax = \lambda_i x\}$. The matrix $f(A)$ has every one of $E_i$ invariant, and $x \in E_i$ implies $Ax = f(\mu_i)x$ for all $i \leq s$.
> Equivalently: given an eigenvalue decomposition $A = U\,\mathrm{Diag}\{\lambda_1, \ldots, \lambda_m\}U^\top$ of $A$, one has $f(A) = U\,\mathrm{Diag}\{f(\lambda_1), \ldots, f(\lambda_m)\}U^\top$.

From this definition it follows immediately that $f(A) = g(A)$ whenever $f$ and $g$ coincide with each other on the spectrum of $A$.

Moreover, it is immediately seen (check it!) that the resulting “calculus of functions of (fixed) symmetric matrix $A$” obeys very natural rules:

1. $f(\cdot) \mapsto f(A)$ is a mapping from the algebra $\mathbf{R}(\sigma(A))$ of real-valued functions on the subset $\sigma(A)$ of the real axis into the space $\mathbf{S}^m$ of $m \times m$ symmetric matrices which preserves linear operations and multiplication. That is, for any $f_1, \ldots, f_k \in \mathbf{R}(\sigma(A))$ and any $a_1, \ldots, a_k \in \mathbf{R}$, we have

$$[a_1 f_1(\cdot) + \ldots + a_k f_k(\cdot)](A) = a_1 f_1(A) + \ldots + a_k f_k(A)$$

and

$$[f_1(\cdot) \cdot f_2(\cdot)](A) = f_1(A) f_2(A).$$

In particular, all matrices of the form $f(A)$, $f \in \mathbf{R}(\sigma(A))$ commute with each other.

Besides this, if $f$ is a real-valued function on $\sigma(A)$, then $\sigma(f(A)) = f(\sigma(A)) :=$

$\{t = f(s) : s \in \sigma(A)\}$, and if $g$ is a real-valued function on $f(\sigma(A))$ and $g \circ f(s) = g(f(s))$, $s \in \Sigma(A)$ is the composition of $g$ and $f$, then

$$(g \circ f)(A) = g(f(A)).$$

2. When $f(\cdot) \equiv 1$, one has $f(A) = I_m$, and when $f(x) \equiv x$, one has $f(A) = A$. More generally, whenever $f$ is a real algebraic polynomial
$$f(x) = a_0 + \sum_{i=1}^{k} a_i x^i,$$
we have
$$f(A) = a_0 I_n + \sum_{i=1}^{k} a_i A^i.$$
3. The mapping $f(\cdot) \to f(A)$ "preserves nonnegativity," i.e., the matrix $f(A)$ is *positive semidefinite* whenever $f(\cdot)$ is nonnegative on $\sigma(A)$.
4. When a sequence $f_i(\cdot) \in \mathbf{R}(\sigma(A))$ pointwise on $\sigma(A)$ converges to $f(\cdot)$ as $i \to \infty$, the matrices $f_i(A)$ converge to $f(A)$ as $i \to \infty$. Moreover, for real-valued on $\sigma(A)$ functions $f, g$ one has
$$\|f(A) - g(A)\| \leq \max_{s \in \sigma(A)} |f(s) - g(s)|,$$
where $\|\cdot\|$ is the spectral norm.

**Illustration: Square root of a positive semidefinite matrix.** When $A$ is a positive semidefinite matrix, its *matrix square root* $A^{1/2}$ is, by definition, $f(A)$, where $f(\mu) = \mu^{1/2}$ is the usual square root on the nonnegative ray. From the preceding rules, it follows that $A^{1/2}$ is symmetric, positive semidefinite, and it squares to $A$: $(A^{1/2})^2 = A$.

Similarly to the arithmetic square root of a nonnegative real number $a$, i.e., the unique *nonnegative* real number which squares to $a$, the square root $A^{1/2}$ of a positive semidefinite matrix $A$ is the unique *positive semidefinite* matrix which squares to $A$.

**Fact** D.23 Let $A$ be a positive semidefinite $m \times m$ matrix, and $B$ be a positive semidefinite matrix such that $B^2 = A$. Then, $B = A^{1/2}$.

**Fact** D.24 Let $f(x)$ be a continuously differentiable real-valued function on an interval $(a, b)$ (where $-\infty \leq a < b \leq +\infty$) of the real axis. The function $\phi(X) = \mathrm{Tr}(f(X))$ is continuously differentiable on the open set $\mathcal{D}$ of $\mathbf{S}^m$ composed of all matrices with spectrum from $(a, b)$, and

$$\left.\frac{d}{dt}\right|_{t=0} \phi(X + tH) = \mathrm{Tr}(f'(X)H) \quad \forall X \in \mathcal{D}, H \in \mathbf{S}^m.$$

In other words, $\phi(\cdot)$ is continuously differentiable on the open set $\mathcal{D}$ and its gradient, taken w.r.t. the Frobenius Euclidean structure on $\mathbf{S}^m$, is $\nabla\phi(X) = f'(X)$.

*Hint:* In order to prove Fact D.24, it makes sense to verify its validity for algebraic polynomials and then use the fact that continuously differentiable real valued function on an interval $(a,b)$ of the real axis is the limit, in the sense of uniform convergence along with the first order derivative on compact subsets of $(a,b)$, of a sequence of polynomials.

**Remark** D.25 Note that when a complex-valued function $f(z)$ of complex-valued variable $z$ can be represented as the sum of everywhere converging power series (these functions are called *entire*):

$$f(z) = \sum_{\nu=0}^{\infty} c_\nu z^\nu, \qquad [\text{where } c_\nu \in \mathbf{C}]$$

we can "extend" $f$ to the function

$$f(Z) = \sum_{\nu=0}^{\infty} c_\nu Z^\nu$$

defined on the space $\mathbf{C}^{n\times n}$ of $n \times n$ matrices with complex entries and taking values in the same $\mathbf{C}^{n\times n}$, ensuring that for entire functions $f, g$ one has

$$\begin{array}{l} f(z) \equiv z^\nu \implies f(Z) = Z^\nu, \qquad \nu = 0, 1, \dots, \\ (f+g)(Z) \equiv f(Z) + g(Z), \\ (f\cdot g)(Z) \equiv f(Z)g(Z), \\ (\lambda f)(Z) \equiv \lambda f(Z), \ \ \lambda \in \mathbf{C}. \end{array}$$

In particular, for fixed $Z$, matrices $f(Z)$ stemming from various entire functions $f$ commute with each other.

Assuming an entire function $f$ real-valued on the real axis (this is the case if and only if the coefficients in the power series representing $f$ are real), the matrix $f(Z)$ for a real $n \times n$ matrix $Z$ also is real, and for real symmetric $Z$ is real symmetric as well. It is immediately seen that when $f$ is an entire function real-valued on the real axis and $Z$ is real symmetric matrix, both definitions of $f(Z)$ – our original definition "keep the eigenbasis intact and replace eigenvalues $\lambda$ with eigenvalues $f(\lambda)$" (this recipe works whenever $f$ is real-valued function well defined on the spectrum of $Z$), and the new definition via power series expansion result in the same matrix $f(Z)$.

**D.1.5.A Continuity of $f(Z)$ in $Z$.** So far we were interested in what happens with the matrix $f(Z)$ of a fixed symmetric $Z$ when $f$ changes. Now let us ask ourselves what happens with this matrix when $Z$ changes. We are about to consider the simplest question – continuity of $f(Z)$ in $Z$. Here is all we need in this respect:

**Proposition** D.26 Let $\Sigma$ be a nonempty closed subset of real axis, $f(\cdot):\Sigma\to\mathbf{R}$ be a continuous function, and $m$ be a positive integer. Then the set $\mathcal{Z}=\mathcal{Z}_\Sigma$ composed of all symmetric $m\times m$ matrices with spectrum from $\Sigma$ is closed, and the function $f(Z)$ of $Z\in\mathcal{Z}$ is continuous on $\mathcal{Z}$.

**Proof.** As we know from Fact D.21, the vector $\lambda(Z)$ of eigenvalues of $Z\in\mathbf{S}^m$ is Lipschitz continuous function of $Z$, implying that $\mathcal{Z}$ is closed. To prove continuity of $f(Z)$ in $Z\in\mathcal{Z}$, it is clearly sufficient to establish this fact then $\Sigma$ is nonempty, close and *bounded*. Thus, let $\Sigma$ be nonempty closed and bounded, $f$ be continuous on $\Sigma$, and let $Z_i\in\mathcal{Z}_\Sigma$ converge, as $i\to\infty$, to $\overline{Z}$; we should verify that $f(Z_i)\to f(\overline{Z})$ as $i\to\infty$. It is well-known that a continuous function on a nonempty compact subset of the real axis can be approximated, within whatever accuracy in the uniform norm, by an algebraic polynomial. Therefore, given $\epsilon>0$, we can find univariate algebraic polynomial $p(s)$ such that $|p(s)-f(s)|\leq\epsilon/3$ for all $s\in\Sigma$, implying, by item 4 of Section D.1.5, that $\|p(Z)-f(Z)\|\leq\epsilon/3$ for all $Z\in\mathcal{Z}$, $\|\cdot\|$ being the spectral norm. Since $p$ is an algebraic polynomial, we clearly have $p(Z_i)\to p(\overline{Z})$ as $i\to\infty$, implying that there exists $i_\epsilon$ such that $\|p(Z_i)-p(\overline{Z})\|\leq\epsilon/3$ when $i\geq i_\epsilon$. We conclude that $\|f(Z_i)-f(\overline{Z})\|\leq\|p(Z_i)-f(Z_i)\|+\|p(Z_i)-p(\overline{Z})\|+\|f(\overline{Z})-p(\overline{Z})\|\leq\epsilon$ when $i\geq i_\epsilon$. Thus, for every $\epsilon>0$ there exists $i_\epsilon$ such that $\|f(Z_i)-f(\overline{Z})\|\leq\epsilon$ for $i\geq i_\epsilon$, so that $f(Z_i)\to f(\overline{Z})$ as $i\to\infty$. □

## D.2 Positive semidefinite matrices and positive semidefinite cone

### *D.2.1 Positive semidefinite matrices.*

Recall that an $n\times n$ matrix $A$ is called *positive semidefinite* [notation: $A\succeq 0$], if $A$ is symmetric and produces nonnegative quadratic form:

$$A\succeq 0 \iff \left\{A=A^\top \quad\text{and}\quad x^\top Ax\geq 0\quad\forall x\right\}.$$

$A$ is called *positive definite* [notation: $A\succ 0$], if it is positive semidefinite and the corresponding quadratic form is positive outside the origin:

$$A\succ 0 \iff \left\{A=A^\top \quad\text{and}\quad x^\top Ax> 0\quad\forall x\neq 0\right\}.$$

The class of positive semidefinite matrices plays an important role in Mathematics. Thus, we next list a number of equivalent definitions of a positive semidefinite matrix.

**Theorem** D.27 For any $A\in\mathbf{S}^n$, the following properties of $A$ are equivalent to each other:

(i) $A\succeq 0$,

(ii) $\lambda(A)\geq 0$,

(iii) $A=D^\top D$ for certain rectangular matrix $D$,

(iv) $A=\Delta^\top\Delta$ for certain upper triangular $n\times n$ matrix $\Delta$,

(v) $A = B^2$ for certain symmetric matrix $B$,
(vi) $A = B^2$ for certain $B \succeq 0$.

Moreover, for any $A \in \mathbf{S}^n$, we also have the following properties equivalent to each other:
(i′) $A \succ 0$,
(ii′) $\lambda(A) > 0$,
(iii′) $A = D^\top D$ for certain rectangular matrix $D$ of rank $n$,
(iv′) $A = \Delta^\top \Delta$ for certain nonsingular upper triangular $n \times n$ matrix $\Delta$,
(v′) $A = B^2$ for certain nonsingular symmetric matrix $B$,
(vi′) $A = B^2$ for certain $B \succ 0$.

**Proof.** (i) $\Longleftrightarrow$ (ii): This equivalence is given by Proposition D.16.

(ii) $\Longrightarrow$ (vi): Suppose that we are in the case of (ii). Let $A = U\Lambda U^\top$ be the eigenvalue decomposition of $A$, so that $U$ is orthogonal and $\Lambda$ is diagonal with nonnegative diagonal entries $\lambda_i(A)$ (as we are in the case of (ii)). Let $\Lambda^{1/2}$ be the diagonal matrix with the diagonal entries $\lambda_i^{1/2}(A)$. Note that $(\Lambda^{1/2})^2 = \Lambda$. Then, the matrix $B = U\Lambda^{1/2}U^\top$ is symmetric with nonnegative eigenvalues $\lambda_i^{1/2}(A)$, and thus $B \succeq 0$ by Proposition D.16. Moreover,

$$B^2 = U\Lambda^{1/2}\underbrace{U^\top U}_{=I}\Lambda^{1/2}U^\top = U(\Lambda^{1/2})^2U^\top = U\Lambda U^\top = A,$$

as required in (vi).

(vi) $\Longrightarrow$ (v): Evident.

(v) $\Longrightarrow$ (iv): Suppose $A = B^2$ with certain symmetric $B$. Let $b_i$ be the $i$-th column of $B$ for all $i \leq n$. Applying the Gram-Schmidt orthogonalization process (see proof of Theorem A.31(ii)), we can find an orthonormal system of vectors $u_1, \ldots, u_n$ and lower triangular matrix $L$ such that

$$b_i = \sum_{j=1}^{i} L_{ij} u_j,$$

or, which is the same, $B^\top = LU$, where $U$ is the orthogonal matrix with the rows $u_1^\top, \ldots, u_n^\top$. Then, we have $A = B^2 = B^\top(B^\top)^\top = LUU^\top L^\top = LL^\top$. Recalling that $L$ is a lower triangular matrix, we see that $A = \Delta^\top\Delta$, where the matrix $\Delta = L^\top$ is upper triangular.

(iv) $\Longrightarrow$ (iii): Evident.

(iii) $\Longrightarrow$ (i): If $A = D^\top D$, then $x^\top Ax = (Dx)^\top(Dx) \geq 0$ for all $x$.

We have proved the equivalence of the properties (i) – (vi). Slightly modifying the reasoning (do it yourself!), one can prove the equivalence of the properties (i′) – (vi′). □

**Remark** D.28 (i) [Checking positive semidefiniteness] Given matrix $A \in \mathbf{S}^n$, we can check whether it is positive semidefinite by a purely algebraic finite algorithm (the so called *Lagrange diagonalization of a quadratic form*) which requires at most $O(n^3)$ arithmetic operations. Positive definiteness of a matrix can be checked also by the *Cholesky factorization algorithm* which finds the decomposition in (iv′), if it exists, in approximately $\frac{1}{6}n^3$ arithmetic operations.

There exists another useful algebraic criterion (*Sylvester's criterion*) for positive semidefiniteness of a matrix. According to this criterion, a symmetric matrix $A$ is positive definite if and only if all of its *angular (leading principal) minors* are positive, and $A$ is positive semidefinite if and only if all its *principal minors* are nonnegative. For example, a symmetric $2 \times 2$ matrix

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

is positive semidefinite if and only if $a \geq 0$, $c \geq 0$, and $\mathrm{Det}(A) := ac - b^2 \geq 0$. Another consequence is that when $A \succeq 0$, one has $A_{ii}A_{jj} \geq A_{ij}^2$ since the principal $2 \times 2$ minors of $A$ should be nonnegative.

(ii) [Square root of a positive semidefinite matrix] By the first chain of equivalences in Theorem D.27, a symmetric matrix $A$ is $\succeq 0$ if and only if $A$ is the square of a positive semidefinite matrix $B$. The latter matrix is uniquely defined by $A \succeq 0$ and is called the *square root* of $A$ [notation: $A^{1/2}$], see section D.1.5.

A symmetric matrix $A$ is called *negative semidefinite* [notation: $A \preceq 0$] if its negative, i.e., the matrix $-A$, is positive semidefinite. Similarly, $A$ is called *negative definite* [notation: $A \prec 0$] if $-A$ is positive definite.

### *D.2.2 The positive semidefinite cone*

When adding symmetric matrices and multiplying them by real numbers, we add, respectively multiply by real numbers, the corresponding quadratic forms. Hence, we arrive at the following basic fact.

**Fact** D.29 Let $A, B \in \mathbf{S}^n$ be positive semidefinite. Then, $A + B$ is also positive semidefinite. For any $\alpha \in \mathbf{R}_+$, the matrix $\alpha A$ is also positive semidefinite.

This fact is also the same as the following one stated in section 1.2.4.

**Fact** D.30 The set of $n \times n$ positive semidefinite matrices form a cone $\mathbf{S}^n_+$ in the Euclidean space $\mathbf{S}^n$ of symmetric $n \times n$ matrices based on the Euclidean structure being given by the Frobenius inner product $\langle A, B \rangle = \mathrm{Tr}(AB) = \sum_{i,j} A_{ij} B_{ij}$.

The cone $\mathbf{S}^n_+$ is called the *positive semidefinite cone* of size $n$. It is immediately seen that the semidefinite cone $\mathbf{S}^n_+$ is quite nice. Specifically, it satisfies the following properties:

- $\mathbf{S}^n_+$ is closed: the limit of a converging sequence of positive semidefinite matrices is positive semidefinite.
- $\mathbf{S}^n_+$ is pointed: the only $n \times n$ matrix $A$ such that both $A$ and $-A$ are positive semidefinite is the $n \times n$ matrix of all zeros.

- $\mathbf{S}^n_+$ possesses a nonempty interior (the subset composed of all interior points of $\mathbf{S}^n_+$) which is exactly the set of positive definite matrices.

In fact, $\mathbf{S}^n_+$ is a *regular* cone, see section 21.3. Note that the relation $A \succeq B$ means exactly that $A - B \in \mathbf{S}^n_+$, while $A \succ B$ is equivalent to $A - B \in \operatorname{int} \mathbf{S}^n_+$. The "matrix inequalities" $A \succeq B$ ($A \succ B$) match the standard properties of the usual scalar inequalities, e.g.,

$$\begin{array}{lr}
A \succeq A & \text{[reflexivity]} \\
A \succeq B,\ B \succeq A \implies A = B & \text{[anti-symmetry]} \\
A \succeq B,\ B \succeq C \implies A \succeq C & \text{[transitivity]} \\
A \succeq B,\ C \succeq D \implies A + C \succeq B + D & \text{[compatibility with linear operations, I]} \\
A \succeq B,\ \lambda \geq 0 \implies \lambda A \succeq \lambda B & \text{[compatibility with linear operations, II]} \\
A_i \succeq B_i,\ A_i \to A,\ B_i \to B \text{ as } i \to \infty \implies A \succeq B & \text{[closedness]}
\end{array}$$

with evident modifications when $\succeq$ is replaced with $\succ$, like

$$A \succeq B,\ C \succ D \implies A + C \succ B + D,$$

etc. Along with these standard properties of inequalities, the inequality $\succeq$ possesses the following nice additional property.

**Fact** D.31 For any $A, B \in \mathbf{S}^n$ such that $A \succeq B$, and for any $V \in \mathbf{R}^{n \times m}$, we always have

$$V^\top A V \succeq V^\top B V.$$

Another important property of the positive semidefinite cone is that its dual cone is equal to itself, i.e., the positive semidefinite cone is *self-dual*.

**Theorem** D.32 A matrix $X$ has nonnegative Frobenius inner products with all positive semidefinite matrices if and only if $X$ itself is positive semidefinite.

**Proof.** We first prove the "if" part. Assume that $X \succeq 0$, and let us prove that then $\operatorname{Tr}(XY) \geq 0$ for every $Y \succeq 0$. Using the eigenvalue decomposition of $X$, we can write it as

$$X = \sum_{i=1}^n \lambda_i(X) e_i e_i^\top,$$

where $e_i$ are the orthonormal eigenvectors of $X$. Then, we arrive at

$$\begin{aligned}
\operatorname{Tr}(XY) = \operatorname{Tr}\left(\left(\sum_{i=1}^n \lambda_i(X) e_i e_i^\top\right) Y\right) &= \sum_{i=1}^n \lambda_i(X) \operatorname{Tr}(e_i e_i^\top Y) \\
&= \sum_{i=1}^n \lambda_i(X) \operatorname{Tr}(e_i^\top Y e_i), \qquad \text{(D.6)}
\end{aligned}$$

where the last equality is given by Fact D.1. Recall that $e_i$s are just vectors, and thus $\mathrm{Tr}(e_i^\top Y e_i) = e_i^\top Y e_i$. Also, as $Y \succeq 0$, we have $e_i^\top Y e_i \geq 0$ for all $i$. Moreover, since $X \succeq 0$ by Proposition D.16, we have $\lambda_i(X) \geq 0$, and thus we conclude

$$\mathrm{Tr}(XY) = \sum_{i=1}^{n} \lambda_i(X)\mathrm{Tr}(e_i^\top Y e_i) \geq 0.$$

In order to prove the "only if" part, suppose that $X$ is such that $\mathrm{Tr}(XY) \geq 0$ for all matrices $Y \succeq 0$. Note that for every vector $y$, the matrix $Y = yy^\top$ is positive semidefinite (Theorem D.27.iii). Then, for any $y \in \mathbf{R}^n$ we have $0 \leq \mathrm{Tr}(Xyy^\top) = \mathrm{Tr}(y^\top Xy) = y^\top Xy$, where the first equality follows from Fact D.1. Thus, $X \succeq 0$ as desired. □

**Graphical illustration.** The single-dimensional positive semidefinite cone $\mathbf{S}_1^n$ is just the nonnegative ray, i.e.,

$$\mathbf{S}_+^1 = \{x \in \mathbf{R} : \; x \geq 0\} = \mathbf{R}_+.$$

The cone

$$\mathbf{S}_+^2 = \left\{ \begin{bmatrix} x & y \\ y & z \end{bmatrix} \in \mathbf{S}^2 : \; \begin{bmatrix} x & y \\ y & z \end{bmatrix} \succeq 0 \right\}$$

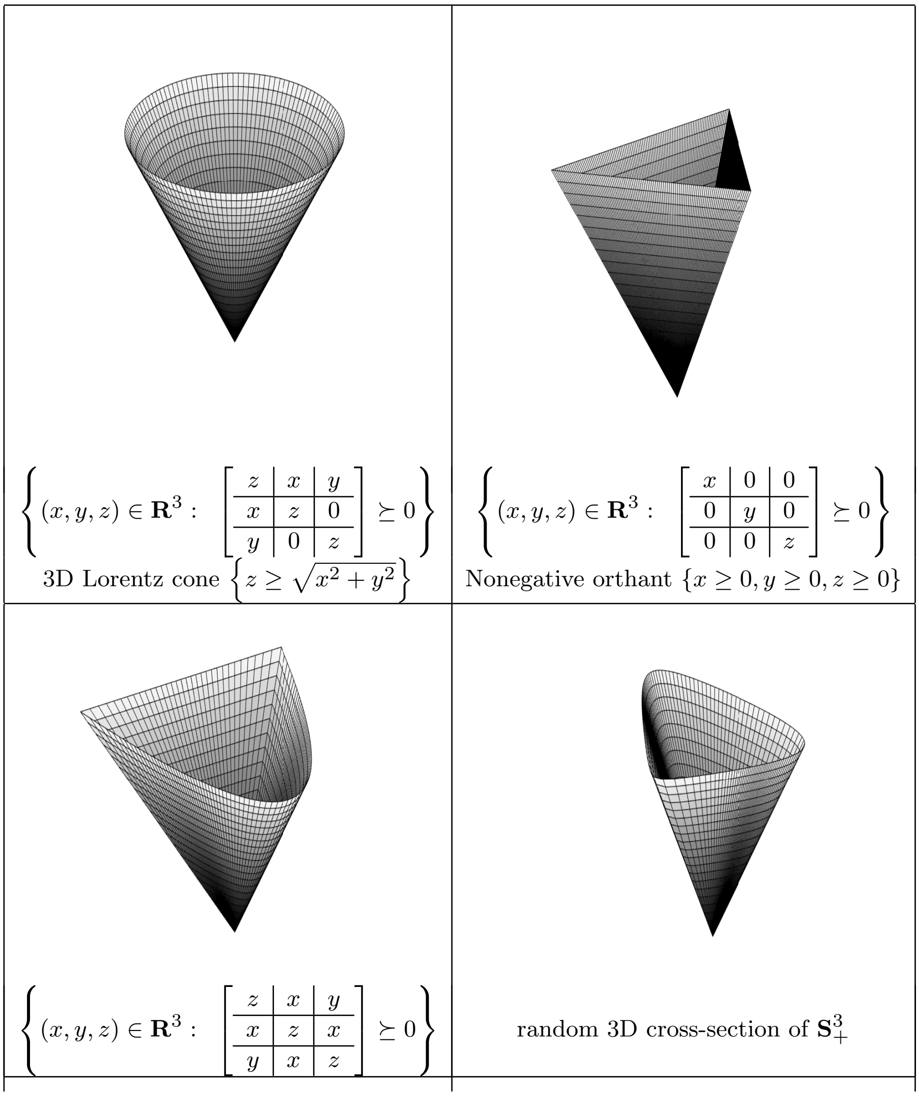
is nothing but the set

$$\begin{aligned} &\{(x,y,z) \in \mathbf{R}^3 : \; x \geq 0, \; z \geq 0, \; xz \geq y^2\} \\ &= \left\{(x,y,z) \in \mathbf{R}^3 : \; x + z \geq \sqrt{(x-z)^2 + 4y^2}\right\}. \end{aligned}$$

After linear invertible substitution of coordinates $w = x + z, \; u = x - z, \; v = 2y$, this last set becomes the *3D second-order*, or *3D Lorentz*, cone

$$\left\{(u,v,w) \in \mathbf{R}^3 : \; w \geq \sqrt{u^2 + v^2}\right\}.$$

The smallest "genuine" – not belonging to a simpler family of cones – semidefinite cone is the cone $\mathbf{S}^3$ of positive semidefinite $3 \times 3$ matrices. This cone, however, lives in 6-dimensional linear space $\mathbf{S}^3$, so that we cannot draw it in 3D. What we can draw, are 3D cross-section of $\mathbf{S}^3$ by various 3D planes, Several cross-sections of this type are shown on Figure V.2 illustrating that the range of convex sets that can be obtained by intersecting even the simplest semidefinite cone with affine planes is quite wide. In this respect it should be noted that the problems of minimizing linear objectives over the intersection of positive semidefinite cones and affine planes have a name – these are *semidefinite programs*, the subject of *Semidefinite Programming*, sometimes referred to as the "Linear Programming of XXI Century." We speak about Semidefinite Programming in more details in sections 23.4, 29.6 and exercises IV.20, IV.21, IV.26, IV.28, IV.35, among others. It is important to note that "expressive abilities" of Semidefinite Programming are extremely strong, and "for all practical purposes," it covers basically all convex problems arising in applications. The ultimate reason for this universality is the second-to-none richness of the family of convex sets which one can get by intersecting positive semidefinite cones with affine planes.



$$\left\{(x,y,z)\in\mathbf{R}^3:\ \left[\begin{array}{c|c|c} z & x & y \\ \hline x & z & 0 \\ \hline y & 0 & z \end{array}\right]\succeq 0\right\}$$

3D Lorentz cone $\left\{z\geq\sqrt{x^2+y^2}\right\}$

$$\left\{(x,y,z)\in\mathbf{R}^3:\ \left[\begin{array}{c|c|c} x & 0 & 0 \\ \hline 0 & y & 0 \\ \hline 0 & 0 & z \end{array}\right]\succeq 0\right\}$$

Nonegative orthant $\{x\geq 0, y\geq 0, z\geq 0\}$

$$\left\{(x,y,z)\in\mathbf{R}^3:\ \left[\begin{array}{c|c|c} z & x & y \\ \hline x & z & x \\ \hline y & x & z \end{array}\right]\succeq 0\right\}$$

random 3D cross-section of $\mathbf{S}^3_+$

Figure V.2. Several 3D cross-sections of $\mathbf{S}^3_+$

**Schur Complement Lemma.** Schur Complement Lemma is the following simple and extremely useful fact:

**Proposition** D.33 [Schur Complement Lemma] Consider a symmetric $2\times 2$ block matrix

$$A=\left[\begin{array}{c|c} P & Q \\ \hline Q^\top & R \end{array}\right]$$

with $R\succ 0$. Then, $A$ is positive definite (semidefinite) if and only if the matrix

$$P - QR^{-1}Q^\top$$

is positive definite (resp. semidefinite).

**Proof.** Suppose $P \in \mathbf{S}^p$, $R \in \mathbf{S}^q$. Then, splitting the vector $x$ from $\mathbf{R}^{p+q}$ into blocks $u \in \mathbf{R}^p$, $v \in \mathbf{R}^q$, we arrive at

$$\begin{aligned}
& P \succeq 0 \\
\Longleftrightarrow\ & u^\top Pu + 2u^\top Qv + v^\top Rv \geq 0 \quad \forall (u,v) \\
\Longleftrightarrow\ & \min_v \left\{ v^\top Rv + 2v^\top Qu + u^\top Pu \right\} \geq 0 \quad \forall u \\
\Longleftrightarrow\ & \min_v \left\{ (v + R^{-1}Qu)^\top R(v + R^{-1}Qu) + u^\top Pu - u^\top QR^{-1}Q^\top u \right\} \geq 0 \quad \forall u \\
\Longleftrightarrow\ & u^\top Pu - u^\top QR^{-1}Q^\top u + \min_v \Big\{ \underbrace{(v + R^{-1}Qu)^\top R(v + R^{-1}Qu)}_{\geq 0 \text{ as } R \succ 0} \Big\} \geq 0 \quad \forall u \\
\Longleftrightarrow\ & P - QR^{-1}Q^\top \succeq 0.
\end{aligned}$$

Here, in the last step we observe that for any $u \in \mathbf{R}^p$, by selecting $v = -R^{-1}Qu$ we see that the optimum value of the minimization problem is equal to zero. As we concluded $P - QR^{-1}Q^\top \succeq 0$, this proves the "positive semidefinite" version of Schur Complement Lemma. The same reasoning, with evident modifications, justifies the "positive definite" version. □

## D.3 Exercises

Exercise 12 1. Find the dimension of $\mathbf{R}^{m \times n}$ and point out a basis in this space.
2. Build an orthonormal basis in $\mathbf{S}^m$.

Exercise 13 In the space $\mathbf{R}^{m \times m}$ of square $m \times m$ matrices, there are two interesting subsets: the set $\mathbf{S}^m$ of *symmetric* matrices $\left\{A : A = A^\top\right\}$ and the set $\mathbf{J}^m$ of *skew-symmetric* matrices $\{A = [A_{ij}] : A_{ij} = -A_{ji},\ \forall i, j\}$.

1. Verify that both $\mathbf{S}^m$ and $\mathbf{J}^m$ are linear subspaces of $\mathbf{R}^{m \times m}$.
2. Find the dimension of $\mathbf{S}^m$ and point out a basis in $\mathbf{S}^m$.
3. Find the dimension of $\mathbf{J}^m$ and point out a basis in $\mathbf{J}^m$.
4. What is the sum of $\mathbf{S}^m$ and $\mathbf{J}^m$? What is the intersection of $\mathbf{S}^m$ and $\mathbf{J}^m$?

Exercise 14 Is the "3-factor" extension of Fact D.1 valid, at least in the case of square matrices $X, Y, Z$ of the same size? That is, for square matrices $X, Y, Z$ of the same size, is it always true that $\mathrm{Tr}(XYZ) = \mathrm{Tr}(YXZ)$?

Exercise 15 Given $P \in \mathbf{S}^p$, $Q \in \mathbf{R}^{r \times p}$, and $R \in \mathbf{S}^r$, consider the matrices

$$A = \begin{bmatrix} P & Q^\top \\ Q & R \end{bmatrix}, \quad B = \begin{bmatrix} P & -Q^\top \\ -Q & R \end{bmatrix}, \quad C = \begin{bmatrix} R & Q \\ Q^\top & P \end{bmatrix}, \quad D = \begin{bmatrix} R & -Q \\ -Q^\top & P \end{bmatrix}.$$

Prove that $\lambda(A) = \lambda(B) = \lambda(C) = \lambda(D)$. Thus, the matrices $A, B, C, D$ simultaneously are/are not positive semidefinite. As a consequence: Schur Complement Lemma says that when $R \succ 0$, one has $A \succeq 0$ iff $P - Q^\top R^{-1}Q \succeq 0$; since $A \succeq 0$ iff $C \succeq 0$, we see that the same Lemma says that when $P \succ 0$, one has $A \succeq 0$ iff $R - QP^{-1}Q^\top \succeq 0$.

Exercise 16 Let $\mathbf{S}^n_{++} := \operatorname{int}\mathbf{S}^n_+ = \{X \in \mathbf{S}^n : X \succ 0\}$, and consider $X, Y \in \mathbf{S}^n_{++}$. Then, $X \preceq Y$ holds if and only if $X^{-1} \succeq Y^{-1}$ ("$\succeq$-antimonotonicity of $X^{-1}$, $X \in \mathbf{S}^n_{++}$). Is it true that from $0 \prec X \preceq Y$ it always follows that $X^{-2} \succeq Y^{-2}$?

Exercise 17 Let $A, B \in \mathbf{S}^n$ be such that $0 \preceq A \preceq B$. For each one of the following, either prove the statement or produce a counter example:

1. $A^2 \preceq B^2$;
2. $0 \preceq A^{1/2} \preceq B^{1/2}$.

Exercise 18 A matrix $A \in \mathbf{S}^n$ is called *diagonally dominant* if it satisfies the relation

$$a_{ii} \geq \sum\nolimits_{j \neq i} |a_{ij}|, \quad i = 1, \ldots, n.$$

Prove that every diagonally dominant matrix $A$ is positive semidefinite.

Exercise 19 Prove the following matrix analogy of the scalar inequality $ab \leq \frac{a^2+b^2}{2}$ for $a, b \in \mathbf{R}$:

$$AB^\top + BA^\top \preceq AA^\top + BB^\top, \qquad \forall A, B \in \mathbf{R}^{m \times n}.$$

Exercise 20 1. Let $I_k$ denote the $k \times k$ identity matrix, and let $A$ be an $m \times n$ matrix. Prove that the following three properties are equivalent to each other:

- $A^\top A \preceq I_n$;
- $AA^\top \preceq I_m$;
- $\begin{bmatrix} I_m & A \\ A^\top & I_n \end{bmatrix} \succeq 0.$

2. Let $A_1, \ldots, A_k$ be $n \times n$ matrices such that

$$A_1^\top A_1 + \ldots + A_k^\top A_k \preceq I_n.$$

For each one of the following, either prove the statement or produce a counter example:

- $A_1 A_1^\top + \ldots + A_k A_k^\top \preceq I_n$;
- $\begin{bmatrix} A_1A_1^\top & A_1A_2^\top & \cdots & A_1A_k^\top \\ A_2A_1^\top & A_2A_2^\top & \cdots & A_2A_k^\top \\ \vdots & \vdots & \ddots & \vdots \\ A_kA_1^\top & A_kA_2^\top & \cdots & A_kA_k^\top \end{bmatrix} \preceq I_{kn}.$

## D.4 Proofs of Facts

**Fact D.1** Let $X, Y$ be rectangular matrices such that $XY$ makes sense and is a square matrix. Then, $\operatorname{Tr}(YX)$ also makes sense and

$$\operatorname{Tr}(XY) = \operatorname{Tr}(YX).$$

Proof. In order for $XY$ to make sense and be a square matrix, the sizes of the matrices should be $m \times n$ (for $X$) and $n \times m$ (for $Y$) with some $m$ and $n$, We now have

$$\operatorname{Tr}(XY) = \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}Y_{ji}, \quad \operatorname{Tr}(YX) = \sum_{j=1}^{n}\sum_{i=1}^{m} Y_{ji}X_{ij},$$

that is, both quantities are the same. □

**Fact D.2.** If $X, Y \in \mathbf{R}^{m \times n}$, the Frobenius inner product of $X$ and $Y$ is equal to the Frobenius inner product of $X^\top$ and $Y^\top$:

$$\operatorname{Tr}(XY^\top) = \operatorname{Tr}\left((X^\top)(Y^\top)^\top\right).$$

Moreover, when $U$ is an orthogonal $m \times m$ matrix (i.e., $UU^\top = U^\top U = I_m$, or, which is the same as $U^{-1} = U^\top$, and $V$ is an orthogonal $n \times n$ matrix (i.e., $VV^\top = V^\top V = I_n$), the Frobenius inner product of $UXV$ and $UYV$ is the same as the Frobenius inner product of $X$ and $Y$:

$$\mathrm{Tr}(XY^\top) = \mathrm{Tr}\left((UXV)(UYV)^\top\right).$$

Proof. We have $\mathrm{Tr}((X^\top)(Y^\top)^\top) = \mathrm{Tr}(X^\top Y) = \mathrm{Tr}((X^\top Y)^\top) = \mathrm{Tr}(Y^\top X) = \mathrm{Tr}(XY^\top)$, where the concluding inequality is given by Fact D.1. Similarly, given orthogonal matrices $U \in \mathbf{R}^{m\times m}$ and $V \in \mathbf{R}^{n\times n}$, we have

$$\mathrm{Tr}((UXV)(UYV)^\top) = \mathrm{Tr}(UX(VV^\top)Y^\top U^\top) = \mathrm{Tr}(XY^\top(U^\top U)) = \mathrm{Tr}(XY^\top),$$

where the first and the last equalities follow from the fact that $U, V$ are orthogonal matrices and thus $V^\top V = VV^\top = I_n$ and $U^\top U = UU^\top = I_m$, and the second equality follows from Fact D.1. □

**Fact D.10.** Let $A, B \in \mathbf{R}^{n\times n}$ be commuting and $\lambda$ be a real eigenvalue of $A$. Then, the spectral subspace $E = \{x \in \mathbf{R}^n : \; Ax = \lambda x\}$ of $A$ corresponding to $\lambda$ is invariant for $B$ (i.e., $Be \in E$ for every $e \in E$).

Proof. For $e \in E$ we have $A(Be) = B(Ae) = B(\lambda e) = \lambda(Be)$, that is $Be \in E$. □

**Fact D.11.** If $A$ is an $n \times n$ matrix and $L$ is an invariant subspace of $A$ (i.e., $L$ is a linear subspace such that $Ae \in L$ whenever $e \in L$), then the orthogonal complement $L^\perp$ of $L$ is invariant for the matrix $A^\top$. In particular, if $A$ is symmetric and $L$ is invariant subspace of $A$, then $L^\perp$ is an invariant subspace of $A^\top$ as well.

Proof. Suppose $L$ is invariant for $A$. Recall that $x \in L^\perp$ if and only if $x^\top y = 0$ for all $y \in L$. Therefore, for all $x \in L^\perp$ we have

$$y \in L \implies Ay \in L \implies 0 = x^\top Ay = (A^\top x)^\top y,$$

where the first implication is due to $L$ being an invariant subspace of $A$, and the second implication is due to $Ay \in L$ and $x \in L^\perp$. Then, from $(A^\top x)^\top y = 0$ for all $x \in L^\perp$ we deduce that $A^\top x \in L^\perp$. □

**Fact D.9.** The matrices $A_1, \ldots, A_k \in \mathbf{S}^n$ commute with each other ($A_iA_j = A_jA_i$ for all $i, j$) if and only if they can be "simultaneously diagonalized," i.e., there exist a single orthogonal matrix $U$ and diagonal matrices $\Lambda_1$,. . . ,$\Lambda_k$ such that

$$A_i = U\Lambda_i U^\top,\ i = 1, \ldots, k.$$

Proof. The fact that when the matrices are simultaneously diagnosable, then they commute, is evident: for diagonal matrices $D$, $E$ and an orthogonal matrix $U$ we have

$$(U^\top DU)(U^\top EU) = U^\top D(U^\top U)EU = U^\top DEU,$$

and similarly $(U^\top EU)(U^\top DU) = U^\top EDU$; it remains to note that $ED = DE$ as $D, E$ are diagonal matrices.

In the opposite direction, let us carry out induction in the number $k$ of commuting symmetric matrices. There is nothing to prove when $k = 1$, or, more precisely, the case of $k = 1$ is covered by the results on eigenvalue decomposition of symmetric matrix. Inductive step: Assuming the statement valid for some $k$ and given $k + 1$ commuting matrices $A_1, \ldots, A_{k+1}$, let $\lambda_1, \ldots, \lambda_s$ be distinct eigenvalues of $A_{k+1}$, and $E_1, \ldots, E_s$ be the corresponding spectral subspaces of $A_{k+1}$, that is, restricted onto $E_\ell$, the mapping $x \mapsto A_{k+1}x$ acts as multiplication by $\lambda_\ell$. By Fact

D.10, subspaces $E_\ell$ are invariant for matrices $A_1,\dots,A_k$. Then, by the inductive hypothesis, we deduce that for every $\ell \le s$ there exists an orthonormal basis in $E_\ell$ with basic vectors being eigenvalues of every one of $A_1,\dots,A_k$. In addition, these vectors, as all vectors from $E_\ell$, are eigenvectors of $A_{k+1}$. So, the union, over $\ell \le s$ of vectors from these orthonormal bases form an orthonormal basis in the entire space (recall that $E_1,...,E_s$ are mutually orthogonal, and their sum is the entire $\mathbf{R}^n$), and in this basis all matrices $A_1,\dots,A_{k+1}$ become diagonal. □

**Fact D.19** (i) Let $E$ and $F$ be finite-dimensional linear spaces. The induced norm $\|\cdot\|_{F,E}$ is indeed a norm on the space $\mathrm{Lin}(E,F)$ of linear mappings from $E$ to $F$.
(ii) Let $E,F,G$ be finite-dimensional linear spaces equipped with the norms $\|\cdot\|_E$, $\|\cdot\|_F$, $\|\cdot\|_G$ respectively. Let $y = Ax : E \to F$ and $z = By : F \to G$ be linear mappings. Then,

$$\|BA\|_{G,E} \le \|B\|_{G,F}\,\|A\|_{F,E}.$$

Proof. (i) Homogeneity and positivity of $\|\cdot\|_{F,E}$ are evident. To prove the Triangle inequality, consider $A,B \in \mathrm{Lin}(E,F)$. Then, for any $x$ we have

$$\|(A+B)x\|_F = \|Ax+Bx\|_F \le \|Ax\|_F + \|Bx\|_F \le (\|A\|_{F,E} + \|B\|_{F,E})\,\|x\|_E.$$

This implies $\|(A+B)\|_{F,E} \le \|A\|_{F,E} + \|B\|_{F,E}$ as desired.
(ii) Indeed, for every $x \in E$ one has $\|BAx\|_G \le \|B\|_{G,F}\,\|Ax\|_F \le \|B\|_{G,F}\,\|A\|_{F,E}\,\|x\|_E$. □

**Fact D.20** Let $\|\cdot\|$ be the spectral norm on $\mathbf{R}^{m\times n}$.
(i) For any $A \in \mathbf{R}^{m\times n}$, we have

$$\|A\| = \max_{x,y}\left\{y^\top Ax :\ \|x\|_2 \le 1,\ \|y\|_2 \le 1\right\},$$

hence also $\|A\| = \|A^\top\|$.
(ii) For any $A \in \mathbf{S}^n$, we have $\|A\| = \max\{|\lambda_{\max}(A)|\,,\,|\lambda_{\min}(A)|\}$. Moreover, for any $A \in \mathbf{R}^{m\times n}$ we have

$$\|A\|^2 = \|A^\top A\| = \lambda_{\max}(A^\top A) = \lambda_{\max}(AA^\top) = \|AA^\top\| = \|A^\top\|^2.$$

Proof.
(i) Indeed, $\|Ax\|_2 = \max\limits_y\left\{y^\top Ax : \|y\|_2 \le 1\right\}$ by Cauchy inequality (Theorem B.1), implying that $\max\limits_x\{\|Ax\|_2 : \|x\|_2 \le 1\} = \max_{x,y}\left\{y^\top Ax :\ \|x\|_2 \le 1,\ \|y\|_2 \le 1\right\}$.

(ii) Indeed, let $A = U\,\mathrm{Diag}\{\lambda\}U^\top$ be the eigenvalue decomposition of the symmetric matrix $A$. Hence, $y^\top Ax = (U^\top y)^\top\,\mathrm{Diag}\{\lambda\}(U^\top x)$ with orthogonal $U$. Thus,

$$\begin{aligned}
&\max_{x,y}\left\{y^\top Ax :\ \|x\|_2 \le 1,\ \|y\|_2 \le 1\right\}\\
&= \max_{u,v}\left\{v^\top\,\mathrm{Diag}\{\lambda\}u :\ \|u\|_2 \le 1,\ \|v\|_2 \le 1\right\}\\
&= \max_{u,v}\left\{\sum_i \lambda_i u_i v_i :\ \|u\|_2 \le 1,\ \|v\|_2 \le 1\right\}\\
&\le (\max_i |\lambda_i|)\max_{u,v}\left\{\sum_i |u_i v_i| :\ \|u\|_2 \le 1,\ \|v\|_2 \le 1\right\}\\
&\le \max_i |\lambda_i|,
\end{aligned}$$

where the last inequality is due to Cauchy inequality. Note also that $y^\top Ax = \max_i |\lambda_i|$ when $x$ is the unit norm eigenvector of $A$ corresponding to the maximum magnitude eigenvalue, $\lambda_{i_*}$, of

$A$, and $y = \mathrm{sign}(\lambda_{i_*})x$. Thus, when $A$ is symmetric, we do have $\|A\| = \max_i |\lambda_i(A)|$. For an arbitrary matrix $A$ we have

$$\|A\|^2 = \max_{x:\|x\|_2\leq 1} \|Ax\|_2^2 = \max_{x:\|x\|_2\leq 1} x^\top(A^\top A)x = \lambda_{\max}(A^\top A),$$

where the last equality is given by the variational characterization of the largest eigenvalue of symmetric positive semidefinite matrix $A^\top A$ (Theorem D.12). Moreover, $\|A^\top A\| = \lambda_{\max}(A^\top A)$ due to the already proved "symmetric" part of (ii). Thus, $\|A\| = \sqrt{\|A^\top A\|} = \sqrt{\lambda_{\max}(A^\top A)}$, Recalling that $\|A\| = \|A^\top\|$, we complete the verification of (ii). □

**Fact D.21.** The vector-valued function $A \mapsto \lambda(A) : \mathbf{S}^n \to \mathbf{R}^n$ is Lipschitz continuous, specifically, denoting by $\|\cdot\|$ the spectral norm, for all $k \leq n$, we have

$$|\lambda_k(A) - \lambda_k(A')| \leq \|A - A'\| \quad \forall (A, A' \in \mathbf{S}^n).$$

Proof. Indeed, by VCE, denoting $\mathcal{E}_k$ the family of all subspaces of codimension $k-1$ in $\mathbf{R}^n$, we have

$$\lambda_k(A) = \min_{E\in\mathcal{E}_k} \max_{e\in E:\|e\|_2=1} e^\top Ae,$$

and the right hand side is clearly Lipschitz continuous, with constant 1 w.r.t. $\|\cdot\|$, function of $A$. [1] □

**Fact D.23** Let $A$ be a positive semidefinite $m\times m$ matrix, and $B$ be a positive semidefinite matrix such that $B^2 = A$. Then, $B = A^{1/2}$.

Proof. When $B^2 = A$, $B$ clearly commutes with $A$, implying by Fact D.9 that the matrices admit simultaneous diagonalization, i.e.,

$$A = U\,\mathrm{Diag}\{\lambda_1,\ldots,\lambda_m\}U^\top, \quad B = U\,\mathrm{Diag}\{\upsilon_1,\ldots,\upsilon_m\}U^\top$$

where $U$ is an orthogonal matrix. Thus, the relation $A = B^2$ reads

$$\begin{aligned} U\,\mathrm{Diag}\{\lambda_1,\ldots,\lambda_m\}U^\top &= (U\,\mathrm{Diag}\{\upsilon_1,\ldots,\upsilon_m\}U^\top)(U\,\mathrm{Diag}\{\upsilon_1,\ldots,\upsilon_m\}U^\top) \\ &= U\,\mathrm{Diag}\{\upsilon_1,\ldots,\upsilon_m\}\underbrace{(U^\top U)}_{=I_m}\mathrm{Diag}\{\upsilon_1,\ldots,\upsilon_m\}U^\top \\ &= U\,\mathrm{Diag}\{\upsilon_1^2,\ldots,\upsilon_m^2\}U^\top, \end{aligned}$$

and so $\lambda_i = \upsilon_i^2$ for all $i$. Moreover, $B$ is positive semidefinite, implying that $\upsilon_i \geq 0$. The bottom line is that $\upsilon_i = \lambda_i^{1/2}$ for all $i$, that is, $B = A^{1/2}$. □

**Fact D.24** Let $f(x)$ be a continuously differentiable real-valued function on an interval $(a,b)$, (where $-\infty \leq a < b \leq +\infty$) of the real axis. The function $\phi(X) = \mathrm{Tr}(f(X))$ is continuously differentiable on the open set $\mathcal{D}$ of $\mathbf{S}^m$ composed of all

[1] Justification of "clearly" is based upon two nearly evident facts (check them!): (1) whenever $e \in \mathbf{R}^m, h \in \mathbf{R}^n$, the function $e^\top Bh$ is Lipschitz continuous, with constant $\|h\|_2\|e\|_2$ w.r.t. $\|\cdot\|$, function of $B \in \mathbf{R}^{m\times n}$, and (2) Let $Q \subset \mathbf{R}^k$, $\|\cdot\|$ be a norm on $\mathbf{R}^k$, and $\{f_\alpha(\cdot) : \alpha \in \mathcal{A} \neq \varnothing\}$ be a family of Lipschitz continuous, with common constant $L$ with respect to $\|\cdot\|$, real-valued functions on $Q$, If the function $f(x) = \sup_{\alpha\in\mathcal{A}} f_\alpha(x)$ is finite at some point of $Q$, it is finite everywhere on $Q$ and is Lipschitz continuous, with constant $L$, w.r.t. $\|\cdot\|$. It should be added that in the latter statement, $Q$ can be replaced with abstract metric space (a nonempty set equipped with distance $d(\cdot,\cdot)$, possessing properties 1 – 3, see page 374, between points of $Q$, with Lipschitz continuity, with constant $L$ w.r.t. $d$, of function $f : Q \to \mathbf{R}$ meaning that $|f(x) - f(x')| \leq Ld(xc, x')$ for all $x, x' \in Q$).

matrices with spectrum from $(a, b)$, and

$$\left.\frac{d}{dt}\right|_{t=0} \phi(X + tH) = \mathrm{Tr}(f'(X)H) \quad \forall X \in \mathcal{D}, H \in \mathbf{S}^m.$$

In other words, $\phi(\cdot)$ is continuously differentiable on the open set $\mathcal{D}$ and its gradient, taken w.r.t. the Frobenius Euclidean structure on $\mathbf{S}^m$, is $\nabla\phi(X) = f'(X)$.

*Hint:* In order to prove Fact D.24, it makes sense to verify its validity for algebraic polynomials and then use the fact that continuously differentiable real valued function on an interval $(a, b)$ of the real axis is the limit, in the sense of uniform convergence along with the first order derivative on compact subsets of $(a, b)$, of a sequence of polynomials.

Proof. Following the hint, let us start with the case when $f(x) = \sum_{i=0}^k a_i x^i$ is an algebraic polynomial. In this case the matrix-valued function $f(X) = a_0 I_m + a_1 X + ... + a_k X^k : \mathbf{S}^m \to \mathbf{S}^m$ is clearly infinitely many times differentiable, and for all $X, H \in \mathbf{S}^m$ we have

$$\left.\frac{d}{dt}\right|_{t=0} f(X + tH) = \sum_{i=1}^k a_i \big(\sum_{j=1}^i X^{j-1} H X^{i-j}\big),$$

and thus

$$\begin{aligned}
&\left.\frac{d}{dt}\right|_{t=0} \phi(X + tH) = \left.\frac{d}{dt}\right|_{t=0} \mathrm{Tr}(f(X + tH)) \\
&= \mathrm{Tr}\left(\left.\frac{d}{dt}\right|_{t=0} f(X + tH)\right) && \text{[as Tr}(Z)\text{ is a linear function of }Z] \\
&= \mathrm{Tr}\left(\sum_{i=1}^k a_i \big(\sum_{j=1}^i X^{j-1} H X^{i-j}\big)\right) \\
&= \sum_{i=1}^k a_i \sum_{j=1}^i \mathrm{Tr}\big(X^{j-1} H X^{i-j}\big) \\
&= \sum_{i=1}^k a_i \sum_{j=1}^i \mathrm{Tr}(H X^{i-1}) && \text{[by Fact D.1]} \\
&= \mathrm{Tr}\left(H\Big(\sum_{i=1}^k a_i i X^{i-1}\Big)\right) \\
&= \mathrm{Tr}(H f'(X)).
\end{aligned}$$

Next, Corollary D.22 implies that $\mathcal{D}$ is an open set. Now let $f$ be continuously differentiable real-valued function on the interval $(a, b) \subseteq \mathbf{R}$, and $f_t(x)$, $t = 1, 2, \ldots$, be polynomials such that $f_t(\cdot)$ and $f_t'(\cdot)$ converge as $t \to \infty$, uniformly on every segment $[a', b'] \subset (a, b)$, to $f(\cdot)$ and $f'(\cdot)$, respectively; it is well known that such a sequence exists. Taking into account that the spectral norm (maximum of magnitudes of eigenvalues) of $g(X)$ is upper-bounded by the uniform norm of $g(\cdot)$ on $\sigma(X)$, we immediately conclude that as $t \to \infty$, the functions $\phi_t(x) := \mathrm{Tr}(f_t(X))$ converge to $\phi(X)$ uniformly on closed and bounded subsets of $\mathcal{D}$, and the gradients $\nabla\phi_t(X) = f_t'(X)$ converge to $f'(X)$ uniformly on closed and bounded subsets of $\mathcal{D}$. By the standard results of Analysis, it follows that $\phi(\cdot)$ is continuously differentiable on $\mathcal{D}$, its gradient is given by $\nabla\phi(X) = f'(X)$. □

**Fact D.31** For any $A, B \in \mathbf{S}^n$ such that $A \succeq B$, and for any $V \in \mathbf{R}^{n\times m}$, we always have

$$V^\top AV \succeq V^\top BV.$$

Proof. Indeed, we should prove that if $A-B \succeq 0$, then also $V^\top(A-B)V \succeq 0$. This is immediate as the quadratic form $y^\top(V^\top(A-B)V)y = (Vy)^\top(A-B)(Vy)$ of $y$ is nonnegative along with the quadratic form $x^\top(A-B)x$ of $x$. □

# References

[Axl15] S. Axler, *Linear algebra done right, 3rd edition*, Undergraduate Texts in Mathematics, Springer, 2015.

[BNO03] D. Bertsekas, A. Nedic, and A. Özdağlar, *Convex analysis and optimization*, Athena Scientific, 2003.

[BTN] A. Ben-Tal and A. Nemirovski, *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, SIAM 2001 and `https://www.isye.gatech.edu/~nemirovs/LMCOLN.pdf` 2023.

[BV04] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[Edw12] C.H. Edwards, *Advanced calculus of several variables*, Courier Corporation, 2012.

[Gel89] I.M. Gel'fand, *Lectures on linear algebra*, Dover Books on Mathematics, Dover Publications, Inc, 1989.

[HUL93] J.-B. Hiriart-Urruty and C. Lemarechal, *Convex analysis and minimization algorithms, I: Fundamentals, II: Advanced theory and bundle methods*, Springer, 1993.

[IT79] A.D. Ioffe and V.M. Tikhomirov, *Theory of extremal problems*, Nauka, 1974 (in Russian) English translation: Studies in Mathematics and its Applications, v.6, North-Holland, 1979.

[Nem24] A. Nemirovski, *Introduction to linear optimization*, World Scientific, 2024 and `https://www.isye.gatech.edu/~nemirovs/WSbook.pdf`, 2024.

[Nes18] Yu. Nesterov, *Lectures on convex optimization, 2nd edition*, Springer, 2018.

[Pas22] D. S. Passman, *Lectures on linear algebra*, World Scientific, 2022.

[Roc70] R.T. Rockafellar, *Convex analysis*, Mathematics, Princeton University Press, 1970.

[Rud13] W. Rudin, *Principles of mathematical analysis, 3rd edition*, McGraw Hill India, 2013.

[Str06] G. Strang, *Linear algebra and its applications.*, Belmont, CA: Thomson, Brooks/Cole, 2006.

# Index

# Exercises from Part I

## Elementaries

**Exercise I.1.** Mark in the following list the sets which are convex:

1. $\{x \in \mathbf{R}^2 : x_1 + i^2 x_2 \leq 1,\, i = 1, \dots, 10\}$

   *Solution:* convex

2. $\{x \in \mathbf{R}^2 : x_1^2 + 2ix_1x_2 + i^2x_2^2 \leq 1,\, i = 1, \dots, 10\}$

   *Solution:* convex. Here is an equivalent description where convexity is evident: $\{x : |x_1 + ix_2| \leq 1,\, i = 1, \dots, 10\}$.

3. $\{x \in \mathbf{R}^2 : x_1^2 + ix_1x_2 + i^2x_2^2 \leq 1,\, i = 1, \dots, 10\}$

   *Solution:* convex (it is the intersection of ellipses)

4. $\{x \in \mathbf{R}^2 : x_1^2 + 5x_1x_2 + 4x_2^2 \leq 1\}$

   *Solution:* nonconvex

5. $\left\{x \in \mathbf{R}^{10} : x_1^2 + 2x_2^2 + 3x_3^2 + \ldots + 10x_{10}^2 \leq 1000x_1 - 999x_2 + 998x_3 - \ldots + 992x_9 - 991x_{10}\right\}$

   *Solution:* convex (ellipsoid)

6. $\{x \in \mathbf{R}^2 : \exp\{x_1\} \leq x_2\}$

   *Solution:* convex

7. $\{x \in \mathbf{R}^2 : \exp\{x_1\} \geq x_2\}$

   *Solution:* nonconvex

8. $\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n x_i^2 = 1\}$

   *Solution:* nonconvex

9. $\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n x_i^2 \leq 1\}$

   *Solution:* convex

10. $\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n x_i^2 \geq 1\}$

    *Solution:* nonconvex

11. $\{x \in \mathbf{R}^n : \max\limits_{i=1,\dots,n} x_i \leq 1\}$

    *Solution:* convex

12. $\{x \in \mathbf{R}^n : \max\limits_{i=1,\dots,n} x_i \geq 1\}$

    *Solution:* nonconvex, except for $n = 1$

13. $\{x \in \mathbf{R}^n : \max\limits_{i=1,\dots,n} x_i = 1\}$

    *Solution:* nonconvex, except for $n = 1$

14. $\{x \in \mathbf{R}^n : \min\limits_{i=1,\dots,n} x_i \leq 1\}$

*Solution:* nonconvex, except for $n = 1$

15. $\{x \in \mathbf{R}^n : \min\limits_{i=1,\dots,n} x_i \geq 1\}$

    *Solution:* convex

16. $\{x \in \mathbf{R}^n : \min\limits_{i=1,\dots,n} x_i = 1\}$

    *Solution:* nonconvex, except for $n = 1$

**Exercise I.2.** Mark by $\mathbf{T}$ those of the following claims which are always true:

1. The linear image $Y = \{Ax : x \in X\}$ of a linear subspace $X$ is a linear subspace. *Solution:* $\mathbf{T}$
2. The linear image $Y = \{Ax : x \in X\}$ of an affine subspace $X$ is an affine subspace. *Solution:* $\mathbf{T}$
3. The linear image $Y = \{Ax : x \in X\}$ of a convex set $X$ is convex. *Solution:* $\mathbf{T}$
4. The affine image $Y = \{Ax + b : x \in X\}$ of a linear subspace $X$ is a linear subspace.
5. The affine image $Y = \{Ax + b : x \in X\}$ of an affine subspace $X$ is an affine subspace. *Solution:* $\mathbf{T}$
6. The affine image $Y = \{Ax + b : x \in X\}$ of a convex set $X$ is convex. *Solution:* $\mathbf{T}$
7. The intersection of two linear subspaces in $\mathbf{R}^n$ is always nonempty. *Solution:* $\mathbf{T}$
8. The intersection of two linear subspaces in $\mathbf{R}^n$ is a linear subspace. *Solution:* $\mathbf{T}$
9. The intersection of two affine subspaces in $\mathbf{R}^n$ is an affine subspace.
10. The intersection of two affine subspaces in $\mathbf{R}^n$, when nonempty, is an affine subspace. *Solution:* $\mathbf{T}$
11. The intersection of two convex sets in $\mathbf{R}^n$ is a convex set. *Solution:* $\mathbf{T}$
12. The intersection of two convex sets in $\mathbf{R}^n$, when nonempty, is a convex set. *Solution:* $\mathbf{T}$

**Exercise I.3.** Prove that the relative interior of a simplex with vertices $y^0, \dots, y^m$ is exactly the set

$$\left\{\sum_{i=0}^{m} \lambda_i y_i : \lambda_i > 0, \sum_{i=0}^{m} \lambda_i = 1\right\}.$$

*Solution:* The claim is evident for the standard simplex $\Delta_m := \{x \in \mathbf{R}^m_+ : \sum_i x_i \leq 1\}$. Moreover, the set $\Delta := \mathrm{Conv}\{y^0, \dots, y^m\}$ is the image of $\Delta_m$ under the affine mapping

$$x \mapsto A(x) = y^0 + \sum_{i=1}^{m} x_i(y^i - y^0) : \mathbf{R}^m \to \mathbf{R}^{\dim(y)},$$

which is a one-to-one affine correspondence between $\mathbf{R}^m$ and $\mathrm{Aff}\{y^0, \dots, y^m\}$, and such a correspondence clearly maps the relative interiors of convex sets in the argument space onto the relative interiors of their images in the image space.

**Exercise I.4** Which of the following claims is true:

1. The set $X = \{x : Ax \leq b\}$ is a cone if and only if $X = \{x : Ax \leq 0\}$.
2. The set $X = \{x : Ax \leq b\}$ is a cone if and only if $b = 0$.

*Solution:* The claim in item 1 is correct, while the claim in item 2 is not. Let us show that the claim in item 1 is correct. We immediately see that if $X = \{x : Ax \leq 0\}$, then $X$ is clearly a cone. To see the other direction, suppose that the set $X = \{x : Ax \leq b\}$ is a cone. Then $0 \in X$, so that $b \geq 0$, and therefore the set $\overline{X} := \{x : Ax \leq 0\}$ is contained in $X$. Moreover, for any $x \in X$, as $X$ is a cone we have that $tx \in X$ for all $t > 0$, and so $Ax \leq t^{-1}b$ for all $t > 0$. Then, by taking the limit of both sides of this latter inequality as $t \to +\infty$ we conclude that $Ax \leq 0$. Therefore, $X \subseteq \overline{X}$, the bottom line being that $X = \overline{X} = \{x : Ax \leq 0\}$.
A counterexample for the claim in item 2 is, e.g., $A = [1; 1]$, $b = [0; 1]$, so that $Ax \leq b$ is the system of two univariate linear inequalities $x \leq 0$, $x \leq 1$; here the solution set is a cone, but $b \neq 0$.

**Exercise I.5** Suppose $\mathbf{K}$ is a closed cone. Prove that the set $X = \{x : Ax - b \in \mathbf{K}\}$ is a cone if and only if $X = \{x : Ax \in \mathbf{K}\}$.

*Solution:* Follows the same argument as in Exercise I.4.1.

**Exercise I.6.** Prove that if $M$ is a nonempty convex set in $\mathbf{R}^n$ and $\epsilon > 0$, then for every norm $\|\cdot\|$ on $\mathbf{R}^n$, the $\epsilon$-neighborhood of $M$, i.e., the set

$$M_\epsilon = \left\{ y \in \mathbf{R}^n : \inf_{x\in M} \|y - x\| \le \epsilon \right\},$$

is convex.

*Solution:* Consider any $y', y'' \in M_\epsilon$ and any $\lambda \in [0; 1]$; we should prove that $y := \lambda y' + (1-\lambda)y'' \in M_\epsilon$. As $y', y'' \in M_\epsilon$, using the definition of the set $M_\epsilon$ we deduce that for every $\delta > 0$ there exist $x'_\delta \in M$ and $x''_\delta \in X$ such that $\|y' - x'_\delta\| \le \epsilon + \delta$ and $\|y'' - x''_\delta\| \le \epsilon + \delta$. Hence,

$$\begin{aligned}\|y - \underbrace{[\lambda x'_\delta + (1-\lambda)x''_\delta]}_{:=x_\delta}\| &= \|\lambda[y' - x'_\delta] + (1-\lambda)[y'' - x''_\delta]\| \\ &\le \lambda\|y' - x'_\delta\| + (1-\lambda)\|y'' - x''_\delta\| \\ &\le \lambda(\epsilon+\delta) + (1-\lambda)(\epsilon+\delta) = \epsilon + \delta.\end{aligned}$$

Also, as $M$ is convex, $x_\delta \in M$. Thus, we see that for every $\delta > 0$ there is a point $x_\delta \in M$ such that $\|y - x_\delta\| \le \epsilon + \delta$, and since $\delta > 0$ is arbitrary, we conclude that $\inf_{x\in M}\|y - x\| \le \epsilon$, that is, $y \in M_\epsilon$.

**Exercise I.7.** Which of the following claims are always true? Explain why/why not.

1. The convex hull of a bounded set in $\mathbf{R}^n$ is bounded.

   *Solution:* yes.

2. The convex hull of a closed set in $\mathbf{R}^n$ is closed.

   *Solution:* not necessarily. Consider the set $X := \{x \in \mathbf{R}^2 : x_2 \ge |x_1|^{-1}, x_1 \ne 0\}$. Note that $X$ is closed yet its convex hull is the open half-plane $\{x \in \mathbf{R}^2 : x_2 > 0\}$.

3. The convex hull of a closed convex set in $\mathbf{R}^n$ is closed.

   *Solution:* yes. And the color of white horse of Alexander the Great is “white.”

4. The convex hull of a closed and bounded set in $\mathbf{R}^n$ is closed and bounded.

   *Solution:* yes, see Corollary I.2.5.

5. The convex hull of an open set in $\mathbf{R}^n$ is open.

   *Solution:* yes

**Exercise I.8.** Let $A$, $B$ be nonempty subsets of $\mathbf{R}^n$. Consider the following claims. If the claim is always (i.e., for every data satisfying premise of the claim) true, give a proof; otherwise, give a counter example.

1. If $A \subseteq B$, then $\mathrm{Conv}(A) \subseteq \mathrm{Conv}(B)$.
   *Solution:* evidently true.
2. If $\mathrm{Conv}(A) \subseteq \mathrm{Conv}(B)$, then $A \subseteq B$.
   *Solution:* evidently false. Consider $n = 1$, $A = \{1, 2, 3\}$, $B = \{1, 3\}$.
3. $\mathrm{Conv}(A \cap B) = \mathrm{Conv}(A) \cap \mathrm{Conv}(B)$.
   *Solution:* evidently false. Consider $n = 1$, $A = \{0, 2\}$, $B = \{1, 3\}$, resulting in $\mathrm{Conv}(A \cap B) = \varnothing$ and $\mathrm{Conv}(A) \cap \mathrm{Conv}(B) = [1, 2]$.
4. $\mathrm{Conv}(A \cap B) \subseteq \mathrm{Conv}(A) \cap \mathrm{Conv}(B)$.
   *Solution:* evidently true, since $A\cap B \subseteq A$, we have $\mathrm{Conv}(A\cap B) \subseteq \mathrm{Conv}(A)$. Similarly, $\mathrm{Conv}(A\cap B) \subseteq \mathrm{Conv}(B)$.
5. $\mathrm{Conv}(A \cup B) \subseteq \mathrm{Conv}(A) \cup \mathrm{Conv}(B)$.
   *Solution:* evidently false. Consider $n = 1$, $A = \{0\}$, $B = \{1\}$.
6. $\mathrm{Conv}(A \cup B) \supseteq \mathrm{Conv}(A) \cup \mathrm{Conv}(B)$.
   *Solution:* evidently true: since $A \cup B$ contains $A$, we have $\mathrm{Conv}(A \cup B) \supseteq \mathrm{Conv}(A)$, and similarly $\mathrm{Conv}(A \cup B) \supseteq \mathrm{Conv}(B)$.
7. If $A$ is closed, so is $\mathrm{Conv}(A)$.
   *Solution:* false, see Remark I.2.6.

8. If $A$ is closed and bounded, so is $\mathrm{Conv}(A)$.
   *Solution:* true, see Corollary I.2.5.
9. If $\mathrm{Conv}(A)$ is closed and bounded, so is $A$.
   *Solution:* evidently false. Consider $A = [0, 1/2) \cup \{1\}$.

**Exercise I.9.** Let $A, B, C$ be nonempty subsets of $\mathbf{R}^n$ and $D$ be a nonempty subset of $\mathbf{R}^m$. Consider the following claims. If the claim is always (i.e., for every data satisfying premise of the claim) true, give a proof; otherwise, give a counter example.

1. $\mathrm{Conv}(A \cup B) = \mathrm{Conv}(\mathrm{Conv}(A) \cup B)$.

   *Solution:* true. Since $A \subseteq \mathrm{Conv}(A)$ and $B \subseteq B$, we have $(A \cup B) \subseteq (\mathrm{Conv}(A) \cup B)$ and so $\mathrm{Conv}(A \cup B) \subseteq \mathrm{Conv}(\mathrm{Conv}(A) \cup B)$. To see the other direction, note that the set $\mathrm{Conv}(A \cup B)$ clearly contains both $\mathrm{Conv}(A)$ and $B$, that is, $\mathrm{Conv}(A \cup B) \supseteq (\mathrm{Conv}(A) \cup B)$. Moreover, $\mathrm{Conv}(A \cup B)$ is convex, implying $\mathrm{Conv}(A \cup B) \supseteq \mathrm{Conv}(\mathrm{Conv}(A) \cup B)$.

2. $\mathrm{Conv}(A \cup B) = \mathrm{Conv}(\mathrm{Conv}(A) \cup \mathrm{Conv}(B))$.

   *Solution:* true. Applying the preceding part twice, we get

$$\mathrm{Conv}(A \cup B) = \mathrm{Conv}(\mathrm{Conv}(A) \cup B) = \mathrm{Conv}(B \cup \mathrm{Conv}(A)) = \mathrm{Conv}(\mathrm{Conv}(B) \cup \mathrm{Conv}(A)).$$

3. $\mathrm{Conv}(A \cup B \cup C) = \mathrm{Conv}(\mathrm{Conv}(A \cup B) \cup C)$.

   *Solution:* true. Applying the first part of this exercise, we get

$$\mathrm{Conv}(A \cup B \cup C) = \mathrm{Conv}((A \cup B) \cup C) = \mathrm{Conv}(\mathrm{Conv}(A \cup B) \cup C).$$

4. $\mathrm{Conv}(A \times D) = \mathrm{Conv}(A) \times \mathrm{Conv}(D)$.

   *Solution:* true. Indeed, $A \times D \subseteq \mathrm{Conv}(A) \times \mathrm{Conv}(D)$. Moreover, $\mathrm{Conv}(A) \times \mathrm{Conv}(D)$ is convex, implying that $\mathrm{Conv}(A \times D) \subseteq \mathrm{Conv}(A) \times \mathrm{Conv}(D)$. To see the reverse direction, consider a point $z \in (\mathrm{Conv}(A) \times \mathrm{Conv}(D))$. Then, $z = [\sum_i \lambda_i a^i; \sum_j \mu_j d^j]$ for some weights $\lambda_i \geq 0$ summing up to 1 and $a^i \in A$, and for some weights $\mu_j \geq 0$ summing up to 1 and $d^j \in D$. Hence, $z = \sum_{i,j} \lambda_i \mu_j [a^i; d^j]$, and since $\lambda_i \mu_j \geq 0$ and $\sum_{i,j} \lambda_i \mu_j = 1$, we see that $z \in \mathrm{Conv}(A \times D)$. Thus, $\mathrm{Conv}(A) \times \mathrm{Conv}(D) \subseteq \mathrm{Conv}(A \times D)$.

5. When $A$ is convex, to get the set $\mathrm{Conv}(A \cup B)$ (which is always the set of convex combinations of several points from $A$ and several points from $B$), it suffices to take convex combinations of points with *at most one of them* taken from $A$, and the rest taken from $B$. Similarly, if $A$ and $B$ are both convex, to get $\mathrm{Conv}(A \cup B)$, it suffices to add to $A \cup B$ all convex combinations of pairs of points, one from $A$ and one from $B$.

   *Solution:* Both claims are true. Indeed, $\mathrm{Conv}(A \cup B)$ is the set of all convex combinations of finite collections of points, some from $A$ and the rest from $B$. Consider such a collection $z = \sum_{i \in I} \lambda_i a^i + \sum_{j \in J} \mu_j b^j$, where $I$, $J$ are sets of indices, $\lambda_i$ are nonnegative and $a^i \in A$, $i \in I$, $\mu_j$ are nonnegative and $b^j \in B$, $j \in J$, and the total sum of all $\lambda_i$ and $\mu_j$ is 1. Justifying the first claim boils down to verifying that when $A$ is convex, we can restrict $I$ to be of cardinality 0 or 1. Indeed, if $\sum_{i \in I} \lambda_i = 0$, $z$ is convex combination of points from $B$, and if $\alpha := \sum_{i \in I} \lambda_i > 0$, we can write $\sum_{i \in I} \lambda_i a^i = \alpha a$, where $a := \sum_{i \in I} \frac{\lambda_i}{\alpha} a^i$ is a point from $A$ (since $A$ is convex), that is, $z$ can be represented as convex combination $\alpha a + \sum_{i \in J} \mu_i b^i$ of a collection where one point is from $A$, and all remaining points are from $B$, as required.

   Similarly, to justify the second claim, we should verify that when $A$ and $B$ are convex, the above $z$ is either a point from $A$, or from $B$, or a convex combination of two points, one from $A$ and one from $B$. When $\alpha := \sum_{i \in I} \lambda_i = 0$ or $\beta := \sum_{i \in J} \mu_i = 0$, the initial representation of $z$ is in fact the representation of the point as convex combination of points from $B$, resp., from $A$, that is, either is a point from $B$, or a point from $A$, or both. And when $\alpha > 0$ and $\beta > 0$, we have, same as in the first claim, $z = \alpha a + \beta b$ with $a \in A$, $b \in B$, and of course $\alpha + \beta = 1$. That is, $z$ is convex combination of a point from $A$ and a point from $B$.

6. Suppose $A$ is a set in $\mathbf{R}^n$. Consider the affine mapping $x \mapsto Px + p : \mathbf{R}^n \to \mathbf{R}^m$, and the image of $A$ under this mapping, i.e., the set $PA + p := \{Px + p : x \in A\}$. Then, $\mathrm{Conv}(PA + p) = P\,\mathrm{Conv}(A) + p$.

*Solution:* trivially true. Here is the justification:

$$\begin{aligned}
\mathrm{Conv}\{PA+p\} &= \left\{\sum_i \lambda_i y^i : \ \lambda_i \geq 0, \ \sum_i \lambda_i = 1, \ y^i \in PA+p, \forall i\right\} \\
&= \left\{\underbrace{\sum_i \lambda_i (Px^i + p)}_{=P(\sum_i \lambda_i x^i)+p} : \ \lambda_i \geq 0, \ \sum_i \lambda_i = 1, \ x^i \in A, \forall i\right\} \\
&= \left\{Px + p : \ x = \sum_i \lambda_i x^i, \lambda \geq 0, \ \sum_i \lambda_i = 1\right\} \\
&= P\,\mathrm{Conv}(A) + p.
\end{aligned}$$

7. Consider an affine mapping $y \mapsto P(y) : \mathbf{R}^m \to \mathbf{R}^n$ where $P(y) := Py+p$. Recall that given a set $X \in \mathbf{R}^n$, its inverse image under the mapping $P(\cdot)$ is given by $P^{-1}(X) := \{y \in \mathbf{R}^m : \ P(y) \in X\}$. Then, $\mathrm{Conv}(P^{-1}(A)) = P^{-1}(\mathrm{Conv}(A))$.

   *Solution:* clearly false. Consider $m = n = 1$, $Px + p \equiv 0$, and $A = \{-1, 1\}$. Note that in this case as $0 \notin A$ we have $P^{-1}(A) = \varnothing$ and so $\mathrm{Conv}(P^{-1}(A)) = \varnothing$. On the other hand, $\mathrm{Conv}(A) = [-1, 1]$ and so $0 \in \mathrm{Conv}(A)$ and $P^{-1}(\mathrm{Conv}(A)) = \mathbf{R}^m$.

8. Consider an affine mapping $y \mapsto P(y) : \mathbf{R}^m \to \mathbf{R}^n$ where $P(y) := Py+p$. Then, $\mathrm{Conv}(P^{-1}(A)) \subseteq P^{-1}(\mathrm{Conv}(A))$.

   *Solution:* clearly true. Consider any $z \in \mathrm{Conv}(P^{-1}(A))$; then $z$ is a convex combination of points from $P^{-1}(A)$, that is, $Pz + p$ is a convex combination of points from $A$.

**Exercise I.10** Let $X_1, X_2 \in \mathbf{R}^n$ be two nonempty sets, and define $Y := X_1 \cup X_2$ and $Z := \mathrm{Conv}(Y)$. Consider the following claims. If the claim is always (i.e., for every data satisfying premise of the claim) true, give a proof; otherwise, give a counter example.

1. Whenever $X_1$ and $X_2$ are both convex, so is $Y$.

   *Solution:* Obviously false. Take $n = 1$, and $X_1 := \{-1\}$ and $X_2 := \{+1\}$.

2. Whenever $X_1$ and $X_2$ are both convex, so is $Z$.

   *Solution:* Obviously true by definition of $Z$.

3. Whenever $X_1$ and $X_2$ are both bounded, so is $Y$.

   *Solution:* Obviously true.

4. Whenever $X_1$ and $X_2$ are both bounded, so is $Z$.

   *Solution:* Obviously true.

5. Whenever $X_1$ and $X_2$ are both closed, so is $Y$.

   *Solution:* Obviously true - closedness is preserved by taking finite unions.

6. Whenever $X_1$ and $X_2$ are both closed, so is $Z$.

   *Solution:* This is false as $Z$ is not necessarily closed. Indeed, this claim is not valid even when $X_1, X_2$ are nonempty polyhedral, but not bounded, sets. For example, by selecting $n = 2$, $X_1 := \{x \in \mathbf{R}^2 : x_1 \geq 0, x_2 = 0\}$ and $X_2 := \{[0; 1]\}$, we see that the set $\mathrm{Conv}(X_1 \cup X_2)$ is not polyhedral, but its closure is.

7. Whenever $X_1$ and $X_2$ are both compact, so is $Y$.

   *Solution:* Obviously true – $Y$ is closed and bounded along with $X_1$ and $X_2$.

8. Whenever $X_1$ and $X_2$ are both compact, so is $Z$.

   *Solution:* Obviously true – by previous item, $Y$ is compact, so that $Z$ is compact by Corollary I.2.5.

9. Whenever $X_1$ and $X_2$ are both polyhedral, so is $Y$.

*Solution:* Obviously false. Take $n = 1$, and $X_1 := \{-1\}$ and $X_2 := \{+1\}$.

10. Whenever $X_1$ and $X_2$ are both polyhedral, so is $Z$.

    *Solution:* This is false as $Z$ is not necessarily closed, see solution to item 6, and closedness for a polyhedral set is a must.

11. Whenever $X_1$ and $X_2$ are both polyhedral and bounded, so is $Y$.

    *Solution:* Obviously false. Take $n = 1$, and $X_1 := \{-1\}$ and $X_2 := \{+1\}$.

12. Whenever $X_1$ and $X_2$ are both polyhedral and bounded, so is $Z$.

    *Solution:* This claim is indeed true, see solution to Exercise I.22.2 for a proof.

**Exercise I.11.** Consider two families of convex sets given by $\{F_i\}_{i\in I}$ and $\{G_j\}_{j\in J}$. Prove that the following relation holds:

$$\mathrm{Conv}\left(\bigcup_{i\in I,\, j\in J}(F_i \cap G_j)\right) \subseteq \mathrm{Conv}\left(\bigcup_{j\in J}[G_j \cap \mathrm{Conv}(\cup_{i\in I} F_i)]\right).$$

*Solution:* Note that for all $j \in J$ and for all $i' \in I$, we have

$$(F_{i'} \cap G_j) \subseteq [G_j \cap (\cup_{i\in I} F_i)] \subseteq [G_j \cap \mathrm{Conv}(\cup_{i\in I} F_i)],$$

and so for all $j \in J$

$$\bigcup_{i'\in I}(F_{i'} \cap G_j) \subseteq [G_j \cap \mathrm{Conv}(\cup_{i\in I} F_i)].$$

By first taking the union of both sides over $j \in J$ and then taking the convex hull of the resulting sets, we arrive at the desired relation.

**Exercise I.12.** Let $C_1, C_2$ be two nonempty conic sets in $\mathbf{R}^n$, i.e., for each $i = 1, 2$, for any $x \in C_i$ and $t \geq 0$, we have $t \cdot x \in C_i$ as well. Note that $C_1, C_2$ are not necessarily convex. Prove that

1. $C_1 + C_2 \neq \mathrm{Conv}(C_1 \cup C_2)$ may happen if either $C_1$ or $C_2$ (or both) is nonconvex.

   *Solution:* Let $C_1$ be the origin in $\mathbf{R}^2$, and $C_2$ be the union of nonnegative rays of the coordinate axes. Here both sets are nonempty and conic, their sum is $C_2$, and the convex hull of their union (which is $C_2$) is the first quadrant.

2. $C_1 + C_2 = \mathrm{Conv}(C_1 \cup C_2)$ always holds if $C_1, C_2$ are both convex.

   *Solution:* When $C_1, C_2$ are nonempty and convex, we have by Exercise I.9.5 that $\mathrm{Conv}(C_1 \cup C_2) = \{x = \alpha y + (1-\alpha)z : y \in C_1, z \in C_2, \alpha \in [0,1]\}$, whence $\mathrm{Conv}(C_1 \cup C_2) = C_1 \cup C_2 \cup \{x = \alpha y + (1-\alpha)z : y \in C_1, z \in C_2, \alpha \in (0,1)\} = C_1 \cup C_2 \cup \big(\cup_{\alpha\in(0,1)}[\alpha C_1 + (1-\alpha)C_2]\big)$. When $C_1, C_2$, in addition to being nonempty and convex, are also conic, for $\alpha \in (0,1)$ it holds $\alpha C_1 + (1-\alpha)C_2 = C_1 + C_2$, so that the above computation results in $\mathrm{Conv}(C_1 \cup C_2) = C_1 \cup C_2 \cup [C_1 + C_2]$. The latter union is just $C_1 + C_2$, since $C_1 + C_2$ contains both $C_1$ and $C_2$ (as a nonempty conic set contains the origin).

3. When the nonempty conic sets $C_1, C_2$ are convex, the equality $C_1 \cap C_2 = \bigcup_{\alpha\in(0,1)\in[0,1]}(\alpha C_1 \cap (1-\alpha)C_2)$ always holds if $C_1, C_2$ are both convex.

   *Solution:* We have

   $\cup_{\alpha\in[0,1]}[\alpha C_1 \cap (1-\alpha)C_2] = [0 \cdot C_1 \cap 1 \cdot C_2] \cup [1 \cdot C_1 \cap 0 \cdot C_2] \cup \big(\cup_{\alpha\in(0,1)}[\alpha C_1 \cap (1-\alpha)C_2]\big)$,
   and the set in parentheses $\big(\big)$ is just $C_1 \cap C_2$ due to the conicity of $C_1$, $C_2$. Besides this, as it was mentioned when solving item 2, $0 \in C_1 \cap C_2$, so that $[0 \cdot C_1 \cap C_2] = [C_1 \cap 0 \cdot C_2] = \{0\} \subset C_1 \cap C_2$. The bottom line is that $\cup_{\alpha\in[0,1]}[\alpha C_1 \cap (1-\alpha)C_2] = C_1 \cap C_2$, as claimed,

**Exercise I.13.** Let $X \subseteq \mathbf{R}^n$ be a convex set with $\mathrm{int}\, X \neq \varnothing$, and consider the following set

$$\mathbf{K} := \mathrm{cl}\,\{[x;t] :\ t > 0,\ x/t \in X\}.$$

Prove that the set $\mathbf{K}$ is a closed cone with a nonempty interior.

*Solution:* $\mathbf{K}$ is what was in section 1.5 called closed conic transform of $X$; it was shown in section 1.5 that $\mathbf{K}$ is a closed cone. When $\overline{x} \in \mathrm{int}\, X$, we clearly have $[\overline{x}; 1] \in \mathrm{int}\, \mathbf{K}$, so that $\mathrm{int}\, \mathbf{K} \neq \varnothing$ whenever $\mathrm{int}\, X \neq \varnothing$.

## Around ellipsoids

**Exercise I.14.** Verify each of the following statements:

1. Any ellipsoid $E \in \mathbf{R}^n$ is the image of the unit Euclidean ball $B_n = \{x \in \mathbf{R}^n : \|x\|_2 \leq 1\}$ under a one-to-one affine mapping. That is, $E \subset \mathbf{R}^n$ can be represented as $E = \{x : (x-c)^\top C(x-c) \leq 1\}$ with $C \succ 0$ and $c \in \mathbf{R}^n$ if and only if it can be represented as $E = \{c + Du : u \in B_n\}$ with nonsingular $D$, and in the latter representation $D$ can be selected to be symmetric positive definite.

   *Solution:* Let $E = \{x : (x-c)^\top C(x-c) \leq 1\}$ with $C \succ 0$. Then, by defining $H := C^{1/2}$ (see section D.1.5) we have

   $$E = \{x : (H(x-c))^\top (\underbrace{H(x-c)}_{:=u}) \leq 1\} = \{x = c + \underbrace{H^{-1}}_{=:D} u : u^\top u \leq 1\}$$

   where $D = H^{-1} \succ 0$ as $C \succ 0$. For the other direction, given a nonsingular $D$, to say that $x = c + Du$ with some $u$ satisfying $\|u\|_2 \leq 1$, is the same as to say that $\|D^{-1}(x-c)\|_2 \leq 1$, that is, the same as to say that $(x-c)^\top \underbrace{D^{-\top} D^{-1}}_{=:C}(x-c) \leq 1$ (by definition, $D^{-\top} = (D^{-1})^\top$), and $C := D^{-\top} D^{-1}$ is symmetric positive definite since $D^{-1}$ is nonsingular.

2. Given $C \succ 0$, $D \succ 0$ and $c, d \in \mathbf{R}^n$, the ellipsoid $E_C := \{x : (x-c)^\top C(x-c) \leq 1\}$ is contained in the ellipsoid $E_D := \{x : (x-c)^\top D(x-c) \leq 1\}$ if and only if $C \succeq D$. If the ellipsoid $E_C$ is contained in the ellipsoid $E'_D = \{x : (x-d)^\top D(x-d) \leq 1\}$, then $C \succeq D$.

   *Solution:* The first claim: Setting $x = y + c$, we should prove that with positive definite $C, D$, the implication $y^\top Cy \leq 1 \implies y^\top Dy \leq 1$ holds true if and only if $C \succeq D$. By homogeneity, the implication in question is the same as the relation

   $$\forall (s, y : s > 0, y^\top Cy \leq s) : \quad y^\top Dy \leq s,$$

   which for $C \succ 0$ is exactly the same as $C \succeq D$.
   The second claim: Suppose $E_C \subseteq E'_D$. Then, using part 1,

   $$\begin{aligned} u^\top u \leq 1 \quad &\iff c + C^{-1/2}u \in E_C \\ &\implies c + C^{-1/2}u \in E'_D \\ &\implies (C^{-1/2}u + c - d)^\top D(C^{-1/2}u + c - d) \leq 1 \\ &\implies (u+f)^\top \underbrace{(C^{-1/2}DC^{-1/2})}_{=:H}(u+f), \; f := C^{1/2}(c-d). \end{aligned}$$

   Applying the resulting inequality to $-u$ in the role of $u$, we conclude that

   $$u^\top u \leq 1 \implies (f \pm u)^\top H(f \pm u) \leq 1,$$

   whence

   $$u^\top u \leq 1 \implies 1 \geq \frac{1}{2}\left((f+u)^\top H(f+u) + (f-u)^\top H(f-u)\right) = u^\top Hu + f^\top Hf \geq u^\top Hu,$$

   where the concluding inequality is due to $H \succ 0$ (implied by $C, D \succ 0$). Hence, we arrive at

   $$I \succeq H = C^{-1/2}DC^{-1/2},$$

   implying, after multiplying both sides from the right and from the left by the symmetric matrix $C^{1/2}$, that $C \succeq D$, as claimed.

3. For a set $U \subset \mathbf{R}^n$, let $\mathrm{Vol}(U)$ be the ratio of the $n$-dimensional volume of $U$ and the $n$-dimensional volume of the unit ball $B_n$. Then, for an $n$-dimensional ellipsoid $E$ represented as $\{x = c + Du : \|u\|_2 \leq 1\}$ with nonsingular $D$ we have

   $$\mathrm{Vol}(E) = |\mathrm{Det}(D)|,$$

and when $E$ is represented as $\{x : (x-c)^\top C(x-c) \leq 1\}$ with $C \succ 0$, we have

$$\mathrm{Vol}(E) = \mathrm{Det}^{-1/2}(C).$$

*Solution:* The first relation is evident – one-to-one affine transformation $u \mapsto c + Du$ multiplies $n$-dimensional volumes by $\mathrm{Det}(D)$. Using item 1, we see that the second representation of $E$ is equivalent to the first representation with $D := C^{-1/2}$, so that the second representation of $\mathrm{Vol}(E)$ is readily given by the first one.

**Exercise I.15.** Given $C \succ 0$, an ellipsoid $\{x : (x-a)^\top C(x-a) \leq 1\}$ is the solution set of quadratic inequality $x^\top Cx - 2(Ca)^\top x + (a^\top Ca - 1) \leq 0$. Prove that the solution set $E$ of any quadratic inequality $f(x) := x^\top Cx - c^\top x + \sigma \leq 0$ with positive *semi*definite matrix $C$ is convex.

*Solution:* Let $x, y \in E$ and $\lambda \in [0,1]$. Then,

$$\begin{aligned}
&f(\lambda x + (1-\lambda)y)\\
&= \left(\lambda^2 x^\top Cx + \lambda(1-\lambda)x^\top Cy + \lambda(1-\lambda)y^\top Cx + (1-\lambda)^2 y^\top Cy\right)\\
&\quad - \lambda c^\top x - (1-\lambda)c^\top y + \lambda\sigma + (1-\lambda)\sigma\\
&= \lambda(x^\top Cx - c^\top x + \sigma) + (1-\lambda)(y^\top Cy - c^\top y + \sigma) - \underbrace{\lambda(1-\lambda)((x-y)^\top C(x-y))}_{\geq 0 \text{ due to } C \succeq 0}\\
&\leq \lambda \underbrace{f(x)}_{\leq 0} + (1-\lambda)\underbrace{f(y)}_{\leq 0} \leq 0.
\end{aligned}$$

That is, $\lambda x + (1-\lambda)y \in E$.

## Truss Topology Design

**Exercise I.16.** [First acquaintance with Truss Topology Design]
**Preamble.** What follows is the first exercise in a "Truss Topology Design" (TTD) series ((other exercises in it are I.18, III.9, IV.11, IV.28). The underlying "real life" mechanical story is simple enough to be told and rich enough to illustrate numerous constructions and results presented in the main body of our textbook – ranging from Caratheodory Theorem to semidefinite duality, demonstrating on a real life example how the theory works.
**Trusses.** Truss is a mechanical construction, like railroad bridge, electric mast, of Eiffel Tower, composed of thin elastic *bars* linked with each other at *nodes* – points from physical space (3D space for spatial, and 2D space for planar trusses).

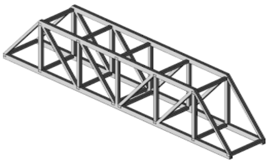


Figure I.7. Pratt Truss Bridge
source: `https://grabcad.com/library/pratt-truss-bridge-2`

When truss is subject to external load – collection of forces acting at the nodes – it starts to deform, so that the nodes move a little bit, leading to elongations/shortenings of bars, which, in turn, result in reaction forces. At the equilibrium, the reaction forces compensate the external ones, and the truss capacitates certain potential energy, called *compliance*. Mechanics models this story as follows.

- The nodes form a finite set $p_1, \ldots, p_K$ of distinct points in physical space $\mathbf{R}^d$ ($d = 2$ for planar, and $d = 3$ for spatial constructions). Virtual displacements of the nodes under the load are somehow restricted by "support conditions;" we will focus on the case when some of the nodes "are fixed" – cannot move at all (think about them as being in the wall), and the remaining "are free" – their virtual displacements form the entire $\mathbf{R}^d$. A virtual displacement $v$ of the nodal set can be identified with a vector of dimension $M = dm$, where $m$ is the number of free nodes; $v$ is block vector with $m$ $d$-dimensional blocks, indexed by the free nodes, representing physical displacements of these nodes.

- There are $N$ bars, $i$-th of them linking the nodes with indexes $\alpha_i$ and $\beta_i$ (with at least one of these nodes free) and with volume (3D or 2D, depending on whether the truss is spatial or planar) $t_i$.
- An external load is a collection of physical forces – vectors from $\mathbf{R}^d$ – acting at the free nodes (forces acting at the fixed nodes are of no interest – they are suppressed by the supports). Thus, an external load $f$ can be identified with block vector of the same structure as a virtual displacement – blocks are indexed by free nodes and represent the external forces acting at these nodes. Thus, displacements $v$ of the nodal set and external loads $f$ are vectors from the space $\mathcal{V}$ of *virtual displacements* – $M$-dimensional block vectors with $m$ $d$-dimensional blocks.
- The bars and the nodes together specify the symmetric positive semidefinite $M \times M$ *stiffness matrix* $A$ of the truss. The role of this matrix is as follows. A displacement $v \in \mathcal{V}$ of the nodal set results in reaction forces at free nodes (those at fixed nodes are of no interest – they are compensated by supports); assembling these forces into $M$-dimensional block-vector, we get a *reaction*, and this reaction is $-Av$. In other words, the potential energy capacitated in truss under displacement $v \in \mathcal{V}$ of nodes is $\frac{1}{2}v^\top Av$, and reaction, as it should be, is the minus gradient of the potential energy as a function of $v$ [2]. At the equilibrium under external load $f$, the total of the reaction and the load should be zero, that is, the equilibrium displacement satisfies

$$Av = f \tag{5.1}$$

Note that (5.1) may be unsolvable, meaning that the truss is crushed by the load in question. Assuming the equilibrium displacement $v$ exists, the truss at equilibrium capacitates potential energy $\frac{1}{2}v^\top Av$; this energy is called *compliance* of the truss w.r.t. the load. Compliance is convenient measure of rigidity of the truss with respect to the load, the less the compliance the better the truss withstands the load.

Let us build the stiffness matrix of a truss. As we have mentioned, the reaction forces originate from elongations/shortenings of bars under displacement of nodes. Consider $i$-th bar linking nodes with initial – prior to the external load being applied – positions $a_i = p_{\alpha_i}$ and $b_i = p_{\beta_i}$, and let us set

$$d_i = \|b_i - a_i\|_2,\ e_i = [b_i - a_i]/d_i.$$

Under displacement $v \in V$ of the nodal set,

- positions of the nodes linked by the bar become $a_i + \underbrace{v^{\alpha_i}}_{da}$, $b_i + \underbrace{v^{\beta_i}}_{db}$, where $v^\gamma$ is $\gamma$-th block in $v$ – the displacement of $\gamma$-th node
- as a result, elongation of the bar becomes, in the first-order in $v$ approximation, $e_i^\top[db - da]$, and the reaction forces caused by this elongation by Hooke's Law[3] are

$$\begin{array}{cl} d_i^{-1}S_ie_ie_i^\top[db-da] & \text{at node } \#\ \alpha_i \\ -d_i^{-1}S_ie_ie_i^\top[db-da] & \text{at node } \#\ \beta_i \\ 0 & \text{at all remaining nodes} \end{array}$$

where $S_i = t_i/d_i$ is the cross-sectional size of $i$-th bar. It follows that *when both nodes linked by i-th bar are free, the contribution of i-th bar to the reaction is*

$$-t_i\mathfrak{b}_i\mathfrak{b}_i^\top v,$$

[2] This is called *linearly elastic* model; it is the linearized in displacements approximation of the actual behavior of a loaded truss. This model works the better the smaller are the nodal displacements as compared to the inter-nodal distances, and is accurate enough to be used in typical real-life applications.

[3] Hooke's Law says that the magnitude of the reaction force caused by elongation/shortening of a bar is proportional to $Sd^{-1}\delta$, where $S$ is bar's cross-sectional size (area for spatial, and thickness for planar truss), $d$ is bar's (pre-deformation) length, and $\delta$ is the elongation. With units of length properly adjusted to bars' material, the proportionality coefficient becomes 1, and this is what we assume from now on.

*where $\mathfrak{b}_i \in \mathcal{V}$ is the vector with just two nonzero blocks:*
*— the block with index $\alpha_i$ – this block is $e_i/d_i = [b_i - a_i]/\|b_i - a_i\|_2^2$, and*
*— the block with index $\beta_i$ – this block is $-e_i/d_i = -[b_i - a_i]/\|b_i - a_i\|_2^2$.*
It is immediately seen that when just one of the nodes linked by $i$-th bar is free, the contribution of $i$-th bar to the reaction is given by similar relations, but with one, rather than 2, blocks in $\mathfrak{b}_i$ – the one corresponding to the free among the nodes linked by the bar.

The bottom line is that *The stiffness matrix of a truss composed of $N$ bars with volumes $t_i$, $1 \leq i \leq N$, is*

$$A = A(t) := \sum_i t_i \mathfrak{b}_i \mathfrak{b}_i^\top,$$

*where $\mathfrak{b}_i \in \mathcal{V} = \mathbf{R}^M$ are readily given by the geometry of nodal set and the indexes of nodes linked by bar $i$.*

**Truss Topology Design problem.** In the simplest Truss Topology Design (TTD) problem, one is given

- a finite *set of tentative nodes* in 2D or 3D along with support conditions indicating which of the nodes are fixed and which are free, and thus specifying the linear space $\mathcal{V} = \mathbf{R}^M$ of virtual displacements of the nodal set,
- the *set of $N$ tentative bars* – unordered pairs of (distinct from each other) nodes which are allowed to be linked by bars, and the total volume $W > 0$ of the truss,
- An external load $f \in \mathcal{V}$.

These data specify, as explained above, vectors $\mathfrak{b}_i \in \mathbf{R}^M$, $i = 1, \ldots, N$, and the stiffness matrix

$$A(t) = \sum_{i=1}^{N} t_i \mathfrak{b}_i \mathfrak{b}_i^\top = B \operatorname{Diag}\{t_1, \ldots, t_N\} B^\top \in \mathbf{S}^M \qquad [B = [\mathfrak{b}_1, \ldots, \mathfrak{b}_N]]$$

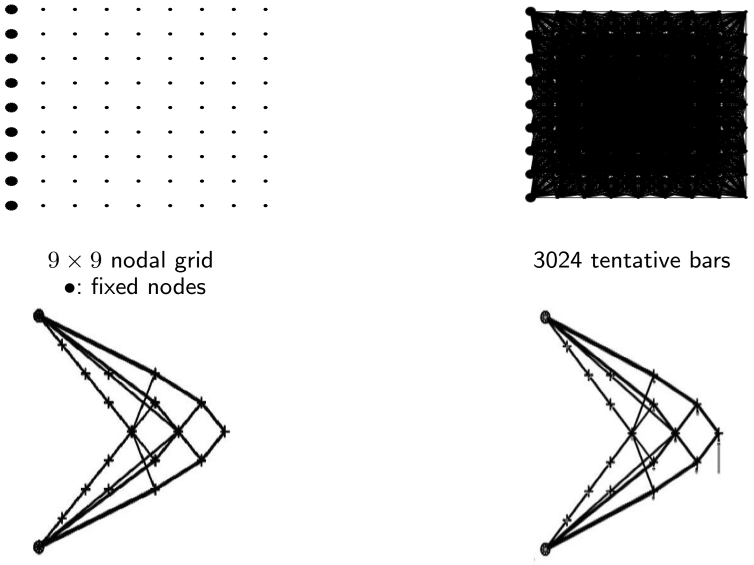
of truss, which under the circumstances can be identified with vector $t \in \mathbf{R}^N_+$ of bar volumes. What we want is to find the truss of given volume capable to "withstand best of all" the given load, that is, the one that minimizes the corresponding compliance.

When applying the TTD model, one starts with dense grid of tentative nodes and broad list of tentative bars (e.g., by allowing to link by a bar every pair of distinct from each other nodes, with at least one of the nodes in the pair free). At the optimal truss yielded by the optimal solution to the TTD problem, many tentative bars (usually vast majority of them) get zero volumes, and significant part of the tentative nodes become unused. Thus, TTD problem in fact is not about sizing – it allows to recover optimal structure of the construction, this is where "Topology Design" comes from.

To illustrate this point, here is a toy example (it will be our guinea pig in the entire series of TTD exercises):

**Console design:** We want to design a 2D truss as follows:

- The set of tentative nodes is the $9\times 9$ grid $\{[p;q]\in \mathbf{R}^2 : p,q\in\{0,1,\dots,8\}\}$ with the 9 most-left nodes fixed and remaining 72 nodes free, resulting in $M=144$-dimensional space $\mathcal{V}$ of virtual displacements
- The external load $f\in\mathcal{V}=\mathbf{R}^{144}$ is a single-force one, with the only nonzero force $[0;-1]$ applied at the 5-th node of the most-right column of nodes.
- We allow for all pairwise connections of pairs of distinct from each other nodes, with at least one of these nodes free, resulting in $N=3204$ tentative bars
- The total volume of truss is $W=1000$.



$9\times 9$ nodal grid
●: fixed nodes

3024 tentative bars

optimal truss, 38 bars
compliance 0.1914

displacement under
load of interest

Figure I.8. Console. Bars and nodes' positions before (crosses) and after (gray dots) deformation. Gray segment starting at the most right node: external force

**Important:** *From now on, speaking about TTD problem, we always make the following assumption:*

$\mathfrak{R}$ : $\sum_{t=1}^N \mathfrak{b}_i\mathfrak{b}_i^\top \succ 0.$

} Under this assumption, the stiffness matrix $A(t)=\sum_i t_i\mathfrak{b}_i\mathfrak{b}_i^\top$ associated with truss $t>0$ is positive definite, so that such a truss can withstand whatever load $f$.

You can verify numerically that this is the case in Console design as stated above.

After this lengthy preamble (to justify its length, note that it is investment to a series of exercises, rather than just one of them), let us pass to the exercise per se. Consider a TTD problem.

1. Prove that truss $t\geq 0$ (recall that we identify truss with the corresponding vector of bar volumes) is capable to carry load $f$ if and only if the quadratic function

$$F(v)=f^\top v-\frac{1}{2}v^\top A(t)v$$

is bounded from above, and that whenever this takes place,

- the maximum of $F$ over $\mathcal{V}$ is achieved

- the maximizers of $F$ are exactly the equilibrium displacements $v$ – those with
$$A(t)v = f,$$
and for such a displacement, one has
$$[\max F =]\ F(v) = \frac{1}{2}v^\top A(t)v = \frac{1}{2}v^\top f$$
- the maximum value of $F$ is exactly the compliance of the truss w.r.t. the load $f$

*Solution:* Observe, first, that a quadratic function
$$G(v) = g^\top v - \frac{1}{2}v^\top Av : \mathbf{R}^M \to \mathbf{R}$$
with $A \succeq 0$ attains its maximum if and only if it is bounded from above, and that the maximizers $v$ of $G$ are exactly the solutions $v$ to the Fermat equation
$$[\nabla G(v) =]\ \ g - Av = 0.$$
Indeed, invoking eigenvalue decomposition $A = U\,\mathrm{Diag}\{\lambda\}U^\top$ of $A$ (here $U$ is orthogonal, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$ are the eigenvalues of $A$; note that $\lambda_M \geq 0$ since $A \succeq 0$) and representing $v \in \mathbf{R}^M$ by the vector $\overline{v} = U^\top v$ of coordinates of $v$ in the eigenbasis of $A$, we get
$$G(v) = \overline{f}^\top\overline{v} - \frac{1}{2}\sum_{i=1}^M \lambda_i\overline{v}_i^2. \qquad [\overline{f} = U^\top f]$$
We conclude that $G$ is bounded from above if and only if $\overline{f}_i = 0$ for all $i$ such that $\lambda_i = 0$, and in this case $G$ attains its maximum, the maximizers being exactly $v$'s such that $\lambda_i\overline{v}_i = \overline{f}_i$ for all $i$, or, which is the same, all $v$'s such that $Av = f$. For such a $v$, if any,
$$G(v) = \sum_i[\overline{f}_i\overline{v}_i - \frac{1}{2}\lambda_i\overline{v}_i^2] = \frac{1}{2}\sum_i\lambda_i\overline{v}_i^2 = \frac{1}{2}v^\top Av = \frac{1}{2}v^\top f.$$
It remains to note that by definition of the compliance of truss $t$ w.r.t. load $f$, this compliance is finite if and only if the equation $Av = f$ in variables $f$ has a solution $v = v_f$, in which case the compliance is $\frac{1}{2}v_f^\top Av_f$. □

**Note:** From the above analysis, it follows that our original definition of compliance indeed makes sense – while the equilibrium displacement $v$ – the one such that $Av = f$ – when exists, not necessarily is uniquely defined by $A$ and $f$, the analysis we have just carried out shows that when $v_f$ exists, the quantity $\frac{1}{2}v_f^\top Av_f$ *is* uniquely defined by $A$ and $f$.

2. Prove that a real $\tau$ is an upper bound on the compliance of truss $t \geq 0$ w.r.t. load $f$ if and only if the symmetric matrix
$$\mathcal{A} = \left[\begin{array}{c|c} B\,\mathrm{Diag}\{t\}B^\top & f \\ \hline f^\top & 2\tau \end{array}\right],\ B = [\mathfrak{b}_1, \ldots, \mathfrak{b}_N]$$
is positive semidefinite. As a result, pose the TTD problem as the optimization problem
$$\mathrm{Opt} = \min_{\tau,r}\left\{\tau : \left[\begin{array}{c|c} B\,\mathrm{Diag}\{t\}B^\top & f \\ \hline f^\top & 2\tau \end{array}\right] \succeq 0, t \geq 0, \sum_i t_i = W\right\} \tag{5.2}$$
Prove that the problem is solvable.

*Solution:* As we have already seen, the compliance $\mathcal{C}$ of truss $t$ w.r.t. $f$ does not exceet a real $\tau$ iff

$\tau \geq \sup_v F(v)$, that is, setting $A = A(t) = B\,\mathrm{Diag}\{t\}B$,

$$\begin{array}{rl}
& \tau \geq \mathcal{C} \\
\Longleftrightarrow & \tau \geq f^\top v - \frac{1}{2}v^\top Av\,\forall v \\
\Longleftrightarrow & \tau - f^\top v + \frac{1}{2}v^\top Av \geq 0\,\forall v \\
\Longleftrightarrow & 2\tau - 2f^\top v + v^\top Av \geq 0\,\forall v \\
\Longleftrightarrow & 2\tau s^2 + 2sf^\top u + u^\top Au \geq 0\,\forall(s \neq 0, u) \text{ [look at } v = -u/s] \\
\Longleftrightarrow & 2\tau s^2 + 2sf^\top u + u^\top Au \geq 0\,\forall[u; s] \in \mathbf{R}^{M+1} \text{ [by continuity]} \\
\Longleftrightarrow & \left[\begin{array}{c|c} A & f \\ \hline f^\top & 2\tau \end{array}\right] \succeq 0 \qquad \square
\end{array}$$

To prove that the problem is solvable, note that $BB^\top \succ 0$, implying that every $t > 0$ such that $\sum_i t_i = W$ can be augmented by large enough $\tau$ to yield a feasible solution. Thus, (5.2) is feasible. Since for every feasible solution to the problem $\tau$ is nonnegative, the objective is below bounded on the (nonempty!) feasible set, so that the infimum Opt of the value of s objective at feasible solutions is nonnegative real. We can find sequence $[t^j, \tau^j]$ of feasible solutions with $\tau^j \to \mathrm{Opt}$ as $j \to \infty$. By feasibility, $t^j$ form a bounded sequence, so that passing to a subsequence, we can assume that $\lim_{j\to\infty}[t^j; \tau^j]$ exists; clearly, this limit is a feasible solution, and the $\tau$-component of this solution is Opt, implying that this solution is optimal. □

3. [computational study]
   - 3.1. Solve the Console problem numerically and reproduce the numerical results presented above.
   - 3.2. Resolve the problem with the set of all possible tentative bars reduced to the subset of "short" bars connecting neighboring nodes only:

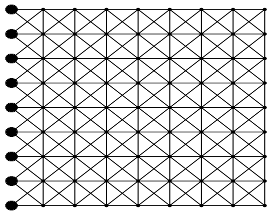


Figure I.9. 262 "short" tentative bars

and compare the resulting design and compliance to those in the previous item.

*Solution:* 3.2: Here are our results:

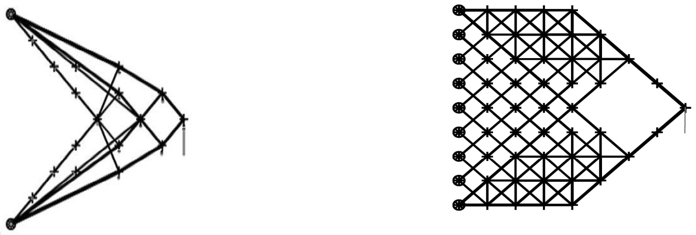


Figure VI.1."Long-bar truss:" 38 bars, compliance 0.1914 "Short-bar truss:" 128 bars, compliance 0.2903
vertical segments at the most-right nodes: external force of interest

## Around Caratheodory Theorem

**Exercise I.17.** Prove the following statement: Let $X \subset \mathbf{R}^n$ be nonempty. Then

1. if a point $x$ can be represented as a convex combination of a collection of vectors from $X$, then the collection can be selected to be affinely independent.
2. if a point $x$ can be represented as a conic combination of a collection of vectors from $X$, then the collection can be selected to be linearly independent,

Note that the claims above are refinements, albeit minor ones, of the Caratheodory Theorem (plain and conic, respectively). Indeed, when $M = \text{Aff}(X)$ and $m$ is the dimension of $M$, every affine independent collection of points from $X$ contains at most $m+1$ points (Proposition A.44), so that the first claim implies that if $x \in \text{Conv}(X)$, then $x$ is a convex combination of at most $m+1$ points from $X$; however, the vectors participating in such a combination are not necessarily affinely independent, so that the first claim provides a bit more information than the plain Caratheodory's Theorem. Similarly, if $L = \text{Lin}(X)$ and $m = \dim L$, then every linearly independent collection of vectors from $X$ contains at most $m \leq n$ points, that is, the second claim implies the Caratheodory's Theorem in conic form, and provides a bit more information than the latter theorem.

*Solution:* 1: For $x \in \text{Conv}(X)$, let $x = \sum_{i\in I} \lambda_i x_i$ be the shortest – with the minimum possible cardinality of $I$ – representation of $x$ as a convex combination of points from $X$, and let us verify that the vectors $x_i$ participating in this representation form an affinely independent collection. Assuming otherwise, there exists a nontrivial collection of reals $\delta_i$, $i \in I$, such that $\sum_i \delta_i x_i = 0$ and $\sum_i \delta_i = 0$, and we can proceed exactly as in the proof of Caratheodory's Theorem: setting $\lambda_i(t) = \lambda_i + t\delta_i$, we have $\sum_i \lambda_i(t) = 1$ and $\sum_i \lambda_i(t)x_i = x$ for all $t$, and since not all $\delta_i$ are zeros and their sum is 0, some of $\lambda_i(t)$ for large $t$ become negative, implying, due to $\lambda_i(0) \geq 0\,\forall i$, that for some $t^*$ all $\lambda_i(t^*)$ are nonnegative, and some of them vanish, contradicting the assumption that number of terms in our initial representation of $x$ as a convex combinations of points from $X$ is the minimum possible.
2: Similarly, for $x \in \text{Cone}(X)$, let $x = \sum_{i\in I} \lambda_i x_i$ be the representation of $x$ as a conic combination of points from $X$ with the minimum possible number of terms, and let us prove that the vectors $x_i$ participating in this representation form a linearly independent collection. This indeed is so when $I = \varnothing$. Now let $I$ be nonempty, and assume, for contradiction, that the vectors $x_i$, $i \in I$, are linearly dependent, so that $\sum_i \delta_i x_i = 0$ for a nontrivial collection $\delta_i, i \in I$. Passing, if necessary, from $\delta_i$ to $-\delta_i, i \in I$, we may assume that some of $\delta_i$ are strictly negative. Setting $\lambda_i(t) = \lambda_i + t\delta_i$, we have that $\sum_i \lambda_i(t)x_i = x$ for all $t$, $\lambda_i(0) \geq 0$, $i \in I$, and some of $\lambda_i(t)$ become negative for large $t$. It follows that there exists the largest $t = t^*$ for which all $\lambda_i(t)$ still are nonnegative, and for $t = t^*$ some of $\lambda_i(t)$ vanish, implying that $x = \sum_{i\in I} \lambda_i(t^*)x_i$ is a representation of $x$ as a conic combination of $x_i$ with some of the coefficients equal to 0, contradicting the minimality of the original representation. □

**Exercise I.18.**[4] Consider TTD problem, and let $N$ be the number of tentative bars, $M$ be the dimension of the corresponding space of virtual displacements $\mathcal{V}$, and $f$ be an external load. Prove that if truss $t \geq 0$ can withstand load $f$ with compliance $\leq \tau$ for some given real $\tau$, then there exists truss $\bar{t}$ of the same total volume as $t$ with compliance w.r.t. $f$ at most $\tau$ and at most $M+1$ bars of positive volume.

*Solution:* Denoting by $v$ the equilibrium displacement of (nodes of) truss $t$ under load $f$, by the results of Exercise I.16.1 we have

$$\sum_{i=1}^{N} t_i \underbrace{[\mathfrak{b}_i \mathfrak{b}_i^\top v]}_{g_i} = f \ \& \ \frac{1}{2} f^\top v \leq \tau$$

Denoting $w$ the volume of $t$ and assuming w.l.o.g. that $w > 0$, we see that $f/w$ is a convex combination, with coefficients $t_i/w$, of vectors $g_i \in \mathbf{R}^M$. By Caratheodory Theorem, $f/w$ is a convex combination of the same vectors $g_i$ with coefficients $s_i$ such that at most $M+1$ of these coefficients are positive. It follows that setting $\bar{t}_i = ws_i$, we get truss $\bar{t}$ of the same volume as $t$, with at most $M+1$ bars of positive volume, such that $\sum_i \bar{t}_i \mathfrak{b}_i \mathfrak{b}_i^\top v = f$, so that $v$ is the equilibrium displacement of truss $\bar{t}$ under load $f$. Consequently, the compliance of truss $\bar{t}$ w.r.t. load $f$ is $\frac{1}{2} f^\top v \leq \tau$. □

**Exercise I.19.**

[4] Preceding exercise in the TTD series is I.16.

1. Prove that if a system of linear equations $Ax = b$ with $n$ variables and $m$ equations has a nonnegative solution, it has a nonnegative solution with at most $m$ positive entries.

   *Solution:* Let $A_1, \ldots, A_n$ be the columns of $A$. Nonnegative solutions to the system $Ax = b$ are exactly the vectors of coefficients in representation of $b \in \mathbf{R}^m$ as a conic combination of $A_1, \ldots, A_n$. Then, by conic version of Caratheodory's Theorem (Fact I.2.7), if $b$ admits such a representation, it admits such a representation with at most $m$ of $A_i$'s involved.

2. Let $V_1, \ldots, V_n$ be $n$ nonempty sets in $\mathbf{R}^m$, and define
$$\overline{V} := \mathrm{Conv}(V_1 + V_2 + \ldots + V_n).$$

   1. Prove that
$$\overline{V} = \mathrm{Conv}(V_1 + \ldots + V_n) = \mathrm{Conv}(V_1) + \ldots + \mathrm{Conv}(V_n).$$

      *Solution:* Suppose $x \in \mathrm{Conv}(V_1 + \ldots + V_n)$. Then, $x = \sum_k \lambda_k \sum_{i=1}^n x_{i,k}$ with $x_{i,k} \in V_i$ and $\lambda_k \geq 0$ such that $\sum_k \lambda_k = 1$, implying that
$$x = \sum_{i=1}^n \underbrace{\left[\sum_k \lambda_k x_{i,k}\right]}_{\in \mathrm{Conv}(V_i)}.$$
      Thus, $\overline{V} \subseteq \mathrm{Conv}(V_1) + \ldots + \mathrm{Conv}(V_n)$.

      To see the reverse containment, suppose $x \in \mathrm{Conv}(V_1) + \ldots + \mathrm{Conv}(V_n)$. Then, for properly selected $K$ we have $x = \sum_{i=1}^n \sum_{k=1}^K \lambda_{i,k} x_{i,k}$ with $x_{i,k} \in V_i$ and $\lambda_{i,k} \geq 0$ such that $\sum_k \lambda_{i,k} = 1$, implying that
$$x = \sum_{1 \leq k_1, \ldots, k_n \leq K} \underbrace{\lambda_{1,k_1} \lambda_{2,k_2} \ldots \lambda_{n,k_n}}_{=:\lambda_{k_1 k_2 \ldots k_n} \geq 0} \underbrace{\left[\sum_{i=1}^n x_{i,k_i}\right]}_{\in V_1 + \ldots + V_n},$$
$$\sum_{1 \leq k_1, \ldots, k_n \leq K} \lambda_{k_1 k_2 \ldots k_n} = \prod_{i=1}^n \left[\sum_{k_i=1}^K \lambda_{i,k_i}\right] = 1.$$
      That is, $x \in \overline{V}$ and so we get $\mathrm{Conv}(V_1) + \ldots + \mathrm{Conv}(V_n) \subseteq \overline{V}$.

   2. Prove *Shapley-Folkman Theorem*:

      *Let $x \in \overline{V}$. Then, there exists a representation of $x$ such that*
$$x = x_1 + \ldots + x_n, \quad x_i \in \mathrm{Conv}(V_i),$$
      *in which at least $n - m$ of $x_i$'s belong to the respective sets $V_i$.*

      *Comment:* Shapley-Folkman Theorem says, informally, that when $n \gg m$, summing up $n$ nonempty sets in $\mathbf{R}^m$ possesses certain "convexification property" – every point from the convex hull $\overline{V}$ of the sum of our sets is the sum of points $x_i$ with all but $m$ of them belonging to $V_i$ rather than to $\mathrm{Conv}(V_i)$, and only $\leq m$ of the points belonging to $V_i$ "fractionally," that is, belonging to $\mathrm{Conv}(V_i)$, but not to $V_i$. This nice fact has numerous useful applications.

      *Solution:* Let $x \in \overline{V}$. Then, by the previous part, $x = \sum_{i=1}^n \left(\sum_{k=1}^K \lambda_{i,k} x_{i,k}\right)$ with some $K$, $x_{i,k} \in V_i$, and $\lambda_{i,k} \geq 0$, $\sum_{k=1}^K \lambda_{i,k} = 1$, $i \leq n$. Hence, $nK$ reals $\lambda_{i,k}$ form a nonnegative solution to the system of $m + n$ linear equations
$$\begin{array}{ll} (a) & \sum_k \lambda_{i,k} = 1, \quad 1 \leq i \leq n \\ (b) & \sum_i \sum_k \lambda_{i,k} x_{i,k} = x. \end{array}$$
      By the first item of this exercise, this system has a nonnegative solution with at most $m + n$

nonzero entries; let us denote this solution by $\{\overline{\lambda}_{i,j}\}$. We can partition the equations in $(a)$ into two groups, the ones in which exactly one of the variables participating in the equation takes a positive value in the solution $\{\overline{\lambda}_{i,j}\}$, and the equations in which two or more variables participating in the equation take positive values in $\{\overline{\lambda}_{i,j}\}$. Let $d$ be the number of equations in the latter set. Every one of the remaining $n-d$ equations in $(a)$ involves at most one, and therefore exactly one, variable which at our solution gets positive value, and this positive values is, of course, 1. Since every one of the variables $\lambda_{i,k}$ enters exactly one of the equations $(a)$, we conclude that the total number of positive $\overline{\lambda}_{i,k}$'s (which, as we remember, is at most $m+n$) is at least $2d+(n-d)=n+d$, implying that $d\leq m$. Thus, for at least $n-m$ values of $i$ all but one of $\overline{\lambda}_{i,k}$'s, $k=1,\dots,K$, are zeros, and the remaining one equals to 1. In other words, all but at most $m$ of the $n$ sums $\sum_k \overline{\lambda}_{i,k}x_{i,k}$, $i=1,\dots,n$, are just points from the respective sets $V_i$. It remains to recall that $x=\sum_{i=1}^n\sum_k \overline{\lambda}_{i,k}x_{i,k}$.

**Exercise I.20.** Caratheodory's Theorem in its plain and its conic forms are "existence" statements: if a point $x\in\mathbf{R}^m$ is a convex, respectively conic, combination of points $x^1,\dots,x^N$, then *there exists* a representation of $x$ of the same type which involves at most $(m+1)$, respectively, $m$, terms. Extract from the proofs of the theorems *algorithms* for finding these "short" representations at the cost of solving at most $N$ solvable systems of linear equations with at most $N$ variables and $m$ equations each.

*Solution:* For the sake of definiteness, consider plain Caratheodory Theorem (conic case can be treated in exactly the same fashion). Proof of Theorem, on immediate inspection, is based on the following observation:

> Given representation $x=\sum_{i=1}^K \lambda_i x^i$ with $\lambda_i>0$, $i\leq K$ and $\sum_i\lambda_i=1$, in the case of $K>m+1$ and finding a nontrivial solution $\delta$ to the homogeneous system of linear equations
>
> $$\sum_{i=1}^K \delta_i x^i=0,\quad \sum_{i=1}^K\delta_i=0.$$
>
> we can convert, by simple computation, the initial representation into a new one, of the same form, which assigns positive weights to at most $K'\leq K-1$ of $x^i$'s.

Recall that we are given representation of $x$ as convex combination of $K=N$ of $x^i$'s. If $K>m+1$, we can apply the above construction to represent $x$ as a convex combination of $K'<K$ of $x^i$'s. If $K'>m+1$, we can iterate this update, with $K'$ in the role of $K$, to represent $x$ as convex combination of at most $K''<K'$ of $x^i$'s. Proceeding in this way, we in at most $N$ steps will represent $x$ as convex combination of at most $m+1$ of $x^i$'s.

**Exercise I.21.** Prove *Kirchberger's Theorem*:

> Consider two sets of finitely many points $X=\{x^1,\dots,x^k\}$ and $Y=\{y^1,\dots,y^m\}$ in $\mathbf{R}^n$ such that $k+m\geq n+2$ and all the points $x^1,\dots,x^k,y^1,\dots,y^m$ are distinct. Assume that for any subset $S\subseteq X\cup Y$ which contains $n+2$ points the convex hulls of the sets $X\cap S$ and $Y\cap S$ do not intersect: $\mathrm{Conv}(X\cap S)\cap\mathrm{Conv}(Y\cap S)=\varnothing$. Then, the convex hulls of $X$ and $Y$ also do not intersect: $\mathrm{Conv}(X)\cap\mathrm{Conv}(Y)=\varnothing$.

*Hint:* Assume for contradiction that $\mathrm{Conv}(X)\cap\mathrm{Conv}(Y)\neq\varnothing$, so that

$$\sum_{i=1}^k\lambda_i x^i=\sum_{j=1}^m\mu_j y^j \tag{$*$}$$

for certain nonnegative $\lambda_i$, $\sum_{i=1}^k\lambda_i=1$, and certain nonnegative $\mu_j$, $\sum_{j=1}^m\mu_j=1$, and look at the expression of this type with the minimum possible total number of nonzero coefficients $\lambda_i$, $\mu_j$.

*Solution:* Following the hint, assume for the contradiction that $\mathrm{Conv}(Y)$ and $\mathrm{Conv}(X)$ do intersect, so that the relation $(*)$ holds for appropriately chosen $\lambda_i$, $\mu_j$ satisfying

$$\lambda_i\geq 0,\quad \mu_j\geq 0,\quad \sum_i\lambda_i=\sum_j\mu_j=1. \tag{$**$}$$

And, among the collection of weights $\lambda_i$, $\mu_j$ satisfying $(*)$ and $(**)$, let us select one that has the smallest in the total number of positive $\lambda_i$, $\mu_j$. Without loss of generality, we may assume that in this collection of weights, the positive weights are the first $p$ of $\lambda_i$'s and the first $q$ of $\mu_j$'s. Note that by the premise of Kirchberger's Theorem, $p+q > n+2$. Now consider the following system of $n+2$ equations with $p+q > n+2$ unknowns:

$$\begin{aligned}\sum_{i=1}^{p}\delta_i x^i - \sum_{j=1}^{q}\theta_j y^j &= 0,\\ \sum_i \delta_i &= 0,\\ \sum_j \theta_j &= 0.\end{aligned}$$

As this is a homogeneous system of linear equations and the number of unknowns is greater than the number of equations, the system has a nontrivial solution $\delta, \theta$. Setting $\lambda_i(t) = \lambda_i + t\delta_i$, $i \leq p$, and $\mu_j(t) = \mu_j + t\theta_j$, $j \leq q$, we have for all $t$:

$$\sum_i \lambda_i(t)x^i = \sum_j \mu_j(t)y^j, \quad \sum_i \lambda_i(t) = 1, \quad \sum_j \mu_j(t) = 1.$$

For $t = 0$, all the coefficients $\lambda_i(t)$, $\mu_j(t)$ are positive. Since $\sum\limits_i \delta_i + \sum\limits_j \theta_j = 0$ and not all $\delta_i$, $\theta_j$ are zeros, among the reals $\delta_i, \theta_j$ at least one should be negative.

Hence, for large enough $t > 0$ some of the coefficients $\lambda_i(t)$, $\mu_j(t)$ will be negative. Consequently, there exists the largest $t = t_*$ for which all $\lambda_i(t)$, $\mu_j(t)$ are nonnegative; among $\lambda_i(t_*)$, $\mu_j(t_*)$, there is clearly at least one zero, and we see that the coefficients $\lambda_i(t_*)$, $\mu_j(t_*)$ satisfy $(*)$, $(**)$, and the total number of positive among them is $< p+q$, which is a contradiction.

**Exercise I.22** [Follow-up to Shapley-Folkman Theorem]

1. Let $X_1, \ldots, X_K$ be nonempty convex sets in $\mathbf{R}^n$, and define $X := \bigcup_{k\leq K} X_k$. Prove that

$$\mathrm{Conv}(X) = \left\{ x = \sum_{k=1}^{K} \lambda_k x^k : \ \lambda_k \geq 0,\ x^k \in X_k,\ \forall k \leq K,\ \sum_{k=1}^{K} \lambda_k = 1 \right\}.$$

*Solution:* Let $\overline{X}$ be the set on the right hand side. As $X = \bigcup_{k\leq K} X_k$, based on the definition of $\overline{X}$ it is clear that $\mathrm{Conv}(X) \supseteq \overline{X}$. So, all we need to show is that $\mathrm{Conv}(X) \subseteq \overline{X}$. To this end consider any $x \in \mathrm{Conv}(X)$, and so $x = \sum_{s=1}^{S} \mu_s y^s$ with $\mu_s \geq 0$, $y^s \in X$, $s \leq S$, and $\sum_s \mu_s = 1$. We can clearly split the index set $\{1, 2, \ldots, S\}$ into $K$ non-overlapping subsets $S_k$, $k \leq K$ (some of these subsets can be empty) in such a way that $s \in S_k$ implies that $\mu_s > 0$ and $y^s \in X_k$. For $k$ with nonempty $S_k$, let us set $\lambda_k := \sum_{s\in S_k} \mu_s$ and define $x^k := \sum_{s\in S_k} \frac{\mu_s}{\lambda_k} y^s$. By definition of $\lambda_k$ and the fact that $y^s \in X_k$ for all $s \in S_k$, we see that $x^k$ is a convex combination of points from $X_k$, and as $X_k$ is a convex set we conclude that $x^k \in X_k$. For $k$ with empty $S_k$, let us set $\lambda_k = 0$ and select somehow $x^k$ in the (nonempty!) set $X_k$. As a result, we get $x = \sum_{s=1}^{S} \mu_s y^s = \sum_{k=1}^{K} \lambda_k x^k$ with $x^k \in X_k$, $\lambda_k \geq 0$, and $\sum_{k=1}^{K} \lambda_k = 1$. This shows that $x \in \overline{X}$ as desired.

2. Let $X_k$, $k \leq K$, be nonempty bounded polyhedral sets in $\mathbf{R}^n$ given by polyhedral representations:

$$X_k = \left\{ x \in \mathbf{R}^n : \ \exists u^k \in \mathbf{R}^{n_k} \text{ such that } P_k x + Q_k u^k \leq r_k \right\}.$$

Define $X := \bigcup_{k\leq K} X_k$. Prove that the set $\mathrm{Conv}(X)$ is a polyhedral set given by the polyhedral representation

$$\mathrm{Conv}(X) = \left\{ x \in \mathbf{R}^n : \begin{array}{ll} \exists x^k \in \mathbf{R}^n,\ u^k \in \mathbf{R}^{n_k},\ \lambda_k \in \mathbf{R},\ \forall k \leq K : & \\ \quad P_k x^k + Q_k u^k - \lambda_k r_k \leq 0,\ k \leq K & (a) \\ \quad \lambda_k \geq 0,\ \sum_{k=1}^{K} \lambda_k = 1 & (b) \\ \quad x = \sum_{k=1}^{K} x^k & (c) \end{array} \right\}. \qquad (*)$$

Does the claim remain true when the assumption of boundedness of the sets $X_k$s is lifted?

*Solution:* Let us temporary denote by $\widehat{X}$ the right hand side set in $(*)$ and set $\overline{X} := \mathrm{Conv}(X)$. We need to show that $\overline{X} = \widehat{X}$. Recall from item 1 that

$$\overline{X} = \left\{ x = \sum_{k=1}^{K} \lambda_k x^k : \ \lambda_k \geq 0, \ x^k \in X_k, \ \forall k \leq K, \ \sum_{k=1}^{K} \lambda_k = 1 \right\}.$$

Let $\widetilde{X}$ be the set of all vectors representable as convex combinations, *with positive coefficients*, of vectors from $X_1, \ldots, X_K$. Note that $\widetilde{X} \subseteq \overline{X}$.
Observe that

$$\widetilde{X} = \left\{ x \in \mathbf{R}^n : \begin{array}{ll} \exists x^k \in \mathbf{R}^n, \ u^k \in \mathbf{R}^{n_k}, \ \lambda_k \in \mathbf{R}, \ \forall k \leq K : & \\ \quad P_k x^k + Q_k u^k - \lambda_k r_k \leq 0, \ k \leq K & (a') \\ \quad \lambda > 0, \ \sum_{k=1}^K \lambda_k = 1 & (b') \\ \quad x = \sum_{k=1}^K x^k & (c') \end{array} \right\}. \qquad (!)$$

Indeed, when $x$ belongs to the right hand side set in (!), we have $y^k := \lambda_k^{-1} x^k \in X_k$ due to $P_k y^k + Q_k[\lambda_k^{-1} u^k] \leq r_k$ and $x = \sum_k \lambda_k y^k$. Vice versa, when $x \in \widetilde{X}$, we have $x = \sum_k \lambda_k y^k$ with positive $\lambda_k$ summing up to 1 and $y^k \in X_k$. The latter means that there exist $v^k$ such that $P_k y^k + Q_k v^k \leq r_k$. Setting $x^k = \lambda_k y^k$, $u^k = \lambda_k v^k$, we ensure validity of $(a')$ – $(c')$, so that $x$ belongs to the right hand side set in (!).
Next, we claim that $\widehat{X} = \mathrm{cl}\,\overline{X}$. First, observe that $\widetilde{X}$ is dense in $\overline{X}$, meaning that every point $x \in \overline{X}$ is the limit of a sequence of points from $\widetilde{X}$. Indeed, consider any $x \in \overline{X}$, i.e., $x = \sum_k \lambda_k x^k$ with nonnegative $\lambda_k$ summing up to 1 and $x^k \in X_k$ for all $k$. Then, we have $x = \lim_{i\to\infty} \sum_k \frac{\lambda_k + 1/i}{1+K/i} x^k$, and the points in the right hand side sequence belong to $\widetilde{X}$. Now, observe that $\widehat{X}$ is closed (it is polyhedrally representable and thus polyhedral) and moreover $\widetilde{X}$ is dense in $\widehat{X}$. Indeed, by (!) we have $\widetilde{X} \subseteq \widehat{X}$. On the other hand, let us fix somehow $\overline{x}^k \in X_k$ and $\overline{\lambda}_k > 0$ such that $\sum_k \overline{\lambda}_k = 1$, and let $\overline{u}^k$ be such that $P_k \overline{x}^k + Q_k \overline{u}^k \leq r_k$. Given $x \in \widehat{X}$, there exist $x^k$, $u^k$ and $\lambda_k$ satisfying $(a)$ – $(c)$. For all $i = 1, 2, \ldots$, setting

$$\begin{aligned} x^{k,i} &:= (1 - 1/i)x^k + (1/i)\overline{x}^k, \\ u^{k,i} &:= (1 - 1/i)u^k + (1/i)\overline{u}^k, \\ \lambda_{k,i} &:= (1 - 1/i)\lambda_k + (1/i)\overline{\lambda}_k, \end{aligned}$$

we ensure that $P_k x^{k,i} + Q_k u^{k,i} - \lambda_{k,i} r_k \leq 0$, $\lambda_{k,i} > 0$, $\sum_i \lambda_{k,i} = 1$, implying that $x^{(i)} := \sum_k x^{k,i} \in \widetilde{X}$. As $i \to \infty$, we clearly have $x^{(i)} \to x$, so that $\widetilde{X}$ indeed is dense in $\widehat{X}$. The latter combines with closedness of $\widehat{X}$ to imply that the $\widehat{X}$ is the closure of $\widetilde{X}$, and the latter set, due to the fact that $\widetilde{X}$ is dense in $\overline{X}$, is the same as the closure of $\overline{X}$. Thus, $\widehat{X} = \mathrm{cl}\,\overline{X}$.
It remains to note that since $X_k$ are bounded, $\overline{X}$ is closed. This is immediate: assuming that $x = \lim_{i\to\infty} \sum_k \lambda_{k,i} x^{k,i}$ with nonnegative $\lambda_{k,i}$, $\sum_k \lambda_{k,i} = 1$, and $x^{k,i} \in X_k$, boundedness of $X_k$, $k \leq K$, allows to find a subsequence $i_1 < i_2 < \ldots$ of indexes such that for some $\lambda_k$ and $x^k$, $k \leq K$, it holds $\lambda_{k,i_s} \to \lambda_k$ and $x^{k,i_s} \to x^k$ for every $k$ as $s \to \infty$. Since $X_k$ are polyhedral and thus closed, we have $x^k \in X_k$, and of course $\lambda_k \geq 0$, $\sum_k \lambda_k = 1$, that is, $x = \lim_{s\to\infty} \sum_k \lambda_{k,i_s} x^{k,i_s} = \sum_k \lambda_k x^k \in \overline{X}$.
Finally, $\mathrm{Conv}(\bigcup_{k\leq K} X_k)$ is not necessarily polyhedral when $X_k$ are nonempty polyhedral, but unbounded, sets. For example, by selecting $K = n = 2$, $X_1 := \{x \in \mathbf{R}^2 : x_1 \geq 0, \ x_2 = 0\}$ and $X_2 := \{[0; 1]\}$, we see that the set $\mathrm{Conv}(X_1 \cup X_2)$ is not polyhedral, but its closure is. On inspection, the above reasoning demonstrates that when $X_k$ are nonempty polyhedral sets given by polyhedral representations, then the polyhedral set $\widehat{X}$ defined as the right hand side set of $(*)$ is the closure of $\mathrm{Conv}(\bigcup_{k\leq K} X_k)$.

After two preliminary items above, let us pass to the essence of the matter. Consider the situation as follows. We are given $n$ nonempty and bounded polyhedral sets $X_j \subset \mathbf{R}^r$, $j = 1, \ldots, n$. We will think of $X_j$ as the "resource set" of the $j$-th production unit: entries in $x \in X_j$ are amounts of various resources, and $X_j$ describes the set of vectors of resources available, in principle, for $j$-th unit. Each production unit $j$ can possibly use any one of its $K_j < \infty$ different production plans. For each $j = 1, \ldots, n$, the vector $y_j \in \mathbf{R}^p$ representing the production of the $j$-th unit depends on the vector $x_j$ of resources consumed by the unit and also on the production plan utilized in the unit.

In particular, the production vector $y_j \in \mathbf{R}^p$ stemming from resources $x_j$ under $k$-th plan can be picked by us, at our will, from the set

$$Y_j^k[x_j] := \left\{ y_j \in \mathbf{R}^p : \; z_j := [x_j; -y_j] \in V_j^k \right\},$$

where $V_j^k$, $k \leq K_j$, are given bounded polyhedral "technological sets" of the units with projections onto the $x_j$-plane equal to $X_j$, so that for every $k \leq K_j$ it holds

$$x_j \in X_j \quad \Longleftrightarrow \quad \exists y_j \text{ such that } [x_j; -y_j] \in V_j^k. \tag{5.3}$$

We assume that all the sets $V_j^k$ are given by polyhedral representations, and we define

$$V_j := \bigcup_{k \leq K_j} V_j^k.$$

Let $R \in \mathbf{R}^r$ be the vector of total resources available to all $n$ units and let $P \in \mathbf{R}^p$ be the vector of total demands for the products. For $j \leq n$, we want to select $x_j \in X_j$, $k_j \leq K_j$, and $y_j \in Y_j^{k_j}[x_j]$ in such a way that

$$\sum_j x_j \leq R \quad \text{and} \quad \sum_j y_j \geq P.$$

That is, we would like to find $z_j = [x_j; v_j] \in V_j$, $j \leq n$, in such a way that $\sum_j z_j \leq [R; -P]$. Note that the presence of "combinatorial part" in our decision – selection of production plans in finite sets – makes the problem difficult.

3. Apply Shapley-Folkman Theorem (Exercise I.19) to overcome, to some extent, the above difficulty and come up with a good and approximately feasible solution.

*Solution:* Let $s := [R; -P]$, and observe that our problem reads

$$\text{Find } z_j \in V_j \text{ such that } \sum_{j=1}^n z_j \leq s. \tag{$P$}$$

Note that given polyhedral representations of $V_j^k$, based on item 2, we can build explicit polyhedral representations of the convex hulls of the sets $V_j$, i.e., we can efficiently compute $\overline{V}_j$, where

$$\overline{V}_j := \mathrm{Conv}(V_j).$$

Let us relax the problem of interest $(P)$ to the problem

$$\text{Find } z_j \in \overline{V}_j \text{ such that } \sum_{j=1}^n z_j \leq s. \tag{$\overline{P}$}$$

By calculus of polyhedral representations, $(\overline{P})$ is the problem of the form

> *Given polyhedral representation of nonempty polyhedral set $Z \subset \mathbf{R}^{r+p}$ and vector $s \in \mathbf{R}^{r+p}$, find $z \in Z$ such that $z \leq s$.*

Note that is an explicit Linear Programming feasibility problem. Thus, we can apply LP algorithms to check whether $(\overline{P})$ is solvable, and if it is the case – find a solution $\{z_j, j \leq n\}$ to $(\overline{P})$. Applying Shapley-Folkman Theorem, we can convert, in a computationally efficient fashion, this solution into another feasible solution, $\{[x_j; v_j], j \leq n\}$, for which for all but at most

$$d := \min\{r + p, \, n\}$$

components $[x_j; v_j]$ belong to $V_j$, that is, "are implementable" – for the corresponding $j$, one has $x_j \in X_j$ and $y_j = -v_j \in Y_j^{k_j}[x_j]$ with properly selected $k_j \leq K_j$. Let $J$ be the set of "bad" indices $j$, i.e., those for which $[x_j; v_j] \in \overline{V}_j \backslash V_j$. Note that for each $j \in J$ we still have $x_j \in X_j$. We can correct the corresponding

$y_j$, passing from $[x_j; v_j]$ to $[x_j; \overline{v}_j]$ with $\overline{v}_j \in -Y^1_j[x_j]$, or, better, $\overline{v}_j$ defined as the optimal solution to the "best" – with the smallest optimal value – among the $K_j$ convex optimization problems

$$\min_{u_k} \left\{ \|v_j - u_k\| : \; [x_j; u_k] \in V_j^k \right\}, \quad k \leq K_j,$$

where $\|\cdot\|$ is some norm. As a result, we get "fully implementable" solution $\{[x_j; \overline{v}_j], j \leq n\}$, where $\overline{v}_j = v_j$ for $j \notin J$, to problem $(P)$. This solution, in general, may not be feasible when $J \neq \varnothing$. However, by selecting somehow norm $\|\cdot\|$, defining

$$D_j := \max_{x,v,x'v'} \left\{ \|v - v'\| : \; [x; v] \in V_j, \; [x'; v'] \in V_j \right\}, \; \forall j \leq n \quad \text{and} \quad D := \max_j D_j,$$

and taking into account that $\text{Card}(J) \leq d = \min\{r + p, n\}$, we have $\sum_j [x_j; \overline{v}_j] \leq s + \delta$, $\|\delta\| \leq dD$, and $\sum_j x_j \leq R$. In the case of "mass production", when $\|P\|$ is large, the violation of the constraint $\sum_j \overline{v}_j \leq -P$ as quantified by $\|\delta\|$ is a small fraction of the magnitude of $P$, and our implementable solution has chances to be a good, from a "practical perspective," surrogate of a feasible solution to $(P)$.

## Around Helly Theorem

**Exercise I.23.** [Alternative proof of Helly Theorem] The goal of this exercise is to build an alternative proof of Helly's Theorem, without using Radon's Theorem.

1. Consider a system $a_i^\top x \leq b_i$, $i \leq N$, of $N$ linear inequalities in variables $x \in \mathbf{R}^n$. Helly's Theorem applied to the sets $A_i := \{x \in \mathbf{R}^n : \; a_i^\top x \leq b_i\}$ gives us that

   > (!) *If a system $a_i^\top x \leq b_i$, $i \leq N$, of linear inequalities in variables $x \in \mathbf{R}^n$ is infeasible, so is a properly selected sub-system composed of at most $n+1$ inequalities from the system.*

   Find an alternative proof of (!) without relying on Helly's or Radon's Theorems.

   *Solution:* Suppose that the system $a_i^\top x \leq b_i$, $i \leq N$, is infeasible. Then, by General Theorem on Alternative there exist nonnegative weights $\lambda_i$ and $\alpha < 0$ such that the vector $[0; \ldots; 0; \alpha]$ is a conic combination, with coefficients $\lambda_i$, of vectors $[a_i; b_i]$, $i \leq N$. Note that $[a_i; b_i] \in \mathbf{R}^{n+1}$, and so by Caratheodory's Theorem in conic form it follows that the vector $[0; \ldots; 0; \alpha]$ is a conic combination of at most $n+1$ of the vectors $[a_i; b_i]$, let $I$ be the set of their indexes. By GTA, the subsystem $a_i^\top x \leq b_i$, $i \in I$, of the original system is infeasible, and the number of inequalities in it is at most $n+1$, as desired.

2. Extract from item 1 Helly's Theorem for polyhedral sets: *If $A_1, \ldots, A_N$, $N \geq n+1$, are polyhedral sets in $\mathbf{R}^n$ and every $n+1$ of these sets have a point in common, then all the sets have a point in common.*

   *Solution:* Let $A_i := \{x \in \mathbf{R}^n : \; P_i x \leq p_i\}$ for $i \leq N$. To justify the claim in question is the same as to prove that if $\cap_i A_i = \varnothing$, then the intersection of properly selected $k \leq n+1$ sets from the collection is empty. Suppose that $\cap_i A_i = \varnothing$, that is, the system

   $$P_i x \leq p_i, i \leq N \qquad (*)$$

   of linear inequalities in variables $x \in \mathbf{R}^n$ has no solutions. By item 1, we can select from $(*)$ $k \leq n+1$ inequalities to get an infeasible subsystem of $(*)$. Denoting by $I$ the set of indices $i$ of the blocks $P_i x \leq p_i$ of $(*)$ containing the $k$ selected inequalities, we conclude that $k \leq n+1$ sets $A_i$, $i \in I$, have no point in common.

3. Extract from item 2 Helly's Theorem (Theorem I.2.10).

   *Solution:* Let $A_1, \ldots, A_N$ be a collection of convex sets in $\mathbf{R}^n$, $N \geq n+1$, such that every $n+1$ sets from the collection have a point in common, and let us prove that all sets have a point in common. For a collection $\iota = \{\iota_1 < \iota_2 < \ldots < \iota_{n+1}\}$ of $n+1$ distinct from each other indices from $\{1, 2, \ldots, N\}$, let $x_\iota$ be a point from the (nonempty!) set $A_{\iota_1} \cap A_{\iota_2} \cap \ldots \cap A_{\iota_{n+1}}$, and let

$\overline{A}_j := \mathrm{Conv}(\{x_\iota : j \in \iota\})$. Note that $\overline{A}_j$ is the convex hull of points from $A_j$ (since $x_\iota \in A_j$ whenever $j \in \iota$), and thus $\overline{A}_j \subset A_j$ (as $A_j$ is convex). The sets $\overline{A}_1, \ldots, \overline{A}_N$ are convex hulls of finite sets and as such are polyhedrally representable and therefore polyhedral. Every $n+1$ sets $\overline{A}_{\iota_1}, \overline{A}_{\iota_2}, \ldots, \overline{A}_{\iota_{n+1}}$, $\iota_1 < \iota_2 < \ldots < \iota_{n+1} \leq N$, have a point in common, namely, $x_{\iota_1,\ldots,\iota_{n+1}}$. Then, by item 2, all the sets $\overline{A}_j$, $j \leq N$, have a point in common, and this point is a common point of $A_1, \ldots, A_N$, since, as we already know, $\overline{A}_j \subset A_j$.

**Exercise I.24.** $A_0, A_1, \ldots, A_m$, $m = 2025$, are nonempty convex subsets of $\mathbf{R}^{2000}$, and $A_0$ is a triangle (convex hull of 3 affinely independent vectors). Which of the claims below are always (that is, for any $A_0, \ldots, A_m$ satisfying the above assumptions) true:

1. If every 3 among the sets $A_0, \ldots, A_m$ have a point in common, all $m+1$ sets have a point in common.
2. If every 4 among the sets $A_0, \ldots, A_m$ have a point in common, all $m+1$ sets have a point in common.
3. If every 2001 among the sets $A_0, \ldots, A_m$ have a point in common, all $m+1$ sets have a point in common.

*Solution:* The true statements are the second and the third ones. To see that the second statement is true, let us define $\overline{A}_i := A_i \cap A_0$. Then, we get $m+1$ convex sets such that every 3 of them intersect (since the intersection of a triple of $\overline{A}$-sets is the same as intersection of four of $A$-sets) and *all of them belong to the affine plane* $\Pi$ *of dimension 2* (namely, the affine span of the triangle $A_0$). Applying Helly's theorem to the sets $\overline{A}_i$ (treated as the subsets of the 2-dimensional affine plane), we conclude that all of them have a point in common, and this point, of course, is a common point of $A_0, \ldots, A_m$. Since the second statement is true, so is the third (the third statement is true even without assumption that $A_0$ is a triangle).

To see that the first statement can be incorrect, consider the following 4 sets in $\mathbf{R}^3$: $B_0$ is a triangle in the plane $L := \{x \in \mathbf{R}^3 : x_3 = 0\}$, and $B_i := \{x \in \mathbf{R}^3 : [x_1, x_2] \in B_0,\ x_3 = \frac{1}{2} - \lambda_i(x_1, x_2)\}$, $1 \leq i \leq 3$, where $\lambda_i(x_1, x_2)$ are the barycentric coordinates of $[x_1; x_2] \in L$, that, is coefficients in the representation of $[x_1; x_2]$ as the linear combination of the 3 vertices of $B_0$ with sum of coefficients equal to 1. Note that we have $B_0 := \{[x_1; x_2; x_3] : \lambda_i(x_1, x_2) \geq 0,\ i \leq 3,\ x_3 = 0\}$. Let us check that every 3 of our 4 sets $B_0, B_1, B_2, B_3$ have a point in common. Indeed, if the triple of sets in question does not contain $B_0$, the common point is $[\bar{x}_1; \bar{x}_2; \bar{x}_3]$, where $[\bar{x}_1; \bar{x}_2]$ is the barycenter of $B_0$ (the average of its vertices), so that $\lambda_i(\bar{x}_1, \bar{x}_2) = 1/3$, $1 \leq i \leq 3$, and $\bar{x}_3 = \frac{1}{2} - \frac{1}{3}$. Now let us verify that if triple of our sets includes $B_0$, the sets from the triple still have a point in common. By symmetry, it suffices to check this for the triple $B_0$, $B_1$, $B_2$, for which the common point is $[\tilde{x}_1; \tilde{x}_2; 0]$, with $\lambda_1([\tilde{x}_1; \tilde{x}_2]) = \lambda_2([\tilde{x}_1; \tilde{x}_2]) = \frac{1}{2}$, $\lambda_3([\tilde{x}_1; \tilde{x}_2]) = 0$ (that is, $[\tilde{x}_1; \tilde{x}_2]$ is the midpoint of a properly selected side of the triangle $B_0$). Thus, every 3 of the four sets $B_0, B_1, B_2, B_3$ have a point in common, while all four sets have no such a point: indeed, such a point $x$ should have $x_3 = 0$ (since it belongs to $B_0$) and therefore $\frac{1}{2} - \lambda_i(x_1, x_2) = 0$, $i = 1, 2, 3$ (since this point belongs to $B_1, B_2, B_3$). Therefore, every 3 of the barycentric coordinates of $[x_1; x_2]$ should be equal to $1/2$, which is impossible, since their sum must be 1.

To show that the first statement is not always true, it suffices to place our 3D sets $B_0, B_1, B_2, B_3$ into 2000-dimensional space by augmenting 3 entries in point from $R^3$ by 1997 zero entries; as a result, we get 4 convex sets $A_0, A_1, A_2, A_3$ in $\mathbf{R}^{2000}$ such that the first of them is triangle, every 3 of the sets have a point in common, but all 4 sets have no such a point. Augmenting the 4 sets $A_i$ we have built by 2021 copies of one of them, say, $A_0$, we get a family of 2025 convex sets in $\mathbf{R}^{2000}$ such that every 3 of them have a point in common, but the intersection of all sets is empty, which is a counterexample for the first statement.

**Exercise I.25.** Let $P_i := \{x \in \mathbf{R}^n : A_i x \leq b_i\}$ for $i \in \{1, \ldots, m\}$ and $C := \{x \in \mathbf{R}^n : Dx \geq d\}$ be nonempty polyhedral sets. Suppose that for any $n+1$ sets, $P_{i_1}, \ldots, P_{i_{n+1}}$, there is a translate of $C$, i.e., the set $C + u$ for some $u \in \mathbf{R}^n$, which is contained in all $P_{i_1}, \ldots, P_{i_{n+1}}$. Prove that there is a translate of $C$, which is contained in all of the sets $P_1, \ldots, P_m$.

*Solution:* For every $i = 1, \ldots, m$, we define the set $C_i := \{u \in \mathbf{R}^n : P_i \supseteq C + u\}$. Note that $C_i$ is a convex set for every $i$. Indeed, if $u+c \in P_i$ and $v+c \in P_i$ for all $c \in C$, then for $\lambda \in [0,1]$ and $c \in C$ one has $[\lambda u+(1-\lambda)v]+c = \lambda[u+c]+(1-\lambda)[v+c] \in P_i$ by convexity of $P_i$, implying that $\lambda u+(1-\lambda)v \in C_i$. From the statement of the problem we know that every $n+1$ sets $C_i$ have a non-empty intersection. From Helly's Theorem, we deduce that all of them have a non-empty intersection. In other words, there is a $u \in \mathbf{R}^n$ such that $P_i \supseteq C+u$ for every $i \in \{1,\dots,m\}$.

**Exercise I.26.** A cake contains 300 g of raisins (you may think of every one of them as of a 3D ball of positive radius). John and Jill are about to divide the cake according to the following rules:

- first, Jill chooses a point $a$ in the cake;
- second, John makes a *cut* through $a$, that is, chooses a 2D plane $\Pi$ passing through $a$ and takes the part of the cake on one side of the plane (both $\Pi$ and the side are up to John, with the only restriction that the plane should pass through $a$); all the rest goes to Jill.

1. Prove that it may happen that Jill cannot guarantee herself 76 g of the raisins.

   *Solution:* Suppose there are 4 raisins, 75 g each, placed in the vertices of large tetrahedron; whatever point Jill chooses, John can cut off 3 of the four raisins.

2. Prove that Jill always can choose $a$ in a way which guarantees her at least 74 g of the raisins.
3. Consider $n$-dimensional version of the problem, where the raisins are $n$-dimensional balls, the cake is a domain in $\mathbf{R}^n$, and "a cut" taken by John is defined as the part of the cake contained in the half-space

$$\left\{x \in \mathbf{R}^n : \ e^\top(x-a) \geq 0\right\},$$

   where $e \neq 0$ is the vector ("inner normal to the cutting hyperplane") chosen by John. Prove that for every $\epsilon > 0$, Jill can guarantee to herself at least $\frac{300}{n+1} - \epsilon$ g of raisins, but in general cannot guarantee to herself $\frac{300}{n+1} + \epsilon$ g.

*Solution:*
(2-3): Let us consider the case of $n=3$ (generalization to arbitrary $n$ will be evident).
For every direction (that is, unit vector) $d \in \mathbf{R}^3$ consider the closed half-spaces

$$\left\{x \in \mathbf{R}^3 : \ d^\top x \leq \alpha\right\},$$

and let us look at the mass of raisins outside of such a half-space. This mass is clearly a continuous function of $\alpha$ (since the distribution of raisins' mass has density) which is close to 300 when $\alpha$ is very negative and close to 0 when $\alpha$ is very positive. It follows that there exists the largest $\alpha = \alpha(d)$ such that the mass of the raisins outside the half-space

$$H_d := \left\{x \in \mathbf{R}^3 : \ d^\top x \leq \alpha(d)\right\}$$

is exactly 74 g. Note that

> (!) *If John takes himself the part of the cake in the half-space* $\{x \in \mathbf{R}^3 : \ d^\top x \leq d^\top \bar{x}\}$ *with* $\bar{x} \in H_d$ (that is, $d$ is exactly the outer normal to the cut chosen by John, and this cut passes through $\bar{x}$), *then Jill gets at least 74 g of raisins.*

In view of (!), it suffices to prove that the intersection of all sets $H_d$ is nonempty. Indeed, in this case Jill can choose, as the point through which the cut should pass, a point in $\bigcap\limits_d H_d$; then whatever John will do, his cut will be as explained in (!) with certain $d$, and therefore Jill will get at least 74 g of raisins. To prove that $\bigcap\limits_d H_d$ is nonempty, we can use Helly Theorem II. Let us check its assumptions: The sets $H_d$ indeed are closed and convex sets in $\mathbf{R}^3$. The intersection of every 4 sets $H_d$ indeed is nonempty, since, assuming the opposite, the complements of the 4 sets with empty intersection would together cover the entire space; but every one of these complements contains 74 g of raisins, and therefore the union of 4 of them can contain at most $4 \cdot 74 = 296$ g of raisins, while the entire space contains 300 g of raisins. It remains to verify that one can choose among the sets $H_d$ finitely many sets with bounded intersection. This is evident, since the intersection of $H_{e_i}$ and $H_{-e_i}$ ($e_1, e_2, e_3$ are basic orth) is a stripe $a_i \leq x_i \leq b_i$ with finite $a_i, b_i$, so that the intersection of the 6 sets $H_{e_i}$, $H_{-e_i}$, $i=1,2,3$, is a bounded box.
Generalization to the $n$-dimensional case is evident.

**Remarks:**

1. With some minor effort, you can prove that Jill can find a point which guarantees her $\frac{300}{n+1}$ g of raisins, and not $\frac{300}{n+1} - \epsilon$ g.
2. If, instead of dividing raisins, John and Jill would divide in the same fashion *uniform and convex* cake (that is, a closed and bounded convex body $X$ with a nonempty interior in $\mathbf{R}^n$, the reward being the $n$-dimensional volume of the part a person gets), the results would change dramatically: choosing as the point the center of masses of the cake
$$\bar{x} := \frac{\int\limits_X x dx}{\int\limits_X dx},$$
Jill would guarantee herself at least $\left(\frac{n}{n+1}\right)^n \approx \frac{1}{e}$ part of the cake. This is a not so easy corollary of the following extremely important and deep result:

> **Brunn-Minkowski Symmetrization Theorem:** *Let $X$ be as above, and let $[a,b]$ be the projection of $X$ on an axis $\ell$, say, on the last coordinate axis. Consider the " symmetrization" $Y$ of $X$, i.e., $Y$ is the set with the same projection $[a,b]$ on $\ell$ and for every hyperplane orthogonal to the axis $\ell$ and crossing $[a,b]$, the intersection of $Y$ with this hyperplane is an $(n-1)$-dimensional ball centered at the axis with precisely the same $(n-1)$-dimensional volume as the one of the intersection of $X$ with the same hyperplane:*
> $$\begin{aligned}&\left\{z \in \mathbf{R}^{n-1} : [z;c] \in Y\right\} = \left\{z \in \mathbf{R}^{n-1} : \|z\|_2 \le \rho(c)\right\}, \quad \forall c \in [a,b], \text{ and}\\ &\mathrm{Vol}_{n-1}\left(\left\{z \in \mathbf{R}^{n-1} : [z;c] \in Y\right\}\right) = \mathrm{Vol}_{n-1}\left(\left\{z \in \mathbf{R}^{n-1} : [z;c] \in X\right\}\right), \quad \forall c \in [a,b].\end{aligned}$$
> *Then, $Y$ is a closed convex set.*

## Around Polyhedral Representations

**Exercise I.27.** Justify the calculus rules for polyhedral representations presented in Section 3.3.

*Solution:* This is straightforward.

**Exercise I.28.** Given two sets $U, V \subseteq \mathbf{R}^m$, we define
$$U + V = \{x \in \mathbf{R}^m : \exists u \in U, \exists v \in V \text{ such that } x = u + v\}.$$
Let $D := \{x \in \mathbf{R}^n : Ax + b + Q_s \subseteq P, \forall s \in S\}$ where the nonempty set $P \subset \mathbf{R}^m$ admits polyhedral representation, the nonempty set $S \subset \mathbf{R}^k$ is given but arbitrary, and the nonempty sets $Q_s \subset \mathbf{R}^m$ are indexed by $s \in S$.

1. Suppose that $S$ is a finite set and for each $s \in S$ we have $Q_s = \{q_s\}$, i.e., is a single point. Then, will the set $D$ be polyhedrally representable?
2. State sufficient conditions on the structure of sets $Q_s$ and $S$ that will guarantee that the resulting set $D$ is polyhedral. Here, the goal is to have conditions as general as possible. Among your sufficient conditions, can you identify at least some of those that are necessary?

*Solution:*

1. This part follows immediately from the next one.
2. Note that $x \in D$ if and only if $Ax + b + q \in P$ holds for all $q \in \bigcup_{s\in S} Q_s$. Since $P$ is polyhedrally representable, let $P = \{y \in \mathbf{R}^m : Gy \le g\}$. Then, $x \in D$ if and only if $G(Ax + b + q) \le g$ for all $q \in \bigcup_{s\in S} Q_s$, i.e., if and only if $x$ satisfies
$$[GA]_i^\top x \le \inf_q \left\{ [(g - G(b+q))]_i : q \in \bigcup_{s\in S} Q_s \right\}, \qquad \text{for all rows } i.$$

Clearly this is a polyhedral representation of $D$ without making any assumptions on the structure of $S$ or $Q_s$.

**Exercise I.29.** For $x \in \mathbf{R}^n$ and integer $k$, $1 \leq k \leq n$, let $s_k(x)$ be the sum of $k$ largest entries in $x$. For example, $s_1(x) = \max_i\{x_i\}$, $s_n(x) = \sum_{i=1}^n x_i$, $s_3([3;1;2;2]) = 3+2+2 = 7$. Now let $1 \leq k \leq n$ be two integers. For any integer $k = 1, \ldots, n$, define

$$X_{k,n} := \{[x;t] \in \mathbf{R}^n \times \mathbf{R} : \; s_k(x) \leq t\}.$$

Observe that $X_{k,n}$ is a polyhedral set. Indeed, $s_k(x) \leq t$ holds if and only if for every $k$ indices $i_1 < i_2 < \ldots < i_k$ from $\{1, 2, \ldots, n\}$ we have $x_{i_1} + x_{i_2} + \ldots + x_{i_k} \leq t$, which is nothing but a linear inequality in variables $x, t$. Since there are $\binom{n}{k}$ possible ways of selecting $k$ indices from $\{1, 2, \ldots, n\}$, the number of linear inequalities describing $X_{k,n}$ is $\binom{n}{k}$, and these linear inequalities give the polyhedral description of $X_{k,n}$. The point of this exercise is to demonstrate that $X_{k,n}$ admits a "short" polyhedral representation, specifically,

$$X_{k,n} = \left\{ [x;t] \in \mathbf{R}^n \times \mathbf{R} : \; \exists z \in \mathbf{R}^n, \, \exists s \in \mathbf{R} \text{ s.t. } x_i \leq z_i + s, \, \forall i, \; z \geq 0, \; \sum_{i=1}^n z_i + ks \leq t \right\}. \tag{$*$}$$

*Solution:* Let $\overline{X}_{k,n}$ be the right hand side set in $(*)$, we will prove that (a) $X_{k,n} \subseteq \overline{X}_{k,n}$ and (b) $\overline{X}_{k,n} \subseteq X_{k,n}$.

(a): Observe first of all that both $X_{k,n}$ and $\overline{X}_{k,n}$ are "permutationally symmetric in $x$", meaning that when $[x;t] \in X_{k,n}$ and $\bar{x}$ is obtained from $x$ by permuting entries, we have $[\bar{x};t] \in X_{k,n}$, and similarly $[x;t] \in \overline{X}_{k,n}$ implies $[\bar{x};t] \in \overline{X}_{k,n}$. It follows that in order to verify (a) it suffices to prove that if $[x;t] \in X_{k,n}$ and $x_1 \geq x_2 \geq \ldots \geq x_n$, then $[x;t] \in \overline{X}_{k,n}$. This is immediate: set $z_i := x_i - x_k$ for $i \leq k$ and $z_i := 0$ for $i > k$, and $s := z_i$. Taking into account that the entries in $x$ form a non-ascending sequence, we immediately see that $z \geq 0$, $x_i \leq z_i + s$ for all $i$, and $s_k(x) = \sum_{i=1}^k x_i = \sum_{i=1}^k z_i + ks = \sum_{i=1}^n z_i + ks$. Recalling that $[x;t] \in X_{k,n}$, that is, $s_k(x) \leq t$, we conclude that $x, t, z, s$ satisfy all inequalities participating in the description of $\overline{X}_{k,n}$, that is, $[x;t] \in \overline{X}_{k,n}$. (a) is proved.

(b): let $x, t, z, s$ satisfy all inequalities in the description of $\overline{X}_{k,n}$. When $i_1 < i_2 < \ldots < i_k$ is an ordered collection of $k$ indices from $\{1, \ldots, n\}$, we have by the inequalities describing $\overline{X}_{k,n}$ that

$$x_{i_1} + x_{i_2} + \ldots + x_{i_k} \leq ks + z_{i_1} + z_{i_2} + \ldots + z_{i_k} \leq ks + \sum_{i=1}^n z_i \leq t,$$

where the second inequality is due to $z \geq 0$. The resulting inequality holds true for all ordered collections of $k$ indices $i_1, \ldots, i_k$, implying that $s_k(x) \leq t$. Thus, $[x;t] \in \overline{X}_{l,n}$ implies $[x;t] \in X_{k,n}$, as claimed in (b).

**Exercise I.30.** [Computational study: Fourier-Motzkin elimination as an LP algorithm] It was mentioned in section 3.2.1 that Fourier-Motzkin elimination provides us with an algorithm for solving LP problems that terminates in finitely many steps. This algorithm, however, is of no computational value due to the potential rapid growth of the number of inequalities one may need to handle when eliminating more and more variables. The goal of this exercise is to get an impression of this phenomenon.

Our "guinea pig" will be transportation problem with $n$ unit capacity suppliers and $n$ unit demand customers:

$$\min_{x,t} \left\{ t : \; t \geq \sum_{i=1}^n \sum_{i=1}^n c_{ij} x_{ij}, \; \sum_i x_{ij} \geq 1, \, \forall j, \; \sum_j x_{ij} \leq 1, \, \forall i, \; x_{ij} \geq 0, \forall i, j \right\}.$$

This problem has $n^2 + 1$ variables and $(n+1)^2$ linear inequality constraints, and let us solve it by applying the Fourier-Motzkin elimination to project the feasible set of the problem onto the axis of the $t$-variable, that is, to build a finite system $\mathcal{S}$ of univariate linear inequalities specifying this projection.

How many inequalities do you think there will be in $\mathcal{S}$ when $n = 1, 2, 3, 4$? Check your intuition by implementing and running the F-M elimination, assuming, for the sake of definiteness, that $c_{ij} = 1$ for all $i, j$.

*Solution:* Our results are as follows (your numbers could be different, since the outcome depends on the serial numbers assigned to the $x$-variables):

| $n$ | $m_{\rm ini}$ | $m_{\rm fin}$ |
|---|---|---|
| 1 | 4 | 4 |
| 2 | 9 | 19 |
| 3 | 16 | 44,854 |
| 4 | 25 | † |

$m_{\rm ini}$ and $m_{\rm fin}$ are the number of inequalities in the initial and the final systems

†When $n = 4$, we are supposed to eliminate $n^2 = 16$ of 17 variables in a system with 25 linear inequality constraints on these 17 variables. Eliminating the last 11 variables results in system of 974,236 constraints with 6 variables. Eliminating in this system the last – the sixth – of the variables would result in a system of 121,226,850 linear inequalities with 5 variables; building this system was terminated after the number of assembled so far inequalities reached our a priori limit $2^{23} = 8,388,608$.

## Around General Theorem on Alternative

**Exercise I.31.**

1. Prove Gordan's Theorem on Alternative:

   > *A system of strict homogeneous linear inequalities $Ax < 0$ in variables $x$ has a solution if and only if the system $A^\top\lambda = 0,\ \lambda \geq 0$ in variables $\lambda$ has only the trivial solution $\lambda = 0$.*

   *Solution:* By GTA, the system has no solutions if and only if by using nonnegative aggregations with weights $\lambda$ of the inequalities in the system we can derive a contradictory consequence inequality, i.e., $[A^\top\lambda]^\top x\,\Omega\,0$, where $\Omega =$ " $<$ " when $\lambda \neq 0$ and $\Omega =$ " $\leq$ " when $\lambda = 0$. Note that the only two possible contradictory linear inequalities are of the form either $0^\top x < \epsilon$ with $\epsilon \leq 0$ or $0^\top x \leq \epsilon'$ with $\epsilon' < 0$. In our case, when $\lambda = 0$ and so $\Omega =$ " $\leq$ ", the right-hand side of the consequence inequality will be zero, so the second option cannot lead to any contradictory inequality. Thus, we deduce that when given the system $Ax < 0$, we can derive a contradictory inequality if and only if $\Omega =$ " $<$ " and $A^\top\lambda = 0$ for some nonzero $\lambda \geq 0$. Thus, $Ax < 0$ has no solutions if and only if there exists a nonzero vector $\lambda \geq 0$ such that $A^\top\lambda = 0$.

2. Prove Motzkin's Theorem on Alternative:

   > *A system $Ax < 0,\ Bx \leq 0$ of strict and nonstrict homogeneous linear inequalities has a solution if and only if the system $A^\top\lambda + B^\top\mu = 0,\ \lambda \geq 0,\ \mu \geq 0$ in variables $\lambda, \mu$ has no solution with $\lambda \neq 0$.*

   *Solution:* Same as above, the infeasibility of the system is equivalent to the existence of nonnegative weights $\lambda, \mu$ resulting in a contradictory consequence inequality $[A^\top\lambda + B^\top\mu]^\top x\,\Omega\,0$ with $\Omega =$ " $<$ " when $\lambda \neq 0$ and $\Omega =$ " $\leq$ " when $\lambda = 0$. Because the given system of constraints is homogeneous, the only one of these two options which can lead to a contradictory consequence inequality is that $A^\top\lambda + B^\top\mu = 0$ and $\lambda \neq 0$.

**Exercise I.32.** For the systems of constraints to follow, write them down equivalently in the standard form $Ax < b, Cx \leq d$ and point out their feasibility status ("feasible – infeasible") along with the corresponding certificates (certificate for feasibility is a feasible solution to the system; certificate for infeasibility is a collection of weights of constraints which leads to a contradictory consequence inequality, as explained in GTA).

1. $x \leq 0$ $(x \in \mathbf{R}^n)$

   *Solution:* already in the standard form, feasible, feasibility certificate $x = 0$.

2. $x \le 0$, and $\sum_{i=1}^n x_i > 0$ $(x \in \mathbf{R}^n)$

   *Solution:* the standard form is given by $-\sum_i x_i < 0$, and $x \le 0$, infeasible, infeasibility certificate $\lambda = [1;\dots;1] \in \mathbf{R}^{n+1}$

3. $-1 \le x_i \le 1,\ 1 \le i \le n,\ \sum_{i=1}^n x_i \ge n$ $(x \in \mathbf{R}^n)$

   *Solution:* the standard form is given by $-\sum_i x_i \le -n,\ x \le [1;\dots;1],\ -x \le [1;\dots;1]$, feasible, feasibility certificate $x = [1;\dots;1]$.

4. $-1 \le x_i \le 1,\ 1 \le i \le n,\ \sum_{i=1}^n x_i > n$ $(x \in \mathbf{R}^n)$

   *Solution:* the standard form is given by $-\sum_i x_i < -n,\ x \le [1;\dots;1],\ -x \le [1;\dots;1]$, infeasible, infeasibility certificate is $\lambda = [1;1;\dots;1;0;\dots;0]$ ($n$ zeros).

5. $-1 \le x_i \le 1,\ 1 \le i \le n,\ \sum_{i=1}^n ix_i \ge \frac{n(n+1)}{2}$ $(x \in \mathbf{R}^n)$

   *Solution:* the standard form is given by $-\sum_i ix_i \le -\frac{n(n+1)}{2},\ x \le [1;\dots;1],\ -x \le [1;\dots;1]$, feasible, feasibility certificate $x = [1;\dots;1]$.

6. $-1 \le x_i \le 1,\ 1 \le i \le n,\ \sum_{i=1}^n ix_i > \frac{n(n+1)}{2}$ $(x \in \mathbf{R}^n)$

   *Solution:* the standard form is given by $-\sum_i ix_i < -\frac{n(n+1)}{2},\ x \le [1;\dots;1],\ -x \le [1;\dots;1]$, infeasible, infeasibility certificate is $\lambda = [1;1;2;3;\dots;n;0;\dots;0]$ ($n$ zeros).

7. $x \in \mathbf{R}^2,\ |x_1| + x_2 \le 1,\ x_2 \ge 0,\ x_1 + x_2 = 1$

   *Solution:* the standard form is given by $-x_1+x_2 \le 1,\ x_1+x_2 \le 1,\ -x_2 \le 0,\ x_1+x_2 \le 1,\ -x_1-x_2 \le -1$, feasible, feasibility certificate $x = [1;0]$.

8. $x \in \mathbf{R}^2,\ |x_1| + x_2 \le 1,\ x_2 \ge 0,\ x_1 + x_2 > 1$

   *Solution:* the standard form is given by $-x_1 + x_2 \le 1,\ x_1 + x_2 \le 1,\ -x_2 \le 0,\ -x_1 - x_2 < -1$, infeasible, infeasibility certificate is $\lambda = [0;1;0;1]$.

9. $x \in \mathbf{R}^4$, $x \ge 0$, the sum of two largest entries in $x$ does not exceed 2, and $x_1 + x_2 + x_3 \ge 3$

   *Solution:* the standard form is given by $-x \le 0,\ x_i + x_j \le 2,\ 1 \le i < j \le 4,\ -x_1 - x_2 - x_3 \le -3$, feasible, feasibility certificate $x = [1;1;1;0]$.

10. $x \in \mathbf{R}^4$, $x \ge 0$, the sum of two largest entries in $x$ does not exceed 2, and $x_1 + x_2 + x_3 > 3$

    *Solution:* the standard form is given by $-x \le 0,\ x_i + x_j \le 2,\ 1 \le i < j \le 4,\ -x_1 - x_2 - x_3 < -3$, infeasible, infeasibility certificate is as follows: sum up inequalities $x_1+x_2 \le 2$, $x_2+x_3 \le 2$, $x_1+x_3 \le 2$ with weights 1/2 and add the inequality $-x_1 - x_2 - x_3 < -3$ with weight 1.

**Exercise I.33.** Let $(\mathcal{S})$ be the following system of linear inequalities in variables $x \in \mathbf{R}^3$

$$x_1 \le 1,\ x_1 + x_2 \le 1,\ x_1 + x_2 + x_3 \le 1 \qquad (\mathcal{S})$$

In the following list, point out which inequalities are/are not consequences of this system, and certify your claims. To certify that a given inequality is a consequence of the given system, you need to provide nonnegative aggregation weights $\lambda \in \mathbf{R}^3_+$ for the inequalities in $(\mathcal{S})$ such that the resulting consequence inequality implies the given inequality. To certify that a given inequality is not a consequence of the given system $(\mathcal{S})$, you need to find a point $x \in \mathbf{R}^3$ that satisfies the given system but violates the given inequality.

1. $3x_1 + 2x_2 + x_3 \le 4$

   *Solution:* This is a consequence of the system with the certificate $\lambda = [1;1;1]$, i.e., when taking weighted sum of the inequalities from the system with weights $\lambda_1, \lambda_2, \lambda_3$, we get the inequality $3x_1 + 2x_2 + x_3 \le 3$, which clearly implies the target inequality.

2. $3x_1 + 2x_2 + x_3 \le 2$

   *Solution:* This is not a consequence of the system. A certificate for this is $x = [1;0;0]$ – this vector is feasible to the system but does not satisfy the inequality $3x_1 + 2x_2 + x_3 \le 2$.

3. $3x_1 + 2x_2 \le 3$

   *Solution:* This is a consequence of the system, the certificate being $\lambda = [1;2;0]$.

4. $3x_1 + 2x_2 \leq 2$

   *Solution:* This is not a consequence of the system, a certificate being $x = [1; 0; 0]$.

5. $3x_1 + 3x_2 + x_3 \leq 3$

   *Solution:* This is a consequence of the system, a certificate being $\lambda = [0; 2; 1]$.

6. $3x_1 + 3x_2 + x_3 \leq 2$

   *Solution:* This is not a consequence of the system, a certificate being $x = [1; 0; 0]$.

Make a generalization: prove that a linear inequality $px_1 + qx_2 + rx_3 \leq s$ is a consequence of $(\mathcal{S})$ if and only if $s \geq p \geq q \geq r \geq 0$.

*Solution:* By Inhomogeneous Farkas Lemma, an inequality is a consequence of the (feasible!) system $(\mathcal{S})$ if and only if there exists a nonnegative vector $\lambda \in \mathbf{R}^3_+$ such that $\lambda_1[1; 0; 0] + \lambda_2[1; 1; 0] + \lambda_3[1; 1; 1] = [p; q; r]$ and $\lambda_1 + \lambda_2 + \lambda_3 \leq s$, which is equivalent to $p = \lambda_1 + \lambda_2 + \lambda_3,\ q = \lambda_1 + \lambda_2,\ r = \lambda_3,\ p \leq s$, which in turn is equivalent to $s \geq p \geq q \geq r \geq 0$.

**Exercise I.34.** Is the inequality $x_1 + x_2 \leq 1$ a consequence of the system $x_1 \leq 1, x_1 \geq 2$? If yes, can it be obtained by taking a legitimate weighted sum of inequalities from the system and the identically true inequality $0^\top x \leq 1$, as it is suggested by the Inhomogeneous Farkas Lemma?

*Solution:* The given system is infeasible, and therefore *every* inequality is a consequence of the system. The consequence in question cannot be obtained by aggregating inequalities from the system and the identically true inequality $0^\top x \leq 1$, since in every aggregation of this type the coefficient at $x_2$ is zero. There is no contradiction with the Inhomogeneous Farkas Lemma, since the latter deals with *feasible* systems of inequalities and this is not applicable in our case.

**Exercise I.35.** Certify the correct statements in the following list:

1. The polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 1\right\}$ is nonempty.

   *Solution:* A certificate is $x = [1/3; 1/3; 1/3] \in X$.

2. The polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 0.99\right\}$ is empty.

   *Solution:* A certificate is $\lambda = [-1; -1; -1; 1]$: by taking weighted sum of the inequalities defining $X$ using these weights $\lambda$ is legitimate and leads to the contradictory inequality $0^\top x \leq -0.01$.

3. The linear inequality $x_1 + x_2 + x_3 \geq 2$ is violated somewhere on the polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 1\right\}$.

   *Solution:* A certificate is $x = [1/3; 1/3; 1/3]$: this point belongs to $X$ but does not satisfy the given inequality.

4. The linear inequality $x_1 + x_2 + x_3 \geq 2$ is violated somewhere on the polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 0.99\right\}$.

   *Solution:* This statement is false: $X$ is empty, and therefore every linear inequality is satisfied everywhere on $X$.

5. The linear inequality $x_1 + x_2 \leq 3/4$ is satisfied everywhere on the polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 1.05\right\}$.

   *Solution:* A certificate is $\lambda = [0; 0; -1; 1]$ – taking weighted sum of the inequalities $x_1 \geq 1/3$, $x_2 \geq 1/3$, $x_3 \geq 1/3$, $x_1 + x_2 + x_3 \leq 1.05$ with the weights $\lambda_1, \ldots, \lambda_4$, we get the inequality $x_1 + x_2 \leq 1.05 - 1/3 < 3/4$.

6. The polyhedral set $Y = \left\{x \in \mathbf{R}^3 :\ x_1 \geq 1/3,\ x_2 \geq 1/3,\ x_3 \geq 1/3\right\}$ is not contained in the polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 1\right\}$.

   *Solution:* A certificate is $x = [1; 1; 1]$: this point is contained in $Y$ but it is not contained in $X$.

7. The polyhedral set $Y = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 1\right\}$ is contained in the polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x_1 + x_2 \leq 2/3,\ x_2 + x_3 \leq 2/3,\ x_1 + x_3 \leq 2/3\right\}$.

*Solution:* It suffices to certify that every one of the constraints defining $X$ is valid for $Y$, i.e., is a consequence of the constraints defining $Y$. The inequality $x_1 + x_2 \leq 2/3$ can be obtained as the weighted sum of the inequalities $x_1 \geq 1/3,\ x_2 \geq 1/3,\ x_3 \geq 1/3,\ x_1 + x_2 + x_3 \leq 1$ using the weights $0, 0, -1, 1$, and thus it is valid for $Y$. Similarly, we can certify that the inequalities $x_1 + x_3 \leq 2/3$ and $x_2 + x_3 \leq 2/3$ are valid for $Y$.

## Around Linear Programming Duality

**Exercise I.36.** Let the polyhedral set $P = \{x \in \mathbf{R}^n : Ax \leq b\}$, where $A = [a_1^\top; ...; a_m^\top]$, be nonempty. Prove that $P$ is bounded if and only if every vector from $\mathbf{R}^n$ can be represented as a linear combination of the vectors $a_i$ with nonnegative coefficients where at most $n$ coefficients are positive. As a result, given $A$, all nonempty sets of the form $\{x \in \mathbf{R}^n : Ax \leq b\}$ simultaneously are/are not bounded.

*Solution:* $P$ is bounded if and only if for every $z \in \mathbf{R}^n$ the feasible LP program

$$\max_x \left\{ z^\top x : \ Ax \leq b \right\}$$

is bounded and thus is solvable, or, which is the same by LP Duality Theorem, if and only if the dual of this problem, i.e.,

$$\min_\lambda \left\{ b^\top \lambda : \ \lambda \geq 0, A^\top \lambda = z \right\}$$

is solvable. Next, since the dual of this dual is feasible ($P$ is nonempty!), the dual automatically is bounded, so that its solvability is the same as its feasibility. We conclude that $P$ is bounded if and only if every $x \in \mathbf{R}^n$ is a conic combination of $a_i$'s.

Applying the Caratheodory Theorem in conic form, the latter is the same as the possibility to represent every $z \in \mathbf{R}^n$ as a conic combination of at most $n$ vectors from the collection $a_1, ..., a_m$.

**Exercise I.37.** Consider the linear program

$$\mathrm{Opt} = \max_{x \in \mathbf{R}^2} \{x_1 : \ x_1 \geq 0, \ x_2 \geq 0, \ ax_1 + bx_2 \leq c\} \qquad (P)$$

where $a, b, c$ are parameters. Answer the following questions:

1. Let $c = 1$. Is the problem feasible?

   *Solution:* The problem is feasible; a certificate is a feasible solution, e.g., $x = 0$.

2. Let $a = b = 1, \ c = -1$. Is the problem feasible?

   *Solution:* The problem is infeasible; a certificate for infeasibility is given by $\lambda = [-1; -1; 1]$: summing up the constraints $x_1 \geq 0, \ x_2 \geq 0, \ x_1 + x_2 \leq -1$ with weights $-1, -1, 1$, we get the contradictory inequality $0^\top x \leq -1$.

3. Let $a = b = 1, \ c = -1$. Is the problem bounded[5]?

   *Solution:* The problem is bounded since it is infeasible; certificate of infeasibility was given in the previous item.

4. Let $a = b = c = 1$. Is the problem bounded?

   *Solution:* The problem is bounded since the weighted sum of the constraints $x_1 \geq 0, \ x_2 \geq 0, \ x_1 + x_2 \leq 1$, the weights being $0, -1, 1$, gives us the consequence inequality $x_1 \leq 1$. Thus, the objective $\max x_1$ is bounded from above on the feasible set.

5. Let $a = 1, \ b = -1, \ c = 1$. Is the problem bounded?

[5] Recall that a maximization problem is called *bounded*, if the objective is bounded from above on the feasible set, which is the same as its optimal value being $< \infty$

*Solution:* The problem is unbounded. Indeed, setting $d = [1; 1]$, we get $d_1 \geq 0$, $d_2 \geq 0$, $ad_1 + bd_2 = d_1 - d_2 \leq 0$, $[1; 0]^\top d = d_1 > 0$, implying that the points $x(t) := td$ are feasible when $t \geq 0$; it remains to note that the objective, as evaluated at $x(t)$, tends to $+\infty$ as $t \to \infty$.

6. Let $a = b = c = 1$. Is it true that $\mathrm{Opt} \geq 0.5$?

   *Solution:* A certificate is a feasible solution with objective value $\geq 0.5$, e.g., the solution $x = [0.5; 0.5]$.

7. Let $a = b = 1,\ c = -1$. Is it true that $\mathrm{Opt} \leq 1$?

   *Solution:* The claim is true due to the fact that the problem is infeasible, for infeasibility certificate see item 2, and thus $\mathrm{Opt} = -\infty$.

8. Let $a = b = c = 1$. Is it true that $\mathrm{Opt} \leq 1$?

   *Solution:* The claim is true, a certificate is the collection of weights (Lagrange multipliers) $0, -1, 1$; taking the corresponding weighted sum of the constraints $x_1 \geq 0,\ x_2 \geq 0,\ x_1 + x_2 \leq 1$, we get the inequality $x_1 \leq 1$. Thus, the objective does not exceed 1 on the feasible set.

9. Let $a = b = c = 1$. Is it true that $x_* = [1; 1]$ is an optimal solution of $(P)$?

   *Solution:* The claim is false since $x_*$ is infeasible (it violates the constraint $x_1 + x_2 \leq 1$).

10. Let $a = b = c = 1$. Is it true that $x_* = [1/2; 1/2]$ is an optimal solution of $(P)$?

    *Solution:* The claim is false since there exists a feasible solution $x = [1; 0]$ with larger objective value.

11. Let $a = b = c = 1$. Is it true that $x_* = [1; 0]$ is an optimal solution of $(P)$?

    *Solution:* The claim is true, the corresponding certificate is the collection of Lagrange multipliers $0, -1, 1$ associated with the constraints $x_1 \geq 0$, $x_2 \geq 0$, $x_1 + x_2 \leq 1$. Indeed, the multipliers are of proper signs, satisfy the complementary slackness condition and the KKT equation $0 \times [1; 0] - 1 \times [0; 1] + 1 \times [1; 1] = [1; 0]$.

**Exercise I.38.** Consider the LP program

$$\max_{x_1, x_2} \left\{ -x_2 : \begin{array}{c} x_1 \leq 0 \\ -x_1 \leq -1 \\ x_2 \leq 1 \end{array} \right\}$$

Write down the dual problem and check whether the optimal values are equal to each other.

*Solution:* The dual problem reads

$$\min_{\lambda_1, \lambda_2, \lambda_3} \left\{ -\lambda_2 + \lambda_3 : \begin{array}{l} \lambda_1 - \lambda_2 = 0 \\ \lambda_3 = -1 \\ \lambda_i \geq 0,\ 1 \leq i \leq 3 \end{array} \right\}$$

Both problems are clearly infeasible, and their optimal values ($-\infty$ and $+\infty$, respectively) differ from each other.

**Exercise I.39.** Write down the problems dual to the following linear programs:

1. $$\max_{x \in \mathbf{R}^3} \left\{ x_1 + 2x_2 + 3x_3 : \begin{array}{l} x_1 - x_2 + x_3 = 0, \\ x_1 + x_2 - x_3 \geq 100, \\ x_1 \leq 0, \\ x_2 \geq 0, \\ x_3 \geq 0 \end{array} \right\}$$

   *Solution:* The dual problem is

   $$\min_{\lambda \in \mathbf{R}^5} \left\{ 100\lambda_2 : \begin{array}{l} \lambda_2 \leq 0,\ \lambda_3 \geq 0,\ \lambda_4 \leq 0,\ \lambda_5 \leq 0, \\ \lambda_1 + \lambda_2 + \lambda_3 = 1, \\ -\lambda_1 + \lambda_2 + \lambda_4 = 2, \\ \lambda_1 - \lambda_2 + \lambda_5 = 3 \end{array} \right\}.$$

2. $\max\limits_{x \in \mathbf{R}^n} \left\{ c^\top x :\ Ax = b,\ x \geq 0 \right\}$

*Solution:* The dual problem is

$$\min_{\lambda=[\lambda_e;\lambda_g]}\left\{b^\top\lambda_e:\ \begin{array}{c}\lambda_g\leq 0,\\ A^\top\lambda_e+\lambda_g=c\end{array}\right\},$$

or, after eliminating $\lambda_g$:

$$\min_{\lambda_e}\left\{b^\top\lambda_e:\ c\leq A^\top\lambda_e\right\}.$$

3. $\max\limits_{x\in\mathbf{R}^n}\left\{c^\top x:\ Ax=b,\ \underline{u}\leq x\leq\overline{u}\right\}$

   *Solution:* The dual problem is

$$\min_{\lambda=[\lambda_e;\lambda_g;\lambda_\ell]}\left\{\overline{u}^\top\lambda_\ell+\underline{u}^\top\lambda_g+b^\top\lambda_e:\ \begin{array}{c}\lambda_\ell\geq 0,\ \lambda_g\leq 0,\\ \lambda_\ell+\lambda_g+A^\top\lambda_e=c\end{array}\right\}.$$

4. $\max\limits_{x,y}\left\{c^\top x:\ Ax+By\leq b,\ x\leq 0,\ y\geq 0\right\}$

   *Solution:* The dual problem is

$$\min_{\lambda=[\lambda_{\ell,b};\lambda_{\ell,0},\lambda_g]}\left\{b^\top\lambda_{\ell,b}:\ \begin{array}{l}\lambda_{\ell,b}\geq 0,\ \lambda_{\ell,0}\geq 0,\ \lambda_g\leq 0,\\ A^\top\lambda_{\ell,b}+\lambda_{\ell,0}=c,\\ B^\top\lambda_{\ell,b}+\lambda_g=0\end{array}\right\},$$

   or, after eliminating $\lambda_{\ell,0}$ and $\lambda_g$,

$$\min_{\lambda_{\ell,b}}\left\{b^\top\lambda_{\ell,b}:\ \lambda_{\ell,b}\geq 0,\ A^\top\lambda_{\ell,b}\leq c,\ B^\top\lambda_{\ell,b}\geq 0\right\}.$$

**Exercise I.40.** Consider a primal-dual pair of linear programs

$$\begin{array}{ll}\mathrm{Opt}(P)=\min\limits_{x}\left\{c^\top x:\ Ax\geq b\right\}, & (P)\\ \mathrm{Opt}(D)=\max\limits_{y}\left\{b^\top y:\ y\geq 0,\ A^\top y=c\right\}. & (D)\end{array}$$

Suppose that both are feasible. Prove that the feasible set of at least one of these problems is unbounded.

*Solution:* See solution to Exercise IV.25.

**Exercise I.41.** Consider the following linear program

$$\mathrm{Opt}=\min_{\{x_{ij}\}_{1\leq i<j\leq 4}}\left\{2\sum_{1\leq i<j\leq 4}x_{ij}:\ x_{ij}\geq 0,\,1\leq i<j\leq 4,\ \sum_{j>i}x_{ij}+\sum_{j<i}x_{ji}\geq i,\ 1\leq i\leq 4\right\}.$$

1. Show that the optimum objective value is at most 20.

   *Solution:* The solution $x_{34}=4$, $x_{23}=3$, $x_{12}=2$, and all other variables equal to 0 is a feasible solution to this LP and has the objective value equal to $2(2+3+4)=18$. Since this is a minimization problem, we deduce that $\mathrm{Opt}\leq 18$.

2. Show that the optimum objective value is at least 10.

   *Solution:* The dual of this LP is given by

$$\max_{y\in\mathbf{R}^4}\left\{\sum_{i=1}^4 iy_i:\ y_i\geq 0\ \forall i,\ y_i+y_j\leq 2,\,1\leq i<j\leq 4\right\}.$$

   The solution $y_1=y_2=y_3=y_4=1$ is feasible to the dual with an objective value of $1+2+3+4=10$. Therefore, by Weak LP Duality, we deduce that $\mathrm{Opt}\geq 10$.

**Exercise I.42.** We say that an $n \times n$ matrix $P$ is *stochastic* if all of its entries are nonnegative and the sum of the entries of each row is equal to 1. Show that if $P$ is a stochastic matrix, then there is a nonzero vector $a \in \mathbf{R}^n$ such that $a^\top P = a^\top$ and $a \geq 0$.

*Solution:* Consider the linear program

$$\min_{x \in \mathbf{R}^n} \left\{ 0^\top x : \ Px \geq x + e \right\},$$

where $e$ is the all-ones vector in $\mathbf{R}^n$. Suppose that there is a feasible solution $x$ to this LP, and let the index $i$ be such that $x_i = \max_j \{x_j\}$. We have that $x_i + 1 \leq \sum_{k=1}^n P_{i,k} x_k \leq \sum_{k=1}^n P_{i,k} x_i = x_i \sum_{k=1}^n P_{i,k} = x_i$, which is a contradiction. Therefore, this LP is infeasible. This means that its dual is either infeasible or unbounded. The dual problem is given by $\max_{y \in \mathbf{R}^n} \left\{ e^\top y : \ y^\top P = y^\top, \ y \geq 0 \right\}$. Clearly, the solution $y = 0$ is feasible for the dual; thus the dual must be unbounded. Therefore, there is an $a \neq 0$, such that $a^\top P = a^\top$ and $a \geq 0$.

**Exercise I.43.** Let $A \in \mathbf{R}^{n \times n}$ be a symmetric matrix. Consider the linear programming problem

$$\min_x \left\{ c^\top x : \ Ax \geq c, \ x \geq 0 \right\}.$$

Prove that if $\bar{x}$ satisfies $A\bar{x} = c$ and $\bar{x} \geq 0$, then $\bar{x}$ is optimal.

*Solution:* Note that the dual of this optimization problem is given by

$$\max_{\lambda, \mu} \left\{ c^\top \lambda : \ A\lambda + \mu = c, \ \lambda \geq 0, \ \mu \geq 0 \right\},$$

where we used $A^\top = A$. The solution $\bar{x}$ along with $\bar{\mu} = 0$ such that $A\bar{x} = c$ and $\bar{x} \geq 0$ is thus feasible for the dual problem, and $\bar{x}$ is feasible for the primal one, with the same objective value. Therefore, by the "zero duality gap" LP optimality condition, Theorem I.4.10, we deduce that $\bar{x}$ is optimal for the primal problem.

**Exercise I.44.** Let $w \in \mathbf{R}^n$, and let $A \in \mathbf{R}^{n \times n}$ be a *skew-symmetric* matrix, i.e., $A^\top = -A$. Consider the following linear program

$$\mathrm{Opt}(P) = \min_{x \in \mathbf{R}^n} \left\{ w^\top x : \ Ax \geq -w, \ x \geq 0 \right\}.$$

Suppose that the problem is solvable. Provide a closed analytical form expression for $\mathrm{Opt}(P)$.

*Solution:* The dual problem is given by

$$\begin{aligned}
\mathrm{Opt}(D) &= \max_{u,v} \left\{ -w^\top u : \ A^\top u + v = w, \ u \geq 0, \ v \geq 0 \right\} \\
&= \max_u \left\{ -w^\top u : \ A^\top u \leq w, \ u \geq 0 \right\} \\
&= \max_u \left\{ -w^\top u : \ -Au \leq w, \ u \geq 0 \right\} \qquad [\text{since } A^\top = -A] \\
&= \max_u \left\{ -w^\top u : \ Au \geq -w, \ u \geq 0 \right\} \\
&= -\min_u \left\{ w^\top u : Au \geq -w, \ u \geq 0 \right\} = -\mathrm{Opt}(P).
\end{aligned}$$

We are given that the primal problem is solvable, thus $\mathrm{Opt}(P) = \mathrm{Opt}(D)$ by LP Duality Theorem, and at the same time, as we just have seen, $\mathrm{Opt}(D) = -\mathrm{Opt}(P)$, implying that $\mathrm{Opt}(P) = 0$.

**Exercise I.45.** [Separation Theorem, polyhedral version] Let $P$ and $Q$ be two nonempty polyhedral sets in $\mathbf{R}^n$ such that $P \cap Q = \varnothing$. Suppose that the polyhedral descriptions of these sets are given as

$$P := \{x \in \mathbf{R}^n : \ Ax \leq b\} \quad \text{and} \quad Q := \{x \in \mathbf{R}^n : \ Dx \geq d\}.$$

Using LP duality show that there exists a vector $c \in \mathbf{R}^n$ such that

$$c^\top x < c^\top y \quad \text{for all } x \in P \text{ and } y \in Q.$$

*Solution:* Consider the following linear program

$$\max_x \left\{0^\top x: \ Ax \leq b, \ Dx \geq d\right\},$$

together with its dual given by

$$\min_{p,q} \left\{b^\top p + d^\top q: \ A^\top p + D^\top q = 0, \ p \geq 0, \ q \leq 0\right\}.$$

Since $P \cap Q = \varnothing$, the primal problem is infeasible, therefore the dual problem can be either infeasible or unbounded. But $p = 0, q = 0$ is a feasible solution to the dual problem therefore, we conclude that the dual problem is unbounded, i.e., there exists, $(z_p, z_q)$ such that $A^\top z_p + D^\top z_q = 0$, $z_p \geq 0$, $z_q \leq 0$ and $b^\top z_p + d^\top z_q < 0$. Let $c := A^\top z_p$. Then, for any $x \in P$ and any $y \in Q$, we have

$$\begin{aligned} c^\top x = z_p^\top Ax \leq z_p^\top b & \qquad [\text{as } z_p \geq 0 \text{ and } Ax \leq b] \\ < -d^\top z_q & \qquad [\text{as } b^\top z_p + d^\top z_q < 0] \\ \leq z_q^\top(-Dy) & \qquad [\text{as } z_q \leq 0 \text{ and } Dy \geq d] \\ \leq c^\top y, & \qquad [\text{as } c = A^\top z_p = -D^\top z_q] \end{aligned}$$

and thus we have proved the result.

**Exercise I.46.** Suppose we are given the following linear program

$$\mathrm{Opt}(P) = \min_x \left\{c^\top x: \ Ax = b, \ x \geq 0\right\} \qquad (P)$$

and its associated *Lagrangian* function given by

$$L(x, \lambda) := c^\top x + \lambda^\top(b - Ax).$$

The LP dual to $(P)$ is (by first replacing $Ax = b$ with $Ax \geq b$, $-Ax \geq -b$)

$$\mathrm{Opt}(D) = \max_{\lambda_\pm, \mu} \left\{b^\top(\lambda_+ - \lambda_-): \ A^\top(\lambda_+ - \lambda_-) + \mu = c, \ \lambda_\pm \geq 0, \ \mu \geq 0\right\},$$

or, after eliminating $\mu$ and setting $\lambda = \lambda_+ - \lambda_-$,

$$\mathrm{Opt}(D) = \max_\lambda \left\{b^\top \lambda: \ A^\top \lambda \leq c\right\}. \qquad (D)$$

Now, let us consider the following "game": Player 1 chooses some $x \geq 0$, and player 2 chooses some $\lambda$ simultaneously; then, player 1 pays to player 2 the amount $L(x, \lambda)$. In this game, player 1 would like to minimize $L(x, \lambda)$ and player 2 would like to maximize $L(x, \lambda)$.

A pair $(x^*, \lambda^*)$ with $x^* \geq 0$, is called an *equilibrium* point (or *saddle point* or *Nash equilibrium* ) if

$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*), \quad \forall x \geq 0 \text{ and } \forall \lambda. \qquad (*)$$

(That is, we have an equilibrium if no player is able to improve her performance by unilaterally modifying her choice.)

Show that a pair $(x^*, \lambda^*)$ is an equilibrium point if and only if $x^*$ and $\lambda^*$ are optimal solutions to the problem $(P)$ and its dual $(D)$ respectively.

*Solution:* First, suppose that $x^*$ and $\lambda^*$ are optimal solutions of $(P)$ and $(D)$, respectively. We will show that they are in equilibrium. Since $x^*$ is primal feasible, we have $Ax^* = b$, and so $L(x^*, \lambda) = c^\top x^* = L(x^*, \lambda^*)$ which proves the left inequality in $(*)$. Moreover, as $\lambda^*$ is dual feasible, we have $c - A^\top \lambda^* \geq 0$. Hence, for every $x \geq 0$, we obtain

$$L(x, \lambda^*) = (c - A^\top \lambda^*)^\top x + b^\top \lambda^* \geq b^\top \lambda^* = c^\top x^* = L(x^*, \lambda^*),$$

where the second equality follows from the LP Duality Theorem which gives us $\mathrm{Opt}(P) = \mathrm{Opt}(D)$, i.e., $c^\top x^* = b^\top \lambda^*$. Justification of $(*)$ is complete.

We now prove the reverse. Suppose that $x^* \geq 0$ and $\lambda^*$ are in equilibrium. The inequality $L(x^*, \lambda) \leq L(x^*, \lambda^*)$ yields $\lambda^\top(b - Ax^*) \leq (\lambda^*)^\top(b - Ax^*)$ for all $\lambda$. This can happen only if $Ax^* = b$, which establishes the primal feasibility of $x^*$. Furthermore, the inequality $L(x^*, \lambda^*) \leq L(x, \lambda^*)$ leads to $c^\top x^* \leq (c - A^\top \lambda^*)^\top x + b^\top \lambda^*$. Since this must be true for all $x \geq 0$, we get $c^\top x^* \leq b^\top \lambda^*$ (set $x = 0$) and $c - A^\top \lambda^* \geq 0$ and therefore $\lambda^*$ is dual feasible. By weak LP duality, we conclude that $c^\top x^* = b^\top \lambda^*$ and it follows that $x^*$ and $\lambda^*$ are optimal solutions of the primal and the dual problems, respectively.

**Exercise I.47.** Given a polyhedral set $X = \left\{x \in \mathbf{R}^n : \, a_i^\top x \leq b_i, \, \forall i = 1, \dots, m\right\}$, consider the associated optimization problem

$$\operatorname{Opt}(X) = \max_{x,t} \left\{t : \, B_\infty(x,t) \subseteq X\right\},$$

where $B_\infty(x,t) := \{y \in \mathbf{R}^n : \, \|y - x\|_\infty \leq t\}$. Is it possible to pose this optimization problem as a linear program with polynomial in $m, n$ number of variables and constraints? If it is possible, give such a representation explicitly. If not, argue why.

*Solution:* Note that in order for $(x,t)$ to be feasible to the given optimization problem, for every $i = 1, \dots, m$, we must have

$$b_i \geq \max_{y \in B_\infty(x,t)} \{a_i^\top y\} = a_i^\top x + t\|a_i\|_1,$$

where the last equality is evident. Hence, we arrive at

$$\operatorname{Opt}(X) = \max_{x,t} \left\{t : \, a_i^\top x + t\|a_i\|_1 \leq b_i, \, i = 1, \dots, m\right\},$$

which clearly is a formulation with polynomially many variables and inequalities.

**Exercise I.48.** Consider the following optimization problem

$$\min_{x \in \mathbf{R}^n} \left\{c^\top x : \, \tilde{a}_i^\top x \leq b_i \text{ for some } \tilde{a}_i \in A_i, \, i = 1, \dots, m, \, x \geq 0\right\}, \tag{*}$$

where $A_i = \{\bar{a}_i + \epsilon_i : \, \|\epsilon_i\|_\infty \leq \rho\}$ for $i = 1, \dots, m$. In this problem, we basically mean that the constraint coefficient $\tilde{a}_{ij}$ ($j$-th component of the $i$-th constraint vector $\tilde{a}_i$) belongs to the interval uncertainty set $[\bar{a}_{ij} - \rho, \, \bar{a}_{ij} + \rho]$, where $\bar{a}_{ij}$ is its nominal value. That is, in $(*)$, we are seeking a solution $x$ such that each constraint is satisfied for *some* coefficient vector from the corresponding uncertainty set.

Note that in its current form $(*)$, this problem is not a linear program (LP). Prove that it can be written as an *explicit* linear program and give the corresponding LP formulation.

*Solution:* This problem is equivalent to

$$\begin{aligned} \min \;\; & c^\top x \\ \text{s.t.} \;\; & \min_{\tilde{a}_i \in A_i} \{\tilde{a}_i^\top x\} \leq b_i \quad i = 1, \dots, m \\ & x \geq 0. \end{aligned}$$

Note that when $x \geq 0$

$$\begin{aligned} \min\{\tilde{a}_i^\top x : \, \tilde{a}_i = \bar{a}_i + \epsilon_i, \, \|\epsilon_i\|_\infty \leq \rho\} &= \bar{a}_i^\top x + \min\{\epsilon_i^\top x : \, \|\epsilon_i\|_\infty \leq \rho\} \\ &= \bar{a}_i^\top x - \rho \sum_{j=1}^n x_j, \end{aligned}$$

where the last equality is evident. Then, the resulting LP formulation for problem in $(*)$ is given by

$$\begin{aligned} \min \;\; & c^\top x \\ \text{s.t.} \;\; & \bar{a}_i^\top x - \rho \sum_{j=1}^n x_j \leq b_i \quad i = 1, \dots, m \\ & x \geq 0. \end{aligned}$$

**Exercise I.49.** Let $S = \{a_1, a_2, \dots, a_n\}$ be a finite set composed of $n$ distinct from each other elements, and let $f$ be a real-valued function defined on the set of all subsets of $S$. We say that $f$ is *submodular* if for every $X, Y \subseteq S$, the following inequality holds

$$f(X) + f(Y) \geq f(X \cup Y) + f(X \cap Y).$$

1. Give an example of a submodular function $f$.

   *Solution:* A simple trivial example is the function $f(X) = 0, \forall X \subseteq S$. Here is another simple slightly less trivial example: given $a \in \mathbf{Z}^n_+$, consider the function $f(X) = \max\{a_i : i \in X\}$ for every $X \subseteq S$. There are quite a lot of other examples of submodular functions; interested readers can refer to books on submodularity and combinatorial optimization.

2. Let $f : 2^S \to \mathbf{Z}$ be an integer-valued submodular function such that $f(\varnothing) = 0$. Consider the polyhedron

   $$P_f := \left\{ x \in \mathbf{R}^{|S|} : \sum_{t \in T} x_t \leq f(T), \ \forall T \subseteq S \right\},$$

   Consider

   $$\bar{x}_{a_k} := f(\{a_1, \dots, a_k\}) - f(\{a_1, \dots, a_{k-1}\}), \quad k = 1, \dots, n.$$

   Show that $\bar{x}$ is feasible to $P_f$.

   *Solution:* To justify this claim, we need to show that $\sum_{t\in T} \bar{x}_t \leq f(T)$ for all subsets $T$ of $S$. If $T = \varnothing$, then by definition $\sum_{t\in T} \bar{x}_t = 0 = f(\varnothing)$. When $T \neq \varnothing$, we will show this by induction on $\mathrm{Card}(T)$. To see the base case, suppose $T$ is a singleton, i.e., $T = \{a_k\}$ for some $k = 1, \dots, n$ . Then, since $f$ is submodular, we always have

   $$f(\{a_1, \dots, a_k\}) - f(\{a_1, \dots, a_{k-1}\}) \leq f(\{a_k\}), \quad \forall k = 1, \dots, n.$$

   Thus, by its definition $\bar{x}_{a_k} \leq f(\{a_k\})$ holds for all $k = 1, \dots, n$. Now, for the inductive hypothesis suppose that $\sum_{t\in T'} \bar{x}_t \leq f(T')$ for all subsets $T'$ of $S$ such that $\mathrm{Card}(T') < k$ for some $k \geq 2$, and let us show that the inequality holds for all subsets $T$ of $S$ of cardinality $k$ as well to complete the induction. So, consider any $T = \{a_{\iota_1}, \dots, a_{\iota_k}\}$, and define $T' := T \setminus \{a_{\iota_k}\}$. Thus, $\mathrm{Card}(T') = k - 1$, and by induction hypothesis we have $\sum_{t\in T'} \bar{x}_t \leq f(T')$ and so

   $$\sum_{t\in T} \bar{x}_t = \bar{x}_{a_{\iota_k}} + \sum_{t\in T'} \bar{x}_t \leq \big(f(\{a_1, \dots, a_{\iota_k}\}) - f(\{a_1, \dots, a_{\iota_k - 1}\})\big) + f(T')$$

   Now, by defining the sets $X := T$ and $Y := \{a_1, \dots, a_{\iota_k - 1}\}$, we see that $X \cup Y = \{a_1, \dots, a_{\iota_k}\}$ and $X \cap Y = \{a_{\iota_1}, \dots, a_{\iota_{k-1}}\} = T'$. Now, combining the previous inequality with the submodularity of $f$ applied to the sets $X$ and $Y$, we obtain

   $$\begin{aligned}\sum_{t\in T} \bar{x}_t &\leq \big(f(\{a_1, \dots, a_{\iota_k}\}) - f(\{a_1, \dots, a_{\iota_k - 1}\})\big) + f(T')\\ &= f(X \cup Y) - f(Y) + f(X \cap Y)\\ &\leq f(X) = f(T),\end{aligned}$$

   as desired. This completes the induction and so $\bar{x}$ is in $P_f$.

3. Consider the following optimization problem associated with $P_f$

   $$\max_x \left\{ c^\top x : \ x \in P_f \right\}.$$

   Write down the dual of this LP.

   *Solution:* The dual of this LP is

   $$\min_y \left\{ \sum_{T: T\subseteq S} f(T) y_T : \ \sum_{T \ni t} y_T = c_t, \forall t \in S, \ y_T \geq 0, \forall T \subseteq S \right\}.$$

4. Assume without loss of generality that $c_{a_1} \geq c_{a_2} \geq \ldots \geq c_{a_n}$. Identify a dual feasible solution and using LP Duality Theorem show that the solution $\bar{x}$ specified in part 2 is optimal to the primal maximization problem associated with $P_f$.

   *Solution:* The following is a feasible solution for the dual problem:

$$\bar{y}_T := \begin{cases} c_{a_k} - c_{a_{k+1}}, & \text{if } T = \{a_1, \ldots, a_k\} \text{ for some } k = 1, \ldots, n \\ 0, & \text{otherwise,} \end{cases}$$

   where we define $c_{a_{n+1}} = 0$. Indeed, as $c_{a_1} \geq c_{a_2} \geq \ldots \geq c_{a_n}$, we immediately see that $\bar{y}_T \geq 0$ for all $T \subseteq S$. Now, consider any $t \in S$, and suppose $i$ is such that $a_i = t$. Note that the only dual variables $\bar{y}_T$ that may take positive values are the ones corresponding to the sets $T = \{a_1, \ldots, a_k\}$ for some $k = 1, \ldots, n$. And among such sets the only ones that contain the given $t = a_i$ are the sets $\bar{T}_i := \{a_1, \ldots, a_i\}$, $\bar{T}_{i+1} := \{a_1, \ldots, a_i, a_{i+1}\}, \ldots, \bar{T}_n := \{a_1, \ldots, a_n\}$. Thus, for any $t \in S$, we have

$$\sum_{T \ni t} \bar{y}_T = \sum_{\ell=i}^{n} \bar{y}_{\bar{T}_\ell} = \sum_{\ell=i}^{n} \left(c_{a_\ell} - c_{a_{\ell+1}}\right) = c_{a_i} - c_{a_{n+1}} = c_{a_i} = c_t,$$

   where we used the fact that $c_{a_{n+1}} = 0$.
   Both of these solutions ($\bar{x}$ and $\bar{y}$) give the same objective value for their corresponding problems as

$$\begin{aligned} c^\top \bar{x} = \sum_{k=1}^{n} c_{a_k} \bar{x}_{a_k} &= \sum_{k=1}^{n} c_{a_k} \left(f(\{a_1, \ldots, a_k\}) - f(\{a_1, \ldots, a_{k-1}\})\right) \\ &= \sum_{k=1}^{n} f(\{a_1, \ldots, a_k\}) \left(c_{a_k} - c_{a_{k+1}}\right) \\ &= \sum_{T : T \subseteq S} f(T) \bar{y}_T. \end{aligned}$$

   Therefore, both solutions are optimal.

**Remark:** Note that when the submodular function $f$ is integer-valued, we immediately see from the characterization of the optimal primal solution $\bar{x}$ that for all integer vectors $c \in \mathbf{Z}^n$ such that there exists an optimum solution to the primal problem, there exists an optimum solution (e.g. $\bar{x}$) where all variables take integer values. A system of linear inequalities $Ax \leq b$ with $b \in \mathbf{Z}^m$ and $A \in \mathbf{Q}^{m \times n}$ satisfying such a property (i.e., whenever $c \in \mathbf{Z}^n$ is such that there is an optimal solution to $\max_x \{c^\top x : Ax \leq b\}$ then there is an integer optimum solution) is called *totally dual integral* (TDI). Thus, we conclude that the polyhedron $P_f$ associated with an integer-valued submodular function $f$ is TDI. TDI property is a well-known sufficient condition that guarantees that every extreme point (see section 8.2) of the associated polyhedron is integral. In particular, TDI property generalizes *total unimodularity* (TU), i.e., the other well-known sufficient condition for integrality of a polyhedron which plays a key role in network-flow based optimization.

# Exercises from Part II

## Separation

**Exercise II.1.** Mark by "Y"/"N" those of the below listed cases where the linear form $f^\top x$ separates/does not separate the sets $S$ and $T$:

- $S = \{0\} \subset \mathbf{R}$, $T = \{0\} \subset \mathbf{R}$, $f^\top x = x$

  *Solution:* N
- $S = \{0\} \subset \mathbf{R}$, $T = [0,1] \subset \mathbf{R}$, $f^\top x = x$

  *Solution:* Y
- $S = \{0\} \subset \mathbf{R}$, $T = [-1,1] \subset \mathbf{R}$, $f^\top x = x$

  *Solution:* N
- $S = \{x \in \mathbf{R}^3 : x_1 = x_2 = x_3\}$, $T = \{x \in \mathbf{R}^3 : x_3 \geq \sqrt{x_1^2 + x_2^2}\}$, $f^\top x = x_1 - x_2$

  *Solution:* N
- $S = \{x \in \mathbf{R}^3 : x_1 = x_2 = x_3\}$, $T = \{x \in \mathbf{R}^3 : x_3 \geq \sqrt{x_1^2 + x_2^2}\}$, $f^\top x = x_3 - x_2$
  *Solution:* Y
- $S = \{x \in \mathbf{R}^3 : -1 \leq x_1 \leq 1\}$, $T = \{x \in \mathbf{R}^3 : x_1^2 \geq 4\}$, $f^\top x = x_1$
  *Solution:* N
- $S = \{x \in \mathbf{R}^2 : x_2 \geq x_1^2, x_1 \geq 0\}$, $T = \{x \in \mathbf{R}^2 : x_2 = 0\}$, $f^\top x = -x_2$
  *Solution:* Y

**Exercise II.2.** Consider the set

$$M = \left\{ x \in \mathbf{R}^{2004} : \begin{array}{rcl} x_1 + x_2 + \ldots + x_{2004} & \geq & 1 \\ x_1 + 2x_2 + 3x_3 \ldots + 2004x_{2004} & \geq & 10 \\ x_1 + 2^2x_2 + 3^2x_3 \ldots + 2004^2x_{2004} & \geq & 10^2 \\ \ldots\ldots\ldots\ldots & & \\ x_1 + 2^{2002}x_2 + 3^{2002}x_3 + \ldots + 2004^{2002}x_{2004} & \geq & 10^{2002} \end{array} \right\}$$

Is it possible to separate this set from the set $\{x_1 = x_2 = \ldots = x_{2004} \leq 0\}$? If yes, what could be a separating plane?

*Solution:* Separation is possible, and a separating plane is, e.g., $\{x : x_1 + \ldots + x_{2004} = 1/2\}$, since the linear form $\sum_i x_i$ is $\geq 1$ on $M$ and clearly is $\leq 0$ on the set $\{x_1 = ... = x_{2004} \leq 0\}$.

**Exercise II.3.** Can the sets $S = \{x \in \mathbf{R}^2 : x_1 > 0, x_2 \geq 1/x_1\}$ and $T = \{x \in \mathbf{R}^2 : x_1 < 0, x_2 \geq -1/x_1\}$ be separated? Can they be strongly separated?

*Solution:* The sets are separated by the line $\{x \in \mathbf{R}^2 : x_1 = 0\}$ They cannot be strongly separated since the distance between the sets is zero (take large $t > 0$ and look at the points $[1/t; t] \in S$ and $[-1/t; t] \in T$).

**Exercise II.4.** Let $M \subset \mathbf{R}^n$ be a nonempty closed convex set. The *metric projection* $\mathrm{Proj}_M(x)$ of a point $x \in \mathbf{R}^n$ onto $M$ is the $\|\cdot\|_2$-closest to $x$ point of $M$, so that

$$\mathrm{Proj}_M(x) \in M \;\&\; \|x - \mathrm{Proj}_M(x)\|_2^2 = \min_{y \in M} \|x - y\|_2^2. \qquad (*)$$

1. Prove that for every $x \in \mathbf{R}^n$ the minimum in the right hand side of $(*)$ is achieved, and $x_+$ is a minimizer if and only if

$$x_+ \in M \ \&\ \forall y \in M : [x - x_+]^\top [x_+ - y] \geq 0. \tag{11.1}$$

Derive from the latter fact that the minimum in $(*)$ is achieved at a unique point, the bottom line being that $\mathrm{Proj}_M(\cdot)$ is well defined

2. Prove that when passing from a point $x \in \mathbf{R}^n$ to its metric projection $x_+ = \mathrm{Proj}_M(x)$, the distance to any point of $M$ does not increase, specifically,

$$\begin{array}{rcl} \forall y & \in & M : \|x_+ - y\|_2^2 \leq \|x - y\|_2^2 - \mathrm{dist}^2(x, M), \\ \mathrm{dist}(x, M) & := & \min_{u \in M} \|x - u\|_2 = \|x - x_+\|_2. \end{array} \tag{11.2}$$

3. Let $x \notin M$, so that, denoting $x_+ = \mathrm{Proj}_M(x)$, the vector $e = \frac{x - x_+}{\|x - x_+\|_2}$ is well defined. Prove that the linear form $e^\top z$ strongly separates $\{x\}$ and $M$, specifically,

$$\forall y \in M : e^\top y \leq e^\top x - \mathrm{dist}(x, M).$$

*Note:* The fact just outlined underlies an alternative proof of Separation Theorem, where the first step is to prove that a point outside a nonempty closed convex set can be strongly separated from the set. In our proof, the first step was similar, but with $M$ restricted to be polyhedral, rather than merely convex and closed.

4. Prove that the mapping $x \mapsto \mathrm{Proj}_M(x) : \mathbf{R}^n \to M$ is *contraction* in $\|\cdot\|_2$:

$$\forall u, u' \in \mathbf{R}^n : \|\mathrm{Proj}_M(u) - \mathrm{Proj}_M(u')\|_2 \leq \|u - u'\|_2.$$

5. Let $M$ be the probabilistic simplex: $M = \{x \in \mathbf{R}^n : x \geq 0, \sum_i x_i = 1\}$, Justify the following recipe for computing $\mathrm{Proj}_M(x)$:

> Let $\psi(t) = \sum_{i=1}^m [x_i - t]_+$. where $[s]_+ = \max[s, 0]$. $\psi$ is piecewise linear, with breakpoints $x_1, x_2, \ldots, x_n$, continuous function of $t \in \mathbf{R}$. $\psi(t) \to +\infty$ as $t \to -\infty$, and $\psi(t) \to 0$ as $t \to +\infty$. Consequently, there exists (and can be easily computed due to piecewise linearity of $\psi$) $t \in \mathbf{R}$ such that $\sum_i [x_i - t]_+ = 1$. The metric projection of $x$ onto $M$ is nothing but the vector $x_+$ with coordinates $[x_i - t]_+$, $1 \leq i \leq n$.

What is metric projection of the point $x = [1; 2; 2.5]$ on the 3-dimensional probabilistic simplex?

*Solution:*

1: Let $d = \inf_{y \in M} \|x - y\|_2$, so that there exists a sequence $\{y_i \in M\}_{i \geq 1}$ such that $\lim_{i \to \infty} \|x - y_i\|_2 = d$. Since $d < \infty$, the sequence $\{y_i\}$ is bounded, so that we can extract from it a converging subsequence $\{y_{i_s}, i_s < i_{s+1}\}_{s \geq 1}$. Since $y_{i_s} \in M$, the limit $\bar{y}$ of the subsequence belongs to $M$, and since $\|y_{i_s} - x\|_2 \to d$ as $s \to \infty$ and $\|\cdot\|_2$ is continuous, we conclude that $\|\bar{y} - x\|_2 = d$. Thus, the minimum in $(*)$ is achieved. Now let us prove that the closest to $x$ points of $M$ are exaclty the points satisfying (11.1), Note that when $x_+ \in M$ and $y \in M$, we have $x_+ + t(y - x) \in M$ when $0 \leq t \leq 1$ due to convexity of $M$. It follows that if $x_+$ is a minimizer of $\|z - y\|_2$ over $y \in M$, then the function $\phi(t) = \|x - [x_+ + t(y - x_+)]\|_2^2$ attains its minimum on the segment $0 \leq t \leq 1$ at $t = 0$. We have

$$\phi(t) = \|x - x_+\|_2^2 - 2t[x - x_+]^\top [y - x_+] + t^2 \|y - x_+\|_2^2;$$

since this smooth function achieves its minimum on $[0, 1]$ at the point $t = 0$, we have $\phi'(0) \geq 0$, which is the inequality in (11.1). As a byproduct, we see that $\|y - x\|_2^2 = \phi(1) \geq \phi(0) + \|y - x_+\|_2^2 = \|x - x_+\|_2^2 + \|y - x_+\|_2^2$, implying that if $y \in M$ and $y \neq x_+$, then $\|x - y\|_2 > \|x - x_+\|_2$, that is, $x_+$ is the unique minimizer of $\|x - y\|_2^2$ over $y \in M$. It remains to prove that if $x_+$ satisfies (11.1), then $x_+$ minimizes $\|y - x\|_2^2$ over $y \in M$. Indeed, assuming that $x_+$ satisfies (11.1) and given $y \in M$, the associated with this $y$ function $\phi(t)$ is quadratic in $t$ and satisfies $\phi(0) \geq 0$, $\phi'(0) \geq 0$, $\phi'' \geq 0$, implying that $\phi(0) \leq \phi(t)$ whenever $t \geq 0$; in particular, $\|x_+ - x\|_2^2 = \phi(0) \leq \phi(1) = \|y - x\|_2^2$. Thus, $\|x_+ - x\|_2^2 \leq \|y - x\|_2^2$ for all $y \in M$ and, in addition, $x_+ \in M$, implying that $x_+$ minimizes $\|y - x\|_2^2$ over $y \in M$. □

2: Let $x \in \mathbf{R}^n$, $x_+ = \mathrm{Proj}_M(x)$, and $y \in M$. We have

$$\begin{array}{rcl}\|x-y\|_2^2 & = & \|[[x-x_+]+[x_+-y]\|_2^2 = \|x-x_+\|_2^2 + \|x_+-y\|_2^2 + 2[x-x_+]^\top[x_+-y] \\ & \geq & \|x-x_+\|_2^2 + \|x_+-y\|_2^2,\end{array}$$

where the concluding $\geq$ is due to (11.1).

3: Assuming $x \notin M$, for $y \in M$ we have

$$[x-x_+]^\top[x-y] = [x-x_+]^\top[x-x_+] + [x-x_+]^\top[x_+-y] \geq \|x-x_+\|_2^2$$

with the inequality given by (11.1). Thus, $[x-x_+]^\top y \leq [x-x_+]^\top x - \|x-x_+\|_2^2$. Recalling what $e$ is, we get $e^\top x \geq e^\top y + \|x-x_+\|_2 = e^\top y + \mathrm{dist}(x,M)\,\forall y \in M$. □

4: Let $u_+ = \mathrm{Proj}_M(u)$, $u'_+ = \mathrm{Proj}_M(u')$. Let us set $e = u - u_+$, $f = u'_+ - u'$, so that $[u_+ - u'_+] + [e+f] = u - u'$ and $e^\top[u_+ - u'_+] \geq 0$, $f^\top[u_+ - u'_+] \geq 0$ by (11.1) as applied with $x = u$ and with $x = u'$. We conclude that $\|u-u'\|_2^2 = \|[u_+ - u'_+] + [e+f]\|_2^2 = \|u_+ - u'_+\|_2^2 + 2[e+f]^\top[u_+ - u'_+] + \|e+f\|_2^2 \geq \|u_+ - u'_+\|_2^2$. □

5: Invoking item 1, all we need is to verify that with $x_+$ given by the construction in question, (11.1) holds true. Indeed, the inclusion $x_+ \in M$ is evident. Besides this,

$$\begin{array}{rl}\forall y \in M: & [x-x_+]^\top[x_+-y] = -\sum_{i:x_i\leq t} x_i y_i + \sum_{i:x_i>t} t([x_i-t]-y_i) \\ & = -\sum_i \min[x_i,t]y_i + t\sum_{i:x_i>t}[x_i-t] = t - \sum_i \min[x_i,t]y_i \geq t - t\sum_i y_i = 0,\end{array}$$

where $\geq$ is due to nonnegativity of $y \in M$.

The metric projection of $[1;2;2.5]$ on 3-dimensional probabilistic simplex is the vector
$[0;0.25;0.75] = [[1-1.75]_+;[2-1.75]_+;[2.5-1.75]_+]$. □

**Exercise II.5.** [Follow-up to Exercise II.4] Let $p(z) = z^n + p_{n-1}z^{n-1} + ... + p_1 z + p_0$, $n \geq 1$ be a polynomial of complex variable $z$. By the Fundamental Theorem of Algebra, $p$ has $n$ roots $\lambda_1, ..., \lambda_n$. Treating complex numbers as 2D real vectors, prove that all roots of the derivative $p'(z) = nz^{n-1} + (n-1)p_{n-1}z^{n-2} + .. + p_1$ belong to the convex hull of $\lambda_1, ..., \lambda_n$.

*Solution:* Let $C = \mathrm{Conv}\{\lambda_1, ..., \lambda_n\}$, and let $\lambda$ be a root of $p'$. Assuming that $\lambda \notin C$, let us lead this assumption to contradiction. Indeed, let $\overline{\lambda} = \mathrm{Proj}_C(\lambda)$ and $e = \lambda - \overline{\lambda}$, so that $e^\top[\lambda - \lambda_i] \geq e^\top e > 0$ by Exercise II.4,3. We have $p(z) = \prod_i (z - \lambda_i)$, whence, setting $f(z) = |p(z)|^2 = \prod_i \|z - \lambda_i\|_2^2 : \mathbf{R}^2 \to \mathbf{R}^2$, one has

$$\tfrac{d}{dt}\big|_{t=0} f(\lambda + te) = 2\Big[e^T[\lambda-\lambda_1]\|\lambda-\lambda_2\|^2...\|\lambda-\lambda_n\|^2 + \|\lambda-\lambda_1\|_2 e^T[\lambda-\lambda_1]\|\lambda-\lambda_3\|^2...\|\lambda-\lambda_n\|^2$$
$$+... + \|\lambda-\lambda_1^2\|...\|\lambda-\lambda_{n-1}\|_2^2 e^\top[\lambda-\lambda_n]\Big] > 0.$$

On the other hand, we have

$$0 = p'(\lambda) = \lim_{\delta\to 0} \frac{p(\lambda+\delta) - p(\lambda)}{\delta},$$

(why?). Denoting by $\imath$ the imaginary unit and setting $\lambda = a + \imath b$, $p(x + \imath y) = u(x,y) + \imath v(x,y)$ with real $a, b, x, y, u, v$, and looking what happens when $\delta \to 0$ stays (a) real, (b) purely imaginary, we get

$$\frac{\partial}{\partial x}u(a,b) = 0, \frac{\partial}{\partial x}v(a,b) = 0 \text{ and } \frac{\partial}{\partial y}u(a,b) = 0, \frac{\partial}{\partial y}v(a,b) = 0,$$

whence

$$\nabla\big|_{x=a,y=b} f(x + \imath y) = \nabla\big|_{x=a,y=b}[u^2(x,y) + v^2(x,y)] = 0,$$

so that $\frac{d}{dt}\big|_{t=0} f(\lambda + te) = 0$, which is a desired contradiction. □

**Exercise II.6.** Derive the statement in Remark I.1.4 from the Separation Theorem.

*Solution:* We already know that the solution set of a whatever system of nonstrict linear inequalities is closed and convex, and all we need to prove is that a closed convex set $M \subset \mathbf{R}^n$ is a solution set of a sequence of nonstrict linear inequalities $a_i^\top x \leq b_i$, $i = 1, 2, \ldots$. There is nothing to prove when $M = \mathbf{R}^n$ (take empty system, or, if you want, single inequality $0^\top x \leq -1$). Similarly, there is nothing to prove when $M = \varnothing$ – take the system of inequalities $x_1 \leq -1, -x_1 \leq -1$. Now let $M$ be nonempty and smaller than $\mathbf{R}^n$. The complement $M^c$ of $C$ is a nonempty open set; note that the set of all rational vectors from $M^c$ can be arranged into sequence $c_1, c_2, \ldots$.

> Indeed, let us look at the set $T_N$ of all rational vectors from $M^c$ with the total of magnitudes of numerators and denominators in representations of their (rational!) coordinates as fractions does not exceed a given integer $N$; for every $N$, this set is finite. We now can list all vectors from $T_1$, then list all unlisted yet vectors from $T_2$, then – all unlisted yet vectors from $T_3$, and so on; as a result, all rational vectors from $M^c$ will be arranged into a sequence.

Now let $r(x) = \min_{y \in M} \|x - y\|_2$ be the distance from $x \in \mathbf{R}^n$ to $M$; since $M$ is closed and nonempty, the minimum is achieved. Again invoking closedness of $M$, $r(x) > 0$ whenever $x \notin M$. Besides this, the function $r(x)$ clearly satisfies the relation $|r(x) - r(x')| \leq \|x - x'\|_2$ and is therefore continuous. Note also that when $x \notin M$, the open ball $B(x)$ of radius $r(x)$ centered at $x$ does not intersect $M$. By Separation Theorem, the balls $B(c_i)$ can be separated from $M$: for properly selected $a_i$ we have $\sup_{x \in M} a_i^\top x \leq \inf_{y \in B(c_i)} a_i^\top y$. We lose nothing by scaling $a_i$ to become a unit vector, in which case the "separation inequality" becomes $\sup_{x \in M} a_i^\top x \leq b_i := a_i^\top c_i - r(c_i)$. We claim that $M$ is exactly the solution set of the resulting sequence of inequalities $a_i^\top x \leq b_i$, $i = 1, 2, \ldots$. Indeed, by construction, every point from $M$ solves this system. All we need to verify is that if $\bar{x}$ solves the system, then $\bar{x} \in M$. Assuming, on the contrary, that this is not the case, $\bar{x} \in M^c$ and therefore for some sequence $i_1 < i_2 < \ldots$ we have $c_{i_j} \to \bar{x}$ as $j \to \infty$, whence $a_{i_j}^\top (c_{i_j} - \bar{x}) \to 0$ as $j \to \infty$. Due to the origin of $\bar{x}$, $a_{i_j}^\top \bar{x} \leq b_{i_j} = a_{i_j}^\top c_{i_j} - r(c_{i_j})$, whence $a_{i_j}^\top (c_{i_j} - \bar{x}) \geq r(c_{i_j})$, which combines with $a_{i_j}^\top (c_{i_j} - \bar{x}) \to 0$ as $j \to \infty$ to imply that $r(c_{i_j}) \to 0$ as $j \to \infty$. On the other hand, $r(\cdot)$ is continuous and $c_{i_j} \to \bar{x}$, $j \to \infty$, implying that $r(c_{i_j}) \to r(\bar{x})$ as $j \to \infty$. The bottom line is that $r(\bar{x}) = 0$, which is not the case, since $\bar{x} \notin M$ and $M$ is closed. Thus, assuming that $M$ is not the solution set of the system $a_i^\top x \leq b_i$, $i = 1, 2, \ldots$, we arrive at contradiction. □

## Extreme points

**Exercise II.7.** Find extreme points of the following sets:

1. $X = \{x \in \mathbf{R}^3 : x_1 + x_2 \leq 1, x_2 + x_3 \leq 1, x_3 + x_1 \leq 1\}$
2. $X = \{x \in \mathbf{R}^4 : x_1 + x_2 \leq 1, x_2 + x_3 \leq 1, x_3 + x_4 \leq 1, x_4 + x_1 \leq 1\}$

*Solution:* 1: The set is polyhedral; by algebraic characterization of extreme point of polyhedral sets, among the inequalities specifying the set, an extreme point $w$, if any, should make equalities 3 inequalities with linearly independent vectors of coefficients, that is, $w$ should make equalities all 3 constraints specifying the set (their vectors of coefficients indeed are linearly independent). As a result, the only extreme point is $[0.5; 0.5; 0.5]$.

2: The same reasoning as in item 1 says that at an extreme point all constraints specifying the set should be satisfied as equalities, and the vectors of coefficients of these constraints should be linearly independent. The latter does *not* take place, so that there are no extreme points.

*Explanation:* when $n \geq 2$, the $n \times n$ matrix $A_n = \left[\begin{array}{c|c|c|c} 1 & 1 & & \\ \hline & 1 & 1 & \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline 1 & & & 1 \end{array}\right]$ is nondegenerate when $n$ is odd and is degenerate, with kernel spanned by the vector $[1; -1; 1; -1; ...; -1]$ when $n$ is even. As a result, the set $X_n = \{x \in \mathbf{R}^n : A_n x \leq [1; ...; 1]\}$ contains lines, and thus has no extreme point, when $n$ is even, and has exactly one extreme point $[0.5; ...; 0.5]$ when $n$ is odd. For odd $n$, $x \mapsto y = A_n x$ is a linear one-to-0one transformation of $\mathbf{R}^n$, and in $y$-variables $X_n$ becomes the set $\{y \in \mathbf{R}^n : y \leq [1; ...; 1]\}$. Thus, for odd $n$, $X_n$ is just a translation of a polyhedral cone – the image of $\mathbf{R}^n_+$ under one-to-one linear transformation.

**Exercise II.8.** Let $M \subset \mathbf{R}^n$ be a nonempty closed convex set not containing lines, and $f^\top x$ be a linear function of $x \in \mathbf{R}^n$ achieving its maximum over $X$. Prove that among maximizers of this function on $M$ there are extreme points of $M$.

*Solution:* Let $\overline{M} = \mathrm{Argmax}_{x\in M} f^\top x$ be the set of maximizers of $f^\top x$ over $x \in M$. By assumption, this set is nonempty; along with $M$, it is convex, closed, and does not contain lines. By item (i) of Krein-Milman Theorem $\overline{M}$ has an extreme point $x_*$; let us prove that $x_*$ is an extreme point of $M$. Indeed, assuming $x \pm h \in M$, we should have $f^\top[x_* \pm h] \leq f^\top x_*$ (since $x_*$ is a maximizer of $f^\top x$ over $x \in M$) which is possible only when $f^\top[x_* \pm h] = f^\top x_*$, Thus, $x_* \pm h \in \overline{M}$, implying that $h = 0$ (since $x_* \in \mathrm{Ext}(\overline{M})$). Thus, $x_*$ is a desired extreme point maximizer of $f^\top x$ over $x \in M$. □

**Exercise II.9.** Mark by $\mathbf{T}$ those of the below claims which always (i.e., for every data satisfying premise of the claim) are true:

1. If $\mathrm{Conv}(A) = \mathrm{Conv}(B)$, then $A = B$.

   *Solution:* evidently false – take $n = 1$, $A = \{0, 1, 2\}$, $B = \{0, 2\}$.

2. If $\mathrm{Conv}(A) = \mathrm{Conv}(B)$ is nonempty and $A, B, \mathrm{Conv}(A)$ are closed, then $A \cap B \neq \varnothing$.

   *Solution:* false – take $n = 1$, $A = \{2k + 1\}_{k=-\infty}^{\infty}$, $B = \{2k\}_{k=-\infty}^{\infty}$.

3. If $\mathrm{Conv}(A) = \mathrm{Conv}(B)$ is nonempty and bounded, then $A \cap B \neq \varnothing$.

   *Solution:* false; take $A = \{\frac{1}{2k}\}_{k=1}^{\infty} \cup \{1 - \frac{1}{2k}\}_{k=1}^{\infty}$, $B = \{\frac{1}{2k+1}\}_{k=1}^{\infty} \cup \{1 - \frac{1}{2k+1}\}_{k=1}^{\infty}$, so that $A \cap B = \varnothing$ and $\mathrm{Conv}(A) = \mathrm{Conv}(B) = (0, 1)$.

4. If $\mathrm{Conv}(A) = \mathrm{Conv}(B)$ is nonempty, closed and bounded, then $A \cap B \neq \varnothing$.

   *Solution:* true. When $\mathrm{Conv}(A)$ is nonempty, closed, and bounded, by Krein-Milman Theorem, $\mathrm{Conv}(A)$ possesses an extreme point $v$, and by Fact II.8.5 $v \in A$. By the same token, $\mathrm{Conv}(A) = \mathrm{Conv}(B)$ implies that $v \in B$, so that $A \cap B$ is nonempty and, moreover, contains all extreme points of $\mathrm{Conv}(A) = \mathrm{Conv}(B)$. Applying Krein-Milman Theorem once more, we conclude that $A \cap B$ is not just nonempty, it is rich enough to ensure that $\mathrm{Conv}(A \cap B) = \mathrm{Conv}(A) = \mathrm{Conv}(B)$.

**Exercise II.10.** As is immediately seen, the only extreme point of the nonnegative orthant $\mathbf{R}^n_+ = \mathbf{R}_+ \times \mathbf{R}_+ \times \ldots \times \mathbf{R}_+$ is the origin, that is, the vector from $\{0\} \times \{0\} \times \ldots \times \{0\}$; as we know, the extreme points of $n$-dimensional unit box $\{x \in \mathbf{R}^n : 0 \leq x_i \leq 1, i \leq n\} = [0, 1] \times [0, 1] \times \ldots \times [0, 1]$ are zero/one vectors, that is, vectors from $\{0, 1\} \times \{0, 1\} \times \ldots \times \{0, 1\}$. Prove the following generalization of these observations:

> Let $X_i \subset \mathbf{R}^{n_i}$, $1 \leq i \leq K$, be closed convex sets. The set of extreme points of the direct product $X = X_1 \times \ldots \times X_K$ of these sets is the direct product of the sets of extreme points of $X_i$.

*Solution:* The vectors from $X$ are the block vectors $x = [x_1; \ldots; x_K]$ with blocks $x_i \in X_i$. If such an $x$ is an extreme point of $X$, that is, $x \pm h \in X$ implies $h = 0$, then for every $i$ the relation $x_i \pm h_i \in X_i$ implies $h_i = 0$, since otherwise, setting $h = [0; ...; 0; h_i; 0; ...; 0]$ we would have $x \pm h \in X$ and $h \neq 0$, which is impossible; thus, $x \in \mathrm{Ext}(X_1) \times ... \times \mathrm{Ext}(X_K)$. Vice versa, if $x \in \mathrm{Ext}(X_1) \times ... \times \mathrm{Ext}(X_K)$ and $x \pm h \in X$, then $x_i \pm h_i \in X_i$ for all $i$, implying that $h_i = 0$ for all $i$, that is, $h = 0$; thus, $x \in \mathrm{Ext}(X)$. □

**Exercise II.11.** Looking at the sets of extreme points of closed convex sets like the unit Euclidean ball, a polytope, the paraboloid $\{[x; t] : t \geq x^\top x\}$, etc., we see that these sets are closed. Do you think this always is the case? Is it true that the set $\mathrm{Ext}(M)$ of extreme points of a closed convex set $M$ always is closed ?

*Solution:* The claim is not true. Indeed, consider the set $X$ in 3D which is the union of the segment $\{[x_1; 0; 0] : -1 \leq x_1 \leq 1\}$ and the arc $\{[0; x_2; x_2^2], 0 \leq x_2 \leq 1\}$; this set is closed and bounded, and therefore so is its convex hull $M := \mathrm{Conv}(X)$ (Corollary I.2.5). We claim that when $t \in (0, 1)$, the point $x_t = [0; t; t^2]$ is an extreme point of $M$. Taking this claim for granted, we conclude that the point $[0; 0; 0]$ is the limit, as $t \to +0$, of extreme points $x_t$ of $M$, but this limit clearly is not an extreme point – it is the midpoint of the segment with the endpoints $[\pm 1; 0; 0] \in M$.

It remains to prove that $x_t$ is an extreme point of $M$. Observe first that $x_t$ is an extreme point of the projection $M_-$ of $M$ onto the plane $L = \{x : x_1 = 0\} \ni x_t$. Indeed, $M_-$ clearly belongs to the convex hull $C$ of the projection of $X$ onto $L$, that is, to the convex hull of the arc $\{[0; s; s^2] : 0 \leq s \leq 1\}$. We clearly have $C = \{x : x_1 = 0, 0 \leq x_2 \leq 1, x_2^2 \leq x_3 \leq x_2\}$, and $x_t$ is an extreme point of $C$. Since this point belongs to $M_- \subset C$, it is extreme point of $M_-$ as well. Now, to prove that $x_t \in \mathrm{Ext}(M)$ is the same as to prove that $x_t \pm h \in M$ implies $h = 0$. Indeed, let $h$ be such that $x_t \pm h \in M$. Looking at the projections of $x_t \pm h$ onto $L$ and taking into account that $x_t$ is an extreme point of the projection of $M$ onto $L$, we see that $h_2 = h_3 = 0$. Thus, we are in the situation $[\pm h_1; t; t^2] \in M$, and should prove that $h_1 = 0$. Recalling what $M$ is, we conclude that $[h_1; t; t^2]$ is convex combination of several points of the type $[s; 0; 0]$, $s \in [-1, 1]$, and several points of the type $[0; r; r^2]$ with $0 \leq r \leq 1$. If the total weight of the points of the first type in this combination is positive, then, projecting the combination onto $L$, we conclude that $x_t$ is a convex combination of several points of the second type and the point $[0; 0; 0]$, the weight of the latter point being positive. This is impossible – all points participating in the latter convex combination belong to $C$, and, as we know, $x_t$ is an extreme point of this set, implying that all points participating, with positive weights, in representation of $x_t$ as a convex combination of points from $C$ should be equal to $x_t$, see Fact II.8.4. The bottom line is that $[h_1; t; t^2]$ can be represented as a convex combination of points of the second type only, that is, $h_1 = 0$, that is, $h = 0$, as claimed. □

**Exercise II.12.** Derive representation $(*)$ in Exercise I.29 from Example II.9.1 in section 9.3.

*Solution:* Given positive integers $k \leq n$ and $x \in \mathbf{R}^n$, consider the LP program

$$\mathrm{Opt} = \max_u \left\{ \sum_i x_i u_i : 0 \leq u_i \leq 1, i \leq n, \sum_i u_i = k \right\}$$

Example II.9.1 in section 9.3 says that extreme points of the bounded feasible set of the problem are 0/1 vectors with exactly $k$ entries equal to 1, implying that $\mathrm{Opt} = s_k(x)$. We now have

$$\begin{array}{rcl} s_k(x) & = & \max_u \sum_i x_i u_i : 0 \leq u_i \leq 1, i \leq n, \sum_i u_i = k\} \\ & = & \min_{z^\pm, s} \left\{ \sum_i z_i^+ + ks : [z_i^+ - z_i^-] + s = x_i, i \leq n, z^\pm \geq 0 \right\} \text{ [LP duality]} \\ & = & \min_{z^+, s} \left\{ \sum_i z_i^+ + ks : z+ \geq 0, x_i \leq z_i^+ + s, i \leq n \right\}, \end{array}$$

or, equivalently,

$$t \geq s_k(x) \Longleftrightarrow \exists (z, s) : x_i \leq z_i + s \,\forall i, z \geq 0, \sum_i z_i + ks \leq t,$$

which is equivalent form of the representation we are justifying. □

**Exercise II.13.** By Birkhoff Theorem, the extreme points of the polytope $\Pi_n = \{[x_{ij}] \in \mathbf{R}^{n \times n} : x_{ij} \geq 0, \sum_i x_{ij} = 1 \,\forall j, \sum_j x_{ij} = 1 \,\forall i\}$ are exactly the Boolean (i.e., with entries 0 and 1) matrices from this set. Prove that the same holds true for the "polytope of sub-doubly stochastic" matrices $\Pi_{m,n} = \{[x_{ij}] \in \mathbf{R}^{m \times n} : x_{ij} \geq 0, \sum_i x_{ij} \leq 1 \,\forall j, \sum_j x_{ij} \leq 1 \,\forall i\}$.

*Solution:* First, every Boolean matrix $[x_{ij}]$ from $\Pi_{m,n}$ is extreme point. Indeed, we know that every Boolean matrix is extreme point of the box $B_{m,n} = \{[x_{ij}] \in \mathbf{R}^{m \times n} : 0 \leq x_{ij} \leq 1 \,\forall i, j\}$, and it remains to refer to the evident fact: *When $Y \subset X$ is a nested pair of convex sets, then every extreme point $v$ of $X$ which happens to be in $Y$ is an extreme point of $Y$.* Indeed, were $v$ the midpoint of a nontrivial segment in $Y$, it would be the midpoint of a nontrivial segment in $X$, which is not the case.

Given that the set $B$ of all Boolean matrices from $\Pi_{m,n}$ belongs to $\mathrm{Ext}(\Pi_{m,n})$, all we need to conclude that $B = \mathrm{Ext}(\Pi_{m,n})$ is to show that $\Pi_{m,n} = \mathrm{Conv}(B)$ (see Fact II.8.5). Our plan is as follows: given a matrix $x \in \Pi_{m,n}$, we will show that $x$ can be made a North-Western $m \times n$ submatrix of $k \times k$ doubly stochastic matrix $\overline{x}$, with properly selected $k$. This is all we need: by Birkhoff Theorem, $\overline{x}$ is convex combination of $k \times k$ permutation matrices, implying that $x$ is a convex combination of the $m \times n$ North-Western submatrices of these permutation matrices, and these submatrices clearly are Boolean matrices from $\Pi_{m,n}$.

Thus, let a matrix $x \in \Pi_{m,n}$ be given; we want to extend it by adding several rows and columns to a larger doubly stochastic matrix. First of all, by adding to $x$ $n - m$ zero rows (if $n > m$) or $m - n$ zero columns (if $m > n$), we can reduce the situation to the one where $m = n$, which we assume from now on. Next, let $S = \sum_{i,j=1}^{n} x_{ij}$. Note that since the row sums in $x$ are $\leq 1$, we have $S \leq n$, so that $\kappa := n - S$ is nonnegative; let $d$ be the smallest integer which is $\geq \kappa$. This is how we can embed $x$, as the North-Western $n \times n$ submatrix, into $(n+d) \times (n+d)$ doubly stochastic matrix. Denote by $r_i$, $i \leq n$, the sum of entries in $i$-th row of $x$, and by $c_j$, $j \leq n$, the sum of entries in $j$-th column of $x$. Note that $0 \leq r_i \leq 1$, $0 \leq c_j \leq 1$ and $\sum_i r_i = \sum_j c_j = S$. Let also $\rho_i = 1 - r_i$, $\sigma_j = 1 - c_j$, so that $\rho_i \geq 0$, $\sigma_j \geq 0$, and $\sum_i \rho_i = \sum_j \sigma_j = \kappa$. Now let $\rho = [\rho_1/d; \rho_2/d; ...; \rho_n/d]$ and $\sigma = [\sigma_1/d; \sigma_2/d; ...; \sigma_n/d]$, so that $\rho$ and $\sigma$ are nonnegative vectors with sums of entries equal to $\kappa/d \leq 1$. Setting $\theta = (1 - \kappa/d)/d$ and specifying $\overline{x}$ as the $(n+d) \times (n+d)$ matrix $\left[\begin{array}{c|c|c|c} x & \rho & \cdots & \rho \\ \hline \sigma^\top & \theta & \cdots & \theta \\ \hline \vdots & \vdots & \ddots & \cdots \\ \hline \sigma^\top & \theta & \cdots & \theta \end{array}\right]$, we, as is immediately seen, get a doubly stochastic matrix, and this is the desired doubly stochastic extension of $x$. □

**Exercise II.14.** [Follow-up to Exercise II.13] Let $m, n$ be two positive integers with $m \leq n$, and $X_{m,n}$ be the set of $m \times n$ matrices $[x_{ij}]$ with $\sum_i |x_{ij}| \leq 1$ for all $j \leq n$ and $\sum_j |x_{ij}| \leq 1$ for all $i \leq m$. Describe the set $\mathrm{Ext}(X_{m,n})$. To get an educated guess, look at the matrices $\left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 0 & -1 \end{array}\right]$, $\left[\begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & -1 \end{array}\right]$, $\left[\begin{array}{ccc} 0.5 & -0.5 & 0 \\ -0.5 & 0.5 & 0 \end{array}\right]$ from $X_{2,3}$.

*Solution:* $\mathrm{Ext}(X_{m,n})$ is the set of all $m \times n$ matrices with entries $-1, 0, 1$ such that in every row there is exactly one nonzero entry, and in every column there is at most one nonzero entry.

In one direction: Let $x = [x_{ij}]$ be $m \times n$ matrix with entries $-1, 0, 1$, at most one nonzero entry per column, and exactly one nonzero entry per row; let us prove that $x \in \mathrm{Ext}(X_{m,n})$. First, $x$ clearly belongs to $X_{m,n}$. It remains to prove that if $x \pm h \in X_{m,n}$, then $h = 0$. Indeed, in the situation in question, denoting $\sigma(i)$ the index $j$ of the column with $x_{ij} \neq 0$ (for our $x$, such $j$ exists for every $i \leq m$), we should have $\sum_j |x_{ij} \pm h_{ij}| \leq 1$. In particular, $|x_{i\sigma(i)} \pm h_{i\sigma(i)}| \leq 1$, implying, in view of $|x_{i\sigma(i)}| = 1$ (all nonzero entries in our $x$ are of magnitude 1!) that $h_{i\sigma(i)} = 0$. Therefore $1 \geq \sum_j |x_{ij} \pm h_{ij}| = \underbrace{|x_{i\sigma(i)}|}_{=1} + \sum_{j \neq \sigma(i)} |x_{ij} \pm h_{ij}|$, implying that $h_{ij} = 0$ for $j \neq \sigma(i)$. Thus, $i$-th row in $h$ is zero; since $i \leq m$ is arbitrary, $h = 0$, as required.

In the opposite direction: Let $x \in \mathrm{Ext}(X_{m,n})$, and let us prove that $x$ has all entries in $\{-1, 0, 1\}$, with exactly one nonzero entry per row and at most one nonzero entry per column. Let $\xi_{ij} \in \{-1, 1\}$ be such that $\xi_{ij} x_{ij} = |x_{ij}|$ for all $i, j$, and let $\Xi$ be the one-to-one linear transformation of the space $\mathbf{R}^{m \times n}$ of $m \times n$ matrices given by entrywise multiplication of a matrix by the matrix $[\xi_{ij}]$. Linear one-to-one transformation $\Xi$ maps the polytope $X_{m,n}$ onto itself and thus maps onto itself the set $\mathrm{Ext}(X_{m,n})$. In particular, the matrix $\overline{x} = [|x_{ij}|]$ composed of magnitudes of entries in $x$ (this is the image of $x$ under the mapping $\Xi$) is an extreme point of $X_{m,n}$. Note that $\overline{x} \in \Pi_{m,n}$, where $\Pi_{m,n}$ is the polytope of entrywise nonnegative $m \times n$ matrices with all column and row sums not exceeding 1. Moreover, $\overline{x}$ is an extreme point of $\Pi_{m,n}$, since from $\overline{x} \pm h \in \Pi_{m,n}$ it clearly follows $\overline{x} \pm h \in X_{m,n}$, and the latter implies that $h = 0$ – $\overline{x}$ is an extreme point of $X_{m,n}$! By the result of Exercise II.13, $\overline{x}$ has entries 0 and 1 only, implying that all nonzero entries in $x$ are $\pm 1$. With this in mind, $\sum_i |x_{ij}| \leq 1$, $j \leq n$, implies that $x$ has at most one nonzero entry per column, and $\sum_j |x_{ij}| \leq 1$, $i \leq m$, implies that $x$ has at most one nonzero entry per row. It remains to verify that every row of $x$ has a nonzero entry. Assume the opposite, say, that the first row of $x$ is zero, and let us lead this assumption to a contradiction. In the case in question $x$ has at most $m - 1$ nonzero entries (since, as we have already seen, there is at most one nonzero entry per row, and the first row is zero). Consequently, among $n > m - 1$ columns of $x$ there is a zero column, w.l.o.g. let it be the first one. Thus, $x$ has zero first column and zero first row, which combines with $x \in X_{m,n}$ to imply that when $h$ is $m \times n$ matrix with the only nonzero entry, equal to 1, in the cell $1, 1$, we have $x \pm h \in X_{m,n}$, contradicting $x$ being an extreme point of $X_{m,n}$. □

**Exercise II.15.** [follow-up to Exercise II.13] Let $x$ be an $n \times n$ entrywise nonnegative matrix with all row and all column sums $\leq 1$. Is it true that for some doubly stochastic matrix $\overline{x}$, the matrix $\overline{x} - x$ is entrywise nonnegative?

*Solution:* Yes. By the result of Exercise II.13, $x$ is a convex combination of Boolean matrices with column and row sums $\leq 1$. Every matrix with the latter property clearly is obtained from appropriate permutation matrix by replacing with zeros some of the unit entries. Thus, every Boolean matrix with row and column sums $\leq 1$ is entrywise $\leq$ a permutation matrix, and therefore a convex combination of the matrices of the former class is entrywise $\leq$ a convex combination of permutation matrices, which is a doubly stochastic matrix. □

**Exercise II.16.** [Assignment problem] Consider the problem as follows:

> *There are $n$ jobs and $n$ workers. When worker $j$ is assigned with job $i$, we get profit $c_{ij}$. We want to assign every worker with a job in such a way that every worker is assigned with exactly one job and every job is assigned to exactly one worker. Under this restriction, we want to maximize the total profit.*

1. Pose the Assignment problem as the Boolean (i.e., with the decision variables restricted to be zeros and ones) Linear Programming problem.

   *Solution:* Encoding a candidate assignment by $n \times n$ matrix $x = [x_{ij}]$ with $x_{ij} = 1$ when job $i$ is assigned to worker $j$ and $x_{ij} = 0$ otherwise, we end up with the problem

$$\max_x \left\{ \sum_{i,j} c_{ij}x_{ij} : x_{ij} \geq 0, \sum_j x_{ij} = 1\,\forall i, \sum_i x_{ij} = 1\,\forall j, x_{ij} \in \{0,1\} \right\} \qquad (!)$$

2. Think how to solve the problem from item 1 via plain Linear Programming

   *Solution:* Removing in (!) the Boolean constraints $x_{ij} \in \{0,1\}$, we arrive at the LP problem of maximizing a linear form over the polytope of doubly stochastic $n \times n$ matrices. The problem clearly is solvable, and among its optimal solutions there are extreme points of the polytope. By Birkhoff Theorem, these extreme points are permutation matrices. Thus, passing from (!) to the LP relaxation of the problem, we preserve the optimal value, and every LP algorithm which produces extreme point solutions will, as applied to relaxation, provide us with an optimal solution to (!).

3. [computational study] Consider the special case of Assignment problem where all profits $c_{ij}$ are zeros or ones; you can interpret $c_{ij} = 1/0$ as the fact that worker $j$ knows/does not know how to execute job $j$. In this situation Assignment problem requires from us to find an assignment which maximizes the total number of executed jobs. Assume now that the matrix $C = [c_{ij}]$ is generated at random, with entries taking, independently of each other, value 1 with probability $\epsilon \in (0,1)$ and value 0 with probability $1 - \epsilon$. For $n \in \{4, 8, 16, 32, 64, 128, 256\}$ and $\epsilon \in \{1/2, 1/4, 1/8, 1/16\}$, run 100 simulations per pair $n, \epsilon$ to find the empirical mean of the ratio "number of executed jobs in optimal assignment"$/n$ and look at the results.

   *Solution:* Our results are as follows:

| | $n = 4$ | $n = 8$ | $n = 16$ | $n = 32$ | $n = 64$ | $n = 128$ | $n = 256$ |
|---|---|---|---|---|---|---|---|
| $\epsilon = 0.5000$ | 0.8800 | 0.9862 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| $\epsilon = 0.2500$ | 0.6025 | 0.8187 | 0.9769 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| $\epsilon = 0.1250$ | 0.3325 | 0.5650 | 0.8094 | 0.9719 | 0.9995 | 1.0000 | 1.0000 |
| $\epsilon = 0.0625$ | 0.2250 | 0.3688 | 0.5387 | 0.7906 | 0.9723 | 0.9993 | 1.0000 |

   The results allow to make an educated guess that with $\epsilon$ fixed and $n \to \infty$, the probability to get *all $n$ jobs executed* in the optimal assignment goes to 1; this guess happens to be true.

**Exercise II.17.** Let $\nu = (\nu_1, ..., \nu_K)$ with positive integer $\nu_i$, and let $\mathbf{S}^\nu = \mathbf{S}^{\nu_1} \times ... \times \mathbf{S}^{\nu_K}$ be the space of block-diagonal, with $K$ diagonal blocks of sizes $\nu_i \times \nu_i$, $i \leq K$, symmetric matrices, let $\mathbf{S}^\nu_+$ be the cone composed of positive semidefinite matrices from $\mathbf{S}^\nu$, and let $E$ be an $m$-dimensional affine plane in $\mathbf{S}^\nu$ which intersects $\mathbf{S}^\nu_+$. The intersection $X = E \cap \mathbf{S}^\nu_+$ is a closed nonempty convex set not containing lines and thus possessing extreme points. Let $W$ be such a point, $W^{ii}$ be the diagonal blocks of $W$, and $r_i$ be the ranks of $\nu_i \times \nu_i$ matrices $W^{ii}$. Prove that

$$\sum_{i=1}^{k} r_i(r_i + 1) \leq \sum_{i=1}^{K} \nu_i(\nu_i + 1) - 2m.$$

What happens in the diagonal case $\nu_1 = ... = \nu_K = 1$ ?

*Solution:* Let $W^{ii} = U_i\Lambda_i U_i^\top$ be eigenvalue decompositions of $W^{ii}$; w.l.o.g. we can assume that the first $r_i$ of eigenvalues of $W^{ii}$ are positive, and the remaining eigenvalues are zero. For every collection of $K$ symmetric $r_i \times r_i$ matrices $D^i$, denoting by $\overline{D}^i$ the $\nu_i \times \nu_i$ matrices obtained by augmenting $D^i$ with zero rows and columns, and setting $\overline{D} = \mathrm{Diag}\{U_1\overline{D}^1 U_1^\top, ..., U_K\overline{D}^K U_K^\top\}$, we get $W \pm t\overline{D} \succeq 0$ for all small positive $t$. Now let us impose on the matrices $D^i$ the requirement

$$\overline{D} \in L, \tag{!}$$

where $L$ is the parallel to $E$ linear subspace in $\mathbf{S}^\nu$. Assuming that

$$\mathrm{codim}\, L := \sum_i \nu_i(\nu_i+1)/2 - m < R := \sum_i r_i(r_i+1)/2,$$

relation (!), which is a system of $\mathrm{codim}\, L$ homogeneous linear equations on $R$ variables $\{D^i_{pq}, p \le q \le r_i, i \le K\}$, has a nontrivial solution, implying that $W \pm tD \in X$ for some nonzero $D$ and positive $t$, which is impossible. Thus, $\mathrm{codim}\, L \ge R$, as claimed. □

In the diagonal case, the result becomes the following fact (perfectly well known to everybody who somehow dealt with the Simplex method in LP): *The number of nonzero entries in any extreme point of the feasible set of a feasible LP problem in the standard form* $\max_{x\in\mathbf{R}^k}\{c^\top x : Ax = b, x \ge 0\}$ *does not exceed the number of equality constraints (i.e., of rows in A).*

**Exercise II.18.** Let $M$ be a closed convex set in $\mathbf{R}^n$ and $\bar{x}$ be a point of $M$.

1. Prove that if there exists a linear form $a^\top x$ such that $\bar{x}$ is the *unique* maximizer of the form on $M$, then $\bar{x}$ is an extreme point of $M$.

2. Is the inverse of 1) true, i.e., is it true that every extreme point $\bar{x}$ of $M$ is the unique maximizer, over $x \in M$, of a properly selected linear form?

*Solution:* 1: the answer is positive, Indeed, let $\bar{x}$ be the unique maximizer over $x \in M$ of a linear form $f^\top x$. Assuming, on the contrary to what should be proved, that $\bar{x} \pm h \in M$ for some $h \neq 0$, the linear function $f^\top x$ attains its maximum on the segment $[\bar{x} - h, \bar{x} + h]$ in the midpoint of this segment, which for a linear function is possible only when the function is constant on the segment. Thus, all points on the segment maximize $f^\top x$ over $x \in M$, contradicting the fact that $\bar{x}$ is the unique maximizer of the function on $M$.

2: The inverse is not true in general. For example, consider the set

$$M = \{(x,y) \in \mathbf{R}^2 : y \ge \left\{\begin{array}{ll} x^2 & , x \le 0 \\ 0 & , 0 \le x \le 1 \\ (x-1)^2 & , x \ge 1 \end{array}\right. \}$$

(draw picture). The origin clearly is an extreme point of the set, but there are no linear forms on $\mathbf{R}^2$ attaining their maximum over $M$ at the origin, and only at it.

**Exercise II.19.** Identify and justify the correct claims in the following list:

1. Let $X \subset \mathbf{R}^n$ be a nonempty closed convex set, $P$ be an $m \times n$ matrix, $Y = PX := \{Px : x \in X\} \subset \mathbf{R}^n$, and $\overline{Y}$ be the closure of $Y$. Then
   - For every $x \in \mathrm{Ext}(X)$, $Px \in \mathrm{Ext}(\overline{Y})$

     *Solution:* Wrong – look at the orthogonal projection of the planar triangle with vertices $(0,0)$, $(1,1)$, $(2,0)$ onto the first coordinate axis.
   - Every extreme point of $\overline{Y}$ which happens to belong to $Y$ is $Px$ for some $x \in \mathrm{Ext}(X)$

     *Solution:* Wrong – look what happens when $X$ is the stripe $0 \le x \le 1$ on the 2D plane
   - When $X$ does not contain lines, then every extreme point of $\overline{Y}$ which happens to belong to $Y$ is $Px$ for some $x \in \mathrm{Ext}(X)$

*Solution:* Correct. Indeed, let $w \in \mathrm{Ext}(\overline{Y}) \cap Y$. Then the set $X_w = \{x \in X : Px = w\}$ is nonempty, convex, closed and does not contain lines, and thus has an extreme point $\overline{x}$. It suffices to verify that $\overline{x} \in \mathrm{Ext}(X)$. Indeed, let $d$ be such that $\overline{x} \pm d \in X$, and let us prove that $d = 0$. We have $P[\overline{x} \pm d] \in Y$, and since $w = P\overline{x} \in \mathrm{Ext}(\overline{Y})$, we get $Pd = 0$, implying that $\overline{x} \pm d \in X_w$; since $\overline{x} \in \mathrm{Ext}(X_w)$, we conclude that $d = 0$. Note that this reasoning demonstrates that when $X$ does not contain lines, every extreme point of $Y$ is $Px$ for some $x \in \mathrm{Ext}(X)$.

2. Let $X, Y$ be nonempty closed convex sets in $\mathbf{R}^n$, and let $Z = X + Y$, $\overline{Z} = \mathrm{cl}\, Z$. Then
   - If $w \in \mathrm{Ext}(\overline{Z}) \cap Z$, then $w = x + y$ for some $x \in \mathrm{Ext}(X)$ and $y \in \mathrm{Ext}(Y)$.

     *Solution:* Correct. Indeed, we have $w = x + y$ for some $x \in X, y \in Y$. If $x \notin \mathrm{Ext}(X)$, then $x \pm d \in X$ for some $d \neq 0$, whence $w \pm d = (w \pm d) + y \in Z$, contradicting $w \in \mathrm{Ext}(\overline{Z})$. By similar reasoning, $y \in \mathrm{Ext}(Y)$.
   - If $x \in \mathrm{Ext}(X)$, $y \in \mathrm{Ext}(Y)$, then $x + y \in \mathrm{Ext}(\overline{Z})$.

     *Solution:* Wrong – look what happens when $X = [0, 1] \subset \mathbf{R}$ and $Y = [2, 3] \subset \mathbf{R}$.

**Exercise II.20.** Let $X = \{x \in \mathbf{R}^n : a_i^\top x \le b_i, i \le m\}$ be a nonempty polyhedral set and $f^\top x$ be a linear form of $x \in \mathbf{R}^n$ which is bounded above on $X$:

$$\mathrm{Opt}(f) = \sup_{x \in X} f^\top x < \infty$$

Prove that

1. $\mathrm{Opt}(f)$ is achieved – the set $\mathop{\mathrm{Argmax}}_{x \in X} f^\top x := \{x \in X : f^\top x = \mathrm{Opt}(f)\}$ is nonempty.

   *Solution:* This is nothing but the claim that bounded and feasible LP program has a solution (section 3.2.1 or Theorem II.10.3).
2. The set $\mathop{\mathrm{Argmax}}_{x \in X} f^\top x$ is as follows: there exists an index set $I \subset \{1, 2, ..., m\}$, perhaps empty, such that

   $$\mathop{\mathrm{Argmax}}_{x \in X} f^\top x = X_I := \{x : a_i^\top x \le b_i \, \forall i, \ a_i^\top x = b_i \, \forall i \in I\}$$

   *Solution:* By Linear Programming Duality Theorem, the problem dual to primal problem $\mathrm{Opt}(f) = \max_x \{f^\top x : a_i^\top x \le b_i, i \le I\}$ reads $\max_\lambda \left\{\lambda^\top b : \lambda \ge 0, \sum_i \lambda_i a_i = f\right\}$ and is solvable with the optimal value $\mathrm{Opt}(f)$ – the same as the one of the primal problem. Let $\lambda^*$ be an optimal solution to the dual problem, and let $I = \{i : \lambda_i^* > 0\}$. We claim that $X_* := \mathop{\mathrm{Argmax}}_{x \in X} f^\top x = X_I$. In one direction: when $x \in X_*$, we have $x \in X$ and

   $$\mathrm{Opt}(f) = f^\top x = [\sum\nolimits_i \lambda_i^* a_i]^\top x = \sum\nolimits_{i \in I} \lambda_i^* [a_i^\top x] \le \sum\nolimits_{i \in I} \lambda_i^* b_i = b^\top \lambda^* = \mathrm{Opt}(f),$$

   where the inequality is due to $\lambda_i^* \ge 0$, and the last equality – due to the fact that $\lambda^*$ is optimal solution to the dual problem, and the dual optimal value is $\mathrm{Opt}(f)$. We conclude that the inequality in the chain is equality, so that $\sum_{i \in I} \lambda_i^* [b_i - a_i^\top x] = 0$. The latter relation implies $a_i^\top x = b_i$, $i \in I$ (since $a_i^\top x \le b_i$ for all $i$ and $\lambda_i^* > 0, i \in I$). In addition, $x \in X$, and we conclude that $x \in X_I$. Thus, $X_* \subset X_I$. Vice versa, if $x \in X_I$, then $x \in X$ and

   $$f^\top x = [\sum\nolimits_{i \in I} \lambda_i^* a_i]^\top x = \sum\nolimits_{i \in I} \lambda_i^* b_i = b^\top \lambda^* = \mathrm{Opt}(f),$$

   that is, $x \in X_*$ □
3. Vice versa, if $I \subset \{1, ..., m\}$ is such that the set $X_I = \{x : a_i^\top x \le b_i \, \forall i, a_i^\top x = b_i \, \forall i \in I\}$ is nonempty, then $X_I = X_* := \mathrm{Argmax}_{x \in X} f^\top x$ for properly selected $f$.
   *Note:* Nonempty sets of the form $X_I$, $I \subset \{1, ..., m\}$, are called *faces* of the polyhedral set $X$. This definition is not geometric – according to it, whether a given set $Y$ is or is not a face in $X$, may depend not on $X$ *per se*, but on its representation as the solution set of a finite system of linear inequalities. Items 2—3, taken together, state that in fact being a face of a polyhedral set is a geometric property – faces are exactly the sets $\mathop{\mathrm{Argmax}}_{x \in X} f^\top x$ of all maximizers of linear forms bounded from above on $X$.

*Solution:* Indeed, given $I \subset \{1, 2, ..., n\}$ such that $X_I$ is nonempty, let us set $f = \sum_{i \in I} a_i$ [6], so that for $x \in X_I$ one has $f^\top x = \sum_{\in I} b_i$. On the other hand, for every $x \in X$ we have $f^\top x = \sum_{i \in} a_i^\top x \leq \sum_{i \in I} b_i$. We conclude that $\mathrm{Opt}(f) = \sum_{i \in I} b_i$ and $X_I \subset X_* := \operatorname*{Argmax}_{x \in X} f^\top x,$. The same reasoning as in the concluding part of the solution to the previous item (where $\lambda_i^*$, $i \in I$, should be set to 1) demonstrates the opposite inclusion $X_* \subset X_I$. Thus, $X_I = \mathrm{Argmax}_{x \in X} f^\top x$. □

4. Extreme points of a face of $X$ are extreme points of $X$.

   *Solution:* Assume that $v$ is an extreme point of a face $X_I$ of $X$; to prove that is an extreme point of $X$ as well, we should show that whenever $v \pm h \in X$, it holds $h = 0$. To this end, note that if $v \pm h \in X$, then $a_i^\top [v \pm h] \leq b_i$ for all $i$; when $i \in I$, the inequalities $a_i^\top [v \pm h] \leq b_i$ imply that $a_i^\top [v \pm h] = b_i$ due to $a_i^\top v = b_i$. We see that in fact $v \pm h \in X_I$; since $v$ is extreme point of $X_I$, we end up with the desired conclusion $h = 0$. □

5. Extreme points of $X$, if any, are exactly the faces of $X$ which are singletons.
   *Note:* As a corollary of 1—3, 5, we see that extreme points of polyhedral set $X$ are exactly the maximizers of those linear forms which achieve their maximum on $X$ at a unique point.

   *Solution:* In one direction: let $v$ be an extreme point of $X$. By Theorem II.9.1, there exists $n$-element set $I \subset \{1, ..., m\}$ such that $a_i^\top v = b_i$ for $i \in I$ and the $n$ vectors $a_i$, $i \in I$, are linearly independent. Since, in addition, $v \in X$, we conclude that $v \in X_I$, and the latter set is a singleton due to linear independence of $a_i$, $i \in I$. In the opposite direction: let $X_I = \{v\}$ for some $I$; then of course, $v$ is an extreme point of $X_I$, which in view of item 4 implies that $v$ is an extreme point of $X$.

**Exercise II.21.** [Follow-up to Exercise II.20]

1. Let $X \subset Y$ be nonempty closed convex sets in $\mathbf{R}^n$. Is it true that $\mathrm{Ext}(Y) \cap X \subset \mathrm{Ext}(X)$ ?

   *Solution:* The answer clearly is positive. Indeed, assuming that $w \in \mathrm{Ext}(Y) \cap X$ is not an extreme point of $X$, $w$ is the midpoint of a nontrivial segment $\Delta \in X$ and thus – a nontrivial segment $\Delta \subset Y$ (since $X \subset Y$), which is impossible.

2. Let $X$ be a nonempty closed convex set contained in the polyhedral set $\{x : Ax \leq b\}$. Assuming that the set $\overline{X} = X \cap \{x : Ax = b\}$ is nonempty, is it true that $\mathrm{Ext}(\overline{X}) = \mathrm{Ext}(X) \cap \overline{X}$ ?

   *Solution:* The answer is positive. Indeed, by item 1 it holds $\mathrm{Ext}(X) \cap \overline{X} \subset \mathrm{Ext}(X)$ due to $\overline{X} \subset X$. To prove the opposite inclusion, assume that an extreme point $w$ of $\overline{X}$ is not extreme point of $X$, and let us lead this assumption to a contradiction. Since $w \in \overline{X} \subset X$, we have $w \in X$, and since $w$ is not an extreme point of $X$, there exists a nontrivial segment $\Delta = [\underline{x}, \overline{x}] \subset X$ with $w$ as the midpoint. By assumption, $A\overline{x} \leq b$ and $A\underline{x} \leq b$, which combines with $A\frac{1}{2}[\underline{x} + \overline{x}] = Aw = b$ to conclude that $A\underline{x} = A\overline{x} = b$, that is, $\Delta \subset \overline{X}$. The bottom line is that $w$ is the midpoint of a nontrivial segment in $\overline{X}$, which is he desired contradiction – $w$ is an extreme point of $\overline{X}$!

3. By the result of Exercise II.13, the extreme points of the polytope $\Pi_{m,n} = \{[x_{ij}] \in \mathbf{R}^{m \times n} : x_{ij} \geq 0, \sum_i x_{ij} \leq 1 \, \forall j, \sum_j x_{ij} \leq 1 \, \forall i\}$ are exactly the Boolean matrices from this polytope. Now let $\widehat{\Pi}_{m,n}$ be the part of $\Pi_{m,n}$ cut off $\Pi_{m,n}$ by imposing on prescribed row and columns of $m \times n$ matrix $x \in \Pi_{m,n}$ the requirement to be equal to 1, rather than to be $\leq 1$. Assuming $\widehat{\Pi}_{m,n}$ nonempty, prove that the extreme points of this polytope are exactly the Boolean matrices contained in it.

   *Solution:* The fact that Boolean matrices contained in $\widehat{\Pi}_{m,n}$ are extreme points of this polytope is readily given by item 1 – we have already mentioned that these matrices are extreme points of the larger polytope $\Pi_{m,n}$. It remains to note that $\widehat{\Pi}_{m,n}$ is cut off $\Pi_{m,n}$ by converting into equalities several inequalities satisfied everywhere on $\Pi_{m,n}$, and thus by item 2 extreme points of $\widehat{X}$ are extreme points of $\Pi_{m,n}$ and thus are Boolean matrices.

[6] recall that by our standard convention, $\sum_{i \in \varnothing} a_i = 0$.

**Exercise II.22.** Let $X \subset \mathbf{R}^m$ be a nonempty polyhedral set, $x \mapsto Px + p : \mathbf{R}^n \to \mathbf{R}^m$ be an affine mapping, and $Y$ be the image of $X$ under this mapping. Mark by $\mathbf{T}$ the statements in the below list which are always (i.e., for all $X, P, p$ compatible with the above assumptions) true:

1. $Y$ is a nonempty polyhedral set.

   *Solution:* True, rule 4 in calculus of polyhedral representations, see section 3.3.

2. If $X$ does not contain lines, so is $Y$.

   *Solution:* False – take $X = \left\{[x; y] \in \mathbf{R}^2 : \; y \geq |x|\right\}$ and consider the affine map $P[x; y] + p \equiv x$. Then, $Y = \mathbf{R}$ and is itself a straight line.

3. If $X$ does contain lines, so does $Y$.

   *Solution:* False – take $X = \{[x; y] \in \mathbf{R}^2 : |x| \leq 1\}$ and $P[x; y] + p \equiv x$, resulting in $Y = [-1, 1]$.

4. If $v$ is an extreme point of $X$, then $Pv + p$ is an extreme point of $Y$.

   *Solution:* False – take $X = \{[x; y] \in \mathbf{R}^2 : |x| + |y| \leq 1\}$ and $P[x; y] + p \equiv x$, resulting in $Y = [-1, 1]$. The image of the extreme point $[0; 1]$ of $X$ under the affine mapping in question is not extreme for $Y$.

5. If $z$ is an extreme point of $Y$, then $z = Pv + p$ for certain extreme point $z$ of $X$.

   *Solution:* False– take $X = \{[x; y] \in \mathbf{R}^2 : |x| \leq 1\}$ and $P[x; y] + p \equiv x$, resulting in $Y = [-1, 1]$. $Y$ has extreme points, and $X$ does not.

6. If $z$ is an extreme point of $Y$ and $X$ does not contain lines, then $z = Pv + p$ for certain extreme point $z$ of $X$.

   *Solution:* True. By Exercise II.20, there is a linear form $f^\top y$ which attains its maximum over $y \in Y$ at $z$, and only at this point. It follows that the form $g^\top x$, $g = P^\top f$, attains its maximum over $x \in X$ exactly at the set $X^z = \{x \in X : Px + p = z\}$. By Exercise II.20, $X^z$ is a face of $X$. Since $X$ does not contain lines, so is $X^z$, implying that $X^z$ has an extreme point, call it $v$. Since $v \in X^z$, we have $Pv + p = z$, and since $v$ is an extreme point of face of $X$, it is extreme point of $X$ by Exercise II.20.4. □

**Exercise II.23.** Find extreme points of the following closed convex sets:

1. The set $\mathcal{S}_n = \{X \in \mathbf{S}^n : -I_n \preceq X \preceq I_n\}$

   *Solution:* $\mathrm{Ext}\{\mathcal{S}_n\}$ is the set of all matrices from $\mathbf{S}^n$ which are orthogonal, or, which is the same, symmetric $n \times n$ matrices with eigenvalues $\pm 1$.

   In one direction: Let $W$ be an orthogonal symmetric matrix; let us prove that this is an extreme point. Indeed, assuming that $W \pm D \in \mathcal{S}_n$ for some $D$, let us prove that $D = 0$. Otherwise there exists $x \in \mathbf{R}^n$ with $Dx \neq 0$; assuming w.l.o.g. that $\|x\|_2 = 1$, we have $\|Wx\|_2 = 1$, and $\|[W \pm D]x\|_2 \leq 1$ (since the spectral norm $\|V\|_{2,2}$ of a symmetric matrix $V \in \mathcal{S}_n$ is the maximum of magnitudes of eigenvalues of $V$ and is therefore $\leq 1$), On the other hand, assuming w.l.o.g. that $[Dx]^\top [Wx] \geq 1$, we have $\|[W + D]x\|_2^2 = \|Wx\|_2^2 + 2[Dx]^\top [Wx] + \|Dx\|_2^2 \geq \|Wx\|_2^2 + \|Dx\|_2^2 = 1 + \|Dx\|_2^2 > 1$, which is a desired contradiction.

   In the opposite direction: Let $W$ be an extreme point of $\mathcal{S}_n$ and $W = U \,\mathrm{Diag}\{\lambda\} U^\top$ be the eigenvalue decomposition of $W$; we should verify that $\lambda$ is a $\pm 1$ vector. We clearly have $\|\lambda\|_\infty \leq 1$, and if $\|\lambda \pm d\|_\infty \leq 1$, then $W \pm U \,\mathrm{Diag}\{d\} U^\top \in \mathcal{S}_n$, implying that $d = 0$. Thus, $\lambda$ is an extreme point of the unit box $\{x \in \mathbf{R}^n : \|x\|_\infty \leq 1\}$, and these points are the $\pm 1$ vectors. □

2. The set $\mathcal{S}_n^+ = \{X \in \mathbf{S}^n : 0 \preceq X \preceq I_n\}$

   *Solution:* The extreme points are exactly the orthogonal projectors – symmetric $n \times n$ matrices with eigenvalues 0 and 1. To see it, note that $\mathcal{S}_n^+$ is the image of $\mathcal{S}_n$ under the one-to-one affine mapping $X \mapsto \frac{1}{2}[X + I_n] : \mathbf{S}^n \to \mathbf{S}^n$.

3. The set $\mathcal{D}_{k,n} = \{X \in \mathbf{S}^n : I_n \succeq X \succeq 0, \mathrm{Tr}(X) = k\}$, where $k$ is a positive integer $\leq n$.

*Solution:* The extreme points are exactly the orthogonal rank $k$ projectors, or, which is the same, the symmetric $n \times n$ matrices with $k$ eigenvalues equal to 1 and the remaining eigenvalues equal to 0.

In one direction: let $W \in \mathrm{Ext}(\mathcal{D}_{k,n})$, and let us prove that $k$ eigenvalues of $W$ are equal to 1, and the remaining – to 0. Indeed, let $W = U\,\mathrm{Diag}\{\lambda\}U^\top$ be the eigenvalue decomposition of $W$; since $W \in \mathcal{D}_{k,n}$, we should have $0 \leq \lambda_i \leq 1$ for $i \leq n$, and $\sum_i \lambda_i = k$. If now $d \in \mathbf{R}^n$ is such that $0 \leq \lambda_i \pm d_i \leq 1$ for $i \leq n$ and $\sum_i d_i = 0$, then $W \pm U\,\mathrm{Diag}\{d\}U^\top \in \mathcal{D}_{k,n}$, implying that $d = 0$ due to $W \in \mathrm{Ext}(\mathcal{D}_{k,n})$. Thus, $\lambda$ should be an extreme point of the set $\{x \in \mathbf{R}^n : 0 \leq x_i \leq 1, i \leq n, \sum_i \lambda_i = k\}$. As we know from Example II.9.1 in section 9.3, this implies that $k$ entries in $\lambda$ are equal to 1, and the remaining to 0.

In the opposite direction: Let $W$ be symmetric $n \times n$ matrix with $k$ eigenvalues equal to 1 and the remaining eigenvalues equal to 0, and let us prove that $W$ is an extreme point of $\mathcal{D}_{k,n}$. Passing to representations of matrices in the eigenbasis of $W$, we lose nothing when assuming that $W = \sum_{i=1}^k e_i e_i^\top$, where $e_1, ..., e_n$ are the standard basic orths in $\mathbf{R}^n$. To see that $W \in \mathrm{Ext}(\mathcal{D}_{k,n})$, we should prove that if $\sum_{i=1}^k e_i e_i^\top \pm D \in \mathcal{D}_{k,n}$ and $D$ is symmetric, then $D = 0$. Indeed, for $D$ satisfying the premise of this claim, the diagonal entries of $\sum_{i=1}^k e_i e_i^\top \pm D$ should be between 0 and 1 and sum up to $k$, implying that $D_{ii} = 0$ for all $i$ (the same Example II.9.1 we have already mentioned). In other words, the diagonals of positive semidefinite symmetric matrices $B^\pm = \sum_{i=1}^k e_i e_i^\top \pm D$ are $(\underbrace{1, ..., 1}_{k}, 0, ..., 0)$, implying that $D_{ij} = D_{ji} = B^\pm_{ij} = B^{ji} = 0$ whenever $\max[i,j] > k$ (since the $2 \times 2$ principal minors in $B^\pm$ should be nonnegative). Thus, all entries in $D$ outside of the $k \times k$ angular submatrix $\overline{D}$ of $D$ are zeros. Next, the matrices $I_k \pm \overline{D}$ are angular submatrices $E^\pm$ of symmetric matrices with eigenvalues between 0 and 1, implying by the Eigenvalue Interlacement Theorem that the eigenvalues of the symmetric $k \times k$ matrices $E^\pm$ are between 0 and 1, so that $E^\pm \in \mathcal{S}_k^+$. From item 2 we know that $I_k$ is an extreme point of the latter set, implying that $\overline{D} = 0$. The bottom line is that $D = 0$, □

4. The set $\mathcal{M}_n = \{X \in \mathbf{R}^{n \times n} : \|X\|_{2,2} \leq 1\}$ ($\|\cdot\|_{2,2}$ is the spectral norm)

*Solution:* The extreme points are exactly the orthogonal $n \times n$ matrices. To see that an orthogonal $n \times n$ matrix $W$ is an extreme point of $\mathcal{M}_n$, you can use exactly the same reasoning as in the proof of the similar fact for $\mathcal{S}_n$, with $D \in \mathbf{R}^{n \times n}$ rather than $D \in \mathbf{S}_n$. To see that if $W$ is an extreme point of $\mathcal{M}_n$, then $W$ is orthogonal, or, which is the same, with all singular values equal to 1, look at the singular value decomposition $W = U\,\mathrm{Diag}\{\sigma\}V^\top$ of $W$, From $\|W\|_{2,2} \leq 1$ it follows that $\|\sigma\|_\infty \leq 1$, and if certain singular value $\sigma_i$ is $< 1$, then the singular values of the matrices $W \pm tU[e_i e_i^\top]V^\top$ ($e_i$ is $i$-th basic orth) for small positive $t$ are $\leq 1$, implying that $W \pm tU[e_i e_i^\top]V^\top \in \mathcal{M}_n$, which is impossible, □

**Exercise II.24.** Prove the following fact (which can be considered as a matrix extension of Birkhoff Theorem):

> For positive integers $d, n$, let $\Pi_{d,n}$ be the set of all $n \times n$ block matrices with $d \times d$ symmetric blocks $X^{ij}$ satisfying
>
> $$X^{ij} \succeq 0, \sum_j \mathrm{Tr}(X^{ij}) = 1 \forall i, \sum_i \mathrm{Tr}(X^{ij}) = 1 \forall j.$$
>
> The extreme points of $\Pi_{d,n}$ are exactly the block matrices $[X^{ij}]_{i,j \leq n}$ as follows: for certain $n \times n$ permutation matrix $P$ and unit vectors $e_{ij} \in \mathbf{R}^d$, one has
>
> $$X^{ij} = P_{ij} e_{ij} e_{ij}^\top \forall i, j.$$

*Solution:* In one direction: Let $[W^{ij}]$ be an extreme point of $\Pi_{d,n}$ and $P_{ij} = \mathrm{Tr}(W_{ij})$, so that $P$ is doubly stochastic. For every $i, j$, $W^{ij}$ should be an extreme point of the set $\mathcal{D}_{ij} = \{X \in \mathbf{S}^d : X \succeq 0, \mathrm{Tr}(X) = P_{ij}\}$ (why?), whence, by item 3 of Exercise II.23, $W^{ij} = P_{ij} e_{ij} e_{ij}^\top$ for some unit $e_{ij}$. Besides this, $P$ should be an extreme point of the polytope of doubly stochastic $n \times n$ matrices, since otherwise

$P \pm D$ will be doubly stochastic for some nonzero $D$, implying that $W \pm \underbrace{[D_{ij}e_{ij}e_{ij}^\top]_{i,j\leq n}}_{\overline{D}} \in \Pi_{d,n}$ for nonzero block-matrix $\overline{D}$ with symmetric blocks, contradicting the fact that $W \in \mathrm{Ext}(\Pi_{d,n})$. Thus, by Birkhoff Theorem, $P$ is a permutation matrix, and $W = [P_{ij}e_{ij}e_{ij}^\top]$ with unit $e_{ij} \in \mathbf{R}^d$. □

In the opposite direction: Let $W = [P_{ij}e_{ij}e_{ij}^\top]$ with unit $e_{ij}$ and permutation matrix $P$, and let $W \pm [D^{ij}] \in \Pi_{d,n}$ for some block-matrix with symmetric blocks $D^{ij}$; we should prove that $D^{ij} = 0$ for all $i, j$. If $i, j$ are such that $P_{ij} = 1$, then the $d \times d$ matrices $e_{ij}e_{ij}^\top \pm D^{ij}$ are $\succeq 0$ with trace not exceeding 1 (as blocks in a matrix from $\Pi_{d,n}$), whence both matrices are $\succeq 0$, $\preceq I_d$, and with trace 1 (the latter – due to $\mathrm{Tr}(e_{ij}e_{ij}^\top) = 1$). Thus, $e_{ij}e_{ij}^\top \pm D^{ij} \in \mathcal{D}_{1,d}$; applying item 3 of Exercise II.23, we conclude that $D^{ij} = 0$. And if $P_{ij} = 0$, then $W^{ij} \pm D^{ij}$ should be $\succeq 0$, again implying that $D^{ij} = 0$. □

**Exercise II.25.** Let $k, n$ be positive integers with $k \leq n$, and let $s_k(\lambda)$ for $\lambda \in \mathbf{R}^n$ be the sum of $k$ largest entries in $\lambda$. From the description of the extreme points of the polytope $X = \{x \in \mathbf{R}^n : 0 \leq x_i \leq 1, i \leq n, \sum_{i=1}^n x_i \leq k\}$, see Example II.9.2 in section 9.3, it follows that when $\lambda \in \mathbf{R}^n_+$, then

$$\max_{x\in X} \sum_{i=1}^n \lambda_i x_i = s_k(\lambda).$$

Prove the following matrix analogy of this fact:

> For $k, n$ as above, let $\mathcal{X} = \{(X_1, ..., X_n) : X_i \in \mathbf{S}^d, 0 \preceq X_i \preceq I_d, i \leq n, \sum_{i=1}^n X_i \preceq kI_d\}$. Then for $\lambda \in \mathbf{R}^n_+$ one has
>
> $$(X_1, ..., X_n) \in \mathcal{X} \implies \sum_{i=1}^n \lambda_i X_i \preceq s_k(\lambda)I_d,$$
>
> with the concluding $\preceq$ being $=$ for properly selected $(X_1, ..., X_n) \in \mathcal{X}$.

*Solution:* Assuming w.l.o.g. that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$, for $X = (X_1, ..., X_n) \in \mathcal{X}$, setting $S_i = \sum_{j=1}^i X_j$, we have $S_i \preceq \min[i, k]I_d$. When $k = n$, we clearly have $\sum_{i=1}^n \lambda_i X_i \preceq \sum_{i=1}^n \lambda_i I_d = s_n(\lambda)I_d$, with $\preceq$ being $=$ when $X_i = I_d$, $i \leq n$ Now let $k < n$, and let $\overline{X}_i = \left\{\begin{array}{ll} I_d, & i \leq k \\ 0, & i > k \end{array}\right.$, $\overline{S}_i = \sum_{j=1}^i \overline{X}_i$; note that $\overline{S}_i = \min[i, k]I_d \succeq S_i$, $i \leq n$. We have

$$\begin{array}{l} \sum_{i=1}^n \lambda_i X_i = \sum_{i=1}^n \lambda_i [S_i - S_{i-1}] = \sum_{i=1}^{n-1} S_i \underbrace{[\lambda_i - \lambda_{i+1}]}_{\geq 0} + \underbrace{\lambda_n}_{\geq 0} S_n \\ \preceq \sum_{i=1}^{n-1} [\lambda_i - \lambda_{i+1}]\overline{S}_i + \lambda_n \overline{S}_n \\ = \sum_{i=1}^n \lambda_i [\overline{S}_i - \overline{S}_{i-1}] = s_k(\lambda)I_d, \end{array}$$

and the resulting inequality $\sum_i \lambda_i X_i \preceq s_k(\lambda)I_d$ is equality when $X_i = \overline{X}_i$, $i \leq n$. □

## Cones and extreme rays

**Exercise II.26.** Let $X$ be a nonempty closed and bounded set in $\mathbf{R}^n$. Which of the following statements are true?

1. $\mathrm{Conv}(X)$ is closed convex set.

   *Solution:* True – see Corollary I.2.5

2. $\mathrm{Cone}(X)$ is a closed cone.

   *Solution:* Wrong in general. When $X = \{x \in \mathbf{R}^2 : x_1^2 + (x_2 - 1)^2 \leq 1\}$ (circle of unit radius in the upper half-plane touching the $x_1$-axis at the origin), $\mathrm{Cone}(X)$ is the open upper half-plane $\{x = [x_1; x_2] : x_2 > 0\}$ with origin added; this cone is not closed

3. When $X$ is convex, $\mathrm{Cone}(X)$ is closed cone.

*Solution:* Wrong in general, see example to item 2.

4. When $0 \notin X$, $\mathrm{Cone}(X)$ is a closed cone.

   *Solution:* Wrong in general. When $X = X^+ \cup X^-$ with $X^+ = \{[x_1; x_2; 1] \in \mathbf{R}^3 : x_1^2 + (x_2-1)^2 \leq 1\}$, $X^- = \{[x_1; x_2; -1] \in \mathbf{R}^3 : x_1^2 + (x_2-1)^2 \leq 1\}$, $\mathrm{Cone}(X)$ contains the circle $\{x_1^2 + (x_2-1)^2 \leq 1\}$ in the plane $x_3 = 0$ and therefore contains the conic hull of this circle. As a result, $\mathrm{cl}\,\mathrm{Cone}(X)$ contains the tangent line $\{x_2 = 0, x_3 = 0\}$ to this circle, and this line clearly does not belong to $\mathrm{Cone}(X)$.

5. When $0 \notin X$ and $X$ is convex, $\mathrm{Cone}(X)$ is closed cone.

   *Solution:* True. The fact that $\mathrm{Cone}(X)$ is a cone holds true for every $X$; all we need is to prove that under the circumstances this cone is closed. Since $X$ is nonempty closed convex set and $0 \notin X$, Separation Theorem applied to $\{0\}$ and $X$ says that these two sets can be strongly separated, so that for properly selected $e$ it holds $0 = e^\top 0 < \alpha := \inf_{x\in X} e^\top x$. Now let $y = \lim_{t\to\infty} y_t$ with $y_t \in \mathrm{Cone}(X)$; we want to prove that $y \in \mathrm{Cone}(X)$. We have $y_t = \sum_{i \leq I_t} \lambda_{ti} x_{ti}$ with $\lambda_{ti} \geq 0$ and $x_{ti} \in X$. Setting $\lambda_t = \sum_i \lambda_{ti}$, we have

   $$e^\top y = \lim_{t\to\infty} e^\top y_t = \lim_{t\to\infty} \sum_i \lambda_{it} \underbrace{e^\top x_{ti}}_{\geq \alpha > 0}.$$

   implying that the sequence of nonnegative reals $\lambda_t$ is bounded. Therefore, passing to a subsequence, we may assume that $\lambda_t \to \bar{\lambda}$ as $t \to \infty$. Taking into account that $\|y_t\|_2 \leq C\lambda_t$ with $C = \max_{x\in X} \|x\|_2 < \infty$, we see that when $\bar{\lambda} = 0$, one has $y = 0$, whence $y \in \mathrm{Cone}(X)$. And when $\bar{\lambda} > 0$, $y = \lim_{t\to\infty} y_t$ implies that $y = \bar{\lambda} \lim_{t\to\infty} \bar{x}_t$ with $\bar{x}_t = \lambda_t^{-1} \sum_i \lambda_{ti} x_{ti}$ (these points are well defined for large enough $t$'s). Since $X$ is convex, the points $\bar{x}_t$ belong to $X$, and since $X$ is closed, the point $\bar{x} := \lim_{t\to\infty} \bar{x}_t$ belongs to $X$ as well. Thus, $y$ is a positive multiple of a point from $X$, so that $y \in \mathrm{Cone}(X)$. □

6. When $X$ is polyhedral, $\mathrm{Cone}(X)$ is a closed cone.

   *Solution:* True. By Krein-Milman Theorem, nonempty bounded polyhedral set is $\mathrm{Conv}\{v_1, ..., v_N\}$ for a finite nonempty set $\{v_1, ..., v_N\}$, whence clearly $\mathrm{Cone}(X) = \mathrm{Cone}(\{v_1, ..., v_N\}) = \{y = \sum_i \lambda_i v_i : \lambda_i \geq 0,\, i \leq N\}$. Thus, $\mathrm{Cone}(X)$ admits polyhedral representation and is therefore polyhedral, and thus closed, set.

**Exercise II.27.** Let $X \subset \mathbf{R}^n$ be a nonempty polyhedral set given by polyhedral representation:

$$X = \{x : \exists u : Ax + Bu \leq r\}$$

and let $K = \mathrm{Cone}(X)$ be the conic hull of $X$.

1. Is it true that $K$ is a closed cone?

   *Solution:* Wrong in general. As every conic hull, $K$ is a cone, but this cone not necessarily is closed. For example, when $X = \{[x_1; 1] : x_1 \in \mathbf{R}\} \subset \mathbf{R}^2$, $\mathrm{Cone}(X)$ is the union of the interior of the upper half-plane and of the origin, and this cone is not closed.

2. Prove that $\overline{K} := \mathrm{cl}\, K$ is a polyhedral cone and find polyhedral representation of $\overline{K}$.

   *Solution:* We claim that $K$ admits polyhedral representation

   $$\overline{K} = \{x : \exists \lambda, u : \lambda \geq 0, A + Bu - \lambda r \leq 0\}$$

   and is therefore a polyhedral cone. To justify the claim, denote the right hand side in the latter relation by $K^+$, so that $K^+$ is polyhedral (and thus closed) cone. To prove that $\overline{K} = K^+$ is the same as to check that, first, $K \subset K^+$ and, second, $K$ is dense in $K^+$.

   To justify the first claim, note that $x \in K$ is of the form $\sum_i \lambda_i x_i$ with $\lambda_i \geq 0$ and $x_i \in X$; the latter means that for properly selected $u_i$ it holds $Ax_i + Bu_i \leq r$. Consequently, $A[\lambda_i x_i] + B[\lambda_i u_i] - \lambda_i r_i \leq 0$. Summing up these vector inequalities, we get

   $$A \underbrace{\sum\nolimits_i \lambda_i x_i}_{x} + B \underbrace{\sum\nolimits_i \lambda_i u_i}_{=:u} + \underbrace{[\sum\nolimits_i \lambda_i]}_{=:\lambda} r \leq 0;$$

implying that $x \in K^+$ due to $\lambda = \sum_i \lambda_i \geq 0$.

To justify the second claim, let us fix $\bar{x} \in X$ ($X$ is nonempty!), so that $A\bar{x} + B\bar{u} - r \leq 0$ for some $\bar{u}$. Now let $x \in K^+$, and let us prove that $x \in \mathrm{cl}\, K$. Indeed, $x \in K^+$ means that $Ax + Bu - \lambda r \leq 0$ for some $u$ and some $\lambda \geq 0$, implying that for $\epsilon > 0$ one has

$$A\underbrace{[x + \epsilon\bar{x}]}_{=:x_\epsilon} + B[u + \epsilon\bar{u}] - \underbrace{[\lambda + \epsilon]}_{>0} r \leq 0.$$

Dividing both sides by $[\lambda + \epsilon]$, we see that $x_\epsilon = [\lambda + \epsilon]x^\epsilon$ with $x^\epsilon = [\lambda + \epsilon]^{-1}x_\epsilon \in X$ Thus, $x_\epsilon \in K = \mathrm{Cone}(X)$; since $x_\epsilon \to x$ as $\epsilon \to +0$, we conclude that $x \in \mathrm{cl}\, K$, as claimed.

3. Assume that $X$ is given by plain – no extra variables – polyhedral representation: $X = \{x : Ax \leq b\}$. Build plain polyhedral representation of $\overline{K} := \mathrm{cl}\,\mathrm{Cone}(X)$.

*Solution:* By the previous item,

$$\overline{K} = \{x : \exists \lambda : \lambda \geq 0, Ax - b\lambda \leq 0\},$$

and to get plain polyhedral representation of $K$, it suffices to subject the above polyhedral representation to one step of Fourier-Motzkin elimination. To this end, let us set $\overline{A} = \left[\begin{array}{c|c} & -1 \\ \hline A & -b \end{array}\right]$, so that $\overline{K} = \{x : \exists \lambda : \overline{A}[x; \lambda] \leq 0\}$. Denoting the transposes of the rows of $\overline{A}$ by $[\alpha_i; \beta_i]$ with $\alpha_i \in \mathbf{R}^n$ and $\beta_i \in \mathbf{R}$ and denoting by $I_0$, $I_+$, $I_-$ the sets of $i$'s with $\beta_i = 0$, $\beta_i > 0$, $\beta_i < 0$, respectively, we have

$$\begin{array}{rcl} \overline{K} & = & \left\{x : \exists \lambda : \overline{A}[x; \lambda] \leq 0\right\} \\ & = & \left\{x : \begin{array}{l} \alpha_i^\top x \leq 0 \,\forall i \in I_0 \\ [\beta_i^{-1}\alpha_i - \beta_j^{-1}\alpha_j]^\top x \leq 0 \;\forall (i \in I_+, j \in I_-) \end{array}\right\} \end{array}$$

and we end up with plain polyhedral representation of $\overline{K}$.

**Exercise II.28.** As we know, the extreme directions of the nonnegative orthant $\mathbf{R}^n_+ = \mathbf{R}_+ \times \mathbf{R}_+ \times \ldots \times \mathbf{R}_+$ are the vectors with single positive entry and remaining entries equal to 0. Prove the following generalization of this observation:

> Let $X_i \subset \mathbf{R}^{n_i}$, $1 \leq i \leq K$, be closed and pointed cones. The extreme directions of the direct product $X = X_1 \times \ldots \times X_K$ of these cones, if any, are the block-vectors $d = [d_1; \ldots; d_K]$ with $d_i \in \mathbf{R}^{n_i}$ of the following structure: all but one blocks in $d$ are zero, and the only nonzero block is an extreme direction of the corresponding factor $X_i$.

*Solution:* In one direction: if $d = [0; \ldots; 0; d_i; 0; \ldots; 0]$ with $d_i$ being extreme direction of $X_i$ and $d = d^1 + d^2$ with $d^1, d^2 \in X$, then $d$ is nonzero and $d^1_j = d^2_j = 0$ for $j \neq i$; indeed, for $j$ in question $d^1_j, d^2_j \in X_j$ and $d^1_j + d^2_j = d_j = 0$, and $X_j$ is pointed. From $d_i = d^1_i + d^2_i$ with $d_i$ being extreme direction of $X_i$ both $d^1_i$ and $d^2_i$ are nonnegative multiples of $d_i$ (indeed, $d^1_i$ and $d^2_i$ belong to $X_i$ and sum up to the extreme direction $d_i$ of $X_i$). Combining our observations, we conclude that $d^1$ and $d^2$ are nonnegative multiples of $d$, and we conclude that $d$ is an extreme direction of $X$. In the opposite direction: let $d = [d_1; \ldots; d_K]$ be an extreme direction of $X$ implying, in particular, that $d_i \in X_i$ for all $i$, and $d \neq 0$, so that $d$ has a nonzero block, say, $d_1$. Since $d = \underbrace{[d_1; 0; \ldots; 0]}_{\in X} + \underbrace{[0; d_2; d_3; \ldots; d_K]}_{\in X}$ and $d$ is extreme direction of $X$, the vector $[d_1; 0; \ldots; 0]$ must be nonnegative multiple of $d$; since $d_1 \neq 0$, $[d_1; 0; 0; \ldots; 0]$ in fact is a positive multiple of $d$, implying that $d_i = 0$, $i \geq 2$. It remains to verify that the only nonzero block, $d_1$, in $d$ is an extreme direction of $X_1$. Assuming the opposite, we can represent $d_1$ as $d^1_1 + d^2_1$ with vectors $d^\chi_1$, $\chi = 1, 2$, belonging to $X_1$ and not both being nonnegative multiples of $d_1$. But then $d = [d_1; 0; \ldots; 0] = d^1 + d^2$, with $d^\chi = [d^\chi_1; 0; \ldots; 0] \in X$, $\chi = 1, 2$, and at least one of $d^1, d^2$ not being a nonnegative multiple of $d$, which contradicts the fact that $d$ is an extreme direction of $X$. □

**Exercise II.29.** Describe all extreme rays of

1. positive semidefinite cone $\mathbf{S}^n_+$

2. Lorentz cone $\mathbf{L}^n$
3. Lorentz cone $\mathbf{L}^n$, $n \geq 2$, is the special case of the following construction: given a norm $\|\cdot\|$ on $\mathbf{R}^{n-1}$ ($n \geq 2$), we associate with it the set
$$\mathbf{K}^n_{\|\cdot\|} = \{[x;t] \in \mathbf{R}^n : t \geq \|x\|\},$$
which is a pointed nontrivial cone with a nonempty interior (why?); note that $\mathbf{L}^n = \mathbf{K}^n_{\|\cdot\|_2}$. Describe the extreme directions of $\mathbf{K}^n_{\|\cdot\|}$.

*Solution:*

1. Extreme rays of $\mathbf{S}^n_+$ are nonnegative multiples $\mathbf{R}_+ \times ee^\top$, $e \in \mathbf{R}^n \setminus \{0\}$, of positive semidefinite rank 1 matrices.
In one direction: When $e \in \mathbf{R}^n \setminus \{0\}$, in every representation $ee^\top = d^1 + d^2$ with $d^1 \succeq 0$, $d^2 \succeq 0$, for every $x$ orthogonal to $e$ we should have $0 = x^\top[ee^\top]x = \underbrace{x^\top d^1 x}_{\geq 0} + \underbrace{x^\top d^2 x}_{\geq 0}$, that is, $[\mathbf{R}\cdot e]^\perp$ is in the kernel of both $d^1$ and $d^2$, implying that the only eigenvector of $d^i$ with nonzero eigenvalue, if any, is proportional to $e$. In other words, eigenvalue decomposition of $d^i$ is $\lambda_i ee^\top$ with nonnegative $\lambda_i$ (since $\lambda_i$ is an eigenvalue of $d^i \succeq 0$). Thus, $d^i$, $i = 1,2,$, are nonnegative multiples of $ee^\top$, so that $ee^\top$ is extreme direction of $\mathbf{S}^n_+$. In the opposite direction: let $E \in \mathbf{S}^n_+$ and $E = \sum_i \lambda_i e_i e_i^\top$ be eigenvalue decomposition of $E$. When the number of nonzero eigenvalues $\lambda_i$ is $> 1$, say, $\lambda_1 > 0$ and $\lambda_2 > 0$, then $E = \lambda_1 e_1 e_1^\top + \sum_{i \geq 2} \lambda_i e_i e_i^\top$ is decomposition of $E$ into sum of two positive semidefinite matrices which are not proportional to $E$, that is, positive semidefinite matrix of rank $> 1$ is not an extreme direction of $\mathbf{S}^n_+$. Since an extreme direction should be nonzero and positive semidefinite, it must be of the form $ee^\top$ with nonzero vector $e$. □
2. The extreme directions of $\mathbf{L}^1 = \mathbf{R}_+$ are positive reals. When $n > 1$, the extreme directions of $\mathbf{L}^n$ are exactly positive multiples of vectors $[e;1]$ with $e \in \mathbf{R}^{n-1}$, $\|e\|_2 = 1$, see solution to item 3.
3. Denoting by $B = \{x \in \mathbf{R}^{n-1} : \|x\| \leq 1\}$ the unit ball of norm $\|\cdot\|$, the extreme directions of $\mathbf{K}^n_{\|\cdot\|}$ are positive multiples of vectors $[x;1]$ with $x \in \mathrm{Ext}(B)$. Indeed, the set
$$Y := \{[x;1] \in \mathbf{K}^n_{\|\cdot\|}\} = B \times \{1\}$$
clearly is a base of the cone $\mathbf{K}^n_{\|\cdot\|}$, see Definition II.8.32. By Fact II.8.33.(iv), the extreme directions of $\mathbf{K}^n_{\|\cdot\|}$ are positive multiples of the vectors from $\mathrm{Ext}(Y)$, and
$$\mathrm{Ext}(Y) = \mathrm{Ext}(B \times \{1\}) = \mathrm{Ext}(B) \times \{1\} = \{[x;1] : x \in \mathrm{Ext}(B)\},$$
where the second equality is due to Exercise II.10. □

## Recessive cone

**Exercise II.30.** Let $M$ be a convex set, and let $\bar{x}$ and $h$ be such that $R_{\bar{x}} := \{\bar{x} + th : t \geq 0\} \subset M$.

1. Is it always true that whenever $x \in M$, the set $R_x = \{x + th, t \geq 0\}$ is contained in $M$ ?

*Solution:* The answer is no, example being $M = \{[x_1;x_2] \in \mathbf{R}^2 : x_1 \geq 0, x_2 > 0\} \cup \{[0;0]\}$. This set clearly is convex and contains the ray $\{[x_1;1] : x_1 \geq 0\}$ (that is, the ray $R_{\bar{x}}$ corresponding to $\bar{x} = [0;1]$ and $h = [1;0]$), but does not contain the parallel ray $R_{[0;0]} = \{[x_1;0] : x_1 \geq 0\}$ emanating from $[0;0] \in M$. □

2. Let $h$ be a recessive direction of $\overline{M} = \mathrm{cl}\, M$, and let $\bar{x}$ be a point from the relative interior of $M$. Is it always true that the set $R_{\bar{x}} = \{\bar{x} + th : t \geq 0\}$ is contained in $M$ ?

*Solution:* The answer is yes. Indeed, by Lemma I.1.30 the ray $R = \{\bar{x} + th : t \geq 0\}$ is contained in $\overline{M}$, and since every point $x = \bar{x} + th$ on this ray is of the form $\frac{1}{2}\bar{x} + \frac{1}{2}x'$ with $x' \in R \subset \mathrm{cl}\, M$ (you can take $x' = \bar{x} + 2th$), $x \in M$ by Lemma I.1.30.

**Exercise II.31.** Let $M \subset \mathbf{R}^n$ be a cone, not necessary closed; recall that pointedness of a cone $M$ means that the only vector $x$ such that $x \in M$ and $-x \in M$ is the zero vector. Which of the following statements are always true:

1. $M$ is pointed if an only if the only representation of 0 as the sum of $k \geq 1$ vectors $x_i \in M$ is the representation with $x_i = 0$, $i \leq k$.

*Solution:* This is true. Indeed, if $M$ is not pointed, so that $\pm x \in M$ for some $x \neq 0$, then setting $k = 2$, $x_1 = x$, $x_2 = -x$, we get a representation of 0 as the sum of two nonzero vectors from $M$. On the other hand, when $M$ is pointed and $0 = x_1 + \ldots + x_k$ with $x_i \in M$, then either $k = 1$ and $x_1 = 0$, or $k > 1$, and then for every $i \leq k$ we have $0 = x_i + \underbrace{\sum_{j \neq i} x_j}_{\in M}$ implying that $\pm x_i \in M$, whence, by pointedness, $x_i = 0$, $i \leq k$. □

2. $M$ is pointed if and only if $M$ does not contain straight lines (one-dimensional affine planes) passing through the origin.

*Solution:* True. In one direction: if $M$ contains straight line passing through the origin, that is, the set $\{th : t \in \mathbf{R}\}$ with some $h \neq 0$ is contained in $M$, then $\pm h \in M$ and $h \neq 0$, contradicting pointedness of $M$. In the opposite direction: if $M$ is not pointed, that is, $\pm h \in M$ for some $h \neq 0$, then $M$, being conic, contains the straight line $\{th : t \in \mathbf{R}\}$ passing through the origin. □

3. $M$ is pointed if and only if $M$ does not contain straight lines.

*Solution:* Wrong – take $M = \{[x_1, x_2] \in \mathbf{R}^2 : x_2 > 0\} \cup \{[0; 0]\}$. This cone is pointed (since all nonzero vectors from $M$ have the second coordinate positive) and contains the line $\{[x_1; 1] : x_1 \in \mathbf{R}\}$. □

4. Assuming $M$ closed, $M$ is pointed if and only if $M$ does not contain straight lines.

*Solution:* True. By Lemma II.8.8, if a closed convex set contains a line, it contains all parallel lines intersecting the set, so that a closed cone $M$ contains lines if and only if it contains lines passing through the origin, and it remains to use item 2. □

5. $M$ is pointed cone if and only if the closure of $M$ is so.

*Solution:* Wrong, the counter-example being the pointed cone $M = \{[x_1; x_2] \in \mathbf{R}^2 : x_2 > 0\} \cup \{[0; 0]\}$. □

6. The closure of $M$ is a pointed cone if and only if $M$ does not contain straight lines.

*Solution:* True. If $M$ contains a line, then this line is contained in the closed cone $\mathrm{cl}\, M$, so that $\mathrm{cl}\, M$ is not pointed by item 4. Vice versa, if $\mathrm{cl}\, M$ is not pointed, it contains a line $\{th : t \in \mathbf{R}\}$ $(h \neq 0)$ passing through the origin by item 2, and therefore by the result stated in Exercise II.30 $M$ contains all lines of the form $\{x + th : t \in \mathbf{R}\}$ with $x \in \mathrm{rint}\, M$ $[\neq \emptyset]$. □

**Exercise II.32.** Literal interpretation of the words "polyhedral cone" is: a polyhedral set $\{x : Ax \leq b\}$ which is a cone. An immediate example is the solution set $\{x : Ax \leq 0\}$ of *homogeneous* system of linear inequalities. Prove that this example is generic: whenever a polyhedral set $K = \{x : Ax \leq b\}$ is a cone, one has $K = \{x : Ax \leq 0\}$.

*Solution:* One way to prove the claim is to note that when the set $K = \{x : Ax \leq b\}$ is a cone, this (clearly closed) set, as every closed cone, coincides with its recessive cone: $K = \mathrm{Rec}(K)$, and Fact II.8.15 states that for a nonempty polyhedral set $M = \{x : Ax \leq b\}$ one has $\mathrm{Rec}(M) = \{x : Ax \leq 0\}$.

A "bare hands" proof of the claim in question can be found in solution to Exercise I.4.

**Exercise II.33.** Prove the following modification of Proposition II.8.18:

> (!) *Let $X \subset \mathbf{R}^N$ be a nonempty closed convex set such that $X \subset V + \mathrm{Rec}(X)$ for some bounded and closed set $V$, let $x \mapsto \mathcal{A}(x) = Ax + b : \mathbf{R}^N \to \mathbf{R}^n$ be an affine mapping, and let $Y = \mathcal{A}(X) := \{y : \exists x \in X : y = \mathcal{A}(x)\}$ be the image of $X$ under this mapping. Let also*
>
> $$K = \{h \in \mathbf{R}^n : \exists g \in \mathrm{Rec}(X) : h = Ag\}.$$
>
> *Then the recessive cone of the closure $\overline{Y}$ of $Y$ is the closure $\overline{K}$ of $K$. In particular, when $K$ is closed (as definitely is the case when $\mathrm{Rec}(X)$ is polyhedral), it holds $\mathrm{Rec}(\overline{Y}) = K$.*

*Solution:* If $y \in Y$ and $h \in K$, so that $y = \mathcal{A}(x)$ and $h = Ag$ for some $x \in X$ and $g \in \mathrm{Rec}(X)$, then $x + tg \in X$ for all $t \geq 0$, so that $y + th = \mathcal{A}(x + tg) \in Y \subset \overline{Y}$ whenever $t \geq 0$. Thus, $h$ is a recessive direction of $\overline{Y}$, so that $K \subset \mathrm{Rec}(\overline{Y})$, and since the cone $\mathrm{Rec}(\overline{Y})$ is closed, $\overline{K}$ belongs to $\mathrm{Rec}(\overline{Y})$ along with $K$.

Vice versa, under the premise of Proposition, let $h \in \mathrm{Rec}(\overline{Y})$; we want to prove that $h \in \overline{K}$. Indeed, selecting somehow $y \in Y$, we have $y + ih \in \overline{Y}$, $i = 1, 2, ...$ Next, from $X \subset V + \mathrm{Rec}(X)$ is follows that $Y \subset \widehat{Y} := \mathcal{A}(V) + A\mathrm{Rec}(X) = \mathcal{A}(V) + K$, and therefore $\overline{Y} \subset \mathrm{cl}(\mathcal{A}(V) + K) = \mathcal{A}(V) + \overline{K}$, where the concluding equality is due to the fact that $\mathcal{A}(V)$ is a compact set along with $V$ [7]. Thus, $y + ih \in \overline{Y} = \mathcal{A}(V) + \overline{K}$, implying that for every $i$ there exists $\delta_i \in \mathbf{R}^n$, $v_i \in V$ and $g_i \in \mathrm{Rec}(X)$ such that $y + ih = \mathcal{A}(v_i) + Ag_i + \delta_i$ and $\delta_i \to 0$ as $i \to \infty$. Setting $h_i = i^{-1}Ag_i$, we have $h = h_i + i^{-1}[\mathcal{A}(v_i) + \delta_i - y]$, and the second term in the right hand side of this equality tends to 0 as $i \to \infty$ due to the boundedness of $V$ and of the sequence $\{\delta_i\}$. We conclude that $h = \lim_{i\to\infty} h_i$ with $h_i \in K$ for all $i$ (due to $g_i \in \mathrm{Rec}(X)$), so that $h \in \overline{K}$. Recalling what $h$ is, we conclude that $\mathrm{Rec}(\overline{Y}) \subset \overline{K}$. The opposite inclusion has already been verified, and we arrive at $\mathrm{Rec}(\overline{Y}) = \overline{K}$. □

**Exercise II.34.** [follow-up to Exercise II.33]

1. Let $K_1 \subset \mathbf{R}^n, K_2 \subset \mathbf{R}^n$ be closed cones, and let $K = K_1 + K_2$.
   - Is it always true that $K$ is a cone?

     *Solution:* $K$ clearly is a cone.
   - Is it always true that $K$ is closed?

     *Solution:* The answer is negative, as is shown by the following example: $K_1 = \{[x; y; z] \in \mathbf{R}^3 : y, z \geq 0, yz \geq x^2\}$ (this, up to one-to-one linear substitution of variables, is the 3D Lorentz cone), $K_2 = \{[x; y; z] : x = y = 0, z \leq 0\}$ (just a ray). In this case $K$ contains all lines $\ell_a = \{[x; y; z] : y = a, z = 0\}$ with $a > 0$; indeed, given $a > 0$ and $x$, the vector $[x; a; x^2/a]$ belongs to $K_1$, and the vector $[0; 0; -x^2/a]$ belongs to $K_2$, so that the sum $[x; a; 0]$ of these vectors belongs to $K$, implying that $\ell_a \subset K$. On the other hand, the only vector of the form $[x; 0; 0]$ belonging to $K$ clearly is the sum of some vector from $K_1$ with the $y$-coordinate equal to 0 and a vector from $K_2$; the only option for the first vector is to be of the form $[0; 0; z]$ with $z \geq 0$, and in this case, $x$ must be zero. We see that $K$ contains all lines $\ell_a$ with $a > 0$, but does not contain the line $\ell_0 \subset \mathrm{cl} \cup_{a>0} \ell_a$.
   - Let $K_2$ be polyhedral. Is it always true that $K$ is closed?

     *Solution:* The answer is negative, as is shown by the example of the previous item, where $K_2$ is a ray.
   - Let both $K_1$ and $K_2$ be polyhedral. Is it always true that $K$ is closed?

     *Solution:* The answer is positive: by evident reasons, $K$ admits polyhedral representation and therefore is polyhedral.
2. Let $X_i$, $i = 1, ..., I$, be closed convex sets in $\mathbf{R}^n$ with nonempty intersection. Is it true that $\cap_i \mathrm{Rec}(X_i) = \mathrm{Rec}(\cap_i X_i)$?

   *Solution:* The answer is positive: selecting $x \in \cap_i X_i$, we have $h \in \mathrm{Rec}(\cap_i X_i)$ iff $x + th \in \cap_i X_i$ for all $t \geq 0$, or, which is the same, iff $h \in \mathrm{Rec}(X_i)$ for every $i$.
3. Let $X_1$, $X_2$ be nonempty closed convex sets in $\mathbf{R}^n$, let $K_1 = \mathrm{Rec}(X_1)$, $K_2 = \mathrm{Rec}(X_2)$, $\overline{X} = \mathrm{cl}(X_1 + X_2)$, $\overline{K} = \mathrm{cl}(K_1 + K_2)$.
   - Is it always true that $\overline{K} \subset \mathrm{Rec}(\overline{X})$ ?

[7] We have used a nearly evident statement (prove it!): *if $A, B$ are nonempty sets in $\mathbf{R}^m$ and $A$ is bounded, then* $\mathrm{cl}(A + B) = \mathrm{cl}(A) + \mathrm{cl}(B)$.

*Solution:* The answer is positive: selecting $x_i \in X_i$ and $h_i \in \mathrm{Rec}(X_i)$, $i = 1, 2$, we have $x_1 + x_2 + t(h_1 + h_2) \in X_1 + X_2$ for all $t \geq 0$, implying that $h_1 + h_2 \in \mathrm{Rec}(\overline{X})$. Thus, the cone $K_1 + K_2$ belongs to the cone $\mathrm{Rec}(\overline{X})$, and since the latter cone is closed, $\overline{K}$ belongs to this cone as well.

- Is is always true that $\overline{K} = \mathrm{Rec}(\overline{X})$ ?

  *Solution:* The answer is negative: take $X_1 = \{[x; t] : x^2 \leq t\}$, $X_2 = -X_1 = \{[x; t] : x^2 \leq -t\}$. Then $K_1 = \{[0; t] : t \geq 0\}$, $K_2 = \{[0; t] : t \leq 0\}$, so that $\overline{K} = \{[0; t], t \in \mathbf{R}\}$. At the same time, we clearly have $X_1 + X_2 = \mathbf{R}^2$, that is, $\mathrm{Rec}(\overline{X}) = \mathbf{R}^2$.

- Assume that $X_i \subset V_i + K_i$ for properly selected closed and bounded set $V_i$, $i = 1, 2$, Is it true that $\overline{K} = \mathrm{Rec}(\overline{X})$ ?

  *Solution:* The answer is positive. Indeed, let $Y = X_1 \times X_2$, $L = K_1 \times K_2$, $V = V_1 \times V_2$. Then clearly $Y$ is a nonempty closed convex set, $L = \mathrm{Rec}(Y)$, and $V$ is a bounded and closed set such that $Y \subset V + L$. Setting $\mathcal{A}(x_1, x_2) = x_1 + x_2$, we get a linear mapping acting from $\mathbf{R}^n \times \mathbf{R}^n$ to $\mathbf{R}^n$ such that $\overline{X} = \mathrm{cl}\,\mathcal{A}(Y)$ and $\overline{K} = \mathrm{cl}\,\mathcal{A}(L)$, so that $\overline{K} = \mathrm{Rec}(\overline{X})$ by the result of Exercise II.33.

**Exercise II.35.** Let $f(x) = x^\top Cx - c^\top x + \sigma$ be quadratic form with $C \succeq 0$. By Exercise I.15, the set $E = \{x : f(x) \leq 0\}$ is convex (and of course closed). Assuming $E \neq \varnothing$, describe $\mathrm{Rec}(E)$.

*Solution:* Let $\bar{x} \in E$. Ray $\{\bar{x} + th : t \geq 0\}$ is contained in $E$ if and only if

$$\forall t \geq 0 : t^2 \underbrace{h^\top Ch}_{\geq 0} + 2t\bar{x}^\top Ch - tc^\top h \leq \underbrace{-[\bar{x}^\top C\bar{x} - c^\top \bar{x} + \sigma]}_{\geq 0},$$

which is possible if and only if $h^\top Ch = 0$ and $c^\top h \geq 0$. Recalling that for $C \succeq 0$ relation $h^\top Ch = 0$ is equivalent to $h \in \mathrm{Ker}C$, we get

$$\mathrm{Rec}(E) = \{h \in \mathrm{Ker}C : c^\top h \geq 0\}.$$

## Around majorization

**Exercise II.36.** Let $x \in \mathbf{R}^m$, let $X[x]$ be the convex hull of all permutations of $x$, and let $X_+[x]$ be the set of all vectors $x'$ dominated by a vector form $X[x]$:

$$X_+[x] = \{y \mid \exists z \in X[x] : y \leq z\}.$$

1) Prove that $X_+[x]$ is a polyhedral set.

2) Prove the following characterization of $X_+[x]$: $X_+[x]$ is exactly the set of solutions of the system of inequalities $s_j(y) \leq s_j(x)$, $j = 1, \dots, m$, in variables $y$, where, as always $s_j(z)$ is the sum of the $j$ largest entries in vector $z$.

*Solution:* 1) The set $X_+[x]$ is the sum of the polyhedral set $X[x]$ and the polyhedral cone $-\mathbf{R}^m_+$ and therefore admits immediate polyhedral representation: denoting by $\Sigma$ the set of all $m!$ permutation matrices of size $m \times m$, we have

$$X_+[x] = \{y : \exists \{\lambda_\sigma, \sigma \in \Sigma\}, z \in \mathbf{R}^m : \lambda \geq 0, \sum_\sigma \lambda_\sigma = 1, z \geq 0, y = \sum_{\sigma \in \Sigma} \lambda_\sigma [\sigma x] - z\}$$

and is therefore polyhedral. □

2) To justify the claim, let us fix $x$, and let $X^+[x]$ be the set of all solutions to the system of constraints $s_j(y) \leq s_j(x)$, $1 \leq j \leq m$. We want to prove that $X^+[x] = X_+[x]$. First, if $y \in X_+[x]$, then $y \leq \overline{y}$ for some $\overline{y} \in X[x]$. By Majorization Principle, we have $s_j(\overline{y}) \leq s_j(x)$, $j \leq m$ (in fact, the last of these inequalities is equality, but this does not matter now). And since $y \leq \overline{y}$ and $s_j(z)$ is monotonically nondecreasing in $z$, we have $s_j(y) \leq s_j(\overline{y}) \leq s_j(x)$, $j \leq m$, so that $y \in X^+[x]$. We conclude that $X_+[x] \subset X^+[x]$. To prove the inverse inclusion, let $y \in X^+[x]$, that is, $s_j(y) \leq s_j(x)$, $j \leq m$. Setting $\Delta = s_m(x) - s_m(y)$, we get $\Delta \geq 0$. Keeping all but the smallest entry in $x$ intact and decreasing the

smallest entry by $\Delta$, we get a vector $\overline{x}$ such that $s_j(x) = s_j(\overline{x})$ for $j < m$ and $s_m(\overline{x}) = s_m(y)$. Thus, $s_j(y) \leq s_j(\overline{x})$ for all $j$, the inequality being equality when $j = m$. By Majorization Principle, $y = D\overline{x}$ for some doubly stochastic matrix $D$, and since by construction $\overline{x} \leq x$, we have $D\overline{x} \leq Dx$, whence $y \leq Dx$. Since $Dx$, by Birkhoff theorem, belongs to $X[x]$, we conclude that $y$ is dominated by some point from $X[x]$, that is, $y \in X_+[x]$. Thus, $X^+[x] \subset X_+[x]$. □

## Around polars

**Exercise II.37.** Justify the last three claims in Example II.8.11.

*Solution:* 5: We have $\sup_{z\in DX} y^\top z = \sup_{x\in X} y^\top Dx = \sup_{x\in X}[D^\top y]^\top x$. Thus, $y \in \mathrm{Polar}\,(DX)$ if and only if $D^\top y \in \mathrm{Polar}\,(X)$. □
6: We have $E = \{x = C^{-1/2}u : u^\top u \leq 1\}$, whence $\max\limits_{x\in E} y^\top x = \max\limits_{u:u^\top u\leq 1} [C^{-1/2}y]^\top u = \|C^{-1/2}y\|_2$, Thus, $\mathrm{Polar}\,(E) = \{y : \|C^{-1/2}y\|_2 \leq 1\} = \{y : [C^{-1/2}y]^\top[C^{-1/2}y] \leq 1\} = \{y : y^\top C^{-1}y \leq 1\}$. □
7: This is evident.

**Exercise II.38.** [more on polars]

1. Recall that for $U \subset \mathbf{R}^n$, $\mathrm{Vol}(U)$ stands for the ratio of the $n$-dimensional volume of $U$ and the volume of the $n$-dimensional unit Euclidean ball. Check that for a centered at the origin ellipsoid $E = \{x : x^\top Cx \leq 1\}$ $(C \succ 0)$ we have $\mathrm{Vol}(E)\mathrm{Vol}(\mathrm{Polar}\,(E)) = 1$.
2. Let $C \succ 0$ and let ellipsoid $E = \{x : (x-c)^\top C(x-c) \leq 1\}$ contain the origin. Compute $\mathrm{Polar}\,(E)$.
3. Let $X_k$, $k \leq K$, be closed convex sets in $\mathbf{R}^n$ containing the origin. Prove that

$$\begin{array}{rcll}\mathrm{Polar}\,(\mathrm{Conv}(\cup_k X_k)) & = & \cap_k \mathrm{Polar}\,(X_k) & (a)\\ \mathrm{Polar}\,(\cap_k X_k) & = & \mathrm{cl}\,\mathrm{Conv}(\cup_k \mathrm{Polar}\,(X_k)) & (b)\end{array}$$

*Solution:* 1: By Example II.8.11.4, $\mathrm{Polar}\,(E) = \{x : x^\top C^{-1}x\}$ , so that by the results of Exercise I.14 one has $\mathrm{Vol}(E) = \mathrm{Det}^{-1/2}(C)$, $\mathrm{Vol}(\mathrm{Polar}\,(E) = \mathrm{Det}^{-1/2}(C^{-1}) = \mathrm{Det}^{1/2}(C)$.
2: We have

$$\begin{array}{l}\mathrm{Polar}\,(E) = \{y : \max_{x=c+C^{-1/2}u:u^\top u\leq 1} y^\top x \leq 1\} = \{y : c^\top y + \max_{u:\|u\|_2\leq 1} u^\top [C^{-1/2}y] \leq 1\}\\ = \{y : \sqrt{yC^{-1}y} \leq 1 - c^\top y\} \subseteq Q := \{y : y^\top C^{-1}y - [1 - c^\top y]^2 \leq 0\}\end{array}$$

Let us prove that the $\subseteq$ above is in fact equality. To this end note that $Q$ is a sublevel set of inhomogeneous quadratic form with the matrix

$$\Theta := C^{-1} - cc^\top = C^{-1/2}[I - dd^\top]C^{-1/2},$$

where $d, = C^{1/2}c$, so that $d^\top d = c^\top Cc \leq 1$ due to $0 \in E$. We conclude that $\Theta \succeq 0$, implying that $Q$ is convex (Exercise I.15). Now, to prove that the $\subseteq$ in question is in fact equality is the same as to prove that the linear function $1 - c^\top y$ is nonnegative everywhere on $Q$. Assuming that the latter is not the case and observing that $0 \in Q$, among the values taken on $Q$ by the linear function in question there are both positive and negative, and since $Q$ is convex, there should be $y \in Q$ with $c^\top y = 1$, and the latter clearly is forbidden by the definition of $Q$.
Thus, the polar of $E$ is

$$Q = \{y : y^\top \Theta y + 2c^\top y \leq 1\}.$$

Geometrically, this is
— either ellipsoid – this is the case when $\Theta \succ 0$, or, which is the same, $0 \in \mathrm{int}\,E$,
— or hyperparaboloid/elliptic cylinder – the set which in coordinates $t = Dx$ with properly selected nonsingular $D$ is given by $\gamma t_1 \geq \alpha + \sum_{i\geq 2}[t_i - \beta_i]^2$ – this is what happens when $0 \in \mathrm{bd}\,E$.
3: A linear form does not exceed a real $a$ on the convex hull of the union of $K$ nonempty convex sets if and only if it does not exceed $a$ on every one of these sets, resulting in $(a)$. Setting $Y_k = \mathrm{Polar}\,(X_k)$, $k \leq K$, so that $X_k = \mathrm{Polar}\,(Y_k)$ by Proposition II.8.37 (recall that $X_k$ are closed, convex, and contain the origin)

and applying $(a)$ to the sets $Y_k$ in the role of $X_k$, we get $\mathrm{Polar}\,(\mathrm{Conv}(\cup_k Y_k)) = \cap_k \mathrm{Polar}\,(Y_k) = \cap_k X_k$, whence also $\mathrm{Polar}\,(\mathrm{cl}\,\mathrm{Conv}(\cup_k Y_k)) = \cap_k X_k$. Since the set $\mathrm{cl}\,\mathrm{Conv}(\cup_k Y_k)$ is closed, convex, and contains the origin, it is the polar of its polar (Proposition II.8.37), that is, $\mathrm{cl}\,\mathrm{Conv}(\cup_k Y_k) = \mathrm{Polar}\,(\cap_k X_k)$. Recalling what $Y_k$ are, we arrive at $(b)$. □

**Exercise II.39.** Let $X \subset \mathbf{R}^n$ be a cone given by polyhedral representation

$$X = \{x \in \mathbf{R}^n : \exists u : Ax + Bu \le r\}$$

Is the dual to $X$ cone $X_*$ polyhedral? If yes, build a polyhedral representation of $X_*$.

*Solution:* The fact that the cone dual to a polyhedral cone is polyhedral as well was explained in Remark II.8.22 and Exercise I.4. An independent reasoning (which as a byproduct yields polyhedral representation of $X_*$ and as such can be considered as an addition to the calculus of polyhedral representations, see section 3.3), is as follows. We have

$$\begin{array}{rcl} y \in X_* & \Longleftrightarrow & y^\top x \ge 0\,\forall x \in X \Longleftrightarrow 0 \le \min_{x,u}\left\{y^\top x : Ax + Bu \le r\right\} \\ & \Longleftrightarrow & 0 \le \max_\lambda \left\{-r^\top \lambda : \lambda \ge 0, A^\top \lambda + y = 0, B^\top \lambda = 0\right\} \text{ [LP Duality]} \\ & \Longleftrightarrow & \exists \lambda : r^\top \lambda \le 0, \lambda \ge 0, A^\top \lambda + y = 0, B^\top \lambda = 0, \end{array}$$

and we end up with polyhedral representation of $X_*$.

**Exercise II.40.**

1. Let $X \subset \mathbf{R}^n$ be a nonempty polyhedral set given by polyhedral representation

$$X = \{x \in \mathbf{R}^n : \exists u : Ax + Bu \le r\}$$

Is the polar $\mathrm{Polar}\,(X)$ of $X$ polyhedral? If yes, point out a polyhedral representation of $\mathrm{Polar}\,(X)$. For non-polyhedral extension, see Exercise IV.36.

*Solution:* We have

$$\begin{array}{rcl} \mathrm{Polar}\,(X) & = & \{y : \mathrm{Opt}(P) := \max_{x,u}\{y^\top x : Ax + Bu \le r\} \le 1\} \\ & = & \{y : \mathrm{Opt}(D) := \min_\lambda \left\{r^\top \lambda : \lambda \ge 0, A^\top \lambda = y, B^\top \lambda = 0\right\} \le 1\} \\ & & \text{[by LP Duality; note that } (P) \text{ is feasible due to } X \neq \varnothing] \\ & = & \{y : \exists \lambda : r^\top \lambda \le 1, \lambda \ge 0, y = A^\top \lambda, B^\top \lambda = 0\}, \\ & & \text{[since by the above, the dual problem is solvable when } y \in \mathrm{Polar}\,(X)] \end{array}$$

and we end up with polyhedral representation of $\mathrm{Polar}\,(X)$, implying polyhedrality of the polar.

2. Compute the polars of

   1. probabilistic simplex $\Delta = \{x \in \mathbf{R}^n : x \ge 0, \sum_i x_i = 1\}$

      *Solution:* $\mathrm{Polar}\,(\Delta) = \{y \in \mathbf{R}^n : y \le [1; ...; 1]\}$

   2. convex hull of nonempty finite set of points $a_1, ..., a_N$ from $\mathbf{R}^n$

      *Solution:* $\mathrm{Polar}\,(\mathrm{Conv}\{a_1, ..., a_N\}) = \{y : a_i^\top y \le 1, i \le N\}$

   3. the set $\{x \in \mathbf{R}^n : x \le b\}$

      *Solution:* $\mathrm{Polar}\,(\{x : x \le b\}) = \{y : y \ge 0, y^\top b \le 1\}$

## Miscellaneous exercises

**Exercise II.41.** Let $X = \{x \in \mathbf{R}^n : Ax \leq b\}$ be a nonempty polyhedral set.

1. Prove that $X$ is bounded if and only if every one of the vectors $\pm e_i$, ($e_i$, $1 \leq i \leq n$, are the standard basic orth) can be represented as conic combination of columns of $A^\top$.

   *Solution:* A nonempty polyhedral set $\{x : Ax \leq b\} \subset \mathbf{R}^n$ is bounded if and only if the optimal values in the $2n$ optimization problems $\max_x\{\pm e_i^\top x : Ax \leq b\}$ are finite, and this, by LP Duality Theorem, boils down to feasibility of their duals, the latter being exactly the possibility to represent $\pm e_i$ as conic combination of the columns of $A^\top$.

2. Certify the correct statements in the following list:

   - The polyhedral set $X = \{x \in \mathbf{R}^3 : x \geq [1/3; 1/3; 1/3], \sum_{i=1}^3 x_i \leq 1\}$ is bounded.

     *Solution:* It suffices to verify that every one of the vectors $\pm e_i$, $i = 1, ..., 3$, is a conic combination of the columns of $A^\top$. The vectors $-e_i$ are among the columns of $A^\top$; to get $e_1$, sum up all columns of $A^\top$ but the first one, and similarly for $e_2$ and $e_3$.

   - The polyhedral set $X = \{x \in \mathbf{R}^3 : x_1 \geq 1/3, x_2 \geq 1/3, \sum_{i=1}^3 x_i \leq 1\}$ is unbounded.

     *Solution:* By Lemma II.8.8 a polyhedral set is unbounded if and only if it is nonempty and its recessive cone is nontrivial. For the set in question, certificate of nonemptiness is, e.g., $x = [1/3; 1/3; 1/3]$, and a nonzero vector in $\mathrm{Rec}(X) = \{x \in \mathbf{R}^3 : x_1 \geq 0, x_2 \geq 0, x_1 + x_2 + x_3 \leq 0\}$ is, e.g., $x = [0; 0; -1]$.

**Exercise II.42.** Prove the easy part of Theorem II.9.9, specifically, that every $n \times n$ permutation matrix is an extreme point of the polytope $\Pi_n$ of $n \times n$ doubly stochastic matrices.

*Solution:* Let $\overline{\Pi}_n$ be the set of all $n \times n$ matrices with entries from $[0, 1]$. As we know, the extreme points of $\overline{\Pi}_n$ are exactly the $n \times n$ matrices with zero and one entries. In view of this, the claim to be proved is readily given by the following statement (evident due to geometric characterization of extreme points): *If $X \subset Y$ are convex sets, then every extreme point of $Y$ which happens to belong to $X$ is an extreme point of $X$.*

**Exercise II.43.** [robust LP] Consider *uncertain* Linear Programming problem – a family

$$\left\{\min_{x \in \mathbf{R}^n} \{c^\top x : [A + \sum\nolimits_{\nu=1}^N \zeta_\nu \Delta_\nu]x \leq b + \sum\nolimits_{\nu=1}^N \zeta_\nu \delta_\nu\} : \zeta \in \mathcal{Z}\right\} \tag{11.3}$$

of LP instances of common sizes ($n$ variables, $m$ constraints). The associated story is as follows: we want to solve an LP program with the data not known exactly when the problem is being solved; what we know at this time, is that the "true problem" belongs to the parametric family given, according to (11.3), by the "nominal data" $c, A, b$, "basic perturbations $\Delta_\nu, \delta_\nu$" and the *perturbation set* $\mathcal{Z}$ through which run the data perturbations $\zeta$ specifying particular instances in the family. In this situation (quite typical for real life applications of LP, where partial data uncertainty is a rule rather than an exception), one way to "immunize" decisions against data uncertainty is to look for *robust solutions* – those remaining feasible for all perturbations of the data from the perturbation set – by solving the *Robust Counterpart* (RC) of our uncertain problem – the optimization problem

$$\min_x \left\{c^\top x : [A + \sum\nolimits_{\nu=1}^N \zeta_\nu \Delta_\nu]x \leq b + \sum\nolimits_{\nu=1}^N \zeta_\nu \delta_\nu \;\forall (\zeta \in \mathcal{Z})\right\} \tag{RC}$$

(RC) is *not* an LP program – it has finitely many decision variables and infinite (when $\mathcal{Z}$ is "massive") system of linear constraints on these variables. Optimization problems of this type are called *semi-infinite* and are, in general, difficult to solve. However, the RC of an uncertain LP is easy, provided that $\mathcal{Z}$ is a "computation-friendly" set, for example, nonempty set given by polyhedral representation:

$$\mathcal{Z} = \{\zeta : \exists u : P\zeta + Qu \leq r\} \tag{11.4}$$

Now goes the exercise *per se*:
Use LP duality to reformulate (RC), (11.4) as an explicit LP program.

*Solution:* The constraints of (RC) are of the form

$$\max_{\zeta\in\mathcal{Z}}\sum\nolimits_{\nu}\zeta_\nu\,[\Delta_\nu x-\delta_\nu]_j\le[b-Ax]_j,\ 1\le j\le m,$$

or, which is the same,

$$\max_{\zeta,u}\left\{\sum\nolimits_{\nu}[\Delta_\nu x-\delta_\nu]_j\,\zeta_\nu : P\zeta+Qu\le r\right\}\le[b-Ax]_j,\ 1\le j\le m.$$

Applying LP Duality, the constraints can be rewritten as

$$\min_{\lambda^j}\left\{r^\top\lambda^j : \begin{array}{l}[P^\top\lambda^j]_\nu=\Delta_\nu x-\delta_\nu,\ 1\le\nu\le N\\ Q^\top\lambda^j=0,\ \lambda^j\ge0\end{array}\right\}\le[b-Ax]_j,\ 1\le j\le m. \tag{$*$}$$

We see that $x$ is robust feasible (i.e., feasible for (RC)) if and only if $x$ can be augmented by properly selected $\lambda^1,...,\lambda^m$ to satisfy system $(*)$ of linear constraints on $x$ and $\lambda^j$'s. As a result, (RC) is equivalent to the explicit LP program

$$\min_{x,\lambda^1,...,\lambda^m}\left\{c^\top x : \left.\begin{array}{l}\Delta_\nu x-[P^\top\lambda^j]_\nu=\delta_\nu, 1\le\nu\le N\\ Q^\top\lambda^j=0,\ \lambda^j\ge0\\ r^\top\lambda^j+[Ax]_j\le b_j\end{array}\right\},\ j=1,...,m\right\}$$

**Exercise II.44.** Consider scalar linear constraint

$$a^\top x\le b \tag{1}$$

with uncertain data $a\in\mathbf{R}^n$ ($b$ is certain) varying in the set

$$\mathcal{U}=\{a:|a_i-a_i^*|/\delta_i\le1,1\le i\le n,\sum\nolimits_{i=1}^n|a_i-a_i^*|/\delta_i\le k\} \tag{2}$$

where $a_i^*$ are given "nominal data," $\delta_i>0$ are given quantities, and $k\le n$ is an integer (in literature, this is called "budgeted uncertainty"). Rewrite the Robust Counterpart

$$a^\top x\le b\,\forall a\in\mathcal{U} \tag{RC}$$

in a tractable LO form (that is, write down an explicit system $(S)$ of linear inequalities in variables $x$ and additional variables such that $x$ satisfies (RC) if and only if $x$ can be extended to a feasible solution of $(S)$).

*Solution:* Let $D$ be diagonal $n\times n$ matrix with diagonal entries $\delta_i$, and let $a_i-a_i^*=\delta_i\epsilon_i$, so that

$$\begin{array}{rcl}\mathcal{U}&=&\{a=a^*+D\epsilon:-1\le\epsilon_i\le1\,\forall i,\sum\nolimits_i|\epsilon_i|\le k\}\\ &=&\{a=a^*+D\epsilon:-u\le\epsilon\le u,u_i\le1\,\forall i,\sum\nolimits_i u_i\le k\}.\end{array}$$

$x$ is robust feasible iff

$$\begin{array}{rcl}b&\ge&\max\limits_a\left\{x^\top a:a\in\mathcal{U}\right\}=\max\limits_{\epsilon,u}\left\{x^\top[a^*+D\epsilon]:-u\le\epsilon\le u,\ u\le[1;...;1],\sum\nolimits_i u_i\le k\right\}\\ &=&x^\top a^*+\max\limits_{\epsilon,u}\left\{[Dx]^\top\epsilon:-u\le\epsilon\le u,u\le[1;...;1],\sum\nolimits_i u_i\le k\right\}\\ &=&x^\top a^*+\min\limits_{\lambda_{\ell,u},\lambda_{g,u},\lambda_{\ell,1},\lambda_{\ell,k}}\left\{[1;...;1]^\top\lambda_{\ell,1}+k\lambda_{\ell,k}:\right.\\ &&\qquad\left.\begin{array}{l}\lambda_{\ell,u}\ge0,\lambda_{\ell,1}\ge0,\lambda_{\ell,k}\ge0,\lambda_{g,u}\le0\\ \lambda_{\ell,u}+\lambda_{g,u}=Dx\\ -\lambda_{\ell,u}+\lambda_{g,u}+\lambda_{\ell,1}+\lambda_{\ell,k}[1;...;1]=0\end{array}\right\}\\ &&\text{[LO duality]}\\ &=&x^\top a^*+\min\limits_{\lambda_{\ell,1},\lambda_{\ell,k}}\left\{[1;...;1]^\top\lambda_{\ell,1}+k\lambda_{\ell,k}:\begin{array}{l}\lambda_{\ell,1}\ge0,\lambda_{\ell,k}\ge0\\ -\lambda_{\ell,1}-\lambda_{\ell,k}[1;...;1]\le Dx\le\lambda_{\ell,1}+\lambda_{\ell,k}[1;...;1]\end{array}\right\}\\ &&\text{[eliminating }\lambda_{\ell,u},\lambda_{g,u}]\end{array}$$

Thus, (RC) can be represented as the system of linear constraints

$$\begin{array}{l}\lambda_{\ell,1}\ge0,\lambda_{\ell,k}\ge0,[a^*]^\top x+[1;...;1]^\top\lambda_{\ell,1}+k\lambda_{\ell,k}\le b,\\ -\lambda_{\ell,1}-\lambda_{\ell,k}[1;...;1]\le Dx\le\lambda_{\ell,1}+\lambda_{\ell,k}[1;...;1],\end{array}$$

in variables $x, \lambda_{\ell,1}, \lambda_{\ell,k}$.

**Exercise II.45.** [computational study, follow-up to Exercise II.43]
*Preliminaries.* Consider oscillator transmitting harmonic wave with unit wavelength and placed at some point $P$ in 3D. Physics says that the electric field generated by the oscillator, when measured at a remote point $A$, is

$$e_A(t) \approx r^{-1} \underbrace{\alpha \cos\left(\omega t - 2\pi r + \theta + 2\pi d \cos(\phi)\right)}_{E_A(t)} \qquad (*)$$

where

- $t$ is time, $\omega$ is the frequency,
- $r$ is the distance from $A$ to the origin $O$, $d$ is the distance from $P$ to the origin, $\phi \in [0,\pi]$ is the angle between the directions $\overrightarrow{OP}$ and $\overrightarrow{OA}$,
- $\alpha$ and $\theta$ are responsible for how the oscillator is actuated.

The difference between the left and the right hand sides in $(*)$ of order of $r^{-2}$ and in all our subsequent considerations can be completely ignored.

It is convenient to assemble $\alpha$ and $\theta$ into the *actuation weight* – the complex number $w = \alpha e^{\imath\theta}$ ($\imath$ is the imaginary unit); with this convention, we have

$$E_A(t) = \Re\left[w D_P(\phi) e^{\imath\omega t - 2\pi r}\right],\ D_P(\phi) = e^{2\pi\imath d\cos(\phi)}.$$

where $\Re[\cdot]$ stands for the real part of a complex number. The complex-valued function $D_P(\phi) : [0,\pi] \to \mathbf{C}$, called *the diagram* of the oscillator, is responsible for the directional density of the energy emitted by the oscillator: when evaluated at certain 3D direction $\vec{e}$, this density is proportional to $|D_p(\phi)|^2$, where $\phi$ is the angle between the direction $\vec{e}$ and the direction $\overrightarrow{OP}$. Physics says that when our transmitting antenna is composed of $K$ harmonic oscillators located at points $P_1, ..., P_K$ and actuated with weights $w_1, ..., w_K$, the directional density of energy emitted by the resulting *antenna array*, as evaluated at a direction $\vec{e}$, is proportional to $|\sum_k w_k D_k(\phi_k(\vec{e}))|^2$, where $\phi_k(\vec{e})$ is the angle between the directions $\vec{e}$ and $\overrightarrow{OP_k}$.

Consider the design problem as follows. We are given linear array of $K$ oscillators placed at the points $P_k = (k-1)\delta\mathbf{e}$, $k \leq K$, where $\mathbf{e}$ is the first basic orth (that is, the unit vector "looking" along the positive direction of the $x$-axis), and $\delta > 0$ is a given distance between consecutive oscillators. Our goal is to specify actuation weights $w_k$, $k \leq K$, in order to send as much of total energy as possible along the directions which make at most a given angle $\gamma$ with $\mathbf{e}$. To this end, we intend to act as follows:

> We want to select actuation weights $w_k$, $k \leq K$, in such a way that the magnitude $|D^w(\phi)|$ of the complex-valued function
>
> $$D^w(\phi) = \sum_{k=1}^K w_k e^{2\pi\imath(k-1)\delta\cos(\phi))}$$
>
> of $\pi \in [0,\pi]$ is "concentrated" on the segment $0 \leq \phi \leq \gamma$. Let us normalize the weights by the requirement
>
> $$D^w(0) = 1$$
>
> and minimize under this restriction the "sidelobe level"
>
> $$\max_{\gamma \leq \phi \leq \pi} |D^w(\phi)|$$
>
> over $w$.

To get a computation-friendly version of this problem, we replace the full range $[0,\pi]$ of values of $\phi$ with $M$-point equidistant grid

$$\Gamma = \{\phi_\ell = \frac{\ell\pi}{M-1} : 0 \leq \ell \leq M-1\},$$

thus converting our design problem into the optimization problem

$$\mathrm{Opt} = \min_{t,w}\left\{t : \begin{array}{l}|\sum_{k=1}^K w_k \mathrm{e}^{2\pi\imath(k-1)\delta\cos(\phi_\ell)}| \leq t\,\forall(\ell:\phi_\ell>\gamma)\\ \sum_{k=1}^K w_k \mathrm{e}^{2\pi\imath(k-1)\delta}=1\end{array}, w_k\in\mathbf{C}, k\leq K\right\} \qquad (P)$$

which is a convex problem in $2k$ real variables – real and imaginary parts of $w_1,...,w_K$.
*Your tasks* are as follows:

1. Process problem $(P)$ numerically and find the optimal design $w^{\mathrm{n}}=\{w^{\mathrm{n}}_k, k\leq K\}$ along with the optimal value $\mathrm{Opt}^{\mathrm{n}}$. Here and in what follows, recommended setup is

   - number of oscillators $K=24$, distance between consecutive oscillators $\delta=0.125$
   - $\gamma=\pi/12$
   - cardinality $M$ of the equidistant grid $\Gamma$ is 512

   Draw the plot of the modulus of the resulting diagram

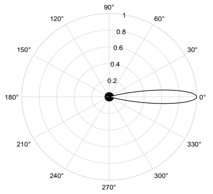
   $$D^{\mathrm{n}}(\phi)=\sum_{k=1}^K w^{\mathrm{n}}_k \mathrm{e}^{2\pi\imath(k-1)\delta\cos(\phi)}$$

   and compute the corresponding "energy concentration" $\mathcal{C}^{\mathrm{n}}$, with concentration of a diagram $D(\cdot)$ defined as

   $$\mathcal{C}=\frac{\sum_{\ell:\phi_\ell\leq\gamma}\sin(\phi_\ell)|D(\phi_\ell)|^2}{\sum_{\ell=1}^M\sin(\phi_\ell)|D(\phi_\ell)|^2}$$

   – up to discretization of $\phi$, this is the ratio of the energy emitted in the "cone of interest" (i.e., along the directions making angle at most $\gamma$ with $\mathbf{e}$) to the total emitted energy. Factors $\sin(\phi_\ell)$ reflect the fact that when computing the energy emitted in a spatial cone, we should integrate $|D(\cdot)|^2$ over the part of the unit sphere in $3D$ cut off the sphere by the cone.

   *Solution:* Our computation yielded diagram with modulus as shown on Figure VI.2.



$\mathrm{Opt}^{\mathrm{n}}=0.053$, $\mathcal{C}^{\mathrm{n}}=74.8\%$

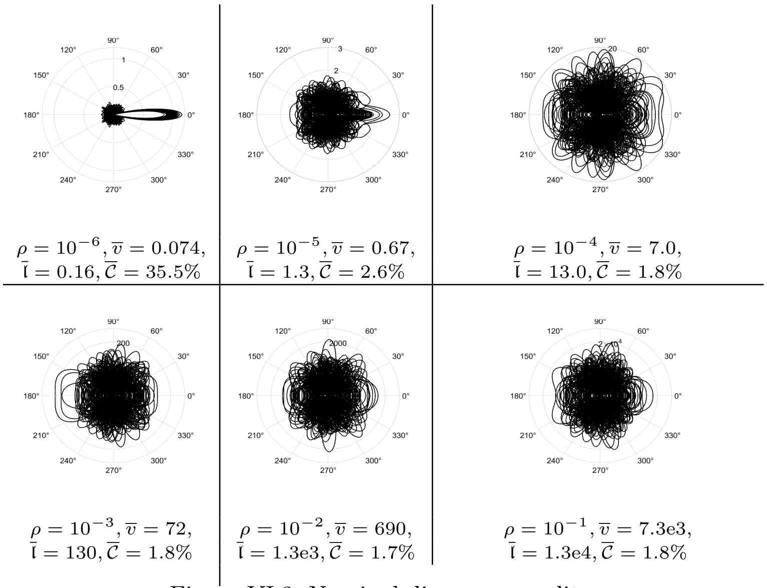
Figure VI.2. Optimal diagram, dream – no actuation errors.

2. Now note that "in reality" the optimal weights $w^{\mathrm{n}}_k$, $k\leq K$ are used to actuate physical devices and as such cannot be implemented with the same 16-digit accuracy with which they are computed; they definitely will be subject to small implementation errors. We can model these errors by assuming that the "real life" diagram is

   $$D(\phi)=\sum_{k=1}^K w^n_k(1+\rho\xi_k)\mathrm{e}^{2\pi\imath(k-1)\delta\cos(\phi)}$$

   where $\rho\geq 0$ is some (perhaps small) perturbation level and $\xi_k\in\mathbf{C}$ are "primitive" perturbations responsible for the implementation errors and running through the unit disk $\{\xi:|\xi|\leq 1\}$. It is not a great sin to assume that $\xi_k$ are independent across $k$ random variables uniformly distributed on the unit circumference in $\mathbf{C}$. Now the diagram becomes random and can violate the constraints of $(P)$ , unless $\rho=0$; in the latter case, the diagram is the "nominal" one given by the optimal weights $w^{\mathrm{n}}$, so that it satisfies the constraints of $(P)$ with $t$ set to $\mathrm{Opt}^{\mathrm{n}}$.
   Now, what happens when $\rho>0$? In this case, the diagram $D(\cdot)$ and its deviation $v$ from the prescribed value 1 at the origin, its sidelobe level $\mathfrak{l}=\max_{\ell:\phi_\ell>\gamma}|D(\phi_\ell)|$, and energy concentration become random. A crucial "real life" question is how large are "typical values" of these quantities. To get impression of what happens, you are asked to carry out the numerical experiment as follows:

   - select perturbation level $\rho\in\{10^{-\ell}, 1\leq\ell\leq 6\}$



$\rho = 10^{-6}, \overline{v} = 0.074, \overline{\mathfrak{l}} = 0.16, \overline{\mathcal{C}} = 35.5\%$

$\rho = 10^{-5}, \overline{v} = 0.67, \overline{\mathfrak{l}} = 1.3, \overline{\mathcal{C}} = 2.6\%$

$\rho = 10^{-4}, \overline{v} = 7.0, \overline{\mathfrak{l}} = 13.0, \overline{\mathcal{C}} = 1.8\%$

$\rho = 10^{-3}, \overline{v} = 72, \overline{\mathfrak{l}} = 130, \overline{\mathcal{C}} = 1.8\%$

$\rho = 10^{-2}, \overline{v} = 690, \overline{\mathfrak{l}} = 1.3\text{e}3, \overline{\mathcal{C}} = 1.7\%$

$\rho = 10^{-1}, \overline{v} = 7.3\text{e}3, \overline{\mathfrak{l}} = 1.3\text{e}4, \overline{\mathcal{C}} = 1.8\%$

Figure VI.3. Nominal diagram – reality,
Magnitudes of 100 actual diagrams stemming from the optimal solution to $(P)$

- for selected $\rho$, simulate and plot 100 realizations of the modulus of the actual diagram, and find empirical averages $\overline{v}$ of $v$, $\overline{\mathfrak{l}}$ of $\mathfrak{l}$, and $\overline{\mathcal{C}}$ of $\mathcal{C}$.

*Solution:* Our experimental results are shown on Figure VI.3. To put the above concentration numerics into proper perspective, note that with our setup, the surface of the "spherical hat" cut off the unit sphere by our cone of interest is 1.7% of the total surface of the sphere, so that energy concentration 1.7% we can get without any trouble by placing just one oscillator at the origin.

Taking into account that in "real life" implementation errors in antenna weights hardly could be less than 0.1% (corresponding to $\rho = 10^{-3}$), we would qualify the *nominal* design yielded by the optimal solution to the nominal problem $(P)$, same as the nominal optimal value, as wishful thinking completely meaningless for actual antenna design.

3. Apply Robust Optimization methodology from Exercise II.43 to build "immunized against implementation errors" solution to $(P)$, compute these solutions for perturbation levels $10^{-\ell}$, $1 \leq \ell \leq 6$, and subject the resulting designs to numerical study similar to the one outlined in the previous item.
*Note:* $(P)$ is *not* a Linear Programming program, so that you cannot formally apply the results stated in Exercise II.43; what you can apply, is the Robust Optimization "philosophy."

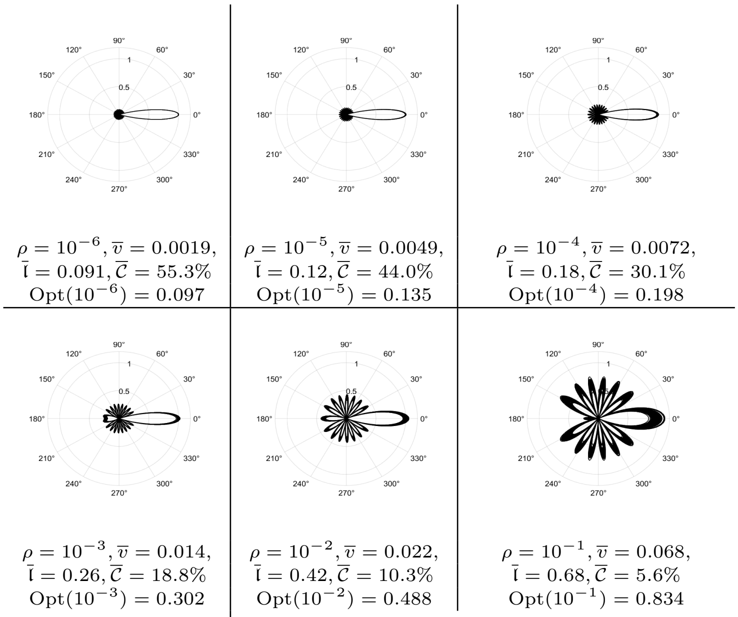
*Solution:*

- With our model of implementation errors, the effect of these errors on the value of the actual diagram $D(\cdot)$ as evaluated at a point $\phi \in \Gamma$ is in adding to the value

$$D_w(\phi) = \sum\nolimits_{k=1}^{k} w_k \mathrm{e}^{2\pi\imath(k-1)\delta\cos(\phi)}$$

of the "no-errors" diagram corresponding to candidate weights $w$ a perturbation which can be whatever complex number of the modulus not exceeding $\rho\sum_k |w_k|$. Thus, the "robust" – worst-case w.r.t. implementation errors, the perturbation level being $\rho$ – sidelobe level corresponding to candidate weights $w$ is

$$\max\nolimits_{\ell:\phi_\ell > \gamma} |D_w(\phi_\ell)| + \rho \sum\nolimits_k |w_k|,$$



$\rho = 10^{-6}, \overline{v} = 0.0019,$
$\bar{\imath} = 0.091, \overline{\mathcal{C}} = 55.3\%$
$\mathrm{Opt}(10^{-6}) = 0.097$

$\rho = 10^{-5}, \overline{v} = 0.0049,$
$\bar{\imath} = 0.12, \overline{\mathcal{C}} = 44.0\%$
$\mathrm{Opt}(10^{-5}) = 0.135$

$\rho = 10^{-4}, \overline{v} = 0.0072,$
$\bar{\imath} = 0.18, \overline{\mathcal{C}} = 30.1\%$
$\mathrm{Opt}(10^{-4}) = 0.198$

$\rho = 10^{-3}, \overline{v} = 0.014,$
$\bar{\imath} = 0.26, \overline{\mathcal{C}} = 18.8\%$
$\mathrm{Opt}(10^{-3}) = 0.302$

$\rho = 10^{-2}, \overline{v} = 0.022,$
$\bar{\imath} = 0.42, \overline{\mathcal{C}} = 10.3\%$
$\mathrm{Opt}(10^{-2}) = 0.488$

$\rho = 10^{-1}, \overline{v} = 0.068,$
$\bar{\imath} = 0.68, \overline{\mathcal{C}} = 5.6\%$
$\mathrm{Opt}(10^{-1}) = 0.834$

Figure VI.4. Robust design.
Magnitudes of 100 actual diagrams stemming from optimal solutions to $(R)$

and the robust counterpart of the system of inequality constraints in $(P)$ is the constraint

$$t \geq \max_{\ell:\phi_\ell>\gamma}|D_w(\phi_\ell)| + \rho\sum\nolimits_k |w_k|. \tag{$C$}$$

As about the normalizing equality constraint in $(P)$, formally its robust counterpart is contradictory – we cannot select $w_k$ such that the actual diagram as evaluated at $\phi = 0$ will be exactly one, whatever be the perturbations. *It makes full sense to keep this constraint as is* – in the presence of implementation errors, it will be violated by at most $\rho\sum_k |w_k|$, the same quantity which we see in $(C)$. Hopefully, this quantity will be made small by minimization over $w$ of the right hand side in $(C)$.

With the outlined approach the robust w.r.t. implementation errors counterpart of $(P)$ is the convex optimization problem

$$\mathrm{Opt}(\rho) = \min_{t,w}\left\{ t: \begin{array}{l} |\sum_{k=1}^K w_k \mathrm{e}^{2\pi\imath(k-1)\delta\cos(\phi_\ell)}| + \rho\sum_k |w_k| \leq t\,\forall(\ell:\phi_\ell>\gamma) \\ \sum_{k=1}^K w_k \mathrm{e}^{2\pi\imath(k-1)\delta} = 1 \\ w_k \in \mathbf{C},\, k \leq K \end{array} \right\} \tag{$R$}$$

- The results of our experiments with robust designs yielded by $(R)$ are shown in Figure VI.4. Comparison with similar results for the nominal design speaks for itself loud and clear.

**Exercise II.46.** Prove the statement "symmetric" the Dubovitski-Milutin Lemma:

> The cone $M_*$ dual to the arithmetic sum of $k$ (close or not) cones $M^i \subset \mathbf{R}^n$, $i \leq k$, is the intersection of the $k$ cones $M^i_*$ dual to $M^i$.

*Solution:* By evident reasons, a linear function $f^\top x$ can be nonnegative everywhere on the arithmetic sum of $k$ nonempty sets $M^1 + ... + M^k$ if and only if it is nonnegative on every one of these sets.

**Exercise II.47.** Prove the following polyhedral version of the Dubovitski-Milutin Lemma:

Let $M^1, ..., M^k$ be polyhedral cones in $\mathbf{R}^n$, and let $M = \cap_i M^i$. The cone $M_*$ dual to $M$ is the sum of cones $M^i_*$, $i \leq k$, dual to $M^i$, so that a linear form $e^\top x$ is nonnegative on $M$ if and only it can be represented as the sum of linear forms $e_i^\top x$ nonnegative on the respective cones $M_i$.

*Solution:* This is immediate consequence of Proposition II.8.25 combined with the fact that by calculus of polyhedral representations from section 3.3 and the result of Exercise II.39, intersections, sums, and duals of polyhedral cones are polyhedral cones and therefore are closed.

**Exercise II.48.** [follow-up to Exercise II.47] Let $A \in \mathbf{R}^{m\times n}$ be a matrix with trivial kernel, $e \in \mathbf{R}^n$, and let the set

$$X = \{x : Ax \geq 0, e^\top x = 1\} \tag{$*$}$$

be nonempty and bounded. Prove that there exists $\lambda \in \mathbf{R}^m$ such that $\lambda > 0$ and $A^\top \lambda = e$.

Prove "partial inverse" of this statement: *if* $\mathrm{Ker} A = \{0\}$ *and* $e = A^\top \lambda$ *for some* $\lambda > 0$, *the set* $(*)$ *is bounded.*

*Solution:* Let $E$ be the image space of $A$, and $P$ be the orthogonal projector of $\mathbf{R}^m$ onto $E$. Since $\mathrm{Ker}A = \{0\}$, there exists $g \in \mathbf{R}^m$ such that $A^\top g = e$, so that $y \in \mathbf{R}^m$ is representable as $Ax$ with $e^\top x = 1$ if and only if $y \in E$ and $g^\top y = 1$. Therefore

$$Y := \{y = Ax : x \in X\} = \{y \in E : y \geq 0, g^\top y = 1\}$$

$Y$ is cut off the cone $M = \mathbf{R}^m_+ \cap E$ by the linear equality constraint $g^\top y = 1$ and is nonempty and bounded. Clearly $M$ is pointed along with $\mathbf{R}^m_+$ and is nontrivial (since $Y$ is nonempty). Treating $M$ as a cone in Euclidean space $E$ equipped with the inner product $\langle \cdot, \cdot \rangle$ inherited from the standard inner product on $\mathbf{R}^m$, and denoting by $f$ the orthoprojection of $g$ onto $E$, we have

$$Y = \{y \in M : \langle f, y \rangle = 1\};$$

since $Y$ is nonempty and bounded, and $M$ is closed, nontrivial and pointed, Fact II.8.23.3 states that $f \in \mathrm{int}\, M_*$, where $M_*$ is the dual of the cone $M \subset E$. In other words, for some $r > 0$ and all $f' \in E$, $\|f' - f\|_2 \leq r$, we have $f' \in M_*$. Now let $\overline{\lambda} \in \mathrm{int}\, \mathbf{R}^m_+$ be such that $\|\overline{\lambda}\|_2 \leq r$, and let $\overline{g} = g - \overline{\lambda}$ and $\overline{f} = P\overline{g} \in E$, so that $\|f - \overline{f}\|_2 \leq \|g - \overline{g}\|_2 \leq r$. The latter inequality, due to the origin of $r$, implies that

$$\langle \overline{f}, y \rangle \geq 0 \ \forall y \in M,$$

or, which is the same,

$$\overline{g}^\top y \geq 0 \ \forall y \in \mathbf{R}^m_+ \cap E.$$

By polyhedral version of Dubovitski-Milutin Lemma (Exercise II.47), there exists $\widetilde{\lambda} \in (\mathbf{R}^m_+)_* = \mathbf{R}^m_+$ and

$$\mu \in E_* := \{\mu : \mu^\top u \geq 0 \, \forall u \in E\} = E^\perp = [\mathrm{Im}A]^\perp = \mathrm{Ker}A^\top$$

such that $\overline{g} = \lambda + \mu$, implying that

$$g = \overline{g} + \overline{\lambda} = \underbrace{[\widetilde{\lambda} + \overline{\lambda}]}_{\lambda > 0} + \mu,$$

whence

$$e = A^\top g = A^\top \lambda + A^\top \mu = A^\top \lambda.$$

We have found $\lambda > 0$ with $A^\top \lambda = e$, as required.
To prove "partial inverse", note that if $\lambda > 0$, then the set

$$Z = \{y \in \mathbf{R}^m_+ : \lambda^\top y = 1\}$$

is bounded; when $A^\top \lambda = e$, $X$ is the inverse linear image of $Z$ under the linear embedding $x \mapsto Ax : \mathbf{R}^n \to \mathbf{R}^m$, and therefore $X$ is bounded along with $Z$.

**Exercise II.49.** Let $E$ be a linear subspace in $\mathbf{R}^n$, $K$ be a closed cone in $\mathbf{R}^n$, and $\ell(x) : E \to \mathbf{R}$ be a linear (linear, not affine!) function which is nonnegative on $K \cap E$. Which of the following claims are always true:

1. $\ell(\cdot)$ can be extended from $E$ onto the entire $\mathbf{R}^n$ to yield a linear function which is nonnegative on $K$
2. Assuming $\mathrm{int}\, K \cap E \neq \varnothing$, $\ell(\cdot)$ can be extended from $E$ onto the entire $\mathbf{R}^n$ to yield a linear function which is nonnegative on $K$.
3. Assuming, in addition to $\ell(x) \geq 0$ for $x \in K \cap E$, that $K = \{x : Px \leq 0\}$ is a polyhedral cone, $\ell(\cdot)$ can be extended from $E$ onto the entire $\mathbf{R}^n$ to yield a linear function which is nonnegative on $K$.

*Solution:* The first claim is wrong in general. The simplest counterexample is $K = \mathbf{L}^3$: take a generator of $K$ - an emanating from the origin ray on the boundary of the cone, say, $R = \{x \in \mathbf{R}^3 : x_3 = x_1, x_2 = 0\}$, and let $E = \{x \in \mathbf{R}^3 : x_1 = x_3\}$ be the 2D plane tangent to the surface of the cone along the ray $R$. Linear function $\ell(x) = x_2 : E \to \mathbf{R}$ is nonnegative on $K \cap E = R$, but cannot be extended to a linear function $f(x) = e^\top x$ on $\mathbf{R}^3$ nonnegative on $K$; indeed, points $x_\delta = [1; -\delta; 1]$ with $\delta > 0$ are in $E$, so that we should have $f(x_\delta) = \ell(x_\delta) = -\delta$; at the same time, $x_\delta$ belongs to the plane $E$ and $\|x_0 - x_\delta\|_2 = \delta$; since $E$ is tangent to the boundary of $K$ at $x_0$, there are points $x_\delta^+ \in K$ with $\|x_\delta - x_\delta^+\|_2 \leq O(1)\delta^2$ (you can take $x_\delta^+ = [1; -\delta; \sqrt{1+\delta^2}]$), so that $f(x_\delta^+) \leq f(x_\delta) + \|e\|_2 \|x_\delta^+ - x_\delta\|_2 \leq -\delta + O(1)\|e\|_2\delta^2$, that is, $f(x_\delta^+) < 0$ for all small $\delta > 0$, which is a desired contradiction, since $x_\delta^+ \in K$, and $f(x)$ on $K$ is nonnegative.

The second claim is true by Dubovitski-Milutin Lemma (DML). Indeed, in the notation of this Lemma, set $M_1 = K$ and $M_2 = E$, thus satisfying the premise of Lemma. Extend $\ell(\cdot)$ to a whatever linear form $e^\top x$ of $x \in \mathbf{R}^n$; this form is nonnegative on $M_1 \cap M_2$ and therefore, by DML, can be represented as $g^\top x + h^\top x$ with $g^\top x$ nonnegative for $x \in M_1 = K$ and $h^\top x$ nonnegative when $x \in M_2 = E$, that is, with $h^\top x = 0$ for $x \in E$ ($E$ is a linear subspace!). The desired extension of $\ell(x)$ from $E$ to a nonnegative on $K$ linear form is $x \mapsto g^\top x$.

The third claim is true. Indeed, $E$ is a linear subspace and thus is a polyhedral cone. We can find a vector $e \in \mathbf{R}^n$ such that $\ell(x) = e^\top x$ for $x \in E$. Applying the result of Exercise II.47 to the polyhedral cones $M^1 = K$ and $M^2 = E$, we conclude that under the premise of item 3 we have $e = e_1 + e_2$ with $e_1 \in M_*^1 = K_*$ and $e_2 \in M_*^2 = E^\perp$, implying that $e_1^\top x$ is the desired linear form nonnegative on $K$ and equal to $\ll (x)$ on $E$. □

**Exercise II.50.** Let $n > 1$. Is the unit $\|\cdot\|_2$-ball $B_n = \{x \in \mathbf{R}^n : \|x\|_2 \leq 1\}$ a polyhedral set? Justify your answer.

*Solution:* The answer, of course, is "no". Indeed, were $B_n$ polyhedral, the set of extreme points of $B_n$ would be finite (Corollary II.9.2), which contradicts the evident fact that $\mathrm{Ext}(B_n) = S_n := \{x \in \mathbf{R}^n : \|x\|_2 = 1\}$, and this set is infinite when $n > 1$. To show that $\mathrm{Ext}(B_n) = S_n$ is clearly the same as to show that every point $e \in S_n$ is extreme; to this end note that when $e \pm h \in B_n$, we have $2\|e\|_2^2 + 2\|h\|_2^2 = \|e+h\|_2^2 + \|e-h\|_2^2 \leq 2$, and the resulting inequality implies that $h = 0$ due to $\|e\|_2 = 1$.

It is worthy of mentioning that *"for all practical purposes," $B_n$ is a simple polyhedral set*. Specifically, it is known (see [Nem24, section 1.4]) that for every $\epsilon \in (0, 1/2)$ and every $n$ one can explicitly write down system of $O(1)n\ln(1/\epsilon)$ linear inequalities with $O(1)n\ln(1/\epsilon)$ variables such that the projection of the solution set of this system onto the plane of the first $n$ variables is in-between $B_n$ and $(1+\epsilon)B_n$. When $\epsilon = 1.0\mathrm{e}{-17}$, usual computer does not distinguish between 1 and $1+\epsilon$, so that for all practical purposes $B_n$ admits explicit polyhedral representation; to get $\epsilon = 1.\mathrm{e}{-17}$, this representation should involve $\approx 79n$ linear inequalities on $\approx 28n$ variables.

**Exercise II.51.** The unit box $\{x \in \mathbf{R}^n : -1 \leq x_i \leq 1, i \leq n\}$ is cut off $\mathbf{R}^n$ by a system of $m = 2n$ linear inequalities and is a nonempty and bounded polyhedral set. However, when we eliminate any inequality from this system, the solution set of the resulting system becomes unbounded. To see that this situation is in a sense extreme, prove the following claim:

*Consider the solution set of a system of $m$ linear inequalities in $n$ variables $x$, i.e., the set*

$$X := \{x \in \mathbf{R}^n : \ Ax \leq b\},$$

*where $A = [a_1^\top; a_2^\top; \ldots; a_m^\top]$. Suppose that $X$ is nonempty and bounded. Then, whenever $m > 2n$, one can drop from this system a properly selected inequality in such a way that the solution set of the resulting subsystem remains bounded.*

A provocative follow-up: *Is it possible to cut off from $\mathbf{R}^{1000}$ a bounded set by using* only *a single linear inequality?*

*Solution:* Suppose $X$ is nonempty and bounded, and let $m > 2n$. Recall from Fact II.8.13 that a nonempty closed convex set is bounded if and only if its recessive cone is trivial. Then, as $X$ is closed (it is polyhedral!), we have $\mathrm{Rec}(X) = \{0\}$. Moreover, based on Fact II.8.15 we have $\mathrm{Rec}(X) = \{x : Ax \leq 0\}$. Thus, the closed cone $K := \{x : Ax \leq 0\}$ is trivial, or, which is the same by Fact II.8.23, the dual $K_*$ of this cone is the entire $\mathbf{R}^n$. On the other hand, as is explained immediately after Proposition II.8.21, $K_* = \mathrm{Cone}(\{-a_1, \ldots, -a_m\})$. Thus, $\mathrm{Cone}(\{-a_1, \ldots, -a_m\}) = \mathbf{R}^n$, implying, in particular, that $\operatorname{rank} A = n$ (since the conic hull of $-a_1, \ldots, -a_m$ belongs to the linear span of the collection $a_1, \ldots, a_m$). Without loss of generality, we can assume that $a_1, \ldots, a_n$ are linearly independent. Hence, the vector $\overline{a} := \sum_{i=1}^n a_i$ belongs to $K_* = \mathbf{R}^n$ and therefore $\overline{a}$ is a conic combination of vectors $-a_1, \ldots, -a_m$. Then, by conic version of Caratheodory's Theorem (Fact I.2.7) we can select $n$ vectors $a_{i_1}, \ldots, a_{i_n}$ from the given $m$ vectors $a_1, \ldots, a_m$ in such a way that $\overline{a}$ is a conic combination of the vectors $-a_{i_1}, \ldots, -a_{i_n}$. As a result, all vectors of the form

$$\sum_{i=1}^n (1 - \mu_i)a_i, \quad \text{where } \mu \geq 0, \tag{$*$}$$

are conic combinations of vectors from the collection $\{-a_1, -a_2, \ldots, -a_n, -a_{i_1}, -a_{i_2}, \ldots, -a_{i_n}\}$. All vectors $\sum_{i=1}^n z_i a_i$ with $\|z\|_\infty \leq 1$ admit a representation of the form $(*)$ and therefore, as we have just seen, they belong to the cone $\mathrm{Cone}(\{-a_1, -a_2, \ldots, -a_n, -a_{i_1}, -a_{i_2}, \ldots, -a_{i_n}\})$. Since $a_1, \ldots, a_n$ are linearly independent, the set of linear combinations of these vectors with coefficients of magnitude $\leq 1$ contains a neighborhood of the origin. Thus, the cone

$$\mathrm{Cone}(\{-a_1, -a_2, \ldots, -a_n, -a_{i_1}, -a_{i_2}, \ldots, -a_{i_n}\})$$

contains a neighborhood of the origin and is therefore the entire $\mathbf{R}^n$. On the other hand, by the same argument as above, this cone is dual to the cone

$$\left\{x : \ a_i^\top x \leq 0, \, i \leq n, \ a_{i_j}^\top x \leq 0, \, j \leq n\right\},$$

so that the latter cone is trivial. Thus, the recessive cone of the set

$$X^+ := \left\{x : \ a_i^\top x \leq b_i, \, i \leq n, \ a_{i_j}^\top x \leq b_{i_j}, \, j \leq n\right\}$$

is trivial, and therefore this set is bounded. Thus, we conclude that we can extract a carefully selected set of $2n$ constraints from the constraints $a_i^\top x \leq b_i$, $i \leq m$, such that they still result in a bounded set in $\mathbf{R}^n$. □

The answer to the follow-up question is positive: the insolvable linear inequality $0^\top x \leq -1$ cuts off $\mathbf{R}^{1000}$ the empty set which of course is bounded.

**Exercise II.52.** [computational study] let $\omega^N = (\omega_1, ..., \omega_N)$ be an $N$-element i.i.d. sample drawn from the standard Gaussian distribution (zero mean, unit covariance) on $\mathbf{R}^d$. How many extreme points are there in the convex hull of the points from the sample?

1. Consider the planar case $d = 2$ and think how to list extreme points of $\mathrm{Conv}\{\omega_1, ..., \omega_N\}$. Fill

the following table:

| $N$ | 2 | 4 | 8 | 16 | 32 | 64 | 128 |
|---|---|---|---|---|---|---|---|
| $U$ | | | | | | | |
| $M$ | | | | | | | |
| $L$ | | | | | | | |

where $U$ is the maximal, $M$ is the mean, and $L$ is the minimal # of extreme points observed when processing 100 samples $\omega^N$ of a given cardinality.

*Solution:* The simplest way to check whether a point, say, $\omega_1$, from $N$-element sample $\omega^N$ of 2D points is or is not extreme, is to look at $N-1$ lines $\ell_j$, $j = 2, ..., N$, linking $\omega_1$ and $\omega_j$. When no triple of points from the sample belong to a common line (which happens with probability 1), $\omega_1$ is extreme point of $\mathrm{Conv}(\omega^N)$ if and only if all points of the sample are on one side of one of these lines.

Here are our results:

| $N$ | 2 | 4 | 8 | 16 | 32 | 64 | 128 |
|---|---|---|---|---|---|---|---|
| $U$ | 2 | 8 | 14 | 18 | 20 | 24 | 26 |
| $M$ | 2.00 | 7.36 | 10.22 | 12.60 | 14.72 | 16.54 | 18.36 |
| $L$ | 2 | 6 | 8 | 8 | 10 | 10 | 10 |
| $E$ | 4 | 7.30 | 10.06 | 12.37 | 14.34 | 17.23 | 17.77 |

(the $E$-row is the answer to the question of item 2).

2. Think how to upper-bound the expected number of extreme points in the set $W = \mathrm{Conv}(\omega^N)$.

*Solution:* Similarly to the previous item, ignoring "degenerate" samples with total probability mass 0, $\omega_1$ is an extreme point of $W$ if one can select $d-1$ points $\omega_{i_2}, ..., \omega_{i_d}$ with $2 \leq i_2 < i_3 < ...$ in such a way that the entire sample is on one side of the hyperplane passing through $\omega_1, \omega_{i_2}, \omega_{i_3}, ..., \omega_{i_d}$. Probability $\pi$ of this outcome clearly is independent of what is the collection $i_2 < i_3 < ... < i_d$ and can be reliably estimated via simulation, namely, as follows: we simulate $M \gg 1$ times $d$-element sample $\omega^d$, measure the Euclidean distance from the hyperplane containing this sample to the origin, and recover the distribution $P$ of this distance. We clearly have

$$\pi = \int_0^\infty [\psi^{N-d}(s) + (1 - \psi(s))^{N-d}] dP(s),$$

where $\psi$ is the cumulative distribution function of $\mathcal{N}(0,1)$ random variable, and we can estimate $\pi$ by substituting the expectation w.r.t. $P$ with expectation w.r.t. the empirical approximation of $P$. After $\pi$ is estimated, we can upper-bound the probability for $\omega_1$ to be an extreme point of $W$ by the quantity $\theta = \binom{N-1}{d-1}\pi$, resulting in the upper bound $\theta N$ on the expected number of extreme points of $W$.

**Exercise II.53.** [computational study] Given positive integers $m, n$, with $n \geq 2$, consider randomly generated system $Ax \leq b$ of $m$ linear inequalities with $n$ variables. We assume that $A$, $b$ are generated by drawing the entries, independently of each other, from $\mathcal{N}(0,1)$.

1. Consider the planar case $n = 2$. For $m = 2, 4, 8, 16$, generate 100 samples of $m \times 2$ systems and fill the following table:

| $m$ | 2 | 4 | 8 | 16 |
|---|---|---|---|---|
| $F$ | | | | |
| $B$ | | | | |

where $F$ is the number of feasible systems, and $U$ is the number of feasible systems with bounded solution sets.

*Solution:* see item 3 below.

Intermezzo: related theoretical results originating from [Nem24, Exercise 2.23] are as follows. Given positive integers $m, n$ with $n \geq 2$, consider homogenous system $Ax \leq 0$ of $m$ inequalities with $n$ variables. We call this system *regular*, if its matrix $A$ is regular, regularity of a matrix $B$ meaning that all square submatrices of $B$ are nonsingular. Clearly, the entries of a regular matrix are nonzero, and when a $p \times q$ matrix $B$ is drawn at random from a probability distribution on $\mathbf{R}^{p\times q}$ which has a density w.r.t the Lebesgue measure, $B$ is regular with probability 1.

Given regular $m\times n$ homogeneous system of inequalities $Ax \leq 0$, let $g_i(x) = \sum_{j=1}^n A_{ij}x_j$, $i \leq m$, so that $g_j$ are nonconstant linear functions. Setting $\Pi_i = \{x : g_i(x) = 0\}$, we get a collection of $m$ hyperplanes in $\mathbf{R}^n$ passing through the origin. For a point $x \in \mathbf{R}^n$, the *signature* of $x$ is, by definition, the $m$-dimensional vector $\sigma(x)$ of signs of the reals $g_i(x)$, $1 \leq i \leq m$. Denoting by $\Sigma$ the set of all $m$-dimensional vectors with entries $\pm 1$, for $\sigma \in \Sigma$ the set $\mathcal{C}_\sigma = \{x : \sigma(x) = \sigma\}$ is either empty, or is a nonempty open convex set; when it is nonempty, let us call it a *cell* associated with $A$, and the corresponding $\sigma$ – an $A$-feasible signature. Clearly, for regular system, $\mathbf{R}^n$ is the union of all hyperplanes $\Pi_i$ and all cells associated with $A$. It turns out that

> *The number $N(m, n)$ of cells associated with a regular homogeneous $m \times n$ system $Ax \leq 0$ is independent of the system and is given by a simple recurrence:*

$$\begin{array}{ccc} N(1,2) & = & 2 \\ m \geq 2, n \geq 2 & \Longrightarrow & N(m,n) = N(m-1,n) + N(m-1,n-1) \quad [N(m,1) = 2,\, m \geq 1]. \end{array}$$

Next, when $A$ is drawn at random from probability distribution $P$ on $\mathbf{R}^{m\times n}$ which possesses *symmetric* density $p$, that is, such that $p([a_1^\top; a_2^\top; ...; a_m^\top]) = p([\epsilon_1 a_1^\top; \epsilon_2 a_2^\top; ...; \epsilon_m a_m^\top])$ for all $A = [a_1^\top; a_2^\top; ...; a_m^\top]$ and all $\epsilon_i = \pm 1$, then *the probability for a vector $\sigma \in \Sigma$ to be an $A$-feasible signature is*

$$\pi(m, n) = N(m, n)/2^m.$$

In particular, the probability for the system $Ax \leq 0$ to have a solution set with a nonempty interior (this is nothing but $A$-feasibility of the signature $[-1; ...; -1]$ is $\pi(m, n)$.

The inhomogeneous version of these results is as follows. An $m \times n$ system of linear inequalities $Ax \leq b$ is called regular, if the matrix $[A, -b]$ is regular. Setting $g_i(x) = \sum_{j=1}^n A_{ij}x_j - b_i$, $i \leq n$, the $[A, b]$-signature of $x$ is, as above, the vector of signs of the reals $g_i(x)$. For $\sigma \in \Sigma$, the set $\mathcal{C}_\sigma = \{x : \sigma(x)) = \sigma\}$ is either empty, or is a nonempty open convex set; in the latter case, we call $\mathcal{C}_\sigma$ an $[A, b]$-cell, and call $\sigma$ an $[A, b]$-feasible signature. Setting $\Pi_i = \{x : g_i(x) = 0\}$, we get $m$ hyperplanes in $\mathbf{R}^n$, and the entire $\mathbf{R}^n$ is the union of those hyperplanes and all $[A, b]$-cells. It turns out that

> *The number $N(m, n)$ of cells associated with a regular $m \times n$ system $Ax \leq b$ is independent of the system and is equal to $\frac{1}{2}N(m + 1, n + 1)$.*

In addition, when $m \times (n + 1)$ matrix $[A, b]$ is drawn at random from a probability distribution on $\mathbf{R}^{m\times(n+1)}$ possessing a symmetric density w.r.t. the Lebesgue distribution, *the probability for every $\sigma \in \Sigma$ to be $[A, b]$-feasible signature is*

$$\overline{\pi}(m, n) = N(m + 1, n + 1)/2^{m+1}.$$

In particular, the probability for the system $Ax \leq b$ to be strictly feasible is $\overline{\pi}(m, n)$.

2. Accompanying exercise: Prove that if $A$ is $m \times n$ regular matrix, then the system $Ax \leq 0$ has a nonzero solution if and only if the system $Ax < 0$ is feasible. Derive from this fact that if $[A, b]$ is regular, then the system $Ax \leq b$ is feasible if and only if it is strictly feasible, and that when the system $Ax \leq 0$ has a nonzero solution, the system $Ax \leq b$ is strictly feasible for every $b$.

*Solution:* Let us start with the first claim. The only nontrivial part of it is that for regular $A$, the existence of nonzero $x$ such that $Ax \leq 0$ implies feasibility of the system $Ax < 0$. Let us lead to contradiction the assumption that $A$ is an $m \times n$ regular matrix such that the system $Ax \leq 0$ has a nonzero solution $\overline{x}$, and at the same time the system $Ax < 0$ is infeasible. By General Theorem on Alternative, infeasibility of the system $Ax < 0$ implies that a nontrivial linear combination of rows of $A$ with nonnegative coefficients is 0, or, which is the same, denoting by $a_i^\top$ the rows of $A$, the origin in $\mathbf{R}^n$ is a convex combination of rows of $A$. W.l.o.g. we can assume that positive coefficients in this combination are associated with $a_1, ..., a_k$, for some $k \leq m$. From the relations $\sum_{i=1}^k \lambda_i a_i = 0$,

$\lambda_i > 0, i \leq k$, and $a_i^\top \overline{x} \leq 0$ it follows that $\overline{x}^\top a_i = 0$, $i = 1, ..., k$. Since $\overline{x} \neq 0$, it follows that $a_i$, $i \leq k$, belong to an $(n-1)$-dimensional subspace of $\mathbf{R}^n$, so that the affine dimension of the affine span of $a_1, ..., a_k$ is at most $n-1$. Since 0 is a convex combination of $a, .., a_k$, by Caratheodory Theorem 0 is a convex combination of $\overline{k} \leq \min[k, n]$ of vectors from the collection $a_1, .., a_k$, implying that properly selected $\overline{k}$ rows in $A$ are linearly dependent, contradicting regularity of $A$. As a corollary, if $A$ is regular and $Ax \leq 0$ for some nonzero $x$, the system $Ax < 0$ is solvable, which, of course, implies that the system $Ax \leq b$ is solvable for every $b$. □

To justify the second claim, it suffices to verify that if the system $Ax \leq b$ is feasible and $[A, b]$ is regular, then the system is strictly feasible. To this end assume that the premise of this claim holds true, so that for some $\bar{x}$ it holds $\bar{b} := A\bar{x} \leq b$. For small $\bar{\epsilon} > 0$ and all $e \in \mathbf{R}^n$, $\|e\|_\infty \leq \bar{\epsilon}$, we have $\underbrace{[A, -\bar{b}; e^\top, -1]}_{B[e]} \underbrace{[\bar{x}; 1]}_{\bar{y}} = [A\bar{x} - \bar{b}; e^\top \bar{x} - 1] \leq 0$. On the other hand, selecting $e$ from the uniform distribution of the box $\|e\|_\infty \leq \epsilon$, with $0 < \epsilon \leq \bar{\epsilon}$, it is easily seen that when $\epsilon > 0$ is small enough, the matrix $B[e]$ is regular with probability 1. Thus, we may assume that $B[e]$ is regular, and, as we have seen, the system $B[e]y \leq 0$ in variables $y$ has a nonzero solution $\bar{y}$. By the already proved first claim, it follows that the system $B[e]y < 0$ has a solution $\widetilde{y}$. As a result, for every $\lambda \in (0, 1)$, the vector $y_\lambda = (1 - \lambda)\bar{y} + \lambda\widetilde{y}$ satisfies $B[e]y_\lambda < 0$. For small positive $\lambda$, $y_\lambda$ is of the form $[x_\lambda; t_\lambda]$ with $t_\lambda > 0$; for such a $\lambda$, relation $B[e]y_\lambda < 0$ implies that $Ax_\lambda - t_\lambda\bar{b} < 0$, whence $A[x_\lambda/t_\lambda] < \bar{b} \leq b$, that is, the system $Ax \leq b$ is strictly feasible. □

Note: by Accompanying Exercise, in the situations described in Intermezzo, probability $\overline{\pi}(m, n)$ for an $m \times n$ system $Ax \leq b$ to be strictly feasible is the same as the probability to be feasible, and the probability to have an unbounded feasible set (i.e., to be feasible and such that $AAh \leq 0$ for some nonzero $h$) is the same as the probability $\pi(m, n)$ for the signature $[-1, ..., -1]$ to be $A$-feasible.

3. Use the results from Intermezzo to compute the expected values of $F$ and $B$, see item 1.

*Solution:* Here are our results:

| $m$ | 2 | 4 | 8 | 16 |
|---|---|---|---|---|
| $F$ | 100 | 72 | 18 | 0 |
| $\mathbf{E}\{F\}$ | 100 | 68.75 | 14.45 | 0.21 |
| $B$ | 0 | 18 | 15 | 0 |
| $\mathbf{E}\{B\}$ | 0 | 18.75 | 8.20 | 0.16 |

**Exercise II.54.** [computational study]

1. For $\nu = 1, 2, ..., 6$, generate 100 systems of linear inequalities $Ax \leq b$ with $n = 2^\nu$ variables and $m = 2n$ inequalities, the entries in $A$, $b$ being drawn, independently of each other, from $\mathcal{N}(0.1)$. Fill the following table:

| $n$ | 2 | 4 | 8 | 16 | 32 | 64 |
|---|---|---|---|---|---|---|
| $F$ | | | | | | |
| $\mathbf{E}\{F\}$ | | | | | | |
| $B$ | | | | | | |

$F$: # of feasible systems in sample;
$B$: # of feasible systems with bounded soultion sets

To compute the expected value of $F$, use the results from [Nem24, Exercise 2.23] cited in item 2 of Exercise II.53.

2. Carry out experiment similar to the one in item 1, but with $m = n + 1$ rather than $m = 2n$.

| $n$ | 2 | 4 | 8 | 16 | 32 | 64 |
|---|---|---|---|---|---|---|
| $F$ | | | | | | |
| $\mathbf{E}\{F\}$ | | | | | | |
| $B$ | | | | | | |
| $\mathbf{E}\{B\}$ | | | | | | |

$F$: # of feasible systems in sample;
$B$: # of feasible systems with bounded soultion sets

*Solution:* Our results, rounded to 2 digits after the dot, are as follows:

1. $m = 2n$:

| $n$ | 2 | 4 | 8 | 16 | 32 | 64 |
|---|---|---|---|---|---|---|
| $F$ | 74 | 72 | 64 | 57 | 50 | 53 |
| $\mathbf{E}\{F\}$ | 68.75 | 63,67 | 59.82 | 57.00 | 54.97 | 53.52 |
| $B$ | 17 | 16 | 7 | 5 | 3 | 3 |
| $\mathbf{E}\{B\}$ | 18.75 | 13.67 | 9.82 | 7.00 | 4.97 | 3.52 |

$F$: # of feasible systems in sample;
$B$: # of feasible systems with bounded solution sets

2. $m = n + 1$:

| $n$ | 2 | 4 | 8 | 16 | 32 | 64 |
|---|---|---|---|---|---|---|
| $F$ | 92 | 96 | 100 | 100 | 100 | 100 |
| $\mathbf{E}\{F\}$ | 87.50 | 96.88 | 100.00 | 100.00 | 100.00 | 100.00 |
| $B$ | 11 | 4 | 0 | 0 | 0 | 0 |
| $\mathbf{E}\{B\}$ | 12.50 | 3.13 | 0.20 | 0.00 | 0.00 | 0.00 |

$F$: # of feasible systems in sample;
$B$: # of feasible systems with bounded solution sets

# Exercises from Part III

## Around convex functions

**Exercise III.1.** Which of the functions below are convex on the indicated domains:

- $f(x) \equiv 1$ on $\mathbf{R}$

  *Solution:* convex
- $f(x) = x$ on $\mathbf{R}$

  *Solution:* convex
- $f(x) = |x|$ on $\mathbf{R}$

  *Solution:* convex
- $f(x) = -|x|$ on $\mathbf{R}$

  *Solution:* nonconvex
- $f(x) = -|x|$ on $\mathbf{R}_+ = \{x \in \mathbf{R} : \; x \geq 0\}$

  *Solution:* convex
- $f(x) = |2x-3|$ on $\mathbf{R}$

  *Solution:* convex
- $f(x) = |2x^2-3|$ on $\mathbf{R}$

  *Solution:* nonconvex
- $\exp\{x\}$ on $\mathbf{R}$

  *Solution:* convex
- $\exp\{x^2\}$ on $\mathbf{R}$

  *Solution:* convex
- $\exp\{-x^2\}$ on $\mathbf{R}$

  *Solution:* nonconvex
- $\exp\{-x^2\}$ on $\{x \in \mathbf{R} : \; x \geq 100\}$

  *Solution:* convex
- $\ln(x)$ on $\{x \in \mathbf{R} : \; x > 0\}$

  *Solution:* nonconvex
- $-\ln(x)$ on $\{x \in \mathbf{R} : \; x > 0\}$

  *Solution:* convex

**Exercise III.2.**

1. Prove the following fact:
   For every $C_i \in \mathbf{S}^m_+$, $i \leq I$, satisfying $\sum_{i \in I} C_i = I_m$ and for every $\lambda_i \in \mathbf{R}$, we have

$$\mathrm{Tr}\left(\left(\sum_{i\in I}\lambda_i C_i\right)^2\right) \leq \mathrm{Tr}\left(\sum_{i\in I}\lambda_i^2 C_i\right).$$

*Solution:* Define $\phi_{ij} := \mathrm{Tr}(C_iC_j)$ so that $\phi_{ij} = \phi_{ji} \geq 0$ as $\mathbf{S}^m_+$ is a self-dual cone (see section D.2.2). Thus,

$$\begin{aligned}
\mathrm{Tr}\left(\left(\sum\nolimits_{i\in I}\lambda_iC_i\right)^2\right) &= \sum\nolimits_{i\in I}\sum\nolimits_{j\in I}\lambda_i\lambda_j\phi_{ij}\\
&= \sum\nolimits_{i\in I}\sum\nolimits_{j\in I}[\lambda_i\sqrt{\phi_{ij}}][\lambda_j\sqrt{\phi_{ij}}]\\
&\leq \left(\sum\nolimits_{i\in I}\sum\nolimits_{j\in I}\phi_{ij}\lambda_i^2\right)^{1/2}\left(\sum\nolimits_{i\in I}\sum\nolimits_{j\in I}\phi_{ij}\lambda_j^2\right)^{1/2}\\
&= \sum\nolimits_{i\in I}\sum\nolimits_{j\in I}\phi_{ij}\lambda_i^2\\
&= \sum\nolimits_{i\in I}\sum\nolimits_{j\in I}\mathrm{Tr}(C_iC_j)\lambda_i^2\\
&= \sum\nolimits_{i\in I}\mathrm{Tr}\left(C_i\sum\nolimits_{j\in I}C_j\right)\lambda_i^2\\
&= \sum\nolimits_{i\in I}\mathrm{Tr}(C_i)\lambda_i^2. \qquad [\text{since} \sum\nolimits_{j\in I}C_j = I_m]
\end{aligned}$$

2. Recall from Example III.14.3 in section 14.2 that for $a_i \geq 0$, $\sum_i a_i > 0$ the function $\ln(\sum_i a_i\exp(\lambda_i))$ is a convex function of $\lambda$. Prove the following matrix analogy of this fact:

   For every $A_i \in \mathbf{S}^m_+$, $1 \leq i \leq I$ such that $\sum_i A_i \succ 0$, the function

   $$f(\lambda) = \ln\mathrm{Det}\left(\sum\nolimits_i \exp(\lambda_i)A_i\right) : \mathbf{R}^I \to \mathbf{R}$$

   is convex.

   *Solution:* Invoking Examples C.7-8 from section C.1.6, we have

   $$\begin{array}{rcl}
   Df(\lambda)[d\lambda] & = & \mathrm{Tr}\left(\left[\sum_i \mathrm{e}^{\lambda_i}A_i\right]^{-1}\left[\sum_i d\lambda_i\mathrm{e}^{\lambda_i}A_i\right]\right)\\
   & & [B_i = \mathrm{e}^{\lambda_i}A_i \succeq 0, B = \sum_i B_i \succ 0, C_i = B^{-1/2}B_iB^{-1/2},\\
   & & \text{ so that } C_i \succeq 0, \sum_i C_i = I_m]\\
   & = & \mathrm{Tr}\left(B^{-1}\sum_i B_id\lambda_i\right) = \mathrm{Tr}\left(\sum_i d\lambda_iC_i\right),\\
   D^2f(\lambda)[d\lambda, d\lambda)] & = & -\mathrm{Tr}\left(B^{-1}\left[\sum_i d\lambda_iB_i\right]B^{-1}\left[\sum_i d\lambda_iB_i\right]\right)\\
   & & +\mathrm{Tr}\left(B^{-1}\sum_i d\lambda_i^2B_i\right)\\
   & = & -\mathrm{Tr}\left(B^{-1/2}\left[\sum_i d\lambda_iB_i\right]B^{-1}\left[\sum_i d\lambda_iB_i\right]B^{-1/2}\right)\\
   & & +\mathrm{Tr}\left(B^{-1/2}\left[\sum_i d\lambda_i^2B_i\right]B^{-1/2}\right)\\
   & = & \mathrm{Tr}\left(\sum_i d\lambda_i^2C_i\right) - \mathrm{Tr}\left(\left[\sum_i d\lambda_iC_i\right]^2\right),\\
   & \geq & 0 \text{ [by item 1]}
   \end{array}$$

   implying that $f$ is convex (Corollary III.14.4).

3. Let $A_i$, $i \leq I$, be as in item 2. Is it true that the function

   $$g(x) = \ln\mathrm{Det}(\sum\nolimits_i x_i^{-1}A_i) : \{x \in \mathbf{R}^I : x > 0\} \to \mathbf{R}$$

   is convex?

   *Solution:* The answer is "yes." Indeed, the function $f$ from item 2 is convex and clearly is nondecreasing in $\lambda_i$; $g$ is obtained from $f$ by convex substitution of the argument $\lambda_i = -\ln(x_i)$, $i \leq I$.

4. Let $B_i$, $i \leq I$, be $m_i \times n$ matrices such that $\sum_i B_i^\top B_i \succ 0$, and let

   $$\Lambda = \{\lambda := (\lambda_1, ..., \lambda_I) : \lambda_i \in \mathbf{S}^{m_i}, \lambda_i \succ 0, i \leq I\}.$$

   Prove that the function

   $$h(\lambda) = \ln\mathrm{Det}\left(\sum\nolimits_i B_i^\top\lambda_i^{-1}B_i\right) : \Lambda \to \mathbf{R}$$

   is convex.

*Solution:* We have

$$
\begin{array}{ll}
& \lambda\in\Lambda, t\geq h(\lambda)\\
\Longleftrightarrow & \exists V\succ 0: V^{-1}\succeq \sum_i B_i^\top\lambda_i^{-1}B_i, -\ln\mathrm{Det}(V)\leq t\\
\Longleftrightarrow & \exists V\succ 0: \left[\begin{array}{c|c|c|c} V^{-1} & B_1^\top & \cdots & B_I^\top\\ \hline B_1 & \lambda_1 & & \\ \hline \vdots & & \ddots & \\ \hline B_I & & & \lambda_I\end{array}\right]\succeq 0, -\ln\mathrm{Det}(V)\leq t\\
& \text{[by Schur Complement Lemma]}\\
\Longleftrightarrow & \exists V\succ 0: -\ln\mathrm{Det}(V)\leq t, \mathrm{Diag}\{\lambda_1,...,\lambda_I\}\succeq [B_1;...;B_I]V[B_1^\top,...,B_I^\top]\\
& \text{[by Schur Complement Lemma]}
\end{array}
$$

Taking into account that the function $-\ln\mathrm{Det}(V):\mathrm{int}\,\mathbf{S}^n_+\to\mathbf{R}$ is convex, we conclude that the epigraph of $h$ is the projection of a convex set in $(t,\lambda,V)$-space onto the subspace of $(t,\lambda)$-variables and is therefore convex. □

5. Let $B_i, i\leq I$, and $\Lambda$ be as in the previous item. Prove that the matrix-valued function

$$F(\lambda)=\left[\sum_i B_i^\top\lambda_i^{-1}B_i\right]^{-1}:\Lambda\to\mathrm{int}\,\mathbf{S}^n_+$$

is $\succeq$-concave, that is, the $\succeq$-hypograph

$$\{(\lambda,Y):\lambda\in\Lambda, Y\preceq F(\lambda)\}$$

of the function is convex.

*Solution:* The values of $F$ on $\lambda$ are positive definite, implying that the set in question is convex if and only if the set

$$\mathcal{E}=\{(\lambda,Y):\lambda\in\Lambda,\exists V\succ 0: Y\preceq V\preceq F(\lambda)\}$$

is convex. When $V\succ 0$ and $\lambda\in\Lambda$, one has

$$V\preceq F(\lambda)\iff V^{-1}\succeq [F(\lambda)]^{-1}=\sum_i B_i^\top\lambda_i^{-1}B_i$$

(Exercise 16), implying that

$$
\begin{array}{l}
\mathcal{E}=\{(\lambda\in\Lambda,Y):\exists V\succ 0: Y\preceq V\preceq F(\lambda)\}=\{(\lambda\in\Lambda,Y):\exists V\succ 0: Y\preceq V, V^{-1}\succeq\sum_i B_i^\top\lambda_i^{-1}B_i\}\\
=\{(\lambda\in\Lambda,Y):\exists V: Y\preceq V, \left[\begin{array}{c|c|c|c} V^{-1} & B_1^\top & \cdots & B_I^\top\\ \hline B_1 & \lambda_1 & & \\ \hline \vdots & & \ddots & \\ \hline B_I & & & \lambda_I\end{array}\right]\succeq 0\},\ \text{[Schur Complement Lemma]}\\
=\{(\lambda\in\Lambda,Y):\exists V\succ 0: V\succeq Y, \mathrm{Diag}\{\lambda_1,...,\lambda_I\}\succeq [B_1;...;B_I]V[B_1^\top,...,B_I^\top]\}\\
\ \text{[Schur Complement Lemma]}
\end{array}
$$

We see that $\mathcal{E}$ is the projection of the convex set in the space of $(\lambda,Y,V)$-variables onto the plane of $(\lambda,Y)$-variables, and thus $\mathcal{E}$ is convex. □

**Exercise III.3.** A function $f$ defined on a convex set $Q$ is called log-convex on $Q$, if it takes real positive values on $Q$ and the function $\ln f$ is convex on $Q$. Prove that

- a log-convex on $Q$ function is convex on $Q$
- the sum (more generally, linear combination with positive coefficients) of two log-convex functions on $Q$ also is log-convex on the set.

*Solution:* If $f(x)=e^{h(x)}$ and $h$ is convex, then so is $f$, as superposition of a convex monotone function $e^x$ and convex function $h$. If $f(x)=\lambda_1f_1(x)+\lambda_2f_2(x)$ with $f_i(x)=e^{h_i(x)}$, where $h_i$ are convex and $\lambda_i>0$, $i=1,2$, then $f(x)=e^{h(x)}$ with $h(x)=\ln(e^{h_1(x)+\ln\lambda_1}+e^{h_2(x)+\ln\lambda_2})$. Since $\ln(e^u+e^v)$ is convex and monotone function of $u,v$, we conclude that $h$ is convex along with $h_1(x)$, $h_2(x)$. □

**Exercise III.4.** [Law of Diminishing Marginal Returns] Consider optimization problem

$$\mathrm{Opt}(r)=\max_x\left\{f(x):G(x)\leq r\ \&\ x\in X\right\}\qquad (P[r])$$

where $X\subset\mathbf{R}^n$ is nonempty convex set, $f(\cdot):X\to\mathbf{R}$ is concave, and $G(x)=[g_1(x);...;g_m(x)]:X\to\mathbf{R}^m$ is vector-function with convex components, and let $\mathcal{R}$ be the set of those $r$ for which $(P[r])$ is feasible. Prove that

1. $\mathcal{R}$ is a convex set with nonempty interior and this set is monotone, meaning that when $r\in\mathcal{R}$ and $r'\geq r$, one has $r'\in\mathcal{R}$.

*Solution:* we clearly have $\mathcal{R}=\cup_{x\in X}\{r:r\geq G(x)\}$, and the right hand side set clearly has nonempty interior and is monotone. To prove that $\mathcal{R}$ is convex, let $r,r'\in\mathcal{R}$ and $\lambda\in[0,1]$. For properly selected $x,x'\in X$ we have $G(x)\leq r$, $G(x')\leq r'$, which combines with convexity of $X$ and $G$ to imply that $\lambda x+(1-\lambda)x'\in X$ and $G(\lambda x+(1-\lambda)x')\leq\lambda r+(1-\lambda)r'$, so that $\lambda r+(1-\lambda)r'\in\mathcal{R}$. □

2. The function $\mathrm{Opt}(r):\mathcal{R}\to\mathbf{R}\cup\{+\infty\}$ satisfies the concavity inequality:

$$\forall(r,r'\in\mathcal{R},\lambda\in[0,1]):\mathrm{Opt}(\lambda r+(1-\lambda)r')\geq\lambda\mathrm{Opt}(r)+(1-\lambda)\mathrm{Opt}(r').\qquad(!)$$

*Solution:* Let $r,r',\lambda$ satisfy the premise in (!). There is nothing to prove when $\lambda=0$ or when $\lambda=1$, so let $\lambda\in(0,1)$. Let us select $s<\mathrm{Opt}(r)$ and $s'<\mathrm{Opt}(r')$, so that there exist $x,x'\in X$ such that $G(x)\leq r$, $f(x)\geq s$. $G(x')\leq r'$, $f(x')\geq s'$. Since $X$ is convex, the components of $G$ are convex, and $f$ is concave, we have

$$\lambda x+(1-\lambda)x'\in X\ \&\ G(\lambda x+(1-\lambda)x')\leq\lambda r+(1-\lambda)r'\ \&\ f(\lambda x+(1-\lambda)x')\geq\lambda s+(1-\lambda)s',$$

implying that $\mathrm{Opt}(\lambda r+(1-\lambda)r')\geq\lambda s+(1-\lambda)s'$. In the resulting inequality, properly selecting $s<\mathrm{Opt}(r)$ and $s'<\mathrm{Opt}(r')$, the right hand side can be made arbitrarily large when $\mathrm{Opt}(r)$ and/or $\mathrm{Opt}(r')$ are $+\infty$, and can be made arbitrarily close to $\lambda\mathrm{Opt}(r)+(1-\lambda)\mathrm{Opt}(r')$ when both $\mathrm{Opt}(r)$ and $\mathrm{Opt}(r')$ are finite, and the concavity inequality follows. □

3. If $\mathrm{Opt}(r)$ is finite at some point $\bar r\in\mathrm{int}\,\mathcal{R}$, then $\mathrm{Opt}(r)$ is real-valued everywhere on $\mathcal{R}$. Moreover, when $X=\mathbf{R}^n$ and $f$ and the components of $G$ are affine, so that $(P[r])$ is an LP program, we can replace in the above claim the inclusion $r\in\mathrm{int}\,\mathcal{R}$ with the inclusion $r\in\mathcal{R}$: in the LP case, the function $\mathrm{Opt}(r)$ is either identically $+\infty$ everywhere on $\mathcal{R}$, or is real-valued at every point of $\mathcal{R}$.

*Solution:* Let $\mathrm{Opt}(r)$ be finite at a point $\bar r\in\mathrm{int}\,\mathcal{R}$ and let $r\in\mathcal{R}$; we need to prove that $\mathrm{Opt}(r)<\infty$. There is nothing to prove when $r=\bar r$, thus assume that $r\neq\bar r$. For properly selected $r_-\in\mathcal{R}$, the point $\bar r$ is a relative interior point of the segment $[r_-,r]$, which combines with the concavity inequality from the previous item to imply that both $\mathrm{Opt}(r_-)$ and $\mathrm{Opt}(r)$ are finite.

Now let $X=\mathbf{R}^n$, $f(x)=f^\top x$ and $G(x)=Ax-b$. Assuming that $\bar r\in\mathcal{R}$ is such that $\mathrm{Opt}(\bar r)<\infty$, we conclude that when $r=\bar r$, the LP program

$$\max_x\{f^\top x:Ax\leq b+r\}\qquad(L[r])$$

is feasible and bounded. By LP duality, it means that the dual problem

$$\min_y\left\{[b+\bar r]^\top y:y\geq0,A^\top y=f\right\}$$

is solvable and therefore feasible. But the feasible set of the LP dual to $(L[r])$ is independent of $r$, implying by Weak Duality that problems $(L[r])$ are above bounded. Thus, in the situation in question $\mathrm{Opt}(r)$ does not take value $+\infty$ at all and therefore is real-valued on $\mathcal{R}$. □

**Comment.** Think about problem $(P[r])$ as about problem where $r$ is the vector of resources you create, and $f(\cdot)$ is your profit, so that the problem is to maximize your profit given your resources and "technological constraints" $x\in X$. Now let $\bar r\in\mathcal{R}$ and $e$ be a nonnegative vector, and let us look what happens when you select your vector of resources on the ray $R=\bar r+\mathbf{R}_+e$, assuming that $\mathrm{Opt}(r)$ on this ray is real-valued. Restricted on this ray, your best profit becomes a function $\phi(t)$ of nonnegative variable $t$:

$$\phi(t)=\mathrm{Opt}(\bar r+te).$$

Since $e \geq 0$, this function is nondecreasing, as it should be: the larger $t$, the more resources you utilize, and the larger is your profit. A not so nice news is that $\phi(t)$ is concave in $t$, meaning that the slope of this function does not increase as $t$ grows. In other words, if it costs you \$1 to pass from resources $\bar{x} + te$ to resources $\bar{x} + (t+1)e$, the return $\phi(t+1) - \phi(t)$ on one extra dollar of your investment goes down (or at least does not go up) as $t$ grows. This is called *The Law of Diminishing Marginal Returns.*

**Exercise III.5.** [follow-up to Exercise III.4] There are $n$ goods $j$ with per-unit prices $c_j > 0$, per-unit utilities $v_j > 0$, and the maximum available amounts $\overline{x}_j$, $j \leq n$. Given budget $R \geq 0$, you want to decide on amounts $x_j$ of goods to be purchased to maximize the total utility of the purchased goods, while respecting the budget and the availability constraints. Pose the problem as LO program and verify that the optimal value $\mathrm{Opt}(R)$ is piecewise linear function of $R$. What are the breakpoints of this function? What are the slopes between breakpoints?

*Solution:* Denoting by $x_j$ the amount of good $j$ we buy, maximizing the total utility becomes the LO program

$$\max_x \left\{ \sum_j v_j x_j : 0 \leq x_j \leq \overline{x}_j \,\forall j, \sum_j c_j x_j \leq R \right\}$$

As is immediately seen (check it!), the optimal solution to the problem is given by the following procedure: we sort the goods to make the ratios $v_j/c_j$ ("utility per \$1 investment") nonincreasing. Assuming this is the case from the very beginning, we start with buying good # 1 until either the available amount of this good, or our budget becomes exhausted, whichever happens first. If this step does not exhaust the budget, we start to buy the second product until either its available amount, or the budget, is exhausted. Then, if we still have money, we start buying product # 3, and proceed in this fashion until either all available goods are bought, or the budget becomes zero. With this strategy, the breakpoints $R_1 < R_2 < ... < R_n$ of $\mathrm{Opt}(R) : [0, \infty) \to \mathbf{R}$ are given by the recurrence

$$R_k = R_{k-1} + \min[R - R_{k-1}, \overline{x}_k / c_k], \ 1 \leq k \leq n,$$

where $R_0 = 0$, and the slope of $\mathrm{Opt}(\cdot)$ on $(R_{k-1}, R_k)$ is $v_k/c_k$; to the right of $R_n$, the slope is zero.

**Exercise III.6**. Let $\beta \in \mathbf{R}^n$ be such that $\beta_1 \geq \beta_2 \geq ... \geq \beta_n$. For $x \in \mathbf{R}^n$, let $x_{(k)}$ be the $k$-th largest entry in $x$. Consider the function

$$f(x) = \sum_k \beta_k x_{(k)} = [\beta_1 - \beta_2] s_1(x) + [\beta_2 - \beta_3] s_2(x) + ... + [\beta_{n-1} - \beta_n] s_{n-1}(x) + \beta_n s_n(x),$$

where, as always, $s_k(x) = \sum_{i=1}^k x_{(i)}$. As we know from Exercise I.29, the functions $s_k(x)$, $k < n$, are polyhedrally representable:

$$t \geq s_k(x) \iff \exists z \geq 0, s : x_i \leq z_i + s, i \leq n, \sum_i z_i + ks \leq t,$$

and $s_n(x)$ is just linear:

$$s_n(x) = \sum_i x_i$$

As a result, $f$ admits the polyhedral representation

$$t \geq f(x) \iff \exists Z = [z_{ik}] \in \mathbf{R}^{n \times (n-1)}, s_k, t_k, k < n :$$
$$\left\{ \begin{array}{l} \forall (i \leq n, k < n) : z_{ik} \geq 0, x_i \leq z_{ik} + s_k, \\ \forall k < n : t_k \geq \sum_i z_{ik} + k s_k \\ t \geq \sum_{k=1}^{n-1} [\beta_k - \beta_{k+1}] t_k + \beta_n \sum_{i=1}^n x_i \end{array} \right.$$

This polyhedral representation has $2n^2 - n$ linear inequalities and $n^2 + n - 2$ extra variables. Now goes the exercise:

1. Find an alternative polyhedral representation of $f$ with $n^2 + 1$ linear inequalities and $2n$ extra variables.

*Solution:* Let $\Pi_n$ be the set of $n \times n$ doubly stochastic matrices. By Birkhoff Theorem, $\Pi_n$ is the convex hull of $n \times n$ permutation matrices, implying that the set $X = \{Px : P \in \Pi_n\}$ is the convex hull of vectors obtained from $x$ by permuting entries, which combines with $\beta_1 \geq,,,. \geq \beta_n$ to imply that

$$f(x) = \max_{P \in \Pi_n} \beta^\top Px,$$

that is, denoting by $\mathbf{e}$ the $n$-dimensional all-ones vector,

$$\begin{array}{rcl} f(x) & = & \max\limits_{P=[P_{ij}]} \left\{\beta^\top Px : P_{ij} \geq 0, P\mathbf{e} = \mathbf{e}, P^\top \mathbf{e} = \mathbf{e}\right\} \\ & = & \min\limits_{\lambda,\mu,[y_{ij}]} \left\{\mathbf{e}^\top[\lambda+\mu] : y_{ij} \geq 0, [\mathbf{e}\lambda^\top + \mu\mathbf{e}^\top]_{ij} - y_{ij} = [\beta x^\top]_{ij}, 1 \leq i,j \leq n\right\} \\ & & \text{[LP Duality]} \\ & = & \min\limits_{\lambda,\mu} \left[\mathbf{e}^\top[\lambda+\mu] : [\mathbf{e}\lambda^\top + \mu\mathbf{e}^\top]_{ij} \geq [\beta x^\top]_{ij},\ 1 \leq i,j \leq n\right\} \end{array}$$

Thus, $f(x)$ admits the polyhedral representation

$$t \geq f(x) \iff \exists \lambda, \mu \in \mathbf{R}^n : t \geq \mathbf{e}^\top[\lambda+\mu], [\mathbf{e}\lambda^\top + \mu\mathbf{e}^\top]_{ij} \geq [\beta x^\top]_{ij},\ 1 \leq i,j \leq n$$

and this representation has $n^2+1$ linear inequalities and $2n$ extra variables.

2. [computational study] Generate at random orthogonal $n \times n$ matrix $U$ and vector $\beta$ with nonincreasing entries and solve numerically the problem

$$\min_x \left\{ f(x) := \sum_k \beta_k x_{(k)} : \|Ux\|_\infty \leq 1 \right\}$$

utilising the above polyhedral representations of $f$. For $n = 8, 16, 32, ..., 1024$, compare the running times corresponding to the 2 representations in question.

*Solution:* In our `CVX` experiments, $U$ and $\beta$ were generated according to

```
[U,D,V]=svd(randn(n,n));beta=-sort(randn(n,1))
```

and the ratio of the CPU time for the "long" polyhedral representation of $f$ in use to the CPU time for the "short" one was as follows:

| $n$ | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |
|---|---|---|---|---|---|---|---|---|
| CPU ratio | 1.67 | 1.35 | 1.71 | 1.92 | 4.26 | 6.39 | 8.19 | 10.29 |

**Exercise III.7.** Let $a \in \mathbf{R}^n$ be a nonzero vector, and let $f(\rho) = \ln(\|a\|_{1/\rho})$, $\rho \in [0,1]$. Moment inequality, see section 17.3.3, states that $f$ is convex. Prove that the function is also nonincreasing and Lipschitz continuous, with Lipschitz constant $\ln n$, or, which is the same, that

$$1 \leq p \leq p' \leq \infty \implies \|a\|_p \geq \|a\|_{p'} \geq n^{\frac{1}{p'}-\frac{1}{p}}\|a\|_p.$$

*Solution:* By homogeneity, it suffices to prove the inequality assuming $\|a\|_p = 1$, and by continuity it suffices to consider the case when $p \leq p' < \infty$. Setting $\alpha_i = |a_i|^p$, we get $1 = \|a\|_p = \left[\sum_i \alpha_i\right]^{1/p}$, $\|a\|_{p'} = \left[\sum_i \alpha_i^{\frac{p'}{p}}\right]^{1/p'}$. In terms of $\alpha_i$ the goal is to prove that when $\alpha_i \geq 0$ sum up to 1, then

$$1 \geq \sum_i \alpha_i^{\frac{p'}{p}} \geq \left[n^{\frac{1}{p'}-\frac{1}{p}}\right]^{p'},$$

which is immediate: due to $p' \geq p$ the function $g(\alpha) = \sum_i \alpha_i^{p'/p}$ is convex on the probabilistic simplex $\{\alpha \in \mathbf{R}^n_+ : \sum_i \alpha_i = 1\}$ and therefore attains its maximum on this simplex at a vertex (Theorem III.15.7), and attains its minimum on the simplex at the barycenter $n^{-1}[1;...;1]$ by Symmetry Principle (Proposition III.15.5 – permutations of coordinates are symmetries of the simplex and of $g(\cdot)$). □

**Exercise III.8.** This Exercise demonstrates power of Symmetry Principle. Consider the situation as follows: you are given noisy observations

$$\omega = Ax + \xi,\ A = \mathrm{Diag}\{\alpha_i, i \leq n\}$$

of unknown signal $x$ known to belong to the unit ball $\mathbf{B} = \{x \in \mathbf{R}^n : \|x\|_2 \leq 1\}$; here $\alpha_i > 0$ are given, and $\xi$ is the standard (zero mean, unit covariance) Gaussian observation noise. Your goal is to recover from this observation the vector $y = Bx$, $B = \mathrm{Diag}\{\beta_i, i \leq n\}$ being given. You intend to recover $y$ by *linear estimate*

$$\widehat{y}_H(\omega) = H\omega,$$

where $H$ is an $n \times n$ matrix you are allowed to choose. For example, selecting $H = BA^{-1} = \mathrm{Diag}\{\beta_i\alpha_i^{-1}\}$, you get an *unbiased* estimate:

$$\mathbf{E}\{\widehat{y}_H(Ax + \xi) - y\} = 0.$$

Let us quantify the quality of a candidate linear estimate $\widehat{y}_H$
— at a particular signal $x \in \mathbf{B}$ - by the quantity

$$\mathrm{Err}_x(H) = \sqrt{\mathbf{E}\{\|\widehat{y}_H(Ax + \xi) - Bx\|_2^2\}},$$

so that $\mathrm{Err}_x^2(H)$ is the expected squared $\|\cdot\|_2$-distance between the estimate and the estimated quantity,
— on the entire set $\mathbf{B}$ of possible signals – by *risk* $\mathrm{Risk}[H] = \max_{x\in\mathbf{B}} \mathrm{Err}_x(H)$.

1. Find closed form expressions for $\mathrm{Err}_x(H)$ and $\mathrm{Risk}(H)$.
2. Formulate the problem of finding the linear estimate with minimal risk as the problem of minimizing a convex function and prove that the problem is solvable, and admits an optimal solution $H^*$ which is diagonal: $H^* = \mathrm{Diag}\{\eta_i, i \leq n\}$.
3. Reduce the problem yielded by item 2 to the problem of minimizing easy-to-compute convex univariate function. Consider the case when $\beta_i = i^{-1}$ and $\alpha_i = [\sigma i^2]^{-1}$, $1 \leq i \leq n$, set $n = 10000$ and fill the following table:

| $\sigma$ | 1.0 | 0.1 | 0.01 | 0.001 | 0.0001 | 0.00001 | 0.000001 |
|---|---|---|---|---|---|---|---|
| $\mathrm{Risk}[H^*]$ | | | | | | | |
| $\mathrm{Risk}[BA^{-1}]$ | | | | | | | |

where $H^*$ is the minimum risk linear estimate as yielded by the solution to univariate problem you end up with, and $\mathrm{Risk}[BA^{-1}]$ is the risk of unbiased linear estimate.
You should see from your numerical results that minimal risk of linear estimation is much smaller than the risk of the unbiased linear estimate. Explain on qualitative level why allowing for bias reduces the risk.

*Solution:* 1: We have

$$\begin{array}{rcl}
\mathrm{Err}_x^2[H] &=& \mathbf{E}\{\|H[Ax+\xi] - Bx\|_2^2\} = \mathbf{E}\{\|[HA-B]x + H\xi\|_2^2\}\\
&=& \mathbf{E}\{\|[B-HA]x\|_2^2 + 2[H\xi]^\top[HA-B]x + [H\xi]^\top[H\xi]\}\\
&=& \|[B-HA]x\|_2^2 + \mathbf{E}\{\xi^\top H^\top H\xi\}\\
&=& \|[B-HA]x\|_2^2 + \mathbf{E}\{\mathrm{Tr}(\xi^\top H^\top H\xi)\} = \|[B-HA]x\|_2^2 + \mathbf{E}\{\mathrm{Tr}(H^\top H\xi\xi^\top)\}\\
&=& \|[B-HA]x\|_2^2 + \mathrm{Tr}(H^\top H\mathbf{E}\{\xi\xi^\top)\} = \|[B-HA]x\|_2^2 + \mathrm{Tr}(H^\top H)
\end{array}$$

and

$$\mathrm{Risk}^2[H] = \max_{x\in\mathbf{B}} \mathrm{Err}_x^2(H) = \mathrm{Tr}(H^\top H) + \max_{x,\|x\|_2\leq 1} \|[B-HA]x\|_2^2 = \mathrm{Tr}(H^\top H) + \|[B-HA]\|^2,$$

where $\|\cdot\|$ is the spectral norm, see section D.1.4.

2: By the solution to item 1, the minimum risk linear estimate is yielded by an optimal solution to the problem

$$\mathrm{Opt} = \min_{H\in\mathbf{R}^{n\times n}} \left[R(H) = \|[B-AH]\|^2 + \mathrm{Tr}(H^\top H)\right]. \qquad (!)$$

the best achievable risk being $\sqrt{\mathrm{Opt}}$. The objective tends to $\infty$ when $\|H\|\to\infty$, implying the existence of solution.

To prove that there exists an optimal solution $H^*$ which is diagonal, let us apply Symmetry Principle (Proposition III.15.5). Consider the set $\mathcal{G}$ of all linear transformations $X\mapsto \overline{G}(X):=GXG:\mathbf{R}^{n\times n}\to\mathbf{R}^{n\times n}$ associated with $2^n$ diagonal $n\times n$ matrices $G$ with diagonal entries $\pm1$. This clearly is a group, and its elements are symmetries of the feasible set $\mathbf{R}^{n\times n}$ of our optimization problem. Let us prove that every transformation $X\to\overline{G}(X)$, $\overline{G}\in\mathcal{G}$, is a symmetry of the objective as well. To this end note that multiplying a matrix from the left and/or from the right by orthonormal matrices, we clearly preserve the spectral norm of the matrix. Therefore for $\overline{G}(\cdot)\in\mathcal{G}$ we have

$$\begin{array}{rcl}
R(\overline{G}(H)) &=& \|B-GHGA\|^2+\mathrm{Tr}([GHG]^\top[GHG])\\
&=& \|GBG-GHG[GAG]\|^2+\mathrm{Tr}(GH^\top G^2HG)\\
&& \text{[due to } B=GBG \text{ and } A=GAG \text{ – } A \text{ and } B \text{ are diagonal, and } G=G^\top]\\
&=& \|G[B-HA]G\|^2+\mathrm{Tr}(GH^\top HG)\ \text{[due to } G^2=I_n]\\
&=& \|B-HA\|^2+\mathrm{Tr}(HH^\top)=R(H)\\
&& \text{[we have used orthonormality of } G \text{ and the relation } \mathrm{Tr}(GH^\top HG)\\
&& =\mathrm{Tr}(H^\top HG^2)=\mathrm{Tr}(H^\top H) \text{ due to } G^2=I_n.]
\end{array}$$

By Symmetry Principle, the (solvable) convex problem in question has an optimal solution $H^*$ which is $\mathcal{G}$-symmetric: $GH^*G=H^*$ for all diagonal $G$ with diagonal entries $\pm1$. Observing that $[GH^*G]_{ij}=G_{ii}H^*_{ij}G_{jj}$, we get that $G_{ii}H^*_{ij}G_{jj}=H^*_{ij}$ for all $i,j$ whenever $G_{kk}=\pm1$ for all $k$, implying that $H^*_{ij}=-H^*_{ij}$ when $i\neq j$, that is, $H^*_{ij}=0$ when $i\neq j$, so that $H^*$ is diagonal. □

3: By item 2, we do not spoil the optimal value in (!) when restricting ourselves with diagonal candidate solutions $H=\mathrm{Diag}\{\eta_i\}$, thus arriving at the problem

$$\begin{array}{rcl}
\mathrm{Opt} &=& \min_{\eta_i,i\leq n}\left[\max_i[\beta_i-\eta_i\alpha_i]^2+\sum_i\eta_i^2\right]\\
&=& \min_{\rho,\{\eta_i\}}\left[\rho^2+\sum_i\eta_i^2:|\beta_i-\alpha_i\eta|\leq\rho\right]\\
&=& \min_{\rho\geq0:}\left\{\rho^2+\sum_i\left[\frac{\max[\rho-|\beta_i|,0]}{\alpha_i}\right]^2\right\}
\end{array}\qquad(*)$$

In the case when $\beta_i=i^{-1}$ and $\alpha_i=[\sigma i^2]^{-1}$, $1\leq i\leq n$, and $n=10000$ one has

| $\sigma$ | 1.0 | 0.1 | 0.01 | 0.001 | 0.0001 | 0.00001 | 0.000001 |
|---|---|---|---|---|---|---|---|
| Risk$[H^*]$ | 0.7071 | 0.28244 | 0.1124 | 0.04474 | 0.01781 | 0.00709 | 0.00282 |
| Risk$[BA^{-1}]$ | 1.827e4 | 1.827e3 | 1.827e2 | 1.827e1 | 1.827e0 | 0.1827 | 0.01827 |

where $H^*$ is the minimum risk linear estimate as yielded by the solution to $(*)$, and Risk$[BA^{-1}]$ is the risk of unbiased linear estimate.

Unbiased recovery in the case of diagonal $B$ and $A$ recovers an entry $y_i$ in $y$ as

$$\widehat{y}_i=\frac{\beta_i}{\alpha_i}\omega_i=y_i+\frac{\beta_i}{\alpha_i}\xi_i.$$

We see that while the recovery is unbiased, it significantly amplifies the noise, provided that $\beta_i/\alpha_i$ is large (with our data, this ratio is $\sigma i$ and indeed is large when $i\gg1/\sigma$). On the other hand, we know in advance that $x$ is bounded by 1 in $\|\cdot\|_2$, so that when $\beta_i$ is small for some $i$, the bias in recovering $y_i$ will be small even when we recover $y_i$ by 0. The optimal linear estimate heavily utilizes our a priori information $\|x\|_2\leq1$ to find optimal tradeoff between the bias and the stochastic component of the recovery error $y_i-\widehat{y}_i$, this is why it not just beats the unbiased linear estimate (this always is the case – the latter estimate is linear!), but may beat it by huge margin, For example, the above table shows that unbiased estimate makes no sense when $\sigma\geq$ 1.e–4 – knowing in advance that $\|x\|_2\leq1$, we can estimate $x$ by 0 with risk 1 (which does not require observations at all), which is better than the risk 1.82... of the unbiased linear estimate when $\sigma=$ 1.e–4 .

**Exercise III.9.** Given the sets of $d$-dimensional tentative nodes ($d=2$ or $d=3$) and of tentative bars of a TTD problem satisfying assumption $\mathfrak{R}$, let $\mathcal{V}=\mathbf{R}^M$ be the space of virtual displacements of the nodes, $N$ be the number of tentative bars, and $W>0$ be the allowed total bar volume, see Exercise I.16. Let, next, $\mathcal{C}(t,f):\mathbf{R}^N_+\times\mathcal{V}\to\mathbf{R}\cup\{+\infty\}$ be the compliance of truss $t\geq0$ w.r.t. load $f$ (we identify trusses with the corresponding vectors $t$ of bar volumes). Prove that

1. $\mathcal{C}(t,f)$ is a convex lsc function, positively homogeneous of homogeneity degree 1, of $[t;f]$ with $\mathbf{R}^N_{++}\times\mathcal{V}\subset\mathrm{Dom}\,\mathcal{C}$, where $\mathbf{R}^N_{++}=\mathrm{int}\,\mathbf{R}^N_+=\{t\in\mathbf{R}^N:t>0\}$. This function is positively homogeneous, with degree -1, in $t$, when $f$ is fixed, and positively homogeneous, of degree 2, in $f$ when $t$ is fixed. Besides this, $\mathcal{C}(t,f)$ is nonincreasing in $t\geq 0$: if $0\leq t'\leq t$, then $\mathcal{C}(t,f)\leq\mathcal{C}(t',f)$ for every $f$.

   *Solution:* As we know from Exercise I.16.2, the epigraph of $\mathcal{C}$ is
$$\mathrm{epi}\{\mathcal{C}\}:=\{[t;f;\tau]:\tau\geq\mathcal{C}(t,f)\}=\{[t;f;\tau]:t\geq 0,\mathcal{A}(t,f,\tau):=\left[\begin{array}{c|c}B\,\mathrm{Diag}(t)B^\top & f\\ \hline f^\top & 2\tau\end{array}\right]\succeq 0\}\qquad(!)$$
   with given by the data $M\times N$ matrix $B$ satisfying $BB^\top\succ 0$ (the latter is our default assumption $\mathfrak{R}$). We see that $\mathrm{epi}\{\mathcal{C}\}$ is closed convex set, implying that $\mathcal{C}$ is a convex lsc function (Proposition III.16.2). The inclusion $\mathbf{R}^N_{++}\times\mathcal{V}\subset\mathrm{Dom}\,\mathcal{C}$ is readily given by positive definiteness of the matrix $A(t)=B\,\mathrm{Diag}(t\}B^\top$ for positive $t$'s; whenever $A(t)\succ 0$, the matrix $\mathcal{A}(t,f,\tau)$ is, for every $f$, positive semidefinite whenever $\tau$ is large enough by the Schur Complement Lemma. Positive homogeneity, of degree 1, of $\mathcal{C}$ clearly follows from the fact that by the above description, $\mathrm{epi}\{\mathcal{C}\}$ is a closed cone. By (!) combined with the Schur Complement Lemma, whenever $[t;f;\tau]\in\mathrm{epi}\{\mathcal{C}\}$ and $\lambda>0,\mu$ are reals, we have $[\lambda t;\mu f;\lambda^{-1}\mu^2\tau]\in\mathrm{epi}\{\mathcal{C}\}$, implying the claims about homogeneity of $\mathcal{C}$ w.r.t. $t$ and w.r.t. $f$. Finally, by the same (!) when $[t';f,\tau]\in\mathrm{epi}\{\mathcal{C}\}$ and $t\geq t'$, we have $[t;f;\tau]\in\mathrm{epi}\{\mathcal{C}\}$ as well, implying that $\mathcal{C}(t,\tau)$ is nonincreasing in $t\geq 0$. □

2. The function $\mathrm{Opt}(W,f)=\inf_t\left\{\mathcal{C}(t,f):t\geq 0,\sum_i t_i=W\right\}$ – the optimal value in the TTD problem (5.2) – with $W$ restricted to reside in $\mathbf{R}_{++}=\{W>0\}$ is convex continuous function with the domain $\mathbf{R}_{++}\times\mathcal{V}$. This function is positively homogeneous, of degree -1, in $W>0$ and homogeneous, of homogeneity degree 2, in $f$:
$$\forall(\lambda>0,\mu):\mathrm{Opt}(\lambda W,\mu f)=\lambda^{-1}\mu^2\mathrm{Opt}(W,f),\ \forall(W,f)\in\mathbf{R}_{++}\times\mathcal{V}.$$
   Moreover, the infimum in $\inf_t\left\{\mathcal{C}(t,f):t\geq 0,\sum_i t_i=W\right\}$ is achieved whenever $W>0$.

   *Solution:* Consider the set
$$\mathcal{G}=\{[t;f;\tau;W]:W=\sum_i t_i,t\geq 0,\mathcal{A}(t,f,\tau)\succeq 0\}$$
   This set clearly is a closed convex cone, and the function
$$F(t,f,\tau,W)=\left\{\begin{array}{ll}\tau, & [t;f;\tau;W]\in\mathcal{G}\ \&\ W>0\\ +\infty, & \text{otherwize}\end{array}\right.$$
   is convex and nonnegative on this set. We clearly have
$$\mathrm{Opt}(W,f)=\inf_{t,\tau}F(t,f,\tau,W),\qquad(!!)$$
   which combines with convexity and nonnegativity of $F$ and the rule on partial minimization (stability of convexity w.r.t. partial minimization, section 14.1) to imply that $\mathrm{Opt}(W,f)$ is convex (and of course nonnegative) function of $(W,f)$. Moreover, by Exercise I.16.3, for every $W>0$
$$\inf_t\{\mathcal{C}(t,f):t\geq 0,\sum_i t_i=W\}=\inf_{t:t\geq 0,\sum_i t_i=W}\ \min_\tau\{\tau:\mathcal{A}(t,f,\tau)\succeq 0\}$$
   is achieved, implying, first, the "Moreover" claim of the statement we are justifying, and, second, the fact that $\mathrm{Opt}(W,f)$ is finite whenever $W>0$. Thus, $\mathrm{Opt}(W,f)$ is convex real-valued function in the domain $\mathbf{R}_{++}\times\mathcal{V}$, and since this domain is convex and open, Opt is continuous in this domain, as claimed. Homogeneity properties of this function we have announced are immediate consequences of the fact that $\mathcal{G}$ is a closed cone and that by Schur Complement Lemma for $\lambda>0,\mu\neq 0$ the matrices $\mathcal{A}(t,f,\tau)$ and $\mathcal{A}(\lambda t,\mu f,\lambda^{-1}\mu^2\tau)$ simultaneously are/are not positive semidefinite. □

3. When on certain bridge there is just one car, of unit weight, the compliance of the bridge does not exceed 1, whatever be the position of the car. How large could the compliance of the bridge when there are 100 cars of total weight 70 on it?

*Solution:* The compliance is at most 4900. Indeed, a 100-force load with the total of forces' magnitudes 1 is a convex combination of loads with single force of magnitude 1 [8]. As we know from item 1, the compliance of a given truss is a convex function of the load, implying by Jensen's inequality that when our bridge is loaded by 100 (or any other number of) cars of total weight 1, the compliance does not exceed the maximum of compliances caused by a single car of unit weight, the maximum being taken over possible positions of this single car. For our bridge, this maximum is $\leq 1$, implying that the compliance of the bridge loaded by a whatever number of cars of total weight 1 does not exceed 1. It remains to note that the compliance is positively homogeneous, of degree 2, function of load, so that with the total weight of cars not exceeding 70, the compliance does not exceed $70^2 = 4900$.

To formulate the next two tasks, let us associate with a free node $p$ the set $\mathcal{F}^p$ of all single-force loads stemming from forces $g$ of magnitude $\|g\|_2$ not exceeding 1 and acting at node $p$. For a set $S$ of free nodes, $\mathcal{F}^S$ is the set of all loads with nonzero forces acting solely at the nodes from $S$ and with the sum of $\|\cdot\|_2$-magnitudes of the forces not exceeding 1, so that

$$\mathcal{F}^S = \mathrm{Conv}(\cup_{p\in S}\mathcal{F}^p)$$

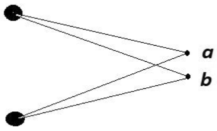
(why?)

4. Let $S = \{p_1, ..., p_K\}$ be a $K$-element collection of free nodes from the nodal set. Assume that for every node $p$ from $S$ and every load $f \in \mathcal{F}^p$ there exists a truss of a given total weight $W$ such that its compliance w.r.t. $f$ does not exceed 1. Which, if any, of the following statements are true?
   (i) For every load $f \in \mathcal{F}^S$, there exists a truss of total volume $W$ with compliance w.r.t. $f$ not exceeding 1
   (ii) There exists a truss of total volume $W$ with compliance w.r.t. every load from $\mathcal{F}^S$ not exceeding 1
   (iii) For properly selected $\gamma$ depending solely on $d$, there exists a truss of total volume $\gamma KW$ with compliance w.r.t. every load from $\mathcal{F}^S$ not exceeding 1

   *Solution:* The first and the third claims are correct, the second, in general, is wrong. Indeed, let $\mathcal{C}(t, f)$ be the compliance of truss $t$ w.r.t. load $f$; as we know from item 1, this is a convex function of $(t, f)$.

   To justify the first claim, given load $f \in \mathcal{F}^S$, we can find its representation $f = \sum_k \lambda_k f^k$ as a convex combination of loads $f^k \in \mathcal{F}^{p_k}$. By assumption on $S$, for every $k$ there exists truss $t^k$ of total volume $W$ such that $\mathcal{C}(t^k, f^k) \leq 1$, implying by convexity that $\mathcal{C}(t := \sum_k \lambda_k t^k, f) \leq 1$. Since the total volume of $t$ is $W$, $t$ is the truss announced in claim 1.

   To demonstrate that the second claim is wrong in general, consider planar sets of tentative nodes and bars depicted on Figure VI.5,



Fugure VI,5. Tentative nodes and bars.

where bold circles are fixed nodes, and $S = \{a, b\}$. Denoting by $\mathcal{T} = \{t \in \mathbf{R}^4_+ : \sum_i t_i = W\}$ the set of all trusses of total volume $W$, we can assume w.l.o.g. that

$$\max_{f\in\mathcal{F}^a}\min_{t\in\mathcal{T}}\mathcal{C}(t, f) = 1,$$

[8] "What is meant is not always put into writing" ("Boris Godunov" by Alexander Pushkin): we tacitly assume that possible locations of cars are among the nodes of the bridge, modeled as a truss, and that the stemming from cars forces acting at the bridge "look down."

note that by symmetry we also have

$$\max_{f\in\mathcal{F}^b}\min_{t\in\mathcal{T}}\mathcal{C}(t,f)=1,$$

so that we are in the situation postulated in item 4. Let $f^a\in\mathcal{F}^1$ and $f^b\in\mathcal{F}^2$ be the "critical loads" – those where the respective $\max_f$ are achieved. Clearly, there is no truss $\bar{t}$ of total volume $W$ with $\mathcal{C}(\bar{t},f^a)\leq 1$ and $\mathcal{C}(\bar{t},f^b)\leq 1$ – were it existing, with our TTD data there clearly would exist truss $t$ of total volume $W/2$ with compliance w.r.t one of the loads, $f^a$ or $f^b$, not exceeding 1. It would imply the existence of truss of total volume $W$ with compliance w.r.t. either $f^a$, or $f^b$ not exceeding 1/2 (recall that $\mathcal{C}(t,f)$ is homogeneous, of degree -1, with respect to $t$), which is not the case, since both loads are critical.

To justify the third claim, note that for every integer $\mu\geq d+1$ the unit $\|\cdot\|_2$-ball $B$ in $\mathbf{R}^d$ is contained in the convex hull $\Delta=\mathrm{Conv}(\{r_{d,\mu}g^\iota, 1\leq\iota\leq d+1\})$ of $\mu$ vectors of the $\|\cdot\|_2$-norm $r_{d,\mu}$ each, $g^\iota$ being unit normalizations of these vectors. For example, when $d=2$, specifying $\Delta$ as $\mu$-side perfect polygon circumscribed around the unit circle and the vectors $g^\iota$ as the unit length normalization of the vertices of the polygon, we get $r_{d,\mu}=1/\cos(\pi/\mu)$.

Let us specify $f^{k\iota}$, $1\leq k\leq K, 1\leq\iota\leq\mu$ as single-force load with force $g^\iota$ acting at $k$-th node, $p_k$, of $S$, $1\leq k\leq K$, $1\leq\iota\leq\mu$, so that $f^{k\iota}\in\mathcal{F}^{p_k}$. Under the premise of item 4, there exist trusses $t^{k\iota}$ of total volume $W$ each such that $\mathcal{C}(t^{k,\iota},f^{k,\iota})\leq 1$. Let $\underline{t}=\sum_{k,\iota}t^{k\iota}$. Since $\mathcal{C}(t,f)$ clearly is nondecreasing in $t\geq 0$, we have $\mathcal{C}(\underline{t},f^{k\iota})\leq 1$ for all $k,\iota$, whence, by convexity of $\mathcal{C}(t,f)$ in $f$, $\mathcal{C}(\underline{t},f)\leq 1$ for all $f\in U:=\mathrm{Conv}(\{f^{k\iota},k\leq K,\iota\leq\mu\})$. Due to the origin of $g^\iota$, we have $U\supset r_{d,\mu}^{-1}\mathcal{F}^{p_k}\ \forall k\leq K$, implying that $U\supset\mathrm{Conv}(\cup_k r_{d,\mu}^{-1}\mathcal{F}^{p_k})=r_{d,\mu}^{-1}\mathcal{F}^S$. Thus, $\mathcal{C}(\underline{t},f)\leq 1$ for all $f\in r_{d,\mu}^{-1}\mathcal{F}^S$. By homogeneity of $\mathcal{C}(t,f)$ w.r.t. $f$ and to $t$, it follows that $\mathcal{C}(r_{d,\mu}^2\underline{t},f)\leq 1$ for all $f\in\mathcal{F}^S$, so that the compliance of the truss $\bar{t}=r_{d,\mu}^2\underline{t}$ w.r.t. every load from $\mathcal{F}^S$ does not exceed 1. It remains to note that the total volume of $\bar{t}$ is $\underbrace{r_{d,\mu}^2\mu}_{\gamma}KW$. We can now try different values of $\mu$ in order to minimize the factor $\gamma$ over $\mu$ (and over geometry of $\Delta$). For $d=2$, restricting ourselves with perfect $\mu$-side polygons $\Delta$ circumscribed around the unit circle, the best $\mu$ is 5, resulting in $\gamma=5/\cos^2(\pi/5)\approx 7.6393$. When $d=3$, we restricted our search with Platonian solids $\Delta$ circumscribed around the unit 3D ball. The best solid was the octahedron, resulting in $\mu=6$ and $\gamma=18$. □

★5. Prove the following statement:

> *In the situation of item 4 above, let $\gamma=4$ when $d=2$ and $\gamma=7$ when $d=3$. For every $k\leq K$ there exists a truss $\widehat{t}^k$ of total volume $\gamma W$ such that the compliance of $t$ w.r.t. every load from $\mathcal{F}^{p_k}$ does not exceed 1. As a result, there exists truss $\widetilde{t}$ of total volume $\gamma KW$ with compliance w.r.t. every load from $\mathcal{F}^S$ not exceeding 1.*

*Solution:* Given $\epsilon\in(0,1)$, let $\mathcal{T}_\epsilon=\{t\in\mathbf{R}^N_+:\sum_i t_i=W, t_i\geq\epsilon W/N\,\forall i\}$.

$\mathbf{1}^o$ Observe that for every $f\in\mathcal{F}^{p_k}$ there exists truss $t\in\mathcal{T}_\epsilon$ such that $\mathcal{C}(t,f)\leq(1-\epsilon)^{-1}$. Indeed, given $f\in\mathcal{F}^{p_k}$, there exists truss $t^f$ of total volume $W$ such that $\mathcal{C}(t^f,f)\leq 1$. Setting $\underline{t}=\epsilon[W/N;W/N;...W/N]$ and $\bar{t}=(1-\epsilon)t^f+\underline{t}$, we get a truss from $\mathcal{T}_\epsilon$ satisfying $\bar{t}\geq(1-\epsilon)t^f$. Since $\mathcal{C}(t,f)$ is nonincreasing and positively homogeneous, of degree -1, in $t>0$, we conclude that $\mathcal{C}(\bar{t},f)\leq(1-\epsilon)^{-1}$, as claimed.

$\mathbf{2}^o$ Let us fix a free node $p$ of the nodal set, and let $f_g$, $g\in\mathbf{R}^d$, stand for single-force load where force $g$ acts at node $p$. Recall that vectors from the space $\mathcal{V}=\mathbf{R}^M$ of nodal displacements are block vectors with $d$-dimensional blocks representing "physical displacements" of free nodes and indexed by these nodes. Let $I_p$ be the set of indexes of those entries in a vector of nodal displacements which correspond to the block indexed by $p$.

When $t\in\mathcal{T}_\epsilon$, the stiffness matrix $A(t)=B\,\mathrm{Diag}\{t\}B^\top$ of truss $t$ is positive definite (assumption $\mathfrak{R}$), so that the equilibrium displacement of truss $t$ under a load $f$ is $v=[A(t)]^{-1}f$, and the compliance is

$$\mathcal{C}(t,f)=\frac{1}{2}v^\top f=\frac{1}{2}f^\top A^{-1}(t)f.$$

As a result, for every $g \in \mathbf{R}^d$ and every truss $t \in \mathcal{T}_\epsilon$ one has

$$\mathcal{C}(t, f_g) = \frac{1}{2} g^\top [A^{-1}(t)]_{I_p} g,$$

where for $Q = [Q_{i,j}]_{i,j \leq M} \in \mathbf{S}^M$, $[Q]_{I_p} = [Q_{ij}]_{i,j \in I_p} \in \mathbf{S}^d$ is the $d \times d$ principal submatrix of $Q$ corresponding to rows and columns with indexes from $I_p$.

Next, let $\underline{A} = A(\underline{t})$, so that $0 \prec \underline{A} \preceq A(t)$ for every $t \in \mathcal{T}_\epsilon$ due to $t \geq \underline{t} > 0$. It follows that for all $t \in \mathcal{T}_\epsilon$ it holds $A^{-1}(t) \preceq \underline{A}^{-1}$, whence

$$\forall t \in \mathcal{T}_\epsilon : [A^{-1}(t)]_{I_p} \preceq \overline{Q} := [\underline{A}^{-1}]_{I_p}.$$

**3**$^o$ Let us associate with node $p$ the set $\mathcal{S}_p \subset \mathbf{S}^d_+$ given by

$$\mathcal{S}_p = \{Q \in \mathbf{S}^d : \exists t \in \mathcal{T}_\epsilon : [A^{-1}(t)]_{I_p} \preceq Q \preceq \overline{Q}\}.$$

We claim that $\mathcal{S}_p$ is a convex compact set in $\mathbf{S}^d_+$. The main component of verification is the following simple observation to be justified at the end of the proof:

> (@) *Given symmetric positive definite $M \times M$ matrix $\underline{A}$ and a $d$-element subset $I$ of its row indexes and denoting by $[C]_I$ the $d \times d$ principal submatrix of $C \in \mathbf{S}^M$ composed of rows and columns with indexes from $I$, the set*
>
> $$\mathcal{S}^I = \{(Q, A) \in \mathbf{S}^d \times \mathbf{S}^M : A \succeq \underline{A} \ \&\ Q \succeq [A^{-1}]_I\}$$
>
> *is closed and convex.*

By (@), the set

$$\mathcal{S}^+_p = \left\{(Q, t) \in \mathbf{S}^d \times \mathcal{T}_\epsilon : Q \succeq [A^{-1}(t)]_{I_p}\right\}$$

is closed and convex (as the inverse image of closed and convex set $\mathcal{S}^{I_p}$ under the linear mapping $(Q, t) \mapsto (Q, A(t))$. Consequently, the projection $\mathcal{S}[p]$ of $\mathcal{S}^+_p$ onto the $Q$-space is convex, so that $\mathcal{S}_p$ is convex as well – this is the intersection of $\mathcal{S}[p]$ with the convex set $\{Q : Q \preceq \overline{Q}\}$. $\mathcal{S}_p$ clearly is bounded – it is contained in the set $\{0 \preceq Q \preceq \overline{Q}\}$. To see that $\mathcal{S}_p$ is closed, let $Q_i \in \mathcal{S}_p$ converge to $Q$ as $i \to \infty$, and let us prove that $Q \in \mathcal{S}_p$, that is, that $Q \preceq \overline{Q}$ (which is evident) and that $Q \succeq [A^{-1}(t)]_{I_p}$ for some $t \in \mathcal{T}_\epsilon$. To see that the latter is the case, recall that for every $i$ there exist $t^i \in \mathcal{T}_\epsilon$ such that $Q_i \succeq [A^{-1}(t^i)]_{I_p}$. Taking into account that $\mathcal{T}_\epsilon$ is compact and passing to a subsequence we can assume that $\lim_{i \to \infty} t^i$ exists; this limit clearly can be taken as the desired $t$. Thus, $\mathcal{S}_p$ is convex compact set.

**4**$^o$ Now – the main step. Assume that $p$ is a node from $S$ and that we are under the premise of item 4, that is, for every $g \in \mathbf{R}^d$ with $\|g\|_2 \leq 1$ there exists a truss $t'$ of total volume $W$ such that $\mathcal{C}(t', f_g) \leq 1$. Invoking 1$^o$, we conclude that

$$\forall (g \in \mathbf{R}^d, \|g\|_2 \leq 1) \exists t \in \mathcal{T}_\epsilon : \mathcal{C}(t, f_g) \leq (1 - \epsilon)^{-1} \qquad (\#)$$

Denoting $B = \{g \in \mathbf{R}^d : \|g\|_2 \leq 1\}$, consider the family of sets

$$\mathcal{Q}[g] = \{Q \in \mathcal{S}_p : \frac{1}{2} g^\top Q g \leq \gamma (1 - \epsilon)^{-1}\}$$

parameterised by vectors $g \in B$. By their origin, the sets from the family are convex and compact. The crucial fact which we are about to prove is that

> (!!) *Every $\gamma$ of sets from the family have a point in common.*

To prove (!!), let $g_\ell \in B$, $1 \leq \ell \leq \gamma$, and let us prove that the sets $\mathcal{Q}[g_\ell]$, $\ell = 1, ..., \gamma$, have a point in common. For every $\ell$, by (#), there exists $t^\ell \in \mathcal{T}_\epsilon$ such that $\mathcal{C}(t^\ell, f_{g_\ell}) \leq (1 - \epsilon)^{-1}$, implying that setting

$$Q_\ell = [A^{-1}(t^\ell)]_{I_p},$$

we get

$$Q_\ell \preceq \overline{Q} \ \& \ \frac{1}{2} g_\ell^\top Q_\ell g_\ell = \frac{1}{2} f_{g_\ell}^\top A^{-1}(t^\ell) f_{g_\ell} \leq (1-\epsilon)^{-1}$$

Now let $t = \frac{1}{\gamma}\sum_{\ell=1}^{\gamma} t^\ell$ and $Q = [A^{-1}(t)]_{I_p}$, so that $t \in \mathcal{T}_\epsilon$, $(Q,t) \in \mathcal{S}_p^+$, and $Q \preceq \overline{Q}$, implying that $Q \in \mathcal{S}_p$. For every $\ell$ we have

$$\begin{array}{l} t \geq \frac{1}{\gamma} t^\ell > 0 \implies A(t) \succeq \frac{1}{\gamma} A(t^\ell) \implies A^{-1}(t) \preceq \gamma A^{-1}(t^\ell) \\ \implies Q \preceq \gamma Q_\ell \implies \frac{1}{2} g_\ell^\top Q g_\ell \leq \gamma g_\ell^\top Q_\ell g_\ell \leq \gamma(1-\epsilon)^{-1} \end{array}$$

Thus, $Q \in \mathcal{S}_p$ and

$$\frac{1}{2} g_\ell^\top Q g_\ell \leq \gamma(1-\epsilon)^{-1} \ \forall \ell \leq \gamma$$

implying that $Q \in \mathcal{Q}[g_\ell]$ for all $\ell \leq \gamma$, that is, $\cap_{\ell \leq \gamma} \mathcal{Q}[g_\ell]$ is nonempty, as claimed.

$\mathbf{5}^o$ The rest of the proof is easy. Note that $\mathcal{Q}[g]$ are convex compact subsets of $\mathbf{S}^d$ and the latter linear space has dimension $\gamma - 1$. Applying Helly Theorem II, there exists a common point $Q$ of all the sets $\mathcal{Q}[g]$, $g \in B$. Due to $\mathcal{Q}[g] \subset \mathcal{S}_p$ for every $g \in B$, we get $Q \in \mathcal{S}_p$, so that there exists $\bar{t} \in \mathcal{T}_\epsilon$ such that

$$Q \succeq [A^{-1}(\bar{t})]_{I_p}$$

Consequently,

$$\forall g \in B : \mathcal{C}(\bar{t}, f_g) = \frac{1}{2} f_g^\top A^{-1}(\bar{t}) f_g = \frac{1}{2} g^\top [A^{-1}(\bar{t})]_{I_p} g \leq \frac{1}{2} g^\top Q g \leq \gamma(1-\epsilon)^{-1},$$

where the concluding inequality is due to $Q \in \mathcal{Q}[g]$. Thus, there exists truss $\bar{t} \in \mathcal{T}_\epsilon$ such that the compliance of this truss w.r.t. every load $f_g$, $g \in B$, is $\leq \gamma(1-\epsilon)^{-1}$.

$\mathbf{6}^o$ The remaining reasoning is quite straightforward. The truss $\bar{t}$ we have build depends on $\epsilon$; setting $\epsilon_i = 1/(i+1)$, $i = 1, 2, ...$ let us denote by $\bar{t}^i$ the truss given by the above construction as applied with $\epsilon = \epsilon_i$. All these trusses are of total volume $W$, and passing to a subsequence, we can assume that $\bar{t}^i \to \bar{t}$ as $t \to \infty$. Due to $\mathcal{C}(\bar{t}^i, f_g) \leq \gamma(1-\epsilon)^{-1}$, we have

$$\forall i \forall g \in B : \left[\begin{array}{c|c} B\,\mathrm{Diag}\{\bar{t}^i\} B^\top & f_g \\ \hline f_g^\top & 2\gamma(1-\epsilon)^{-1} \end{array}\right] \succeq 0,$$

implying that

$$\forall g \in B : \left[\begin{array}{c|c} B\,\mathrm{Diag}\{\bar{t}\} B^\top & f_g \\ \hline f_g^\top & 2\gamma \end{array}\right] \succeq 0$$

that is, $\mathcal{C}(\bar{t}, f_g) \leq \gamma$ for all $g \in B$. Besides this, truss $\bar{t}$, same as all trusses $\bar{t}^i$, is of total volume $W$. Setting $\widehat{t} = \gamma t$, we get truss of total volume $\gamma W$ and compliance, w.r.t. every load $f_g$ with $\|g\|_2 \leq 1$, not exceeding 1.

Summing up the $K$ trusses given by the above construction as applied to every one of the $K$ free nodes composing the set $S$, we get a truss $\widetilde{t}$ of total volume $\gamma K W$ with compliance w.r.t. every load from $\mathcal{F}^S$ not exceeding 1.

**Paying debts: proof of (@).** Closedness of $\mathcal{S}^I$ is evident; all we need is to prove that the set is convex.

Let us make the following

> **Observation:** The mapping $X \mapsto X^{-1} : \mathrm{int}\,\mathbf{S}^M_+ \to \mathrm{int}\,\mathbf{S}^M_+$ is $\succeq$-convex: whenever $X, Y \in \mathrm{int}\,\mathbf{S}^M_+$ and $\lambda \in [0,1]$, one has $[\lambda X + ((1-\lambda)Y]^{-1} \preceq \lambda X^{-1} + (1-\lambda)Y^{-1}$.

The "'bare hands" proof of this important fact (to be put into perspective in Part IV) is as follows. For $X, Y \succ 0$ we have, by Schur Complement Lemma, $\left[\begin{array}{c|c} X^{-1} & I \\ \hline I & X \end{array}\right] \succeq 0$, $\left[\begin{array}{c|c} Y^{-1} & I \\ \hline I & Y \end{array}\right] \succeq 0$, whence

$\left[\begin{array}{c|c}\lambda X^{-1}+(1-\lambda)Y^{-1} & I\\ \hline I & \lambda X+(1-\lambda)Y\end{array}\right]\succeq 0$, implying, by the same Schur Complement Lemma, that $\lambda X^{-1}+(1-\lambda)Y^{-1}\succeq[\lambda X+(1-\lambda)Y]^{-1}$.

Due to Observation, the mapping $X\mapsto[X^{-1}]_I:\mathrm{int}\,\mathbf{S}^M_+\to\mathrm{int}\,\mathbf{S}^d_+$ is $\succeq$-convex:

$$\begin{array}{l}X\succ 0,Y\succ 0,\lambda\in[0.1]\implies[\lambda X+(1-\lambda)Y]^{-1}\preceq\lambda X^{-1}+(1-\lambda)Y^{-1}\\ \implies[(\lambda X+(1-\lambda)Y)^{-1}]_I\preceq[\lambda X^{-1}+(1-\lambda)Y^{-1}]_I=\lambda[X^{-1}]_I+(1-\lambda)[Y^{-1}]_I\\ \text{[since principal submatrix of a positive semidefinite matrix is positive semidefinite as well]}\end{array}$$

which clearly implies the convexity of $\mathcal{S}^I=\{(Q,A):A\succeq\underline{A},Q\succeq[A^{-1}]_I\}$. □

Some remarks are in order.

1. A careful reader hopefully recognizes the "driving force" behind the above proof – it is the same as the one used in essentially less technical, and thus much more transparent section 2.3.1.
2. In the above proof, it was completely unimportant that $B$ was the unit ball of $\|\cdot\|_2$ – it could be a whatever nonempty subset of $\mathbf{R}^d$. In fact the proof justifies the following claim:

   *Given (1) the data of a TTD problem satisfying assumption $\mathfrak{R}$ and with nodes "living" in $\mathbf{R}^d$ ($d=2$ or $d=3$), (2) a collection of $K$ free nodes $p_k$, $k\leq K$, from the nodal set, and (3) $K$ nonempty subsets $B_k\subset\mathbf{R}^d$, let $\mathcal{F}^k$ be the set of all single-force loads where a force from $B_k$ acts at the node $p_k$, and let $\mathcal{F}=\mathrm{Conv}\{\cup_{k\leq K}\mathcal{F}^k\}$. Assume that for every $k$ and every load $f\in\mathcal{F}^k$ there exists a truss of total volume $W$ with compliance w.r.t. $f$ not exceeding 1. Then there exists truss of total volume $\gamma KW$ with compliance w.r.t. every load from $\mathcal{F}$ not exceeding 1, with $\gamma=4$ when $d=2$ and $\gamma=7$ when $d=3$.*
3. Finally we remark that the values of $\gamma$ yielded by the above proof are essentially better than the values yielded by much more transparent (and fully adjusted to the unit Euclidean balls in the role of $B_k$'s) solution to item 4. There is no reason to believe that these values are the smallest ones for which the above claim is true. This being said, it is easy to demonstrate that for $d=2$ the best possible in this respect value of $\gamma$ is at least 2.

## Support, characteristic, and Minkowski functions

**Exercise III.10.** [characteristic and support functions of convex sets] Let $X\subset\mathbf{R}^n$ be a nonempty convex set. *Characteristic* (a.k.a *indicator*) *function* of $X$ is, by definition, the function

$$\chi_X(x)=\left\{\begin{array}{cl}0 & ,x\in X\\ +\infty & ,x\notin X\end{array}\right.$$

As is immediately seen, this function is convex and proper. The Legendre transform of this function is called the *support function* $\phi_X(x)$ of $X$:

$$\phi_X(x)=\sup_u[x^\top u-\chi_X(u)]=\sup_{u\in X}x^\top u.$$

1. Prove that $\chi_X$ is lower semicontinuous (lsc) if and only if $X$ is closed, and that the support functions of $X$ and $\mathrm{cl}\,X$ are the same.

*Solution:* Lower semicontinuity of convex function is, by Proposition III.16.2, exactly the same as closedness of the epigraph of the function. The epigraph of $\chi_X(\cdot)$ is exactly $X\times\mathbf{R}_+$, and this set is closed if and only if $X$ is so. And of course

$$\phi_X(x)=\sup_{u\in X}x^\top u=\sup_{u\in\mathrm{cl}\,X}x^\top u,$$

so that the support functions of $X$ and $\mathrm{cl}\,X$ are the same.

In the remaining part of Exercise, we are interested in properties of support functions, and in view of item 1, it makes sense to assume from now on that $X$, on the top of being nonempty and convex, is also closed.

Prove the following facts:

2. $\phi_X(\cdot)$ is proper lsc convex function which is positively homogeneous of degree 1:

$$\forall(x \in \mathrm{Dom}\,\phi_x, \lambda \geq 0) : \phi_X(\lambda x) = \lambda\phi_X(x).$$

In particular, the domain of $\phi_X$ is a cone. Demonstrate by example that this cone not necessarily is closed (look at the support function of the closed convex set $\{[v; w] \in \mathbf{R}^2 : v > 0, w \leq \ln v\}$).

*Solution:* $\phi_X(\cdot)$ is convex, proper and lsc, as Legendre transform of a whatever proper convex function. And of course whenever $x$ is such that $\sup_{u\in X} x^\top u < \infty$, we have $\sup_{u\in X}[\lambda x]^\top u = \lambda \sup_{u\in X} x^\top u$ for all $\lambda \geq 0$.
Finally, for $X = \{[v; w] \in \mathbf{R}^2 : v > 0, w \leq \ln v\}$ we have

$$\begin{array}{rcl}\phi_X([x_1; x_2]) & = & \sup_{v,w} \{x_1 v + x_2 w : v > 0, w \leq \ln v\} \\ & = & \left\{\begin{array}{lll} +\infty & , x_1 > 0 & (a) \\ +\infty & , x_1 \leq 0, x_2 < 0 & (b) \\ +\infty & , x_1 = 0, x_2 \neq 0 & (c) \\ < +\infty & , x_1 < 0, x_2 \geq 0 & (d) \\ < +\infty & , x_1 = x_2 = 0 & (e) \end{array}\right.\end{array}$$

(to justify $(a)$ and $(c)$, set $[v; w] = [v; \ln v]$ and look what happens when $v \to \infty$ and when $v \to +0$, to justify $(b)$, look what happens when $[v; w] = [1; w]$ and $w \to -\infty$). We see that $\mathrm{Dom}\,\phi_X$ is the second quadrant $\{x_1 \leq 0, x_2 \geq 0\}$ with eliminated open ray $\{[0; x_2] : x_2 > 0\}$, and this set is just a cone, not a closed one.

3. Vice versa, every proper convex lsc function $\phi$ which is positively homogeneous of degree 1,

$$(x \in \mathrm{Dom}\, f, \lambda \geq 0) \Longrightarrow \phi(\lambda x) = \lambda\phi(x)$$

is the support function of a nonempty closed convex set, specifically, its subdifferential $\partial\phi(0)$ taken at the origin. In particular, $\phi_X(\cdot)$ "remembers" $X$: if $X, Y$ are nonempty closed convex sets, then $\phi_X(\cdot) \equiv \phi_Y(\cdot)$ if and only if $X = Y$.

*Solution:* Let $\phi$ be proper lsc convex and positively homogeneous, of degree 1, function, and let $\chi(x)$ be the Legendre transform of $\phi$. As every Legendre transform of proper convex function, $\chi$ is proper, convex and lsc. In addition, from properness and positive homogeneity of $\phi$ it follows that $0 \in \mathrm{Dom}\,\phi$ and $\phi(0) = 0$, whence

$$\chi(u) = \sup_x \left\{u^\top x - \phi(x)\right\} \geq u^\top 0 - \phi(0) = 0.$$

It remains to prove that $\chi$ takes just two values, 0 and $+\infty$; given this, we immediately conclude that $\chi$ is the characteristic function of its (nonempty, convex, and closed due to properness, convexity and lower semicontinuity of $\chi$, see item 1 of Exercise) domain. Indeed, we already know that $\chi(\cdot) \geq 0$; what remains to prove is that if $\chi(u) > 0$ for some $u$, then in fact $\chi(u) = \infty$. Relation $\chi(u) > 0$ amounts to existence of $x$ such that $u^\top x - \phi(x) > 0$; but then, due to positive homogeneity of $\phi$, for $\lambda > 0$ if holds $u^\top[\lambda x] - \phi(\lambda x) = \lambda[u^\top x - \phi(x)] \to \infty$, $\lambda \to \infty$, that is, $\chi(u) = +\infty$, as claimed.
Finally, from Proposition III.17.3 it follows that $\phi$, being proper convex lsc function, is the Legendre transform $\chi^*$ of $\chi$, that is, of the characteristic function of nonempty closed convex domain. By item **B** from chapter 17, see p. 211, the subdifferential of $\phi \equiv \chi^*$ taken at the origin is the set of all minimizers of $\chi$, and this set for characteristic function is nothing but its domain. □

4. Let $X, Y$ be two nonempty closed convex sets. Then $\phi_X(\cdot) \geq \phi_Y(\cdot)$ if and only if $Y \subset X$.

*Solution:* For proper lsc convex functions $f, g$ and their Legendre transforms $x^*$, $g^*$ the relation $f(\cdot) \leq g(\cdot)$ clearly implies that $f^*(\cdot) \geq g^*(\cdot)$; since $f$ and $g$ are the Legendre transforms of their Legendre transforms (Proposition III.17.3), the latter relation, in turn, implies that $f \leq g$. Thus, for proper lsc convex functions $f, g$ the relation $f \leq g$ is equivalent to $f^* \geq g^*$. In particular, $\phi_X(\cdot) \geq \phi_Y(\cdot)$ if and only if $\chi_X(\cdot) \leq \chi_Y(\cdot)$, and the latter relation clearly takes place if and only if $Y \subset X$. □

5. $\mathrm{Dom}\,\phi_X = \mathbf{R}^n$ if and only if $X$ is bounded.

*Solution:* When $X$ is bounded, $\phi_X(\cdot)$ clearly is real-valued on the entire space. Vice versa, if the convex function $\phi_X(\cdot)$ is real valued on the entire space, $\partial\phi_X(0)$ is bounded by Proposition III.16.10; it remains to note that by item 3 of Exercise, $X = \partial\phi_X(0)$. □

6. Let $X$ be the unit ball of some norm $\|\cdot\|$. Then $\phi_X$ is nothing but the norm $\|\cdot\|_*$ conjugate to $\|\cdot\|$. In particular, when $p \in [1,\infty]$ and $X = \{x \in \mathbf{R}^n : \|x\|_p \leq 1\}$, we have $\phi_X(x) \equiv \|x\|_q$, $\frac{1}{q} + \frac{1}{p} = 1$.

*Solution:* This is nothing but straightforward rewording of Fact III.17.4.

7. Let $x \mapsto Ax + b : \mathbf{R}^n \to \mathbf{R}^m$ be an affine mapping, and let $Y = AX + b = \{Ax + b : x \in X\}$. Then
$$\phi_Y(v) = \phi_X(A^\top v) + b^\top v.$$

*Solution:* Indeed, $\phi_Y(v) = \sup_{y\in Y} u^\top y = \sup_{x\in X} u^\top[Ax + b] = b^\top u + \sup_{x\in X}[A^\top v]^\top x = b^\top u + \phi_X(A^\top v)$. □

**Exercise III.11.** [Minkowski functions of convex sets] The goal of this Exercise is to acquaint the reader with important special family of convex functions – Minkowski functions of convex sets.

Consider a proper *nonnegative* lower semicontinuous function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ which is *positively homogeneous of degree 1*, meaning that
$$x \in \operatorname{Dom} f, t \geq 0 \implies tx \in \operatorname{Dom} f \ \&\ f(tx) = tf(x).$$

Note that from the latter property of $f$ and its properness it follows that $0 \in \operatorname{Dom} f$ and $f(0) = 0$.

We can associate with $f$ its *basic sublevel set*
$$X = \{x \in \mathbf{R}^n : f(x) \leq 1\}.$$

Note that $X$ "remembers" $f$, specifically
$$\forall t > 0 : f(x) \leq t \iff f(t^{-1}x) \leq 1 \iff t^{-1}x \in X,$$

whence also
$$\begin{array}{c}\forall x \in \mathbf{R}^n : f(x) = \inf\left\{t : t > 0, t^{-1}x \in X\right\}\\ [\inf\{t : t > 0, t \in \varnothing\} = +\infty \text{ by definition}]\end{array} \tag{19.1}$$

Note that the basic sublevel set of our $f$ cannot be arbitrary: it is convex and closed (since $f$ is convex lsc) and contains the origin (since $f(0) = 0$).

Now, given a closed convex set $X \subset \mathbf{R}^n$ containing the origin, we can associate with it a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ by construction from (19.1), specifically, as
$$f(x) = \inf\left\{t : t > 0, t^{-1}x \in X\right\} \tag{19.2}$$

This function is called *the Minkowski function* (M.f.) of $X$.

Here goes your first task:

1. Prove that when $X \subset \mathbf{R}^n$ is convex, closed, bounded, and contains the origin, function $f$ given by (19.2) is proper, nonnegative, convex lsc function positively homogeneous of degree 1, and $X$ is the basic sublevel set of $f$. Moreover, $f$ is nothing but the support function $\phi_{X_*}$ of the polar $X_*$ of $X$.

*Solution:* The polar $X_*$ of $X$ is closed convex set containing the origin, and therefore its support function $\overline{f}$, as the support function of any nonempty convex set, is convex lsc and positively homogeneous of degree 1 by Exercise III.10.2. Nonnegativity of $\overline{f}$ is readily given by the inclusion $0 \in X_*$. Thus, all which remains to verify is that in fact $f = \overline{f}$ and $X$ is the basic sublevel set of $\overline{f}$. Verification of the equality $f = \overline{f}$ is immediate:
$$\begin{array}{l}\forall t > 0 : \overline{f}(x) \leq t \iff \overline{f}(t^{-1}x) \leq 1 \iff \sup_{y\in X_*}[t^{-1}x]^\top y \leq 1\\ \iff t^{-1}x \in \operatorname{Polar}(X_*) = X,\end{array}$$

which combines with (19.2) to imply that whenever $t > 0$, relation $t \geq f(x)$ is the same as the relation $t \geq \overline{f}(x)$; since $f$ and $\overline{f}$ are nonnegative, it follows that $f \equiv \overline{f}$. To see that $X$ is the basic sublevel set of $f \equiv \overline{f}$, note that the basic sublevel set of the support function of $X_*$ clearly is $\operatorname{Polar}(X_*) = X$.

Your next tasks are as follows:

2. What are the Minkowski funcions of
   - the singleton $\{0\}$ ?
   - a linear subspace ?
   - a closed cone $\mathbf{K}$ ?
   - the unit ball of a norm $\|\cdot\|$ ?
3. Prove that the Minkowski functions $f_X$, $f_Y$ of closed convex and containing the origin sets $X, Y$ are linked by the relation $f_X \geq f_Y$ if and only if $X \subset Y$
4. When the Minkowski function of a set $X$ (convex, closed, bounded, and containing the origin) does not take value $+\infty$?
5. What is the set of zeros of the Minkowski function of a set $X$ (convex, closed, bounded, and c0ntaining the origin)?
6. What is the M.f. of the intersection $\cap_{k\leq K} X_k$ of closed convex sets containing the origin?

*Solution:* 2: The M.f. of a closed cone (in particular, of a linear subspace) is nothing but the characteristic function of this set. The M.f. of the unit $\|\cdot\|$-ball is the norm $\|\cdot\|$.

3: This is immediate consequence of the fact that $f_X$, $f_Y$ are the support functions of the polars $X_*$, $Y_*$ of $X, Y$ combined with the the result of Exercise III.10.4 and the fact that passing to polars reverses inclusions.

4: The M.f. $f_X$ of a closed convex set $X$ containing the origin is real-valued if and only if $X$ contains a neighbourhood of the origin. Indeed, if $X$ contains a centered at the origin $\|\cdot\|_2$-ball of raduis $r > 0$, $f_X$, by item 3, does not exceed the real-valued M.f. $r^{-1}\|\cdot\|_2$ of this ball. And if $f_X$ is real-valued, small enough positive multiples of $\pm e_i$ ($e_i$ are the standard basic orths) belong to $X$, so that the origin is an interior point of $X$.

5: The set of zeros of the M.f. of $X$ is exactly the recessive cone of $X$.

6: The M.f. in question is the maximum of the M.f.'s of $X_k$.

### Exercise III.12.

1. Recall that the closed conic transform
$$\overline{\mathrm{ConeT}}(X) = \mathrm{cl}\,\{[x;t] \in \mathbf{R}^n \times \mathbf{R} :\ t > 0,\ x/t \in X\},$$
of a nonempty convex set $X \subset \mathbf{R}^n$ (see section 1.5) is a closed cone such that
$$\mathrm{cl}(X) = \{x : [x;1] \in \overline{\mathrm{ConeT}}(X).$$
What is the cone dual to $\overline{\mathrm{ConeT}}(X)$ ?
2. Let $X \subset \mathbf{R}^n$ be a nonempty closed convex set and $X^+ = \overline{\mathrm{ConeT}}(X)$. Prove that
$$X_t^+ := \{x : [x;t] \in X^+\} = \left\{\begin{array}{lll} tX, & t > 0 & (a) \\ \mathrm{Rec}(X), & t = 0 & (b) \\ \varnothing, & t < 0 & (c) \end{array}\right.$$
3. Let $X_1, ..., X_K$ be closed convex sets in $\mathbf{R}^n$ with nonempty intersection $X$. Prove that
$$\overline{\mathrm{ConeT}}(X) = \cap_k \overline{\mathrm{ConeT}}(X_k).$$
4. Let $X = \cap_{k\leq K} X_k$, where $X_1, ..., X_K$ are closed convex sets in $\mathbf{R}^n$ such that $X_K \cap \mathrm{int}\, X_1 \cap \mathrm{int}\, X_2 ... \cap \mathrm{int}\, X_{K-1} \neq \varnothing$. Prove that $\phi_X(y) \leq a$ if and only if there exist $y_k$, $k \leq K$, such that
$$y = \sum_k y_k \ \&\ \sum_k \phi_{X_k}(y_k) \leq a. \qquad (*)$$
In words: *the supremum of a linear form on $\cap_k X_k$ does not exceed some $a$ if and only if the form can be decomposed into the sum of $K$ forms with the sum of their suprema over the respective sets $X_k$ not exceeding $a$.*

5. Prove the following polyhedral version of the claim in item 4:
*Let $X_k = \{x \in \mathbf{R}^n : A_k x \leq b_k\}$, $k \leq K$, be polyhedral sets with nonempty intersection $X$. A linear form does not exceed some $a \in \mathbf{R}$ everywhere on $X$ if and only if the form can be decomposed into the sum of $K$ linear forms with the sum of their maxima on respective sets $X_k$ not exceeding $a$.*

*Solution:* 1: This is the cone

$$\{[y;s] \in \mathbf{R}^n_y \times \mathbf{R}_s : s \geq \phi_X(-y)\},$$

where

$$\phi_X(y) = \sup_{x \in X} y^\top x$$

is the support function of $X$.

Indeed, from the definition of the closed conic transform it immediately follows that $[y;s]^\top [x;t] \geq 0$ for all $[x;t] \in \overline{\mathrm{ConeT}}(X)$ if and only if $[y;s]^\top [x;1] \geq 0$ for all $x \in X$, that is, if and only if

$$0 \leq s + \inf_{x \in X} y^\top x = s - \sup_{x \in X} [-y]^\top x = s - \phi_X(-y). \quad \square$$

2: A point $[x;t]$ belongs to $X^+$ if and only if there exist a sequence $[x_i;t_i]$ converging to $[x;t]$ and such that $t_i > 0$ and $x_i = t_i y_i$ with $y_i \in X$. When $t > 0$, the points $y_i = x_i/t_i$ have the limit $y = x/t$ as $i \to \infty$, and $y \in X$ since $X$ is closed; vice versa, if $y \in X$ and $t > 0$, the point $[ty;t]$ clearly belongs to $X^+ \cap \Pi_t$; we have proved $(a)$. $(c)$ is evident. $(b)$ is stated in Fact II.8.14.

3: Let $\Pi_s$ be the hyperplane $\{[x;s] : x \in \mathbf{R}^n\}$ in $\mathbf{R}^{n+1}$. By $(a)$ and $(c)$ in the previous item, we have $\overline{\mathrm{ConeT}}(X) \cap \Pi_t = \cap_k [\overline{\mathrm{ConeT}}(X_k) \cap \Pi_t]$ for $t \neq 0$. Since $X_k$ are closed convex sets with nonempty intersection, we have $\mathrm{Rec}(\cap_k X_k) = \cap_k \mathrm{Rec}(X_k)$, whence $\overline{\mathrm{ConeT}}(X) \cap \Pi_0 = \cap_k [\overline{\mathrm{ConeT}}(X_k) \cap \Pi_0]$ as well.

4: There is nothing to prove when $a = \infty$. Now let $a \in \mathbf{R}$. When $(*)$ takes place and $x \in X$, for every $k$ we have $x \in X_k$, so that $y_k^\top x \leq \phi_{X_k}(y_k) \leq a_k$. Summing up the resulting inequalities and taking into account that $\sum_k y_k = y$, we get $y^\top x \leq \sum_k a_k \leq a$. The resulting inequality holds true for every $x \in X$, implying $\phi_X(y) \leq a$.

Vice versa, let $\phi_X(y) \leq a \in \mathbf{R}$. Let $X_k^+ = \overline{\mathrm{ConeT}}(X_k)$, $k \leq K$. By item 3, $\overline{\mathrm{ConeT}}(X) = \cap_k \overline{\mathrm{ConeT}}(X_k)$, and by item 1 we have $[-y;a] \in [\overline{\mathrm{ConeT}}(X)]_*$ The cones $M^k = \overline{\mathrm{ConeT}}(X_k)$ are closed, and their interiors clearly contain the sets $\{[x;1] : x \in \mathrm{int}\, X_k\}$. It follows $M^K \cap \mathrm{int}\, M^1 \cap ... \cap \mathrm{int}\, M^{K-1} \neq \varnothing$. We see that the linear from with the vector of coefficients $[-y;a]$ and cones $M^1, ..., M^K$ satisfy the premise of the Dubovitski-Milutin Lemma. By this lemma, there exists a decomposition

$$[-y;a] = \sum_k [-y_k; a_k]$$

with $[-y_k;a_k] \in M^k_*$, that is, invoking item 1, with $\phi_{X_k}(y_k) \leq a_k$, $k \leq K$. □

5: We could prove this claim by slight modification of the reasoning for item 4, but it is easier to get it immediately from LP duality: for $\bar{y} \in \mathbf{R}^n$ and $a \in \mathbf{R}$ we have

$$\begin{array}{ll} & a \geq \sup_{x \in X} \bar{y}^\top x \implies a \geq \max_x \{\bar{y}^\top x : A_k x \leq b_k, k \leq K\} \\ \implies & a \geq \min_{z_1,...,z_K} \{\sum_k b_k^\top z_k : z_k \geq 0 \,\forall k, \sum_k A_k^\top z_k = \bar{y}\} \text{ [LP Duality Theorem]} \\ \implies & \exists (z_1 \geq 0, ..., z_K \geq 0) : \sum_k \underbrace{A_k^\top z_k}_{y_k} = \bar{y}, \sum_k \underbrace{b_k^\top z_k}_{a_k} \leq a \\ \implies & \exists y_1, ..., y_K, a_1, ..., a_K : a_k \geq \max_x \{y_k^\top x_k : A_k x \leq b_k\},\ k \leq K, \sum_k y_k = \bar{y} \\ & \text{[LP Duality Theorem]} \end{array}$$

We see that *if* $\bar{y}^\top x \leq a \in \mathbf{R}\, \forall x \in X$, *then* the linear form $\bar{y}^\top x$ can be decomposed into the sum of linear forms $y_k^\top x$ with the sum of maxima of the forms on the respective sets $X_k$ not exceeding $a$. The inverse statement is evident.

**Exercise III.13.** Let $X \subset \mathbf{R}^n$ be a nonempty polyhedral set given by polyhedral representation

$$X = \{x : \exists u : Ax + Bu \le r\}.$$

Build polyhedral representation of the epigraph of the support function of $X$. For non-polyhedral extension, see Exercise IV.36.

*Solution:* We have

$$\begin{array}{rcl} t \ge \phi_X(y) & \Longleftrightarrow & t \ge \mathrm{Opt}(P) := \max_{x,u}\{y^\top x : Ax + Bu \le r\} \\ & \Longleftrightarrow & t \ge \mathrm{Opt}(D) := \min_\lambda \left\{r^\top\lambda : \lambda \ge 0, A^\top\lambda = y, B^\top\lambda = 0\right\} \\ & & [\text{LP Duality Theorem; note that } (P) \text{ is feasible due to } X \ne \varnothing] \\ & \Longleftrightarrow & \exists \lambda : r^\top\lambda \le t, \lambda \ge 0, A^\top\lambda = y, B^\top\lambda = 0 \\ & & [\text{since by the above, } (D) \text{ is solvable whenever } t \ge \phi_X(y)] \end{array}$$

and we end up with polyhedral representation of epi$\{\phi_X\}$.

**Exercise III.14.** Compute in closed analytic form the support functions of the following sets:

1. The ellipsoid $\{x \in \mathbf{R}^n : (x-c)^\top C(x-c) \le 1\}$ with $C \succ 0$

   *Solution:* $\phi(y) = \sqrt{y^\top C^{-1}y}+c^\top y$

2. The probabilistic simplex $\{x \in \mathbf{R}^n_+ : \sum_i x_i = 1\}$

*Solution:* $\phi(y) = \max_{i\le n} y_i$.

3. The nonnegative part of the unit $\|\cdot\|_p$-ball: $X = \{x \in \mathbf{R}^n_+ : \|x\|_p \le 1\}$, $p \in [1,\infty]$

*Solution:* $\phi(y) = \|[y]_+\|_q$, where $\frac{1}{p}+\frac{1}{q} = 1$ and $[[y_1;...;y_n]]_+ = [\max[y_1;0];...;\max[y_n,0]]$.

4. The positive semidefinite part of the unit $\|\cdot\|_{p,\mathrm{Sh}}$ norm: $X = \{x \in \mathbf{S}^n_+ : \|x\|_{p,\mathrm{Sh}} \le 1\}$

*Solution:* $\phi(y) = \|[y]_+\|_{q,\mathrm{Sh}}$, where $\frac{1}{p}+\frac{1}{q} = 1$ and $[y]_+$ is the "positive semidefinite part of $y$" – the matrix obtained from symmetric matrix $y$ by keeping intact all eigenvectors and nonnegative eigenvalues, and zeroing out the negative eigenvalues (what is called the function $[\cdot]_+$ as applied to a symmetric matrix, see section D.1.5).

5. The paraboloid $\{x \in \mathbf{R}^{n+1} : x_{n+1} \ge \frac{1}{2}\sum_{i=1}^n x_i^2\}$ $(n \ge 1)$.

*Solution:* $$\phi(y) = \left\{\begin{array}{ll} -\frac{\sum_{i=1}^n y_i^2}{2y_{n+1}} & , y_{n+1} < 0 \\ 0 & , y = 0 \\ +\infty & , \text{all other cases} \end{array}\right.$$

## Around subdifferentials

**Exercise III.15.** Let $f$ be a convex function and $\bar{x} \in \mathrm{Dom}\, f \subset \mathbf{R}^n$. Prove that the property of $g \in \mathbf{R}^n$ to be a subgradient of $f$ at $\bar{x}$ is local: the inequality

$$f(x) \ge f(\bar{x}) + g^\top(x - \bar{x}) \qquad (*)$$

hods true for all $x \in \mathbf{R}^n$ iff it holds true for all $x$ in a neighborhood of $\bar{x}$.

*Solution:* In one direction the claim is evident. Now assume that $(*)$ holds true for all $x$ in a neighborhood pf $\bar{x}$, and let us prove that it holds true for all $x$. Indeed, let $\bar{f}(x) = f(x) - f(\bar{x}) - g^\top(x-\bar{x})$, so that $\bar{f}$ is convex along with $f$ by calculus of convexity. Validity of $(*)$ is a neighborhood of $\bar{x}$ means that $\bar{x}$ is a local minimizer of $\bar{f}$: $\bar{f}(x) \ge \bar{f}(\bar{x}) = 0$ for all $x$ from a neighborhood of $\bar{x}$. By unimodality (Theorem III.15.1 applied with $Q = \mathbf{R}^n$ and $x^* = \bar{x}$ to $\bar{f}$ in the role of $f$) $\bar{x}$ is a global minimizer of $\bar{f}$, so that $\bar{f}(x) \ge \bar{f}(\bar{x}) = 0$ for all $x$. Recalling what $\bar{f}$ is, we see that $(*)$ holds true for all $x$. □

**Exercise III.16.** [subdifferentials of norms] Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$, and $\|\cdot\|_*$ be its conjugate (see Fact III.17.4). Prove that

1. The subdifferential of $\|\cdot\|$ taken at the origin is the unit ball $B_*$ of $\|\cdot\|_*$, or, which is the same, the polar
$$\{u : u^\top x \leq 1\,\forall(x : \|u\| \leq 1)\}$$
of the unit ball $B$ of the norm $\|\cdot\|$.
2. When $x \neq 0$, the subdifferential of $\|\cdot\|$ taken at $x$ is the set $\{u \in B_* : u^\top x = \|x\|\}$. In particular, the subdifferential of $\|\cdot\|$ remains intact when replacing $x$ with $tx$, $t > 0$, and is reflected w.r.t. the origin when $x$ is replaced with $tx$, $t < 0$.

*Solution:* 1: By Fact III.17.4, the Legendre transform of $\|\cdot\|$ is the characteristic function of $B_*$, so that $\|\cdot\|$, being real-valued convex continuous (and thus lsc) function, by Proposition III.17.3 is the Legendre transform of the characteristic function of the closed nonempty convex set $B_*$, or, which is the same, $\|\cdot\|$ is the support function of $B_*$. By item 3 of Exercise III.10, $\|\cdot\|$ is the support function of its subdifferential, taken at the origin, which also is a closed nonempty convex set. As we know from Exercise III.10.4, the support functions of nonempty closed convex sets coincide iff the sets coincide, so that the subdifferential of $\|\cdot\|$ taken at the origin is $B_*$. □

2: Let $x \neq 0$ and $X$ be the subdifferential of $\|\cdot\|$ taken at $x$; this is a nonempty convex compact set (Proposition III.16.10). When $g \in X$, we should have
$$g^\top(ty - x) \leq \|ty\| - \|x\|, \, t > 0,$$
which, after dividing both sides by $t$ and passing to limit as $t \to \infty$, implies that $g^\top y \leq \|y\|$ for all $y$, so that $g \in B_* = \{h : h^\top y \leq 1\,\forall(y, \|y\| \leq 1)\}$ (see Fact III.17.4). On the other hand, for $\epsilon \in (0,1)$ one has $g^\top[(1-\epsilon)x - x] \leq \|(1-\epsilon)x\| - \|x\| = -\epsilon\|x\|$, implying thai $g^\top x \geq \|x\|$. Strict inequality is impossible due to already proved $g \in B_*$, and we conclude that $g^\top x = \|x\|$. Thus, $X \subset \{g \in B_* : g^\top x = \|x\|\}$. On the other hand, when $g \in B_*$ is such that $g^\top x = \|x\|$, we have for every $y \in \mathbf{R}^n$
$$\|y\| \geq g^\top y = g^\top(y - x) + g^\top x = g^\top(y - x) + \|x\|,$$
where the first inequality is due to $g \in B_*$. Thus, $\|x\| + g^\top(y - x) \leq \|y\|$ for all $y$, implying that $g \in X$. □

**Exercise III.17.** [Shatten norms] Let $p \in [1, \infty]$. The space $\mathbf{S}^n$ of symmetric $n \times n$ matrices can be equipped with *Shatten p-norms* – matrix analogies of the standard $\|\cdot\|_p$-norms on $\mathbf{R}^n$. Specifically, Shatten $p$-norm $\|\cdot\|_{p,\mathrm{Sh}}$ of symmetric matrix $X$ is defined as
$$\|X\|_{p,\mathrm{Sh}} = \|\lambda(X)\|_p,$$
where $\lambda(X)$, as always, is the vector of eigenvalues of $X$.

1. Prove that Shatten norms indeed are norms, and the norm conjugate to $\|\cdot\|_{p,\mathrm{Sh}}$ is $\|\cdot\|_{q,\mathrm{Sh}}$, $\frac{1}{p} + \frac{1}{q} = 1$:
$$\|X\|_{q,\mathrm{Sh}} = \max_Y\{\mathrm{Tr}(XY) : \|Y\|_{p,\mathrm{Sh}} \leq 1\} \tag{19.3}$$
2. Verify that $\|\cdot\|_{2,\mathrm{Sh}}$ is nothing but the Frobenius norm of $X$, and $\|\mathbb{X}\|_{\infty,\mathrm{Sh}}$ is the same as the spectral norm of $X$.

*Solution:* 1: The facts that $\|\cdot\|_{p,\mathrm{Sh}}$ is positive outside of the origin and satisfies $\|\lambda X\|_{p,\mathrm{Sh}} = |\lambda|\|X\|_{p,\mathrm{Sh}}$ are evident. Therefore, all we need to justify all claims in item 1 is to justify (19.3), which, as a byproduct, implies convexity of $\|\cdot\|_{p,\mathrm{Sh}}$, which for positively homogeneous, of degree 1, functions implies the Triangle inequality. To justify (19.3), let $X = U\Lambda U^\top$ be the eigenvalue decomposition of $X \in \mathbf{S}^n$, so that $\Lambda = \mathrm{Diag}\{\lambda(X)\}$. Denoting by $\mathrm{Dg}\{Z\}$ the vector of diagonal entries in matrix $Z$, we have
$$\begin{array}{rcl}\forall(Y \in \mathbf{S}^n) : \mathrm{Tr}(XY) & = & \mathrm{Tr}(U\Lambda U^\top Y) = \mathrm{Tr}(\Lambda[U^\top YU]) = \lambda^\top(X)\mathrm{Dg}\{U^\top YU\} \\ & \leq & \|\lambda(X)\|_q\|\mathrm{Dg}\{U^\top YU\}\|_p \leq \|\lambda(X)\|_q\|\lambda(Y)\|_p,\end{array}$$
where the concluding inequality is due to Proposition III.18.3 as applied with $f(x) = \|x\|_p$. We see that the right hand side in (19.3) is $\leq$ the left hand side. On the other hand, we can select $g \in \mathbf{R}^n$ in such a way that $\|g\|_p = 1$ and $\lambda^\top(X)g = \|\lambda(X)\|_q$ (since $\|\cdot\|_q$ is the conjugate of $\|\cdot\|_p$). Setting

$Y = U\,\mathrm{Diag}\{g\}U^\top$, we get $\lambda(Y) = g$, $\|Y\|_{p,\mathrm{Sh}} = 1$, and $U^\top Y U = \mathrm{Diag}\{g\}$ , whence by the above computation, $\mathrm{Tr}(XY) = \lambda^\top(X)\lambda(Y) = \|\lambda(X)\|_q$, so that the right hand side in (19.3) is $\geq$ the left hand side. Thus, (19.3) does hold true. □

2: Taken together, eigenvalue decomposition and the fact that multiplication of matrix from the left and from the right by orthogonal matrices preserves the Frobenius norm (Fact D.2) demonstrate that the Frobenius norm of a symmetric matrix is the same as $\|\cdot\|_2$-norm of its vector of eigenvalues. The fact that $\|\cdot\|_{\infty,\mathrm{Sh}}$ is the spectral norm is evident. □

**Exercise III.18.** [chain rule for subdifferentials] Let $Y \subset \mathbf{R}^m$ and $X \subset \mathbf{R}^n$ be nonempty convex sets, $\overline{y} \in Y$, $\overline{x} \in X$, let $f(\cdot) : Y \to \mathbf{R}$ be a convex function, and $A(\cdot) : X \to Y$ with $A(\overline{x}) = \overline{y}$. Let, further, $\mathbf{K}$ be a closed cone in $\mathbf{R}^n$. Function $f$ is called $\mathbf{K}$-monotone on $Y$, if for $y, y' \in Y$ such that $y' - y \in \mathbf{K}$ it holds $f(y') \geq f(y)$, and $A$ is called $\mathbf{K}$-convex on $X$ if for all $x, x' \in X$ and $\lambda \in [0,1]$ it holds $\lambda A(X) + (1-\lambda)A(x') - A(\lambda x + (1-\lambda)x') \in \mathbf{K}$.

Prove that

1. $A$ is $\mathbf{K}$-convex on $X$ if and only if for every $\phi \in \mathbf{K}_*$ the real-valued function $\phi^\top A(x)$ is convex on $X$.

   *Solution:* Indeed, since $\mathbf{K}$ is closed, we have $\mathbf{K} = (\mathbf{K}_*)_*$, so that $\lambda A(X) + (1-\lambda)A(x') - A(\lambda x + (1-\lambda)x') \in \mathbf{K}$ if and only if $\lambda\phi^\top A(x) + (1-\lambda)\phi^\top A(x') - \phi^\top A(\lambda x + (1-\lambda)x') \geq 0$ for all $\phi \in \mathbf{K}_*$.

2. Let $A$ be $\mathbf{K}$-convex on $X$ and differentiable at $\overline{x}$. Prove that
$$\forall x \in X : A(x) - [A(\overline{x}) + A'(\overline{x})[x - \overline{x}]] \in \mathbf{K}. \qquad (*)$$

   *Solution:* By item 1, for $\phi \in \mathbf{K}_*$ the function $\phi^\top A(x)$ is convex on $X$, and by the standard Calculus it is differentiable at $\overline{x}$ with the derivative $\phi^\top A'(x)$. Therefore by Gradient inequality one has
$$\forall x \in X : \phi^\top \left[A(x) - [A(\overline{x}) + A'(\overline{x})[x-\overline{x}]]\right] = \phi^\top A(x) - [\phi^\top A(\overline{x}) + \phi^\top A'(\overline{x})[x - \overline{x}]] \geq 0,$$
   and $(*)$ follows.

3. Let $f$ be $\mathbf{K}$-monotone on $Y$ and $A$ be $\mathbf{K}$-convex on $X$. Prove that the real valued on $X$ function $f{\circ}A\,(x) = f(A(x))$ is convex.

   *Solution:* Indeed, for $x, x' \in X$ and $\lambda \in [0,1]$ the points $y = A(x)$, $y' = A(x')$, $w = \lambda y + (1-\lambda)y'$ and $z = A(\lambda x + (1-\lambda)x')$ belong to $Y$ since $Y$ is convex and $A$ maps $X$ into $Y$, and $w - z \in \mathbf{K}$ since $A$ is $\mathbf{K}$-convex. Since $f$ is $\mathbf{K}$-monotone, $w - z \in \mathbf{K}$ implies that $f(w) \geq f(z)$. Besides this, recalling what $w$ is and that $f$ is convex, $\lambda f(y) + (1-\lambda)f(y') \geq f(w)$. The bottom line is that $\lambda f(y) + (1-\lambda)f(y') \geq f(z)$, that is,
$$f{\circ}A\,(\lambda x + (1-\lambda)x') = f(z) \leq \lambda f(y) + (1-\lambda)f(y') = \lambda f{\circ}A\,(x) + (1-\lambda)f{\circ}A\,(x').$$
   The resulting inequality hods true for all $x, x' \in X$ and $\lambda \in [0,1]$, so that $f{\circ}A$ is convex on $X$.

4. Let $f$ be $\mathbf{K}$-monotone on $Y$. Prove that $\partial f(\overline{y}) \subset \mathbf{K}_*$, provided $\overline{y} \in \mathrm{int}\,Y$.

   *Solution:* Indeed, let $g \in \partial f(\overline{y})$ and $h \in \mathbf{K}$. Since $\overline{y} \in \mathrm{int}\,Y$, we have $\overline{y} - th \in Y$ for small positive $t$, and $f(\overline{y} - th) \leq f(\overline{y})$ by $\mathbf{K}$-monotonicity of $f$. Besides this, $f(\overline{y} - th) \geq f(\overline{y}) - tg^\top h$ due to $g \in \partial f(\overline{y})$. Thus, for all small positive $t$ it holds
$$f(\overline{y}) \geq f(\overline{y}) - tg^\top h,$$
   implying that $g^\top h \geq 0$. This relation holds true for every $g \in \partial f(\overline{y})$ and $h \in \mathbf{K}$, implying that $\partial f(\overline{y}) \subset \mathbf{K}_*$.

5. [chain rule] Let $\overline{y} \in \mathrm{int}\,Y$, $\overline{x} \in \mathrm{int}\,X$, let $f$ be $\mathbf{K}$-monotone on $Y$, $A$ be $\mathbf{K}$-convex on $X$ and differentiable at $\overline{x}$. Prove that
$$\partial f{\circ}A\,(\overline{x}) = [A'(\overline{x})]^\top \partial f(\overline{y}) = \{[A'(\overline{x})]^\top g : g \in \partial f(\overline{y})\} \qquad (!)$$

*Solution:* Let us first verify that the right hand side set in (!) is contained in the left hand side one. Indeed, let $g \in \partial f(\overline{y})$, $x \in X$, and $y = A(x)$. We have

$$\begin{array}{l} f{\circ}A\,(x) = f(y) \geq f(\overline{y}) + g^\top[y-\overline{y}] = f(\overline{y}) + g^\top[A(x) - A(\overline{x})] \\ \geq f(\overline{y}) + g^\top A'(\overline{x})[x-\overline{x}] \text{ [since } g \in \mathbf{K}_* \text{ and due to } (*)] \\ = f{\circ}A\,(\overline{x}) + g^\top A'(\overline{x})[x-\overline{x}]. \end{array}$$

The resulting inequality holds true for all $x \in X$ and $g \in \partial f(\overline{y})$, implying that $[A'(\overline{x})]^\top \partial f(\overline{y})) \subset \partial f{\circ}A\,(\overline{x})$.

Now let us prove that the left hand side set in (!) is contained in the right hand side set, let it be called $D$. $\overline{y} \in \operatorname{int} Y$, so that $\partial f(\overline{y})$ is a nonempty convex compact set; therefore $D$ also is nonempty convex compact set. Assume, on the contrary to what should be proved, that there exists $e \in \partial f{\circ}A\,(\overline{x}) \setminus D$. By Separation Theorem, there exists $h \in \mathbf{R}^n$ such that

$$h^\top e > \alpha = \max_{z \in D} h^\top z = \max_{g \in \partial f(\overline{y})} g^\top A'(\overline{x})h.$$

For small positive $t$ from differentiability of $A$ at $\overline{x}$ it follows that

$$y_t := A(\overline{x} + th) = \overline{y} + tA'(\overline{x})h + \epsilon_t, \;\; \|\epsilon_t\|_2/t \to 0, t \to +0.$$

Since $f$ is convex and real-valued in a neighbourhood of $\overline{y}$, it is Lipschitz continuous, with some constant $L$, in such a neighbourhood, which combines with the above relation to imply that

$$f{\circ}A\,(\overline{x} + th) = f(y_t) = f(\overline{y} + tA'(\overline{x})h) + \delta_t, \;\; \delta_t/t \to 0, \; t \to +0,$$

whence

$$\lim_{t \to +0} \frac{f{\circ}A\,(\overline{x}+th) - f{\circ}A\,(\overline{x})}{t} = \lim_{t \to +0} \frac{f(\overline{y} + tA'(\overline{x})h) - f(\overline{y})}{t}.$$

By Theorem III.16.12, the left hand side in this equality is $\max_{d \in \partial f \circ A(\overline{x})} d^\top h$ (and is therefore $\geq e^\top h$ due to the origin of $e$), and the right hand side is $\max_{g \in \partial f(\overline{y})} g^\top A'(\overline{x})h = \alpha$. Thus, $e^\top h \leq \alpha$, which is the desired contradiction.

**Exercise III.19.** Recall that the sum $S_k(X)$ of $k \leq n$ largest eigenvalues of $X \in \mathbf{S}^n$ is a convex function of $X$, see Remark III.18.4. Point out a subgradient of $S_k(\cdot)$ at a point $\overline{X} \in \mathbf{S}^n$. As a special case, find a subgradient of the maximal eigenvalue $\lambda_{\max}(X)$ of $X \in \mathbf{S}^n$ treated as a function of $X$.

*Solution:* Let $\overline{X} = \overline{U} \operatorname{Diag}\{\lambda(\overline{X}\}\overline{U}^\top$ be the eigenvalue decomposition of $\overline{X}$. Setting

$$P = \overline{U} \operatorname{Diag}\{\underbrace{1, ..., 1}_{k}, 0, ..., 0\}\overline{U}^\top,$$

we get $\operatorname{Tr}(\overline{X}P) = \sum_{i=1}^k \lambda_k(\overline{X}) = S_k(\overline{X})$. On the other hand, for $X \in \mathbf{S}^n$ we have

$$\begin{array}{rcl} \operatorname{Tr}(XP) & = & \operatorname{Tr}(X\overline{U} \operatorname{Diag}\{1, ..., 1, 0, .., 0\}\overline{U}^\top) = \operatorname{Tr}([\overline{U}^\top X\overline{U}] \operatorname{Diag}\{1, ..., 1, 0, ..., 0\}) \\ & \leq & s_k(\operatorname{Dg}\{\overline{U}^\top X\overline{U}\}), \end{array}$$

where, as always, $s_k(x)$ is the sum of $k$ largest entries in a vector $x$. By Proposition III.18.3, $s_k(\operatorname{Dg}\{\overline{U}^\top X\overline{U}\}) \leq s_k(\lambda(X)) = S_k(X)$. Thus,

$$\forall X \in \mathbf{S}^n : S_k(X) \geq \operatorname{Tr}(XP) = \operatorname{Tr}(\overline{X}P) + \operatorname{Tr}(P[X - \overline{X}]) = S_k(\overline{X}) + \operatorname{Tr}(P[X - \overline{X}]).$$

Recalling what is the inner product on $\mathbf{S}^n$, we conclude that $P \in \partial S_k(\overline{X})$.

To get a subgradient of $\lambda_{\max}(X)$, note that $\lambda_{\max}(X) \equiv S_1(X)$, so that the above computation says that if $e(X)$ is leading eigenvector of $X$ (i.e., unit $\|\cdot\|_2$-norm eigenvector of $X$ with eigenvalue $\lambda_{\max}(X)$), then $e(X)e^\top(X) \in \partial\lambda_{\max}(X)$.

## Around Legendre transform

**Exercise III.20.** Compute Legendre transforms of the following univariate functions:

1. $f(x) = -\ln x$, $\mathrm{Dom}\, f = (0, \infty)$

*Solution:* $f^*(y) = \sup_{x>0}[xy + \ln x]$. When $y \geq 0$, the supremum is $+\infty$ (look what happens when $x \to +\infty$). When $y < 0$, the sufficient condition for some $x > 0$ to maximize $\phi_y(x) = xy + \ln x$ is to be a root of $\phi'_y(x)$ (Theorem III.15.2 as applied to convex differentiable function $-\phi_y(\cdot)$). The equation $\phi'_y(x) = 0$ reads $y + 1/x = 0$, resulting in $x = -1/y$ and $\max_{x>0} \phi_y(x) = \phi_y(-1/y) = -\ln(-y) - 1$. Thus,

$$f^*(y) = -\ln(-y) - 1, \ \mathrm{Dom}\, f^* = (-\infty, 0).$$

2. $f(x) = \mathrm{e}^x$, $\mathrm{Dom}\, f = \mathbf{R}$.

*Solution:* Setting $\phi_y(x) = xy - \mathrm{e}^x$, we have $\sup_x \phi_y(x) = +\infty$ when $y < 0$ (look what happens when $x \to -\infty$). When $y = 0$, we clearly have $\sup_x \phi_y(x) = 0$. Finally, when $y > 0$, the maximizer of $\phi_y(\cdot)$, same as in the previous item, can be found via Fermat rule – as a root of the equation $\phi'_y(x) = 0$. This equation reads $y - \mathrm{e}^x = 0$, resulting in $x = \ln y$ and $\sup_x \phi_y(x) = y \ln y - y$. Thus,

$$f^*(y) = y \ln y - y, \ \mathrm{Dom}\, f^* = [0, \infty); \text{ here, as always, } 0 \ln 0 = 0 \text{ by definition.}$$

3. $f(x) = x \ln x$, $\mathrm{Dom}\, f = [0, \infty)$ ($0 \ln 0 = 0$ by definition).

*Solution:* To maximize $xy - x \ln x$ over $x \geq 0$ we can use the Fermat rule resulting in the equation $y - 1 - \ln x = 0$. Thus, the maximizer is $x = \mathrm{e}^{y-1}$, resulting in

$$f^*(y) = \mathrm{e}^{y-1}, \ \mathrm{Dom}\, f^* = \mathbf{R}.$$

We could get the same result without computation: from item 2 we know that the Legendre transform of $\mathrm{e}^x$ is $y \ln y - y$, implying that the Legendre transform of $x \ln x - x$ is $\mathrm{e}^y$; and linear perturbation of a function (in our case, adding $x$ to $x \ln x - x$) results in shift of the Legendre transform.

4. $f(x) = x^p/p$, $\mathrm{Dom}\, f = [0, \infty)$; here $p > 1$.

*Solution:* $f^*(y) = \sup_{x \geq 0}[\phi_y(x) := xy - x^p/p]$. When $y \leq 0$, we clearly have $\sup_{x \geq 0} \phi_y(x) = \phi_y(0) = 0$. When $y > 0$, the maximizer of $\phi_y(x)$ over $x \geq 0$ is given by Fermat rule resulting in the equation $y = x^{p-1}$. Thus, for $y > 0$ we have $\sup_{x \geq 0} \phi_y(x) = y^{1 + \frac{1}{p-1}} - y^{\frac{p}{p-1}}/p = y^q/q$, where $q = \frac{p}{p-1}$, or, which is the same, $\frac{1}{p} + \frac{1}{q} = 1$. We end up with

$$f^*(y) = [y_+]^q/q, \ \mathrm{Dom}\, f^* = \mathbf{R}; \text{ here } y_+ = \max[y, 0], \ q = \tfrac{p}{p-1}.$$

**Exercise III.21.** Compute Legendre transforms of the following functions:

- [log-barrier for nonnegative orthant $\mathbf{R}^n_+$] $f(x) = -\sum_{i=1}^n \ln x_i : \mathrm{int}\, \mathbf{R}^n_+ \to \mathbf{R}$

*Solution:*

$$f^*(z) = \sup_{x>0} \sum_i [z_i x_i + \ln x_i] = \left\{ \begin{array}{ll} -n - \sum_i \ln(-z_i), & z < 0 \\ +\infty & \text{, otherwise} \end{array} \right. ,$$

thus, $f^*(z) = f(-z) - n$.

- [log-det barrier for semidefinite cone $\mathbf{S}^n_+$] $f(x) = -\ln \mathrm{Det}(x) : \mathrm{int}\, \mathbf{S}^n_+ \to \mathbf{R}$ (start with proving convexity of $f$).

*Solution:* Convexity of $f$ was already established twice – first time via computing second order directional derivative in section C.2.2, second time in chapter 18. We have

$$f^*(z) = \sup_{x \succ 0} [\mathrm{Tr}(zx) + \ln \mathrm{Det}(x)] \,.$$

It is immediately seen that $f^*(z) = +\infty$ unless $z \in -\mathrm{int}\, \mathbf{S}^n_+$. Indeed, restricting maximization over $x \succ 0$

by maximization over $x \succ 0$ commuting with $z$ and looking what happens when $x$ and $z$ are represented in the orthonormal eigenbasis of $z$, we get

$$f^*(z) \geq \sup_{\xi > 0} \left[ \sum\nolimits_i \xi_i \zeta_i + \sum\nolimits_i \ln \xi_i \right],$$

where $\zeta_i$ are the eigenvalues of $z$, and from item 1 we know that the right hand side sup is $+\infty$ unless all $\xi_i$ are negative. Now let $z \prec 0$. In this case we can maximize the concave function $\mathrm{Tr}(zx) - f(x) = \mathrm{Tr}(zx) + \ln \mathrm{Det}(x)$ over $x \succ 0$ by solving the Fermat equation; as we know from Example C.9 in section C.1.6, $\nabla f(x) = -x^{-1}$, so that the Fermat rule results in $x = -z$ and $f^*(z) = -\ln \mathrm{Det}(-z) - n = f(-z) - n$.

**Exercise III.22.** [computing the Legendre transform of the log-barrier $-\ln(x_n^2 - x_1^2 - ... - x_{n-1}^2)$ for Lorentz cone] Consider the optimization problem

$$\max_{x,t} \left\{ \xi^\top x + \tau t + \ln(t^2 - x^\top x) : (t,x) \in X = \{t > \sqrt{x^\top x}\} \right\}$$

where $\xi \in \mathbf{R}^n$, $\tau \in \mathbf{R}$ are parameters. Is the problem convex[9)]? What is the domain in the space of parameters where the problem is solvable? What is the optimal value? Is it convex in the parameters?

*Solution:* Problem is convex, since the function $f(t,x) = -\ln(t^2 - x^\top x)$ is convex (direct computation of the second order directional derivative[10]); the domain of the problem is open. Therefore the problem is solvable if and only if the Fermat system

$$\begin{array}{lcl} \xi - f'_x(t,x) = 0 & \Longleftrightarrow & \frac{2x}{t^2 - x^\top x} = \xi \\ \tau - f'_\tau(t,x) = 0 & \Longleftrightarrow & -\frac{2t}{t^2 - x^\top x} = \tau \end{array} \qquad (*)$$

in variables $t, x$ has a solution with $t > \sqrt{x^\top x}$; it follows that $\tau$ should be negative. Assuming that it is the case, the second equation says that $\frac{1}{t^2 - x^\top x} = -\frac{\tau}{2t}$, whence the first equation says that $x = -\frac{t}{\tau}\xi$. It follows that

$$-\frac{2t}{\tau} = t^2 - x^\top x = t^2 - \frac{t^2}{\tau^2}\xi^\top \xi = \frac{t^2}{\tau^2}(\tau^2 - \xi^\top \xi). \qquad (1)$$

In order for this equation be solvable one should have $\tau^2 > \xi^\top \xi$, which combines with $\tau < 0$ to yield that $-\tau > \sqrt{\xi^\top \xi}$. Under the latter assumption, (1) implies that

$$t = -\frac{2\tau}{\tau^2 - \xi^\top \xi}, \qquad (2)$$

whence also

$$x = \frac{2\xi}{\tau^2 - \xi^\top \xi} \qquad (3)$$

Thus, the space of parameters for which the problem is solvable is given by

$$-\tau > \sqrt{\xi^\top \xi},$$

the solution is given by (2) - (3), and the optimal value is (direct computation)

$$-\ln(\tau^2 - \xi^\top \xi) + 2\ln 2 - 2.$$

[9] A *maximization* problem with objective $f(\cdot)$ and certain constraints and domain is called convex if the equivalent minimization problem with the objective $(-f)$ and the original constraints and domain is convex.

[10] intelligent reasoning: in the domain $t > \sqrt{x^\top x}$ we have $f(t,x) = -\ln t - \ln(t - t^{-1}x^\top x) = -\ln t + g(t^{-1}x^\top x - t)$, where the function $g(s) = \left\{ \begin{array}{ll} -\ln(-s), & s < 0 \\ +\infty, & s \geq 0 \end{array} \right.$ is convex and nondecreasing. The function $t^{-1}x^\top x - t$ is convex in the domain $t > 0$ as the perspective transform of $x^\top x - 1$. Now convexity of $f$ is readily given by calculus of convexity-preserving operations.

The optimal value is convex in the parameters $\tau, \xi$ (by its origin, it is supremum of linear forms, parameterized by $x, t$, of the parameters $\tau, \xi$).

**Exercise III.23.** Consider the optimization problem

$$\max_{x,y} \{f(x,y) = ax + by + \ln(\ln y - x) + \ln(y) : (x,y) \in X = \{(x,y) : y > \exp\{x\}\}\},$$

where $a, b \in \mathbf{R}$ are parameters. Is the problem convex? What is the domain in space of parameters where the problem is solvable? What is the optimal value? Is it convex in the parameters?

*Solution:* The objective is concave (direct computation), the domain is convex, so that the problem is convex; the domain of the problem is open. Therefore $a, b$ correspond to a solvable problem if and only if the Fermat system

$$\begin{array}{lcl} f'_x(x,y) = 0 & \Longleftrightarrow & a = \frac{1}{\ln y - x} \\ f'_y(x,y) = 0 & \Longleftrightarrow & b = -\frac{1}{y}\left[1 + \frac{1}{\ln y - x}\right] \end{array} \tag{4}$$

in variables $x, y$ has a solution with $y > 0$, $\ln y > x$. From the first equation, $a$ should be positive, and if this is the case, the second equation says that $b$ should be negative and $y = -\frac{1+a}{b}$. Thus, $a$ should be positive, $b$ should be negative, and in this case the solution to (4) is

$$x = \ln\left(-\frac{1+a}{b}\right) - \frac{1}{a},\ y = -\frac{1+a}{b},$$

whence the optimal value is

$$(a+1)\ln\left(-\frac{1+a}{b}\right) - \ln a - a - 2.$$

This quantity, due to its origin, is supremum of linear forms of $a, b$ and therefore is convex in the domain $a > 0, b < 0$.

**Exercise III.24.** Compute Legendre transforms of the following functions:

- ["geometric mean"] $f(x) = -\prod_{i \le n} x_i^{\pi_i} : \mathbf{R}^n_+ \to \mathbf{R}$, where $\pi_i > 0$ sum up to 1 and $n > 1$.

  *Solution:* Convexity of $f$ was established in section 14.2.A.(2). The Legendre transform is

  $$f^*(y) = \sup_{x \ge 0}\{\sum_i y_i x_i + \prod_i x_i^{\pi_i}\} \tag{$*$}$$

  The right hand side is $+\infty$ unless $y < 0$ (assuming that, say, $y_1 \ge 0$, look what happens when $x$ runs through the ray $\{[t;1;...;1] : t \ge 0\}$). Assuming $y < 0$ and setting $z = -y$, we have $f^*(-z) = \sup_{x \ge 0}\{\prod_i x_i^{\pi_i} - \sum_i z_i x_i\}$. What we are maximizing over $x$, is a homogeneous, of homogeneity degree 1, function of $x \ge 0$ (recall that $\sum_i \pi_i = 1$); therefore the supremum is either 0 or $+\infty$, depending on whether what we are maximizing is or is not nonpositive on $\mathbf{R}^n_+$, or, which is the same, is or is not nonpositive on the set $X_z = \{x \ge 0 : \sum_i z_i x_i = 1\}$. Making educated guess that the maximizer $x_z$ of $\prod_i x_i^{\pi_i}$ over $X_z$ is positive, Karush-Kuhn-Tucker optimality conditions (see discussion after Proposition III.15.3) as applied to our maximization problem (rewritten as $\min_{x \in X_z}[\sum_i z_i x_i + f(x)]$) result in the system

  $$\pi_i \underbrace{[\prod_j x_j^{\pi_j}]}_{\alpha} x_i^{-1} = \lambda z_i,\ i \le n, \sum_i z_i x_i = 1$$

  in variables $x, \lambda$; $x$-component of a solution to this system, if positive, is the desired $x_z$ by Proposition III.15.3. From the system, $\sum_i z_i x_i = \lambda^{-1}\alpha$, that is, $\lambda = \alpha$ and therefore $x_i = \pi_i / z_i$. The vector $[\pi_1/z_1; ...; \pi_n/z_n]$ indeed is positive and is therefore the desired maximizer $x_z$ of $\phi$ over $X_z$ the maximum being $\prod_i [\pi_i/z_i]^{\pi_i} - 1$. As we remember, $f^*(-z)$ is $+\infty$ when this maximum is positive and is zero otherwise. The bottom line is that the domain of $f^*$ is $\{y \in \mathbf{R}^n : y < 0, \prod_i [-\pi_i/y_i]^{\pi_i} \le 1\}$ and in this domain $f^*$ is identically equal to 0.

- ["inverse geometric mean"] $f(x)=\prod_{i\leq n}x_i^{-\pi_i}:\operatorname{int}\mathbf{R}^n_+\to\mathbf{R}$, where $\pi_i>0$.

*Solution:* Convexity of $f$ is stated in section 14.2.A.(3). We have $f^*(y)=\sup_{x>0}[\sum_i y_ix_i-\prod_i x_i^{-\pi_i}]$. This supremum is $+\infty$ when some of $y_i$ are positive (look what happens when, say, $y_1>0$, $x_2=x_3=...=x_n=1$ and $x_1\to\infty$). Assuming $y\leq 0$, a necessary and sufficient condition for $x>0$ to maximize the concave function $\phi(x)=\sum_i y_ix_i-\underbrace{\prod_i x_i^{-\pi_i}}_{\psi(x)}$ on its domain $\operatorname{int}\mathbf{R}^n_+$ is to solve the Fermat equation $\nabla\phi(x)=0$, that is, to satisfy

$$\pi_i\psi(x)x_i^{-1}=-y_i,\ i\leq n,$$

resulting in

$$f^*(y)=\left\{\begin{array}{ll}-(1+\sum_i\pi_i)\left[\prod_i[-y_i/\pi_i]^{\pi_i}\right]^{\frac{1}{1+\sum_i\pi_i}}, & y\leq 0\\ +\infty, & \text{otherwise}\end{array}\right..$$

## Miscellaneous exercises

**Exercise III.25.** [multi-factor Hölder inequality]

Given positive reals $q_1,...,q_n$ and $p\in[1,\infty)$, we define the weighted $p$-norm of a vector $x\in\mathbf{R}^n$ as

$$|x|_p=\left(\sum_{j=1}^n q_j|x_j|^p\right)^{1/p}$$

This clearly is a norm which becomes the standard norm $\|\cdot\|_p$ when $q_j=1$, $j\leq n$. Same as $\|x\|_p$, the quantity $|x|_p$ has limit, namely, $\|x\|_\infty$, as $p\to\infty$, and we define $|\cdot|_\infty$ as this limit.

Now let $p_i$, $i\leq k$, be positive reals such that

$$\sum_{i=1}^k\frac{1}{p_i}=1.$$

1. Prove that for nonnegative reals $a_1,...,a_k$ one has

$$a_1a_2...a_k\leq\frac{a_1^{p_1}}{p_1}+\ldots+\frac{a_k^{p_k}}{p_k}$$

or, equivalently (set $b_i=a_i^{p_i}$)

$$\forall b\geq 0: b_1^{1/p_1}b_2^{1/p_2}...b_k^{1/p_k}\leq\frac{b_1}{p_1}+\frac{b_2}{p_2}+...+\frac{b_k}{p_k}.$$

Note: the special case $p_i=k$, $i\leq k$, of this inequality is the inequality between the geometric and the arithmetic means.

*Solution:* Set $\lambda_i=1/p_i$, so that $\lambda_i\geq 0$ and $\sum_i\lambda_i=1$. The claim is evident if some of $a_i\geq 0$ are equal to 0. Assuming $a_i>0$ for all $i$ and taking into account that $\ln(s)$ is concave on the positive ray, we have

$$\sum_i\lambda_i\ln(a_i^{p_i})\leq\ln\left(\sum_i\lambda_ia_i^{p_i}\right)$$

whence, taking the exponents of both sides and recalling what $\lambda_i$ are,

$$a_1...a_k\leq\sum_i\frac{a_i^{p_i}}{p_i}.$$ □

2. Let $x^1, ..., x^k \in \mathbf{R}^n$, and let $x^1x^2 \ldots x^k$ be the entrywise product of $x^1, ..., x^k$:

$$[x^1x^2 \ldots x^k]_j = x^1_j x^2_j \cdots x^k_j,\ 1 \leq j \leq n.$$

Prove that

$$|x^1x^2 \ldots x^k|_1 \leq \sum_{i=1}^k \frac{|x^i_i|^{p_i}_{p_i}}{p_i}. \qquad (*)$$

*Solution:* Set $x = x^1x^2 \ldots x^k$ and $y^i_j = q_j^{1/p_i}|x^i_j|$, so that $\prod_{i=1}^k y^i_j = q_j \prod_{i=1}^k |x^i_j|$ due to $\sum_i 1/p_i = 1$. We have

$$\begin{array}{rcl} q_j|x_j| & = & q_j \prod_i |x^i_j| = \prod_i y^i_j \\ & \leq & \sum_i [y^i_j]^{p_i}/p_i \text{ [by item 1]} \\ & = & \sum_i q_j |x^i_j|^{p_i}/p_i, \text{ [by definition of } y^i_j] \end{array}$$

that is, $q_j|x_j| \leq \sum_{i=1}^k q_j|x^i_j|^{p_i}/p_i$. Summing up these inequalities over $j = 1, ..., n$, we get $(*)$. □

3. Prove *multi-factor Hölder inequality*: for vectors $x^i \in \mathbf{R}^n$, $i \leq k$, one has

$$|x^1x^2 \ldots x^k|_1 \leq |x^1|_{p_1}|x^2|_{p_2} \cdots |x^k|_{p_k} \qquad (\#)$$

*Solution:* (#) clearly holds true when some of $x^i$ are zero vectors. Assuming $|x^i|_{p_i} > 0$ for all $i$, observe that both sides in (#) are positively homogeneous, of degree 1, w.r.t. every one of $x^i$: when multiplying $x^i$ by $t$, both sides are multiplied by $|t|$. As a result, to verify (#) for nonzero $x^i$ is the same as to verify this inequality when $|x^i|_{p_i} = 1$ for all $i$. But in this case, invoking item 2,

$$|x^1x^2 \ldots x^k|_1 \leq \sum_{i=1}^k |x^i|^{p_i}_{p_i}/p_i = \sum_{i=1}^k 1/p_i = 1,$$

exactly as stated by (#) in the case of $|x^i|_{p_i} = 1$, $i \leq k$. □

Note: we have proved (#) for positive reals $p_1, ..., p_k$ with $\sum_i 1/p_i = 1$. From the reasoning it is immediately seen that the (#) remains true when $p_i = \infty$ for some $i$ (and, of course, $1/p_i$ is set to 0 for these $i$).

**Exercise III.26.** [Muirhead's inequality]
For any $u \in \mathbf{R}^n$ and $z \in \mathbf{R}^n_{++} := \{z \in \mathbf{R}^n : z > 0\}$ define

$$f_z(u) = \frac{1}{n!}\sum_\sigma z^{u_1}_{\sigma(1)} \cdots z^{u_n}_{\sigma(n)},$$

where the sum is over all permutations $\sigma$ of $\{1, \ldots, n\}$. Show that if $P$ is a doubly stochastic $n \times n$ matrix, then

$$f_z(Pu) \leq f_z(u)\ \forall (u \in \mathbf{R}^n, z \in \mathbf{R}^n_{++}).$$

*Solution:* For $z \in \mathbf{R}^n_{++}$, $f_z(u)$ clearly is convex and permutation symmetric function of $u$; it remains to apply Lemma III.18.1.

**Exercise III.27.** Prove that a convex lsc function $f$ with polyhedral domain is continuous on its domain. Does the conclusion remain true when lifting either one of the assumptions that (a) convex $f$ is lsc, and (b) $\operatorname{Dom} f$ is polyhedral?

*Solution:* We should prove that if $\operatorname{Dom} f$ is polyhedral, $x_i \in \operatorname{dom} f$ converge to $\bar{x}$ as $i \to \infty$, then $f(x) = \lim_{i\to\infty} f(x_i)$. Passing to a subsequence, it suffices to prove this relation when the sequence $f(x_i)$ has a limit (finite or infinite) as $i \to \infty$. Finally, restricting $f$ from $\operatorname{Dom} f$ onto the intersection of $\operatorname{Dom} f$ with appropriate box, the situation reduces to the one where $\operatorname{Dom} f$ is polyhedral and bounded. Let $V = \operatorname{Ext}(\operatorname{Dom} f)$; then $V$ is nonempty finite set: $V = \{v_1, ..., v_N\}$, and $\operatorname{Dom} f = \operatorname{Conv}(V)$ (by Krein-Milman Theorem). Since $f$ is lsc, we have $s := \lim_{i\to\infty} f(x_i) \geq f(\bar{x})$. We want to prove that in fact $s = f(\bar{x})$; given that $s \geq f(\bar{x})$, all we need is to lead to a contradiction the assumption that $s > f(\bar{x})$. Assume that $s > f(\bar{x})$; then for some $\delta > 0$ we have $f(x_i) \geq f(\bar{x}) + \delta$ for all but finitely many values of $i$. Representing $x_i$ as a convex combination $\sum_{j=1}^N \lambda^i_j v_j$ of $v_j$ and passing to a subsequence, we can assume that the $N$ sequences $\{\lambda^i_j\}_{i\geq 1}$ have limits $\lambda_j$ as $i\to\infty$, so that $\lambda_j\geq 0$, $\sum_j\lambda_j=1$, and $\bar{x}=\lim_{i\to\infty}x_i=\sum_j\lambda_j v_j$, and, in addition, that $f(x_i)\geq f(\bar{x})+\delta$ for all $i$. Now let $J=\{j:\lambda_j>0\}$. For every $\theta>1$, we have

$$x_i^\theta := \bar{x}+\theta(x_i-\bar{x})=\sum_{j=1}^N \lambda^i_{j,\theta}v_j,\ \lambda^i_{j,\theta}=\lambda_j+\theta[\lambda^i_j-\lambda_j].$$

Note that $\sum_j\lambda^i_{j,\theta}=1$ for all $i$, same as $\lambda^i_{j,\theta}\geq 0$ for all $i$ provided that $j\not\in J$. When $j\in J$, we have $\lambda_j>0$, and therefore $\lambda^i_{j,\theta}\geq 0$ for all large enough values of $i$, due to $\lambda_j-\lambda^i_j\to 0$, $i\to\infty$. The bottom line is that for fixed $\theta$, all coefficients $\lambda^i_{j,\theta}$, $1\leq j\leq N$, are nonnegative for all large enough values of $i$. Consequently, $x_i^\theta$ for large $i$ is a convex combination of $v_j$ and therefore belongs to $\mathrm{Dom}\, f$. For $i$ such that $x_i^\theta\in\mathrm{Dom}\, f$ by convexity of $f$ we have

$$\delta\leq f(x_i)-f(\bar{x})\leq\theta^{-1}[f(x_i^\theta)-f(\bar{x})]$$

due to $x_i^\theta-\bar{x}=\theta[x_i-\bar{x}]$. We conclude that for every $\theta>1$ and all large enough values of $i$ we have $x_i^\theta\in\mathrm{Dom}\, f$ and $f(x_i^\theta)\geq f(\bar{x})+\theta\delta$. As a result, $f$ is not bounded from above on $\mathrm{Dom}\, f$, which is the desired contradiction, since $\max_{x\in\mathrm{Dom}\, f}f(x)=\max_{j\leq N}f(v_j)<\infty$ by convexity of $f$ combined with $\mathrm{Dom}\, f=\mathrm{Conv}\{v_1,...,v_N\}$. □

A convex non-lsc function with polyhedral domain can be discontinuous, e.g., $f(x)=\left\{\begin{array}{ll}1 & ,x=0\\ 0 & ,0<x\leq 1\end{array}\right.$, $\mathrm{Dom}\, f=[0,1]$. Similarly, a convex lsc function with non-polyhedral domain, even a closed one, can be discontinuous. To give an example, consider the following construction: we take the convex hull $E$ of the set $\{[x;y;0]:(x-1)^2+y^2\leq 1\}\cup\{[0;0;-1]\}$ and set $E^+=\{[x;y;t]:\exists\tau:[x;y;\tau]\in E\ \&\ t\geq\tau\}$. Clearly, $E^+$ is closed and convex and is the epigraph of some function $f$ with the domain $D=\{[x;y]:(x-1)^2+y^2\leq 1\}$. Since $E^+=\mathrm{epi}\{f\}$ is closed and convex, $f$ is convex lsc. At the same time, the intersection of $E^+$ and the line $\{[0;0;t]:t\in\mathbf{R}\}$ is the ray $\{[0;0;t]:t\geq -1\}$, so that $f(0,0)=-1$, and the intersection of $E^+$ and a line $\{[a;b;t]:t\in\mathbf{R}\}$ with $a>0$, $b$ satisfying $(a-1)^2+b^2=1$ is the ray $\{[a;b;t]:t\geq 0\}$, that is, $f(a,b)=0$ whenever $[a;b]$ is a boundary point of $D$ distinct from $[0;0]$. Since the boundary point $[0;0]$ of $D$ is the limit of a sequence of distinct from it boundary points of $D$, $f$ is not continuous on $D$.

**Exercise III.28.** Let $a_1,...,a_n>0$, $\alpha,\beta>0$. Solve the optimization problem

$$\min_x\left\{\sum_{i=1}^n\frac{a_i}{x_i^\alpha}:x>0,\sum_i x_i^\beta\leq 1\right\}$$

*Solution:* Passing to variables $y_i=x_i^\beta$, we convert the problem to a *convex* program

$$\min_y\left\{\sum_i a_iy_i^{-\alpha/\beta}:y>0,\sum_i y_i\leq 1\right\}$$

KKT conditions (where we guess that the constraint is active) read

$$\begin{array}{l}-\frac{\alpha}{\beta}a_iy_i^{-\frac{\alpha}{\beta}-1}+\lambda=0,\ i=1,...,n\\ \sum\limits_i y_i=1\end{array}$$

whence

$$y_i=\frac{a_i^{\frac{\beta}{\alpha+\beta}}}{\sum\limits_j a_j^{\frac{\beta}{\alpha+\beta}}}\Longrightarrow x_i=\frac{a_i^{\frac{1}{\alpha+\beta}}}{\left(\sum\limits_j a_j^{\frac{\beta}{\alpha+\beta}}\right)^{1/\beta}}$$

Since the problem in $y$-variables is convex, the KKT point we have found is a globally optimal solution. The optimal value is

$$\left(\sum_j a_j^{\frac{\beta}{\alpha+\beta}}\right)^{\frac{\alpha+\beta}{\beta}}.$$

**Exercise III.29.** [computational study] Consider the following situation: there are $K$ "radars" with $k$-th of them capable to locate targets within ellipsoid $E_k = \{x \in \mathbf{R}^n : (x-c_k)^\top C_k(x-c_k) \le 1\}$ ($C_k \succ 0$); the measured position of target is

$$y_k = x + \sigma_k \zeta_k,$$

where $x$ is the actual position of the target, and $\zeta_k$ is the standard (zero mean, unit covariance) Gaussian observation noise; $\zeta_k$'s are independent across $k$. Given measurements $y_1, ..., y_K$ of target's location $x$ known to belong to the "common field of view" $E = \cap_k E_k$ of the radars, which we assume to possess a nonempty interior, we want to estimate a given linear form $e^\top x$ of $x$ by using linear estimate

$$\widehat{x} = \sum_k h_k^\top y_k + h.$$

We are interested in finding the estimate (e.g., the parameters $h_1, ..., h_K$, $h$) minimizing the risk

$$\mathrm{Risk2} = \max_{x \in E} \sqrt{\mathbf{E}\left\{\left[e^\top x - \sum\nolimits_k h_k^\top [x + \sigma_k \zeta_k] - h\right]^2\right\}}$$

1. Pose the problem as convex optimization program

   *Solution:* We have

   $$\begin{array}{l}\mathrm{Risk2}^2 = \max_{x\in E} \mathbf{E}\left\{\left[([e - \sum_k h_k]^\top x - h) - (\sum_k \sigma_k h_k^\top \zeta_k)\right]^2\right\|\\ = \max_{x\in E}\left[\left[e - \sum_k h_k]^\top x - h\right]^2 + \sum_k \sigma_k^2 h_k^\top h_k\right]\\ = \sum_k \sigma_k^2 h_k^\top h_k + \max_{x\in E}\left[[e - \sum_k h_k]^\top x - h\right]^2.\end{array}$$

   As a result, denoting by $\phi_E$ the support function of $E$, the problem of minimizing $\mathrm{Risk2}^2$ can be posed as convex optimization problem

   $$\mathrm{Opt} = \min_{h_1,...,h_K,h,t}\left\{t^2 + \sum_k \sigma_k^2 h_k^\top h_k : \phi_E(e - \sum_k h_k) \le t + h, \phi_E(\sum_k h_k - e) \le t - h\right\}$$

   By Exercise III.12.4, we have

   $$\phi_E(g) = \min_{g_1,\cdots,g_K}\left\{\sum_k \phi_{E_k}(g_k) : \sum_k g_k = g\right\},$$

   and by Exercise III.14.1,

   $$\phi_{E_k}(g) = \sqrt{gC_k^{-1}g} + g^\top c_k,$$

   so that the problem of interest becomes

   $$\mathrm{Risk2}^2 = \min_{h_k, g_k, f_k, k \le k, h, t}\left\{t^2 + \sum_k \sigma_k^2 h_k^\top h_k : \begin{array}{l}\sum_k [\|C_k^{-1/2} g_k\|_2 + c_k^\top g_k] \le t + h,\\ \sum_k g_k + \sum_k h_k = e\\ \sum_k [\|C_k^{-1/2} f_k\|_2 + c_k^\top f_k] \le t - h,\\ -\sum_k f_k + \sum_k h_k = e\end{array}\right\}$$

2. Process the problem numerically and look at the results.
   Recommended setup:

- $K=3$, $n=2$, $[c_1,c_2,c_3]=\left[\begin{array}{ccc}1.000 & -0.500 & -0.500\\ 0 & 0.866 & -0.866\end{array}\right]$,

$$C_1=\left[\begin{array}{cc}0.2500 & 0\\ 0 & 1.5000\end{array}\right],\ C_2=\left[\begin{array}{cc}1.1875 & 0.5413\\ 0.5413 & 0.5625\end{array}\right],\ C_3=\left[\begin{array}{cc}1.1875 & -0.5413\\ -0.5413 & 0.5625\end{array}\right]$$

- $\sigma_1=0.1, \sigma_2=0.2, \sigma_3=0.3$
- $e=[1;1]/\sqrt{2}$.

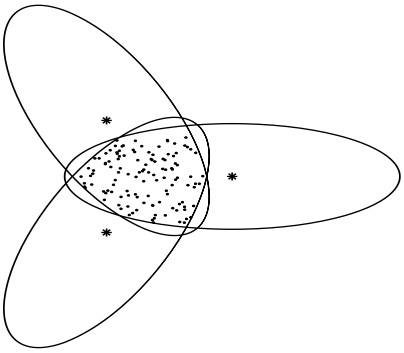


Figure III.5. 3 radars and their common filed of view (dotted)

*Solution:* Our results are as follows:

- $\text{Risk2}=0.0852$, $[h_1,h_2,h_3]=\left[\begin{array}{ccc}0.5130 & 0.1283 & 0.0570\\ 0.5136 & 0.1284 & 0.0571\end{array}\right]$, $h=0.0010$

**Exercise III.30.** For any $k\leq m$ and $X\in\mathbf{S}^m$, recall that $S_k(X)$ denotes the sum of $k$ largest eigenvalues of the matrix $X$. Given $X\in\mathbf{S}^m$, define $R[X]:=\left\{V^\top XV:\ V\in\mathcal{O}_m\right\}$ where $\mathcal{O}_m=\{V\in\mathbf{R}^{m\times m}: VV^\top=I_m\}$ is the set of all $m\times m$ orthogonal matrices. Prove that for any two symmetric matrices $X,Y\in\mathbf{S}^m$, we have

$$Y\in\operatorname{Conv}(R[X]) \text{ if and only if } S_k(Y)\leq S_k(X) \text{ for all } k<m \text{ and } \operatorname{Tr}(Y)=\operatorname{Tr}(X).$$

*Solution:* In one direction: suppose $Y\in\operatorname{Conv}(R[X])$, and let us prove that $S_k(Y)\leq S_k(X)$, $k\leq M$, with $S_m(Y)=S_m(X)$. Observe that a rotation $X\mapsto V^\top XV$, $V\in\mathcal{O}_m$, as every similarity transformation $X\mapsto Z^{-1}XZ$, preserves the vector of eigenvalues. It follows that the linear function $\operatorname{Tr}(Z)=S_m(Z)$ is equal to $S_m(X)$ on the entire $R[X]$ and is therefore equal to the same $S_m(X)$ on $\operatorname{Conv}(R[X])$. The bottom line is that $\operatorname{Tr}(Y)=S_m(Y)=S_m(X)=\operatorname{Tr}(X)$. Next, by the above argument $S_k(Z)$ is identically equal to $S_k(X)$ on the entire $R[X]$, and since $S_k(\cdot)$ is convex (see chapter 18), we conclude that $S_k(\cdot)$ is $\leq S_k(X)$ everywhere on $\operatorname{Conv}(R[X])$, whence $S_k(Y)\leq S_k(X)$, $k\leq m$.

In the opposite direction: let $S_k(Y)\leq S_k(X)$, $k\leq m$, and $S_m(Y)=S_m(X)$. By Majorization Principle, $\lambda(Y)=\pi\lambda(X)$ with doubly stochastic $\pi$. By Birkhoff Theorem (Theorem II.9.9), $\pi$ is a convex combination of permutation matrices $P_i$: $\pi=\sum_i\alpha_iP_i$ with $\alpha_i\geq0$ summing up to 1. Consequently, $\lambda(Y)=\sum_i\alpha_iP_i\lambda(X)$, or, which is clearly the same, $\operatorname{Diag}\{\lambda(Y)\}=\sum_i\alpha_iP_i\operatorname{Diag}\{\lambda(X)\}P_i^\top$. Next, $X=U\operatorname{Diag}\{\lambda(X)\}U^\top$ and $Y=V\operatorname{Diag}\{\lambda(Y)\}V^\top$ with $U,V\in\mathcal{O}_m$. The bottom line is

$$\begin{array}{l}Y=V\operatorname{Diag}\{\lambda(Y)\}V^\top=V\left[\sum_i\alpha_iP_i\operatorname{Diag}\{\lambda(X)\}P_i^\top\right]V^\top=V\left[\sum_i\alpha_iP_iU^\top XUP_i^\top\right]V^\top\\ =\sum_i\alpha_i\underbrace{VP_iU^\top}_{=:W_i^\top}XW_i\end{array}$$

The matrices $W_i$ are products of matrices form $\mathcal{O}_m$ and thus $W_i\in\mathcal{O}_m$, and we conclude that $Y\in\operatorname{Conv}(R[X])$. □

# Exercises from Part IV

## Around Conic Duality

**Exercise IV.1.** Given Linear Dynamical System

$$\begin{array}{rcl} x_0 &=& 0 \\ x_{t+1} &=& Ax_t + Bu_t,\ t = 0, 1, \dots, N-1 \end{array} \tag{LDS}$$

$(A : n \times n, B : n \times m)$ with controls $u_t$ subject to the "energy constraints"

$$\|u_t\|_2 \leq 1,\ 0 \leq t < N, \tag{EN}$$

pose the problem of minimizing $f^\top x_N$ ($f$ is a given vector) as a conic problem on the product of Lorentz cones, write down the conic dual of this problem and answer the following questions:

1. Is the problem essentially strictly feasible?
2. Is the problem bounded?
3. Is the problem solvable?
4. Is the dual problem feasible?
5. Is the dual problem solvable?
6. Are the optimal values equal to each other?
7. What do optimality conditions say?

*Solution:* Relation $\|u\|_2 \leq 1$ with $u \in \mathbf{R}^m$ is equivalent to $[u; 1] \in \mathbf{L}^{m+1}$, so that the problem of interest in the conic form reads

$$\min_{x,u,r} \{f^\top x_N : x_0 = 0,\ x_{t+1} = Ax_t + Bu_t,\ 0 \leq t \leq N-1, [u_t; 1] \in \mathbf{L}^{m+1}, 0 \leq t \leq N-1\} \tag{$P$}$$

To get the dual problem, we denote by $s_t \in \mathbf{R}^n$ the vectors of Lagrange multipliers for the state constraints $x_{t+1} - Ax_t - Bu_t = 0$, by $s_{-1}$ the vector of Lagrange multipliers for the equality constraints $x_0 = 0$ and by $[y_t; z_t] \in \mathbf{L}_*^{m+1} = \mathbf{L}^{m+1}$ – the Lagrange multipliers for the conic constraints. Aggregating constraints of $(P)$ with multipliers as the weights, we get the aggregated constraint

$$s_{-1}^\top x_0 + \sum\nolimits_{t=0}^{N-1} s_t^\top [x_{t+1} - Ax_t - Bu_t] + \sum\nolimits_{t=0}^{N-1} [y_t^\top u_t + z_t] \geq 0,$$

or, which is the same, the constraint

$$\begin{array}{l} s_{N-1}^\top x_N + [s_{N-2} - A^\top s_{N-1}]^\top x_{N-1} + [s_{N-3} - A^\top s_{N-2}]^\top x_{N-2} + \dots + [s_{-1} - A^\top s_0]^\top x_0 \\ + \sum_{t=0}^{N-1} [y_t - B^\top s_t]^\top u_t \geq -\sum_{t=0}^{N-1} z_t \end{array} \tag{$*$}$$

To get the dual problem, we add to the restrictions $\|y_t\|_2 \leq z_t$ (that is, restrictions $[y_t; z_t] \in \mathbf{L}^{m+1}$) the restriction that the left hand side in $(*)$ identically in $x$'s and $u$'s is $f^\top x_N$ and maximize under this restriction the right hand side in $(*)$. Thus, the dual problem is

$$\max_{y_t, z_t, s_t} \left\{ -\sum\nolimits_{t=0}^{N-1} z_t : \begin{array}{l} s_{N-1} = f, A^\top s_{t+1} = s_t,\ -1 \leq t \leq N-2, \\ y_t = B^\top s_t, 0 \leq t \leq N-1, \|y_t\|_2 \leq z_t, 0 \leq t \leq N-1 \end{array} \right\} \tag{$D$}$$

An optimal solution to the dual problem is evident:

$$s_t = [A^{N-1-t}]^\top f,\ -1 \leq t \leq N-1, y_t = B^\top [A^{N-1-t}]^\top f,\ z_t = \|B^\top [A^{N-1-t}]^\top f\|_2,$$

the optimal value is

$$-\sum\nolimits_{t=0}^{N-1}\|B^\top[A^{N-1-t}]^\top f\|_2.$$

The answers to the remaining questions are as follows:

1. Is the problem essentially strictly feasible? – Yes, $(P)$ is essentially strictly feasible, an essentially strictly feasible solution being, e.g. $u_t=0$, $0\leq t\leq N-1$, $x_t=0$, $0\leq t\leq N$
2. Is the problem bounded? – Yes, since the feasible set clearly is bounded
3. Is the problem solvable? – Yes, as every feasible problem with bounded feasible set (this set is automatically closed, and therefore the linear – and thus continuous – objective attains its minimum on this set)
4. Is the dual problem feasible? – Yes, by Conic Duality Theorem (not speaking about the fact that we see feasible solution by naked eyes)
5. Is the dual problem solvable? – Yes, by Conic Duality Theorem (not speaking about the fact that we see the optimal solution by naked eyes)
6. Are the optimal values equal to each other? – Yes, by Conic Duality Theorem
7. What do optimality conditions say? – They say that at the primal-dual optimum the primal slacks $[u_t;1]$ are orthogonal to the vectors $[y_t;z_t]=[B^\top[A^{N-1-t}]^\top f;\|B^\top[A^{N-1-t}]^\top f\|_2]$, that is,

   $$\|B^\top[A^{N-1-t}]^\top f\|_2+u_t^\top B^\top[A^{N-1-t}]^\top f=0,$$

   which combines with $\|u_t\|_2\leq 1$ and the Cauchy inequality to imply that whenever the vector $e_t=B^\top[A^{N-1-t}]^\top f$ is nonzero, we have $u_t=-e_t/\|e_t\|_2$, and when $e_t=0$, $u_t$ can be a whatever vector of norm not exceeding 1. Note that we got "closed form" solutions to both $(P)$ and $(D)$.

**Exercise IV.2.** Consider conic constraint $Ax-b\in K$ where $K\subset\mathbf{R}^m$ is a regular cone and matrix $A$ is of full column rank (i.e., has linearly independent columns, or, which is the same, has trivial kernel). Suppose that the constraint is feasible. Show that the following properties are all equivalent to each other:

(i) the feasible region $\{x\in\mathbf{R}^n: Ax-b\in K\}$ is bounded;
(ii) $\mathrm{Im}(A)\cap K=\{\ 0\ \}$, where $\mathrm{Im}(A):=\{Ax: x\in\mathbf{R}^n\}$;
(iii) the following system of vector inequalities is solvable

$$A^\top\lambda=0,\quad \lambda\in\mathrm{int}\,K_*.$$

Using these conclude that the property of whether a conic problem $\min_x\{c^\top x: Ax-b\in K\}$ has a bounded feasible region or not is independent of the choice of $b$, provided that the problem is feasible.

*Solution:* (i) $\Longleftrightarrow$ (ii): We are in the case when the feasible set $X=\{x:Ax-b\in K\}$ is nonempty (and clearly is closed). By Fact II.8.13 $X$ is bounded iff $X$ has no nonzero recessive directions, that is, iff the recessive cone of $X$ (which is $\{h:Ah\in -K\}$ (why?)) is trivial. Since $h\mapsto Ah$ is an embedding, the latter happens iff $\mathrm{Im}(A)\cap[-K]=\{0\}$, or, which is the same, iff $\mathrm{Im}(A)\cap K=\{0\}$. □

(iii) $\Longrightarrow$ (ii): With (iii) in force, there exists $\lambda\in\mathrm{int}\,K_*$ such that $A^\top\lambda=0$. If now $x\in\mathbf{R}^n$ is such that $y=Ax\in K$, we have $\lambda^\top y=[A^\top\lambda]x=0$, and since $y\in K$ and $\lambda\in\mathrm{int}K_*$, we conclude that $y=0$. The bottom line is that $\mathrm{Im}(A)\cap K=\{0\}$, that is, (ii) takes place. □

(ii) $\Longrightarrow$ (iii): Assume, on the contrary to what should be proved, that (ii) does take place, and (iii) does not. Then the convex nonempty set $\{\lambda: A^\top\lambda=0\}$ does not intersect $\mathrm{int}\,K_*$, which also is a nonempty convex set, implying, by Separation Theorem, that there exists $y\in\mathbf{R}^m$,$y\neq 0$, such that

$$\sup_{\lambda:A^\top\lambda=0} y^\top\lambda\leq\inf_{u\in\mathrm{int}\,K_*} y^\top u.$$

since the right hand side infimum is finite and $K_*$ is a cone, this infimum is 0, implying that $y\in(\mathrm{int}\,K_*)_*=K$. On the other hand, the supremum in the left hand side is taken over a linear subspace $\mathrm{Ker}A^\top$; it can be finite iff $y\in[\mathrm{Ker}A^\top]^\perp=\mathrm{Im}(A)$. Thus, $y$ is a nonzero vector from $\mathrm{Im}(A)\cap K$, which is impossible by (ii). □

Finally, consider a feasible conic constraint $Ax-b\in K$ with tegular cone $K$. If $\mathrm{Ker}A\neq\{0\}$, the feasible set of this constraint is unbounded independently of what $b$ is, since Ker$A$ is the recessive subspace of the feasible set, provided the latter is nonempty. And if Ker$A = \{0\}$, we, as we just have seen, are in the case where (i) is equivalent to (ii), and the validity status of (ii) is independent of what $b$ is.

**Exercise IV.3.**

Given a cone $K$ in a Euclidean space $E$ with an inner product $\langle\cdot,\cdot\rangle$, we call a pair of elements $x \in K$ and $y \in K_*$ *complementary* if $\langle x, y\rangle = 0$.

In this question, we will examine complementarity relations for the second-order cones $\mathbf{L}^n$ and the positive semidefinite cone $\mathbf{S}^n_+$.

1. Consider $\mathbf{L}^n := \left\{x = [\tilde{x}; x_n] \in \mathbf{R}^{n-1} \times \mathbf{R} : \ x_n \geq \|\tilde{x}\|_2\right\}$; as we know, this cone is self-dual (Example II.8.8). Prove that $x, s \in \mathbf{L}^n$ satisfy $\langle x, s\rangle = 0$ iff $x_n\tilde{s} + s_n\tilde{x} = 0$ holds.
2. Consider the space of $n \times n$ symmetric matrices, i.e., $E = \mathbf{S}^n$ equipped with the Frobenius inner product $\langle X, Y\rangle = \text{Tr}(XY) = \sum_{i=1}^n\sum_{j=1}^n X_{ij}Y_{ij}$. Let $K = \mathbf{S}^n_+ := \{X \in \mathbf{S}^n : x^\top Xx \geq 0\, \forall x \in \mathbf{R}^n\}$ be the positive semidefinite cone; recall that this cone is self-dual (Example II.8.9). Prove that $X, Y \in \mathbf{S}^n_+$ are complementary, i.e., $\langle X, Y\rangle = 0$, iff their matrix product is zero, i.e., $XY = YX = 0$. In particular, matrices from a complementary pair commute and therefore share a common orthonormal eigenbasis.

*Solution:*

1. The statement clearly is true when $s_n = 0$ or $x_n = 0$. Assuming $x_n > 0$, $s_n > 0$, the relation $[\tilde{x}; x_n]^\top[\tilde{s}; s_n] = 0$ for $[\tilde{x}; x_n]$ and $[\tilde{s}; s_n]$ from $\mathbf{L}^n$ means that $s_nx_n = -\tilde{s}^\top\tilde{x}$ together with $\|\tilde{x}\|_2 \leq x_n$ and $\|\tilde{s}\|_2 \leq s_n$, which, by Cauchy inequality (Theorem B.1) may happen iff $\tilde{x} = x_ne$ and $\tilde{s} = -s_ne$ for unit vector $e$, or, which is the same, when $x_n\tilde{s} + s_n\tilde{x} = 0$. □
2. For $X, Y \in \mathbf{S}^n$ we have $[XY]^\top = YX$, whence $XY = 0$ iff $YX = 0$. Clearly when $XY = YX = 0$, we have $\text{Tr}(XY) = 0$. So we will prove the reverse direction. Assume that $X \succeq 0, Y \succeq 0$ and $\text{Tr}(XY) = 0$, and let us prove that $XY = 0$. Indeed, we have
$$0 = \text{Tr}(XY) = \text{Tr}(X^{1/2}[X^{1/2}Y^{1/2}]Y^{1/2}) = \text{Tr}([X^{1/2}Y^{1/2}][X^{1/2}Y^{1/2}]^\top) = \sum_{i,j}[X^{1/2}Y^{1/2}]^2_{ij},$$
whence $X^{1/2}Y^{1/2} = 0$, so that $XY = X^{1/2}[X^{1/2}Y^{1/2}]Y^{1/2} = 0$. □

**Exercise IV.4.** By General Theorem on Alternative, a system of $m$ scalar linear constraints $Ax \geq b$ in variables $x \in \mathbf{R}^n$ (or, which is the same, the conic inequality $Ax \geq_{\mathbf{R}^m_+} b$) has no solutions iff it can be led to contradiction by aggregation: there exist nonnegative weights $\lambda_1, ..., \lambda_m$ such that the associated weighted sum $\lambda^\top Ax \geq \lambda^\top b$ of inequalities from the system is a contradictory inequality, that is, $A^\top\lambda = 0$ and $b^\top\lambda > 0$. For a general conic constraint of the form
$$Ax \geq_{\mathbf{K}} b \tag{I}$$
where $\mathbf{K} \subset \mathbf{R}^m$ is a regular cone, similar recipe for certifying infeasibility would read
$$\exists\lambda \in \mathbf{K}_* : A^\top\lambda = 0 \ \&\ b^\top\lambda > 0. \tag{II}$$
The goal of Exercise is to investigate relation between feasibility statuses of (I) and of (II).

Your first task is easy:

1. Prove that if (II) is feasible, then (I) is infeasible.

*Solution:* Let $\lambda$ satisfy (II). To prove that (I) has no solutions, assume, on the contrary, that $x$ solves (I). Then $\lambda^\top Ax \geq \lambda^\top b$ due to $\lambda \in \mathbf{K}_*$, which with our $\lambda$ results in $0 \geq b^\top\lambda$, which is not the case due to $\lambda^\top b > 0$; this is the desired contradiction.

The rest of your effort is aimed at investigating to which extent item 1 can be inverted: if and when it is true that when (II) has no solutions, then (I) is feasible? General Theorem on Alternative says that this indeed is the case when $\mathbf{K}$ is the nonnegative orthant $\mathbf{R}^m_+$. In the general case, the situation is different.

2. Let (I) be the univariate conic inequality

$$Ax := [1;0;1]x \geq_{\mathbf{L}^3} b := [0;1;0] \qquad \text{(i)}$$

where $\mathbf{L}^3$ is the 3D Lorentz cone. Write down the associated system (II) and check that both this system and (i) are infeasible. Conclude from this example that in general, solvability of (II) is only sufficient, but not necessary, condition for infeasibility of (I).

*Solution:* Recalling what $\mathbf{L}^3$ is, (i) is the scalar inequality $x \geq \sqrt{x^2+1}$; of course, it is infeasible. Now, the associated system (II) reads

$$\lambda_3 \geq \sqrt{\lambda_1^2+\lambda_2^2}, \lambda_1+\lambda_3=0, \lambda_2>0$$

in real variables $\lambda_1,\lambda_2,\lambda_3$; $\lambda_1$ can be immediately eliminated, resulting in clearly infeasible system $\lambda_3 \geq \sqrt{\lambda_3^2+\lambda_2^2}$, $\lambda_2>0$. □

3. Prove that (II) is infeasible iff (I) is *nearly feasible*, meaning that for every $\epsilon>0$ there exists $b'$ such that $\|b'-b\|_2 \leq \epsilon$ and the conic constraint $Ax \geq_{\mathbf{K}} b'$ is feasible. Equivalently: (II) is infeasible iff $b$ belongs to the closure $\overline{B}$ of the set $B = A\mathbf{R}^n - \mathbf{K}$ of those right hand side vectors in (I) for which (I) is feasible.

*Solution:* Taking into account item 1, the claim we want to prove is that (II) has no solutions iff $b \in \overline{B}$, or, which is the same,

(!) (II) has a solution iff $b \notin \overline{B}$.

Justification of (!) is immediate. In one direction: if (II) has a solution $\lambda$, then $A^\top\lambda=0$ and $A^\top b'>0$ for all $b'$ close enough to $b$, implying, by item 1, that all these close enough to $b$ vectors $b'$ , when treated as the right hand sides in (I), result in infeasible conic constraint, that is, do not belong to $B$. Thus, there is a neighbourhood of $b$ which does not intersect $B$, implying that $b \notin \overline{B}$.
In the opposite direction: assume that $b \notin \overline{B}$, and let us prove that (II) is feasible. Since $b \notin \overline{B}$, $b$ is at a positive distance from the nonempty convex set $B = \{z : z = Ax - y, x \in \mathbf{R}^n, y \in \mathbf{K}\}$, implying by the Separation Theorem that $\{b\}$ can be strongly separated from $B$: for properly selected $\lambda$ it holds

$$\lambda^\top b > \sup_{z\in B} \lambda^\top z = \sup_{x\in\mathbf{R}^n, y\in\mathbf{K}} \lambda^\top[Ax-y]. \qquad (*)$$

the concluding supremum here is finite, implying that $A^\top\lambda=0$ (otherwise we could make $\lambda^\top[Ax-y]$ arbitrarily large by properly selecting $x$ and setting $y=0$) and $\lambda \in \mathbf{K}_*$ (otherwise we could make $\lambda^\top[Ax-y]$ arbitrarily large by properly selecting $y \in \mathbf{K}$ and setting $x=0$). Thus, $A^\top\lambda=0$ and $\lambda\in\mathbf{K}_*$, implying that the supremum in $(*)$ is 0, that is, $\lambda^\top b>0$; we conclude that $\lambda$ solves (II), so that the latter system is feasible. □

Conclusion: *Feasibility of* (II) *is necessary and sufficient for infeasibility of* (I) *iff the set* $B = A\mathbf{R}^n - \mathbf{K}$ *of the right hand sides in the conic constraint* (I) *resulting in constraint's feasibility is closed; in fact, feasibility of* (II) *is necessary and sufficient condition for* $b$ *not to belong to the closure* $\overline{B}$ *of* $B$. Now, when $\mathbf{K} = \{y : Py \geq 0\}$ is a polyhedral cone, e.g., $\mathbf{R}^m_+$, $B$ is polyhedral (since its definition in the case under consideration is its polyhedral representation as well) and therefore is closed, which explains why when the cone $\mathbf{K}$ is polyhedral infeasibility of (II) is equivalent to feasibility of (I). At the same time, when $\mathbf{K}$ is not polyhedral, $B$ can be non-closed, as is the case in example from item 2. Let us look at the geometry of this example. (i) wants of us to find a point in the intersection of the cone $\mathbf{L}^3 = \{x \in \mathbf{R}^3 : x_3 \geq \sqrt{x_1^2+x_2^2}\}$ with the line $\ell = \{[t;-1;t] \in \mathbf{R}^3 : t \in \mathbf{R}\}$. $\ell$ belongs to the 2D plane $L = \{x \in \mathbf{R}^3 : x_2 = -1\}$, and the intersection of $\mathbf{L}^3$ with this plane is the set $\{[x_1;-1;x_3] : x_3^2 - x_1^2 \geq 1, x_3 \geq 0\}$, or, which is the same, the set $\{[x_1;-1;x_3] : (x_3-x_1)(x_3+x_1) \geq 1, x_3-x_1 \geq 0\}$; introducing the coordinates $u = x_3+x_1$, $v = x_3-x_1$ on the 2D plane $L$, the intersection of $L$ and $\mathbf{L}^3$ in these coordinates becomes the inner part $H = \{[u;v] : u \geq 1/v, v>0\}$ of the branch $\Gamma = \{[u;v] : uv=1, v>0\}$ of hyperbola. In $u,v$-coordinates the line $\ell$ is just the line $v=0$. Thus, geometrically the situation is as follows: to intersect $\ell$ and $\mathbf{L}^3$ is the same as to intersect $H$ with the $v$-axis of the $[u;v]$-plane; the intersection clearly is empty, so that (i) is infeasible. At the same time, our line is an asymptote of $\Gamma$, so that the shift $v=\epsilon$ of the line $v=0$ makes the intersection of the shifted line with $H$ nonempty, whatever small $\epsilon > 0$ be. The outlined shift of $\ell$ in our original $x$-coordinates reduces to passing from $b = [0; 1; 0]$ to $b_\epsilon = [0; 1; -\epsilon]$. The bottom line is that $b \notin B$ and $b \in \overline{B}$, since $b = \lim_{\epsilon \to +0} b_\epsilon$ and $b_\epsilon \in B$.

The result of item 3 attracts our attention to the following question: *What are natural sufficient conditions which guarantee the closedness of the set $A\mathbf{R}^n - \mathbf{K}$ ?* Here is a simple answer:

4. Prove that when the only common point of the image space $L := \{y \in \mathbf{R}^m : \exists x : y = Ax\}$ of $A$ and of $\mathbf{K}$ is the origin, the set $B := A\mathbf{R}^n - \mathbf{K} = L - \mathbf{K}$ is closed. Prove that the same holds true when the condition $L \cap \mathbf{K} = \{0\}$ is "heavily violated," meaning that $L \cap \mathrm{int}\, \mathbf{K} \neq \varnothing$.

*Solution:* Assume that $L \cap \mathbf{K} = \{0\}$, and let us prove that $B$ is closed. Thus, let $b_i = y_i - z_i$ with $y_i \in L$ and $z_i \in \mathbf{K}$, and let $b_i \to b$ as $i \to \infty$; we need to prove that then $b \in B$. Consider two cases: (a) the sequence $\{z_i\}$ is bounded, and (b) the sequence $\{z_i\}$ is unbounded. In the case of (a), passing to a subsequence, we can assume that $z_i \to \overline{z}$ as $i \to \infty$; since $z_i \to \overline{z}$ and $b_i \to b$ as $i \to \infty$, we conclude that $y_i = b_i + z_i \to b + \overline{z} =: \overline{y}$ as $i \to \infty$. As $\mathbf{K}$ is closed and $z_i \in \mathbf{K}$, we have $\overline{z} \in \mathbf{K}$. By its origin, $\overline{y}$ is the limit of a converging sequence of points from $L$ and thus $\overline{y} \in L$. We see that $b = \overline{y} - \overline{z} \in B$, as claimed. In the case of (b), passing to a subsequence, we can assume that $r_i := \|z_i\|_2 \to \infty$ as $i \to \infty$; since $z_i = y_i - b_i$ and the sequence $\{b_i\}$ converges and is therefore bounded, we conclude that $r_i^{-1}\|y_i\|_2 \to 1$ as $i \to \infty$. Passing to a subsequence, we can further assume that the unit vectors $\overline{z}_i := r_i^{-1} z_i$ converge as $i \to \infty$ to some unit vector $\overline{z}$, and the sequence of vectors $\overline{y}_i = r_i^{-1} y_i$ converges as $i \to \infty$ to some vector $\overline{y}$, which also is unit due to $r_i^{-1}\|y_i\|_2 \to 1$, $i \to \infty$. We have

$$r_i^{-1} y_i - r_i^{-1} z_i = r_i^{-1} b_i; \qquad (*)$$

since $\{b_i\}$ is a bounded sequence and $r_i \to \infty$, $i \to \infty$, passing to limit in $(*)$ we get $\overline{y} = \overline{z}$. By its origin, $\overline{y}$ is the limit of sequence of points from $L$ and thus $\overline{y} \in L$, and $\overline{z}$ is the limit of a sequence of points from the closed cone $\mathbf{K}$ and therefore $\overline{z} \in \mathbf{K}$. The bottom line is that in the case of (b) the set $L \cap \mathbf{K}$ contains the unit vector $\overline{z} = \overline{y}$, which is impossible due to $L \cap \mathbf{K} = \{0\}$. Thus, (b) is impossible, and we are done.

Finally, in the case of $L \cap \mathrm{int}\, \mathbf{K} \neq \varnothing$ $B$ is closed by a very simple reason – in this case $B = \mathbf{R}^m$. Indeed, if $a \in \mathrm{int}\, \mathbf{K} \cap L$, then $\lambda a - b \in \mathbf{K}$ for all large enough positive $\lambda$, that is, $b = \lambda a - z$ for certain $\lambda > 0$ and $z \in \mathbf{K}$. And since $a \in L$, we have $\lambda a \in L$ as well, that is, $b \in L - \mathbf{K}$. □

**Exercise IV.5.** [follow-up to Exercise IV.4] Let $\mathbf{K} \subset \mathbf{R}^m$ be a regular cone, $P \in \mathbf{R}^{m \times n}$, $Q \in \mathbf{R}^{m \times k}$, and $p \in \mathbf{R}^m$. Consider the set

$$\overline{K} \quad = \quad \{x \in \mathbf{R}^n : \exists u \in \mathbf{R}^k : Px + Qu + p \in \mathbf{K}\}$$

This set clearly is convex. When the cone $\mathbf{K}$ is polyhedral, the above description of $\overline{K}$ is its polyhedral representation, so that the set $\overline{K}$ is polyhedral and as such is closed.

The goal of this exercise is to understand what happens with closedness of $\overline{K}$ when $\mathbf{K}$ is a general-type regular cone.

1. Is it true that $\overline{K}$ is closed whenever $\mathbf{K}$ is a regular cone?
   *Hint:* Look what happens when $\mathbf{K} = \mathbf{L}^3$, $P = I_3$, $Q = [0; 1; 1] \in \mathbf{R}^{3 \times 1}$, and $p = [0; 0; 0]$
2. Prove when $\mathbf{K}$ is a regular cone and $\mathrm{Im} Q \cap \mathbf{K} = \{0\}$, $\overline{K}$ is closed.

*Solution:* 1: In the situation of Hint. denoting by $L$ the linear subspace $\{x \in \mathbf{R}^3 : x_2 = 0\}$ we have

$$\overline{K} \cap L = \{x = [x_1; 0; x_3] : \exists u \in \mathbf{R} : [x_1; u; x_3 + u] \in \mathbf{L}^3\} = \{x = [x_1; 0; x_3] : \exists u : x_3 + u \geq \sqrt{x_1^2 + u^2}\}.$$

From the concluding description of $\overline{K} \cap L$ we see that this set contains all triples $[1; 0; \epsilon]$ with $\epsilon > 0$ and does not contain the triple $[1; 0; 0]$ and therefore is not closed; consequently, $\overline{K}$ is not closed as well ($L$ is closed!).

2: Assuming $\mathbf{K}$ regular and $L \cap \mathbf{K} = \{0\}$, where $L = \mathrm{Im} Q$, the set $Z = \{b \in \mathbf{R}^m : \exists u \in L : u + b \in \mathbf{K}\}$ is closed (by Exercise IV.4.4); it remains to note that $\overline{K}$ is the inverse image of the closed set $Z$ under the continuous mapping $x \mapsto Px + p$ and as such is closed along with $Z$. □

**Exercise IV.6.** Let $\mathfrak{n}(x)$ be a norm on $\mathbf{R}^n$ such that $\mathfrak{n}$ is continuously differentiable outside of the origin, and let

$$\mathfrak{n}_*(y) = \max_x \{y^\top x : \mathfrak{n}(x) \leq 1\}.$$

be the norm conjugate to $\mathfrak{n}$ (see Fact III.17.4), so that $\mathfrak{n}_*(\cdot)$ is a norm such that

$$x^\top y \leq \mathfrak{n}(x)\mathfrak{n}_*(y)\,\forall x, y \in \mathbf{R}^n$$

and $(\mathfrak{n}_*)_* = \mathfrak{n}$, implying that for every $x \neq 0$ there exists $y \neq 0$ such that

$$x^\top y = \mathfrak{n}(x)\mathfrak{n}_*(y).$$

Here are your tasks:

1. Let $M$ be a $d \times d$ matrix, $d \geq 2$, with diagonal entries equal to 1. Assume that $M\lambda \leq 0$ for some nonzero vector $\lambda \geq 0$. How large could be $\min_{i,j} M_{ij}$ ?

   *Solution:* $\mu := \min_{i,j} M_{i,j} \leq -\frac{1}{d-1}$. Indeed, assuming that $M\lambda \leq 0$ with nonzero $\lambda \geq 0$, let $k$ be the index of largest entry in $\lambda$. We have

   $$0 \geq \sum_j M_{kj}\lambda_j = \lambda_k + \sum_{j \neq k} M_{kj}\lambda_j \geq \lambda_k + \lambda_k(d-1)\mu \implies \mu \leq -\frac{1}{d-1}.$$

   For the matrix $M$ with diagonal entries equal to 1 and off-diagonal entries equal to $-1/(d-1)$ we have $M[1;...;1] = 0$, that is, the bound $\min_{i,j} M_{ij} \leq -\frac{1}{d-1}$ is unimprovable.

2. For $d \geq 2$, let $p_1, ..., p_d$ be $\mathfrak{n}_*(\cdot)$-unit vectors, $w_1, ..., w_d$ be $\mathfrak{n}(\cdot)$-unit vectors, and let $p_i^\top w_i = 1$, $1 \leq i \leq d$. Assume that $0 \in \mathrm{Conv}\{p_1, ..., p_d\}$. How small could be $\max_{i \neq j} \mathfrak{n}(w_i - w_j)$ ?

   *Solution:* $\max_{i,j \leq d} \mathfrak{n}(w_i - w_j) \geq \frac{d}{d-1}$. Indeed, consider the $d \times d$ matrix $M = [M_{ij} = w_i^\top p_j]_{i,j \leq d}$, We have $0 = \sum_j \lambda_j p_j$ with properly selected $\lambda \geq 0$ such that $\sum_j \lambda_j = 1$, so that $M\lambda = 0$. Besides this, the diagonal entries in $M$ are equal to 1. By the previous item, we have $M_{ij} \leq -\frac{1}{d-1}$ for some $i, j$, that is,

   $$\mathfrak{n}(w_j - w_i) = \mathfrak{n}_*(p_j)\mathfrak{n}(w_j - w_i) \geq p_j^\top [w_j - w_i] = 1 - M_{ij} \geq 1 + \frac{1}{d-1} = \frac{d}{d-1}.$$

3. Let $x \in \mathbf{R}^n$ be nonzero.
   1. Let $g = \nabla \mathfrak{n}(x)$.
      1. What is $\mathfrak{n}_*(g)$ ?

         *Solution:* For every $h \in \mathbf{R}^n$ we have $\mathfrak{n}(x+h) \geq \mathfrak{n}(x) + g^\top h$, whence also $\mathfrak{n}(x) + \mathfrak{n}(h) \geq \mathfrak{n}(x) + g^\top h$, that is, $\mathfrak{n}(h) \geq g^\top h$, implying that $\mathfrak{n}_*(g) = \max_h\{g^\top h : \mathfrak{n}(h) \leq 1\} \leq 1$. On the other hand, $0 = \mathfrak{n}(0) \geq \mathfrak{n}(x) - g^\top x$, that is, $g^\top x \geq \mathfrak{n}(x)$. Besides this, $g^\top x \leq \mathfrak{n}_*(g)\mathfrak{n}(x)$, and we get $\mathfrak{n}(x) \leq g^\top x \leq \mathfrak{n}_*(g)\mathfrak{n}(x)$, implying that $\mathfrak{n}_*(g) \geq 1$. The bottom line is that $\mathfrak{n}_*(g) = 1$.
      2. What is $g^\top x$ ?

         *Solution:* $g^\top x = \mathfrak{n}(x)$ – differentiate the identity $\mathfrak{n}(tx) = t\mathfrak{n}(x)$, $t > 0$, in $t$ at $t = 1$.
      3. Let $e$ be such that $\mathfrak{n}_*(e) \leq \mathfrak{n}_*(g)$ and $e^\top x = g^\top x$. Is it true that $e = g$ ?

         *Solution:* Yes. $e$ in question should satisfy $\mathfrak{n}_*(e) \leq \mathfrak{n}_*(g)$, that is, $\mathfrak{n}_*(e) \leq 1$ by item 3.1.1. Next, $e^\top x = g^\top x$, that is, $e^\top x = \mathfrak{n}(x)$ by item 3.1.2. Therefore for every $h$ it holds $e^\top h = e^\top(x+h) - e^\top x \leq \mathfrak{n}_*(e)\mathfrak{n}(x+h) - \mathfrak{n}(x) \leq \mathfrak{n}(x+h) - \mathfrak{n}(x)$, that is, $\mathfrak{n}(x) + e^\top h \leq \mathfrak{n}(x+h)$ for all $h$, so that $e$ is a subgradient of $f$ at $x$. Since $x \neq 0$ and $\mathfrak{n}$ is differentiable outside of the origin, we have $e = \nabla \mathfrak{n}(x) = g$.
   2. Given $N$ points $y_i \in \mathbf{R}^n$, consider the problem of finding the smallest $\mathfrak{n}(\cdot)$-ball containing $y_1, ..., y_N$.

1. Write down the problem as a conic one, and write down the conic dual of this problem. Are both the problems solvable with equal optimal values?

*Solution:* The problem in question is

$$\begin{array}{rcl}\mathrm{Opt}(P) & = & \min_{x,t}\{t : \mathfrak{n}(x-y_i)\leq t, 1\leq i\leq N\}\\ & = & \min_{[x;t]}\{t : [x-y_i;t]\in \mathbf{K}, i\leq N\}\\ & & \mathbf{K}=\{[u;t]\in\mathbf{R}^n\times\mathbf{R} : \mathfrak{n}(u)\leq t\}\end{array}\qquad (P)$$

its dual is

$$\begin{array}{rcl}\mathrm{Opt}(D) & = & \max_{z_1,\dots,z_N,s_1,\dots,s_N}\left\{\sum_i z_i^\top y_i : \begin{array}{l}\sum_i z_i=0, \sum_i s_i=1\\ {[z_i;s_i]}\in\mathbf{K}_*, i\leq N\end{array}\right\}\\ & & \mathbf{K}_*=\{[z;s]\in\mathbf{R}^n\times\mathbf{R} : \mathfrak{n}_*(z)\leq s\}\end{array}\qquad (D)$$

and both problems are solvable with equal optimal values.

Indeed, $(P)$ is self-explanatory; the fact that $\mathbf{K}$ is a regular cone is evident. The fact that the dual cone is as indicated in $(D)$ is immediate: denoting a vector from $\mathbf{R}^n\times\mathbf{R}$ by $[z;s]$, this vector is in the cone dual to $\mathbf{K}$ iff for every $t\geq 0$ one has

$$0\leq \min_u\{[u;t]^\top[z;s] : [u;t]\in\mathbf{K}\}=\min_u\{st+u^\top z : \mathfrak{n}(u)\leq t\}=st-t\mathfrak{n}_*(z),$$

implying that the dual cone is as in $(D)$. Now, to get the dual problem, we should equip the constraints $[x-y_i;t]\in\mathbf{K}$ of $(P)$ with Lagrange multipliers $[z_i;s_i]\in\mathbf{K}_*$ in such a way, that the left hand side in the aggregated constraint $\sum_i[z_i;s_i]^\top[x-y_i;t]\geq 0$, that is, in the inequality

$$\sum_i ts_i+\sum_i z_i^\top x\geq\sum_i z_i^\top y_i$$

is identically in $t,x$ equal to the primal objective $t$, and to maximize under this restriction (taken along with the restrictions $[z_i;s_i]\in\mathbf{K}_*$) the right hand side of the aggregated constraint, which results in $(D)$.

Problem $(P)$ clearly is strictly feasible (to get a strictly feasible solution, set $x=0$ and take $t>\max_i\mathfrak{n}(y_i)$) and solvable (since the sets of feasible solutions where the objective is upper-bounded by a given real are compact); by Conic Duality Theorem, the dual problem is solvable with the same optimal value as $(P)$.

2. Assume that the data are such that the optimal value in $(P)$ is equal to 1. How small can be $\max_{i,j}\mathfrak{n}(y_i-y_j)$ ?
*Hint:* write down and analyze optimality conditions.

*Solution:* $\max_{i,j}\mathfrak{n}(y_i-y_j)\geq\frac{n+1}{n}$.

Indeed, let $[x;1]$ be primal optimal, and $[z_i;s_i]$, $i\leq N$, be dual optimal. By optimality conditions (complementary slackness) the primal slacks $[x-y_i;1]$ should be orthogonal to the respective dual solutions $[z_i;s_i]$, that is,

$$s_i=[y_i-x]^\top z_i, i\leq N, \qquad (\#)$$

and since $\mathfrak{n}(y_i-x)\leq 1$ and $\mathfrak{n}_*(z_i)\leq s_i$ for all $i$ by primal and dual feasibility, we have $\mathfrak{n}(y_i-x)\leq\mathrm{Opt}(P)=1$, so that the right hand side in (#) is $\leq\mathfrak{n}(y_i-x)\mathfrak{n}_*(z_i)\leq\mathfrak{n}_*(z_i)$, and since $\mathfrak{n}_*(z_i)\leq s_i$ by dual feasibility, (#) implies that $\mathfrak{n}_*(z_i)=s_i$ for all $i$. Next, setting $w_i=y_i-x$, we have $\mathfrak{n}(w_i)\leq 1$, so that the right hand side in (#) is $\leq\mathfrak{n}(w_i)\mathfrak{n}_*(z_i)=\mathfrak{n}(w_i)s_i$; therefore (#) implies that $\mathfrak{n}(w_i)=1$ for all $i\in\mathcal{I}=\{i : s_i>0\}$. Note that the set $\mathcal{I}$ is nonempty, since otherwise $\mathrm{Opt}(D)$ would be 0 and not 1. Now, from the constraints of $(D)$ we have

$$\sum_{i\in\mathcal{I}}z_i=0,$$

and $\mathfrak{n}_*(z_i) = s_i > 0$ for $i \in \mathcal{I}$, so that setting $p_i = s_i^{-1} z_i$, $i \in \mathcal{I}$, we have

$$\mathfrak{n}(w_i) = 1,\, i \in \mathcal{I} \;\&\; \mathfrak{n}_*(p_i) = 1, i \in \mathcal{I} \;\&\; p_i^\top w_i = 1, i \in \mathcal{I} \;\&\; \sum_{i\in\mathcal{I}} s_i p_i = 0, \qquad (!)$$

where the relations $p_i^\top w_i = 1$, $i \in \mathcal{I}$, stem from (#) due to $s_i > 0$, $i \in \mathcal{I}$.

Since $0 < s_i$ for $i \in \mathcal{I} \neq \varnothing$, the last relation in (!) means that $0 \in \mathrm{Conv}\{p_i : i \in \mathcal{I}\}$. By Caratheodory Theorem, we can find a subset $I \subset \mathcal{I}$ of cardinality $d$, $2 \le d \le n+1$, such that $0 \in \mathrm{Conv}\{p_i, i \in I\}$. Assuming w.l.o.g. that $I = \{1, ..., d\}$, for $M = [w_i^\top p_j]_{i,j\le d}$ and some $\lambda \in \mathbf{R}^d_+$ with $\sum_i \lambda_i = 1$ we have

$$\mathfrak{n}_*(p_i) = 1, \mathfrak{n}(w_i) = 1, p_i^\top w_i = 1, i \le d \;\& \lambda \ge 0, \lambda \ne 0, M\lambda = 0 \qquad (!!)$$

By item 2, we have $\max_{i,j\le d} \mathfrak{n}(w_i - w_j) \ge \frac{d}{d-1} \ge \frac{n+1}{n}$, and it remains to note that $w_i - w_j = y_i - y_j$.

3. In the situation of item 3.2.2, assume that $\mathfrak{n}(x) = \|x\|_2$ is the standard Euclidean norm. How small can be $\max_{i,j} \mathfrak{n}(y_i - y_j)$ now?

   *Solution:* The concluding relation in the solution to the previous item now reads $\|p_i\|_2 = \|w_i\|_2 = 1$ and $p_i^\top w_i = 1$, $1 \le i \le d \le n+1$, whence $p_i = w_i$, $i \le d$. By item 1, (!!) implies that $\min_{i,j\le d} w_i p_j^\top \le -\frac{1}{d-1}$, that is, there exist $i, j \le d$ such that $w_i^\top p_j \le -\frac{1}{d-1}$, that is, $w_i^\top w_j \le -\frac{1}{d-1}$. Consequently,

$$\|w_i - w_j\|^2 = w_i^\top w_i + w_j^\top w_j - 2w_i^\top w_j \ge 2(1 + \frac{1}{d-1}) = \frac{2d}{d-1},$$

   that is, $\max_{i,j} \|y_i - y_j\|_2 \ge \max_{i,j\le d} \|w_i - w_j\|_2 \ge \sqrt{\frac{2d}{d-1}} \ge \sqrt{\frac{2(n+1)}{n}}$.

   Note: instead of asking how large is the maximum of pairwise distances between $y_i \in \mathbf{R}^n$ given that the smallest Euclidean ball containing $y_1, ..., y_N$ is of radius 1, we could ask how large could be radius of the smallest Euclidean ball containing the points $y_1, ..., y_N \in \mathbf{R}^n$ with pairwise $\|\cdot\|_2$-distances not exceeding 1, and in terms of the latter question, the above result states that this radius is at most $\sqrt{\frac{n}{2(n+1)}}$. This is called "Jung's Theorem;" the result is sharp, since the smallest radius Euclidean ball containing the $n+1$ vertices of the perfect simplex (simplex in $\mathbf{R}^n$ with distances 1 between every two vertices) is exactly $\sqrt{\frac{n}{2(n+1)}}$; to see this, realize the perfect simplex as $\{x \in \mathbf{R}^{n+1}_+ : \sum_i x_i = 1/\sqrt{2}\}$, and $\mathbf{R}^n$ - as the hyperplane $\sum_i x_i = 1/\sqrt{2}$ in $\mathbf{R}^{n+1}$.

## Geometry of primal-dual pair of conic problems

**Exercise IV.7.** [geometry of primal-dual pair of conic problem] The goal of the Exercise is to reveal notable geometry of primal-dual pair of conic problem.

It is convenient to work with the primal problem in the form

$$\mathrm{Opt}(P) = \min_x \left\{ c^\top x : Ax - b \ge_{\mathbf{K}} 0,\; Px = p \right\} \qquad (P)$$

where $\mathbf{K}$ is a regular cone in certain $\mathbf{R}^N$. As is immediately seen, the conic dual of $(P)$ reduces to the problem

$$\mathrm{Opt}(D) = \max_{y,z} \left\{ b^\top y + p^\top z : y \in \mathbf{K}_*,\; A^\top y + P^\top z = c \right\} \quad ^{11} \qquad (D)$$

From now on we make the following, in fact, rather weak,

[11] building conic dual to a conic problem is a purely mechanical process; however, this process as presented in section 23.4 operates with conic problem in a form slightly different from the one of $(P)$, namely, with linear inequality constraints instead of linear equalities. To apply this process to

**Assumption:** *The systems of linear equality constraints in $(P)$ and $(D)$ are solvable.*

Let us fix $\overline{x}$ and $(\overline{y}, \overline{z})$ such that

$$P\overline{x} = p \;\&\; A^\top \overline{y} + P^\top \overline{z} = c. \tag{\#}$$

Your first task is as follows:

1. Pass in $(P)$ from variables $x$ to *primal slack* $\xi = Ax - b$. Specifically, prove that in terms of primal slack $(P)$ becomes the problem

$$\begin{array}{c}\mathrm{Opt}(\mathcal{P}) = \min_\xi \left\{\overline{y}^\top \xi : \xi \in \mathbf{K} \cap [\mathcal{L} - \overline{\xi}]\right\} \\ \left[\mathcal{L} = \{\xi : \exists x : \xi = Ax, Px = 0\},\; \overline{\xi} = b - A\overline{x}\right]\end{array} \tag{$\mathcal{P}$}$$

namely, prove that

(i) Every feasible solution $x$ to $(P)$ induces feasible solution $\xi = Ax - b$ to $(\mathcal{P})$, and the value of the objective of $(P)$ at $x$ differs from the value of the objective of $(\mathcal{P})$ at $Ax - b$ by the independent of $x$ constant:

$$\overline{y}^\top \xi = c^\top x - \left[\overline{y}^\top b + \overline{z}^\top p\right]. \tag{$A$}$$

(ii) Vice versa, every feasible solution $\xi$ to $(\mathcal{P})$ is of the form $Ax - b$ for some feasible solution $x$ to $(P)$.

The bottom line is that $(P)$ can be reformulated equivalently as $(\mathcal{P})$, and the optimal values of these two problems are linked by the relation

$$\mathrm{Opt}(\mathcal{P}) = \mathrm{Opt}(P) - \left[\overline{y}^\top b + \overline{z}^\top p\right].$$

*Solution:* Let $x$ be feasible for $(P)$ and $\xi = Ax - b$. Then $\xi$ satisfies the inclusion $\xi \in \mathbf{K}$ and

$$\xi = A[x - \overline{x}] + [A\overline{x} - b] = A[x - \overline{x}] - \overline{\xi}$$

and $P[x - \overline{x}] = 0$, that is, $\xi \in \mathbf{K} \cap [\mathcal{L} - \overline{\xi}]$. This reasoning can be easily reversed to demonstrate that if $\xi \in \mathbf{K} \cap [\mathcal{L} - \overline{\xi}]$, then $\xi = Ax - b$ for some $x$ feasible for $(P)$. Besides this,

$$c^\top x = [A^\top \overline{y} + P^\top \overline{z}]^\top x = \overline{y}^\top [Ax - b] + \overline{y}^\top b + \overline{z}^\top Px = \overline{y}^\top \xi + \left[\overline{y}^\top b + \overline{z}^\top p\right],$$

as claimed in $(A)$.

On the other hand, when $\xi$ is feasible for $(\mathcal{P})$, we have $\xi \in \mathbf{K}$ and $\xi = Ax' - \overline{\xi}$ for some $x'$ with $Px' = 0$, whence

$$\mathbf{K} \ni \xi = Ax' - \overline{\xi} = A[x' + \overline{x}] - b = Ax - b,$$

where $x = x' + \overline{x}$ satisfies $Px = p$. We conclude that $\xi = Ax - b$ with $x$ feasible for $(P)$. □

Next task is as follows:

2. Pass from problem $(D)$ in variables $y$, $z$ to problem

$$\begin{array}{c}\max\limits_y \left\{\overline{\xi}^\top y : y \in \mathbf{K}_* \cap [\mathcal{L}^\perp + \overline{y}]\right\} \\ \left[\mathcal{L}^\perp := \{y : y^\top \xi = 0 \,\forall \xi \in \mathcal{L}\} = \{y : \exists z : A^\top y + P^\top z = 0\}\right]\end{array} \tag{$\mathcal{D}$}$$

in variable $y$ only, specifically, prove that

(i) The orthogonal complement $\mathcal{L}^\perp$ of $\mathcal{L}$ indeed is the linear subspace $\{y : \exists z : A^\top y + P^\top z = 0\}$.

(ii) $y$-component of feasible solution $(y, z)$ to $(D)$ is a feasible solution to $(\mathcal{D})$, and vice versa –

$(P)$, it suffices to represent the linear equalities $Px = p$ by a pair of opposite linear inequalities $Px - p \geq 0, -Px + p \geq 0$. Applying the recipe from section 23.4 to the resulting problem, the dual reads

$$\max_{y,z',z''} \left\{b^\top y + [z' - z'']^\top p : A^\top y + P^\top [z' - z''] = c,\, y \in \mathbf{K}_*, z' \geq 0, z'' \geq 0\right\}.$$

Passing from $z', z''$ to $z = z' - z''$, we reduce the latter problem to $(D)$.

every feasible solution $y$ to $(\mathcal{D})$ can be augmented by $z$ to yield a feasible solution $(y,z)$ to $(D)$. Besides this, whenever $(y,z)$ is feasible for $(D)$, we have

$$b^\top y + p^\top z = \overline{\xi}^\top y + c^\top \overline{x}. \qquad (B)$$

The bottom line is that $(D)$ can be reformulated equivalently as $(\mathcal{D})$, and the optimal values of these two problems are linked by the relation

$$\mathrm{Opt}(\mathcal{D}) = \mathrm{Opt}(D) - c^\top \overline{x}.$$

*Solution:* (i): To prove that $\mathcal{L}^\perp = \{y : \exists z : A^\top y + P^\top z = 0\}$ is the same as to prove that the necessary and sufficient condition for equality $y^\top \xi = 0$ treated as equality in variables $\xi, x$ to be consequence of the system of linear equalities $\xi - Ax = 0, Px = 0$ in variables $\xi, x$ is for $y$ to admit selection of $z$ such that $A^\top y + P^\top z = 0$, but this is what Linear Algebra (not speaking about Homogeneous Farkas Lemma) says: a homogeneous linear equation is a consequence of a system of homogeneous linear equations if and only is the vector of coefficients of this equation (in our case, the vector $[y^\top, 0_{1\times n}]$) is linear combination of the vectors of coefficients of the equations from the system, which in the case in question boils down to $y^\top A + z^\top P = 0$ for certain $z$. □

(ii) If $(y,z)$ is feasible for $(D)$, then $[A^\top, P^\top][y - \overline{y}; z - \overline{z}] = 0$, that is, $y \in \mathcal{L}^\perp + \overline{y}$ by already proved (i), and $y \in \mathbf{K}_*$, that is, $y$ is feasible for $(\mathcal{D})$. Besides this, $A^\top y + P^\top z = c$, $\overline{\xi} = b - A\overline{x}$, and $P\overline{x} = p$, whence

$$\overline{\xi}^\top y - [b^\top y + p^\top z] = [b - A\overline{x}]^\top y - [b^\top y + p^\top z] = -\overline{x}^\top A^\top y - p^\top z = \overline{x}^\top [P^\top z - c] - p^\top z = -\overline{x}^\top c,$$

as required in $(B)$.

Vice versa, if $y$ is feasible for $(\mathcal{D})$, then $y \in \mathbf{K}_*$ and $y - \overline{y} \in \mathcal{L}^\perp$, that is, by (i), for properly selected $w$ one has $A^\top [y - \overline{y}] + P^\top w = 0$. This, due to the origin of $\overline{y}$ implies that

$$A^\top y + P^\top w = A^\top \overline{y} = c - P^\top \overline{z},$$

so that $y$ can be augmented by $z = \overline{z} + w$ to yield a feasible solution to $(D)$. (ii) is proved. □

The summary of items 1 and 2 is as follows:

- Primal-dual pair $(P)$, $(D)$ of conic problems reduces to pair of problems $(\mathcal{P})$, $(\mathcal{D})$, "reduces" meaning that feasible solutions $x$ and $(y,z)$ to $(P)$, $(D)$ induce feasible solutions $\xi = Ax - b$ and $y$ to $(\mathcal{P})$, $(\mathcal{D})$, and every pair of feasible solutions to the latter problems can be obtained, in the fashion just described, from a pair of feasible solutions to $(P)$, $(D)$;
- Geometrically, $(\mathcal{P})$, $(\mathcal{D})$ are as follows:

- Problems' data are (a) primal-dual pair of regular cones $\mathbf{K}$, $\mathbf{K}_*$ in some $\mathbf{R}^N$, (b) pair of linear subspaces $\mathcal{L}_\mathcal{P}$, $\mathcal{L}_\mathcal{D}$ in $\mathbf{R}^N$ which are orthogonal complements to each other, and (c) pair of vectors $\overline{y}, \overline{\xi}$ in $\mathbf{R}^N$.
- $(\mathcal{P})$ is the problem of minimizing linear objective $\overline{y}^\top \xi$ over the intersection of the *primal feasible plane* $\mathcal{M}_\mathcal{P} := \mathcal{L}_\mathcal{P} - \overline{\xi}$ with the cone $\mathbf{K}$, while $(\mathcal{D})$ is the problem of maximizing the linear objective $\overline{\xi}^\top y$ over the intersection of the *dual feasible plane* $\mathcal{M}_\mathcal{D} := \mathcal{L}_\mathcal{D} + \overline{y}$ with the dual cone $\mathbf{K}_*$.

Pay attention to the "nearly perfect" primal-dual symmetry; the only asymmetry is that in the primal feasible plane the shift vector is $-\overline{\xi}$ – minus the vector of coefficients of the objective in $(\mathcal{D})$, while in the dual feasible plane the shift vector is $\overline{y}$ – the vector of coefficients of the objective in $(\mathcal{P})$. This minor asymmetry stems from the fact that by tradition one of the problems (in our presentation, $(\mathcal{P})$) is written as a minimization program, and the other problem from the pair as a maximization one.

In fact, the symmetry can be made perfect, and the objectives – eliminated at all.

3. Consider pairs of problems $(P)$, $(D)$ along with problems $(\mathcal{P})$, $(\mathcal{D})$, and let $x$, $(y,z)$ be feasible solutions to $(P)$, $(D)$, and $\xi$, $y$ – the feasible solutions to $(\mathcal{P})$, $(\mathcal{D})$ induced by $x$ and $(y,z)$, respectively. Prove that the *duality gap*

$$\mathrm{DualityGap}(x; y, z) := c^\top x - [b^\top y + p^\top z]$$

– the difference between the objective of primal problem $(P)$ evaluated at primal feasible solution $x$ and the objective of the dual problem $(D)$ evaluated at the dual feasible solution $(y, z)$ – is nothing but the inner product $\xi^\top y$ of $\xi$ and $y$.

*Solution:* Here is the computation: Let $x$, $(y, z)$ be feasible for $(P)$, $(D)$, and $\xi = Ax - b$, $y$ be the induced by $x$, $(y, z)$ feasible solutions to $(\mathcal{P})$, $(\mathcal{D})$. Then

$$\begin{array}{rcl}
0 & = & [y - \overline{y}]^\top [\xi + \overline{\xi}] \text{ [since } y - \overline{y} \in \mathcal{L} \text{ and } \xi + \overline{\xi} \in \mathcal{L}^\perp] \\
\Longrightarrow y^\top \xi & = & [\overline{y}^\top \xi - \overline{\xi}^\top y] + \overline{y}^\top \overline{\xi} \\
& = & [c^\top x - [b^\top y + p^\top z]] - b^\top \overline{y} - p^\top \overline{z} + c^\top \overline{x} + \overline{y}^\top \overline{\xi} \text{ [by } (A) \text{ and } (B)] \\
& = & \text{DualityGap}(x; y, z) - b^\top \overline{y} - p^\top \overline{z} + [A^\top \overline{y} + P^\top \overline{z}]^\top \overline{x} + \overline{y}^\top [b - A\overline{x}] \\
& & \qquad\qquad\qquad\qquad \text{[by origin of } \overline{y}, \overline{z}, \overline{\xi}] \\
& = & \text{DualityGap}(x; y, z) - p^\top \overline{z} + [A^\top \overline{y} + P^\top \overline{z}]^\top \overline{x} - \overline{y}^\top A\overline{x} \\
& = & \text{DualityGap}(x; y, z) \text{ [since } P\overline{x} = p]
\end{array}$$

## Around $\mathcal{S}$-Lemma

**Exercise IV.8.** Recall that $\mathcal{S}$-Lemma guarantees that the validity of the implication

$$x^T Ax \geq 0 \implies x^T Bx \geq 0 \qquad [A, B \in \mathbf{S}^n]$$

is the same as the existence of $\lambda \geq 0$ such that $B \succeq \lambda A$ only under the assumption that the inequality $x^T Ax \geq 0$ is strictly feasible. Does the lemma remain true when this assumption is lifted?

*Solution:* The answer is negative. When $n = 2$, $x^T Ax = -x_2^2$ and $x^T Bx = 2x_1x_2$, the above implication holds true, but the quadratic form $x^T(B - \lambda A)x = 2x_1x_2 + \lambda x_2^2$ is not everywhere nonnegative whatever be $\lambda \in \mathbf{R}$.

**Exercise IV.9.** Given $A \in \mathbf{S}^n$, consider the set $Q_A = \{x \in \mathbf{R}^n : x^\top Ax \leq 0\}$.

1. Let $B \in \mathbf{S}^n$ be such that $B \neq A$ and $Q_B = Q_A$. Then, is it always true that there exists $\rho > 0$ such that $B = \rho A$?
2. Suppose that $A \in \mathbf{S}^n$ satisfies $A_{ij} \geq 0$ for all $i, j$. Under this condition, does your answer to item 1 change?
3. Suppose that $A \in \mathbf{S}^n$ satisfies $\lambda_{\min}(A) < 0 < \lambda_{\max}(A)$. Under this condition, does your answer to item 1 change?

*Solution:*

1 : A counter-example is given by $A = -I$ and $B = \text{Diag}\{-1, -2, \ldots, -n\}$ where $Q_A = Q_B = \mathbf{R}^n$.

2 : A counter-example is given by $A = 0$ and $B = -I$, where $Q_A = Q_B = \mathbf{R}^n$.

3 : Suppose $x^\top Ax \leq 0 \iff x^\top Bx \leq 0$. Then $x^\top(-A)x \geq 0 \iff x^\top(-B)x \geq 0$. Since $\lambda_{\min}(A) < 0 < \lambda_{\max}(A)$, $Q_A \neq \mathbf{R}^n$ and $Q_A \neq \{0\}$. Therefore, $\lambda_{\min}(B) < 0 < \lambda_{\max}(B)$ also. Furthermore, the same eigenvalue condition holds for both $-A, -B$, which means $x^\top(-A)x \geq 0$ and $x^\top(-B)x \geq 0$ are both strictly feasible. By the $\mathcal{S}$-lemma, this implies that there exist $\lambda_1, \lambda_2 \geq 0$ such that

$$-B \succeq -\lambda_1 A \implies \lambda_1 A \succeq B$$
$$-A \succeq -\lambda_2 B \implies \lambda_2 B \succeq A.$$

Note that $\lambda_1, \lambda_2 > 0$, otherwise one of $Q_A, Q_B$ will be $\mathbf{R}^n$, which we have already established is not true. Therefore, we can multiply the first inequality by $1/\lambda_1 > 0$ to get

$$A \succeq \frac{1}{\lambda_1} B \implies \lambda_2 B \succeq \frac{1}{\lambda_1} B \implies (\lambda_1 \lambda_2 - 1)B \succeq 0.$$

Since $B$ is not positive semidefinite or negative semidefinite, we must have $\lambda_1\lambda_2 = 1 \implies \lambda_2 = 1/\lambda_1$. But this means

$$\lambda_1 A \succeq B$$
$$\lambda_2 B = \frac{1}{\lambda_1} B \succeq A \implies B \succeq \lambda_1 A$$

which combines with $\lambda_1 A \succeq B$ to imply that $B = \lambda_1 A$ with $\lambda_1 > 0$. □

**Exercise IV.10.** For two nonzero reals $a, b$, one has $2|ab| = \min_{\lambda>0}[\lambda^{-1}a^2 + \lambda b^2]$, implying by the Schur Complement Lemma that $2|ab| \leq c$ iff there exists $\lambda > 0$ such that $\left[\begin{array}{c|c} c - \lambda b^2 & a \\ \hline a & \lambda \end{array}\right] \succeq 0$. Assuming $b \neq 0$, we have also $2|ab| \leq c$ iff there exists $\lambda \geq 0$ such that $\left[\begin{array}{c|c} c - \lambda b^2 & a \\ \hline a & \lambda \end{array}\right] \succeq 0$. Note also that $c \geq 2|ab|$ is the same as $c \geq 2a\delta b$ for all $\delta \in [-1, 1]$.

Prove the following matrix analogy of the above observation:

Let $A \in \mathbf{R}^{p\times r}$, $B \in \mathbf{R}^{p\times s}$, let $B \neq 0$, and let $\mathcal{D} = \{\Delta \in \mathbf{R}^{r\times s} : \|\Delta\| \leq 1\}$, where $\|\cdot\|$ is the spectral norm. Then $C \succeq [A\Delta B^\top + B\Delta^\top A^\top]$ for all $\Delta \in \mathcal{D}$ iff there exists $\lambda \geq 0$ such that $\left[\begin{array}{c|c} C - \lambda BB^\top & A \\ \hline A^\top & \lambda I_r \end{array}\right] \succeq 0$. In particular, when $a, b \in \mathbf{R}^p$ and $b \neq 0$, one has $C \succeq \pm[ab^\top + ba^\top]$ iff there exists $\lambda \geq 0$ such that $\left[\begin{array}{c|c} C - \lambda bb^\top & a \\ \hline a^\top & \lambda \end{array}\right] \succeq 0$.

*Solution:* We have

$$\begin{array}{l} C \succeq [A\Delta B^\top + B\Delta^\top A^\top] \Longleftrightarrow x^\top Cx - 2x^\top A[\Delta B^\top x] \geq 0 \ \forall (x \in \mathbf{R}^p, \Delta \in \mathcal{D}) \\ \Longleftrightarrow x^\top Cx - 2x^\top A\xi \geq 0 \ \forall (x \in \mathbf{R}^p, \xi : \exists \Delta \in \mathcal{D} : \xi = \Delta B^\top x) \\ \Longleftrightarrow x^\top Cx - 2x^\top A\xi \geq 0 \ \forall (x \in \mathbf{R}^p, \xi \in \mathbf{R}^r : \xi^\top\xi \leq x^\top BB^\top x) \\ \Longleftrightarrow \exists \lambda \geq 0 : x^\top Cx - 2x^\top A\xi \geq \lambda[x^\top BB^\top x - \xi^\top \xi] \,\forall (x, \xi) \ [\mathcal{S}\text{-Lemma}] \\ \Longleftrightarrow \exists \lambda \geq 0 : \left[\begin{array}{c|c} C - \lambda BB^\top & A \\ \hline A^\top & \lambda I_r \end{array}\right] \succeq 0. \end{array}$$

Note that the assumption $B \neq 0$ implies that the quadratic form $x^\top BB^\top x - \xi^\top \xi$ of $x, \xi$ is positive at certain point, thus making $\mathcal{S}$-Lemma applicable. □

**Exercise IV.11.** [Robust TTD] Let us come back to TTD problem (5.2). Assume we have solved this problem and have at our disposal the resulting *nominal truss* withstanding best of all, the total truss volume being a given $W > 0$, the load of interest $f$. Now, we cannot ignore the possibility that "in real life" the truss can be affected, aside of the load of interest $f$, by perhaps small, but still nonzero, occasional load composed of forces acting at the free nodes utilized by the nominal truss (think of railroad bridge and wind). In order for our truss to be useful, it should withstand well all small enough occasional loads of this type. Note that our design gives no guarantees of this type – when building the nominal truss, we took into account just one loading scenario $f$.

1. To get impression of potential dangers of "small occasional loads," run numerical study as follows:

- Compute the optimal console $t^*$ (see "Console design" in Exercise I.16)
- Looking one by one at the free nodes $p^1, ..., p^\mu$ actually used by the nominal console, associate with every one of them single-force occasional load, the corresponding force acting at node under consideration, generate this force as random 2D vector of Euclidean length 0.01 (that is, 1% of the magnitude of the single nonzero force in the load of interest), and compute the compliance of the nominal truss w.r.t. to the resulting occasional load. Conclude that the nominal console can be crushed by small occasional load and is therefore completely impractical.

*Solution:* Were the nominal truss be able to withstand occasional loads as well as it withstands the load of interest, we could expect the compliances w.r.t. occasional loads to be of order of $10^{-5}$ (the nominal compliance is $\approx 0.191$, and reducing the load by factor $\alpha$, we reduce the compliance by factor $\alpha^2$). In our experiments, the actual compliance w.r.t. the worst small occasional external force was as large as 0.344; the corresponding equilibrium displacement is shown on Figure VI.6:

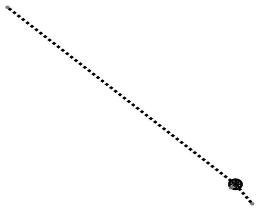


Figure VI.6. Deformation of nominal truss under small occasional load.

The dotted line on Figure VI.6 is the equilibrium displacement under small – just 1% of the force of interest – "badly placed" occasional external force. In the scale of this displacement, we merely do not see the original truss – it is represented by the black area on the figure. Thus, for all practical purposes, *the nominal truss can be completely crushed by a small occasional load and as such is completely impractical.*

2. Proposed cure is, of course, to use Robust Optimization methodology – to immunize the truss against small occasional loads, that is, to control its compliance w.r.t. the load of interest *and* all small occasional loads. An immediate question is where the occasional loads should be applied. There is no sense to allow them to act at all free nodes from the original set of tentative nodes – we have all reasons to believe that some, if not most, of these nodes will not be used in the optimal truss, so that we should not bother about forces acting at these nodes. On the other hand, we should take into account occasional loads acting at the nodes actually used by the optimal robust truss, and we do *not* know in advance what these nodes are. A reasonable compromise here as follows. After the nominal optimal truss is built, we can reduce the nodal set to the nodes actually used in this truss, allow for all pair connection of these nodes and resolve the TTD problem on this reduced sets of tentative nodes and tentative bars, now taking into account not only the load of interest, but all small occasional loads distributed along the nodes of our new nodal set. This approach can be implemented as follows.
   - We specify $\overline{\mathcal{V}}$ as the set of virtual displacements of nodes of our reduced nodal set, preserving the original status ("fixed" – "free") of these nodes, and denote by $\overline{f}$ the natural projection of the load of interest on $\overline{\mathcal{V}}$; note that all nonzero blocks in $f$ – those representing nonzero physical forces from the collection specifying $f$ – are inherited by $\overline{f}$, since the free nodes where these nonzero forces are applied should clearly be used by the nominal truss.
   - We specify $\mathcal{F}$ as the "ellipsoidal envelope" of $\overline{f}$ and all small in magnitude (measured in $\|\cdot\|_2$-norm) loads from $\overline{\mathcal{V}}$. Specifically, we use $\overline{f}$ as one of the half-axes of $\mathcal{F}$; the other $\overline{M}-1$ half-axes of $\mathcal{F}$ ($\overline{M}=\dim\overline{\mathcal{V}}$) are orthogonal to each other and to $\overline{f}$ vectors from $\overline{\mathcal{V}}$ of $\|\cdot\|_2$-norm $\rho\|\overline{f}\|_2$, where the "uncertainty level" $\rho\in[0,1]$ is a parameter of our construction. Note that
   $$\mathcal{F}=\{g=Ph:h^\top h\leq 1\}$$
   for properly selected $\overline{M}\times\overline{M}$ matrix $P$.
   - We define the *robust compliance* $\overline{\mathcal{C}}(\bar{t})$ of a truss $\bar{t}\in\mathbf{R}^{\overline{N}}_+$ ($\overline{N}$ is the number of bars in our new – reduced – set of tentative bars), as the supremum, over $g\in\mathcal{F}$, of the usual compliances (computed for the new nodal set) of $\bar{t}$ w.r.t. load $g$, and pose the Robust Counterpart of the TTD problem as the problem of minimizing this robust compliance over trusses $\bar{t}\geq 0$ of total volume $W$. Solving this problem, we arrive at the *robust truss.*

An immediate question is how to solve the Robust Counterpart. Those who solved Exercise I.16.3 know that as stated right now, the Robust Counterpart is the *semiinfinite* – with infinitely many convex constraints – optimization program

$$\overline{\mathrm{Opt}}=\min_{\bar{t},\tau}\left\{\tau:\bar{t}\in\mathbf{R}^{\overline{N}}_+,\sum_{i=1}^{\overline{N}}\bar{t}_i=W,\left[\begin{array}{c|c}\overline{B}\,\mathrm{Diag}\{\bar{t}\}\overline{B}^\top & g\\ \hline g^\top & 2\tau\end{array}\right]\succeq 0,\forall g\in\mathcal{F}\right\}\qquad(\#)$$

where $\overline{B}$ is the matrix built for the new TTD data in the same fashion as the matrix $B$ was built for the original data.

Here go your tasks:

1. Reformulate ($\#$) as a "normal" convex optimization problem – one with efficiently computable convex objective and finitely many explicitly verifiable convex constraints.
2. Solve the Console design version of the latter problem and subject the resulting robust truss to the same tests as those proposed above for quantifying the "real-life" quality of the nominal truss.

*Solution:* As we know from the solution to Exercise I.16.2-3, a real $\tau$ is an upper bound on the robust compliance of truss $t$ iff

$$\forall g \in \mathcal{F}: \quad 2\tau \geq 2g^\top v - v^\top \overline{A}(\bar{t})v \;\forall v \in \mathbf{R}^{\overline{M}} \qquad [\overline{A}(\bar{t}) = \overline{B}\,\mathrm{Diag}\{\bar{t}\}\overline{B}^\top]$$

Note that when $h$ runs through the $\|\cdot\|_2$-unit ball in $\overline{\mathcal{V}} = \mathbf{R}^{\overline{M}}$, vector $g = -Ph$ runs through the entire ellipsoid $\mathcal{F}$, so that the above relation is equivalent to

$$[u;h]^\top \overline{Q}[u;h] := -2h^\top P^\top u - u^\top \overline{A}(\bar{t})u \leq 2\tau \,\forall([u;h] \in \mathbf{R}^{2\overline{M}} : [u;h]^\top \overline{P}[u;h] := h^\top h \leq 1)$$

Applying Inhomogeneous $\mathcal{S}$-Lemma (Lemma IV.23.9), the latter relation takes place iff

$$\exists \lambda \geq 0 : \left[\begin{array}{c|c|c} \overline{A}(\bar{t}) & P & \\ \hline P^\top & \lambda I_{\overline{M}} & \\ \hline & & 2\tau - \lambda \end{array}\right] \succeq 0,$$

or, which is clearly the same, iff

$$\left[\begin{array}{c|c} \overline{A}(\bar{t}) & P \\ \hline P^\top & 2\tau I_{\overline{M}} \end{array}\right] \succeq 0.$$

The bottom line is that problem ($\#$) is equivalent to the "normal" convex optimization problem

$$\overline{\mathrm{Opt}} = \min_{\bar{t},\tau} \left\{ \tau : \bar{t} \geq 0, \sum_i \bar{t}_i = W, \left[\begin{array}{c|c} \overline{B}\,\mathrm{Diag}\{\bar{t}\}\overline{B}^t & P \\ \hline P^\top & 2\tau I_{\overline{M}} \end{array}\right] \succeq 0 \right\}.$$

We solved the latter problem with $\rho = 0.1$ and tested the resulting robust truss against the load of interest $\overline{f}$ and 100 randomly selected occasional loads of magnitude 1% of the nominal load. The results are presented at Figure VI.7. Pay attention to the low cost of robustness: optimal *robust* compliance corresponding to the rather high (10%) uncertainty level is just by 10% larger than the optimal nominal compliance; compliance of the robust truss w.r.t. the load of interest is just by 0.6% larger than the compliance of the nominal truss.

## Miscellaneous exercises

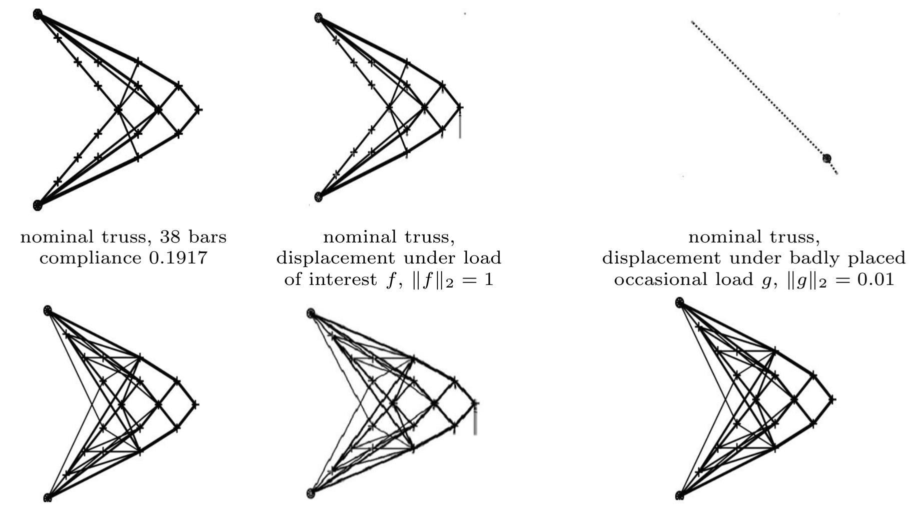
**Exercise IV.12.** Find the minimizer of a linear function

$$f(x) = c^\top x$$

on the set

$$V_p = \{x \in \mathbf{R}^n \mid \sum_{i=1}^n |x_i|^p \leq 1\};$$

here $p$, $1 < p < \infty$, is a parameter. What happens with the solution when the parameter becomes $0.5$?



nominal truss, 38 bars
compliance 0.1917

nominal truss,
displacement under load
of interest $f$, $\|f\|_2 = 1$

nominal truss,
displacement under badly placed
occasional load $g$, $\|g\|_2 = 0.01$

robust truss, 152 bars
robust compliance 0.1992

robust truss,
displacement under load
of interest $f$, $\|f\|_2 = 1$

robust truss,
displacement under badly placed
occasional load $g$, $\|g\|_2 = 0.01$

Figure VI.7. Nominal and robust consoles. Positions of nodes before (crosses) and after (dots) deformation. Vertical segment starting at the most-right node: external force

*Solution:* Let us find a KKT point of the problem where the constraint is active. The KKT condition reads

$$\begin{array}{l} c_i + \lambda p|x_i|^{p-1}\text{sign}(x_i) = 0,\ i = 1, ..., n \\ \sum\limits_i |x_i|^p = 1 \end{array}$$

whence

$$x_i = -\frac{|c_i|^{q-1}\text{sign}(c_i)}{\|c\|_q^{q-1}},\ q = \frac{p}{p-1}$$

(we have assumed that $c \neq 0$, otherwise every feasible point is optimal). When $1 < p < \infty$, the problem is convex, so that the KKT point we have found is global optimal solution to the problem.

When $p = 0.5$, the solution is as follows. W.l.o.g. assume that $c_i \leq 0$; then at the optimum one clearly has $x_i \geq 0$; we lose nothing by adding these inequalities to the list of constraints. Assuming $x \geq 0$ and passing to new variables $y_i = \sqrt{x_i}$, our problem becomes

$$\min \sum_i c_i y_i^2 \text{ s.t. } y \geq 0, \sum_i y_i \leq 1.$$

Since $c_i \leq 0$, this is the problem of minimizing a *concave* function over the standard simplex; the solution is at the vertex of the simplex, and clearly this vertex is $y_{i_*} = 1$, $y_i = 0$, $i \neq i_*$, where $i_*$ is the index of the most negative $c_i$. Thus, an optimal solution is the basic orth corresponding to the most negative $c_i$. In general (that is, without the assumption $c_i \leq 0$), the solution is $\epsilon e_{i_*}$, where $i_*$ is the index of the maximal, in absolute value, coordinate of $c$, $\epsilon = \pm 1$ is the minus sign of this coordinate, and $e_i$ are standard basic orths. The optimal value is $-\|c\|_\infty$.

**Exercise IV.13** Every one of 3 random variables $\xi_1$, $\xi_2$, $\xi_3$ takes values 0 and 1 with probabilities 0.5, and every two of these 3 variables are independent. Is it true that all 3 variables are mutually independent? If not, how large could be probability of the event $\xi_1 = \xi_2 = \xi_3 = 1$?

*Solution:* We know that 8-dimensional probabilistic vector $p$ representing the probability distribution of the random 3D Boolean vector $\xi = [\xi_1; \xi_2; \xi_3]$ satisfies a bunch of linear equalities and inequalities (specifically, is nonnegative and induces uniform on $\{0, 1\}^2$ distributions of pairs of entries; we could add also "induces uniform on $\{0,1\}$ marginal distributions of entries," but this requirement is covered by the one on marginal distributions of pairs of entries), and ask what is under the circumstances the maximin allowed value of a particular entry. This is a simple LP program, and its optimal value turns out to be $1/4$ – twice the value of this entry in the distribution corresponding to the case of $\xi_i$ independent across $i = 1, 2, 3$.

This simple example illustrates potential difficulties in recovering multivariate distributions from samples – with no a priori information on a probability distribution on, say, $\{0,1\}^d$ - a priori, it could be a whatever probabilistic vector $p$ of dimension $2^d$ – statistically reliable recovery of $p$ by sampling the corresponding random vector would require exponential in $d$, and thus unrealistic already for moderate $d$'s, sample sizes. An alternative could be to try to "reconstruct" $p$ from something we can estimate by sampling reliably, e.g., from low-dimensional marginal distributions induced by $p$. Our example is the simplest illustration of the difficulties which cold be met along this road.

**Exercise IV.14.** [computational study] Consider situation as follows: at discrete time instants $t = 1, 2, ..., T$ we observe the states $y_t \in \mathbf{R}^\nu$ of dynamical system; our observations are

$$y_t + \sigma\xi_t, t = 1, 2, ..., T,$$

where $\sigma > 0$ is a given noise intensity and $\xi_t$ are independent across $t$ zero mean Gaussian noises with unit covariance matrix. All we know about the trajectory of the system is that

$$\|y_{t+1} - 2y_t + y_{t-1}\|_2 \leq dt^2\alpha, \tag{!}$$

where $dt > 0$ is the continuous time interval between consecutive discrete time instants; in other words, the Euclidean norm of the (finite-difference approximation of the) acceleration of the system is $\leq \alpha$. Given time delay $d$, we want to estimate the linear form $f^\top y_{T+d}$ of the system's state at time $T + d \geq 1$, and we intend to use a linear estimate

$$\widehat{y} = \sum_{t=1}^T h_t^\top \omega_t.$$

1. Write down optimization problem specifying the minimum risk linear estimate, with the risk of an estimate defined as

   $$\mathrm{Risk}[\widehat{y}] = \sqrt{\sup_{y\in\mathcal{Y}} \mathbf{E}\{|\widehat{y} - f^\top y_{T+d}|^2\}},$$

   where $\mathcal{Y}$ is the set of all trajectories $y = \{y_t, -\infty < t < \infty\}$ satisfying all constraints (!).

   *Solution:* Let $E_t y = y_{t+1} - 2y_t + y_{t-1}$, so that (!) reads $\|E_t y\|_2 \leq \beta = dt^2\alpha$, $t = 0, \pm1, \pm2, ....$ For a linear estimate, we clearly have

   $$\mathbf{E}\{|\widehat{y} - y_{T+d}|_2^2\} = |\sum_{t=1}^T h_t^\top y_t - f^\top y_{T+d}|^2 + \sigma^2 \sum_{t=1}^T \|h_t\|_2^2$$

   Consequently,

   $$\begin{array}{rcl}\mathrm{Risk}^2[\widehat{y}] & = & \sigma^2 \sum_{t=1}^T \|h_t\|_2^2 + \Phi^2(h),\\ \Phi(h) & := & \max\limits_{y:\|E_t y\|_2\leq\beta\,\forall t} |\sum_{t=1}^T h_t^\top y_t - f^\top y_{T+d}|\\ & = & \max\limits_{y:\|E_t y\|_2\leq\beta\,\forall t} [\sum_{t=1}^T h_t^\top y_t - f^\top y_{T+d}]\end{array}$$

   where the concluding equality follows from the fact that trajectories $y$ and $-y$ simultaneously satisfy/do not satisfy the acceleration bound. Next, when computing $\sup_y$, we clearly can restrict ourselves with $\sup_{y:\|E_t y\|_2\leq\beta, 1\leq t\leq\overline{T}}$, where $\overline{T} = \max[T, T+d]$, and in this case we lose nothing when thinking of $y$ as of finite sequence: $y = [y_0; y_1; ...; y_{\overline{T}+1}]$. Thus, the optimization problem responsible for the minimum risk linear estimate is

   $$\mathrm{Opt} = \min_{h_1,...,h_T} \left[F(h) = \sigma^2 \sum_{t=1}^T \|h_t\|_2^2 + \left[\max_{\substack{y=[y_0;y_1;...;y_{\overline{T}+1}]:\\ \|E_t y\|_2\leq\beta, 1\leq t\leq\overline{T}}} [\sum_{t=1}^T h_t^\top y_t - f^\top y_{T+d}]\right]^2\right], \tag{$*$}$$

This is a convex optimization problem with albeit implicitly given, but efficiently computable objective. The optimal value Opt of the problem is the squared risk of the minimum risk linear estimate.

2. Use Conic Duality to convert the problem from the previous item into a Conic Quadratic problem.

*Solution:* By Conic Duality,

$$\begin{array}{rcl}\Phi(h) & := & \max\limits_{y=[y_0;y_1;...;y_{\overline{T}+1}]} \left\{G^\top[h]y := \sum_{t=1}^T h_t^\top y_t - f^\top y_{T+d} : \|E_t y\|_2 \leq \beta, 1 \leq t \leq \overline{T}\right\} \\ & = & \min\limits_{\lambda_1,...,\lambda_{\overline{T}}\in\mathbf{R}^\kappa} \left\{\beta \sum_{t=1}^{\overline{T}} \|\lambda_t\|_2 : \sum_{t=1}^{\overline{T}} E_t^\top \lambda_t = G[h],\right\}\end{array}$$

making $(*)$ the Conic Quadratic problem

$$\sqrt{\mathrm{Opt}} = \min_{\lambda_1,...,\lambda_{\overline{T}},h_1,...,h_T,r,s,\gamma} \left\{\gamma : \begin{array}{rcl} \|[r;s]\|_2 & \leq & \gamma \\ \|h_t\|_2 & \leq & r_t/\sigma, 1 \leq t \leq T \\ \|\lambda_t\|_2 & \leq & s_t/\beta, 1 \leq t \leq \overline{T} \\ \sum_{t=1}^{\overline{T}} E_t^\top \lambda_t & = & [0_{\nu\times 1}; h_1; ...; h_T; 0_{\nu(\overline{T}-T)\times 1}] \\ & & -[0_{\nu(T+d)\times 1}; f; 0_{\nu(\overline{T}-T-d)\times 1}] \end{array}\right\}$$

3. Carry out numerical experimentation with minimum risk linear estimate.

*Solution:* In our experiments, we used $\nu = 3$, $T = 100$, $dt = 0.25$, $\alpha = 1$, $\sigma = 0.1$, $d = 10$. We built linear estimates for every one of the 3 coordinates in $y_{T+d}$. The risks of these estimates and their empirical, over 300 simulations, risks are as follows:

| | | |
|---|---|---|
| 3.73 | 3.73 | 3.73 |
| 1.80 | 1.75 | 1.76 |

Top row: Theoretical risk bounds; bottom row: empirical risks

The plot of a sample simulation is presented on Figure VI.8.

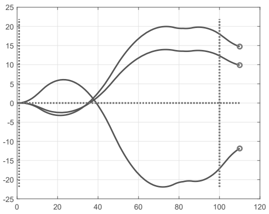


Figure VI.8. Coordinates of system's state vs. time. Circles: forecasts; dotted segment between dotted vertical lines: instants where observations are taken.

**Exercise IV.15.** [computational study] Consider the following problem:

> A particle is moving through $\mathbf{R}^d$. Given positions and velocities of the particle at times $t = 0$ and $t = 1$, find the trajectory of the particle on $[0,1]$ with minimum possible (upper bound) on acceleration.

1. Formulate the (discretized in time version of the) problem as a Conic Quadratic problem and write down its conic dual. Are the problems solvable? Are the optimal values equal to each other? What is said by optimality conditions?
2. Run numerical experiments in 2D and 3D and look at the results.

*Solution:* 1: Let us discretize time interval $[0,1]$, splitting it into $n+1$ consecutive segments of length $dt = 1/(n+1)$ each, and set $t_i = i/(n+1)$. We discretize a candidate trajectory by looking at the sequence $u = \{u_1, ..., u_n\}$ of positions of the particle at time instants $t_i$, $i = 1, ..., n$, and augment this sequence with two initial terms, $u_{-1}, u_0$, and two concluding terms $u_{n+1}, u_{n+2}$ to model the boundary conditions; specifically, we set $u_0$ to be the given position of the particle at time 0, and set $u_{-1} = u_0 - dtv^0$, where $v^0$ is the velocity of the particle at time 0. Similarly, we set $u_{n+1}$ to be the given position of particle at time 1, and set $u_{n+2} = u_{n+1} + dtv^1$, where $v^1$ is the velocity of the particle at time 1. Finally, we approximate the acceleration of the particle at time $t_i$ by the finite difference $[u_i - 2u_{i-1} + u_{i-2}]/dt^2$. As a result, the discretized model of our problem becomes

$$\mathrm{Opt}(P) = \min_{\tau,u} \left\{\tau : \|u_i - 2u_{i-1} + u_{i-2}\|_2 \leq dt^2\tau, 1 \leq i \leq n+2\right\} \qquad (P)$$

Note that in this problem, the variables are $u_1, ..., u_n$, while $u_{-1}, u_0, u_{n+1}, u_{n+2}$ are data. $(P)$ is a Conic Quadratic problem; its "canonical" form is

$$\min_{\tau,u} \left\{\tau : [2u_i - u_{i-1} + u_{i+2}; dt^2\tau] \in \mathbf{L}^{d+1},\ 1 \leq i \leq n+2\right\}$$

Equipping the conic constraints with Lagrange multipliers $[y_i; s_i] \in \mathbf{L}^{d+1}$, $1 \leq i \leq n+2$, the conic dual of $(P)$ is built as follows: we aggregate the constraints with the "weights" $[y_i; s_i]$, thus arriving at the relation

$$\sum_{i=1}^{n} [dt^2\tau s_i + y_i^\top [u_i - 2u_{i-1} + u_{i=2}]] \geq 0$$

which, due to its origin, is a consequence of the constraints of $(P)$, rewrite this relation equivalently as

$$[\text{homogeneous linear function of } \tau, u_1, ..., u_n] \geq [\text{linear function of } y_i, s_i,\ 1 \leq i \leq n+2] \qquad (*)$$

and impose on the Lagrange multipliers, in addition to the constrains $[y_i; s_i] \in \mathbf{L}^{d+1}$, the restriction that the left hand side linear function in $(*)$ is identically in $\tau, u_1, ..., u_n$ equal to the objective of $(P)$. The dual problem is to maximize under these restrictions the right hand side of $(*)$.
Executing this strategy (which is a fully mechanical process) results in the dual problem

$$\begin{array}{rcl}\mathrm{Opt}(D) & = & \max_{y_i,s_i} \left\{ y_1^\top[2u_0 - u_{-1}] - y_2^\top u_0 - y_{n+1}^\top u_{n+1} + y_{n+2}^\top[2u_{n+1} - u_{n+2}] : \right. \\ & & \left. \left\{\begin{array}{lll} y_{i+2} - 2y_{i+1} - y_i = 0 & , i = 1, ..., n & (a) \\ \sum_{i=1}^{n+2} s_i = 1/dt^2 & , & (b) \\ [y_i; s_i] \in \mathbf{L}^{d+1} & , 1 \leq i \leq n+2 & (c) \end{array}\right. \right\} \end{array} \qquad (D)$$

Clearly, $(P)$ is strictly feasible and bounded, implying that $(D)$ is solvable and that the optimal values are equal to each other. Besides this, $(D)$ clearly is essentially strictly feasible, so that $(P)$ is solvable as well. Optimality conditions in their complementary slackness form say that a pair $(u_i^*, i \leq n, \tau^*; y_i^*, s_i^*, i \leq n+2)$ of primal-dual feasible solutions is composed of optimal solutions iff

$$[u_i^* - 2u_{i-1}^* + u_{i-2}^*; dt^2\tau^*]^\top [y_i^*; s_i^*] = 0,\ 1 \leq i \leq n+2. \qquad (**)$$

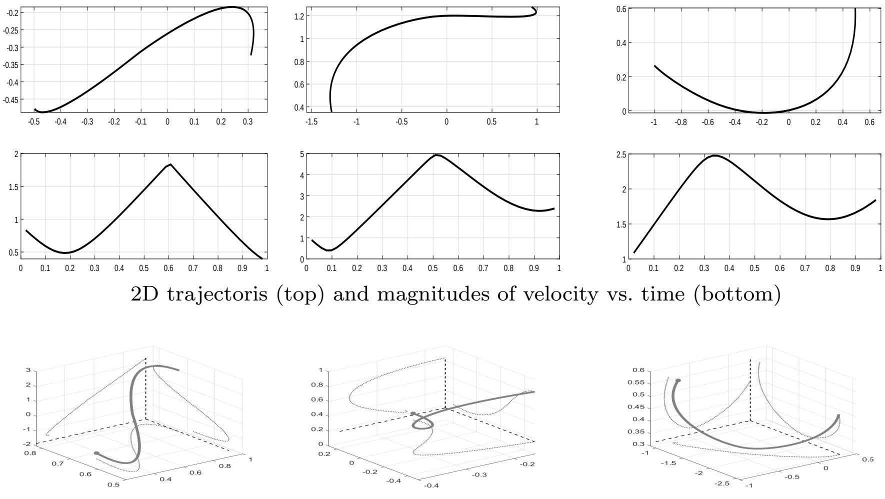
Observe that the equality constraints $(a)$ in $(D)$ say that entries in $y_i$ are linear functions of $i$: $y_i = g + (i-1)h$ for some $g, h$. As a result, $(D)$ simplifies to

$$\begin{array}{l}\mathrm{Opt} = \min_{g,h,s_i} \left\{ \quad [-u_{-1} + u_0 + u_{n+1} - u_{n+2}]^\top g + [-u_0 + (n+2)u_{n+1} - (n+1)u_{n+2}]^\top h : \right. \\ \qquad\qquad \left. \|g + (i-1)h\|_2 \leq s_i, 1 \leq i \leq n+2, \sum_i s_i = 1/dt^2 \right\}, \end{array} \qquad (D')$$

and Opt $=$ Opt$(P) =$ Opt$(D)$. This combines with complementary slackness $(**)$ to conclude that in the only nontrivial case Opt $> 0$ an optimal solution to $(P)$ is readily given by an optimal solution $g^*, h^*, s_i^*, i \leq n+2$ to $(D')$ via the relation

$$u_i^* - 2u_{i-1}^* + u_{i-2}^* = -dt^2 \mathrm{Opt} \frac{g^* + (i-1)h^*}{\|g^* + (i-1)h^*\|_2} \qquad (!)$$

which holds true for all $i$, $1 \leq i \leq n+2$, such that $g^* + (i-1)h^* \neq 0$; in this relation, by definition $u_{-1}^* = u_{-1}$, $u_0^* = u_0$, $u_{n+1}^* = u_{n+1}$, $u_{n+2}^* = u_{n+2}$. Note that if there is no $i \leq n+2$ such that $g^* + (i-1)h^* = 0$, recurrence (!) fully determines $u_i^*$, $1 \leq i \leq n$. If $g^* + (i-1)h^* = 0$ (if such an $i = i_*$ exists, it is unique, since otherwise we would have $g^* = h^* = 0$ and therefore Opt $= 0$, which is not the case), we could specify $u_i^*$ for $1 \leq i < i_*$ via recurrence (!), and for $i = i_*, i_* + 1, ..., n$ - running the same recurrence backward, starting with $i = n+2$.



2D trajectoris (top) and magnitudes of velocity vs. time (bottom)

3D trajetories

Figure VI.9. Sample trajectories (bold) in Exercise IV.15 and their 2D projections (dotted).

2: In or computations, we used $n = 50$. Sample trajectories in 2D and 3D are shown at Figure VI.9.

**Exercise IV.16.** [computational study] The study offered to you in this Exercise is as follows:

> *A steel rod is heated at time $t = 0$, the magnitude of the temperature being $\leq R$, and is left to cool, the temperature at the endpoints being all the time kept 0. We measure the temperature of the rod at locations $s_i$ and times $t_i > 0$, $1 \leq i \leq m$; the measurements are affected by Gaussian noise with zero mean and covariance matrix $\sigma^2 I_m$. Given the measurements, we want to recover the distribution of temperature of the rod at time $\bar{t} > 0$.*

**Building the model.** With properly selected units of temperature and length (so that the rod becomes the segment $[0, 1]$), evolution of the temperature $u(t, s)$ ($t \geq 0$ is time, $s \in [0, 1]$ is location) is governed by the *Heat equation*

$$\frac{\partial}{\partial t} u(t, s) = \frac{\partial^2}{\partial s^2} u(t, s) \qquad [u(t, 0) = u(t, 1) \equiv 0]$$

It is convenient to represent functions on $[0, 1]$ as

$$f(s) = \sum_{k=1}^{\infty} f_k \phi_k(s),\ \phi_k(s) = \sqrt{2}\sin(\pi k s).$$

Functions $\phi_k$ form an orthonormal basis in the space $L_2 = L_2[0, 1]$ of square summable real-valued functions on $[0, 1]$ equipped with the inner product

$$\langle f, g \rangle = \int_0^1 f(s) g(s) ds,$$

the corresponding norm being $\|f\|_2 = \sqrt{\int_0^1 f^2(s) ds}$.

Functions $\phi_k$ form an orthonormal basis in $L_2$, meaning that for every $f \in L_2$ the series

$$\sum_{k=1}^{\infty} f_k \phi_k(s),\ f_k = \langle f, \phi_k \rangle$$

converges in $\|\cdot\|_2$ to $f$, $f \in L_2$ iff $\sum_k f_k^2 < \infty$, and

$$\langle \sum_k f_k\phi_k(\cdot), \sum_k g_k\phi_k(\cdot)\rangle = \sum_k f_k g_k \,\forall f, g \in L_2.$$

In particular,

$$u(t,s) = \sum_{k=1}^{\infty} u_k(t)\phi_k(s),\; u_k(t) = \int_0^1 u(t,s)\phi_k(s)ds.$$

Assuming $|u(0,\cdot)| \leq R$, we have

$$\sum_k u_k^2(0) \leq R^2, \tag{29.1}$$

and in terms of the coefficients $u_k(t)$ of the rod's temperature, the Heat equation becomes very simple:

$$\frac{d}{dt}u_k(t) = -\pi^2k^2u_k(t) \implies u_k(t) = \exp\{-\pi^2k^2t\}u_k(0).$$

As a result, when $t > 0$, the coefficients $u_k(t)$ go to 0 exponentially fast as $k \to \infty$, so that the series

$$\sum_k u_k(t)\phi_k(s)$$

converges to the solution $(t,s)$ of the heat equation not only in $\|\cdot\|_2$, but uniformly on $[0,1]$ as well, implying, due to $\phi_k(0) = \phi_k(1) = 0$, that the series does satisfy the boundary conditions $u(t,0) = u(t,1) = 0$, $t > 0$.

Now our problem can be posed as follows:

*The sequence of coefficients $\{u_k^t\}_{k=1}^{\infty}$ of $u(t,\cdot)$ in the orthonormal basis $\{\phi_k(\cdot)\}_{k\geq 1}$ of $L_2$ evolves according to*

$$u_k^t = \exp\{-\pi^2k^2t\}u_k^0,$$

*with*

$$u^0 := \{u_k^0\}_{k\geq 1} \in \mathbf{B} := \{\{c_k\}_{k\geq 1} : \sum_k c_k^2 \leq R^2\}.$$

*Given $m$ noisy observations*

$$\omega_i = \Omega_i[u^0] + \sigma\xi_i,\; \Omega_i[u^0] = \sum_{k=1}^{\infty} \exp\{-\pi^2k^2t_i\}u_k^0\phi_k(s_i),$$

*where $\xi_1, ..., \xi_m$ are independent of each other $\mathcal{N}(0,1)$ observation noises, and $t_i > 0$, $s_i \in [0,1]$ are given, we want to recover the sequence $\{u_k^{\bar{t}}\}_{k\geq 1}$.*
*We quantify the performance of a candidate estimate $\omega := (\omega_1, ..., \omega_m) :\mapsto \widehat{u} = \{\widehat{u}_k(\omega)\}_{k\geq 1}$ by the risk*

$$\mathrm{Risk}[\widehat{u}] = \sqrt{\max_{u^0\in\mathbf{B}} \mathbf{E}_\xi \left\{ \sum_{k\geq 1} [\widehat{u}_k(\Omega_1[u^0] + \xi_1, ..., \Omega_m[u^0] + \xi_m) - \exp\{-\pi^2k^2\bar{t}\}u_k^0]^2 \right\}}$$

*that is,* $\mathrm{Risk}^2$ *is the worst, w.r.t. the distribution of temperature at time $t = 0$ of $\|\cdot\|_2$-norm not exceeding $R$, expected squared norm $\|\cdot\|_2^2$ of the recovery error.*

Our last modeling step is to replace infinite sequences $\{u_k^0\}_{k\geq 1}$ with their finite initial segments $\{u_k^0\}_{1\leq k\leq K}$, that is, to approximate the situation by the one where $u_0^k = 0$ when $k > K$. The simplest way to do it is as follows. Let $\underline{t} = \min[\min_i t_i, \bar{t}]]$, so that $\underline{t} > 0$. For $u_0 \in \mathbf{B}$ and $K \geq 1$, the magnitude of the total contribution of the coefficients $u_0^k$, $k > K$, to $u(t,s)$ with $t \geq \underline{t}$ does not exceed

$$\sum_{k=K+1}^{\infty} \max_s |\phi_k(s)| \exp\{-\pi^2 k^2 \underline{t}\} |u_0^k| \leq \delta := \sqrt{2}R \sum_{k=K+1}^{\infty} \exp\{-\pi^2 k^2 \underline{t}\}.$$

Given a "really small" tolerance $\bar{\delta} > 0$, say, $\bar{\delta} = 10^{-10}$, we can easily find $K = K(\bar{\delta})$ such that $\delta \leq \bar{\delta}$. Thus, as far as the temperatures we measure and the temperatures we want to recover are concerned, zeroing out coefficients $u_0^k$ with $k > K(\bar{\delta})$ changes these temperatures by at most $\bar{\delta}$. Common sense (which can be easily justified by formal analysis) says, that with $\bar{\delta}$ as small as $10^{-10}$, these changes have no effect on the quality of our recovery, at least when $\sigma \gg \bar{\delta}$.

Now goes your task:

1. Assuming $u_0^k = 0$ for $k > K$, model the problem of interest as the following estimation problem:

   > *"In the nature" there exists $K$-dimensional signal $u$ known to belong to the centered at the origin Euclidean ball $B^R = \{u \in \mathbf{R}^K : u^\top u \leq R^2\}$ of a given radius $R$. Given noisy observations*
   >
   > $$\omega = Au + \sigma\xi, \qquad\qquad [A : m \times K, \xi \sim \mathcal{N}(0, I_m)]$$
   >
   > *we want to recover $Bu$, quantifying the recovery error of a candidate estimate $\omega \mapsto \widehat{u}(\omega)$ by its risk*
   >
   > $$\mathrm{Risk2}[\widehat{u}] = \sqrt{\sup_{u \in B^R} \mathbf{E}_{\xi \sim \mathcal{N}(0,I_m)} \{[\widehat{u}(Au + \sigma\xi) - Bu]^\top [\widehat{u}(Au + \sigma\xi) - Bu]\}}$$
   >
   > *where $B$ is a given $K \times K$ matrix.*

   Write down the expressions for the matrices $A$ and $B$.
2. Build convex optimization problem responsible for the minimum risk *linear estimate* – estimate of the form $\widehat{u}(\omega) = H^\top \omega$.
3. Compute the minimum risk linear estimate and run simulations to test its performance. Recommended setup:
   - $\bar{t} \in \{0.01, 0.001, 0.0001, 0.00001\}$
   - $m = 100$, $t_i$ are drawn at random from the uniform distribution on $[\bar{t}, 2\bar{t}]$, $s_i$ are drawn at random from the uniform distribution on $[0, 1]$;
   - $R = 10^4$, $\sigma = 10$, $\bar{\delta} = 10^{-10}$;
   - To accelerate computations, truncate $K(\bar{\delta})$ at the level 100.

*Solution:* 1: $A_{ik} = \sqrt{2}\exp\{-\pi^2 k^2 t_i\} \sin(\pi k s_i)$, $1 \leq i \leq m, 1 \leq k \leq K$. $B$ is diagonal $K \times K$ matrix with diagonal entries $\exp\{-\pi^2 k^2 \bar{t}\}$, $1 \leq k \leq K$.

2: The problem is

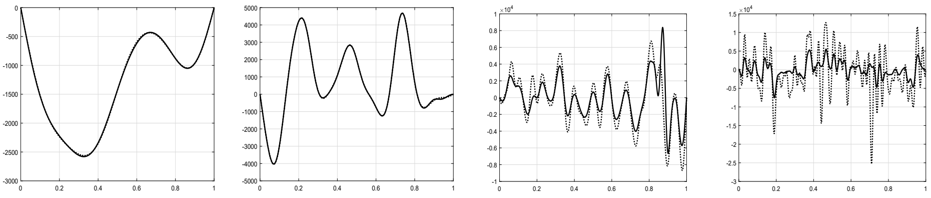
$$\mathrm{Opt} = \min_{H \in \mathbf{R}^{m \times K}} \sqrt{R^2 \|B - H^\top A\|_{2,2}^2 + \sigma^2 \mathrm{Tr}(H^\top H)},$$

where $\|\cdot\|_{2,2}$ is the spectral norm (the largest singular value) of a matrix.

3: Our results are as follows:

| | | | empirical errors | | |
|---|---|---|---|---|---|
| $\bar{t}$ | $K$ | Risk2 | mean | median | max |
| 0.01 | 18 | 6.47 | 6.35 | 11.24 | 7.17 |
| 0.001 | 58 | 15.60 | 14.06 | 13.84 | 22.71 |
| 0.0001 | 100 | 1186.8 | 1154.3 | 1638.9 | 2192.3 |
| 0.00001 | 100 | 4332.2 | 4332.2 | 4762.9 | 9294.7 |

Risks and empirical recovery errors,
data over 100 simulations



$\bar{t} = 0.01$ $\bar{t} = 0.001$ $\bar{t} = 0.0001$ $\bar{t} = 0.00001$

Figure VI.10. Sample recoveries: dotted line - true; solid line – recovery

To put the results in proper perspective, pay attention to the range of true signals on Figure VI.10.

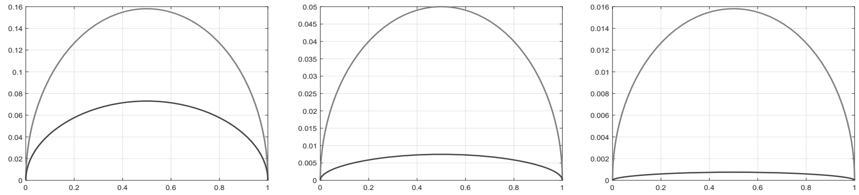
**Exercise IV.17.** Given positive definite $A \in \mathbf{S}^n$, let us set

$$P[A] = \{X \in \mathbf{S}^n : X \succeq 0, X^2 \preceq A\},\ Q[A] = \{X \in \mathbf{S}^n : X \succeq 0, X \preceq A^{1/2}\}.$$

From $\succeq$-monotonicity of the matrix square root on $\mathbf{S}^n_+$ (Example IV.26.5 in section 26.2) it follows that $P[A] \subseteq Q[A]$. Your task is to answer the following question:

> Are $P[A]$ and $Q[A]$ "comparable," meaning that for some $c$ independent of $A$ (but perhaps depending on $n$) one has
> $$Q[A] \subset c \cdot P[A] \qquad ?$$

*Solution:* The answer is negative, unless $n = 1$. To justify the claim, it suffices to consider the case of $n = 2$. Given $\epsilon \in (0,1)$, let us set $A = \left[\begin{array}{c|c} \epsilon & \\ \hline & 1 \end{array}\right]$, so that $A^{1/2} = \left[\begin{array}{c|c} \epsilon^{1/2} & \\ \hline & 1 \end{array}\right]$, implying that the matrix $\overline{X} = \frac{1}{2}\left[\begin{array}{c|c} \epsilon^{1/2} & \epsilon^{1/4} \\ \hline \epsilon^{1/4} & 1 \end{array}\right]$ belongs to $Q[A]$. On the other hand, for $X = [x_{ij}]_{i,j=1,2} \in P[A]$ we should have $[X^2]_{1,1} = X_{1,1}^2 + X_{1,2}^2 \leq A_{1,1} = \epsilon$, that is, $X_{1,2} \leq \sqrt{\epsilon}$. By looking at off-diagonal entries, we conclude that if $Q[A] \subset c \cdot P[A]$, so that $\overline{X} = cX$ for some $X \in P[A]$, we should have $c \geq \frac{1}{2}\epsilon^{-1/4}$, and the right hand side here tends to $+\infty$ when $\epsilon \to +0$. $\square$



$\epsilon = 10^{-2}$ $\epsilon = 10^{-4}$ $\epsilon = 10^{-6}$

Figure VI.11. Results for Exercise IV.17

For a given $\epsilon$, setting $A = \left[\begin{array}{c|c} \epsilon & \\ \hline & 1 \end{array}\right]$, the matrices from $P[A]$ are of the form $\left[\begin{array}{c|c} t\sqrt{\epsilon} & \delta \\ \hline \delta & s \end{array}\right]$ with $0 \leq t, s \leq 1$ and $-\gamma_\epsilon(t,s) \leq \delta \leq \gamma_\epsilon(t,s)$, and the matrices from $Q[A]$ are of the form $\left[\begin{array}{c|c} t\sqrt{\epsilon} & \delta \\ \hline \delta & s \end{array}\right]$ with $0 \leq t, s \leq 1$ and $-\theta_\epsilon(t,s) \leq \delta \leq \theta_\epsilon(t,s)$. What you see on Figure VI.11 are the plots of the functions $\gamma_\epsilon(t, 1-t)$ (lower curves) and $\theta_\epsilon(t, 1-t)$ (upper curves) vs. $t \in [0,1]$.

**Exercise IV.18.** Find the optimal value in the convex optimization problem

$$\mathrm{Opt}(a) = \min_x \left\{\sum\nolimits_{i=1}^n [-(1+a_i)x_i + x_i \ln x_i] : x \geq 0, \sum\nolimits_i x_i \leq 1\right\}$$

where $0 \ln 0 = 0$ by definition, so that the function $x \ln x$ is well defined and continuous on the nonnegative ray $x \geq 0$.

*Solution:* The problem is of the form $\min_{x\in X} f(x)$ with $X = \{x \in \mathbf{R}^n : x \geq 0, \sum_i x_i \leq 1\}$ and

$$f(x) = \sum\nolimits_i [-(1+a_i)x_i + x_i \ln x_i];$$

$f$ is convex on $X$ and is differentiable on the part $x > 0$ of $X$. Let us make an educated guess that there is an optimal solution $x$ to the problem with $x > 0$. The necessary and sufficient condition for such a solution $x$ to be optimal is that

1. either $x \in \operatorname{int} X$ and $\nabla f(x) = 0$, resulting in

   $$x_i = \exp\{a_i\}, 1 \leq i \leq n;$$

   this solution indeed belongs to the interior of $X$ provided that

   $$\sum\nolimits_{i=1}^n \exp\{a_i\} < 1 \tag{!}$$

2. or $x > 0$ belongs to the face $\sum_i x_i = 1$ of $X$, in which case $x$ is optimal iff $\nabla f(x)$ has nonnegative inner products with all directions leading from $x$ to points of $X$, or, which is the same, $\nabla f(x)$ is a nonpositive multiple of the all-ones vector. Thus, what we want of $x$ now is to be positive, to have $\sum_i x = 1$, and to have

   $$\nabla f(x) = -\lambda[1; \ldots; 1]$$

   with some $\lambda \geq 0$. Looking at what $\nabla f(x)$ is, this boils down to

   $$x_i = \exp\{a_i - \lambda\}$$

   and $\sum_i x_i = 1$, implying

   $$\lambda = \ln(\sum\nolimits_j \exp\{a_j\}).$$

   Thus, $\lambda \geq 0$ whenever (!) fails to be true, and in this case

   $$x_i = \frac{\exp\{a_i\}}{\sum_j \exp\{a_j\}}$$

The bottom line is that an optimal solution is given by

$$x_i = \frac{\exp\{a_i\}}{\max[1, \sum_j \exp\{a_j\}]}, \ 1 \leq i \leq n,$$

and the optimal value is

$$\begin{cases} -\sum_i \exp\{a_i\}, & \sum_i \exp\{a_i\} \leq 1, \\ -\ln\left(\sum_i \exp\{a_i\}\right) - 1, & \text{otherwise} \end{cases}$$

**Exercise IV.19.** Given $m \times n$ matrix $A$ with trivial kernel, consider the matrix-valued function $F(X) = [A^\top X^{-1} A]^{-1} : \operatorname{Dom} F := \{X \in \mathbf{S}^m, X \succ 0\} \to \mathbf{S}^n_+$. Prove that $F$ is $\succeq$-concave on its domain.

*Solution:* We should prove that the $\succeq$-hypograph of $F$ – the set

$$\mathcal{E} = \{X, Y : X \in \mathbf{S}^m, X \succ 0, Y \in \mathbf{S}^n, Y \preceq F(X)\}$$

is convex. To this end it suffices to show that the set

$$\mathcal{F} = \{X, Y : X \in \mathbf{S}^m, X \succ 0, 0 \prec Y \preceq F(X)\}$$

is convex, since $\mathcal{E}$ is the sum of $\mathcal{F}$ and the convex set $\{0_{m\times m}\} \times [-\mathbf{S}^n_+]$. We have

$$\begin{array}{rl} & Y \succ 0 \ \&\ X \succ 0 \ \&\ Y \preceq [A^\top X^{-1} A]^{-1} \\ \Longleftrightarrow & Y \succ 0 \ \&\ X \succ 0 \ \&\ A^\top X^{-1} A \preceq Y^{-1} \ \text{[Exercise 16]} \\ \Longleftrightarrow & Y \succ 0 \ \&\ X \succ 0 \ \&\ \left[\begin{array}{c|c} Y^{-1} & A^\top \\ \hline A & X \end{array}\right] \succeq 0 \ \text{[Schur Complement Lemma]} \\ \Longleftrightarrow & Y \succ 0 \ \&\ X \succ 0 \ \& X \succeq AYA^\top \ \text{[Schur Complement Lemma]} \end{array}$$

and the resulting description of $\mathcal{F}$ clearly states that $\mathcal{F}$ is convex. □

Note: the result we have just proved is the special case of the one stated in Exercise III.2.5.

**Exercise IV.20.** [cone-constrained semidefinite problems]

1. Let $X, Y \in \mathbf{S}^m_+$. Prove that $\mathrm{Tr}(XY) = 0$ is and only if $XY = YX = 0$.
2. Given an ordered collection $\nu = \{n_1, ..., n_k\}$ of positive integers, let $\mathbf{S}^\nu$ be the space of block-diagonal symmetric matrices with $k$ diagonal blocks of sizes $n_1 \times n_1$,...,$n_k \times n_k$, and let $\mathbf{S}^\nu_+$ be the cone of positive semidefinite matrices from $\mathbf{S}^\nu$. Equipping $\mathbf{S}^\nu$ with the Frobenius inner product, $\mathbf{S}^\nu_+$ clearly is a self-dual regular cone in the resulting Euclidean space.

   Convex cone-constrained problem on the cone $\mathbf{S}^\nu_+$ is of the generic form

   $$\mathrm{Opt}(\mathrm{SDP}) = \min_{x \in X} \left\{ f(x) : \overline{g}(x) := Ax - b \leq 0,\, \widehat{g}(x) := \mathrm{Diag}\{g_1(x), ..., g_k(x)\} \leq_{\mathbf{S}^\nu_+} 0 \right\} \tag{SDP}$$

   where $X$ is a nonempty convex set in some $\mathbf{R}^n$, the function $f : X \to \mathbf{R}$ is convex, and the mapping $\widehat{g} : X \to \mathbf{S}^\nu$ is $\mathbf{S}^\nu_+$-convex.

   Prove that in the case of convex cone-constrained semidefinite problem (SDP) Theorem IV.24.7 reads

> **Theorem IV.24.7.SDP** *Consider a convex cone-constrained semidefinite problem* (SDP), *let $x^* \in X$ be a feasible solution to the problem, and let $f$ and $\widehat{g}$ be differentiable at $x^*$.*
>
> (i) *If $x^*$ is a KKT point of* (SDP), *the Lagrange multipliers being $\overline{\lambda}^* \geq 0$ and $\widehat{\lambda}^* \in \mathbf{S}^\nu_+$, meaning that*
>
> $$\begin{array}{ll} \overline{\lambda}^*_i [\overline{g}(x^*)]_i = 0\, \forall i \;\&\; \widehat{\lambda}^* \widehat{g}(x^*) = 0 & \text{[sdp complementary slackness]} \\ \nabla_x \left[ f(x) + [\overline{\lambda}^*]^\top \overline{g}(x) + \mathrm{Tr}(\widehat{\lambda}^* \widehat{g}(x)) \right] \big|_{x=x^*} \in -N_X(x^*) & \text{[ KKT equation]} \end{array}$$
>
> *(here, as always, $N_X(x)$ is the normal cone of $X$, see (15.5)), then $x^*$ is an optimal solution to* (SDP).
>
> (ii) *If $x^*$ is optimal solution to* (SDP) *and, if addition to the above premise,* (SDP) *satisfies the cone-constrained Relaxed Slater condition, then $x^*$ is an sdp KKT point, as defined in item* (i).

*Solution:* 1. This claim is Exercise IV.3.2.

2. Straightforward application of Theorem IV.24.7 to the convex cone-constrained problem (SDP) differs from Theorem IV.24.7.SDP in the only point: in the complementary slackness part of the former Theorem, the $\widehat{g}$-related equality reads $\mathrm{Tr}(\widehat{\lambda}^* \widehat{g}(x^*)) = 0$, while in Theorem IV.24.7.SDP the corresponding equality is $\widehat{\lambda}^* \widehat{g}_j(x^*) = 0$. Taking into account that we are in the case $\widehat{\lambda}^* \succeq 0,, \widehat{g}(x^*) \preceq 0$ and invoking item 1 of Exercise, in the situation in question both equalities are satisfied/not satisfied simultaneously.

**Exercise IV.21.** [follow-up to Exercise IV.20] In the sequel, we fix the dimension $n$ of the embedding space and denote by $E_C = \{x \in \mathbf{R}^n : x^\top C x \leq 1\}$ the centered at the origin ellipsoid associated with positive definite $n \times n$ matrix $C$. Given positive $K$ and $K$ ellipsoids $E_{A_k}$, $k \leq K$, consider two optimization problems:

— $\mathcal{O}$: find the smallest volume centered at the origin ellipsoid containing $\cup_{k \leq K} E_{A_k}$,

— $\mathcal{I}$: find the largest volume centered at the origin ellipsoid contained in $\cap_{k \leq K} E_{A_k}$.

1. Pose $\mathcal{O}$ as a solvable convex cone-constrained semidefinite program
2. Prove that problems $\mathcal{O}$ and $\mathcal{I}$ reduce to each other at the cost of appropriate modification of the data
3. Prove that there exist matrices $\Lambda_k \succeq 0$ such that $\Lambda := \sum_k \Lambda_k \succ 0$ and

$$\Lambda_k = \Lambda_k A_k \Lambda,\ k \leq K.$$

*Solution:* Let $X$ be the set of positive definite $n \times n$ matrices, and $\nu = \{\underbrace{n, ..., n}_{K}\}$.

1: By the result of Exercise I.14.2, we have $E_P \subset E_Q$ iff $P \succeq Q$. Specifying a candidate solution to $\mathcal{O}$

as $E_U$, the constraint $E_U \supset \cup_k E_{A_k}$, that is, $E_{A_k} \subset E_U$ for all $k$, becomes $U \preceq A_k$ and $U \in X$. By the result of Exercise I.14.3, $\mathrm{Vol}(E_U) = \mathrm{Det}^{-1/2}(U)$, so that $\mathcal{O}$ can be posed as the optimization problem

$$\min_{U \in X} \left\{ -\ln \mathrm{Det}(U) : \widehat{g}(U) := \mathrm{Diag}\{U - A_1, ..., U - A_K\} \leq_{\mathbf{S}^\nu_+} 0 \right\} \tag{$\mathcal{O}$}$$

By Fact III.18.6 applied to the convex symmetric function $g(t) = -\sum_i \ln t_i : \mathrm{int}\, \mathbf{R}^n_+ \to \mathbf{R}$, the objective in $(\mathcal{O})$ is convex on $X$; thus, $(\mathcal{O})$ is a convex cone-constrained semidefinite program,

It is immediately seen that the problem is solvable. Indeed, its feasible set $F$ is nonempty and bounded, and the sublevel sets $\{U \in F : -\ln \mathrm{Det}(U) \leq a\}$ of the objective on this set clearly are closed, so that on the feasible set the objective attains its minimum.

2: As we know from Example II.8.11, $\mathrm{Polar}\,(E_P) = E_{P^{-1}}$, and from Exercise II.38

$$\mathrm{Vol}(\mathrm{Polar}\,(E_P)) = 1/\mathrm{Vol}(E_P).$$

Besides this, passing to polars reverses inclusions. It follows that an ellipsoid $E_U$ is a feasible solution to problem $\mathcal{O}$ with data $A_1, ..., A_K$ iff the ellipsoid $E_{U^{-1}}$ is a feasible solution to problem $\mathcal{I}$ with the data $A_1^{-1}, ..., A_K^{-1}$, and the volumes Vol of these two ellipsoids are reciprocals of each other. The bottom line is that problem $\mathcal{O}$ with data $A_1, ..., A_K$ reduces straightforwardly to problem $\mathcal{I}$ with data $A_1^{-1}, ..., A_K^{-1}$, and vice versa.

3: The objective $-\ln \mathrm{Det}(U)$ in the cone-constrained Problem $(\mathcal{O})$ is differentiable everywhere on $X$ with the gradient $-U^{-1}$, see Example C.9. Besides this, $(\mathcal{O})$ is strictly feasible due to $A_k \succ 0$, $k \leq K$. Let $U_*$ be an optimal solution to the problem (it exists by item 1). The cone-constrained Lagrange function of $(\mathcal{O})$ is

$$-\ln \mathrm{Det}(U) + \sum_{k=1}^{K} \mathrm{Tr}(\Lambda_k [U - A_k]).$$

Invoking Theorem IV.24.7.SDP.ii and taking into account that $U_* \in \mathrm{int}\, X = X$, there exist Lagrange multipliers $\Lambda_k \in \mathbf{S}^n_+$, $k \leq K$, such that

$$-U_*^{-1} + \sum_k \Lambda_k = 0 \;\&\; \Lambda_k (U_* - A_k) = 0,\; k \leq K.$$

Augmenting $\Lambda_1, ..., \lambda_K$ with $\Lambda := \sum_k \Lambda_k = U_*^{-1} \succ 0$, we meet the requirements from item 3. □

**Exercise IV.22.** Recall convex cone-constrained problem in Example IV.23.1, section 23.1

$$\mathrm{Opt}(P) = \min_{x=(t,y) \in \mathbf{R} \times \mathbf{S}^n} \{ t : \underbrace{t \geq \mathrm{Tr}(y)}_{\Longleftrightarrow\, \langle y, I_n \rangle - t \leq 0}, y^2 \preceq B \} \tag{23.1}$$

where $B$ is a positive definite matrix.

1. Verify (23.2)

*Solution:* We have

$$\begin{array}{rl} & L(t, y; \lambda, \Lambda) = t + \overline{\lambda}[\mathrm{Tr}(y) - t] + \mathrm{Tr}(\widehat{\lambda}[y^2 - B]) : [\mathbf{R} \times \mathbf{S}^n] \times [\mathbf{R}_+ \times \mathbf{S}^n_+] \to \mathbf{R} \\ \Longrightarrow & \underline{L}(\overline{\lambda}, \widehat{\lambda}) := \inf_{t \in \mathbf{R}, y \in \mathbf{S}^n} \left[ t + \overline{\lambda}[\mathrm{Tr}(y) - t] + \mathrm{Tr}(\widehat{\lambda}[y^2 - B]) \right] \\ & = \left\{ \begin{array}{ll} -\infty & , \overline{\lambda} \neq 1 \\ \inf_{y \in \mathbf{S}^n} \left[ \mathrm{Tr}(y) + \mathrm{Tr}(\widehat{\lambda}[y^2 - B]) \right] & , \overline{\lambda} = 1 \end{array} \right. \\ & = \left\{ \begin{array}{ll} -\infty & , \overline{\lambda} \neq 1 \\ -\infty & , \overline{\lambda} = 1 \;\&\; \widehat{\lambda} \in \mathrm{bd}\, \mathbf{S}^n_+ \\ -\frac{1}{4}\mathrm{Tr}(\widehat{\lambda}^{-1}) - \mathrm{Tr}(\widehat{\lambda} B) & , \overline{\lambda} = 1 \;\&\; \widehat{\lambda} \in \mathrm{int}\, \mathbf{S}^n_+ \end{array} \right. \end{array}$$

where the concluding equality stems from the following observations:

- when $\widehat{\lambda} \in \mathrm{bd}\,\mathbf{S}^n_+$, $\widehat{\lambda}$ has a nontrivial kernel; taking as $f$ a nonzero vector from this kernel and setting $y(s) = -sff^\top$, we get

$$L(\mathrm{Tr}(y(s)), y(s); 1, \widehat{\lambda}) = -sf^\top f + \mathrm{Tr}(s^2[f^\top f]\underbrace{\widehat{\lambda} f f^\top}_{=0} - \widehat{\lambda}B) = -sf^\top f - \mathrm{Tr}(\widehat{\lambda}B) \to -\infty,\ s \to \infty,$$

that is, $\underline{L}(1, \widehat{\lambda}) = -\infty$;
- when $\widehat{\lambda} \succ 0$, the minimum in $t, y$ of the convex in $(t, y)$ function

$$L(t, y; 1, \widehat{\lambda}) = \mathrm{Tr}(y) + \mathrm{Tr}(\widehat{\lambda}[y^2 - B])$$

can be found from the Fermat equation

$$0 = \nabla_y \left[\mathrm{Tr}(y) + \mathrm{Tr}(\widehat{\lambda}[y^2 - B])\right] = I_n + \widehat{\lambda}y + y\widehat{\lambda}$$

resulting in $y = -\frac{1}{2}\widehat{\lambda}^{-1}$ and

$$\underline{L}(1, \widehat{\lambda}) = -\frac{1}{4}\mathrm{Tr}(\widehat{\lambda}^{-1}) - \mathrm{Tr}(\widehat{\lambda}B),$$

as claimed in (23.2). □

2. Find Lagrange multipliers certifying that $t_* = -\mathrm{Tr}(B^{1/2})$, $y_* = -B^{1/2}$ is a cone-constrained KKT point of problem (23.1) (and thus, by Theorem IV.24.7, is an optimal solution to the problem).

*Solution:* As it should be, the Lagrange multipliers, if any, certifying that $(t_*, y_*)$ is a KKT point of (23.1) should be an optimal solution to the cone-constrained Lagrange dual (23.3) of (23.1). This solution

$$\overline{\lambda} = 1, \widehat{\lambda} = \frac{1}{2}B^{-1/2}$$

was found in section 23.3, and augmenting $(t_*, y_*)$ with these Lagrange multipliers, we clearly meet the complementary slackness

$$\overline{\lambda}[t_* - \mathrm{Tr}(y_*)] = 0,\ \mathrm{Tr}(\widehat{\lambda}[y_*^2 - B]) = 0$$

and the KKT equation

$$\nabla_t L(t, y; \overline{\lambda}, \widehat{\lambda})\Big|_{t=t_*, y=y_*} = 0,\ \nabla_y L(t, y; \overline{\lambda}, \widehat{\lambda})\Big|_{t=t_*, y=y_*} = I_n + [y_*\widehat{\lambda} + \widehat{\lambda}y_*] = 0$$

3. Consider the parametric family

$$\mathrm{Opt}(p := (v, w)) = \min_{t \in \mathbf{R}, y \in \mathbf{S}^n} \left\{t : t \geq \mathrm{Tr}(y), yv^{-1}y \preceq w\right\} \qquad (\mathrm{P}[p])$$

of convex cone-constrained problems, with $p \in P = \{p = (v, w) : v \in \mathrm{int}\,\mathbf{S}^n_+, w \in \mathrm{int}\,\mathbf{S}^n_+\}$, so that (23.1) is problem (P[$\overline{p}$]) corresponding to

$$\overline{p} = (I_n, B).$$

Prove that $\mathrm{Opt}(p)$ is convex function of $p \in P$ and find a subgradient of this function at the point $\overline{p}$.

*Solution:* Observe that setting $x = (t, y)$, the scalar function $\overline{g}(x, p) = \mathrm{Tr}(y) - t$ clearly is $\mathbf{R}_+$-convex, and the $\mathbf{S}^n$-valued function $\widehat{g}(x, p) = yv^{-1}y - w$ is $\succeq$-convex in $(x, p) \in [\mathbf{R} \times \mathbf{S}^n] \times P$; indeed, by Fact IV.26.1 $\succeq$-convexity of $\widehat{g}(x, p)$ on the indicated domain is the same as the convexity of the $\succeq$-epigraph of $\widehat{g}$, and this epigraph is, by the Schur Complement Lemma, the set

$$\{(t, y, v, w, z) \in \left[\mathbf{R} \times \mathbf{S}^n \times [\mathrm{int}\,\mathbf{S}^n_+] \times \mathbf{S}^n\right] \times \mathbf{S}^n : \left[\begin{array}{c|c} z + w & y \\ \hline y & v \end{array}\right] \succeq 0\}$$

which clearly is convex. Applying Proposition IV.24.8, we conclude that $\mathrm{Opt}(p)$ is convex on $P$. Recalling that by item 2 of Exercise, $(t_* = -\mathrm{Tr}(B^{1/2}), y_* = -B^{1/2})$ is a cone-constrained KKT point of (P[$\overline{p}$]), the corresponding Lagrange multipliers being $\overline{\lambda} = 1$, $\widehat{\lambda} = \frac{1}{2}B^{-1/2}$, and taking into account that in the case in question in the notation from Proposition IV.24.8 we have

$$F_p = 0,\ G_p[\delta v, \delta w] = -y_* v^{-1}\delta v v^{-1} y_*\Big|_{v=I_n} - \delta w$$

(see Example C.8 in section C.1.6), the pair $[-\frac{1}{2}B^{1/2}, -\frac{1}{2}B^{-1/2}]$ is a subgradient of $\mathrm{Opt}(\cdot)$ at $p = \overline{p} = [I_n, B]$:

$$\begin{array}{rcl}\mathrm{Opt}(p=(v,w)) & \geq & \underbrace{-\mathrm{Tr}(B^{1/2})}_{\mathrm{Opt}(I_n,B)} + \mathrm{Tr}(\widehat{\lambda}[G_p[v-I_n, w-B]]) \\ & = & -\mathrm{Tr}(B^{1/2}) - \frac{1}{2}\mathrm{Tr}([v-I_n]B^{1/2}) - \frac{1}{2}\mathrm{Tr}(B^{-1/2}[w-B]).\end{array}$$

**Exercise IV.23.** [follow-up to Exercise IV.4] Given positive integers $m, n$, consider two parametric families of convex sets:

- $S_1[P] = \{(X,Y) \in \mathcal{R}_1 := \mathbf{S}^m \times \mathbf{S}^n : \left[\begin{array}{c|c} X & P \\ \hline P^\top & Y \end{array}\right] \succeq 0\}$, where the "parameter" $P$ runs through the space $\mathbf{R}^{m\times n}$ of $m \times n$ matrices, let it be temporarily denoted $\mathcal{P}_1$;
- $S_2[P] = \{(X,Y) \in \mathcal{R}_2 := \mathbf{S}^m \times \mathbf{R}^{m\times n} : \left[\begin{array}{c|c} X & Y \\ \hline Y^\top & P \end{array}\right] \succeq 0\}$, where the "parameter" $P$ runs through the positive semidefinite cone $\mathbf{S}^n_+$, let it be temporarily denoted $\mathcal{P}_2$.

Prove that for $\chi = 1, 2$ the set-valued mappings $P \to S_\chi[P]$ are super-additive on their domains:

$$P, Q \in \mathcal{P}_\chi \implies P + Q \in \mathcal{P}_\chi \ \&\ \underbrace{S_\chi[P] + S_\chi[Q] \subset S_\chi[P+Q]}_{(*)}.$$

and that the concluding inclusion
— not necessarily is equality for $\chi = 1$, and
— is equality for $\chi = 2$.

*Solution:* The super-additivity is evident due to the following evident fact:

> *Let $\mathbf{K} \subset \mathbf{R}^M$, $\mathcal{P} \subset \mathbf{R}^N$ be cones, and $\mathcal{A}(U,P)$ be a linear (linear, not affine!) mapping from $\mathbf{R}^K_U \times \mathbf{R}^N_P$ into $\mathbf{R}^M$ such that the set*
>
> $$S[P] = \{U \in \mathbf{R}^K_U : \mathcal{A}(U,P) \in \mathbf{K}\}$$
>
> *is nonempty when $P \in \mathcal{P}$. Then the set-valued mapping $P \mapsto S[P]$ is super-additive on $\mathcal{P}$.*
> Indeed, when $P_i \in \mathcal{P}$ and $U_i \in S[P_i]$, $i = 1, 2$, we have $\mathbf{K} \ni \mathcal{A}(U_1, P_1) + \mathcal{A}(U_2, P_2) = \mathcal{A}(U_1 + U_2, P_1 + P_2)$, implying that $U_1 + U_2 \in S[P_1 + P_2]$.

The fact that inclusion $(*)$ can be strict when $\chi = 1$ is nearly evident: take a nonzero $P \in \mathbf{R}^{m\times n}$ and note that all pairs $(X,Y) \in S_1[P]$, same as all pairs $(X,Y) \in S_1[-P]$, have nonzero positive semidefinite components $X, Y$, while $S[0_{m\times n}]$ contains the pair $(X = 0_{m\times m}, Y = 0_{n\times n})$ which clearly cannot be represented as the sum of two pairs of matrices with nonzero positive semidefinite components in every pair. Thus, when $Q = -P \neq 0$ and $\chi = 1$, inclusion $(*)$ is strict.
The only nontrivial part of the exercise is to prove that when $\chi = 2$, the inclusion $(*)$ is equality whenever $P, Q$ are positive semidefinite. In other words, we want to demonstrate that when

$$P \in \mathbf{S}^n_+ \ \&\ Q \in \mathbf{S}^n_+ \ \&\ R \in \mathbf{S}^m, S \in \mathbf{R}^{m\times n} \ \&\ D := \left[\begin{array}{c|c} R & S \\ \hline S^\top & P+Q \end{array}\right] \succeq 0 \tag{1}$$

the matrix $D$ can be decomposed into the sum of two positive semidefinite matrices, one of the form $\left[\begin{array}{c|c} U & V \\ \hline V^\top & P \end{array}\right]$, and the other – of the form $\left[\begin{array}{c|c} W & Z \\ \hline Z^\top & Q \end{array}\right]$. This is the same as to verify that setting

$$\begin{array}{c} A_+(U,V) = \left[\begin{array}{c|c} U & V \\ \hline V^\top & \end{array}\right],\ A_-(U,V) = \left[\begin{array}{c|c} -U & -V \\ \hline -V^\top & \end{array}\right], \\ b_+ = -\left[\begin{array}{c|c} & \\ \hline & P \end{array}\right],\ b_- = -\left[\begin{array}{c|c} R & S \\ \hline S^\top & Q \end{array}\right] \end{array}$$

the conic constraint

$$\underbrace{(A_+(U,V),A_-(U,V))}_{=:A(U,V)}-\underbrace{(b_+,b_-)}_{=:b}\in\underbrace{\mathbf{S}^{m+n}_+\times\mathbf{S}^{m+n}_+}_{=:\mathbf{K}} \tag{2}$$

in variables $U,V$ is feasible.

Assume that the constraint is infeasible, and let us lead this assumption to contradiction. The image space of the linear mapping $(U,V)\mapsto A(U,V)$ is composed of pairs

$$\left(\left[\begin{array}{c|c}U&V\\ \hline V^\top&\end{array}\right],\left[\begin{array}{c|c}-U&-V\\ \hline -V^\top&\end{array}\right]\right)$$

of symmetric $(m+n)\times(m+n)$ matrices with $U\in\mathbf{S}^m$ and $V\in\mathbf{R}^{m\times n}$; such a pair can belong to $\mathbf{K}$ (that is, has both components positive semidefinite) iff $V=0$ and both $U$ and $-U$ are positive semidefinite, that is, iff $U=0$ and $V=0$. With this in mind, the results of Exercise IV.4 say that infeasibility of (2) implies existence of $\lambda=(\lambda_+,\lambda_-)\in\mathbf{K}_*$, that is, $\lambda_\pm\in\mathbf{S}^{m+n}_+$, such that

$$\mathrm{Tr}(\lambda_+A_+(U,V))+\mathrm{Tr}(\lambda_-A_-(U,V))=0\,\forall(U\in\mathbf{S}^m,V\in\mathbf{R}^{m\times n})\ \&\ \mathrm{Tr}(\lambda_+b_+)+\mathrm{Tr}(\lambda_-b_-)>0. \tag{3}$$

Representing

$$\lambda_\pm=\left[\begin{array}{c|c}E_\pm&F_\pm\\ \hline F_\pm^\top&G_\pm\end{array}\right]\quad[E_\pm\in\mathbf{S}^m,G_\pm\in\mathbf{S}^n)$$

the first relation in (3) boils down to

$$\mathrm{Tr}(U[E_+-E_-])+2\mathrm{Tr}(V[F_+-F_-]^\top)=0\,\forall(U\in\mathbf{S}^m,V\in\mathbf{R}^{m\times n}),$$

that is, $E_+=E_-=:E$, $F_+=F_-=:F$. Now the second relation in (3) reads

$$\mathrm{Tr}(G_+P)+\mathrm{Tr}(ER)+2\mathrm{Tr}(FS^\top)+\mathrm{Tr}(G_-Q)<0. \tag{4}$$

and, besides this $\left[\begin{array}{c|c}E&F\\ \hline F^\top&G_\pm\end{array}\right]\succeq0$. When replacing $E$ in (4) with $E'=E+\epsilon I_m$ with small enough positive $\epsilon$, the strict inequality remains valid:

$$\mathrm{Tr}(G_+P)+\mathrm{Tr}(E'R)+2\mathrm{Tr}(FS^\top)+\mathrm{Tr}(G_-Q)<0. \tag{5}$$

On the other hand, we have $\left[\begin{array}{c|c}E'&F\\ \hline F^\top&G_\pm\end{array}\right]\succeq0$ and $E'\succ0$, whence, by Schur Complement Lemma, $G_\pm\succeq G:=F^\top[E']^{-1}F$. When replacing $G_\pm$ with $G\preceq G_\pm$ in the left hand side of (5), we can only decrease the value of the left hand side due to $P\succeq0$, $Q\succeq0$, so that

$$\mathrm{Tr}(GP)+\mathrm{Tr}(E'R)+2\mathrm{Tr}(FS^\top)+\mathrm{Tr}(GQ)<0.$$

The left hand side here is nothing but

$$\mathrm{Tr}\left(\left[\begin{array}{c|c}R&S\\ \hline S^\top&P+Q\end{array}\right]\left[\begin{array}{c|c}E'&F\\ \hline F^\top&G\end{array}\right]\right)$$

i.e., it is the Frobenius inner product of two positive semidefinite (by (1) and due to the origin of $G$) matrices, and this product cannot be negative, which is the desired contradiction. □

**Exercise IV.24.** In the simplest Steiner problem, one is given $m$ distinct points $a_1,\dots,a_m$ in $\mathbf{R}^n$ and is looking for a point $x_*$ such that the sum of Euclidean distances between the points and $x_*$ is as small as possible (think, e.g., about $m$ oil wells on 2D plane and the problem of locating collector to be linked to the wells by pipes in a way minimizing the total length of the pipes).

1. Pose the problem as conic problem, the cone being direct product of $m$ Lorentz cones.
2. Build the dual problem. Are the primal and the dual problems solvable? Are the primal and the dual optimal values equal to each other?
3. Write down optimality conditions and see what they say
*Hint:* You are advised to consider separately the cases where optimal solution differs from all of the points $a_1,\dots,a_m$, and the case when it is one of the points.

4. Solve the problem in the case when $n = 2$, $m = 3$ and $a_1, a_2, a_3$ are vertices of triangle on 2D plane.

*Solution:* 1: The "maiden" form of the problem is

$$\min_{x\in\mathbf{R}^n} \sum_{i=1}^m \|x - a_i\|_2, \tag{$P_{\rm ini}$}$$

the conic reformulation is

$$\min_{t,x} \left\{ \sum_i t_i : \|x - a_i\|_2 \leq t_i, i \leq m \right\};$$

the constraints in this reformulation say that the vectors $[x - a_i; t_i]$ belong to the Lorentz cone $\mathbf{L}^{n+1}$, and the objective is linear. To obey our standards on writing down conic problems, we should rewrite the above problem as

$$\begin{array}{c}\mathrm{Opt}(P) = \min_{x,t} \left\{\sum_i t_i : A\,[x;t] - b := [[x - a_1; t_1]; [x - a_2; t_2]; \ldots ; [x - a_m; t_m]] \geq_{\mathbf{K}} 0\right\}\\ \mathbf{K} = \underbrace{\mathbf{L}^{n+1} \times \ldots \times \mathbf{L}^{n+1}}_{m \text{ times}}\end{array} \tag{$P$}$$

2: To build the dual problem, we equip the conic constraint with Lagrange multiplier restricted to belong to the dual to $\mathbf{K}$ cone $\mathbf{K}_*$. This dual cone is $\mathbf{K}$ itself due to self-duality of the Lorentz cone, so that the Lagrange multiplier is $[[y_1; s_1]; \ldots ; [y_m; s_m]]$ with $[y_i; s_i] \in \mathbf{L}^{n+1}$. We then take the sidewise inner product of the conic constraint of the primal problem with the Lagrange multiplier, thus arriving at the scalar linear inequality

$$\sum_i [t_i s_i + y_i^\top x] \geq \sum_i y_i^\top a_i; \tag{$*$}$$

whenever $[y_i; s_i] \in \mathbf{L}^{n+1}$, $i \leq m$, this inequality is consequence of the constraints of the primal problem. To build the dual problem, we impose on the Lagrange multipliers, aside of the restrictions $[y_i; s_i] \in \mathbf{L}^{n+1}$, the restriction that the left hand side in $(*)$ is equal to the objective of $(P)$ identically in $x, t_1, \ldots , t_m$, which in our case reads

$$s_i = 1, i \leq m, \sum_i y_i = 0.$$

The dual problem is to maximize the right hand side of $(*)$ in $[y_i; s_i]$ under the resulting constraints, so that the dual problem reads

$$\mathrm{Opt}(D) = \max_{y_i,s_i} \left\{ \sum_i a_i^\top y_i : \|y_i\|_2 \leq s_i = 1, ., i \leq m, \sum_i y_i = 0 \right\} \tag{$D$}$$

The primal and the dual problems clearly satisfy the Slater and the Relaxed Slater conditions, respectively, so that by Conic Duality Theorem problems are solvable with equal optimal values.

3: Optimality conditions from Theorem IV.24.9 in their "complementary slackness" form state that the (under the circumstances, necessary and sufficient) condition for feasible solutions $(x, \{t_i\})$ to $(P)$ and $\{y_i, s_i\}$ to $(D)$ to be optimal for the respective problems is

$$\sum_i [x - a_i; t_i]^\top [y_i; s_i] = 0,$$

or, which is the same due to $s_i = 1$ (by dual feasibility) and $[x - a_i; t_i] \in \mathbf{L}^{n+1}$, $[y_i; s_i] \in \mathbf{L}^{n+1}$,

$$y_i^\top [a_i - x] = t_i, \ i \leq m$$

Since $\|y_i\|_2 \leq 1$ and $\|a_i - x\|_2 \leq t_i$, the above equality implies that $y_i^\top [a_i - x] \geq \|y_i\|_2 \|a_i - x\|_2$. This, taken together with what we know about equality case of Cauchy inequality (Theorem B.1) means that for every $i$,

— either $x \neq a_i$, and then $y_i = \frac{a_i - x}{\|a_i - x\|_2}$ and $t_i = \|a_i - x\|_2$

— or $x = a_i$, $t_i = 0$, and $\|y_i\|_2 \leq 1$.

We conclude that there are two possible cases:
(A): the $x$-component of optimal solution to $(P)$ differs from every one of $a_i$. In this case we should have $y_i = \frac{a_i-x}{\|a_i-x\|_2}$ for all $i$;
(B): the $x$-component of optimal solution to $(P)$ is some $a_{i_*}$. In this case $y_i = \frac{a_i-x}{\|a_i-x\|_2}$ for all $i \neq i_*$ and $y_{i_*}$ can be arbitrary vector of $\|\cdot\|_2$-norm $\leq 1$.
Taking into account that we should have $\sum_i y_i = 0$, we arrive at the following conclusions:

> (A): If $x$ is different from all $a_i$ and satisfies the relation $\sum_i \frac{a_i-x}{\|a_i-x\|_2} = 0$, then $x$ and $t_i = \|x-a_i\|_2$, $i \leq m$, form an optimal solution to $(P)$.
> (B): If $x = a_{i_*}$ for $i_*$ such that $\|\sum_{i\neq i_*} \frac{a_i-a_{i_*}}{\|a_i-a_{i_*}\|_2}\|_2 \leq 1$, then $x$ and $t_i = \|x-a_i\|_2$, $i \leq m$, form an optimal solution to $(P)$.
> Moreover, every optimal solution to $(P)$ is either given by (A), or by (B), and optimal solutions do exist.

4: When $m = 3$ and $a_1, a_2, a_3$ are the vertices of a triangle in 2D plane ($n = 2$), (A) says that a point distinct from $a_1, a_2, a_3$ solves $(P_{\text{ini}})$ iff all three sides of $\triangle a_1a_2a_3$ are seen from the point $x$ under angles $120^o$ — the unit vectors "looking" from $x$ at the vertices of the triangle should sum up to 0. Elementary geometry says that such a point does exist when all three angles of $\triangle a_1a_2a_3$ are $< 120^o$. If the latter is not the case, optimal $x$ is given by (B) and is just the vertex of $\triangle a_1a_2a_3$ with angle at the vertex $\geq 120^o$.

**Exercise IV.25.** Consider a primal-dual pair of conic problems

$$\begin{aligned}\mathrm{Opt}(P) &= \min_x \left\{c^\top x : \ Ax \geq_{\mathbf{K}} b\right\} && (P)\\ \mathrm{Opt}(D) &= \max_y \left\{b^\top y : \ y \geq_{\mathbf{K}_*} 0, \ A^\top y = c\right\} && (D)\end{aligned}$$

($\mathbf{K} \subset \mathbf{R}^n$ is a regular cone) and assume that both problems are feasible.

1. Find the recessive cones $\mathrm{Rec}(P)$ and $\mathrm{Rec}(D)$ of the primal and the dual feasible sets.
2. Prove that the feasible set of at least one of the problems is unbounded.

*Solution:* The feasible sets of $(P)$ and $(D)$ are nonempty, convex, and clearly closed. Let $\bar{x}$ be primal feasible and $\bar{y}$ be dual feasible. Then, we have

$$\begin{aligned}\mathrm{Rec}(P) &= \{h : \ A[\bar{x}+th] - b \in \mathbf{K}, \forall t \geq 0\} = \left\{h : \ Ah - t^{-1}[A\bar{x}-b] \in \mathbf{K}, \forall t > 0\right\}\\ &= \{h : \ Ah \in \mathbf{K}\}\\ \mathrm{Rec}(D) &= \left\{g : \ \bar{y} + tg \in \mathbf{K}_*, tA^\top g = 0, t \geq 0\right\} = \left\{g \in \mathbf{K}_* : \ A^\top g = 0\right\}.\end{aligned}$$

Now assume that the feasible set of $(D)$ is bounded, and let us prove that the feasible set of $(P)$ is unbounded. As $\mathbf{K}$ is a regular cone, we can select $f \in \operatorname{int}\mathbf{K}$. Consider two convex sets $S = \left\{y : \ A^\top y = 0\right\}$ and $T = \left\{y \in \mathbf{K}_* : \ f^\top y = 1\right\}$; note that $T \neq \varnothing$ due to $f \in \operatorname{int}\mathbf{K}$. We are in the situation where the dual feasible set is nonempty, closed, and bounded, implying by Fact II.8.13 that $\mathrm{Rec}(D) = \{0\}$, that is, $S \cap \mathbf{K}_* = \{0\}$, whence the closed nonempty convex sets $S$ and $T$ do not intersect. Recall also that as $\mathbf{K}$ is a regular cone so is its dual $\mathbf{K}_*$, and thus by Fact II.8.33 $T$ is compact. Then, the convex sets $S, T$ are at positive distance from each other and one of them is compact, hence they can be strongly separated: there exists a vector $a$ such that

$$\sup_{y\in S} a^\top y < \min_{y \in T} a^\top y.$$

As $S$ is a linear space and $T$ is a base of $\mathbf{K}_*$ (Fact II.8.33), this relationship is equivalent to

$$a \in S^\perp = [\mathrm{Ker} A^\top]^\perp = \mathrm{Im} A, \qquad \text{and} \qquad a \in \operatorname{int}(\mathbf{K}_*)_* = \operatorname{int}\mathbf{K}.$$

Here the inclusion $a \in \operatorname{int}(\mathbf{K}_*)_*$ is given by Fact II.8.33.ii as applied to the regular cone $\mathbf{K}_*$ in the role of $M$ combined with the fact that $a^\top y > 0$ for all $y \in T$ and therefore for all $y \in \mathbf{K}_* \setminus \{0\}$. Thus, we conclude that $a = Ah$ for some $h$ such that $Ah \in \operatorname{int}\mathbf{K}$ and thus $h \neq 0$, which clearly shows that

$\text{Rec}(P) = \{h : Ah \in \mathbf{K}\} \neq \{0\}$ and thus the primal feasible set is unbounded. Note that we have shown $\{h : Ah \in \text{int}\, \mathbf{K}\} \neq \varnothing$, which is indeed something even stronger than what was desired. □

**Exercise IV.26.** [semidefinite duality] A *semidefinite program* is a conic program involving the positive semidefinite cone. As a matter of fact, *Semidefinite programming* – the family of semidefinite programs – possesses extremely powerful "expressive abilities." It is prudent to say that *for all practical purposes*, whatever it means, Semidefinite programming is "the same" as the entire Convex programming. In this exercise we would like to acquaint the reader with the specific form taken by Conic duality when the cone involved is the positive semidefinite cone.

Formally, a semidefinite program is of the form

$$\text{Opt}(P) = \min_{x \in \mathbf{R}^n} \left\{ c^\top x : \begin{array}{c} Ax - b := \sum_j a_j x_j - b \geq 0 \\ \mathcal{A}x - B := x_1 A_1 + \ldots + x_n A_n - B \succeq 0 \end{array} \right\}, \tag{$P$}$$

where $a_j, b$ are vectors from some $\mathbf{R}^p$, and $A_j, B$ are matrices from some $\mathbf{S}^q$. "Real life" form of a semidefinite program usually is a bit different, namely,

$$\text{Opt}(\mathcal{P}) = \min_{x \in \mathbf{R}^n} \left\{ c^\top x : \begin{array}{l} Ax - b := \sum_j x_j a_j - b \geq 0 \\ \mathcal{A}_i x - B^i := x_1 A_1^i + \ldots + x_n A_n^i - B^i \succeq 0, \ \forall i \leq m \end{array} \right\}, \tag{$\mathcal{P}$}$$

where $A_j^i, B^i \in \mathbf{S}^{q_i}$. In the formulation $(\mathcal{P})$ as opposed to the formulation $(P)$ we have a bunch of positive semidefinite cone constraints, i.e., $\mathcal{A}_i x - B^i \succeq 0$, $i \leq m$, instead of a single constraint $\mathcal{A}x - B \succeq 0$. We can always rewrite $(\mathcal{P})$ in the form of $(P)$ by assembling $A_j^i$, $B^i$ into block-diagonal matrices $A_j = \text{Diag}\{A_j^1, \ldots, A_j^m\}$, $B = \text{Diag}\{B^1, \ldots, B^m\}$. Taking into account that a block-diagonal symmetric matrix is positive semidefinite iff all the diagonal blocks are positive semidefinite, we deduce that $(\mathcal{P})$ is equivalent to the problem

$$\min_{x \in \mathbf{R}^n} \left\{ c^\top x : \begin{array}{c} Ax - b := \sum_j x_j a_j - b \geq 0 \\ \mathcal{A}x - B := \sum_j x_j A_j - B \succeq 0 \end{array} \right\}$$

of the form $(P)$. When proving theorems, it is usually better to work with program in the form of $(P)$ – it saves notation; in contrast, when working with "real life" semidefinite programs, it is usually better to operate with problems in more detailed form $(\mathcal{P})$.

Your task is as follows:

1. Verify that the conic dual of $(\mathcal{P})$ is the semidefinite program

$$\max_{\lambda, \{\Lambda_i, i \leq m\}} \left\{ b^\top \lambda + \sum_{i=1}^m \text{Tr}(\Lambda_i B^i) : \begin{array}{c} \lambda \in \mathbf{R}_+^p, \Lambda_i \in \mathbf{S}_+^{q_i}, i \leq m \\ A^\top \lambda + \sum_{i=1}^m \mathcal{A}_i^* \Lambda_i = c, \end{array} \right\}, \tag{$\mathcal{D}$}$$

where for the linear mapping $x \mapsto \sum_j x_j A_j : \mathbf{R}^n \to \mathbf{S}^q$ its *conjugate* linear mapping $X \mapsto \mathcal{A}^* X : \mathbf{S}^q \to \mathbf{R}^n$ is given by the identity

$$\text{Tr}(X[\mathcal{A}x]) \equiv [\mathcal{A}^* X]^\top x \quad \forall (x \in \mathbf{R}^n, X \in \mathbf{S}^q),$$

or, which is the same,

$$\mathcal{A}^* X = [\text{Tr}(A_1 X); \ldots; \text{Tr}(A_n X)].$$

*Solution:* $(\mathcal{P})$ is the cone-constrained problem

$$\begin{array}{l} \min\limits_{x \in \mathbf{R}^n} \left\{ f(x) := c^\top x : \ \overline{g}(x) := b - Ax \leq 0, \ \widehat{g}(x) \right. \\ \left. := \text{Diag}\left\{ B^1 - \sum_j x_j A_j^1, \ldots, B^m - \sum_j x_j A_j^m \right\} \in -\mathbf{K} \right\} \end{array} \tag{$*$}$$

where $\mathbf{K}$ is the cone composed of the positive semidefinite block-diagonal symmetric matrices with $m$ diagonal blocks of sizes $q_1, \ldots, q_m$. Then, the cone $\mathbf{K}$ lives in the space $\mathbf{S}^{\{q_1, \ldots, q_m\}}$ of block-diagonal symmetric matrices with $m$ diagonal blocks of sizes $q_1, \ldots, q_m$. Equipping $\mathbf{S}^{\{q_1, \ldots, q_m\}}$ with Frobenius inner product and taking into account that positive semidefinite cone is self-dual, we immediately conclude that $\mathbf{K}$ is self-dual as well. As a result,

- the Lagrange multipliers $\Lambda \in \mathbf{K}_*$ are exactly block-diagonal matrices $\Lambda = \mathrm{Diag}\{\Lambda_1, \dots \Lambda_m\}$ with diagonal blocks $\Lambda_i \in \mathbf{S}^{q_i}_+$, for all $i \leq m$;
- the cone-constrained Lagrange function of $(*)$ is the function

$$L(x; \lambda, \Lambda) = f(x) + \lambda^\top \overline{g}(x) + \mathrm{Tr}(\widehat{g}(x)\Lambda), \tag{!}$$

where the last term in the right hand side is precisely what is prescribed by our general description of cone-constrained Lagrange function, i.e., it is the inner product of the Lagrange multiplier $\Lambda$ for the cone constraint $\widehat{g}(x) \leq_{\mathbf{K}} 0$ and the left hand side of this constraint[12]. In other words,

$$\begin{array}{r} L(x; \lambda, \Lambda) = c^\top x + \lambda^\top [b - Ax] + \sum_{i=1}^m \mathrm{Tr}\left(\Lambda_i [B^i - \sum_j A^i_j x_j]\right) : \\ \mathbf{R}^n_x \times \left[\mathbf{R}^p_+ \times \mathbf{S}^{q_1}_+ \times \ldots \times \mathbf{S}^{q_m}_+\right] \to \mathbf{R}. \end{array}$$

Consequently, the cone-constrained Lagrange dual of $(*)$ is the problem

$$\max_{\lambda \in \mathbf{R}^p_+, \Lambda = \{\Lambda_i \in \mathbf{S}^{q_i}_+\}} \left\{ \begin{array}{l} \underline{L}(\lambda, \Lambda) \\ := \inf\limits_{x \in \mathbf{R}^n} \left[\lambda^\top b + \sum_i \mathrm{Tr}(\Lambda_i B^i) + \sum_j x_j [c_j - \lambda^\top a_j - \sum_i \mathrm{Tr}(\Lambda_i A^i_j)]\right]. \end{array} \right\}$$

Note also that

$$\underline{L}(\lambda, \Lambda) = \left\{ \begin{array}{ll} \lambda^\top b + \sum_i \mathrm{Tr}(\Lambda_i B_i), & \text{if } A^\top \lambda + \sum_i \left[\mathrm{Tr}(A^i_1 \Lambda_i); \ldots; \mathrm{Tr}(A^i_n \Lambda_i)\right] = c \\ -\infty, & \text{otherwise} \end{array} \right. .$$

Therefore, the conic dual of $(\mathcal{P})$ is given by

$$\max_{\lambda, \{\Lambda_i, i \leq m\}} \left\{ \lambda^\top b + \sum_i \mathrm{Tr}(\Lambda_i B_i) : \begin{array}{l} \lambda \in \mathbf{R}^p_+,\ \Lambda_i \in \mathbf{S}^{q_i}_+,\ i \leq m \\ A^\top \lambda + \sum_i \left[\mathrm{Tr}(A^i_1 \Lambda_i); \ldots; \mathrm{Tr}(A^i_n \Lambda_i)\right] = c \end{array} \right\} \tag{$\mathcal{D}$}$$

In words, the recipe for building the dual to the semidefinite program $(\mathcal{P})$ is as follows:

1. We equip the constraints of $(\mathcal{P})$ with Lagrange multipliers, specifically, the linear constraints $Ax - b \geq 0$ with the multiplier $\lambda \in \mathbf{R}^p$ such that $\lambda \geq 0$, and the semidefinite constraints $\mathcal{A}_i x - B_i := \sum_j x_j A^i_j - B^i \succeq 0$ with the multipliers $\Lambda_i \in \mathbf{S}^{q_i}$ such that $\Lambda_i \succeq 0$.
2. We take the inner products of the left hand sides of the constraints in $(\mathcal{P})$ and the associated Lagrange multipliers (the standard inner product for the linear constraint $Ax - b \geq 0$, and the Frobenius inner products for the semidefinite constraints $\mathcal{A}_i x - B^i \succeq 0$) and sum up the results, arriving at the aggregated constraint

$$\left[A^\top \lambda + \sum_i \mathcal{A}^*_i \Lambda_i\right]^\top x \geq b^\top \lambda + \sum_i \mathrm{Tr}(B^i \Lambda_i),$$
$$\text{where } \mathcal{A}^*_i X = [\mathrm{Tr}(A^i_1 X); \ldots; \mathrm{Tr}(A^i_n X)].$$

By its origin, this constraint is a consequence of the system of constraints in $(\mathcal{P})$.

[12] in our general description of cone-constrained Lagrange function, the cone in the cone constraint lived in some $\mathbf{R}^N$, and the product of the Lagrange multiplier and the body of the constraint was the standard inner product in $\mathbf{R}^N$. Our present situation can be reduced to the standard one by identifying $\mathbf{S} = \mathbf{S}^{\{q_1, \ldots, q_m\}}$ equipped with the Frobenius inner product with appropriate $\mathbf{R}^N$ equipped with the standard inner product, identification being given by selecting orthonormal, w.r.t. the Frobenius inner product, basis in $\mathbf{S}$ and identifying $X \in \mathbf{S}$ with the vector of coordinates of $X$ in this basis. There, however, is no necessity to carry out this identification explicitly, since all we are interested in is what will be the standard inner product of vectors of coordinates of $\Lambda$ and of $\widehat{g}(x)$ in this orthonormal basis, and we know the answer in advance – this will be the Frobenius inner product of $\Lambda$ and $\widehat{g}(x)$, the entity we see in (!).

3. We impose on the Lagrange multipliers, aside of the restrictions mentioned in item 1, the restriction that the left hand side in the aggregated constraint is equal to $c^\top x$ identically in $x \in \mathbf{R}^n$, so that the right hand side in this constraint is a lower bound on Opt($\mathcal{P}$), The dual program ($\mathcal{D}$) is nothing but the problem of maximizing this lower bound over Lagrange multipliers satisfying the restrictions just listed.

**Exercise IV.27.** [example of semidefinite relaxation] Let $T_k \succeq 0, k \leq K$, be positive semidefinite $m \times m$ matrices such that $\sum_k T_k \succ 0$, $\mathcal{T} \subset \mathbf{R}^K_+$ be a convex compact set intersecting the interior of $\mathbf{R}^K_+$, and $A$ be a symmetric $m \times m$ matrix. Let also $\phi_{\mathcal{T}}(z) = \max_{t\in\mathcal{T}} z^\top t$ be the support function of $\mathcal{T}$. Prove that

$$\begin{array}{rcll}\mathrm{Opt} & := & \min_z \left\{\phi_{\mathcal{T}}(z) : z \geq 0, A \preceq \sum_k z_k T_k\right\} & (a)\\ & = & \max_{\Lambda,t} \left\{\mathrm{Tr}(A\Lambda) : \Lambda \succeq 0, t \in \mathcal{T}, \mathrm{Tr}(T_k\Lambda) \leq t_k, k \leq K\right\} & (b)\end{array}$$

and that both minimization and maximization problems above are solvable.

*Solution:* Since $\mathcal{T}$ is bounded, $\phi_{\mathcal{T}}$ is real-valued and continuous, and since $\mathcal{T} \subset \mathbf{R}^K_+$ contains a positive vector, the sets $\{z \geq 0 : \phi_{\mathcal{T}}(z) \leq a\}$ are closed and bounded for every $a \in \mathbf{R}$. The problem specifying Opt is cone constrained problem which is strictly feasible (due to $\sum_k T_k \succ 0$), and by the above, denoting by $\mathcal{Z}$ the feasible set of the problem, the feasible sublevel sets $\{z \in \mathcal{Z} : \phi_{\mathcal{T}}(z) \leq a\}$ of $\phi_{\mathcal{T}}$ are closed and bounded for every $a$; since the objective is continuous, it follows that the problem is solvable (Theorem B.32). The minimization problem specifying Opt is cone constrained strictly feasible and below bounded problem. Thus, by cone constrained version of Convex Programming Duality Theorem (Theorem IV.23.1), the cone constrained Lagrange dual of problem ($a$) is solvable with optimal value Opt. The cone constrained Lagrange function of ($a$) is

$$L(z;\lambda,\Lambda) = \phi_{\mathcal{T}}(z) - \lambda^\top z + \mathrm{Tr}(\Lambda[A - \sum_k z_k T_k]) : \mathbf{R}^K_z \times [\mathbf{R}^K_+ \times \mathbf{S}^m_+] \to \mathbf{R},$$

so that the objective in the dual problem is

$$\underline{L}(\lambda,\Lambda) = \mathrm{Tr}(A\Lambda) + \inf_{z\in\mathbf{R}^K}\left[\phi_{\mathcal{T}}(z) - z^\top \underbrace{[\mathrm{Tr}(\Lambda T_1) + \lambda_1; ...; \mathrm{Tr}(\Lambda T_K) + \lambda_K]}_{=:\ell(\lambda,\Lambda)}\right],$$

that is, $\underline{L}(\lambda,\Lambda) - \mathrm{Tr}(\Lambda A)$ is the minus Legendre transform $\phi^*_{\mathcal{T}}$ of $\phi_{\mathcal{T}}(\cdot)$ as evaluated at $\ell(\lambda,\Lambda)$. Since $\mathcal{T}$ is convex, nonempty, and closed, $\phi^*_{\mathcal{T}}$ is just the characteristic function of $\mathcal{T}$ (Exercise III.10), that is,

$$\underline{L}(\lambda,\Lambda) = \left\{\begin{array}{ll}\mathrm{Tr}(A\Lambda) & , \ell(\lambda,\Lambda) \in \mathcal{T}\\ -\infty & , \text{otherwise}\end{array}\right\}$$

so that the cone constrained Lagrange dual of ($a$), which is the problem of maximizing $\underline{L}$ over the set $\{\lambda \geq 0, \Lambda \succeq 0\}$ is equivalent to ($b$). □

**Exercise IV.28**. What follows is the concluding exercise in the "Truss Topology Design" series. We have already used TTD problem to present instructive "real life" illustrations of the power of several results of Convex Analysis, specifically, Caratheodory Theorem (Exercise I.18), epigraph description of convexity and Helly Theorem (Exercise III.9) and $\mathcal{S}$-Lemma (Exercise IV.11), not speaking about the Schur Complement Lemma which was instrumental in all these exercises. Now it is time to illustrate the power of conic duality.

In the sequel, we assume that the reader is reasonably well acquainted with Truss Topology Design story as told in Exercise I.16 and use without additional comments the notions, notation, and the results presented in this Exercise, including the default assumption $\mathfrak{R}$ which remains in force below. In addition, we assume from now on that the load of interest $f$ is nonzero – this is the only nontrivial case in TTD.

Recall that the TTD problem as posed in Exercise I.16.2 reads

$$\mathrm{Opt} = \min_{\tau,r}\left\{\tau : \left[\begin{array}{c|c}B\,\mathrm{Diag}\{t\}B^\top & f\\ \hline f^\top & 2\tau\end{array}\right] \succeq 0, t \geq 0, \sum_i t_i = W\right\} \qquad (P)$$

In our present language, this is a semidefinite program, and we know from Exercise I.16 that this problem is solvable.

Your first task is easy:

1. Build the semidefinite dual of $(P)$ and prove that the dual problem is solvable with the same optimal value Opt as the primal problem $(P)$.

Since passing from a semidefinite problem to its dual is a purely mechanical process, on one hand, and the subsequent tasks will be formulated in terms of the dual problem, here is the dual as given by Conic Duality:

$$\max_{V,g,\theta,\lambda,\mu}\left\{-2f^\top g-W\mu: 2\theta=1, \mathfrak{b}_i^\top V\mathfrak{b}_i+\lambda_i-\mu=0\,\forall i, \lambda\geq 0, \left[\begin{array}{c|c}V & g\\\hline g^\top & \theta\end{array}\right]\succeq 0\right\}$$

Eliminating variable $\theta$ (which is fixed by the corresponding constraint), we rewrite the dual as

$$\max_{V,g,\lambda,\mu}\left\{-2f^\top g-W\mu: \mathfrak{b}_i^\top V\mathfrak{b}_i+\lambda_i-\mu=0\,\forall i, \lambda\geq 0, \left[\begin{array}{c|c}V & g\\\hline g^\top & \frac{1}{2}\end{array}\right]\succeq 0\right\} \tag{$D$}$$

What is left to you, is to verify the derivation and to prove that $(D)$ is solvable with the same optimal value Opt as $(P)$.

*Solution:* Assumption $\mathfrak{R}$ states that every $t>0$ satisfying the linear equality $\sum_i t_i = W$ results in positive definite matrix $B\,\mathrm{Diag}\{t\}B^\top$, implying by the Schur Complement Lemma that augmenting $t$ with large enough $\tau$, we get a feasible solution to $(P)$ which strictly satisfies all $\geq$- and $\succeq$-constraints of $(P)$. Thus, $(P)$ is essentially strictly feasible (and of course bounded – the objective is nonnegative on the feasible set, not speaking about already known to us solvability of $(P)$). Applying Conic Duality Theorem, we conclude that $(D)$ is solvable with the same optimal value Opt as the primal problem $(P)$. □

Your next task still is easy:

2. Verify that eliminating, by partial optimization, variables $V$ and $\lambda$, problem $(D)$ reduces to the problem

$$\max_{g,\mu}\left\{-2f^\top g-W\mu: \left[\begin{array}{c|c}\mu & \mathfrak{b}_i^\top g\\\hline \mathfrak{b}_i^\top g & \frac{1}{2}\end{array}\right]\succeq 0\,\forall i\right\} \tag{$\overline{D}$}$$

and the latter problem is solvable with the same optimal value Opt as $(P)$ and $(D)$.

Pay attention to the first surprising fact: semidefinite constraints in $(\overline{D})$ involve the cone $\mathbf{S}^2_+$ of $2\times 2$ positive semidefinite matrices, and this cone, as we know, is, up to one-to-one linear transformation, just the Lorentz cone $\mathbf{L}^3$. Thus, $(\overline{D})$ is a conic quadratic problem.

*Solution:* Eliminating variables $\lambda_i$ is immediate – all we need is to replace the linear equality constraints $\mathfrak{b}_i^\top V\mathfrak{b}_i+\lambda_i-\mu=0$ with inequality constraints

$$\mathfrak{b}_i^\top V\mathfrak{b}_i\leq\mu,\ i\leq N,$$

reducing $(D)$ to the problem

$$\max_{V,g,\mu}\left\{-2f^\top g-W\mu: \underbrace{\mathfrak{b}_i^\top V\mathfrak{b}_i\leq\mu\,\forall i,}_{(*)} \left[\begin{array}{c|c}V & g\\\hline g^\top & \frac{1}{2}\end{array}\right]\succeq 0\right\} \tag{$D'$}$$

Next, by the Schur Complement Lemma, semidefinite constraint in $(D')$ is equivalent to the constraint $V\succeq\overline{V}:=2gg^\top$, and replacing $V$ with $\overline{V}$, we clearly preserve validity of constraints $(*)$. It follows that if $(V,g,\mu)$ is feasible for $(D')$, so is $(\overline{V},g,\mu)$. As a result, $(D')$ is equivalent to the problem

$$\max_{g,\mu}\left\{-2f^\top g-W\mu: 2(\mathfrak{b}_i^\top g)^2\leq\mu\,\forall i\right\}. \tag{$D''$}$$

Due to its origin, $(D'')$ is solvable along with $(D)$ and shares with $(D)$ and with $(P)$ the optimal value Opt. It remains to note that by the Schur Complement Lemma $(D'')$ is exactly the same as $(\overline{D})$. □

Your next task is

3. Pass from problem $(\overline{D})$ to its semidefinite dual $(\overline{P})$ and prove that the latter problem is solvable with optimal value $\mathrm{Opt}$.

At the first glance, the task seems crazy: the dual of the dual is the primal! Note, however, that $(\overline{D})$ is *not* the plain conic dual to $(P)$ problem $(D)$ – it is obtained from $(D)$ by eliminating part of variables, and nobody told us that this elimination keeps the dual to $(\overline{D})$ equivalent to the dual of $(D)$, that is, to $(P)$.

By the same reasons as in item 1, we take upon ourselves writing down $(\overline{P})$:

$$\min_{s,t,q}\left\{\frac{1}{2}\sum_i s_i : \sum_i t_i = W, \sum_i q_i\mathfrak{b}_i = f, \left[\begin{array}{c|c} t_i & q_i \\ \hline q_i & s_i\end{array}\right]\succeq 0\,\forall i\right\} \qquad (\overline{P})$$

What is left to you is to prove that $(\overline{P})$ is solvable with optimal value $\mathrm{Opt}$.

*Solution:* Problem $(\overline{D})$ clearly is strictly feasible, and we already know that it is solvable (and thus bounded) with optimal value Opt. By Conic Duality, $(\overline{P})$ is solvable with the same optimal value. □

Now – the main surprise:

4. Verify that $(\overline{P})$ allows eliminating, by partial optimization, variables $t_i$ and $s_i$, which reduces $(\overline{P})$ to solvable optimization problem

$$\min_q\left\{\frac{1}{2W}\left(\sum_i |q_i|\right)^2 : \sum_i q_i\mathfrak{b}_i = f\right\} \qquad (\#.1)$$

with the same optimal value $\mathrm{Opt}$ as all preceding problems, $(P)$ included.

This indeed is a great surprise – $(\#.1)$ is equivalent to *Linear Programming* problem

$$\min_q\left\{\|q\|_1 : \sum_i q_i\mathfrak{b}_i = f\right\}. \qquad (\#.2)$$

*Solution:* Let $s,t,q$ be a feasible solution to $(\overline{P})$, and let $I$ be the set of indexes $i$ with nonzero $q_i$; note that $I\neq\varnothing$ since, as we have assumed from the very beginning, $f\neq 0$. Note that zeroing out $s_i$ and $t_i$ with $i\notin I$ and increasing somehow $t_i$ with $i\in I$ to keep $\sum_i t_i$ intact, we preserve feasibility and do not spoil the value of the objective. In the resulting feasible solution $q,t',s'$ we have $t'_i=s'_i=0$, $i\notin I$, $t'_i>0$ for $i\in I$ (due to $\left[\begin{array}{c|c} t_i & q_i \\ \hline q_i & s_i\end{array}\right]\succeq 0$) and $s'_i\geq q_i^2/t'_i$ for $i\in I$ (Schur Complement Lemma); when replacing in $s'$ entries with indexes from $I$ with $q_i^2/t'_i$, we again preserve feasibility and do not spoil the objective. The bottom line is that partial optimization over $s,t$-components of a feasible solution $(q,t,s)$ reduces to solving the optimization problem

$$\min_{t_i, i\in I}\left\{\frac{1}{2}\sum_{i\in I} q_i^2/t_i : t_i>0, i\in I, \sum_{i\in I} t_i = W\right\}$$

This problem is easy to solve (see Exercise III.28); its optimal solution is given by

$$t_i = W|q_i|/\sum_{j\in I}|q_j|,\ i\in I,$$

and optimal value is $\frac{1}{2}W^{-1}(\sum_{i\in I}|q_i|)^2$. Thus, problem $(\overline{P})$ reduces to the optimization problem

$$\min_q\left\{\frac{1}{2W}(\sum_i |q_i|)^2 : \sum_i q_i\mathfrak{b}_i = f\right\}.$$

As follows from our analysis, the latter problem is solvable with optimal value Opt. □

The challenge is, of course, to extract from optimal solution to $(\#.2)$ an optimal truss $t^*$ – one with total bar volume $W$ and compliance, w.r.t. load $f$, equal to $\mathrm{Opt}$, and this is your final task:

5.1. Prove the following

**Observation** Let $t \geq 0$ be a nontrivial ($t \neq 0$) truss and $I = \{i : t_i > 0\}$. Consider the convex optimization problem

$$\min_q \left\{ \frac{1}{2}\sum_{i\in I} q_i^2/t_i : q_i = 0, i \notin I, \sum_i q_i \mathfrak{b}_i = f \right\} \qquad (\#.3)$$

and assume that the problem is feasible. Then

1. The problem is solvable
2. A feasible solution $q$ to the problem is optimal iff for some nodal displacement $v \in \mathcal{V}$ one has
$$q_i = t_i \mathfrak{b}_i^\top v \, \forall i \qquad (\#.4)$$
3. The optimal value in the problem is nothing but the compliance of truss $t$ w.r.t. load $f$.

*Solution:* Solvability of (#.3) is evident - the problem is feasible with bounded sublevel sets of the objective. By optimality conditions in convex minimization under linear equality constraints (see the second example after Proposition III.15.3) a feasible solution $q$ is optimal iff for some $v \in \mathbf{R}^M$ one has

$$q_i/t_i = \mathfrak{b}_i^\top v, i \in I,$$

which is the same as (#.4). Assuming $q$ optimal, (#.4) combines with $\sum_i q_i \mathfrak{b}_i = f$ to imply that

$$\sum_i t_i \mathfrak{b}_i \mathfrak{b}_i^\top v = f.$$

We see that $v$ is the equilibrium displacement of truss $t$ loaded by $f$, implying that the compliance of this truss under the load $f$ is (see Exercise I.16.1)

$$\begin{array}{l} \frac{1}{2} v^\top f = \frac{1}{2}\sum_i t_i (\mathfrak{b}_i^\top v)^2 \\ = \frac{1}{2}\sum_{i\in I} t_i (\mathfrak{b}_i^\top v)^2 \text{ [since } t_i = 0 \text{ for } i \notin I] \\ = \frac{1}{2}\sum_{i\in I} q_i^2/t_i \text{ [since } \mathfrak{b}_i^\top v = q_i/t_i, \text{ for } i \in I] \end{array}$$

and the concluding quantity is the optimal value of (#.3). □

5.2. Extract from optimal solution to (#.2) an optimal truss.

*Solution:* From our preceding considerations (#.1) is solvable with the same optimal value Opt as $(P)$ and $(\overline{P})$ and is obtained from $(\overline{P})$ by partial optimization in $s, t$-variables. Let $q^*$ be an optimal solution to (#.2), or, which is the same, to (#.1). Due to the origin of (#.1), the value Opt of its objective at $q^*$ satisfies

$$\mathrm{Opt} = \min_{s,t} \left\{ \frac{1}{2}\sum_i s_i : \sum_i t_i = W, \left[\begin{array}{c|c} t_i & q_i^* \\ \hline q_i^* & s_i \end{array}\right] \succeq 0 \, \forall i \leq N \right\}$$

and we know what is an optimal solution $s^*, t^*$ to the right hand side problem: setting $I = \{i : q_i^* \neq 0\}$, we have

$$t_i^* = \left\{ \begin{array}{ll} 0, & i \notin I \\ W\frac{|q_i^*|}{\sum_{j\in I}|q_j^*|}, & i \in I \end{array} \right., s_i^* = \left\{ \begin{array}{ll} 0, & i \notin I \\ \frac{(q_i^*)^2}{t_i^*}, & i \in I \end{array} \right. \qquad (\#.4)$$

Thus,

$$\mathrm{Opt} = \frac{1}{2}\sum_{i\in I} [q_i^*]^2/t_i^* \qquad (!)$$

Now consider the optimization problem (#.3) stemming from $t = t^*$. $q^*$ is a feasible solution to this problem with the value of the objective Opt (by (!)). By Observation, the optimal value in this problem is the compliance of $t^*$ w.r.t. $f$, and since the total bar volume of $t^*$ is $W$, this optimal value is $\geq$ Opt due to the origin of Opt. Thus, $q^*$ is a feasible solution to the stemming from $t = t^*$problem (#.3), the value of the problem's objective at this solution is Opt, and the optimal value in the problem is $\geq$ Opt.

We conclude that $q^*$ is an optimal solution to the problem in question with the value of the objective Opt, implying by Observation that Opt is the compliance of truss $t^*$ w.r.t. $f$. Recalling that Opt is the optimal value in $(P)$ and the total bar volume of $t^*$ is $W$, we conclude that $t^*$ is the $t$-component of an optimal solution of $(P)$. □

**Explanation of LP miracle.** Problem $(\#.1)$ was obtained from problem $(\overline{P})$ by eliminating $t$- and $s$-variables. When eliminating in $(\overline{P})$ $s$-variables only, we arrive at the problem

$$\min_{q,t}\left\{\sum_i \frac{q_i^2}{2t_i} : t \geq 0, \sum_i t_i = W, \sum_i q_i \mathfrak{b}_i = f\right\} \qquad (\widetilde{P})$$

where, by definition, $\frac{q_i^2}{t_i}$ is $0$ when $q_i = 0$ and $+\infty$ otherwise. LP reformulation of the problem is an immediate consequence of formulation $(\widetilde{P})$. The question we address here is: can we derive $(\widetilde{P})$ directly from the first principles of Mechanics (as was the case with our initial TTD problem $(P)$), thus avoiding twice passing to dual which led us from $(P)$ to $(\widetilde{P})$? As we shall see in a while, the answer is both "yes" and "no."

To interpret $(\widetilde{P})$ in terms of Mechanics, we need first of all to interpret in this way the decision variables of the problem. Looking at $(\widetilde{P})$, we can guess that $t$ plays the role of a tentative truss; at least the constraints on $t$ are exactly those imposed on a truss with total bar volume $W$. To interpret $q$, consider a displacement $v$ of nodes in truss $t$. As we remember from the derivation of the TTD model in the preamble to Exercise I.16, the vector

$$-\sum_i [t_i \mathfrak{b}_i^\top v]\mathfrak{b}_i$$

is the reaction (block-vector of reaction forces at different nodes) resulting from nodal displacement $v$, and

$$t_i \mathfrak{b}_i^\top v = -S_i \delta_i, \qquad (\#.5)$$

where $S_i$ is the cross-sectional size of $i$-th bar, and $\delta_i$ is the change in the bar's length caused by the displacement $v$ of the nodes[13]. Recall that by Hooke's Law the *tension* in a bar of (pre-deformation) length $d$ and cross-sectional size $S$ caused by elongation/shortening of the bar by $\delta$ (that is, the reaction force caused by this deformation at bar's endpoint) is $-S\delta/d$, so that the quantities $t_i \mathfrak{b}_i^\top v$ admit, according to $(\#.5)$, transparent mechanical interpretation – these are *scaled tensions*, products of (pre-deformation) bar lengths and tensions in bars of truss $t$ caused by displacement $v$ of the nodes. Moreover, Mechanics says that the potential energy capacitated in elastic bar of length $d$ and cross-sectional size $S$ as a result of bar's elongation/shortening by $\delta$ is $\frac{1}{2}S\delta^2/d$. It follows that given a truss $t$ and a nodal displacement $v$ and setting $q_i = t_i \mathfrak{b}_i^\top v$, the reaction of the truss caused by nodal displacement $v$ is $-\sum_i q_i \mathfrak{b}_i$, and the potential energy capacitated in the truss as a result of nodal displacement $v$ is $\frac{1}{2}\sum_i q_i^2/t_i$. We see that when $t$ is a truss, and vector $q$ is linked to $t$ and to some nodal displacement $v$ by the relations

$$q_i = t_i \mathfrak{b}_i^\top v \qquad (\#.6)$$

then $q_i$, $-\sum_i q_i \mathfrak{b}_i$ and $\frac{1}{2}\sum_i q_i^2/t_i$ are, respectively, the scaled tensions, the reaction, and the potential energy capacitated in the truss as a result of displacement $v$ of its nodes. Consequently, if $(q,t)$ is a feasible solution to $(\widetilde{P})$ *and* $q, t$ *and some nodal displacement* $v$ *are linked by* $(\#.6)$, then $v$ is the equilibrium displacement of truss $t$ under load $f$, and the value of the objective of $(\widetilde{P})$ at the feasible solution $(q,t)$ is the compliance of truss $t$ w.r.t. the load $f$.

Our observations suggest the following mechanical interpretation of candidate solutions to $(\widetilde{P})$: $t_i$ are bar volumes, and $q_i$ are scaled tensions in bars. With this interpretation, the linear constraints $\sum_i q_i \mathfrak{b}_i = f$ say that the reaction compensates the external load, and the value of the objective at a feasible solution $(q,t)$ is the compliance of truss $t$ w.r.t load $f$, so that $(\widetilde{P})$ indeed is the problem of minimizing, over trusses of total volume $W$, the compliance of the truss w.r.t. load $f$. Unfortunately, this mechanical interpretation of $(\widetilde{P})$ is *completely wrong*. Indeed, the dimension of vector $q$ is $N$, and typically it is much larger than the dimension $M$ of those vectors $q$ which could be linked to $M$-dimensional nodal displacements $v$ according to $(\#.6)$ (think about Console design where $N = 3024$ and $M = 144$). In order for our guessed mechanical interpretation of $(\widetilde{P})$ to make sense, $(\widetilde{P})$ should include additional constraints stating that $q$ is linked to $t$ and some nodal displacement $v$ by relations $(\#.6)$, but $(\widetilde{P})$ does *not* include this sine qua non, from the viewpoint of Mechanics, restriction! As a result, "most" of feasible solutions to $(\widetilde{P})$ make no mechanical sense – what pretends to be the vector of scaled tensions does *not* come from any deformation of the truss! Note that a straightforward attempt to include into the problem the above sine qua non restriction by adding to the design variables $t$, $q$ additional design variables $v$, and to the constraints – equality constraints $(\#.6)$ , fails – it recovers "mechanical validity" of $(\widetilde{P})$ at the disastrous, from the computational viewpoint, price – constraints $(\#.6)$ are *nonconvex* in the design variables $q, t, v$!

All this being said, how happens that $(\widetilde{P})$ does allow to recover the optimal truss? The explanation is: *at the optimum*, $q$ and $t$ indeed are linked by relations $(\#.6)$ with certain nodal displacement $v$; this displacement stems from the Lagrange multipliers certifying optimality of $q, t$ (look at the justification of Observation from item 5.1). Thus, $(\widetilde{P})$ can be treated as *precise relaxation* of the "true" TTD problem: formulating the latter problem in terms of scaled tensions, bar volumes, *and* nodal displacements, which is fully legitimate from the viewpoint of Mechanics, we then relax the problem by throwing

[13] All this corresponds to the Hooke's Law in the form "reaction force caused by elongation/shortening by $\delta$ of bar with length $d$ and cross-sectional size $S$ is $-S\delta/d$" – the form corresponding to the linearly elastic model of truss's deformation.

away variables $v$ and constraints $(\#.6)$, thus arriving at problem $(\widetilde{P})$. This relaxation is precise in the sense that the optimal solution to the relaxed problem provably is the $(q,t)$-component of optimal solution to the "true" TTD problem in variables $q,t,v$.

Finally, we remark that while the "LP miracle" stemming from $(\widetilde{P})$ has rather restricted scope – it disappears when passing from single-load TTD with the simplest possible constraints on tentative $t$'s to more general problems of structural design (multi-load TTD, Shape Design, etc.), these more general problems still admit "precise relaxations" of type $(\widetilde{P})$, see [BTN], and one arrives at these reformulations by strategy similar to the one we have used – start with the "natural" conic formulation $(P)$ of the problem, pass to the conic dual $(D)$ of $(P)$, process $(D)$ "on paper" by eliminating variables allowing for easy elimination, and end up by passing from the resulting reformulation $(\overline{D})$ of $(D)$ to the conic dual $(\overline{P})$ of $(\overline{D})$.

## Cone-convexity

**Exercise IV.29.** [elementary properties of cone-convex functions] The goal of this Exercise is to extend elementary properties of convex functions onto cone-convex mappings.

**A.** Let $\mathcal{X},\mathcal{Y}$ be Euclidean spaces equipped with norms $\|\cdot\|_{\mathcal{X}}$, $\|\cdot\|_{\mathcal{Y}}$. Let, next, $\mathbf{X}$ be a closed pointed cone in $\mathcal{X}$, $\mathbf{Y}$ be a closed *and pointed* cone in $\mathcal{Y}$, and $f: X\to\mathcal{Y}$ be a mapping defined on a nonempty convex set $X\subset\mathcal{X}$. Recall that for a closed and pointed cone $\mathbf{K}$ in Euclidean space $\mathcal{K}$ and $x,x'\in\mathcal{K}$, relation $x\leq_{\mathbf{K}} x'$, same as $x'\geq_{\mathbf{K}} x$, means that $x'-x\in\mathbf{K}$.

Recall that $f$ is called

- $(\mathbf{X},\mathbf{Y})$-monotone on $X$, if
$$\left\{x,x'\in X \text{ and } x\leq_{\mathbf{X}} x'\right\} \implies f(x)\leq_{\mathbf{Y}} f(x');$$
- $\mathbf{Y}$-convex on $X$, if
$$f(\lambda x+(1-\lambda)x')\leq_{\mathbf{Y}} \lambda f(x)+(1-\lambda)f(x')$$

for every $x,x'\in X$ and $\lambda\in[0,1]$.

For example,

— an affine mapping $f(x)=Ax+a:\mathcal{X}\to\mathcal{Y}$ is $\mathbf{Y}$-convex, whatever be pointed closed cone $\mathbf{Y}$;

— when $\mathcal{Y}=\mathbf{R}$ and $\mathbf{Y}=\mathbf{R}_+$, $\mathbf{Y}$-convex on $X$ functions are the convex, in the standard definition, real-valued functions on $X$.

A.1. In the situation of $\mathbf{A}$, let $\mathbf{Y}^*$ be the cone dual to $\mathbf{Y}$. For $e\in\mathcal{Y}$, let $f_e(x)=\langle e,f(x)\rangle_{\mathcal{Y}}:X\to\mathbf{R}$. Prove that $f$ is
— $\mathbf{Y}$-convex on $X$ iff the function $f_e$ is convex on $X$ whenever $e\in\mathbf{Y}^*$
— $(\mathbf{X},\mathbf{Y})$-monotone on $X$ iff the function $f_e$ is $\mathbf{X}$-monotone on $X$ (i.e., $x,x'\in X, x\leq_{\mathbf{X}} x' \implies f_e(x)\leq f_e(x')$) for every $e\in\mathbf{Y}^*$.

*Solution:* Evident due to the fact that $y\in\mathbf{Y}$ iff $\langle e,y\rangle\geq 0$ for all $e\in\mathbf{Y}^*$; indeed, the cone $\mathbf{Y}$ is closed and therefore is dual to $\mathbf{Y}^*$.

A.2. In the situation of $\mathbf{A}$, let $f$ be $\mathbf{Y}$-convex. Prove that $f$ is locally bounded and locally Lipschitz continuous on the interior of $X$, meaning that if $\bar{X}\subset\operatorname{int}X$ is a closed and bounded set, then there exists $M<\infty$ such that $\|f(x)\|_{\mathcal{Y}}\leq M$ holds for all $x\in\bar{X}$ (this is local boundedness) and there exists $L<\infty$ such that $\|f(x)-f(z')\|_{\mathcal{Y}}\leq L\|x-x'\|_{\mathcal{X}}$ holds for all $x,x'\in\bar{X}$ (this is local Lipschitz continuity).

*Solution:* Since $\mathbf{Y}$ is pointed closed cone, the cone $\mathbf{Y}^*$ has a nonempty interior. Selecting once for ever $N:=\dim\mathcal{Y}$ linearly independent vectors $e^1,...,e^N$ in $\operatorname{int}\mathbf{Y}^*$, let us set $y^i:=\langle y,e^i\rangle_{\mathcal{Y}}$. Then, the linear mapping $y\mapsto\bar{y}(y):=[y^1;...;y^N]$ is a one-to-one linear map from $\mathcal{Y}$ onto $\mathbf{R}^N$, so that the function $|y|_\infty:=\|\bar{y}(y)\|_\infty$ is a norm on $\mathcal{Y}$. By A.1, the real valued functions $f_{e^i}(x)$ are convex on $X$, and therefore are locally bounded and locally Lipschitz continuous on $\operatorname{int}X$, $1\leq i\leq N$, implying similar properties of $f$ w.r.t. $|\cdot|_\infty$ on $\mathcal{Y}$, and therefore w.r.t. $\|\cdot\|_{\mathcal{Y}}$. □

**B.** Now let us look at elementary operations preserving cone convexity. From now on, $\operatorname{Lin}(\mathcal{X},\mathcal{Y})$ denotes the linear space of linear mappings acting from Euclidean space $\mathcal{X}$ to Euclidean space $\mathcal{Y}$. Prove the following statements:

B.1. ["nonnegative linear combinations"] Let $X$ be a nonempty convex subset of Euclidean space $\mathcal{X}$, $\mathcal{Y}_j$, $j \leq J$, and $\mathcal{Y}$ be Euclidean spaces equipped with pointed closed cones $\mathbf{Y}_j$, $\mathbf{Y}$, and $\alpha_j \in \mathrm{Lin}(\mathcal{Y}_j, \mathcal{Y})$ be "nonnegative coefficients", meaning that $\alpha_j y_j \in \mathbf{Y}$ whenever $y_j \in \mathbf{Y}_j$. When mappings $f_j(x) : X \to \mathcal{Y}_j$. are $\mathbf{Y}_j$-convex, $j \leq J$, their "linear combination with coefficients $\alpha_j$" – the mapping

$$f(x) = \sum_j \alpha_j f_j(x) : X \to \mathcal{Y}$$

– is $\mathbf{Y}$-convex.

*Solution:* For $x, x' \in X$ and $\lambda \in [0,1]$ we have

$$\lambda f(x) + (1-\lambda) f(x') - f(\lambda x + (1-\lambda)x') = \sum_j \alpha_j \underbrace{[\lambda f_j(x) + (1-\lambda) f_j(x') - f_j(\lambda x + (1-\lambda)x')]}_{\in \mathbf{Y}_j} \geq_{\mathbf{Y}} 0,$$

where the concluding $\geq_{\mathbf{Y}}$ is due to $\alpha_j y_j \geq_{\mathbf{Y}} 0$ whenever $y_j \geq_{\mathbf{Y}_j} 0$. □

B.2. [affine substitution of variables] In the situation of $\mathbf{A}$, let $z \mapsto Az + a : \mathcal{Z} \to \mathcal{X}$ be an affine mapping, and let $f$ be $\mathbf{Y}$-convex on $X$. Then, the function $g(z) := f(Az + a)$ is $\mathbf{Y}$-convex on the set $Z = \{z : Az + a \in X\}$.

*Solution:* evident.

B.3. [monotone composition] Let $\mathcal{U}_j$, $j \leq J$, be Euclidean spaces equipped with closed pointed cones $\mathbf{U}_j$, let $\mathcal{U} = \mathcal{U}_1 \times ... \times \mathcal{U}_J$, $\mathbf{U} = \mathbf{U}_1 \times ... \times \mathbf{U}_J$, and let $\mathcal{Y}$ be an Euclidean space equipped with closed pointed cone $\mathbf{Y}$. Next, let $X$ be nonempty convex set in Euclidean space $\mathcal{X}$, $U$ be a nonempty convex set in $\mathcal{U}$, let $f_j(x) : X \to \mathcal{U}_j$ be $\mathbf{U}_j$-convex functions, $j \leq J$, such that $f(x) = [f_1(x); ...; f_J(x)] \in U$ whenever $x \in X$. Finally, let mapping $F : U \to \mathcal{Y}$ be $(\mathbf{U}, \mathbf{Y})$-monotone and $\mathbf{Y}$-convex on $U$. Then the composition

$$G(x) = F(f(x)) : X \to \mathcal{Y}$$

is $\mathbf{Y}$-convex on $X$.

*Solution:* Indeed, when $x, x' \in X$ and $\lambda \in [0,1]$, setting $\bar{x} = \lambda x + (1-\lambda)x'$, $u = f(x)$, $u' = f(x')$, $\bar{u} = f(\bar{x})$, we get $\bar{x} \in X$, $u, u', \bar{u} \in U$ (since $f$ maps $X$ into $U$) and $\bar{u} \leq_{\mathbf{U}} \widehat{u} := \lambda u + (1-\lambda)u'$ due to $\mathbf{U}_j$-convexity of $f_j$ and the origin of $\mathbf{U}$. Consequently,

$$\begin{aligned} G(\bar{x}) = F(\bar{u}) &\leq_{\mathbf{Y}} F(\widehat{u}) \text{ [since } \bar{u}, \widehat{u} \in U,\ \bar{u} \leq_{\mathbf{U}} \widehat{u} \text{ and } F \text{ is } (\mathbf{U},\mathbf{Y})\text{-monotone]} \\ &\leq_{\mathbf{Y}} \lambda F(u) + (1-\lambda)F(u') \text{ [since } F \text{ is } \mathbf{Y}\text{-convex on } U] \\ &= \lambda G(x) + (1-\lambda)G(x'). \quad \square \end{aligned}$$

$\mathbf{C.}$ The gradient inequality and existence of directional derivative can be extended from the usual convex functions (i.e., $\mathbf{R}_+$-convex functions taking values in $\mathbf{R}$) to the cone-convex ones. Prove the following statements:

C.1. ["gradient inequality"] In the situation of $\mathbf{A}$, let $\bar{x} \in X$ and $f$ be $\mathbf{Y}$-convex on $X$ and differentiable at $\bar{x}$. Then

$$\forall y \in X : f(y) \geq_{\mathbf{Y}} f(\bar{x}) + f'(\bar{x})(y - x),$$

where $f'(\bar{x})$ is the Jacobian of $f$ at $\bar{x}$.

*Solution:* it suffices to apply the standard gradient inequality to convex functions $f_e$, $e \in \mathbf{Y}^*$, and use the same argument as when processing A.1. □

C.2. [existence of directional derivative] In the situation of $\mathbf{A}$, let $f$ be $\mathbf{Y}$-convex on $X$, let $\bar{x} \in \mathrm{int}\, X$ and $d \in \mathcal{X}$. Then

$$\exists Df(\bar{x})[d] := \lim_{t \to +0} \frac{f(\bar{x} + td) - f(\bar{x})}{t}$$

and

$$(t \geq 0\ \&\ \bar{x} + td \in X) \implies f(\bar{x} + td) \geq_{\mathbf{Y}} f(\bar{x}) + tDf(\bar{x})[d]. \qquad (\#)$$

Besides this, as a function of $d \in \mathcal{X}$, $Df(\bar{x})[d]$ is positively homogeneous of degree 1 (i.e., $Df(\bar{x})[td] = tDf(\bar{x})[d]$ when $t \geq 0$) and $\mathbf{Y}$-convex.

*Solution:* By arguments completely similar to those used when justifying A.1-3, this is immediate consequence of the standard results on directional derivatives of the usual convex functions, see section 16.3.

**D.** Subdifferentials of the usual convex functions admit natural extensions to the cone-convex mappings. Specifically, in the situation of **A**, let $\bar{x} \in X$. Let us say that $g \in \mathrm{Lin}(\mathcal{X}, \mathcal{Y})$ is a *sub-Jacobian* of $f$ at $\bar{x}$, if

$$\forall y \in X : f(y) \geq_{\mathbf{Y}} f(\bar{x}) + g[y - x].$$

For example, C.1 says that if $f$ is $\mathbf{Y}$-convex on $X$ and differentiable at $\bar{x} \in X$, then the taken at $x$ Jacobian $f'(\bar{x})$ of $f$ is a sub-Jacobian of $f$ at $\bar{x}$. Clearly, for a usual convex function its sub-Jacobians at a point are exactly the linear forms on $\mathcal{X}$ given by subgradients $f'(x)$ of $f$ at $x$ according to

$$gh = \langle f'(x), h \rangle_{\mathcal{X}},\ h \in \mathcal{X}.$$

Let $\mathcal{J}f(x)$ be the set of all sub-Jacobians of $f$ at $x \in X$. Prove the following statements:

D.1. In the situation of **A**, for $x \in X$ one has $g \in \mathcal{J}f(x)$ iff for every $e \in \mathbf{Y}^*$ the vector $g^*e \in \mathcal{X}$ is a subgradient of $f_e$ at $x$; here for $g \in \mathrm{Lin}(\mathcal{X}, \mathcal{Y})$, $g^* \in \mathrm{Lin}(\mathcal{Y}, \mathcal{X})$ is the conjugate of $g$: $\langle gu, v \rangle_{\mathcal{Y}} = \langle u, g^*v \rangle_{\mathcal{X}}$ for all $u \in \mathcal{X}$, $v \in \mathcal{Y}$.

*Solution:* Evident due to the same argument as used when processing A.1. □

D.2. In the situation of **A**, let $f$ be $\mathbf{Y}$-convex on $X$. Then
— D.2.1. For every $x \in X$, the set $\mathcal{J}f(x)$ is a closed convex subset of $\mathrm{Lin}(\mathcal{X}, \mathcal{Y})$;
— D.2.2. The mapping $x \mapsto \mathcal{J}f(x)$ is locally bounded on the interior of $X$, that is, for every closed and bounded set $\bar{X} \subset \mathrm{int}\, X$, the induced norms $\|g\|_{\mathcal{X},\mathcal{Y}} = \max_z \{\|gz\|_{\mathcal{Y}} : \|z\|_{\mathcal{X}} \leq 1\}$ of linear mappings $g \in \mathcal{J}f(x)$, $x \in \bar{X}$ are bounded away from $+\infty$;
— D.2.3. The multivalued mapping $x \mapsto \mathcal{J}f(x)$ is closed on $\mathrm{int}\, X$: if $x_i \in \mathrm{int}\, X$ converge as $i \to \infty$ to $\bar{x} \in \mathrm{int}\, X$ and linear mappings $g_i \in \mathcal{J}f(x_i)$ converge as $i \to \infty$ to some $\bar{g} \in \mathrm{Lin}(\mathcal{X}, \mathcal{Y})$, then $\bar{g} \in \mathcal{J}f(\bar{x})$.

*Solution:* D.2.1 is immediate consequence of the fact that $\mathbf{Y}$ is a closed cone; D.2.2-3 are readily given by local Lipschitz continuity of $f$ on $\mathrm{int}\, X$, see A.2. □

The most attractive property of subgradients of the usual convex function is their existence, at least at interior points of the function's domain. This fact extends to the cone-convex mappings. Prove the following statements:

D.3. [existence of sub-Jacobians] In the situation of **A**, let $\bar{x} \in \mathrm{int}\, X$ and $f$ be $\mathbf{Y}$-convex on $X$. Then $\mathcal{J}f(\bar{x})$ is nonempty.

*Solution:* This is the only claim which seemingly cannot be extracted more or less automatically from standard facts about the usual convex functions. Moreover, the Separation Theorem underlying the existence of subgradients of the usual convex functions at interior points of their domains seemingly does not help now. Fortunately, there is an easily implementable alternative as follows.

For $\epsilon > 0$, let $X_\epsilon = \{x \in X : \|y - x\|_{\mathcal{X}} \leq \epsilon \implies y \in \mathrm{int}\, X\}$ and $\delta_\epsilon(x)$ be a nonnegative $\mathrm{C}^\infty$ function such that $\delta_\epsilon(x) = 0$ when $\|x\|_{\mathcal{X}} \geq \epsilon$ and $\int_{\mathcal{X}} \delta_\epsilon(x)dx = 1$. Clearly, for small $\epsilon > 0$ $X_\epsilon$ is a nonempty open convex set, and the function

$$f_\epsilon(x) := \int_{\mathcal{X}} f(x - y)\delta_\epsilon(y)dy$$

with the domain $X_\epsilon$ is well defined, continuously differentiable and $\mathbf{Y}$-convex on its domain. Besides this, for a convex compact set $\bar{X} \subset \mathrm{int}\, X$ such that $\bar{x} \in \mathrm{int}\, \bar{X}$ we have $\bar{X} \subset X_\epsilon$ for all small enough positive $\epsilon$, and for those $\epsilon$ the functions $f_\epsilon$ are uniformly in $\epsilon$ Lipschitz continuous on $\bar{X}$. From this latter observation it follows that the Jacobians $f'_\epsilon(\bar{x})$ are uniformly in $\epsilon$ bounded, which in turn implies that for a properly selected $\epsilon_i \to +0$, $i \to \infty$, the linear mappings $g_i := f'_{\epsilon_i}(\bar{x})$ converge as $i \to \infty$ to some $\bar{g} \in \mathrm{Lin}(\mathcal{X}, \mathcal{Y})$. Let us prove that $\bar{g} \in \mathcal{J}f(\bar{x})$, implying that $\mathcal{J}f(\bar{x})$ is nonempty. Indeed, in view of A.2 the functions $f^i := f_{\epsilon_i}$ converge as $i \to \infty$, uniformly on compact subsets of $\operatorname{int} X$, to $f$. Then, by C.1, we have

$$y \in X_{\epsilon_i} \implies f_i(y) \geq_{\mathbf{Y}} f(\bar{x}) + g_i(y - \bar{x}),$$

implying in view of the outlined convergencies that

$$f(y) \geq_{\mathbf{Y}} f(\bar{x}) + g[y - \bar{x}]\ \forall y \in \operatorname{int} X.$$

The only remaining task is to extend the latter relation from $y \in \operatorname{int} X$ to $y \in X$. Passing from $f : X \to \mathcal{Y}$ to $\overline{f}(x) := f(x) - [f(\bar{x}) + g[x - \bar{x}]]$, which, of course, is $\mathbf{Y}$-convex on $X$ along with $f$, we get

$$\overline{f}(y) \geq_{\mathbf{Y}} \overline{f}(\bar{x}) = 0,\ \forall y \in \operatorname{int} X, \qquad (!)$$

and what we need to prove is that $\overline{f}(y) \geq_{\mathbf{Y}} \overline{f}(\bar{x})$ for all $y \in Y$. Let $y \in X$. Using the definition of the directional derivative, we observe that (!) implies that $D\overline{f}(\bar{x})[y - \bar{x}] \geq_{\mathbf{Y}} 0$, whence by (#) one has $\overline{f}(y) \geq_{\mathbf{Y}} \overline{f}(\bar{x})$. □

For a real-valued convex function $f$ and $x \in \operatorname{int} \operatorname{Dom} f$, $d \in \mathcal{X}$, one has $Df(x)[d] = \max_{y \in \partial f(x)} \langle y, d\rangle_{\mathcal{X}}$. A similar fact holds true for cone-convex functions:

D.4. In the situation of $\mathbf{A}$, let $f$ be $\mathbf{Y}$-convex on $X$. Let also $\bar{x} \in \operatorname{int} X$ and $d \in \mathcal{X}$. Then for properly selected $g \in \mathcal{J}f(\bar{x})$ one has

$$Df(\bar{x})[d] = gd,$$

while for every $g' \in \mathcal{J}f(\bar{x})$ one has

$$Df(\bar{x})[d] \geq_{\mathbf{Y}} g'd.$$

*Solution:* For $g' \in \mathcal{J}f(\bar{x})$ and $t > 0$ such that $\bar{x} + td \in X$ we have

$$f(\bar{x} + td) - f(\bar{x}) \geq_{\mathbf{Y}} tg'd,$$

whence, dividing by $t$ and passing to limit as $t \to +0$, we get

$$Df(\bar{x})[d] \geq_{\mathbf{Y}} g'd.$$

On the other hand, let $t_0 > t_1 > t_2 > ... > 0$ be such that $x_i := \bar{x} + t_i d \in \operatorname{int} X$ and $t_i \to 0$, $i \to \infty$. By D.3, there exists $g_i \in \mathcal{J}f(x_i)$; by D.2.2 the sequence $g_i$ is bounded, so that passing to a subsequence, we can assume that $g_i \to g$ as $i \to \infty$; by D.2.3, $g \in \mathcal{J}f(\bar{x})$. Since $g_i \in \mathcal{J}f(x_i)$, we have

$$f(x_i) - f(\bar{x}) \leq_{\mathbf{Y}} g_i[x_i - \bar{x}],$$

whence

$$g_i d \geq_{\mathbf{Y}} t_i^{-1}[f(x_i) - f(\bar{x})].$$

As $i \to \infty$, the left hand side in this $\geq_{\mathbf{Y}}$-inequality tends to $gd$, and the right hand side to $Df(\bar{x})[d]$. Thus, $g'd \leq_{\mathbf{Y}} Df(\bar{x})[d]$ for all $g' \in \mathcal{J}f(\bar{x})$ and $gd \geq_{\mathbf{K}} Df(\bar{x})[d]$ for some $g \in Df(\bar{x})[d]$; in particular, $gd = Df(\bar{x})[d]$ (recall that $\mathbf{Y}$ is pointed). □

There is a natural relation between sub-Jacobians of $\mathbf{Y}$-convex function $f$ and subgradients of functions $f_e = \langle e, f\rangle_{\mathcal{Y}}$, $e \in \mathbf{Y}^*$:

D.5. In the situation of $\mathbf{A}$, let $f$ be $\mathbf{Y}$-convex on $X$ and $\bar{x} \in \operatorname{int} X$. For $e \in \mathbf{Y}^*$, $h \in \partial f_e(\bar{x})$ (that is, $f_e(y) \geq f_e(\bar{x}) + \langle h, y - \bar{x}\rangle_{\mathcal{X}}$ for all $y \in X$) iff $h = g^*e$ for some $g \in \mathcal{J}f(\bar{x})$.

*Solution:* In one direction: when $e \in \mathbf{Y}^*$ and $h = g^*e$ for $g \in \mathcal{J}f(\bar{x})$, we have for every $y \in X$:

$$f(y) \geq_{\mathbf{Y}} f(\bar{x}) + g[y - x] \implies \langle e, f(y)\rangle_{\mathcal{Y}} \geq \langle e, f(\bar{x})\rangle_{\mathcal{Y}} + \langle e, g[y - \bar{x}]\rangle_{\mathcal{Y}} \iff f_e(y) \geq f_e(\bar{x}) + \langle g^*e, y - \bar{x}\rangle_{\mathcal{X}}.$$

In the opposite direction: let $e \in \mathbf{Y}^*$ and $h \in \partial f_e(\bar{x})$. By D.2 and D.3, the set $\mathcal{J} = \mathcal{J}f(\bar{x})$ is a nonempty closed and bounded convex set in $\operatorname{Lin}(\mathcal{X}, \mathcal{Y})$. Thus, the set $\mathcal{I} := \{g^*e : g \in \mathcal{J}\}$ is a nonempty closed and bounded convex set in $\mathcal{X}$. Assume for contradiction that $h \notin \mathcal{I}$. Then, by Separation Theorem there exists $d \in \mathcal{X}$ such that

$$\langle h, d\rangle_{\mathcal{X}} > \max_{g\in\mathcal{J}} \langle g^* e, d\rangle_{\mathcal{X}} = \max_{g\in\mathcal{J}} \langle e, gd\rangle_{\mathcal{Y}}. \qquad (*)$$

As $h \in \partial f_e(\bar{x})$, for all small enough $t > 0$ we have

$$f_e(\bar{x} + td) - f_e(\bar{x}) \geq t\langle h, d\rangle_{\mathcal{X}},$$

whence $Df_e(\bar{x})[d] \geq \langle h, d\rangle_{\mathcal{X}}$. We clearly have $Df_e(\bar{x})[d] = \langle e, Df(\bar{x})[d]\rangle_{\mathcal{Y}}$, and we arrive at

$$\langle h, d\rangle_{\mathcal{X}} \leq \langle e, Df(\bar{x})[d]\rangle_{\mathcal{Y}}.$$

By D.4, we have $Df(\bar{x})[d] = \bar{g}d$ for some $\bar{g} \in \mathcal{J}f(\bar{x})$, so that $\langle h, d\rangle_{\mathcal{X}} \leq \langle e, \bar{g}d\rangle_{\mathcal{Y}}$, contradicting $(*)$. □

Finally, the chain rule:

D.6. [chain rule] Let $\mathcal{U}_j$, $j \leq J$, be Euclidean spaces equipped with closed pointed cones $\mathbf{U}_j$, let $\mathcal{U} = \mathcal{U}_1 \times ... \times \mathcal{U}_J$, $\mathbf{U} = \mathbf{U}_1 \times ... \times \mathbf{U}_J$, and let $\mathcal{Y}$ be an Euclidean space equipped with closed pointed cone $\mathbf{Y}$. Next, let $X$ be nonempty convex set in Euclidean space $\mathcal{X}$, $U$ be a nonempty convex set in $\mathcal{U}$, let $f_j(x) : X \to \mathcal{U}_j$ be $\mathbf{U}_j$-convex on $X$ functions, $j \leq J$, such that $f(x) = [f_1(x); ...; f_J(x)] \in U$ whenever $x \in X$. Finally, let mapping $F : U \to \mathcal{Y}$ be $(\mathbf{U}, \mathbf{Y})$-monotone and $\mathbf{Y}$-convex on $U$. As we know from B.3, the composition

$$G(x) = F(f(x)) : X \to \mathcal{Y}$$

is $\mathcal{Y}$-convex on $X$. Now let $\overline{x} \in \operatorname{int} X$, $\overline{u}_j = f_j(\overline{x})$ be such that $\overline{u} = [\overline{u}_1; ...; \overline{u}_J] \in \operatorname{int} U$. Finally, let $g_j \in \mathcal{J}f_j(\overline{x})$, $j \leq J$, and $g \in \mathcal{J}F(\overline{u})$. Then the linear mapping $[u_1; ...; u_J] \mapsto g[u_1; ...; u_J]$ is $(\mathbf{U}, \mathbf{Y})$-monotone, and the linear mapping

$$h \mapsto \widehat{g}h := g[g_1 h; ...; g_J h] : \mathcal{X} \to \mathcal{Y}$$

is sub-Jacobian of $G$ at $\overline{x}$.

*Solution:* Indeed, let $\overline{V}$ be a convex neighborhood of $\overline{x}$ such that the images of $\overline{V}$ under the mapping $f(\cdot)$ and under the linear mapping

$$\overline{f}(x) := f(\overline{x}) + [g_1[x - \overline{x}]; ...; g_J[x - \overline{x}]]$$

belong to $\operatorname{int} U$ (such a neighborhood exists due to $f(\overline{x}) = \overline{f}(\overline{x}) \in \operatorname{int} U$ combined with continuity of $\overline{f}$ (evident) and $f$ (by A.2) at $\overline{x}$). Let also $V^j$, $j = 1, ..., J$, be convex neighborhoods of origins in $\mathcal{U}_j$ such that $\overline{u} + V^1 \times ... \times V^J \subset U$. For $d = [d_1; ...; d_J]$ with $d_j \in V^j \cap \mathbf{U}_j$ for $j \leq J$ we have

$$-gd + F(\overline{u}) \leq_{\mathbf{Y}} F(\overline{u} - d),$$

whence $gd \geq_{\mathbf{Y}} 0$ by $(\mathbf{U}, \mathbf{Y})$-monotonicity of $F$. Thus, $gd \geq_{\mathbf{Y}} 0$ for all $d \in \mathbf{U}$ of small enough norm, implying that $gd \geq_{\mathbf{Y}} 0$ for all $d \in \mathbf{U}$, as claimed.

When $x \in \overline{V}$, we have $\overline{f}(x) \leq_{\mathbf{U}} f(x)$, since $g_j$ are sub-Jacobians of $f_j$ at $\overline{x}$. Due to the $(\mathbf{U}, \mathbf{Y})$-monotonicity of $F$, we conclude that

$$\begin{aligned} G(x) = F(f(x)) &\geq_{\mathbf{Y}} F(\overline{f}(x)) = F\left(\overline{u} + [g_1[x - \overline{x}]; ...; g_J[x - \overline{x}]]\right) \\ &\geq_{\mathbf{Y}} F(f(\overline{x})) + g[g_1[x - \overline{x}]; ...; g_J[x - \overline{x}]] \text{ [since } g \in \mathcal{J}F(\overline{u})] \\ &= G(\overline{x}) + \widehat{g}[x - \overline{x}]. \end{aligned}$$

Thus, for $x$ in a neighborhood $\overline{V}$ of $\overline{x} \in \operatorname{int} X$ we have

$$G(x) \geq_{\mathbf{Y}} G(\overline{x}) + \widehat{g}[x - \overline{x}].$$

It remains to prove that the latter relation holds true for all $x \in X$, not for only $x \in \overline{V}$. This can be done in the same way as when justifying D.3: the mapping $\overline{G}(x) := G(x) - [G(\overline{x}) + \widehat{g}[x - \overline{x}]] : X \to \mathcal{Y}$ which is $\mathbf{Y}$-convex along with $G$ satisfies

$$\overline{G}(x) \geq_{\mathbf{Y}} \overline{G}(\overline{x}) = 0 \qquad (!!)$$

for $x$ from a neighborhood of $\overline{x}$, and we want to prove that in fact (!!) holds true for all $x \in X$. Indeed, (!!) implies that for every $x \in X$ we have $D\overline{G}(\overline{x})[x-\overline{x}] \geq_{\mathbf{Y}} 0$, whence $\overline{G}(x) \geq_{\mathbf{Y}} \overline{G}(\overline{x}) = 0$ for $x \in X$ by (#), so that $G(x) \geq_{\mathbf{Y}} G(\overline{x}) + \widehat{g}[x-\overline{x}]$ for all $x \in X$. □

**Exercise IV.30.** Univariate function $f(x) = x^{-1/2} : \{x > 0\} \to \mathbf{R}$ is nonincreasing and convex, and $\nabla f(x) = -x^{-3/2}/2$, $x > 0$. Now let $P$ be $m \times n$ matrix of rank $m$.

1. Prove that the mapping $F(X) = [PXP^\top]^{-1/2} : \mathbf{S}^n_{++} \to \mathbf{S}^m$, where $\mathbf{S}^n_{++} = \operatorname{int} \mathbf{S}^n_+ = \{X \in \mathbf{S}^n : X \succ 0\}$, is $(\mathbf{S}^n_+, \mathbf{S}^m_+)$-antimonotone and $\mathbf{S}^m_+$-convex
2. Assuming $P = I_2$, compute numerically $F(X)$ and $dF(X)[dX]$ for $X = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$ and $dX = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. For the above $X$, compute also the Jacobian $J$ of $F$ at $X$ – the matrix of the linear mapping $dX \mapsto DF(X)[dX] : \mathbf{S}^2 \to \mathbf{S}^2$ – in the basis $[1,0;0,0]$, $[0,0;0,1]$, $[0,1/\sqrt{2};1/\sqrt{2},0]$ of $\mathbf{S}^2$.
3. How the "Gradient inequality" (Exercise IV.29.C.1) for the $\mathbf{S}^n_+$-convex mapping $F$ looks like?

*Solution:*

1: Let $G(X) = -(PXP^\top)^{1/2} : \mathbf{S}^n_{++} \to \mathbf{S}^m$ and $H(U) = [-U]^{-1} : -\mathbf{S}^m_{++} \to \mathbf{S}^m$. The mapping $U \mapsto \mathcal{P}(X) := PXP^\top$ with the domain $\mathbf{S}^n_{++}$ maps this domain into $\mathbf{S}^m_{++}$ (since $\mathrm{Ker}P^\top = \{0\}$ due to $\mathrm{rank}(P) = m$), clearly is $(\mathbf{S}^n_+, \mathbf{S}^m_+)$-monotone and, as any linear mapping, is $\mathbf{S}^m_+$-convex; the mapping $U \mapsto \mathcal{G}(U) = -U^{1/2} : \mathbf{S}^m_+ \to \mathbf{S}^m$ is $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-antimonotone and $\mathbf{S}^m_+$-convex by Example IV.26.5. Consequently, the map $X \mapsto G(X) = \mathcal{G}(\mathcal{P}(X)) : \mathbf{S}^n_+ \to \mathbf{S}^m$ is $\mathbf{S}^m_+$-convex (Rule **A.3** in section 26.3) and clearly is $(\mathbf{S}^n_+, \mathbf{S}^m_+)$-antimonotone (as superposition of $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-antimonotone mapping $\mathcal{G}$ and $(\mathbf{S}^n_+, \mathbf{S}^m_+)$-monotone mapping $\mathcal{P}$). Next, mapping $U \mapsto U^{-1} : \mathbf{S}^m_{++} \to \mathbf{S}^m$ is $\mathbf{S}^m_+$-convex and $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-antimonotone (Example IV.26.4), whence the mapping $U \mapsto H(U) := [-U]^{-1} : [-\mathbf{S}^m_{++}] \to \mathbf{S}^m$ is $\mathbf{S}^m_+$-convex (as the superposition of $\mathbf{S}^m_+$-convex mapping $U \mapsto U^{-1} : \mathbf{S}^m_{++} \to \mathbf{S}^m$ and linear mapping $U \mapsto -U : \mathbf{S}^m \to \mathbf{S}^m$). In addition $H(U)$ is $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-monotone on its domain $-\mathbf{S}^m_{++}$ in view of $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-antimonotonicity of the mapping $U \mapsto -U$ and $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-antimonotonicity of the mapping $U \mapsto U^{-1}$ on the domain $\mathbf{S}^m_{++}$. The indicated cone-convexity and cone-monotonicity properties of the mapping $G(\cdot)$ and $H(\cdot)$ imply, in view of Rule **B** in section 26.3, that $F(X) = H(G(X))$ is $\mathbf{S}^m_+$-convex and $(\mathbf{S}^n_+, \mathbf{S}^m_+)$-antimonotone.

2: When justifying Examples IV.26.4 and IV.26.5, we have seen that the mappings $H(\cdot)$ and $\mathcal{G}(\cdot)$ are differentiable on the domains $-\mathbf{S}^m_{++}$, resp., $\mathbf{S}^m_{++}$, and

$$\begin{array}{rcl} DH(U)[dU] & = & U^{-1}dUU^{-1},\, U \in -\mathbf{S}^m_{++}, dU \in \mathbf{S}^m, \\ D\mathcal{G}(V)[dV] & = & -\int\limits_0^\infty \exp\{-V^{1/2}t\}dV\exp\{-V^{1/2}t\}dt,\, V \in \mathbf{S}^m_{++}, dV \in \mathbf{S}^m, \end{array}$$

implying by Chain rule that for $X \in \mathbf{S}^n_{++}, dX \in \mathbf{S}^n$ one has

$$DF(X)[dX] = -\left[PXP^\top\right]^{-1/2}\left[\int\limits_0^\infty \exp\{-[PXP^\top]^{1/2}t\}PdXP^\top\exp\{-[PXP^\top]^{1/2}t\}dt\right]\left[PXP^\top\right]^{-1/2}$$

3: Our computation yields the following results (rounded to 4 digits after the dot):

$$F = \begin{bmatrix} 0.8944 & 0.4472 \\ 0.4472 & 1.3416 \end{bmatrix},\; DF(X)[dX] = \begin{bmatrix} -0.6265 & -0.9846 \\ -0.9846 & -1.1634 \end{bmatrix},$$
$$J = \begin{bmatrix} -0.4025 & -0.2683 & -0.4427 \\ -0.2683 & -1.2970 & -0.8222 \\ -0.4427 & -0.8222 & -0.9839 \end{bmatrix}.$$

4: $\forall (X, Y \in \mathbf{S}^n_{++}) : [PYP^\top]^{-1/2} \succeq [PXP^\top]^{-1/2} + DF(X)[Y-X]$ with $DF(X)[\cdot]$ as described in item 2.

## Around conic representations of sets and functions

### Conic representations: definitions

Let $\mathfrak{K}$ be a family of regular cones in Euclidean spaces which contains the nonnegative ray $\mathbf{R}_+$ and is closed with respect to taking finite direct products and passing from a cone to its dual. Instructive examples are the families $\mathfrak{N}$ of nonnegative orthants, $\mathfrak{L}$ of finite direct products of Lorentz cones, and $\mathfrak{S}$ of finite direct products of semidefinite cones.

- A $\mathfrak{K}$-*representation* ($\mathfrak{K}$-r.) of a set $X \subset \mathbf{R}^n$ is its representation of the form

$$X = \{x \in \mathbf{R}^n : \exists u : Px + Qu - r \in \mathbf{K}\} \tag{29.2}$$

with $\mathbf{K} \in \mathfrak{K}$ – representation of $X$ as the projection of the solution set of conic inequality $Px+Qu \geq_{\mathbf{K}} r$ in variables $x, u$ onto the plane of $x$-variables where $X$ lives. A set $X$ admitting conic representation with cone from $\mathfrak{K}$ is called $\mathfrak{K}$-*representable* ($\mathfrak{K}$-r for short).

- A $\mathfrak{K}$-*representation of a function* $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is, by definition, $\mathfrak{K}$-representation of the epigraph of $f$:

$$[t; x] \in \mathrm{epi}\{f\} := \{[x; t] : t \geq f(x)\} \iff \exists u : Px + tp + Qu - r \in \mathbf{K} \text{ with } \mathbf{K} \in \mathfrak{K}.$$

Functions admitting $\mathfrak{K}$-representation are called $\mathfrak{K}$-*representable* ($\mathfrak{K}$-r for short)

We are already acquainted with $\mathfrak{N}$-representability – it is that was called polyhedral representability. By Fourier-Motzkin elimination, polyhedral representable sets $X \subset \mathbf{R}^n$ admit polyhedral representations not involving additional variables $u$, and similarly for $\mathfrak{N}$-representable functions; this is not the case for more general families $\mathfrak{K}$, like families $\mathfrak{L}$ of Lorentz- and $\mathfrak{S}$ of semidefinite-representable sets.

The following exercise explains what is the rationale underlying the above restrictions on $\mathfrak{K}$ and why we are interested in $\mathfrak{K}$-representations.

**Exercise IV.31.** Check that

1. Every finite system $P_0 y \geq r_0$, $P_i y - r_i \in \mathbf{K}_i$, $i \leq I$, of scalar linear inequalities and conic inequalities, involving cones from $\mathfrak{K}$, in variables $y$ is equivalent to a single conic inequality, with cone from $\mathfrak{K}$, in these variables:

$$\begin{array}{c}\{P_0 y - r_0 \geq 0, P_i y - r_i \in \mathbf{K}_i, 1 \leq i \leq I\} \\ \iff \left\{[P_0; P_1; ...; P_I]y - [r_0; r_1; ...; r_I] \in \mathbf{K} := \underbrace{\mathbf{R}_+ \times ... \times \mathbf{R}_+}_{\dim r_0 \text{ times}} \times \mathbf{K}_1 \times \mathbf{K}_2 \times ... \times \mathbf{K}_I\right\}\end{array}$$

and $\mathbf{K} \in \mathfrak{K}$ (since $\mathbf{R}_+ \in \mathfrak{K}$ and $\mathfrak{K}$ is closed w.r.t. taking finite direct products).
As a result, representation of a set $X$ as

$$X = \{x : \exists u : P_0 x + Q_0 u - r_0 \geq 0, P_i x + Q_i u - r_i \in \mathbf{K}_i, 1 \leq i \leq I\} \qquad [\mathbf{K}_i \in \mathfrak{K}] \quad (!)$$

– as the projection of the solution set of a finite system of linear and $\mathfrak{K}$-conic inequalities in variables $x, u$ onto the plane of $x$-variables where $X$ lives, can be straightforwardly converted into a $\mathfrak{K}$-r. of $X$.

**Important:** Item 1 allows us from now on to refer to representations of the form (!) as to $\mathfrak{K}$-representations of $X$, skipping (always straightforward and purely mechanical) conversion of such a representation into the "canonical" representation (29.2).

2. $\mathfrak{K}$-r. of a function straightforwardly induces $\mathfrak{K}$-r.'s of its sublevel sets:

$$\begin{array}{c}\left\{\{t \geq f(x)\} \iff \{\exists u : Px + tp + Qu - r \in \mathbf{K}\}\right\} \\ \Longrightarrow X_a := \{x : f(x) \leq a\} = \{x : \exists u : Px + Qu - [r - ap] \in \mathbf{K}\}\end{array} \qquad [a \in \mathbf{R}, \mathbf{K} \in \mathfrak{K}]$$

3. Given $\mathfrak{K}$-representations of a set $X \subset \mathbf{R}^n$ and a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$:

$$\begin{array}{rcl} X & = & \{x \in \mathbf{R}^n : \exists u : P_X x + Q_X u - r_X \in \mathbf{K}_X\}, \\ \mathrm{epi}\{f\} & = & \{[x; t] : \exists v : P_f x + tp_f + Q_f v - r_f \in \mathbf{K}_f\} \end{array} \qquad [\mathbf{K}_X \in \mathfrak{K}, \mathbf{K}_f \in \mathfrak{K}]$$

we can straightforwardly convert the optimization problem

$$\min_{x\in X} f(x) \tag{$*$}$$

into conic problem on a cone from $\mathfrak{K}$, namely, the problem

$$\min_{x,t,u,v}\left\{t:\quad A[x;t;u;v]-b \right.$$
$$\left. := [P_Xx+Q_Xu;P_fx+tp_f+Q_fv]-[r_X;r_f]\in\underbrace{\mathbf{K}:=\mathbf{K}_X\times\mathbf{K}_f}_{\in\mathfrak{K}}\right\}$$

As a result, a solver $\mathcal{S}$ capable to solve conic problems on cones from $\mathfrak{K}$ can be straightforwardly utilized when solving problems $(*)$ with $X$ and $f$ given by $\mathfrak{K}$-r.'s.

4. Given a conic problem

$$\min_x\left\{c^\top x: Ax-b\in\mathbf{K}, Rx\geq r\right\} \tag{$P$}$$

on a cone from $\mathfrak{K}$, its conic dual – the conic problem

$$\max\nolimits_{y,z}\left\{\langle b,y\rangle+r^\top z: A^*y+R^\top z=c, y\in\mathbf{K}_*, z\geq 0\right\}$$
$$\left[\begin{array}{c}\langle\cdot,\cdot\rangle \text{ is the inner product in the Euclidean space where } \mathbf{K} \text{ lives, } \mathbf{K}_* \text{ is the cone dual to } \mathbf{K},\\ A^* \text{ is the conjugate of } A\text{: } \langle Ax,y\rangle\equiv x^\top A^*y\ \forall x,y\end{array}\right] \tag{$D$}$$

also is a conic problem on a cone from $\mathfrak{K}$ (since $\mathfrak{K}$ is closed w.r.t. passing from a cone to its dual and contains nonnegative orthants).

*Solution:* This is straightforward – substitute "$\mathfrak{K}$-representation" with the definition of this notion.

Note that the option mentioned in the last item of Exercise IV.31 is implemented in "CVX: MATLAB software for disciplined convex programming" due to M. Grant and S. Boyd `http://cvxr.com/cvx` – second to none in its scope and user-friendliness tool for numerical processing of well-structured convex problems, the underlying family $\mathfrak{K}$ being the semidefinite family $\mathfrak{S}$. We conclude that it makes sense to develop a kind of calculus allowing to recognize $\mathfrak{K}$-representability of sets/functions and to build, when possible, their $\mathfrak{K}$-representations. The desired calculus exists and is pretty simple, general and fully algorithmic. The goal of subsequent exercises is to make you acquainted with the most frequently used elements of this calculus; for more on this subject, see [BTN].

## Conic representability: elementary calculus

Elementary calculus of conic representability is completely similar to calculus of polyhedral representations from section 3.3.

**Exercise IV.32.** [elementary calculus of $\mathfrak{K}$-representable sets] Check that basic convexity-preserving[14] operations with sets preserve $\mathfrak{K}$-representability. Specifically,

1. Finite intersection of $\mathfrak{K}$-r sets $X_i=\{x\in\mathbf{R}^n:\exists u^i: P_ix+Q_iu^i-r_i\in\mathbf{K}_i\}$, $i\leq I$ (here and in what follows all cones involved are from $\mathfrak{K}$) is $\mathfrak{K}$-r.:

$$\bigcap_{i\leq I}X_i = \{x\in\mathbf{R}^n:\exists u=[u^1;...;u^I]: Px+Qu-r$$
$$:= [P_1x+Q_1u^1;...;P_Ix+Q_Iu^I]-[r_1;...;r_I]\in\underbrace{\mathbf{K}:=\mathbf{K}_1\times...\times\mathbf{K}_I}_{\in\mathfrak{K}}\}$$

2. Direct product of finitely many $\mathfrak{K}$-r sets $X_i=\{x\in\mathbf{R}^n:\exists u^i: P_ix+Q_iu^i-r_i\in\mathbf{K}_i\}$, $i\leq I$ is $\mathfrak{K}$-r:

$$X_1\times...\times X_I=\{x=[x^1;...;x^I]:\exists u=[u^1;...;u^I]:$$
$$Px+Qu-r:=[P_1x^1+Q_1u^1;...;P_Ix^I+Q_Iu^I]-[r_1;...;r_I]\in\underbrace{\mathbf{K}:=\mathbf{K}_1\times...\times\mathbf{K}_I}_{\in\mathfrak{K}}\}$$

[14] "convexity-preserving" is crucial – clearly, $\mathfrak{K}$-r sets and functions must be convex!

3. Affine image $Y = \{y = Ax + b : x \in X\}$ of $\mathfrak{K}$-r set $X = \{x \in \mathbf{R}^n : \exists u : Px + Qu - r \in \mathbf{K}\}$ is $\mathfrak{K}$-r:
$$Y = \{y : \exists [x; u] : Ax + b = y, Px + Qu - r \in \mathbf{K}\}$$
is the projection onto the $y$-plane of a set given by explicit finite system of linear and $\mathfrak{K}$-conic inequalities and as such admits an explicit $\mathfrak{K}$-r. by item 1 of Exercise IV.31.
4. Inverse affine image $Y = \{y : Ay + b \in X\}$ of $\mathfrak{K}$-r. set $X = \{x \in \mathbf{R}^n : \exists u : Px + Qu - r \in \mathbf{K}\}$ is $\mathfrak{K}$-r.:
$$Y = \{y : \exists u : PAy + Qy - [r - Pb] \in \mathbf{K}\}.$$
5. The arithmetic sum $X = X_1 + ... + X_I$ of $\mathfrak{K}$-r sets $X_i = \{x \in \mathbf{R}^n : \exists u^i : P_i x + Q_i u^i - r_i \in \mathbf{K}_i\}$, $i \leq I$, is $\mathfrak{K}$-r:
$$X = \{x : \exists [x^1; ...; x^I; u^1; ...; u^I] : x - \sum_i x^i = 0, P_i x^i + Q_i u^i - r_i \in \mathbf{K}_i, i \leq I\}$$
and it remains to apply item 1 of Exercise IV.31.

*Solution:* This is straightforward – substitute "$\mathfrak{K}$-representation" with the definition of this notion.

**Exercise IV.33.** [elementary calculus of $\mathfrak{K}$-representable functions] Check that the following convexity-preserving operations with functions preserve $\mathfrak{K}$-representability:

0. Restricting onto $\mathfrak{K}$-r set: $\mathfrak{K}$-r. $t \geq f(x) \iff \exists u : P_f x + t_f p + Q_f u - r_f \in \mathbf{K}_f$ of a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ taken together with $\mathfrak{K}$-r. $X = \{x \in \mathbf{R}^n : \exists v : P_X x + Q_X v - r_X \in \mathbf{K}_X\}$ of a set $X \subset \mathbf{R}^n$ induce $\mathfrak{K}$-r.
$$t \geq f\big|_X(x) \iff \exists u, v : P_f x + t p_f + Q_f u - r_f \in \mathbf{K}_f,\ P_X x + Q_X v - r_X \in \mathbf{K}_X$$
of the restriction $f\big|_X(x) = \left\{\begin{array}{ll} f(x) & , x \in X \\ +\infty & , x \notin X \end{array}\right.$ of $f$ onto $X$
1. Taking linear combination $\sum_{i=1}^I \lambda_i f_i$ with positive coefficients:
$$\begin{array}{c} t \geq f_i(x) \iff \exists u^i : P_i x + t p_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \\ \Downarrow \\ t \geq f(x) := \sum_{i=1}^I \lambda_i f_i(x) \iff \exists [t_1; ...; t_I; u^1; ...; u^i] : t \geq \sum_i \lambda_i t_i, P_i x + t_i p_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \end{array}$$
2. Direct summation:
$$\begin{array}{c} t \geq f_i(x^i) \iff \exists u^i : P_i x^i + t p_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \\ \Downarrow \\ t \geq f(x^1, ..., x^I) := \sum_{i=1}^I f_i(x^i) \iff \exists [t_1; ...; t_I; u^1; ...; u^i] : t \geq \sum_i t_i, P_i x^i + t_i p_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \end{array}$$
3. Taking finite maxima:
$$\begin{array}{c} t \geq f_i(x) \iff \exists u^i : P_i x + t p_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \\ \Downarrow \\ t \geq f(x) := \max_{i \leq I} f_i(x) \iff \exists [u^1; ...; u^i] : P_i x + t p_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \end{array}$$
4. Affine substitution of variables:
$$\begin{array}{c} t \geq f(x) \iff \exists u : Px + tp + Qu - r \in \mathbf{K} \\ \Downarrow \\ t \geq g(y) := f(Ay + b) \iff \exists u : PAu + tp + Qu - [r - Pb] \in \mathbf{K} \end{array}$$
In fact, claims in items 1–4 are special cases of the following observation:
5. Monotone superposition: let functions $f_i(x)$, $i \leq I$, be $\mathfrak{K}$-r, with the first $K$ of the functions being affine, and let $F(y) : \mathbf{R}^I \to \mathbf{R} \cup \{+\infty\}$ be $\mathfrak{K}$-r and monotonically nondecreasing in $y_{K+1}, ..., y_I$.
$$y, y' \in \mathbf{R}^I, y \geq y', y_i = y'_i, i \leq K \implies F(y) \geq F(y').$$
Then the functions
$$g(x) = \left\{\begin{array}{ll} F(f_1(x), ..., f_I(x)) & , f_i(x) < \infty\, \forall i \\ +\infty & \text{, otherwise.} \end{array}\right.$$

is $\mathfrak{K}$-r, specifically,

$$\left\{\begin{array}{c} f_i \text{ are affine}, i \leq K, \ \& \ t \geq f_i(x) \iff \exists u^i : P_i x + tp_i + Q_i u^i - r_i \in \mathbf{K}_i, \ K < i \leq I \\ t \geq F(y) \iff \exists u : Py + tp + Qu - r \in \mathbf{K} \end{array}\right\}$$
$$\Downarrow$$
$$t \geq g(x) \iff \exists t_i, 1 \leq i \leq I, u^i, K < i \leq I, u : \left\{\begin{array}{ll} \underbrace{t_i - f_i(x) = 0}_{\text{linear equations}} & , i \leq K \\ P_i x + t_i p_i + Q_i u^i - r_i \in \mathbf{K}_i & , K < i \leq I \\ P[t_1; ...; t_k] + tp + Qu - r \in \mathbf{K} & \end{array}\right.$$

*Solution:* This is straightforward – substitute "$\mathfrak{K}$-representation" with the definition of this notion.

## $\mathfrak{R}/\mathfrak{L}/\mathfrak{S}$ hierarchy

**Exercise IV.34.**

1. Let $\mathfrak{K}$ and $\mathfrak{M}$ be two families of regular cones, each containing nonnegative rays and closed w.r.t. taking finite direct products and passing from a cone to its dual cone. Assume that every cone $\mathbf{M} \in \mathfrak{M}$ admits $\mathfrak{K}$-representation:
$$\mathbf{M} = \{y : \exists v : P_{\mathbf{M}} y + Q_{\mathbf{M}} v - r_{\mathbf{M}} \in \underbrace{\mathbf{K}_{\mathbf{M}}}_{\in \mathfrak{K}}\}.$$
Show that a $\mathfrak{M}$-r. $X = \{x \in \mathbf{R}^n : \exists u : Px + Qu - r \in \underbrace{\mathbf{M}}_{\in \mathfrak{M}}\}$ of a set $X$ can be straightforwardly converted into $\mathfrak{K}$-r. of $X$.
2. [Cf. Exercise IV.35] Note that $\mathbf{R}^n_+$ belongs to $\mathfrak{L}$ (same as to every other family of cones we are considering here – all these families contain nonnegative rays and are closed w.r.t. taking finite direct products), thus, every polyhedral representable set/function is Lorentz-representable as well by item 1. Check that the Lorentz cone $\mathbf{L}^m$ is semidefinite-representable as well, specifically,
$$\begin{array}{rcl} \mathbf{L}^m & := & \{x \in \mathbf{R}^m : x_m \geq \sqrt{\sum_{i=1}^{m-1} x_i^2}\} \\ & = & \{x \in \mathbf{R}^m : \text{Arrow}(x) := \left[\begin{array}{c|c|c|c} x_m & x_1 & \cdots & x_{m-1} \\ \hline x_1 & x_m & & \\ \hline \vdots & & \ddots & \\ \hline x_{m-1} & & & x_m \end{array}\right] \succeq 0\} \end{array}$$
implying by item 1 that cones from $\mathfrak{L}$ admit explicit $\mathfrak{S}$-representations and thus that Lorentz-representable sets and functions are semidefinite representable as well, with $\mathfrak{S}$-r.'s readily given by $\mathfrak{L}$-r.'s.

*Solution:* 1: When $X = \{x \in \mathbf{R}^n : \exists u : Px + Qu - r \in \mathbf{M}\}$ and $\mathbf{M} = \{y : \exists v : P_{\mathbf{M}} y + Q_{\mathbf{M}} v - r_{\mathbf{M}} \in \mathbf{K}_{\mathbf{M}} \in \mathfrak{K}\}$, we clearly have
$$\begin{array}{rcl} X & = & \{x : \exists u : Px + Qu - r \in \mathbf{M}\} = \{x : \exists u, y : y = Px + Qu - r, y \in \mathbf{M}\} \\ & = & \{x : \exists u, y, v : y = Px + Qy - r, P_{\mathbf{M}} y + Q_{\mathbf{M}} v - r_{\mathbf{M}} \in \underbrace{\mathbf{K}_{\mathbf{M}}}_{\in \mathfrak{K}}\}, \end{array}$$
and we end up with $\mathfrak{K}$-representation of $X$. □

2: See solution to Exercise IV.35.1.

□

**Exercise IV.35.** It is easy "to see" the nonnegative orthant $\mathbf{R}^n_+$ in the semidefinite cone $\mathbf{S}^n_+$ – $\mathbf{R}^n_+$ is nothing but the intersection of $\mathbf{S}^n_+$ with the linear subspace $L$ of diagonal matrices from $\mathbf{S}^n$. Formally: Let $A$ be the embedding of $\mathbf{R}^n$ into $\mathbf{S}^n$ which maps vector $a$ into diagonal matrix $\text{Diag}\{a\}$; then $z \in \mathbf{R}^n_+$ iff $Az \in \mathbf{S}^n_+$. Alternatively, you can get $\mathbf{R}^n_+$ as the linear image of the positive semidefinite cone, namely, its image under the linear mapping which maps a symmetric $n \times n$ matrix $Z$ into the vector $\text{Dg}\{Z\}$ composed of diagonal entries of $Z$. As a result, a Linear Programming problem $\min_{x \in \mathbf{R}^n}\{c^\top z : Ax \leq b\}$ can be converted into equivalent semidefinite problem $\min_{X\in\mathbf{S}^n}\{\sum_i c_i X_{ii} : X \succeq 0, A\mathrm{Dg}\{X\} \leq b\}$. As it happens, similar possibilities exist for the Lorentz cone $\mathbf{L}^n$, including possibility to reformulate a conic problem involving direct products of Lorentz cones as a semidefinite program. Specifically,

1. Prove that $x \in \mathbf{L}^n$ iff the "arrow" matrix

$$\mathrm{Arrow}(x) = \left[\begin{array}{c|c|c|c|c} x_n & x_1 & x_2 & \dots & x_{n-1} \\ \hline x_2 & x_n & & & \\ \hline \vdots & & \ddots & & \\ \hline x_{n-1} & & & & x_n \end{array}\right]$$

is positive semidefinite.

2. Represent $\mathbf{L}^n$ as the image of $\mathbf{S}^n_+$ under a linear mapping.

*Solution:* 1: The case of $n = 1$ is trivial. Now let $n \geq 2$. In one direction: Assume that $x \in \mathbf{L}^n$, and let us verify that $\mathrm{Arrow}(x) \in \mathbf{S}^n_+$. Indeed, from $x \in \mathbf{L}^n$ it follows that $x_n \geq 0$. If $x_n = 0$, then $x = 0$ due to $\sum_{i=1}^{n-1} x_i^2 \leq x_n^2$, and therefore $\mathrm{Arrow}(x) = 0_{n\times n} \succeq 0$. If $x_n > 0$, then $x_n - \sum_{i=1}^{n-1} x_i^2/x_n \geq 0$, or, which is the same, $x_n - [x_1;\dots;x_{n-1}]^\top[x_n I_{n-1}]^{-1}[x_1;\dots;x_{n-1}] \geq 0$, and $\mathrm{Arrow}(x) \succeq 0$ by the Schur Complement Lemma applied with the $1 \times 1$ North-Western block. In the opposite direction: Assume that $\mathrm{Arrow}(x) \succeq 0$, and let us prove that $x \in \mathbf{L}^n$. Indeed, $x_n$ is diagonal element in positive semidefinite matrix and as such is nonnegative. If $x_n = 0$, then the diagonal of positive semidefinite matrix $\mathrm{Arrow}(x)$ is zero, whence the matrix itself is zero[15], so that $x = 0 \in \mathbf{L}^n$. And if $x_n > 0$, then $\sum_{i=1}^{n-1} x_i^2/x_n \leq x_n$ by the Schur Complement Lemma, the bottom line being that $x_n \geq \sqrt{\sum_{i=1}^{n-1} x_i^2}$, that is, $x \in \mathbf{L}^n$. □

2: The required linear mapping is the mapping $X \mapsto A^*(X)$ conjugate to the mapping $x \mapsto \mathrm{Arrow}(x)$, that is, the mapping $A(X)$ given by the identity

$$[A^*(X)]^\top x = \mathrm{Tr}(X\mathrm{Arrow}(x))\,\forall x \in \mathbf{R}^n, X \in \mathbf{S}^n$$

or, which is the same,

$$A^*(X) = [2X_{12}; 2X_{13}; \dots; 2X_{1n}; \mathrm{Tr}(X)]$$

Indeed, let $L$ be the linear subspace in $\mathbf{S}^n$ composed of arrow matrices, that is, the image space of the linear mapping $x \mapsto \mathrm{Arrow}(x) : \mathbf{R}^n \to \mathbf{S}^n$. Since $L$ intersects $\mathrm{int}\,\mathbf{S}^n_+$, Dubovitski-Milutin Lemma says that restricting onto $L$ nonnegative on $\mathbf{S}^n_+$ linear forms, we get exactly the set of all linear forms on $L$ which are nonnegative on $\mathbf{S}^n_+ \cap L$ (see Exercise II.49.2). Taking into account that $x \mapsto \mathrm{Arrow}(x)$ is one-to-one linear mapping of $\mathbf{R}^n$ onto $L$, we conclude that *linear form $g^\top x$ is nonnegative whenever $x \in \mathbf{L}^n$ iff it is of the form* $\mathrm{Tr}(X\mathrm{Arrow}(x))$ *for some $X \in \mathbf{S}^n$ such that* $\mathrm{Tr}(XY) \geq 0$ *for all $Y \in \mathbf{S}^n_+$*. Since both $\mathbf{S}^n_+$ and $\mathbf{L}^n_+$ are self-dual, the latter observation can be reformulated as "$g \in \mathbf{L}^n$ iff there exists $X \in \mathbf{S}^n_+$ such that $g^\top x = \mathrm{Tr}(X\mathrm{Arrow}(x))$ identically in $x \in \mathbf{R}^n$," or, which is the same, iff $g = A^*(X)$ for some $X \in \mathbf{S}^n_+$. □

## More calculus

The calculus rules to follow are less trivial:

**Exercise IV.36** [passing from a set to its support function and polar] Let $X \subset \mathbf{R}^n$ be a nonempty closed convex set given by essentially strictly feasible $\mathfrak{K}$-representation:

$$\begin{array}{c} X = \{x \in \mathbf{R}^n : \exists u : Ax + Bu - c \geq 0, Px + Qu - r \in \mathbf{K}\} \\ \&\ \exists \bar{x}, \bar{u} : A\bar{x} + B\bar{u} - c \geq 0, P\bar{x} + Q\bar{u} - r \in \mathrm{int}\,\mathbf{K}. \end{array} \qquad (*)$$

[15] due to immediate observation: *if a diagonal entry $A_{ii}$ of $A \succeq 0$ vanishes, then all entries in $i$-th row and $i$-th column of $A$ vanish as well*, due to the inequality $A_{ij}^2 \leq A_{ii}A_{jj}$, see Remark D.28

This representation induces $\mathfrak{K}$-r. of the support function $\phi_X(y) = \sup_{x\in X} y^\top x$, specifically,

$$t \geq \phi_X(y) \iff \exists(\lambda,\xi): \begin{array}{l} A^\top\lambda + P^*\xi + y = 0, B^\top\lambda + Q^*\xi = 0 \\ c^\top\lambda + \langle r,\xi\rangle + t \geq 0, \lambda \geq 0, \xi \in \mathbf{K}_* \end{array}.$$

where $\langle\cdot,\cdot\rangle$ is the inner product in the Euclidean space where $\mathbf{K}$ lives and, as always, $\mathbf{K}_*$ is the cone dual to $\mathbf{K}$. In addition, $(*)$ induces $\mathfrak{K}$-r. of the polar $\mathrm{Polar}\,(X)$ of $X$:

$$\begin{array}{rcl} \mathrm{Polar}\,(X) & := & \{y : y^\top x \leq 1\,\forall x \in X\} \\ & = & \left\{ y : \exists(\lambda,\xi): \begin{array}{l} A^\top\lambda + P^*\xi + y = 0, B^\top\lambda + Q^*\xi = 0 \\ c^\top\lambda + \langle r,\xi\rangle + 1 \geq 0, \lambda \geq 0, \xi \in \mathbf{K}_* \end{array} \right\} \end{array}$$

*Solution:* By definition, $t \geq \phi_X(y)$ iff the optimization problem

$$\max_{x\in X} y^\top x$$

is bounded with optimal value $\leq t$, or, which is the same under the circumstances, the conic problem

$$\max_{x,u} \left\{ y^\top x : Ax + Bu - c \geq 0, Px + Qu - r \in \mathbf{K} \right\} \qquad (\#)$$

is bounded with the optimal value $\leq t$. We are in the case when the latter problem is essentially strictly feasible; applying Conic Duality Theorem, we conclude that $(\#)$ is bounded with optimal value $\leq t$ iff the optimization problem

$$\min_{\lambda,\xi} \left\{ -c^\top\lambda - \langle r,\xi\rangle : A^\top\lambda + P^*\xi = -y, B^\top\lambda + Q^*\xi = 0, \lambda \geq 0, \xi \in \mathbf{K}_* \right\}$$

has a feasible solution with the value of the objective $\leq t$. Thus,

$$t \geq \phi_X(y) \iff \exists(\lambda,\xi): \begin{array}{l} A^\top\lambda + P^*\xi + y = 0, B^\top\lambda + Q^*\xi = 0 \\ b^\top\lambda + \langle r,\xi\rangle + t \geq 0, \lambda \geq 0, \xi \in \mathbf{K}_* \end{array}.$$

The resulting representation of the epigraph of $\phi_X$, by item 1 of Exercise IV.31, straightforwardly induces a $\mathfrak{K}$-r. of $\phi_X$.

Now, $\mathrm{Polar}\,(X) = \{Y : \phi_X(y) \leq 1\}$; applying item 2 of Exercise IV.31, we conclude that

$$\mathrm{Polar}\,(X) = \left\{ y : \exists(\lambda,\xi): \begin{array}{l} A^\top\lambda + P^*\xi + y = 0, B^\top\lambda + Q^*\xi = 0 \\ c^\top\lambda + \langle r,\xi\rangle + 1 \geq 0, \lambda \geq 0, \xi \in \mathbf{K}_* \end{array} \right\} \qquad \square$$

**Exercise IV.37.** Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a proper convex lower semiconscious function given by essentially strictly feasible $\mathfrak{K}$-representation:

$$\begin{array}{c} t \geq f(x) \iff \exists u : Ax + tq + Bu \geq c, Px + tp + Qu - r \in \mathbf{K} \\ \& \ \exists \overline{x}, \overline{t}, \overline{u} : A\overline{x} + \overline{t}q + B\overline{u} \geq c, P\overline{x} + \overline{t}p + Q\overline{u} - r \in \mathrm{int}\,\mathbf{K} \end{array}$$

Build $\mathfrak{K}$-r. of the Legendre transform

$$f^*(y) = \sup_x \left[ y^\top x - f(x) \right]$$

of $f$.

*Solution:* We clearly have

$$f^*(y) = \sup_{x,t} \left\{ y^\top x - t : t \geq f(x) \right\} = \sup_{x,t,u} \left\{ y^\top x - t : Ax + tq + Bu \geq c, Px + tp + Qu - r \in \mathbf{K} \right\},$$

that is, $f^*(y)$ is the optimal value in the conic problem

$$\sup_{x,t,u} \left\{ y^\top x - t : Ax + tq + Bu \geq c, Px + tp + Qu - r \in \mathbf{K} \right\} \qquad (P)$$

Under the circumstances, the problem is essentially strictly feasible, implying by the Conic Duality Theorem that $f^*(y) \leq \tau$ iff the conic dual of $(P)$ – the problem

$$\min_{\lambda,\xi}\left\{-c^\top\lambda - \langle r,\xi\rangle : \begin{array}{l} A^\top\lambda + P^*\xi + y = 0 \\ q^\top\lambda + \langle p,\xi\rangle = 1 \\ B^\top\lambda + Q^*\xi = 0 \\ \lambda \geq 0, \xi \in \mathbf{K}_* \end{array}\right\}$$

– has a feasible solution with the value of the objective $\leq \tau$, that is,

$$\tau \geq f^*(y) \iff \exists \lambda,\xi : \begin{array}{l} c^\top\lambda + \langle r,\xi\rangle + \tau \geq 0 \\ A^\top\lambda + P^*\xi + y = 0 \\ q^\top\lambda + \langle p,\xi\rangle = 1 \\ B^\top\lambda + Q^*\xi = 0 \\ \lambda \geq 0, \xi \in \mathbf{K}_* \end{array}$$

which is a $\mathfrak{K}$-r. of $f^*$. □

**Raw materials.** Rules of grammar become useful only after we have at our disposal words in "dictionary form" which we can combine using these rules. Similarly, calculus of conic representations becomes useful only after a rich enough dictionary of "raw materials," "atoms" – specific $\mathfrak{K}$-representable sets and functions – is built. In contrast to calculus rules which are, basically, independent of what is the family $\mathfrak{K}$ of cones in question, raw materials do depend on $\mathfrak{K}$. Here we restrict ourselves with few instructive examples of Lorentz- and Semidefinite-representable sets and functions; for in-depth acquaintance with this topic, we refer the reader to [BTN].

We understand well what are the "atomic" $\mathfrak{N}$-representable functions and sets – these are half-spaces and affine functions. Other polyhedrally representable sets are intersections of finite families of half-spaces, and other polyhedrally representable functions – maxima of finitely many affine functions restricted on a polyhedral domain. In other words, all $\mathfrak{N}$-representable functions and sets are obtained from the above atoms via the calculus we have just outlined.

In the next two exercises we present instructive examples of $\mathfrak{L}$-r functions and sets.

**Exercise IV.38.** [$\mathfrak{L}$-representability of $\|\cdot\|_2$ and $\|\cdot\|_2^2$] Check that the functions $\|x\|_2$ and $x^\top x$ on $\mathbf{R}^n$ admits $\mathfrak{L}$-r.'s as follows:

$$\begin{array}{l} \{[x;t] \in \mathbf{R}^n_x \times \mathbf{R}_t : t \geq \|x\|_2\} = \left\{[x;t] \in \mathbf{R}^n \times \mathbf{R} : [x;t] \in \mathbf{L}^{n+1}\right\} \\ \{[x;t] \in \mathbf{R}^n_x \times \mathbf{R}_t : t \geq x^\top x\} = \left\{[x;t] \in \mathbf{R}^n \times \mathbf{R} : [2x;t-1;t+1] \in \mathbf{L}^{n+2}\right\} \end{array}$$

*Solution:* evident.

**Exercise IV.39.** [$\mathfrak{L}$-representability of power functions] Justify the following claims

1. Let $k$ be a positive integer. Then the set

$$\mathfrak{G}_k = \left\{[t;x_1;x_2;...;x_{2^k}] \geq 0 : t \leq \left[\prod_{i=1}^{2^k} x_i\right]^{1/2^k}\right\}$$

– the intersection of the hypograph of the geometric mean of $2^k$ nonnegative variables $x_1,...,x_{2^k}$ with the half-space $\{[t;x] \in \mathbf{R}^{2^k}_x \times \mathbf{R}_t : t \geq 0\}$ – admits $\mathfrak{L}$-representation, specifically,

$$\mathfrak{G}_k = \left\{\begin{array}{l} [t;x_1;x_2;...;x_{2^k}] \geq 0 : \exists\{u_{i,\ell} \geq 0, 1 \leq \ell \leq k, 1 \leq i \leq 2^\ell\} : \\ \qquad u_{ik} = x_i, 1 \leq i \leq 2^k \\ \qquad [2u_{i\ell}; u_{2i-1,\ell+1} - u_{2i,\ell+1}; u_{2i-1,\ell+1} + u_{2i,\ell+1}] \in \mathbf{L}^3, \\ \qquad\qquad\qquad\qquad 1 \leq i \leq 2^\ell, 1 \leq \ell < k \\ \qquad [2t; u_{1,1} - u_{2,1}; u_{1,1} + u_{2,1}] \in \mathbf{L}^3. \end{array}\right\} \qquad (*)$$

*Solution:* For a triple of nonnegative reals $u, v, w$, relation $[2u; w - v; v + w] \in \mathbf{L}^3$ is equivalent to $u \leq \sqrt{vw}$. Thus, the inequalities on $x, t, u_{i,\ell}$ in $(*)$ tell us the following story:

> We split $2^k$ nonnegative variables $x_i$, $i \leq 2^k$ of "generation 0" into $2^{k-1}$ consecutive pairs and associate with $i$-th of these pairs its "child" – nonnegative variable $u_{i,k-1}$ "of generation 1" linked to its parents $x_{2i-1}, x_{2i}$ by the inequality $u_{i,k-1} \leq \sqrt{x_{2i-1}x_{2i}}$. Similarly, we split $2^{k-1}$ variables $u_{i,k-1}$ of generation 1 into $2^{k-2}$ consecutive pairs and associate with every pair its child, nonnegative variable $u_{i,k-2}$ of generation 2, and link it to its parents by the inequality $u_{i,k-2} \leq \sqrt{u_{2i-1,k-1}u_{2i,k-1}}$.
>
> We proceed in the same fashion until 2 variables, $u_{1,1}$, $u_{2,1}$ of generation $k-1$ are built, and link these two variables to variable $t$ by the inequality $t \leq \sqrt{u_{1,1}u_{2,1}}$.

Note that the constraints on all our variables are the linear nonnegativity constraints and the constraints stating that specific linear images of the vector of these variables belong to $\mathbf{L}^3$, that is, the solution set $S$ of the system of constraints specifying all our variables is given by explicit system of linear and $\mathbf{L}^3$-conic inequalities, and this system provides an explicit $\mathfrak{L}$-r. of the projection $\overline{S}$ of $S$ onto the plane of variables $t, x_i$. On the other hand, it is clear that what our story says about relation between (nonnegative!) variables $x_i$ and $t$ is *exactly* the inequality $t \leq \left[\prod\limits_{i=1}^{2^k} x_i\right]^{1/2^k}$, so that $\overline{S}$ is nothing but the set $\mathfrak{G}_k$.

Surprisingly, item 1 paves road to $\mathfrak{L}$-representations of power functions.

2. Build explicit $\mathfrak{L}$-r's of the univariate functions as follows:

2.1. $f(x) = \max[0, x]^\theta$ with rational $\theta = p/q \geq 1$ ($p \geq q$ are positive integers).

2.2. $f(x) = \left\{\begin{array}{ll} x^{p_+/q_+} & , x \geq 0 \\ |x|^{p_-/q_-} & , x \leq 0 \end{array}\right.$, where $p_\pm$, $q_\pm$ are positive integers with $p_+/q_+ \geq 1$, $p_-/q_- \geq$ 1

2.3. $f(x) = \left\{\begin{array}{ll} -x^{p/q} & , x \geq 0 \\ +\infty & , x < 0 \end{array}\right.$ with positive integers $p, q$ such that $p/q \leq 1$

2.4. $f(x) = \left\{\begin{array}{ll} x^{-p/q} & , x > 0 \\ +\infty & , x \leq 0 \end{array}\right.$ with positive integers $p, q$

*Solution:* 2.1: Given positive integers $p \geq q$, let us select positive integer $k$ such that $p + q \leq 2^k$ and consider the affine mapping

$$(y, t) \to [y; \overbrace{y; ...; y}^{2^k-p}; \overbrace{t; ...; t}^{q}; \overbrace{1; ...; 1}^{p-q}] : \mathbf{R}^2 \to \mathbf{R}^{1+2^k}.$$

Our calculus of conic representations allows to convert the $\mathfrak{L}$-r. of $\mathfrak{G}_k$ built in item 1 into an explicit $\mathfrak{L}$-r. for the inverse image of the set $\mathfrak{G}_k$ under the above affine mapping, that is, for the set

$$F = \{[y; t] \in \mathbf{R}^2_+ : t \geq y^{p/q}\}.$$

The epigraph $E$ of $f$ is obtained from $F$ by operations covered by our calculus:

$$E = \{[t; x] : t \geq \max[x; 0]^{p/q}\} = \{[t; x] : \exists y : [t; y] \in F, y \geq x\},$$

so that our calculus allows to convert the $\mathfrak{L}$-r. of $F$ we have already built into $\mathfrak{L}$-r. for $E$.

2.2: Construction from item 2.1 allows us to build an explicit $\mathfrak{L}$-r. for the function $\max[0, x]^{p_+/q_+}$ and, after evident modification, for the function $\max[0, -x]^{p_-/q_-}$. These $\mathfrak{L}$-r.'s via calculus provide explicit $\mathfrak{L}$-r. for the sum of these two functions, that is, for our now target function $f$.

2.3: The epigraph of our $f$ is obtained from the one of the function $g(z) = \max[0, z]^{q/p}$ by one-to-one linear transformation, and we can convert the explicit $\mathfrak{L}$-r. of $g$ given in item 2.1 into $\mathfrak{L}$-r. of our current $f$.

2.4: Given $p, q$, let us find positive integer $k$ such that $2^k \geq p + q$, and consider the affine mapping

$$[t; x] \mapsto [1; \overbrace{t; ...; t}^{q}; \overbrace{x; ...; x}^{p}; \overbrace{1; ...; 1}^{2^k-p-q}] : \mathbf{R}^2 \to \mathbf{R}^{1+2^k}.$$

The inverse affine image of $\mathfrak{G}_k$ under this mapping is exactly the epigraph of our current $f$, so that calculus of $\mathfrak{L}$-r.'s provides us with explicit $\mathfrak{L}$-r. of $f$ inherited from the $\mathfrak{L}$-r. of $\mathfrak{G}_k$ built in item 1.

3. Build $\mathfrak{L}$-r's of the following sets:

   3.1. The hypograph

   $$\{[x;t] \in \mathbf{R}^n_+ \times \mathbf{R}_t : t \leq f(x) := x_1^{\pi_1} x_2^{\pi_2} ... x_n^{\pi_n}\}$$

   of algebraic monomial of $n$ nonnegative variables, where $\pi_i$ are positive rationals such that $\sum_i \pi_i \leq 1$ (the latter inequality for nonnegative $\pi_i$'s is a necessary and sufficient for $f$ to be concave on $\mathbf{R}^n_+$).

   3.2. The epigraph of algebraic monomial $f(x) = x_1^{-\pi_1} x_2^{-pi_2} ... x_n^{-\pi_n}$ of $n$ positive variables, where $\pi_i$ are positive rationals.

   3.3. The epigraph of $\|\cdot\|_\pi$ on $\mathbf{R}^n$ with rational $\pi \geq 1$.

*Solution:* 3.1: Let $\pi_i = p_i/q$ with positive integers $p_i$ and $q$ and $k$ be positive integer such that $2^k \geq q$. Consider the affine mapping

$$[x;t] \mapsto [t; \overbrace{x_1;...;x_1}^{p_1}; \overbrace{x_2;...;x_2}^{p_2}; ...; \overbrace{x_n;...;x_n}^{p_n}; \overbrace{t;...;t}^{2^k-q}; \quad \overbrace{1;...;1}^{q-p_1-...-p_n} \quad ] : \mathbf{R}^{1+n} \to \mathbf{R}^{1+2^k};$$

note that the right hand side makes sense due to $p_1+...+p_n \leq q$ in view of $\sum_i \pi_i \leq 1$. As is immediately seen, the inverse image of $\mathfrak{G}_k$ under this mapping is the set

$$F = \{[x;t] \geq 0 : t \leq f(x)\},$$

and the $\mathfrak{L}$-r. of $\mathfrak{G}_k$ built in item 1 combines with the calculus of $\mathfrak{L}$-representations to yield an explicit $\mathfrak{L}$-r. for $F$. It remains to note that, similarly to what happens in item 2.1, the hypograph $E$ of $f$ is obtained from $F$ by operations covered by our calculus:

$$E = \{[x;t] : \exists \tau : [x;\tau] \in F \ \& \ t \leq \tau\}.$$

3.2: Representing $\pi_i = p_i/q$ with positive integers $p_i$, $q$ and selecting positive integer $k$ such that $2^k \geq q + \sum_i p_i$, consider the affine mapping

$$[x;t] \mapsto [1; \overbrace{x_1;...;x_1}^{p_1}; \overbrace{x_2;...;x_2}^{p_2}; ...; \overbrace{x_n;...;x_n}^{p_n}; \overbrace{t;...;t}^{q}; \quad \overbrace{1;...;1}^{2^k-q-\sum_i p_i} \quad ] : \mathbf{R}^{n+1} \to \mathbf{R}^{1+2^k}.$$

as is immediately seen, the inverse image of $\mathfrak{G}_k$ under this mapping is exactly the epigraph of $f$, so that a $\mathfrak{L}$-r. for $f$ is readily given by our calculus as applied to the $\mathfrak{L}$-r. of $\mathfrak{G}_k$ built in item 1.

3.3: The case of $\pi = 1$ is trivial. Now let $\pi \in (1, \infty)$. It is immediately seen (check it) that

$$t \geq \|x\|_\pi \iff t \geq 0 \ \& \ \exists u_i, v_i : \pm x_i \leq u_i, u_i \leq v_i^{1/\pi} t^{1-1/\pi}, \sum_i v_i \leq t.$$

The sets $\{(t, u_i, v_i) \geq 0 : u_i \leq v_i^{1/\pi} t^{1-1/\pi}\}$ admit explicit $\mathfrak{L}$-r.'s by item 3.1, and these $\mathfrak{L}$-r.'s via our calculus yield an explicit $\mathfrak{L}$-r. for the epigraph of $\|x\|_\pi$

By Exercise IV.34, expressive abilities of semidefinite representations are at least as strong as those of Lorentz representability. In fact, $\mathfrak{S}$-representability is strong enough to bring, "for all practical purposes," the entire Convex Optimization within the grasp of Semidefinite Optimization. In our next exercise we are just touching the tip of the "semidefinite iceberg."

**Exercise IV.40.**

1. For starters, build $\mathfrak{S}$-r.'s of the maximum eigenvalue of a symmetric matrix and of the spectral norm $\|\cdot\|_{2,2}$ (the maximum singular value) of a rectangular matrix.
   *Hint:* Note that for a $p \times q$ matrix $A$, the eigenvalues of the symmetric $(p+q) \times (p+q)$ matrix $\left[\begin{array}{c|c} & A \\ \hline A^\top & \end{array}\right]$ are the singular values of $A$, minus these singular values, and perhaps a number of zeros.

*Solution:* $\mathfrak{S}$-r. of the maximal eigenvalue $\lambda_{\max}(X)$ of symmetric $m \times m$ matrix $X$ is immediate:

$$t \geq \lambda_{\max}(X) \iff tI_m - X \succeq 0,$$

This observation combines with Hint to yield $\mathfrak{S}$-r. of the spectral norm of $p \times q$ matrix:

$$t \geq \|X\|_{2,2} \iff \left[\begin{array}{c|c} tI_p & X \\ \hline X^\top & tI_q \end{array}\right] \succeq 0.$$

As a matter of fact, the single most valuable $\mathfrak{S}$-representation is the one for the sums $S_k(X)$ of $k$ largest eigenvalues of a symmetric matrix $X$; convexity of these sums in $X$ was established in chapter 18.

2. Build $\mathfrak{S}$-r. of the sum $S_k(X)$ of $k \leq m$ largest eigenvalues of $m \times m$ symmetric matrix $X$.
   *Hint:* Recall the polyhedral representation, built in Exercise I.29, of the "vector analogy" of $S_k(X)$ – the sum $s_k(x)$ of $k$ largest entries in $m$-dimensional vector $x$:

$$t \geq s_k(x) \iff \exists z \geq 0, s : x \leq z + s\mathbf{1}, \sum_i z_i + ks \leq t,$$

   where $\mathbf{1}$ is the all-ones vector.

*Solution:* The matrix analogy of the representation of $s_k(x)$ is

$$\exists Z \succeq 0, s : X \preceq Z + sI_m, \mathrm{Tr}(Z) + ks \leq t,$$

and we arrive at the "educated guess" stating that for symmetric $m \times m$ matrices $X$ it holds

$$t \geq S_k(X) \iff \exists Z \succeq 0, s : X \preceq Z + sI_m, \mathrm{Tr}(Z) + ks \leq t.$$

Let us verify that this educated guess is true.
In one direction: assume that $Z \succeq 0$ and $s$ are such that $X \preceq Z + sI_m$ and $\mathrm{Tr}(Z) + ks \leq t$, and let us prove that $S_k(X) \leq t$. Denoting by $\lambda(U)$ the vector of eigenvalues, taken with their multiplicities and written down in the non-ascending order, of a symmetric matrix $U$, recall that $U \succeq U'$ implies that $\lambda(U) \geq \lambda(U')$ (by Variational Characterisation of Eigenvalues). Consequently,

$$S_k(X) = s_k(\lambda(X)) \leq s_k(\lambda(Z + sI_m)) = s_k(\lambda(Z) + s\mathbf{1}) = s_k(\lambda(Z)) + sk \leq \mathrm{Tr}(Z) + sk,$$

where the last inequality is due to $Z \succeq 0$. The concluding quantity in the above chain is $\leq t$, that is , $S_k(X) \leq t$, as claimed.
In the opposite direction: let $S_k(X) \leq t$, and let $X = U\,\mathrm{Diag}\{\lambda_1, \lambda_2, ..., \lambda_m\}U^\top$ be the eigenvalue decomposition of $X$, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$ being the eigenvalues of $X$. Let us set $s = \lambda_k$ and $Z = U\,\mathrm{Diag}\{\lambda_1 - \lambda_k, \lambda_2 - \lambda_k, ..., \lambda_{k-1} - \lambda_k, 0, ..., 0\}U^\top$, so that $Z \succeq 0$ and $\mathrm{Tr}(Z) = S_k(X) - k\lambda_k = S_k(X) - ks$, that is, $t \geq S_k(X) = \mathrm{Tr}(Z) + ks$. It remains to note that $X \preceq Z + sI_m$ due to

$$\begin{array}{l} U^\top[sI_m + Z - X]U = \lambda_k I_m + \mathrm{Diag}\{\lambda_1 - \lambda_k, ..., \lambda_{k-1} - \lambda_k, 0, ..., 0\} - \mathrm{Diag}\{\lambda_1, \lambda_2, ...\lambda_m) \\ = \mathrm{Diag}\{0, ..., 0, \lambda_k - \lambda_{k+1}, \lambda_k - \lambda_{k+2}, ..., \lambda_k - \lambda_m\} \succeq 0. \end{array}$$

The importance of $\mathfrak{S}$-representability of $S_k(\cdot)$ becomes clear from the following

3. Let $f(x) : \mathbf{R}^m \to \mathbf{R} \cup \{+\infty\}$ be a convex function symmetric w.r.t. permutations of entries in the argument, and let

$$F(X) = f(\lambda(X)) : \mathbf{S}^m \to \mathbf{R} \cup \{+\infty\};$$

   recall that $F$ is convex by Proposition III.18.3. Show that $F(X)$ admits the following representation:

$$t \geq F(x) \iff \exists u \in \mathbf{R}^m : \begin{array}{ll} f(u) \leq t & (a) \\ u_1 \geq u_2 \geq ... \geq u_m & (b) \\ S_k(X) \leq u_1 + ... + u_k,\ 1 \leq k < m & (c_k) \\ \mathrm{Tr}(X) = u_1 + ... + u_m & (c_m) \end{array} \qquad (29.3)$$

   Combine this fact with $\mathfrak{S}$-representability of $S_k(\cdot)$ to arrive at the following

> *Corollary* In the situation of item 3, assume that $f$ is not just symmetric, but is $\mathfrak{S}$-representable as well. A $\mathfrak{S}$-r. of $f$ gives rise to explicit $\mathfrak{S}$-r. of $F(X)$.

Corollary underlies $\mathfrak{S}$-representations of numerous highly important functions and sets, e.g., *Shatten norms* of rectangular matrices – $p$-norms of the vector of matrix's singular values, or the hypograph $t \le \mathrm{Det}^{1/m}(X)$ of the (appropriate power of the) determinant of $X \in \mathbf{S}^m_+$, or the epigraph of the function $\mathrm{Det}^{-1}(X)$ of $X \succ 0$.

*Solution:* All we need is to justify (29.3). In one direction: when $t \ge F(X)$, setting $u = \lambda(X)$, we satisfy $(a)-(c)$. In the opposite direction: Let $u$, $X$ satisfy $(a)-(c)$. From $(b)$, $(c)$ it follows that $s_k(\lambda(X)) \le s_k(u)$ for all $k \le m$, with $s_m(\lambda(X)) = s_m(u)$. Invoking Majorization Principle (section 9.4), we conclude that $\lambda(X) = Pu$ for a properly selected doubly stochastic matrix $P$. The latter relation, by permutational symmetry and convexity of $f$, implies that $f(\lambda(X)) \le f(u)$ (see Lemma III.18.1), which combines with $(a)$ to imply the desired relation $F(X) \le t$. □

**Exercise IV.41.** A rather interesting example of $\mathfrak{S}$-representable sets deals with matrix square and marix square root:

1. [$\succeq$-epigraph of the matrix square] Prove that the function $F(X) = X^\top X : \mathbf{R}^{m\times n} \to \mathbf{S}^n$ is $\succeq$-convex and find a $\mathfrak{S}$-r. of its $\succeq$-epigraph $\{(X,Y) \in \mathbf{R}^{m\times n} \times \mathbf{S}^n : Y \succeq X^\top X\}$.

   *Solution:* This is immediate: by Schur Compelent Lemma,

   $$\{(X,Y) \in \mathbf{R}^{m\times n} \times \mathbf{S}^n : Y \succeq X^\top X\} = \{(X,Y) : \left[\begin{array}{c|c} Y & X^\top \\ \hline X & I_m \end{array}\right] \succeq 0\}.$$

   In particular,

   $$\{(X,Y) \in \mathbf{S}^n \times \mathbf{S}^n : Y \succeq X^2\} = \{(X,Y) : \left[\begin{array}{c|c} Y & X \\ \hline X & I_n \end{array}\right] \succeq 0\}.$$

2. [$\succeq$-hypograph of the matrix square root] Prove that the set $\{(X,Y) \in \mathbf{S}^n \times \mathbf{S}^n : X \succeq 0, Y \preceq X^{1/2}\}$ is convex and find its $\mathfrak{S}$-r.

   *Solution:* The function $X^{1/2} : \mathbf{S}^n_+ \to \mathbf{S}^n_+$ is $\succeq$-concave and $\succeq$-monotone (Example IV.26.5), and therefore

   $$\begin{array}{rcl}\{(X,Y) \in \mathbf{S}^n \times \mathbf{S}^n : X \succeq 0, Y \preceq X^{1/2}\} & = & \{(X,Y) : \exists V : 0 \preceq V, V^2 \preceq X, Y \preceq V\} \\ & = & \{(X,Y) : \exists V : \left\{\begin{array}{l} X \succeq 0, V \succeq 0, Y \preceq V \\ \left[\begin{array}{c|c} X & V \\ \hline V & I_n \end{array}\right] \succeq 0 \end{array}\right. \quad \}\end{array}$$

Note: Solutions to items 1–2 provide us with $\mathfrak{S}$-r.'s of the sets $\{(X,Y) \in \mathbf{S}^n \times \mathbf{S}^n : X \succeq 0, 0 \preceq X \preceq Y^{1/2}\}$ and $\{(X,Y) \in \mathbf{S}^n \times \mathbf{S}^n : X \succeq 0, X^2 \preceq Y\}$. These sets are different, and the second is "essentially smaller" than the first one, see Exercise IV.17.

**Exercise IV.42.** [important example of $\mathfrak{S}$-representation] Consider the situation as follows. Given a *basic set* $\mathcal{B} \subset \mathbf{R}^n$ which is the solution set of a strictly feasible quadratic inequality:

$$\mathcal{B} = \{u \in \mathbf{R}^n : u^\top Q u + 2q^\top u + \kappa \le 0\},$$

we consider *target set*

$$\mathcal{Q} = \{x \in \mathbf{R}^m : x^\top S x + 2s^\top x + \sigma \le 0\} \qquad [S \in \mathbf{S}^m, s \in \mathbf{R}^m, \sigma \in \mathbf{R}]$$

and affine mapping

$$u \mapsto P(x) := Pu + p : \mathbf{R}^n \to \mathbf{R}^m.$$

We are interested in the situation when the image of the basic set under the mapping $P(\cdot)$ is contained in the target set, and want to describe this situation in terms of the parameters $S, s, \sigma, P, p$. Your task is as follows. Let us set

$$\mathcal{M}(S,s,\sigma;P,p;\lambda)=[P,p]^\top S[P,p]+\left[\begin{array}{c|c}-\lambda Q & P^\top s-\lambda q\\ \hline s^\top P-\lambda q^\top & 2s^\top p+\sigma-\lambda\kappa\end{array}\right].$$

Prove that the inclusion $P(\mathcal{B})\subset\mathcal{Q}$ is equivalent to the existence of $\lambda\geq 0$ such that

$$\mathcal{M}(S,s,\sigma;P,p;\lambda)\preceq 0. \tag{!}$$

*Solution:*

1. Observe that $P(\mathcal{B})\subset\mathcal{Q}$ iff the strictly feasible quadratic inequality

$$u^\top Qu+2q^\top u+\kappa\leq 0$$

on variables $u\in\mathbf{R}^n$ implies validity of the quadratic inequality

$$[Pu+p]^\top S[Pu+p]+2s^\top[Pu+p]+\sigma\leq 0,$$

By Inhomogeneous $\mathcal{S}$-Lemma this is the case if and only if there exists $\lambda\geq 0$ such that

$$\forall(u\in\mathbf{R}^n,t\in\mathbf{R}):[Pu+tp]^\top S[Pu+tp]+2ts^\top[Pu+tp]+\sigma t^2-\lambda[u^\top Qu+2tq^\top u+\kappa t^2]\leq 0,$$

and immediate computation shows that the matrix of the left hand side homogeneous quadratic function of $[u;t]$ is exactly $\mathcal{M}(S,s,\sigma;P,p;\lambda)$. □

2. Here are the results of our experiments with the inscribed ellipsoid method:

- $n=5$: # of iterations: $I=71$, $f(x^I)=37.36223$, cpu 68 sec
- $n=10$: # of iterations: $I=131$, $f(x^I)=41.30913$, cpu 177 sec

Note that the convex optimization problems in question are well-structured: from the results of Exercise IV.39 it follows that the objectives are $\mathfrak{L}$-r, so that the problems can be solved via Conic Quadratic Programming. With this tool (as implemented in CVX), solving the instance with $n=5$ took just 1.28 sec with reported optimal value 37.36220; similar numbers for the instance with $n=10$ are 1.99 and 41.30908. We see that, on one hand, just exploiting convexity *per se* already allows to solve optimization problems, at lest low-dimensional ones, to high accuracy in reasonable time, and, on the other hand, utilizing problem's structure via the machinery of $\mathfrak{R}/\mathfrak{L}/\mathfrak{S}$ representations reduces dramatically the computational effort.

# Exercises from Appendix A

**Exercise 1.**

1. Mark in the list below those subsets of $\mathbf{R}^n$ which are linear subspaces. For the ones that are linear subspaces, find their dimensions and point out bases. For the ones that are not linear subspaces provide counterexamples.

   1. $\mathbf{R}^n$

      *Solution:* linear subspace, dimension is $n$, basis, e.g., the collection of $n$ standard basic orth.

   2. $\{0\}$

      *Solution:* linear subspace, dimension is 0, basis is empty.

   3. $\varnothing$

      *Solution:* not a linear subspace (linear subspace by definition must be nonempty).

   4. $\left\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n ix_i = 0\right\}$

      *Solution:* linear subspace, dimension is $n-1$, basis, e.g., the collection of vectors

      $$f_i := [\underbrace{0;\ldots;0}_{i-1}; i+1; -i; 0; \ldots; 0], \quad \text{for } 1 \le i \le n-1.$$

   5. $\left\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n ix_i^2 = 0\right\}$

      *Solution:* linear subspace, dimension is 0, basis is empty.

   6. $\left\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n ix_i = 1\right\}$

      *Solution:* not a linear subspace (e.g., does not contain the origin).

   7. $\left\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n ix_i^2 = 1\right\}$

      *Solution:* not a linear subspace (e.g., contains the first basic orth, but does not contain twice this orth).

2. Suppose that we know $L$ is a subspace of $\mathbf{R}^n$ with exactly one basis. What is $L$?

   *Solution:* $L = \{0\}$, basis is empty.

**Exercise 2.** Consider the sets given in Exercise 1 and identify the ones that are affine subspaces. For the ones that are affine subspaces, find their affine dimensions and point out their linear subspaces parallel to them. For the ones that are not affine subspaces, provide counterexamples.

*Solution:* All of the sets that are marked as linear subspaces are also affine subspaces. Their affine dimension is equal to their linear dimension, and the corresponding linear subspace parallel to them is just themselves.

Among the ones that are not linear subspaces, we have the following for their affine subspace status:

- $\varnothing$: not an affine subspace since affine subspace by definition needs to be nonempty.

- $\left\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n i x_i = 1\right\}$: affine subspace, affine dimension is $n-1$, and the corresponding linear subspace parallel to it is given by $\left\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n i x_i = 0\right\}$.
- $\left\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n i x_i^2 = 1\right\}$: not an affine subspace, e.g., it contains the points $a_\pm = [\pm 1; 0; \dots; 0]$, but does not contain their average (which is their affine combination!).

**Exercise 3.**

1. Find the orthogonal complement (w.r.t. the standard inner product) of the following subspace of $\mathbf{R}^n$:

$$\left\{x \in \mathbf{R}^n : \sum_{i=1}^n x_i = 0\right\}.$$

*Solution:* The orthogonal complement in question is $\mathbf{R} \cdot [1; \dots; 1]$, i.e., the one-dimensional linear subspace spanned by the all-ones vector.

2. Given vectors $a_1, \dots, a_m \in \mathbf{R}^n$, find the orthogonal complement (w.r.t. the standard inner product) of the linear subspace $\{x \in \mathbf{R}^n : a_i^\top x = 0, \forall i = 1, \dots, n\}$.

*Solution:* The orthogonal complement to the linear subspace $\{x \in \mathbf{R}^n : Ax = 0\}$ is spanned by the transposes of rows of $A$.

3. Find an orthonormal basis (w.r.t. the standard inner product) of the linear subspace $\{x \in \mathbf{R}^n : x_1 = 0\}$ of $\mathbf{R}^n$

*Solution:* An orthonormal basis is, e.g., $\{e_2, e_3, \dots, e_n\}$, where $e_i$ are the standard basic orth in $\mathbf{R}^n$.

**Exercise 4.** Suppose $a \in \mathbf{R}^n$ where $a_i > 0$ for all $i = 1, \dots, n$, and consider the affine subspace

$$M = \left\{x \in \mathbf{R}^n : \sum_{i=1}^n a_i x_i = 1\right\}$$

Point out the linear subspace parallel to $M$ and find an affine basis in $M$.

*Solution:* The parallel linear subspace is $\{x \in \mathbf{R}^n : \sum_{i=1}^n a_i x_i = 0\}$. An example of an affine basis is the collection $\left\{\frac{1}{a_1} e_1, \dots, \frac{1}{a_n} e_n\right\}$, where $e_i$ is the $i$-th standard basic orth.

**Exercise 5.** Let $\varnothing \neq C \subseteq \mathbf{R}^n$ and $x \in \mathbf{R}^n$ be given.

1. Is it always true that $\mathrm{Aff}(C - \{x\}) = \mathrm{Aff}(C) - \{x\}$?

*Solution:* This is always true. Let $y \in \mathrm{Aff}(C - \{x\})$. Then, there are $\lambda_i$'s with $\sum_i \lambda_i = 1$ and $z_i \in C - \{x\}$, such that $y = \sum_i \lambda_i z_i$. Since $z_i \in C - \{x\}$, there are $x_i \in C$ such that $z_i = x_i - x$. Therefore, $y = \sum_i \lambda_i (x_i - x) = \sum_i \lambda_i x_i - x \in \mathrm{Aff}(C) - \{x\}$. Similarly, if $y \in \mathrm{Aff}(C) - \{x\}$, then there are $\lambda_i$'s with $\sum_i \lambda_i = 1$ and $x_i \in C$, such that $y = \sum_i \lambda_i x_i - x = \sum_i \lambda_i (x_i - x) \in \mathrm{Aff}(C - \{x\})$. Therefore, $\mathrm{Aff}(C - \{x\}) = \mathrm{Aff}(C) - \{x\}$.

2. Is it always true that $\mathrm{Lin}(C - \{x\}) = \mathrm{Aff}(C) - \{x\}$?

*Solution:* The equality $\mathrm{Lin}(C - \{x\}) = \mathrm{Aff}(C) - \{x\}$ is not always true, because if $x \notin \mathrm{Aff}(C)$, the set $\mathrm{Aff}(C) - \{x\}$ does not contain the zero vector, but the set $\mathrm{Lin}(C - \{x\})$ always contains the zero vector.

3. Do your answers to the previous questions change if you further assume $x \in \mathrm{Aff}(C)$?

*Solution:* The answer to the first question does not depend on whether or not $x \in \mathrm{Aff}(C)$ holds. On the other hand, when $x \in \mathrm{Aff}(C)$, the answer to the second question changes and the relation $\mathrm{Lin}(C - \{x\}) = \mathrm{Aff}(C) - \{x\}$ always holds. This is because $\mathrm{Aff}(C) - \{x\}$ is an affine subspace that contains the zero vector, therefore it is a linear subspace. Since it also contains all the elements of $C - \{x\}$, it holds that $\mathrm{Lin}(C - \{x\}) \subseteq \mathrm{Aff}(C) - \{x\}$. For the other direction, we can use the equality

$\mathrm{Aff}(C-\{x\}) = \mathrm{Aff}(C)-\{x\}$ we have shown before. Therefore, we have $\mathrm{Aff}(C)-\{x\} = \mathrm{Aff}(C-\{x\}) \subseteq \mathrm{Lin}(C-\{x\})$. Thus, $\mathrm{Lin}(C-\{x\}) = \mathrm{Aff}(C)-\{x\}$ when $x \in \mathrm{Aff}(C)$.

**Exercise 7.** Prove that the *Triangle inequality in Euclidean norm*, i.e., $\|x+y\|_2 \leq \|x\|_2 + \|y\|_2$, holds true as *equality* if and only if $x$ and $y$ are nonnegative multiples of some vector (which always can be taken to be $x+y$).

*Solution:* Observe, first, that $x$, $y$ are nonnegative multiples of some vector iff they are nonnegative multiples of $x+y$. Next, the Triangle inequality in $\|\cdot\|_2$ holds true as equality if and only if $x^\top x + 2x^\top y + y^\top y = \|x+y\|_2^2 = (\|x\|_2 + \|y\|_2)^2 = \|x\|_2^2 + 2\|x\|_2\|y\|_2 + \|y\|_2^2$, or, which is the same, $x^\top y = \|x\|_2\|y\|_2$. The latter relation clearly holds true when $x, y$ are nonnegative multiples of some vector. Now let $x^\top y = \|x\|_2\|y\|_2$, and let us prove that $x$ and $y$ are nonnegative multiples of some vector. There is nothing to prove when either $x$, or $y$, or both, are zero. Now assume that $x \neq 0$, $y \neq 0$. Setting $f = x/\|x\|_2$, $g = y/\|y\|_2$, we arrive at the situation when $\|f\|_2 = \|g\|_2 = 1$, and $x^\top y = \|x\|_2\|y\|_2$ translates to $f^\top g = 1$. Consequently, $\|f-g\|_2^2 = \|f\|_2^2 + \|g\|_2^2 - 2f^\top g = 0$, that is, $f = g$, so that $x$ and $y$ are positive multiples of $f = g$. □

# Exercises from Appendix B

**Exercise 8.** Mark in the list below those sets which are closed and those which are open (sets are in $\mathbf{R}^n$, $\|\cdot\|$ is a norm on $\mathbf{R}^n$, $n > 0$):

1. All vectors with integer coordinates.

   *Solution:* closed

2. All vectors with rational coordinates.

   *Solution:* neither closed, nor open

3. All vectors with positive coordinates.

   *Solution:* open

4. All vectors with nonnegative coordinates.

   *Solution:* closed

5. $\{x \in \mathbf{R}^n : \ \|x\| < 1\}$.

   *Solution:* open

6. $\{x \in \mathbf{R}^n : \ \|x\| = 1\}$.

   *Solution:* closed

7. $\{x \in \mathbf{R}^n : \ \|x\| \leq 1\}$

   *Solution:* closed

8. $\{x \in \mathbf{R}^n : \ \|x\| \geq 1\}$

   *Solution:* closed

9. $\{x \in \mathbf{R}^n : \ \|x\| > 1\}$

   *Solution:* open

10. $\{x \in \mathbf{R}^n : \ 1 < \|x\| \leq 2\}$

    *Solution:* neither closed, nor open

**Exercise 9.** Consider the function $f(x_1, x_2) : \mathbf{R}^2 \to \mathbf{R}$ defined as

$$f(x_1, x_2) = \begin{cases} \frac{x_1^2 - x_2^2}{x_1^2 + x_2^2}, & \text{if } (x_1, x_2) \neq 0 \\ 0, & \text{if } x_1 = x_2 = 0. \end{cases}$$

Check whether this function is continuous on the following sets:

1. $\mathbf{R}^2$

   *Solution:* $f$ is not continuous on the set

2. $\mathbf{R}^2 \setminus \{0\}$

   *Solution:* $f$ is continuous on the set

3. $\{x \in \mathbf{R}^2 : x_1 = 0\}$

   *Solution:* $f$ is not continuous on the set (note that in this domain we have $f(x) = -1$ whenever $x_2 \neq 0$ and $f(x) = 0$ whenever $x_2 = 0$)

4. $\{x \in \mathbf{R}^2 : x_2 = 0\}$

   *Solution:* $f$ is not continuous on the set (note that in this domain we have $f(x) = 1$ whenever $x_1 \neq 0$ and $f(x) = 0$ whenever $x_1 = 0$)

5. $\{x \in \mathbf{R}^2 : x_1 + x_2 = 0\}$

   *Solution:* $f$ is continuous on the set

6. $\{x \in \mathbf{R}^2 : x_1 - x_2 = 0\}$

   *Solution:* $f$ is continuous on the set

7. $\{x \in \mathbf{R}^2 : |x_1 - x_2| \leq x_1^4 + x_2^4\}$

   *Solution:* $f$ is continuous on the set.

**Exercise 10.** Let $f : \mathbf{R}^n \to \mathbf{R}^m$ be a continuous mapping. Among the following statements, mark those which are always true:

1. If $U$ is an open set in $\mathbf{R}^m$, then so is the set $f^{-1}(U) := \{x \in \mathbf{R}^n : \ f(x) \in U\}$.

   *Solution:* true

2. If $U$ is an open set in $\mathbf{R}^n$, then so is the set $f(U) = \{f(x) : \ x \in U\}$.

   *Solution:* not always true (take $f \equiv 0$)

3. If $F$ is a closed set in $\mathbf{R}^m$, then so is the set $f^{-1}(F) = \{x \in \mathbf{R}^n : \ f(x) \in F\}$.

   *Solution:* true

4. If $F$ is a closed set in $\mathbf{R}^n$, then so is the set $f(F) = \{f(x) : \ x \in F\}$.

   *Solution:* not always true (take $f(x) = \exp\{x\} : \mathbf{R} \to \mathbf{R}$ and look at $f(\mathbf{R})$).

**Exercise 11.** Prove that in general *neither one* of Theorems B.25, B.29, and B.31 remains valid when

1. $X$ is closed, but not bounded;

   *Solution:* Take the mapping $x \mapsto \exp\{x\} : X := \mathbf{R} \to \mathbf{R}$, so that $X$ is closed and $f$ is continuous on $X$. Here:

   - $f$ is unbounded on $X$, and $f(X)$ is not closed, in contrast to the conclusion of Theorem B.25
   - $f$ is not uniformly continuous on $X$, in contrast to the conclusion of Theorem B.29
   - $f$ does not achieve its minimum on $X$, in contrast to the conclusion of Theorem B.31

2. $X$ is bounded, but not closed.

   *Solution:* Take the mapping $x \mapsto \frac{1}{x} : X := (0, 1) \to \mathbf{R}$, so that $X$ is bounded and $f$ is continuous on $X$. Here:

   - $f$ is unbounded on $X$, and $f(X)$ is not closed, in contrast to the conclusion of Theorem B.25
   - $f$ is not uniformly continuous on $X$, in contrast to the conclusion of Theorem B.29
   - $f$ does not achieve its minimum on $X$, in contrast to the conclusion of Theorem B.31

# Exercises from Appendix D

## Exercise 12.

1. Find the dimension of $\mathbf{R}^{m\times n}$ and point out a basis in this space.

   *Solution:* The dimension is $mn$, and a basis is, e.g., the basis $\left\{e_i f_j^\top : i \leq m,\, j \leq n\right\}$, where $e_1, ..., e_m$ and $f_1, ..., f_n$ are the standard basic orths in $\mathbf{R}^m$, resp., $\mathbf{R}^n$

2. Build an orthonormal basis in $\mathbf{S}^m$.

   *Solution:* An orthonormal basis in $\mathbf{S}^m$ is composed of $m$ matrices $e_i e_i^\top$ and $\frac{m(m-1)}{2}$ matrices $\frac{1}{\sqrt{2}}[e_i e_j^\top + e_j e_i^\top]$, $1 \leq i < j \leq m$, where $e_i$ are the standard basic orths in $\mathbf{R}^m$.

**Exercise 13.** In the space $\mathbf{R}^{m\times m}$ of square $m \times m$ matrices, there are two interesting subsets: the set $\mathbf{S}^m$ of *symmetric* matrices $\left\{A : A = A^\top\right\}$ and the set $\mathbf{J}^m$ of *skew-symmetric* matrices $\{A = [A_{ij}] : A_{ij} = -A_{ji},\ \forall i, j\}$.

1. Verify that both $\mathbf{S}^m$ and $\mathbf{J}^m$ are linear subspaces of $\mathbf{R}^{m\times m}$.

   *Solution:* This is evident.

2. Find the dimension of $\mathbf{S}^m$ and point out a basis in $\mathbf{S}^m$.

   *Solution:* The dimension of $\mathbf{S}^m$ is $\frac{m(m+1)}{2}$, the basis for $\mathbf{S}^m$ was built in Exercise 12.2.

3. Find the dimension of $\mathbf{J}^m$ and point out a basis in $\mathbf{J}^m$.

   *Solution:* The dimension of $\mathbf{J}^m$ is $\frac{m(m-1)}{2}$ (note that all of the diagonal entries of the matrices in $\mathbf{J}^m$ must be zero), an orthonormal basis for $\mathbf{J}^m$ is, e.g., $\frac{1}{\sqrt{2}}[e_i e_j^\top - e_j e_i^\top]$, $1 \leq i < j \leq m$.

4. What is the sum of $\mathbf{S}^m$ and $\mathbf{J}^m$? What is the intersection of $\mathbf{S}^m$ and $\mathbf{J}^m$?

   *Solution:* Their sum is the entire $\mathbf{R}^{m\times m}$, and their intersection is $\{0\}$.

**Exercise 14.** Is the "3-factor" extension of Fact D.1 valid, at least in the case of square matrices $X, Y, Z$ of the same size? That is, for square matrices $X, Y, Z$ of the same size, is it always true that $\mathrm{Tr}(XYZ) = \mathrm{Tr}(YXZ)$?

*Solution:* Beyond the trivial case of $1 \times 1$ matrices, this is wrong, as is immediately shown by numerical experimentation.

**Exercise 15.** Given $P \in \mathbf{S}^p$, $Q \in \mathbf{R}^{r\times p}$, and $R \in \mathbf{S}^r$, consider the matrices

$$A = \begin{bmatrix} P & Q^\top \\ Q & R \end{bmatrix},\quad B = \begin{bmatrix} P & -Q^\top \\ -Q & R \end{bmatrix},\quad C = \begin{bmatrix} R & Q \\ Q^\top & P \end{bmatrix},\quad D = \begin{bmatrix} R & -Q \\ -Q^\top & P \end{bmatrix}.$$

Prove that $\lambda(A) = \lambda(B) = \lambda(C) = \lambda(D)$. Thus, the matrices $A, B, C, D$ simultaneously are/are not positive semidefinite. As a consequence: Schur Complement Lemma says that when $R \succ 0$, one has $A \succeq 0$ iff $P - Q^\top R^{-1} Q \succeq 0$; since $A \succeq 0$ iff $C \succeq 0$, we see that the same Lemma says that when $P \succ 0$, one has $A \succeq 0$ iff $R - QP^{-1}Q^\top \succeq 0$.

*Solution:* Indeed, the matrices are rotations of each other:

$$B = UAU^\top, \quad C = VAV^\top, \quad D = WAW^\top$$

where

$$U = \begin{bmatrix} -I_p & \\ & I_r \end{bmatrix}, \quad V = \begin{bmatrix} & I_r \\ I_p & \end{bmatrix}, \quad W = \begin{bmatrix} & -I_r \\ I_p & \end{bmatrix},$$

and clearly the matrices $U, V, W$ are orthogonal.

**Exercise 16.** Let $\mathbf{S}^n_{++} := \operatorname{int} \mathbf{S}^n_+ = \{X \in \mathbf{S}^n : X \succ 0\}$, and consider $X, Y \in \mathbf{S}^n_{++}$. Then, $X \preceq Y$ holds if and only if $X^{-1} \succeq Y^{-1}$ ("$\succeq$-antimonotonicity of $X^{-1}$, $X \in \mathbf{S}^n_{++}$). Is it true that from $0 \prec X \preceq Y$ it always follows that $X^{-2} \succeq Y^{-2}$?

*Solution:* For $Z \succ 0$, we clearly have $Z \preceq I_n$ if and only if $Z^{-1} \succeq I_n$, and therefore for $X \succ 0$, $Y \succ 0$ we have

$$X \preceq Y \iff Y^{-1/2} X Y^{-1/2} \preceq I_n \iff Y^{1/2} X^{-1} Y^{1/2} \succeq I_n \iff X^{-1} \succeq Y^{-1}.$$

Numerical experimentation shows that $0 \prec X \preceq Y$ not always implies that $X^{-2} \succeq Y^{-2}$.

**Exercise 17.** Let $A, B \in \mathbf{S}^n$ be such that $0 \preceq A \preceq B$. For each one of the following, either prove the statement or produce a counter example:

1. $A^2 \preceq B^2$;

   *Solution:* We can verify (with Mathematica) that for $n = 2$, taking

   $$A = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

   gives a counterexample to the claim.

2. $0 \preceq A^{1/2} \preceq B^{1/2}$.

   *Solution:* This is always true.
   Note that

   $$B - A = \frac{1}{2}\left[(B^{1/2} + A^{1/2})(B^{1/2} - A^{1/2}) + (B^{1/2} - A^{1/2})(B^{1/2} + A^{1/2})\right].$$

   Hence, $B - A \in \mathbf{S}^n_+$ implies $(B^{1/2} + A^{1/2})(B^{1/2} - A^{1/2}) + (B^{1/2} - A^{1/2})(B^{1/2} + A^{1/2}) \in \mathbf{S}^n_+$. Because $B^{1/2} - A^{1/2}$ is a symmetric matrix, we can rewrite it in terms of its eigenvector decomposition as

   $$B^{1/2} - A^{1/2} = UDU^\top,$$

   where $U$ is an orthogonal matrix and $D$ is a diagonal matrix. Then, by defining $X := 2U^\top(B - A)U$ and $Y := U^\top(B^{1/2} + A^{1/2})U$, we observe that

   $$X = YD + DY \tag{$*$}$$

   holds. Because $B - A \in \mathbf{S}^n_+$, we have $X \in \mathbf{S}^n_+$ (see Fact D.31). Likewise $Y \in \mathbf{S}^n_+$ because $B^{1/2} + A^{1/2} \in \mathbf{S}^n_+$ (since both $A$ and $B$ are positive semidefinite). In addition, observe that

   $$\begin{array}{rcl} A' & := & Y - D = 2U^\top A^{1/2} U \\ B' & := & Y + D = 2U^\top B^{1/2} U. \end{array} \tag{$**$}$$

   Therefore, both $A'$ and $B'$ are in $\mathbf{S}^n_+$. Finally, let us consider the diagonal elements of the matrices $X$, $Y$, $A'$ and $B'$. From $(*)$, $(**)$ we see that

   $$X_{ii} = 2Y_{ii}D_{ii}$$
   $$A'_{ii} = Y_{ii} - D_{ii}$$
   $$B'_{ii} = Y_{ii} + D_{ii}$$

Because all of $X$, $Y$, $A'$ and $B'$ are in $\mathbf{S}^n_+$, we have all these diagonal elements are nonnegative and $Y_{ii} \geq 0$ for all $i \in [m]$. In particular, we have $Y_{ii} \geq |D_{ii}|$ for all $i$. Then for any $j \in [m]$ such that $D_{jj} \neq 0$, we have $Y_{jj} > 0$. Moreover, from $X_{jj} = 2Y_{jj}D_{jj}$ and $Y_{jj} > 0$, we deduce that $D_{jj} \geq 0$ as well. This then implies that $D$ has a nonnegative diagonal, and hence $B^{1/2} - A^{1/2} \in \mathbf{S}^n_+$ as desired.

An alternative proof of "$\succeq$-monotonicity" of the square root of a positive semidefinite matrix is given in Example IV.26.5 in section 26.2.

**Exercise 18.** A matrix $A \in \mathbf{S}^n$ is called *diagonally dominant* if it satisfies the relation

$$a_{ii} \geq \sum\nolimits_{j \neq i} |a_{ij}|, \quad i = 1, \ldots, n.$$

Prove that every diagonally dominant matrix $A$ is positive semidefinite.

*Solution:* Let $x$ be an eigenvector of $A$ with eigenvalue $\lambda$, and let $x_i$ be the entry of $x$ with the maximum absolute value. As $x$ is an eigenvector, $x \neq 0$ and so $x_i \neq 0$. Replacing, if necessary, $x$ with $-x$, we can assume that $x_i > 0$. Then, as $x$ is an eigenvector of $A$ with eigenvalue $\lambda$, we deduce from $Ax = \lambda x$ that

$$a_{ii}x_i + \sum\nolimits_{j \neq i} a_{ij}x_j = \lambda x_i.$$

Moreover, using the fact that $x_i > 0$ is the largest magnitude coordinate in $x$, we get

$$\sum\nolimits_{j \neq i} a_{ij}x_j \leq \left|\sum\nolimits_{j \neq i} a_{ij}x_j\right| \leq \sum\nolimits_{j \neq i} |a_{ij}x_j| \leq x_i \sum\nolimits_{j \neq i} |a_{ij}| \leq a_{ii}x_i.$$

Combining these two relations, we arrive at

$$\lambda x_i = a_{ii}x_i + \sum\nolimits_{j \neq i} a_{ij}x_j \geq a_{ii}x_i - \left|\sum\nolimits_{j \neq i} a_{ij}x_j\right| \geq 0.$$

This means that $\lambda \geq 0$, and since the eigenvalue $\lambda$ was arbitrary, all eigenvalues of $A$ are non-negative, and hence $A \succeq 0$.

**Exercise 19.** Prove the following matrix analogy of the scalar inequality $ab \leq \frac{a^2+b^2}{2}$ for $a, b \in \mathbf{R}$:

$$AB^\top + BA^\top \preceq AA^\top + BB^\top, \qquad \forall A, B \in \mathbf{R}^{m \times n}.$$

*Solution:* Note that we can rewrite this expression as

$$AA^\top - AB^\top - BA^\top + BB^\top = (A - B)(A - B)^\top.$$

Then, the positive semidefiniteness of this matrix is immediate.

**Exercise 20.**

1. Let $I_k$ denote the $k \times k$ identity matrix, and let $A$ be an $m \times n$ matrix. Prove that the following three properties are equivalent to each other:
   - $A^\top A \preceq I_n$;
   - $AA^\top \preceq I_m$;
   - $\begin{bmatrix} I_m & A \\ A^\top & I_n \end{bmatrix} \succeq 0.$

   *Solution:* By the Schur Complement Lemma,

   $$X = \begin{bmatrix} I_m & A \\ A^\top & I_n \end{bmatrix} \succeq 0 \iff I_m - AA^\top \succeq 0.$$

   Invoking the concluding comment in Exercise 15, $X \succeq 0 \iff I_n - A^\top A \succeq 0$.
2. Let $A_1, \ldots, A_k$ be $n \times n$ matrices such that

   $$A_1^\top A_1 + \ldots + A_k^\top A_k \preceq I_n.$$

   For each one of the following, either prove the statement or produce a counter example:
   - $A_1 A_1^\top + \ldots + A_k A_k^\top \preceq I_n$;

*Solution:* When $n = 1$, the claim clearly is true; when $k = 1$, it is true due to item 1 of Exercise. When $k > 1$ and $n > 1$, the claim is wrong in general: set $\kappa = \min[k, n]$, $A_i = ee_i^\top$ with unit $e$, the first $\kappa$ basic orths of $\mathbf{R}^n$ in the role of $e_i$ when $i \leq \kappa$, and $e_i = 0$ for $\kappa < i \leq k$. With this setup, $\sum_i A_i^\top A_i = \sum_{i=1}^\kappa e_i e_i^\top \preceq I_n$, while $\sum_i A_i A_i^\top = \kappa ee^\top \not\preceq I_n$.

- $\begin{bmatrix} A_1A_1^\top & A_1A_2^\top & \cdots & A_1A_k^\top \\ A_2A_1^\top & A_2A_2^\top & \cdots & A_2A_k^\top \\ \vdots & \vdots & \ddots & \vdots \\ A_kA_1^\top & A_kA_2^\top & \cdots & A_kA_k^\top \end{bmatrix} \preceq I_{kn}.$

*Solution:* Observe that by the Schur Complement Lemma and the concluding comment in Exercise 15 we have

$$\begin{aligned} & I_n - (A_1^\top A_1 + \ldots + A_k^\top A_k) \succeq 0 \\ \Longleftrightarrow & \begin{bmatrix} I_n & A_1^\top & \ldots & A_k^\top \\ A_1 & & & \\ \vdots & & I_{kn} & \\ A_k & & & \end{bmatrix} \succeq 0 \\ \Longleftrightarrow & \; I_{kn} - \begin{bmatrix} A_1 \\ \vdots \\ A_k \end{bmatrix} \begin{bmatrix} A_1^\top & \ldots & A_k^\top \end{bmatrix} = I_{kn} - \begin{bmatrix} A_1A_1^\top & A_1A_2^\top & \cdots & A_1A_k^\top \\ A_2A_1^\top & A_2A_2^\top & \cdots & A_2A_k^\top \\ \vdots & \vdots & \ddots & \vdots \\ A_kA_1^\top & A_kA_2^\top & \cdots & A_kA_k^\top \end{bmatrix} \succeq 0, \end{aligned}$$

which is exactly what is required.