Convexity and Optimization

Lars-Åke Lindahl

2016

Contents

Preface

As promised by the title, this book has two themes, convexity and optimization, and convex optimization is the common denominator. Convexity plays a very important role in many areas of mathematics, and the book's first part, which deals with finite dimensional convexity theory, therefore contains significantly more of convexity than is then used in the subsequent three parts on optimization, where Part II provides the basic classical theory for linear and convex optimization, Part III is devoted to the simplex algorithm, and Part IV describes Newton's algorithm and an interior point method with self-concordant barriers.

We present a number of algorithms, but the emphasis is always on the mathematical theory, so we do not describe how the algorithms should be implemented numerically. Anyone who is interested in this important aspect should consult specialized literature in the field.

Mathematical optimization methods are today used routinely as a tool for economic and industrial planning, in production control and product design, in civil and military logistics, in medical image analysis, etc., and the development in the field of optimization has been tremendous since World War II. In 1945, George Stigler studied a diet problem with 77 foods and 9 constraints without being able to determine the optimal diet − today it is possible to solve optimization problems containing hundreds of thousands of variables and constraints. There are two factors that have made this possible − computers and efficient algorithms. Of course it is the rapid development in the computer area that has been most visible to the common man, but the algorithm development has also been tremendous during the past 70 years, and computers would be of little use without efficient algorithms.

Maximization and minimization problems have of course been studied and solved since the beginning of the mathematical analysis, but optimization theory in the modern sense started around 1948 with George Dantzig, who introduced and popularized the concept of linear programming (LP) and proposed an efficient solution algorithm, the simplex algorithm, for such problems. The simplex algorithm is an iterative algorithm, where the number of iterations empirically is roughly proportional to the number of variables for normal real world LP problems. Its worst-case behavior, however, is bad; an example of Victor Klee and George Minty 1972 shows that there are LP

problems in n variables, which for their solution require 2^n iterations. A natural question in this context is therefore how difficult it is to solve general LP problems.

An algorithm for solving a class $\mathcal{K}$ of problems is called *polynomial* if there is a polynomial P, such that the algorithm solves every problem of size s in $\mathcal{K}$ with a maximum of $P(s)$ arithmetic operations; here the size of a problem is defined as the number of binary bits needed to represent it. The class $\mathcal{K}$ is called *tractable* if there is a polynomial algorithm that solves all the problems in the class, and *intractable* if there is no such algorithm.

Klee–Minty's example demonstrates that (their variant of) the simplex algorithm is not polynomial. Whether LP problems are tractable or intractable, however, was an open question until 1979, when Leonid Khachiyan showed that LP problems can be solved by a polynomial algorithm, the ellipsoid method. LP problems are thus, in a technical sense, easy to solve.

The ellipsoid method, however, did not have any practical significance because it behaves worse than the simplex algorithm on normal LP problems. The simplex algorithm was therefore unchallenged as practicable solution tool for LP problems until 1984, when Narendra Karmarkar introduced a polynomial interior-point algorithm with equally good performance as the simplex algorithm, when applied to LP problems from the real world.

Karmarkar's discovery became the starting point for an intensive development of various interior-point methods, and a new breakthrough occurred in the late 1980's, when Yurii Nesterov and Arkadi Nemirovski introduced a special type of convex barrier functions, the so-called self-concordant functions. Such barriers will cause a classical interior-point method to convergence polynomially, not only for LP problems but also for a large class of convex optimization problems. This makes it possible today to solve optimization problems that were previously out of reach.

The embryo of this book is a compendium written by Christer Borell and myself 1978–79, but various additions, deletions and revisions over the years, have led to a completely different text. The most significant addition is Part IV which contains a description of self-concordant functions based on the works of Nesterov and Nemirovski,

The presentation in this book is complete in the sense that all theorems are proved. Some of the proofs are quite technical, but none of them requires more previous knowledge than a good knowledge of linear algebra and calculus of several variables.

Uppsala, April 2016
Lars-Åke Lindahl

List of symbols

$\operatorname{aff} X$	affine hull of X, p. 22
$\operatorname{bdry} X$	boundary of X, p. 12
$\operatorname{cl} f$	closure of the function f, p. 149
$\operatorname{cl} X$	closure of X, p. 12
$\operatorname{con} X$	conic hull of X, p. 40
$\operatorname{cvx} X$	convex hull of X, p. 32
$\dim X$	dimension of X, p. 23
$\operatorname{dom} f$	the effective domain of f: $\{x \mid -\infty < f(x) < \infty\}$, p. 5
$\operatorname{epi} f$	epigraph of f, p. 91
$\operatorname{exr} X$	set of extreme rays of X, p. 68
$\operatorname{ext} X$	set of extreme points of X, p. 67
$\operatorname{int} X$	interior of X, p. 12
$\operatorname{lin} X$	recessive subspace of X, p. 46
$\operatorname{rbdry} X$	relative boundary of X, p. 35
$\operatorname{recc} X$	recession cone of X, p. 43
$\operatorname{rint} X$	relative interior of X, p. 34
$\operatorname{sublev}_\alpha f$	α-sublevel set of f, p. 91
$\mathbf{e}_i$	ith standard basis vector $(0, \ldots, 1, \ldots, 0)$, p. 6
f'	derivate or gradient of f, p. 16
$f'(x; v)$	direction derivate of f at x in direction v, p. 156
f''	second derivative or hessian of f, p. 18
f^*	conjugate function of f, p. 150
$v_{\max}$, $v_{\min}$	optimal values, p. 166
$B(a; r)$	open ball centered at a with radius r, p. 11
$\overline{B}(a; r)$	closed ball centered at a with radius r, p. 11
$Df(a)[v]$	differential of f at a, p. 16
$D^2 f(a)[u, v]$	$\sum_{i,j=1}^n \frac{\partial^2 f}{\partial x_i \partial x_j}(a) u_i v_j$, p. 18
$D^3 f(a)[u, v, w]$	$\sum_{i,j,k=1}^n \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}(a) u_i v_j w_k$, p. 19
$\mathcal{E}(x; r)$	ellipsoid $\{y \mid \|y - x\|_x \leq r\}$, p. 365
$I(x)$	set of active constraints at x, p. 199
L	input length, p. 388
$L(x, \lambda)$	Lagrange function, p. 191
$M_{\hat{r}}[x]$	object obtained by replacing the element in M at location r by x, p. 246

$\mathbf{R}_+$, $\mathbf{R}_{++}$	$\{x \in \mathbf{R} \mid x \geq 0\}$, $\{x \in \mathbf{R} \mid x > 0\}$, p. 3
$\mathbf{R}_-$	$\{x \in \mathbf{R} \mid x \leq 0\}$, p. 3
$\overline{\mathbf{R}}$, $\underline{\mathbf{R}}$, $\overline{\underline{\mathbf{R}}}$	$\mathbf{R} \cup \{\infty\}$, $\mathbf{R} \cup \{-\infty\}$, $\mathbf{R} \cup \{\infty, -\infty\}$, p. 3
S_X	support function of X, p. 118
$S_{\mu,L}(X)$	class of μ-strongly convex functions on X with L-Lipschitz continuous derivative, p. 136
$\mathrm{Var}_X(v)$	$\sup_{x \in X} \langle v, x \rangle - \inf_{x \in X} \langle v, x \rangle$, p. 369
X^+	dual cone of X, p. 58
$\mathbf{1}$	the vector $(1, 1, \ldots, 1)$, p. 6
$\partial f(a)$	subdifferential of f at a, p. 141
$\lambda(f, x)$	Newton decrement of f at x, p. 304, 319
π_y	translated Minkowski functional, p. 366
$\rho(t)$	$-t - \ln(1-t)$, p. 333
ϕ_X	Minkowski functional of X, p. 121
$\phi(\lambda)$	dual function $\inf_x L(x, \lambda)$, p. 192
Δx_{nt}	Newton direction at x, p. 303, 319
∇f	gradient of f, p. 16
$\overrightarrow{x}$	ray from 0 through x, p. 37
$[x, y]$	line segment between x and y, p. 8
$]x, y[$	open line segment between x and y, p. 8
$\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_\infty$	ℓ^1-norm, Euclidean norm, maximum norm, p. 10
$\|\cdot\|_x$	the local seminorm $\sqrt{\langle \cdot\,, f''(x)\cdot \rangle}$, p. 305
$\|v\|_x^*$	the dual local seminorm $\sup_{\|w\|_x \leq 1} \langle v, w \rangle$, p. 368

Part I

Convexity

Chapter 1

Preliminaries

The purpose of this chapter is twofold − to explain certain notations and terminologies used throughout the book and to recall some fundamental concepts and results from calculus and linear algebra.

Real numbers

We use the standard notation $\mathbf{R}$ for the set of real numbers, and we let

$$\begin{aligned}\mathbf{R}_+ &= \{x \in \mathbf{R} \mid x \geq 0\},\\ \mathbf{R}_- &= \{x \in \mathbf{R} \mid x \leq 0\},\\ \mathbf{R}_{++} &= \{x \in \mathbf{R} \mid x > 0\}.\end{aligned}$$

In other words, $\mathbf{R}_+$ consists of all nonnegative real numbers, and $\mathbf{R}_{++}$ denotes the set of all positive real numbers.

The extended real line

Each nonempty set A of real numbers that is bounded above has a least upper bound, denoted by $\sup A$, and each nonempty set A that is bounded below has a greatest lower bound, denoted by $\inf A$. In order to have these two objects defined for arbitrary subsets of $\mathbf{R}$ (and also for other reasons) we extend the set of real numbers with the two symbols $-\infty$ and ∞ and introduce the notation

$$\overline{\mathbf{R}} = \mathbf{R} \cup \{\infty\}, \quad \underline{\mathbf{R}} = \mathbf{R} \cup \{-\infty\} \quad \text{and} \quad \overline{\underline{\mathbf{R}}} = \mathbf{R} \cup \{-\infty, \infty\}.$$

We furthermore extend the order relation $<$ on $\mathbf{R}$ to the extended real line $\overline{\underline{\mathbf{R}}}$ by defining, for each real number x,

$$-\infty < x < \infty.$$

The arithmetic operations on $\mathbf{R}$ are partially extended by the following "natural" definitions, where x denotes an arbitrary real number:

$$
\begin{aligned}
&x + \infty = \infty + x = \infty + \infty = \infty \\
&x + (-\infty) = -\infty + x = -\infty + (-\infty) = -\infty \\
&x \cdot \infty = \infty \cdot x = \begin{cases} \infty & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -\infty & \text{if } x < 0 \end{cases} \\
&x \cdot (-\infty) = -\infty \cdot x = \begin{cases} -\infty & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ \infty & \text{if } x < 0 \end{cases} \\
&\infty \cdot \infty = (-\infty) \cdot (-\infty) = \infty \\
&\infty \cdot (-\infty) = (-\infty) \cdot \infty = -\infty.
\end{aligned}
$$

It is now possible to define in a consistent way the least upper bound and the greatest lower bound of an arbitrary subset of the extended real line. For nonempty sets A which are not bounded above by any real number, we define $\sup A = \infty$, and for nonempty sets A which are not bounded below by any real number we define $\inf A = -\infty$. Finally, for the empty set $\emptyset$ we define $\inf \emptyset = \infty$ and $\sup \emptyset = -\infty$.

Sets and functions

We use standard notation for sets and set operations that are certainly well known to all readers, but the intersection and the union of an arbitrary family of sets may be new concepts for some readers.

So let $\{X_i \mid i \in I\}$ be an arbitrary family of sets X_i, indexed by the set I; their *intersection*, denoted by

$$\bigcap \{X_i \mid i \in I\} \quad \text{or} \quad \bigcap_{i \in I} X_i,$$

is by definition the set of elements that belong to all the sets X_i. The *union*

$$\bigcup \{X_i \mid i \in I\} \quad \text{or} \quad \bigcup_{i \in I} X_i$$

consists of the elements that belong to X_i for at least one $i \in I$.

We write $f\colon X \to Y$ to indicate that the function f is defined on the set X and takes its values in the set Y. The set X is then called the *domain*

of the function and Y is called the *codomain*. Most functions in this book have domain equal to $\mathbf{R}^n$ or to some subset of $\mathbf{R}^n$, and their codomain is usually $\mathbf{R}$ or more generally $\mathbf{R}^m$ for some integer $m \geq 1$, but sometimes we also consider functions whose codomain equals $\overline{\mathbf{R}}$, $\underline{\mathbf{R}}$ or $\overline{\underline{\mathbf{R}}}$.

Let A be a subset of the domain X of the function f. The set

$$f(A) = \{f(x) \mid x \in A\}$$

is called the *image of* A under the function f. If B is a subset of the codomain of f, then

$$f^{-1}(B) = \{x \in X \mid f(x) \in B\}$$

is called the *inverse image of* B under f. There is no implication in the notation $f^{-1}(B)$ that the inverse f^{-1} exists.

For functions $f\colon X \to \overline{\underline{\mathbf{R}}}$ we use the notation $\operatorname{dom} f$ for the inverse image of $\mathbf{R}$, i.e.

$$\operatorname{dom} f = \{x \in X \mid -\infty < f(x) < \infty\}.$$

The set $\operatorname{dom} f$ thus consists of all $x \in X$ with finite function values $f(x)$, and it is called the *effective domain* of f.

The vector space $\mathbf{R}^n$

The reader is assumed to have a solid knowledge of elementary linear algebra and thus, in particular, to be familiar with basic vector space concepts such as linear subspace, linear independence, basis and dimension.

As usual, $\mathbf{R}^n$ denotes the vector space of all n-tuples $(x_1, x_2, \ldots, x_n)$ of real numbers. The elements of $\mathbf{R}^n$, interchangeably called points and vectors, are denoted by lowercase letters from the beginning or the end of the alphabet, and if the letters are not numerous enough, we provide them with sub- or superindices. Subindices are also used to specify the coordinates of a vector, but there is no risk of confusion, because it will always be clear from the context whether for instance x_1 is a vector of its own or the first coordinate of the vector x.

Vectors in $\mathbf{R}^n$ will interchangeably be identified with *column matrices*. Thus, to us

$$(x_1, x_2, \ldots, x_n) \qquad \text{and} \qquad \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

denote the same object.

The vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ in $\mathbf{R}^n$, defined as

$$\mathbf{e}_1 = (1, 0, \ldots, 0), \quad \mathbf{e}_2 = (0, 1, 0, \ldots, 0), \quad \ldots, \quad \mathbf{e}_n = (0, 0, \ldots, 0, 1),$$

are called the natural basis vectors in $\mathbf{R}^n$, and $\mathbf{1}$ denotes the vector whose coordinates are all equal to one, so that

$$\mathbf{1} = (1, 1, \ldots, 1).$$

The *standard scalar product* $\langle \cdot\,, \cdot \rangle$ on $\mathbf{R}^n$ is defined by the formula

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n,$$

and, using matrix multiplication, we can write this as

$$\langle x, y \rangle = x^T y = y^T x,$$

where ^{T} denotes transposition. In general, A^T denotes the transpose of the matrix A.

The solution set to a homogeneous system of linear equations in n unknowns is a linear subspace of $\mathbf{R}^n$. Conversely, every linear subspace of $\mathbf{R}^n$ can be presented as the solution set to some homogeneous system of linear equations:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = 0 \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = 0 \end{cases}$$

Using matrices we can of course write the system above in a more compact form as

$$Ax = 0,$$

where the matrix A is called the *coefficient matrix* of the system.

The dimension of the solution set of the above system is given by the number $n - r$, where r equals the rank of the matrix A. Thus in particular, for each linear subspace X of $\mathbf{R}^n$ of dimension $n - 1$ there exists a nonzero vector $c = (c_1, c_2, \ldots, c_n)$ such that

$$X = \{x \in \mathbf{R}^n \mid c_1 x_1 + c_2 x_2 + \cdots + c_n x_n = 0\}.$$

Sum of sets

If X and Y are nonempty subsets of $\mathbf{R}^n$ and α is a real number, we let

$$\begin{aligned} X + Y &= \{x + y \mid x \in X,\, y \in Y\}, \\ X - Y &= \{x - y \mid x \in X,\, y \in Y\}, \\ \alpha X &= \{\alpha x \mid x \in X\}. \end{aligned}$$

The set $X + Y$ is called the *(vector) sum* of X and Y, $X - Y$ is the *(vector) difference* and αX is the product of the number α and the set X.

It is convenient to have sums, differences and products defined for the empty set $\emptyset$, too. Therefore, we extend the above definitions by defining

$$X \pm \emptyset = \emptyset \pm X = \emptyset$$

for all sets X, and

$$\alpha\emptyset = \emptyset.$$

For singleton sets $\{a\}$ we write $a + X$ instead of $\{a\} + X$, and the set $a + X$ is called a *translation* of X.

It is now easy to verify that the following rules hold for arbitrary sets X, Y and Z and arbitrary real numbers α and β:

$$\begin{gathered} X + Y = Y + X \\ (X + Y) + Z = X + (Y + Z) \\ \alpha X + \alpha Y = \alpha(X + Y) \\ (\alpha + \beta)X \subseteq \alpha X + \beta X\,. \end{gathered}$$

In connection with the last inclusion one should note that the converse inclusion $\alpha X + \beta X \subseteq (\alpha + \beta)X$ does **not** hold for general sets X.

Inequalites in $\mathbf{R}^n$

For vectors $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ in $\mathbf{R}^n$ we write $x \geq y$ if $x_j \geq y_j$ for all indices j, and we write $x > y$ if $x_j > y_j$ for all j. In particular, $x \geq 0$ means that all coordinates of x are nonnegative.

The set

$$\mathbf{R}^n_+ = \mathbf{R}_+ \times \mathbf{R}_+ \times \cdots \times \mathbf{R}_+ = \{x \in \mathbf{R}^n \mid x \geq 0\}$$

is called the *nonnegative orthant* of $\mathbf{R}^n$.

The order relation $\geq$ is a partial order on $\mathbf{R}^n$. It is thus, in other words, reflexive ($x \geq x$ for all x), transitive ($x \geq y$ & $y \geq z \Rightarrow x \geq z$) and antisymmetric ($x \geq y$ & $y \geq x \Rightarrow x = y$). However, the order is not a complete order when $n > 1$, since two vectors x and y may be unrelated.

Two important properties, which will be used now and then, are given by the following two trivial implications:

$$\begin{aligned} x \geq 0 \;\&\; y \geq 0 &\;\Rightarrow\; \langle x, y\rangle \geq 0 \\ x \geq 0 \;\&\; y \geq 0 \;\&\; \langle x, y\rangle = 0 &\;\Rightarrow\; x = y = 0. \end{aligned}$$

Line segments

Let x and y be points in $\mathbf{R}^n$. We define

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

and

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

and we call the set $[x, y]$ the *line segment* and the set $]x, y[$ the *open line segment* between x and y, if the two points are distinct. If the two points coincide, i.e. if $y = x$, then obviously $[x, x] =]x, x[= \{x\}$.

Linear maps and linear forms

Let us recall that a map $S\colon \mathbf{R}^n \to \mathbf{R}^m$ is called *linear* if

$$S(\alpha x + \beta y) = \alpha Sx + \beta Sy$$

for all vectors x, $y \in \mathbf{R}^n$ and all scalars (i.e. real numbers) α, β. A linear map $S\colon \mathbf{R}^n \to \mathbf{R}^n$ is also called a *linear operator* on $\mathbf{R}^n$.

Each linear map $S\colon \mathbf{R}^n \to \mathbf{R}^m$ gives rise to a unique $m \times n$-matrix $\tilde{S}$ such that

$$Sx = \tilde{S}x,$$

which means that the function value Sx of the map S at x is given by the matrixproduct $\tilde{S}x$. (Remember that vectors are identified with column matrices!) For this reason, the same letter will be used to denote a map and its matrix. We thus interchangeably consider Sx as the value of a map and as a matrix product.

By computing the scalar product $\langle x, Sy\rangle$ as a matrix product we obtain the following relation

$$\langle x, Sy\rangle = x^T Sy = (S^T x)^T y = \langle S^T x, y\rangle$$

between a linear map $S\colon \mathbf{R}^n \to \mathbf{R}^m$ (or $m \times n$-matrix S) and its *transposed map* $S^T\colon \mathbf{R}^m \to \mathbf{R}^n$ (or transposed matrix S^T).

An $n \times n$-matrix $A = [a_{ij}]$, and the corresponding linear map, is called *symmetric* if $A^T = A$, i.e. if $a_{ij} = a_{ji}$ for all indices i, j.

A linear map $f\colon \mathbf{R}^n \to \mathbf{R}$ with codomain $\mathbf{R}$ is called a *linear form.* A linear form on $\mathbf{R}^n$ is thus of the form

$$f(x) = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n,$$

where $c = (c_1, c_2, \ldots, c_n)$ is a vector in $\mathbf{R}^n$. Using the standard scalar product we can write this more simply as

$$f(x) = \langle c, x \rangle,$$

and in matrix notation this becomes

$$f(x) = c^T x.$$

Let $f(x) = \langle c, y \rangle$ be a linear form on $\mathbf{R}^m$ and let $S\colon \mathbf{R}^n \to \mathbf{R}^m$ be a linear map with codomain $\mathbf{R}^m$. The composition $f \circ S$ is then a linear form on $\mathbf{R}^n$, and we conclude that there exists a unique vector $d \in \mathbf{R}^n$ such that $(f \circ S)(x) = \langle d, x \rangle$ for all $x \in \mathbf{R}^n$. Since $f(Sx) = \langle c, Sx \rangle = \langle S^T c, x \rangle$, it follows that $d = S^T c$.

Quadratic forms

A function $q\colon \mathbf{R}^n \to \mathbf{R}$ is called a *quadratic form* if there exists a symmetric $n \times n$-matrix $Q = [q_{ij}]$ such that

$$q(x) = \sum_{i,j=1}^{n} q_{ij} x_i x_j,$$

or equivalently

$$q(x) = \langle x, Qx \rangle = x^T Q x.$$

The quadratic form q determines the symmetric matrix Q uniquely, and this allows us to identify the form q with its matrix (or operator) Q.

An arbitrary quadratic polynomial $p(x)$ in n variables can now be written in the form

$$p(x) = \langle x, Ax \rangle + \langle b, x \rangle + c,$$

where $x \mapsto \langle x, Ax \rangle$ is a quadratic form determined by a symmetric operator (or matrix) A, $x \mapsto \langle b, x \rangle$ is a linear form determined by a vector b, and c is a real number.

Example. In order to write the quadratic polynomial

$$p(x_1, x_2, x_3) = x_1^2 + 4x_1x_2 - 2x_1x_3 + 5x_2^2 + 6x_2x_3 + 3x_1 + 2x_3 + 2$$

in this form we first replace the terms dx_ix_j for $i < j$ with $\frac{1}{2}dx_ix_j + \frac{1}{2}dx_jx_i$. This yields

$$\begin{aligned} p(x_1, x_2, x_3) &= (x_1^2 + 2x_1x_2 - x_1x_3 + 2x_2x_1 + 5x_2^2 + 3x_2x_3 - x_3x_1 + 3x_3x_2) \\ &\quad + (3x_1 + 2x_3) + 2 = \langle x, Ax \rangle + \langle b, x \rangle + c \end{aligned}$$

with $A = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 5 & 3 \\ -1 & 3 & 0 \end{bmatrix}$, $b = \begin{bmatrix} 3 \\ 0 \\ 2 \end{bmatrix}$ and $c = 2$. □

A quadratic form q on $\mathbf{R}^n$ (and the corresponding symmetric operator and matrix) is called *positive semidefinite* if $q(x) \geq 0$ and *positive definite* if $q(x) > 0$ for all vectors $x \neq 0$ in $\mathbf{R}^n$.

Norms and balls

A *norm* $\|\cdot\|$ on $\mathbf{R}^n$ is a function $\mathbf{R}^n \to \mathbf{R}_+$ that satisfies the following three conditions:

(i)	$\|x + y\| \leq \|x\| + \|y\|$	for all x, y
(ii)	$\|\lambda x\| = \|\lambda\| \, \|x\|$	for all $x \in \mathbf{R}^n$, $\lambda \in \mathbf{R}$
(iii)	$\|x\| = 0 \Leftrightarrow x = 0$.	

The most important norm to us is the *Euclidean norm*, defined via the standard scalar product as

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

This is the norm that we use unless the contrary is stated explicitely. We use the notation $\|\cdot\|_2$ for the Euclidean norm whenever we for some reason have to emphasize that the norm in question is the Euclidean one.

Other norms, that will occur now and then, are the *maximum norm*

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

and the ℓ^1*-norm*

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

It is easily verified that these really are norms, that is that conditions (i)–(iii) are satisfied.

All norms on $\mathbf{R}^n$ are equivalent in the following sense: If $\|\cdot\|$ and $\|\cdot\|'$ are two norms, then there exist two positive constants c and C such that

$$c\|x\|' \leq \|x\| \leq C\|x\|'$$

for all $x \in \mathbf{R}^n$.

For example, $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \, \|x\|_\infty$.

Given an arbitrary norm $\|\cdot\|$ we define the corresponding *distance* between two points x and a in $\mathbf{R}^n$ as $\|x-a\|$. The set

$$B(a;r) = \{x \in \mathbf{R}^n \mid \|x-a\| < r\},$$

consisting of all points x whose distance to a is less than r, is called the *open ball* centered at the point a and with radius r. Of course, we have to have $r > 0$ in order to get a nonempty ball. The set

$$\overline{B}(a;r) = \{x \in \mathbf{R}^n \mid \|x-a\| \leq r\}$$

is the corresponding *closed ball*.

The geometric shape of the balls depends on the underlying norm. The ball $\overline{B}(0;1)$ in $\mathbf{R}^2$ is a square with corners at the points $(\pm 1, \pm 1)$ when the norm is the maximum norm, it is a square with corners at the points $(\pm 1, 0)$ and $(0, \pm 1)$ when the norm is the ℓ^1-norm, and it is the unit disc when the norm is the Euclidean one.

If B denotes balls defined by one norm and B' denotes balls defined by a second norm, then there are positive constants c and C such that

$$B'(a;cr) \subseteq B(a;r) \subseteq B'(a;Cr) \tag{1.1}$$

for all $a \in \mathbf{R}^n$ and all $r > 0$. This follows easily from the equivalence of the two norms.

All balls that occur in the sequel are assumed to be Euclidean, i.e. defined with respect to the Euclidean norm, unless otherwise stated.

Topological concepts

We now use balls to define a number of topological concepts. Let X be an arbitrary subset of $\mathbf{R}^n$. A point $a \in \mathbf{R}^n$ is called

- an *interior point* of X if there exists an $r > 0$ such that $B(a;r) \subseteq X$;
- a *boundary point* of X if $X \cap B(a;r) \neq \emptyset$ and $\complement X \cap B(a;r) \neq \emptyset$ for all $r > 0$;
- an *exterior point* of X if there exists an $r > 0$ such that $X \cap B(a;r) = \emptyset$.

Observe that because of property (1.1), the above concepts do not depend on the kind of balls that we use.

A point is obviously either an interior point, a boundary point or an exterior point of X. Interior points belong to X, exterior points belong to the complement of X, while boundary points may belong to X but must not do so. Exterior points of X are interior points of the complement $\complement X$, and vice versa, and the two sets X and $\complement X$ have the same boundary points.

The set of all interior points of X is called the *interior* of X and is denoted by $\operatorname{int} X$. The set of all boundary points is called the *boundary* of X and is denoted by $\operatorname{bdry} X$.

A set X is called *open* if all points in X are interior points, i.e. if $\operatorname{int} X = X$.

It is easy to verify that the union of an arbitrary family of open sets is an open set and that the intersection of finitely many open sets is an open set. The empty set $\emptyset$ and $\mathbf{R}^n$ are open sets

The interior $\operatorname{int} X$ is a (possibly empty) open set for each set X, and $\operatorname{int} X$ is the biggest open set that is included in X.

A set X is called *closed* if its complement $\complement X$ is an open set. It follows that X is closed if and only if X contains all its boundary points, i.e. if and only if $\operatorname{bdry} X \subseteq X$.

The intersection of an arbitrary family of closed sets is closed, the union of finitely many closed sets is closed, and $\mathbf{R}^n$ and $\emptyset$ are closed sets.

For arbitrary sets X we set

$$\operatorname{cl} X = X \cup \operatorname{bdry} X.$$

The set $\operatorname{cl} X$ is then a closed set that contains X, and it is called the *closure* (or *closed hull*) of X. The closure $\operatorname{cl} X$ is the smallest closed set that contains X as a subset.

For example, if $r > 0$ then

$$\operatorname{cl} B(a; r) = \{x \in \mathbf{R}^n \mid \|x - a\| \leq r\} = \overline{B}(a; r),$$

which makes it consistent to call the set $\overline{B}(a; r)$ a closed ball.

For nonempty subsets X of $\mathbf{R}^n$ and numbers $r > 0$ we define

$$X(r) = \{y \in \mathbf{R}^n \mid \exists x \in X \colon \|y - x\| < r\}.$$

The set $X(r)$ thus consists of all points whose distance to X is less than r.

A point x is an exterior point of X if and only if the distance from x to X is positive, i.e. if and only if there is an $r > 0$ such that $x \notin X(r)$. This means that a point x belongs to the closure $\operatorname{cl} X$, i.e. x is an interior point or a boundary point of X, if and only if x belongs to the sets $X(r)$ for all $r > 0$. In other words,

$$\operatorname{cl} X = \bigcap_{r>0} X(r).$$

A set X is said to be *bounded* if it is contained in some ball centered at 0, i.e. if there is a number $R > 0$ such that $X \subseteq B(0; R)$.

A set X that is both closed and bounded is called *compact*.

An important property of compact subsets X of $\mathbf{R}^n$ is given by the Bolzano–Weierstrass theorem: *Every infinite sequence* $(x_n)_{n=1}^{\infty}$ *of points* x_n *in a compact set* X *has a subsequence* $(x_{n_k})_{k=1}^{\infty}$ *that converges to a point in* X.

The cartesian product $X \times Y$ of a compact subset X of $\mathbf{R}^m$ and a compact subset Y of $\mathbf{R}^n$ is a compact subset of $\mathbf{R}^m \times \mathbf{R}^n \quad (= \mathbf{R}^{m+n})$.

Continuity

A function $f\colon X \to \mathbf{R}^m$, whose domain X is a subset of $\mathbf{R}^n$, is defined to be *continuous at the point* $a \in X$ if for each $\epsilon > 0$ there exists an $r > 0$ such that

$$f(X \cap B(a; r)) \subseteq B(f(a); \epsilon).$$

(Here, of course, the left B stands for balls in $\mathbf{R}^n$ and the right B stands for balls in $\mathbf{R}^m$.) The function is said to be *continuous on* X, or simply *continuous*, if it is continuous at all points $a \in X$.

The inverse image $f^{-1}(I)$ of an open interval under a continuous function $f\colon \mathbf{R}^n \to \mathbf{R}$ is an open set in $\mathbf{R}^n$. In particular, the sets $\{x \mid f(x) < a\}$ and $\{x \mid f(x) > a\}$, i.e. the sets $f^{-1}(]-\infty, a[)$ and $f^{-1}(]a, \infty[)$, are open for all $a \in \mathbf{R}$. Their complements, the sets $\{x \mid f(x) \geq a\}$ and $\{x \mid f(x) \leq a\}$, are thus closed.

Sums and (scalar) products of continuous functions are continuous, and quotients of real-valued continuous functions are continuous at all points where the quotients are well-defined. Compositions of continuous functions are continuous.

Compactness is preserved under continuous functions, that is the image $f(X)$ is compact if X is a compact subset of the domain of the continuous function f. For continuous functions f with codomain $\mathbf{R}$ this means that f is bounded on X and has a maximum and a minimum, i.e. there are two points x_1, $x_2 \in X$ such that $f(x_1) \leq f(x) \leq f(x_2)$ for all $x \in X$.

Lipschitz continuity

A function $f\colon X \to \mathbf{R}^m$ that is defined on a subset X of $\mathbf{R}^n$, is called *Lipschitz continuous* with Lipschitz constant L if

$$\|f(y) - f(x)\| \leq L\|y - x\| \qquad \text{for all } x, y \in X.$$

Note that the definition of Lipschitz continuity is norm independent, since all norms on $\mathbf{R}^n$ are equivalent, but the value of the Lipschitz constant L is obviously norm dependent.

Operator norms

Let $\|\cdot\|$ be a given norm on $\mathbf{R}^n$. Since the closed unit ball is compact and linear operators S on $\mathbf{R}^n$ are continuous, we get a finite number $\|S\|$, called the *operator norm*, by the definition

$$\|S\| = \sup_{\|x\|\leq 1} \|Sx\|.$$

That the operator norm really is a norm on the space of linear operators, i.e. that it satisfies conditions (i)–(iii) in the norm definition, follows immediately from the corresponding properties of the underlying norm on $\mathbf{R}^n$.

By definition, $S(x/\|x\|) \leq \|S\|$ for all $x \neq 0$, and consequently

$$\|Sx\| \leq \|S\|\|x\|$$

for all $x \in \mathbf{R}^n$.

From this inequality follows immediately that

$$\|STx\| \leq \|S\|\|Tx\| \leq \|S\|\|T\|\|x\|,$$

which gives us the important inequality

$$\|ST\| \leq \|S\|\|T\|$$

for the norm of a product of two operators.

The identity operator I on $\mathbf{R}^n$ clearly has norm equal to 1. Therefore, if the operator S is invertible, then, by choosing $T = S^{-1}$ in the above inequality, we obtain the inequality

$$\|S^{-1}\| \geq 1/\|S\|.$$

The operator norm obviously depends on the underlying norm on $\mathbf{R}^n$, but again, different norms on $\mathbf{R}^n$ give rise to equivalent norms on the space of operators. However, when speaking about the operator norm we shall in this book always assume that the underlying norm is the Euclidean norm even if this is not stated explicitely.

Symmetric operators, eigenvalues and norms

Every symmetric operator S on $\mathbf{R}^n$ is diagonizable according to the spectral theorem. This means that there is an ON-basis $e_1, e_2, \ldots, e_n$ consisting of eigenvectors of S. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ denote the corresponding eigenvalues.

The largest and the smallest eigenvalue $\lambda_{\max}$ and $\lambda_{\min}$ are obtained as maximum and minimum values, respectively, of the quadratic form $\langle x, Sx\rangle$ on the unit sphere $\|x\| = 1$:

$$\lambda_{\max} = \max_{\|x\|=1} \langle x, Sx\rangle \quad \text{and} \quad \lambda_{\min} = \min_{\|x\|=1} \langle x, Sx\rangle.$$

For, by using the expansion $x = \sum_{i=1}^n \xi_i e_i$ of x in the ON-basis of eigenvectors, we obtan the inequality

$$\langle x, Sx\rangle = \sum_{i=1}^n \lambda_i \xi_i^2 \le \lambda_{\max} \sum_{i=1}^n \xi_i^2 = \lambda_{\max} \|x\|^2,$$

and equality prevails when x is equal to the eigenvector e_i that corresponds to the eigenvalue $\lambda_{\max}$. An analogous inequality in the other direction holds for $\lambda_{\min}$, of course.

The operator norm (with respect to the Euclidean norm) moreover satisfies the equality

$$\|S\| = \max_{1\le i\le n} |\lambda_i| = \max\{|\lambda_{\max}|, |\lambda_{\min}|\}.$$

For, by using the above expansion of x, we have $Sx = \sum_{i=1}^n \lambda_i \xi_i e_i$, and consequently

$$\|Sx\|^2 = \sum_{i=1}^n \lambda_i^2 \xi_i^2 \le \max_{1\le i\le n} |\lambda_i|^2 \sum_{i=1}^n \xi_i^2 = (\max_{1\le i\le n} |\lambda_i|)^2 \, \|x\|^2,$$

with equality when x is the eigenvector that corresponds to $\max_i |\lambda_i|$.

If all eigenvalues of the symmetric operator S are nonzero, then S is invertible, and the inverse S^{-1} is symmetric with eigenvalues $\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_n^{-1}$. The norm of the inverse is given by

$$\|S^{-1}\| = 1/\min_{1\le i\le n} |\lambda_i|.$$

A symmetric operator S is positive semidefinite if all its eigenvalues are nonnegative, and it is positive definite if all eigenvalues are positive. Hence, if S is positive definite, then

$$\|S\| = \lambda_{\max} \qquad \text{and} \qquad \|S^{-1}\| = 1/\lambda_{\min}.$$

It follows easily from the diagonizability of symmetric operators on $\mathbf{R}^n$ that every positive semidefinite symmetric operator S has a unique positive semidefinite symmetric square root $S^{1/2}$. Moreover, since

$$\langle x, Sx\rangle = \langle x, S^{1/2}(S^{1/2}x)\rangle = \langle S^{1/2}x, S^{1/2}x\rangle = \|S^{1/2}x\|$$

we conclude that the two operators S and $S^{1/2}$ have the same null space $\mathcal{N}(S)$ and that

$$\mathcal{N}(S) = \{x \in \mathbf{R}^n \mid Sx = 0\} = \{x \in \mathbf{R}^n \mid \langle x, Sx\rangle = 0\}.$$

Differentiability

A function $f\colon U \to \mathbf{R}$, which is defined on an open subset U of $\mathbf{R}^n$, is called *differentiable at the point* $a \in U$ if the partial derivatives $\frac{\partial f}{\partial x_i}$ exist at the point x and the equality

$$f(a+v) = f(a) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a)\, v_i + r(v) \tag{1.2}$$

holds for all v in some neighborhood of the origin with a remainder term $r(v)$ that satisfies the condition

$$\lim_{v\to 0} \frac{r(v)}{\|v\|} = 0.$$

The linear form $Df(a)[v]$, defined by

$$Df(a)[v] = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a)\, v_i,$$

is called the *differential* of the function f at the point a. The coefficient vector

$$\Big(\frac{\partial f}{\partial x_1}(a), \frac{\partial f}{\partial x_2}(a), \ldots, \frac{\partial f}{\partial x_n}(a)\Big)$$

of the differential is called the *derivative* or the *gradient* of f at the point a and is denoted by $f'(a)$ or $\nabla f(a)$. We shall mostly use the first mentioned notation.

The equation (1.2) can now be written in a compact form as

$$f(a+v) = f(a) + Df(a)[v] + r(v),$$

with

$$Df(a)[v] = \langle f'(a), v\rangle.$$

A function $f\colon U \to \mathbf{R}$ is called *differentiable (on U)* if it is differentiable at each point in U. In particular, this implies that U is an open set.

For functions of one variable, differentiability is clearly equivalent to the existence of the derivative, but for functions of several variables, the mere existence of the partial derivatives is no longer a guarantee for differentiability. However, if a function f has partial derivatives and these are continous on an open set U, then f is differentiable on U.

The Mean Value Theorem

Suppose $f\colon U \to \mathbf{R}$ is a differentiable function and that the line segment $[a, a+v]$ lies in U. Let $\phi(t) = f(a+tv)$. The function ϕ is then defined and differentiable on the interval $[0,1]$ with derivative

$$\phi'(t) = Df(a+tv)[v] = \langle f'(a+tv), v\rangle.$$

This is a special case of the chain rule but also follows easily from the definition of the derivative. By the usual mean value theorem for functions of one variable, there is a number $s \in]0,1[$ such that $\phi(1) - \phi(0) = \phi'(s)(1-0)$. Since $\phi(1) = f(a+v)$, $\phi(0) = f(a)$ and $a+sv$ is a point on the open line segment $]a, a+v[$, we have now deduced the following *mean value theorem* for functions of several variables.

Theorem 1.1.1. *Suppose the function $f\colon U \to \mathbf{R}$ is differentiable and that the line segment $[a, a+v]$ lies in U. Then there is a point $c \in]a, a+v[$ such that*

$$f(a+v) = f(a) + Df(c)[v].$$

Functions with Lipschitz continuous derivative

We shall sometimes need more precise information about the remainder term $r(v)$ in equation (1.2) than what follows from the definition of differentiability. We have the following result for functions with a Lipschitz continuous derivative.

Theorem 1.1.2. *Suppose the function $f\colon U \to \mathbf{R}$ is differentiable, that its derivative is Lipschitz continuous, i.e. that $\|f'(y) - f'(x)\| \le L\|y-x\|$ for all $x, y \in U$, and that the line segment $[a, a+v]$ lies in U. Then*

$$|f(a+v) - f(a) - Df(a)[v]| \le \frac{L}{2}\,\|v\|^2.$$

Proof. Define the function Φ on the interval $[0,1]$ by

$$\Phi(t) = f(a+tv) - t\,Df(a)[v].$$

Then Φ is differentiable with derivative

$$\Phi'(t) = Df(a+tv)[v] - Df(a)[v] = \langle f'(a+tv) - f'(a), v\rangle,$$

and by using the Cauchy–Schwarz inequality and the Lipschitz continuity, we obtain the inequality

$$|\Phi'(t)| \le \|f'(a+tv) - f'(a)\| \cdot \|v\| \le Lt\,\|v\|^2.$$

Since $f(a+v) - f(a) - Df(a)[v] = \Phi(1) - \Phi(0) = \int_0^1 \Phi'(t)\,dt$, it now follows that

$$|f(a+v) - f(a) - Df(a)[v]| \le \int_0^1 |\Phi'(t)|\,dt \le L\|v\|^2 \int_0^1 t\,dt = \frac{L}{2}\,\|v\|^2. \qquad \square$$

Two times differentiable functions

If the function f together with all its partial derivatives $\frac{\partial f}{\partial x_i}$ are differentiable on U, then f is said to be two times differentiable on U. The mixed partial second derivatives are then automatically equal, i.e.

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(a) = \frac{\partial^2 f}{\partial x_j \partial x_i}(a)$$

for all i, j and all $a \in U$.

A sufficient condition for the function f to be two times differentiable on U is that all partial derivatives of order up to two exist and are continuous on U.

If $f\colon U \to \mathbf{R}$ is a two times differentiable function and a is a point in U, we define a symmetric bilinear form $D^2 f(a)[u,v]$ on $\mathbf{R}^n$ by

$$D^2 f(a)[u,v] = \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(a) u_i v_j, \quad u, v \in \mathbf{R}^n.$$

The corresponding symmetric linear operator is called the *second derivative* of f at the point a and it is denoted by $f''(a)$. The matrix of the second derivative, i.e. the matrix

$$\Big[\frac{\partial^2 f}{\partial x_i \partial x_j}(a)\Big]_{i,j=1}^{n},$$

is called the *hessian* of f (at the point a). Since we do not distinguish between matrices and operators, we also denote the hessian by $f''(a)$.

The above symmetric bilinear form can now be expressed in the form

$$D^2 f(a)[u,v] = \langle u, f''(a)v\rangle = u^T f''(a)v,$$

depending on whether we interpret the second derivative as an operator or as a matrix.

Let us recall *Taylor's formula*, which reads as follows for two times differentiable functions.

Theorem 1.1.3. *Suppose the function f is two times differentiable in a neighborhood of the point a. Then*

$$f(a+v) = f(a) + Df(a)[v] + \tfrac{1}{2}D^2 f(a)[v,v] + r(v)$$

with a remainder term that satisfies $\lim_{v\to 0} r(v)/\|v\|^2 = 0$.

Three times differentiable functions

To define self-concordance we also need to consider functions that are three times differentiable on some open subset U of $\mathbf{R}^n$. For such functions f and points $a \in U$ we define a trilinear form $D^3 f(a)[u,v,w]$ in the vectors $u, v, w \in \mathbf{R}^n$ by

$$D^3 f(a)[u,v,w] = \sum_{i,j,k=1}^{n} \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}(a) u_i v_j w_k.$$

We leave to the reader to formulate Taylor's formula for functions that are three times differentiable. We have the following differentiation rules, which follow from the chain rule and will be used several times in the final chapters:

$$\begin{aligned}
\frac{d}{dt} f(x+tv) &= Df(x+tv)[v] \\
\frac{d}{dt}\Big(Df(x+tv)[u]\Big) &= D^2 f(x+tv)[u,v], \\
\frac{d}{dt}\Big(D^2 f(x+tw)[u,v]\Big) &= D^3 f(x+tw)[u,v,w].
\end{aligned}$$

As a consequence we get the following expressions for the derivatives of the restriction ϕ of the function f to the line through the point x with the direction given by v:

$$\begin{aligned}
\phi(t) &= f(x+tv), \\
\phi'(t) &= Df(x+tv)[v], \\
\phi''(t) &= D^2 f(x+tv)[v,v], \\
\phi'''(t) &= D^3 f(x+tv)[v,v,v].
\end{aligned}$$

Chapter 2

Convex sets

2.1 Affine sets and affine maps

Affine sets

Definition. A subset of $\mathbf{R}^n$ is called *affine* if for each pair of distinct points in the set it contains the entire line through the points.

Thus, a set X is affine if and only if

$$x, y \in X,\ \lambda \in \mathbf{R} \ \Rightarrow\ \lambda x + (1-\lambda)y \in X.$$

The empty set $\emptyset$, the entire space $\mathbf{R}^n$, linear subspaces of $\mathbf{R}^n$, singleton sets $\{x\}$ and lines are examples of affine sets.

Definition. A linear combination $y = \sum_{j=1}^m \alpha_j x_j$ of vectors $x_1, x_2, \ldots, x_m$ is called an *affine combination* if $\sum_{j=1}^m \alpha_j = 1$.

Theorem 2.1.1. *An affine set contains all affine combination of its elements.*

Proof. We prove the theorem by induction on the number of elements in the affine combination. So let X be an affine set. An affine combination of one element is the element itself. Hence, X contains all affine combinations that can be formed by one element in the set.

Now assume inductively that X contains all affine combinations that can be formed out of $m-1$ elements from X, where $m \geq 2$, and consider an arbitrary affine combination $x = \sum_{j=1}^m \alpha_j x_j$ of m elements $x_1, x_2, \ldots, x_m$ in X. Since $\sum_{j=1}^m \alpha_j = 1$, at least one coefficient α_j must be different from 1; assume without loss of generality that $\alpha_m \neq 1$, and let $s = 1 - \alpha_m = \sum_{j=1}^{m-1} \alpha_j$.

Then $s \neq 0$ and $\sum_{j=1}^{m-1} \alpha_j/s = 1$, which means that the element

$$y = \sum_{j=1}^{m-1} \frac{\alpha_j}{s} x_j$$

is an affine combination of $m-1$ elements in X. Therefore, y belongs to X, by the induction assumption. But $x = sy+(1-s)x_m$, and it now follows from the definition of affine sets that x lies in X. This completes the induction step, and the theorem is proved. □

Definition. Let A be an arbitrary nonempty subset of $\mathbf{R}^n$. The set of all affine combinations $\lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_m a_m$ that can be formed of an arbitrary number of elements $a_1, a_2, \ldots, a_m$ from A, is called the *affine hull* of A and is denoted by aff A .

In order to have the affine hull defined also for the empty set, we put $\operatorname{aff} \emptyset = \emptyset$.

Theorem 2.1.2. *The affine hull* aff A *is an affine set containing* A *as a subset, and it is the smallest affine subset with this property, i.e. if the set* X *is affine and* $A \subseteq X$*, then* aff $A \subseteq X$*.*

Proof. The set aff A is an affine set, because any affine combination of two elements in aff A is obviously an affine combination of elements from A, and the set A is a subset of its affine hull, since any element is an affine combination of itself.

If X is an affine set, then aff $X \subseteq X$, by Theorem 2.1.1, and if $A \subseteq X$, then obviously aff $A \subseteq$ aff X. Thus, aff $A \subseteq X$ whenever X is an affine set and A is a subset of X. □

Characterisation of affine sets

Nonempty affine sets are translations of linear subspaces. More precisely, we have the following theorem.

Theorem 2.1.3. *If* X *is an affine subset of* $\mathbf{R}^n$ *and* $a \in X$*, then* $-a+X$ *is a linear subspace of* $\mathbf{R}^n$*. Moreover, for each* $b \in X$ *we have* $-b+X = -a+X$*.*

Thus, to each nonempty affine set X *there corresponds a uniquely defined linear subspace* U *such that* $X = a + U$*.*

Proof. Let $U = -a + X$. If $u_1 = -a + x_1$ and $u_2 = -a + x_2$ are two elements in U and α_1, α_2 are arbitrary real numbers, then the linear combination

$$\alpha_1 u_1 + \alpha_2 u_2 = -a + (1 - \alpha_1 - \alpha_2)a + \alpha_1 x_1 + \alpha_2 x_2$$

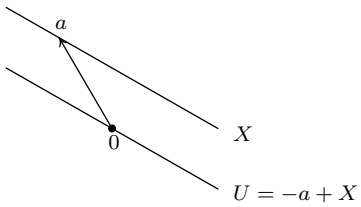

Figure 2.1. Illustration for Theorem 2.1.3: An affine set X and the corresponding linear subspace U.

is an element in U, because $(1-\alpha_1-\alpha_2)a+\alpha_1 x_1+\alpha_2 x_2$ is an affine combination of elements in X and hence belongs to X, according to Theorem 2.1.1. This proves that U is a linear subspace.

Now assume that $b \in X$, and let $v = -b + x$ be an arbitrary element in $-b + X$. By writing v as $v = -a + (a - b + x)$ we see that v belongs to $-a + X$, too, because $a - b + x$ is an affine combination of elements in X. This proves the inclusion $-b + X \subseteq -a + X$. The converse inclusion follows by symmetry. Thus, $-a + X = -b + X$. □

Dimension

The following definition is justified by Theorem 2.1.3.

Definition. The *dimension* $\dim X$ of a nonempty affine set X is defined as the dimension of the linear subspace $-a+X$, where a is an arbitrary element in X.

Since every nonempty affine set has a well-defined dimension, we can extend the dimension concept to arbitrary nonempty sets as follows.

Definition. The *(affine) dimension* $\dim A$ of a nonempty subset A of $\mathbf{R}^n$ is defined to be the dimension of its affine hull $\operatorname{aff} A$.

The dimension of an open ball $B(a; r)$ in $\mathbf{R}^n$ is n, and the dimension of a line segment $[x, y]$ is 1.

The dimension is *invariant under translation* i.e. if A is a nonempty subset of $\mathbf{R}^n$ and $a \in \mathbf{R}^n$ then

$$\dim(a + A) = \dim A,$$

and it is *increasing* in the following sense:

$$A \subseteq B \Rightarrow \dim A \leq \dim B.$$

Affine sets as solutions to systems of linear equations

Our next theorem gives a complete description of the affine subsets of $\mathbf{R}^n$.

Theorem 2.1.4. *Every affine subset of $\mathbf{R}^n$ is the solution set of a system of linear equations*

$$\begin{cases} c_{11}x_1 + c_{12}x_2 + \cdots + c_{1n}x_n = b_1 \\ c_{21}x_1 + c_{22}x_2 + \cdots + c_{2n}x_n = b_2 \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ c_{m1}x_1 + c_{m2}x_2 + \cdots + c_{mn}x_n = b_m \end{cases}$$

and conversely. The dimension of a nonempty solution set equals $n-r$, where r is the rank of the coefficient matrix C.

Proof. The empty affine set is obtained as the solution set of an inconsistent system. Therefore, we only have to consider nonempty affine sets X, and these are of the form $X = x_0 + U$, where x_0 belongs to X and U is a linear subspace of $\mathbf{R}^n$. But each linear subspace is the solution set of a homogeneous system of linear equations. Hence there exists a matrix C such that

$$U = \{x \mid Cx = 0\},$$

and $\dim U = n - \operatorname{rank} C$. With $b = Cx_0$ it follows that $x \in X$ if and only if $Cx - Cx_0 = C(x - x_0) = 0$, i.e. if and only if x is a solution to the linear system $Cx = b$.

Conversely, if x_0 is a solution to the above linear system so that $Cx_0 = b$, then x is a solution to the same system if and only if the vector $z = x - x_0$ belongs to the solution set U of the homogeneous equation system $Cz = 0$. It follows that the solution set of the equation system $Cx = b$ is of the form $x_0 + U$, i.e. it is an affine set. □

Hyperplanes

Definition. Affine subsets of $\mathbf{R}^n$ of dimension $n - 1$ are called *hyperplanes*.

Theorem 2.1.4 has the following corollary:

Corollary 2.1.5. *A subset X of $\mathbf{R}^n$ is a hyperplane if and only if there exist a nonzero vector $c = (c_1, c_2, \ldots, c_n)$ and a real number b so that*

$$X = \{x \in \mathbf{R}^n \mid \langle c, x\rangle = b\}.$$

It follows from Theorem 2.1.4 that every affine proper subset of $\mathbf{R}^n$ can be expressed as an intersection of hyperplanes.

Affine maps

Definition. Let X be an affine subset of $\mathbf{R}^n$. A map $T\colon X \to \mathbf{R}^m$ is called *affine* if

$$T(\lambda x + (1-\lambda)y) = \lambda Tx + (1-\lambda)Ty$$

for all $x, y \in X$ and all $\lambda \in \mathbf{R}$.

Using induction, it is easy to prove that if $T\colon X \to \mathbf{R}^m$ is an affine map and $x = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m$ is an affine combination of elements in X, then

$$Tx = \alpha_1 Tx_1 + \alpha_2 Tx_2 + \cdots + \alpha_m Tx_m.$$

Moreover, the image $T(Y)$ of an affine subset Y of X is an affine subset of $\mathbf{R}^m$, and the inverse image $T^{-1}(Z)$ of an affine subset Z of $\mathbf{R}^m$ is an affine subset of X.

The composition of two affine maps is affine. In particular, a linear map followed by a translation is an affine map, and our next theorem shows that each affine map can be written as such a composition.

Theorem 2.1.6. *Let X be an affine subset of $\mathbf{R}^n$, and suppose the map $T\colon X \to \mathbf{R}^m$ is affine. Then there exist a linear map $C\colon \mathbf{R}^n \to \mathbf{R}^m$ and a vector v in $\mathbf{R}^m$ so that*

$$Tx = Cx + v$$

for all $x \in X$.

Proof. Write the domain of T in the form $X = x_0 + U$ with $x_0 \in X$ and U as a linear subspace of $\mathbf{R}^n$, and define the map C on the subspace U by

$$Cu = T(x_0 + u) - Tx_0.$$

Then, for each $u_1, u_2 \in U$ and $\alpha_1, \alpha_2 \in \mathbf{R}$ we have

$$\begin{aligned} C(\alpha_1 u_1 + \alpha_2 u_2) &= T(x_0 + \alpha_1 u_1 + \alpha_2 u_2) - Tx_0 \\ &= T\big(\alpha_1(x_0 + u_1) + \alpha_2(x_0 + u_2) + (1 - \alpha_1 - \alpha_2)x_0\big) - Tx_0 \\ &= \alpha_1 T(x_0 + u_1) + \alpha_2 T(x_0 + u_2) + (1 - \alpha_1 - \alpha_2)Tx_0 - Tx_0 \\ &= \alpha_1\big(T(x_0 + u_1) - Tx_0\big) + \alpha_2\big(T(x_0 + u_2) - Tx_0\big) \\ &= \alpha_1 Cu_1 + \alpha_2 Cu_2. \end{aligned}$$

So the map C is linear on U and it can, of course, be extended to a linear map on all of $\mathbf{R}^n$.

For $x \in X$ we now obtain, since $x - x_0$ belongs to U,

$$Tx = T(x_0 + (x - x_0)) = C(x - x_0) + Tx_0 = Cx - Cx_0 + Tx_0,$$

which proves the theorem with v equal to $Tx_0 - Cx_0$. □

2.2 Convex sets

Basic definitions and properties

Definition. A subset X of $\mathbf{R}^n$ is called *convex* if $[x, y] \subseteq X$ for all x, $y \in X$.

In other words, a set X is convex if and only if it contains the line segment between each pair of its points.

Figure 2.2. A convex set and a non-convex set

Example 2.2.1. Affine sets are obviously convex. In particular, the empty set $\emptyset$, the entire space $\mathbf{R}^n$ and linear subspaces are convex sets. Open line segments and closed line segments are clearly convex. □

Example 2.2.2. Open balls $B(a; r)$ (with respect to arbitrary norms $\|\cdot\|$) are convex sets. This follows from the triangle inequality and homogenouity, for if $x, y \in B(a; r)$ and $0 \leq \lambda \leq 1$, then

$$\begin{aligned}\|\lambda x + (1-\lambda)y - a\| &= \|\lambda(x-a) + (1-\lambda)(y-a)\| \\ &\leq \lambda\|x-a\| + (1-\lambda)\|y-a\| < \lambda r + (1-\lambda)r = r,\end{aligned}$$

which means that each point $\lambda x + (1-\lambda)y$ on the segment $[x, y]$ lies in $B(a; r)$.

The corresponding closed balls $\overline{B}(a; r) = \{x \in \mathbf{R}^n \mid \|x - a\| \leq r\}$ are of course convex, too. □

Definition. A linear combination $y = \sum_{j=1}^m \alpha_j x_j$ of vectors $x_1, x_2, \ldots, x_m$ is called a *convex combination* if $\sum_{j=1}^m \alpha_j = 1$ and $\alpha_j \geq 0$ for all j.

Theorem 2.2.1. *A convex set contains all convex combinations of its elements.*

Proof. Let X be an arbitrary convex set. A convex combination of one element is the element itself, and hence X contains all convex combinations formed by just one element of the set. Now assume inductively that X contains all convex combinations that can be formed by $m-1$ elements of X, and consider an arbitrary convex combination $x = \sum_{j=1}^m \alpha_j x_j$ of $m \geq 2$

elements $x_1, x_2, \ldots, x_m$ in X. Since $\sum_{j=1}^{m} \alpha_j = 1$, some coefficient α_j must be strictly less than 1, and assume without loss of generality that $\alpha_m < 1$, and let $s = 1 - \alpha_m = \sum_{j=1}^{m-1} \alpha_j$. Then $s > 0$ and $\sum_{j=1}^{m-1} \alpha_j/s = 1$, which means that

$$y = \sum_{j=1}^{m-1} \frac{\alpha_j}{s} x_j$$

is a convex combination of $m-1$ elements in X. By the induction hypothesis, y belongs to X. But $x = sy+(1-s)x_m$, and it now follows from the convexity definition that x belongs to X. This completes the induction step and the proof of the theorem. □

2.3 Convexity preserving operations

We now describe a number of ways to construct new convex sets from given ones.

Image and inverse image under affine maps

Theorem 2.3.1. *Let $T\colon V \to \mathbf{R}^m$ be an affine map.*

(i) The image $T(X)$ of a convex subset X of V is convex.

(ii) The inverse image $T^{-1}(Y)$ of a convex subset Y of $\mathbf{R}^m$ is convex.

Proof. (i) Suppose y_1, $y_2 \in T(X)$ and $0 \leq \lambda \leq 1$. Let x_1, x_2 be points in X such that $y_i = T(x_i)$. Since

$$\lambda y_1 + (1-\lambda)y_2 = \lambda T x_1 + (1-\lambda) T x_2 = T(\lambda x_1 + (1-\lambda)x_2)$$

and $\lambda x_1 + (1-\lambda)x_2$ lies X, it follows that $\lambda y_1 + (1-\lambda)y_2$ lies in $T(X)$. This proves that the image set $T(X)$ is convex.

(ii) To prove the convexity of the inverse image $T^{-1}(Y)$ we instead assume that $x_1, x_2 \in T^{-1}(Y)$, i.e. that $Tx_1, Tx_2 \in Y$, and that $0 \leq \lambda \leq 1$. Since Y is a convex set,

$$T(\lambda x_1 + (1-\lambda)x_2) = \lambda T x_1 + (1-\lambda) T x_2$$

is an element of Y, and this means that $\lambda x_1 + (1-\lambda)x_2$ lies in $T^{-1}(Y)$. □

As a special case of the preceding theorem it follows that translations $a + X$ of a convex set X are convex.

EXAMPLE 2.3.1. The sets

$$\{x \in \mathbf{R}^n \mid \langle c, x\rangle \geq b\} \quad \text{and} \quad \{x \in \mathbf{R}^n \mid \langle c, x\rangle \leq b\},$$

where b is an arbitrary real number and $c = (c_1, c_2, \ldots, c_n)$ is an arbirary nonzero vector, are called opposite *closed halfspaces*. Their complements, i.e.

$$\{x \in \mathbf{R}^n \mid \langle c, x\rangle < b\} \quad \text{and} \quad \{x \in \mathbf{R}^n \mid \langle c, x\rangle > b\},$$

are called *open halfspaces*.

The halfspaces $\{x \in \mathbf{R}^n \mid \langle c, x\rangle \geq b\}$ and $\{x \in \mathbf{R}^n \mid \langle c, x\rangle > b\}$ are inverse images of the real intervals $[b, \infty[$ and $]b, \infty[$, respectively, under the linear map $x \mapsto \langle c, x\rangle$. It therefore follows from Theorem 2.3.1 that *halfspaces are convex sets.* □

Intersection and union

Theorem 2.3.2. *Let $\{X_i \mid i \in I\}$ be a family of convex subsets of $\mathbf{R}^n$. The intersection $\bigcap\{X_i \mid i \in I\}$ is a convex set.*

Proof. Suppose x, y are points in the intersection Y. The definition of an intersection implies that x and y lie in X_i for all indices $i \in I$, and convexity implies that $[x, y] \subseteq X_i$ for all $i \in I$. Therefore, $[x, y] \subseteq Y$, again by the definition of set intersection. This proves that the intersection is a convex set. □

A union of convex sets is, of course, in general not convex. However, there is a trivial case when convexity is preserved, namely when the sets can be ordered in such a way as to form an "increasing chain".

Theorem 2.3.3. *Suppose $\{X_i \mid i \in I\}$ is a family of convex sets X_i and that for each pair $i, j \in I$ either $X_i \subseteq X_j$ or $X_j \subseteq X_i$. The union $\bigcup\{X_i \mid i \in I\}$ is then a convex set.*

Proof. The assumptions imply that, for each pair of points x, y in the union there is an index $i \in I$ such that both points belong to X_i. By convexity, the entire segment $[x, y]$ lies in X_i, and thereby also in the union. □

EXAMPLE 2.3.2. The convexity of closed balls follows from the convexity of open balls, because $\overline{B}(a; r_0) = \bigcap\{B(a; r) \mid r > r_0\}$.

Conversely, the convexity of open balls follows from the convexity of closed balls, since $B(a; r_0) = \bigcup\{\overline{B}(a; r) \mid r < r_0\}$ and the sets $\overline{B}(a; r)$ form an increasing chain. □

Definition. A subset X of $\mathbf{R}^n$ is called a *polyhedron* if X can be written as an intersection of finitely many closed halfspaces or if $X = \mathbf{R}^n$.[†]

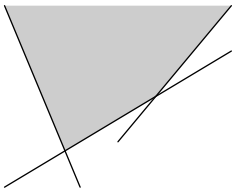

Figure 2.3. A polyhedron in $\mathbf{R}^2$

Polyhedra are convex sets because of Theorem 2.3.2, and they can be represented as solution sets to systems of linear inequalities. By multiplying some of the inequalities by -1, if necessary, we may without loss of generality assume that all inequalities are of the form $c_1x_1 + c_2x_2 + \cdots + c_nx_n \geq d$. This means that every polyedron is the solution set to a system of the following form

$$\begin{cases} c_{11}x_1 + c_{12}x_2 + \cdots + c_{1n}x_n \geq b_1 \\ c_{21}x_1 + c_{22}x_2 + \cdots + c_{2n}x_n \geq b_2 \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ c_{m1}x_1 + c_{m2}x_2 + \cdots + c_{mn}x_n \geq b_m, \end{cases}$$

or in matrix notation

$$Cx \geq b.$$

The intersection of finitely many polyhedra is clearly a polyhedron. Since each hyperplane is the intersection of two opposite closed halfspaces, and each affine set (except the entire space) is the intersection of finitely many hyperplanes, it follows especially that affine sets are polyhedra. In particular, the empty set is a polyhedron.

Cartesian product

Theorem 2.3.4. *The Cartesian product $X \times Y$ of two convex sets X and Y is a convex set.*

Proof. Suppose X lies in $\mathbf{R}^n$ and Y lies in $\mathbf{R}^m$. The projections

$$P_1 \colon \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n \quad \text{and} \quad P_2 \colon \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^m,$$

[†]The intersection of an empty family of sets is usually defined as the entire space, and using this convention the polyhedron $\mathbf{R}^n$ can also be viewed as an intersection of halfspaces.

defined by $P_1(x, y) = x$ and $P_2(x, y) = y$, are linear maps, and

$$X \times Y = (X \times \mathbf{R}^m) \cap (\mathbf{R}^n \times Y) = P_1^{-1}(X) \cap P_2^{-1}(Y).$$

The assertion of the theorem is therefore a consequence of Theorem 2.3.1 and Theorem 2.3.2. □

Sum

Theorem 2.3.5. *The sum $X + Y$ of two convex subsets X and Y of $\mathbf{R}^n$ is convex, and the product αX of a number α and a convex set X is convex.*

Proof. The maps $S\colon \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}^n$ and $T\colon \mathbf{R}^n \to \mathbf{R}^n$, defined by $S(x, y) = x + y$ and $Tx = \alpha x$, are linear. Since $X + Y = S(X \times Y)$ and $\alpha X = T(X)$, our assertions follow from Theorems 2.3.1 and 2.3.4. □

Example 2.3.3. The set $X(r)$ of all points whose distance to a given set X is less than the positive number r, can be written as a sum, namely

$$X(r) = X + B(0; r).$$

Since open balls are convex, we conclude from Theorem 2.3.5 that the set $X(r)$ is convex if X is a convex set. □

Image and inverse image under the perspective map

Definition. The *perspective map* $P\colon \mathbf{R}^n \times \mathbf{R}_{++} \to \mathbf{R}^n$ is defined by

$$P(x, t) = t^{-1}x$$

for $x \in \mathbf{R}^n$ and $t > 0$.

The perspective map thus first rescales points in $\mathbf{R}^n \times \mathbf{R}_{++}$ so that the last coordinate becomes 1 and then throws the last coordinate away. Figure 2.4 illustrates the process.

Theorem 2.3.6. *Let X be a convex subset of $\mathbf{R}^n \times \mathbf{R}_{++}$ and Y be a convex subset of $\mathbf{R}^n$. The image $P(X)$ of X and the inverse image $P^{-1}(Y)$ of Y under the perspective map $P\colon \mathbf{R}^n \times \mathbf{R}_{++} \to \mathbf{R}^n$ are convex sets.*

Proof. To prove that the image $P(X)$ is convex we assume that $y, y' \in P(X)$ and have to prove that the point $\lambda y + (1 - \lambda)y'$ lies in $P(X)$ if $0 < \lambda < 1$.

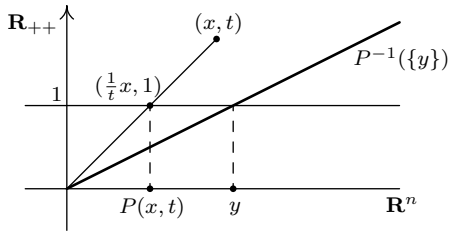

Figure 2.4. The perspective map P. The inverse image of a point $y \in \mathbf{R}^n$ is a halfline.

To achieve this we first note that there exist numbers $t, t' > 0$ such that the points (ty, t) and $(t'y', t')$ belong to X, and then define

$$\alpha = \frac{\lambda t'}{\lambda t' + (1-\lambda)t}.$$

Clearly $0 < \alpha < 1$, and it now follows from the convexity of X that the point

$$z = \alpha(ty, t) + (1-\alpha)(t'y', t') = \Big(\frac{tt'(\lambda y + (1-\lambda)y')}{\lambda t' + (1-\lambda)t}, \frac{tt'}{\lambda t' + (1-\lambda)t}\Big)$$

lies X. Thus, $P(z) \in P(X)$, and since $P(z) = \lambda y + (1-\lambda)y'$, we are done.

To prove that the inverse image $P^{-1}(Y)$ is convex, we instead assume that (x,t) and (x',t') are points in $P^{-1}(Y)$ and that $0 < \lambda < 1$. We will prove that the point $\lambda(x,t) + (1-\lambda)(x',t')$ lies in $P^{-1}(Y)$.

To this end we note that the the points $\frac{1}{t}x$ and $\frac{1}{t'}x'$ belong to Y and that

$$\alpha = \frac{\lambda t}{\lambda t + (1-\lambda)t'}$$

is a number between 0 and 1. Thus,

$$z = \alpha\frac{1}{t}x + (1-\alpha)\frac{1}{t'}x' = \frac{\lambda x + (1-\lambda)x'}{\lambda t + (1-\lambda)t'}$$

is a point in Y by convexity, and consequently $\big((\lambda t + (1-\lambda)t')z, \lambda t + (1-\lambda)t'\big)$ is a point in $P^{-1}(Y)$. But

$$\big((\lambda t + (1-\lambda)t')z, \lambda t + (1-\lambda)t'\big) = \lambda(x,t) + (1-\lambda)(x',t')$$

and this completes the proof. □

EXAMPLE 2.3.4. The set $\{(x, x_{n+1}) \in \mathbf{R}^n \times \mathbf{R} \mid \|x\| < x_{n+1}\}$ is the inverse image of the unit ball $B(0;1)$ under the perspective map, and it is therefore a convex set in $\mathbf{R}^{n+1}$ for each particular choice of norm $\|\cdot\|$. The following convex sets are obtained by choosing the ℓ^1-norm, the Euclidean norm and the maximum norm, respectively, as norm:

$$\{x \in \mathbf{R}^{n+1} \mid x_{n+1} > |x_1| + |x_2| + \cdots + |x_n|\},$$
$$\{x \in \mathbf{R}^{n+1} \mid x_{n+1} > (x_1^2 + x_2^2 + \cdots + x_n^2)^{1/2}\} \quad \text{and}$$
$$\{x \in \mathbf{R}^{n+1} \mid x_{n+1} > \max_{1 \le i \le n} |x_i|\}.$$ □

2.4 Convex hull

Definition. Let A be a nonempty set in $\mathbf{R}^n$. The set of all convex combinations $\lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_m a_m$ of an arbitrary number of elements $a_1, a_2, \ldots, a_m$ in A is called the *convex hull* of A and is denoted by $\operatorname{cvx} A$.

Moreover, to have the convex hull defined for the empty set, we define $\operatorname{cvx} \emptyset = \emptyset$.

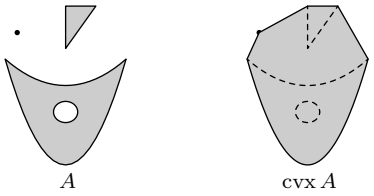

Figure 2.5. A set and its convex hull

Theorem 2.4.1. *The convex hull* $\operatorname{cvx} A$ *is a convex set containing* A*, and it is the smallest set with this property, i.e. if* X *is a convex set and* $A \subseteq X$*, then* $\operatorname{cvx} A \subseteq X$*.*

Proof. $\operatorname{cvx} A$ is a convex set, because convex combinations of two elements of the type $\sum_{j=1}^m \lambda_j a_j$, where $m \ge 1$, $\lambda_1, \lambda_2, \ldots, \lambda_m \ge 0$, $\sum_{j=1}^m \lambda_j = 1$ and $a_1, a_2, \ldots, a_m \in A$, is obviously an element of the same type. Moreover, $A \subseteq \operatorname{cvx} A$, because each element in A is a convex combination of itself ($a = 1a$).

A convex set X contains all convex combinations of its elements, according to Theorem 2.2.1. If $A \subseteq X$, then in particular X contains all convex combinations of elements in A, which means that $\operatorname{cvx} A \subseteq X$. □

The convex hull of a set in $\mathbf{R}^n$ consists of all convex combinations of an arbitrary number of elements in the set, but each element of the hull is actually a convex combination of at most $n+1$ elements.

Theorem 2.4.2. *Let $A \subseteq \mathbf{R}^n$ and suppose that $x \in \operatorname{cvx} A$. Then A contains a subset B with at most $n+1$ elements such that $x \in \operatorname{cvx} B$.*

Proof. According to the definition of convex hull there exists a finite subset B of A such that $x \in \operatorname{cvx} B$. Choose such a subset $B = \{b_1, b_2, \ldots, b_m\}$ with as few elements as possible. By the minimality assumption, $x = \sum_{j=1}^m \lambda_j b_j$ with $\sum_{j=1}^m \lambda_j = 1$ and $\lambda_j > 0$ for all j.

Let $c_j = b_j - b_m$ for $j = 1, 2, \ldots, m-1$. We will show that the set $C = \{c_1, c_2, \ldots, c_{m-1}\}$ is a linearly independent subset of $\mathbf{R}^n$, and this obviously implies that $m \leq n+1$.

Suppose on the contrary that the set C is linearly dependent. Then there exist real numbers μ_j, not all of them equal to 0, such that $\sum_{j=1}^{m-1} \mu_j c_j = 0$. Now let $\mu_m = -\sum_{j=1}^{m-1} \mu_j$; then $\sum_{j=1}^m \mu_j = 0$ and $\sum_{j=1}^m \mu_j b_j = 0$. Moreover, at least one of the m numbers $\mu_1, \mu_2, \ldots, \mu_m$ is positive.

Consider the numbers $\nu_j = \lambda_j - t\mu_j$ for $t > 0$. We note that

$$\sum_{j=1}^m \nu_j = \sum_{j=1}^m \lambda_j - t\sum_{j=1}^m \mu_j = 1 \quad \text{and} \quad \sum_{j=1}^m \nu_j b_j = \sum_{j=1}^m \lambda_j b_j - t\sum_{j=1}^m \mu_j b_j = x.$$

Moreover, $\nu_j \geq \lambda_j > 0$ if $\mu_j \leq 0$, and $\nu_j \geq 0$ if $\mu_j > 0$ and $t \leq \lambda_j/\mu_j$. Therefore, by choosing t as the smallest number of the numbers λ_j/μ_j with positive denominator μ_j, we obtain numbers ν_j such that $\nu_j \geq 0$ for all j and $\nu_{j_0} = 0$ for at least one index j_0. This means that x is a convex combination of elements in the set $B \setminus \{b_{j_0}\}$, which consists of $m-1$ elements. This contradicts the minimality assumption concerning the set B, and our proof by contradiction is finished. □

2.5 Topological properties

Closure

Theorem 2.5.1. *The closure $\operatorname{cl} X$ of a convex set X is convex.*

Proof. We recall that $\operatorname{cl} X = \bigcap_{r>0} X(r)$, where $X(r)$ denotes the set of all points whose distance from X is less than r. The sets $X(r)$ are convex when the set X is convex (see Example 2.3.3), and an intersection of convex sets is convex. □

Interior and relative interior

The interior of a convex set may be empty. For example, line segments in $\mathbf{R}^n$ have no interior points when $n \geq 2$. A necessary and sufficient condition for nonempty interior is given by the following theorem.

Theorem 2.5.2. *A convex subset X of $\mathbf{R}^n$ has interior points if and only if* $\dim X = n$.

Proof. If X has an interior point a, then there exists an open ball $B = B(a; r)$ around a such that $B \subseteq X$, which implies that $\dim X \geq \dim B = n$, i.e. $\dim X = n$.

To prove the converse, let us assume that $\dim X = n$; we will prove that $\operatorname{int} X \neq \emptyset$. Since the dimension of a set is invariant under translations and $\operatorname{int}(a + X) = a + \operatorname{int} X$, we may assume that $0 \in X$.

Let $\{a_1, a_2, \ldots, a_m\}$ be a maximal set of linearly independent vectors in X; then X is a subset of the linear subspace of dimension m which is spanned by these vectors, and it follows from the dimensionality assumption that $m = n$. The set X contains the convex hull of the vectors $0, a_1, a_2, \ldots, a_n$ as a subset, and, in particular, it thus contains the nonempty open set

$$\{\lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_n a_n \mid 0 < \lambda_1 + \cdots + \lambda_n < 1,\ \lambda_1 > 0, \ldots, \lambda_n > 0\}.$$

All points in this latter set are interior points of X, so $\operatorname{int} X \neq \emptyset$. □

A closed line segment $[a, b]$ in the two-dimensional space $\mathbf{R}^2$ has no interior points, but if we consider the line segment as a subset of a line and identify the line with the space $\mathbf{R}$, then it has interior points and its interior is equal to the corresponding open line segment $]a, b[$. A similar remark holds for the convex hull $T = \operatorname{cvx}\{a, b, c\}$ of three noncolinear points in three-space; the triangle T has interior points when viewed as a subset of $\mathbf{R}^2$, but it lacks interior points as a subset of $\mathbf{R}^3$. This conflict is unsatisfactory if we want a concept that is independent of the dimension of the surrounding space, but the dilemma disappears if we use the relative topology that the affine hull of the set inherits from the surrounding space $\mathbf{R}^n$.

Definition. Let X be an m-dimensional subset of $\mathbf{R}^n$, and let V denote the affine hull of X, i.e. V is the m-dimensional affine set that contains X.

A point $x \in X$ is called a *relative interior point* of X if there exists an $r > 0$ such that $B(x; r) \cap V \subseteq X$, and the set of all relative interior points of X is called the *relative interior* of X and is denoted by $\operatorname{rint} X$.

A point $x \in \mathbf{R}^n$ is called a *relative boundary point* of X if, for every $r > 0$, the intersection $B(x; r) \cap V$ contains at least one point from X and at least

one point from $V \setminus X$. The set of all relative boundary points is called the *relative boundary* of X and is denoted by $\operatorname{rbdry} X$.

The relative interior of X is obviously a subset of X, and the relative boundary of X is a subset of the boundary of X. It follows that

$$\operatorname{rint} X \cup \operatorname{rbdry} X \subseteq X \cup \operatorname{bdry} X = \operatorname{cl} X.$$

Conversely, if x is a point in the closure $\operatorname{cl} X$, then for each $r > 0$

$$B(x, r) \cap V \cap X = B(x, r) \cap X \neq \emptyset.$$

Thus, x is either a relative boundary point or a relative interior point of X. This proves the converse inclusion, and we conclude that

$$\operatorname{rint} X \cup \operatorname{rbdry} X = \operatorname{cl} X,$$

or equivalently, that

$$\operatorname{rbdry} X = \operatorname{cl} X \setminus \operatorname{rint} X.$$

It follows from Theorem 2.5.2, with $\mathbf{R}^n$ replaced by $\operatorname{aff} X$, that every nonempty convex set has a nonempty relative interior.

Note that the relative interior of a line segment $[a, b]$ is the corresponding open line segment $]a, b[$, which is consistent with calling $]a, b[$ an open segment. The relative interior of a set $\{a\}$ consisting of just one point is the set itself.

Theorem 2.5.3. *The relative interior* $\operatorname{rint} X$ *of a convex set* X *is convex.*

Proof. The theorem follows as a corollary of the following theorem, since $\operatorname{rint} X \subseteq \operatorname{cl} X$. □

Theorem 2.5.4. *Suppose that* X *is a convex set, that* $a \in \operatorname{rint} X$ *and that* $b \in \operatorname{cl} X$. *The entire open line segment* $]a, b[$ *is then a subset of* $\operatorname{rint} X$.

Proof. Let $V = \operatorname{aff} X$ denote the affine set of least dimension that includes X, and let $c = \lambda a + (1 - \lambda)b$, where $0 < \lambda < 1$, be an arbitrary point on the open segment $]a, b[$. We will prove that c is a relative interior point of X by constructing an open ball B which contains c and whose intersection with V is contained in X.

To this end, we choose $r > 0$ such that $B(a; r) \cap V \subseteq X$ and a point $b' \in X$ such that $\|b' - b\| < \lambda r/(1 - \lambda)$; this is possible since a is a relative interior point of X and b is a point in the closure of X. Let

$$B = \lambda B(a; r) + (1 - \lambda)b',$$

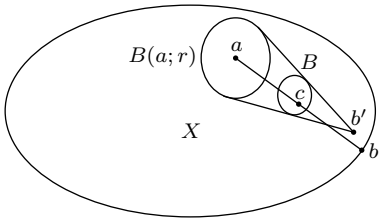

Figure 2.6. Illustration of the proof of Theorem 2.5.4. The convex hull of the ball $B(a;r)$ and the point b' forms a "cone" with nonempty interior that contains the point c.

and observe that $B = B(\lambda a + (1-\lambda)b'; \lambda r)$. The open ball B contains the point c because

$$\|c - (\lambda a + (1-\lambda)b')\| = \|(1-\lambda)(b-b')\| = (1-\lambda)\|b-b'\| < \lambda r.$$

Moreover, $B \cap V = \lambda(B(a;r) \cap V) + (1-\lambda)b' \subseteq \lambda X + (1-\lambda)X \subseteq X$, due to convexity. This completes the proof. □

Theorem 2.5.5. *Let X be a convex set. Then*

(i) $\operatorname{cl}(\operatorname{rint} X) = \operatorname{cl} X$;

(ii) $\operatorname{rint}(\operatorname{cl} X) = \operatorname{rint} X$;

(iii) $\operatorname{rbdry}(\operatorname{cl} X) = \operatorname{rbdry}(\operatorname{rint} X) = \operatorname{rbdry} X$.

Proof. The equalities in (iii) for the relative boundaries follow from the other two and the definition of the relative boundary.

The inclusions $\operatorname{cl}(\operatorname{rint} X) \subseteq \operatorname{cl} X$ and $\operatorname{rint} X \subseteq \operatorname{rint}(\operatorname{cl} X)$ are both trivial, because it follows, for arbitrary sets A and B, that $A \subseteq B$ implies $\operatorname{cl} A \subseteq \operatorname{cl} B$ and $\operatorname{rint} A \subseteq \operatorname{rint} B$.

It thus only remains to prove the two inclusions

$$\operatorname{cl} X \subseteq \operatorname{cl}(\operatorname{rint} X) \quad \text{and} \quad \operatorname{rint}(\operatorname{cl} X) \subseteq \operatorname{rint} X.$$

So fix a point $x_0 \in \operatorname{rint} X$.

If $x \in \operatorname{cl} X$, then every point on the open segment $]x_0, x[$ lies in $\operatorname{rint} X$, by Theorem 2.5.4, and it follows from this that the point x is either an interior point or a boundary point of $\operatorname{rint} X$, i.e. a point in the closure $\operatorname{cl}(\operatorname{rint} X)$. This proves the inclusion $\operatorname{cl} X \subseteq \operatorname{cl}(\operatorname{rint} X)$.

To prove the remaining inclusion $\operatorname{rint}(\operatorname{cl} X) \subseteq \operatorname{rint} X$ we instead assume that $x \in \operatorname{rint}(\operatorname{cl} X)$ and define $y_t = (1-t)x_0 + tx$ for $t > 1$. Since $y_t \to x$ as $t \to 1$, the points y_t belong to $\operatorname{cl} X$ for all t sufficiently close to 1. Choose

a number $t_0 > 1$ such that y_{t_0} belongs to $\operatorname{cl} X$. According to Theorem 2.5.4, all points on the open segment $]x_0, y_{t_0}[$ are relative interior points in X, and x is such a point since $x = \frac{1}{t_0} y_{t_0} + (1 - \frac{1}{t_0}) x_0$. This proves the implication $x \in \operatorname{rint}(\operatorname{cl} X) \Rightarrow x \in \operatorname{rint} X$, and the proof is now complete. □

Compactness

Theorem 2.5.6. *The convex hull* $\operatorname{cvx} A$ *of a compact subset* A *in* $\mathbf{R}^n$ *is compact.*

Proof. Let $S = \{\lambda \in \mathbf{R}^{n+1} \mid \lambda_1, \lambda_2, \ldots, \lambda_{n+1} \geq 0,\ \sum_{j=1}^{n+1} \lambda_j = 1\}$, and let $f \colon S \times \mathbf{R}^n \times \mathbf{R}^n \times \cdots \times \mathbf{R}^n \to \mathbf{R}^n$ be the function

$$f(\lambda, x_1, x_2, \ldots, x_{n+1}) = \sum_{j=1}^{n+1} \lambda_j x_j.$$

The function f is of course continuous, and the set S is compact, since it is closed and bounded. According to Theorem 2.4.2, every element $x \in \operatorname{cvx} A$ can be written as a convex combination $x = \sum_{j=1}^{n+1} \lambda_j a_j$ of at most $n+1$ elements $a_1, a_2, \ldots, a_{n+1}$ from the set A. This means that the convex hull $\operatorname{cvx} A$ coincides with the image $f(S \times A \times A \times \cdots \times A)$ under f of the compact set $S \times A \times A \times \cdots \times A$. Since compactness is preserved by continuous functions, we conclude that the convex hull $\operatorname{cvx} A$ is compact. □

2.6 Cones

Definition. Let x be a point in $\mathbf{R}^n$ different from 0. The set

$$\overrightarrow{x} = \{\lambda x \mid \lambda \geq 0\}$$

is called the *ray* through x, or the *halfline* from the origin through x.

A *cone* X in $\mathbf{R}^n$ is a non-empty set which contains the ray through each of its points.

A cone X is in other words a non-empty set which is closed under multiplication by nonnegative numbers, i.e. which satisfies the implication

$$x \in X, \lambda \geq 0 \Rightarrow \lambda x \in X.$$

In particular, all cones contain the point 0.

We shall study convex cones. Rays and linear subspaces of $\mathbf{R}^n$ are convex cones, of course. In particular, the entire space $\mathbf{R}^n$ and the trivial subspace $\{0\}$ are convex cones. Other simple examples of convex cones are provided by the following examples.

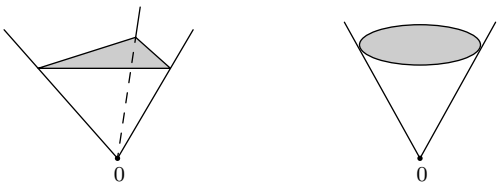

Figure 2.7. A plane cut through two proper convex cones in $\mathbf{R}^3$

EXAMPLE 2.6.1. A closed halfspace $\{x \in \mathbf{R}^n \mid \langle c, x\rangle \geq 0\}$, which is bounded by a hyperplane through the origin, is a convex cone and is called a *conic halfspace*.

The union $\{x \in \mathbf{R}^n \mid \langle c, x\rangle > 0\} \cup \{0\}$ of the corresponding open halfspace and the origin is also a convex cone. □

EXAMPLE 2.6.2. The nonnegative orthant

$$\mathbf{R}^n_+ = \{x = (x_1, \ldots, x_n) \in \mathbf{R}^n \mid x_1 \geq 0, \ldots, x_n \geq 0\}$$

in $\mathbf{R}^n$ is a convex cone. □

Definition. A cone that does not contain any line through 0, is called a *proper* cone.[‡]

That a cone X does not contain any line through 0 is equivalent to the condition

$$x, -x \in X \;\Rightarrow\; x = 0.$$

In other words, a cone X is a proper cone if and only if $X \cap (-X) = \{0\}$.

Closed conic halfspaces in $\mathbf{R}^n$ are non-proper cones if $n \geq 2$. The nonnegative orthant $\mathbf{R}^n_+$ is a proper cone. The cones $\{x \in \mathbf{R}^n \mid \langle c, x\rangle > 0\} \cup \{0\}$ are also proper cones.

We now give two alternative ways to express that a set is a convex cone.

Theorem 2.6.1. *The following three conditions are equivalent for a nonempty subset X of $\mathbf{R}^n$:*

(i) X is a convex cone.

(ii) X is a cone and $x + y \in X$ for all $x, y \in X$.

(iii) $\lambda x + \mu y \in X$ for alla $x, y \in X$ and all $\lambda, \mu \in \mathbf{R}_+$.

[‡]The terminology is not universal. A proper cone is usually called a *salient cone*, while the term *proper cone* is sometimes reserved for cones that are closed, have a nonempty interior and do not contain any lines through the origin.

Proof. (i) ⇒ (ii): If X is a convex cone and $x, y \in X$, then $z = \frac{1}{2}x + \frac{1}{2}y$ belongs to X because of convexity, and $x + y$ $(= 2z)$ belongs to X since X is cone.

(ii) ⇒ (iii): If (ii) holds, $x, y \in X$ and $\lambda, \mu \in \mathbf{R}_+$, then λx and μy belong to X by the cone condition, and $\lambda x + \mu y \in X$ by additivity.

(iii) ⇒ (i): If (iii) holds, then we conclude that X is a cone by choosing $y = x$ and $\mu = 0$, and that the cone is convex by choosing $\lambda + \mu = 1$. □

Definition. A linear combination $\sum_{j=1}^{m} \lambda_j x_j$ of vectors $x_1, x_2, \ldots, x_m$ in $\mathbf{R}^n$ is called a *conic combination* if all coefficients $\lambda_1, \lambda_2, \ldots, \lambda_m$ are nonnegative.

Theorem 2.6.2. *A convex cone contains all conic combinations of its elements.*

Proof. Follows immediately by induction from the characterization of convex cones in Theorem 2.6.1 (iii). □

Cone preserving operations

The proofs of the four theorems below are analogous to the proofs of the corresponding theorems on convex sets, and they are therefore left as exercises.

Theorem 2.6.3. *Let $T\colon \mathbf{R}^n \to \mathbf{R}^m$ be a linear map.*

(i) The image $T(X)$ of a convex cone X in $\mathbf{R}^n$ is a convex cone.

(ii) The inverse image $T^{-1}(Y)$ of a convex cone in $\mathbf{R}^m$ is a convex cone.

Theorem 2.6.4. *The intersection $\bigcap_{i\in I} X_i$ of an arbitrary family of convex cones X_i in $\mathbf{R}^n$ is a convex cone.*

Theorem 2.6.5. *The Cartesian product $X \times Y$ of two convex cones X and Y is a convex cone.*

Theorem 2.6.6. *The sum $X + Y$ of two convex cones X and Y in $\mathbf{R}^n$ is a convex cone, and $-X$ is a convex cone if X is a convex cone.*

Example 2.6.3. An intersection

$$X = \bigcap_{i=1}^{m} \{x \in \mathbf{R}^n \mid \langle c_i, x\rangle \geq 0\}$$

of finitely many closed conic halfspaces is called a *polyhedral cone* or a *conic polyhedron.*

By defining C as the $m \times n$-matrix with rows c_i^T, $i = 1, 2, \ldots, m$, we can write the above polyhedral cone X in a more compact way as

$$X = \{x \in \mathbf{R}^n \mid Cx \geq 0\}.$$

A polyhedral cone is in other words the solution set of a system of homogeneous linear inequalities. □

Conic hull

Definition. Let A be an arbitrary nonempty subset of $\mathbf{R}^n$. The set of all conic combinations of elements of A is called the *conic hull* of A, and it is denoted by $\operatorname{con} A$. The elements of A are called *generators* of $\operatorname{con} A$.

We extend the concept to the empty set by defining $\operatorname{con} \emptyset = \{0\}$.

Theorem 2.6.7. *The set* $\operatorname{con} A$ *is a convex cone that contains* A *as a subset, and it is the smallest convex cone with this property, i.e. if* X *is a convex cone and* $A \subseteq X$*, then* $\operatorname{con} A \subseteq X$*.*

Proof. A conic combination of two conic combinations of elements from A is clearly a new conic combination of elements from A, and hence $\operatorname{con} A$ is a convex cone. The inclusion $A \subseteq \operatorname{con} A$ is obvious. Since a convex cone contains all conic combinations of its elements, a convex cone X that contains A as a subset must in particular contain all conic combinations of elements from A, which means that $\operatorname{con} A$ is a subset of X. □

Theorem 2.6.8. *Let* $X = \operatorname{con} A$ *be a cone in* $\mathbf{R}^n$*,* $Y = \operatorname{con} B$ *be a cone in* $\mathbf{R}^m$ *and* $T\colon \mathbf{R}^n \to \mathbf{R}^m$ *be a linear map. Then*

(i) $T(X) = \operatorname{con} T(A)$*;*

(ii) $X \times Y = \operatorname{con}\big((A \times \{0\}) \cup (\{0\} \times B)\big)$*;*

(iii) $X + Y = \operatorname{con}(A \cup B)$*, provided that* $m = n$ *so that the sum* $X + Y$ *is well-defined.*

Proof. (i) The cone X consists of all conic combinations $x = \sum_{j=1}^{p} \lambda_j a_j$ of elements a_j in A. For such a conic combination $Tx = \sum_{j=1}^{p} \lambda_j T a_j$. The image cone $T(X)$ thus consists of all conic combinations of the elements $Ta_j \in T(A)$, which means that $T(X) = \operatorname{con} T(A)$.

(ii) The cone $X \times Y$ consists of all pairs $(x, y) = \big(\sum_{j=1}^{p} \lambda_j a_j, \sum_{k=1}^{q} \mu_k b_k\big)$ of conic combinations of elements in A and B, respectively. But

$$(x, y) = \sum_{j=1}^{p} \lambda_j (a_j, 0) + \sum_{k=1}^{q} \mu_k (0, b_k),$$

and hence (x, y) is a conic combination of elements in $(A \times \{0\}) \cup (\{0\} \times B)$, that is (x, y) is an element of the cone $Z = \operatorname{con}\big((A \times \{0\}) \cup (\{0\} \times B)\big)$. This proves the inclusion $X \times Y \subseteq Z$.

The converse inclusion $Z \subseteq X \times Y$ follows at once from the trivial inclusion $(A \times \{0\}) \cup (\{0\} \times B) \subseteq X \times Y$, and the fact that $X \times Y$ is a cone.

(iii) A typical element of $X + Y$ has the form $\sum_{j=1}^{p} \lambda_j a_j + \sum_{k=1}^{q} \mu_k b_k$, which is a conic combination of elements in $A \cup B$. This proves the assertion. □

Finitely generated cones

Definition. A convex cone X is called *finitely generated* if $X = \operatorname{con} A$ for some finite set A.

Example 2.6.4. The nonnegative orthant $\mathbf{R}^n_+$ is finitely generated by the standard basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ of $\mathbf{R}^n$. □

Theorem 2.6.8 has the following immediate corollary.

Corollary 2.6.9. *Cartesian products $X \times Y$, sums $X + Y$ and images $T(X)$ under linear maps T of finitely generated cones X and Y, are themselves finitely generated cones.*

Intersections $X \cap Y$ and inverse images $T^{-1}(Y)$ of finitely generated cones are finitely generated, too, but the proof of this fact has to wait until we have shown that finitely generated cones are polyhedral, and vice versa. See Chapter 5.

Theorem 2.6.10. *Suppose that $x \in \operatorname{con} A$, where A is a subset of $\mathbf{R}^n$. Then $x \in \operatorname{con} B$ for some linearly independent subset B of A. The number of elements in B is thus at most equal to n.*

Proof. Since x is a conic combination of elements of A, x is per definition a conic combination of finitely many elements chosen from A. Now choose a subset B of A with as few elements as possible and such that $x \in \operatorname{con} B$. We will prove that the set B is linearly independent.

If $B = \emptyset$ (i.e. if $x = 0$), then we are done, because the empty set is linearly independent. So assume that $B = \{b_1, b_2, \ldots, b_m\}$, where $m \geq 1$. Then $x = \sum_{j=1}^{m} \lambda_j b_j$, where each $\lambda_j > 0$ due to the minimality assumption.

We will prove our assertion by contradiction. So, suppose that the set B is linearly dependent. Then there exist scalars $\mu_1, \mu_2, \ldots, \mu_m$, at least

one of them being positive, such that $\sum_{j=1}^{m} \mu_j b_j = 0$, and it follows that $x = \sum_{j=1}^{m} (\lambda_j - t\mu_j) b_j$ for all $t \in \mathbf{R}$.

Now let $t_0 = \min \lambda_j/\mu_j$, where the minimum is taken over all indices such that $\mu_j > 0$, and let j_0 be an index where the minimum is achieved. Then $\lambda_j - t_0\mu_j \geq 0$ for all indices j, and $\lambda_{j_0} - t_0\mu_{j_0} = 0$. This means that x belongs to the cone generated by the set $B \setminus \{b_{j_0}\}$, which contradicts the minimality assumption about B. Thus, B is linearly independent. □

Theorem 2.6.11. *Every finitely generated cone is closed.*

Proof. Let X be a finitely generated cone in $\mathbf{R}^n$ so that $X = \operatorname{con} A$ for some finite set A.

We first treat the case when $A = \{a_1, a_2, \ldots, a_m\}$ is a linearly independent set. Then $m \leq n$, and it is possible to extend the set A, if necessary, with vectors $a_{m+1}, \ldots, a_n$ to a basis for $\mathbf{R}^n$. Let $(c_1(x), c_2(x), \ldots, c_n(x))$ denote the coordinates of the vector x with respect to the basis $a_1, a_2, \ldots, a_n$, so that $x = \sum_{j=1}^{n} c_j(x)\, a_j$. The coordinate functions $c_j(x)$ are linear forms on $\mathbf{R}^n$.

A vector x belongs to X if and only if x is a conic combination of the first m basis vectors, and this means that

$$X = \{x \in \mathbf{R}^n \mid c_1(x) \geq 0, \ldots, c_m(x) \geq 0,\ c_{m+1}(x) = \cdots = c_n(x) = 0\}.$$

We conclude that X is equal to the intersection of the closed halfspaces $\{x \in \mathbf{R}^n \mid c_j(x) \geq 0\}$, $1 \leq j \leq m$, and the hyperplanes $\{x \in \mathbf{R}^n \mid c_j(x) = 0\}$, $m + 1 \leq j \leq n$. This proves that X is a closed cone in the present case and indeed a polyhedral cone.

Let us now turn to the general case. Let A be an arbitrary finite set. By the previous theorem, there corresponds to each $x \in \operatorname{con} A$ a linearly independent subset B of A such that $x \in \operatorname{con} B$, and this fact implies that $\operatorname{con} A = \bigcup \operatorname{con} B$, where the union is to be taken over all linearly independent subsets B of A. Of course, there are only finitely many such subsets, and hence $\operatorname{con} A$ is a union of finitely many cones $\operatorname{con} B$, each of them being closed, by the first part of the proof. A union of finitely many closed sets is closed. Hence, $\operatorname{con} A$ is a closed cone. □

2.7 The recession cone

The recession cone of a set consists of the directions in which the set is unbounded and in this way provides information about the behavior of the set at infinity. Here is the formal definition of the concept.

Definition. Let X be a subset of $\mathbf{R}^n$ and let v be a nonzero vector in $\mathbf{R}^n$. We say that the set X *recedes* in the direction of v and that v is a *recession vector* of X if X contains all halflines with direction vector v that start from an arbitrary point of X.

The set consisting of all recession vectors of X and the zero vector is called the *recession cone* and is denoted by $\operatorname{recc} X$. Hence, if X is a nonempty set then

$$\operatorname{recc} X = \{v \in \mathbf{R}^n \mid x + tv \in X \text{ for all } x \in X \text{ and all } t > 0\},$$

whereas $\operatorname{recc} \emptyset = \{0\}$.

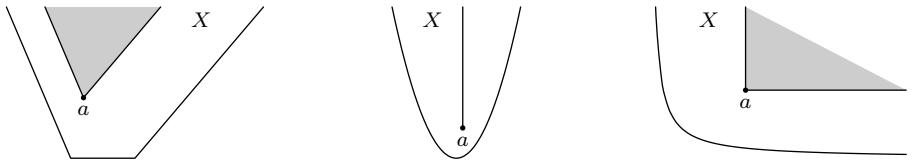

Figure 2.8. Three convex sets X and the corresponding translated recession cones $a + \operatorname{recc} X$.

Theorem 2.7.1. *The recession cone of an arbitrary set X is a convex cone and*

$$X = X + \operatorname{recc} X.$$

Proof. That $\operatorname{recc} X$ is a cone follows immediately from the very definition of recession cones, and the same holds for the inclusion $X + \operatorname{recc} X \subseteq X$. The converse inclusion $X \subseteq X + Y$ is trivially true for all sets Y containing 0 and thus in particular for the cone $\operatorname{recc} X$.

If v and w are two recession vectors of X, x is an arbitrary point in X and t is an arbitrary positive number, then first $x + tv$ belongs to X by definition, and then $x + t(v + w) = (x + tv) + tw$ belongs to X. This means that the sum $v + w$ is also a recession vector. So the recession cone has the additivity property $v, w \in \operatorname{recc} X \Rightarrow v + w \in \operatorname{recc} X$, which implies that the cone is convex according to Theorem 2.6.1. □

Example 2.7.1. Here are some simple examples of recession cones:

$$\operatorname{recc}(\mathbf{R}_+ \times [0,1]) = \operatorname{recc}(\mathbf{R}_+ \times]0,1[) = \mathbf{R}_+ \times \{0\},$$
$$\operatorname{recc}(\mathbf{R}_+ \times]0,1[\cup \{(0,0)\}) = \{(0,0)\},$$

$$\operatorname{recc}\{x \in \mathbf{R}^2 \mid x_1^2 + x_2^2 \leq 1\} = \{(0,0)\},$$
$$\operatorname{recc}\{x \in \mathbf{R}^2 \mid x_2 \geq x_1^2\} = \{0\} \times \mathbf{R}_+,$$
$$\operatorname{recc}\{x \in \mathbf{R}^2 \mid x_2 \geq 1/x_1,\ x_1 > 0\} = \mathbf{R}_+^2. \qquad \square$$

The computation of the recession cone of a convex set is simplified by the following theorem.

Theorem 2.7.2. *A vector v is a recession vector of a nonempty convex set X if and only if $x + v \in X$ for all $x \in X$.*

Proof. If v is a recession vector, then obviously $x + v \in X$ for all $x \in X$.

To prove the converse, assume that $x + v \in X$ for all $x \in X$, and let x be an arbitrary point in X. Then, $x + nv \in X$ for all natural numbers n, by induction, and since X is a convex set, we conclude that the closed line segment $[x, x+nv]$ lies in X for all n. Of course, this implies that $x+tv \in X$ for all positive numbers t, and hence v is a recession vector of X. $\square$

Corollary 2.7.3. *If X is a convex cone, then* $\operatorname{recc} X = X$.

Proof. The inclusion $\operatorname{recc} X \subseteq X$ holds for all sets X containing 0 and thus in particular for cones X. The converse inclusion $X \subseteq \operatorname{recc} X$ is according to Theorem 2.7.2 a consequence of the additivity property $x, v \in X \Rightarrow x+v \in X$ for convex cones. $\square$

EXAMPLE 2.7.2. $\operatorname{recc} \mathbf{R}_+^2 = \mathbf{R}_+^2$, $\operatorname{recc}(\mathbf{R}_{++}^2 \cup \{(0,0)\}) = \mathbf{R}_{++}^2 \cup \{(0,0)\}$. $\square$

The recession vectors of a closed convex set are characterized by the following theorem.

Theorem 2.7.4. *Let X be a nonempty closed convex set. The following three conditions are equivalent for a vector v.*

(i) v is a recession vector of X.

(ii) There exists a point $x \in X$ such that $x + nv \in X$ for all $n \in \mathbf{Z}_+$.

(iii) There exist a sequence $(x_n)_1^\infty$ of points x_n in X and a sequence $(\lambda_n)_1^\infty$ of positive numbers such that $\lambda_n \to 0$ and $\lambda_n x_n \to v$ as $n \to \infty$.

Proof. (i) $\Rightarrow$ (ii): Trivial, since $x + tv \in X$ for all $x \in X$ and all $t \in \mathbf{R}_+$, if v is a recession vector of X.

(ii) $\Rightarrow$ (iii): If (ii) holds, then condition (iii) is satisfied by the points $x_n = x + nv$ and the numbers $\lambda_n = 1/n$.

(iii) $\Rightarrow$ (i): Assume that $(x_n)_1^\infty$ and $(\lambda_n)_1^\infty$ are sequences of points in X and positive numbers such that $\lambda_n \to 0$ and $\lambda_n x_n \to v$ as $n \to \infty$, and let x be an arbitrary point in X. The points $z_n = (1 - \lambda_n)x + \lambda_n x_n$ then lie in X for

all sufficiently large n, and since $z_n \to x + v$ as $n \to \infty$ and X is a closed set, it follows that $x + v \in X$. Hence, v is a recession vector of X according to Theorem 2.7.2. □

Theorem 2.7.5. *The recession cone* $\operatorname{recc} X$ *of a closed convex set* X *is a closed convex cone.*

Proof. The case $X = \emptyset$ is trivial, so assume that X is a nonempty closed convex set. To prove that the recession cone $\operatorname{recc} X$ is closed, we assume that v is a boundary point of the cone and choose a sequence $(v_n)_1^\infty$ of recession vectors that converges to v as $n \to \infty$. If x is an arbitrary point in X, then the points $x + v_n$ lie in X for each natural number n, and this implies that their limit point $x + v$ lies in X, since X is a closed set. Hence, v is a recession vector, i.e. v belongs to the recession cone $\operatorname{recc} X$. This proves that the recession cone contains all its boundary points. □

Theorem 2.7.6. *Let* $\{X_i \mid i \in I\}$ *be a family of closed convex sets, and assume that their intersection is nonempty. Then* $\operatorname{recc}(\bigcap_{i\in I} X_i) = \bigcap_{i\in I} \operatorname{recc} X_i$.

Proof. Let x_0 be a point in $\bigcap_i X_i$. By Theorem 2.7.4, $v \in \operatorname{recc}(\bigcap_i X_i)$ if and only if $x_0 + nv$ lies in X_i for all positive integers n and all $i \in I$, and this holds if and only if $v \in \operatorname{recc} X_i$ for all $i \in I$. □

The recession cone of a polyhedron is given by the following theorem.

Theorem 2.7.7. *If* $X = \{x \in \mathbf{R}^n \mid Cx \geq b\}$ *is a nonempty polyhedron, then* $\operatorname{recc} X = \{x \in \mathbf{R}^n \mid Cx \geq 0\}$.

Proof. The recession cone of a closed halfspace is obviously equal to the corresponding conical halfspace. The theorem is thus an immediate consequence of Theorem 2.7.6. □

Note that the recession cone of a subset Y of X can be bigger than the recession cone of X. For example,

$$\operatorname{recc} \mathbf{R}^2_{++} = \mathbf{R}^2_{+} \supsetneq \mathbf{R}^2_{++} \cup \{(0,0)\} = \operatorname{recc}(\mathbf{R}^2_{++} \cup \{(0,0)\}).$$

However, this cannot occur if the superset X is closed.

Theorem 2.7.8. *(i) Suppose that* X *is a closed convex set and that* $Y \subseteq X$. *Then* $\operatorname{recc} Y \subseteq \operatorname{recc} X$.

(ii) If X *is a convex set, then* $\operatorname{recc}(\operatorname{rint} X) = \operatorname{recc}(\operatorname{cl} X)$.

Proof. (i) The case $Y = \emptyset$ being trivial, we assume that Y is a nonempty subset of X and that y is an arbitrary point in Y. If v is a recession vector of Y, then $y + nv$ are points in Y and thereby also in X for all natural numbers n. We conclude from Theorem 2.7.4 that v is a recession vector of X.

(ii) The inclusion $\operatorname{recc}(\operatorname{rint} X) \subseteq \operatorname{recc}(\operatorname{cl} X)$ follows from part (i), because $\operatorname{cl} X$ is a closed convex subset.

To prove the converse inclusion, assume that v is a recession vector of $\operatorname{cl} X$, and let x be an arbitrary point in $\operatorname{rint} X$. Then $x + 2v$ belongs to $\operatorname{cl} X$, so it follows from Theorem 2.5.4 that the open line segment $]x, x + 2v[$ is a subset of $\operatorname{rint} X$, and this implies that the point $x + v$ belongs to $\operatorname{rint} X$. Thus, $x \in \operatorname{rint} X \Rightarrow x + v \in \operatorname{rint} X$, and we conclude from Theorem 2.7.2 that v is a recession vector of $\operatorname{rint} X$. □

Theorem 2.7.9. *Let X be a closed convex set. Then X is bounded if and only if* $\operatorname{recc} X = \{0\}$.

Proof. Obviously, $\operatorname{recc} X = \{0\}$ if X is a bounded set. So assume that X is unbounded. Then there exists a sequence $(x_n)_1^\infty$ of points in X such that $\|x_n\| \to \infty$ as $n \to \infty$. The bounded sequence $(x_n/\|x_n\|)_1^\infty$ has a convergent subsequence, and by deleting elements, if necessary, we may as well assume that the original sequence is convergent. The limit v is a vector of norm 1, which guarantees that $v \neq 0$. With $\lambda_n = 1/\|x_n\|$ we now have a sequence of points x_n in X and a sequence of positive numbers λ_n such that $\lambda_n \to 0$ and $\lambda_n x_n \to v$ as $n \to \infty$, and this means that v is a recession vector of X according to Theorem 2.7.4. Hence, $\operatorname{recc} X \neq \{0\}$. □

Definition. Let X be an arbitary set. The intersection $\operatorname{recc} X \cap (-\operatorname{recc} X)$ is a linear subspace, which is called the *recessive subspace of* X and is denoted $\operatorname{lin} X$.

A closed convex set is called *line-free* if $\operatorname{lin} X = \{0\}$. The set X is in other words line-free if and only if $\operatorname{recc} X$ is a proper cone.

If X is a nonempty closed convex subset of $\mathbf{R}^n$ and $x \in X$ is arbitrary, then clearly

$$\operatorname{lin} X = \{x \in \mathbf{R}^n \mid a + tx \in X \text{ for all } t \in \mathbf{R}\}.$$

The image $T(X)$ of a closed convex set X under a linear map T is not necessarily closed. A counterexample is given by $X = \{x \in \mathbf{R}^2_+ \mid x_1 x_2 \geq 1\}$ and the projection $T(x_1, x_2) = x_1$ of $\mathbf{R}^2$ onto the first factor, the image being $T(X) =]0\,, \infty[$. The reason why the image is not closed in this case is the fact that X has a recession vector $v = (0, 1)$ which is mapped on 0 by T.

However, we have the following general result, where $\mathcal{N}(T)$ denotes the null space of the map T, i.e. $\mathcal{N}(T) = \{x \mid Tx = 0\}$.

Theorem 2.7.10. *Let $T\colon \mathbf{R}^n \to \mathbf{R}^m$ be a linear map, let X be a closed convex subset of $\mathbf{R}^n$, and suppose that*

$$\mathcal{N}(T) \cap \operatorname{recc} X \subseteq \operatorname{lin} X.$$

The image $T(X)$ is then a closed set, and

$$\operatorname{recc} T(X) = T(\operatorname{recc} X).$$

In particular, the image $T(X)$ is closed if X is a closed convex set and $x = 0$ is the only vector in $\operatorname{recc} X$ such that $Tx = 0$.

Proof. The intersection

$$L = \mathcal{N}(T) \cap \operatorname{lin} X = \mathcal{N}(T) \cap \operatorname{recc} X$$

is a linear subspace of $\mathbf{R}^n$. Let $L^\perp$ denote its orthogonal complement. Then $X = X \cap L + X \cap L^\perp$, and

$$T(X) = T(X \cap L^\perp),$$

since $Tx = 0$ for all $x \in L$.

Let y be an arbitrary boundary point of the image $T(X)$. Due to the equality above, there exists a sequence $(x_n)_1^\infty$ of points $x_n \in X \cap L^\perp$ such that $\lim_{n\to\infty} Tx_n = y$.

We claim that the sequence $(x_n)_1^\infty$ is bounded. Assume the contrary. The sequence $(x_n)_1^\infty$ has then a subsequence $(x_{n_k})_1^\infty$ such that $\|x_{n_k}\| \to \infty$ as $k \to \infty$ and the bounded sequence $(x_{n_k}/\|x_{n_k}\|)_1^\infty$ converges. The limit v is, of course, a vector of norm 1 in the linear subspace $L^\perp$. Moreover, since $x_{n_k} \in X$ and $1/\|x_{n_k}\| \to 0$, it follows from Theorem 2.7.4 that $v \in \operatorname{recc} X$. Finally,

$$Tv = \lim_{k\to\infty} T(x_{n_k}/\|x_{n_k}\|) = \lim_{k\to\infty} \|x_{n_k}\|^{-1} Tx_{n_k} = 0 \cdot y = 0,$$

and hence v belongs to $\mathcal{N}(T)$, and thereby also to L. This means that $v \in L \cap L^\perp$, which contradicts tha fact that $v \neq 0$, since $L \cap L^\perp = \{0\}$.

The sequence $(x_n)_1^\infty$ is thus bounded. Let $(x_{n_k})_1^\infty$ be a convergent subsequence, and let $x = \lim_{k\to\infty} x_{n_k}$. The limit x lies in X since X is closed, and $y = \lim_{k\to\infty} Tx_{n_k} = Tx$, which implies that $y \in T(X)$. This proves that the image $T(X)$ contains its boundary points, so it is a closed set.

The inclusion $T(\operatorname{recc} X) \subseteq \operatorname{recc} T(X)$ holds for all sets X. To prove this, assume that v is a recession vector of X and let y be an arbitrary point in $T(X)$. Then $y = Tx$ for some point $x \in X$, and since $x + tv \in X$ for all $t > 0$ and $y + tTv = T(x + tv)$, we conclude that the points $y + tTv$ lie in $T(X)$ for all $t > 0$, and this means that Tv is a recession vector of $T(X)$.

To prove the converse inclusion $\operatorname{recc} T(X) \subseteq T(\operatorname{recc} X)$ for closed convex sets X and linear maps T fulfilling the assumptions of the theorem, we assume that $w \in \operatorname{recc} T(X)$ and shall prove that there is a vector $v \in \operatorname{recc} X$ such that $w = Tv$.

We first note that there exists a sequence $(y_n)_1^\infty$ of points $y_n \in T(X)$ and a sequence $(\lambda_n)_1^\infty$ of positive numbers such that $\lambda_n \to 0$ and $\lambda_n y_n \to w$ as $n \to \infty$. For each n, choose a point $x_n \in X \cap L^\perp$ such that $y_n = T(x_n)$.

The sequence $(\lambda_n x_n)_1^\infty$ is bounded. Because assume the contrary; then there is a subsequence such that $\|\lambda_{n_k} x_{n_k}\| \to \infty$ and $(x_{n_k}/\|x_{n_k}\|)_1^\infty$ converges to a vector z as $k \to \infty$. It follows from Theorem 2.7.4 that $z \in \operatorname{recc} X$, because the x_{n_k} are points in X and $\|x_{n_k}\| \to \infty$ as $k \to \infty$. The limit z belongs to the subspace $L^\perp$, and since

$$\begin{aligned} Tz &= \lim_{k\to\infty} T(x_{n_k}/\|x_{n_k}\|) = \lim_{k\to\infty} T(\lambda_{n_k} x_{n_k}/\|\lambda_{n_k} x_{n_k}\|) \\ &= \lim_{k\to\infty} \lambda_{n_k} y_{n_k}/\|\lambda_{n_k} x_{n_k}\| = 0 \cdot w = 0, \end{aligned}$$

we also have $z \in \mathcal{N}(T) \cap \operatorname{recc} X = L$. Hence, $z \in L \cap L^\perp = \{0\}$, which contradicts the fact that $\|z\| = 1$.

The sequence $(\lambda_n x_n)_1^\infty$, being bounded, has a subsequence that converges to a vector v, which belongs to $\operatorname{recc} X$ according to Theorem 2.7.4. Since $T(\lambda_n x_n) = \lambda_n y_n \to w$, we conclude that $Tv = w$. Hence, $w \in T(\operatorname{recc} X)$. □

Theorem 2.7.11. *Let X and Y be nonempty closed convex subsets of $\mathbf{R}^n$ and suppose that*

$$x \in \operatorname{recc} X \;\&\; y \in \operatorname{recc} Y \;\&\; x + y = 0 \quad\Rightarrow\quad x \in \operatorname{lin} X \;\&\; y \in \operatorname{lin} Y.$$

The sum $X + Y$ is then a closed convex set and

$$\operatorname{recc}(X + Y) = \operatorname{recc} X + \operatorname{recc} Y.$$

Remark. The assumption of the theorem is fulfilled if $\operatorname{recc} X$ and $-\operatorname{recc} Y$ have no common vector other than the zero vector.

Proof. Let $T\colon \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}^n$ be the linear map $T(x, y) = x + y$. We leave as an easy exercise to show that $\operatorname{recc}(X \times Y) = \operatorname{recc} X \times \operatorname{recc} Y$ and that $\operatorname{lin}(X \times Y) = \operatorname{lin} X \times \operatorname{lin} Y$. Since $\mathcal{N}(T) = \{(x, y) \mid x + y = 0\}$, the assumption

of the theorem yields the inclusion $\mathcal{N}(T) \cap \operatorname{recc}(X \times Y) \subseteq \operatorname{lin}(X \times Y)$, and it now follows from Theorem 2.7.10 that $T(X \times Y)$, i.e. the sum $X + Y$, is closed and that

$$\begin{aligned}\operatorname{recc}(X + Y) &= \operatorname{recc} T(X \times Y) = T(\operatorname{recc}(X \times Y)) = T(\operatorname{recc} X \times \operatorname{recc} Y) \\ &= \operatorname{recc} X + \operatorname{recc} Y. \qquad \square\end{aligned}$$

Corollary 2.7.12. *The sum $X + Y$ of a nonempty closed convex set X and a nonempty compact convex set Y is a closed convex set and*

$$\operatorname{recc}(X + Y) = \operatorname{recc} X.$$

Proof. The assumptions of Theorem 2.7.11 are trivially fulfilled, because $\operatorname{recc} Y = \{0\}$. $\square$

Theorem 2.7.13. *Suppose that C is a closed convex cone and that Y is a nonempty compact convex set. Then* $\operatorname{recc}(C + Y) = C$.

Proof. The corollary is a special case of Corollary 2.7.12 since $\operatorname{recc} C = C$. $\square$

Exercises

2.1 Prove that the set $\{x \in \mathbf{R}^2_+ \mid x_1 x_2 \geq a\}$ is convex and, more generally, that the set $\{x \in \mathbf{R}^n_+ \mid x_1 x_2 \cdots x_n \geq a\}$ is convex.
Hint: Use the inequality $x_i^\lambda y_i^{1-\lambda} \leq \lambda x_i + (1 - \lambda) y_i$; see Theorem 6.4.1.

2.2 Determine the convex hull $\operatorname{cvx} A$ for the following subsets A of $\mathbf{R}^2$:
a) $A = \{(0,0), (1,0), (0,1)\}$ b) $A = \{x \in \mathbf{R}^2 \mid \|x\| = 1\}$
c) $A = \{x \in \mathbf{R}^2_+ \mid x_1 x_2 = 1\} \cup \{(0,0)\}$.

2.3 Give an example of a closed set with a non-closed convex hull.

2.4 Find the inverse image $P^{-1}(X)$ of the convex set $X = \{x \in \mathbf{R}^2_+ \mid x_1 x_2 \geq 1\}$ under the perspective map $P\colon \mathbf{R}^2 \times \mathbf{R}_{++} \to \mathbf{R}^2$.

2.5 Prove that the set $\{x \in \mathbf{R}^{n+1} \mid \left(\sum_{j=1}^n x_j^2\right)^{1/2} \leq x_{n+1}\}$ is a cone.

2.6 Let $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ denote the standard basis in $\mathbf{R}^n$ and let $\mathbf{e}_0 = -\sum_1^n \mathbf{e}_j$. Prove that $\mathbf{R}^n$ is generated as a cone by the $n + 1$ vectors $\mathbf{e}_0$, $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$.

2.7 Prove that each conical halfspace in $\mathbf{R}^n$ is the conic hull of a set consisting of $n + 1$ elements.

2.8 Prove that each closed cone in $\mathbf{R}^2$ is the conic hull of a set consisting of at most three elements.

2.9 Prove that the sum of two closed cones in $\mathbf{R}^2$ is a closed cone.

2.10 Find $\operatorname{recc} X$ and $\operatorname{lin} X$ for the following convex sets:

a) $X = \{x \in \mathbf{R}^2 \mid -x_1 + x_2 \leq 2,\ x_1 + 2x_2 \geq 2,\ x_2 \geq -1\}$
b) $X = \{x \in \mathbf{R}^2 \mid -x_1 + x_2 \leq 2,\ x_1 + 2x_2 \leq 2,\ x_2 \geq -1\}$
c) $X = \{x \in \mathbf{R}^3 \mid 2x_1 + x_2 + x_3 \leq 4,\ x_1 + 2x_2 + x_3 \leq 4\}$
d) $X = \{x \in \mathbf{R}^3 \mid x_1^2 - x_2^2 \geq 1,\ x_1 \geq 0\}$.

2.11 Let X and Y be arbitrary nonempty sets. Prove that

$$\operatorname{recc}(X \times Y) = \operatorname{recc} X \times \operatorname{recc} Y$$

and that

$$\operatorname{lin}(X \times Y) = \operatorname{lin} X \times \operatorname{lin} Y.$$

2.12 Let $P\colon \mathbf{R}^n \times \mathbf{R}_{++} \to \mathbf{R}^n$ be the perspective map. Suppose X is a convex subset of $\mathbf{R}^n$, and let $c(X) = P^{-1}(X) \cup \{(0,0)\}$.

a) Prove that $c(X)$ is a cone and, more precisely, that $c(X) = \operatorname{con}(X \times \{1\})$.
b) Find an explicit expression for the cones $c(X)$ and $\operatorname{cl}(c(X))$ if
   (i) $n = 1$ and $X = [2,3]$;
   (ii) $n = 1$ and $X = [2,\infty[$;
   (iii) $n = 2$ and $X = \{x \in \mathbf{R}^2 \mid x_1 \geq x_2^2\}$.
c) Find $c(X)$ if $X = \{x \in \mathbf{R}^n \mid \|x\| \leq 1\}$ and $\|\cdot\|$ is an arbitrary norm on $\mathbf{R}^n$.
d) Prove that $\operatorname{cl}(c(X)) = c(\operatorname{cl} X) \cup (\operatorname{recc}(\operatorname{cl} X) \times \{0\})$.
e) Prove that $\operatorname{cl}(c(X)) = c(\operatorname{cl} X)$ if and only if X is a bounded set.
f) Prove that the cone $c(X)$ is closed if and only if X is compact.

2.13 $Y = \{x \in \mathbf{R}^3 \mid x_1 x_3 \geq x_2^2,\ x_3 > 0\} \cup \{x \in \mathbf{R}^3 \mid x_1 \geq 0,\ x_2 = x_3 = 0\}$ is a closed cone. (Cf. problem 2.12 b) (iii)). Put

$$Z = \{x \in \mathbf{R}^3 \mid x_1 \leq 0,\ x_2 = x_3 = 0\}.$$

Show that

$$Y + Z = \{x \in \mathbf{R}^3 \mid x_3 > 0\} \cup \{x \in \mathbf{R}^3 \mid x_2 = x_3 = 0\},$$

with the conclusion that the sum of two closed cones in $\mathbf{R}^3$ is not necessarily a closed cone.

2.14 Prove that the sum $X + Y$ of an arbitrary closed set X and an arbitrary compact set Y is closed.

Chapter 3

Separation

3.1 Separating hyperplanes

Definition. Let X and Y be two sets in $\mathbf{R}^n$. We say that the hyperplane H *separarates* the two sets if the following two conditions[†] are satisfied:

(i) X is contained in one of the two opposite closed halvspaces defined by H and Y is contained in the other closed halfspace;
(ii) X and Y are not both subsets of the hyperplane H.

The separation is called *strict* if there exist two parallell hyperplanes to H, one on each side of H, that separates X and Y.

The hyperplane $H = \{x \mid \langle c, x\rangle = b\}$ thus separates the two sets X and Y, if $\langle c, x\rangle \leq b \leq \langle c, y\rangle$ for all $x \in X$ and all $y \in Y$ and $\langle c, x\rangle \neq b$ for some element $x \in X \cup Y$.

The separation is strict if there exist numbers $b_1 < b < b_2$ such that $\langle c, x\rangle \leq b_1 < b_2 \leq \langle c, y\rangle$ for all $x \in X$, $y \in Y$.

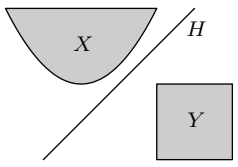

Figure 3.1. A strictly separating hyperplane H

[†]The second condition is usually not included in the definition of separation, but we have included it in order to force a hyperplane H that separates two subsets of a hyperplane H' to be different from H'.

The existence of separating hyperplanes is in a natural way connected to extreme values of linear functions.

Theorem 3.1.1. *Let X and Y be two nonempty subsets of $\mathbf{R}^n$.*

(i) There exists a hyperplane that separates X and Y if and only if there exists a vector c such that

$$\sup_{x\in X}\langle c,x\rangle \le \inf_{y\in Y}\langle c,y\rangle \quad \textit{and} \quad \inf_{x\in X}\langle c,x\rangle < \sup_{y\in Y}\langle c,y\rangle.$$

(ii) There exists a hyperplane that separates X and Y strictly if and only if there exists a vector c such that

$$\sup_{x\in X}\langle c,x\rangle < \inf_{y\in Y}\langle c,y\rangle.$$

Proof. A vector c that satisfies the conditions in (i) or (ii) is nonzero, of course.

Suppose that c satisfies the conditions in (i) and choose the number b so that $\sup_{x\in X}\langle c,x\rangle \le b \le \inf_{y\in Y}\langle c,y\rangle$. Then $\langle c,x\rangle \le b$ for all $x\in X$ and $\langle c,y\rangle \ge b$ for all $y\in Y$. Moreover, $\langle c,x\rangle \ne b$ for some $x\in X\cup Y$ because of the second inequality in (i). The hyperplane $H=\{x \mid \langle c,x\rangle = b\}$ thus separates the two sets X and Y.

If c satisfies the condition in (ii), we choose instead b so that

$$\sup_{x\in X}\langle c,x\rangle < b < \inf_{y\in Y}\langle c,y\rangle,$$

and now conclude that the hyperplane H separates X and Y strictly.

Conversely, if the hyperplane H separates X and Y, then, by changing the signs of c and b if necessary, we may assume that $\langle c,x\rangle \le b$ for all $x\in X$ and $\langle c,y\rangle \ge b$ for all $y\in Y$, and this implies that $\sup_{x\in X}\langle c,x\rangle \le b \le \inf_{y\in Y}\langle c,y\rangle$. Since H does not contain both X and Y, there are points $x_1\in X$ and $y_1\in Y$ with $\langle c,x_1\rangle < \langle c,y_1\rangle$, and this gives us the second inequality in (i).

If the separation is strict, then there exist two parallel hyperplanes $H_i=\{x \mid \langle c,x\rangle = b_i\}$ with $b_1<b<b_2$ that separate X and Y. Assuming that X lies in the halfspace $\{x \mid \langle c,x\rangle \le b_1\}$, we conclude that

$$\sup_{x\in X}\langle c,x\rangle \le b_1 < b < b_2 \le \inf_{y\in Y}\langle c,y\rangle,$$

i.e. the vector c satisfies the condition in (ii). □

The following simple lemma reduces the problem of separating two sets to the case when one of the sets consists of just one point.

Lemma 3.1.2. *Let X and Y be two nonempty sets.*

(i) If there exists a hyperplane that separates 0 from the set $X-Y$, then there exists a hyperplane that separates X and Y.

(ii) If there exists a hyperplane that strictly separates 0 *from the set* $X - Y$*, then there exists a hyperplane that strictly separates* X *and* Y*.*

Proof. (i) If there exists a hyperplane that separates 0 from $X - Y$, then by Theorem 3.1.1 there exists a vector c such that

$$\begin{cases} 0 = \langle c, 0\rangle \leq \inf\limits_{x\in X,\, y\in Y} \langle c, x - y\rangle = \inf\limits_{x\in X} \langle c, x\rangle - \sup\limits_{y\in Y} \langle c, y\rangle \\ 0 = \langle c, 0\rangle < \sup\limits_{x\in X,\, y\in Y} \langle c, x - y\rangle = \sup\limits_{x\in X} \langle c, x\rangle - \inf\limits_{y\in Y} \langle c, y\rangle \end{cases}$$

i.e. $\sup_{y\in Y}\langle c, y\rangle \leq \inf_{x\in X}\langle c, x\rangle$ and $\inf_{y\in Y}\langle c, y\rangle < \sup_{x\in X}\langle c, x\rangle$, and we conclude that there exists a hyperplane that separates X and Y.

(ii) If instead there exists a hyperplane that strictly separates 0 from $X - Y$, then there exists a vector c such that

$$0 = \langle c, 0\rangle < \inf_{x\in X,\, y\in Y} \langle c, x - y\rangle = \inf_{x\in X} \langle c, x\rangle - \sup_{y\in Y} \langle c, y\rangle$$

and it now follows that $\sup_{y\in Y}\langle c, y\rangle < \inf_{x\in X}\langle c, x\rangle$, which shows that Y and X can be strictly separated by a hyperplane. □

Our next theorem is the basis for our results on separation of convex sets.

Theorem 3.1.3. *Suppose* X *is a convex set and that* $a \notin \operatorname{cl} X$*. Then there exists a hyperplane* H *that strictly separates* a *and* X*.*

Proof. The set $\operatorname{cl} X$ is closed and convex, and a hyperplane that strictly separates a and $\operatorname{cl} X$ will, of course, also strictly separate a and X, since X is a subset of $\operatorname{cl} X$. Hence, it suffices to prove that we can strictly separate a point a from each closed convex set that does not contain the point.

We may therefore assume, without loss of generality, that the convex set X is closed and nonempty. Define $d(x) = \|x - a\|^2$, i.e. $d(x)$ is the square of the Euclidean distance between x and a. We start by proving that the restriction of the continuous function $d(\cdot)$ to X has a minimum point.

To this end, choose a positive real number r so big that the closed ball $\overline{B}(a; r)$ intersects the set X. Then $d(x) > r^2$ for all $x \in X \setminus \overline{B}(a; r)$, and $d(x) \leq r^2$ for all $x \in X \cap \overline{B}(a; r)$. The restriction of d to the compact set $X \cap \overline{B}(a; r)$ has a minimum point $x_0 \in X \cap \overline{B}(a; r)$, and this point is clearly also a minimum point for d restricted to X, i.e. the inequality $d(x_0) \leq d(x)$ holds for all $x \in X$.

Now, let $c = a - x_0$. We claim that $\langle c, x - x_0\rangle \leq 0$ for all $x \in X$. Therefore, assume the contrary, i.e. that there is a point $x_1 \in X$ such that $\langle c, x_1 - x_0\rangle > 0$. We will prove that this assumption yields a contradiction.

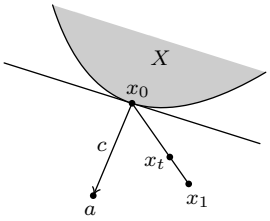

Figure 3.2. Illustration for the proof of Theorem 3.1.3.

Consider the points $x_t = tx_1+(1-t)x_0$. They belong to X when $0 \leq t \leq 1$ because of convexity. Let $f(t) = d(x_t) = \|x_t - a\|^2$. Then

$$\begin{aligned} f(t) = \|x_t - a\|^2 &= \|t(x_1 - x_0) + (x_0 - a)\|^2 = \|t(x_1 - x_0) - c\|^2 \\ &= t^2\|x_1 - x_0\|^2 - 2t\,\langle c, x_1 - x_0\rangle + \|c\|^2. \end{aligned}$$

The function $f(t)$ is a quadratic polynomial in t, and its derivative at 0 satisfies $f'(0) = -2\,\langle c, y_1 - y_0\rangle < 0$. Hence, $f(t)$ is strictly decreasing in a neighbourhood of $t = 0$, which means that $d(x_t) < d(x_0)$ for all sufficiently small positive numbers t.

This is a contradiction to x_0 being the minimum point of the function and proves our assertion that $\langle c, x - x_0\rangle \leq 0$ for all $x \in X$. Consequently, $\langle c, x\rangle \leq \langle c, x_0\rangle = \langle c, a - c\rangle = \langle c, a\rangle - \|c\|^2$ for all $x \in X$, and this implies that $\sup_{x\in X}\langle c, x\rangle < \langle c, a\rangle$. So there exists a hyperplane that strictly separates a from X according to Theorem 3.1.1. □

Definition. Let X be a subset of $\mathbf{R}^n$ and let x_0 be a point in X. A hyperplane H through x_0 is called a *supporting hyperplane* of X, if it separates x_0 and X. We then say that the hyperplane H *supports* X *at the point* x_0.

The existence of a supporting hyperplane of X at x_0 is clearly equivalent to the condition that there exists a vector c such that

$$\langle c, x_0\rangle = \inf_{x\in X}\langle c, x\rangle \qquad \text{and} \qquad \langle c, x_0\rangle < \sup_{x\in X}\langle c, x\rangle.$$

The hyperplane $\{x \mid \langle c, x\rangle = \langle c, x_0\rangle\}$ is then a supporting hyperplane.

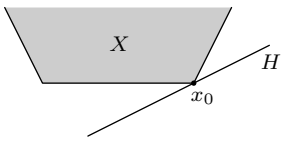

Figure 3.3. A supporting hyperplane of X at the point x_0

If a hyperplane supports the set X at the point x_0, then x_0 is necessarily a relative boundary point of X. For convex sets the following converse holds.

Theorem 3.1.4. *Suppose that X is a convex set and that $x_0 \in X$ is a relative boundary point of X. Then there exists a hyperplane H that supports X at the point x_0.*

Proof. First suppose that the dimension of X equals the dimension of the surrounding space $\mathbf{R}^n$. Since x_0 is then a boundary point of X, there exists a sequence $(x_n)_1^\infty$ of points $x_n \notin \operatorname{cl} X$ that converges to x_0 as $n \to \infty$, and by Theorem 3.1.3 there exists, for each $n \geq 1$, a hyperplane which strictly separates x_n and X. Theorem 3.1.1 thus gives us a sequence $(c_n)_1^\infty$ of vectors such that

$$\langle c_n, x_n \rangle < \langle c_n, x \rangle \qquad \text{for all } x \in X \tag{3.1}$$

and all $n \geq 1$, and we can obviously normalize the vectors c_n so that $\|c_n\| = 1$ for all n.

The unit sphere $\{x \in \mathbf{R}^n \mid \|x\| = 1\}$ is compact. Hence, by the Bolzano–Weierstrass theorem, the sequence $(c_n)_1^\infty$ has a subsequence $(c_{n_k})_{k=1}^\infty$ which converges to some vector c of length $\|c\| = 1$. Clearly $\lim_{k\to\infty} x_{n_k} = x_0$, so by going to the limit in the inequality (3.1) we conclude that $\langle c, x_0 \rangle \leq \langle c, x \rangle$ for all $x \in X$. The set X is therefore a subset of one of the two closed halfspaces determined by the hyperplane $H = \{x \in \mathbf{R}^n \mid \langle c, x \rangle = \langle c, x_0 \rangle\}$, and X is not a subset of H, since $\dim X = n$. The hyperplane H is consequently a supporting hyperplane of X at the point x_0.

Next suppose that $\dim X < n$. Then there exists an affine subspace $a+U$ that contains X, where U is a linear subspace of $\mathbf{R}^n$ and $\dim U = \dim X$. Consider the set $Y = X + U^\perp$, where $U^\perp$ is the orthogonal complement of U. Compare with figure 3.4. Y is a "cylinder" with X as "base", and each $y \in Y$ has a unique decomposition of the form $y = x + v$ with $x \in X$ and $v \in U^\perp$.

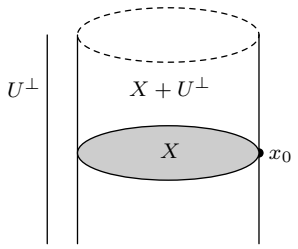

Figure 3.4. Illustration for the proof of Theorem 3.1.4.

The set Y is a convex set of dimension n with x_0 as a boundary point. By the already proven case of the theorem, there exists a hyperplane which supports Y at the point x_0, i.e. there exists a vector c such that

$$\langle c, x_0\rangle = \inf_{y\in Y}\langle c, y\rangle = \inf_{x\in X,\, v\in U^\perp}\langle c, x+v\rangle = \inf_{x\in X}\langle c, x\rangle + \inf_{v\in U^\perp}\langle c, v\rangle$$

and

$$\langle c, x_0\rangle < \sup_{y\in Y}\langle c, y\rangle = \sup_{x\in X,\, v\in U^\perp}\langle c, x+v\rangle = \sup_{x\in X}\langle c, x\rangle + \sup_{v\in U^\perp}\langle c, v\rangle.$$

It follows from the first equation that $\inf_{v\in U^\perp}\langle c, v\rangle$ is a finite number, and since $U^\perp$ is a vector space, this is possible if and only if $\langle c, v\rangle = 0$ for all $v \in U^\perp$. The conditions above are therefore reduced to the conditions

$$\langle c, x_0\rangle = \inf_{x\in X}\langle c, x\rangle \quad \text{and} \quad \langle c, x_0\rangle < \sup_{x\in X}\langle c, x\rangle,$$

which show that X has indeed a supporting hyperplane at x_0. □

We are now able to prove the following necessary and sufficient condition for separation of convex sets.

Theorem 3.1.5. *Two convex sets X and Y can be separated by a hyperplane if and only if their relative interiors are disjoint.*

Proof. A hyperplane that separates two sets A and B clearly also separates their closures $\operatorname{cl} A$ and $\operatorname{cl} B$ and thereby also all sets C and D that satisfy the inclusions $A \subseteq C \subseteq \operatorname{cl} A$ and $B \subseteq D \subseteq \operatorname{cl} B$.

To prove the existence of a hyperplane that separates the two convex sets X and Y provided $\operatorname{rint} X \cap \operatorname{rint} Y = \emptyset$, it hence suffices to prove that there exists a hyperplane that separates the two convex sets $A = \operatorname{rint} X$ and $B = \operatorname{rint} Y$, because $\operatorname{rint} X \subseteq X \subseteq \operatorname{cl}(\operatorname{rint} X) = \operatorname{cl} X$, and the corresponding inclusions are of course also true for Y.

Since the sets A and B are disjoint, 0 does not belong to the convex set $A - B$. Thus, the point 0 either lies in the complement of $\operatorname{cl}(A - B)$ or belongs to $\operatorname{cl}(A - B)$ and is a relative boundary point of $\operatorname{cl}(A - B)$, because

$$\begin{aligned}\operatorname{cl}(A-B) \setminus (A-B) &\subseteq \operatorname{cl}(A-B) \setminus \operatorname{rint}(A-B)\\ &= \operatorname{rbdry}(A-B) = \operatorname{rbdry}(\operatorname{cl}(A-B)).\end{aligned}$$

In the first case it follows from Theorem 3.1.3 that there is a hyperplane that strictly separates 0 and $A - B$, and in the latter case Theorem 3.1.4 gives us a hyperplane that separates 0 from the set $\operatorname{cl}(A - B)$, and thereby afortiori also 0 from $A - B$. The existence of a hyperplane that separates A and B then follows from Lemma 3.1.2.

Now, let us turn to the converse. Assume that the hyperplane H separates the two convex sets X and Y. We will prove that there is no point that is a relative interior point of both sets. To this end, let us assume that x_0 is a point in the intersection $X \cap Y$. Then, x_0 lies in the hyperplane H because X and Y are subsets of opposite closed halfspaces determined by H. According to the separability definition, at least one of the two convex sets, X say, has points that lie outside H, and this clearly implies that the affine hull $V = \operatorname{aff} X$ is not a subset of H. Hence, there are points in V from each side of H. Therefore, the intersection $V \cap B(x_0; r)$ between V and an arbitrary open ball $B(x_0; r)$ centered at x_0 also contains points from both sides of H, and consequently surely points that do not belong to X. This means that x_0 must be a relative boundary point of X.

Hence, every point in the intersection $X \cap Y$ is a relative boundary point of either of the two sets X and Y. The intersection $\operatorname{rint} X \cap \operatorname{rint} Y$ is thus empty. □

Let us now consider the possibility of strict separation. A hyperplane that strictly separates two sets obviously also strictly separates their closures, so it suffices to examine when two closed convex subsets X and Y can be strictly separated. Of course, the two sets have to be disjoint, i.e. $0 \notin X - Y$ is a necessary condition, and Lemma 3.1.2 now reduces the problem of separating X strictly from Y to the problem of separating 0 strictly from $X - Y$. So it follows at once from Theorem 3.1.3 that there exists a separating hyperplane if the set $X - Y$ is closed. This gives us the following theorem, where the sufficient conditions follow from Theorem 2.7.11 and Corollary 2.7.12.

Theorem 3.1.6. *Two disjoint closed convex sets X and Y can be strictly separated by a hyperplane if the set $X - Y$ is closed, and a sufficient condition for this to be the case is* $\operatorname{recc} X \cap \operatorname{recc} Y = \{0\}$. *In particular, two disjoint closed convex set can be separated strictly by a hyperplane if one of the sets is bounded.*

We conclude this section with a result that shows that proper convex cones are proper subsets of conic halfspaces. More precisely, we have:

Theorem 3.1.7. *Let $X \neq \{0\}$ be a proper convex cone in $\mathbf{R}^n$, where $n \geq 2$. Then X is a proper subset of some conic halfspace $\{x \in \mathbf{R}^n \mid \langle c, x\rangle \geq 0\}$, whose boundary $\{x \in \mathbf{R}^n \mid \langle c, x\rangle = 0\}$ does not contain X as a subset.*

Proof. The point 0 is a relative boundary point of X, because no point on the line segment $]0, -a[$ belongs to X when a is a point $\neq 0$ in X. Hence, by Theorem 3.1.4, there exists a hyperplane $H = \{x \in \mathbf{R}^n \mid \langle c, x\rangle = 0\}$ through 0 such that X lies in the closed halfspace $K = \{x \in \mathbf{R}^n \mid \langle c, x\rangle \geq 0\}$ without

X being a subset of H. K is a conic halfspace, and the proper cone X must be different from K, since no conic halfspaces in $\mathbf{R}^n$ are proper cones when $n \geq 2$. □

3.2 The dual cone

To each subset A of $\mathbf{R}^n$ we associate a new subset A^+ of $\mathbf{R}^n$ by letting

$$A^+ = \{x \in \mathbf{R}^n \mid \langle a, x\rangle \geq 0 \text{ for all } a \in A\}.$$

In particular, for sets $\{a\}$ consisting of just one point we have

$$\{a\}^+ = \{x \in \mathbf{R}^n \mid \langle a, x\rangle \geq 0\},$$

which is a conic closed halfspace. For general sets A, $A^+ = \bigcap_{a\in A}\{a\}^+$, and this is an intersection of conic closed halfspaces. The set A^+ is thus in general a closed convex cone, and it is a polyhedral cone if A is a finite set.

Definition. The closed convex cone A^+ is called the *dual cone* of the set A.

The dual cone A^+ of a set A in $\mathbf{R}^n$ has an obvious geometric interpretation when $n \leq 3$; it consists of all vectors that form an acute angle or are perpendicular to all vectors in A.

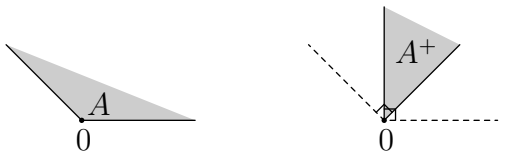

Figure 3.5. A cone A and its dual cone A^+.

Theorem 3.2.1. *The following properties hold for subsets A and B of $\mathbf{R}^n$.*

(i) $A \subseteq B \;\Rightarrow\; B^+ \subseteq A^+$;

(ii) $A^+ = (\operatorname{con} A)^+$;

(iii) $A^+ = (\operatorname{cl} A)^+$.

Proof. Property (i) is an immediate consequence of the definition of the dual cone.

To prove (ii) and (iii), we first observe that

$$(\operatorname{con} A)^+ \subseteq A^+ \text{ and } (\operatorname{cl} A)^+ \subseteq A^+,$$

because of property (i) and the obvious inclusions $A \subseteq \operatorname{con} A$ and $A \subseteq \operatorname{cl} A$.

It thus only remains to prove the converse inclusions. So let us assume that $x \in A^+$. Then

$$\langle \lambda_1 a_1 + \cdots + \lambda_k a_k, x\rangle = \lambda_1\langle a_1, x\rangle + \cdots + \lambda_k\langle a_k, x\rangle \geq 0$$

for all conic combinations of elements a_i in A. This proves the implication $x \in A^+ \Rightarrow x \in (\operatorname{con} A)^+$, i.e. the inclusion $A^+ \subseteq (\operatorname{con} A)^+$.

For each $a \in \operatorname{cl} A$ there exists a sequence $(a_k)_1^\infty$ of elements in A such that $a_k \to a$ as $k \to \infty$. If $x \in A^+$, then $\langle a_k, x\rangle \geq 0$ for all k, and it follows, by passing to the limit, that $\langle a, x\rangle \geq 0$. Since $a \in \operatorname{cl} A$ is arbitrary, this proves the implication $x \in A^+ \Rightarrow x \in (\operatorname{cl} A)^+$ and the inclusion $A^+ \subseteq (\operatorname{cl} A)^+$. □

EXAMPLE 3.2.1. Clearly, $(\mathbf{R}^n)^+ = \{0\}$ and $\{0\}^+ = \mathbf{R}^n$. □

EXAMPLE 3.2.2. Let, as usual, $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ denote the standard basis of $\mathbf{R}^n$. Then

$$\{\mathbf{e}_j\}^+ = \{x \in \mathbf{R}^n \mid \langle \mathbf{e}_j, x\rangle \geq 0\} = \{x \in \mathbf{R}^n \mid x_j \geq 0\}.$$

Since $\mathbf{R}^n_+ = \operatorname{con}\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, it follows that

$$(\mathbf{R}^n_+)^+ = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}^+ = \bigcap_{j=1}^n \{\mathbf{e}_j\}^+ = \{x \in \mathbf{R}^n \mid x_1 \geq 0, \ldots, x_n \geq 0\} = \mathbf{R}^n_+.$$

□

The bidual cone

Definition. The dual cone A^+ of a set A in $\mathbf{R}^n$ is a new set in $\mathbf{R}^n$, and we may therefore form the dual cone $(A^+)^+$ of A^+. The cone $(A^+)^+$ is called the *bidual cone* of A, and we write A^{++} instead of $(A^+)^+$.

Theorem 3.2.2. *Let A be an arbitrary set in $\mathbf{R}^n$. Then*

$$A \subseteq \operatorname{con} A \subseteq A^{++}.$$

Proof. The definitions of dual and bidual cones give us the implications

$$\begin{aligned} a \in A &\Rightarrow \langle x, a\rangle = \langle a, x\rangle \geq 0 \text{ for all } x \in A^+ \\ &\Rightarrow a \in A^{++}, \end{aligned}$$

which show that $A \subseteq A^{++}$. Since A^{++} is a cone and $\operatorname{con} A$ is the smallest cone containing A, we conclude that $\operatorname{con} A \subseteq A^{++}$. □

Because of the previous theorem, it is natural to ask when $\operatorname{con} A = A^{++}$. Since A^{++} is a closed cone, a necessary condition for this to be the case is that the cone $\operatorname{con} A$ be closed. Our next theorem shows that this condition is also sufficient.

Theorem 3.2.3. *Let X be a convex cone. Then $X^{++} = \operatorname{cl} X$, and consequently, $X^{++} = X$ if and only if the cone X is a closed.*

Proof. It follows from the inclusion $X \subseteq X^{++}$ and the closedness of the bidual cone X^{++} that $\operatorname{cl} X \subseteq X^{++}$.

To prove the converse inclusion $X^{++} \subseteq \operatorname{cl} X$, we assume that $x_0 \notin \operatorname{cl} X$ and will prove that $x_0 \notin X^{++}$.

By Theorem 3.1.3, there exists a hyperplane that strictly separates x_0 from $\operatorname{cl} X$. Hence, there exist a vector $c \in \mathbf{R}^n$ and a real number b such that the inequality $\langle c, x\rangle \geq b > \langle c, x_0\rangle$ holds for all $x \in X$. In particular, $t\langle c, x\rangle = \langle c, tx\rangle \geq b$ for all $x \in X$ and all numbers $t \geq 0$, since X is a cone, and this clearly implies that $b \leq 0$ and that $\langle c, x\rangle \geq 0$ for all $x \in X$. Hence, $c \in X^+$, and since $\langle c, x_0\rangle < b \leq 0$, we conclude that $x_0 \notin X^{++}$. □

By Theorem 2.6.11, finitely generated cones are closed, so we have the following immediate corollary.

Corollary 3.2.4. *If the cone X is finitely generated, then $X^{++} = X$.*

Example 3.2.3. The dual cone of the polyhedral cone

$$X = \bigcap_{i=1}^{m} \{x \in \mathbf{R}^n \mid \langle a_i, x\rangle \geq 0\}$$

is the cone

$$X^+ = \operatorname{con}\{a_1, a_2, \ldots, a_m\}.$$

This follows from the above corollary and Theorem 3.2.1, for

$$X = \{a_1, a_2, \ldots, a_m\}^+ = (\operatorname{con}\{a_1, a_2, \ldots, a_m\})^+.$$

If we use matrices to write the above cone X as $\{x \in \mathbf{R}^n \mid Ax \geq 0\}$, then the vector a_i corresponds to the ith column of the *transposed* matrix A^T (cf. Example 2.6.3), and the dual cone X^+ is consequently generated by the *columns* of A^T. Thus,

$$\{x \in \mathbf{R}^n \mid Ax \geq 0\}^+ = \{A^T y \mid y \in \mathbf{R}^m_+\}.$$ □

3.3 Solvability of systems of linear inequalities

Corollary 3.2.4 can be reformulated as a criterion for the solvability of systems of linear inequalities. The proof of this criterion uses the following lemma about dual cones.

Lemma 3.3.1. *Let X and Y be closed convex cones in $\mathbf{R}^n$. Then*

(i) $X \cap Y = (X^+ + Y^+)^+$;

(ii) $X + Y = (X^+ \cap Y^+)^+$, *provided that the cone $X + Y$ is closed.*

Proof. We have $X^+ \subseteq (X \cap Y)^+$ and $Y^+ \subseteq (X \cap Y)^+$, by Theorem 3.2.1 (i). Hence, $X^+ + Y^+ \subseteq (X \cap Y)^+ + (X \cap Y)^+ = (X \cap Y)^+$.

Another application of Theorem 3.2.1 in combination with Theorem 3.2.3 now yields $X \cap Y = (X \cap Y)^{++} \subseteq (X^+ + Y^+)^+$.

To obtain the converse inclusion we first deduce from $X^+ \subseteq X^+ + Y^+$ that $(X^+ + Y^+)^+ \subseteq X^{++} = X$, and the inclusion $(X^+ + Y^+)^+ \subseteq Y$ is of course obtained in the same way. Consequently, $(X^+ + Y^+)^+ \subseteq X \cap Y$. This completes the proof of property (i).

By replacing X and Y in (i) by the closed cones X^+ and Y^+, we obtain the equality $X^+ \cap Y^+ = (X^{++} + Y^{++})^+ = (X + Y)^+$, and since the cone $X + Y$ is assumed to be closed, we conclude that

$$X + Y = (X + Y)^{++} = (X^+ \cap Y^+)^+. \qquad \square$$

We are now ready for the promised result on the solvability of certain systems of linear inequalities, a result that will be used in our proof of the duality theorem in linear programming.

Theorem 3.3.2. *Let U be a finitely generated cone in $\mathbf{R}^n$, V be a finitely generated cone in $\mathbf{R}^m$, A be an $m \times n$-matrix and c be an $n \times 1$-matrix. Then the system*

$$\text{(S)} \qquad \begin{cases} Ax \in V^+ \\ \quad x \in U^+ \\ c^T x < 0 \end{cases}$$

has a solution x if and only if the system

$$\text{(S}^*\text{)} \qquad \begin{cases} c - A^T y \in U \\ \qquad\quad y \in V \end{cases}$$

has no solution y.

Proof. The system (S*) clearly has a solution if and only if $c \in (A^T(V) + U)$, and consequently, there is no solution if and only if $c \notin (A^T(V) + U)$. Therefore, it is worthwhile to take a closer look at the cone $A^T(V) + U$.

The cones $A^T(V)$, U and $A^T(V) + U$ are closed, since they are finitely generated. We may therefore apply Lemma 3.3.1 with

$$A^T(V) + U = \left(A^T(V)^+ \cap U^+\right)^+$$

as conclusion. The condition $c \notin (A^T(V)+U)$ is now seen to be equivalent to the existence of a vector $x \in A^T(V)^+ \cap U^+$ satisfying the inequality $c^T x < 0$, i.e. to the existence of an x such that

$$(\dagger) \qquad \begin{cases} x \in A^T(V)^+ \\ x \in U^+ \\ c^T x < 0\,. \end{cases}$$

It now only remains to translate the condition $x \in A^T(V)^+$; it is equivalent to the condition

$$\langle y, Ax\rangle = \langle A^T y, x\rangle \geq 0 \qquad \text{for all } y \in V,$$

i.e. to $Ax \in V^+$. The two systems (†) and (S) are therefore equivalent, and this observation completes the proof. □

By choosing $U = \{0\}$ and $V = \mathbf{R}^m_+$ with dual cones $U^+ = \mathbf{R}^n$ and $V^+ = \mathbf{R}^m_+$, we get the following special case of Theorem 3.3.2.

Corollary 3.3.3 (Farkas's lemma)**.** *Let A be an $m \times n$-matrix and c be an $n \times 1$-matrix, and consider the two systems:*

$$(\mathrm{S}) \quad \begin{cases} Ax \geq 0 \\ c^T x < 0 \end{cases} \qquad \textit{and} \qquad (\mathrm{S}^*) \quad \begin{cases} A^T y = c \\ y \geq 0 \end{cases}$$

The system (S) *has a solution if and only if the system* (S*) *has no solution.*

Example 3.3.1. The system

$$\begin{cases} x_1 - x_2 + 2x_3 \geq 0 \\ -x_1 + x_2 - x_3 \geq 0 \\ 2x_1 - x_2 + 3x_3 \geq 0 \\ 4x_1 - x_2 + 10x_3 < 0 \end{cases}$$

has no solution, because the dual system

$$\begin{cases} y_1 - y_2 + 2y_3 = 4 \\ -y_1 + y_2 - y_3 = -1 \\ 2y_1 - y_2 + 3y_3 = 10 \end{cases}$$

has a nonnegative solution $y = (3, 5, 3)$. □

EXAMPLE 3.3.2. The system

$$\begin{cases} 2x_1 + x_2 - x_3 \geq 0 \\ x_1 + 2x_2 - 2x_3 \geq 0 \\ x_1 - x_2 + x_3 \geq 0 \\ x_1 - 4x_2 + 4x_3 < 0 \end{cases}$$

is solvable, because the solutions of the dual system

$$\begin{cases} 2y_1 + y_2 + y_3 = 1 \\ y_1 + 2y_2 - y_3 = -4 \\ -y_1 - 2y_2 + y_3 = 4 \end{cases}$$

are of the form $y = (2-t, -3+t, t)$, $t \in \mathbf{R}$, and none of those is nonnegative since $y_1 < 0$ for $t > 2$ and $y_2 < 0$ for $t < 3$. □

The following generalization of Example 3.2.3 will be needed in Chapter 10.

Theorem 3.3.4. *Let $a_1, a_2, \ldots, a_m$ be vectors in $\mathbf{R}^n$, and let I, J be a partition of the index set $\{1, 2, \ldots, m\}$, i.e. $I \cap J = \emptyset$ and $I \cup J = \{1, 2, \ldots, m\}$. Let*

$$X = \bigcap_{i \in I} \{x \in \mathbf{R}^n \mid \langle a_i, x \rangle \geq 0\} \cap \bigcap_{i \in J} \{x \in \mathbf{R}^n \mid \langle a_i, x \rangle > 0\},$$

and suppose that $X \neq \emptyset$. Then

$$X^+ = \operatorname{con}\{a_1, a_2, \ldots, a_m\}.$$

Proof. Let

$$Y = \bigcap_{i=1}^{m} \{x \in \mathbf{R}^n \mid \langle a_i, x \rangle \geq 0\}.$$

The set Y is closed and contains X, and we will prove that $Y = \operatorname{cl} X$ by showing that every neighborhood of an arbitrary point $y \in Y$ contains points from X.

So, fix a point $x_0 \in X$, and consider the points $y + tx_0$ for $t > 0$. These points lie in X, for

$$\langle a_i, y + tx_0 \rangle = \langle a_i, y \rangle + t\langle a_i, x_0 \rangle = \begin{cases} \geq 0 & \text{if } i \in I \\ > 0 & \text{if } i \in J, \end{cases}$$

and since $y + tx_0 \to y$ as $t \to 0$, there are indeed points in X arbitrarily close to y.

Hence $X^+ = (\operatorname{cl} X)^+ = Y^+$, by Theorem 3.2.1, and the conclusion of the theorem now follows from the result in Example 3.2.3. □

How do we decide whether the set X in Theorem 3.3.4 is nonempty? If just one of the m linear inequalities that define X is strict (i.e. if the index set J consists of one element), then Farkas's lemma gives a necessary and sufficient condition for X to be nonempty. A generalization to the general case reads as follows.

Theorem 3.3.5. *The set X in Theorem 3.3.4 is nonempty if and only if*

$$\begin{cases} \sum_{i=1}^{m} \lambda_i a_i = 0 \\ \lambda_i \geq 0 \ \textit{for all } i \end{cases} \quad \Rightarrow \quad \lambda_i = 0 \ \textit{for all } i \in J.$$

Proof. Let the vectors $\hat{a}_i$ in $\mathbf{R}^{n+1}$ $(= \mathbf{R}^n \times \mathbf{R})$ be defined by

$$\hat{a}_i = \begin{cases} (a_i, 0) & \text{if } i \in I \\ (a_i, 1) & \text{if } i \in J. \end{cases}$$

Write $\tilde{x} = (x, x_{n+1})$, and let $\tilde{X}$ be polyhedral cone

$$\tilde{X} = \bigcap_{i=1}^{m} \{\tilde{x} \in \mathbf{R}^{n+1} \mid \langle \hat{a}_i, \tilde{x} \rangle \geq 0\} = (\operatorname{con}\{\hat{a}_1, \ldots, \hat{a}_m\})^+.$$

Since

$$\langle \hat{a}_i, (x, x_{n+1}) \rangle = \begin{cases} \langle a_i, x \rangle & \text{if } i \in I \\ \langle a_i, x \rangle + x_{n+1} & \text{if } i \in J, \end{cases}$$

and $\langle a_i, x \rangle > 0$ for all $i \in J$ if and only if there exists a negative real number x_{n+1} such that $\langle a_i, x \rangle + x_{n+1} \geq 0$ for all $i \in J$, we conclude that the point x lies in X if and only if there exists a negative number x_{n+1} such that $\langle \hat{a}_i, (x, x_{n+1}) \rangle \geq 0$ for all i, i.e. if and only if there exists a negative number x_{n+1} such that $(x, x_{n+1}) \in \tilde{X}$. This is equivalent to saying that the set X is empty if and only if the implication $\tilde{x} \in \tilde{X} \Rightarrow x_{n+1} \geq 0$ is true, i.e. if and only if $\tilde{X} \subseteq \mathbf{R}^n \times \mathbf{R}_+$. Using the results on dual cones in Theorems 3.2.1 and 3.2.3 we thus obtain the following chain of equivalences:

$$\begin{aligned} X = \emptyset \ &\Leftrightarrow \ \tilde{X} \subseteq \mathbf{R}^n \times \mathbf{R}_+ \\ &\Leftrightarrow \ \{0\} \times \mathbf{R}_+ = (\mathbf{R}^n \times \mathbf{R}_+)^+ \subseteq \tilde{X}^+ = \operatorname{con}\{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_m\} \\ &\Leftrightarrow \ (0,1) \in \operatorname{con}\{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_m\} \\ &\Leftrightarrow \ \text{there are numbers } \lambda_i \geq 0 \text{ such that } \sum_{i=1}^{m} \lambda_i a_i = 0 \text{ and } \sum_{i \in J} \lambda_i = 1 \\ &\Leftrightarrow \ \text{there are numbers } \lambda_i \geq 0 \text{ such that } \sum_{i=1}^{m} \lambda_i a_i = 0 \text{ and } \lambda_i > 0 \text{ for} \\ &\qquad \text{some } i \in J. \end{aligned}$$

(The last equivalence holds because of the homogenouity of the condition $\sum_{i=1}^{m} \lambda_i a_i = 0$. If the condition is fulfilled for a set of nonnegative numbers λ_i with $\lambda_i > 0$ for at least one $i \in J$, then we can certainly arrange so that $\sum_{i \in J} \lambda_i = 1$ by multiplying with a suitable constant.)

Since the first and the last assertion in the above chain of equivalences are equivalent, so are their negations, and this is the statement of the theorem. □

The following corollary is an immediate consequence of Theorem 3.3.5.

Corollary 3.3.6. *The set X in Theorem 3.3.4 is nonempty if the vectors $a_1, a_2, \ldots, a_m$ are linearly independent.*

The following equivalent matrix version of Theorem 3.3.5 is obtained by considering the vectors a_i, $i \in I$ and a_i, $i \in J$ in Theorems 3.3.4 and 3.3.5 as rows in two matrices A and C, respectively.

Theorem 3.3.7. *Let A be a $p \times n$-matrix and C be $q \times n$-matris. Then exactly one of the two dual systems*

$$\begin{cases} Ax \geq 0 \\ Cx > 0 \end{cases} \qquad \textit{and} \qquad \begin{cases} A^T y + C^T z = 0 \\ y, z \geq 0,\ z \neq 0 \end{cases}$$

has a solution.

Theorem 3.3.7 will be generalized in Chapter 6.5, where we prove a theorem on the solvability of systems of convex and affine inequalities.

Exercises

3.1 Find two disjoint closed convex sets in $\mathbf{R}^2$ that are not strictly separable by a hyperplane (i.e. by a line in $\mathbf{R}^2$).

3.2 Let X be a convex proper subset of $\mathbf{R}^n$. Show that X is an intersection of closed halfspaces if X is closed, and an intersection of open halfspaces if X is open.

3.3 Prove the following converse of Lemma 3.1.2: If two sets X and Y are (strictly) separable, then $X - Y$ and 0 are (strictly) separable.

3.4 Find the dual cones of the following cones in $\mathbf{R}^2$:
a) $\mathbf{R}_+ \times \{0\}$ b) $\mathbf{R} \times \{0\}$ c) $\mathbf{R} \times \mathbf{R}_+$ d) $(\mathbf{R}_{++} \times \mathbf{R}_{++}) \cup \{(0,0)\}$
e) $\{x \in \mathbf{R}^2 \mid x_1 + x_2 \geq 0,\ x_2 \geq 0\}$

3.5 Prove for arbitrary sets X and Y that $(X \times Y)^+ = X^+ \times Y^+$.

3.6 Determine the cones X, X^+ and X^{++}, if $X = \operatorname{con} A$ and

a) $A = \{(1,0),(1,1),(-1,1)\}$ b) $A = \{(1,0),(-1,1),(-1,-1)\}$

c) $A = \{x \in \mathbf{R}^2 \mid x_1 x_2 = 1,\ x_1 > 0\}$.

3.7 Let $A = \{a_1, a_2, \ldots, a_m\}$ be a subset of $\mathbf{R}^n$, and suppose $0 \notin A$. Prove that the following three conditions are equivalent:

(i) $\operatorname{con} A$ is a proper cone.

(ii) $\sum_{j=1}^m \lambda_j a_j = 0$, $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m) \geq 0 \ \Rightarrow \ \lambda = 0$.

(iii) There is a vector c such that $\langle c, a\rangle > 0$ for all $a \in A$.

3.8 Is the following system consistent?

$$\begin{cases} x_1 - 2x_2 - 7x_3 \geq 0 \\ 5x_1 + x_2 - 2x_3 \geq 0 \\ x_1 + 2x_2 + 5x_3 \geq 0 \\ 18x_1 + 5x_2 - 3x_3 < 0 \end{cases}$$

3.9 Show that

$$\begin{cases} x_1 + x_2 - x_3 \geq 2 \\ x_1 - x_2 \geq 1 \\ x_1 + x_3 \geq 3 \end{cases} \quad \Rightarrow \quad 6x_1 - 2x_2 + x_3 \geq 11.$$

3.10 For which values of the parameter $\alpha \in \mathbf{R}$ is the system

$$\begin{cases} x_1 + x_2 + \alpha x_3 \geq 0 \\ x_1 + \alpha x_2 + x_3 \geq 0 \\ \alpha x_1 + x_2 + x_3 \geq 0 \\ x_1 + \alpha x_2 + \alpha^2 x_3 < 0 \end{cases}$$

solvable?

3.11 Let A be an $m \times n$-matrix. Prove that exactly one of the two systems (S) and (S*) has a solution if

a) (S) $\begin{cases} Ax = 0 \\ x \geq 0 \\ x \neq 0 \end{cases}$ and (S*) $A^T y > 0$

b) (S) $\begin{cases} Ax = 0 \\ x > 0 \end{cases}$ and (S*) $\begin{cases} A^T y \geq 0 \\ A^T y \neq 0\,. \end{cases}$

3.12 Prove that the following system of linear inequalities is solvable:

$$\begin{cases} Ax = 0 \\ x \geq 0 \\ A^T y \geq 0 \\ A^T y + x > 0\,. \end{cases}$$

Chapter 4

More on convex sets

4.1 Extreme points and faces

Extreme point

Polyhedra, like the one in figure 4.1, have vertices. A vertex is characterized by the fact that it is not an interior point of any line segment that lies entirely in the polyhedron. This property is meaningful for arbitrary convex sets.

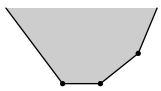

Figure 4.1. A polyhedron with vertices.

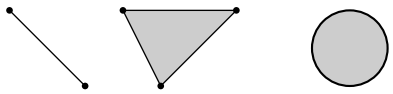

Figure 4.2. Extreme points of a line segment, a triangle and a circular disk.

Definition. A point x in a convex set X is called an *extreme point* of the set if it does not lie in any open line segment joining two points of X, i.e. if

$$a_1,\, a_2 \in X \;\&\; a_1 \neq a_2 \;\;\Rightarrow\;\; x \notin]a_1, a_2[.$$

The set of all extreme points of X will be denoted by $\operatorname{ext} X$.

A point in the relative interior of a convex set is clearly never an extreme point, except when the convex set consists just one point.[†] With an exception for this trivial case, $\operatorname{ext} X$ is consequently a subset of the relative boundary of X. In particular, open convex sets have no extreme points.

[†]For if $X = \{x_0\}$, then $\operatorname{rint} X = \{x_0\}$, $\operatorname{rbdry} X = \emptyset$ and $\operatorname{ext} X = \{x_0\}$.

EXAMPLE 4.1.1. The two endpoints are the extreme points of a closed line segment. The three vertices are the extreme points of a triangle. All points on the boundary $\{x \mid \|x\|_2 = 1\}$ are extreme points of the Euclidean closed unit ball $\overline{B}(0;1)$ in $\mathbf{R}^n$. □

Extreme ray

The extreme point concept is of no interest for convex cones, because non-proper cones have no extreme points, and proper cones have 0 as their only extreme point. Instead, for cones the correct extreme concept is about rays, and in order to define it properly we first have to define what it means for a ray to lie between to rays.

Definition. We say that the ray $R = \overrightarrow{a}$ *lies between* the two rays $R_1 = \overrightarrow{a_1}$ and $R_2 = \overrightarrow{a_2}$ if the two vectors a_1 and a_2 are linearly independent and there exist two positive numbers λ_1 and λ_2 so that $a = \lambda_1 a_1 + \lambda_2 a_2$.

It is easy to convince oneself that the concept "lie between" only depends on the rays R, R_1 and R_2, and not on the vectors a, a_1 and a_2 chosen to represent them. Furthermore, a_1 and a_2 are linearly independent if and only if the rays R_1 and R_2 are different and not opposite to each other, i.e. if and only if $R_1 \neq \pm R_2$.

Definition. A ray R in a convex cone X is called an *extreme ray* of the cone if the following two conditions are satisfied:

(i) the ray R does not lie between any rays in the cone X;
(ii) the opposite ray $-R$ does not lie in X.

The set of all extreme rays of X is denoted by $\operatorname{exr} X$.

The second condition (ii) is automatically satisfied for all proper cones, and it implies, as we shall see later (Theorem 4.2.4), that non-proper cones have no extreme rays.

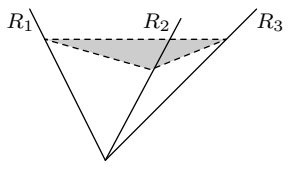

Figure 4.3. A polyhedral cone in $\mathbf{R}^3$ with three extreme rays.

It follows from the definition that no extreme ray of a convex cone with dimension greater than 1 can pass through a relative interior point of the cone. The extreme rays of a cone of dimension greater than 1 are in other words subsets of the relative boundary of the cone.

EXAMPLE 4.1.2. The extreme rays of the four subcones of $\mathbf{R}$ are as follows: $\operatorname{exr}\{0\} = \operatorname{exr}\mathbf{R} = \emptyset$, $\operatorname{exr}\mathbf{R}_+ = \mathbf{R}_+$ and $\operatorname{exr}\mathbf{R}_- = \mathbf{R}_-$.

The non-proper cone $\mathbf{R} \times \mathbf{R}_+$ in $\mathbf{R}^2$ (the "upper halfplane") has no extreme rays, since the two boundary rays $\mathbf{R}_+ \times \{0\}$ and $\mathbf{R}_- \times \{0\}$ are disqualified by condition (ii) of the extreme ray definition. □

Face

Definition. A subset F of a convex set X is called a *proper face* of X if $F = X \cap H$ for some supporting hyperplane H of X. In addition, the set X itself and the empty set $\emptyset$ are called *non-proper faces* of X.‡

The reason for including the set itself and and the empty set among the faces is that it simplifies the wording of some theorems and proofs.

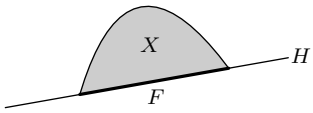

Figure 4.4. A convex set X with F as one its faces.

The faces of a convex set are obviously convex sets. And the proper faces of a convex cone are cones, since the supporting hyperplanes of a cone must pass through the origin and thus be linear subspaces.

EXAMPLE 4.1.3. Every point on the boundary $\{x \mid \|x\|_2 = 1\}$ is a face of the closed unit ball $\overline{B}(0;1)$, because the tangent plane at a boundary point is a supporting hyperplane and does not intersect the unit ball in any other point. □

EXAMPLE 4.1.4. A cube in $\mathbf{R}^3$ has 26 proper faces: 8 vertices, 12 edges and 6 sides. □

‡There is an alternative and more general definition of the face concept, see exercise 4.7. Our proper faces are called *exposed faces* by Rockafellar in his standard treatise *Convex Analysis*. Every exposed face is also a face according to the alternative definition, but the two definitions are not equivalent, because there are convex sets with faces that are not exposed.

Theorem 4.1.1. *The relative boundary of a closed convex set X is equal to the union of all proper faces of X.*

Proof. We have to prove that $\operatorname{rbdry} X = \bigcup F$, where the union is taken over all proper faces F of X. So suppose that $x_0 \in F$, where $F = X \cap H$ is a proper face of X, and H is a supporting hyperplane. Since H supports X at x_0, and since, by definition, X is not contained in H, it follows that x_0 is a relative boundary point of X. This proves the inclusion $\bigcup F \subseteq \operatorname{rbdry} X$.

Conversely, if x_0 is a relative boundary point of X, then there exists a hyperplane H that supports X at x_0, and this means that x_0 lies in the proper face $X \cap H$. □

Theorem 4.1.2. *The intersection of two faces of a convex set is a face of the set.*

Proof. Let F_1 and F_2 be two faces of the convex set X, and let $F = F_1 \cap F_2$. That F is a face is trivial if the two faces F_1 and F_2 are identical, or if they are disjoint, or if one of them is non-proper.

So suppose that the two faces F_1 and F_2 are distinct and proper, i.e. that they are of the form $F_i = X \cap H_i$, where H_1 och H_2 are distinct supporting hyperplanes of the set X, and $F \neq \emptyset$. Let

$$H_i = \{x \in \mathbf{R}^n \mid \langle c_i, x\rangle = b_i\},$$

where the normal vectors c_i of the hyperplanes are chosen so that X lies in the two halfspaces $\{x \in \mathbf{R}^n \mid \langle c_i, x\rangle \leq b_i\}$, and let $x_1 \in X$ be a point satisfying the condition $\langle c_1, x_1\rangle < b_1$.

The hyperplanes H_1 and H_2 must be non-parallel, since $X \cap H_1 \cap H_2 = F \neq \emptyset$. Hence $c_2 \neq -c_1$, and we obtain a new hyperplane

$$H = \{x \in \mathbf{R}^n \mid \langle c, x\rangle = b\}$$

by defining $c = c_1 + c_2$ and $b = b_1 + b_2$. We will show that H is a supporting hyperplane of X and that $F = X \cap H$, which proves our claim that the intersection F is a face of X.

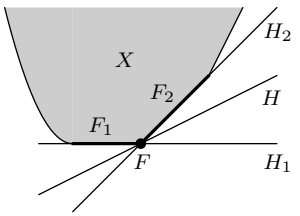

Figure 4.5. Illustration for the proof of Theorem 4.1.2.

For all $x \in X$, we have the inequality

$$\langle c, x\rangle = \langle c_1, x\rangle + \langle c_2, x\rangle \leq b_1 + b_2 = b,$$

and the inequality is strict for the particular point $x_1 \in X$, since

$$\langle c, x_1\rangle = \langle c_1, x_1\rangle + \langle c_2, x_1\rangle < b_1 + b_2 = b.$$

So X lies in one of the two closed halfspaces determined by H without being a subset of H. Moreover, for all $x \in F = X \cap H_1 \cap H_2$,

$$\langle c, x\rangle = \langle c_1, x\rangle + \langle c_2, x\rangle = b_1 + b_2 = b.$$

which implies that H is a supporting hyperplane of X and that $F \subseteq X \cap H$.

Conversely, if $x \in X \cap H$, then $\langle c_1, x\rangle \leq b_1$, $\langle c_2, x\rangle \leq b_2$ and

$$\langle c_1, x\rangle + \langle c_2, x\rangle = b_1 + b_2,$$

and this implies that $\langle c_1, x\rangle = b_1$ and $\langle c_2, x\rangle = b_2$. Hence, $x \in X \cap H_1 \cap H_2 = F$. This proves the inclusion $X \cap H \subseteq F$. □

Theorem 4.1.3. *(i) Suppose F is a face of the convex set X. A point x in F is an extreme point of F if and only if x is an extreme point of X.*

(ii) Suppose F is a face of the convex cone X. A ray R in F is an extreme ray of F if and only if R is an extreme ray of X.

Proof. Since the assertions are trivial for non-proper faces, we may assume that $F = X \cap H$ for some supporting hyperplane H of X.

No point in a hyperplane lies in the interior of a line segment whose endpoints both lie in the same halfspace, unless both endpoints lie in the hyperplane, i.e. unless the line segment lies entirely in the hyperplane.

Analogously, no ray in a hyperplane H (through the origin) lies between two rays in the same closed halfspace determined by H, unless both these rays lie in the hyperplane H. And the opposite ray $-R$ of a ray R in a hyperplane clearly lies in the same hyperplane.

(i) If $x \in F$ is an interior point of some line segment with both endpoints belonging to X, then x is in fact an interior point of a line segment whose

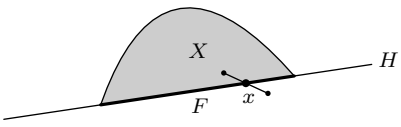

Figure 4.6. The two endpoints of an open line segment that intersects the hyperplane H must either both belong to H or else lie in opposite open halfspaces.

endpoints both belong to F. This proves the implication

$$x \notin \operatorname{ext} X \Rightarrow x \notin \operatorname{ext} F.$$

The converse implication is trivial, since every line segment in F is a line segment i X. Hence, $x \notin \operatorname{ext} X \Leftrightarrow x \notin \operatorname{ext} F$, and this is of course equivalent to assertion (i).

(ii) Suppose R is a ray in F and that R is not an extreme ray of the cone X. Then there are two possibilities: R lies between two rays R_1 and R_2 in X, or the opposite ray $-R$ lies in X. In the first case, R_1 and R_2 will necessarily lie in F, too. In the second case, the ray $-R$ will lie in F. Thus, both cases lead to the conclusion that R is not an extreme ray of the cone F, and this proves the implication $R \notin \operatorname{exr} X \Rightarrow R \notin \operatorname{exr} F$.

The converse implication is again trivial, and this observation concludes the proof of assertion (ii). □

4.2 Structure theorems for convex sets

Theorem 4.2.1. *Let X be a line-free closed convex set with* $\dim X \geq 2$. *Then*

$$X = \operatorname{cvx}(\operatorname{rbdry} X).$$

Proof. Let $n = \dim X$. By identifying the affine hull of X with $\mathbf{R}^n$, we may without loss of generality assume that X is a subset of $\mathbf{R}^n$ of full dimension, so that $\operatorname{rbdry} X = \operatorname{bdry} X$. To prove the theorem it is now enough to prove that every point in X lies in the convex hull of the boundary $\operatorname{bdry} X$, because the inclusionen $\operatorname{cvx}(\operatorname{bdry} X) \subseteq X$ is trivially true.

The recession cone $C = \operatorname{recc} X$ is a proper cone, since X is supposed to be line-free. Hence, there exists a closed halfspace

$$K = \{x \in \mathbf{R}^n \mid \langle c, x\rangle \geq 0\},$$

which contains C as a proper subset, by Theorem 3.1.7. Since C is a closed cone, we conclude that the corresponding open halfspace

$$K_+ = \{x \in \mathbf{R}^n \mid \langle c, x\rangle > 0\}.$$

contains a vector v that does not belong to C. The opposite vector $-v$, which lies in the opposite open halfspace, does not belong to C, either. Compare figure 4.7.

We have produced two opposite vectors $\pm v$, both lying outside the recession cone. The two opposite halflines $x + \overrightarrow{v}$ and $x - \overrightarrow{v}$ from a point $x \in X$ therefore both intersect the complement of X. The intersection between X

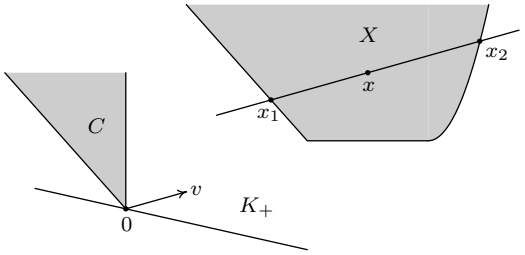

Figure 4.7. An illustration for the proof of Theorem 4.2.1.

and the line through x with direction vector v, which is a closed convex set, is thus either a closed line segment $[x_1, x_2]$ containing x and with endpoints belonging to the boundary of X, or the singleton set $\{x\}$ with x belonging to the boundary of X. In the first case, x is a convex combination of the boundary points x_1 and x_2. So, x lies in the convex hull of the boundary in both cases. This completes the proof of the theorem. □

It is now possible to give a complete description of line-free closed convex sets in terms of extreme points and recession cones.

Theorem 4.2.2. *A nonempty closed convex set X has extreme points if and only if X is line-free, and if X is line-free, then*

$$X = \operatorname{cvx}(\operatorname{ext} X) + \operatorname{recc} X.$$

Proof. First suppose that the set X is not line-free. Its recessive subspace will then, by definition, contain a nonzero vector y, and this implies that the two points $x \pm y$ lie in X for each point $x \in X$. Therefore, x being the midpoint of the line segment $]x - y, x + y[$, is not an extreme point. This proves that the set $\operatorname{ext} X$ of extreme points is empty.

Next suppose that X is line-free. We claim that $\operatorname{ext} X \neq \emptyset$ and that $X = \operatorname{cvx}(\operatorname{ext} X) + \operatorname{recc} X$, and we will prove this by induction on the dimension of the set X.

Our claim is trivially true for zero-dimensional sets X, i.e. sets consisting of just one point. If $\dim X = 1$, then either X is a halfline $a + \overrightarrow{v}$ with one extreme point a and recession cone equal to $\overrightarrow{v}$, or a line segment $[a, b]$ with two extreme points a, b and recession cone equal to $\{0\}$, and the equality in the theorem is clearly satisfied in both cases.

Now assume that $n = \dim X \geq 2$ and that our claim is true for all line-free closed convex sets X with dimension less than n. By Theorems 4.1.1 and 4.2.1,

$$X = \operatorname{cvx}\bigl(\bigcup F\bigr),$$

where the union is taken over all proper faces F of X. Each proper face F is a nonempty line-free closed convex subset of a supporting hyperplane H and has a dimension which is less than or equal to $n-1$. Therefore, $\operatorname{ext} F \neq \emptyset$ and

$$F = \operatorname{cvx}(\operatorname{ext} F) + \operatorname{recc} F,$$

by our induction hypothesis.

Since $\operatorname{ext} F \subseteq \operatorname{ext} X$ (by Theorem 4.1.3), it follows that $\operatorname{ext} X \neq \emptyset$. Moreover, $\operatorname{recc} F$ is a subset of $\operatorname{recc} X$, so we have the inclusion

$$F \subseteq \operatorname{cvx}(\operatorname{ext} X) + \operatorname{recc} X$$

for each face F. The union $\bigcup F$ is consequently included in the convex set $\operatorname{cvx}(\operatorname{ext} X) + \operatorname{recc} X$. Hence

$$X = \operatorname{cvx}\big(\bigcup F\big) \subseteq \operatorname{cvx}(\operatorname{ext} X) + \operatorname{recc} X \subseteq X + \operatorname{recc} X = X,$$

so $X = \operatorname{cvx}(\operatorname{ext} X) + \operatorname{recc} X$, and this completes the induction and the proof of the theorem. □

The recession cone of a compact set is equal to the null cone, and the following result is therefore an immediate corollary of Theorem 4.2.2.

Corollary 4.2.3. *Each nonempty compact convex set has extreme points and is equal to the convex hull of its extreme points.*

We shall now formulate and prove the anaologue of Theorem 4.2.2 for convex cones, and in order to simplify the notation we shall use the following convention: If $\mathcal{A}$ is a family of rays, we let $\operatorname{con}\mathcal{A}$ denote the cone

$$\operatorname{con}\Big(\bigcup_{R\in\mathcal{A}} R\Big),$$

i.e. $\operatorname{con}\mathcal{A}$ is the cone that is generated by the vectors on the rays in the family $\mathcal{A}$. If we choose a nonzero vector a_R on each ray $R \in \mathcal{A}$ and let $A = \{a_R \mid R \in \mathcal{A}\}$, then of course $\operatorname{con}\mathcal{A} = \operatorname{con} A$.

The cone $\operatorname{con}\mathcal{A}$ is clearly finitely generated if $\mathcal{A}$ is a finite family of rays, and we obtain a set of generators by choosing one nonzero vector from each ray.

Theorem 4.2.4. *A closed convex cone X has extreme rays if and only if the cone is proper and not equal to the null cone $\{0\}$. If X is a proper closed convex cone, then*

$$X = \operatorname{con}(\operatorname{exr} X).$$

Proof. First suppose that the cone X is not proper, and let $R = \overrightarrow{x}$ be an arbitrary ray in X. We will prove that R can not be an extreme ray.

Since X is non-proper, there exists a nonzero vector a in the intersection $X \cap (-X)$. First suppose that R is equal to $\overrightarrow{a}$ or to $-\overrightarrow{a}$. Then both R and its opposite ray $-R$ lie in X, and this means that R is not an extreme ray.

Next suppose $R \neq \pm\overrightarrow{a}$. The vectors x and a are then linearly independent, and the two rays $R_1 = \overrightarrow{x+a}$ and $R_2 = \overrightarrow{x-a}$ are consequently distinct and non-opposite rays in the cone X. Since $x = \frac{1}{2}(x+a) + \frac{1}{2}(x-a)$, we conclude that R lies between R_1 and R_2. Thus, R is not an extreme ray in this case either, and this proves that non-proper cones have no extreme rays.

The equality $X = \operatorname{con}(\operatorname{exr} X)$ is trivially true for the null cone, since $\operatorname{exr}\{0\} = \emptyset$ and $\operatorname{con}\emptyset = \{0\}$. To prove that the equality holds for all non-trivial proper closed convex cones and that these cones do have extreme rays, we only have to modify slightly the induction proof for the corresponding part of Theorem 4.2.2.

The start of the induction is of course trivial, since one-dimensional proper cones are rays. So suppose our assertion is true for all cones of dimension less than or equal to $n-1$, and let X be a proper closed n-dimensional convex cone. X is then, in particular, a line-free set, whence $X = \operatorname{cvx}\big(\bigcup F\big)$, where the union is taken over all proper faces F of the cone. Moreover, since X is a convex cone, $\operatorname{cvx}\big(\bigcup F\big) \subseteq \operatorname{con}\big(\bigcup F\big) \subseteq \operatorname{con} X = X$, and we conclude that

$$X = \operatorname{con}\big(\bigcup F\big). \tag{4.1}$$

We may of course delete the trivial face $F = \{0\}$ from the above union without destroying the identity, and every remaining face F is a proper closed convex cone of dimension less than or equal to $n-1$ with $\operatorname{exr} F \neq \emptyset$ and $F = \operatorname{con}(\operatorname{exr} F)$, by our induction assumption. Since $\operatorname{exr} F \subseteq \operatorname{exr} X$, it now follows that the set $\operatorname{exr} X$ is nonempty and that $F \subseteq \operatorname{con}(\operatorname{exr} X)$.

The union $\bigcup F$ of the faces is thus included in the cone $\operatorname{con}(\operatorname{exr} X)$, so it follows from equation (4.1) that $X \subseteq \operatorname{con}(\operatorname{exr} X)$. Since the converse inclusion is trivial, we have equality $X = \operatorname{con}(\operatorname{exr} X)$, and the induction step is now complete. □

The recession cone of a line-free convex set is a proper cone. The following structure theorem for convex sets is therefore an immediate consequence of Theorems 4.2.2 and 4.2.4.

Theorem 4.2.5. *If X is a nonempty line-free closed convex set, then*

$$X = \operatorname{cvx}(\operatorname{ext} X) + \operatorname{con}(\operatorname{exr}(\operatorname{recc} X)).$$

The study of arbitrary closed convex sets is reduced to the study of line-free such sets by the following theorem, which says that every non-line-free closed convex set is a cylinder with a line-free convex set as its base and with the recessive subspace $\lin X$ as its "axis".

Theorem 4.2.6. *Suppose X is a closed convex set in $\mathbf{R}^n$. The intersection $X \cap (\lin X)^\perp$ is then a line-free closed convex set and*

$$X = \lin X + X \cap (\lin X)^\perp.$$

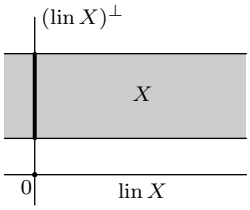

Figure 4.8. Illustration for Theorem 4.2.6.

Proof. Each $x \in \mathbf{R}^n$ has a unique decomposition $x = y + z$ with $y \in \lin X$ and $z \in (\lin X)^\perp$. If $x \in X$, then z lies in X, too, since

$$z = x - y \in X + \lin X = X.$$

This proves the inclusion $X \subseteq \lin X + X \cap (\lin X)^\perp$, and the converse inclusion follows from $\lin X + X \cap (\lin X)^\perp \subseteq \lin X + X = X$. □

Exercises

4.1 Find $\ext X$ and decide whether $X = \operatorname{cvx}(\ext X)$ when

a) $X = \{x \in \mathbf{R}^2_+ \mid x_1 + x_2 \geq 1\}$

b) $X = \big([0,1] \times [0,1[\,\big) \cup \big([0, \tfrac{1}{2}] \times \{1\}\big)$

c) $X = \operatorname{cvx}\big(\{x \in \mathbf{R}^3 \mid (x_1 - 1)^2 + x_2^2 = 1,\, x_3 = 0\} \cup \{(0,0,1), (0,0,-1)\}\big)$.

4.2 Prove that $\ext(\operatorname{cvx} A) \subseteq A$ for each subset A of $\mathbf{R}^n$.

4.3 Let $X = \operatorname{cvx} A$ and suppose the set A is minimal in the following sense: If $B \subseteq A$ och $X = \operatorname{cvx} B$, then $B = A$. Prove that $A = \ext X$.

4.4 Let x_0 be a point in a convex set X. Prove that $x_0 \in \ext X$ if and only if the set $X \setminus \{x_0\}$ is convex.

4.5 Give an example of a compact convex subset of $\mathbf{R}^3$ such that the set of extreme points is not closed.

4.6 A point x_0 in a convex set X is called an *exposed point* if the singleton set $\{x_0\}$ is a face, i.e. if there exists a supporting hyperplane H of X such that $X \cap H = \{x_0\}$.

a) Prove that every exposed point is an extreme point of X.

b) Give an example of a closed convex set in $\mathbf{R}^2$ with an extreme point that is not exposed.

4.7 There is a more general definition of the face concept which runs as follows:

A *face* of a convex set X is a convex subset F of X such that every closed line segment in X with a relative interior point in F lies entirely in F, i.e.

$$(a, b \in X \;\&\;]a, b[\cap F \neq \emptyset) \implies a, b \in F.$$

Let us call faces according to this definition *general faces* in order to distinguish them from faces according to our old definition, which we call *exposed faces*, provided they are proper, i.e. different from the faces X and $\emptyset$.

The empty set $\emptyset$ and X itself are apparently general faces of X, and all extreme points of X are general faces, too.

Prove that the general faces of a convex set X have the following properties.

a) Each exposed face is a general face.

b) There is a convex set with a general face that is not an exposed face.

c) If F is a general face of X and F' is a general face of F, then F' is a general face of X, but the coresponding result is not true in general for exposed faces.

d) If F is a general face of X and C is an arbitrary convex subset of X such that $F \cap \operatorname{rint} C \neq \emptyset$, then $C \subseteq F$.

e) If F is a general face of X, then $F = X \cap \operatorname{cl} F$. In particular, F is closed if X is closed.

f) If F_1 and F_2 are two general faces of X and $\operatorname{rint} F_1 \cap \operatorname{rint} F_2 \neq \emptyset$, then $F_1 = F_2$.

g) If F is a general face of X and $F \neq X$, then $F \subseteq \operatorname{rbdry} X$.

Chapter 5

Polyhedra

We have already obtained some isolated results on polyhedra, but now is the time to collect these and to complement them in order to get a complete description of this important class of convex sets.

5.1 Extreme points and extreme rays

Polyhedra and extreme points

Each polyhedron X in $\mathbf{R}^n$, except for the entire space, is an intersection of finitely many closed halfspaces and may therefore be written in the form

$$X = \bigcap_{j=1}^{m} K_j,$$

with

$$K_j = \{x \in \mathbf{R}^n \mid \langle c_j, x\rangle \geq b_j\}$$

for suitable nonzero vectors c_j in $\mathbf{R}^n$ and real numbers b_j. Using matrix notation,

$$X = \{x \in \mathbf{R}^n \mid Cx \geq b\},$$

where C is an $m \times n$-matrix with ${c_j}^T$ as rows, and $b = \begin{bmatrix} b_1 & b_2 & \dots & b_m \end{bmatrix}^T$.

Let

$$K_j^\circ = \{x \in \mathbf{R}^n \mid \langle c_j, x\rangle > b_j\} = \operatorname{int} K_j, \quad \text{and}$$
$$H_j = \{x \in \mathbf{R}^n \mid \langle c_j, x\rangle = b_j\} = \operatorname{bdry} K_j.$$

The sets K_j° are open halfspaces, and the H_j are hyperplanes.

If $b = 0$, i.e. if all hyperplanes H_j are linear subspaces, then X is a polyhedral cone.

The polyhedron X is clearly a subset of the closed halfspace K_j, which is bounded by the hyperplane H_j. Let

$$F_j = X \cap H_j.$$

If there is a point in common between the hyperplane H_j and the polyhedron X, without X being entirely contained in H, then H is a supporting hyperplane of X, and the set F_j is a proper face of X. But F_j is a face of X also in the cases when $X \cap H_j = \emptyset$ or $X \subseteq H_j$, due to our convention regarding non-proper faces. Of course, the faces F_j are polyhedra.

All points of a face F_j (proper as well as non-proper) are boundary points of X. Since

$$X = \bigcap_{j=1}^{m} K_j^\circ \cup \bigcup_{j=1}^{m} F_j,$$

and all points in the open set $\bigcap_{j=1}^{m} K_j^\circ$ are interior points of X, we conclude that

$$\operatorname{int} X = \bigcap_{j=1}^{m} K_j^\circ \qquad \text{and} \qquad \operatorname{bdry} X = \bigcup_{j=1}^{m} F_j.$$

The set $\operatorname{ext} X$ of extreme points of the polyhedron X is a subset of the boundary $\bigcup_{j=1}^{m} F_j$, and the extreme points are characterized by the following theorem.

Theorem 5.1.1. *A point x_0 in the polyhedron $X = \bigcap_{j=1}^{m} K_j$ is an extreme point if and only if there exists a subset I of the index set $\{1, 2, \ldots, m\}$ such that $\bigcap_{j \in I} H_j = \{x_0\}$.*

Proof. Suppose there exists such an index set I. The intersection

$$F = \bigcap_{j \in I} F_j = X \cap \bigcap_{j \in I} H_j = \{x_0\}$$

is a face of X, by Theorem 4.1.2, and x_0 is obviously an extreme point of F. Therefore, x_0 is also an extreme point of X, by Theorem 4.1.3.

Now suppose, conversely, that there is no such index set I, and let J be an index set that is maximal with respect to the property $x_0 \in \bigcap_{j \in J} H_j$. (Remember that the intersection over an empty index set is equal to the entire space $\mathbf{R}^n$, so $J = \emptyset$ if x_0 is an interior point of X.) The intersection $\bigcap_{j \in J} H_j$ is an affine subspace, which by assumption consists of more than one point and, therefore, contains a line $\{x_0 + tv \mid t \in \mathbf{R}\}$ through x_0. The line is obviously also contained in the larger set $\bigcap_{j \in J} K_j$.

Since x_0 is an interior point of the halfspace K_j for all indices $j \notin J$, we conclude that the points $x_0 + tv$ belong to all these halfspaces for all sufficiently small values of $|t|$. Consequently, there is a number $\delta > 0$ such that the line segment $[x_0 - \delta v, x_0 + \delta v]$ lies in $X = \bigcap_{j\in J} K_j \cap \bigcap_{j\notin J} K_j$, which means that x_0 is not an extreme point. □

The condition $\bigcap_{j\in I} H_j = \{x_0\}$ means that the corresponding system of linear equations

$$\langle c_j, x\rangle = b_j, \quad j \in I,$$

in n unknowns has a unique solution. A necessary condition for this to be true is that the index set I contains at least n elements. And if the system has a unique solution and there are more than n equations, then it is always possible to obtain a quadratic subsystem with a uniqe solution by eliminating suitably selected equations.

Hence, the condition $m \geq n$ is necessary for the polyhedron $X = \bigcap_{j=1}^m K_j$ to have at least one extreme point. (This also follows from Theorem 2.7.7, for if $m < n$, then

$$\dim \operatorname{lin} X = \dim\{x \in \mathbf{R}^n \mid Cx = 0\} = n - \operatorname{rank} C \geq n - m > 0,$$

which means that X is not line-free.)

Theorem 5.1.1 gives us the following method for finding all extreme points of the polyhedron X when $m \geq n$:

Solve for each subset J of $\{1, 2, \ldots, m\}$ with n elements the corresponding linear system $\langle c_j, x\rangle = b_j$, $j \in J$. If the system has a unique solution x_0, and the solution lies in X, i.e. satisfies the remaining linear inequalities $\langle c_j, x\rangle \geq b_j$, then x_0 is an extreme point of X.

The number of extreme points of X is therefore bounded by $\binom{m}{n}$, which is the number of subsets J of $\{1, 2, \ldots, m\}$ with n elements. In particular, we have proved the following theorem.

Theorem 5.1.2. *Polyhedra have finitely many extreme points.*

Polyhedral cones and extreme rays

A polyhedral cone in $\mathbf{R}^n$ is an intersection $X = \bigcap_{j=1}^m K_j$ of conic halfspaces K_j which are bounded by hyperplanes H_j through the origin, and the faces $F_j = X \cap H_j$ are polyhedral cones. Our next theorem is a direct analogue of Theorem 5.1.1.

Theorem 5.1.3. *A point x_0 in the polyhedral cone X generates an extreme ray $R = \overrightarrow{x_0}$ of the cone if and only if $-x_0 \notin X$ and there exists a subset I of the index set $\{1, 2, \ldots, m\}$ such that $\bigcap_{j\in I} H_j = \{tx_0 \mid t \in \mathbf{R}\}$.*

Proof. Suppose there exists such an index set I and that $-x_0$ does not belong to the cone X. Then

$$\bigcap_{j\in I} F_j = X \cap \bigcap_{j\in I} H_j = R.$$

By Theorem 4.1.2, this means that R is a face of the cone X. The ray R is an extreme ray of the face R, of course, so it follows from Theorem 4.1.3 that R is an extreme ray of X.

If $-x_0$ belongs to X, then X is not a proper cone, and hence X has no extreme rays according to Theorem 4.2.4.

It remains to show that R is not an extreme ray in the case when $-x_0 \notin X$ and there is no index set I with the property that the intersection $\bigcap_{j\in I} H_j$ is equal to the line through 0 and x_0. So let J be a maximal index set satisfying the condition $x_0 \in \bigcap_{j\in J} H_j$. Due to our assumption, the intersection $\bigcap_{j\in J} H_j$ is then a linear subspace of dimension greater than or equal to two, and therefore it contains a vector v which is linearly independent of x_0. The vectors $x_0 + tv$ and $x_0 - tv$ both belong to $\bigcap_{j\in J} H_j$, and consequently also to $\bigcap_{j\in J} K_j$, for all real numbers t. When $|t|$ is a sufficiently small number, the two vectors also belong to the halfspaces K_j for indices $j \notin J$, because x_0 is an interior point of K_j for these indices j. Therefore, there exists a positive number δ such that the vectors $x_+ = x_0 + \delta v$ and $x_- = x_0 - \delta v$ both belong to the cone X. The two vectors x_+ and x_- are linearly independent and $x_0 = \frac{1}{2}x_+ + \frac{1}{2}x_-$, so it follows that the ray $R = \overrightarrow{x_0}$ lies between the two rays $\overrightarrow{x_+}$ and $\overrightarrow{x_-}$ in X, and R is therefore not an extreme ray. □

Thus, to find all the extreme rays of the cone

$$X = \bigcap_{j=1}^{m} \{x \in \mathbf{R}^n \mid \langle c_j, x\rangle \geq 0\}$$

we should proceed as follows. First choose an index set J consisting of $n-1$ elements from the set $\{1, 2, \ldots, m\}$. This can be done in $\binom{m}{n-1}$ different ways. Then solve the corresponding homogeneous linear system $\langle c_j, x\rangle = 0$, $j \in J$. If the solution set is one-dimensional, than pick a solution x_0. If x_0 satisfies the remaining linear inequalities and $-x_0$ does not, then $R = \overrightarrow{x_0}$ is an extreme ray. If, instead, $-x_0$ satisfies the remaining linear inequalities and x_0 does not, then $-R$ is an extreme ray. Since this is the only way to obtain extreme rays, we conclude that the number of extreme rays is bounded by the number $\binom{m}{n-1}$. In particular, we get the following corollary.

Theorem 5.1.4. *Polyhedral cones have finitely many extreme rays.*

5.2 Polyhedral cones

Theorem 5.2.1. *A cone is polyhedral if and only if it is finitely generated.*

Proof. We first show that every polyhedral cone is finitely generated.

By Theorem 4.2.6, every polyhedral cone X can be written in the form

$$X = \operatorname{lin} X + X \cap (\operatorname{lin} X)^{\perp},$$

and $X \cap (\operatorname{lin})^{\perp}$ is a line-free, i.e. proper, polyhedral cone. Let B be a set consisting of one point from each extreme ray of $X \cap (\operatorname{lin} X)^{\perp}$; then B is a finite set and

$$X \cap (\operatorname{lin} X)^{\perp} = \operatorname{con} B,$$

according to Theorems 5.1.4 and 4.2.4.

Let $e_1, e_2, \ldots, e_d$ be a basis for the linear subspace $\operatorname{lin} X$, and put $e_0 = -(e_1 + e_2 + \cdots + e_d)$. The cone $\operatorname{lin} X$ is generated as a cone by the set $A = \{e_0, e_1, \ldots, e_d\}$, i.e.

$$\operatorname{lin} X = \operatorname{con} A.$$

Summing up,

$$X = \operatorname{lin} X + X \cap (\operatorname{lin} X)^{\perp} = \operatorname{con} A + \operatorname{con} B = \operatorname{con}(A \cup B),$$

which shows that the cone X is finitely generated by the set $A \cup B$.

Next, suppose that X is a finitely generated cone so that $X = \operatorname{con} A$ for some finite set A. We start by the observation that the dual cone X^+ is polyhedral. Indeed, if $A \neq \emptyset$ then

$$X^+ = A^+ = \{x \in \mathbf{R}^n \mid \langle x, a\rangle \geq 0 \text{ for all } a \in A\} = \bigcap_{a \in A} \{x \in \mathbf{R}^n \mid \langle a, x\rangle \geq 0\}$$

is an intersection of finitely many conical halfspaces, i.e. a polyhedral cone. And if $A = \emptyset$, then $X = \{0\}$ and $X^+ = \mathbf{R}^n$.

The already proven part of the theorem now implies that the dual cone X^+ is finitely generated. But the dual cone of X^+, i.e. the bidual cone X^{++}, is then polyhedral, too. Since the bidual cone X^{++} coincides with the original cone X, by Corollary 3.2.4, we conclude the X is a polyhedral cone. □

We are now able two prove two results that were left unproven in Chapter 2.6; compare Corollary 2.6.9.

Theorem 5.2.2. *(i) The intersection $X \cap Y$ of two finitely generated cones X and Y is a finitely generated cone.*

(ii) The inverse image $T^{-1}(X)$ of a finitely generated cone X under a linear map T is a finitely generated cone.

Proof. The intersection of two conical polyhedra is obviously a conical polyhedron, and the same holds for the inverse image of a conical polyhedron under a linear map. The theorem is therefore a corollary of Theorem 5.2.1. □

5.3 The internal structure of polyhedra

Polyhedra are by definition intersections of finite collections of closed halfspaces, and this can be viewed as an external description of polyhedra. We shall now give an internal description of polyhedra in terms of extreme points and extreme rays, and the following structure theorem is the main result of this chapter.

Theorem 5.3.1. *A nonempty subset X of $\mathbf{R}^n$ is a polyhedron if and only if there exist two finite subsets A and B of $\mathbf{R}^n$ with $A \neq \emptyset$ such that*

$$X = \operatorname{cvx} A + \operatorname{con} B.$$

The cone $\operatorname{con} B$ *is then equal to the recession cone* $\operatorname{recc} X$ *of* X*. If the polyhedron is line-free, we may choose for A the set* $\operatorname{ext} X$ *of all extreme points of X, and for B a set consisting of one nonzero point from each extreme ray of the recession cone* $\operatorname{recc} X$*.*

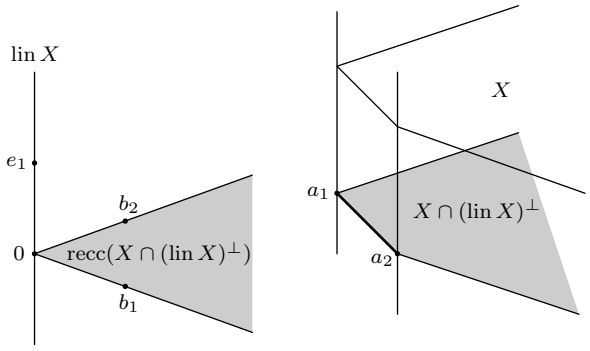

Figure 5.1. An illustration for Theorem 5.3.1. The right part of the figure depicts an unbounded polyhedron X in $\mathbf{R}^3$. Its recessive subspace $\operatorname{lin} X$ is one-dimensional and is generated as a cone by e_1 and -e_1. The intersection $X \cap (\operatorname{lin} X)^\perp$, which is shadowed, is a line-free polyhedron with two extreme points a_1 and a_2. The recession cone $\operatorname{recc}(X \cap (\operatorname{lin} X)^\perp)$ is generated by b_1 and b_2. The representation $X = \operatorname{cvx} A + \operatorname{con} B$ is obtained by taking $A = \{a_1, a_2\}$ and $B = \{e_1, -e_1, b_1, b_2\}$.

Proof. We first prove that polyhedra have the stated decomposition. So let X be a polyhedron and put $Y = X \cap (\operatorname{lin} X)^{\perp}$. Then, Y is a line-free polyhedron, and

$$X = \operatorname{lin} X + Y = \operatorname{lin} X + \operatorname{recc} Y + \operatorname{cvx}(\operatorname{ext} Y),$$

by Theorems 4.2.6 and 4.2.2. The two polyhedral cones $\operatorname{lin} X$ and $\operatorname{recc} Y$ are, according to Theorem 5.2.1, generated by two finite sets B_1 and B_2, respectively, and their sum is generated by the finite set $B = B_1 \cup B_2$. The set $\operatorname{ext} Y$ is finite, by Theorem 5.1.2, so the representation

$$X = \operatorname{cvx} A + \operatorname{con} B$$

is now obtained by taking $A = \operatorname{ext} Y$.

The cone $\operatorname{con} B$ is closed and the convex set $\operatorname{cvx} A$ is compact, since the sets A and B are finite. Hence, $\operatorname{con} B = \operatorname{recc} X$ by Corollary 2.7.13.

If X is a line-free polyhedron, then

$$X = \operatorname{cvx}(\operatorname{ext} X) + \operatorname{con}(\operatorname{exr}(\operatorname{recc} X)),$$

by Theorems 4.2.2 and 4.2.4, and this gives us the required representation of X with $A = \operatorname{ext} X$ and with B as a set consisting of one nonzero point from each extreme ray of $\operatorname{recc} X$.

To prove the converse, suppose that $X = \operatorname{cvx} A + \operatorname{con} B$, where $A = \{a_1, \ldots, a_p\}$ and $B = \{b_1, \ldots, b_q\}$ are finite sets. Consider the cone Y in $\mathbf{R}^n \times \mathbf{R}$ that is generated by the finite set $(A \times \{1\}) \cup (B \times \{0\})$. The cone Y is polyhedral according to Theorem 5.2.1, which means that there is an $m \times (n+1)$-matrix C such that

$$(x, x_{n+1}) \in Y \;\Leftrightarrow\; C \begin{bmatrix} x \\ x_{n+1} \end{bmatrix} \geq 0. \tag{5.1}$$

(Here $\begin{bmatrix} x \\ x_{n+1} \end{bmatrix}$ denotes the vector $(x_1, \ldots, x_n, x_{n+1})$ written as a column matrix.)

Let C' denote the submatrix of C which consists of all columns but the last, and let c' be the last column of the matrix C. Then

$$C \begin{bmatrix} x \\ x_{n+1} \end{bmatrix} = C'x + x_{n+1}c',$$

which means that the equivalence (5.1) may be written as

$$(x, x_{n+1}) \in Y \;\Leftrightarrow\; C'x + x_{n+1}c' \geq 0.$$

By definition, a vector $(x, 1) \in \mathbf{R}^n \times \mathbf{R}$ belongs to the cone Y if and only if there exist nonnegative numbers $\lambda_1, \lambda_2, \ldots, \lambda_p$ and $\mu_1, \mu_2, \ldots, \mu_q$ such that

$$\begin{cases} x = \lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_p a_p + \mu_1 b_1 + \mu_2 b_2 + \cdots + \mu_q b_q \\ 1 = \lambda_1 + \lambda_2 + \cdots + \lambda_p \end{cases}$$

i.e. if and only if $x \in \operatorname{cvx} A + \operatorname{con} B$. This yields the equivalences

$$x \in X \;\Leftrightarrow\; (x, 1) \in Y \;\Leftrightarrow\; C'x + c' \geq 0,$$

which means that $X = \{x \in \mathbf{R}^n \mid C'x \geq -c'\}$. Thus, X is a polyhedron. □

5.4 Polyhedron preserving operations

Theorem 5.4.1. *The intersection of finitely many polyhedra in $\mathbf{R}^n$ is a polyhedron.*

Proof. Trivial. □

Theorem 5.4.2. *Suppose X is a polyhedron in $\mathbf{R}^n$ and that $T\colon \mathbf{R}^n \to \mathbf{R}^m$ is an affine map. The image $T(X)$ is then a polyhedron in $\mathbf{R}^m$.*

Proof. The assertion is trivial if the polyhedron is empty, so suppose it is nonempty and write it in the form

$$X = \operatorname{cvx} A + \operatorname{con} B,$$

where $A = \{a_1, \ldots, a_p\}$ and $B = \{b_1, \ldots, b_q\}$ are finite sets. Each $x \in X$ has then a representation of the form

$$x = \sum_{j=1}^{p} \lambda_j a_j + \sum_{j=1}^{q} \mu_j b_j = \sum_{j=1}^{p} \lambda_j a_i + \sum_{j=1}^{q} \mu_j b_j - (\sum_{j=1}^{q} \mu_j)0$$

with nonnegative coefficients λ_j and μ_j and $\sum_{j=1}^{p} \lambda_j = 1$, i.e. each $x \in X$ is an affine combination of elements in the set $A \cup B \cup \{0\}$. Since T is an affine map,

$$Tx = \sum_{j=1}^{p} \lambda_j Ta_j + \sum_{j=1}^{q} \mu_j Tb_j - (\sum_{j=1}^{q} \mu_j)T0 = \sum_{j=1}^{p} \lambda_j Ta_j + \sum_{j=1}^{q} \mu_j (Tb_j - T0).$$

This shows that the image $T(X)$ is of the form

$$T(X) = \operatorname{cvx} A' + \operatorname{con} B'$$

with $A' = T(A)$ and $B' = -T0 + T(B) = \{Tb_1 - T0, \ldots, Tb_q - T0\}$. So the image $T(X)$ is a polyhedron, by Theorem 5.3.1. □

Theorem 5.4.3. *Suppose Y is a polyhedron in $\mathbf{R}^m$ and that $T\colon \mathbf{R}^n \to \mathbf{R}^m$ is an affine map. The inverse image $T^{-1}(Y)$ is then a polyhedron in $\mathbf{R}^n$.*

Proof. First assume that Y is a closed halfspace in $\mathbf{R}^m$ (or the entire space $\mathbf{R}^m$), i.e. that $Y = \{y \in \mathbf{R}^m \mid \langle c, y\rangle \geq b\}$. (The case $Y = \mathbf{R}^m$ is obtained by $c = 0$ and $b = 0$.) The affine map T can be written in the form $Tx = Sx + y_0$, with S as a linear map and y_0 as a vector in $\mathbf{R}^m$. This gives us

$$T^{-1}(Y) = \{x \in \mathbf{R}^n \mid \langle c, Tx\rangle \geq b\} = \{x \in \mathbf{R}^n \mid \langle S^T c, x\rangle \geq b - \langle c, y_0\rangle\}.$$

So $T^{-1}(Y)$ is a closed halfspace in $\mathbf{R}^n$ if $S^T c \neq 0$, the entire space $\mathbf{R}^n$ if $S^T c = 0$ and $b \leq \langle c, y_0\rangle$, and the empty set $\emptyset$ if $S^T c = 0$ and $b > \langle c, y_0\rangle$.

In the general case, $Y = \bigcap_{j=1}^p K_j$ is an intersection of finitely many closed halfspaces. Since $S^{-1}(Y) = \bigcap_{j=1}^p S^{-1}(K_j)$, the inverse image $S^{-1}(Y)$ is an intersection of closed halfspaces, the empty set, or the entire space $\mathbf{R}^n$. Thus, $S^{-1}(Y)$ is a polyhedron. □

Theorem 5.4.4. *The Cartesian product $X \times Y$ of two polyhedra X and Y is a polyhedron.*

Proof. Suppose X lies in $\mathbf{R}^m$ and Y lies in $\mathbf{R}^n$. The set $X \times \mathbf{R}^n$ is a polyhedron since it is the inverse image of X under the projection $(x, y) \mapsto x$, and $\mathbf{R}^m \times Y$ is a polyhedron for a similar reason. It follows that $X \times Y$ is a polyhedron, because $X \times Y = (X \times \mathbf{R}^n) \cap (\mathbf{R}^m \times Y)$. □

Theorem 5.4.5. *The sum $X + Y$ of two polyhedra in $\mathbf{R}^n$ is a polyhedron.*

Proof. The sum $X + Y$ is equal to the image of $X \times Y$ under the linear map $(x, y) \to x + y$, so the theorem is a consequence of the previous theorem and Theorem 5.4.2. □

5.5 Separation

It is possible to obtain sharper separation results for polyhedra than for general convex sets. Compare the following two theorems with Theorems 3.1.6 and 3.1.5.

Theorem 5.5.1. *If X and Y are two disjoint polyhedra, then there exists a hyperplane that strictly separates the two polyhedra.*

Proof. The difference $X - Y$ of two polyhedra X and Y is a closed set, since it is a polyhedron according to Theorem 5.4.5. So it follows from Theorem 3.1.6 that there exists a hyperplane that strictly separates the two polyhedra, if they are disjoint. □

Theorem 5.5.2. *Let X be a convex set, and let Y be a polyhedron that is disjoint from X. Then there exists a hyperplane that separates X and Y and does not contain X as a subset.*

Proof. We prove the theorem by induction over the dimension n of the surrounding space $\mathbf{R}^n$.

The case $n = 1$ is trivial, so suppose the assertion of the theorem is true when the dimension is $n - 1$, and let X be a convex subset of $\mathbf{R}^n$ that is disjoint from the polyhedron Y. An application of Theorem 3.1.5 gives us a hyperplane H that separates X and Y and, as a consequence, does not contain both sets as subsets. If X is not contained in H, then we are done. So suppose that X is a subset of H. The polyhedron Y then lies in one of the two closed halfspaces defined by the hyperplane H. Let us denote this closed halfspace by H_+, so that $Y \subseteq H_+$, and let H_{++} denote the corresponding open halfspace.

If $Y \subseteq H_{++}$, then Y and H are disjoint polyhedra, and an application of Theorem 5.5.1 gives us a hyperplane that strictly separates Y and H. Of course, this hyperplane also strictly separates Y and X, since X is a subset of H.

This proves the case $Y \subseteq H_{++}$, so it only remains to consider the case when Y is a subset of the closed halfspace H_+ without being a subset of the corresponding open halfspace, i.e. the case

$$Y \subseteq H_+, \; Y \cap H \neq \emptyset.$$

Due to our induction hypothesis, it is possible to separate the nonempty polyhedron $Y_1 = Y \cap H$ and X inside the $(n-1)$-dimensional hyperplane H using an affine $(n-2)$-dimensional subset L of H which does not contain X as a subset.

L divides the hyperplane H into two closed halves L_+ and L_- with L as their common relative boundary, and with X as a subset of L_- and Y_1 as a subset of L_+. Let us denote the relative interior of L_- by L_{--}, so that $L_{--} = L_- \setminus L$. The assumption that X is not a subset of L implies that $X \cap L_{--} \neq \emptyset$.

Observe that $Y \cap L_- = Y_1 \cap L$. If $Y_1 \cap L = \emptyset$, then there exists a hyperplane that strictly separates the polyhedra Y och L_-, by Theorem 5.5.1, and since $X \subseteq L_-$, we are done in this case, too.

What remains is to treat the case $Y_1 \cap L \neq \emptyset$, and by performing a translation, if necessary, we may assume that the origin lies in $Y_1 \cap L$, which implies that L is a linear subspace. See figure 5.2.

Note that the set $H_{++} \cup L_+$ is a cone and that Y is a subset of this cone.

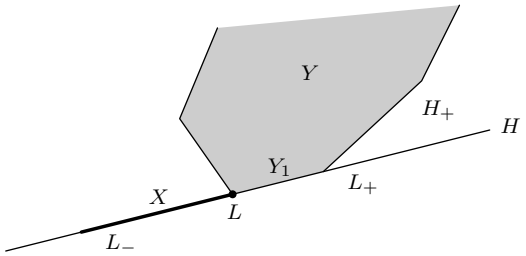

Figure 5.2. Illustration for the proof of Theorem 5.5.2.

Now, consider the cone $\operatorname{con} Y$ generated by the polyhedron Y, and let

$$C = L + \operatorname{con} Y.$$

C is a cone, too, and a subset of the cone $H_{++} \cup L_+$, since Y and L are both subsets of the last mentioned cone. The cone $\operatorname{con} Y$ is polyhedral, because if the polyhedron Y is written as $Y = \operatorname{cvx} A + \operatorname{con} B$ with finite sets A and B, then $\operatorname{con} Y = \operatorname{con}(A \cup B)$ due to the fact that 0 lies in Y. Since the sum of two polyhedral cones is polyhedral, it follows that the cone C is also polyhedral.

The cone C is disjoint from the set L_{--}, since the sets L_{--} and $H_{++} \cup L_+$ are disjoint.

Now write the polyhedral cone C as an intersection $\bigcap K_i$ of finitely many closed halfspaces K_i which are bounded by hyperplanes H_i through the origin. Each halfspace K_i is a cone containing Y as well as L. If a given halfspace K_i contains in addition a point from L_{--}, then it contains the cone generated by that point and L, that is all of L_-. Therefore, since $C = \bigcap K_i$ and $C \cap L_{--} = \emptyset$, we conclude that there exists a halfspace K_i that does not contain any point from L_{--}. In other words, the corresponding boundary hyperplane H_i separates L_- and the cone C and is disjoint from L_{--}. Since $X \subseteq L_-$, $Y \subseteq C$ and $X \cap L_{--} \neq \emptyset$, H_i separates the sets X and Y and does not contain X. This completes the induction step and the proof of the theorem. □

Exercises

5.1 Find the extreme points of the following polyhedra X:

a) $X = \{x \in \mathbf{R}^2 \mid -x_1 + x_2 \leq 2,\ x_1 + 2x_2 \geq 2,\ x_2 \geq -1\}$
b) $X = \{x \in \mathbf{R}^2 \mid -x_1 + x_2 \leq 2,\ x_1 + 2x_2 \leq 2,\ x_2 \geq -1\}$
c) $X = \{x \in \mathbf{R}^3 \mid 2x_1 + x_2 + x_3 \leq 4,\ x_1 + 2x_2 + x_3 \leq 4,\ x \geq 0\}$
d) $X = \{x \in \mathbf{R}^4 \mid x_1 + x_2 + 3x_3 + x_4 \leq 4,\ 2x_2 + 3x_3 \geq 5,\ x \geq 0\}$.

5.2 Find the extreme rays of the cone

$$X = \{x \in \mathbf{R}^3 \mid x_1 - x_2 + 2x_3 \geq 0,\ x_1 + 2x_2 - 2x_3 \geq 0,\ x_2 + x_3 \geq 0,\ x_3 \geq 0\}.$$

5.3 Find a matrix C such that

$$\operatorname{con}\{(1,-1,1),(-1,0,1),(3,2,1),(-2,-1,0)\} = \{x \in \mathbf{R}^3 \mid Cx \geq 0\}.$$

5.4 Find finite sets A and B such that $X = \operatorname{con} A + \operatorname{cvx} B$ for the following polyhedra:

a) $X = \{x \in \mathbf{R}^2 \mid -x_1 + x_2 \leq 2,\ x_1 + 2x_2 \geq 2,\ x_2 \geq -1\}$
b) $X = \{x \in \mathbf{R}^2 \mid -x_1 + x_2 \leq 2,\ x_1 + 2x_2 \leq 2,\ x_2 \geq -1\}$
c) $X = \{x \in \mathbf{R}^3 \mid 2x_1 + x_2 + x_3 \leq 4,\ x_1 + 2x_2 + x_3 \leq 4,\ x \geq 0\}$
d) $X = \{x \in \mathbf{R}^4 \mid x_1 + x_2 + 3x_3 + x_4 \leq 4,\ 2x_2 + 3x_3 \geq 5,\ x \geq 0\}$.

5.5 Suppose 0 lies in the polyhedron $X = \operatorname{cvx} A + \operatorname{con} B$, where A and B are finite sets. Prove that $\operatorname{con} X = \operatorname{con}(A \cup B)$.

Chapter 6

Convex functions

6.1 Basic definitions

Epigraph and sublevel set

Definition. Let $f\colon X \to \overline{\mathbf{R}}$ be a function with domain $X \subseteq \mathbf{R}^n$ and codomain $\overline{\mathbf{R}}$, i.e. the real numbers extended with ∞. The set

$$\operatorname{epi} f = \{(x,t) \in X \times \mathbf{R} \mid f(x) \leq t\}$$

is called the *epigraph* of the function.

Let α be a real number. The set

$$\operatorname{sublev}_\alpha f = \{x \in X \mid f(x) \leq \alpha\}$$

is called a *sublevel set* of the function, or more precisely, the *α-sublevel set*.

The epigraph is a subset of $\mathbf{R}^{n+1}$, and the word 'epi' means above. So epigraph means above the graph.

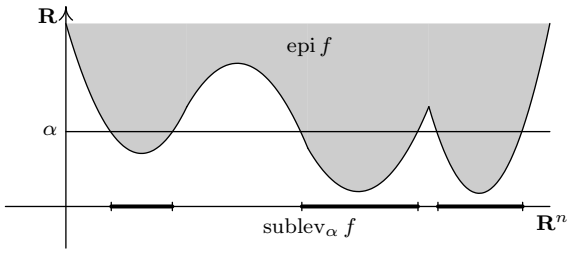

Figure 6.1. Epigraph and a sublevel set

We remind the reader of the notation $\operatorname{dom} f$ for the effective domain of f, i.e. the set of points where the function $f\colon X \to \overline{\mathbf{R}}$ is finite. Obviously,

$$\operatorname{dom} f = \{x \in X \mid f(x) < \infty\}$$

is equal to the union of all the sublevel sets of f, and these form an increasing family of sets, i.e.

$$\operatorname{dom} f = \bigcup_{\alpha \in \mathbf{R}} \operatorname{sublev}_\alpha f \qquad \text{and} \qquad \alpha < \beta \;\Rightarrow\; \operatorname{sublev}_\alpha f \subseteq \operatorname{sublev}_\beta f.$$

This implies that $\operatorname{dom} f$ is a convex set if all the sublevel sets are convex.

Convex functions

Definition. A function $f\colon X \to \overline{\mathbf{R}}$ is called *convex* if its domain X and epigraph $\operatorname{epi} f$ are convex sets.

A function $f\colon X \to \underline{\mathbf{R}}$ is called *concave* if the function $-f$ is convex.

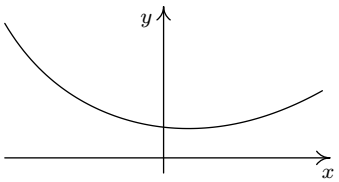

Figure 6.2. The graph of a convex function

Example 6.1.1. The epigraph of an affine function is a closed halfspace. All affine functions, and in particular all linear functions, are thus convex and concave. □

Example 6.1.2. The exponential funcion e^x with $\mathbf{R}$ as domain of definition is a convex function.

To see this, we replace x with $x-a$ in the elementary inequality $\mathrm{e}^x \geq x+1$ and obtain the inequality $\mathrm{e}^x \geq (x-a)\mathrm{e}^a + \mathrm{e}^a$, which implies that the epigraph of the exponential function can be expressed as the intersection

$$\bigcap_{a \in \mathbf{R}} \{(x,y) \in \mathbf{R}^2 \mid y \geq (x-a)\mathrm{e}^a + \mathrm{e}^a\}$$

of a family of closed halfspaces in $\mathbf{R}^2$. The epigraph is thus convex. □

Theorem 6.1.1. *The effective domain* $\operatorname{dom} f$ *and the sublevel sets* $\operatorname{sublev}_\alpha f$ *of a convex function* $f\colon X \to \overline{\mathbf{R}}$ *are convex sets.*

Proof. Suppose that the domain X is a subset of $\mathbf{R}^n$ and consider the projection $P_1\colon \mathbf{R}^n \times \mathbf{R} \to \mathbf{R}^n$ of $\mathbf{R}^n \times \mathbf{R}$ onto its first factor, i.e. $P_1(x,t) = x$. Let furthermore K_α denote the closed halfspace $\{x \in \mathbf{R}^{n+1} \mid x_{n+1} \le \alpha\}$. Then $\operatorname{sublev}_\alpha f = P_1(\operatorname{epi} f \cap K_\alpha)$, for

$$\begin{aligned} f(x) \le \alpha &\Leftrightarrow \exists t\colon f(x) \le t \le \alpha \Leftrightarrow \exists t\colon (x,t) \in \operatorname{epi} f \cap K_\alpha \\ &\Leftrightarrow x \in P_1(\operatorname{epi} f \cap K_\alpha). \end{aligned}$$

The intersections $\operatorname{epi} f \cap K_\alpha$ are convex sets, and since convexity is preserved by linear maps, it follows that the sublevel sets $\operatorname{sublev}_\alpha f$ are convex. Consequently, their union $\operatorname{dom} f$ is also convex. □

Quasiconvex functions

Many important properties of convex functions are consequences of the mere fact that their sublevel sets are convex. This is the reason for paying special attention to functions with convex sublevel sets and motivates the following definition.

Definition. A function $f\colon X \to \overline{\mathbf{R}}$ is called *quasiconvex* if X and all its sublevel sets $\operatorname{sublev}_\alpha f$ are convex.

A function $f\colon X \to \underline{\mathbf{R}}$ is called *quasiconcave* if $-f$ is quasiconvex.

Convex functions are quasiconvex since their sublevel sets are convex. The converse is not true, because a function f that is defined on some subinterval I of $\mathbf{R}$ is quasiconvex if it is increasing on I, or if it is decreasing on I, or more generally, if there exists a point $c \in I$ such that f is decreasing to the left of c and increasing to the right of c. There are, of course, non-convex functions of this type.

Convex extensions

The effective domain $\operatorname{dom} f$ of a convex (quasiconvex) function $f\colon X \to \overline{\mathbf{R}}$ is convex, and since

$$\begin{aligned} \operatorname{epi} f &= \{(x,t) \in \operatorname{dom} f \times \mathbf{R} \mid f(x) \le t\} \qquad \text{and} \\ \operatorname{sublev}_\alpha f &= \{x \in \operatorname{dom} f \mid f(x) \le \alpha\}, \end{aligned}$$

the restriction $f|_{\operatorname{dom} f}$ of f to $\operatorname{dom} f$ is also a convex (quasiconvex) function, and the restriction has the same epigraph and the same α-sublevel sets as f.

So what is the point of allowing ∞ as a function value of a convex function? We are of course primarily interested in functions with finite values but functions with infinite values arise naturally as suprema or limits of sequences of functions with finite values.

Another benefit of allowing ∞ as a function value of (quasi)convex functions is that we can without restriction assume that they are defined on the entire space $\mathbf{R}^n$. For if $f\colon X \to \overline{\mathbf{R}}$ is a (quasi)convex function defined on a proper subset X of $\mathbf{R}^n$, and if we define the function $\tilde{f}\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ by

$$\tilde{f}(x) = \begin{cases} f(x) & \text{if } x \in X \\ \infty & \text{if } x \notin X, \end{cases}$$

then f and $\tilde{f}$ have the same epigraphs and the same α-sublevel sets. The extension $\tilde{f}$ is therefore also (quasi)convex. Of course, $\operatorname{dom} \tilde{f} = \operatorname{dom} f$.

(Quasi)concave functions have an analogous extension to functions with values in $\underline{\mathbf{R}} = \mathbf{R} \cup \{-\infty\}$.

Alternative characterization of convexity

Theorem 6.1.2. *A function $f\colon X \to \overline{\mathbf{R}}$ with a convex domain of definition X is*

(a) convex if and only if

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) \tag{6.1}$$

for all points $x, y \in X$ and all numbers $\lambda \in]0,1[$;

(b) quasiconvex if and only if

$$f(\lambda x + (1-\lambda)y) \le \max\{f(x), f(y)\} \tag{6.2}$$

for all points $x, y \in X$ and all numbers $\lambda \in]0,1[$.

Proof. (a) Suppose f is convex, i.e. that the epigraph $\operatorname{epi} f$ is convex, and let x and y be two points in $\operatorname{dom} f$. Then the points $(x, f(x))$ and $(y, f(y))$ belong to the epigraph, and the convexity of the epigraph implies that the convex combination

$$\big(\lambda x + (1-\lambda)y, \lambda f(x) + (1-\lambda) f(y)\big)$$

of these two points also belong to the epigraph. This statement is equivalent to the inequality (6.1) being true. If any of the points $x, y \in X$ lies outside $\operatorname{dom} f$, then the inequality is trivially satisfied since the right hand side is equal to ∞ in that case.

To prove the converse, we assume that the inequality (6.1) holds. Let (x, s) and (y, t) be two points in the epigraph, and let $0 < \lambda < 1$. Then $f(x) \leq s$ and $f(y) \leq t$, by definition, and it therefore follows from the inequality (6.1) that

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \leq \lambda s + (1-\lambda)t,$$

so the point $(\lambda x + (1-\lambda)y, \lambda s + (1-\lambda)t)$, i.e. the point $\lambda(x, s) + (1-\lambda)(y, t)$, lies in the epigraph. In orther words, the epigraph is convex.

(b) The proof is analogous and is left to the reader. □

A function $f : X \to \overline{\mathbf{R}}$ is clearly (quasi)convex if and only if the restriction $f|_L$ is (quasi)convex for each line L that intersects X. Each such line has an equation of the form $x = x_0 + tv$, where x_0 is a point in X and v is a vector in $\mathbf{R}^n$, and the corresponding restriction is a one-variable function $g(t) = f(x_0 + tv)$ (with the set $\{t \mid x_0 + tv \in X\}$ as its domain of definition). To decide whether a function is (quasi)convex or not is thus essentially a one-variable problem.

Definition. Let $f\colon X \to \overline{\mathbf{R}}$ be a function defined on a convex cone X. The function is called

- *subadditive* if $f(x + y) \leq f(x) + f(y)$ for all $x, y \in X$;
- *positive homogeneous* if $f(\alpha x) = \alpha f(x)$ for all $x \in X$ and all $\alpha \in \mathbf{R}_+$.

Every positive homogeneous, subadditive function is clearly convex. Conversely, every convex, positive homogeneous function f is subadditive, because

$$f(x + y) = 2f(\tfrac{1}{2}x + \tfrac{1}{2}y) \leq 2(\tfrac{1}{2}f(x) + \tfrac{1}{2}f(y)) = f(x) + f(y).$$

A *seminorm* on $\mathbf{R}^n$ is a function $f\colon \mathbf{R}^n \to \mathbf{R}$, which is subadditive, positive homogeneous, and *symmetric*, i.e. satisfies the condition

$$f(-x) = f(x) \quad \text{for all } x \in \mathbf{R}^n.$$

The symmetry and homogenouity conditions may of course be merged to the condition

$$f(\alpha x) = |\alpha| f(x) \text{ for all } x \in \mathbf{R}^n \text{ and all } \alpha \in \mathbf{R}.$$

If f is a seminorm, then $f(x) \geq 0$ for all x, since

$$0 = f(0) = f(x - x) \leq f(x) + f(-x) = 2f(x).$$

A seminorm f is called a *norm* if $f(x) = 0$ implies $x = 0$. The usual notation for a norm is $\|\cdot\|$.

Seminorms, and in particular norms, are convex functions.

EXAMPLE 6.1.3. The Euclidean norm and the ℓ^1-norm, that were defined in Chapter 1, are special cases of the ℓ^p-norms $\|\cdot\|_p$ on $\mathbf{R}^n$. They are defined for $1 \le p < \infty$ by

$$\|x\|_p = \Big(\sum_{i=1}^{n} |x_i|^p\Big)^{1/p},$$

and for $p = \infty$ by

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

The maximum norm $\|\cdot\|_\infty$ is a limiting case, because $\|x\|_p \to \|x\|_\infty$ as $p \to \infty$.

The ℓ^p-norms are obviously positive homogeneous and symmetric and equal to 0 only if $x = 0$. Subadditivity is an immediate consequence of the triangle inequality $|x + y| \le |x| + |y|$ for real numbers when $p = 1$ or $p = \infty$, and of the Cauchy–Schwarz inequality when $p = 2$. For the remaining values of p, subadditivity will be proved in Section 6.4 (Theorem 6.4.3). □

Strict convexity

By strengthening the inequalities in the alternative characterization of convexity, we obtain the following definitions.

Definition. A convex function $f\colon X \to \overline{\mathbf{R}}$ is called *strictly convex* if

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y)$$

for all pairs of distinct points x, $y \in X$ and all $\lambda \in]0,1[$.

A quasiconvex function f is called *strictly quasiconvex* if inequality (6.2) is strict for all pairs of distinct points x, $y \in X$ and all $\lambda \in]0,1[$.

A function f is called *strictly concave* (*strictly quasiconcave*) if the function $-f$ is strictly convex (strictly quasiconvex).

EXAMPLE 6.1.4. A quadratic form $q(x) = \langle x, Qx\rangle = \sum_{i,j=1}^{n} q_{ij} x_i x_j$ on $\mathbf{R}^n$ is convex if and only if it is positive semidefinite, and the form is strictly convex if and only if it is positive definite. This follows from the identity

$$\big(\lambda x_i + (1-\lambda)y_i\big)\big(\lambda x_j + (1-\lambda)y_j\big) = \lambda x_i x_j + (1-\lambda) y_i y_j - \lambda(1-\lambda)(x_i - y_i)(x_j - y_j)$$

which after multiplication by q_{ij} and summation yields the equality

$$q(\lambda x + (1-\lambda)y) = \lambda q(x) + (1-\lambda) q(y) - \lambda(1-\lambda) q(x-y).$$

The right hand side is $\le \lambda q(x) + (1-\lambda) q(y)$ for all $0 < \lambda < 1$ if and only if $q(x-y) \ge 0$, which holds for all $x \ne y$ if and only if q is positive semidefinite. Strict inequality requires q to be positive definite. □

Jensen's inequality

The inequalities (6.1) and (6.2) are easily extended to convex combinations of more than two points.

Theorem 6.1.3. *Let f be a function and suppose $x = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_m x_m$ is a convex combination of the points $x_1, x_2, \ldots, x_m$ in the domain of f.*
(a) If f is convex, then

$$f(x) \le \sum_{j=1}^{m} \lambda_j f(x_j). \qquad \textbf{(Jensen's inequality)} \tag{6.3}$$

If f is strictly convex and $\lambda_j > 0$ for all j, then equality prevails in (6.3) *if and only if $x_1 = x_2 = \cdots = x_m$.*

(b) If f is quasiconvex, then

$$f(x) \le \max_{1 \le j \le m} f(x_j). \tag{6.4}$$

If f is strictly quasiconvex and $\lambda_j > 0$ for all j, then equality prevails in (6.4) *if and only if $x_1 = x_2 = \cdots = x_m$.*

Proof. (a) To prove the Jensen inequality we may assume that all coefficients λ_j are positive and that all points x_j lie in $\operatorname{dom} f$, because the right hand side of the inequality is infinite if some point x_j lies outside $\operatorname{dom} f$. Then

$$\big(x, \sum_{j=1}^{m} \lambda_j f(x_j)\big) = \sum_{j=1}^{m} \lambda_j \big(x_j, f(x_j)\big),$$

and the right sum, being a convex combination of elements in the epigraph $\operatorname{epi} f$, belongs to $\operatorname{epi} f$. So the left hand side is a point in $\operatorname{epi} f$, and this gives us inequality (6.3).

Now assume that f is strictly convex and that we have equality in Jensen's inequality for the convex combination $x = \sum_{j=1}^{m} \lambda_j x_j$, with positive coefficents λ_j and $m \ge 2$. Let $y = \sum_{j=2}^{m} \lambda_j (1-\lambda_1)^{-1} x_j$. Then $x = \lambda_1 x_1 + (1-\lambda_1) y$, and y is a convex combination of x_2, x_3, …, x_m, so it follows from Jensen's inequality that

$$\begin{aligned} \sum_{j=1}^{m} \lambda_j f(x_j) = f(x) &\le \lambda_1 f(x_1) + (1-\lambda_1) f(y) \\ &\le \lambda_1 f(x_1) + (1-\lambda_1) \sum_{j=2}^{m} \lambda_j (1-\lambda_1)^{-1} f(x_j) = \sum_{j=1}^{m} \lambda_j f(x_j). \end{aligned}$$

Since the left hand side and right hand side of this chain of inequalities and equalities are equal, we conclude that equality holds everywhere. Thus, $f(x) = \lambda_1 f(x_1) + (1 - \lambda_1) f(y)$, and since f is strictly convex, this implies that $x_1 = y = x$.

By symmetri, we also have $x_2 = x$, ..., $x_m = x$, and hence $x_1 = x_2 = \cdots = x_m$.

(b) Suppose f is quasiconvex, and let $\alpha = \max_{1 \leq j \leq m} f(x_j)$. If any of the points x_j lies outside $\operatorname{dom} f$, then there is nothing to prove since the right hand side of the inequality (6.4) is infinite. In the opposite case, α is a finite number, and each point x_j belongs to the convex sublevel set $\operatorname{sublev}_\alpha f$, and it follows that so does the point x. This proves inequality (6.4).

The proof of the assertion about equality for strictly quasiconvex functions is analogous with the corresponding proof for strictly convex functions. □

6.2 Operations that preserve convexity

We now describe some ways to construct new convex functions from given convex functions.

Conic combination

Theorem 6.2.1. *Suppose that $f\colon X \to \overline{\mathbf{R}}$ and $g\colon X \to \overline{\mathbf{R}}$ are convex functions and that α and β are nonnegative real numbers. Then $\alpha f + \beta g$ is also a convex function.*

Proof. Follows directly from the alternative characterization of convexity in Theorem 6.1.2. □

The set of convex functions on a given set X is, in other words, a convex cone. So every conic combination $\alpha_1 f_1 + \alpha_2 f_2 + \cdots + \alpha_m f_m$ of convex functions on X is convex.

Note that there is no counterpart of this statement for quasiconvex functions − a sum of quasiconvex functions is not necessarily quasiconvex.

Pointwise limit

Theorem 6.2.2. *Suppose that the functions $f_i\colon X \to \overline{\mathbf{R}}$, $i = 1, 2, 3, \ldots$, are convex and that the limit*

$$f(x) = \lim_{i \to \infty} f_i(x)$$

exists as a finite number or ∞ for each $x \in X$. The limit function $f\colon X \to \overline{\mathbf{R}}$ is then also convex.

Proof. Let x and y be two points in X, and suppose $0 < \lambda < 1$. By passing to the limit in the inequality $f_i(\lambda x + (1-\lambda)y) \le \lambda f_i(x) + (1-\lambda)f_i(y)$ we obtain the following inequality

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y),$$

which tells us that the limit function f is convex. □

Using Theorem 6.2.2, we may extend the result in Theorem 6.2.1 to infinite sums and integrals. For example, a pointwise convergent infinite sum $f(x) = \sum_{i=1}^{\infty} f_i(x)$ of convex functions is convex.

And if $f(x, y)$ is a function that is convex with respect to the variable x on some set X for each y in a set Y, α is a nonnegative function defined on Y, and the integral $g(x) = \int_Y \alpha(y) f(x, y)\, dy$ exists for all $x \in X$, then g is a convex function on X. This follows from Theorem 6.2.2 by writing the integral as a limit of Riemann sums, or more directly, by integrating the inequalites that characterize the convexity of the functions $f(\cdot, y)$.

Composition with affine maps

Theorem 6.2.3. *Suppose $A\colon V \to \mathbf{R}^n$ is an affine map, that Y is a convex subset of $\mathbf{R}^n$, and that $f\colon Y \to \overline{\mathbf{R}}$ is a convex function. The composition $f \circ A$ is then a convex function on its domain of definition $A^{-1}(Y)$*

Proof. Let $g = f \circ A$. Then, for $x_1, x_2 \in A^{-1}(Y)$ and $0 < \lambda < 1$,

$$\begin{aligned} g(\lambda x_1 + (1-\lambda)x_2) &= f(\lambda A x_1 + (1-\lambda)A x_2) \le \lambda f(Ax_1) + (1-\lambda) f(Ax_2) \\ &= \lambda g(x_1) + (1-\lambda) g(x_2), \end{aligned}$$

which shows that the function g is convex. □

The composition $f \circ A$ of a quasiconvex function f and an affine map A is quasiconvex.

Example 6.2.1. The function $x \mapsto \mathrm{e}^{c_1 x_1 + \cdots + c_n x_n}$ is convex on $\mathbf{R}^n$ since it is a composition of a linear map and the convex exponential function $t \mapsto \mathrm{e}^t$. □

Pointwise supremum

Theorem 6.2.4. *Let $f_i\colon X \to \overline{\mathbf{R}}$, $i \in I$, be a family of functions, and define the function $f\colon X \to \overline{\mathbf{R}}$ for $x \in X$ by*

$$f(x) = \sup_{i \in I} f_i(x).$$

Then

(i) f is convex if the functions f_i are all convex;
(ii) f is quasiconvex if the functions f_i are all quasiconvex.

Proof. By the least upper bound definition, $f(x) \le t$ if and only if $f_i(x) \le t$ for all $i \in I$, and this implies that

$$\operatorname{epi} f = \bigcap_{i\in I} \operatorname{epi} f_i \quad \text{and} \quad \operatorname{sublev}_t f = \bigcap_{i\in I} \operatorname{sublev}_t f_i$$

for all $t \in \mathbf{R}$. The assertions of the theorem are now immediate consequences of the fact that intersections of convex sets are convex. □

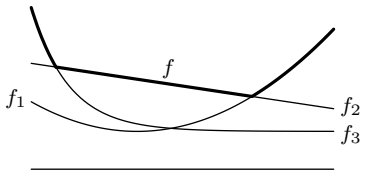

Figure 6.3. $f = \sup f_i$ for a family consisting of three functions.

Example 6.2.2. A pointwise maximum of finitely many affine functions, i.e. a function of the form

$$f(x) = \max_{1\le i\le m} (\langle c_i, x\rangle + a_i),$$

is a convex function and is called a convex *piecewise affine* function. □

Example 6.2.3. Examples of convex piecewise affine functions f on $\mathbf{R}^n$ are:

(a) The absolute value of the i:th coordinate of a vector
$$f(x) = |x_i| = \max\{x_i, -x_i\}.$$
(b) The maximum norm
$$f(x) = \|x\|_\infty = \max_{1\le i\le n} |x_i|.$$
(c) The sum of the m largest coordinates of a vector
$$f(x) = \max\{x_{i_1} + \cdots + x_{i_m} \mid 1 \le i_1 < i_2 < \cdots < i_m \le n\}.$$ □

Composition

Theorem 6.2.5. *Suppose that the function $\phi\colon I \to \overline{\mathbf{R}}$ is defined on a real intervall I that contains the range $f(X)$ of the function $f\colon X \to \mathbf{R}$. The composition $\phi \circ f\colon X \to \overline{\mathbf{R}}$ is convex*

(i) if f is convex and ϕ is convex and increasing;

(ii) if f is concave and ϕ is convex and decreasing.

Proof. The inequality

$$\phi\big(f(\lambda x + (1-\lambda)y)\big) \le \phi\big(\lambda f(x) + (1-\lambda)f(y)\big)$$

holds for $x, y \in X$ and $0 < \lambda < 1$ if either f is convex and ϕ is increasing, or f is concave and ϕ is decreasing. If in addition ϕ is convex, then

$$\phi\big(\lambda f(x) + (1-\lambda)f(y)\big) \le \lambda\phi(f(x)) + (1-\lambda)\phi(f(y)),$$

and by combining the two inequalities above, we obtain the inequality that shows that the function $\phi \circ f$ is a convex. □

There is a corresponding result for quasiconvexity: The composition $\phi \circ f$ is quasiconvex if either f is quasiconvex and ϕ is increasing, or f is quasiconcave and ϕ is decreasing.

Example 6.2.4. The function

$$x \mapsto \mathrm{e}^{x_1^2 + x_2^2 + \cdots + x_k^2},$$

where $1 \le k \le n$, is convex on $\mathbf{R}^n$, since the exponential function is convex and increasing, and positive semidefinite quadratic forms are convex. □

Example 6.2.5. The two functions $t \mapsto 1/t$ and $t \mapsto -\ln t$ are convex and decreasing on the interval $]0, \infty[$. So the function $1/g$ is convex and the function $\ln g$ is concave, if g is a concave and positive function. □

Infimum

Theorem 6.2.6. *Let C be a convex subset of $\mathbf{R}^{n+1}$, and let g be the function defined for $x \in \mathbf{R}^n$ by*

$$g(x) = \inf\{t \in \mathbf{R} \mid (x,t) \in C\},$$

with the usual convention $\inf \emptyset = +\infty$. Suppose there exists a point x_0 in the relative interior of the set

$$X_0 = \{x \in \mathbf{R}^n \mid g(x) < \infty\} = \{x \in \mathbf{R}^n \mid \exists t \in \mathbf{R}\colon (x,t) \in C\}$$

with a finite function value $g(x_0)$. Then $g(x) > -\infty$ for all $x \in \mathbf{R}^n$, and $g\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ is a convex function with X_0 as its effective domain.

Proof. Let x be an arbitrary point in X_0. To show that $g(x) > -\infty$, i.e. that the set

$$T_x = \{t \in \mathbf{R} \mid (x,t) \in C\}$$

is bounded below, we first choose a point $x_1 \in \operatorname{rint} X_0$ such that x_0 lies on the open line segment $]x, x_1[$, and write $x_0 = \lambda x + (1-\lambda)x_1$ with $0 < \lambda < 1$. We then fix a real number t_1 such that $(x_1, t_1) \in C$, and for $t \in T_x$ define the number t_0 as $t_0 = \lambda t + (1-\lambda)t_1$. The pair (x_0, t_0) is then a convex combination of the points (x,t) and (x_1,t_1) in C, so

$$g(x_0) \le t_0 = \lambda t + (1-\lambda)t_1,$$

by convexity and the definition of g. We conclude that

$$t \ge \frac{1}{\lambda}\big(g(x_0) - (1-\lambda)t_1\big),$$

and this inequality shows that the set T_x is bounded below.

So the function g has $\overline{\mathbf{R}}$ as codomain, and $\operatorname{dom} g = X_0$. Now, let x_1 and x_2 be arbitrary points in X_0, and let λ_1 and λ_2 be two positive numbers with sum 1. To each $\epsilon > 0$ there exist two real numbers t_1 and t_2 such that the two points (x_1,t_1) and (x_2,t_2) lie in C and $t_1 < g(x_1) + \epsilon$ and $t_2 < g(x_2) + \epsilon$. The convex combination $(\lambda_1 x_1 + \lambda_2 x_2, \lambda_1 t_1 + \lambda_2 t_2)$ of the two points lies in C, too, and

$$g(\lambda_1 x_1 + \lambda_2 x_2) \le \lambda_1 t_1 + \lambda_2 t_2 \le \lambda_1 g(x_1) + \lambda_2 g(x_2) + \epsilon.$$

This means that the point $\lambda_1 x_1 + \lambda_2 x_2$ lies in X_0, and by letting ϵ tend to 0 we conclude that $g(\lambda_1 x_1 + \lambda_2 x_2) \le \lambda_1 g(x_1) + \lambda_2 g(x_2)$. Hence, the set X_0 is convex, and the function g is convex. □

We have seen that the pointwise supremum $f(x) = \sup_{i\in I} f_i(x)$ of an arbitrary family of convex functions is convex. So if $f\colon X \times Y \to \overline{\mathbf{R}}$ is a function with the property that the functions $f(\cdot, y)$ are convex on X for each $y \in Y$, and we define the function g on X by $g(x) = \sup_{y\in Y} f(x,y)$, then g is convex, and this is true without any further conditions on the set Y. Our next theorem shows that the corresponding infimum is a convex function, provided f is convex as a function on the product set $X \times Y$.

Theorem 6.2.7. *Suppose $f\colon X \times Y \to \mathbf{R}$ is a convex function, and for each $x \in X$ define*

$$g(x) = \inf_{y\in Y} f(x,y).$$

If there is a point $x_0 \in \operatorname{rint} X$ such that $g(x_0) > -\infty$, then $g(x)$ is a finite number for each $x \in X$, and $g\colon X \to \mathbf{R}$ is a convex function.

Proof. Suppose X is a subset of $\mathbf{R}^n$ and let

$$C = \{(x,t) \in X \times \mathbf{R} \mid \exists y \in Y\colon f(x,y) \le t\}.$$

C is a convex subset of $\mathbf{R}^{n+1}$, because given two points (x_1,t_1) and (x_2,t_2) in C, and two positive numbers λ_1 and λ_2 with sum 1, there exist two points y_1 and y_2 in the convex set Y such that $f(x_i,y_i) \le t_i$ for $i = 1,2$, and this implies that

$$f(\lambda_1 x_1 + \lambda_2 x_2, \lambda_1 y_1 + \lambda_2 y_2) \le \lambda_1 f(x_1,y_1) + \lambda_2 f(x_2,y_2) \le \lambda_1 t_1 + \lambda_2 t_2,$$

which shows that the convex combination $\lambda_1(x_1,t_1) + \lambda_2(x_2,t_2)$ lies in C. Moreover, $g(x) = \inf\{t \mid (x,t) \in C\}$, so the corollary follows immediately from Theorem 6.2.6. □

Perspective

Definition. Let $f\colon X \to \overline{\mathbf{R}}$ be a function defined on a cone X in $\mathbf{R}^n$. The function $g\colon X \times \mathbf{R}_{++} \to \overline{\mathbf{R}}$, defined by

$$g(x,s) = sf(x/s),$$

is called the *perspective* of f.

Theorem 6.2.8. *The perspective g of a convex function $f\colon X \to \overline{\mathbf{R}}$ with a convex cone X as domain is a convex function.*

Proof. Let (x,s) and (y,t) be two points in $X \times \mathbf{R}_{++}$, and let α, β be two positive numbers with sum 1. Then

$$\begin{aligned} g\big(\alpha(x,s) + \beta(y,t)\big) &= g(\alpha x + \beta y, \alpha s + \beta t) = (\alpha s + \beta t) f\Big(\frac{\alpha x + \beta y}{\alpha s + \beta t}\Big) \\ &= (\alpha s + \beta t) f\Big(\frac{\alpha s}{\alpha s + \beta t} \cdot \frac{x}{s} + \frac{\beta t}{\alpha s + \beta t} \cdot \frac{y}{t}\Big) \\ &\le \alpha s f\Big(\frac{x}{s}\Big) + \beta t f\Big(\frac{y}{t}\Big) = \alpha g(x,s) + \beta g(y,t). \end{aligned}$$ □

Example 6.2.6. By the previous theorem, $f(x) = x_n q(x/x_n)$ is a convex function on $\mathbf{R}^{n-1} \times \mathbf{R}_{++}$ whenever $q(x)$ is a positive semidefinite quadratic form on $\mathbf{R}^{n-1}$. In particular, by choosing the Euclidean norm as quadratic form, we see that the function

$$x \mapsto (x_1^2 + x_2^2 + \cdots + x_{n-1}^2)/x_n$$

is convex on the open halfspace $\mathbf{R}^{n-1} \times \mathbf{R}_{++}$. □

6.3 Maximum and minimum

Minimum points

For an arbitrary function to decide whether a given point is a global minimum point is an intractable problem, but there are good numerical methods for finding local minimum points if we impose some regularity conditions on the function. This is the reason why convexity plays such an important role in optimization theory. A local minimum of a convex function is namely automatically a global minimum.

Let us recall that a point $x_0 \in X$ is a *local minimum point* of the function $f\colon X \to \overline{\mathbf{R}}$ if there exists an open ball $B = B(x_0; r)$ with center at x_0 such that $f(x) \geq f(x_0)$ for all $x \in X \cap B$. The point is a *(global) minimum point* if $f(x) \geq f(x_0)$ for all $x \in X$.

Theorem 6.3.1. *Suppose that the function $f\colon X \to \overline{\mathbf{R}}$ is convex and that $x_0 \in \operatorname{dom} f$ is a local minimum point of f. Then x_0 is a global minimum point. The minimum point is unique if f is strictly convex.*

Proof. Let $x \in X$ be an arbitrary point different from x_0. Since f is a convex function and $\lambda x + (1-\lambda)x_0 \to x_0$ as $\lambda \to 0$, the following inequalities hold for $\lambda > 0$ sufficiently close to 0:

$$f(x_0) \leq f(\lambda x + (1-\lambda)x_0) \leq \lambda f(x) + (1-\lambda) f(x_0)$$

(with strict inequality in the last place if f is strictly convex). From this follows at once that $f(x) \geq f(x_0)$ (and $f(x) > f(x_0)$, respectively), which proves that x_0 is a global minimum point (and that there are no other minimum points if the convexity is strict) □

Theorem 6.3.2. *The set of minimum points of a quasiconvex function is convex.*

Proof. The assertion is trivial for functions with no minimum point, since the empty set is convex, and for the function which is identically equal to ∞ on X. So, suppose that the quasiconvex function $f\colon X \to \overline{\mathbf{R}}$ has a minimum point $x_0 \in \operatorname{dom} f$. The set of minimum points is then equal to the sublevel set $\{x \in X \mid f(x) \leq f(x_0)\}$, which is convex by definition. □

Maximum points

Theorem 6.3.3. *Suppose $X = \operatorname{cvx} A$ and that the function $f\colon X \to \overline{\mathbf{R}}$ is quasiconvex. Then*

$$\sup_{x \in X} f(x) = \sup_{a \in A} f(a).$$

If the function has a maximum, then there is a maximum point in A.

Proof. Let $x \in X$. Since x is a convex combination $x = \sum_{j=1}^{m} \lambda_j a_j$ of elements $a_j \in A$,

$$f(x) = f(\sum_{j=1}^{m} \lambda_j a_j) \leq \max_{1 \leq j \leq m} f(a_j) \leq \sup_{a \in A} f(a),$$

and it follows that

$$\sup_{x \in X} f(x) \leq \sup_{a \in A} f(a).$$

The converse inequality being trivial, since A is a subset of X, we conclude that equality holds.

Moreover, if x is a maximum point, then $f(x) \geq \max_{1 \leq j \leq m} f(a_j)$, and combining this with the inequality above, we obtain $f(x) = \max_{1 \leq j \leq m} f(a_j)$, which means that the maximum is certainly attained at some of the points $a_j \in A$. □

Thus, we can find the maximum of a quasiconvex function whose domain is the convex hull of a finite set A, by just comparing finitely many function values. Of course, this may be infeasible if the set A is very large.

Since compact convex sets coincide with the convex hull of their extreme points, we have the following corollary of the previous theorem.

Corollary 6.3.4. *Suppose that X is a compact convex set and that $f \colon X \to \overline{\mathbf{R}}$ is a quasiconvex function. If f has a maximum, then there is a maximum point among the extreme points of X.*

Example 6.3.1. The quadratic form $f(x_1, x_2) = x_1^2 + 2x_1x_2 + 2x_2^2$ is strictly convex, since it positive definite. The maximum of f on the traingle with vertices at the points $(1, 1)$, $(-2, 1)$ and $(0, 2)$ is attained at some of the vertices. The function values at the vertices are 5, 2, and 8, respectively. The maximum value is hence equal to 8, and it is attained at $(0, 2)$. □

A non-constant realvalued convex function can not attain its maximum at an interior point of its domain, because of the following theorem.

Theorem 6.3.5. *A convex function $f \colon X \to \mathbf{R}$ that attains its maximum at a relative interior point of X, is necessarily constant on X.*

Proof. Suppose f has a maximum at the point $a \in \operatorname{rint} X$, and let x be an arbitrary point in X. Since a is a relative interior point, there exists a point $y \in X$ such that a belongs to the open line segment $]x, y[$, i.e. $a = \lambda x + (1-\lambda)y$ for some number λ satisfying $0 < \lambda < 1$. By convexity and since $f(y) \leq f(a)$,

$$f(a) = f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \leq \lambda f(x) + (1-\lambda) f(a),$$

with $f(x) \geq f(a)$ as conclusion. Since the converse inequality holds trivially, we have $f(x) = f(a)$. The function f is thus equal to $f(a)$ everywhere. □

6.4 Some important inequalities

Many inequalities can be proved by convexity arguments, and we shall give three important examples.

Arithmetic and geometric mean

Definition. Let $\theta_1, \theta_2, \ldots, \theta_n$ be given positive numbers with $\sum_{j=1}^{n} \theta_j = 1$. The *weighted arithmetic mean* A and the *weighted geometric mean* G of n positive numbers $a_1, a_2, \ldots, a_n$ with the given numbers $\theta_1, \theta_2, \ldots, \theta_n$ as *weights* are defined as

$$A = \sum_{j=1}^{n} \theta_j a_j \qquad \text{and} \qquad G = \prod_{j=1}^{n} a_j^{\theta_j}.$$

The usual arithmetic and geometric means are obtained as special cases by taking all weights equal to $1/n$.

We have the following well-known inequality between the arithmetic and the geometric means.

Theorem 6.4.1. *For all positive numbers* $a_1, a_2, \ldots, a_n$

$$G \leq A$$

with equality if and only if $a_1 = a_2 = \cdots = a_n$.

Proof. Let $x_j = \ln a_j$, so that $a_j = \mathrm{e}^{x_j} = \exp(x_j)$. The inequality $G \leq A$ is now transformed to the inequality

$$\exp\Big(\sum_{j=1}^{n} \theta_j x_j\Big) \leq \sum_{j=1}^{n} \theta_j \exp(x_j),$$

which is Jensen's inequality for the strictly convex exponential function, and equality holds if and only if $x_1 = x_2 = \cdots = x_n$, i.e. if and only if $a_1 = a_2 = \cdots = a_n$. □

Example 6.4.1. A lot of maximum and minimum problems can be solved by use of the inequality of arithmetic and geometric means. Here follows a general example.

Let f be a function of the form

$$f(x) = \sum_{i=1}^{m} c_i \Big(\prod_{j=1}^{n} x_j^{\alpha_{ij}} \Big), \qquad x \in \mathbf{R}^n$$

where $c_i > 0$ and α_{ij} are real numbers for all i, j.

The function $g(x) = 16x_1 + 2x_2 + x_1^{-1}x_2^{-2}$, corresponding to $n = 2$, $m = 3$, $c = (16, 2, 1)$ and

$$\alpha = [\alpha_{ij}] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -2 \end{bmatrix},$$

serves as a typical example of such a function.

Suppose that we want to minimize $f(x)$ over the set $\{x \in \mathbf{R}^n \mid x > 0\}$. This problem can be attacked in the following way. Let $\theta_1, \theta_2, \ldots, \theta_m$ be positive numbers with sum equal to 1, and write

$$f(x) = \sum_{i=1}^{m} \theta_i \Big(\frac{c_i}{\theta_i} \prod_{j=1}^{n} x_j^{\alpha_{ij}} \Big).$$

The inequality of arithmetic and geometric means now gives us the following inequality

$$f(x) \geq \prod_{i=1}^{m} \Big(\Big(\frac{c_i}{\theta_i} \Big)^{\theta_i} \Big(\prod_{j=1}^{n} x_j^{\theta_i \alpha_{ij}} \Big) \Big) = C(\theta) \cdot \prod_{j=1}^{n} x_j^{\beta_j}, \tag{6.5}$$

with

$$C(\theta) = \prod_{i=1}^{m} \Big(\frac{c_i}{\theta_i} \Big)^{\theta_i} \qquad \text{and} \qquad \beta_j = \sum_{i=1}^{m} \theta_i \alpha_{ij}.$$

If it is possible to choose the weights $\theta_i > 0$ so that $\sum_{i=1}^{m} \theta_i = 1$ and

$$\beta_j = \sum_{i=1}^{m} \theta_i \alpha_{ij} = 0 \quad \text{for all } j,$$

then inequality (6.5) becomes

$$f(x) \geq C(\theta),$$

and equality occurs if and only if all the products $\dfrac{c_i}{\theta_i} \displaystyle\prod_{j=1}^{n} x_j^{\alpha_{ij}}$ are equal, a condition that makes it possible to determine x. □

Hölder's inequality

Theorem 6.4.2 (Hölder's inequallity). *Suppose $1 \le p \le \infty$ and let q be the dual index defined by the equality $1/p + 1/q = 1$. Then*

$$|\langle x, y\rangle| = \Big|\sum_{j=1}^{n} x_j y_j\Big| \le \|x\|_p \|y\|_q$$

for all x, $y \in \mathbf{R}^n$. Moreover, to each x there corresponds a y with norm $\|y\|_q = 1$ such that $\langle x, y\rangle = \|x\|_p$.

Remark. Observe that $q = 2$ when $p = 2$. Thus, the Cauchy–Schwarz inequality is a special case of Hölder's inequality.

Proof. The case $p = \infty$ follows directly from the triangle inequality for sums:

$$\Big|\sum_{j=1}^{n} x_j y_j\Big| \le \sum_{j=1}^{n} |x_j||y_j| \le \sum_{j=1}^{n} \|x\|_\infty |y_j| = \|x\|_\infty \|y\|_1.$$

So assume that $1 \le p < \infty$. Since $\big|\sum_1^n x_j y_j\big| \le \sum_1^n |x_j||y_j|$, and the vector $(|x_1|, \dots, |x_n|)$ has the same ℓ^p-norm as $(x_1, \dots, x_n)$ and the vector $(|y_1|, \dots, |y_n|)$ has the same ℓ^q-norm as $(y_1, \dots, y_n)$, we can without loss of generality assume that the numbers x_j and y_j are positive.

The function $t \mapsto t^p$ is convex on the interval $[0, \infty[$. Hence,

$$\Big(\sum_{j=1}^{n} \lambda_j t_j\Big)^p \le \sum_{j=1}^{n} \lambda_j t_j^p. \tag{6.6}$$

for all positive numbers $t_1, t_2, \dots, t_n$ and all positive numbers $\lambda_1, \lambda_2, \dots, \lambda_n$ with $\sum_1^n \lambda_j = 1$. Now, let us make the particular choice

$$\lambda_j = \frac{y_j^q}{\sum_{j=1}^n y_j^q} \qquad \text{and} \qquad t_j = \frac{x_j y_j}{\lambda_j}.$$

Then

$$\lambda_j t_j = x_j y_j \qquad \text{and} \qquad \lambda_j t_j^p = \frac{x_j^p y_j^p}{y_j^{(p-1)q}} \Big(\sum_{j=1}^{n} y_j^q\Big)^{p-1} = x_j^p \Big(\sum_{j=1}^{n} y_j^q\Big)^{p-1},$$

which inserted in the inequality (6.6) gives

$$\Big(\sum_{j=1}^{n} x_j y_j\Big)^p \le \sum_{j=1}^{p} x_j^p \Big(\sum_{j=1}^{n} y_j^q\Big)^{p-1},$$

and we obtain Hölder's inequality by raising both sides to $1/p$.

It is easy to verify that Hölder's inequality holds with equality and that $\|y\|_q = 1$ if we choose y as follows:

$x = 0$: All y with norm equal to 1.

$x \neq 0,\ 1 \leq p < \infty$: $y_j = \begin{cases} \|x\|_p^{-p/q}|x_j|^p/x_j & \text{if } x_j \neq 0, \\ 0 & \text{if } x_j = 0. \end{cases}$

$x \neq 0,\ p = \infty$: $y_j = \begin{cases} |x_j|/x_j & \text{if } j = j_0, \\ 0 & \text{if } j \neq j_0, \end{cases}$

where j_0 is an index such that $|x_{j_0}| = \|x\|_\infty$. □

Theorem 6.4.3 (Minkowski's inequality)**.** *Suppose $p \geq 1$ and let x and y be arbitrary vectors in $\mathbf{R}^n$. Then*

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p.$$

Proof. Consider the linear forms $x \mapsto f_a(x) = \langle a, x\rangle$ for vectors $a \in \mathbf{R}^n$ satisfying $\|a\|_q = 1$. By Hölder's inequality,

$$f_a(x) \leq \|a\|_q\|x\|_p \leq \|x\|_p,$$

and for each x there exists a vector a with $\|a\|_q = 1$ such that Hölder's inequality holds with equality, i.e. such that $f_a(x) = \|x\|_p$. Thus

$$\|x\|_p = \sup\{f_a(x) \mid \|a\|_q = 1\},$$

and hence, $f(x) = \|x\|_p$ is a convex function by Theorem 6.2.4. Positive homogenouity is obvious, and positive homogeneous convex functions are subadditive, so the proof of Minkowski's inequality is now complete. □

6.5 Solvability of systems of convex inequalities

The solvability of systems of linear inequalities was discussed in Chapter 3. Our next theorem is kind of a generalization of Theorem 3.3.7 and treats the solvability of a system of convex and affine inequalities.

Theorem 6.5.1. *Let $f_i\colon \Omega \to \mathbf{R}$, $i = 1, 2, \ldots, m$, be a family of convex functions defined on a convex subset Ω of $\mathbf{R}^n$.*

Let p be an integer in the interval $1 \leq p \leq m$, and suppose if $p < m$ that the functions f_i are restrictions to Ω of affine *functions for $i \geq p + 1$ and that the set*

$$\{x \in \operatorname{rint}\Omega \mid f_i(x) \leq 0 \text{ for } i = p + 1, \ldots, m\}$$

is nonempty. The following two assertions are then equivalent:

(i) The system

$$\begin{cases} f_i(x) < 0, & i = 1, 2, \ldots, p \\ f_i(x) \leq 0, & i = p+1, \ldots, m \end{cases}$$

has no solution $x \in \Omega$.

(ii) There exist nonnegative numbers $\lambda_1, \lambda_2, \ldots, \lambda_m$, with at least one of the numbers $\lambda_1, \lambda_2, \ldots, \lambda_p$ being nonzero, such that

$$\sum_{i=1}^{m} \lambda_i f_i(x) \geq 0$$

for all $x \in \Omega$.

Remark. The system of inequalities must contain at least one strict inequality, and all inequalities are allowed to be strict (the case $p = m$).

Proof. If the system (i) has a solution x, then the sum in (ii) is obviously negative for the same x, since at least one of its terms is negative and the others are non-positive. Thus, (ii) implies (i).

To prove the converse implication, we assume that the system (i) has no solution and define M to be the set of all $y = (y_1, y_2, \ldots, y_m) \in \mathbf{R}^m$ such that the system

$$\begin{cases} f_i(x) < y_i, & i = 1, 2, \ldots, p \\ f_i(x) = y_i, & i = p+1, \ldots, m \end{cases}$$

has a solution $x \in \Omega$.

The set M is convex, for if y' and y'' are two points in M, $0 \leq \lambda \leq 1$, and x', x'' are solutions in Ω of the said systems of inequalities and equalities with y' and y'', respectively, as right hand members, then $x = \lambda x' + (1-\lambda)x'' \in \Omega$ will be a solution of the system with $\lambda y' + (1-\lambda)y''$ as its right hand member, due to the convexity and affinity of the functions f_i for $i \leq p$ and $i > p$, respectively.

Our assumptions concerning the system (i) imply that $M \cap \mathbf{R}^m_- = \emptyset$. Since $\mathbf{R}^m_-$ is a polyhedron, there exist, by the separation theorem 5.5.2, a nonzero vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ and a real number α such that the hyperplane $H = \{y \mid \langle \lambda, y \rangle = \alpha\}$ separates M and $\mathbf{R}^m_-$ and does not contain M as subset. We may assume that

$$\lambda_1 y_1 + \lambda_2 y_2 + \cdots + \lambda_m y_m \begin{cases} \geq \alpha & \text{for all } y \in M, \\ \leq \alpha & \text{for all } y \in \mathbf{R}^m_-. \end{cases}$$

By first choosing $y = 0$, we see that $\alpha \geq 0$, and by then choosing $y = t\mathbf{e}_i$, where $\mathbf{e}_i$ is the i:th standard basis vector in $\mathbf{R}^m$, and letting t tend to $-\infty$, we conclude that $\lambda_i \geq 0$ for all i.

For each $x \in \Omega$ and $\epsilon > 0$,

$$y = (f_1(x) + \epsilon, \ldots, f_p(x) + \epsilon, f_{p+1}(x), \ldots, f_m(x))$$

is a point in M. Consequently,

$$\lambda_1(f_1(x) + \epsilon) + \cdots + \lambda_p(f_p(x) + \epsilon) + \lambda_{p+1}f_{p+1}(x) + \cdots + \lambda_m f_m(x) \geq \alpha \geq 0,$$

and by letting ϵ tend to zero, we obtain the inequality

$$\lambda_1 f_1(x) + \lambda_2 f_2(x) + \cdots + \lambda_m f_m(x) \geq 0$$

for all $x \in \Omega$.

If $p = m$, we are done since the vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ is then nonzero, but it remains to prove that some of the coefficients $\lambda_1, \lambda_2, \ldots, \lambda_p$ is nonzero when $p < m$. Assume the contrary, i.e. that $\lambda_1 = \lambda_2 = \cdots = \lambda_p = 0$, and let

$$h(x) = \sum_{i=p+1}^{m} \lambda_i f_i(x).$$

The function h is affine, and $h(x) \geq 0$ for all $x \in \Omega$. Furthermore, by the assumptions of the theorem, there exists a point x_0 in the relative interior of Ω such that $f_i(x_0) \leq 0$ for all $i \geq p + 1$, which implies that $h(x_0) \leq 0$. Thus, $h(x_0) = 0$. This means that the restriction $h|_\Omega$, which is a concave function since h is affine, attains its minimum at a relative interior point, and according to Theorem 6.3.5 (applied to the function $-h|_\Omega$), this implies that the function h is constant and equal to 0 on Ω.

But to each $y \in M$ there corresponds a point $x \in \Omega$ such that $y_i = f_i(x)$ for $i = p + 1, \ldots m$, and this implies that

$$\langle \lambda, y \rangle = \textstyle\sum_{i=p+1}^{m} \lambda_i f_i(x) = h(x) = 0.$$

We conclude that $\alpha = 0$ and that the hyperplane H contains M, which is a contradiction. Thus, at least one of the coefficients $\lambda_1, \lambda_2, \ldots, \lambda_p$ has to be nonzero, and the theorem is proved. □

6.6 Continuity

A real-valued convex function is automatically continuous at all relative interior points of the domain. More precisely, we have the following theorem.

Theorem 6.6.1. *Suppose $f\colon X \to \overline{\mathbf{R}}$ is a convex function and that a is a point in the relative interior of* $\operatorname{dom} f$*. Then there exist a relative open neighborhood U of a in* $\operatorname{dom} f$ *and a constant M such that*

$$|f(x) - f(a)| \leq M\|x - a\|$$

for all $x \in U$. Hence, f is continuous on the relative interior of $\operatorname{dom} f$*.*

Proof. We start by proving a special case of the theorem and then show how to reduce the general case to this special case.

1. So first assume that X is an open subset of $\mathbf{R}^n$, that $\operatorname{dom} f = X$, i.e. that f is a real-valued convex function, that $a = 0$, and that $f(0) = 0$.

We will show that if we choose the number $r > 0$ such that the closed hypercube

$$K(r) = \{x \in \mathbf{R}^n \mid \|x\|_\infty \leq r\}$$

is included in X, then there is a constant M such that

$$|f(x)| \leq M\|x\| \tag{6.7}$$

for all x in the closed ball $\overline{B}(0;r) = \{x \in \mathbf{R}^n \mid \|x\| \leq r\}$, where $\|\cdot\|$ is the usual Euclidean norm.

The hypercube $K(r)$ has 2^n extreme points (vertices). Let L denote the largest of the function values of f at these extreme points. Since the convex hull of the extreme points is equal to $K(r)$, it follows from Theorem 6.3.3 that

$$f(x) \leq L$$

for all $x \in K(r)$, and thereby also for all $x \in \overline{B}(0;r)$, because $\overline{B}(0;r)$ is a subset of $K(r)$.

We will now make this inequality sharper. To this end, let x be an arbitrary point in $\overline{B}(0;r)$ different from the center 0. The halfline from 0 through x intersects the boundary of $\overline{B}(0;r)$ at the point

$$y = \frac{r}{\|x\|}x,$$

and since x lies on the line segment $[0, y]$, x is a convex combination of its end points. More precisely, $x = \lambda y + (1-\lambda)0$ with $\lambda = \|x\|/r$. Therefore, since f is convex,

$$f(x) \leq \lambda f(y) + (1-\lambda)f(0) = \lambda f(y) \leq \lambda L = \frac{L}{r}\|x\|.$$

The above inequality holds for all $x \in \overline{B}(0;r)$. To prove the same inequality with $f(x)$ replaced by $|f(x)|$, we use the fact that the point $-x$ belongs to $\overline{B}(0;r)$ if x does so, and the equality $0 = \frac{1}{2}x + \frac{1}{2}(-x)$. By convexity,

$$0 = f(0) \leq \frac{1}{2}f(x) + \frac{1}{2}f(-x) \leq \frac{1}{2}f(x) + \frac{L}{2r}\|x\|,$$

which simplifies to the inequality

$$f(x) \geq -\frac{L}{r}\|-x\| = -\frac{L}{r}\|x\|.$$

This proves that inequality (6.7) holds for $x \in \overline{B}(0;r)$ with $M = L/r$.

2. We now turn to the general case. Let n be the dimension of the set $\operatorname{dom} f$. The affine hull of $\operatorname{dom} f$ is equal to the set $a + V$ for some n-dimensional linear subspace V, and since V is isomorphic to $\mathbf{R}^n$, we can obtain a bijective linear map $T\colon \mathbf{R}^n \to V$ by choosing a coordinate system in V.

The inverse image Y of the relative interior of $\operatorname{dom} f$ under the map $y \mapsto a + Ty$ of $\mathbf{R}^n$ onto $\operatorname{aff}(\operatorname{dom} f)$ is an open convex subset of $\mathbf{R}^n$, and Y contains the point 0. Define the function $g\colon Y \to \mathbf{R}$ by

$$g(y) = f(a + Ty) - f(a).$$

Then, g is a convex function, since g is composed by a convex function and an affine function, and $g(0) = 0$.

For $x = a + Ty \in \operatorname{rint}(\operatorname{dom} f)$ we now have $f(x) - f(a) = g(y)$ and $x - a = Ty$, so in order to prove the general case of our theorem, we have to show that there is a constant M such that $|g(y)| \leq M\|Ty\|$ for all y in some neighborhood of 0. But the map $y \mapsto \|Ty\|$ is a norm on $\mathbf{R}^n$, and since all norms are equivalent, it suffices to show that there is a constant M such that

$$|g(y)| \leq M\|y\|$$

for all y in some neighborhood of 0, and that is exactly what we did in step 1 of the proof. So the theorem is proved. □

The following corollary follows immediately from Theorem 6.6.1, because affine sets have no relative boundary points.

Corollary 6.6.2. *A convex function $f\colon X \to \mathbf{R}$ with an affine subset X as domain is continuous.*

For functions f with a closed interval $I = [a, b]$ as domain, convexity imposes no other restrictions on the function value $f(b)$ than that it has to be greater than or equal to $\lim_{x \to b-} f(x)$. Thus, a convex function need not be continuous at the endpoint b, and a similar remark holds for the left endpoint, of course. For example, a function f, that is identically equal to zero on $I \setminus \{a, b\}$, is convex if $f(a) \geq 0$ and $f(b) \geq 0$. Cf. exercise 7.6.

6.7 The recessive subspace of convex functions

Example 6.7.1. Let $f\colon \mathbf{R}^2 \to \mathbf{R}$ be the convex function

$$f(x_1, x_2) = x_1 + x_2 + \mathrm{e}^{(x_1 - x_2)^2}.$$

The restrictions of f to lines with direction given by the vector $v = (1, 1)$ are affine functions, since

$$f(x + tv) = f(x_1 + t, x_2 + t) = x_1 + x_2 + 2t + \mathrm{e}^{(x_1 - x_2)^2} = f(x) + 2t.$$

Let $V = \{x \in \mathbf{R}^2 \mid x_1 = x_2\}$ be the linear subspace of $\mathbf{R}^2$ spanned by the vector v, and consider the the orthogonal decomposition $\mathbf{R}^2 = V^\perp + V$. Each $x \in \mathbf{R}^2$ has a corresponding unique decomposition $x = y + z$ with $y \in V^\perp$ and $z \in V$, namely

$$y = \tfrac{1}{2}(x_1 - x_2, x_2 - x_1) \text{ and } z = \tfrac{1}{2}(x_1 + x_2, x_1 + x_2).$$

Moreover, since $z = \frac{1}{2}(x_1 + x_2)v = z_1 v$,

$$f(x) = f(y + z) = f(y) + 2z_1 = f|_{V^\perp}(y) + 2z_1.$$

So there is a corresponding decomposition of f as a sum of the restriction of f to $V^\perp$ and a linear function on V. It is easy to verify that the vector $(v, 2) = (1, 1, 2)$ spans the recessive subspace $\operatorname{lin}(\operatorname{epi} f)$, and that V is equal to the image $P_1(\operatorname{lin}(\operatorname{epi} f))$ of $\operatorname{lin}(\operatorname{epi} f)$ under the projection P_1 of $\mathbf{R}^2 \times \mathbf{R}$ onto the first factor $\mathbf{R}^2$. □

The result in the previous example can be generalized, and in order to describe this generalization we need a definition.

Definition. Let $f\colon X \to \overline{\mathbf{R}}$ be a function defined on a subset X of $\mathbf{R}^n$. The linear subspace

$$V_f = P_1(\operatorname{lin}(\operatorname{epi} f)),$$

where $P_1\colon \mathbf{R}^n \times \mathbf{R} \to \mathbf{R}^n$ is the projection of $\mathbf{R}^n \times \mathbf{R}$ onto its first factor $\mathbf{R}^n$, is called the *recessive subspace* of the function f.

Theorem 6.7.1. *Let f be a convex function with recessive subspace V_f.*

(i) A vector v belongs to V_f if and only if there is a unique number α_v such that (v, α_v) belongs to the recessive subspace $\operatorname{lin}(\operatorname{epi} f)$ *of the epigraph of the function.*

(ii) The map $g\colon V_f \to \mathbf{R}$, defined by $g(v) = \alpha_v$ for $v \in V_f$, is linear.

(iii) $\operatorname{dom} f = \operatorname{dom} f + V_f$.

(iv) $f(x + v) = f(x) + g(v)$ for all $x \in \operatorname{dom} f$ and all $v \in V_f$.

(v) If the function f is differentiable at $x \in \operatorname{dom} f$ then $g(v) = \langle f'(x), v\rangle$ for all $v \in V_f$.

(vi) Suppose V is a linear subspace, that $h\colon V \to \mathbf{R}$ is a linear map, that $\operatorname{dom} f + V \subseteq \operatorname{dom} f$, *and that $f(x+v) = f(x) + h(v)$ for all $x \in \operatorname{dom} f$ and all $v \in V$. Then, $V \subseteq V_f$.*

Proof. (i) By definition, $v \in V_f$ if and only if there is a real number α_v such that $(v, \alpha_v) \in \operatorname{lin}(\operatorname{epi} f)$. To prove that the number α_v is uniquely determined

by $v \in V_f$, we assume that the pair (v, β) also lies in $\operatorname{lin}(\operatorname{epi} f)$.

The point $(x+tv, f(x)+t\alpha_v)$ belongs to the epigraph for each $x \in \operatorname{dom} f$ and each $t \in \mathbf{R}$, i.e.

$$x + tv \in \operatorname{dom} f \quad \text{and} \quad f(x+tv) \leq f(x) + t\alpha_v. \tag{6.8}$$

Hence, $(x + tv, f(x + tv))$ is a point in the epigraph, and our assumption $(v, \beta) \in \operatorname{lin}(\operatorname{epi} f)$ now implies that $(x + tv - tv, f(x + tv) - t\beta)$ is a point in i epi f, too. We conclude that

$$f(x) \leq f(x + tv) - t\beta \tag{6.9}$$

for all $t \in \mathbf{R}$. By combining the two inequalities (6.8) and (6.9), we obtain the inequality $f(x) \leq f(x) + (\alpha_v - \beta)t$, which holds for all $t \in \mathbf{R}$. This is possible only if $\beta = \alpha_v$, and proves the uniqueness of the number α_v.

(ii) Let, as before, P_1 be the projection of $\mathbf{R}^n \times \mathbf{R}$ onto $\mathbf{R}^n$, and let P_2 be the projection of $\mathbf{R}^n \times \mathbf{R}$ onto the second factor $\mathbf{R}$. The uniqueness result (i) implies that the restriction of P_1 to the linear subspace $\operatorname{lin}(\operatorname{epi} f)$ is a bijective linear map onto V_f. Let Q denote the inverse of this restriction; the map g is then equal to the composition $P_2 \circ Q$ of the two linear maps P_2 and Q, and this implies that g is a linear function.

(iii) The particular choice of $t = 1$ in (6.8) yields the implication

$$x \in \operatorname{dom} f \;\&\; v \in V_f \;\Rightarrow\; x + v \in \operatorname{dom} f,$$

which proves the inclusion $\operatorname{dom} f + V_f \subseteq \operatorname{dom} f$, and the converse inclusion is of course trivial.

(iv) By choosing $t = 1$ in the inequalities (6.8) and(6.9) and using the fact that $\alpha_v = \beta = g(v)$, we obtain the two inequalities $f(x + v) \leq f(x) + g(v)$ and $f(x) \leq f(x + v) - g(v)$, which when combined prove assertion (iv).

(v) Consider the restriction $\phi(t) = f(x + tv)$ of the function f to the line through the point x with direction $v \in V_f$. By (iii), ϕ is defined for all $t \in \mathbf{R}$, and by (iv), $\phi(t) = f(x) + tg(v)$. Hence, $\phi'(0) = g(v)$. But if f is differentiable at x, then we also have $\phi'(0) = \langle f'(x), v\rangle$ according to the chain rule, and this proves our assertion (v).

(vi) Suppose $v \in V$. If (x, s) is an arbitrary point in the epigraph epi f, then $f(x + tv) = f(x) + h(tv) \leq s + th(v)$, which means that the point $(x + tv, s + th(v))$ lies in epi f for every real number t. This proves that $(v, h(v))$ belongs to $\operatorname{lin}(\operatorname{epi} f)$ and, consequently, that v is a vector in V_f. □

By our next theorem, every convex function is the sum of a convex function with a trivial recessive subspace and a linear function.

Theorem 6.7.2. *Suppose that f is a convex function with recessive subspace V_f. Let $\tilde{f}$ denote the restriction of f to $\operatorname{dom} f \cap V_f^\perp$, and let $g\colon V_f \to \mathbf{R}$ be the linear function defined in Theorem 6.7.1. The recessive subspace $V_{\tilde{f}}$ of $\tilde{f}$ is then trivial, i.e. equal to $\{0\}$, $\operatorname{dom} f = \operatorname{dom} f \cap V_f^\perp + V_f$, and*

$$f(y+z) = \tilde{f}(y) + g(z)$$

for all $y \in \operatorname{dom} f \cap V_f^\perp$ and all $z \in V_f$.

Proof. Each $x \in \mathbf{R}^n$ has a unique decomposition $x = y + z$ with $y \in V_f^\perp$ and $z \in V_f$, and if $x \in \operatorname{dom} f$ then $y = x - z \in \operatorname{dom} f + V_f = \operatorname{dom} f$, by Theorem 6.7.1, and hence $y \in \operatorname{dom} f \cap V_f^\perp$. This proves that $\operatorname{dom} f = \operatorname{dom} f \cap V_f^\perp + V_f$.

The equality $f(y+z) = \tilde{f}(y) + g(z)$ now follows from (iv) in Theorem 6.7.1, so it only remains to prove that $V_{\tilde{f}} = \{0\}$. Suppose $v \in V_{\tilde{f}}$, and let x_0 be an arbitrary point in $\operatorname{dom} \tilde{f}$. Then $x_0 + v$ lies in $\operatorname{dom} \tilde{f}$, too, and since $\operatorname{dom} \tilde{f} \subseteq V_f^\perp$ and $V_f^\perp$ is a linear subspace, we conclude that $v = (x_0 + v) - x_0$ is a vector in $V_f^\perp$. This proves the inclusion $V_{\tilde{f}} \subseteq V_f^\perp$.

Theorem 6.7.1 gives us two linear functions $g\colon V_f \to \mathbf{R}$ and $\tilde{g}\colon V_{\tilde{f}} \to \mathbf{R}$ such that $f(x+v) = f(x) + g(v)$ for all $x \in \operatorname{dom} f$ and all $v \in V_f$, and $\tilde{f}(y+w) = \tilde{f}(y) + \tilde{g}(w)$ for all $y \in \operatorname{dom} f \cap V_f^\perp$ and all $w \in V_{\tilde{f}}$.

Now, let w be an arbitrary vector in $V_{\tilde{f}}$ and x be an arbitrary point in $\operatorname{dom} f$, and write x as $x = y + v$ with $y \in \operatorname{dom} f \cap V_f^\perp$ and $v \in V_f$. The point $y + w$ lies in $\operatorname{dom} f \cap V_f^\perp$, and we get the following identities:

$$\begin{aligned} f(x+w) &= f(y+v+w) = f(y+w+v) = f(y+w) + g(v) \\ &= \tilde{f}(y+w) + g(v) = \tilde{f}(y) + \tilde{g}(w) + g(v) \\ &= f(y) + g(v) + \tilde{g}(w) = f(x) + \tilde{g}(w). \end{aligned}$$

Therefore, $V_{\tilde{f}} \subseteq V_f$, by Theorem 6.7.1 (v). Hence, $V_{\tilde{f}} \subseteq V_f^\perp \cap V_f = \{0\}$, which proves that $V_{\tilde{f}} = \{0\}$. □

6.8 Closed convex functions

Definition. A convex function is called *closed* if it has a closed epigraph.

Theorem 6.8.1. *A convex function $f\colon X \to \overline{\mathbf{R}}$ is closed if and only if all its sublevel sets are closed.*

Proof. Suppose that X is a subset of $\mathbf{R}^n$ and that f is a closed function. Let

$$X_\alpha = \operatorname{sublev}_\alpha f = \{x \in X \mid f(x) \leq \alpha\}$$

be an arbitrary nonempty sublevel set of f, and define Y_α to be the set

$$Y_\alpha = \operatorname{epi} f \cap \{(x, x_{n+1}) \in \mathbf{R}^n \times \mathbf{R} \mid x_{n+1} \leq \alpha\}.$$

The set Y_α is closed, being the intersection between the closed epigraph $\operatorname{epi} f$ and a closed halfspace, and $X_\alpha = P(Y_\alpha)$, where $P\colon \mathbf{R}^n \times \mathbf{R} \to \mathbf{R}^n$ is the projection $P(x, x_{n+1}) = x$.

Obviously, the recession cone $\operatorname{recc} Y_\alpha$ contains no nonzero vector of the form $v = (0, v_{n+1})$, i.e. no nonzero vector in the null space $\mathcal{N}(P) = \{0\} \times \mathbf{R}$ of the projection P. Hence, $(\operatorname{recc} Y_\alpha) \cap \mathcal{N}(P) = \{0\}$, so it follows from Theorem 2.7.10 that the sublevel set X_α is closed.

To prove the converse, assume that all sublevel sets are closed, and let (x_0, y_0) be a boundary point of $\operatorname{epi} f$. Let $\big((x_k, y_k)\big)_1^\infty$ be a sequence of points in $\operatorname{epi} f$ that converges to (x_0, y_0), and let ϵ be an arbitrary positive number. Then, since $y_k \to y_0$ as $k \to \infty$, $f(x_k) \leq y_k \leq y_0 + \epsilon$ for all sufficiently large k, so the points x_k belong to the sublevel set $\{x \in X \mid f(x) \leq y_0 + \epsilon\}$ for all sufficiently large k. The sublevel set being closed, it follows that the limit point x_0 lies in the same sublevel set, i.e. $x_0 \in X$ and $f(x_0) \leq y_0 + \epsilon$, and since $\epsilon > 0$ is arbitrary, we conclude that $f(x_0) \leq y_0$. Hence, (x_0, y_0) is a point in $\operatorname{epi} f$. So $\operatorname{epi} f$ contains all its boundary points and is therefore a closed set. □

Corollary 6.8.2. *Continuous convex functions $f\colon X \to \mathbf{R}$ with closed domains X are closed functions.*

Proof. Follows immediately from Theorem 6.8.1, because the sublevel sets of real-valued continuous functions with closed domains are closed sets. □

Theorem 6.8.3. *All nonempty sublevel sets of a closed convex function have the same recession cone and the same recessive subspace. Hence, all sublevel sets are bounded if one of the nonempty sublevel sets is bounded.*

Proof. Let $f\colon X \to \overline{\mathbf{R}}$ be a closed convex function, and suppose that x_0 is a point in the sublevel set $X_\alpha = \{x \in X \mid f(x) \leq \alpha\}$. Since X_α and $\operatorname{epi} f$ are closed convex sets and (x_0, α) is a point in $\operatorname{epi} f$, we obtain the following equivalences:

$$\begin{aligned} v \in \operatorname{recc} X_\alpha &\Leftrightarrow x_0 + tv \in X_\alpha \quad \text{for all } t \in \mathbf{R}_+ \\ &\Leftrightarrow f(x_0 + tv) \leq \alpha \quad \text{for all } t \in \mathbf{R}_+ \\ &\Leftrightarrow (x_0 + tv, \alpha) \in \operatorname{epi} f \quad \text{for all } t \in \mathbf{R}_+ \\ &\Leftrightarrow (x_0, \alpha) + t(v, 0) \in \operatorname{epi} f \quad \text{for all } t \in \mathbf{R}_+ \\ &\Leftrightarrow (v, 0) \in \operatorname{recc}(\operatorname{epi} f), \end{aligned}$$

with the conclusion that the recession cone

$$\operatorname{recc} X_\alpha = \{v \in \mathbf{R}^n \mid (v, 0) \in \operatorname{recc}(\operatorname{epi} f)\}$$

does not depend on α as long as $X_\alpha \neq \emptyset$. Of course, the same is then true for the recessive subspace

$$\operatorname{lin} X_\alpha = \operatorname{recc} X_\alpha \cap (-\operatorname{recc} X_\alpha) = \{v \in \mathbf{R}^n \mid (v, 0) \in \operatorname{lin}(\operatorname{epi} f)\}.$$

The statement concerning bounded sublevel sets follows from the fact that a closed convex set is bounded if and only if its recession cone is equal to the zero cone $\{0\}$. □

Theorem 6.8.4. *A convex function f, which is bounded on an affine subset M, is constant on M.*

Proof. Let $M = a + U$, where U is a linear subspace, and consider the restriction $g = f|_M$ of f to M. The function g is continuous since all points of M are relative interior points, and closed since the domain M is a closed set. Let $\alpha = \sup\{g(x) \mid x \in M\}$; then $\{x \mid g(x) \leq \alpha\} = M$, so by the previous theorem, all nonempty sublevel sets of g has $\operatorname{lin} M$, that is the subspace U, as their recessive subspace.

Let now x_0 be an arbitrary point in M. Since the recessive subspace of the particular sublevel set $\{x \mid g(x) \leq g(x_0\}$ is equal to U, we conclude that $g(x_0 + u) \leq g(x_0)$ for all $u \in U$. Hence, $g(x) \leq g(x_0)$ for all $x \in M$, which means that x_0 is a maximum point of g. Since $x_0 \in M$ is arbitrary, all points in M are maximum points, and this implies that g is constant on M. □

6.9 The support function

Definition. Let A be a nonempty subset of $\mathbf{R}^n$. The function $S_A \colon \mathbf{R}^n \to \overline{\mathbf{R}}$, defined by

$$S_A(x) = \sup\{\langle y, x\rangle \mid y \in A\}$$

(with the usual convention that $S_A(x) = \infty$ if the function $y \mapsto \langle y, x\rangle$ is unbounded above on A) is called the *support function* of the set A.

Theorem 6.9.1. *(a) The support function S_A is a closed convex function.*

(b) Suppose A and B are nonempty subsets of $\mathbf{R}^n$, that $\alpha > 0$ and that $C\colon \mathbf{R}^n \to \mathbf{R}^m$ is a linear map. Then

(i) $$S_A = S_{\operatorname{cvx} A} = S_{\operatorname{cl}(\operatorname{cvx} A)}$$

(ii) $$S_{\alpha A} = \alpha S_A$$

(iii) $$S_{A+B} = S_A + S_B$$

(iv) $$S_{A\cup B} = \max\{S_A, S_B\}$$

(v) $$S_{C(A)} = S_A \circ C^T.$$

Proof. (a) The support function S_A is closed and convex, because its epigraph

$$\operatorname{epi} S_A = \{(x,t) \mid \langle y, x\rangle \le t \text{ for all } y \in A\} = \bigcap_{y\in A}\{(x,t) \mid \langle y, x\rangle \le t\}$$

is closed, being the intersection of a family of closed halfspaces in $\mathbf{R}^n \times \mathbf{R}$.

(b) Since linear forms are convex, it follows from Theorem 6.3.3 that

$$S_A(x) = \sup\{\langle x, y\rangle \mid y \in A\} = \sup\{\langle x, y\rangle \mid y \in \operatorname{cvx} A\} = S_{\operatorname{cvx} A}(x)$$

for all $x \in \mathbf{R}^n$. Moreover, if a function f is continuous on the closure of a set X, then $\sup_{y\in X} f(y) = \sup_{y\in \operatorname{cl} X} f(y)$, and linear forms are of course continuous. Therefore, $S_{\operatorname{cvx} A}(x) = S_{\operatorname{cl}(\operatorname{cvx} A)}(x)$ for all x.

This proves the identity (i), and the remaining identities are obtained as follows:

$$S_{\alpha A}(x) = \sup_{y\in \alpha A}\langle y, x\rangle = \sup_{y\in A}\langle \alpha y, x\rangle = \alpha \sup_{y\in A}\langle y, x\rangle = \alpha S_A(x).$$

$$\begin{aligned} S_{A+B}(x) &= \sup_{y\in A+B}\langle y, x\rangle = \sup_{y_1\in A,\, y_2\in B}\langle y_1 + y_2, x\rangle \\ &= \sup_{y_1\in A,\, y_2\in B}(\langle y_1, x\rangle + \langle y_2, x\rangle) = \sup_{y_1\in A}\langle y_1, x\rangle + \sup_{y_2\in B}\langle y_2, x\rangle \\ &= S_A(x) + S_B(x). \end{aligned}$$

$$\begin{aligned} S_{A\cup B}(x) &= \sup_{y\in (A\cup B)}\langle y, x\rangle = \max\{\sup_{y\in A}\langle y, x\rangle, \sup_{y\in B}\langle y, x\rangle\} \\ &= \max\{S_A(x), S_B(x)\}. \end{aligned}$$

$$S_{C(A)}(x) = \sup_{y\in C(A)}\langle y, x\rangle = \sup_{z\in A}\langle Cz, x\rangle = \sup_{z\in A}\langle z, C^T x\rangle = S_A(C^T x). \qquad \square$$

Example 6.9.1. The support function of a closed interval $[a,b]$ on the real line is given by

$$S_{[a,b]}(x) = S_{\{a,b\}}(x) = \max\{ax, bx\},$$

since $[a,b] = \operatorname{cvx}\{a,b\}$. $\square$

EXAMPLE 6.9.2. In order to find the support function of the closed unit ball $\overline{B}_p = \{x \in \mathbf{R}^n \mid \|x\|_p \leq 1\}$ with respect to the ℓ^p-norm, we use Hölder's inequality, obtaining

$$S_{\overline{B}_p}(x) = \sup\{\langle x, y\rangle \mid \|y\|_p \leq 1\} = \|x\|_q,$$

where the relation between p and q is given by the equation $1/p+1/q = 1$. □

Closed convex sets are completely characterized by their support functions, due to the following theorem.

Theorem 6.9.2. *Suppose that X_1 and X_2 are two nonempty closed convex subsets of $\mathbf{R}^n$ with support functions S_{X_1} and S_{X_2}, respectively. Then*

$$(a) \qquad X_1 \subseteq X_2 \;\Leftrightarrow\; S_{X_1} \leq S_{X_2}$$
$$(b) \qquad X_1 = X_2 \;\Leftrightarrow\; S_{X_1} = S_{X_2}.$$

Proof. Assertion (b) is an immediate consequence of (a), and the implication $X_1 \subseteq X_2 \;\Rightarrow\; S_{X_1} \leq S_{X_2}$ is trivial, so it only remains to prove the converse implication, or equivalently, the implication $X_1 \not\subseteq X_2 \Rightarrow S_{X_1} \not\leq S_{X_2}$.

To prove the latter implication we assume that $X_1 \not\subseteq X_2$, i.e. that there exists a point $x_0 \in X_1 \setminus X_2$. The point x_0 is strictly separable from the closed convex set X_2, which means that there exist a vector $c \in \mathbf{R}^n$ and a number b such that $\langle x, c\rangle \leq b$ for all $x \in X_2$ while $\langle x_0, c\rangle > b$. Consequently,

$$S_{X_1}(c) \geq \langle x_0, c\rangle > b \geq \sup\{\langle x, c\rangle \mid x \in X_2\} = S_{X_2}(c),$$

which shows that $S_{X_1} \not\leq S_{X_2}$. □

By combining the previous theorem with property (i) of Theorem 6.9.1, we obtain the following corollary.

Corollary 6.9.3. *Let A and B be two nonempty subsets of $\mathbf{R}^n$. Then,*

$$S_A = S_B \;\Leftrightarrow\; \mathrm{cl}(\mathrm{cvx}\, A) = \mathrm{cl}(\mathrm{cvx}\, B).$$

6.10 The Minkowski functional

Let X be a convex subset of $\mathbf{R}^n$ with 0 as an interior point of X. Consider the sets tX for $t \geq 0$. This is an increasing family of sets, whose union equals all of $\mathbf{R}^n$, i.e.

$$0 \leq s < t \Rightarrow sX \subseteq tX \qquad \text{and} \qquad \bigcup_{t\geq 0} tX = \mathbf{R}^n.$$

The family is increasing, because using the convexity of the sets tX and the fact that they contain 0, we obtain the following inclusions for $0 \le s < t$:

$$sX = \frac{s}{t}(tX) + (1 - \frac{s}{t})\,0 \subseteq \frac{s}{t}(tX) + (1 - \frac{s}{t})(tX) \subseteq tX.$$

That the union equals $\mathbf{R}^n$ only depends on 0 being an interior point of X. For let $\overline{B}(0; r_0)$ be a closed ball centered at 0 and contained in X. An arbitrary point $x \in \mathbf{R}^n$ will then belong to the set $r_0^{-1}\|x\|X$ since $r_0\|x\|^{-1}x$ lies in $\overline{B}(0; r_0)$.

Now fix $x \in \mathbf{R}^n$ and consider the set $\{t \ge 0 \mid x \in tX\}$. This set is an unbounded subinterval of $[0, \infty[$, and it contains the number $r_0^{-1}\|x\|$. We may therefore define a function

$$\phi_X \colon \mathbf{R}^n \to \mathbf{R}_+$$

by letting

$$\phi_X(x) = \inf\{t \ge 0 \mid x \in tX\}.$$

Obviously,

$$\phi_X(x) \le r_0^{-1}\|x\| \qquad \text{for all } x.$$

Definition. The function $\phi_X \colon \mathbf{R}^n \to \mathbf{R}_+$ is called the *Minkowski functional* of the set X.

Theorem 6.10.1. *The Minkowski functional ϕ_X has the following properties:*

(i) For all $x, y \in \mathbf{R}^n$ and all $\lambda \in \mathbf{R}_+$,
  (a) $\phi_X(\lambda x) = \lambda\phi_X(x)$,
  (b) $\phi_X(x + y) \le \phi_X(x) + \phi_X(y)$.

(ii) There exists a constant C such that

$$|\phi_X(x) - \phi_X(y)| \le C\|x - y\|$$

for all $x, y \in \mathbf{R}^n$.

(iii) $\operatorname{int} X = \{x \in \mathbf{R}^n \mid \phi_X(x) < 1\}$ *and* $\operatorname{cl} X = \{x \in \mathbf{R}^n \mid \phi_X(x) \le 1\}$.

The Minkowski functional is, in other words, *positive homogeneous*, *subadditive*, and *Lipschitz continuous*. So it is in particular a convex function.

Proof. (i) The equivalence $x \in tX \Leftrightarrow \lambda x \in \lambda tX$, which holds for $\lambda > 0$, together with the fact that $\phi_X(0) = 0$, implies positive homogenouity.

To prove subadditivity we choose, given $\epsilon > 0$, two positive numbers $s < \phi_X(x) + \epsilon$ and $t < \phi_X(y) + \epsilon$ such that $x \in sX$ and $y \in tX$. The point

$$\frac{1}{s+t}(x + y) = \frac{s}{s+t}\,\frac{x}{s} + \frac{t}{s+t}\,\frac{y}{t}$$

is a point in X, by convexity, and it follows that the point $x + y$ belongs to the set $(s + t)X$. This implies that

$$\phi_X(x + y) \leq s + t < \phi_X(x) + \phi_X(y) + 2\epsilon,$$

and since this inequality is true for all $\epsilon > 0$, we conclude that

$$\phi_X(x + y) \leq \phi_X(x) + \phi_X(y).$$

(ii) We have already noted that the inequality $\phi_X(x) \leq C\|x\|$ holds for all x with $C = r_0^{-1}$. By subadditivity,

$$\phi_X(x) = \phi_X(x - y + y) \leq \phi_X(x - y) + \phi_X(y),$$

and hence

$$\phi_X(x) - \phi_X(y) \leq \phi_X(x - y) \leq C\|x - y\|.$$

For symmetry reasons

$$\phi_X(y) - \phi_X(x) \leq C\|y - x\| = C\|x - y\|,$$

and hence $|\phi_X(x) - \phi_X(y)| \leq C\|x - y\|$.

(iii) The sets $\{x \in \mathbf{R}^n \mid \phi_X(x) < 1\}$ and $\{x \in \mathbf{R}^n \mid \phi_X(x) \leq 1\}$ are open and closed, respectively, since ϕ_X is continuous. Therefore, to prove assertion (iii) it suffices, due to the characterization of $\operatorname{int} X$ as the largest open set contained in X and of $\operatorname{cl} X$ as the smallest closed set containing X, to prove the inclusions

$$\operatorname{int} X \subseteq \{x \in \mathbf{R}^n \mid \phi_X(x) < 1\} \subseteq X \subseteq \{x \in \mathbf{R}^n \mid \phi_X(x) \leq 1\} \subseteq \operatorname{cl} X.$$

Suppose $x \in \operatorname{int} X$. Since $tx \to x$ as $t \to 1$, the points tx belong to the interior of X for all numbers t that are sufficiently close to 1. Thus, there exists a number $t_0 > 1$ such that $t_0 x \in X$, i.e. such that $x \in t_0^{-1}X$, which means that $\phi_X(x) \leq t_0^{-1} < 1$, and this proves the inclusion

$$\operatorname{int} X \subseteq \{x \in \mathbf{R}^n \mid \phi_X(x) < 1\}.$$

The implications $\phi_X(x) < t \Rightarrow x \in tX \Rightarrow \phi_X(x) \leq t$ are direct consequences of the definition of $\phi_X(x)$, and by choosing $t = 1$ we obtain the inclusions

$$\{x \in \mathbf{R}^n \mid \phi_X(x) < 1\} \subseteq X \subseteq \{x \in \mathbf{R}^n \mid \phi_X(x) \leq 1\}.$$

To prove the remaining inclusion it is now enough to prove the inclusion

$$\{x \in \mathbf{R}^n \mid \phi_X(x) = 1\} \subseteq \operatorname{cl} X.$$

So, suppose $\phi_X(x) = 1$. Then there is a sequence $(t_n)_1^\infty$ of numbers > 1 such that $t_n \to 1$ as $n \to \infty$ and $x \in t_n X$ for all n. The points $t_n^{-1}x$ belong to X for all n, and since $t_n^{-1}x \to x$ as $n \to \infty$, x is a point in the closure $\operatorname{cl} X$. □

Exercises

6.1 Find two quasiconvex functions f_1, f_2 with a non-quasiconvex sum $f_1 + f_2$.

6.2 Prove that the following functions $f\colon \mathbf{R}^3 \to \mathbf{R}$ are convex:

a) $f(x) = x_1^2 + 2x_2^2 + 5x_3^2 + 3x_2x_3$

b) $f(x) = 2x_1^2 + x_2^2 + x_3^2 - 2x_1x_2 + 2x_1x_3$

c) $f(x) = \mathrm{e}^{x_1 - x_2} + \mathrm{e}^{x_2 - x_1} + x_3^2 - 2x_3$.

6.3 For which values of the real number a is the function

$$f(x) = x_1^2 + 2x_2^2 + ax_3^2 - 2x_1x_2 + 2x_1x_3 - 6x_2x_3$$

convex and strictly convex?

6.4 Prove that the function $f(x) = x_1x_2\cdots x_n$ with $\mathbf{R}^n_+$ as domain is quasiconcave, and that the function $g(x) = (x_1x_2\cdots x_n)^{-1}$ with $\mathbf{R}^n_{++}$ as domain is convex.

6.5 Let $x_{[k]}$ denote the k:th biggest coordinate of the point $x = (x_1, x_2, \ldots, x_n)$. In other words, $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$ are the coordinates of x in decreasing order. Prove for each k that the function $f(x) = \sum_{i=1}^{k} x_{[i]}$ is convex.

6.6 Suppose $f\colon \mathbf{R}_+ \to \mathbf{R}$ is convex. Prove that

$$f(x_1) + f(x_2) + \cdots + f(x_n) \le f(x_1 + x_2 + \cdots + x_n) + (n-1)f(0)$$

for all $x_1, x_2, \ldots, x_n \ge 0$. Note the special case $f(0) = 0$!

6.7 The function f is defined on a convex subset of $\mathbf{R}^n$. Suppose that the function $f(x) + \langle c, x\rangle$ is quasiconvex for each $c \in \mathbf{R}^n$. Prove that f is convex.

6.8 We have derived Corollary 6.2.7 from Theorem 6.2.6. Conversely, prove that Theorem 6.2.6 follows easily from Corollary 6.2.7.

6.9 X is a convex set in $\mathbf{R}^n$ with a nonempty interior, and $f\colon X \to \mathbf{R}$ is a continuous function, whose restriction to $\operatorname{int} X$ is convex. Prove that f is convex.

6.10 Suppose that the function $f\colon X \to \overline{\mathbf{R}}$ is convex. Prove that

$$\inf\{f(x) \mid x \in X\} = \inf\{f(x) \mid x \in \operatorname{rint}(\operatorname{dom} f)\}.$$

6.11 Use the method in Example 6.4.1 to determine the minimum of the function

$$g(x_1, x_2) = 16x_1 + 2x_2 + x_1^{-1}x_2^{-2}$$

over the set $x_1 > 0$, $x_2 > 0$.

6.12 Find the Minkowski functional of

a) the closed unit ball $\overline{B}(0;1)$ in $\mathbf{R}^n$ with respect to the ℓ^p-norm $\|\cdot\|_p$;

b) the halfspace $\{x \in \mathbf{R}^n \mid x_1 \le 1\}$.

6.13 Let X be a convex set with 0 as interior point and suppose that the set is symmetric with respect to 0, i.e. $x \in X \Rightarrow -x \in X$. Prove that the Minkowski functional ϕ_X is a norm.

Chapter 7

Smooth convex functions

This chapter is devoted to the study of smooth convex functions, i.e. convex functions that are differentiable. A prerequisite for differentiability at a point is that the function is defined and finite in a neighborhood of the point. Hence, it is only meaningful to study differentiability properties at interior points of the domain of the function, and by passing to the restriction of the function to the interior of its domain, we may as well assume from the beginning that the domain of definition is open. That is the reason for assuming all domains to be open and all function values to be finite in this chapter.

7.1 Convex functions on R

Let f be a real-valued function that is defined in a neighborhood of the point $x \in \mathbf{R}$. The one-sided limit

$$f'_+(x) = \lim_{t \to 0+} \frac{f(x+t) - f(x)}{t},$$

if it exists, is called the *right derivative* of f at the point x. The *left derivative* $f'_-(x)$ is similarly defined as the one-sided limit

$$f'_-(x) = \lim_{t \to 0-} \frac{f(x+t) - f(x)}{t}.$$

The function is obviously differentiable at the point x if and only if the right and the left derivatives both exist and are equal, and the derivative $f'(x)$ is in that case equal to their common value.

The left derivative of the function $f\colon I \to \mathbf{R}$ can be expressed as a right derivative of the function $\check{f}$, defined by

$$\check{f}(x) = f(-x) \quad \text{for all } x \in -I,$$

because

$$f'_-(x) = \lim_{t\to 0+} \frac{f(x-t)-f(x)}{-t} = -\lim_{t\to 0+} \frac{\check{f}(-x+t)-\check{f}(-x)}{t}$$

and hence

$$f'_-(x) = -\check{f}'_+(-x).$$

Observe that the function $\check{f}$ is convex if f is convex.

The basic differentiability properties of convex functions are consequences of the following lemma, which has an obvious interpretation in terms of slopes of various chords. Cf. figure 7.1.

Lemma 7.1.1. *Suppose f is a real-valued convex function that is defined on a subinterval of $\mathbf{R}$ containing the points $x_1 < x_2 < x_3$. Then*

$$\frac{f(x_2)-f(x_1)}{x_2-x_1} \le \frac{f(x_3)-f(x_1)}{x_3-x_1} \le \frac{f(x_3)-f(x_2)}{x_3-x_2}.$$

The above inequalities are strict if f is strictly convex.

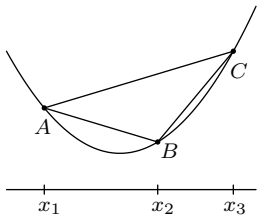

Figure 7.1. A geometric interpretation of Lemma 7.1.1: If k_{PQ} denotes the slope of the chord PQ, then $k_{AB} \le k_{AC} \le k_{BC}$.

Proof. Write $x_2 = \lambda x_3 + (1-\lambda)x_1$; then $\lambda = \dfrac{x_2-x_1}{x_3-x_1}$ is a number in the interval $]0,1[$. By convexity,

$$f(x_2) \le \lambda f(x_3) + (1-\lambda) f(x_1),$$

which simplifies to $f(x_2) - f(x_1) \le \lambda(f(x_3) - f(x_1))$, and this is equivalent to the leftmost of the two inequalities in the lemma.

The rightmost inequality is obtained by applying the already proven inequality to the convex function $\check{f}$. Since $-x_3 < -x_2 < -x_1$,

$$\frac{f(x_2)-f(x_3)}{x_3-x_2} = \frac{\check{f}(-x_2)-\check{f}(-x_3)}{-x_2-(-x_3)} \le \frac{\check{f}(-x_1)-\check{f}(-x_3)}{-x_1-(-x_3)} = \frac{f(x_1)-f(x_3)}{x_3-x_1},$$

and multiplication by -1 gives the desired result.

The above inequalities are strict if f is strictly convex. □

The differentiability properties of convex one-variable functions are given by the following theorem.

Theorem 7.1.2. *Suppose $f\colon I \to \mathbf{R}$ is a convex function with an open subinterval I of $\mathbf{R}$ as its domain. Then:*

(a) The function f has right and left derivatives at all points $x \in I$, and $f'_-(x) \le f'_+(x)$.

(b) If $f'_-(x) \le a \le f'_+(x)$, then

$$f(y) \ge f(x) + a(y - x) \qquad \text{for all } y \in I.$$

The above inequality is strict for $y \ne x$, if f is strictly convex.

(c) If $x < y$, then $f'_+(x) \le f'_-(y)$, and the inequality is strict if f is strictly convex.

(d) The functions $f'_+\colon I \to \mathbf{R}$ and $f'_-\colon I \to \mathbf{R}$ are increasing, and they are strictly increasing if f is strictly convex.

(e) The set of points $x \in I$ where the function is not differentiable, is finite or countable.

Proof. Fix $x \in I$ and let

$$F(t) = \frac{f(x+t) - f(x)}{t}.$$

The domain of F is an open interval J_x with the point 0 removed.

We start by observing that if $s, t, u \in J_x$ and $u < 0 < t < s$, then

$$F(u) \le F(t) \le F(s) \tag{7.1}$$

(and the inequalities are strict if f is strictly convex).

The right inequality $F(t) \le F(s)$ follows directly from the left inequality in Lemma 7.1.1 by choosing $x_1 = x$, $x_2 = x + t$ and $x_3 = x + s$, and the left inequality $F(u) \le F(t)$ follows from the inequality between the extreme ends in the same lemma by instead choosing $x_1 = x + u$, $x_2 = x$ and $x_3 = x + t$.

It follows from inequality (7.1) that the function $F(t)$ is increasing for $t > 0$ (strictly increasing if f is strictly convex) and bounded below by $F(u_0)$, where u_0 is an arbitrary negative number in the domain of F. Hence, the limit

$$f'_+(x) = \lim_{t \to 0+} F(t)$$

exists and

$$F(t) \ge f'_+(x)$$

for all $t > 0$ in the domain of F (with strict inequality if f is strictly convex).

By replacing t with $y-x$, we obtain the following implication for $a \le f'_+(x)$:

$$y > x \Rightarrow f(y) - f(x) \ge f'_+(x)(y-x) \ge a(y-x) \tag{7.2}$$

(with strict inequality if f is strictly convex).

The same argument, applied to the function $\check{f}$ and the point $-x$, shows that $\check{f}'_+(-x)$ exists, and that

$$-y > -x \Rightarrow \check{f}(-y) - \check{f}(-x) \ge -a(-y-(-x))$$

if $-a \le \check{f}'_+(-x)$. Since $f'_-(x) = -\check{f}'_+(-x)$, this means that the left derivative $f'_-(x)$ exists and that the implication

$$y < x \Rightarrow f(y) - f(x) \ge a(y-x) \tag{7.3}$$

is true for all constants a satisfying $a \ge f'_-(x)$. The implications (7.2) and (7.3) are both satisfied if $f'_-(x) \le a \le f'_+(x)$, and this proves assertion (b).

Using inequality (7.1) we conclude that $F(-t) \le F(t)$ for all sufficiently small values of t. Hence

$$f'_-(x) = \lim_{t\to 0+} F(-t) \le \lim_{t\to 0+} F(t) = f'_+(x),$$

and this proves assertion (a).

As a special case of assertion (b), we have the two inequalities

$$f(y) - f(x) \ge f'_+(x)(y-x) \quad \text{and} \quad f(x) - f(y) \ge f'_-(y)(x-y),$$

and division by $y-x$ now results in the implication

$$y > x \Rightarrow f'_+(x) \le \frac{f(y)-f(x)}{y-x} \le f'_-(y).$$

(If f is strictly convex, we may replace $\le$ with $<$ at both places.) This proves assertion (c).

By combining (c) with the inequality in (a) we obtain the implication

$$x < y \Rightarrow f'_+(x) \le f'_-(y) \le f'_+(y),$$

which shows that the right derivative f'_+ is increasing. That the left derivative is increasing is proved in a similar way. (And the derivatives are strictly increasing if f is strictly convex.)

To prove the final assertion (e) we define I_x to be the open interval $]f'_-(x), f'_+(x)[$. This interval is empty if the derivative $f'(x)$ exists, and it is nonempty if the derivative does not exist, and intervals I_x and I_y belonging to different points x and y are disjoint because of assertion (c). Now choose,

for each point x where the derivative does not exist, a rational number r_x in the interval I_x. Since different intervals are pairwise disjoint, the chosen numbers will be different, and since the set of rational numbers is countable, there are at most countably many points x at which the derivative does not exist. □

Definition. The line $y = f(x_0) + a(x - x_0)$ is called a *supporting line* of the function $f\colon I \to \mathbf{R}$ at the point $x_0 \in I$ if

$$f(x) \geq f(x_0) + a(x - x_0) \tag{7.4}$$

for all $x \in I$.

A supporting line at the point x_0 is a line which passes through the point $(x_0, f(x_0))$ and has the entire function curve $y = f(x)$ above (or on) itself. It is, in other words, a (one-dimensional) supporting hyperplane of the epigraph of f at the point $(x_0, f(x_0))$. The concept will be generalized for functions of several variables in the next chapter.

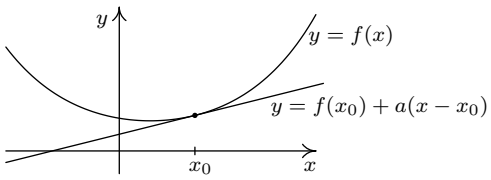

Figure 7.2. A supporting line.

Assertion (b) of the preceding theorem shows that convex functions with open domains have supporting lines at each point, and that the tangent is a supporting line at points where the derivative exists. By our next theorem, the existence of supporting lines is also a sufficient condition for convexity.

Theorem 7.1.3. *Suppose that the function $f\colon I \to \mathbf{R}$, where I is an open interval, has a supporting line at each point in I. Then, f is a convex function.*

Proof. Suppose that $x, y \in I$ and that $0 < \lambda < 1$, and let a be the constant belonging to the point $x_0 = \lambda x + (1-\lambda)y$ in the definition (7.4) of a supporting line. Then we have $f(x) \geq f(x_0) + a(x - x_0)$ and $f(y) \geq f(x_0) + a(y - x_0)$. By multiplying the first inequality by λ and the second inequality by $(1-\lambda)$, and then adding the two resulting inequalities, we obtain

$$\lambda f(x) + (1-\lambda) f(y) \geq f(x_0) + a\big(\lambda x + (1-\lambda)y - x_0\big) = f(x_0).$$

So the function f is convex.

Observe that if the inequality (7.4) is strict for all $x \neq x_0$ and for all $x_0 \in I$, then f is strictly convex. □

For differentiable functions we now obtain the following necessary and sufficient condition for convexity.

Theorem 7.1.4. *A differentiable function $f\colon I \to \mathbf{R}$ is convex if and only if its derivative f' is increasing. And it is strictly convex if and only if the derivative is strictly increasing.*

Proof. Assertion (d) in Theorem 7.1.2 shows that the derivative of a (strictly) convex function is (strictly) increasing.

To prove the converse, we assume that the derivative f' is increasing. By the mean value theorem, if x and x_0 are distinct points in I, there exists a point ξ between x and x_0 such that

$$\frac{f(x) - f(x_0)}{x - x_0} = f'(\xi) \begin{cases} \geq f'(x_0) & \text{if } x > x_0, \\ \leq f'(x_0) & \text{if } x < x_0. \end{cases}$$

Multiplication by $x - x_0$ results, in both cases, in the inequality

$$f(x) - f(x_0) \geq f'(x_0)(x - x_0),$$

which shows that $y = f(x_0) + f'(x_0)(x - x_0)$ is a supporting line of the function f at the point x_0. Therefore, f is convex by Theorem 7.1.3.

The above inequalites are strict if the derivative is strictly increasing, and we conclude that f is strictly convex in that case. □

For two times differentiable functions we obtain the following corollary.

Corollary 7.1.5. *A two times differentiable function $f\colon I \to \mathbf{R}$ is convex if and only if $f''(x) \geq 0$ for all $x \in I$. The function is strictly convex if $f''(x) > 0$ for all $x \in I$.*

Proof. The derivative f' is increasing (strictly increasing) if the second derivative f'' is nonnegative (positive). And the second derivative is nonnegative if the derivative is increasing. □

Remark. A continuous function $f\colon J \to \mathbf{R}$ with a non-open interval J as domain is convex if (and only if) the restriction of f to the interior of J is convex. Hence, if the derivative exists and is increasing in the interior of J, or if the second derivative exists and $f''(x) \geq 0$ for all interior points x of the interval, then f is convex on J. Cf. exercise 7.7.

EXAMPLE 7.1.1. The functions $x \mapsto \mathrm{e}^x$, $x \mapsto -\ln x$ and $x \mapsto x^p$, where $p > 1$, are strictly convex on their domains $\mathbf{R}$, $]0, \infty[$ and $[0, \infty[$, respectively, because their first derivatives are strictly increasing functions. □

7.2 Differentiable convex functions

A differentiable one-variable function f is convex if and only if its derivative is an increasing function. In order to generalize this result to functions of several variables it is necessary to express the condition that the derivative is increasing in a generalizable way. To this end, we note that the derivative f' is increasing on an interval if and only if $f'(x+h)h \geq f'(x)h$ for all numbers x and $x+h$ in the interval, and this inequality is also meaningful for functions f of several variables if we interpret $f'(x)h$ as the value of the linear form $Df(x)$ at h. The inequality generalizing that the derivative of a function of several variables is increasing will thus be written $Df(x+h)[h] \geq Df(x)[h]$, or using gradient notation, $\langle f'(x+h), h\rangle \geq \langle f'(x), h\rangle$.

Theorem 7.2.1. *Let X be an open convex subset of $\mathbf{R}^n$, and let $f\colon X \to \mathbf{R}$ be a differentiable function. The following three conditions are equivalent:*

(i) *f is a convex function.*
(ii) *$f(x+v) \geq f(x) + Df(x)[v]$ for all $x, x+v \in X$.*
(iii) *$Df(x+v)[v] \geq Df(x)[v]$ for all $x, x+v \in X$.*

The function f is strictly convex if and only if the inequalities in (ii) and (iii) can be replaced by strict inequalities when $v \neq 0$.

Proof. Let us for given points x and $x+v$ in X consider the restriction $\phi_{x,v}$ of f to the line through x with direction v, i.e. the one-variable function

$$\phi_{x,v}(t) = f(x+tv)$$

with the open interval $I_{x,v} = \{t \in \mathbf{R} \mid x+tv \in X\}$ as domain. The functions $\phi_{x,v}$ are differentiable with derivative $\phi'_{x,v}(t) = Df(x+tv)[v]$, and f is convex if and only if all restrictions $\phi_{x,v}$ are convex.

(i) ⇒ (ii) So if f is convex, then $\phi_{x,v}$ is a convex function, and it follows from Theorem 7.1.2 (b) that $\phi_{x,v}(t) \geq \phi_{x,v}(0) + \phi'_{x,v}(0)t$ for all $t \in I_{x,v}$, which means that $f(x+tv) \geq f(x) + Df(x)[v]\,t$ for all t such that $x+tv \in X$. We now obtain the inequality in (ii) by choosing $t = 1$.

(ii) ⇒ (iii) We obtain inequality (iii) by adding the two inequalities

$$f(x+v) \geq f(x) + Df(x)[v] \quad \text{and} \quad f(x) \geq f(x+v) + Df(x+v)[-v].$$

(iii) ⇒ (i) Suppose (iii) holds, and let $y = x + sv$ and $w = (t-s)v$. If $t > s$, then

$$\begin{aligned}\phi'_{x,v}(t) - \phi'_{x,v}(s) &= Df(x+tv)[v] - Df(x+sv)[v] \\ &= Df(y+w)[v] - Df(y)[v] \\ &= \frac{1}{t-s}\big(Df(y+w)[w] - Df(y)[w]\big) \geq 0,\end{aligned}$$

which means that the derivative $\phi'_{x,v}$ is increasing. The functions $\phi_{x,v}$ are thus convex.

This proves the equivalence of assertions (i), (ii) and (iii), and by replacing all inequalities in the proof by strict inequalities, we obtain the corresponding equivalent assertions for strictly convex functions. □

The derivative of a differentiable function is equal to zero at a local minimum point. For convex functions, the converse is also true.

Theorem 7.2.2. *Suppose $f\colon X \to \mathbf{R}$ is a differentiable convex function. Then $\hat{x} \in X$ is a global minimum point if and only if $f'(\hat{x}) = 0$.*

Proof. That the derivative equals zero at a minimum point is a general fact, and the converse is a consequence of property (ii) in the previous theorem, for if $f'(\hat{x}) = 0$, then $f(x) \geq f(\hat{x}) + Df(\hat{x})[x - \hat{x}] = f(\hat{x})$ for all $x \in X$. □

Convexity can also be expressed by a condition on the second derivative, and the natural substitute for the one-variable condition $f''(x) \geq 0$ is that the second derivative should be positive semidefinite.

Theorem 7.2.3. *Let X be an open convex subset of $\mathbf{R}^n$, and suppose that the function $f\colon X \to \mathbf{R}$ is two times differentiable. Then f is convex if and only if the second derivative $f''(x)$ is positive semidefinite for all $x \in X$.*

If $f''(x)$ is positive definite for all $x \in X$, then f is strictly convex.

Proof. The one-variable functions $\phi_{x,v}(t) = f(x + tv)$ are now two times differentiable with second derivative

$$\phi''_{x,v}(t) = D^2 f(x+tv)[v,v] = \langle v, f''(x+tv)v\rangle.$$

Since f is convex if and only if all functions $\phi_{x,v}$ are convex, f is convex if and only if all second derivatives $\phi''_{x,v}$ are nonnegative functions .

If the second derivative $f''(x)$ is positive semidefinite for all $x \in X$, then $\phi''_{x,v}(t) = \langle v, f''(x+tv)v\rangle \geq 0$ for all $x \in X$ and all $v \in \mathbf{R}^n$, which means that the second derivatives $\phi''_{x,v}$ are nonnegative funtions. Conversely, if the second

derivatives $\phi''_{x,v}$ are nonnegative, then in particular $\langle v, f''(x)\rangle = \phi''_{x,v}(0) \geq 0$ for all $x \in X$ and all $v \in \mathbf{R}^n$, and we conclude that the second derivative $f''(x)$ is positive semidefinite for all $x \in X$.

If the second derivatives $f''(x)$ are all positive definite, then $\phi''_{x,v}(t) > 0$ for $v \neq 0$, which implies that the functions $\phi_{x,v}$ are strictly convex, and then f is strictly convex, too. □

7.3 Strong convexity

The function surface of a convex functions bends upwards, but there is no lower positive bound on the curvature. By introducing such a bound we obtain the notion of strong convexity.

Definition. Let μ be a positive number. A function $f\colon X \to \overline{\mathbf{R}}$ is called *μ-strongly convex* if the function $f(x) - \frac{1}{2}\mu\|x\|^2$ is convex, and the function f is called *strongly convex* if it is μ-strongly convex for some positive number μ.

Theorem 7.3.1. *A differentiable function $f\colon X \to \mathbf{R}$ with a convex domain is μ-strongly convex if and only if the following two mutually equivalent inequalities are satisfied for all $x, x+v \in X$:*

(i) $$Df(x+v)[v] \geq Df(x)[v] + \mu\|v\|^2$$

(ii) $$f(x+v) \geq f(x) + Df(x)[v] + \tfrac{1}{2}\mu\|v\|^2.$$

Proof. Let $g(x) = f(x) - \frac{1}{2}\mu\|x\|^2$ and note that $g'(x) = f'(x) - \mu x$ and that consequently $Df(x)[v] = Dg(x)[v] + \mu\langle x, v\rangle$.

If f is μ-strongly convex, then g is a convex function, and so it follows from Theorem 7.2.1 that

$$\begin{aligned} Df(x+v)[v] - Df(x)[v] &= Dg(x+v)[v] - Dg(x)[v] + \mu\langle x+v, v\rangle - \mu\langle x, v\rangle \\ &\geq \mu\langle v, v\rangle = \mu\|v\|^2, \end{aligned}$$

i.e. inequality (i) is satisfied.

(i) ⇒ (ii): Assume (i) holds, and define the function Φ for $0 \leq t \leq 1$ by

$$\Phi(t) = f(x+tv) - f(x) - Df(x)[v]\,t.$$

Then $\Phi'(t) = Df(x+tv)[v] - Df(x)[v] = \dfrac{1}{t}\big(Df(x+tv)[tv] - Df(x)[tv]\big)$, and it now follows from inequality (i) that

$$\Phi'(t) \geq t^{-1}\mu\|tv\|^2 = \mu\|v\|^2\,t.$$

By integrating the last inequality over the interval $[0,1]$ we obtain

$$\Phi(1) = \Phi(1) - \Phi(0) \geq \tfrac{1}{2}\mu\|v\|^2,$$

which is the same as inequality (ii).

If inequality (ii) holds, then

$$\begin{aligned} g(x+v) &= f(x+v) - \tfrac{1}{2}\mu\|x+v\|^2 \geq f(x) + Df(x)[v] + \tfrac{1}{2}\mu\|v\|^2 - \tfrac{1}{2}\mu\|x+v\|^2 \\ &= g(x) + \tfrac{1}{2}\mu\|x\|^2 + Dg(x)[v] + \mu\langle x, v\rangle + \tfrac{1}{2}\mu\|v\|^2 - \tfrac{1}{2}\mu\|x+v\|^2 \\ &= g(x) + Dg(x)[v]. \end{aligned}$$

The function g is thus convex, by Theorem 7.2.1, and $f(x) = g(x) + \frac{1}{2}\mu\|x\|^2$ is consequently μ-strongly convex. □

Theorem 7.3.2. *A twice differentiable function $f\colon X \to \mathbf{R}$ with a convex domain is μ-strongly convex if and only if*

$$\langle v, f''(x)v\rangle = D^2 f(x)[v,v] \geq \mu\|v\|^2 \tag{7.5}$$

for all $x \in X$ and all $v \in \mathbf{R}^n$.

Remark. If A is a symmetric operator, then

$$\min_{v\neq 0} \frac{\langle v, Av\rangle}{\|v\|^2} = \lambda_{\min},$$

where $\lambda_{\min}$ is the smallest eigenvalue of the operator. Thus, a two times differentiable function f with a convex domain is μ-strongly convex if and only if the eigenvalues of the hessian $f''(x)$ are greater than or equal to μ for each x in the domain.

Proof. Let $\phi_{x,v}(t) = f(x+tv)$. If condition (7.5) holds, then

$$\phi''_{x,v}(t) = D^2 f(x+tv)[v,v] \geq \mu\|v\|^2$$

for all t in the domain of the function. Using Taylor's formula with remainder term, we therefore conclude that

$$\phi_{x,v}(t) = \phi_{x,v}(0) + \phi'_{x,v}(0)t + \tfrac{1}{2}\phi''_{x,v}(\xi)t^2 \geq \phi_{x,v}(0) + \phi'_{x,v}(0)t + \tfrac{1}{2}\mu\|v\|^2\, t^2.$$

For $t = 1$ this amounts to inequality (ii) in Theorem 7.3.1, and hence f is a μ-strongly convex function.

Conversely, if f is μ-strongly convex, then by Theorem 7.3.1 (i)

$$\frac{\phi'_{x,v}(t) - \phi'_{x,v}(0)}{t} = \frac{Df(x+tv)[tv] - Df(x)[tv]}{t^2} \geq \mu\|v\|^2.$$

Taking the limit as $t \to 0$ we obtain

$$D^2 f(x)[v,v] = \phi''_{x,v}(0) \geq \mu\|v\|^2.$$ □

7.4 Convex functions with Lipschitz continuous derivatives

The rate of convergence of classical iterative algorithms for minimizing functions depends on the variation of the deriviative − the more the derivative varies in a neighborhood of the minimum point, the slower the convergence. The size of the Lipschitz constant is a measure of the variation of the derivative for functions with a Lipschitz continuous derivative. Therefore, we start with a result which for arbitrary functions connects Lipschitz continuity of the first derivative to bounds on the second derivative.

Theorem 7.4.1. *Suppose f is a twice differentiable function and that X is a convex subset of its domain.*

(i) If $\|f''(x)\| \leq L$ for all $x \in X$, then the derivative f' is Lipschitz continuous on X with Lipschitz constant L.

(ii) If the derivative f' is Lipschitz continuous on the set X with constant L, then $\|f''(x)\| \leq L$ for all $x \in \operatorname{int} X$.

Proof. (i) Suppose that $\|f''(x)\| \leq L$ for all $x \in X$, and let x and y be two points in X. Put $v = y - x$, let w be an arbitrary vector with $\|w\| = 1$, and define the function ϕ for $0 \leq t \leq 1$ by

$$\phi(t) = Df(x + tv)[w] = \langle f'(x + tv), w\rangle.$$

Then ϕ is differentiable with derivative

$$\phi'(t) = D^2 f(x + tv)[w, v] = \langle w, f''(x + tv)v\rangle$$

so it follows from the Cauchy-Schwarz inequality that

$$|\phi'(t)| \leq \|w\|\|f''(x + tv)v\| \leq \|f''(x + tv)\|\|v\| \leq L\|v\|,$$

since $x + tv$ is a point in X. By the mean value theorem, $\phi(1) - \phi(0) = \phi'(s)$ for some point $s \in]0, 1[$. Consequently,

$$|\langle f'(y) - f'(x), w\rangle| = |\phi(1) - \phi(0)| = |\phi'(s)| \leq L\|y - x\|.$$

Since w is an arbitrary vector of norm 1, we conclude that

$$\|f'(y) - f'(x)\| = \sup_{\|w\|=1} \langle f'(y) - f'(x), w\rangle \leq L\|y - x\|,$$

i.e. the derivative f' is Lipschitz continuous on X with constant L.

(ii) Assume conversely that the first derivative f' is Lipschitz continuous on the set X with constant L. Let x be a point in the interior of X, and let v and w be arbitrary vectors with norm 1. The function

$$\phi(t) = Df(x+tv)[w] = \langle f'(x+tv), w\rangle$$

is then defined and differentiable and the point $x+tv$ lies in X for all t in a neighborhood of 0, and it follows that

$$\begin{aligned}|\phi(t)-\phi(0)| &= |\langle f'(x+tv)-f'(x), w\rangle| \le \|f'(x+tv)-f'(x)\|\|w\| \\ &\le L\|tv\| = L|t|.\end{aligned}$$

Division by t and passing to the limit as $t \to 0$ results in the inequality

$$|\langle w, f''(x)v\rangle| = |\phi'(0)| \le L$$

with the conclustion that

$$\|f''(x)\| = \sup_{\|v\|=1} \|f''(x)v\| = \sup_{\|v\|,\|w\|=1} \langle w, f''(x)v\rangle \le L. \qquad \square$$

Definition. A differentiable function $f\colon X \to \mathbf{R}$ belongs to the class $\mathcal{S}_{\mu,L}(X)$ if f is μ-strongly convex and the derivative f' is Lipschitz continuous with constant L. The quotient $Q = L/\mu$ is called the *condition number* of the class.

Due to Theorem 7.3.1, a differentiable function f with a convex domain X belongs to the class $\mathcal{S}_{\mu,L}(X)$ if and only if it satisfies the following two inequalities for all $x, x+v \in X$:

$$\langle f'(x+v)-f'(x), v\rangle \ge \mu\|v\|^2 \quad \text{and} \quad \|f'(x+v)-f'(x)\| \le L\|v\|.$$

If we combine the first of these two inequalities with the Cauchy–Schwarz inequality, we obtain the inequality $\mu\|v\| \le \|f'(x+v)-f'(x)\|$, so we conclude that $\mu \le L$ and $Q \ge 1$.

Example 7.4.1. Strictly convex quadratic functions

$$f(x) = \tfrac{1}{2}\langle x, Px\rangle + \langle q, x\rangle + r$$

belong to the class $\mathcal{S}_{\lambda_{\min},\lambda_{\max}}(\mathbf{R}^n)$, where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and the largest eigenvalue, respectively, of the positive definite matrix P.

For $f'(x) = Px + q$ and $f''(x) = P$, whence

$$\begin{aligned}D^2f(x)[v,v] = \langle v, Pv\rangle &\ge \lambda_{\min}\|v\|^2 \qquad \text{and} \\ \|f'(x+v)-f'(x)\| = \|Pv\| &\le \|P\|\,\|v\| = \lambda_{\max}\|v\|.\end{aligned}$$

The condition number of the quadratic function f is thus equal to the quotient $\lambda_{\max}/\lambda_{\min}$ between the largest and the smallest eigenvalue. $\square$

The sublevel sets $\{x \mid f(x) \leq \alpha\}$ of a strictly convex quadratic function f are ellipsoids for all values of α greater than the minimum value of the function, and the ratio of the longest and the shortest axes of any of these ellipsoids is equal to $\sqrt{\lambda_{\max}/\lambda_{\min}}$, i.e. to the square root of the condition number Q. This ratio is obviously also equal to the ratio of the radii of the smallest ball containing and the largest ball contained in the ellipsoid. As we shall see, something similar applies to all functions in the class $\mathcal{S}_{\mu,L}(\mathbf{R}^n)$.

Theorem 7.4.2. *Let f be a function in the class $\mathcal{S}_{\mu,L}(\mathbf{R}^n)$ with minimum point $\hat{x}$, and let α be a number greater than the minimum value $f(\hat{x})$. Then*

$$B(\hat{x}; r) \subseteq \{x \in X \mid f(x) \leq \alpha\} \subseteq B(\hat{x}; R),$$

where $r = \sqrt{2L^{-1}(\alpha - f(\hat{x}))}$ and $R = \sqrt{2\mu^{-1}(\alpha - f(\hat{x}))}$.

Remark. Note that $R/r = \sqrt{L/\mu} = \sqrt{Q}$.

Proof. Since $f'(\hat{x}) = 0$ we obtain the following inequalities from Theorems 1.1.2 and 7.3.1 (by replacing a and x respectively with $\hat{x}$ and v with $x - \hat{v}$):

$$f(\hat{x}) + \tfrac{1}{2}\mu\|x - \hat{x}\|^2 \leq f(x) \leq f(\hat{x}) + \tfrac{1}{2}L\|x - \hat{x}\|^2.$$

Hence, $x \in S = \{x \in X \mid f(x) \leq \alpha\}$ implies

$$\tfrac{1}{2}\mu\|x - \hat{x}\|^2 \leq f(x) - f(\hat{x}) \leq \alpha - f(\hat{x}) = \tfrac{1}{2}\mu R^2,$$

which means that $\|x - \hat{x}\| \leq R$ and proves the inclusion $S \subseteq B(\hat{x}; R)$.

And if $x \in B(\hat{x}; r)$, then $f(x) \leq f(\hat{x}) + \frac{1}{2}Lr^2 = \alpha$, which means that $x \in S$ and proves the inclusion $B(\hat{x}; r) \subseteq S$. □

Convex functions on $\mathbf{R}^n$ with Lipschitz continuous derivatives are characterized by the following theorem.

Theorem 7.4.3. *A differentiable function $f\colon \mathbf{R}^n \to \mathbf{R}$ is convex and its derivative is Lipschitz continuous with Lipschitz constant L if and only if the following mutually equivalent inequalities are fulfilled for all $x, v \in \mathbf{R}^n$:*

(i) $$f(x) + Df(x)[v] \leq f(x+v) \leq f(x) + Df(x)[v] + \frac{L}{2}\|v\|^2$$

(ii) $$f(x+v) \geq f(x) + Df(x)[v] + \frac{1}{2L}\|f'(x+v) - f'(x)\|^2$$

(iii) $$Df(x+v)[v] \geq Df(x)[v] + \frac{1}{L}\|f'(x+v) - f'(x)\|^2$$

Proof. That inequality (i) has to be satisfied for convex functions with a Lipschitz continuous derivative is a consequence of Theorems 1.1.2 and 7.2.1.

(i) ⇒ (ii): Let $w = f'(x+v) - f'(x)$, and apply the right inequality in (i) with x replaced by $x+v$ and v replaced by $-L^{-1}w$; this results in the inequality

$$f(x+v-L^{-1}w) \le f(x+v) - L^{-1}Df(x+v)[w] + \tfrac{1}{2}L^{-1}\|w\|^2.$$

The left inequality in (i) with $v - L^{-1}w$ instead of v yields

$$f(x+v-L^{-1}w) \ge f(x) + Df(x)[v-L^{-1}w].$$

By combining these two new inequalities, we obtain

$$\begin{aligned} f(x+v) &\ge f(x) + Df(x)[v-L^{-1}w] + L^{-1}Df(x+v)[w] - \tfrac{1}{2}L^{-1}\|w\|^2 \\ &= f(x) + Df(x)[v] + L^{-1}\big(Df(x+v)[w] - Df(x)[w]\big) - \tfrac{1}{2}L^{-1}\|w\|^2 \\ &= f(x) + Df(x)[v] + L^{-1}\langle f'(x+v) - f'(x), w\rangle - \tfrac{1}{2}L^{-1}\|w\|^2 \\ &= f(x) + Df(x)[v] + L^{-1}\langle w, w\rangle - \tfrac{1}{2}L^{-1}\|w\|^2 \\ &= f(x) + Df(x)[v] + \tfrac{1}{2}L^{-1}\|w\|^2, \end{aligned}$$

and that is inequality (ii).

(ii) ⇒ (iii): Add inequality (ii) to the inequality obtained by changing x to $x+v$ and v to $-v$. The result is inequality (iii).

Let us finally assume that inequality (iii) holds. The convexity of f is then a consequence of Theorem 7.2.1, and by combining (iii) with the Cauchy–Schwarz inequality, we obtain the inequality

$$\begin{aligned} \frac{1}{L}\|f'(x+v) - f'(x)\|^2 \le Df(x+v)[v] - Df(x)[v] &= \langle f'(x+v) - f'(x), v\rangle \\ &\le \|f'(x+v) - f'(x)\| \cdot \|v\|, \end{aligned}$$

which after division by $\|f'(x+v) - f'(x)\|$ gives us the desired conclusion: the derivative is Lipschitz continuous with Lipschitz constant L. □

Theorem 7.4.4. *If $f \in \mathcal{S}_{\mu,L}(\mathbf{R}^n)$, then*

$$Df(x+v)[v] \ge Df(x)[v] + \frac{\mu L}{\mu+L}\|v\|^2 + \frac{1}{\mu+L}\|f'(x+v) - f'(x)\|^2$$

for all $x, v \in \mathbf{R}^n$.

Proof. Let $g(x) = f(x) - \frac{1}{2}\mu\|x\|^2$; the function g is then convex, and since $Dg(x)[v] = Df(x)[v] - \mu\langle x, v\rangle$, it follows from Theorem 1.1.2 that

$$\begin{aligned} g(x+v) &= f(x+v) - \tfrac{1}{2}\mu\|x+v\|^2 \\ &\le f(x) + Df(x)[v] + \tfrac{1}{2}L\|v\|^2 - \tfrac{1}{2}\mu\|x+v\|^2 \\ &= g(x) + \tfrac{1}{2}\mu\|x\|^2 + Dg(x)[v] + \mu\langle x, v\rangle + \tfrac{1}{2}L\|v\|^2 - \tfrac{1}{2}\mu\|x+v\|^2 \\ &= g(x) + Dg(x)[v] + \tfrac{1}{2}(L-\mu)\|v\|^2. \end{aligned}$$

This shows that g satisfies condition (i) in Theorem 7.4.3 with L replaced by $L - \mu$. The derivative g' is consequently Lipschitz continuous with constant $L - \mu$. The same theorem now gives us the inequality

$$Dg(x+v)[v] \ge Dg(x)[v] + \frac{1}{L-\mu}\|g'(x+v) - g'(x)\|^2,$$

which is just a reformulation of the inequality in Theorem 7.4.4. □

Exercises

7.1 Show that the following functions are convex.

a) $f(x_1, x_2) = \mathrm{e}^{x_1} + \mathrm{e}^{x_2} + x_1x_2, \quad x_1 + x_2 > 0$

b) $f(x_1, x_2) = \sin(x_1 + x_2), \quad -\pi < x_1 + x_2 < 0$

c) $f(x_1, x_2) = -\sqrt{\cos(x_1 + x_2)}, \quad -\frac{\pi}{2} < x_1 + x_2 < \frac{\pi}{2}$.

7.2 Is the function

$$f(x_1, x_2) = \frac{x_1^2}{x_2} + \frac{x_2^2}{x_1}$$

convex in the first quadrant $x_1 > 0$, $x_2 > 0$?

7.3 Show that the function

$$f(x) = \sum_{j=1}^{n-1} x_j^2/x_n$$

is convex in the halfspace $x_n > 0$.

7.4 Show that the following function is concave on the set $[0,1[\times[0,1[\times[0,1[$:

$$\begin{aligned} f(x_1, x_2, x_3) = \ln(1-x_1) + \ln(1-x_2) + \ln(1-x_3) \\ - (x_1^2 + x_2^2 + x_3^2 + x_1x_2 + x_1x_3 + x_2x_3). \end{aligned}$$

7.5 Let I be an interval and suppose that the function $f\colon I \to \mathbf{R}$ is convex. Show that f is either increasing on the interval, or decreasing on the interval, or there exists a point $c \in I$ such that f is decreasing to the left of c and increasing to the right of c.

7.6 Suppose $f\colon]a,b[\to \mathbf{R}$ is a convex function.

a) Prove that the two one-sided limits $\lim_{x\to a+} f(x)$ and $\lim_{x\to b-} f(x)$ exist (as finite numbers or $\pm\infty$).

b) Suppose that the interval is finite, and extend the function to the closed interval $[a,b]$ by defining $f(a) = \alpha$ and $f(b) = \beta$. Prove that the extended function is convex if and only if $\alpha \geq \lim_{x\to a+} f(x)$ and $\beta \geq \lim_{x\to b-} f(x)$.

7.7 Prove that a continuous function $f\colon [a,b] \to \mathbf{R}$ is convex if and only if its restriction to the open interval $]a,b[$ is convex.

7.8 $\mathcal{F}$ is a family of differentiable functions on $\mathbf{R}^n$ with the following two properties:

(i) $f \in \mathcal{F} \;\Rightarrow\; f + g \in \mathcal{F}$ for all affine functions $g\colon \mathbf{R}^n \to \mathbf{R}$.
(ii) If $f \in \mathcal{F}$ and $f'(x_0) = 0$, then x_0 is a minimum point of f.

Prove that all functions in $\mathcal{F}$ are convex.

7.9 Suppose that $f\colon X \to \mathbf{R}$ is a twice differentiable convex function. Prove that its recessive subspace V_f is a subset of $\mathcal{N}(f''(x))$ for each $x \in X$.

7.10 Let $f\colon X \to \mathbf{R}$ be a differentiable function with a convex domain X. Prove that f is quasiconvex if and only if

$$f(x+v) \leq f(x) \;\Rightarrow\; Df(x)[v] \leq 0$$

for all $x, x+v \in X$.
[Hint: It suffices to prove the assertion for functions on $\mathbf{R}$; the general result then follows by taking restrictions to lines.]

7.11 Let $f\colon X \to \mathbf{R}$ be a twice differentiable function with a convex domain X. Prove the following assertions:

a) If f is quasiconvex, then

$$Df(x)[v] = 0 \;\Rightarrow\; D^2f(x)[v,v] \geq 0$$

for all $x \in X$ and all $v \in \mathbf{R}^n$.

b) If

$$Df(x)[v] = 0 \;\Rightarrow\; D^2f(x)[v,v] > 0$$

for all $x \in X$ and all $v \neq 0$, then f is quasiconvex.
[Hint: It is enough to prove the results for functions defined on $\mathbf{R}$.]

7.12 Prove that the function $\alpha_1 f_1 + \alpha_2 f_2$ is $(\alpha_1\mu_1 + \alpha_2\mu_2)$-strongly convex if f_1 is μ_1-strongly convex, f_2 is μ_2-strongly convex and $\alpha_1, \alpha_2 > 0$.

7.13 Prove that if a differentiable μ-strongly convex function $f\colon X \to \mathbf{R}$ has a minimum at the point $\hat{x}$, then $\|x - \hat{x}\| \leq \mu^{-1}\|f'(x)\|$ for all $x \in X$.

Chapter 8

The subdifferential

We will now generalize a number of results from the previous chapter to convex functions that are not necessarily differentiable everywhere. However, real-valued convex functions with open domains can not be too irregular – they are, as already noted, continuous, and they have direction derivatives.

8.1 The subdifferential

If f is a differentiable function, then $y = f(a) + \langle f'(a), x - a\rangle$ is the equation of a hyperplane that is tangent to the surface $y = f(x)$ at the point $(a, f(a))$. And if f is also convex, then $f(x) \geq f(a) + \langle f'(a), x - a\rangle$ for all x in the domain of the function (Theorem 7.2.1), so the tangent plane lies below the graph of the function and is a supporting hyperplane of the epigraph.

The epigraph of an arbitrary convex function is a convex set, by definition. Hence, through each boundary point belonging to the epigraph of a convex function there passes a supporting hyperplane. The supporting hyperplanes of a convex one-variable function f, defined on an open interval, are given by Theorem 7.1.2, which says that the line $y = f(x_0) + a(x - x_0)$ supports the epigraph at the point $(x_0, f(x_0))$ if (and only if) $f'_-(x_0) \leq a \leq f'_+(x_0)$.

The existence of supporting hyperplanes characterizes convexity, and this is a reason for a more detailed study of this concept.

Definition. Let $f\colon X \to \overline{\mathbf{R}}$ be a function defined on a subset X of $\mathbf{R}^n$. A vector $c \in \mathbf{R}^n$ is called a *subgradient of f at the point $a \in X$* if the inequality

$$f(x) \geq f(a) + \langle c, x - a\rangle \tag{8.1}$$

holds for all $x \in X$.

The set of all subgradients of f at a is called the *subdifferential* of f at a and is denoted by $\partial f(a)$.

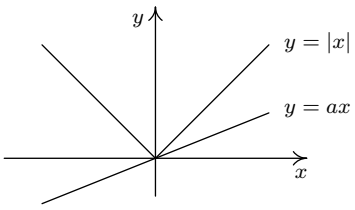

Figure 8.1. The line $y = ax$ is a supporting line of the function $f(x) = |x|$ at the origin if $-1 \leq a \leq 1$.

Remark. The inequality (8.1) is of course satisfied by all points $a \in X$ and all vectors $c \in \mathbf{R}^n$ if x is a point in the set $X \setminus \operatorname{dom} f$. Hence, to verify that c is a subgradient of f at a it suffices to verify that the inequality holds for all $x \in \operatorname{dom} f$.

The inequality (8.1) does not hold for any vector c if a is a point in $X \setminus \operatorname{dom} f$ and x is a point in $\operatorname{dom} f$. Hence, $\partial f(a) = \emptyset$ for all $a \in X \setminus \operatorname{dom} f$, except in the trivial case when $\operatorname{dom} f = \emptyset$, i.e. when f is equal to ∞ on the entire set X. In this case we have $\partial f(a) = \mathbf{R}^n$ for all $a \in X$ since the inequality (8.1) is now trivially satisfied by all $a, x \in X$ and all $c \in \mathbf{R}^n$.

Example 8.1.1. The subdifferentials of the one-variable function $f(x) = |x|$ are

$$\partial f(a) = \begin{cases} \{-1\} & \text{if } a < 0, \\ [-1, 1] & \text{if } a = 0, \\ \{1\} & \text{if } a > 0. \end{cases}$$

□

Theorem 8.1.1. *The subdifferentials of an arbitrary function $f\colon X \to \overline{\mathbf{R}}$ are closed and convex sets.*

Proof. For points $a \in \operatorname{dom} f$,

$$\partial f(a) = \textstyle\bigcap_{x \in \operatorname{dom} f} \{c \in \mathbf{R}^n \mid \langle c, x - a\rangle \leq f(x) - f(a)\}$$

is convex and closed, since it is an intersection of closed halfspaces, and the case $a \in X \setminus \operatorname{dom} f$ is trivial. □

Theorem 8.1.2. *A point $a \in X$ is a global minimum point of the function $f\colon X \to \overline{\mathbf{R}}$ if and only if $0 \in \partial f(a)$.*

Proof. The assertion follows immediately from the subgradient definition. □

Our next theorem tells us that the derivative $f'(a)$ is the only subgradient candidate for functions f that are differentiable at a. Geometrically this means that the tangent plane at a is the only possible supporting hyperplane.

Theorem 8.1.3. *Suppose that the function $f\colon X \to \overline{\mathbf{R}}$ is differentiable at the point $a \in \operatorname{dom} f$. Then either $\partial f(a) = \{f'(a)\}$ or $\partial f(a) = \emptyset$.*

Proof. Suppose $c \in \partial f(a)$. By the differentiability definition,

$$f(a+v) - f(a) = \langle f'(a), v\rangle + r(v)$$

with a remainder term $r(v)$ satisfying the condition

$$\lim_{v\to 0} \frac{r(v)}{\|v\|} = 0,$$

and by the subgradient definition, $f(a+v) - f(a) \geq \langle c, v\rangle$ for all v such that $a+v$ belongs to X. Consequently,

$$\frac{\langle c, v\rangle}{\|v\|} \leq \frac{\langle f'(a), v\rangle + r(v)}{\|v\|} \tag{8.2}$$

for all v with a sufficiently small norm $\|v\|$.

Let $\mathbf{e}_j$ be the j:th unit vector. Then $\langle c, \mathbf{e}_j\rangle = c_j$ and $\langle f'(a), \mathbf{e}_j\rangle = \dfrac{\partial f}{\partial x_j}(a)$, so by choosing $v = t\mathbf{e}_j$ in inequality (8.2), noting that $\|t\mathbf{e}_j\| = |t|$, and letting $t \to 0$ from the right and from the left, respectively, we obtain the following two inequalities

$$c_j \leq \frac{\partial f}{\partial x_j}(a) \quad \text{and} \quad -c_j \leq -\frac{\partial f}{\partial x_j}(a),$$

which imply that $c_j = \dfrac{\partial f}{\partial x_j}(a)$. Hence, $c = f'(a)$, and this proves the inclusion $\partial f(a) \subseteq \{f'(a)\}$. □

We can now reformulate Theorem 7.2.1 as follows: A differentiable function with a convex domain is convex if and only if it has a subgradient (which is then equal to the derivative) everywhere. Our next theorem generalizes this result.

Theorem 8.1.4. *Let $f\colon X \to \overline{\mathbf{R}}$ be a function with a convex domain X.*

(a) If $\operatorname{dom} f$ is a convex set and $\partial f(x) \neq \emptyset$ for all $x \in \operatorname{dom} f$, then f is a convex function.

(b) If f is a convex function, then $\partial f(x) \neq \emptyset$ for all $x \in \operatorname{rint}(\operatorname{dom} f)$.

Proof. (a) Let x and y be two arbitrary points in $\operatorname{dom} f$ and consider the point $z = \lambda x + (1-\lambda)y$, where $0 < \lambda < 1$. By assumption, f has a subgradient c at the point z. Using the inequality (8.1) at the point $a = z$ twice, one time with x replaced by y, we obtain the inequality

$$\begin{aligned}\lambda f(x) + (1-\lambda)f(y) &\geq \lambda\big(f(z) + \langle c, x - z\rangle\big) + (1-\lambda)\big(f(z) + \langle c, y - z\rangle\big)\\ &= f(z) + \langle c, \lambda x + (1-\lambda)y - z\rangle = f(z) + \langle c, 0\rangle = f(z),\end{aligned}$$

which shows that the restriction of f to $\operatorname{dom} f$ is a convex function, and this implies that f itself is convex.

(b) Conversely, assume that f is a convex function, and let a be a point in $\operatorname{rint}(\operatorname{dom} f)$. We will prove that the subdifferential $\partial f(a)$ is nonempty.

The point $(a, f(a))$ is a relative boundary point of the convex set $\operatorname{epi} f$. Therefore, there exists a supporting hyperplane

$$H = \{(x, x_{n+1}) \in \mathbf{R}^n \times \mathbf{R} \mid \langle c, x - a\rangle + c_{n+1}(x_{n+1} - f(a)) = 0\}$$

of $\operatorname{epi} f$ at the point $(a, f(a))$, and we may choose the normal vector (c, c_{n+1}) in such a way that

$$\langle c, x - a\rangle + c_{n+1}(x_{n+1} - f(a)) \geq 0 \tag{8.3}$$

for all points $(x, x_{n+1}) \in \operatorname{epi} f$. We shall see that this implies that $c_{n+1} > 0$.

By applying inequality (8.3) to the point $(a, f(a)+1)$ in the epigraph, we first obtain the inequality $c_{n+1} \geq 0$.

Now suppose that $c_{n+1} = 0$, and put $L = \operatorname{aff}(\operatorname{dom} f)$. Since $\operatorname{epi} f \subseteq L \times \mathbf{R}$ and the supporting hyperplane $H = \{(x, x_{n+1}) \in \mathbf{R}^n \times \mathbf{R} \mid \langle c, x - a\rangle = 0\}$ by definition does not contain $\operatorname{epi} f$ as a subset, it does not contain $L \times \mathbf{R}$ either. We conclude that there exists a point $y \in L$ such that $\langle c, y - a\rangle \neq 0$. Consider the points $y_\lambda = (1-\lambda)a + \lambda y$ for $\lambda \in \mathbf{R}$; these points lie in the afine set L, and $y_\lambda \to a$ as $\lambda \to 0$. Since a is a point in the relative interior of $\operatorname{dom} f$, the points y_λ lie in $\operatorname{dom} f$ if $|\lambda|$ is sufficiently small, and this implies that the inequality (8.3) can not hold for all points $(y_\lambda, f(y_\lambda))$ in the epigraph, because the expression $\langle c, y_\lambda - a\rangle$ $(= \lambda\langle c, y - a\rangle)$ assumes both positive and negative values depending on the sign of λ.

This is a contradiction and proves that $c_{n+1} > 0$, and by dividing inequality (8.3) by c_{n+1} and letting $d = -(1/c_{n+1})c$, we obtain the inequality

$$x_{n+1} \geq f(a) + \langle d, x - a\rangle$$

for all $(x, x_{n+1}) \in \operatorname{epi} f$. In particular, $f(x) \geq f(a) + \langle d, x - a\rangle$ for all $x \in \operatorname{dom} f$, which means that d is a subgradient of f at a. □

It follows from Theorem 8.1.4 that a real-valued function f with an open convex domain X is convex if and only if $\partial f(x) \neq \emptyset$ for all $x \in X$.

Theorem 8.1.5. *The subdifferential $\partial f(a)$ of a convex function f is a compact nonempty set if a is an interior point of* $\operatorname{dom} f$.

Proof. Suppose a is a point in $\operatorname{int}(\operatorname{dom} f)$. The subdifferential $\partial f(a)$ is closed by Theorem 8.1.1 and nonempty by Theorem 8.1.4, so it only remains to prove that it is a bounded set.

Theorem 6.6.1 yields two positive constants M and δ such that the closed ball $\overline{B}(a;\delta)$ lies in $\operatorname{dom} f$ and

$$|f(x) - f(a)| \leq M\|x - a\| \quad \text{for } x \in \overline{B}(a;\delta).$$

Now suppose that $c \in \partial f(a)$ and that $c \neq 0$. By choosing $x = a + \delta c/\|c\|$ in inequality (8.1), we conclude that

$$\delta\|c\| = \langle c, x - a\rangle \leq f(x) - f(a) \leq M\|x - a\| = \delta M$$

with the bound $\|c\| \leq M$ as a consequence. The subdifferential $\partial f(a)$ is thus included in the closed ball $\overline{B}(0;M)$. □

Theorem 8.1.6. *The sublevel sets of a strongly convex function $f\colon X \to \overline{\mathbf{R}}$ are bounded sets.*

Proof. Suppose that f is μ-strongly convex. Let x_0 be a point in the relative interior of $\operatorname{dom} f$, and let c be a subgradient at the point x_0 of the convex function $g(x) = f(x) - \frac{1}{2}\mu\|x\|^2$. Then, for each x belonging to the sublevel set $S = \{x \in X \mid f(x) \leq \alpha\}$,

$$\begin{aligned}\alpha \geq f(x) &= g(x) + \tfrac{1}{2}\mu\|x\|^2 \geq g(x_0) + \langle c, x - x_0\rangle + \tfrac{1}{2}\mu\|x\|^2\\ &= f(x_0) - \tfrac{1}{2}\mu\|x_0\|^2 + \langle c, x - x_0\rangle + \tfrac{1}{2}\mu\|x\|^2\\ &= f(x_0) + \tfrac{1}{2}\mu\big(\|x + \mu^{-1}c\|^2 - \|x_0 + \mu^{-1}c\|^2\big),\end{aligned}$$

which implies that

$$\|x + \mu^{-1}c\|^2 \leq \|x_0 + \mu^{-1}c\|^2 + 2\mu^{-1}(\alpha - f(x_0)).$$

The sublevel set S is thus included in a closed ball with center at the point $-\mu^{-1}c$ and radius $R = \sqrt{\|x_0 + \mu^{-1}c\|^2 + 2\mu^{-1}(\alpha - f(x_0))}$. □

Corollary 8.1.7. *If a continuous and strongly convex function has a nonempty closed sublevel set, then it has a unique minimum point.*

In particular, every strongly convex function $f\colon \mathbf{R}^n \to \mathbf{R}$ has a unique minimum point.

Proof. Let f be a continuous, strongly convex function with a nonempty closed sublevel set S. Then S is compact by the previous theorem, so the restriction of f to S assumes a minimum at some point in S, and this point is obviously a global minimum point of f. The minimum point is unique, because strongly convex functions are strictly convex.

A convex function $f\colon \mathbf{R}^n \to \mathbf{R}$ is automatically continuous, and continuous functions on $\mathbf{R}^n$ are closed. Hence, all sublevel sets of a strongly convex function on $\mathbf{R}^n$ are closed, so it follows from the already proven part of the theorem that there is a unique minimum point. □

8.2 Closed convex functions

In this section, we will use the subdifferential to supplement the results on closed convex functions in chapter 6.8 with some new results. We begin with an alternative characterization of closed convex functions.

Theorem 8.2.1. *A convex function $f\colon X \to \overline{\mathbf{R}}$ is closed if and only if, for all convergent sequences $(x_k)_1^\infty$ of points in* dom f *with limit x_0,*

$$\varliminf_{k\to\infty} f(x_k) \begin{cases} \geq f(x_0) & \text{if } x_0 \in \operatorname{dom} f, \\ = +\infty & \text{if } x_0 \in \operatorname{cl}(\operatorname{dom} f) \setminus \operatorname{dom} f. \end{cases} \tag{8.4}$$

Proof. Suppose that f is closed, i.e. that epi f is a closed set, and let $(x_k)_1^\infty$ be a sequence in dom f which converges to a point $x_0 \in \operatorname{cl}(\operatorname{dom} f)$, and put

$$L = \varliminf_{k\to\infty} f(x_k).$$

Let a be an arbitrary point in the relative interior of dom f and let c be a subgradient of f at the point a. Then $f(x_k) \geq f(a) + \langle c, x_k - a\rangle$ for all k, and since the right hand side converges (to $f(a) + \langle c, x_0 - a\rangle$) as $k \to \infty$, we conclude that the sequence $(f(x_k))_1^\infty$ is bounded below. Its least limit point, i.e. L, is therefore a real number or $+\infty$.

Inequality (8.4) is trivially satisfied if $L = +\infty$, so assume that L is a finite number, and let $(x_{k_j})_{j=1}^\infty$ be a subsequence of the given sequence with the property that $f(x_{k_j}) \to L$ as $j \to \infty$. The points $(x_{k_j}, f(x_{k_j}))$, which belong to epi f, then converge to the point (x_0, L), and since the epigraph is assumed to be closed, we conclude that the limit point (x_0, L) belongs to the epigraph, i.e. $x_0 \in \operatorname{dom} f$ and $L \geq f(x_0)$.

So if x_0 does not belong to dom f but to $\operatorname{cl}(\operatorname{dom} f) \setminus \operatorname{dom} f$, then we must have $L = +\infty$. This proves that (8.4) holds.

Conversely, suppose (8.4) holds for all convergent sequences, and let $((x_k.t_k))_1^\infty$ be a sequence of points in epi f which converges to a point (x_0, t_0). Then, $(x_k)_1^\infty$ converges to x_0 and $(t_k)_1^\infty$ converges to t_0, and since $f(x_k) \le t_k$ for all k, we conclude that

$$\underline{\lim}_{k\to\infty} f(x_k) \le \underline{\lim}_{k\to\infty} t_k = t_0.$$

In particular, $\underline{\lim}_{k\to\infty} f(x_k) < +\infty$, so it follows from inequality (8.4) that $x_0 \in \operatorname{dom} f$ and that $f(x_0) \le t_0$. This means that the limit point (x_0, t_0) belongs to epi f. Hence, epi f contains all its boundary points and is therefore a closed set, and this means that f is a closed function. □

Corollary 8.2.2. *Suppose that $f\colon X \to \overline{\mathbf{R}}$ is a convex function and that its effective domain* dom f *is relative open. Then, f is closed if and only if $\lim_{k\to\infty} f(x_k) = +\infty$ for each sequence $(x_k)_1^\infty$ of points in* dom f *that converges to a relative boundary point of* dom f.

Proof. Since a convex function is continuous at all points in the relative interior of its effective domain, we conclude that $\lim_{k\to\infty} f(x_k) = f(x_0)$ for each sequence $(x_k)_1^\infty$ of points in dom f that converges to a point x_0 in dom f. Condition (8.4) of the previous theorem is therefore fulfilled if and only if $\lim_{k\to\infty} f(x_k) = +\infty$ for all sequences $(x_k)_1^\infty$ in dom f that converge to a point in rbdry(dom f). □

So a convex function with an affine set as effective domain is closed (and continuous), because affine sets lack relative boundary points.

Example 8.2.1. The convex function $f(x) = -\ln x$ with $\mathbf{R}_{++}$ as domain is closed, since $\lim\limits_{x\to 0} f(x) = +\infty$. □

Theorem 8.2.3. *If the function $f\colon X \to \overline{\mathbf{R}}$ is convex and closed, then*

$$f(x) = \lim_{\lambda\to 1^-} f(\lambda x + (1-\lambda)y)$$

for all $x, y \in$ dom f.

Proof. The inequality

$$\overline{\lim_{\lambda\to 1^-}} f(\lambda x + (1-\lambda)y) \le \overline{\lim_{\lambda\to 1^-}} (\lambda f(x) + (1-\lambda) f(y)) = f(x)$$

holds for all convex functions f, and the inequality

$$\underline{\lim}_{\lambda\to 1^-} f(\lambda x + (1-\lambda)y) \ge f(x)$$

holds for all closed convex functions f according to Theorem 8.2.1. □

Theorem 8.2.4. *Suppose that f and g are two closed convex functions, that*

$$\operatorname{rint}(\operatorname{dom} f) = \operatorname{rint}(\operatorname{dom} g)$$

and that

$$f(x) = g(x)$$

for all $x \in \operatorname{rint}(\operatorname{dom} f)$. Then $f = g$.

We remind the reader that the equality $f = g$ should be interpreted as $\operatorname{dom} f = \operatorname{dom} g$ and $f(x) = g(x)$ for all points x in the common effective domain.

Proof. If $\operatorname{rint}(\operatorname{dom} f) = \emptyset$, then $\operatorname{dom} f = \operatorname{dom} g = \emptyset$, and there is nothing to prove, so suppose that x_0 is a point in $\operatorname{rint}(\operatorname{dom} f)$. Then, $\lambda x + (1-\lambda)x_0$ lies in $\operatorname{rint}(\operatorname{dom} f)$, too, for each $x \in \operatorname{dom} f$ and $0 < \lambda < 1$, and it follows from our assumptions and Theorem 8.2.3 that

$$g(x) = \lim_{\lambda \to 1^-} g(\lambda x + (1-\lambda)x_0) = \lim_{\lambda \to 1^-} f(\lambda x + (1-\lambda)x_0) = f(x).$$

Hence, $g(x) = f(x)$ for all $x \in \operatorname{dom} f$, and it follows that $\operatorname{dom} f \subseteq \operatorname{dom} g$. The converse inclusion holds by symmetry, so $\operatorname{dom} f = \operatorname{dom} g$. □

Theorem 8.2.5. *Let $f\colon X \to \overline{\mathbf{R}}$ and $g\colon Y \to \overline{\mathbf{R}}$ be two closed convex functions with $X \cap Y \neq \emptyset$. The sum $f + g\colon X \cap Y \to \overline{\mathbf{R}}$ is then a closed convex function.*

Proof. The theorem follows from the characterization of closedness in Theorem 8.2.1. Let $(x_k)_1^\infty$ be a convergent sequence of points in $\operatorname{dom}(f+g)$ with limit point x_0. If x_0 belongs to $\operatorname{dom}(f+g)$ $(= \operatorname{dom} f \cap \operatorname{dom} g)$, then

$$\varliminf_{k\to\infty} (f(x_k) + g(x_k)) \geq \varliminf_{k\to\infty} f(x_k) + \varliminf_{k\to\infty} g(x_k) \geq f(x_0) + g(x_0),$$

and if x_0 does not belong to $\operatorname{dom}(f+g)$, then we use the trivial inclusion

$$\operatorname{cl}(A \cap B) \setminus A \cap B \subseteq (\operatorname{cl} A \setminus A) \cup (\operatorname{cl} B \setminus B),$$

with $A = \operatorname{dom} f$ and $B = \operatorname{dom} g$, to conclude that the sum $f(x_k) + g(x_k)$ tends to $+\infty$, because one of the two sequences $(f(x_k))_1^\infty$ and $(g(x_k))_1^\infty$ tends to $+\infty$ while the other either tends to $+\infty$ or has a finite limes inferior. □

The closure

Definition. Let $f\colon X \to \overline{\mathbf{R}}$ be a function defined on a subset of $\mathbf{R}^n$ and define $(\operatorname{cl} f)(x)$ for $x \in \mathbf{R}^n$ by

$$(\operatorname{cl} f)(x) = \inf\{t \mid (x,t) \in \operatorname{cl}(\operatorname{epi} f)\}.$$

The function $\operatorname{cl} f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ is called the *closure of* f.

Theorem 8.2.6. *The closure* $\operatorname{cl} f$ *of a convex function* f*, whose effective domain is a nonempty subset of* $\mathbf{R}^n$*, has the following properties:*

(i) *The closure* $\operatorname{cl} f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ *is a convex function.*

(ii) $\operatorname{dom} f \subseteq \operatorname{dom}(\operatorname{cl} f) \subseteq \operatorname{cl}(\operatorname{dom} f)$.

(iii) $\operatorname{rint}(\operatorname{dom}(\operatorname{cl} f)) = \operatorname{rint}(\operatorname{dom} f)$.

(iv) $(\operatorname{cl} f)(x) \le f(x)$ *for all* $x \in \operatorname{dom} f$.

(v) $(\operatorname{cl} f)(x) = f(x)$ *for all* $x \in \operatorname{rint}(\operatorname{dom} f)$.

(vi) $\operatorname{epi}(\operatorname{cl} f) = \operatorname{cl}(\operatorname{epi} f)$.

Proof. (i) Let x_0 be an arbitrary point in $\operatorname{rint}(\operatorname{dom} f)$, and let c be a subgradient of f at the point x_0. Then $f(x) \ge f(x_0) + \langle c, x - x_0\rangle$ for all $x \in \operatorname{dom} f$, which means that the epigraph $\operatorname{epi} f$ is a subset of the closed set $K = \{(x,t) \in \operatorname{cl}(\operatorname{dom} f) \times \mathbf{R} \mid \langle c, x - x_0\rangle + f(x_0) \le t\}$. It follows that $\operatorname{cl}(\operatorname{epi} f) \subseteq K$, and hence

$$\begin{aligned}(\operatorname{cl} f)(x) &= \inf\{t \mid (x,t) \in \operatorname{cl}(\operatorname{epi} f)\} \ge \inf\{t \mid (x,t) \in K\}\\ &= f(x_0) + \langle c, x - x_0\rangle > -\infty\end{aligned}$$

for all $x \in \mathbf{R}^n$. So $\overline{\mathbf{R}}$ is a codomain of the function $\operatorname{cl} f$, and since $\operatorname{cl}(\operatorname{epi} f)$ is a convex set, it now follows from Theorem 6.2.6 that $\operatorname{cl} f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ is a convex function.

(ii), (iv) and (v) It follows from the inclusion $\operatorname{epi} f \subseteq \operatorname{cl}(\operatorname{epi} f) \subseteq K$ that

$$(\operatorname{cl} f)(x) \begin{cases} \le \inf\{t \mid (x,t) \in \operatorname{epi} f\} = f(x) < +\infty & \text{if } x \in \operatorname{dom} f,\\ \ge \inf\{t \mid (x,t) \in K\} = \inf \emptyset = +\infty & \text{if } x \notin \operatorname{cl}(\operatorname{dom} f),\end{cases}$$

$$(\operatorname{cl} f)(x_0) \ge \inf\{t \mid (x_0,t) \in K\} = f(x_0).$$

This proves that $\operatorname{dom} f \subseteq \operatorname{dom}(\operatorname{cl} f) \subseteq \operatorname{cl}(\operatorname{dom} f)$, that $(\operatorname{cl} f)(x) \le f(x)$ for all $x \in \operatorname{dom} f$, and that $(\operatorname{cl} f)(x_0) = f(x_0)$, and since x_0 is an arbitrary point in $\operatorname{rint}(\operatorname{dom} f)$, we conclude that $(\operatorname{cl} f)(x) = f(x)$ for all $x \in \operatorname{rint}(\operatorname{dom} f)$.

(iii) Since $\operatorname{rint}(\operatorname{cl} X) = \operatorname{rint} X$ for arbitrary convex sets X, it follows in particular from (ii) that

$$\operatorname{rint}(\operatorname{dom} f) \subseteq \operatorname{rint}(\operatorname{dom}(\operatorname{cl} f)) \subseteq \operatorname{rint}(\operatorname{cl}(\operatorname{dom} f)) = \operatorname{rint}(\operatorname{dom} f),$$

with the conclusion that $\operatorname{rint}(\operatorname{dom}(\operatorname{cl} f)) = \operatorname{rint}(\operatorname{dom} f)$.

(vi) The implications

$$(x,t) \in \operatorname{cl}(\operatorname{epi} f) \Rightarrow (\operatorname{cl} f)(x) \leq t \Rightarrow (x,t) \in \operatorname{epi}(\operatorname{cl} f)$$

follow immediately from the closure and epigraph definitions. Conversely, suppose that (x,t) is a point in $\operatorname{epi}(\operatorname{cl} f)$, i.e. that $(\operatorname{cl} f)(x) \leq t$, and let $U \times I$ be an open neighborhood of (x,t). The neighborhood I of t contains a number s such that $(x,s) \in \operatorname{cl}(\operatorname{epi} f)$, and since $U \times I$ is also an open neighborhood of (x,s), it follows that $\operatorname{epi} f \cap (U \times I) \neq \emptyset$. This proves that $(x,t) \in \operatorname{cl}(\operatorname{epi} f)$, so we have the implication

$$(x,t) \in \operatorname{epi}(\operatorname{cl} f) \Rightarrow (x,t) \in \operatorname{cl}(\operatorname{epi} f).$$

Thus, $\operatorname{epi}(\operatorname{cl} f) = \operatorname{cl}(\operatorname{epi} f)$. □

Theorem 8.2.7. *If f is a closed convex function, then $\operatorname{cl} f = f$.*

Proof. We have $\operatorname{rint}(\operatorname{dom}(\operatorname{cl} f)) = \operatorname{rint}(\operatorname{dom} f)$ and $(\operatorname{cl} f)(x) = f(x)$ for all $x \in \operatorname{rint}(\operatorname{dom} f)$, by the previous theorem. Therefore it follows from Theorem 8.2.4 that $\operatorname{cl} f = f$. □

8.3 The conjugate function

Definition. Let $f\colon X \to \overline{\mathbf{R}}$ be an arbitrary function defined on a subset X of $\mathbf{R}^n$ and define a function f^* on $\mathbf{R}^n$ by

$$f^*(y) = \sup\{\langle y, x\rangle - f(x) \mid x \in X\}$$

for $y \in \mathbf{R}^n$. The function f^* is called the *conjugate function* or the *Fenchel transform* of f.

We use the shorter notation f^{**} for the conjugate function of f^*, i.e. $f^{**} = (f^*)^*$.

The conjugate function f^* of a function $f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ with a nonempty effective domain is obviously a function $\mathbf{R}^n \to \overline{\mathbf{R}}$, and

$$f^*(y) = \sup\{\langle y, x\rangle - f(x) \mid x \in \operatorname{dom} f\}.$$

There are two trivial cases: If the effective domain of $f\colon X \to \overline{\mathbf{R}}$ is empty, then $f^*(y) = -\infty$ for all $y \in \mathbf{R}^n$, and if $f\colon X \to \overline{\mathbf{R}}$ assumes the value $-\infty$ at some point, then $f^*(y) = +\infty$ for all $y \in \mathbf{R}^n$.

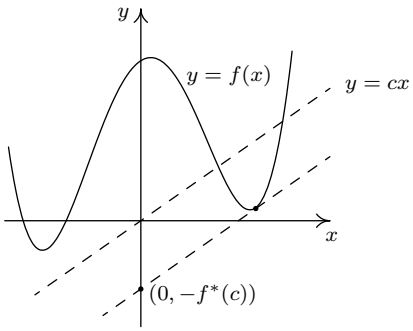

Figure 8.2. A graphical illustration of the conjugate function f^* when f is a one-variable function. The function value $f^*(c)$ is equal to the maximal vertical distance between the line $y = cx$ and the curve $y = f(x)$. If f is differentiable, then $f^*(c) = cx_0 - f(x_0)$ for some point x_0 with $f'(x_0) = c$.

EXAMPLE 8.3.1. The support functions that were defined in Section 6.9, are conjugate functions. To see this, define for a given subset A of $\mathbf{R}^n$ the function $\chi_A \colon \mathbf{R}^n \to \overline{\mathbf{R}}$ by

$$\chi_A(x) = \begin{cases} 0 & \text{if } x \in A, \\ +\infty & \text{if } x \notin A. \end{cases}$$

The function χ_A is called the *indicator function* of the set A, and it is a convex function if A is a convex set. Obviously,

$$\chi_A^*(y) = \sup\{\langle y, x\rangle \mid x \in A\} = S_A(y)$$

for all $y \in \mathbf{R}^n$, so the support function of A coincides with the conjugate function χ_A^* of the indicator function of A. □

We are primarily interested in conjugate functions of convex functions $f \colon X \to \overline{\mathbf{R}}$, but we start with some general results.

Theorem 8.3.1. *The conjugate function f^* of a function $f \colon X \to \overline{\mathbf{R}}$ with a nonempty effective domain is convex and closed.*

Proof. The epigraph $\operatorname{epi} f^*$ consists of all points $(y, t) \in \mathbf{R}^n \times \mathbf{R}$ that satisfy the inequalities $\langle x, y\rangle - t \le f(x)$ for all $x \in \operatorname{dom} f$, which means that it is the intersection of a family of closed halfspaces in $\mathbf{R}^n \times \mathbf{R}$. Hence, $\operatorname{epi} f^*$ is a closed convex set, so the conjugate function f^* is closed and convex. □

Theorem 8.3.2 (Fenchel's inequality)**.** *Let $f \colon X \to \overline{\mathbf{R}}$ be a function with a nonempty effective domain. Then*

$$\langle x, y\rangle \le f(x) + f^*(y)$$

for all $x \in X$ and all $y \in \mathbf{R}^n$. Moreover, the two sides are equal for a given $x \in \operatorname{dom} f$ if and only if $y \in \partial f(x)$.

Proof. The inequality follows immediately from the definition of $f^*(y)$ as a least upper bound if $x \in \operatorname{dom} f$, and it is trivially true if $x \in X \setminus \operatorname{dom} f$. Moreover, by the subgradient definition, if $x \in \operatorname{dom} f$ then

$$\begin{aligned} y \in \partial f(x) &\Leftrightarrow f(z) - f(x) \geq \langle y, z - x\rangle \quad \text{for all } z \in \operatorname{dom} f \\ &\Leftrightarrow \langle y, z\rangle - f(z) \leq \langle y, x\rangle - f(x) \quad \text{for all } z \in \operatorname{dom} f \\ &\Leftrightarrow f^*(y) \leq \langle y, x\rangle - f(x) \\ &\Leftrightarrow f(x) + f^*(y) \leq \langle x, y\rangle, \end{aligned}$$

and by combining this with the already proven Fenchel inequality, we obtain the equivalence $y \in \partial f(x) \Leftrightarrow f(x) + f^*(y) = \langle x, y\rangle$. □

By the previous theorem, for all points y in the set $\bigcup\{\partial f(x) \mid x \in \operatorname{dom} f\}$

$$f^*(y) = \langle x_y, y\rangle - f(x_y),$$

where x_y is a point satisfying the condition $y \in \partial f(x_y)$. For differentiable functions f we obtain the points x_y as solutions to the equation $f'(x) = y$. Here follows a concrete example.

EXAMPLE 8.3.2. Let $f\colon]-1, \infty[\to \mathbf{R}$ be the function

$$f(x) = \begin{cases} -x(x+1)^{-1} & \text{if } -1 < x \leq 0, \\ 2x & \text{if } 0 \leq x < 1, \\ (x-2)^2 + 1 & \text{if } 1 \leq x < 2, \\ 2x - 3 & \text{if } x \geq 2. \end{cases}$$

Its graph is shown in the left part of Figure 8.3.

A look at the figure shows that the curve $y = f(x)$ lies above all lines that are tangent to the curve at a point (x, y) with $-1 < x < 0$, lies above all lines through the origin with a slope between $f'_-(0) = -1$ and the slope of the chord that connects the origin and the point $(2, 1)$ on the curve, and lies above all lines through the point $(2, 1)$ with a slope between $\frac{1}{2}$ and $f'_+(2) = 2$. This means that

$$\begin{aligned} \bigcup\{\partial f(x)\} &= \bigcup_{-1<x<0} \{f'(x)\} \cup \partial f(0) \cup \partial f(2) \cup \bigcup_{x>2} \{f'(x)\} \\ &=]-\infty, -1[\cup [-1, \tfrac{1}{2}] \cup [\tfrac{1}{2}, 2] \cup \{2\} =]-\infty, 2]. \end{aligned}$$

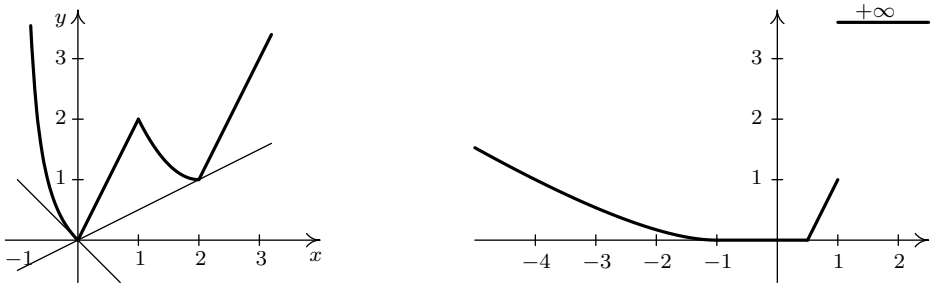

Figure 8.3. To the left the graph of the function $f\colon]-1,\infty[\to \mathbf{R}$, and to the right the graph of the conjugate function $f^*\colon \mathbf{R}\to\overline{\mathbf{R}}$.

The equation $f'(x)=c$ has for $c<-1$ the solution $x=-1+\sqrt{-1/c}$ in the interval $-1<x<0$. Let

$$x_c=\begin{cases}-1+\sqrt{-1/c} & \text{if } c<-1,\\ 0 & \text{if } -1\le c\le \tfrac12,\\ 2 & \text{if } \tfrac12\le c\le 2.\end{cases}$$

Then $c\in\partial f(x_c)$, and it follows from the remark after Theorem 8.3.2 that

$$f^*(c)=cx_c-f(x_c)=\begin{cases}-c-2\sqrt{-c}+1 & \text{if } c<-1,\\ 0 & \text{if } -1\le c\le \tfrac12,\\ 2c-1 & \text{if } \tfrac12\le c\le 2.\end{cases}$$

Since

$$f^*(c)=\sup_{x>-1}\{cx-f(x)\}\ge\sup_{x\ge2}\{cx-f(x)\}=\sup_{x\ge2}\{(c-2)x+3\}=+\infty$$

if $c>2$, we conclude that $\operatorname{dom} f^*=]-\infty,2]$. The graph of f^* is shown in the right part of Figure 8.3. □

Theorem 8.3.3. *Let $f\colon X\to\overline{\mathbf{R}}$ be an arbitrary function. Then*

$$f^{**}(x)\le f(x)$$

for all $x\in X$. Furthermore, $f^{**}(x)=f(x)$ if $x\in\operatorname{dom} f$ and $\partial f(x)\neq\emptyset$.

Proof. If $f(x)=+\infty$ for all $x\in X$, then $f^*\equiv-\infty$ and $f^{**}\equiv+\infty$, according to the remarks following the definition of the conjugate function, so the inequality holds with equality for all $x\in X$ in this trivial case.

Suppose, therefore, that $\operatorname{dom} f \neq \emptyset$. Then $\langle x, y\rangle - f^*(y) \leq f(x)$ for all $x \in X$ and all $y \in \operatorname{dom} f^*$ because of Fenchel's inequality, and hence $f^{**}(x) = \sup\{\langle x, y\rangle - f^*(y) \mid y \in \operatorname{dom} f^*\} \leq f(x)$.

If $\partial f(x) \neq \emptyset$, then Fenchel's inequality holds with equality for $y \in \partial f(x)$. This means that $f(x) = \langle x, y\rangle - f^*(y) \leq f^{**}(x)$ and implies that $f(x) = f^{**}(x)$. □

The following corollary follows immediately from Theorem 8.3.3, because convex functions have subgradients at all relative interior points of their effective domains.

Corollary 8.3.4. *If $f\colon X \to \overline{\mathbf{R}}$ is a convex function, then $f^{**}(x) = f(x)$ for all x in the relative interior of* $\operatorname{dom} f$.

We will prove that $f^{**} = \operatorname{cl} f$ if f is a convex function, and for this purpose we need the following lemma.

Lemma 8.3.5. *Suppose that f is a convex function and that (x_0, t_0) is a point in $\mathbf{R}^n \times \mathbf{R}$ which does not belong to* $\operatorname{cl}(\operatorname{epi} f)$. *Then there exist a vector $c \in \mathbf{R}^n$ and a real number d such that the "non-vertical" hyperplane*

$$H = \{(x, x_{n+1}) \mid x_{n+1} = \langle c, x\rangle + d\}$$

strictly separates the point (x_0, t_0) from $\operatorname{cl}(\operatorname{epi} f)$.

Proof. By the Separation Theorem 3.1.3, there exists a hyperplane

$$H = \{(x, x_{n+1}) \mid c_{n+1}x_{n+1} = \langle c, x\rangle + d\}$$

which strictly separates the point from $\operatorname{cl}(\operatorname{epi} f)$. If $c_{n+1} \neq 0$, then we can without loss of generality assume that $c_{n+1} = 1$, and there is nothing more to prove. So assume that $c_{n+1} = 0$, and choose the signs of c and d so that $\langle c, x_0\rangle + d > 0$ and $\langle c, x\rangle + d < 0$ for all $x \in \operatorname{dom} f$.

Using the subgradient c' at some point in the relative interior of $\operatorname{dom} f$ we obtain an affine function $\langle c', x\rangle + d'$ such that $f(x) \geq \langle c', x\rangle + d'$ for all $x \in \operatorname{dom} f$. This implies that

$$f(x) \geq \langle c', x\rangle + d' + \lambda(\langle c, x\rangle + d) = \langle c' + \lambda c, x\rangle + d' + \lambda d$$

for all $x \in \operatorname{dom} f$ and all positive numbers λ, while

$$\langle c' + \lambda c, x_0\rangle + d' + \lambda d = \langle c', x_0\rangle + d' + \lambda(\langle c, x_0\rangle + d) > t_0$$

for all sufficiently large numbers λ. So the epigraph $\operatorname{epi} f$ lies above the hyperplane

$$H_\lambda = \{(x, x_{n+1}) \mid x_{n+1} = \langle c' + \lambda c, x\rangle + d' + \lambda d\}.$$

and the point (x_0, t_0) lies strictly below the same hyperplane, if the number λ is big enough. By moving the hyperplane H_λ slightly downwards, we obtain a parallel non-vertical hyperplane which strictly separates (x_0, t_0) and $\mathrm{cl}(\mathrm{epi}\, f)$. □

Lemma 8.3.6. *If $f\colon X \to \overline{\mathbf{R}}$ is a convex function, then*

$$\mathrm{rint}(\mathrm{dom}\, f^{**}) = \mathrm{rint}(\mathrm{dom}\, f).$$

Proof. Since $\mathrm{rint}(\mathrm{dom}\, f) = \mathrm{rint}(\mathrm{cl}(\mathrm{dom}\, f))$, it suffices to prove the inclusion

$$\mathrm{dom}\, f \subseteq \mathrm{dom}\, f^{**} \subseteq \mathrm{cl}(\mathrm{dom}\, f).$$

The left inclusion follows immediately from the inequality in Theorem 8.3.3. To prove the right inclusion, we assume that $x_0 \notin \mathrm{cl}(\mathrm{dom}\, f)$ and shall prove that this implies that $x_0 \notin \mathrm{dom}\, f^{**}$.

It follows from our assumption that the points (x_0, t_0) do not belong to $\mathrm{cl}(\mathrm{epi}\, f)$ for any number t_0. Thus, given $t_0 \in \mathbf{R}$ there exists, by the previous lemma, a hyperplane $H = \{(x, x_{n+1}) \in \mathbf{R}^n \times \mathbf{R} \mid x_{n+1} = \langle c, x\rangle + d\}$ which strictly separates (x_0, t_0) and $\mathrm{cl}(\mathrm{epi}\, f)$. Hence, $t_0 < \langle c, x_0\rangle + d$ and $\langle c, x\rangle + d < f(x)$ for all $x \in \mathrm{dom}\, f$. Consequently,

$$-d \geq \sup\{\langle c, x\rangle - f(x) \mid x \in \mathrm{dom}\, f\} = f^*(c),$$

and hence

$$t_0 < \langle c, x_0\rangle + d \leq \langle c, x_0\rangle - f^*(c) \leq f^{**}(x_0).$$

Since this holds for all real numbers t_0, it follows that $f^{**}(t_0) = +\infty$, which means that $x_0 \notin \mathrm{dom}\, f^{**}$. □

Theorem 8.3.7. *If f is a convex function, then $f^{**} = \mathrm{cl}\, f$.*

Proof. It follows from Lemma 8.3.6 and Theorem 8.2.6 (iii) that

$$\mathrm{rint}(\mathrm{dom}\, f^{**}) = \mathrm{rint}(\mathrm{dom}(\mathrm{cl}\, f)),$$

and from Theorem 8.3.4 and Theorem 8.2.6 (v) that

$$f^{**}(x) = (\mathrm{cl}\, f)(x)$$

for all $x \in \mathrm{rint}(\mathrm{dom}\, f^{**})$. So the two functions f^{**} and $\mathrm{cl}\, f$ are equal, according to Theorem 8.2.4, because both of them are closed and convex . □

Corollary 8.3.8. *If f is a closed convex function, then $f^{**} = f$.*

8.4 The direction derivative

Definition. Suppose the function $f\colon X \to \mathbf{R}$ is defined in a neighborhood of x, and let v be an arbitrary vector in $\mathbf{R}^n$. The limit

$$f'(x;v) = \lim_{t\to 0+} \frac{f(x+tv) - f(x)}{t},$$

provided it exists, is called the *direction derivative* of f at the point x in the direction v.

If f is differentiable at x, then obviously $f'(x;v) = Df(x)[v]$.

Example 8.4.1. If f is a one-variable function, then

$$f'(x;v) = \begin{cases} f'_+(x)v & \text{if } v > 0, \\ 0 & \text{if } v = 0, \\ f'_-(v)v & \text{if } v < 0. \end{cases}$$

So, the direction derivative is a generalization of left- and right derivatives. □

Theorem 8.4.1. *Let $f\colon X \to \mathbf{R}$ be a convex function with an open domain. The direction derivative $f'(x;v)$ exists for all $x \in X$ and all directions v, and*

$$f(x+v) \geq f(x) + f'(x;v)$$

if $x+v$ lies in X.

Proof. Let $\phi(t) = f(x+tv)$; then $f'(x;v) = \phi'_+(0)$, which exists since convex one-variable functions do have right derivatives at each point by Theorem 7.1.2. Moreover,

$$\phi(t) \geq \phi(0) + \phi'_+(0)\, t$$

for all t in the domain of ϕ, and we obtain the inequality of the theorem by choosing $t = 1$. □

Theorem 8.4.2. *The direction derivative $f'(x;v)$ of a convex function is a positively homogeneous and convex function of the second variable v, i.e.*

$$f'(x;\alpha v) = \alpha f'(x;v) \quad \text{if } \alpha \geq 0$$
$$f'(x;\alpha v + (1-\alpha)w) \leq \alpha f'(x;v) + (1-\alpha) f'(x;w) \quad \text{if } 0 \leq \alpha \leq 1.$$

Proof. The homogenouity follows directly from the definition (and holds for arbitrary functions). Moreover, for convex functions f

$$\begin{aligned} f(x+t(\alpha v+(1-\alpha)w))-f(x) &= f(\alpha(x+tv)+(1-\alpha)(x+tw))-f(x)\\ &\le \alpha\big(f(x+tv)-f(x)\big)+(1-\alpha)\big(f(x+tw)-f(x)\big). \end{aligned}$$

The required convexity inequality is now obtained after division by $t>0$ by passing to the limit as $t\to 0+$. □

Theorem 7.1.2 gives a relation between the subgradient and the direction derivative for convex one-variable functions f – the number c is a subgradient at x if and only if $f'_-(x)\le c\le f'_+(x)$. The subdifferential $\partial f(x)$ is in other words equal to the interval $[f'_-(x), f'_+(x)]$.

We may express this relation using the support function of the subdifferential. Let us recall that the support function S_X of a set X in $\mathbf{R}^n$ is defined as

$$S_X(x)=\sup\{\langle y,x\rangle \mid y\in X\}.$$

For one-variable functions f this means that

$$\begin{aligned} S_{\partial f(x)}(v)=S_{[f'_-(x),f'_+(x)]}(v)&=\max\{f'_+(x)v, f'_-(x)v\}=\begin{cases} f'_+(x)v & \text{if } v>0,\\ 0 & \text{if } v=0,\\ f'_-(x)v & \text{if } v<0 \end{cases}\\ &= f'(x;v). \end{aligned}$$

We will generalize this identity, and to achieve this we need to consider the subgradients of the function $v\mapsto f'(x;v)$. We denote the subdifferential of this function at the point v_0 by $\partial_2 f'(x;v_0)$.

If the function $f\colon X\to\mathbf{R}$ is convex, then so is the function $v\mapsto f'(x;v)$, according to our previous theorem, and the subdifferentials $\partial_2 f'(x;v)$ are thus nonempty sets for all $x\in X$ and all $v\in\mathbf{R}^n$.

Lemma 8.4.3. *Let $f\colon X\to\mathbf{R}$ be a convex function with an open domain X and let x be a point in X. Then:*

(i) $\quad c\in\partial_2 f'(x;0)\Leftrightarrow f'(x;v)\ge\langle c,v\rangle$ *for all* $v\in\mathbf{R}^n$

(ii) $\quad \partial_2 f'(x;v)\subseteq\partial_2 f'(x;0)$ *for all* $v\in\mathbf{R}^n$

(iii) $\quad c\in\partial_2 f'(x;v)\Rightarrow f'(x;v)=\langle c,v\rangle$

(iv) $\quad \partial f(x)=\partial_2 f'(x;0).$

Proof. The equivalence (i) follows directly from the definition of the subgradient and the fact that $f'(x;0)=0$.

(ii) and (iii): Suppose $c \in \partial_2 f'(x;v)$ and let $w \in \mathbf{R}^n$ be an arbitrary vector. Then, by homogenouity and the definition of the subgradient, we have the following inequality for $t \geq 0$:

$$tf'(x;w) = f'(x;tw) \geq f'(x;v) + \langle c, tw - v\rangle = f'(x;v) + t\langle c, w\rangle - \langle c, v\rangle,$$

and this is possible for all $t > 0$ only if $f'(x;w) \geq \langle c, w\rangle$. So we conclude from (i) that $c \in \partial_2 f'(x;0)$, and this proves the inclusion (ii). By choosing $t = 0$ we obtain the inequality $f'(x;v) \leq \langle c, v\rangle$, which implies that $f'(x;v) = \langle c, v\rangle$, and completes the proof of the implication (iii).

(iv) Suppose $c \in \partial_2 f'(x;0)$. By (i) and Theorem 8.4.1,

$$f(y) \geq f(x) + f'(x;y-x) \geq f(x) + \langle c, y - x\rangle$$

for all $y \in X$, which proves that c is a subgradient of f at the point x and gives us the inclusion $\partial_2 f'(x;0) \subseteq \partial f(x)$.

Conversely, suppose $c \in \partial f(x)$. Then $f(x+tv) - f(x) \geq \langle c, tv\rangle = t\langle c, v\rangle$ for all sufficiently small numbers t. Division by $t > 0$ and passing to the limit as $t \to 0+$ results in the inequality $f'(x;v) \geq \langle c, v\rangle$, and it now follows from (i) that $c \in \partial_2 f'(x;0)$. This proves the inclusion $\partial f(x) \subseteq \partial_2 f'(x;0)$, and the proof is now complete. □

Theorem 8.4.4. *Suppose that $f\colon X \to \mathbf{R}$ is a convex function with an open domain. Then*

$$f'(x;v) = S_{\partial f(x)}(v) = \max\{\langle c, v\rangle \mid c \in \partial f(x)\}$$

for all $x \in X$ and all $v \in \mathbf{R}^n$.

Proof. It follows from (i) and (iv) in the preceding lemma that

$$\langle c, v\rangle \leq f'(x;v)$$

for all $c \in \partial f(x)$, and from (ii), (iii) and (iv) in the same lemma that $\langle c, v\rangle$ is equal to $f'(x;v)$ for all subgradients c in the nonempty subset $\partial_2 f'(x;v)$ of $\partial f(x)$. □

8.5 Subdifferentiation rules

Theorem 8.5.1. *Let X be an open convex set, and suppose that $f_i\colon X \to \mathbf{R}$ are convex functions and α_i are nonnegative numbers for $i = 1, 2 \ldots, m$. Define*

$$f = \sum_{i=1}^{m} \alpha_i f_i.$$

Then

$$\partial f(x) = \sum_{i=1}^{m} \alpha_i \partial f_i(x).$$

Proof. A sum of compact, convex sets is compact and convex. Therefore, $\sum_{i=1}^{m} \alpha_i \partial f_i(x)$ is a closed and convex set, just as the set $\partial f(x)$. Hence, by Theorem 6.9.2 it suffices to prove that the two sets have the same support function. And this follows from Theorems 8.4.4 and 6.9.1, according to which

$$S_{\partial f(x)}(v) = f'(x;v) = \sum_{i=1}^{m} \alpha_i f_i'(x;v) = \sum_{i=1}^{m} \alpha_i S_{\partial f_i}(x) = S_{\sum_{i=1}^{m} \alpha_i \partial f_i(x)}(v). \quad \square$$

Theorem 8.5.2. *Suppose that the functions $f_i \colon X \to \mathbf{R}$ are convex for $i = 1$, 2,..., m, and that their domain X is open, and let*

$$f = \max_{1 \le i \le m} f_i.$$

Then

$$\partial f(x) = \operatorname{cvx}\Big(\bigcup_{i \in I(x)} \partial f_i(x) \Big),$$

for all $x \in X$, where $I(x) = \{i \mid f_i(x) = f(x)\}$.

Proof. The functions f_i are continuous at x and $f_j(x) < f(x)$ for all $j \notin I(x)$. Hence, for all sufficiently small numbers t,

$$f(x+tv) - f(x) = \max_{i \in I(x)} f_i(x+tv) - f(x) = \max_{i \in I(x)} (f_i(x+tv) - f_i(x)),$$

and it follows after division by t and passing to the limit that

$$f'(x;v) = \max_{i \in I(x)} f_i'(x;v).$$

We use Theorem 6.9.1 to conclude that

$$\begin{aligned} S_{\partial f(x)}(v) = f'(x;v) &= \max_{i \in I(x)} f_i'(x;v) = \max_{i \in I(x)} S_{\partial f_i(x)}(v) = S_{\bigcup_{i \in I(x)} \partial f_i(x)}(v) \\ &= S_{\operatorname{cvx}(\bigcup_{i \in I(x)} \partial f_i(x))}(v), \end{aligned}$$

and the equality for $\partial f(x)$ is now a consequence of Theorem 6.9.2. $\square$

Our next theorem shows how to compute the subdifferential of a composition with affine functions.

Theorem 8.5.3. *Suppose C is a linear map from $\mathbf{R}^n$ to $\mathbf{R}^m$, that b is a vector in $\mathbf{R}^m$, and that g is a convex function with an open domain in $\mathbf{R}^m$, and let f be the function defined by $f(x) = g(Cx+b)$. Then, for each x in the domain of f,*

$$\partial f(x) = C^T(\partial g(Cx+b)).$$

Proof. The sets $\partial f(x)$ and $C^T(\partial g(Cx+b))$ are convex and compact, so it suffices to show that their support functions are identical. But for each $v \in \mathbf{R}^n$

$$\begin{aligned} f'(x;v) &= \lim_{t\to 0+} \frac{g(C(x+tv)+b) - g(Cx+b)}{t} \\ &= \lim_{t\to 0+} \frac{g(Cx+b+t\,Cv) - g(Cx+b)}{t} = g'(Cx+b;Cv), \end{aligned}$$

so it follows because of Theorem 6.9.1 that

$$S_{\partial f(x)}(v) = f'(x;v) = g'(Cx+b;Cv) = S_{\partial g(Cx+b)}(Cv) = S_{C^T(\partial g(Cx+b))}(v). \ \square$$

The Karush–Kuhn–Tucker theorem

As an application of the subdifferentiation rules we now prove a variant of a theorem by Karush–Kuhn–Tucker on minimization of convex functions with convex constraints. A more thorough treatment of this theme will be given in Chapters 10 and 11.

Theorem 8.5.4. *Suppose that the functions f, $g_1, g_2, \dots, g_m$ are convex and defined on an open convex set Ω, and let*

$$X = \{x \in \Omega \mid g_i(x) \le 0 \ \textit{for } i = 1,2,\dots,m.\}$$

Moreover, suppose that there exists a point $\overline{x} \in \Omega$ such that $g_i(\overline{x}) < 0$ for $i = 1,2,\dots,m$. (Slater's condition)

Then, $\hat{x} \in X$ is a minimum point of the restriction $f|_X$ if and only if for each $i = 1,2,\dots,m$ there exist a subgradient $c_i \in \partial g_i(\hat{x})$ and a scalar $\hat{\lambda}_i \in \mathbf{R}_+$ with the following properties:

(i) $$-\sum_{i=1}^m \hat{\lambda}_i c_i \in \partial f(\hat{x}) \qquad \textit{and}$$

(ii) $$\hat{\lambda}_i g_i(\hat{x}) = 0 \quad \textit{for } i = 1,2,\dots,m.$$

Remark. If the functions are differentiable, then condition (i) simplifies to

$$\nabla f(\hat{x}) + \sum_{i=1}^m \hat{\lambda}_i \nabla g_i(\hat{x}) = 0.$$

Cf. Theorem 11.2.1.

Proof. Let $\hat{x}$ be a point in X and consider the convex function

$$h(x) = \max\{f(x) - f(\hat{x}), g_1(x), \ldots, g_m(x)\}$$

with Ω as its domain. Clearly, $h(\hat{x}) = 0$. By defining

$$I(\hat{x}) = \{i \mid g_i(\hat{x}) = 0\},$$

we obtain $I(\hat{x}) = \{i \mid g_i(\hat{x}) = h(\hat{x})\}$, and it follows from Theorem 8.5.2 that

$$\partial h(\hat{x}) = \operatorname{cvx}\big(\partial f(\hat{x}) \cup \bigcup\{\partial g_i(\hat{x}) \mid i \in I(\hat{x})\}\big).$$

Now assume that $\hat{x}$ is a minimum point of the restriction $f|_X$. Then $h(x) = f(x) - f(\hat{x}) \geq 0$ for all $x \in X$ with equality when $x = \hat{x}$. And if $x \notin X$, then $h(x) > 0$ since $g_i(x) > 0$ for at least one i. Thus, $\hat{x}$ is a global minimum point of h.

Conversely, if $\hat{x}$ is a global minimum point of h, then $h(x) \geq 0$ for all $x \in \Omega$. In particular, for $x \in X$ this means that $h(x) = f(x) - f(\hat{x}) \geq 0$, and hence $\hat{x}$ is a mimimum point of the restriction $f|_X$, too.

Using Theorem 8.1.2 we therefore obtain the following equivalences:

$$\begin{aligned}
\hat{x} \text{ is a minimum point of } f|_X \quad &\Leftrightarrow \quad \hat{x} \text{ is a minimum point of } h \\
&\Leftrightarrow \quad 0 \in \partial h(\hat{x}) \\
&\Leftrightarrow \quad 0 \in \operatorname{cvx}\big(\partial f(\hat{x}) \cup \bigcup\{\partial g_i(\hat{x}) \mid i \in I(\hat{x})\}\big) \\
&\Leftrightarrow \quad 0 = \lambda_0 c_0 + \sum_{i \in I(\hat{x})} \lambda_i c_i \\
&\Leftrightarrow \quad \lambda_0 c_0 = -\sum_{i \in I(\hat{x})} \lambda_i c_i, \qquad (8.5)
\end{aligned}$$

where $c_0 \in \partial f(\hat{x})$, $c_i \in \partial g_i(\hat{x})$ for $i \in I(\hat{x})$, and the scalars λ_i are nonnegative numbers with sum equal to 1.

We now claim that $\lambda_0 > 0$. To prove this, assume the contrary. Then $\sum_{i \in I(\hat{x})} \lambda_i c_i = 0$, and it follows that

$$\sum_{i \in I(\hat{x})} \lambda_i g_i(\overline{x}) \geq \sum_{i \in I(\hat{x})} \lambda_i \big(g_i(\hat{x}) + \langle c_i, \overline{x} - \hat{x}\rangle\big) = \langle \sum_{i \in I(\hat{x})} \lambda_i c_i, \overline{x} - \hat{x}\rangle = 0,$$

which is a contradiction, since $g_i(\overline{x}) < 0$ for all i and $\lambda_i > 0$ for some $i \in I(\hat{x})$.

We may therefore divide the equality in (8.5) by λ_0, and conditions (i) and (ii) in our theorem are now fulfilled if we define $\hat{\lambda}_i = \lambda_i/\lambda_0$ for $i \in I(\hat{x})$, and $\hat{\lambda}_i = 0$ for $i \notin I(\hat{x})$, and choose arbitrary subgradients $c_i \in \partial g_i(\hat{x})$ for $i \notin I(\hat{x})$. □

Exercises

8.1 Suppose $f\colon \mathbf{R}^n \to \mathbf{R}$ is a strongly convex function. Prove that

$$\lim_{\|x\|\to\infty} f(x) = \infty$$

8.2 Find $\partial f(-1,1)$ for the function $f(x_1,x_2) = \max(|x_1|,|x_2|)$.

8.3 Determine the subdifferential $\partial f(0)$ at the origin for the following functions $f\colon \mathbf{R}^n \to \mathbf{R}$:

a) $f(x) = \|x\|_2$ b) $f(x) = \|x\|_\infty$ c) $f(x) = \|x\|_1$.

8.4 Determine the conjugate functions of the following functions:

a) $f(x) = ax + b,\ \operatorname{dom} f = \mathbf{R}$ b) $f(x) = -\ln x,\ \operatorname{dom} f = \mathbf{R}_{++}$
c) $f(x) = \mathrm{e}^x,\ \operatorname{dom} f = \mathbf{R}$ d) $f(x) = x\ln x,\ \operatorname{dom} f = \mathbf{R}_{++}$
e) $f(x) = 1/x,\ \operatorname{dom} f = \mathbf{R}_{++}$.

8.5 Use the relation between the support function S_A and the indicator function χ_A and the fact that $S_A = S_{\mathrm{cl}(\mathrm{cvx}\,A)}$ to prove Corollary 6.9.3, i.e. that

$$\mathrm{cl}(\mathrm{cvx}\,A) = \mathrm{cl}(\mathrm{cvx}\,B) \ \Leftrightarrow\ S_A = S_B.$$

Part II

Optimization – basic theory

Chapter 9

Optimization

The Latin word *optimum* means 'the best'. The optimal alternative among a number of different alternatives is the one that is the best in some way. Optimization is therefore, in a broad sense, the art of determining the best.

Optimization problems occur not only in different areas of human planning, but also many phenomena in nature can be explained by simple optimization principles. Examples are light propagation and refraction in different media, thermal conductivity and chemical equilibrium.

In everyday optimization problems, it is often difficult, if not impossible, to compare and evaluate different alternatives in a meaningful manner. We shall leave this difficulty aside, for it can not be solved by mathematical methods. Our starting point is that the alternatives are ranked by means of a function, for example a profit or cost function, and that the option that gives the maximum or minimum function value is the best one.

The problems we will address are thus purely mathematical − to minimize or maximize given functions over sets that are given by a number of constraints.

9.1 Optimization problems

Basic notions

For the problem of minimizing a function $f\colon \Omega \to \overline{\mathbf{R}}$ over a subset X of the domain Ω of the function, we use the notation

$$\begin{aligned} &\min \quad f(x) \\ &\text{s.t.} \quad x \in X. \end{aligned}$$

Here, s.t. is an abbreviation for the phrase *subject to the condition.*

The elements of the set X are called the *feasible points* or *feasible solutions* of the optimization problem. The function f is the *objective function*.

Observe that vi allow ∞ as a function value of the objective function in a minimization problem.

The *(optimal) value* $v_{\min}$ of the minimization problem is by definition

$$v_{\min} = \begin{cases} \inf\{f(x) \mid x \in X\} & \text{if } X \neq \emptyset, \\ \infty & \text{if } X = \emptyset. \end{cases}$$

The optimal value is thus a real number if the objective function is bounded below and not identically equal to ∞ on the set X, the value is $-\infty$ if the function is not bounded below on X, and the value is ∞ if the objective function is identically equal to ∞ on X or if $X = \emptyset$.

Of course, we will also study maximization problems, and the problem of maximizing a function $f\colon \Omega \to \underline{\mathbf{R}}$ over X will be written

$$\begin{aligned} &\max\ f(x) \\ &\text{s.t.}\quad x \in X. \end{aligned}$$

The *(optimal) value* $v_{\max}$ of the maximization problem is defined by

$$v_{\max} = \begin{cases} \sup\{f(x) \mid x \in X\} & \text{if } X \neq \emptyset, \\ -\infty & \text{if } X = \emptyset. \end{cases}$$

The optimal value of a minimization or maximization problem is in this way always defined as a real number, $-\infty$ or ∞, i.e. as an element of the extended real line $\overline{\underline{\mathbf{R}}}$. If the value is a real number, we say that the optimization problem has a *finite* value.

A feasible point x_0 for an optimization problem with objective function f is called an *optimal point* or *optimal solution* if the value of the problem is finite and equal to $f(x_0)$. An optimal solution of a minimization problem is, in other words, the same as a global minimum point. Of course, problems with finite optimal values need not necessarily have any optimal solutions.

From a mathematical point of view, there is no difference in principle between maximization problems and minimization problems, since the optimal values $v_{\max}$ and $v_{\min}$ of the problems

$$\begin{aligned} &\max\ f(x) \\ &\text{s.t.}\quad x \in X \end{aligned} \qquad \text{and} \qquad \begin{aligned} &\min\ -f(x), \\ &\text{s.t.}\quad x \in X \end{aligned}$$

respectively, are connected by the simple relation $v_{\max} = -v_{\min}$, and x_0 is a maximum point of f if and only if x_0 is a minimum point of $-f$. For this reason, we usually only formulate results for minimization problems.

Finally, a comment as to why we allow ∞ and $-\infty$ as function values of the objective functions as this seems to complicate matters. The most important reason is that sometimes we have to consider functions that are defined as pointwise suprema of an infinite family of functions, and the supremum function may assume infinite values even if all functions in the family assume only finite values. The alternative to allowing functions with values in the extended real line would be to restrict the domain of these supremum functions, and this is neither simpler nor more elegant.

General comments

There are some general and perhaps completely obvious comments that are relevant for many optimization problems.

Existence of feasible points

This point may seem trivial, for if a problem has no feasible points then there is not much more to be said. It should however be remembered that the set of feasible points is seldom given explicitly. Instead it is often defined by a system of equalities and inequalities, which may not be consistent. If the problem comes from the "real world", simplifications and defects in the mathematical model may lead to a mathematical problem that lacks feasible points.

Existence of optimal solutions

Needless to say, a prerequisite for the determination of the optimal solution of a problem is that there is one. Many theoretical results are of the form "If x_0 is an optimal solution, then x_0 satisfies these conditions." Although this usually restricts the number of potential candidates for optimal points, it does not prove the existence of such points.

From a practical point of view, however, the existence of an optimal solution − and its exact value, if such a solution exists − may not be that important. In many applications one is often satisfied with a feasible solutions that is good enough.

Uniqueness

Is the optimal solution, if such a solution exists, unique? The answer is probably of little interest for somebody looking for the solution of a practical problem − he or she should be satisfied by having found a best solution even if there are other solutions that are just as good. And if he or she would

consider one of the optimal solutions better than the others, then we can only conclude that the optimization problem is not properly set from the start, because the objective function apparently does not include everything that is required to sort out the best solution.

However, uniqueness of an optimal solution may sometimes lead to interesting properties that can be of use when looking for the solution.

Dependence on parameters and sensitivity

Sometimes, and in particular in problems that come directly from "reality", objective functions and constraints contain parameters, which are only given with a certain accuracy and, in the worst case, are more or less coarse estimates. In such cases, it is not sufficient to determine the optimal solution, but it is at least as important to know how the solution changes when parameters are changed. If a small perturbation of one parameter alters the optimal solution very much, there is reason to consider the solution with great skepticism.

Qualitative aspects

Of course, it is only for a small class of optimization problems that one can specify the optimum solution in exact form, or where the solution can be described by an algorithm that terminates after finitely many iterations. The mathematical solution to an optimization problem often consists of a number of necessary and/or sufficient conditions that the optimal solution must meet. At best, these can be the basis for useful numerical algorithms, and in other cases, they can perhaps only be used for qualitative statements about the optimal solutions, which however in many situations can be just as interesting.

Algorithms

There is of course no numerical algorithm that solves all optimization problems, even if we restrict ourselves to problems where the constraint set is defined by a a finite number of inequalities and equalities. However, there are very efficient numerical algorithms for certain subclasses of optimization problems, and many important applied optimization problems happen to belong to these classes. We shall study some algorithms of this type in Part III and Part IV of this book.

The development of good algorithms has been just as important as the computer development for the possibility of solving big optimization problems, and much of the algorithm development has occurred in recent decades.

9.2 Classification of optimization problems

To be able to say anything sensible about the minimization problem

$$\text{(P)} \qquad \begin{aligned} &\min \ f(x) \\ &\text{s.t.} \ \ x \in X \end{aligned}$$

we must make various assumptions about the objective function $f\colon \Omega \to \overline{\mathbf{R}}$ and about the set X of feasible points.

We will always assume that Ω is a subset of $\mathbf{R}^n$ and that the set X can be expressed as the solution set of a number of inequalities and equalities, i.e. that

$$X = \{x \in \Omega \mid g_1(x) \le 0, \ldots, g_p(x) \le 0, g_{p+1}(x) = 0, \ldots, g_m(x) = 0\}$$

where $g_1, g_2, \ldots, g_m$ are real valued functions defined on Ω.

We do not exclude the possibility that all constraints are equalities, i.e. that $p = 0$, or that all constraints are inequalities, i.e. that $p = m$, or that there are no constraints at all, i.e. that $m = 0$.

Since the equality $h(x) = 0$ can be replaced by the two inequalities $\pm h(x) \le 0$, we could without loss of generality assume that all constraints are inequalities, but it is convenient to formulate results for optimization problems with equalities among the constraints without first having to make such rewritings.

If $\hat{x}$ is a feasible point and $g_i(\hat{x}) = 0$, we say that the i:th constraint is *active* at the point $\hat{x}$. All constraints in the form of equalities are, of course, active at all feasible points.

The condition $x \in \Omega$ is (in the case $\Omega \neq \mathbf{R}^n$) of course also a kind of constraint, but it plays a different role than the other constraints. We will sometimes call it the *implicit* constraint in order to distinguish it from the other *explicit* constraints. If Ω is given as the solution set of a number of inequalities of type $h_i(x) \le 0$ and the functions h_i, the objective function and the explicit constraint functions are defined on the entire space $\mathbf{R}^n$, we can of course include the inequalities $h_i(x) \le 0$ among the explicit conditions and omit the implicit constraint.

The domain Ω will often be clear from the context, and it is in these cases not mentioned explicitly in the formulation of the optimization problem. The minimization problem (P) will therefore often be given in the following form

$$\begin{aligned} &\min \ f(x) \\ &\text{s.t.} \ \begin{cases} g_i(x) \le 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m. \end{cases} \end{aligned}$$

Linear programming

The problem of maximizing or minimizing a linear form over a polyhedron, which is given in the form of an intersection of closed halvspaces in $\mathbf{R}^n$, is called *linear programming*, abbreviated LP. The problem (P) is, in other words, an LP problem if the objective function f is linear and X is the set of solutions to a finite number of linear equalities and inequalities.

We will study LP problems in detail in Chapter 12.

Convex optimization

The minimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & \begin{cases} g_i(x) \leq 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m \end{cases} \end{array}$$

with implicit constraint $x \in \Omega$ is called *convex*, if the set Ω is convex, the objective function $f \colon \Omega \to \overline{\mathbf{R}}$ is convex, and the constraint functions g_i are convex for $i = 1, 2, \ldots, p$ and affine for $i = p+1, \ldots, m$.

The affine conditions $g_{p+1}(x) = 0, \ldots, g_m(x) = 0$ in a convex problem can of course be summarized as $Ax = b$, where A is an $(m-p) \times n$-matrix.

The set X of feasible points is convex in a convex minimization problem, for

$$X = \bigcap_{i=1}^{p} \{x \in \Omega \mid g_i(x) \leq 0\} \cap \bigcap_{i=p+1}^{m} \{x \mid g_i(x) = 0\},$$

and this expresses X as an intersection of sublevel sets of convex functions and hyperplanes.

A maximization problem

$$\begin{array}{ll} \max & f(x) \\ \text{s.t.} & x \in X \end{array}$$

is called convex if the corresponding equivalent minimization problem

$$\begin{array}{ll} \min & -f(x) \\ \text{s.t.} & x \in X \end{array}$$

is convex, which means that the objective function f has to be concave.

LP problems are of course convex optimization problems. General convex optimization problems are studied in Chapter 11.

Convex quadratic programming

We get a special case of convex optimization if X is a polyhedron and the objective function f is a sum of a linear form and a positive semidefinite quadratic form, i.e. has the form $f(x) = \langle c, x\rangle + \langle x, Qx\rangle$, where Q is a positive semidefinite matrix. The problem (P) is then called *convex quadratic programming.* LP problems constitute a subclass of the convex quadratic problems, of course.

Non-linear optimization

Non-linear optimization is about optimization problems that are not supposed to be LP problems. Since non-linear optimization includes almost everything, there is of course no general theory that can be applied to an arbitrary non-linear optimization problem.

If f is a differentiable function and X is a "decent" set in $\mathbf{R}^n$, one can of course use differential calculus to attack the minimization problem (P). We recall in this context the Lagrange theorem, which gives a necessary condition for the minimum (and maximum) when

$$X = \{x \in \mathbf{R}^n \mid g_1(x) = g_2(x) = \cdots = g_m(x) = 0\}.$$

A counterpart of Lagrange's theorem for optimization problems with constraints in the form of inequalities is given in Chapter 10.

Integer programming

An *integer programming* problem is a mathematical optimization problem in which some or all of the variables are restricted to be integers. In particular, a *linear integer problem* is a problem of the form

$$\begin{array}{ll} \min & \langle c, x\rangle \\ \text{s.t.} & x \in X \cap (\mathbf{Z}^m \times \mathbf{R}^{n-m}) \end{array}$$

where $\langle c, x\rangle$ is a linear form and X is a polyhedron in $\mathbf{R}^n$.

Many problems dealing with flows in networks, e.g. commodity distribution problems and maximum flow problems, are linear integer problems that can be solved using special algorithms.

Simultaneous optimization

The title refers to a type of problems that are not really optimization problems in the previous sense. There are many situations, where an individual may affect the outcome through his actions without having full control over

the situation. Some variables may be in the hands of other individuals with completely different desires about the outcome, while other variables may be of a completely random nature. The problem to in some sense optimize the outcome could then be called *simultaneous optimization*.

Simultaneous optimization is the topic of *game theory*, which deals with the behavior of the various agents in conflict situations. Game theoretical concepts and results have proved to be very useful in various contexts, e.g. in economics.

9.3 Equivalent problem formulations

Let us informally call two optimization problems *equivalent* if it is possible to determine in an automatical way an optimal solution to one of the problems, given an optimal solution to the other, and vice versa.

A trivial example of equivalent problems are, as already mentioned, the problems

$$\begin{array}{ll} \max & f(x) \\ \text{s.t.} & x \in X \end{array} \qquad \text{and} \qquad \begin{array}{ll} \min & -f(x). \\ \text{s.t.} & x \in X \end{array}$$

We now describe some useful transformations that lead to equivalent optimization problem

Elimination of equalities

Consider the problem

$$\text{(P)} \qquad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & \begin{cases} g_i(x) \le 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m. \end{cases} \end{array}$$

If it is possible to solve the subsystem of equalities and express the solution in the form $x = h(y)$ with a parameter y running over some subset of $\mathbf{R}^d$, then we can eliminate the equalities and rewrite problem (P) as

$$\text{(P}') \qquad \begin{array}{ll} \min & f(h(y)) \\ \text{s.t.} & g_i(h(y)) \le 0, \quad i = 1, 2, \ldots, p \end{array}$$

If $\hat{y}$ is an optimal solution to (P$'$), then $h(\hat{y})$ is of course an optimal solution to (P). Conversely, if $\hat{x}$ is an optimal solution to (P), then $\hat{x} = h(\hat{y})$ for some value $\hat{y}$ of the parameter, and this value is an optimal solution to (P$'$).

The elimination is always possible (by a simple algorithm) if all constraint equalities are affine, i.e. if the system can be written in the form $Ax = b$ for

some $(m-p) \times n$-matrix A. Assuming that the system is consistent, the solution set is an affine subspace of dimension $d = n - \operatorname{rank} A$, and there exists an $n \times d$-matrix C of rank d and a particular solution x_0 to the system such that $Ax = b$ if and only if $x = Cy + x_0$ for some $y \in \mathbf{R}^d$. The problem (P) is thus in this case equivalent to the problem

$$\begin{array}{ll} \min & f(Cy + x_0) \\ \text{s.t.} & g_i(Cy + x_0) \le 0, \qquad i = 1, 2, \ldots, p \end{array}$$

(with implicit constraint $Cy + x_0 \in \Omega$).

In convex optimization problems, and especially in LP problems, we can thus, in principle, eliminate the equalities from the constraints and in this way replace the problem by an equivalent optimization problem without any equality constraints.

Slack variables

The inequality $g(x) \le 0$ holds if and only if there is a number $s \ge 0$ such that $g(x) + s = 0$. By thus replacing all inequalities in the problem

$$\text{(P)} \qquad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & \begin{cases} g_i(x) \le 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m \end{cases} \end{array}$$

with equalities, we obtain the following equivalent problem

$$\text{(P}') \qquad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & \begin{cases} g_i(x) + s_i = 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m \\ s_i \ge 0, & i = 1, 2, \ldots, p \end{cases} \end{array}$$

with $n + p$ variables, m equality constraints and p simple inequality constraints. The new variables s_i are called *slack variables*.

If $\hat{x}$ is an optimal solution to (P), we get an optimal solution $(\hat{x}, \hat{s})$ to (P′) by setting $\hat{s}_i = -g_i(\hat{x})$. Conversely, if $(\hat{x}, \hat{s})$ is an optimal solution to the last mentioned problem, then $\hat{x}$ is of course an optimal solution to the original problem.

If the original constraints are affine, then so are all new constraints. The transformation thus transforms LP problems to LP problems.

Inequalities of the form $g(x) \ge 0$ can of course similarly be written as equalities $g(x) - s = 0$ with nonnegative variables s. These new variables are usually called *surplus variables*.

Nonnegative variables

Every real number can be written as a difference between two nonnegative numbers. In an optimization problem, we can thus replace an unrestricted variable x_i, i.e. a variable that a priori may assume any real value, with two nonnegative variables x_i' and x_i'' by setting

$$x_i = x_i' - x_i'', \quad x_i' \geq 0, \; x_i'' \geq 0.$$

The number of variables increases with one and the number of inequalities increases with two for each unrestricted variable that is replaced, but the transformation leads apparently to an equivalent problem. Moreover, convex problems are transfered to convex problems and LP problems are transformed to LP problems.

Example 9.3.1. The LP problem

$$\begin{array}{ll} \min & x_1 + 2x_2 \\ \text{s.t.} & \begin{cases} x_1 + x_2 \geq 2 \\ 2x_1 - x_2 \leq 3 \\ x_1 \geq 0 \end{cases} \end{array}$$

is transformed, using two slack/surplus variables and by replacing the unrestricted variable x_2 with a difference of two nonnegative variables, to the following equivalent LP problem in which all variables are nonnegative and all remaining constraints are equalities.

$$\begin{array}{ll} \min & x_1 + 2x_2' - 2x_2'' + 0s_1 + 0s_2 \\ \text{s.t.} & \begin{cases} x_1 + x_2' - x_2'' - s_1 = 2 \\ 2x_1 - x_2' + x_2'' + s_2 = 3 \\ x_1, x_2', x_2'', s_1, s_2 \geq 0. \end{cases} \end{array}$$

□

Epigraph form

Every optimization problem can be replaced by an equivalent problem with a linear objective function, and the trick to accomplish this is to utilize the epigraph of the original objective function. The two problems

$$\text{(P)} \quad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in X \end{array} \qquad \text{and} \qquad \text{(P}') \quad \begin{array}{ll} \min & t \\ \text{s.t.} & \begin{cases} f(x) \leq t \\ x \in X \end{cases} \end{array}$$

are namely equivalent, and the objective function in (P′) is linear. If $\hat{x}$ is an optimal solution to (P), then $(\hat{x}, f(\hat{x}))$ is an optimal solution to (P′), and if $(\hat{x}, \hat{t})$ is an optimal solution to (P′), then $\hat{x}$ is an optimal solution to (P).

If problem (P) is convex, i.e. has the form

$$\begin{array}{ll}\min & f(x)\\ \text{s.t.} & \begin{cases} g_i(x) \le 0, & i = 1, 2, \dots, p\\ g_i(x) = 0, & i = p+1, \dots, m \end{cases}\end{array}$$

with convex functions f and g_i for $1 \le i \le p$, and affine functions g_i for $i \ge p+1$, then the epigraph variant

$$\begin{array}{ll}\min & t\\ \text{s.t.} & \begin{cases} f(x) - t \le 0, & \\ g_i(x) \le 0, & i = 1, 2, \dots, p\\ g_i(x) = 0, & i = p+1, \dots, m \end{cases}\end{array}$$

is also a convex problem.

So there is no restriction to assume that the objective function of a convex program is linear when we are looking for general properties of such programs.

Piecewise affine objective functions

Suppose that X is a polyhedron (given as an intersection of closed halfspaces) and consider the convex optimization problem

$$\text{(P)} \qquad \begin{array}{ll}\min & f(x)\\ \text{s.t.} & x \in X\end{array}$$

where the objective function $f(x)$ is piecewise affine and given as

$$f(x) = \max\{\langle c_i, x\rangle + b_i \mid i = 1, 2, \dots, m\}.$$

The epigraph transformation results in the equivalent convex problem

$$\begin{array}{ll}\min & t\\ \text{s.t.} & \begin{cases} \max\limits_{1\le i\le m} (\langle c_i, x\rangle + b_i) \le t\\ x \in X, \end{cases}\end{array}$$

and since $\max_{1\le i\le m} \alpha_i \le t$ if and only if $\alpha_i \le t$ for all i, this problem is in turn equivalent to the LP problem

$$\text{(P}') \qquad \begin{array}{ll}\min & t\\ \text{s.t.} & \begin{cases} \langle c_i, x\rangle - t + b_i \le 0, & i = 1, 2, \dots, m\\ x \in X. \end{cases}\end{array}$$

The constraint set of this LP problem is a polyhedron in $\mathbf{R}^n \times \mathbf{R}$.

If instead the objective function in problem (P) is a sum

$$f(x) = f_1(x) + f_2(x) + \cdots + f_k(x)$$

of piecewise affine functions f_i, then problem (P) is equivalent to the convex problem

$$\begin{array}{ll} \min & t_1 + t_2 + \cdots + t_k \\ \text{s.t.} & \begin{cases} f_i(x) \le t_i & i = 1, 2, \ldots, k \\ x \in X \end{cases} \end{array}$$

and this problem becomes an LP problem if every inequality $f_i(x) \le t_i$ is expressed as a system of linear inequalities in a similar way as above.

9.4 Some model examples

Diet problem

Let us start with a classical LP problem that was formulated and studied during the childhood of linear programming. The goal of the diet problem is to select a set of foods that will satisfy a set of daily nutritional requirements at minimum cost. There are n foods $L_1, L_2, \ldots, L_n$ available at a cost of $c_1, c_2, \ldots, c_n$ dollars per unit. The foods contain various nutrients $N_1, N_2, \ldots, N_m$ (proteins, carbohydrates, fats, vitamins, etc.). The number of units of nutrients per unit of food is shown by the following table:

	L_1	L_2	$\ldots$	L_n
N_1	a_{11}	a_{12}	$\ldots$	a_{1n}
N_2	a_{21}	a_{22}	$\ldots$	a_{2n}
$\vdots$				
N_m	a_{m1}	a_{m2}	$\ldots$	a_{mn}

Buying $x_1, x_2, \ldots, x_n$ units of the foods, one thus obtains

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n$$

units of nutrient N_i at a cost of

$$c_1x_1 + c_2x_2 + \cdots + c_nx_n.$$

Suppose that the daily requirement of the different nutrients is b_1, b_2, $\ldots$, b_m and that it is not harmful to have too much of any substance. The

problem to meet the daily requirement at the lowest possible cost is called the *diet problem*. Mathematically, it is of the form

$$\begin{aligned} \min \ & c_1x_1 + c_2x_2 + \cdots + c_nx_n \\ \text{s.t.} \ & \begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \geq b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \geq b_2 \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \geq b_m \\ x_1,\, x_2, \ldots, x_n \geq 0. \end{cases} \end{aligned}$$

The diet problem is thus an LP problem. In addition to determining the optimal diet and the cost of this, it would be of interest to answer the following questions:

1. How does a price change of one or more of the foods affect the optimal diet and the cost?
2. How is the optimal diet affected by a change of the daily requirement of one or more nutrients?
3. Suppose that pure nutrients are available on the market. At what price would it be profitable to buy these and satisfy the nutritional needs by eating them instead of the optimal diet? Hardly a tasty option for a gourmet but perhaps possible in animal feeding.

Assume that the cost of the optimal diet is z, and that its cost changes to $z + \Delta z$ when the need for nutrient N_1 is changed from b_1 to $b_1 + \Delta b_1$, ceteris paribus. It is obvious that the cost can not be reduced when demand increases, so therefore $\Delta b_1 > 0$ entails $\Delta z \geq 0$. If it is possible to buy the nutrient N_1 in completely pure form to the price p_1, then it is economically advantageous to meet the increased need by taking the nutrient in pure form, provided that $p_1\Delta b_1 \leq \Delta z$. The maximum price of N_1 which makes nutrient in pure form an economical alternative is therefore $\Delta z/\Delta b_1$, and the limit as $\Delta b_1 \to 0$, i.e. the partial derivative $\frac{\partial z}{\partial b_1}$, is called the *dual price* or the *shadow price* in economic literature.

It is possible to calculate the nutrient shadow prices by solving an LP problem closely related to the diet problem. Assume again that the market provides nutrients in pure form and that their prices are $y_1, y_2, \ldots, y_m$. Since one unit of food L_i contains $a_{1i}, a_{2i}, \ldots, a_m$ units of each nutrient, we can "manufacture" one unit of food L_i by buying just this set of nutrients, and hence it is economically advantageous to replace all foods by pure nutrients if

$$a_{1i}y_1 + a_{2i}y_2 + \cdots + a_my_m \leq c_i$$

for $i = 1, 2, \ldots, n$. Under these conditions the cost of the required daily ration

$b_1, b_2, \ldots, b_m$ is at most equal to the maximum value of the LP problem

$$\begin{aligned} \max\ & b_1y_1 + b_2y_2 + \cdots + b_my_m \\ \text{s.t.}\ & \begin{cases} a_{11}y_1 + a_{21}y_2 + \ldots + a_{m1}y_m \le c_1 \\ a_{12}y_1 + a_{22}y_2 + \ldots + a_{m2}y_m \le c_2 \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{1n}y_1 + a_{2n}y_2 + \ldots + a_{mn}y_m \le c_n \\ \qquad\qquad\qquad\qquad y_1,\ y_2, \ldots, y_m \ge 0. \end{cases} \end{aligned}$$

We will show that this so called *dual problem* has the same optimal value as the original diet problem and that the optimal solution is given by the shadow prices.

Production planning

Many problems related to production planning can be formulated as LP problems, and a pioneer in the field was the Russian mathematician and economist Leonid Kantorovich, who studied and solved such problems in the late 1930s. Here is a typical such problem.

A factory can manufacture various goods $V_1, V_2, \ldots, V_n$. This requires various inputs (raw materials and semi-finished goods) and different types of labor, something which we collectively call production factors $P_1, P_2, \ldots, P_m$. These are available in limited quantities $b_1, b_2, \ldots, b_m$. In order to manufacture, market and sell one unit of the respective goods, production factors are needed to an extent given by the following table:

	V_1	V_2	$\ldots$	V_n
P_1	a_{11}	a_{12}	$\ldots$	a_{1n}
P_2	a_{21}	a_{22}	$\ldots$	a_{2n}
$\vdots$				
P_m	a_{m1}	a_{m2}	$\ldots$	a_{mn}

Every manufactured product V_j can be sold at a profit which is c_j dollars per unit, and the goal now is to plan the production $x_1, x_2, \ldots, x_n$ of the various products so that the profit is maximized.

Manufacturing $x_1, x_2, \ldots, x_n$ units of the goods consumes $a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n$ units of production factor P_i and results in a profit equal to $c_1x_1 + c_2x_2 + \cdots + c_nx_n$. The optimization problem that we need to solve is thus the LP problem

$$\max\ c_1x_1 + c_2x_2 + \cdots + c_nx_n$$
$$\text{s.t.}\ \begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \le b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \le b_2 \\ \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \le b_m \\ x_1,\, x_2, \ldots, x_n \ge 0. \end{cases}$$

Here it is reasonable to ask similar questions as for the diet problem, i.e. how is the optimal solution and the optimal profit affected by

1. altered pricing $c_1, c_2, \ldots, c_n$;
2. changes in the resource allocation.

If we increase a resource P_i that is already fully utilized, so does (normally) the profit. What will the price of this resource be for the expansion to pay off? The critical price is called the shadow price, and it can be interpreted as a partial derivative, and as the solution to a dual problem.

Transportation problem

The transportation problem is another classical LP problem that was formulated and solved before the invention of the simplex algorithm

A commodity (e.g. gasoline) is stored at m places $S_1, S_2, \ldots, S_m$ and demanded at n other locations $D_1, D_2, \ldots, D_n$. The quantity of the commodity available at S_i is a_i units, while b_j units are demanded at D_j. To ship 1 unit from storage place S_i to demand center D_j costs c_{ij} dollars.

The total supply, i.e. $\sum_{i=1}^{m} a_i$, is assumed for simplicity to be equal to the total demand $\sum_{j=1}^{n} b_j$, so it is possible to meet the demand by distributing x_{ij} units from S_i to D_j. To do this at the lowest transportation cost gives

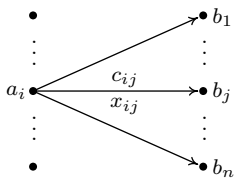

Figure 9.1. The transportation problem

rise to the LP problem

$$\min \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij}$$
$$\text{s.t.} \begin{cases} \sum_{j=1}^{n} x_{ij} = a_i, & i = 1, 2, \ldots, m \\ \sum_{i=1}^{m} x_{ij} = b_j, & j = 1, 2, \ldots, n \\ x_{ij} \geq 0, & \text{all } i, j. \end{cases}$$

An investment problem

An investor has 1 million dollars, which he intends to invest in various projects, and he has found m interesting candidates $P_1, P_2, \ldots, P_m$ for this. The return will depend on the projects and the upcoming economic cycle. He thinks he can identify n different economic situations $E_1, E_2, \ldots, E_n$, but it is impossible for him to accurately predict what the economy will look like in the coming year, after which he intends to collect the return. However, one can accurately assess the return of each project during the various economic cycles; each invested million dollars in project P_i will yield a return of a_{ij} million dollars during business cycle E_j. We have, in other words, the following table of return for various projects and business cycles:

	E_1	E_2	$\ldots$	E_n
P_1	a_{11}	a_{12}	$\ldots$	a_{1n}
P_2	a_{21}	a_{22}	$\ldots$	a_{2n}
$\vdots$				
P_m	a_{m1}	a_{m2}	$\ldots$	a_{mn}

Our investor intends to invest $x_1, x_2, \ldots, x_m$ million dollars in the various projects, and this will give him the return

$$a_{1j}x_1 + a_{2j}x_2 + \cdots + a_{mj}x_m$$

million dollars, assuming that the economy will be in state E_j. Since our investor is a very cautious person, he wants to guard against the worst possible outcome, and the worst possible outcome for the investment $x_1, x_2, \ldots, x_m$ is

$$\min_{1\leq j\leq n} \sum_{i=1}^{m} a_{ij}x_i.$$

He therefore wishes to maximize this outcome, which he does by solving the problem

$$\begin{aligned} &\max \ \min_{1\le j\le n} \sum_{i=1}^{m} a_{ij}x_i \\ &\text{s.t.} \quad x \in X \end{aligned}$$

where X is the set $\{(x_1, x_2, \ldots, x_m) \in \mathbf{R}^m_+ \mid \sum_{i=1}^m x_i = 1\}$ of all possible ways to distribute one million on the various projects.

In this formulation, the problem is a convex maximization problem with a piecewise affine concave objective function. However, we can transform it into an equivalent LP problem by making use of a hypograph formulation. Utilizing the techniques of the previous section, we see that the investor's problem is equivalent to the LP problem

$$\begin{aligned} &\max \ v \\ &\text{s.t.} \ \begin{cases} a_{11}x_1 + a_{21}x_2 + \ldots + a_{m1}x_m \ge v \\ a_{12}x_1 + a_{22}x_2 + \ldots + a_{m2}x_m \ge v \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{1n}x_1 + a_{2n}x_2 + \ldots + a_{mn}x_m \ge v \\ x_1 + x_2 + \ldots + x_m = 1 \\ x_1,\, x_2, \ldots, x_m \ge 0. \end{cases} \end{aligned}$$

Two-person zero-sum game

Two persons, row player Rick and column player Charlie, each choose, independently of each other, an integer. Rick chooses a number i in the range $1 \le i \le m$ and Charlie a number j in the range $1 \le j \le n$. If they choose the pair (i, j), Rick wins a_{ij} dollars of Charlie, and to win a negative amount is of course the same as to loose the corresponding positive amount.

The numbers m, n and a_{ij} are supposed to be known by both players, and the objective of each player is to win as much as possible (or equivalently, to loose as little as possible). There is generally no best choice for any of the players, but they could try to maximize their expected winnings by selecting their numbers at random with a certain probability distribution.

Suppose Rick chooses the number i with probability x_i, and Charlie chooses the number j with probability y_j. All probabilities are of course nonnegative numbers, and $\sum_{i=1}^m x_i = \sum_{j=1}^n y_j = 1$. Let

$$X = \{x \in \mathbf{R}^m_+ \mid \sum_{i=1}^{m} x_i = 1\} \quad \text{and} \quad Y = \{y \in \mathbf{R}^n_+ \mid \sum_{j=1}^{n} y_j = 1\}.$$

The elements in X are called the row player's *mixed strategies*, and the elements in Y are the column player's mixed strategies.

Since the players choose their numbers independently of each other, the outcome (i, j) will occur with probability $x_i y_j$. Rick's pay-off is therefore a random variable with expected value

$$f(x, y) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} x_i y_j.$$

Row player Rick can now conceivably argue like this: "The worst that can happen to me, if I choose the probability distribution x, is that my opponent Charlie happens to choose a probability distribution y that minimizes my expected profit $f(x, y)$". In this case, Rick will obtain the amount

$$g(x) = \min_{y \in Y} f(x, y) = \min_{y \in Y} \sum_{j=1}^{n} y_j \big(\sum_{i=1}^{m} a_{ij} x_i \big).$$

The sum $\sum_{j=1}^{n} y_j \big(\sum_{i=1}^{m} a_{ij} x_i \big)$ is a weighted arithmetic mean of the n numbers $\sum_{i=1}^{m} a_{ij} x_i,\ j = 1, 2, \ldots, n$, with the weights $y_1, y_2, \ldots, y_n$, and such a mean is greater than or equal to the smallest of the n numbers, and equality is obtained by putting all weight on this smallest number. Hence,

$$g(x) = \min_{1 \le j \le n} \sum_{i=1}^{m} a_{ij} x_i.$$

Rick, who wants to maximize his outcome, should therefore choose to maximize $g(x)$, i.e. Rick's problem becomes

$$\begin{aligned} &\max\ g(x) \\ &\text{s.t.}\quad x \in X. \end{aligned}$$

This is exactly the same problem as the investor's problem. Hence, Rick's optimal strategy, i.e. optimal choice of probabilities, coincides with the optimal solution to the LP problem

$$\begin{aligned} &\max\ v \\ &\text{s.t.} \begin{cases} a_{11}x_1 + a_{21}x_2 + \ldots + a_{m1}x_m \ge v \\ a_{12}x_1 + a_{22}x_2 + \ldots + a_{m2}x_m \ge v \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{1n}x_1 + a_{2n}x_2 + \ldots + a_{mn}x_m \ge v \\ x_1 + \quad x_2 + \ldots + \quad x_m = 1 \\ x_1,\, x_2, \ldots, x_m \ge 0. \end{cases} \end{aligned}$$

The column player's problem is analogous, but he will of course minimize the maximum expected outcome $f(x, y)$. Charlie must therefore solve the problem

$$\begin{aligned} \min \quad & \max_{1\le i\le m} \sum_{j=1}^{n} a_{ij} y_j \\ \text{s.t.} \quad & y \in Y \end{aligned}$$

to find his optimal strategy, and this problem is equivalent to the LP problem

$$\begin{aligned} \min \quad & u \\ \text{s.t.} \quad & \begin{cases} a_{11}y_1 + a_{12}y_2 + \ldots + a_{1n}y_n \le u \\ a_{21}y_1 + a_{22}y_2 + \ldots + a_{2n}y_n \le u \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{m1}y_1 + a_{m2}y_2 + \ldots + a_{mn}y_n \le u \\ y_1 + y_2 + \ldots + y_n = 1 \\ y_1,\, y_2, \ldots, y_n \ge 0. \end{cases} \end{aligned}$$

The two players' problems are examples of dual problems, and it follows from results that will appear in Chapter 12 that they have the same optimal value.

Consumer Theory

The behavior of consumers is studied in a branch of economics known as microeconomics. Assume that there are n commodities $V_1, V_2, \ldots, V_n$ on the market and that the price of these goods is given by the price vector $p = (p_1, p_2, \ldots, p_n)$. A basket x consisting of $x_1, x_2, \ldots, x_n$ units of the goods thus costs $\langle p, x\rangle = p_1x_1 + p_2x_2 + \cdots + p_nx_n$.

A consumer values her benefit of the commodity bundle x by using a subjective utility function f, where $f(x) > f(y)$ means that she prefers x to y. A reasonable assumption about the utility function is that every convex combination $\lambda x + (1-\lambda)y$ of two commodity bundles should be valued as being at least as good as the worst of the two bundles x and y, i.e. that $f(\lambda x + (1-\lambda)y) \ge \min\big(f(x), f(y)\big)$. The utility function f is assumed, in other words, to be quasiconcave, and a stronger assumption, which is often made in the economic literature and that we are making here, is that f is concave.

Suppose now that our consumer's income is I, that the entire income is disposable for consumption, and that she wants to maximize her utility.

Then, the problem that she needs to solve is the convex optimization problem

$$\begin{aligned} &\max \quad f(x) \\ &\text{s.t.} \quad \begin{cases} \langle p, x \rangle \leq I \\ x \geq 0. \end{cases} \end{aligned}$$

To determine empirically a consumer's utility function is of course almost impossible, so microtheory is hardly useful for quantitative calculations. However, one can make qualitative analyzes and answer questions of the type: How does an increase in income change the consumer behavior? and How does changes in the prices of the goods affect the purchasing behavior?

Portfolio optimization

A person intends to buy shares in n different companies $C_1, C_2, \ldots, C_n$ for S dollars. One dollar invested in the company C_j gives a return of R_j dollars, where R_j is a random variable with known expected value

$$\mu_j = \mathrm{E}[R_j].$$

The covariances

$$\sigma_{ij} = \mathrm{E}[(R_i - \mu_i)(R_j - \mu_j)]$$

are also assumed to be known.

The expected total return $e(x)$ from investing $x = (x_1, x_2, \ldots, x_n)$ dollars in the companies $C_1, C_2, \ldots, C_n$ is given by

$$e(x) = \mathrm{E}\big[\sum_{j=1}^{n} x_j R_j\big] = \sum_{j=1}^{n} \mu_j x_j,$$

and the variance of the total return is

$$v(x) = \mathrm{Var}\,\big[\sum_{j=1}^{n} x_j R_j\big] = \sum_{i,j=1}^{n} \sigma_{ij} x_i x_j.$$

Note that $v(x)$ is a positive semi-definite quadratic form.

It is not possible for our person to maximize the total return, because the return is a random variable, i.e. depends on chance. However, he can maximize the expected total return under appropriate risk conditions, i.e. requirements for the variance. Alternatively, he can minimize the risk with the investment given certain requirements on the expected return. Thus there are several possible strategies, and we will formulate three such.

(i) The strategy to maximize the expected total return, given an upper bound B on the variance, leads to the convex optimization problem

$$\begin{array}{ll} \max & e(x) \\ \text{s.t.} & \begin{cases} v(x) \le B \\ x_1 + x_2 + \cdots + x_n = S \\ x \ge 0. \end{cases} \end{array}$$

(ii) The strategy to minimize the variance of the total return, given a lower bound b on the expected return, gives rise to the convex quadratic programming problem

$$\begin{array}{ll} \min & v(x) \\ \text{s.t.} & \begin{cases} e(x) \ge b \\ x_1 + x_2 + \cdots + x_n = S \\ x \ge 0. \end{cases} \end{array}$$

(iii) The two strategies can be considered together in the following way. Let $\epsilon \ge 0$ be a (subjective) parameter, and consider the convex quadratic problem

$$\begin{array}{ll} \min & \epsilon v(x) - e(x) \\ \text{s.t.} & \begin{cases} x_1 + x_2 + \cdots + x_n = S \\ x \ge 0 \end{cases} \end{array}$$

with optimal solution $x(\epsilon)$. We leave as an exercise to show that

$$v(x(\epsilon_1)) \ge v(x(\epsilon_2)) \qquad \text{and} \qquad e(x(\epsilon_1)) \ge e(x(\epsilon_2))$$

if $0 \le \epsilon_1 \le \epsilon_2$. The parameter ϵ is thus a measure of the person's attitude towards risk; the smaller the ϵ, the greater the risk (= variance) but also the greater expected return.

Snell's law of refraction

We will study the path of a light beam which passes through n parallel transparent layers. The j:th slice S_j is assumed to be a_j units wide and to consist of a homogeneous medium in which the speed of light is v_j. We choose a coordinate system as in figure 9.2 and consider a light beam on its path from the origin on the surface of the first slice to a point with y-coordinate b on the outer surface of the last slice.

According to Fermat's principle, the light chooses the fastest route. The path of the beam is therefore determined by the optimal solution to the

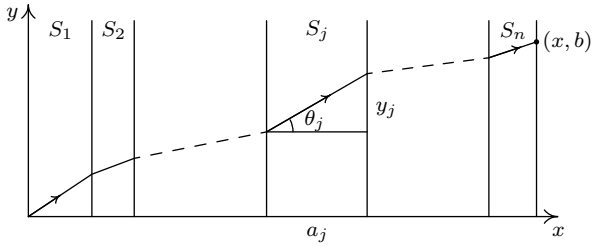

Figure 9.2. The path of a light beam through layers with different refraction indices.

convex optimization problem

$$\begin{aligned} \min \quad & \sum_{j=1}^{n} v_j^{-1} \sqrt{y_j^2 + a_j^2} \\ \text{s.t.} \quad & \sum_{j=1}^{n} y_j = b, \end{aligned}$$

and we obtain *Snell's law of refraction*

$$\frac{\sin \theta_i}{\sin \theta_j} = \frac{v_i}{v_j}$$

by solving the problem.

Overdetermined systems

If a system of linear equations

$$Ax = b$$

with n unknowns and m equations is inconsistent, i.e. has no solutions, you might want to still determine the best approximate solution, i.e. the n-tuple $x = (x_1, x_2, \ldots, x_n)$ that makes the error as small as possible. The error is by definition the difference $Ax - b$ between the left and the right hand side of the equation, and as a measure of the size of the error we use $\|Ax - b\|$ for some suitably chosen norm.

The function $x \mapsto \|Ax - b\|$ is convex, so the problem of minimizing $\|Ax - b\|$ over all $x \in \mathbf{R}^n$ is a convex problem regardless of which norm is used, but the solution depends on the norm, of course. Let as usual a_{ij} denote the element at location i, j in the matrix A, and let $b = (b_1, b_2, \ldots, b_m)$.

1. The so-called *least square solution* is obtained by using the Euclidean norm $\|\cdot\|_2$. Since $\|Ax-b\|_2^2 = \sum_{i=1}^m (a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n - b_i)^2$, we get the least square solution as the solution of the convex quadratic problem

$$\text{minimize} \sum_{i=1}^{m} (a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n - b_i)^2.$$

The gradient of the objective function is equal to zero at the optimal point, which means that the optimal solution is obtained as the solution to the linear system

$$A^T Ax = A^T b.$$

2. By instead using the $\|\cdot\|_\infty$ norm, one obtains the solution that gives the smallest maximum deviation between the left and the right hand side of the linear system $Ax = b$. Since

$$\|Ax-b\|_\infty = \max_{1\le i\le m} |a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n - b_i|,$$

the objective function is now piecewise affine, and the problem is therefore equivalent to the LP problem

$$\begin{array}{ll} \min & t \\ \text{s.t.} & \left\{ \begin{array}{l} \pm(a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n - b_1) \le t \\ \qquad\qquad \vdots \\ \pm(a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n - b_m) \le t. \end{array} \right. \end{array}$$

3. Instead of minimizing the sum of squares of the differences between left and right sides, we can of course minimize the sum of the absolute value of the differences, i.e. use the $\|\cdot\|_1$-norm. Since the objective function

$$\|Ax-b\|_1 = \sum_{i=1}^{m} |a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n - b_i|$$

is a sum of convex piecewise affine functions, our convex minimization problem is in this case equivalent to the LP problem

$$\begin{array}{ll} \min & t_1 + t_2 + \cdots + t_m \\ \text{s.t.} & \left\{ \begin{array}{l} \pm(a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n - b_1) \le t_1 \\ \qquad\qquad \vdots \\ \pm(a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n - b_m) \le t_m. \end{array} \right. \end{array}$$

Largest inscribed ball

A convex set X with nonempty interior is given in $\mathbf{R}^n$, and we want to determine a ball $B(x,r)$ in X (with respect to a given norm) with the largest possible radius r. We assume that X can be described as the solution set to a system of inequalities, i.e. that

$$X = \{x \in \mathbf{R}^n \mid g_i(x) \leq 0,\ i = 1, 2, \ldots, m\},$$

with convex functions g_i.

The ball $B(x,r)$ lies in X if and only if $g_i(x + ry) \leq 0$ for all y with $\|y\| \leq 1$ and $i = 1, 2, \ldots, m$, which makes it natural to consider the functions

$$h_i(x,r) = \sup_{\|y\| \leq 1} g_i(x + ry), \quad i = 1, 2, \ldots, m.$$

The functions h_i are convex since they are defined as suprema of convex functions in the variables x and r.

The problem of determining the ball with the largest possible radius has now been transformed into the convex optimization problem

$$\begin{aligned} &\max\ r \\ &\text{s.t.}\quad h_i(x,r) \leq 0, \quad i = 1, 2, \ldots, m. \end{aligned}$$

For general convex sets X, it is of course impossible to determine the functions h_i explicitly, but if X is a polyhedron, $g_i(x) = \langle c_i, x\rangle - b_i$, and the norm in question is the ℓ^p-norm, then it follows from Hölder's inequality that

$$h_i(x,r) = \sup_{\|y\|_p \leq 1} (\langle c_i, x\rangle + r\langle c_i, y\rangle - b_i) = \langle c_i, x\rangle + r\|c_i\|_q - b_i$$

for $r \geq 0$, where $\|\cdot\|_q$ denotes the dual norm.

The problem of determining the center x and the radius r of the largest ball that is included in the polyhedron

$$X = \{x \in \mathbf{R}^n \mid \langle c_i, x\rangle \leq b_i,\ i = 1, 2, \ldots, m\}$$

has now been reduced to the LP problem

$$\begin{aligned} &\max\ r \\ &\text{s.t.}\quad \langle c_i, x\rangle + r\|c_i\|_q \leq b_i, \quad i = 1, 2, \ldots, m. \end{aligned}$$

Exercises

9.1 In a chemical plant one can use four different processes P_1, P_2, P_3, and P_4 to manufacture the products V_1, V_2, and V_3. Produced quantities of the various products, measured in tons per hour, for the various processes are shown in the following table:

	P_1	P_2	P_3	P_4
V_1	-1	2	2	1
V_2	4	1	0	2
V_3	3	1	2	1

(Process P_1 thus consumes 1 ton of V_1 per hour!) Running processes P_1, P_2, P_3, and P_4 costs 5 000, 4 000, 3 000, and 4 000 dollars per per hour, respectively. The plant intends to produce 16, 40, and 24 tons of products V_1, V_2, and V_3 at the lowest possible cost. Formulate the problem of determining an optimal production schedule.

9.2 Bob has problems with the weather. The weather occurs in the three states pouring rain, drizzle and sunshine. Bob owns a raincoat and an umbrella, and he is somewhat careful with his suit. The raincoat is difficult to carry, and the same applies − though to a lesser degree − to the umbrella; the latter, however, is not fully satisfactory in case of pouring rain. The following table reveals how happy Bob considers himself in the various situations that can arise (the numbers are related to his blood pressure, with 0 corresponding to his normal state).

	Pouring rain	Drizzle	Sunshine
Raincoat	2	1	-2
Umbrella	1	2	-1
Only suit	-4	-2	2

In the morning, when Bob goes to work, he does not know what the weather will be like when he has to go home, and he would therefore choose the clothes that optimize his mind during the walk home. Formulate Bob's problem as an LP problem.

9.3 Consider the following two-person game in which each player has three alternatives and where the payment to the row player is given by the following payoff matrix.

	1	2	3
1	1	0	5
2	3	3	4
3	2	4	0

In this case, it is obvious which alternatives both players must choose. How will they play?

9.4 Charlie and Rick have three cards each. Both have the ace of diamonds and the ace of spades. Charlie also has the two of diamonds, and Rick has the two of spades. The players play simultaneously one card each. Charlie wins if both these cards are of the same color and loses in the opposite case. The winner will receive as payment the value of his winning card from the opponent, with ace counting as 1. Write down the payoff matrix for this two-person game, and formulate column player Charlie's problem to optimize his expected profit as an LP problem.

9.5 The overdetermined system

$$\begin{cases} x_1 + x_2 = 2 \\ x_1 - x_2 = 0 \\ 3x_1 + 2x_2 = 4 \end{cases}$$

has no solution.

a) Determine the least square solution.
b) Formulate the problem of determining the solution that minimizes the maximum difference between the left and the right hand sides of the system.
c) Formulate the problem of determining the solution that minimizes the sum of the absolut values of the differences between the left and the right hand sides.

9.6 Formulate the problem of determining

a) the largest circular disc,
b) the largest square with sides parallel to the coordinate axes,

that is contained in the triangle bounded by the lines $x_1 - x_2 = 0$, $x_1 - 2x_2 = 0$ and $x_1 + x_2 = 1$.

Chapter 10

The Lagrange function

10.1 The Lagrange function and the dual problem

The Lagrange function

To the minimization problem

$$\text{(P)} \qquad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & \begin{cases} g_i(x) \le 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m \end{cases} \end{array}$$

with $x \in \Omega$ as implicit condition and m explicit constraints, the first p of which in the form of inequalities, we shall associate a dual maximization problem, and the tool to accomplish this is the Lagrange function defined below. To avoid trivial matters we assume that $\operatorname{dom} f \neq \emptyset$, i.e. that the objective function $f\colon \Omega \to \overline{\mathbf{R}}$ is not identically equal to ∞ on Ω.

X denotes as before the set of feasible points in the problem (P), i.e.

$$X = \{x \in \Omega \mid g_1(x) \le 0, \ldots, g_p(x) \le 0, g_{p+1}(x) = 0, \ldots, g_m(x) = 0\},$$

and $v_{\min}(P)$ is the optimal value of the problem.

Definition. Let

$$\Lambda = \mathbf{R}_+^p \times \mathbf{R}^{m-p}.$$

The function $L\colon \Omega \times \Lambda \to \overline{\mathbf{R}}$, defined by

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x),$$

is called the *Lagrange function* of the minimization problem (P), and the variables $\lambda_1, \lambda_2, \ldots, \lambda_m$ are called *Lagrange multipliers*.

For each $x \in \operatorname{dom} f$, the expression $L(x, \lambda)$ is the sum of a real number and a linear form in $\lambda_1, \lambda_2, \ldots, \lambda_m$. Hence, the function $\lambda \mapsto L(x, \lambda)$ is affine (or rather, the restriction to Λ of an affine function on $\mathbf{R}^m$). The Lagrange function is thus especially *concave in the variable* λ for each fixed $x \in \operatorname{dom} f$.

If $x \in \Omega \setminus \operatorname{dom} f$, then obviously $L(x, \lambda) = \infty$ for all $\lambda \in \Lambda$. Hence,

$$\inf_{x \in \Omega} L(x, \lambda) = \inf_{x \in \operatorname{dom} f} L(x, \lambda) < \infty$$

for all $\lambda \in \Lambda$.

Definition. For $\lambda \in \Lambda$, we define

$$\phi(\lambda) = \inf_{x \in \Omega} L(x, \lambda)$$

and call the function $\phi \colon \Lambda \to \underline{\mathbf{R}}$ the *dual function* associated to the minimization problem (P).

It may of course happen that the domain

$$\operatorname{dom} \phi = \{\lambda \in \Lambda \mid \phi(\lambda > -\infty\}$$

of the dual function is empty; this occurs if the functions $x \mapsto L(x, \lambda)$ are unbounded below on Ω for all $\lambda \in \Lambda$.

Theorem 10.1.1. *The dual function ϕ of the minimization problem (P) is concave and*

$$\phi(\lambda) \leq v_{\min}(P)$$

for all $\lambda \in \Lambda$.

Hence, $\operatorname{dom} \phi = \emptyset$ if the objective function f in the original problem (P) is unbounded below on the constraint set, i.e. if $v_{\min}(P) = -\infty$.

Proof. The functions $\lambda \to L(x, \lambda)$ are concave for $x \in \operatorname{dom} f$, which means that the function ϕ is the infimum of a family of concave functions. It therefore follows from Theorem 6.2.4 that ϕ is concave.

Suppose $\lambda \in \Lambda$ and $x \in X$; then $\lambda_i g_i(x) \leq 0$ for $i \leq p$ and $\lambda_i g_i(x) = 0$ for $i > p$, and it follows that

$$L(x, \lambda) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x) \leq f(x),$$

and that consequently

$$\phi(\lambda) = \inf_{x \in \Omega} L(x, \lambda) \leq \inf_{x \in X} L(x, \lambda) \leq \inf_{x \in X} f(x) = v_{\min}(P). \qquad \square$$

The following optimality criterion is now an immediate consequence of the preceding theorem.

Theorem 10.1.2 (Optimality criterion)**.** *Suppose $\hat{x}$ is a feasible point for the minimization problem (P) and that there is a point $\hat{\lambda} \in \Lambda$ such that*

$$\phi(\hat{\lambda}) = f(\hat{x}).$$

Then $\hat{x}$ is an optimal solution.

Proof. The common value $f(\hat{x})$ belongs to the intersection $\overline{\mathbf{R}} \cap \underline{\mathbf{R}} = \mathbf{R}$ of the codomains of f and ϕ, and it is thus a real number, and by Theorem 10.1.1, $f(\hat{x}) \leq v_{\min}(P)$. Hence, $f(\hat{x}) = v_{\min}(P)$. □

Example 10.1.1. Let us consider the simple minimization problem

$$\begin{aligned} \min \quad & f(x) = x_1^2 - x_2^2 \\ \text{s.t.} \quad & x_1^2 + x_2^2 \leq 1. \end{aligned}$$

The Lagrange function is

$$\begin{aligned} L(x_1, x_2, \lambda) &= x_1^2 - x_2^2 + \lambda(x_1^2 + x_2^2 - 1) \\ &= (\lambda + 1)x_1^2 + (\lambda - 1)x_2^2 - \lambda \end{aligned}$$

with $(x_1, x_2) \in \mathbf{R}^2$ and $\lambda \in \mathbf{R}_+$.

The Lagrange function is unbounded below when $0 \leq \lambda < 1$, and it attains the minimum value $-\lambda$ for $x_1 = x_2 = 0$ when $\lambda \geq 1$, so the dual function ϕ is given by

$$\phi(\lambda) = \begin{cases} -\infty\,, & \text{if } 0 \leq \lambda < 1 \\ -\lambda\,, & \text{if } \lambda \geq 1. \end{cases}$$

We finally note that the optimality condition $\phi(\hat{\lambda}) = f(\hat{x})$ is satisfied by the point $\hat{x} = (0, 1)$ and the Lagrange multiplier $\hat{\lambda} = 1$. Hence, $(0, 1)$ is an optimal solution. □

The optimality criterion gives a sufficient condition for optimality, but it is not necessary, as the following trivial example shows.

Example 10.1.2. Consider the problem

$$\begin{aligned} \min \quad & f(x) = x \\ \text{s.t.} \quad & x^2 \leq 0. \end{aligned}$$

There is only one feasible point, $\hat{x} = 0$, which is therefore the optimal solution. The Lagrange function $L(x, \lambda) = x + \lambda x^2$ is bounded below for $\lambda > 0$ and

$$\phi(\lambda) = \inf_{x \in \mathbf{R}} (x + \lambda x^2) = \begin{cases} -1/4\lambda, & \text{if } \lambda > 0 \\ -\infty, & \text{if } \lambda = 0. \end{cases}$$

But $\phi(\lambda) < 0 = f(\hat{x})$ for all $\lambda \in \Lambda = \mathbf{R}_+$, so the optimality criterion in Theorem 10.1.2 is not satisfied by the optimal point. □

For the converse of Theorem 10.1.2 to hold, some extra condition is thus needed, and we describe such a condition in Chapter 11.1.

The dual problem

In order to obtain the best possible lower estimate of the optimal value of the minimization problem (P), we should, in the light of Theorem 10.1.1, maximize the dual function. This leads to the following definition.

Definition. The optimization problem

$$\text{(D)} \qquad \begin{aligned} &\max\ \phi(\lambda) \\ &\text{s.t.}\quad \lambda \in \Lambda \end{aligned}$$

is called the *dual problem* of the minimization problem (P).

The dual problem is a convex problem, irrespective of whether the problem (P) is convex or not, because the dual function is concave. The value of the dual problem will be denoted by $v_{\max}(D)$ with the usual conventions for $\pm\infty$-values.

Our next result is now an immediate corollary of Theorem 10.1.1.

Theorem 10.1.3 (Weak duality)**.** *The following inequality holds between the optimal values of the problem(P) and its dual problem (D):*

$$v_{\max}(D) \leq v_{\min}(P).$$

The inequality in the above theorem is called *weak duality.* If the two optimal values are equal, i.e. if

$$v_{\max}(D) = v_{\min}(P)$$

then we say that *strong duality* holds for problem (P).

Weak duality thus holds for all problems while strong duality only holds for special types of problems. Of course, strong duality prevails if the optimality criterion in Theorem 10.1.2 is satisfied.

EXAMPLE 10.1.3. Consider the minimization problem

$$\begin{aligned} &\min \ \ x_1^3 + 2x_2 \\ &\text{s.t.} \ \ x_1^2 + x_2^2 \leq 1. \end{aligned}$$

It is easily verified that the minimum is attained for $x = (0, -1)$ and that the optimal value is $v_{\min}(P) = -2$. The Lagrange function

$$L(x_1, x_2, \lambda) = x_1^3 + 2x_2 + \lambda(x_1^2 + x_2^2 - 1) = x_1^3 + \lambda x_1^2 + 2x_2 + \lambda x_2^2 - \lambda$$

tends, for each fixed $\lambda \geq 0$, to $-\infty$ as $x_2 = 0$ and $x_1 \to -\infty$. The Lagrange function is in other words unbounded below on $\mathbf{R}^2$ for each λ, and hence $\phi(\lambda) = -\infty$ for all $\lambda \in \Lambda$. The value of the dual problem is therefore $v_{\max}(D) = -\infty$, so strong duality does not hold in this problem. □

The Lagrange function, the dual function and the dual problem of a minimization problem of the type (P) are defined in terms of the constraint functions of the problem. Therefore, it may be worth emphasizing that problems that are equivalent in the sense that they have the same objective function f and the same set X of feasible points do not necessarily have equivalent dual problems. Thus, strong duality may hold for one way of framing a problem but fail to hold for other ways. See exercise 10.2.

EXAMPLE 10.1.4. Let us find the dual problem of the LP problem

$$\text{(LP-P)} \qquad \begin{aligned} &\min \ \ \langle c, x \rangle \\ &\text{s.t.} \ \ \begin{cases} Ax \geq b \\ \ \ x \geq 0. \end{cases} \end{aligned}$$

Here A is an $m \times n$-matrix, c is a vector in $\mathbf{R}^n$ and b a vector in $\mathbf{R}^m$. Let us rewrite the problem in the form

$$\begin{aligned} &\min \ \ \langle c, x \rangle \\ &\text{s.t.} \ \ \begin{cases} b - Ax \leq 0 \\ \ \ x \in \mathbf{R}_+^n \end{cases} \end{aligned}$$

with $x \in \mathbf{R}_+^n$ as an implicit constraint. The matrix inequality $b - Ax \leq 0$ consists of m linear inequalities, and the Lagrangefunction is therefore defined on the product set $\mathbf{R}_+^n \times \mathbf{R}_+^m$, and it is given by

$$L(x, \lambda) = \langle c, x \rangle + \langle \lambda, b - Ax \rangle = \langle c - A^T\lambda, x \rangle + \langle b, \lambda \rangle.$$

For fixed λ, $L(x, \lambda)$ is bounded below on the set $\mathbf{R}_+^n$ if and only if $c - A^T\lambda \geq 0$, with minimum value equal to $\langle b, \lambda \rangle$ attained at $x = 0$. The dual function $\phi\colon \mathbf{R}_+^m \to \underline{\mathbf{R}}$ is thus given by

$$\phi(\lambda) = \begin{cases} \langle b, \lambda \rangle, & \text{if } A^T\lambda \leq c \\ -\infty, & \text{otherwise.} \end{cases}$$

The dual problem to the LP problem (LP-P) is therefore also an LP problem, namely (after renaming the parameter λ to y) the LP problem

$$\text{(LP-D)} \qquad \begin{array}{ll} \max & \langle b, y\rangle \\ \text{s.t.} & \begin{cases} A^T y \le c \\ \quad y \ge 0. \end{cases} \end{array}$$

Note the beautiful symmetry between the two problems.

By weak duality, we know for sure that the optimal value of the maximization problem is less than or equal to the optimal value of the minimization problem. As we shall see later, strong duality holds for LP problems, i.e. the two problems above have the same optimal value, provided at least one of the problems has feasible points. □

We now return to the general minimization problem

$$\text{(P)} \qquad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & \begin{cases} g_i(x) \le 0, & i = 1, 2, \dots, p \\ g_i(x) = 0, & i = p+1, \dots, m \end{cases} \end{array}$$

with X as the set of feasible points, Lagrange function $L\colon \Omega \times \Lambda \to \overline{\mathbf{R}}$, and dual function ϕ. Our next theorem shows that the optimality criterion in Theorem 10.1.2 can be formulated as a saddle point condition on the Lagrange function.

Theorem 10.1.4. *Suppose $(\hat{x}, \hat{\lambda}) \in \Omega \times \Lambda$. The following three conditions are equivalent for the optimization problem (P):*

(i) $\hat{x} \in X$ and $f(\hat{x}) = \phi(\hat{\lambda})$, i.e. the optimality criterion is satisfied.

(ii) For all $(x, \lambda) \in \Omega \times \Lambda$,

$$L(\hat{x}, \lambda) \le L(\hat{x}, \hat{\lambda}) \le L(x, \hat{\lambda}),$$

i.e. $(\hat{x}, \hat{\lambda})$ is a saddle point *for the Lagrange function.*

(iii) $\hat{x} \in X$, $\hat{x}$ minimizes the function $x \mapsto L(x, \hat{\lambda})$ when x runs through Ω, and

$$\hat{\lambda}_i g_i(\hat{x}) = 0$$

for $i = 1, 2, \dots, p$.

Thus, $\hat{x}$ is an optimal solution to the problem (P) if any of the equivalent conditions (i)–(iii) is satisfied.

The condition in (iii) that $\hat{\lambda}_i g_i(\hat{x}) = 0$ for $i = 1, 2, \dots, p$ is called *complementarity*. An equivalent way to express this, which explains the name, is

$$\hat{\lambda}_i = 0 \quad \text{or} \quad g_i(\hat{x}) = 0.$$

A constraint with a positive Lagrange multiplier is thus necessarily active at the point $\hat{x}$.

Proof. (i) $\Rightarrow$ (ii): For $\hat{x} \in X$ and arbitrary $\lambda \in \Lambda$ $(= \mathbf{R}_+^p \times \mathbf{R}^{n-p})$ we have

$$L(\hat{x}, \lambda) = f(\hat{x}) + \sum_{i=1}^{m} \lambda_i g_i(\hat{x}) = f(\hat{x}) + \sum_{i=1}^{p} \lambda_i g_i(\hat{x}) \le f(\hat{x}),$$

since $\lambda_i \ge 0$ and $g_i(\hat{x}) \le 0$ for $i = 1, 2, \dots, p$. Moreover,

$$\phi(\hat{\lambda}) = \inf_{z \in \Omega} L(z, \hat{\lambda}) \le L(x, \hat{\lambda}) \qquad \text{for all } x \in \Omega.$$

If $f(\hat{x}) = \phi(\hat{\lambda})$, then consequently

$$L(\hat{x}, \lambda) \le f(\hat{x}) = \phi(\hat{\lambda}) \le L(x, \hat{\lambda})$$

for all $(x, \lambda) \in \Omega \times \Lambda$, and by the particular choice of $x = \hat{x}$, $\lambda = \hat{\lambda}$ in this inequality, we see that $f(\hat{x}) = L(\hat{x}, \hat{\lambda})$. This proves the saddle point inequality in (ii) with $L(\hat{x}, \hat{\lambda}) = f(\hat{x})$.

(ii) $\Rightarrow$ (iii): It is obvious that $\hat{x}$ minimizes the function $L(\,\cdot\,, \hat{\lambda})$ if and only if the right part of the saddle point inequality holds. The minimum value is moreover finite (due to our tacit assumption $\operatorname{dom} f \neq \emptyset$), and hence $f(\hat{x})$ is a finite number.

The left part of the saddlepoint inequality means that

$$f(\hat{x}) + \sum_{i=1}^{m} \lambda_i g_i(\hat{x}) \le f(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i g_i(\hat{x})$$

for all $\lambda \in \Lambda$, or equivalently that

$$\sum_{i=1}^{m} (\lambda_i - \hat{\lambda}_i) g_i(\hat{x}) \le 0$$

for all $\lambda \in \Lambda$.

Now fix the index k and choose in the above inequality the number λ so that $\lambda_i = \hat{\lambda}_i$ for all i except $i = k$. It follows that

$$(\lambda_k - \hat{\lambda}_k) g_k(\hat{x}) \le 0 \tag{10.1}$$

for all such λ.

If $k > p$, we choose $\lambda_k = \hat{\lambda}_k \pm 1$ with the conclusion that $\pm g_k(\hat{x}) \le 0$, i.e. that $g_k(\hat{x}) = 0$. For $k \le p$ we instead choose $\lambda_k = \hat{\lambda}_k + 1$, with the conclusion that $g_k(\hat{x}) \le 0$. Thus, $\hat{x}$ satisfies all the constraints, i.e. $\hat{x} \in X$.

For $k \leq p$ we finally choose $\lambda_k = 0$ and $\lambda_k = 2\hat{\lambda}_k$, respectively, in the inequality (10.1) with $\pm\hat{\lambda}_k g_k(\hat{x}) \leq 0$ as result. This means that $\hat{\lambda}_k g_k(\hat{x}) = 0$ for $k \leq p$, and the implication (ii) $\Rightarrow$ (iii) is now proved.

(iii) $\Rightarrow$ (i): From (iii) follows at once

$$\phi(\hat{\lambda}) = \inf_{x\in\Omega} L(x, \hat{\lambda}) = L(\hat{x}, \hat{\lambda}) = f(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i g_i(\hat{x}) = f(\hat{x}),$$

which is condition (i). □

If the objective and constraint functions f and $g_1, g_2, \ldots, g_m$ are differentiable, so is the Lagrange function $L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$, and we use $L'_x(x_0, \lambda)$ as the notation for the value of the derivative of the function $x \mapsto L(x, \lambda)$ at the point x_0, i.e.

$$L'_x(x_0, \lambda) = f'(x_0) + \sum_{i=1}^{m} \lambda_i g'_i(x_0).$$

If the differentiable function $x \mapsto L(x, \lambda)$ has a minimum at an interior point x_0 in Ω, then $L'_x(x_0, \lambda) = 0$. The following corollary is thus an immediate consequence of the implication (i) $\Rightarrow$ (iii) in Theorem 10.1.4.

Corollary 10.1.5. *Suppose that $\hat{x}$ is an optimal solution to the minimization problem (P), that $\hat{x}$ is an interior point of the domain Ω, that the objective and constraint functions are differentiable at $\hat{x}$, and that the optimality criterion $f(\hat{x}) = \phi(\hat{\lambda})$ is satisfied by some Lagrange multiplier $\hat{\lambda} \in \Lambda$. Then*

$$\text{(KKT)} \qquad \begin{cases} L'_x(\hat{x}, \hat{\lambda}) = 0 & \text{and} \\ \hat{\lambda}_i g_i(\hat{x}) = 0 & \text{for } i = 1, 2, \ldots, p. \end{cases}$$

The system (KKT) is called the *Karush–Kuhn–Tucker condition.*

The equality $L'_x(\hat{x}, \hat{\lambda}) = 0$ means that

$$f'(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i g'_i(\hat{x}) = 0,$$

which written out in more detail becomes

$$\begin{cases} \dfrac{\partial f}{\partial x_1}(\hat{x}) + \displaystyle\sum_{i=1}^{m} \hat{\lambda}_i \frac{\partial g_i}{\partial x_1}(\hat{x}) = 0 \\ \qquad\qquad \vdots \\ \dfrac{\partial f}{\partial x_n}(\hat{x}) + \displaystyle\sum_{i=1}^{m} \hat{\lambda}_i \frac{\partial g_i}{\partial x_n}(\hat{x}) = 0. \end{cases}$$

EXAMPLE 10.1.5. In Example 10.1.1 we found that $\hat{x} = (0, 1)$ is an optimal solution to the minimization problem

$$\begin{array}{ll} \min & x_1^2 - x_2^2 \\ \text{s.t.} & x_1^2 + x_2^2 \leq 1 \end{array}$$

and that the optimality criterion is satisfied with $\hat{\lambda} = 1$. The Lagrange function is $L(x, \lambda) = x_1^2 - x_2^2 + \lambda(x_1^2 + x_2^2 - 1)$, and indeed, $x = (0, 1)$ and $\lambda = 1$ satisfy the KKT-system

$$\begin{cases} \dfrac{\partial L(x,\lambda)}{\partial x_1} = 2(\lambda + 1)x_1 = 0 \\ \dfrac{\partial L(x,\lambda)}{\partial x_1} = 2(\lambda - 1)x_2 = 0 \\ \qquad\qquad \lambda(x_1^2 + x_2^2 - 1) = 0. \end{cases}$$

□

10.2 John's theorem

Conditions which guarantee that the KKT condition is satisfied at an optimal point, are usually called *constraint qualification conditions*, and in the next chapter we will describe such a condition for convex problems. In this section we will study a different qualifying condition, John's condition, for general optimization problems with constraints in the form of inequalities.

Let us therefore consider a problem of the form

$$\text{(P)} \qquad \begin{array}{lll} \min & f(x) & \\ \text{s.t.} & g_i(x) \leq 0, & i = 1, 2, \ldots, m \end{array}$$

with implicit constraint set Ω, i.e. domain for the objective and the constraint functions.

Whether a constraint is active or not at an optimal point plays a major role, and affine constraints are thereby easier to handle than other constraints. Therefore, we introduce the following notations:

$$\begin{aligned} I_{\text{aff}}(x) &= \{i \mid \text{the function } g_i \text{ is affine and } g_i(x) = 0\}, \\ I_{\text{oth}}(x) &= \{i \mid \text{the function } g_i \text{ is not affine and } g_i(x) = 0\}, \\ I(x) &= I_{\text{aff}}(x) \cup I_{\text{oth}}(x). \end{aligned}$$

So $I_{\text{aff}}(x)$ consists of the indices of all active affine constraints at the point x, $I_{\text{oth}}(x)$ consists of the indices of all other active constraints at the point, and $I(x)$ consists of the indices of all active constraints at the point.

Theorem 10.2.1 (John's theorem). *Suppose $\hat{x}$ is a local minimum point for the problem (P), that $\hat{x}$ is an interior point in Ω, and that the functions f and $g_1, g_2, \ldots, g_m$ are differentiable at the point $\hat{x}$. If there exists a vector $z \in \mathbf{R}^n$ such that*

$$\text{(J)} \qquad \begin{cases} \langle g_i'(\hat{x}), z\rangle \geq 0 & \text{for all } i \in I_{\text{aff}}(\hat{x}) \\ \langle g_i'(\hat{x}), z\rangle > 0 & \text{for all } i \in I_{\text{oth}}(\hat{x}), \end{cases}$$

then there exist Lagrange parameters $\hat{\lambda} \in \mathbf{R}_+^m$ such that

$$\text{(KKT)} \qquad \begin{cases} L_x'(\hat{x}, \hat{\lambda}) = 0 \\ \hat{\lambda}_i g_i(\hat{x}) = 0 \qquad \text{for } i = 1, 2, \ldots, m. \end{cases}$$

Remark 1. According to Theorem 3.3.5, the system (J) is solvable if and only if

$$\text{(J}') \qquad \begin{cases} \sum_{i \in I(\hat{x})} u_i g_i'(\hat{x}) = 0 \\ u \geq 0 \end{cases} \Rightarrow \quad u_i = 0 \quad \text{for all } i \in I_{\text{oth}}(\hat{x}).$$

The system (J) is thus in particular solvable if the gradient vectors $\nabla g_i(\hat{x})$ are linearly independent for $i \in I(\hat{x})$.

Remark 2. If $I_{\text{oth}}(\hat{x}) = \emptyset$, then (J) is trivially satisfied by $z = 0$.

Proof. Let Z denote the set of solutions to the system (J). The first part of the proof consists in showing that Z is a subset of the conic halfspace $\{z \in \mathbf{R}^n \mid -\langle f'(\hat{x}), z\rangle \geq 0\}$.

Assume therefore that $z \in Z$ and consider the halfline $\hat{x} - tz$ for $t \geq 0$. We claim that $\hat{x} - tz \in X$ for all sufficiently small $t > 0$.

If g is an affine function, i.e. has the form $g(x) = \langle c, x\rangle + b$, then $g'(x) = c$ and $g(x + y) = \langle c, x + y\rangle + b = \langle c, x\rangle + b + \langle c, y\rangle = g(x) + \langle g'(x), y\rangle$ for all x and y. Hence, for all indices $i \in I_{\text{aff}}(\hat{x})$,

$$g_i(\hat{x} - tz) = g_i(\hat{x}) - t\langle g_i'(\hat{x}), z\rangle = -t\langle g_i'(\hat{x}), z\rangle \leq 0$$

for all $t \geq 0$.

For indices $i \in I_{\text{oth}}(\hat{x})$, we obtain instead, using the chain rule, the inequality

$$\frac{d}{dt}\, g_i(\hat{x} - tz)|_{t=0} = -\langle g_i'(\hat{x}), z\rangle < 0.$$

The function $t \mapsto g_i(\hat{x} - tz)$ is in other words decreasing at the point $t = 0$, whence $g_i(\hat{x} - tz) < g_i(\hat{x}) = 0$ for all sufficiently small $t > 0$.

If the i:th constraint is inactive at $\hat{x}$, i.e. if $i \notin I(\hat{x})$, then $g_i(\hat{x}) < 0$, and it follows from continuity that $g_i(\hat{x} - tz) < 0$ for all sufficiently small $t > 0$.

We have thus proved that the points $\hat{x} - tz$ belong to the constraint set X if $t > 0$ is sufficiently small. Since $\hat{x}$ is a local minimum point of f, it follows that $f(\hat{x} - tz) \geq f(\hat{x})$ for all sufficiently small $t > 0$. Consequently,

$$-\langle f'(\hat{x}), z\rangle = \frac{d}{dt}\, f(\hat{x} - tz)\big|_{t=0} = \lim_{t\to 0^+} \frac{f(\hat{x} - tz) - f(\hat{x})}{t} \geq 0.$$

This proves the alleged inclusion

$$Z \subseteq \{z \in \mathbf{R}^n \mid -\langle f'(\hat{x}), z\rangle \geq 0\} = \{-f'(\hat{x})\}^+ = \big(\mathrm{con}\{-f'(\hat{x})\}\big)^+,$$

and it now follows from Theorem 3.2.1, Corollary 3.2.4 and Theorem 3.3.4 that

$$\mathrm{con}\{-f'(\hat{x})\} \subseteq Z^+ = \mathrm{con}\{g_i'(\hat{x}) \mid i \in I(\hat{x})\}.$$

So the vector $-f'(\hat{x})$ belongs to the cone generated by the vektors $g_i'(\hat{x})$, $i \in I(\hat{x})$, which means that there are nonnegative integers $\hat{\lambda}_i$, $i \in I(\hat{x})$, such that

$$-f'(\hat{x}) = \sum_{i\in I(\hat{x})} \hat{\lambda}_i g_i'(\hat{x}).$$

If we finally define $\hat{\lambda}_i = 0$ for $i \notin I(\hat{x})$, then

$$f'(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i g_i'(\hat{x}) = 0$$

and $\hat{\lambda}_i g_i(\hat{x}) = 0$ for $i = 1, 2, \ldots, m$. This means that the KKT-condition is satisfied. □

The condition in John's statement that the system (J) has a solution can be replaced with other qualifying constraints but can not be completely removed without the conclusion being lost. This is shown by the following example.

Example 10.2.1. Consider the problem

$$\begin{array}{ll} \min & f(x) = x_1 \\ \text{s.t.} & \begin{cases} g_1(x) = -x_1^3 + x_2 \leq 0 \\ g_2(x) = \qquad -x_2 \leq 0 \end{cases} \end{array}$$

with Lagrange function $L(x, \lambda) = x_1 + \lambda_1(x_2 - x_1^3) - \lambda_2 x_2$. The unique optimal solution is $\hat{x} = (0, 0)$, but the system $L_x'(\hat{x}, \lambda) = 0$, i.e.

$$\begin{cases} 1 = 0 \\ \lambda_1 - \lambda_2 = 0, \end{cases}$$

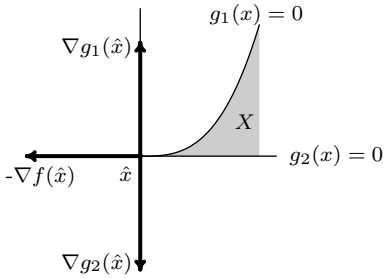

Figure 10.1. Illustration for Example 10.2.1: The vector $-\nabla f(\hat{x})$ does not belong to the cone generated by the gradients $\nabla g_1(\hat{x})$ and $\nabla g_2(\hat{x})$.

has no solutions. This is explained by the fact that the system (J), i.e.

$$\begin{cases} -z_2 \geq 0 \\ z_2 > 0, \end{cases}$$

has no solutions. □

Example 10.2.2. We will solve the problem

$$\begin{array}{ll} \min & x_1x_2 + x_3 \\ \text{s.t.} & \begin{cases} 2x_1 - 2x_2 + x_3 + 1 \leq 0 \\ x_1^2 + x_2^2 - x_3 \leq 0 \end{cases} \end{array}$$

using John's theorem. Note first that the constraints define a compact set X, for the inequalities

$$x_1^2 + x_2^2 \leq x_3 \leq -2x_1 + 2x_2 - 1$$

imply that $(x_1 + 1)^2 + (x_2 - 1)^2 \leq 1$, and consequently, $-2 \leq x_1 \leq 0$, $0 \leq x_2 \leq 2$, and $0 \leq x_3 \leq 7$. Since the objective function is continuous, there is indeed an optimal solution.

Let us now first investigate whether the system (J) is solvable. We use the equivalent version (J′) in the remark after the theorem. First note that the gradients of the constraint functions are never equal to zero. The condition (J′) is thus met in the points where only one of the constraints is active.

Assume therefore that x is a point where $I(x) = \{1, 2\}$, i.e. where both constraints are active, and that $u_1(2, -2, 1) + u_2(2x_1, 2x_2, -1) = (0, 0, 0)$. If $u_2 > 0$, we conclude from the above equation that $u_1 = u_2$, $x_1 = -1$ and $x_2 = 1$. Inserting $x_1 = -1$ and $x_2 = 1$ into the two active constraints yields $x_3 = 3$ and $x_3 = 2$, respectively, which is contradictory. Thus, $u_2 = 0$, which means that the condition (J′) is fulfilled at all feasible points.

We conclude that the optimal point satisfies the KKT-condition, which in this instance is as follows

$$\begin{cases} x_2 + 2\lambda_1 + 2x_1\lambda_2 = 0 & \text{(i)} \\ x_1 - 2\lambda_1 + 2x_2\lambda_2 = 0 & \text{(ii)} \\ 1 + \lambda_1 - \lambda_2 = 0 & \text{(iii)} \\ \lambda_1(2x_1 - 2x_2 + x_3 + 1) = 0 & \text{(iv)} \\ \lambda_2(x_1^2 + x_2^2 - x_3) = 0 & \text{(v)} \end{cases}$$

The further investigation is divided into two cases.

$\lambda_1 = 0$: Equation (iii) implies that $\lambda_2 = 1$, which inserted into (i) and (ii) gives $x_1 = x_2 = 0$, and from (v) now follows $x_3 = 0$. But this is a false solution, since $(0, 0, 0) \notin X$.

$\lambda_1 > 0$: Equation (iv) now implies that

$$2x_1 - 2x_2 + x_3 + 1 = 0. \qquad \text{(vi)}$$

From (i) and (ii) follows $(x_1 + x_2)(1 + 2\lambda_2) = 0$, and since $\lambda_2 \geq 0$,

$$x_1 + x_2 = 0. \qquad \text{(vii)}$$

By (iii,) $\lambda_2 > 0$. Condition (v) therefore implies that

$$x_1^2 + x_2^2 - x_3 = 0. \qquad \text{(viii)}$$

The system consisting of equations (vi), (vii), (viii) has two solutions, namely

$$\hat{x} = (-1+\sqrt{1/2}, 1-\sqrt{1/2}, 3-2\sqrt{2}) \text{ and } \overline{x} = (-1-\sqrt{1/2}, 1+\sqrt{1/2}, 3+2\sqrt{2}).$$

Using (i) and (iii), we compute the corresponding λ and obtain

$$\hat{\lambda} = (-1/2 + \sqrt{1/2}, 1/2 + \sqrt{1/2}) \text{ and } \overline{\lambda} = (-1/2 - \sqrt{1/2}, 1/2 - \sqrt{1/2}),$$

respectively. Note that $\hat{\lambda} \geq 0$ and $\overline{\lambda} < 0$. The system KKT thus has a unique solution (x, λ) with $\lambda \geq 0$, namely $x = \hat{x}$, $\lambda = \hat{\lambda}$. By John's theorem, $\hat{x}$ is the unique optimal solution of our minimization problem, and the optimal value is $3/2 - \sqrt{2}$. □

Exercises

10.1 Determine the dual function for the optimization problem

$$\begin{array}{ll} \min & x_1^2 + x_2^2 \\ \text{s.t.} & x_1 + x_2 \geq 2, \end{array}$$

and prove that $(1, 1)$ is an optimal solution by showing that the optimality criterion is satisfied by $\hat{\lambda} = 2$. Also show that the KKT-condition is satisfied at the optimal point.

10.2 Consider the two minimization problems

$$(P_a)\quad \begin{array}{l}\min\ e^{-x_1}\\ x_1^2/x_2 \le 0\end{array} \qquad \text{and} \qquad (P_b)\quad \begin{array}{l}\min\ e^{-x_1}\\ |x_1| \le 0\end{array}$$

both with $\Omega = \{(x_1, x_2) \mid x_2 > 0\}$ as implicit domain. The two problems have the same set $X = \{(0, x_2) \mid x_2 > 0\}$ of feasible points and the same optimal value $v_{\min} = 1$. Find their dual functions and dual problems, and show that strong duality holds for (P_b) but not for (P_a).

10.3 Suppose the function $f\colon X \times Y \to \mathbf{R}$ has two saddle points $(\hat{x}_1, \hat{y}_1)$ and $(\hat{x}_2, \hat{y}_2)$. Prove that

a) $f(\hat{x}_1, \hat{y}_1) = f(\hat{x}_2, \hat{y}_2)$;

b) $(\hat{x}_1, \hat{y}_2)$ and $(\hat{x}_2, \hat{y}_1)$ are saddle points, too.

10.4 Let $f\colon X \times Y \to \mathbf{R}$ be an arbitrary function.

a) Prove that

$$\sup_{y\in Y} \inf_{x\in X} f(x, y) \le \inf_{x\in X} \sup_{y\in Y} f(x, y).$$

b) Suppose there is a point $(\hat{x}, \hat{y}) \in X \times Y$ such that

$$\sup_{y\in Y} \inf_{x\in X} f(x, y) = \inf_{x\in X} f(x, \hat{y}) \quad \text{and} \quad \inf_{x\in X} \sup_{y\in Y} f(x, y) = \sup_{y\in Y} f(\hat{x}, y).$$

Prove that $(\hat{x}, \hat{y})$ is a saddle point of the function f if and only if

$$\inf_{x\in X} f(x, \hat{y}) = \sup_{y\in Y} f(\hat{x}, y),$$

and that the common value then is equal to $f(\hat{x}, \hat{y})$.

10.5 Consider a minimization problem

$$\begin{array}{ll}\min & f(x)\\ \text{s.t.} & g_i(x) \le 0, \qquad i = 1, 2, \dots, m\end{array}$$

with *convex* differentiable constraint functions $g_1, g_2, \dots, g_m$, and suppose there is a point $x_0 \in X = \{x \mid g_1(x) \le 0, \dots, g_m(x) \le 0\}$ which satisfies all non-affine constraints with strict inequality. Show that the system (J) is solvable at all points $\hat{x} \in X$.

[Hint: Show that $z = \hat{x} - x_0$ satisfies (J).]

10.6 Solve the following optimization problems

a) $\min\ x_1^3 + x_1 x_2^2$ s.t. $\begin{cases} x_1^2 + 2x_2^2 \le 1 \\ x_2 \ge 0 \end{cases}$

b) $\max\ x_1^2 + x_2^2 + \arctan x_1 x_2$ s.t. $\begin{cases} x_1^2 + x_2^2 \le 2 \\ 0 \le x_1 \le x_2 \end{cases}$

c) $\min\ x_1 x_2$ s.t. $\begin{cases} x_1^2 + x_1 x_2 + 4x_2^2 \le 1 \\ x_1 + 2x_2 \ge 0 \end{cases}$

d) $\max\ x_1^2 x_2 x_3$ s.t. $\begin{cases} 2x_1 + x_1 x_2 + x_3 \le 1 \\ x_1,\ x_2,\ x_3 \ge 0. \end{cases}$

Chapter 11

Convex optimization

11.1 Strong duality

We recall that the minimization problem

$$\text{(P)} \qquad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & \begin{cases} g_i(x) \le 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m \end{cases} \end{array}$$

is called convex if

- the implicit constraint set Ω is convex,
- the objective function f is convex,
- the constraint functions g_i are convex for $i = 1, 2, \ldots, p$ and affine for $i = p+1, \ldots, m$.

The set X of feasible points is convex in a convex optimization problem, and the Lagrange function

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$$

is convex in the variable x for each fixed $\lambda \in \Lambda = \mathbf{R}_+^p \times \mathbf{R}^{m-p}$, since it is a conic combination of convex functions.

We have already noted that the optimality criterion in Theorem 10.1.2 need not be fulfilled at an optimal point, not even for convex problems, because of the trivial counterexample in Example 10.1.2. For the criterion to be met some additional condition is needed, and a weak one is given in the next definition.

Definition. The problem (P) satisfies *Slater's condition* if there is a feasible point $\overline{x}$ in the relative interior of Ω such that $g_i(\overline{x}) < 0$ for each *non-affine* constraint function g_i.

Slater's condition is of course vacously fulfilled if all constraint functions are affine.

For convex problems that satisfy Slater's condition, the optimality criterion is both sufficient and necessary for optimality. We have namely the following result.

Theorem 11.1.1 (Duality theorem)**.** *Suppose that the problem (P) is convex and satisfies Slater's condition, and that the optimal value $v_{\min}$ is finite. Let $\phi\colon \Lambda \to \underline{\mathbf{R}}$ denote the dual function of the problem. Then there is a point $\hat{\lambda} \in \Lambda$ such that*

$$\phi(\hat{\lambda}) = v_{\min}.$$

Proof. First suppose that all constraints are inequalities, i.e. that $p = m$, and renumber the constraints so that the functions g_i are convex and non-affine for $i = 1, 2, \ldots, k$ and affine for $i = k+1, \ldots, m$.

Because of Slater's condition, the system

$$\begin{cases} g_i(x) < 0, & i = 1, 2, \ldots, k \\ g_i(x) \leq 0, & i = k+1, \ldots, m \end{cases}$$

has a solution in the relative interior of Ω, whereas the system

$$\begin{cases} f(x) - v_{\min} < 0 & \\ g_i(x) < 0, & i = 1, 2, \ldots, k \\ g_i(x) \leq 0, & i = k+1, \ldots, m \end{cases}$$

lacks solutions in Ω, due to the definition of $v_{\min}$. Therefore, it follows from Theorem 6.5.1 that there exist nonnegative scalars $\hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_m$ such that at least one of the numbers $\hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_k$ is positive and

$$\hat{\lambda}_0(f(x) - v_{\min}) + \hat{\lambda}_1 g_1(x) + \hat{\lambda}_2 g_2(x) + \cdots + \hat{\lambda}_m g_m(x) \geq 0$$

for all $x \in \Omega$. Here, the coefficient $\hat{\lambda}_0$ has to be positive, because if $\hat{\lambda}_0 = 0$ then $\hat{\lambda}_1 g_1(x) + \cdots + \hat{\lambda}_m g_m(x) \geq 0$ for all $x \in \Omega$, which contradicts the fact that the first mentioned system of inequalities has a solution in Ω. We may therefore assume, by dividing by $\hat{\lambda}_0$ if necessary, that $\hat{\lambda}_0 = 1$, and this gives us the inequality

$$L(x, \hat{\lambda}) = f(x) + \sum_{i=1}^{m} \hat{\lambda}_i g_i(x) \geq v_{\min}$$

for all $x \in \Omega$. It follows that

$$\phi(\hat{\lambda}) = \inf_{x \in \Omega} L(x, \hat{\lambda}) \geq v_{\min},$$

which combined with Theorem 10.1.1 yields the desired equality $\phi(\hat{\lambda}) = v_{\min}$.

If the problem has affine equality constraints, i.e. if $p < m$, we replace each equality $g_i(x) = 0$ with the two inequalities $\pm g_i(x) \leq 0$, and it follows from the already proven case of the theorem that there exist nonnegative Lagrange multipliers $\hat{\lambda}_1, \ldots, \hat{\lambda}_p, \hat{\mu}_{p+1}, \ldots, \hat{\mu}_m, \hat{\nu}_{p+1}, \ldots, \hat{\nu}_m$ such that

$$f(x) + \sum_{i=1}^{p} \hat{\lambda}_i g_i(x) + \sum_{i=p+1}^{m} (\hat{\mu}_i - \hat{\nu}_i) g_i(x) \geq v_{\min}$$

for all $x \in \Omega$, By defining $\hat{\lambda}_i = \hat{\mu}_i - \hat{\nu}_i$ for $i = p+1, \ldots, m$, we obtain a point $\hat{\lambda} \in \Lambda = \mathbf{R}_+^p \times \mathbf{R}^{m-p}$ which satisfies $\phi(\hat{\lambda}) \geq v_{\min}$, and this completes the proof of the theorem. □

By combining Theorem 11.1.1 with Theorem 10.1.2 we get the following corollary.

Corollary 11.1.2. *Suppose that the problem (P) is convex and that it satisfies Slater's condition. Then, a feasible point $\hat{x}$ is optimal if and only if it satisfies the optimality criterion, i.e. if and only if there exists a $\hat{\lambda} \in \Lambda$ such that $\phi(\hat{\lambda}) = f(\hat{x})$.*

11.2 The Karush–Kuhn–Tucker theorem

Variants of the following theorem were first proved by Karush and Kuhn–Tucker, and the theorem is therefore usually called the Karush–Kuhn–Tucker theorem.

Theorem 11.2.1. *Let*

$$(P) \qquad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & \begin{cases} g_i(x) \leq 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m \end{cases} \end{array}$$

be a convex problem, and suppose that the objective and constraint functions are differentiable at the feasible point $\hat{x}$.

(i) If $\hat{\lambda}$ is a point in Λ and the pair $(\hat{x}, \hat{\lambda})$ satisfies the KKT-condition

$$\begin{cases} L'_x(\hat{x}, \hat{\lambda}) = 0 \\ \hat{\lambda}_i g_i(\hat{x}) = 0 \qquad \text{for } i = 1, 2, \ldots, p \end{cases}$$

then strong duality prevails; $\hat{x}$ is an optimal solution to the problem (P) and $\hat{\lambda}$ is an optimal solution to the dual problem.

(ii) Conversely, if Slater's condition is fulfilled and $\hat{x}$ is an optimal solution, then there exist Lagrange multipliers $\hat{\lambda} \in \Lambda$ such that $(\hat{x}, \hat{\lambda})$ satisfies the KKT-condition.

Proof. (i) The KKT-condition implies that $\hat{x}$ is a stationary point of the convex function $x \mapsto L(x, \hat{\lambda})$, and an interior stationary point of a convex function is a minimum point, according to Theorem 7.2.2. Condition (iii) in Theorem 10.1.4 is thus fulfilled, and this means that the optimality criterion is satisfied by the pair $(\hat{x}, \hat{\lambda})$.

(ii) Conversely, if Slater's condition is satisfied and $\hat{x}$ is an optimal solution, then the optimality criterion $f(\hat{x}) = \phi(\hat{\lambda})$ is satisfied by some $\hat{\lambda} \in \Lambda$, according to Theorem 11.1.1. The KKT-condition is therefore met because of Corollary 10.1.5. □

The KKT-condition has a natural geometrical interpretation. Assume for simplicity that all constraints are inequalities, i.e. that $p = m$, and let $I(\hat{x})$ denote the index set for the constraints that are active at the optimal point $\hat{x}$. The KKT-condition means that $\hat{\lambda}_i = 0$ for all indices $i \notin I(\hat{x})$ and that

$$-\nabla f(\hat{x}) = \sum_{i \in I(\hat{x})} \hat{\lambda}_i \nabla g_i(\hat{x}),$$

where all coefficients $\hat{\lambda}_i$ occuring in the sum are nonnegative. The geometrical meaning of the above equality is that the vector $-\nabla f(\hat{x})$ belongs to the cone generated by the gradients $\nabla g_i(\hat{x})$ of the active inequality constraints. Cf. figure 11.1 and figure 11.2.

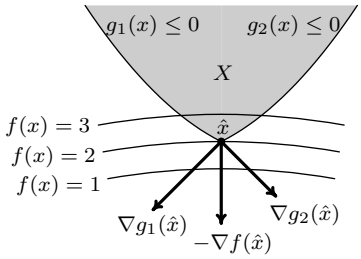

Figure 11.1. The point $\hat{x}$ is optimal since both constraints are active at the point and $-\nabla f(\hat{x}) \in \text{con}\{\nabla g_1(\hat{x}), \nabla g_2(\hat{x})\}$.

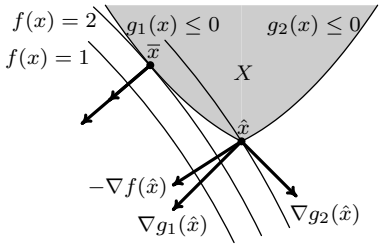

Figure 11.2. Here the point $\hat{x}$ is not optimal since $-\nabla f(\hat{x}) \notin \text{con}\{\nabla g_1(\hat{x}), \nabla g_2(\hat{x})\}$. The optimum is instead attained at $\overline{x}$, where $-\nabla f(\overline{x}) = \lambda_1 \nabla g_1(\overline{x})$ for some $\lambda_1 > 0$.

EXAMPLE 11.2.1. Consider the problem

$$\begin{cases} \min\ \mathrm{e}^{x_1-x_3}+\mathrm{e}^{-x_2} \\ (x_1-x_2)^2-x_3\le 0 \\ x_3-4\le 0. \end{cases}$$

The objective and the constraint functions are convex. Slater's condition is satisfied, since for instance $(1,1,1)$ satisfies both constraints strictly. According to Theorem 11.2.1, x is therefore an optimal solution to the problem if and only if x solves the system

$$\begin{cases} \mathrm{e}^{x_1-x_3}+2\lambda_1(x_1-x_2)=0 & \text{(i)} \\ -\mathrm{e}^{-x_2}-2\lambda_1(x_1-x_2)=0 & \text{(ii)} \\ -\mathrm{e}^{x_1-x_3}-\lambda_1+\lambda_2=0 & \text{(iii)} \\ \lambda_1\big((x_1-x_2)^2-x_3\big)=0 & \text{(iv)} \\ \lambda_2(x_3-4)=0 & \text{(v)} \\ \lambda_1,\ \lambda_2\ge 0 & \text{(vi)} \end{cases}$$

It follows from (i) and (vi) that $\lambda_1>0$, from (iii) and (vi) that $\lambda_2>0$, and from (iv) and (v) that $x_3=4$ and $x_1-x_2=\pm 2$. But $x_1-x_2<0$, because of (i) and (vi), and hence $x_1-x_2=-2$. By comparing (i) and (ii) we see that $x_1-x_3=-x_2$, i.e. $x_1+x_2=4$. It follows that $x=(1,3,4)$ and $\lambda=(\mathrm{e}^{-3}/4,5\mathrm{e}^{-3}/4)$ is the unique solution of the system. The problem therefore has a unique optimal solution, namely $(1,3,4)$. The optimal value is equal to $2\mathrm{e}^{-3}$. □

11.3 The Lagrange multipliers

In this section we will study how the optimal value $v_{\min}(b)$ of an arbitrary minimization problem of the type

$$(\mathrm{P}_b)\qquad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & \begin{cases} g_i(x)\le b_i, & i=1,2,\dots,p \\ g_i(x)=b_i, & i=p+1,\dots,m \end{cases} \end{array}$$

depends on the constraint parameters $b_1,b_2,\dots,b_m$. The functions f and $g_1,g_2,\dots,g_m$ are, as previously, defined on a subset Ω of $\mathbf{R}^n$, $b=(b_1,\dots,b_m)$ is a vector in $\mathbf{R}^m$, and

$$X(b)=\{x\in\Omega \mid g_i(x)\le b_i \text{ for } 1\le i\le p \text{ and } g_i(x)=b_i \text{ for } p<i\le m\}$$

is the set of feasible points.

The Lagrange function and the dual function associated to the minimization problem (P_b) are denoted by L_b and ϕ_b, respectively. By definition,

$$L_b(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i (g_i(x) - b_i),$$

and the relationship between the Lagrange functions L_b and $L_{\overline{b}}$ belonging to two different parameter vectors b and $\overline{b}$, is therefore given by the equation

$$L_b(x,\lambda) = L_{\overline{b}}(x,\lambda) + \sum_{i=1}^{m} \lambda_i (\overline{b}_i - b_i) = L_{\overline{b}}(x,\lambda) + \langle \lambda, \overline{b} - b \rangle.$$

By forming the infimum over $x \in \Omega$, we immediately get the following relation for the dual functions:

$$\phi_b(\lambda) = \phi_{\overline{b}}(\lambda) + \langle \lambda, \overline{b} - b \rangle. \tag{11.1}$$

The following theorem gives an interpretation of the Lagrange parameters in problems which satisfy the optimality criterion in Theorem 10.1.2, and thus especially for convex problems which satisfy Slater's condition.

Theorem 11.3.1. *Suppose that the minimization problem $(P_{\overline{b}})$ has an optimal solution $\overline{x}$ and that the optimality criterion is satisfied at the point, i.e. that there are Lagrange multipliers $\overline{\lambda}$ such that $\phi_{\overline{b}}(\overline{\lambda}) = f(\overline{x})$. Then:*

(i) The objective function f is bounded below on $X(b)$ for each $b \in \mathbf{R}^m$, so the optimal value $v_{\min}(b)$ of problem (P_b) is finite if the set $X(b)$ of feasible points is nonempty, and equal to $+\infty$ if $X(b) = \emptyset$.

(ii) The vector $-\overline{\lambda}$ is a subgradient at the point $\overline{b}$ of the optimal value function $v_{\min} \colon \mathbf{R}^m \to \overline{\mathbf{R}}$.

(iii) Suppose that the optimality criterion is satisfied in the problem (P_b) for all b in an open convex set U. The restriction of the function $v_{\min}$ to U is then a convex function.

Proof. By using weak duality for problem (P_b), the identity (11.1) and the optimality criterion for problem $(P_{\overline{b}})$, we obtain the following inequality:

$$\begin{aligned} v_{\min}(b) &= \inf_{x \in X(b)} f(x) \geq \phi_b(\overline{\lambda}) = \phi_{\overline{b}}(\overline{\lambda}) + \langle \overline{\lambda}, \overline{b} - b \rangle = f(\overline{x}) + \langle \overline{\lambda}, \overline{b} - b \rangle \\ &= v_{\min}(\overline{b}) - \langle \overline{\lambda}, b - \overline{b} \rangle. \end{aligned}$$

It follows, first, that the optimal value $v_{\min}(b)$ can not be equal to $-\infty$, and second, that $-\overline{\lambda}$ is a subgradient of the function $v_{\min}$ at the point $\overline{b}$.

If the optimality criterion is satisfied at all $b \in U$, then $v_{\min}$ has a subgradient at all points in U, and such a function is convex. □

Now suppose that the function $v_{\min}$ is differentiable at the point $\overline{b}$. The gradient at the point $\overline{b}$ is then, by Theorem 8.1.3, the unique subgradient at the point, so it follows from (ii) in the above theorem that $v'_{\min}(\overline{b}) = -\overline{\lambda}$. This gives us the approximation

$$v_{\min}(\overline{b}_1 + \Delta b_1, \ldots, \overline{b}_m + \Delta b_m) \approx v_{\min}(\overline{b}_1, \ldots, \overline{b}_m) - \overline{\lambda}_1 \Delta b_1 \cdots - \overline{\lambda}_m \Delta b_m$$

for small increments Δb_j. So the Lagrange multipliers provide information about how the optimal value is affected by small changes in the parameters.

EXAMPLE 11.3.1. As an illustration of Theorem 11.3.1, let us study the convex problem

$$\begin{aligned} &\min \;\; x_1^2 + x_2^2 \\ &\text{s.t.} \begin{cases} \;\; x_1 + 2x_2 \le b_1 \\ 2x_1 + \;\; x_2 \le b_2. \end{cases} \end{aligned}$$

Since it is about minimizing the distance squared from the origin to a polyhedron, there is certainly an optimal solution for each right-hand side b, and since the constraints are affine, it follows from the Karush–Kuhn–Tucker theorem that the optimal solution satisfies the KKT-condition, which in the present case is the system

$$\begin{cases} \;\; 2x_1 + \;\; \lambda_1 + 2\lambda_2 = 0 & \text{(i)} \\ \;\; 2x_2 + 2\lambda_1 + \;\; \lambda_2 = 0 & \text{(ii)} \\ \lambda_1(x_1 + 2x_2 - b_1) = 0 & \text{(iii)} \\ \lambda_2(2x_1 + x_2 - b_2) = 0 & \text{(iv)} \\ \qquad\qquad \lambda_1,\, \lambda_2 \ge 0. & \end{cases}$$

We now solve this system by considering four separate cases:

$\lambda_1 = \lambda_2 = 0$: In this case, $x_1 = x_2 = 0$ is the unique solution to the KKT-system. Thus, the point $(0, 0)$ is optimal provided it is feasible, and so is the case if and only if $b_1 \ge 0$ and $b_2 \ge 0$. The optimal value for these parameter values is $v_{\min}(b) = 0$.

$\lambda_1 > 0$, $\lambda_2 = 0$: From (i) and (ii), it follows first that $x_2 = 2x_1 = -\lambda_1$, and (iii) then gives $x = \frac{1}{5}(b_1, 2b_1)$. This point is feasible if $2x_1 + x_2 = \frac{4}{5}b_1 \le b_2$, and for the Lagrange multiplier $\lambda_1 = -\frac{2}{5}b_1$ to be positive, we must also have $b_1 < 0$. Thus, the point $x = \frac{1}{5}(b_1, 2b_1)$ is optimal if $b_1 < 0$ and $4b_1 \le 5b_2$, and the corresponding value is $v_{\min}(b) = \frac{1}{5}b_1^2$.

$\lambda_1 = 0$, $\lambda_2 > 0$: From (i) and (ii), it now follows that $x_1 = 2x_2 = -\lambda_2$, which inserted into (iv) gives $x = \frac{1}{5}(2b_2, b_2)$. This is a feasible point if $x_1 + 2x_2 = \frac{4}{5}b_2 \le b_1$. The Lagrange multiplier $\lambda_2 = -\frac{2}{5}b_2$ is positive if $b_2 < 0$.

Hence, the point $x = \frac{1}{5}(2b_2, b_2)$ is optimal and the optimal value is $v(b) = \frac{1}{5}b_2^2$, if $b_2 < 0$ och $4b_2 \leq 5b_1$.

$\lambda_1 > 0$, $\lambda_2 > 0$: By solving the subsystem obtained from (iii) and (iv), we get $x = \frac{1}{3}(2b_2 - b_1, 2b_1 - b_2)$, and the equations (i) and (ii) then result in $\lambda = \frac{2}{9}(4b_2 - 5b_1, 4b_1 - 5b_2)$. The two Lagrange multipliers are positive if $\frac{5}{4}b_1 < b_2 < \frac{4}{5}b_1$. For these parameter values, x is the optimal point and $v_{\min}(b) = \frac{1}{9}(5b_1^2 - 8b_1b_2 + 5b_2^2)$ is the optimal value.

The result of our investigation is summarized in the following table:

	$v_{\min}(b)$	$-\lambda_1 = \dfrac{\partial v}{\partial b_1}$	$-\lambda_2 = \dfrac{\partial v}{\partial b_2}$
$b_1 \geq 0,\ b_2 \geq 0$	0	0	0
$b_1 < 0,\ b_2 \geq \frac{4}{5}b_1$	$\frac{1}{5}b_1^2$	$\frac{2}{5}b_1$	0
$b_2 < 0,\ b_2 \leq \frac{5}{4}b_1$	$\frac{1}{5}b_2^2$	0	$\frac{2}{5}b_2$
$\frac{5}{4}b_1 < b_2 < \frac{4}{5}b_1$	$\frac{1}{9}(5b_1^2 - 8b_1b_2 + 5b_2^2)$	$\frac{2}{9}(5b_1 - 4b_2)$	$\frac{2}{9}(5b_2 - 4b_1)$

□

Exercises

11.1 Let $b > 0$ and consider the following trivial convex optimization problem

$$\begin{aligned} \min\quad & x^2 \\ \text{s.t.}\quad & x \geq b. \end{aligned}$$

Slater's condition is satisfied and the optimal value is attained at the point $\hat{x} = b$. Find the number $\hat{\lambda}$ which, according to Theorem 11.1.1, satisfies the optimality criterion.

11.2 Verify in the previous exercise that $v'(b) = \hat{\lambda}$.

11.3 Consider the minimization problem

$$\text{(P)} \qquad \begin{aligned} \min\quad & f(x) \\ \text{s.t.}\quad & \begin{cases} g_i(x) \leq 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m \end{cases} \end{aligned}$$

with $x \in \Omega$ as implicit constraint, and the equivalent epigraph formulation

$$\text{(P}') \qquad \begin{aligned} \min\quad & t \\ \text{s.t.}\quad & \begin{cases} f(x) - t \leq 0, & \\ \quad g_i(x) \leq 0, & i = 1, 2, \ldots, p \\ \quad g_i(x) = 0, & i = p+1, \ldots, m \end{cases} \end{aligned}$$

of the problem with $(t, x) \in \mathbf{R} \times \Omega$ as implicit constraint.

a) Show that (P′) satisfies Slater's condition if and only if (P) does.

b) Determine the relation between the Lagrange functions of the two problems and the relation between their dual functions.

c) Prove that the two dual problems have the same optimal value, and that the optimality criterion is satisfied in the minimization problem (P) if and only if it is satisfied in the problem (P′).

11.4 Prove for convex problems that Slater's condition is satisfied if and only if, for each non-affine constraint $g_i(x) \leq 0$, there is a feasible point $\overline{x}_i$ in the relative interior of Ω such that $g_i(\overline{x}_i) < 0$.

11.5 Let

$$(\mathrm{P}_b) \qquad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & \begin{cases} g_i(x) \leq b_i, & i = 1, 2, \ldots, p \\ g_i(x) = b_i, & i = p+1, \ldots, m \end{cases} \end{array}$$

be a convex problem, and suppose that its optimal value $v_{\min}(b)$ is $> -\infty$ for all right-hand sides b that belong to some convex subset U of $\mathbf{R}^m$. Prove that the restriction of $v_{\min}$ to U is a convex function.

11.6 Solve the following convex optimization problems.

a) $\begin{array}{ll} \min & \mathrm{e}^{x_1 - x_2} + \mathrm{e}^{x_2} - x_1 \\ \text{s.t.} & x \in \mathbf{R}^2 \end{array}$

b) $\begin{array}{ll} \min & \mathrm{e}^{x_1 - x_2} + \mathrm{e}^{x_2} - x_1 \\ \text{s.t.} & \begin{cases} x_1^2 + x_2^2 \leq 1 \\ x_1 + x_2 \geq -1 \end{cases} \end{array}$

c) $\begin{array}{ll} \min & -x_1 - 2x_2 \\ \text{s.t.} & \begin{cases} \mathrm{e}^{x_1} + x_2 \leq 1 \\ x_2 \geq 0 \end{cases} \end{array}$

d) $\begin{array}{ll} \min & x_1 + 2x_2 \\ \text{s.t.} & \begin{cases} x_1^2 + x_2^2 \leq 5 \\ x_1 - x_2 \leq 1 \end{cases} \end{array}$

e) $\begin{array}{ll} \min & x_1 - x_2 \\ \text{s.t.} & \begin{cases} 0 < x_1 \leq 2 \\ 0 \leq x_2 \leq \ln x_1 \end{cases} \end{array}$

f) $\begin{array}{ll} \min & \mathrm{e}^{x_1} + \mathrm{e}^{x_2} + x_1 x_2 \\ \text{s.t.} & \begin{cases} x_1 + x_2 \geq 1 \\ x_1,\, x_2 \geq 0 \end{cases} \end{array}$

11.7 Solve the convex optimization problem

$$\begin{array}{ll} \min & x_1^2 + x_2^2 - \ln(x_1 + x_2) \\ \text{s.t.} & \begin{cases} (x_1 - 1)^2 + x_2^2 \leq 9 \\ x_1 + x_2 \geq 2 \\ x_1,\, x_2 \geq 0. \end{cases} \end{array}$$

11.8 Solve the convex optimization problem

$$\begin{array}{ll} \min & \displaystyle\sum_{j=1}^{n} v_j^{-1} \sqrt{y_j^2 + a_j^2} \\ \text{s.t.} & \begin{cases} \sum_{j=1}^{n} y_j = b \\ y \in \mathbf{R}^n \end{cases} \end{array}$$

that occurred in our discussion of light refraction in Section 9.4, and verify *Snell's law of refraction:* $\sin\theta_i / \sin\theta_j = v_i/v_j$, where $\theta_j = \arctan y_j/a_j$.

11.9 Lisa has inherited 1 million dollars that she intends to invest by buying shares in three companies: A, B and C. Company A manufactures mobile phones, B manufactures antennas for mobile phones, and C manufactures ice cream. The annual return on an investment in the companies is a random variable, and the expected return for each company is estimated to be

	A	B	C
Expected return:	20%	12%	4%

Lisa's expected return if she invests x_1, x_2, x_3 million dollars in the three companies, is thus equal to

$$0.2\,x_1 + 0.12\,x_2 + 0.04\,x_3.$$

The investment risk is by definition the variance of the return. To calculate this we need to know the variance of each company's return and the correlation between the returns of the various companies. For obvious reasons, there is a strong correlation between sales in companies A and B, while sales of the company C only depend on whether the summer weather is beautiful or not, and not on the number of mobile phones sold. The so-called covariance matrix is in our case the matrix

$$\begin{bmatrix} 50 & 40 & 0 \\ 40 & 40 & 0 \\ 0 & 0 & 10 \end{bmatrix}$$

For those who know some basic probability theory, it is now easy to calculate the risk − it is given by the expression

$$50x_1^2 + 80x_1x_2 + 40x_2^2 + 10x_3^2.$$

Lisa, who is a careful person, wants to minimize her investment risk but she also wants to have an expected return of at least 12 %. Formulate and solve Lisa's optimization problem.

11.10 Consider the consumer problem

$$\begin{aligned} \max\ & f(x) \\ \text{s.t.}\ & \begin{cases} \langle p, x\rangle \le I \\ x \ge 0 \end{cases} \end{aligned}$$

discussed in Section 9.4, where $f(x)$ is the consumer's utility function, assumed to be concave and differentiable, I is her disposable income, $p = (p_1, p_2, \ldots, p_n)$ is the price vector and $x = (x_1, x_2, \ldots, x_n)$ denotes a consumption bundle.

Suppose that $\hat{x}$ is an optimal solution. The optimal utility v, as well as $\hat{x}$, depends on the income I, of course; let us assume that $v = v(I)$ is a differentiable function. Show that under these assumptions

$$\begin{aligned} \hat{x}_j,\, \hat{x}_k > 0 \;\; &\Rightarrow \;\; \frac{1}{p_j}\frac{\partial f}{\partial x_j}\Big|_{\hat{x}} = \frac{1}{p_k}\frac{\partial f}{\partial x_k}\Big|_{\hat{x}} = \frac{dv}{dI} \\ \hat{x}_j = 0,\, \hat{x}_k > 0 \;\; &\Rightarrow \;\; \frac{1}{p_j}\frac{\partial f}{\partial x_j}\Big|_{\hat{x}} \leq \frac{1}{p_k}\frac{\partial f}{\partial x_k}\Big|_{\hat{x}}. \end{aligned}$$

In words, this means:

The ratio between the marginal utility and the price of a commodity is for the optimal solution the same for all goods that are actually purchased, and it equals the marginal increase of utility at an increase of income. For goods that are not purchased, the corresponding ratio is not larger.

The conclusion is rather trivial, for it $x_k > 0$ and $\dfrac{1}{p_j}\dfrac{\partial f}{\partial x_j} > \dfrac{1}{p_k}\dfrac{\partial f}{\partial x_k}$, then the consumer benefits from changing a small quantity ϵ/p_k of commodity no. k to the quantity ϵ/p_j of commodity no. j.

Chapter 12

Linear programming

Linear programming (LP) is the art of optimizing linear functions over polyhedra, described as solution sets to systems of linear inequalities. In this chapter, we describe and study the basic mathematical theory of linear programming, above all the very important duality concept.

12.1 Optimal solutions

The optimal value of a general optimization problem was defined in Chapter 9. In particular, each LP problem

$$\text{(P)} \qquad \begin{array}{ll} \min & \langle c, x \rangle \\ \text{s.t.} & x \in X \end{array}$$

has an optimal value, which in this section will be denoted by $v_{\min}(c)$ to indicate its dependence of the objective function.

LP problems with finite optimal values always have optimal solutions. The existence of an optimal solution is of course obvious if the polyhedron of feasible points is bounded, i.e. compact, since the objective function is continuous. For arbitrary LP problems, we rely on the representation theorem for polyhedra to prove the existence of optimal solutions.

Theorem 12.1.1. *Suppose that the polyhedron X of feasible solutions in the LP problem (P) is nonempty and a subset of $\mathbf{R}^n$. Then we have:*

(i) The value function $v_{\min}\colon \mathbf{R}^n \to \underline{\mathbf{R}}$ is concave with effective domain

$$\operatorname{dom} v_{\min} = (\operatorname{recc} X)^+.$$

The objective function $\langle c, x \rangle$ is, in other words, bounded below on X if and only if c belongs to the dual cone of the recession cone of X.

(ii) The problem has optimal solutions for each $c \in (\operatorname{recc} X)^+$, and the set of optimal solutions is a polyhedron. Moreover, the optimum is attained at some extreme point of X if X is a line-free polyhedron.

Proof. By definition, the optimal value $v_{\min}(c) = \inf\{\langle c, x\rangle \mid x \in X\}$ is the pointwise infimum of a family of concave functions, namely the linear functions $c \mapsto \langle c, x\rangle$, with x running through X. So the value function $v_{\min}$ is concave by Theorem 6.2.4.

Let us now determine $\operatorname{dom} v_{\min}$, i.e. the set of c such that $v_{\min}(c) > -\infty$. By the structure theorem for polyhedra (Theorem 5.3.1), there is a finite nonempty set A such that $X = \operatorname{cvx} A + \operatorname{recc} X$, where $A = \operatorname{ext} X$ if the polyhedron is line-free. The optimal value $v_{\min}(c)$ can therefore be calculated as follows:

$$\begin{aligned} v_{\min}(c) &= \inf\{\langle c, y + z\rangle \mid y \in \operatorname{cvx} A,\ z \in \operatorname{recc} X\} \\ &= \inf\{\langle c, y\rangle \mid y \in \operatorname{cvx} A\} + \inf\{\langle c, z\rangle \mid z \in \operatorname{recc} X\} \\ &= \min\{\langle c, y\rangle \mid y \in A\} + \inf\{\langle c, z\rangle \mid z \in \operatorname{recc} X\}, \end{aligned} \tag{12.1}$$

The equality $\inf\{\langle c, y\rangle \mid y \in \operatorname{cvx} A\} = \min\{\langle c, y\rangle \mid y \in A\}$ holds because of Theorem 6.3.3, since linear functions are concave.

If c belongs to the dual cone $(\operatorname{recc} X)^+$, then $\langle c, z\rangle \geq 0$ for all vectors $z \in \operatorname{recc} X$ with equality for $z = 0$, and it follows from equation (12.1) that

$$v_{\min}(c) = \min\{\langle c, y\rangle \mid y \in A\} > -\infty.$$

This proves the inclusion $(\operatorname{recc} X)^+ \subseteq \operatorname{dom} v_{\min}$, and that the optimal value is attained at a point in A, and then in particular at some extreme point of X if the polyhedron X is line-free.

If $c \notin (\operatorname{recc} X)^+$, then $\langle c, z_0\rangle < 0$ for some vector $z_0 \in \operatorname{recc} X$. Since $tz_0 \in \operatorname{recc} X$ for $t > 0$ and $\lim_{t\to\infty}\langle c, tz_0\rangle = -\infty$, it follows that

$$\inf\{\langle c, z\rangle \mid z \in \operatorname{recc} X\} = -\infty,$$

and equation (12.1) now implies that $v_{\min}(c) = -\infty$. This concludes the proof of the equality $\operatorname{dom} v_{\min} = (\operatorname{recc} X)^+$.

The set of minimum points to an LP problem with finite value $v_{\min}$ is equal to the intersection

$$X \cap \{x \in \mathbf{R}^n \mid \langle c, x\rangle = v_{\min}\}$$

between the polyhedron X and a hyperplane, and it is consequently a polyhedron. □

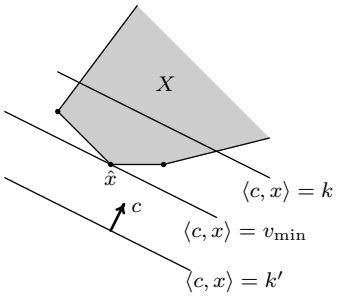

Figure 12.1. The minimum of $\langle c, x\rangle$ over the line-free polyhedron X is attained at an extreme point.

EXAMPLE 12.1.1. The polyhedron X of feasible points for the LP problem

$$\begin{array}{ll} \min & x_1 + x_2 \\ \text{s.t.} & \begin{cases} x_1 - x_2 \geq -2 \\ x_1 + x_2 \geq 1 \\ -x_1 \qquad \geq -3 \end{cases} \end{array}$$

has three extreme points, namely $(3, 5)$, $(-\frac{1}{2}, \frac{3}{2})$ and $(3, -2)$. The values of the objective function $f(x) = x_1 + x_2$ at these points are $f(3, 5) = 8$ and $f(-\frac{1}{2}, \frac{3}{2}) = f(3, -2) = 1$. The least of these is 1, which is the optimal value. The optimal value is attained at two extreme points, $(\frac{1}{2}, \frac{3}{2})$ och $(3, -2)$, and thus also at all points on the line segment between those two points.

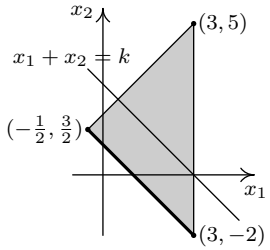

Figure 12.2. Illustration for Example 12.1.1. □

Suppose that $X = \{x \in \mathbf{R}^n \mid Ax \geq b\}$ is a line-free polyhedron and that we want to minimize a given linear function over X. To determine the optimal value of this LP problem, we need according to the previous theorem, assuming that the objective function is bounded below on X, only calculate function values at the finitely many extreme points of X. In theory, this

is easy, but in practice it can be an insurmountable problem, because the number of extreme points may be extremely high. The number of potential extreme points of X when A is an $m \times n$-matrix, equals $\binom{m}{n}$, which for $m = 100$ and $n = 50$ is a number that is greater than 10^{29}. The simplex algorithm, which we will study in Chapter 13, is based on the idea that it is not necessary to search through all the extreme points; the algorithm generates instead a sequence $x_1, x_2, x_3, \ldots$ of extreme points with decreasing objective function values $\langle c, x_1\rangle \geq \langle c, x_2\rangle \geq \langle c, x_3\rangle \geq \ldots$ until the minimum point is found. The number of extreme points that needs to be investigated is therefore generally relatively small.

Sensitivity analysis

Let us rewrite the polyhedron of feasible points in the LP problem

$$\text{(P)} \qquad \begin{array}{ll} \min & \langle c, x\rangle \\ \text{s.t.} & x \in X \end{array}$$

as

$$X = \operatorname{cvx} A + \operatorname{con} B$$

with finite sets A and B. We know from the preceding theorem and its proof that a feasible point $\overline{x}$ is optimal for the LP problem if and only if

$$\begin{cases} \langle c, a\rangle \geq \langle c, \overline{x}\rangle & \text{for all } a \in A \\ \langle c, b\rangle \geq 0 & \text{for all } b \in B, \end{cases}$$

and these inequalities define a convex cone $C_{\overline{x}}$ in the variable c. The set of all c for which a given feasible point is optimal, is thus a convex cone.

Now suppose that $\overline{x}$ is indeed an optimal solution to (P). How much can we change the coefficients of the objective function without changing the optimal solution? The study of this issue is an example of *sensitivity analysis.*

Expressed in terms of the cone $C_{\overline{x}}$, the answer is simple: If we change the coefficients of the objective function to $c + \Delta c$, then $\overline{x}$ is also an optimal solution to the perturbed LP problem

$$\text{(P}') \qquad \begin{array}{ll} \min & \langle c + \Delta c, x\rangle \\ \text{s.t.} & x \in X \end{array}$$

if and only if $c + \Delta c$ belongs to the cone $C_{\overline{x}}$, i.e. if and only if Δc lies in the polyhedron $-c + C_{\overline{x}}$.

In summary, we have thus come to the following conclusions.

Theorem 12.1.2. *(i) The set of all c for which a given feasible point is optimal in the LP problem (P), is a convex cone.*

(ii) If $\overline{x}$ is an optimal solution to problem (P), then there is a polyhedron such that $\overline{x}$ is also an optimal solution to the perturbed LP problem (P′) for all Δc in the polyhedron.

The set $\{\Delta c_k \mid \Delta c \in -c + C_{\overline{x}} \text{ and } \Delta c_j = 0 \text{ for } j \neq k\}$ is a (possibly unbounded) closed interval $[-d_k, e_k]$ around 0. An optimal solution to the problem (P) is therefore also optimal for the perturbed problem that is obtained by only varying the objective coefficient c_k, provided that the perturbation Δc_k lies in the interval $-d_k \leq \Delta c_k \leq e_k$. Many computer programs for LP problems, in addition to generating the optimal value and the optimal solution, also provide information about these intervals.

Sensitivity analysis will be studied in connection with the simplex algorithm in Chapter13.7.

EXAMPLE 12.1.2. The printout of a computer program that was used to solve an LP problem with $c = (20, 30, 40, \dots)$ contained among other things the following information:

Optimal value: 4000 **Optimal solution:** $\overline{x} = (50, 40, 10, \dots)$

Sensitivity report:

Variable	Value	Objective coeff.	Allowable decrease	Allowable increase
x_1	50	20	15	5
x_2	40	30	10	10
x_3	10	40	15	20
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Use the printout to determine the optimal solution and the optimal value if the coefficients c_1, c_2 and c_3 are changed to 17, 35 and 45, respectively, and the other objective coefficients are left unchanged.

Solution: The columns "Allowable decrease" and "Allowable increase" show that the polyhedron of changes Δc that do not affect the optimal solution contains the vectors $(-15, 0, 0, 0, \dots)$, $(0, 10, 0, 0, \dots)$ and $(0, 0, 20, 0, \dots)$. Since

$$(-3, 5, 5, 0, \dots) = \tfrac{1}{5}(-15, 0, 0, 0, \dots) + \tfrac{1}{2}(0, 10, 0, 0, \dots) + \tfrac{1}{4}(0, 0, 20, 0, \dots)$$

and $\frac{1}{5} + \frac{1}{2} + \frac{1}{4} = \frac{19}{20} < 1$, $\Delta c = (-3, 5, 5, 0, \dots)$ is a convex combination of changes that do not affect the optimal solutions, namley the three changes mentioned above and $(0, 0, 0, 0, \dots)$. The solution $\overline{x} = (50, 40, 10, \dots)$ is therefore still optimal for the LP problem with $c = (17, 35, 45, \dots)$. However, the new optimal value is of course $4000 - 20 \cdot 3 + 30 \cdot 5 + 40 \cdot 5 = 4290$. □

12.2 Duality

Dual problems

By describing the polyhedron X in a linear minimization problem

$$\begin{aligned} &\min \ \langle c, x\rangle \\ &\text{s.t.} \ \ x \in X \end{aligned}$$

as the solution set of a system of linear inequalities, we get a problem with a corresponding Lagrange function, and hence also a dual function and a dual problem. The description of X as a solution set is of course not unique, so the dual problem is not uniquely determined by X as a polyhedron, but whichever description we choose, we get, according to Theorem 11.1.1, a dual problem, where strong duality holds, because Slater's condition is satisfied for convex problems with affine constraints.

In this section, we describe the dual problem for some commonly occurring polyhedron descriptions, and we give an alternative proof of the duality theorem. Our premise is that the polyhedron X is given as

$$X = \{x \in U^+ \mid Ax - b \in V^+\},$$

where

- U and V are finitely generated cones in $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively;
- A is an $m \times n$-matrix;
- b is a vector in $\mathbf{R}^m$.

As usual, we identify vectors with column matrices and matrices with linear transformations. The set X is of course a polyhedron, for by writing

$$X = U^+ \cap A^{-1}(b + V^+)$$

we see that X is an intersection of two polyhedra − the conical polyhedron U^+ and the inverse image $A^{-1}(b + V^+)$ under the linear map A of the polyhedron $b + V^+$.

The LP problem of minimizing $\langle c, x\rangle$ over the polyhedron X with the above description will now be written

$$\text{(P)} \qquad \begin{aligned} &\min \ \langle c, x\rangle \\ &\text{s.t.} \ \ Ax - b \in V^+, \ x \in U^+ \end{aligned}$$

and in order to form a suitable dual problem we will perceive the condition $x \in U^+$ as an implicit constraint and express the other condition $Ax - b \in V^+$ as a system of linear inequalities. Assume therefore that the finitely generated cone V is generated by the columns of the $m \times k$-matrix D, i.e. that

$$V = \{Dz \mid z \in \mathbf{R}^k_+\}.$$

The dual cone V^+ can then be written as

$$V^+ = \{y \in \mathbf{R}^m \mid D^T y \geq 0\},$$

and the constraint $Ax - b \in V^+$ can now be expressed as a system of inequalities, namely $D^T Ax - D^T b \geq 0$.

Our LP problem (P) has thus been transformed into

$$\begin{aligned} &\min \ \langle c, x\rangle \\ &\text{s.t.} \ \ D^T b - D^T Ax \leq 0, \ x \in U^+. \end{aligned}$$

The associated Lagrange function $L\colon U^+ \times \mathbf{R}^k_+ \to \mathbf{R}$ is defined by

$$L(x, \lambda) = \langle c, x\rangle + \langle \lambda, D^T b - D^T Ax\rangle = \langle c - A^T D\lambda, x\rangle + \langle b, D\lambda\rangle,$$

and the corresponding dual function $\phi\colon \mathbf{R}^k_+ \to \underline{\mathbf{R}}$ is given by

$$\phi(\lambda) = \inf_{x \in U^+} L(x, \lambda) = \begin{cases} \langle b, D\lambda\rangle\,, & \text{if } c - A^T D\lambda \in U \\ -\infty\,, & \text{otherwise.} \end{cases}$$

This gives us a dual problem of the form

$$\begin{aligned} &\max \ \langle b, D\lambda\rangle \\ &\text{s.t.} \ \ c - A^T D\lambda \in U, \ \lambda \in \mathbf{R}^k_+. \end{aligned}$$

Since $D\lambda$ describes the cone V as λ runs through $\mathbf{R}^k_+$, we can by setting $y = D\lambda$ reformulate the dual problem so that it becomes

$$\begin{aligned} &\max \ \langle b, y\rangle \\ &\text{s.t.} \ \ c - A^T y \in U, \ y \in V. \end{aligned}$$

It is therefore natural to define duality for LP problems of the form (P) as follows.

Definition. Given the LP problem

$$\text{(P)} \qquad \begin{aligned} &\min \ \langle c, x\rangle \\ &\text{s.t.} \ \ Ax - b \in V^+, \ x \in U^+, \end{aligned}$$

which we call the *primal* problem, we call the problem

$$\text{(D)} \qquad \begin{aligned} &\max \ \langle b, y\rangle \\ &\text{s.t.} \ \ c - A^T y \in U, \ y \in V \end{aligned}$$

the *dual* LP problem.

The optimal values of the two problems are denoted by $v_{\min}(P)$ and $v_{\max}(D)$. The polyhedron of feasible points will be denoted by X for the primal problem and by Y for the dual problem.

EXAMPLE 12.2.1. Different choices of the cones U and V give us different concrete dual problems (P) and (D). We exemplify with four important special cases.

1. The choice $U = \{0\}$, $U^+ = \mathbf{R}^n$ and $V = V^+ = \mathbf{R}^m_+$ gives us the following dual pair:

$$\begin{array}{lll} (\mathrm{P}_1) & \min \ \langle c, x\rangle & \text{and} \\ & \text{s.t.} \ \ Ax \geq b & \end{array} \qquad \begin{array}{ll} (\mathrm{D}_1) & \max \ \langle b, y\rangle \\ & \text{s.t.} \ \ A^T y = c, \ y \geq 0. \end{array}$$

Every LP problem can be expressed in the form (P_1), because every polyhedron can be expressed as an intersection of halfspaces, i.e. be written as $Ax \geq b$.

2. The choice $U = U^+ = \mathbf{R}^n_+$ and $V = V^+ = \mathbf{R}^m_+$ gives instead the dual pair:

$$\begin{array}{lll} (\mathrm{P}_2) & \min \ \langle c, x\rangle & \text{and} \\ & \text{s.t.} \ \ Ax \geq b, \ x \geq 0 & \end{array} \qquad \begin{array}{ll} (\mathrm{D}_2) & \max \ \langle b, y\rangle \\ & \text{s.t.} \ \ A^T y \leq c, \ y \geq 0. \end{array}$$

This is the most symmetric formulation of duality, and the natural formulation for many application problems with variables that represent physical quantities or prices, which of course are nonnegative. The diet problem and the production planning problem in Chapter 9.4 are examples of such problems.

3. $U = U^+ = \mathbf{R}^n_+$, $V = \mathbf{R}^m$ and $V^+ = \{0\}$ result in the dual pair:

$$\begin{array}{lll} (\mathrm{P}_3) & \min \ \langle c, x\rangle & \text{and} \\ & \text{s.t.} \ \ Ax = b, \ x \geq 0 & \end{array} \qquad \begin{array}{ll} (\mathrm{D}_3) & \max \ \langle b, y\rangle \\ & \text{s.t.} \ \ A^T y \leq c. \end{array}$$

The formulation (P_3) is the natural starting point for the simplex algorithm.

4. The choice $U = \{0\}$, $U^+ = \mathbf{R}^n$, $V = \mathbf{R}^m$ and $V^+ = \{0\}$ gives us the pair

$$\begin{array}{lll} (\mathrm{P}_4) & \min \ \langle c, x\rangle & \text{and} \\ & \text{s.t.} \ \ Ax = b & \end{array} \qquad \begin{array}{ll} (\mathrm{D}_4) & \max \ \langle b, y\rangle \\ & \text{s.t.} \ \ A^T y = c. \end{array}$$

□

EXAMPLE 12.2.2. A trivial example of dual LP problems in one variable is

$$\begin{array}{lll} \min \ 5x & \text{and} & \max \ 4y \\ \text{s.t.} \ \ 2x \geq 4 & & \text{s.t.} \ \ 2y = 5, \ y \geq 0 \end{array}$$

Both problems have the optimal value 10. □

EXAMPLE 12.2.3. The problems

$$\begin{array}{lll} \min\ x_1 + x_2 & \text{and} & \max\ -2y_1 + y_2 - 3y_3 \\ \text{s.t.} \begin{cases} x_1 - x_2 \geq -2 \\ x_1 + x_2 \geq 1 \\ -x_1 \quad\ \geq -3 \end{cases} & & \text{s.t.} \begin{cases} y_1 + y_2 - y_3 = 1 \\ -y_1 + y_2 \quad\ = 1 \\ y_1, y_2, y_3 \geq 0 \end{cases} \end{array}$$

are dual. The optimal solutions to the primal minimization problem were determined in Example 12.1.1 and the optimal value was found to be 1. The feasible points for the dual maximization problem are of the form $y = (t, 1+t, 2t)$ with $t \geq 0$, and the corresponding values of the objective function are $1 - 7t$. The maximum value is attained for $t = 0$ at the point $(0, 1, 0)$, and the maximum value is equal to 1. □

The Duality Theorem

The primal and dual problems in Examples 12.2.2 and 12.2.3 have the same optimal value, and this is no coincidence but a consequence of the duality theorem, which is formulated below and is a special case of the duality theorem for general convex problems (Theorem 11.1.1). In this section we give an alternative proof of this important theorem, and we start with the trivial result about weak duality.

Theorem 12.2.1 (Weak duality)**.** *The optimal values of the two dual LP problems (P) and (D) satisfy the inequality*

$$v_{\max}(D) \leq v_{\min}(P).$$

Proof. The inequality is trivially satisfied if any of the two polyhedra X and Y of feasible points is empty, because if $Y = \emptyset$ then $v_{\max}(D) = -\infty$, by definition, and if $X = \emptyset$ then $v_{\min}(P) = +\infty$, by definition.

Assume therefore that both problems have feasible points. If $x \in X$ and $y \in Y$, then $y \in V$, $(Ax - b) \in V^+$, $(c - A^T y) \in U$ and $x \in U^+$, by definition, and hence $\langle Ax - b, y\rangle \geq 0$ and $\langle c - A^T y, x\rangle \geq 0$. It follows that

$$\begin{aligned} \langle b, y\rangle &\leq \langle b, y\rangle + \langle c - A^T y, x\rangle = \langle b, y\rangle + \langle c, x\rangle - \langle y, Ax\rangle \\ &= \langle c, x\rangle + \langle b, y\rangle - \langle Ax, y\rangle = \langle c, x\rangle - \langle Ax - b, y\rangle \leq \langle c, x\rangle. \end{aligned}$$

The objective function $\langle b, y\rangle$ in the maximization problem (D) is in other words bounded above on Y by $\langle c, x\rangle$ for each $x \in X$, and hence

$$v_{\max}(D) = \sup_{y \in Y} \langle b, y\rangle \leq \langle c, x\rangle.$$

The objective function $\langle c, x\rangle$ in the minimization problem (P) is therefore bounded below on X by $v_{\max}(D)$. This implies that $v_{\max}(D) \leq v_{\min}(P)$ and completes the proof of the theorem. □

The following optimality criterion follows from weak duality.

Theorem 12.2.2 (Optimality criterion)**.** *Suppose that $\hat{x}$ is a feasible point for the minimization problem (P), that $\hat{y}$ is a feasible point for the dual maximization problem (D), and that*

$$\langle c, \hat{x}\rangle = \langle b, \hat{y}\rangle.$$

Then $\hat{x}$ and $\hat{y}$ are optimal solutions of the respective problems.

Proof. The assumptions on $\hat{x}$ and $\hat{y}$ combined with Theorem 12.2.1 give us the following chain of inequalities

$$v_{\max}(D) \geq \langle b, \hat{y}\rangle = \langle c, \hat{x}\rangle \geq v_{\min}(P) \geq v_{\max}(D).$$

Since the two extreme ends are equal, there is equality everywhere, which means that $\hat{y}$ is a maximum point and $\hat{x}$ is a minimum point. □

Theorem 12.2.3 (Duality theorem)**.** *Suppose that at least one of the two dual LP problems*

$$\text{(P)} \qquad \begin{aligned} &\min \ \langle c, x\rangle \\ &\text{s.t.} \quad Ax - b \in V^+,\ x \in U^+ \end{aligned}$$

and

$$\text{(D)} \qquad \begin{aligned} &\max \ \langle b, y\rangle \\ &\text{s.t.} \quad c - A^T y \in U,\ y \in V \end{aligned}$$

has feasible points. Then, the two problem have the same optimal value.

Thus, provided that at least one of the two dual problems has feasible points:

(i) $X = \emptyset \ \Leftrightarrow$ the objective function $\langle b, y\rangle$ is not bounded above on Y.
(ii) $Y = \emptyset \ \Leftrightarrow$ the objective function $\langle c, x\rangle$ is not bounded below on X.
(iii) If $X \neq \emptyset$ and $Y \neq \emptyset$, then there exist points $\hat{x} \in X$ and $\hat{y} \in Y$ such that $\langle b, y\rangle \leq \langle b, \hat{y}\rangle = \langle c, \hat{x}\rangle \leq \langle c, x\rangle$ for all $x \in X$ and all $y \in Y$.

The duality theorem for linear programming problems is a special case of the general duality theorem for convex problems, but we give here an alternative proof based directly on the following variant of Farkas's lemma.

Lemma. *The system*

$$\text{(12.2)} \qquad \begin{cases} \langle c, x\rangle \leq \alpha \\ x \in X \end{cases}$$

has a solution if and only if the systems

$$\text{(12.3-A)} \qquad \begin{cases} \langle b, y\rangle > \alpha \\ y \in Y \end{cases} \qquad \textit{and} \qquad \text{(12.3-B)} \qquad \begin{cases} \langle b, y\rangle = 1 \\ -A^T y \in U \\ y \in V \end{cases}$$

both have no solutions,

Proof. The system (12.2), i.e.

$$\begin{cases} \langle c, x\rangle \leq \alpha \\ Ax - b \in V^+ \\ x \in U^+, \end{cases}$$

is solvable if and only if the following homogenized system is solvable:

$$\text{(12.2}') \qquad \begin{cases} \langle c, x\rangle \leq \alpha t \\ Ax - bt \in V^+ \\ x \in U^+ \\ t \in \mathbf{R} \\ t > 0. \end{cases}$$

(If x solves the system (12.2), then $(x, 1)$ solves the system (12.2′), and if (x, t) solves the system (12.2′), then x/t solves the system (12.2).) We can write the system (12.2′) more compactly by introducing the matrix

$$\tilde{A} = \begin{bmatrix} \alpha & -c^T \\ -b & A \end{bmatrix}$$

and the vectors $\tilde{x} = (t, x) \in \mathbf{R} \times \mathbf{R}^n$ and $d = (-1, 0) \in \mathbf{R} \times \mathbf{R}^n$, namely as

$$\text{(12.2}'') \qquad \begin{cases} \tilde{A}\tilde{x} \in \mathbf{R}_+ \times V^+ \\ \tilde{x} \in \mathbf{R} \times U^+ \\ d^T\tilde{x} < 0. \end{cases}$$

By Theorem 3.3.2, the system (12.2″) is solvable if and only if the following dual system has no solutions:

$$\text{(12.3}'') \qquad \begin{cases} d - \tilde{A}^T\tilde{y} \in \{0\} \times U \\ \tilde{y} \in \mathbf{R}_+ \times V. \end{cases}$$

Since

$$\tilde{A}^T = \begin{bmatrix} \alpha & -b^T \\ -c & A^T \end{bmatrix},$$

we obtain the following equivalent system from (12.3″) by setting $\tilde{y} = (s, y)$ with $s \in \mathbf{R}$ and $y \in \mathbf{R}^m$:

$$\text{(12.3}'\text{)} \qquad \begin{cases} -1 - \alpha s + \langle b, y\rangle = 0 \\ cs - A^T y \in U \\ y \in V \\ s \geq 0. \end{cases}$$

The system (12.2) is thus solvable if and only if the system (12.3′) has no solutions, and by considering the cases $s > 0$ and $s = 0$ separately, we see that the system (12.3′) has no solutions if and only if the two systems

$$\begin{cases} \langle b, y/s\rangle = \alpha + 1/\mathrm{s} \\ c - A^T(y/s) \in U \\ y/s \in V \\ s > 0 \end{cases} \quad \text{and} \quad \begin{cases} \langle b, y\rangle = 1 \\ -A^T y \in U \\ y \in V \end{cases}$$

have no solutions, and this is obviously the case if and only if the systems (12.3-A) and (12.3-B) both lack solutions. □

Proof of the duality theorem. We now return to the proof of the duality theorem, and because of weak duality, we only need to show the inequality

$$\text{(12.4)} \qquad v_{\min}(P) \leq v_{\max}(D).$$

We divide the proof of this inequality in three separate cases.

Case 1. $Y \neq \emptyset$ and the system (12.3-B) has no solution.

The inequality (12.4) is trivially true if $v_{\max}(D) = \infty$. Therefore, assume that $v_{\max}(D) < \infty$. Then, because of the definition of $v_{\max}(D)$, the system (12.3-A) has no solution when $\alpha = v_{\max}(D)$. So neither of the two systems in (12.3) has a solution for $\alpha = v_{\max}(D)$. Thus, the system (12.2) has a solution for this α-value by the lemma, which means that there is a feasible point $\hat{x}$ such that $\langle c, \hat{x}\rangle \leq v_{\max}(D)$. Consequently, $v_{\min}(P) \leq \langle c, \hat{x}\rangle \leq v_{\max}(D)$.

Note that it follows from the proof that the minimization problem actually has an optimal solution $\hat{x}$.

Case 2. $Y = \emptyset$ and the system (12.3-B) has no solution.

The system (12.3-A) now lacks solutions for all values of α, so it follows from the lemma that the system (12.2) is solvable for all α-values, and this means that the objective function $\langle c, x\rangle$ is unbounded below on X. Hence, $v_{\min}(P) = -\infty = v_{\max}(D)$ in this case.

Case 3. The system (12.3-B) has a solution

It now follows from the lemma that the system (12.2) has no solution for all values of α, and this implies that the set X of feasible solutions is empty. The polyhedron Y of feasible points in the dual problem is consequently nonempty. Choose a point $y_0 \in Y$, let $\overline{y}$ be a solution to the system (12.3-B) and consider the points $y^t = y_0 + t\overline{y}$ for $t > 0$. The vectors y^t belong to V, because they are conical combinations of vectors in V. Moreover, the vectors $c - A^T y^t = (c - A^T y_0) - tA^T\overline{y}$ are conic combinations of vectors in U and thus belong to U. This means that the vector y^t lies in Y for $t > 0$, and since

$$\langle b, y^t \rangle = \langle b, y_0 \rangle + t\langle b, \overline{y} \rangle = \langle b, y_0 \rangle + t \to +\infty$$

as $t \to \infty$, we conclude that $v_{\max}(D) = \infty$. The inequality (12.4) is in other words trivially fulfilled. □

The Complementary Theorem

Theorem 12.2.4 (Complementary theorem)**.** *Suppose that $\hat{x}$ is a feasible point for the LP problem (P) and that $\hat{y}$ is a feasible point for the dual LP problem (D). Then, the two points are optimal for their respective problems if and only if*

$$\langle c - A^T\hat{y}, \hat{x} \rangle = \langle A\hat{x} - b, \hat{y} \rangle = 0.$$

Proof. Note first that due to the definition of the polyhedra X and Y of feasible points, we have $\langle Ax - b, y \rangle \geq 0$ for all points $x \in X$ and $y \in V$, while $\langle c - A^T y, x \rangle \geq 0$ for all points $y \in Y$ and $x \in U$.

In particular, $\langle A\hat{x} - b, \hat{y} \rangle \geq 0$ and $\langle c - A^T\hat{y}, \hat{x} \rangle \geq 0$ if $\hat{x}$ is an optimal solution to the primal problem (P) and $\hat{y}$ is an optimal solution to the dual problem (D). Moreover, $\langle c, \hat{x} \rangle = \langle b, \hat{y} \rangle$ because of the Duality theorem, so it follows that

$$\langle c, \hat{x} \rangle - \langle A\hat{x} - b, \hat{y} \rangle \leq \langle c, \hat{x} \rangle = \langle b, \hat{y} \rangle \leq \langle b, \hat{y} \rangle + \langle c - A^T\hat{y}, \hat{x} \rangle = \langle c, \hat{x} \rangle - \langle A\hat{x} - b, \hat{y} \rangle.$$

Since the two extreme ends of this inequality are equal, we have equality everywhere, i.e. $\langle A\hat{x} - b, \hat{y} \rangle = \langle c - A^T\hat{y}, \hat{x} \rangle = 0$.

Conversely, if $\langle c - A^T\hat{y}, \hat{x} \rangle = \langle A\hat{x} - b, \hat{y} \rangle = 0$, then $\langle c, \hat{x} \rangle = \langle A^T\hat{y}, \hat{x} \rangle$ and $\langle b, \hat{y} \rangle = \langle A\hat{x}, \hat{y} \rangle$, and since $\langle A^T\hat{y}, \hat{x} \rangle = \langle A\hat{x}, \hat{y} \rangle$, we conclude that $\langle c, \hat{x} \rangle = \langle b, \hat{y} \rangle$. The optimality of the two points now follows from the Optimality criterion. □

Let us for clarity formulate the Complementarity theorem in the important special case when the primal and dual problems have the form described as Case 2 in Example 12.2.1.

Corollary 12.2.5. *Suppose that $\hat{x}$ and $\hat{y}$ are feasible points for the dual problems*

$$(\mathrm{P}_2) \qquad \begin{aligned} &\min \ \langle c, x\rangle \\ &\text{s.t.} \quad Ax \geq b,\ x \geq 0 \end{aligned}$$

and

$$(\mathrm{D}_2) \qquad \begin{aligned} &\max \ \langle b, y\rangle \\ &\text{s.t.} \quad A^T y \leq c,\ y \geq 0. \end{aligned}$$

respectively. Then, they are optimal solutions if and only if

$$\begin{cases} (A\hat{x})_i > b_i \ \Rightarrow \ \hat{y}_i = 0 \\ \hat{x}_j > 0 \ \Rightarrow \ (A^T\hat{y})_j = c_j \end{cases} \tag{12.5}$$

In words we can express condition (12.5) as follows, which explains the term 'complementary slackness': If $\hat{x}$ satisfies an individual inequality in the system $Ax \geq b$ strictly, then the corresponding dual variable $\hat{y}_i$ has to be equal to zero, and if $\hat{y}$ satisfies an individual inequality in the system $A^T y \leq c$ strictly, then the corresponding primal variable x_j has to be equal to zero.

Proof. Since $\langle A\hat{x} - b, \hat{y}\rangle = \sum_{i=1}^{m}((A\hat{x})_i - b_i)\hat{y}_i$ is a sum of nonnegative terms, we have $\langle A\hat{x} - b, \hat{y}\rangle = 0$ if and only if all the terms are equal to zero, i.e. if and only if $(A\hat{x})_i > b_i \Rightarrow \hat{y}_i = 0$.

Similarly, $\langle c - A^T\hat{y}, \hat{x}\rangle = 0$ if and only if $\hat{x}_j > 0 \Rightarrow (A^T\hat{y})_j = c_j$. The corollary is thus just a reformulation of Theorem 12.2.4 for dual problems of type (P_2)–(D_2). □

The curious reader may wonder whether the implications in the condition (12.5) can be replaced by equivalences. The following trivial example shows that this is not the case.

Example 12.2.4. Consider the dual problems

$$\begin{aligned} &\min \ x_1 + 2x_2 \\ &\text{s.t.} \quad x_1 + 2x_2 \geq 2,\ \ x \geq 0 \end{aligned} \qquad \text{and} \qquad \begin{aligned} &\max \ 2y \\ &\text{s.t.} \ \begin{cases} y \leq 1 \\ 2y \leq 2,\ y \geq 0 \end{cases} \end{aligned}$$

with $A = c^T = \begin{bmatrix} 1 & 2 \end{bmatrix}$ and $b = [2]$. The condition (12.5) is not fulfilled with equivalence at the optimal points $\hat{x} = (2, 0)$ and $\hat{y} = 1$, because $\hat{x}_2 = 0$ and $(A^T\hat{y})_2 = 2 = c_2$.

However, there are other optimal solutions to the minimization problem; all points on the line segment between $(2, 0)$ and $(0, 1)$ are optimal, and the optimal pairs $\hat{x} = (2 - 2t, t)$ and $\hat{y} = 1$ satisfy the condition (12.5) with equivalence for $0 < t < 1$. □

The last conclusion in the above example can be generalized. All dual problems with feasible points have a pair of optimal solutions $\hat{x}$ and $\hat{y}$ that satisfy the condition (12.5) with implications replaced by equivalences. See exercise 12.8.

EXAMPLE 12.2.5. The LP problem

$$\begin{aligned} \min\ & -x_1 + 2x_2 + x_3 + 2x_4 \\ \text{s.t.}\ & \begin{cases} -x_1 - x_2 - 2x_3 + x_4 \geq 4 \\ -2x_1 + x_2 + 3x_3 + x_4 \geq 8 \\ x_1, x_2, x_3, x_4 \geq 0 \end{cases} \end{aligned}$$

is easily solved by first solving the dual problem

$$\begin{aligned} \max\ & 4y_1 + 8y_2 \\ \text{s.t.}\ & \begin{cases} -y_1 - 2y_2 \leq -1 \\ -y_1 + y_2 \leq 2 \\ -2y_1 + 3y_2 \leq 1 \\ y_1 + y_2 \leq 2 \\ y_1, y_2 \geq 0 \end{cases} \end{aligned}$$

graphically and then using the Complementary theorem.

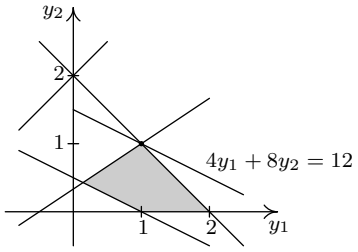

Figure 12.3. A graphical solution to the maximization problem in Ex. 12.2.5.

A graphical solution is obtained from figure 12.3, which shows that $\hat{y} = (1, 1)$ is the optimal point and that the value is 12. Since $\hat{y}$ satisfies the first two constraints with strict inequality and $\hat{y}_1 > 0$ and $\hat{y}_2 > 0$, we obtain the optimal solution $\hat{x}$ to the minimization problem as a solution to the system

$$\begin{cases} -x_1 - x_2 - 2x_3 + x_4 = 4 \\ -2x_1 + x_2 + 3x_3 + x_4 = 8 \\ x_1 = 0 \\ x_2 = 0 \\ x_1, x_2, x_3, x_4 \geq 0. \end{cases}$$

The solution to this system is $\hat{x} = \left(0, 0, \frac{4}{5}, \frac{28}{5}\right)$, and the optimal value is 12, which it of course has to be according to the Duality theorem. □

Exercises

12.1 The matrix A and the vector c are assumed to be fixed in the LP problem

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

but the right hand side vector b is allowed to vary. Suppose that the problem has a finite value for some right hand side b. Prove that for each b, the value is either finite or there are no feasible points. Show also that the optimal value is a convex function of b.

12.2 Give an example of dual problems which both have no feasible points.

12.3 Use duality to show that $(3, 0, 1)$ is an optimal solution to the LP problem

$$\begin{array}{ll} \min & 2x_1 + 4x_2 + 3x_3 \\ \text{s.t.} & \begin{cases} 2x_1 + 3x_2 + 4x_3 \geq 10 \\ x_1 + 2x_2 \geq 3 \\ 2x_1 + 7x_2 + 2x_3 \geq 5, \quad x \geq 0. \end{cases} \end{array}$$

12.4 Show that the column player's problem and the row player's problem in a two-person zero-sum game (see Chapter 9.4) are dual problems.

12.5 Investigate how the optimal solution to the LP problem

$$\begin{array}{ll} \max & x_1 + x_2 \\ \text{s.t.} & \begin{cases} tx_1 + x_2 \geq -1 \\ x_1 \leq 2 \\ x_1 - x_2 \geq -1 \end{cases} \end{array}$$

depends on the parameter t.

12.6 The Duality theorem follows from Farkas's lemma (Corollary 3.3.3). Show conversely that Farkas's lemma follows from the Duality theorem by considering the dual problems

$$\begin{array}{lll} \min & \langle c, x \rangle & \\ \text{s.t.} & Ax \geq 0 & \end{array} \qquad \text{and} \qquad \begin{array}{ll} \max & \langle 0, y \rangle \\ \text{s.t.} & A^T y = c, \ y \geq 0 \end{array}$$

12.7 Let $Y = \{y \in \mathbf{R}^m \mid c - A^T y \in U, \, y \in V\}$, where U and V are closed convex cones, and suppose that $Y \neq \emptyset$.

a) Show that $\operatorname{recc} Y = \{y \in \mathbf{R}^m \mid -A^T y \in U, \, y \in V\}$.

b) Show that the system (12.3-B) has a solution if and only if the vector $-b$ does not belong to the dual cone of $\operatorname{recc} Y$.

c) Show, using the result in b), that the conclusion in case 3 of the proof of the Duality theorem follows from Theorem 12.1.1, i.e. that $v_{\max}(D) = \infty$ if (and only if) the system (12.3-B) has a solution.

12.8 Suppose that the dual problems

$$\begin{array}{lll} \min & \langle c, x\rangle & \\ \text{s.t.} & Ax \geq b,\ x \geq 0 & \end{array} \qquad \text{and} \qquad \begin{array}{lll} \max & \langle b, y\rangle & \\ \text{s.t.} & A^T y \leq c,\ y \geq 0 & \end{array}$$

both have feasible points. Prove that there exist optimal solutions $\hat{x}$ and $\hat{y}$ to the problems that satisfy

$$\begin{cases} (A\hat{x})_i > b_i & \Leftrightarrow \quad \hat{y}_i = 0 \\ \hat{x}_j > 0 & \Leftrightarrow \quad (A^T\hat{y})_j = c_j. \end{cases}$$

[Hint: Because of the Complementarity theorem it suffices to show that the following system of inequalities has a solution: $Ax \geq b$, $x \geq 0$, $A^T y \leq c$, $y \geq 0$, $\langle b, y\rangle \geq \langle c, x\rangle$, $Ax + y > b$, $Ay - c < x$. And this system is solvable if and only if the following homogeneous system is solvable: $Ax - bt \geq 0$, $x \geq 0$, $-A^T y + ct \geq 0$, $y \geq 0$, $-\langle c, x\rangle + \langle b, y\rangle \leq 0$, $Ax + y - bt > 0$, $x - A^T y + ct > 0$, $t > 0$. The solvability can now be decided by using Theorem 3.3.7.]

Part III

The simplex algorithm

Chapter 13

The simplex algorithm

For practical purposes, there are somewhat simplified two kinds of methods for solving LP problems. Both generate a sequence of points with progressively better objective function values. Simplex methods, which were introduced by Dantzig in the late 1940s, generate a sequence of extreme points of the polyhedron of feasible points in the primal (or dual) problem by moving along the edges of the polyhedron. Interior-point methods generate instead, as the name implies, points in the interior of the polyhedron. These methods are derived from techniques for non-linear programming, developed by Fiacco and McCormick in the 1960s, but it was only after Karmarkars innovative analysis in 1984 that the methods began to be used for LP problems.

In this chapter, we describe and analyze the simplex algorithm.

13.1 Standard form

The simplex algorithm requires that the LP problem is formulated in a special way, and the variant of the algorithm that we will study assumes that the problem is a minimization problem, that all variables are nonnegative and that all other constraints are formulated as equalities.

Definition. An LP problem has *standard form* if it has the form

$$\begin{aligned} \min\ & c_1x_1 + c_2x_2 + \cdots + c_nx_n \\ \text{s.t.}\ & \begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \\ x_1, x_2, \ldots, x_n \geq 0. \end{cases} \end{aligned}$$

By introducing the matrices

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \quad \text{and} \quad c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}$$

we get the following compact writing for an LP problem in standard form:

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

We noted in Chapter 9 that each LP problem can be transformed into an equivalent LP problem in standard form by using slack/surplus variables and by replacing unrestricted variables with differences of nonnegative variables.

Duality

We gave a general definition of the concept of duality in Chapter 12.2 and showed that dual LP problems have the same optimal value, except when both problems have no feasible points. In our description of the simplex algorithm, we will need a special case of duality, and to make the presentation independent of the results in the previous chapter, we now repeat the definition for this special case.

Definition. The LP problem

$$\text{(D)} \qquad \begin{aligned} &\max \ \langle b, y \rangle \\ &\text{s.t.} \ \ A^T y \leq c \end{aligned}$$

is said to be *dual* to the LP problem

$$\text{(P)} \qquad \begin{aligned} &\min \ \langle c, x \rangle \\ &\text{s.t.} \quad Ax = b, \ x \geq 0. \end{aligned}$$

We shall use the following trivial part of the Duality theorem.

Theorem 13.1.1 (Weak duality). *If x is a feasible point for the minimization problem (P) and y is a feasible point for the dual maximization problem (D), i.e. if $Ax = b$, $x \geq 0$ and $A^T y \leq c$, then*

$$\langle b, y \rangle \leq \langle c, x \rangle.$$

Proof. The inequalities $A^T y \leq c$ and $x \geq 0$ imply that att $\langle x, A^T y \rangle \leq \langle x, c \rangle$, and hence

$$\langle b, y \rangle = \langle Ax, y \rangle = \langle x, A^T y \rangle \leq \langle x, c \rangle = \langle c, x \rangle. \qquad \square$$

Corollary 13.1.2 (Optimality criterion). *Suppose that $\hat{x}$ is a feasible point for the minimization problem (P), that $\hat{y}$ is a feasible point for the dual maximization problem (D), and that $\langle c, \hat{x}\rangle = \langle b, \hat{y}\rangle$. Then $\hat{x}$ and $\hat{y}$ are optimal solutions to the respective problems.*

Proof. It follows from the assumptions and Theorem 13.1.1, applied to the point $\overline{y}$ and an arbitrary feasible point x for the minimization problem, that

$$\langle c, \overline{x}\rangle = \langle b, \overline{y}\rangle \leq \langle c, x\rangle$$

for all feasible points x. This shows that $\overline{x}$ is a minimum point, and an analogous argument shows that $\overline{y}$ is a maximum point. □

13.2 Informal description of the simplex algorithm

In this section we describe the main features of the simplex algorithm with the help of some simple examples. The precise formulation of the algorithm and the proof that it works is given in sections 13.4 and 13.5.

Example 13.2.1. We start with a completely trivial problem, namely

$$\begin{array}{ll} \min & f(x) = x_3 + 2x_4 \\ \text{s.t.} & \begin{cases} x_1 \quad + 2x_3 - x_4 = 2 \\ \quad x_2 - \ x_3 + x_4 = 3, \ x \geq 0. \end{cases} \end{array}$$

Since the coefficients of the objective function $f(x)$ are positive and $x \geq 0$, it is clear that $f(x) \geq 0$ for all feasible points x. There is also a feasible point x with $x_3 = x_4 = 0$, namely $x = (2, 3, 0, 0)$. The minimum is therefore equal to 0, and $(2, 3, 0, 0)$ is the (unique) minimum point. □

Now consider an arbitrary problem of the form

$$\begin{array}{ll} \min & f(x) = c_{m+1}x_{m+1} + \cdots + c_n x_n + d \\ \text{s.t.} & \begin{cases} x_1 + \ a_{1\,m+1}x_{m+1} + \ldots + \ a_{1n}x_n = b_1 \\ x_2 + \ a_{2\,m+1}x_{m+1} + \ldots + \ a_{2n}x_n = b_2 \\ \qquad\qquad\qquad\qquad\qquad\qquad\quad \vdots \\ x_m + a_{m\,m+1}x_{m+1} + \ldots + a_{mn}x_n = b_m, \ \ x \geq 0 \end{cases} \end{array} \tag{13.1}$$

where

$$b_1, b_2, \ldots, b_m \geq 0.$$

If $c_{m+1}, c_{m+2}, \dots, c_n \geq 0$, then obviously $f(x) \geq d$ for all feasible points x, and since $\overline{x} = (b_1, \dots, b_m, 0, \dots, 0)$ is a feasible point and $f(\overline{x}) = d$, it follows that d is the optimal value.

The constraint system in LP problem (13.1) has a very special form, for it is solved with respect to the basic variables $x_1, x_2, \dots, x_m$, and these variables are not present in the objective function. Quite generally, we shall call a set of variables *basic* to a given system of linear equations if it is possible to solve the system with respect to the variables in the set.

EXAMPLE 13.2.2. Let us alter the objective function in Example 13.2.1 by changing the sign of the x_3-coefficient. Our new problem thus reads as follows:

$$\begin{aligned} &\min \ f(x) = -x_3 + 2x_4 \\ &\text{s.t.} \ \begin{cases} x_1 \quad + 2x_3 - x_4 = 2 \\ \quad x_2 - \ x_3 + x_4 = 3, \ x \geq 0. \end{cases} \end{aligned} \tag{13.2}$$

The point $(2, 3, 0, 0)$ is of course still feasible and the corresponding value of the objective function is 0, but we can get a smaller value by choosing $x_3 > 0$ and keeping $x_4 = 0$. However, we must ensure that $x_1 \geq 0$ and $x_2 \geq 0$, so the first constraint equation yields the bound $x_1 = 2 - 2x_3 \geq 0$, i.e. $x_3 \leq 1$.

We now transform the problem by solving the system (13.2) with respect to the variables x_2 and x_3, i.e. by changing basic variables from x_1, x_2 to x_2, x_3. Using Gaussian elimination, we obtain

$$\begin{cases} \frac{1}{2}x_1 \quad + x_3 - \frac{1}{2}x_4 = 1 \\ \frac{1}{2}x_1 + x_2 \quad + \frac{1}{2}x_4 = 4. \end{cases}$$

The new basic variable x_3 is then eliminated from the objectiv function by using the first equation in the new system. This results in

$$f(x) = \tfrac{1}{2}x_1 + \tfrac{3}{2}x_4 - 1,$$

and our problem has thus been reduced to a problem of the form (13.1), namely

$$\begin{aligned} &\min \ \tfrac{1}{2}x_1 + \tfrac{3}{2}x_4 - 1 \\ &\text{s.t.} \ \begin{cases} \frac{1}{2}x_1 \quad + x_3 - \frac{1}{2}x_4 = 1 \\ \frac{1}{2}x_1 + x_2 \quad + \frac{1}{2}x_4 = 4, \ x \geq 0 \end{cases} \end{aligned}$$

with x_2 and x_3 as basic variables and with nonnegative coefficients for the other variables in the objectiv function. Hence, the optimal value is equal to -1, and $(0, 4, 1, 0)$ is the optimal point. □

The strategy for solving a problem of the form (13.1), where some coefficient c_{m+k} is negative, consists in replacing one of the basic variables $x_1, x_2, \ldots, x_m$ with x_{m+k} so as to obtain a new problem of the same form. If the new c-coefficients are nonnegative, then we are done. If not, we have to repeat the procedure. We illustrate with another example.

EXAMPLE 13.2.3. Consider the problem

$$\begin{aligned} &\min \ f(x) = 2x_1 - x_2 + x_3 - 3x_4 + x_5 \\ &\text{s.t.} \begin{cases} x_1 \qquad\quad + 2x_4 - \ x_5 = 5 \\ \quad x_2 \quad + \ x_4 + 3x_5 = 4 \\ \qquad x_3 - \ x_4 + \ x_5 = 3, \ x \geq 0. \end{cases} \end{aligned} \tag{13.3}$$

First we have to eliminate the basic variables x_1, x_2, x_3 from the objective function with

$$f(x) = -5x_4 + 5x_5 + 9 \tag{13.4}$$

as result. Since the coefficient of x_4 is negative, x_4 has to be eliminated from the objective function and from two constraint equations in such a way that the right hand side of the transformed system remains nonnegative. The third equation in (13.3) can not be used for this elimination, since the coefficient of x_4 is negative. Eliminating x_4 from the first equation by using the second equation results in the equation $x_1 - 2x_2 - 7x_5 = 5 - 2 \cdot 4 = -3$, which has an illegal right-hand side. It therefore only remains to use the first of the constraints in (13.3) for the elimination. We then get the following equivalent system

$$\begin{cases} \frac{1}{2}x_1 \qquad\quad + x_4 - \frac{1}{2}x_5 = \frac{5}{2} \\ -\frac{1}{2}x_1 + x_2 \qquad\quad + \frac{7}{2}x_5 = \frac{3}{2} \\ \frac{1}{2}x_1 \quad + x_3 \quad + \frac{1}{2}x_5 = \frac{11}{2}, \ x \geq 0 \end{cases} \tag{13.5}$$

with x_2, x_3, x_4 as new basic variables.

The reason why the right-hand side of the system remains positive when the first equation of (13.3) is used for the elimination of x_4, is that the ratio of the right-hand side and the x_4-coefficient is smaller for the first equation than for the second ($5/2 < 4/1$).

We now eliminate x_4 from the objective function, using equation (13.4) and the first equation of the system (13.5), and obtain

$$f(x) = \tfrac{5}{2}x_1 + \tfrac{5}{2}x_5 - \tfrac{7}{2}$$

which is to be minimized under the constraints (13.5). The minimum value is clearly equal to $-\frac{7}{2}$, and $(0, \frac{3}{2}, \frac{11}{2}, \frac{5}{2}, 0)$ is the minimum point.

To reduce the writing it is customary to omit the variables and only work with coefficients in tabular form. The problem (13.3) is thus represented by the following *simplex tableau:*

$$\begin{array}{rrrrr|r} 1 & 0 & 0 & 2 & -1 & 5 \\ 0 & 1 & 0 & 1 & 3 & 4 \\ 0 & 0 & 1 & -1 & 1 & 3 \\ \hline 2 & -1 & 1 & -3 & 1 & f \end{array}$$

The upper part of the tableau represents the system of equations, and the lower row represents the objective function f. The vertical line corresponds to the equality signs in (13.3).

To eliminate the basic variables x_1, x_2, x_3 from the objective function we just have to add -2 times row 1, row 2 and -1 times row 3 to the objective function row in the above tableau. This gives us the new tableau

$$\begin{array}{rrrrr|r} 1 & 0 & 0 & \underline{2} & -1 & 5 \\ 0 & 1 & 0 & 1 & 3 & 4 \\ 0 & 0 & 1 & -1 & 1 & 3 \\ \hline \underline{0} & \underline{0} & \underline{0} & -5 & 5 & f-9 \end{array}$$

The bottom row corresponds to equation (13.4). Note that the constant term 9 appears on the other side of the equality sign compared to (13.4), and this explains the minus sign in the tableau. We have also highlighted the basic variables by underscoring.

Since the x_4-coefficient of the objective function is negative, we have to transform the tableau in such a way that x_4 becomes a new basic variable. By comparing the ratios $5/2$ and $4/1$ we conclude that the first row has to be the *pivot row*, i.e. has to be used for the eliminations. We have indicated this by underscoring the coefficient in the first row and the fourth column of the tableau, the so-called *pivot element.*

Gaussian elimination gives rise to the new simplex tableau

$$\begin{array}{rrrrr|r} \frac{1}{2} & 0 & 0 & 1 & -\frac{1}{2} & \frac{5}{2} \\ -\frac{1}{2} & 1 & 0 & 0 & \frac{7}{2} & \frac{3}{2} \\ \frac{1}{2} & 0 & 1 & 0 & \frac{1}{2} & \frac{11}{2} \\ \hline \frac{5}{2} & \underline{0} & \underline{0} & \underline{0} & \frac{5}{2} & f+\frac{7}{2} \end{array}$$

Since the coefficients of the objective function are now nonnegative, we can read the minimum, with *reversed sign*, in the lower right corner of the tableau. The minimum point is obtained by assigning the value 0 to the non-basic variables x_1 and x_5, which gives $x = (0, \frac{3}{2}, \frac{11}{2}, \frac{5}{2}, 0)$. □

EXAMPLE 13.2.4. Let us solve the LP problem

$$\begin{array}{ll}\min & x_1 - 2x_2 + x_3 \\ \text{s.t.} & \begin{cases} x_1 + 2x_2 + 2x_3 + x_4 = 5 \\ x_1 + x_3 + x_5 = 2 \\ x_2 - 2x_3 + x_6 = 1, \ x \geq 0. \end{cases}\end{array}$$

The corresponding simplex tableau is

$$\begin{array}{cccccc|c} 1 & 2 & 2 & 1 & 0 & 0 & 5 \\ 1 & 0 & 1 & 0 & 1 & 0 & 2 \\ 0 & \underline{1} & -2 & 0 & 0 & 1 & 1 \\ \hline 1 & -2 & 1 & \underline{0} & \underline{0} & \underline{0} & f \end{array}$$

with x_4, x_5, x_6 as basic variables, and these are already eliminated from the objective function. Since the x_2-coefficient of the objective function is negative, we have to introduce x_2 as a new basic variable, and we have to use the underscored element as pivot element, since $1/1 < 5/2$. Using the third row, the tableau is transformed into

$$\begin{array}{cccccc|c} 1 & 0 & \underline{6} & 1 & 0 & -2 & 3 \\ 1 & 0 & 1 & 0 & 1 & 0 & 2 \\ 0 & 1 & -2 & 0 & 0 & 1 & 1 \\ \hline 1 & \underline{0} & -3 & \underline{0} & \underline{0} & 2 & f+2 \end{array}$$

and this tableau corresponds to the problem

$$\begin{array}{ll}\min & x_1 - 3x_3 + 2x_6 - 2 \\ \text{s.t.} & \begin{cases} x_1 + 6x_3 + x_4 - 2x_6 = 3 \\ x_1 + x_3 + x_5 = 2 \\ x_2 - 2x_3 + x_6 = 1, \ x \geq 0. \end{cases}\end{array}$$

Since the x_3-coefficient in the objective function is now negative, we have to repeat the procedure. We must thus introduce x_3 as a new basic variable, and this time we have to use the first row as pivot row, for $3/6 < 2/1$. The new tableau has the following form

$$\begin{array}{cccccc|c} \frac{1}{6} & 0 & 1 & \frac{1}{6} & 0 & -\frac{1}{3} & \frac{1}{2} \\ \frac{5}{6} & 0 & 0 & -\frac{1}{6} & 1 & \frac{1}{3} & \frac{3}{2} \\ \frac{1}{3} & 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 2 \\ \hline \frac{3}{2} & \underline{0} & \underline{0} & \frac{1}{2} & \underline{0} & 1 & f+\frac{7}{2} \end{array}$$

We can now read off the minimum $-\frac{7}{2}$ and the minimum point $(0, 2, \frac{1}{2}, 0, \frac{3}{2}, 0)$. □

So far, we have written the function symbol f in the lower right corner of our simplex tableaux. We have done this for pedagogical reasons to explain why the function value in the box gets a reverse sign. Remember that the last row of the previous simplex tableau means that

$$\tfrac{3}{2}x_1 + \tfrac{1}{2}x_4 + x_6 = f(x) + \tfrac{7}{2}.$$

Since the symbol has no other function, we will omit it in the future.

EXAMPLE 13.2.5. The problem

$$\begin{aligned} \min\ & f(x) = -2x_1 + x_2 \\ \text{s.t.}\ & \begin{cases} x_1 - x_2 + x_3 = 3 \\ -x_1 + x_2 + x_4 = 4, \ x \geq 0 \end{cases} \end{aligned}$$

gives rise to the following simplex tableaux:

$$\begin{array}{rrrr|r} \underline{1} & -1 & 1 & 0 & 3 \\ -1 & 1 & 0 & 1 & 4 \\ \hline -2 & 1 & \underline{0} & \underline{0} & 0 \end{array}$$

$$\begin{array}{rrrr|r} 1 & -1 & 1 & 0 & 3 \\ 0 & 0 & 1 & 1 & 7 \\ \hline \underline{0} & -1 & 2 & \underline{0} & 6 \end{array}$$

Since the objective function has a negative x_2-coefficient, we are now supposed to introduce x_2 as a basic variable, but no row will work as a pivot row since the entire x_2-column is non-positive. This implies that the objective function is unbounded below, i.e. there is no minimum. To see this, we rewrite the last tableau with variables in the form

$$\begin{aligned} \min\ & f(x) = -x_2 + 2x_3 - 6 \\ \text{s.t.}\ & \begin{cases} x_1 = x_2 - x_3 + 3 \\ x_4 = \quad - x_3 + 7. \end{cases} \end{aligned}$$

By choosing $x_2 = t$ and $x_3 = 0$ we get a feasible point $x^t = (3 + t, t, 0, 7)$ for each $t \geq 0$, and since $f(x^t) = -t - 6 \to -\infty$ as $t \to \infty$, we conclude that the objective function is unbounded below. □

Examples 13.2.4 and 13.2.5 are typical for LP problems of the form (13.1). In Section 13.5, namely, we show that one can always perform the iterations so as to obtain a final tableau similar to the one in Example 13.2.4 or the one in Example 13.2.5, and in Section 13.6 we will show how to get started, i.e. how to transform an arbitrary LP problem in standard form into a problem of the form (13.1).

13.3 Basic solutions

In order to describe and understand the simplex algorithm it is necessary first to know how to produce solutions to a linear system of equations. We assume that Gaussian elimination is familiar and concentrate on describing how to switch from one basic solution to another. We begin by reviewing the notation that we will use in the rest of this chapter.

The columns of an $m \times n$-matrix A will be denoted A_{*1}, A_{*2}, ..., A_{*n} so that

$$A = \left[A_{*1} \; A_{*2} \; \ldots \; A_{*n} \right].$$

We will often have to consider submatrices comprised of certain columns in an $m \times n$-matrix A. So if $1 \leq k \leq n$ and

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$$

is a permutation of k elements chosen from the set $\{1, 2, \ldots, n\}$, we let $A_{*\alpha}$ denote the $m \times k$-matrix consisting of the columns $A_{*\alpha_1}$, $A_{*\alpha_2}$, ..., $A_{*\alpha_k}$ in the matrix A, i.e.

$$A_{*\alpha} = \left[A_{*\alpha_1} \; A_{*\alpha_2} \; \ldots \; A_{*\alpha_k} \right].$$

And if

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

is a column matrix with n entries, then x_α denotes the column matrix

$$\begin{bmatrix} x_{\alpha_1} \\ x_{\alpha_2} \\ \vdots \\ x_{\alpha_k} \end{bmatrix}.$$

As usual, we make no distinction between column matrices with n entries and vectors in $\mathbf{R}^n$.

We consider permutations $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ as ordered sets and allow us therefore to write $j \in \alpha$ if j is any of the numbers $\alpha_1, \alpha_2, \ldots, \alpha_k$. This also allows us to write sums of the type

$$\sum_{i=1}^{k} x_{\alpha_i} A_{*\alpha_i}$$

as

$$\sum_{j\in\alpha} x_j A_{*j},$$

or with matrices as

$$A_{*\alpha}x_\alpha.$$

Definition. Let A be an $m\times n$-matrix of rank m, and let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ be a permutation of m numbers from the set $\{1, 2, \ldots, n\}$. The permutation α is called a *basic index set* of the matrix A if the columns of the $m\times m$-matrix $A_{*\alpha}$ form a basis for $\mathbf{R}^m$.

The condition that the columns $A_{*\alpha_1}, A_{*\alpha_2}, \ldots, A_{*\alpha_m}$ form a basis is equivalent to the condition that the submatrix

$$A_{*\alpha} = \begin{bmatrix} A_{*\alpha_1} & A_{*\alpha_2} & \ldots & A_{*\alpha_m} \end{bmatrix}$$

is invertible. The inverse of the matrix $A_{*\alpha}$ will be denoted by $A_{*\alpha}^{-1}$. This matrix, which thus means $(A_{*\alpha})^{-1}$, will appear frequently in the sequel − do not confuse it with $(A^{-1})_{*\alpha}$, which is not generally well defined.

If $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ is a basic index set, so too is of course every permutation of α.

Example 13.3.1. The matrix

$$\begin{bmatrix} 3 & 1 & 1 & -3 \\ 3 & -1 & 2 & -6 \end{bmatrix}$$

has the following basic index sets: $(1,2)$, $(2,1)$, $(1,3)$, $(3,1)$, $(1,4)$, $(4,1)$ $(2,3)$, $(3,2)$, $(2,4)$, and $(4,2)$. □

We also need a convenient way to show the result of replacing an element in an ordered set with some other element. Therefore, let $M = (a_1, a_2, \ldots, a_n)$ be an arbitrary n-tuple (ordered set). The n-tuple obtained by replacing the item a_r at location r with an arbitrary object x will be denoted by $M_{\hat{r}}[x]$. In other words,

$$M_{\hat{r}}[x] = (a_1, \ldots, a_{r-1}, x, a_{r+1}, \ldots, a_n).$$

An $m \times n$-matrix can be regarded as an ordered set of columns. If b is a column matrix with m entries and $1 \le r \le n$, we therefore write $A_{\hat{r}}[b]$ for the matrix

$$\begin{bmatrix} A_{*1} & \ldots & A_{*r-1} & b & A_{*r+1} & \ldots & A_{*n} \end{bmatrix}.$$

Another context in which we will use the above notation for replacement of elements, is when $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ is a permutation of m elements

taken from the set $\{1, 2, \dots, n\}$. If $1 \le r \le m$, $1 \le k \le n$ and $k \notin \alpha$, then $\alpha_{\hat{r}}[k]$ denotes the new permutation

$$(\alpha_1, \dots, \alpha_{r-1}, k, \alpha_{r+1}, \dots, \alpha_m).$$

Later we will need the following simple result, where the above notation is used.

Lemma 13.3.1. *Let E be the unit matrix of order m, and let b be a column matrix with m elements. The matrix $E_{\hat{r}}[b]$ is invertible if and only if $b_r \neq 0$, and in this case*

$$E_{\hat{r}}[b]^{-1} = E_{\hat{r}}[c],$$

where

$$c_j = \begin{cases} -b_j/b_r & \text{for } j \neq r, \\ 1/b_r & \text{for } j = r. \end{cases}$$

Proof. The proof is left to the reader as a simple exercise. □

Example 13.3.2.

$$\begin{bmatrix} 1 & 4 & 0 \\ 0 & 3 & 0 \\ 0 & 5 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -4/3 & 0 \\ 0 & 1/3 & 0 \\ 0 & -5/3 & 1 \end{bmatrix}$$ □

Systems of linear equations and basic solutions

Consider a system of linear equations

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{cases} \tag{13.6}$$

with coefficient matrix A of rank m and right-hand side matrix b. Such a system can equivalently be regarded as a vector equation

$$\sum_{j=1}^{n} x_j A_{*j} = b \tag{13.6′}$$

or as a matrix equation

$$Ax = b. \tag{13.6″}$$

Both alternative approaches are, as we shall see, fruitful.

We solve the system (13.6), preferably using Gaussian elimination, by expressing m of the variables, $x_{\alpha_1}, x_{\alpha_2}, \ldots, x_{\alpha_m}$ say, as linear combinations of the remaining $n-m$ variables $x_{\beta_1}, x_{\beta_2}, \ldots, x_{\beta_{n-m}}$ and $b_1, b_2, \ldots, b_m$. Each assignment of values to the latter β-variables results in a unique set of values for the former α-variables. In particular, we get a unique solution by setting all β-variables equal to 0.

This motivates the following definition.

Definition. Let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ be a permutation of m numbers chosen from the set $\{1, 2, \ldots, n\}$, and let $\beta = (\beta_1, \beta_2, \ldots, \beta_{n-m})$ be a permutation of the remaining $n-m$ numbers. The variables $x_{\alpha_1}, x_{\alpha_2}, \ldots, x_{\alpha_m}$ are called *basic variables* and the variables x_{β_1}, $x_{\beta_2}, \ldots, x_{\beta_{n-m}}$ are called *free variables* in the system (13.6), if for each $c = (c_1, c_2, \ldots, c_{n-m}) \in \mathbf{R}^{n-m}$ there is a unique solution x to the system (13.6) such that $x_\beta = c$. The unique solution obtained by setting all free variables equal to 0 is called a *basic solution*.

Any m variables can not be chosen as basic variables; to examine which ones can be selected, let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ be a permutation of m numbers from the set $\{1, 2, \ldots, n\}$ and let $\beta = (\beta_1, \beta_2, \ldots, \beta_{n-m})$ be an arbitrary permutation of the remaining $n-m$ numbers, and rewrite equation (13.6′) as

$$\sum_{j=1}^{m} x_{\alpha_j} A_{*\alpha_j} = b - \sum_{j=1}^{n-m} x_{\beta_j} A_{*\beta_j}. \tag{13.6‴}$$

If α is a basic index set, i.e. if the columns $A_{*\alpha_1}, A_{*\alpha_2}, \ldots, A_{*\alpha_m}$ form a basis of $\mathbf{R}^m$, then equation (13.6‴) has clearly a unique solution for each assignment of values to the β-variables, and $(x_{\alpha_1}, x_{\alpha_2}, \ldots, x_{\alpha_m})$ is in fact the coordinates of the vector $b - \sum_{j=1}^{n-m} x_{\beta_j} A_{*\beta_j}$ in this basis. In particular, the coordinates of the vector b are equal to $(\overline{x}_{\alpha_1}, \overline{x}_{\alpha_2}, \ldots, \overline{x}_{\alpha_m})$, where $\overline{x}$ is the corresponding basic solution, defined by the condition that $\overline{x}_{\beta_j} = 0$ for all j.

Conversely, suppose that each assignment of values to the β-variables determines uniquely the values of the α-variables. In particular, the equation

$$\sum_{j=1}^{m} x_{\alpha_j} A_{*\alpha_j} = b \tag{13.7}$$

has then a unique solution, and this implies that the equation

$$\sum_{j=1}^{m} x_{\alpha_j} A_{*\alpha_j} = 0 \tag{13.8}$$

has no other solution then the trivial one, $x_{\alpha_j} = 0$ for all j, because we would otherwise get several solutions to equation (13.7) by to a given one adding a non-trivial solution to equation (13.8). The columns $A_{*\alpha_1}, A_{*\alpha_2}, \ldots, A_{*\alpha_m}$ are in other words linearly independent, and they form a basis for $\mathbf{R}^m$ since they are m in number. Hence, α is a basic index set.

In summary, we have proved the following result.

Theorem 13.3.2. *The variables $x_{\alpha_1}, x_{\alpha_2}, \ldots, x_{\alpha_m}$ are basic variables in the system* (13.6) *if and only if α is a basic index set of the coefficient matrix A.*

Let us now express the basic solution corresponding to the basic index set α in matrix form. By writing the matrix equation (13.6″) in the form

$$A_{*\alpha}x_\alpha + A_{*\beta}x_\beta = b$$

and multiplying from the left by the matrix $A_{*\alpha}^{-1}$, we get

$$\begin{aligned} x_\alpha + A_{*\alpha}^{-1}A_{*\beta}x_\beta &= A_{*\alpha}^{-1}b, \qquad \text{i.e.} \\ x_\alpha &= A_{*\alpha}^{-1}b - A_{*\alpha}^{-1}A_{*\beta}x_\beta, \end{aligned}$$

which expresses the basic variables as linear combinations of the free variables and the coordinates of b. The basic solution is obtained by setting $x_\beta = 0$ and is given by

$$\overline{x}_\alpha = A_{*\alpha}^{-1}b\,, \;\; \overline{x}_\beta = 0.$$

We summarize this result in the following theorem.

Theorem 13.3.3. *Let α be a basic index set of the matrix A. The corresponding basic solution $\overline{x}$ to the system $Ax = b$ is given by the conditions*

$$\overline{x}_\alpha = A_{*\alpha}^{-1}b \quad \textit{and} \quad \overline{x}_k = 0 \quad \textit{for } k \notin \alpha.$$

The $n-m$ free variables in a basic solution are equal to zero by definition. Of course, some basic variable may also happen to be equal to zero, and since this results in certain complications for the simplex algorithm, we make the following definition.

Definition. A basic solution $\overline{x}$ is called *non-degenerate* if $\overline{x}_i \neq 0$ for m indices i and *degenerate* if $\overline{x}_i \neq 0$ for less than m indices i.

Two basic index sets α and α', which are permutations of each other, naturally give rise to the same basic solution $\overline{x}$. So the number of different basic solutions to a system $Ax = b$ with m equations and n unknowns is at most equal to the number of subsets with m elements that can be chosen from the set $\{1, 2, \ldots, n\}$, i.e. at most equal to $\binom{n}{m}$. The number is smaller if the matrix A contains m linearly dependent columns.

EXAMPLE 13.3.3. The system

$$\begin{cases} 3x_1 + x_2 + \ x_3 - 3x_4 = 3 \\ 3x_1 - x_2 + 2x_3 - 6x_4 = 3 \end{cases}$$

has − apart from permutations − the following basic index sets: $(1,2)$, $(1,3)$, $(1,4)$, $(2,3)$ and $(2,4)$, and the corresponding basic solutions are in turn $(1,0,0,0)$, $(1,0,0,0)$, $(1,0,0,0)$, $(0,1,2,0)$ and $(0,1,0,-\frac{2}{3})$. The basic solution $(1,0,0,0)$ is degenerate, and the other two basic solutions are non-degenerate. □

The reason for our interest in basic index sets and basic solutions is that optimal values of LP problems are attained at extreme points, and these points are basic solutions, because we have the following characterisation of extreme points.

Theorem 13.3.4. *Suppose that A is an $m \times n$-matrix of rank m, that $b \in \mathbf{R}^m$ and that $c \in \mathbf{R}^n$. Then:*

(i) $\overline{x}$ is an extreme point of the polyhedron $X = \{x \in \mathbf{R}^n \mid Ax = b,\ x \geq 0\}$ if and only if $\overline{x}$ is a nonnegative basic solution to the system $Ax = b$, i.e. if and only if there is a basic index set α of the matrix A such that $\overline{x}_\alpha = A_{*\alpha}^{-1} b \geq 0$ and $\overline{x}_k = 0$ for $k \notin \alpha$.

(ii) $\overline{y}$ is an extreme point of the polyhedron $Y = \{y \in \mathbf{R}^m \mid A^T y \leq c\}$ if and only if $A^T \overline{y} \leq c$ and there is a basic index set α of the matrix A such that $\overline{y} = (A_{*\alpha}^{-1})^T c_\alpha$.

Proof. (i) According to Theorem 5.1.1, $\overline{x}$ is an extreme point of the polyhedron X if and only if $\overline{x} \geq 0$ and $\overline{x}$ is the unique solution of a system of linear equations consisting of the equation $Ax = b$ and $n - m$ equations out of the n equations $x_1 = 0$, $x_2 = 0$, …, $x_n = 0$. Let $\alpha_1, \alpha_2, \ldots, \alpha_m$ be the indices of the m equations $x_i = 0$ that are not used in this system. Then, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ is a basic index set and $\overline{x}$ is the corresponding basic solution.

(ii) Because of the same theorem, $\overline{y}$ is an extreme point of the polyhedron Y if and only if $\overline{y} \in Y$ and $\overline{y}$ is the unique solution of a quadratic system of linear equations obtained by selecting m out of the n equations in the system $A^T y = c$. Let $\alpha_1, \alpha_2, \ldots, \alpha_m$ denote the indices of the selected equations. The quadratic system is then of the form $(A_{*\alpha})^T y = c_\alpha$, and this system of equations has a unique solution $\overline{y} = (A_{*\alpha}^{-1})^T c_\alpha$ if and only if $A_{*\alpha}$ is an invertible matrix, i.e. if and only if $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ is a basic index set of A. □

EXAMPLE 13.3.4. It follows from Theorem 13.3.4 and Example 13.3.3 that the polyhedron X of solutions to the system

$$\begin{cases} 3x_1 + x_2 + \ x_3 - 3x_4 = 3 \\ 3x_1 - x_2 + 2x_3 - 6x_4 = 3, \ x \geq 0 \end{cases}$$

has two extreme points, namely $(1, 0, 0, 0)$ and $(0, 1, 2, 0)$.

The "dual" polyhedron Y of solutions to the system

$$\begin{cases} 3y_1 + 3y_2 \leq \ 2 \\ y_1 - \ y_2 \leq \ 1 \\ y_1 + 2y_2 \leq \ 1 \\ -3y_1 - 6y_2 \leq -1 \end{cases}$$

has three extreme points, namely $(\frac{5}{6}, -\frac{1}{6})$, $(\frac{1}{3}, \frac{1}{3})$ and $(\frac{7}{9}, -\frac{2}{9})$, corresponding to the basic index sets $(1, 2)$, $(1, 3)$ and $(2, 4)$. (The points associated with the other two basic index sets $(1, 4)$ and $(2, 3)$, $y = (1, -\frac{1}{3})$ and $y = (1, 0)$, respectively, are not extreme points since they lie outside Y.) □

Changing basic index sets

We will now discuss how to generate a suite of basic solutions by successively replacing one element at a time in the basic index set.

Theorem 13.3.5. *Suppose that $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ is a basic index set of the system $Ax = b$ and let $\overline{x}$ denote the corresponding basic solution. Let k be a column index not belonging to the basic index set α, and let $v \in \mathbf{R}^n$ be the column vector defined by*

$$v_\alpha = A_{*\alpha}^{-1} A_{*k}, \quad v_k = -1 \quad \text{and} \quad v_j = 0 \quad \text{for } j \notin \alpha \cup \{k\}.$$

(i) Then $Av = 0$, so it follows that $\overline{x} - tv$ is a solution to the system $Ax = b$ for all $t \in \mathbf{R}$.

(ii) Suppose that $1 \leq r \leq m$ and define a new ordered set α' by replacing the element α_r in α with the number k, i.e.

$$\alpha' = \alpha_{\hat{r}}[k] = (\alpha_1, \ldots, \alpha_{r-1}, k, \alpha_{r+1}, \ldots, \alpha_m).$$

Then, α' is a basic index set if and only if $v_{\alpha_r} \neq 0$. In this case,

$$A_{*\alpha'}^{-1} = E_{\hat{r}}[v_\alpha]^{-1} A_{*\alpha}^{-1}$$

and if $\overline{x}'$ is the basic solution corresponding to the basic index set α', then

$$\overline{x}' = \overline{x} - \tau v,$$

where $\tau = \overline{x}_{\alpha_r} / v_{\alpha_r}$.

(iii) The two basic solutions $\overline{x}$ and $\overline{x}'$ are identical if and only if $\tau = 0$. So if $\overline{x}$ is a non-degenerate basic solution, then $\overline{x} \neq \overline{x}'$.

We will call v the *search vector* associated with the basic index set α and the index k, since we obtain the new basic solution $\overline{x}'$ from the old one $\overline{x}$ by searching in the direction of minus v.

Proof. (i) It follows immediately from the definition of v that

$$Av = \sum_{j\in\alpha} v_j A_{*j} + \sum_{j\notin\alpha} v_j A_{*j} = A_{*\alpha}v_\alpha - A_{*k} = A_{*k} - A_{*k} = 0.$$

(ii) The set α' is a basic index set if and only if $A_{*\alpha'}$ is an invertible matrix. But

$$\begin{aligned} A_{*\alpha}^{-1}A_{*\alpha'} &= A_{*\alpha}^{-1}\left[\, A_{*\alpha_1} \;\ldots\; A_{*\alpha_{r-1}}\; A_{*k}\; A_{*\alpha_{r+1}} \;\ldots\; A_{*\alpha_m}\,\right] \\ &= \left[\, A_{*\alpha}^{-1}A_{*\alpha_1} \;\ldots\; A_{*\alpha}^{-1}A_{*\alpha_{r-1}}\; A_{*\alpha}^{-1}A_{*k}\; A_{*\alpha}^{-1}A_{*\alpha_{r+1}} \;\ldots\; A_{*\alpha}^{-1}A_{*\alpha_m}\,\right] \\ &= \left[\, E_{*1} \;\ldots\; E_{*r-1}\; v_\alpha\; E_{*r+1} \;\ldots\; E_{*m}\,\right] = E_{\hat{r}}[v_\alpha], \end{aligned}$$

where of course E denotes the unit matrix of order m. Hence

$$A_{*\alpha'} = A_{*\alpha}E_{\hat{r}}[v_\alpha].$$

The matrix $A_{*\alpha'}$ is thus invertible if and only if the matrix $E_{\hat{r}}[v_\alpha]$ is invertible, and this is the case if and only if $v_{\alpha_r} \neq 0$, according to Lemma 13.3.1. If the inverse exists, then

$$A_{*\alpha'}^{-1} = \big(A_{*\alpha}E_{\hat{r}}[v_\alpha]\big)^{-1} = E_{\hat{r}}[v_\alpha]^{-1}A_{*\alpha}^{-1}.$$

Now, define $x^\tau = \overline{x} - \tau v$. Then x^τ is a solution to the equation $Ax = b$, by part (i) of the theorem, so in order to prove that x^τ is the basic solution corresponding to the basic index set α', it suffices to show that $x_j^\tau = 0$ for all $j \notin \alpha'$, i.e. for $j = \alpha_r$ and for $j \notin \alpha \cup \{k\}$.

But $x_{\alpha_r}^\tau = \overline{x}_{\alpha_r} - \tau v_{\alpha_r} = 0$, because of the definition of τ, and if $j \notin \alpha\cup\{k\}$ then $\overline{x}_j$ and v_j are both equal to 0, whence $x_j^\tau = \overline{x}_j - \tau v_j = 0$.

(iii) Since $v_k = -1$, we have $\tau v = 0$ if and only if $\tau = 0$. Hence , $\overline{x}' = \overline{x}$ if and only if $\tau = 0$.

If the basic solution $\overline{x}$ is non-degenerate, then $\overline{x}_j \neq 0$ for all $j \in \alpha$ and in particular $\overline{x}_{\alpha_r} \neq 0$, which implies that $\tau \neq 0$, and that $\overline{x}' \neq \overline{x}$. □

Corollary 13.3.6. *Keep the asumptions of Theorem 13.3.5 and suppose in addition that $\overline{x} \geq 0$, that the index set*

$$I_+ = \{j \in \alpha \mid v_j > 0\}$$

is nonempty, and that the index r is chosen so that $\alpha_r \in I_+$ and

$$\tau = \overline{x}_{\alpha_r}/v_{\alpha_r} = \min\{\overline{x}_j/v_j \mid j \in I_+\}.$$

Then $\overline{x}' \geq 0$.

Proof. Since $\overline{x}'_j = 0$ for all $j \notin \alpha'$, it suffices to show that $\overline{x}'_j \geq 0$ for all $j \in \alpha \cup \{k\}$.

We begin by noting that $\tau \geq 0$ since $\overline{x} \geq 0$, and therefore

$$\overline{x}'_k = \overline{x}_k - \tau v_k = 0 + \tau \geq 0.$$

For indices $j \in \alpha \setminus I_+$ we have $v_j \leq 0$, and this implies that

$$\overline{x}'_j = \overline{x}_j - \tau v_j \geq \overline{x}_j \geq 0.$$

Finally, if $j \in I_+$, then $\overline{x}_j/v_j \geq \tau$, and it follows that

$$\overline{x}'_j = \overline{x}_j - \tau v_j \geq 0.$$

This completes the proof. □

13.4 The simplex algorithm

The variant of the simplex algorithm that we shall describe assumes that the LP problem is given in standard form. So we start from the problem

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

where A is an $m \times n$-matrix, $b \in \mathbf{R}^m$ and $c \in \mathbf{R}^n$.

We assume that

$$\operatorname{rank} A = m = \text{the number of rows in } A.$$

Of course, this is no serious restriction, because if $\operatorname{rank} A < m$ and the system $Ax = b$ is consistent, then we can delete $(m - \operatorname{rank} A)$ constraint equations without changing the set of solutions, and this leaves us with an equivalent system $A'x = b$, where the rank of A' is equal to the number of rows in A'.

Let us call a basic index set α of the matrix A and the corresponding basic solution $\overline{x}$ to the system $Ax = b$ *feasible*, if $\overline{x}$ is a feasible point for our standard problem, i.e. if $\overline{x} \geq 0$.

The simplex algorithm starts from a feasible basic index set α of the matrix A, and we shall show in Section 13.6 how to find such an index set by applying the simplex algorithm to a so-called artificial problem.

First compute the corresponding feasible basic solution $\overline{x}$, i.e.

$$\overline{x}_\alpha = A_{*\alpha}^{-1} b \geq 0,$$

and then the number $\lambda \in \mathbf{R}$ and the column vectors $\overline{y} \in \mathbf{R}^m$ and $z \in \mathbf{R}^n$, defined as

$$\begin{aligned} \lambda &= \langle c, \overline{x} \rangle = \langle c_\alpha, \overline{x}_\alpha \rangle \\ \overline{y} &= (A_{*\alpha}^{-1})^T c_\alpha \\ z &= c - A^T \overline{y}. \end{aligned}$$

The number λ is thus equal to the value of the objective function at $\overline{x}$.

Note that $z_\alpha = c_\alpha - (A^T \overline{y})_{\alpha *} = c_\alpha - (A_{*\alpha})^T \overline{y} = c_\alpha - c_\alpha = 0$, so in order to compute the vector z we only have to compute its coordinates

$$z_j = c_j - (A_{*j})^T \overline{y} = c_j - \langle A_{*j}, \overline{y} \rangle$$

for indices $j \notin \alpha$. The numbers z_j are usually called *reduced costs*.

Lemma 13.4.1. *The number λ and the vectors $\overline{x}$, $\overline{y}$ and z have the following properties:*

(i) $\langle z, \overline{x} \rangle = 0$, *i.e. the vectors z and $\overline{x}$ are orthogonal.*

(ii) $Ax = 0 \Rightarrow \langle c, x \rangle = \langle z, x \rangle$.

(iii) $Ax = b \Rightarrow \langle c, x \rangle = \lambda + \langle z, x \rangle$.

(iv) If v is the search vector corresponding to the basic index set α and the index $k \notin \alpha$, then $\langle c, \overline{x} - tv \rangle = \lambda + t z_k$.

Proof. (i) Since $z_j = 0$ for $j \in \alpha$ and $\overline{x}_j = 0$ for $j \notin \alpha$,

$$\langle z, \overline{x} \rangle = \sum_{j \in \alpha} z_j \overline{x}_j + \sum_{j \notin \alpha} z_j \overline{x}_j = 0 + 0 = 0.$$

(ii) It follows immediately from the definition of z that

$$\langle z, x \rangle = \langle c, x \rangle - \langle A^T \overline{y}, x \rangle = \langle c, x \rangle - \langle \overline{y}, Ax \rangle = \langle c, x \rangle$$

for all x satisfying the equation $Ax = 0$.

(iii) If $Ax = b$, then

$$\begin{aligned} \langle c, x \rangle - \langle z, x \rangle = \langle A^T \overline{y}, x \rangle = \langle \overline{y}, Ax \rangle = \langle (A_{*\alpha}^{-1})^T c_\alpha, b \rangle = \langle c_\alpha, A_{*\alpha}^{-1} b \rangle \\ = \langle c_\alpha, \overline{x}_\alpha \rangle = \lambda. \end{aligned}$$

(iv) Since $Av = 0$, it follows from (ii) that

$$\langle c, \overline{x} - tv \rangle = \langle c, \overline{x} \rangle - t \langle c, v \rangle = \lambda - t \langle z, v \rangle = \lambda + t z_k. \qquad \square$$

The following theorem contains all the essential ingredients of the simplex algorithm.

Theorem 13.4.2. *Let α, $\overline{x}$, λ, $\overline{y}$ and z be defined as above.*

(i) ***(Optimality)*** *If $z \geq 0$, then $\overline{x}$ is an optimal solution to the minimization problem*

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

and $\overline{y}$ is an optimal solution to the dual maximization problem

$$\begin{aligned} &\max \ \langle b, y\rangle \\ &\text{s.t.} \quad A^T y \leq c \end{aligned}$$

with λ as the optimal value. The optimal solution $\overline{x}$ to the minimization problem is unique if $z_j > 0$ for all $j \notin \alpha$.

(ii) Suppose that $z \ngeq 0$, and let k be an index such that $z_k < 0$. Let further v be the search vector associated to α and k, i.e.

$$v_\alpha = A_{*\alpha}^{-1} A_{*k}, \quad v_k = -1, \quad v_j = 0 \quad \text{for } j \notin \alpha \cup \{k\},$$

and set $x^t = \overline{x} - tv$ for $t \geq 0$. Depending on whether $v \leq 0$ or $v \nleq 0$, the following applies:

(ii a) ***(Unbounded objective function)*** *If $v \leq 0$, then the points x^t are feasible for the minimization problem for all $t \geq 0$ and $\langle c, x^t\rangle \to -\infty$ as $t \to \infty$. The objective function is thus unbounded below, and the dual maximization problem has no feasible points.*

(ii b) ***(Iteration step)*** *If $v \nleq 0$, then define a new basic index set α' and the number τ as in Theorem 13.3.5 (ii) with the index r chosen as in Corollary 13.3.6. The basic index set α' is feasible with $\overline{x}' = \overline{x} - \tau v$ as the corresponding feasible basic solution, and*

$$\langle c, \overline{x}'\rangle = \langle c, \overline{x}\rangle + \tau z_k \leq \langle c, \overline{x}\rangle.$$

Hence, $\langle c, \overline{x}'\rangle < \langle c, \overline{x}\rangle$, if $\tau > 0$.

Proof. (i) Suppose that $z \geq 0$ and that x is an arbitrary feasible point for the minimization problem. Then $\langle z, x\rangle \geq 0$ (since $x \geq 0$), and it follows from part (iii) of Lemma 13.4.1 that $\langle c, x\rangle \geq \lambda = \langle c, \overline{x}\rangle$. The point $\overline{x}$ is thus optimal and the optimal value is equal to λ.

The condition $z \geq 0$ also implies that $A^T\overline{y} = c - z \leq c$, i.e. $\overline{y}$ is a feasible point for the dual maximization problem, and

$$\langle b, \overline{y}\rangle = \langle \overline{y}, b\rangle = \langle (A_{*\alpha}^{-1})^T c_\alpha, b\rangle = \langle c_\alpha, A_{*\alpha}^{-1} b\rangle = \langle c_\alpha, \overline{x}_\alpha\rangle = \langle c, \overline{x}\rangle,$$

so if follows from the optimality criterion (Corollary 13.1.2) that $\overline{y}$ is an optimal solution to the dual problem.

Now suppose that $z_j > 0$ for all $j \notin \alpha$. If x is a feasible point $\neq \overline{x}$, then $x_{j_0} > 0$ for some index $j_0 \notin \alpha$, and it follows that $\langle z, x\rangle = \sum_{j\notin\alpha} z_j x_j \geq z_{j_0} x_{j_0} > 0$. Hence, $\langle c, x\rangle = \lambda + \langle z, x\rangle > \lambda = \langle c, \overline{x}\rangle$, by Lemma 13.4.1 (iii). This proves that the minimum point is unique.

(ii a) According to Theorem 13.3.5, x^t is a solution to the equation $Ax = b$ for all real numbers t, and if $v \leq 0$ then $x^t = \overline{x} - tv \geq \overline{x} \geq 0$ for $t \geq 0$. So the points x^t are feasible for all $t \geq 0$ if $v \leq 0$, and by Lemma 13.4.1 (iv),

$$\lim_{t\to\infty} \langle c, x^t\rangle = \lambda + \lim_{t\to\infty} z_k t = -\infty.$$

The objective function is thus not bounded below.

Suppose that the dual maximization problem has a feasible point y. Then, $\langle b, y\rangle \leq \langle c, x^t\rangle$ for all $t \geq 0$, by the weak duality theorem, and this is contradictory since the right hand side tends to $-\infty$ as $t \to \infty$. So it follows that the dual maximization problem has no feasible points.

(ii b) By Corollary 13.3.6, α' is a feasible basic solution with x^τ as the corresponding basic solution, and the inequality $\langle c, \overline{x}'\rangle \leq \langle c, \overline{x}\rangle$ now follows directly from Lemma 13.4.1 (iv), because $\tau \geq 0$. □

Theorem 13.4.2 gives rise to the following algorithm for solving the standard problem

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

The simplex algorithm

Given a feasible basic index set α.

1. Compute the matrix $A_{*\alpha}^{-1}$, the corresponding feasible basic solution $\overline{x}$, i.e. $\overline{x}_\alpha = A_{*\alpha}^{-1} b$ and $\overline{x}_j = 0$ for $j \notin \alpha$, and the number $\lambda = \langle c_\alpha, \overline{x}_\alpha\rangle$.

Repeat steps 2–8 until a stop occurs.

2. Compute the vector $\overline{y} = (A_{*\alpha}^{-1})^T c_\alpha$ and the numbers $z_j = c_j - \langle A_{*j}, \overline{y}\rangle$ for $j \notin \alpha$.
3. *Stopping criterion:* **quit** if $z_j \geq 0$ for all $j \notin \alpha$.
   Optimal solution: $\overline{x}$. Optimal value: λ. Optimal dual solution: $\overline{y}$.
4. Choose otherwise an index k such that $z_k < 0$, compute the corresponding search vector v, i.e. $v_\alpha = A_{*\alpha}^{-1} A_{*k}$, $v_k = -1$ and $v_j = 0$ for $j \notin \alpha \cup \{k\}$, and put $I_+ = \{j \in \alpha \mid v_j > 0\}$.
5. *Stopping criterion:* **quit** if $I_+ = \emptyset$.
   Optimal value: $-\infty$.
6. Define otherwise $\tau = \min\{\overline{x}_j / v_j \mid j \in I_+\}$ and determine an index r so that $\alpha_r \in I_+$ and $\overline{x}_{\alpha_r} / v_{\alpha_r} = \tau$.

7. Put $\alpha' = \alpha_{\hat{r}}[k]$ and compute the inverse $A_{*\alpha'}^{-1} = E_{\hat{r}}[v_\alpha]^{-1} A_{*\alpha}^{-1}$.
8. *Update:* $\alpha := \alpha'$, $A_{*\alpha}^{-1} := A_{*\alpha'}^{-1}$, $\overline{x} := \overline{x} - \tau v$, and $\lambda := \lambda + \tau z_k$.

Before we can call the above procedure an algorithm in the sense of a mechanical calculation that a machine can perform, we need to specify how to choose k in step 4 in the case when $z_j < 0$ for several indices j, and r in step 6 when $\overline{x}_j / v_j = \tau$ for more than one index $j \in I_+$.

A simple rule that works well most of the time, is to select the index j that minimizes z_j (and if there are several such indices the least of these) as the index k, and the smallest of all indices i for which $\overline{x}_{\alpha_i} / v_{\alpha_i} = \tau$ as the index r. We shall return to the choice of k and r later; for the immediate discussion of the algorithm, it does not matter how to make the choice.

We also need a method to find an initial feasible basic index set to start the simplex algorithm from. We shall treat this problem and solve it in Section 13.6.

Now suppose that we apply the simplex algorithm to an LP problem in standard form, starting from a feasible basic index set. It follows from Theorem 13.4.2 that the algorithm delivers an optimal solution if it stops during step 3, and that the objective function is unbounded from below if the algorithm stops during step 5.

So let us examine what happens if the algorithm does not stop. Since a feasible basic index set is generated each time the algorithm comes to step 7, we will obtain in this case an infinite sequence $\alpha^1, \alpha^2, \alpha^3, \ldots$ of feasible basic index sets with associated feasible basic solutions $\overline{x}^1, \overline{x}^2, \overline{x}^3, \ldots$. As the number of different basic index sets is finite, some index set α^p has to be repeated after a number of additional, say q, iterations. This means that $\alpha^p = \alpha^{p+q}$ and $\overline{x}^p = \overline{x}^{p+q}$ and in turn implies that the sequence $\alpha^p, \alpha^{p+1}, \ldots, \alpha^{p+q-1}$ is repeated periodically in all infinity. We express this by saying that the algorithm *cycles*. According to (ii) in Theorem 13.4.2,

$$\langle c, \overline{x}^p \rangle \geq \langle c, \overline{x}^{p+1} \rangle \geq \cdots \geq \langle c, \overline{x}^{p+q} \rangle = \langle c, \overline{x}^p \rangle,$$

and this implies that

$$\langle c, \overline{x}^p \rangle = \langle c, \overline{x}^{p+1} \rangle = \cdots = \langle c, \overline{x}^{p+q-1} \rangle.$$

The number τ is hence equal to 0 for all the iterations of the cycle, and this implies that the basic solutions $\overline{x}^p, \overline{x}^{p+1}, \ldots, \overline{x}^{p+q-1}$ are identical and degenerate. If the simplex algorithm does not stop, but continues indefinitely, it is so because the algortihm has got stuck in a degenerate basic solution.

The following theorem is now an immediate consequence of the above discussion.

Theorem 13.4.3. *The simplex algorithm stops when applied to an LP problem in which all feasible basic solutions are non-degenerate.*

Cycling can occur, and we shall give an example of this in the next section. Theoretically, this is a bit troublesome, but cycling seems to be a rare phenomenon in practical problems and therefore lacks practical significance. The small rounding errors introduced during the numerical treatment of an LP problem also have a beneficial effect since these errors usually turn degenerate basic solutions into non-degenerate solutions and thereby tend to prevent cycling. There is also a simple rule for the choice of indices k and r, *Bland's rule*, which prevents cycling and will be described in the next section.

Example

EXAMPLE 13.4.1. We now illustrate the simplex algorithm by solving the minimization problem

$$\begin{array}{ll} \min & x_1 - x_2 + x_3 \\ \text{s.t.} & \begin{cases} -2x_1 + x_2 + \ x_3 \leq 3 \\ -x_1 + x_2 - 2x_3 \leq 3 \\ 2x_1 - x_2 + 2x_3 \leq 1, \ x \geq 0. \end{cases} \end{array}$$

We start by writing the problem in standard form by introducing three slack variables:

$$\begin{array}{ll} \min & x_1 - x_2 + x_3 \\ \text{s.t.} & \begin{cases} -2x_1 + x_2 + \ x_3 + x_4 = 3 \\ -x_1 + x_2 - 2x_3 + x_5 = 3 \\ 2x_1 - x_2 + 2x_3 + x_6 = 1, \ x \geq 0. \end{cases} \end{array}$$

Using matrices, this becomes

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

with

$$A = \begin{bmatrix} -2 & 1 & 1 & 1 & 0 & 0 \\ -1 & 1 & -2 & 0 & 1 & 0 \\ 2 & -1 & 2 & 0 & 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 3 \\ 3 \\ 1 \end{bmatrix} \quad \text{and}$$
$$c^T = \begin{bmatrix} 1 & -1 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

We note that we can start the simplex algorithm with

$$\alpha = (4,5,6), \quad A_{*\alpha}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \overline{x}_\alpha = \begin{bmatrix} 3 \\ 3 \\ 1 \end{bmatrix},$$

$$\lambda = \langle c_\alpha, \overline{x}_\alpha \rangle = c_\alpha^T \overline{x}_\alpha = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 3 \\ 3 \\ 1 \end{bmatrix} = 0.$$

1st iteration:

$$\overline{y} = (A_{*\alpha}^{-1})^T c_\alpha = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$z_{1,2,3} = c_{1,2,3} - (A_{*1,2,3})^T \overline{y} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} - \begin{bmatrix} -2 & -1 & 2 \\ 1 & 1 & -1 \\ 1 & -2 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}.$$

Since $z_2 = -1 < 0$, we have to select $k = 2$ and then

$$v_\alpha = A_{*\alpha}^{-1} A_{*k} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \quad v_2 = -1$$

$$I_+ = \{j \in \alpha \mid v_j > 0\} = \{4,5\}$$

$$\tau = \min\{\overline{x}_j / v_j \mid j \in I_+\} = \min\{\overline{x}_4/v_4, \overline{x}_5/v_5\} = \min\{3/1, 3/1\} = 3$$

for $\alpha_1 = 4$, i.e.

$$r = 1$$

$$\alpha' = \alpha_{\hat{r}}[k] = (4,5,6)_{\hat{1}}[2] = (2,5,6)$$

$$E_{\hat{r}}[v_\alpha]^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$A_{*\alpha'}^{-1} = E_{\hat{r}}[v_\alpha]^{-1} A_{*\alpha}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\overline{x}'_{\alpha'} = \overline{x}_{\alpha'} - \tau v_{\alpha'} = \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix} - 3 \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \\ 4 \end{bmatrix}$$

$$\lambda' = \lambda + \tau z_k = 0 + 3\,(-1) = -3.$$

Update: $\alpha := \alpha'$, $A_{*\alpha}^{-1} := A_{*\alpha'}^{-1}$, $\overline{x}_\alpha := \overline{x}'_{\alpha'}$ and $\lambda := \lambda'$.

2nd iteration:

$$\overline{y} = (A_{*\alpha}^{-1})^T c_\alpha = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}$$

$$z_{1,3,4} = c_{1,3,4} - (A_{*1,3,4})^T \overline{y} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} - \begin{bmatrix} -2 & -1 & 2 \\ 1 & -2 & 2 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}.$$

Since $z_1 = -1 < 0$,

$$\begin{aligned} k &= 1 \\ v_\alpha = A_{*\alpha}^{-1} A_{*k} &= \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} -2 \\ -1 \\ 2 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}, \quad v_1 = -1 \\ I_+ &= \{j \in \alpha \mid v_j > 0\} = \{5\} \\ \tau &= \overline{x}_5 / v_5 = 0/1 = 0 \quad \text{for } \alpha_2 = 5, \text{ i.e.} \\ r &= 2 \\ \alpha' &= \alpha_{\hat{r}}[k] = (2,5,6)_{\hat{2}}[1] = (2,1,6) \end{aligned}$$

$$E_{\hat{r}}[v_\alpha]^{-1} = \begin{bmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$A_{*\alpha'}^{-1} = E_{\hat{r}}[v_\alpha]^{-1} A_{*\alpha}^{-1} = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 2 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\overline{x}'_{\alpha'} = \overline{x}_{\alpha'} - \tau v_{\alpha'} = \begin{bmatrix} 3 \\ 0 \\ 4 \end{bmatrix} - 0 \begin{bmatrix} -2 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \\ 4 \end{bmatrix}$$

$$\lambda' = \lambda + \tau z_k = -3 + 0\,(-1) = -3.$$

Update: $\alpha := \alpha'$, $A_{*\alpha}^{-1} := A_{*\alpha'}^{-1}$, $\overline{x}_\alpha := \overline{x}'_{\alpha'}$ and $\lambda := \lambda'$.

3rd iteration:

$$\overline{y} = (A_{*\alpha}^{-1})^T c_\alpha = \begin{bmatrix} -1 & -1 & 1 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix}$$

$$z_{3,4,5} = c_{3,4,5} - (A_{*3,4,5})^T \overline{y} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 & -2 & 2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}.$$

Since $z_3 = -1 < 0$,

$$
\begin{aligned}
k &= 3 \\
v_\alpha = A_{*\alpha}^{-1} A_{*k} &= \begin{bmatrix} -1 & 2 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \\ 2 \end{bmatrix} = \begin{bmatrix} -5 \\ -3 \\ 3 \end{bmatrix}, \quad v_3 = -1 \\
I_+ &= \{j \in \alpha \mid v_j > 0\} = \{6\} \\
\tau &= \overline{x}_6 / v_6 = 4/3 \quad \text{for } \alpha_3 = 6, \text{ i.e.} \\
r &= 3 \\
\alpha' &= \alpha_{\hat{r}}[k] = (2,1,6)_{\hat{3}}[3] = (2,1,3) \\
E_{\hat{r}}[v_\alpha]^{-1} &= \begin{bmatrix} 1 & 0 & -5 \\ 0 & 1 & -3 \\ 0 & 0 & 3 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & \frac{5}{3} \\ 0 & 1 & 1 \\ 0 & 0 & \frac{1}{3} \end{bmatrix} \\
A_{*\alpha'}^{-1} = E_{\hat{r}}[v_\alpha]^{-1} A_{*\alpha}^{-1} &= \begin{bmatrix} 1 & 0 & \frac{5}{3} \\ 0 & 1 & 1 \\ 0 & 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} -1 & 2 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} & 2 & \frac{5}{3} \\ 0 & 1 & 1 \\ \frac{1}{3} & 0 & \frac{1}{3} \end{bmatrix} \\
\overline{x}'_{\alpha'} = \overline{x}_{\alpha'} - \tau v_{\alpha'} &= \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix} - \frac{4}{3} \begin{bmatrix} -5 \\ -3 \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{29}{3} \\ 4 \\ \frac{4}{3} \end{bmatrix} \\
\lambda' = \lambda + \tau z_k &= -3 + \frac{4}{3}(-1) = -\frac{13}{3}.
\end{aligned}
$$

Update: $\alpha := \alpha'$, $A_{*\alpha}^{-1} := A_{*\alpha'}^{-1}$, $\overline{x}_\alpha := \overline{x}'_{\alpha'}$ and $\lambda := \lambda'$.

4th iteration:

$$
\begin{aligned}
\overline{y} = (A_{*\alpha}^{-1})^T c_\alpha &= \begin{bmatrix} \frac{2}{3} & 0 & \frac{1}{3} \\ 2 & 1 & 0 \\ \frac{5}{3} & 1 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{3} \\ -1 \\ -\frac{1}{3} \end{bmatrix} \\
z_{4,5,6} = c_{4,5,6} - (A_{*4,5,6})^T \overline{y} &= \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -\frac{1}{3} \\ -1 \\ -\frac{1}{3} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ 1 \\ \frac{1}{3} \end{bmatrix}.
\end{aligned}
$$

The solution $\overline{x} = (4, \frac{29}{3}, \frac{4}{3}, 0, 0, 0)$ is optimal with optimal value $-\frac{13}{3}$ since $z_{4,5,6} > 0$. The original minimization problem has the same optimal value, of course, and $(x_1, x_2, x_3) = (4, \frac{29}{3}, \frac{4}{3})$ is the optimal solution. □

The version of the simplex algorithm that we have presented is excellent for computer calculations, but it is unnecessarily complicated for calculations by hand. Then it is better to use the tableau form which we utilized in

Section 13.2, even if this entails performing unnecessary calculations. To the LP problem

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

we associate the following simplex tableau:

$$\begin{array}{c|c|c} A & b & E \\ \hline c^T & 0 & 0^T \end{array} \tag{13.9}$$

We have included the column on the far right of the table only to explain how the tableau calculations work; it will be omitted later on.

Let α be a feasible basic index set with $\overline{x}$ as the corresponding basic solution. The upper part $[\,A \mid b \mid E\,]$ of the tableau can be seen as a matrix, and by multiplying this matrix from the left by $A_{*\alpha}^{-1}$, we obtain the following new tableau:

$$\begin{array}{c|c|c} A_{*\alpha}^{-1}A & A_{*\alpha}^{-1}b & A_{*\alpha}^{-1} \\ \hline c^T & 0 & 0^T \end{array}$$

Now subtract the upper part of this tableau multiplied from the left by c_α^T from the bottom row of the tableau. This results in the tableau

$$\begin{array}{c|c|c} A_{*\alpha}^{-1}A & A_{*\alpha}^{-1}b & A_{*\alpha}^{-1} \\ \hline c^T - c_\alpha^T A_{*\alpha}^{-1}A & -c_\alpha^T A_{*\alpha}^{-1}b & -c_\alpha^T A_{*\alpha}^{-1} \end{array}$$

Using the notation introduced in the definition of the simplex algorithm, we have $A_{*\alpha}^{-1}b = \overline{x}_\alpha$, $c_\alpha^T A_{*\alpha}^{-1} = ((A_{*\alpha}^{-1})^T c_\alpha)^T = \overline{y}^T$, $c^T - c_\alpha^T A_{*\alpha}^{-1}A = c^T - \overline{y}^T A = z^T$ and $c_\alpha^T A_{*\alpha}^{-1}b = c_\alpha^T \overline{x}_\alpha = \langle c_\alpha, \overline{x}_\alpha\rangle = \lambda$, which means that the above tableau can be written in the form

$$\begin{array}{c|c|c} A_{*\alpha}^{-1}A & \overline{x}_\alpha & A_{*\alpha}^{-1} \\ \hline z^T & -\lambda & -\overline{y}^T \end{array} \tag{13.10}$$

Note that the columns of the unit matrix appear as columns in the matrix $A_{*\alpha}^{-1}A$, because column number α_j in $A_{*\alpha}^{-1}A$ is identical with unit matrix column E_{*j}. Moreover, $z_{\alpha_j} = 0$.

When performing the actual calculations, we use Gaussian elimination to get from tableau (13.9) to tableau (13.10).

If $z^T \geq 0$, which we can determine with the help of the bottom line in (13.10), then $\overline{x}$ is an optimal solution, and we can also read off the optimal solution $\overline{y}$ to the dual maximization problem. (The matrix A will in many cases contain the columns of the unit matrix, and if so then it is of course possible to read off the solution to the dual problem in the final simplex tableau without first having to add the unit matrix on the right side of tableau (13.9).)

If $z^T \not\geq 0$, then we choose a column index k with $z_k < 0$, and consider the corresponding column $a = A_{*\alpha}^{-1} A_{*k}$ $(= v_\alpha)$ in the upper part of the tableau.

The minimization problem is unbounded if $a \leq 0$. In the opposite case, we choose an index $i = r$ that minimizes $\overline{x}_{\alpha_i}/a_i$ $(= \overline{x}_{\alpha_i}/v_{\alpha_i})$ among all ratios with positive a_i. This means that r is the index of a row with the least ratio $\overline{x}_{\alpha_i}/a_i$ among all rows with positive a_i. Finally, we transform the simplex tableau by pivoting around the element at location (r, k).

Example 13.4.2. We solve Example 13.4.1 again − this time by performing all calculations in tabular form. Our first tableau has the form

$$\left.\begin{array}{rrrrrr|r} -2 & \underline{1} & 1 & 1 & 0 & 0 & 3 \\ -1 & 1 & -2 & 0 & 1 & 0 & 3 \\ 2 & -1 & 2 & 0 & 0 & 1 & 1 \\ \hline 1 & -1 & 1 & \underline{0} & \underline{0} & \underline{0} & 0 \end{array}\right.$$

and in this case it is of course not necessary to repeat the columns of the unit matrix in a separate part of the tableau in order also to solve the dual problem.

The basic index set $\alpha = (4, 5, 6)$ is feasible, and since $A_{*\alpha} = E$ and $c_\alpha^T = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$, we can directly read off $z^T = \begin{bmatrix} 1 & -1 & 1 & 0 & 0 & 0 \end{bmatrix}$ and $-\lambda = 0$ from the bottom line of the tableau.

The optimality criterion is not satisfied since $z_2 = -1 < 0$, so we proceed by choosing $k = 2$. The positive ratios of corresponding elements in the right-hand side column and the second column are in this case the same and equal to 3/1 for the first and the second row. Therefore, we can choose $r = 1$ or $r = 2$, and we decide to use the smaller of the two numbers, i.e. we put $r = 1$. The tableau is then transformed by pivoting around the element at location $(1, 2)$. By then continuing in the same style, we get the following sequence of tableaux:

$$\begin{array}{cccccc|c}
-2 & 1 & 1 & 1 & 0 & 0 & 3 \\
\underline{1} & 0 & -3 & -1 & 1 & 0 & 0 \\
0 & 0 & 3 & 1 & 0 & 1 & 4 \\
\hline
-1 & \underline{0} & 2 & 1 & \underline{0} & \underline{0} & 3
\end{array}$$

$\alpha = (2, 5, 6), \quad k = 1, \quad r = 2$

$$\begin{array}{cccccc|c}
0 & 1 & -5 & -1 & 2 & 0 & 3 \\
1 & 0 & -3 & -1 & 1 & 0 & 0 \\
0 & 0 & \underline{3} & 1 & 0 & 1 & 4 \\
\hline
\underline{0} & \underline{0} & -1 & 0 & 1 & \underline{0} & 3
\end{array}$$

$\alpha = (2, 1, 6), \quad k = 3, \quad r = 3$

$$\begin{array}{cccccc|c}
0 & 1 & 0 & \frac{2}{3} & 2 & \frac{5}{3} & \frac{29}{3} \\
1 & 0 & 0 & 0 & 1 & 1 & 4 \\
0 & 0 & 1 & \frac{1}{3} & 0 & \frac{1}{3} & \frac{4}{3} \\
\hline
\underline{0} & \underline{0} & \underline{0} & \frac{1}{3} & 1 & \frac{1}{3} & \frac{13}{3}
\end{array}$$

$\alpha = (2, 1, 3)$

The optimality criterion is now satisfied with $\overline{x} = (4, \frac{29}{3}, \frac{4}{3}, 0, 0, 0)$ as optimal solution and $-\frac{13}{3}$ as optimal value. The dual problem has the optimal solution $(-\frac{1}{3}, -1, -\frac{1}{3})$. □

Henceforth, we will use the tableau variant of the simplex algorithm to account for our calculations, because it is the most transparent method.

The optimality condition in step 2 of the simplex algorithm is a sufficient condition for optimality, but the condition is not necessary. A degenerate basic solution can be optimal without the optimality condition being satisfied. Here is a trivial example of this.

Example 13.4.3. The problem

$$\begin{array}{ll} \min & -x_2 \\ \text{s.t.} & x_1 + x_2 = 0, \;\; x \geq 0 \end{array}$$

has only one feasible point, $x = (0, 0)$, which is therefore optimal. There are two feasible basic index sets, $\alpha = (1)$ and $\alpha' = (2)$, both with $(0, 0)$ as the corresponding degenerate basic solution.

The optimality condition is not fulfilled at the basic index set α, because $\overline{y} = 1 \cdot 0 = 0$ and $z_2 = -1 - 1 \cdot 0 = -1 < 0$. At the other basic index set α',

$\overline{y} = 1 \cdot (-1) = -1$ and $z_2 = 0 - 1 \cdot (-1) = 1 > 0$, and the optimality criterion is now satisfied.

The corresponding simplex tableaux are

$$\begin{array}{cc|c} 1 & 1 & 0 \\ \hline \underline{0} & -1 & 0 \end{array} \qquad \text{and} \qquad \begin{array}{cc|c} 1 & 1 & 0 \\ \hline 1 & \underline{0} & 0 \end{array}$$

$$\alpha = (1) \qquad\qquad \alpha = (2)$$

□

We shall now study a simple example with a non-unique optimal solution.

EXAMPLE 13.4.4. The simplex tableaux associated with the problem

$$\begin{array}{ll} \min & x_1 + x_2 \\ \text{s.t.} & \begin{cases} x_1 + x_2 - x_3 = 1 \\ 2x_2 - x_3 + x_4 = 1, \ x \geq 0 \end{cases} \end{array}$$

are as follows:

$$\begin{array}{cccc|c} 1 & 1 & -1 & 0 & 1 \\ 0 & 2 & -1 & 1 & 1 \\ \hline 1 & 1 & 0 & 0 & 0 \end{array}$$

$$\alpha = (1, 4)$$

$$\begin{array}{cccc|c} 1 & 1 & -1 & 0 & 1 \\ 0 & 2 & -1 & 1 & 1 \\ \hline \underline{0} & 0 & 1 & \underline{0} & -1 \end{array}$$

$$\alpha = (1, 4)$$

The optimality condition is met; $\overline{x} = (1, 0, 0, 1)$ is an optimal solution, and the optimal value is 1. However, coefficient number 2 in the last row, i.e. z_2, is equal to 0, so we can therefore perform another iteration of the simplex algorithm by choosing the second column as the pivot column and the second row as the pivot row, i.e. $k = 2$ and $r = 2$. This gives rise to the following new tableau:

$$\begin{array}{cccc|c} 1 & 0 & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ 0 & 1 & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \hline \underline{0} & \underline{0} & 1 & 0 & -1 \end{array}$$

$$\alpha = (1, 2)$$

The optimality condition is again met, now with $\hat{x} = (\frac{1}{2}, \frac{1}{2}, 0, 0)$ as optimal solution. Since the set of optimal solutions is convex, each point on the line segment between $\hat{x}$ and $\overline{x}$ is also an optimal point. □

13.5 Bland's anti cycling rule

We begin with an example of Kuhn showing that cycling can occur in degenerate LP problems if the column index k and the row index r are not properly selected.

EXAMPLE 13.5.1. Consider the problem

$$\begin{array}{ll}\min & -2x_1 - 3x_2 + x_3 + 12x_4 \\ \text{s.t.} & \begin{cases} -2x_1 - 9x_2 + x_3 + 9x_4 + x_5 = 0 \\ \frac{1}{3}x_1 + x_2 - \frac{1}{3}x_3 - 2x_4 + x_6 = 0 \\ 2x_1 + 3x_2 - x_3 - 12x_4 + x_7 = 2, \ x \geq 0. \end{cases}\end{array}$$

We use the simplex algorithm with the additional rule that the column index k should be chosen so as to make z_k as negative as possible and the row index r should be the least among all allowed row indices. Our first tableau is

$$\begin{array}{rrrrrrr|r} -2 & -9 & 1 & 9 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & \underline{1} & -\frac{1}{3} & -2 & 0 & 1 & 0 & 0 \\ 2 & 3 & -1 & -12 & 0 & 0 & 1 & 2 \\ \hline -2 & -3 & 1 & 12 & \underline{0} & \underline{0} & \underline{0} & 0 \end{array}$$

with $\alpha = (5, 6, 7)$ as feasible basic index set. According to our rule for the choice of of k, we must choose $k = 2$. There is only one option for the row index r, namely $r = 2$, so we use the element located at $(2, 2)$ as pivot element and obtain the following new tableau

$$\begin{array}{rrrrrrr|r} \underline{1} & 0 & -2 & -9 & 1 & 9 & 0 & 0 \\ \frac{1}{3} & 1 & -\frac{1}{3} & -2 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & -6 & 0 & -3 & 1 & 2 \\ \hline -1 & \underline{0} & 0 & 6 & \underline{0} & 3 & \underline{0} & 0 \end{array}$$

with $\alpha = (5, 2, 7)$. This time $k = 1$, but there are two row indices i with the same least value of the ratios $\overline{x}_{\alpha_i}/v_{\alpha_i}$, namely 1 and 2. Our additional rule tells us to choose $r = 1$. Pivoting around the element at location $(1, 1)$ results in the next tableau

$$\begin{array}{rrrrrrr|r} 1 & 0 & -2 & -9 & 1 & 9 & 0 & 0 \\ 0 & 1 & \frac{1}{3} & \underline{1} & -\frac{1}{3} & -2 & 0 & 0 \\ 0 & 0 & 2 & 3 & -1 & -12 & 1 & 2 \\ \hline \underline{0} & \underline{0} & -2 & -3 & 1 & 12 & \underline{0} & 0 \end{array}$$

with $\alpha = (1, 2, 7)$, $k = 4$, $r = 2$.

The algorithms goes on with the following sequence of tableaux:

$$\begin{array}{rrrrrrr|r} 1 & 9 & \underline{1} & 0 & -2 & -9 & 0 & 0 \\ 0 & 1 & \frac{1}{3} & 1 & -\frac{1}{3} & -2 & 0 & 0 \\ 0 & -3 & 1 & 0 & 0 & -6 & 1 & 2 \\ \hline \underline{0} & 3 & -1 & \underline{0} & 0 & 6 & \underline{0} & 0 \end{array}$$

$$\alpha = (1, 4, 7), \quad k = 3, \quad r = 1$$

$$\begin{array}{rrrrrrr|r} 1 & 9 & 1 & 0 & -2 & -9 & 0 & 0 \\ -\frac{1}{3} & -2 & 0 & 1 & \frac{1}{3} & \underline{1} & 0 & 0 \\ -1 & -12 & 0 & 0 & 2 & 3 & 1 & 2 \\ \hline \underline{1} & 12 & \underline{0} & \underline{0} & -2 & -3 & \underline{0} & 0 \end{array}$$

$$\alpha = (3, 4, 7), \quad k = 6, \quad r = 2$$

$$\begin{array}{rrrrrrr|r} -2 & -9 & 1 & 9 & \underline{1} & 0 & 0 & 0 \\ -\frac{1}{3} & -2 & 0 & 1 & \frac{1}{3} & 1 & 0 & 0 \\ 0 & -6 & 0 & -3 & 1 & 0 & 1 & 2 \\ \hline 0 & 6 & \underline{0} & 3 & -1 & \underline{0} & \underline{0} & 0 \end{array}$$

$$\alpha = (3, 6, 7), \quad k = 5, \quad r = 1$$

$$\begin{array}{rrrrrrr|r} -2 & -9 & 1 & 9 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & \underline{1} & -\frac{1}{3} & -2 & 0 & 1 & 0 & 0 \\ 2 & 3 & -1 & -12 & 0 & 0 & 1 & 2 \\ \hline -2 & -3 & 1 & 12 & \underline{0} & \underline{0} & \underline{0} & 0 \end{array}$$

$$\alpha = (5, 6, 7)$$

After six iterations we are back to the starting tableau. The simplex algorithm cycles! □

We now introduce a rule for the choice of indices k and r that prevents cycling.

Bland's rule: *Choose k in step 4 of the simplex algorithm so that*

$$k = \min\{j \mid z_j < 0\}$$

and r in step 6 so that

$$\alpha_r = \min\{j \in I_+ \mid \overline{x}_j / v_j = \tau\}.$$

EXAMPLE 13.5.2. Consider again the minimization problem in the previous example and now use the simplex algorithm with Bland's rule. This results in the following sequence of tableaux:

$$\begin{array}{rrrrrrr|r}
-2 & -9 & 1 & 9 & 1 & 0 & 0 & 0 \\
\underline{\frac{1}{3}} & 1 & -\frac{1}{3} & -2 & 0 & 1 & 0 & 0 \\
2 & 3 & -1 & -12 & 0 & 0 & 1 & 2 \\
\hline
-2 & -3 & 1 & 12 & \underline{0} & \underline{0} & \underline{0} & 0
\end{array}$$

$$\alpha = (5, 6, 7), \quad k = 1, \quad r = 2$$

$$\begin{array}{rrrrrrr|r}
0 & -3 & -1 & -3 & 1 & 6 & 0 & 0 \\
1 & 3 & -1 & -6 & 0 & 3 & 0 & 0 \\
0 & -3 & \underline{1} & 0 & 0 & -6 & 1 & 2 \\
\hline
\underline{0} & 3 & -1 & 0 & \underline{0} & 6 & \underline{0} & 0
\end{array}$$

$$\alpha = (5, 1, 7), \quad k = 3, \quad r = 3$$

$$\begin{array}{rrrrrrr|r}
0 & -6 & 0 & -3 & 1 & 0 & 0 & 2 \\
1 & 0 & 0 & -6 & 0 & -3 & 1 & 2 \\
0 & -3 & 1 & 0 & 0 & -6 & 1 & 2 \\
\hline
\underline{0} & 0 & \underline{0} & 0 & \underline{0} & 12 & 1 & 2
\end{array}$$

$$\alpha = (5, 1, 3)$$

The optimality criterion is met with $\overline{x} = (2, 0, 2, 0, 2, 0, 0)$ as optimal solution and -2 as optimal value. □

Theorem 13.5.1. *The simplex algorithm always stops if Bland's rule is used.*

Proof. We prove the theorem by contradiction. So suppose that the simplex algorithm cycles when applied to some given LP problem, and let $\overline{x}$ be the common basic solution during the iterations of the cycle.

Let $\mathcal{C}$ denote the set of indices k of the varibles x_k that change from being basic to being free during the iterations of the cycle. Since these variables have to return as basic variables during the cycle, $\mathcal{C}$ is of course also equal to the set of indices of the variables x_k that change from being free to being basic during the cycle. Moreover, $\overline{x}_k = 0$ for all $k \in \mathcal{C}$.

Let

$$q = \max\{j \mid j \in \mathcal{C}\},$$

and let α be the basic index set which is in use during the iteration in the cycle when the variabel x_q changes from being basic to being free, and let x_k be the free variable that replaces x_q. The index q is in other words replaced by k in the basic index set that follows after α. The corresponding search vector v and reduced cost vector z satisfy the inequalities

$$z_k < 0 \quad \text{and} \quad v_q > 0,$$

and

$$z_j \geq 0 \quad \text{for } j < k.$$

since the index k is chosen according to Bland's rule. Since $k \in \mathcal{C}$, we also have $k < q$, because of the definition of q.

Let us now consider the basic index set α' that belongs to an iteration when x_q returns as a basic variables after having been free. Because of Bland's rule for the choice of incoming index, in this case q, the corresponding reduced cost vector z' has to satisfy the following inequalities:

$$z'_j \geq 0 \quad \text{for } j < q \quad \text{and} \quad z'_q < 0. \tag{13.11}$$

Especially, thus $z'_k \geq 0$.

Since $Av = 0$, $v_k = -1$ and $v_j = 0$ for $j \notin \alpha \cup \{k\}$, and $z_j = 0$ for $j \in \alpha$, it follows from Lemma 13.4.1 that

$$\sum_{j\in\alpha} z'_j v_j - z'_k = \langle z', v\rangle = \langle c, v\rangle = \langle z, v\rangle = \sum_{j\in\alpha} z_j v_j + z_k v_k = -z_k > 0,$$

and hence

$$\sum_{j\in\alpha} z'_j v_j > z'_k \geq 0.$$

There is therefore an index $j_0 \in \alpha$ such that $z'_{j_0} v_{j_0} > 0$. Hence $z'_{j_0} \neq 0$, which means that j_0 can not belong to the index set α'. The variable x_{j_0} is in other words basic during one iteration of the cycle and free during another iteration. This means that j_0 is an index in the set $\mathcal{C}$, and hence $j_0 \leq q$, by the definition of q. The case $j_0 = q$ is impossible since $v_q > 0$ and $z'_q < 0$. Thus $j_0 < q$, and it now follows from (13.11) that $z'_{j_0} > 0$. This implies in turn that $v_{j_0} > 0$, because the product $z'_{j_0} v_{j_0}$ is positive. So j_0 belongs to the set $I_+ = \{j \in \alpha \mid v_j > 0\}$, and since $\overline{x}_{j_0}/v_{j_0} = 0 = \tau$, it follows that

$$\min\{j \in I_+ \mid \overline{x}_j/v_j = \tau\} \leq j_0 < q.$$

The choice of q thus contradicts Bland's rule for how to choose index to leave the basic index set α, and this contradiction proves the theorem. □

Remark. It is not necessary to use Bland's rule all the time in order to prevent cycling; it suffices to use it in iterations with $\tau = 0$.

13.6 Phase 1 of the simplex algorithm

The simplex algorithm assumes that there is a feasible basic index set to start from. For some problems we will automatically get one when the problem is written in standard form. This is the case for problems of the type

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

where A is an $m \times n$-matrix and the right-hand side vector b is nonnegative. By introducing m slack variables $s_{n+1}, s_{n+2}, \ldots, s_{n+m}$ and defining

$$s = (s_{n+1}, s_{n+2}, \ldots, s_{n+m}),$$

we obtain the standard problem

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

and it is now obvious how to start; the slack variables will do as basic variables, i.e. $\alpha = (n+1, n+2, \ldots, n+m)$ is a feasible basic index set with $\overline{x} = 0$, $\overline{s} = b$ as the corresponding basic solution.

In other cases, it is not at all obvious how to find a feasible basic index set to start from, but one can always generate such a set by using the simplex algorithm on a suitable artificial problem.

Consider an arbitrary standard LP problem

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

where A is an $m \times n$-matrix. We can assume without restriction that $b \ge 0$, for if any b_j is negative, we just multiply the corresponding equation by -1.

We begin by choosing an $m \times k$-matris B so that the matrix

$$A' = \begin{bmatrix} A & B \end{bmatrix}$$

gets rank equal to m and the system

$$A' \begin{bmatrix} x \\ y \end{bmatrix} = Ax + By = b$$

gets an obvious feasible basic index set α^0. The new y-variables are called *artificial variables*, and we number them so that $y = (y_{n+1}, y_{n+2}, \ldots, y_{n+k})$.

A trivial way to achieve this is to choose B equal to the unit matrix E of order m, for $\alpha^0 = (n+1, n+2, \ldots, n+m)$ is then a feasible basic index

set with $(\overline{x}, \overline{y}) = (0, b)$ as the corresponding feasible basic solution. Often, however, A already contains a number of unit matrix columns, and then it is sufficient to add the missing unit matrix columns to A.

Now let

$$\mathbf{1} = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^T$$

be the $k \times 1$-matrix consisting of k ones, and consider the following artificial LP problems:

$$\begin{array}{ll} \min & \langle \mathbf{1}, y \rangle = y_{n+1} + \cdots + y_{n+k}. \\ \text{s.t.} & Ax + By = b,\ x, y \geq 0 \end{array}$$

The optimal value is obviously ≥ 0, and *the value is equal to zero if and only if there is a feasible solution of the form* $(x, 0)$*, i.e. if and only if there is a nonnegative solution to the system* $Ax = b$.

Therefore, we solve the artificial problem using the simplex algorithm with α^0 as the first feasible basic index set. Since the objective function is bounded below, the algorithm stops after a finite number of iterations (perhaps we need to use Bland's supplementary rule) in a feasible basic index set α, where the optimality criterion is satisfied. Let $(\overline{x}, \overline{y})$ denote the corresponding basic solution.

There are now two possibilities.

Case 1. *The artificial problem's optimal value is greater than zero.*
In this case, the original problem has no feasible solutions.

Case 2. *The artificial problem's value is equal to zero.*
In this case, $\overline{y} = 0$ and $A\overline{x} = b$.

If $\alpha \subseteq \{1, 2, \dots, n\}$, then α is also a feasible basic index set of the matrix A, and we can start the simplex algorithm on our original problem from α and the corresponding feasible basic solution $\overline{x}$.

If $\alpha \not\subseteq \{1, 2, \dots, n\}$, we set

$$\alpha' = \alpha \cap \{1, 2, \dots, n\}.$$

The matrix columns $\{A_{*k} \mid k \in \alpha'\}$ are now linearly independent, and we can construct an index set $\beta \supseteq \alpha'$, which is maximal with respect to the property that the columns $\{A_{*k} \mid k \in \beta\}$ are linearly independent.

If $\operatorname{rank} A = m$, then β will consist of m elements, and β is then a basic index set of the matrix A. Since $\overline{x}_j = 0$ for all $j \notin \alpha'$, and thus especially for all $j \notin \beta$, it follows that $\overline{x}$ is the basic solution of the system $Ax = b$ that corresponds to the basic index set β. Hence, β is a feasible basic index set for our original problem. We can also note that $\overline{x}$ is a degenerate basic solution.

If $\operatorname{rank} A < m$, then β will consist of just $p = \operatorname{rank} A$ elements, but we can now delete $m - p$ equations from the system $Ax = b$ without changing the set of solutions. This results in a new equivalent LP problem with a coefficient matrix of rank p, and β is a feasible basic index set with $\overline{x}$ as the corresponding basic solution in this problem.

To solve a typical LP problem, one thus normally needs to use the simplex algorithm twice. In Phase 1, we use the simplex algorithm to generate a feasible basic index set α for the original LP problem by solving an artificial LP problem, and in phase 2, the simplex algorithm is used to solve the original problem starting from the basic index set α.

EXAMPLE 13.6.1. We illustrate the technique on the simple problem

$$\begin{array}{ll} \min & x_1 + 2x_2 + x_3 - 2x_4 \\ \text{s.t.} & \left\{ \begin{array}{r} x_1 + x_2 + x_3 - \ x_4 = 2 \\ 2x_1 + x_2 - x_3 + 2x_4 = 3 \\ x_1, x_2, x_3, x_4 \geq 0. \end{array} \right. \end{array}$$

Phase 1 consists in solving the artificial problem

$$\begin{array}{ll} \min & y_5 + y_6 \\ \text{s.t.} & \left\{ \begin{array}{r} x_1 + x_2 + x_3 - \ x_4 + y_5 \qquad = 2 \\ 2x_1 + x_2 - x_3 + 2x_4 \qquad + y_6 = 3 \\ x_1, x_2, x_3, x_4, y_5, y_6 \geq 0. \end{array} \right. \end{array}$$

The computations are shown in tabular form, and the first simplex tableau is the following one.

$$\begin{array}{cccccc|c} 1 & 1 & 1 & -1 & 1 & 0 & 2 \\ 2 & 1 & -1 & 2 & 0 & 1 & 3 \\ \hline 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{array}$$

We begin by eliminating the basic variables from the objective function and then obtain the following sequence of tableaux:

$$\begin{array}{cccccc|c} 1 & 1 & 1 & -1 & 1 & 0 & 2 \\ \underline{2} & 1 & -1 & 2 & 0 & 1 & 3 \\ \hline -3 & -2 & 0 & -1 & \underline{0} & \underline{0} & -5 \end{array}$$

$$\alpha = (5, 6), \quad k = 1, \quad r = 2$$

$$\begin{array}{ccccc|c} 0 & \frac{1}{2} & \underline{\frac{3}{2}} & -2 & 1 & -\frac{1}{2} & \frac{1}{2} \\ 1 & \frac{1}{2} & -\frac{1}{2} & 1 & 0 & \frac{1}{2} & \frac{3}{2} \\ \hline \underline{0} & -\frac{1}{2} & -\frac{3}{2} & 2 & \underline{0} & \frac{3}{2} & -\frac{1}{2} \end{array}$$

$$\alpha = (5, 1), \quad k = 3, \quad r = 1$$

$$\begin{array}{ccccc|c} 0 & \frac{1}{3} & 1 & -\frac{4}{3} & \frac{2}{3} & -\frac{1}{3} & \frac{1}{3} \\ 1 & \frac{2}{3} & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{5}{3} \\ \hline \underline{0} & 0 & \underline{0} & 0 & 1 & 1 & 0 \end{array}$$

$$\alpha = (3, 1)$$

The above final tableau for the artificial problem shows that $\alpha = (3, 1)$ is a feasible basic index set for the original problem with $\overline{x} = (\frac{5}{3}, 0, \frac{1}{3}, 0)$ as corresponding basic solution. We can therefore proceed to phase 2 with the following tableau as our first tableau.

$$\begin{array}{cccc|c} 0 & \frac{1}{3} & 1 & -\frac{4}{3} & \frac{1}{3} \\ 1 & \frac{2}{3} & 0 & \frac{1}{3} & \frac{5}{3} \\ \hline 1 & 2 & 1 & -2 & 0 \end{array}$$

By eliminating the basic variables from the objective function, we obtain the following tableau:

$$\begin{array}{cccc|c} 0 & \frac{1}{3} & 1 & -\frac{4}{3} & \frac{1}{3} \\ 1 & \frac{2}{3} & 0 & \underline{\frac{1}{3}} & \frac{5}{3} \\ \hline \underline{0} & 1 & \underline{0} & -1 & -2 \end{array}$$

$$\alpha = (3, 1), \quad k = 4, \quad r = 2$$

One iteration is enough to obtain a tableau satisfying the optimality criterion.

$$\begin{array}{cccc|c} 4 & 3 & 1 & 0 & 7 \\ 3 & 2 & 0 & 1 & 5 \\ \hline 3 & 3 & \underline{0} & \underline{0} & 3 \end{array}$$

$$\alpha = (3, 4)$$

The optimal value is thus equal to -3, and $\overline{x} = (0, 0, 7, 5)$ is the optimal solution. □

Since the volume of work grows with the number of artificial variables, one should not introduce more artificial variables than necessary. The minimization problem

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

requires no more than one artificial variable. By introducing slack variables $s = (s_{n+1}, s_{n+2}, \ldots, s_{n+m})$, we first obtain an equivalent standard problem

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

If $b \geq 0$, this problem can be solved, as we have already noted, without artificial variables. Let otherwise i_0 be the index of the most negative coordinate of the right-hand side b, and subtract equation no. i_0 in the system $Ax + Es = b$ from all other equations with negative right-hand side, and change finally the sign of equation no. i_0.

The result is a system of equations of the form $A'x + E's = b'$, which is equivalent to the system $Ax+Es = b$ and where $b' \geq 0$ and all the columns of the matrix E', except column no. i_0, are equal to the corresponding columns of the unit matrix E. Phase 1 of the simplex algorithm applied to the problem

$$\begin{array}{ll} \min & \langle c, x\rangle \\ \text{s.t.} & A'x + E's = b',\ x, s \geq 0 \end{array}$$

therefore requires only one artificial variable.

Existence of optimal solutions and the duality theorem

The simplex algorithm is of course first and foremost an efficient algorithm for solving concrete LP problems, but we can also use it to provide alternative proofs of important theoretical results. These are corollaries to the following theorem.

Theorem 13.6.1. *Each standard LP problem with feasible points has a feasible basic index set where one of the two stopping criteria in the simplex algorithm is satisfied.*

Proof. Bland's rule ensures that phase 1 of the simplex algorithm stops with a feasible basic index set from where to start phase 2, and Bland's rule also ensures that this phase stops in a feasible basic index set, where one of the two stopping criteria is satisfied. □

As first corollary we obtain a new proof that every LP problem with finite value has optimal solutions (Theorem 12.1.1).

Corollary 13.6.2. *Each linear minimization problem with feasible solutions and downwards bounded objective function has an optimal solution.*

Proof. Since each LP problem can be replaced by an equivalent LP problem in standard form, it is sufficient to consider such problems. The only way for the simplex algorithm to stop, when the objective function is bounded below, is to stop at a basic solution which satisfies the optimality criterion. So it follows at once from the above theorem that there exists an optimal solution if the objective function is bounded below and the set of feasible solutions is nonempty. □

We can also give an algorithmic proof of the Duality theorem.

Corollary 13.6.3 (Duality theorem)**.** *If the linear optimization problem*

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

has feasible solutions, then it has the same optimal value as the dual maximization problem

$$\begin{aligned} &\max \ \langle b, y\rangle \\ &\text{s.t.} \quad A^T y \leq c. \end{aligned}$$

Proof. Let α be the feasible basic index set where the simplex algorithm stops. If the optimality criterion is satisfied at α, then it follows from Theorem 13.4.2 that the minimization problem and the dual maximization problem have the same finite optimal value. If instead the algorithm stops because the objective function is unbounded below, then the dual problem has no feasible points according to Theorem 13.4.2, and the value of both problems is equal to $-\infty$, by definition. □

By writing general minimization problems in standard form, one can also deduce the general form of the duality theorem from the above special case.

13.7 Sensitivity analysis

In Section 12.1, we studied how the optimal value and the optimal solution depend on the coefficients of the objective function. In this section we shall

study the same issue in connection with the simplex algorithm and also study how the solution to the LP problem

$$\text{(P)} \qquad \begin{aligned} &\min \ \langle c, x\rangle \\ &\text{s.t.} \quad Ax = b,\ x \geq 0 \end{aligned}$$

depends on the right-hand side b. In real LP problems, the coefficients of the objective function and the constraints are often not exactly known, some of them might even be crude estimates, and it is then of course important to know how sensitive the optimal solution is to errors in input data. And even if the input data are accurate, it is of course interesting to know how the optimum solution is affected by changes in one or more of the coefficients.

Let α be a basic index set of the matrix A, and let $\overline{x}(b)$ denote the corresponding basic solution to the system $Ax = b$, i.e.

$$\overline{x}(b)_\alpha = A_{*\alpha}^{-1} b \quad \text{and} \quad \overline{x}(b)_j = 0 \text{ for all } j \notin \alpha.$$

Suppose that the LP problem (P) has an optimal solution for certain given values of b and c, and that this solution has been obtained because the simplex algorithm stopped at the basic index set α. For that to be the case, the basic solution $\overline{x}(b)$ has to be feasible, i.e.

$$\text{(i)} \qquad A_{*\alpha}^{-1} b \geq 0,$$

and the optimality criterion $z \geq 0$ in the simplex algorithm has to be satisfied. Since

$$z = c - A^T \overline{y} \quad \text{and} \quad \overline{y} = (A_{*\alpha}^{-1})^T c_\alpha,$$

we have $z = c - (A_{*\alpha}^{-1} A)^T c_\alpha$, which means that the optimality criterion can be written as

$$\text{(ii)} \qquad z(c) = c - (A_{*\alpha}^{-1} A)^T c_\alpha \geq 0.$$

Conversely, $\overline{x}(b)$ is an optimal solution to the LP problem (P) for all b and c that satisfy the conditions (i) and (ii), because the optimality criterion in the simplex algorithm is then satisfied.

Condition (i) is a system of homogeneous linear inequalities in the variables $b_1, b_2, \ldots, b_m$, and it defines a polyhedral cone B_α in $\mathbf{R}^m$, while (ii) is a system of homogeneous linear inequalities in the variables $c_1, c_2, \ldots, c_n$ and defines a polyhedral cone C_α in $\mathbf{R}^n$. In summary, we have the following result:

$\overline{x}(b)$ is an optimal solution to the LP problem (P) for all $b \in B_\alpha$ and all $c \in C_\alpha$.

Now suppose that we have solved the problem (P) for given values of b and c with $\overline{x} = \overline{x}(b)$ as optimal solution and λ as optimal value. Condition (ii) determines how much we are allowed to change the coefficients of the objective function without changing the optimal solution; $\overline{x}$ is still an optimal solution to the perturbed problem

$$(\mathrm{P}') \qquad \begin{array}{ll} \min & \langle c + \Delta c, x\rangle \\ \text{s.t.} & Ax = b,\ x \geq 0 \end{array}$$

if $z(c + \Delta c) = z(c) + z(\Delta c) \geq 0$, i.e. if

$$\Delta c - (A_{*\alpha}^{-1}A)^T(\Delta c)_\alpha \geq -z(c). \tag{13.12}$$

The optimal value is of course changed to $\lambda + \langle \Delta c, \overline{x}\rangle$.

Inequality (13.12) defines a polyhedron in the variables Δc_1, Δc_2, ..., Δc_n. If for instance $\Delta c_j = 0$ for all j except $j = k$, i.e. if only the c_k-coefficient of the objective function is allowed to change, then inequality (13.12) determines a (possibly unbounded) closed interval $[-d_k, e_k]$ around 0 for Δc_k.

If instead we change the right-hand side of the constraints replacing the vector b by $b+\Delta b$, then $\overline{x}(b+\Delta b)$ becomes an optimal solution to the problem

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

as long as the solution is feasible, i.e. as long as $A_{*\alpha}^{-1}(b + \Delta b) \geq 0$. After simplification, this results in the condition

$$A_{*\alpha}^{-1}(\Delta b) \geq -\overline{x}(b)_\alpha,$$

which is a system of linear inequalities that determines how to choose Δb. If $\Delta b_i = 0$ for all indices except $i = k$, then the set of solutions for Δb_k is an interval around 0 of the form $[-d_k, e_k]$.

The printouts of softwares for the simplex algorithm generally contain information on these intervals.

Example 13.7.1. A person is studying the diet problem

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

in a specific case with six foods and four nutrient requirements. The following computer printout is obtained when $c^T = (1, 2, 3, 4, 1, 6)$ and $b^T = (10, 15, 20, 18)$.

Optimal value: 8.52

Optimal solution:

Food 1: 5.73
Food 2: 0.00
Food 3: 0.93
Food 4: 0.00
Food 5: 0.00
Food 6: 0.00

Sensitivity report

Variable	Value	Objective-coeff.	Allowable decrease	Allowable increase
Food 1:	5.73	1.00	0.14	0.33
Food 2:	0.00	2.00	1.07	∞
Food 3:	0.93	3.00	2.00	0.50
Food 4:	0.00	4.00	3.27	∞
Food 5:	0.00	1.00	0.40	∞
Food 6:	0.00	6.00	5.40	∞

Constraint	Final value	Shadow price	Bounds r.h. side	Allowable decrease	Allowable increase
Nutrient 1:	19.07	0.00	10.00	∞	9.07
Nutrient 2:	31.47	0.00	15.00	∞	16.47
Nutrient 3:	20.00	0.07	20.00	8.00	7.00
Nutrient 4:	18.00	0.40	18.00	4.67	28.67

The sensitivity report shows that the optimal solution remains unchanged as long as the price of food 1 stays in the interval $[5.73 - 0.14, 5.73 + 0.33]$, ceteris paribus. A price change of z units in this range changes the optimal value by $5.73\,z$ units.

A price reduction of food 4 with a maximum of 3.27, or an unlimited price increase of the same food, ceteris paribus, does not affect the optimal solution, nor the optimal value.

The set of price changes that leaves the optimal solution unchanged is a convex set, since it is a polyhedron according to inequality (13.12). The optimal solution of our example is therefore unchanged if for example the prices of foods 1, 2 and 3 are increased by 0.20, 1.20 and 0.10, respectively, because $\Delta c = (0.20, 1.20, 0.10, 0, 0, 0)$ is a convex combination of allowable increases, since

$$\frac{0.20}{0.33} + \frac{1.20}{\infty} + \frac{0.10}{0.50} \leq 1.$$

The sensitivity report also shows how the optimal solution is affected by certain changes in the right-hand side b. The optimal solution remains unchanged, for example, if the need for nutrient 1 would increase from 10 to 15, since the constraint is not binding and the increase 5 is less than the permitted increase 9.07.

The sensitivity report also tells us that the new optimal solution will still be derived from the same basic index set as above, if b_4 is increased by say 20 units from 18 to 38, an increase that is within the scope of the permissible. So in this case, the optimal diet will also only consist of foods 1 and 3, but the optimal value will increase by $20 \cdot 0.40$ to 16.52 since the shadow price of nutrient 4 is equal to 0.40. □

13.8 The dual simplex algorithm

The simplex algorithm, applied to a problem

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

with a bounded optimal value, starts from a given feasible basic index set α^0 and then generates a finite sequence $(\alpha^k, \overline{x}^k, \overline{y}^k)_{k=0}^p$ of basic index sets α^k, corresponding basic solutions $\overline{x}^k$ and vectors $\overline{y}^k$ with the following properties:

(i) The basic solutions $\overline{x}^k$ are extreme points of the polyhedron

$$X = \{x \in \mathbf{R}^n \mid Ax = b,\ x \geq 0\}$$

of feasible solutions.

(ii) The line segments $[\overline{x}^k, \overline{x}^{k+1}]$ are edges of the polyhedron X.

(iii) The objective function values $(\langle c, \overline{x}^k\rangle)_{k=0}^p$ form a decreasing sequence.

(iv) $\langle b, \overline{y}^k\rangle = \langle c, \overline{x}^k\rangle$ for all k.

(v) The algorithm stops after p iterations when the optimality criterion is met, and $\overline{y}^p$ is then an extreme point of the polyhedron

$$Y = \{y \in \mathbf{R}^m \mid A^T y \leq c\}.$$

(vi) $\overline{x}^p$ is an optimal solution, and $\overline{y}^p$ is an optimal solution to the dual problem

$$\begin{aligned} &\max \ \langle b, y\rangle \\ &\text{s.t.} \quad A^T y \leq c. \end{aligned}$$

(vii) The vectors $\overline{y}^k$ do not, however, belong to Y for $0 \leq k \leq p-1$.

The optimal solution $\overline{x}^p$ is obtained by moving along edges of the polyhedron X until an extreme point has been reached that also corresponds to

an extreme point of the polyhedron Y. Instead, we could move along edges of the polyhedron Y, and this observation leads to the following method for solving the minimization problem.

The dual simplex algorithm

Given a basic index set α such that $z = c - A^T\overline{y} \geq 0$, where $\overline{y} = (A_{*\alpha}^{-1})^T c_\alpha$.

Repeat steps 1–4 until a stop occurs.

1. Compute the basic solution $\overline{x}$ corresponding to α.
2. *Stopping criterion:* **quit** if $\overline{x} \geq 0$.
   Optimal solution: $\overline{x}$. Optimal dual solution: $\overline{y}$.
   Also **quit** if any of the constraint equations has the form $a'_{i1}x_1 + a'_{i2}x_2 + \cdots + a'_{in}x_n = b'_i$ with $b'_i > 0$ and $a'_{ij} \leq 0$ for all j, because then there are no feasible solutions to the primal problem.
3. Generate a new basic index set α' by replacing one of the indices of α in such a way that the new reduced cost vector z' remains nonnegative and $\langle b, \overline{y}' \rangle \geq \langle b, \overline{y} \rangle$, where $\overline{y}' = (A_{*\alpha'}^{-1})^T c_{\alpha'}$.
4. *Update:* $\alpha := \alpha'$, $\overline{y} := \overline{y}'$.

We refrain from specifying the necessary pivoting rules. Instead, we consider a simple example.

EXAMPLE 13.8.1. We shall solve the minimization problem

$$\begin{array}{ll} \min & x_1 + 2x_2 + 3x_3 \\ \text{s.t.} & \left\{ \begin{array}{rrrl} 2x_1 & + & x_3 & \geq 9 \\ x_1 + 2x_2 & & & \geq 12 \\ & x_2 + 2x_3 & & \geq 15, \ x \geq 0 \end{array} \right. \end{array}$$

by using the dual simplex algorithm, and we begin by reformulating the problem in standard form as follows:

$$\begin{array}{ll} \min & x_1 + 2x_2 + 3x_3 \\ \text{s.t.} & \left\{ \begin{array}{rrrrrl} 2x_1 & + & x_3 - x_4 & & = & 9 \\ x_1 + 2x_2 & & & - x_5 & = & 12 \\ & x_2 + 2x_3 & & - x_6 & = & 15, \ x \geq 0. \end{array} \right. \end{array}$$

The corresponding simplex tableau now looks like this:

$$\begin{array}{cccccc|c} 2 & 0 & 1 & -1 & 0 & 0 & 9 \\ 1 & 2 & 0 & 0 & -1 & 0 & 12 \\ 0 & 1 & 2 & 0 & 0 & -1 & 15 \\ \hline 1 & 2 & 3 & \underline{0} & \underline{0} & \underline{0} & 0 \end{array}$$

For comparison, we also state the corresponding dual maximization problem:

$$\begin{aligned} &\max\ 9y_1 + 12y_2 + 15y_3 \\ &\text{s.t.} \begin{cases} 2y_1 + y_2 \leq 1 \\ 2y_2 + y_3 \leq 2 \\ y_1 + 2y_3 \leq 3, \ y \geq 0. \end{cases} \end{aligned}$$

We can start the dual simplex algorithm from the basic index set $\alpha = (4, 5, 6)$, and as usual, we have underlined the basic columns. The corresponding basic solution $\overline{x}$ is not feasible since the coordinates of $\overline{x}_\alpha = (-9, -12, -15)$ are negative. The bottom row $[\,1\ 2\ 3\ 0\ 0\ 0\,]$ of the tableau is the reduced cost vector $z^T = c^T - \overline{y}^T A$. The row vector $\overline{y}^T = c_\alpha^T A_{*\alpha}^{-1} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ can also be read in the bottom row; it is found below the matrix $-E$, and $\overline{y}$ belongs to the polyhedron Y of feasible solutions to the dual problem, since $z^T \geq 0$.

We will now gradually replace one element at a time in the basic index set. As pivot row r, we choose the row that corresponds to the most negative coordinate of $\overline{x}_\alpha$, and in the first iteration, this is the third row in the above simplex tableau. To keep the reduced cost vector nonnegative, we must select as pivot column a column k, where the matrix element a_{rk} is positive and the ratio z_k/a_{rk} is as small as possible. In the above tableau, this is the third column, so we pivot around the element at location $(3, 3)$. This leads to the following tableau:

$$\begin{array}{cccccc|c} 2 & -\frac{1}{2} & 0 & -1 & 0 & \frac{1}{2} & \frac{3}{2} \\ 1 & 2 & 0 & 0 & -1 & 0 & 12 \\ 0 & \frac{1}{2} & 1 & 0 & 0 & -\frac{1}{2} & \frac{15}{2} \\ \hline 1 & \frac{1}{2} & \underline{0} & \underline{0} & \underline{0} & \frac{3}{2} & -\frac{45}{2} \end{array}$$

In this new tableau, $\alpha = (4, 5, 3)$, $\overline{x}_\alpha = (-\frac{3}{2}, -12, \frac{15}{2})$ and $\overline{y} = (0, 0, \frac{3}{2})$. The most negative element of $\overline{x}_\alpha$ is to be found in the second row, and the least ratio z_k/a'_{2k} with a positive denominator a'_{2k} is obtained for $k = 2$. Pivoting around the element at location $(2, 2)$ leads to the following simplex tableau:

$$\begin{array}{cccccc|c} \frac{9}{4} & 0 & 0 & -1 & -\frac{1}{4} & \frac{1}{2} & \frac{9}{2} \\ \frac{1}{2} & 1 & 0 & 0 & -\frac{1}{2} & 0 & 6 \\ -\frac{1}{4} & 0 & 1 & 0 & \frac{1}{4} & -\frac{1}{2} & \frac{9}{2} \\ \hline \frac{3}{4} & \underline{0} & \underline{0} & \underline{0} & \frac{1}{4} & \frac{3}{2} & -\frac{51}{2} \end{array}$$

Now, $\alpha = (4, 2, 3)$, $\overline{x}_\alpha = (-\frac{9}{2}, 6, \frac{9}{2})$ and $\overline{y} = (0, \frac{1}{4}, \frac{3}{2})$. This time, we should select the element in the first row and the first column as pivot element, which leads to the next tableau.

$$\begin{array}{cccccc|c} 1 & 0 & 0 & -\frac{4}{9} & -\frac{1}{9} & \frac{2}{9} & 2 \\ 0 & 1 & 0 & \frac{2}{9} & -\frac{4}{9} & -\frac{1}{9} & 5 \\ 0 & 0 & 1 & -\frac{1}{9} & \frac{2}{9} & -\frac{4}{9} & 5 \\ \hline \underline{0} & \underline{0} & \underline{0} & \frac{1}{3} & \frac{1}{3} & \frac{4}{3} & -27 \end{array}$$

Here, $\alpha = (1, 2, 3)$, $\overline{x}_\alpha = (2, 5, 5)$ and $\overline{y} = (\frac{1}{3}, \frac{1}{3}, \frac{4}{3})$, and the optimality criterion is met since $\overline{x}_\alpha \geq 0$. The optimal value is 27 and $(2, 5, 5, 0, 0, 0)$ is the optimal point. The dual maximization problem attains its maximum at $(\frac{1}{3}, \frac{1}{3}, \frac{4}{3})$. The optimal solution to our original minimization problem is of course $x = (2, 5, 5)$. □

13.9 Complexity

How many iterations are needed to solve an LP problem using the simplex algorithm? The answer will depend, of course, on the size of the problem. Experience shows that the number of iterations largely grows linearly with the number of rows m and sublinearly with the number of columns n for realistic problems, and in most real problems, n is a small multiple of m, usually not more than $10m$. The number of iterations is therefore usually somewhere between m and $4m$, which means that the simplex algorithm generally performs very well.

The worst case behavior of the algorithm is bad, however (for all known pivoting rules). Klee and Minty has constructed an example where the number of iterations grows exponentially with the size of the problem.

Example 13.9.1 (Klee and Minty, 1972). Consider the following LP problem in n variables and with n inequality constraints:

$$\begin{aligned} \max\ & 2^{n-1}x_1 + 2^{n-2}x_2 + \cdots + 2x_{n-1} + x_n \\ \text{s.t.}\ & \begin{cases} x_1 & & & \leq & 5 \\ 4x_1 + & x_2 & & \leq & 25 \\ 8x_1 + & 4x_2 + & x_3 & \leq & 125 \\ \vdots & & & & \vdots \\ 2^n x_1 + 2^{n-1}x_2 + \ldots + 4x_{n-1} + x_n & \leq & 5^n \end{cases} \end{aligned}$$

The polyhedron of feasible solutions has in this case 2^n extreme points.

Suppose that we apply the simplex algorithm to the equivalent standard problem, in each iteration choosing as pivot column the column with the most negative value of the reduced cost. If we start from the feasible basic solution that corresponds to $x = 0$, then we have to go through all the 2^n feasible basic solutions before we finally reach the optimal solution $(0, 0, \ldots, 5^n)$. The number of iterations is therefore equal to 2^n and thus increases exponentially with n. □

An algorithm for solving a problem in n variables is called *strictly polynomial* if there exists a positive integer k such that the number of elementary arithmetic operations in the algorithm grows with n as at most $O(n^k)$. In many algorithms, the number of operations also depends on the size of the input data. An algorithm is called *polynomial* if the number of arithmetic operations is growing as a polynomial in L, where L is the number of binary bits needed to represent all input (i.e. the matrices A, b and c in linear programming).

Gaussian elimination is a strictly polynomial algorithm, because a system of linear equations with n equations and n unknowns is solved with $O(n^3)$ arithmetic operations.

Klee–Minty's example and other similar examples demonstrate that the simplex algorithm is not strictly polynomial. But all experience shows that the simplex algorithm works very well, even if the worst case behavior is bad. This is also supported by probabilistic analyzes, made by Borgwardt (1987), Smale (1983), Adler and Megiddo (1985), among others. Such an analysis shows, for example, that (a variant of) the simplex algorithm, given a certain special probability distribution of the input data, on average converges after $O(m^2)$ iterations, where m is the number of constraints.

The existence of a polynomial algorithm that solves LP problems (with rational coefficients as input data) was first demonstrated in 1979 by Leonid Khachiyan. His so-called *ellipsoid algorithm* reduces LP problems to the problem of finding a solution to a system $Ax > b$ of strict inequalities with a bounded set of solutions, and the algorithm generates a sequence of shrinking ellipsoids, all guaranteed to contain all the solutions to the system. If the center of an ellipsoid satisfies all inequalities of the system, then a solution has been found, of course. Otherwise, the process stops when a generated ellipsoid has too small volume to contain all solutions, if there are any, with the conclusion that there are no solutions.

LP problems in standard form with n variables and input size L are solved by the ellipsoid method in $O(n^4 L)$ arithmetic operations. However, in spite of this nice theoretical result, it was soon clear that the ellipsoid method could not compete with the simplex algorithm on real problems of moderate

size due to slow convergence. (The reason for this is, of course, that the implicit constant in the O-estimate is very large.)

A new polynomial algorithm was discovered in 1984 by Narendra Karmarkar. His algorithm generates a sequence of points, which lie in the interior of the set of feasible points and converge towards an optimal point. The algorithm uses repeated centering of the generated points by a projective scaling transformation. The theoretical complexity bound of the original version of the algorithm is also $O(n^4L)$.

Karmarkar's algorithm turned out to be competitive with the simplex algorithm on practical problems, and his discovery was the starting point for an intensive development of alternative interior point methods for LP problems and more general convex problems. We will study such an algorithm in Chapter 18.

It is still an open problem whether there exists any strictly polynomial algorithm for solving LP problems.

Exercises

13.1 Write the following problems in standard form.

a) $\min \ 2x_1 - 2x_2 + x_3$ s.t. $\begin{cases} x_1 + x_2 - x_3 \geq 3 \\ x_1 + x_2 - x_3 \leq 2 \\ x_1, x_2, x_3 \geq 0 \end{cases}$

b) $\min \ x_1 + 2x_2$ s.t. $\begin{cases} x_1 + x_2 \geq 1 \\ x_2 \geq -2 \\ x_1 \geq 0. \end{cases}$

13.2 Find all nonnegative basic solutions to the following systems of equations.

a) $\begin{cases} 5x_1 + 3x_2 + x_3 = 40 \\ x_1 + x_2 + x_3 = 10 \end{cases}$

b) $\begin{cases} x_1 - 2x_2 - x_3 + x_4 = 3 \\ 2x_1 + 5x_2 - 3x_3 + 2x_4 = 6. \end{cases}$

13.3 State the dual problem to

$$\min \ x_1 + x_2 + 4x_3 \quad \text{s.t.} \ \begin{cases} x_1 - x_3 = 1 \\ x_1 + 2x_2 + 7x_3 = 7, \ x \geq 0 \end{cases}$$

and prove that $(1, 3, 0)$ is an optimal solution and that $(\frac{1}{2}, \frac{1}{2})$ is an optimal solution to the dual problem.

13.4 Solve the following LP problems using the simplex algorithm.

a) $\min \ -x_4$ s.t. $\begin{cases} x_1 + x_4 = 1 \\ x_2 + 2x_4 = 2 \\ x_3 - x_4 = 3, \ x \geq 0 \end{cases}$

b) $\max\ 2x_1 - x_2 + x_3 - 3x_4 + x_5$
$$\text{s.t.} \begin{cases} x_1 \quad + 2x_4 - x_5 = 15 \\ x_2 + x_4 + x_5 = 12 \\ x_3 - 2x_4 + x_5 = 9,\ x \geq 0 \end{cases}$$

c) $\max\ 15x_1 + 12x_2 + 14x_3$
$$\text{s.t.} \begin{cases} 3x_1 + 2x_2 + 5x_3 \leq 6 \\ x_1 + 3x_2 + 3x_3 \leq 3 \\ 5x_3 \leq 4,\ x \geq 0 \end{cases}$$

d) $\max\ 2x_1 + x_2 + 3x_3 + x_4 + 2x_5$
$$\text{s.t.} \begin{cases} x_1 + 2x_2 + x_3 \quad + x_5 \leq 10 \\ x_2 + x_3 + x_4 + x_5 \leq 8 \\ x_1 \quad + x_3 + x_4 \quad \leq 5,\ x \geq 0 \end{cases}$$

e) $\min\ x_1 - 2x_2 + x_3$
$$\text{s.t.} \begin{cases} x_1 + x_2 - 2x_3 \leq 3 \\ x_1 - x_2 + x_3 \leq 2 \\ -x_1 - x_2 + x_3 \leq 0,\ x \geq 0 \end{cases}$$

f) $\min\ x_1 - x_2 + 2x_3 - 3x_4$
$$\text{s.t.} \begin{cases} 2x_1 + 3x_2 + x_3 \quad = 2 \\ x_1 + 3x_2 + x_3 + 5x_4 = 4,\ x \geq 0. \end{cases}$$

13.5 Carry out in detail all the steps of the simplex algorithm for the problem

$$\min\ -x_2 + x_4$$
$$\text{s.t.} \begin{cases} x_1 \quad + x_4 + x_5 = 1 \\ x_2 \quad - 2x_4 - x_5 = 1 \\ x_3 + 2x_4 + x_5 = 3,\ x \geq 0. \end{cases}$$

Is the optimal solution unique?

13.6 Use artificial variables to solve the LP problem

$$\max\ x_1 + 2x_2 + 3x_3 - x_4$$
$$\text{s.t.} \begin{cases} x_1 + 2x_2 + 3x_3 \quad = 15 \\ 2x_1 + x_2 + 5x_3 \quad = 20 \\ x_1 + 2x_2 + x_3 + x_4 = 10,\ x \geq 0. \end{cases}$$

13.7 Use the simplex algorithm to show that the following systems of equalities and inequalities are consistent.

a) $\begin{cases} 3x_1 + x_2 + 2x_3 + x_4 + x_5 = 2 \\ 2x_1 - x_2 + x_3 + x_4 + 4x_5 = 3,\ x \geq 0 \end{cases}$

b) $\begin{cases} x_1 - x_2 + 2x_3 + x_4 \geq 6 \\ -2x_1 + x_2 - 2x_3 + 7x_4 \geq 1 \\ x_1 - x_2 + x_3 - 3x_4 \geq -1,\ x \geq 0. \end{cases}$

13.8 Solve the LP problem

$$\begin{array}{ll}\min & x_1 + 2x_2 + 3x_3 \\ \text{s.t.} & \begin{cases} 2x_1 \qquad + \ x_3 \geq 3 \\ x_1 + 2x_2 \qquad \geq 4 \\ \qquad x_2 + 2x_3 \geq 5, \ x \geq 0. \end{cases}\end{array}$$

13.9 Write the following problem in standard form and solve it using the simplex algorithm.

$$\begin{array}{ll}\min & 8x_1 - x_2 \\ \text{s.t.} & \begin{cases} 3x_1 + \ x_2 \geq \ 1 \\ x_1 - \ x_2 \leq \ 2 \\ x_1 + 2x_2 = 20, \ x \geq 0. \end{cases}\end{array}$$

13.10 Solve the following LP problems using the dual simplex algorithm.

a) $$\begin{array}{ll}\min & 2x_1 + x_2 + 3x_3 \\ \text{s.t.} & \begin{cases} x_1 + x_2 + \ x_3 \geq 2 \\ 2x_1 - x_2 \qquad \geq 1 \\ \qquad x_2 + 2x_3 \geq 2, \ x \geq 0 \end{cases}\end{array}$$

b) $$\begin{array}{ll}\min & x_1 + 2x_2 \\ \text{s.t.} & \begin{cases} \ x_1 \qquad - 2x_3 \geq -5 \\ -2x_1 + 3x_2 - \ x_3 \geq -4 \\ -2x_1 + 5x_2 - \ x_3 \geq \ 2, \ x \geq 0 \end{cases}\end{array}$$

c) $$\begin{array}{ll}\min & 3x_1 + 2x_2 + 4x_3 \\ \text{s.t.} & \begin{cases} 4x_1 \qquad + 2x_3 \geq 5 \\ x_1 + 3x_2 + 2x_3 \geq 4, \ x \geq 0. \end{cases}\end{array}$$

13.11 Suppose $b_2 \geq b_1 \geq 0$. Show that $\overline{x} = \big(b_1, \frac{1}{2}(b_2 - b_1), 0\big)$ is an optimal solution to the problem

$$\begin{array}{ll}\min & x_1 + x_2 + 4x_3 \\ \text{s.t.} & \begin{cases} x_1 \qquad - \ x_3 = b_1 \\ x_1 + 2x_2 + 7x_3 = b_2, \ x \geq 0. \end{cases}\end{array}$$

13.12 Investigate how the optimal solution to the LP problem

$$\begin{array}{ll}\max & 2x_1 + tx_2 \\ \text{s.t.} & \begin{cases} x_1 + x_2 \leq 5 \\ 2x_1 + x_2 \leq 7, \ x \geq 0 \end{cases}\end{array}$$

varies as the real parameter t varies.

13.13 A shoe manufacturer produces two shoe models A and B. Due to limited supply of leather, the manufactured number of pairs x_A and x_B of the two models must satisfy the inequalities

$$x_A \leq 1000, \quad 4x_A + 3x_B \leq 4100, \quad 3x_A + 5x_B \leq 5000.$$

The sale price of A and B is 500 SEK and 350 SEK, respectively per pair. It costs 200 SEK to manufacture a pair of shoes of model B. However, the cost of producing a pair of shoes of model A is uncertain due to malfunctioning machines, and it can only be estimated to be between 300 SEK and 410 SEK. Show that the manufacturer may nevertheless decide how many pairs of shoes he shall manufacture of each model to maximize his profit.

13.14 Joe wants to meet his daily requirements of vitamins P, Q and R by only living on milk and bread. His daily requirement of vitamins is 6, 12 and 4 mg, respectively. A liter of milk costs 7.50 SEK and contains 2 mg of P, 2 mg of Q and nothing of R; a loaf of bread costs 20 SEK and contains 1 mg of P, 4 mg of Q and 4 mg of R. The vitamins are not toxic, so a possible overdose does not harm. Joe wants to get away as cheaply as possible. Which daily bill of fare should he choose? Suppose that the price of milk begins to rise. How high can it be without Joe having to change his bill of fare?

13.15 Using the assumptions of Lemma 13.4.1, show that the reduced cost z_k is equal to the direction derivative of the objective function $\langle c, x\rangle$ in the direction $-v$.

13.16 This exercise outlines an alternative method to prevent cycling in the simplex algorithm. Consider the problem

$$\text{(P)} \qquad \begin{aligned} &\min \quad \langle c, x\rangle \\ &\text{s.t.} \quad Ax = b,\ x \geq 0 \end{aligned}$$

and let α be an arbitrary feasible basic index set with corresponding basic solution $\overline{x}$. For each positive number ϵ, we define new vectors $\overline{x}(\epsilon) \in \mathbf{R}^n$ and $b(\epsilon) \in \mathbf{R}^m$ as follows:

$$\begin{aligned} \overline{x}(\epsilon)_\alpha &= \overline{x}_\alpha + (\epsilon, \epsilon^2, \ldots, \epsilon^m) \quad \text{and} \quad \overline{x}(\epsilon)_j = 0 \text{ for all } j \notin \alpha, \\ b(\epsilon) &= A\overline{x}(\epsilon). \end{aligned}$$

Then $\overline{x}(\epsilon)$ is obviously a nonnegative basic solution to the system $Ax = b(\epsilon)$ with α as the corresponding basic index set, and the coordinates of the vector $b(\epsilon)$ are polynomials of degree m in the variable ϵ.

a) Prove that all basic solutions to the system $Ax = b(\epsilon)$ are non-degenerate except for finitely many numbers $\epsilon > 0$. Consequently, there is a number $\epsilon_0 > 0$ so that all basic solution are non-degenerate if $0 < \epsilon < \epsilon_0$.

b) Prove that if $0 < \epsilon < \epsilon_0$, then all feasible basic index sets for the problem

$$(\mathrm{P}_\epsilon) \qquad \begin{array}{ll} \min & \langle c, x \rangle \\ \text{s.t.} & Ax = b(\epsilon),\ x \geq 0 \end{array}$$

are also feasible basic index sets for the original problem (P).

c) The simplex algorithm applied to problem (P_ϵ) will therefore stop at a feasible basic index set β, which is also feasible for problem (P), provided ϵ is a sufficiently small number. Prove that β also satisfies the stopping condition for problem (P).

Cycling can thus be avoided by the following method: Perturb the right-hand side by forming $\overline{x}(\epsilon)$ and the column matrix $b(\epsilon)$, where ϵ is a small positive number. Use the simplex algorithm on the perturbed problem. The algorithm stops at a basic index set β. The corresponding unperturbed problem stops at the same basic index set.

13.17 Suppose that $\mathcal{A}$ is a polynomial algorithm for solving systems $Cx \geq b$ of linear inequalities. When applied to a solvable system, the algorithm finds a solution $\overline{x}$ and stops with the output $\mathcal{A}(C, b) = \overline{x}$. For unsolvable systems, it stops with the output $\mathcal{A}(C, b) = \emptyset$. Use the algorithm $\mathcal{A}$ to construct a polynomial algorithm for solving arbitrary LP problems

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

13.18 Perform all the steps of the simplex algorithm for the example of Klee and Minty when $n = 3$.

Part IV

Interior-point methods

Chapter 14

Descent methods

The most common numerical algorithms for minimization of differentiable functions of several variables are so-called *descent algorithms*. A descent algorithm is an iterative algorithm that from a given starting point generates a sequence of points with decreasing function values, and the process is stopped when one has obtained a function value that approximates the minimum value good enough according to some criterion. However, there is no algorithm that works for arbitrary functions; special assumptions about the function to be minimized are needed to ensure convergence towards the minimum point. Convexity is such an assumption, which makes it also possible in many cases to determine the speed of convergence.

This chapter describes descent methods in general terms, and we exemplify with the simplest descent method, the gradient descent method.

14.1 General principles

We shall study the optimization problem

$$\text{(P)} \qquad \min f(x)$$

where f is a function which is defined and differentiable on an open subset Ω of $\mathbf{R}^n$. We assume that the problem has a solution, i.e. that there is an optimal point $\hat{x} \in \Omega$, and we denote the optimal value $f(\hat{x})$ as $f_{\min}$. A convenient assumption which, according to Corollary 8.1.7, guarantees the existence of a (unique) optimal solution is that f is strongly convex and has some closed nonempty sublevel set.

Our aim is to generate a sequence $x_1, x_2, x_3, \ldots$ of points in Ω from a given *starting point* $x_0 \in \Omega$, with decreasing function values and with the property that $f(x_k) \to f_{\min}$ as $k \to \infty$. In the iteration leading from the

point x_k to the next point x_{k+1}, except when x_k is already optimal, one first selects a vector v_k such that the one-variable function $\phi_k(t) = f(x_k + tv_k)$ is strictly decreasing at $t = 0$. Then, a *line search* is performed along the half-line $x_k + tv_k$, $t > 0$, and a point $x_{k+1} = x_k + h_k v_k$ satisfying $f(x_{k+1}) < f(x_k)$ is selected according to specific rules.

The vector v_k is called the *search direction*, and the positive number h_k is called the *step size*. The algorithm is terminated when the difference $f(x_k) - f_{\min}$ is less than a given tolerance.

Schematically, we can describe a typical descent algorithm as follows:

Descent algorithm

Given a starting point $x \in \Omega$.
Repeat

1. Determine (if $f'(x) \neq 0$) a search direction v and a step size $h > 0$ such that $f(x + hv) < f(x)$.
2. *Update:* $x := x + hv$.

until stopping criterion is satisfied.

Different strategies for selecting the search direction, different ways to perform the line search, as well as different stop criteria, give rise to different algorithms, of course.

Search direction

Permitted search directions in iteration k are vectors v_k which satisfy the inequality

$$\langle f'(x_k), v_k \rangle < 0,$$

because this ensures that the function $\phi_k(t) = f(x_k + tv_k)$ is decreasing at the point $t = 0$, since $\phi_k'(0) = \langle f'(x_k), v_k \rangle$. We will study two ways to select the search direction.

The *gradient descent method* selects $v_k = -f'(x_k)$, which is a permissible choice since $\langle f'(x_k), v_k \rangle = -\|f'(x_k)\|^2 < 0$. Locally, this choice gives the fastest decrease in function value.

Newton's method assumes that the second derivative exists, and the search direction at points x_k where the second derivative is positive definite is

$$v_k = -f''(x_k)^{-1} f'(x_k).$$

This choice is permissible since $\langle f'(x_k), v_k \rangle = -\langle f'(x_k), f''(x_k)^{-1} f'(x_k) \rangle < 0$.

Line search

Given the search direction v_k there are several possible strategies for selecting the step size h_k.

1. *Exact line search.* The step size h_k is determined by minimizing the one-variable function $t \mapsto f(x_k + tv_k)$. This method is used for theoretical studies of algorithms but almost never in practice due to the computational cost of performing the one-dimensional minimization.

2. The step size sequence $(h_k)_{k=1}^{\infty}$ is given *a priori*, for example as $h_k = h$ or as $h_k = h/\sqrt{k+1}$ for some positive constant h. This is a simple rule that is often used in convex optimization.

3. The step size h_k at the point x_k is defined as $h_k = \rho(x_k)$ for some given function ρ. This technique is used in the analysis of Newton's method for self-concordant functions.

4. *Armijo's rule.* The step size h_k at the point x_k depends on two parameters α, $\beta \in]0, 1[$ and is defined as

$$h_k = \beta^m,$$

where m is the smallest nonnegative integer such that the point $x_k + \beta^m v_k$ lies in the domain of f and satisfies the inequality

$$f(x_k + \beta^m v_k) \leq f(x_k) + \alpha\beta^m \langle f'(x_k), v_k \rangle. \tag{14.1}$$

Such an m certainly exists, since $\beta^n \to 0$ as $n \to \infty$ and

$$\lim_{t \to 0} \frac{f(x_k + tv_k) - f(x_k)}{t} = \langle f'(x_k), v_k \rangle < \alpha \langle f'(x_k), v_k \rangle.$$

The number m is determined by simple backtracking: Start with $m = 0$ and examine whether $x_k + \beta^m v_k$ belongs to the domain of f and inequality (14.1) holds. If not, increase m by 1 and repeat until the conditions are fulfilled. Figure 14.1 illustrates the process.

The decrease in iteration k of function value per step size, i.e. the ratio $(f(x_k) - f(x_{k+1}))/h_k$, is for convex functions less than or equal to $-\langle f'(x_k), v_k \rangle$ for any choice of step size h_k. With step size h_k selected according to Armijo's rule the same ratio is also $\geq -\alpha \langle f'(x_k), v_k \rangle$. With Armijo's rule, the decrease per step size is, in other words, at least α of what the maximum might be. Typical values of α in practical applications lie in the range between 0.01 and 0.3.

The parameter β determines how many backtracking steps are needed. The larger β, the more backtracking steps, i.e. the finer the line search. The parameter β is often chosen between 0.1 and 0.8.

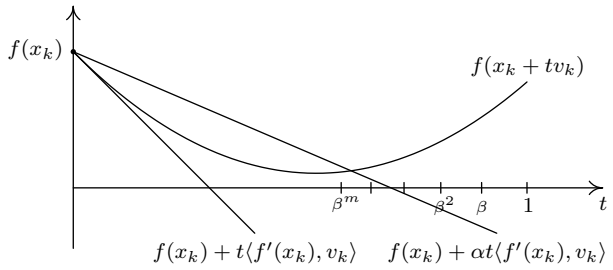

Figure 14.1. Armijo's rule: The step size is $h_k = \beta^m$, where m is the smallest nonnegative integer such that $f(x_k + \beta^m v_k) \leq f(x_k) + \alpha\beta^m \langle f'(x_k), v_k\rangle$.

Armijo's rule exists in different versions and is used in several practical algorithms.

Stopping criteria

Since the optimum value is generally not known beforehand, it is not possible to formulate the stopping criterion directly in terms of the minimum. Intuitively, it seems reasonable that x should be close to the minimum point if the derivative $f'(x)$ is comparatively small, and the next theorem shows that this is indeed the case, under appropriate conditions on the objective function.

Theorem 14.1.1. *Suppose that the function $f\colon \Omega \to \mathbf{R}$ is differentiable, μ-strongly convex and has a minimum at $\hat{x} \in \Omega$. Then, for all $x \in \Omega$*

(i) $$f(x) - f(\hat{x}) \leq \frac{1}{2\mu}\|f'(x)\|^2 \qquad \textit{and}$$

(ii) $$\|x - \hat{x}\| \leq \frac{1}{\mu}\|f'(x)\|.$$

Proof. Due to the convexity assumption,

(14.2) $$f(y) \geq f(x) + \langle f'(x), y - x\rangle + \tfrac{1}{2}\mu\|y - x\|^2$$

for all $x, y \in \Omega$. The right-hand side of inequality (14.2) is a convex quadratic function in the variable y, which is minimized by $y = x - \mu^{-1}f'(x)$, and the minimum is equal to $f(x) - \frac{1}{2}\mu^{-1}\|f'(x)\|^2$. Hence,

$$f(y) \geq f(x) - \tfrac{1}{2}\mu^{-1}\|f'(x)\|^2$$

for all $y \in \Omega$, and we obtain the inequality (i) by choosing y as the minimum point $\hat{x}$.

Now, replace y with x and x with $\hat{x}$ in inequality (14.2). Since $f'(\hat{x}) = 0$, the resulting inequality becomes

$$f(x) \geq f(\hat{x}) + \tfrac{1}{2}\mu\|x - \hat{x}\|^2,$$

which combined with inequality (i) gives us inequality (ii). □

We now return to the descent algorithm and our discussion of the the stopping criterion. Let

$$S = \{x \in \Omega \mid f(x) \leq f(x_0)\},$$

where x_0 is the selected starting point, and assume that the sublevel set S is convex and that the objective function f is μ-strongly convex on S. All the points $x_1, x_2, x_3, \ldots$ that are generated by the descent algorithm will of course lie in S since the function values are decreasing. Therefore, it follows from Theorem 14.1.1 that $f(x_k) < f_{\min} + \epsilon$ if $\|f'(x_k)\| < (2\mu\epsilon)^{1/2}$.

As a stopping criterion, we can thus use the condition

$$\|f'(x_k)\| \leq \eta,$$

which guarantees that $f(x_k) - f_{\min} \leq \eta^2/2\mu$ and that $\|x_k - \hat{x}\| \leq \eta/\mu$. A problem here is that the convexity constant μ is known only in rare cases. So the stopping condition $\|f'(x_k)\| \leq \eta$ can in general not be used to give precise bounds on $f(x_k) - f_{\min}$. But Theorem 14.1.1 verifies our intuitive feeling that the difference between $f(x)$ and $f_{\min}$ is small if the gradient of f at x is small enough.

Convergence rate

Let us say that a convergent sequence $x_0, x_1, x_2, \ldots$ of points with limit $\hat{x}$ converges *at least linearly* if there is a constant $c < 1$ such that

$$\|x_{k+1} - \hat{x}\| \leq c\|x_k - \hat{x}\| \tag{14.3}$$

for all k, and that the convergence is *at least quadratic* if there is a constant C such that

$$\|x_{k+1} - \hat{x}\| \leq C\|x_k - \hat{x}\|^2 \tag{14.4}$$

for all k. We also say that the convergence is *no better than linear* and *no better than quadratic* if

$$\varliminf_{k\to\infty} \frac{\|x_{k+1} - \hat{x}\|}{\|x_k - \hat{x}\|^\alpha} > 0$$

for $\alpha = 1$ and $\alpha = 2$, respectively.

Note that inequality (14.3) implies that the sequence $(x_k)_0^\infty$ converges to $\hat{x}$, because it follows by induction that

$$\|x_k - \hat{x}\| \leq c^k \|x_0 - \hat{x}\|$$

for all k.

Similarly, inequality (14.4) implies that the sequence $(x_k)_0^\infty$ convergences to $\hat{x}$ if the starting point x_0 satisfies the condition $\|x_0 - \hat{x}\| < C^{-1}$, because we now have

$$\|x_k - \hat{x}\| \leq C^{-1}\big(C\|x_0 - \hat{x}\|\big)^{2^k}$$

for all k.

If an iterative method, when applied to functions in a given class of functions, always generates sequences that are at least linearly (quadratic) convergent and there is a sequence which does not converge better than linearly (quadratic), then we say that the method is *linearly (quadratic) convergent* for the function class in question.

14.2 The gradient descent method

In this section we analyze the gradient descent algorithm with constant step size. The iterative formulation of the variant of the algorithm that we have in mind looks like this:

Gradient descent algorithm with constant step size

Given a starting point x and a step size h.
Repeat
1. Compute the search direction $v = -f'(x)$.
2. *Update:* $x := x + hv$.

until stopping criterion is satisfied.

The algorithm converges linearly to the minimum point for strongly convex functions with Lipschitz continuous derivatives provided that the step size is small enough and the starting point is chosen sufficiently close to the minimum point. This is the main content of the following theorem (and Example 14.2.1).

Theorem 14.2.1. *Let f be a function with a local minimum point $\hat{x}$, and suppose that there is an open neighborhood U of $\hat{x}$ such that the restriction $f|_U$ of f to U is μ-strongly convex and differentiable with a Lipschitz continuous derivative and Lipschitz constant L. The gradient descent algorithm with constant step size h then converges at least linearly to $\hat{x}$ provided that the*

step size is sufficiently small and the starting point x_0 lies sufficiently close to $\hat{x}$.

More precisely: If the ball centered at $\hat{x}$ and with radius equal to $\|x_0 - \hat{x}\|$ lies in U and if $h \leq \mu/L^2$, and $(x_k)_0^\infty$ is the sequence of points generated by the algorithm, then x_k lies in U and

$$\|x_{k+1} - \hat{x}\| \leq c\|x_k - \hat{x}\|,$$

for all k, where $c = \sqrt{1 - h\mu}$.

Proof. Suppose inductively that the points $x_0, x_1, \ldots, x_k$ lie in U and that $\|x_k - \hat{x}\| \leq \|x_0 - \hat{x}\|$. Since the restriction $f|_U$ is assumed to be μ-strongly convex and since $f'(\hat{x}) = 0$,

$$\langle f'(x_k), x_k - \hat{x}\rangle = \langle f'(x_k) - f'(\hat{x}), x_k - \hat{x}\rangle \geq \mu\|x_k - \hat{x}\|^2$$

according to Theorem 7.3.1, and since the derivative is assumed to be Lipschitz continuous, we also have the inequality

$$\|f'(x_k)\| = \|f'(x_k) - f'(\hat{x})\| \leq L\|x_k - \hat{x}\|.$$

By combining these two inequalities, we obtain the inequality

$$\begin{aligned}\langle f'(x_k), x_k - \hat{x}\rangle &\geq \mu\|x_k - \hat{x}\|^2 = \frac{\mu}{2}\|x_k - \hat{x}\|^2 + \frac{\mu}{2}\|x_k - \hat{x}\|^2 \\ &\geq \frac{\mu}{2}\|x_k - \hat{x}\|^2 + \frac{\mu}{2L^2}\|f'(x_k)\|^2.\end{aligned}$$

Our next point $x_{k+1} = x_k - hf'(x_k)$ therefore satisfies the inequality

$$\begin{aligned}\|x_{k+1} - \hat{x}\|^2 &= \|x_k - hf'(x_k) - \hat{x}\|^2 = \|(x_k - \hat{x}) - hf'(x_k)\|^2 \\ &= \|x_k - \hat{x}\|^2 - 2h\langle f'(x_k), x_k - \hat{x}\rangle + h^2\|f'(x_k)\|^2 \\ &\leq \|x_k - \hat{x}\|^2 - h\mu\|x_k - \hat{x}\|^2 - h\frac{\mu}{L^2}\|f'(x_k)\|^2 + h^2\|f'(x_k)\|^2 \\ &= (1 - h\mu)\|x_k - \hat{x}\|^2 + h\big(h - \frac{\mu}{L^2}\big)\|f'(x_k)\|^2.\end{aligned}$$

Hence, $h \leq \mu/L^2$ implies that $\|x_{k+1} - \hat{x}\|^2 \leq (1 - h\mu)\|x_k - \hat{x}\|^2$, and this proves that the inequality of the theorem holds with $c = \sqrt{1 - h\mu} < 1$, and that the induction hypothesis is satisfied by the point x_{k+1}, too, since it lies closer to $\hat{x}$ than the point x_k does. So the gradient descent algorithm converges at least linearly for f under the given conditions on h and x_0. □

We can obtain a slightly sharper result for μ-strongly convex functions that are defined on the whole $\mathbf{R}^n$ and have a Lipschitz continuous derivative.

Theorem 14.2.2. *Let f be a function in the class $\mathcal{S}_{\mu,L}(\mathbf{R}^n)$. The gradient descent method, with arbitrary starting point x_0 and constant step size h, generates a sequence $(x_k)_0^\infty$ of points that converges at least linearly to the function's minimum point $\hat{x}$, if*

$$0 < h \leq \frac{2}{\mu + L}.$$

More precisely,

$$\|x_k - \hat{x}\| \leq \Big(1 - \frac{2h\mu L}{\mu + L}\Big)^{k/2} \|x_0 - \hat{x}\|. \tag{14.5}$$

Moreover, if $h = \dfrac{2}{\mu + L}$ then

$$\|x_k - \hat{x}\| \leq \Big(\frac{Q-1}{Q+1}\Big)^k \|x_0 - \hat{x}\| \qquad \textit{and} \tag{14.6}$$

$$f(x_k) - f_{\min} \leq \frac{L}{2}\Big(\frac{Q-1}{Q+1}\Big)^{2k} \|x_0 - \hat{x}\|^2, \tag{14.7}$$

where $Q = L/\mu$ is the condition number of the function class $\mathcal{S}_{\mu,L}(\mathbf{R}^n)$.

Proof. The function f has a unique minimum point $\hat{x}$, according to Corollary 8.1.7, and

$$\|x_{k+1} - \hat{x}\|^2 = \|x_k - \hat{x}\|^2 - 2h\langle f'(x_k), x_k - \hat{x}\rangle + h^2\|f'(x_k)\|^2,$$

just as in the proof of Theorem 14.2.1. Since $f'(\hat{x}) = 0$, it now follows from Theorem 7.4.4 (with $x = \hat{x}$ and $v = x_k - \hat{x}$) that

$$\langle f'(x_k), x_k - \hat{x}\rangle \geq \frac{\mu L}{\mu + L}\|x_k - \hat{x}\|^2 + \frac{1}{\mu + L}\|f'(x_k)\|^2,$$

which inserted in the above equation results in the inequality

$$\|x_{k+1} - \hat{x}\|^2 \leq \Big(1 - \frac{2h\mu L}{\mu + L}\Big)\|x_k - \hat{x}\|^2 + h\Big(h - \frac{2}{\mu + L}\Big)\|f'(x_k)\|^2.$$

So if $h \leq 2/(\mu + L)$, then

$$\|x_{k+1} - \hat{x}\| \leq \Big(1 - \frac{2h\mu L}{\mu + L}\Big)^{1/2}\|x_k - \hat{x}\|,$$

and inequality (14.5) now follows by iteration.

The particular choice of $h = 2(\mu + L)^{-1}$ in inequality (14.5) gives us inequality (14.6), and the last inequality (14.7) follows from inequality (14.6) and Theorem 1.1.2, since $f'(\hat{x}) = 0$. □

The rate of convergence in Theorems 14.2.1 and 14.2.2 depends on the condition number $Q \geq 1$. The smaller the Q, the faster the convergence. The constants μ and L, and hence the condition number Q, are of course rarely known in practical examples, so the two theorems have a qualitative character and can rarely be used to predict the number of iterations required to achieve a certain precision.

Our next example shows that inequality (14.6) can not be sharpened.

EXAMPLE 14.2.1. Consider the function

$$f(x) = \tfrac{1}{2}(\mu x_1^2 + L x_2^2),$$

where $0 < \mu \leq L$. This function belongs to the class $\mathcal{S}_{\mu,L}(\mathbf{R}^2)$, $f'(x) = (\mu x_1, L x_2)$, and $\hat{x} = (0, 0)$ is the minimum point.

The gradient descent algorithm with constant step size $h = 2(\mu + L)^{-1}$, starting point $x^{(0)} = (L, \mu)$, and $\alpha = \frac{Q-1}{Q+1}$ proceeds as follows

$$\begin{aligned}
x^{(0)} &= (L, \mu) \\
f'(x^{(0)}) &= (\mu L, \mu L) \\
x^{(1)} &= x^{(0)} - h f'(x^{(0)}) = \alpha(L, -\mu) \\
f'(x^{(1)}) &= \alpha(\mu L, -\mu L) \\
x^{(2)} &= x^{(1)} - h f'(x^{(1)}) = \alpha^2(L, \mu) \\
&\vdots \\
x^{(k)} &= \alpha^k(L, (-1)^k \mu)
\end{aligned}$$

Consequently,

$$\|x^{(k)} - \hat{x}\| = \alpha^k \sqrt{L^2 + \mu^2} = \alpha^k \|x^{(0)} - \hat{x}\|,$$

so inequality (14.6) holds with equality in this case. Cf. with 14.2.

Finally, it is worth noting that $2(\mu+L)^{-1}$ coincides with the step size that we would obtain if we had used exact line search in each iteration step. □

The gradient descent algorithm is not invariant under affine coordinate changes. The speed of convergence can thus be improved by first making a coordinate change that reduces the condition number.

EXAMPLE 14.2.2. We continue with the function $f(x) = \frac{1}{2}(\mu x_1^2 + L x_2^2)$ in the previous example. Make the change of variables $y_1 = \sqrt{\mu}\, x_1$, $y_2 = \sqrt{L}\, x_2$, and define the function g by

$$g(y) = f(x) = \tfrac{1}{2}(y_1^2 + y_2^2).$$

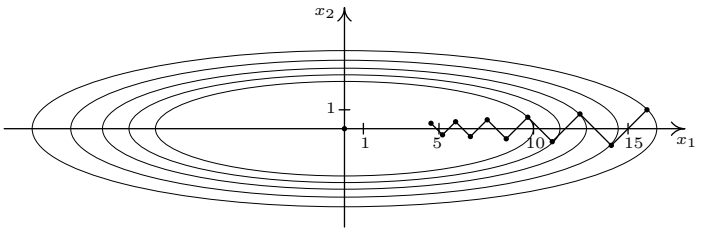

Figure 14.2. Some level curves for the function $f(x) = \frac{1}{2}(x_1^2 + 16x_2^2)$ and the progression of the gradient descent algorithm with $x^{(0)} = (16, 1)$ as starting point. The function's condition number Q is equal to 16, so the convergence to the minimum point $(0, 0)$ is relatively slow. The distance from the generated point to the origin is improved by a factor of 15/17 in each iteration.

The condition number Q of the function g is equal to 1, so the gradient descent algorithm, started from an arbitrary point $y^{(0)}$, hits the minimum point $(0, 0)$ after just one iteration. □

The gradient descent algorithm converges too slowly to be of practical use in realistic problems. In the next chapter we shall therefore study in detail a more efficient method for optimization, Newton's method.

Exercises

14.1 Perform three iterations of the gradient descent algorithm with $(1, 1)$ as starting point on the minimization problem

$$\min x_1^2 + 2x_2^2.$$

14.2 Let $X = \{x \in \mathbf{R}^2 \mid x_1 > 1\}$, let $x^{(0)} = (2, 2)$, and let $f\colon X \to \mathbf{R}$ be the function defined by $f(x) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2$.

a) Show that the sublevel set $\{x \in X \mid f(x) \leq f(x^{(0)})\}$ is not closed.

b) Obviously, $f_{\min} = \inf f(x) = \frac{1}{2}$, but show that the gradient descent method, with $x^{(0)}$ as starting point and with line search according to Armijo's rule with parameters $\alpha \leq \frac{1}{2}$ and $\beta < 1$, generates a sequence $x^{(k)} = (a_k, a_k)$, $k = 0, 1, 2, \ldots$, of points that converges to the point $(1, 1)$. So the function values $f(x^{(k)})$ converge to 1 and not to $f_{\min}$.
[Hint: Show that $a_{k+1} - 1 \leq (1 - \beta)(a_k - 1)$ for all k.]

14.3 Suppose that the gradient descent algorithm with constant step size converges to the point $\hat{x}$ when applied to a continuously differentiable function f. Prove that $\hat{x}$ is a stationary point of f, i.e. that $f'(\hat{x}) = 0$.

Chapter 15

Newton's method

In Newton's method for minimizing a function f, the search direction at a point x is determined by minimizing the function's Taylor polynomial of degree two, i.e. the polynomial

$$P(v) = f(x) + Df(x)[v] + \tfrac{1}{2}D^2f(x)[v,v] = f(x) + \langle f'(x), v\rangle + \tfrac{1}{2}\langle v, f''(x)v\rangle,$$

and since $P'(v) = f'(x) + f''(x)v$, we obtain the minimizing search vector as a solution to the equation

$$f''(x)v = -f'(x).$$

Each iteration is of course more laborious in Newton's method than in the gradient descent method, since we need to compute the second derivative and solve a quadratic equation to determine the search vector. However, as we shall see, this is more than compensated by a much faster convergence to the minimum value.

15.1 Newton decrement and Newton direction

Since the search directions in Newton's method are obtained by minimizing quadratic polynomials, we start by examining when such polynomials have minimum values, and since convexity is a necessary condition for quadratic polynomials to be bounded below, we can restrict ourself to the study of convex quadratic polynomials.

Theorem 15.1.1. *A quadratic polynomial*

$$P(v) = \tfrac{1}{2}\langle v, Av\rangle + \langle b, v\rangle + c$$

in n variables, where A is a positive semidefinite symmetric operator, is bounded below on $\mathbf{R}^n$ if and only if the equation

$$Av = -b \tag{15.1}$$

has a solution.

The polynomial has a minimum if it is bounded below, and $\hat{v}$ is a minimum point if and only if $A\hat{v} = -b$.

If $\hat{v}$ is a minimum point of the polynomial P, then

$$P(v) - P(\hat{v}) = \tfrac{1}{2}\langle v - \hat{v}, A(v - \hat{v})\rangle \tag{15.2}$$

for all $v \in \mathbf{R}^n$.

If $\hat{v}_1$ and $\hat{v}_2$ are two minimum points, then $\langle \hat{v}_1, A\hat{v}_1\rangle = \langle \hat{v}_2, A\hat{v}_2\rangle$.

Remark. Another way to state that equation (15.1) has a solution is to say that the vector $-b$, and of course also the vector b, belongs to the range of the operator A. But the range of an operator on a finite dimensional space is equal to the orthogonal complement of the null space of the operator. Hence, equation (15.1) is solvable if and only if

$$Av = 0 \Rightarrow \langle b, v\rangle = 0.$$

Proof. First suppose that equation (15.1) has no solution. Then, by the remark above there exists a vector v such that $Av = 0$ and $\langle b, v\rangle \neq 0$. It follows that

$$P(tv) = \tfrac{1}{2}\langle v, Av\rangle t^2 + \langle b, v\rangle t + c = \langle b, v\rangle t + c$$

for all $t \in \mathbf{R}$, and since the t-coefficient is nonzero, we conclude that the polynomial $P(t)$ is unbounded below.

Next suppose that $A\hat{v} = -b$. Then

$$\begin{aligned} P(v) - P(\hat{v}) &= \tfrac{1}{2}(\langle v, Av\rangle - \langle \hat{v}, A\hat{v}\rangle) + \langle b, v\rangle - \langle b, \hat{v}\rangle \\ &= \tfrac{1}{2}(\langle v, Av\rangle - \langle \hat{v}, A\hat{v}\rangle) - \langle A\hat{v}, v\rangle + \langle A\hat{v}, \hat{v}\rangle \\ &= \tfrac{1}{2}(\langle v, Av\rangle + \langle \hat{v}, A\hat{v}\rangle - \langle A\hat{v}, v\rangle - \langle \hat{v}, Av\rangle) \\ &= \tfrac{1}{2}\langle v - \hat{v}, A(v - \hat{v})\rangle \geq 0 \end{aligned}$$

for all $v \in \mathbf{R}^n$. This proves that the polynomial $P(t)$ is bounded below, that $\hat{v}$ is a minimum point, and that the equality (15.2) holds.

Since every positive semidefinite symmetric operator A has a unique positive semidefinite symmetric square root $A^{1/2}$, we can rewrite equality (15.2) as follows:

$$P(v) = P(\hat{v}) + \tfrac{1}{2}\langle A^{1/2}(v - \hat{v}), A^{1/2}(v - \hat{v})\rangle = P(\hat{v}) + \tfrac{1}{2}\|A^{1/2}(v - \hat{v})\|^2.$$

If v is another minimum point of P, then $P(v) = P(\hat{v})$, and it follows that

$$A^{1/2}(v - \hat{v}) = 0.$$

Consequently, $A(v - \hat{v}) = A^{1/2}(A^{1/2}(v - \hat{v})) = 0$, i.e. $Av = A\hat{v} = -b$. Hence, every minimum point of P is obtained as a solution to equation (15.1).

Finally, if $\hat{v}_1$ and $\hat{v}_2$ are two minimum points of the polynomial, then $A\hat{v}_1 = A\hat{v}_2$ $(= -b)$, and it follows that $\langle \hat{v}_1, A\hat{v}_1 \rangle = \langle \hat{v}_1, A\hat{v}_2 \rangle = \langle A\hat{v}_1, \hat{v}_2 \rangle = \langle A\hat{v}_2, \hat{v}_2 \rangle = \langle \hat{v}_2, A\hat{v}_2 \rangle$. □

The problem to solve a convex quadratic optimization problem in $\mathbf{R}^n$ is thus reduced to solving a quadratic system of linear equations in n variables (with a positive semidefinite coefficient matrix), which is a rather trivial numerical problem that can be performed with $O(n^3)$ arithmetic operations.

We are now ready to define the main ingredients of Newton's method.

Definition. Let $f\colon X \to \mathbf{R}$ be a twice differentiable function with an open subset X of $\mathbf{R}^n$ as domain, and let $x \in X$ be a point where the second derivative $f''(x)$ is positive semidefinite.

By a *Newton direction* Δx_{nt} of the function f at the point x we mean a solution v to the equation

$$f''(x)v = -f'(x).$$

Remark. It follows from the remark after Theorem 15.1.1 that there exists a Newton direction at x if and only if

$$f''(x)v = 0 \Rightarrow \langle f'(x), v \rangle = 0.$$

The nonexistence of Newton directions at x is thus equivalent to the existence of a vector w such that $f''(x)w = 0$ and $\langle f'(x), w \rangle = 1$.

The Newton direction Δx_{nt} is of course uniquely determined as

$$\Delta x_{\mathrm{nt}} = -f''(x)^{-1} f'(x)$$

if the second derivative $f''(x)$ is non-singular, i.e. positive definite.

A Newton direction Δx_{nt} is according to Theorem 15.1.1, whenever it exists, a minimizing vector for the Taylor polynomial

$$P(v) = f(x) + \langle f'(x), v \rangle + \tfrac{1}{2}\langle v, f''(x)v \rangle,$$

and the difference $P(0) - P(\Delta x_{\mathrm{nt}})$ is given by

$$P(0) - P(\Delta x_{\mathrm{nt}}) = \tfrac{1}{2}\langle 0 - \Delta x_{\mathrm{nt}}, f''(x)(0 - \Delta x_{\mathrm{nt}}) \rangle = \tfrac{1}{2}\langle \Delta x_{\mathrm{nt}}, f''(x)\Delta x_{\mathrm{nt}} \rangle.$$

Using the Taylor approximation $f(x+v) \approx P(v)$, we conclude that

$$f(x) - f(x+\Delta x_{\mathrm{nt}}) \approx P(0) - P(\Delta x_{\mathrm{nt}}) = \tfrac{1}{2}\langle \Delta x_{\mathrm{nt}}, f''(x)\Delta x_{\mathrm{nt}}\rangle.$$

Hence, $\frac{1}{2}\langle \Delta x_{\mathrm{nt}}, f''(x)\Delta x_{\mathrm{nt}}\rangle$ is (for small Δx_{nt}) an approximation of the decrease in function value which is obtained by replacing $f(x)$ with $f(x+\Delta x_{\mathrm{nt}})$. This motivates our next definition.

Definition. The *Newton decrement* $\lambda(f,x)$ of the function f at the point x is a quantity defined as

$$\lambda(f,x) = \sqrt{\langle \Delta x_{\mathrm{nt}}, f''(x)\Delta x_{\mathrm{nt}}\rangle}$$

if f has a Newton direction Δx_{nt} at x, and as

$$\lambda(f,x) = +\infty$$

if there is no Newton direction at x.

Note that the definition is independent of the choice of Newton direction at x in case of nonuniqueness of Newton direction. This follows immediately from the last statement in Theorem 15.1.1.

In terms of the Newton decrement, we thus have the following approximation

$$f(x) - f(x+\Delta x_{\mathrm{nt}}) \approx \tfrac{1}{2}\lambda(f,x)^2$$

for small values of Δx_{nt}.

By definition $f''(x)\Delta x_{\mathrm{nt}} = -f'(x)$, so it follows that the Newton decrement, whenever finite, can be computed using the formula

$$\lambda(f,x) = \sqrt{-\langle \Delta x_{\mathrm{nt}}, f'(x)\rangle}.$$

In particular, if x is a point where the second derivative is positive definite, then

$$\lambda(f,x) = \sqrt{\langle f''(x)^{-1}f'(x), f'(x)\rangle}.$$

Example 15.1.1. The convex one-variable function

$$f(x) = -\ln x, \;\; x > 0$$

has Newton decrement

$$\lambda(f,x) = \sqrt{\langle x^2(-x^{-1}), -x^{-1}\rangle} = \sqrt{(-x)\cdot(-x^{-1})} = 1$$

at all points $x > 0$. □

At points x with a Newton direction it is also possible to express the Newton decrement in terms of the Euclidean norm $\|\cdot\|$ as follows, by using the fact that $f''(x)$ har a positive definite symmetric square root:

$$\lambda(f,x) = \sqrt{\langle f''(x)^{1/2}\Delta x_{\mathrm{nt}}, f''(x)^{1/2}\Delta x_{\mathrm{nt}}\rangle} = \|f''(x)^{1/2}\Delta x_{\mathrm{nt}}\|.$$

The improvement in function value obtained by taking a step in the Newton direction Δx_{nt} is thus proportional to $\|f''(x)^{1/2}\Delta x_{\mathrm{nt}}\|^2$ and not to $\|\Delta x_{\mathrm{nt}}\|^2$, a fact which motivates our introduction of the following seminorm.

Definition. Let $f\colon X \to \mathbf{R}$ be a twice differentiable function with an open subset X of $\mathbf{R}^n$ as domain, and let $x \in X$ be a point where the second derivative $f''(x)$ is positive semidefinite. The function $\|\cdot\|_x\colon \mathbf{R}^n \to \mathbf{R}_+$, defined by

$$\|v\|_x = \sqrt{\langle v, f''(x)v\rangle} = \|f''(x)^{1/2}v\|$$

for all $v \in \mathbf{R}^n$, is called the *local seminorm* at x of the function f.

It is easily verified that $\|\cdot\|_x$ is indeed a seminorm on $\mathbf{R}^n$. Since

$$\{v \in \mathbf{R}^n \mid \|v\|_x = 0\} = \mathcal{N}(f''(x)),$$

where $\mathcal{N}(f''(x))$ is the null space of $f''(x)$, $\|\cdot\|_x$ is a norm if and only if the positive definite second derivative $f''(x)$ is nonsingular, i.e. positive definite.

At points x with a Newton direction, we now have the following simple relation between direction and decrement:

$$\lambda(f,x) = \|\Delta x_{\mathrm{nt}}\|_x.$$

Example 15.1.2. Let us study the Newton decrement $\lambda(f,x)$ when f is a convex quadratic polynomial, i.e. a function of the form

$$f(x) = \tfrac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c$$

with a positive semidefinite operator A. We have $f'(x) = Ax + b$, $f''(x) = A$ and $\|v\|_x = \sqrt{\langle v, Av\rangle}$, so the seminorms $\|\cdot\|_x$ are the same for all $x \in \mathbf{R}^n$.

If Δx_{nt} is a Newton direction of f at x, then

$$A\Delta x_{\mathrm{nt}} = -(Ax + b),$$

by definition, and it follows that $A(x + \Delta x_{\mathrm{nt}}) = -b$. This implies that the function f is bounded below, according to Theorem 15.1.1.

So if f is not bounded below, then there are no Newton directions at any point x, which means that $\lambda(f,x) = +\infty$ for all x.

Conversely, assume that f is bounded below. Then there exists a vector v_0 such that $Av_0 = -b$, and it follows that

$$f''(x)(v_0 - x) = Av_0 - Ax = -b - Ax = -f'(x).$$

The vector $v_0 - x$ is in other words a Newton direction of f at the point x, which means that the Newton decrement $\lambda(f, x)$ is finite at all points x and is given by

$$\lambda(f, x) = \|v_0 - x\|_x.$$

If f is bounded below without being constant, then necessarily $A \neq 0$ and we can choose a vector w such that $\|w\|_x = \sqrt{\langle w, Aw\rangle} = 1$. Let $x_k = kw + v_0$, where k is a positive number. Then

$$\lambda(f, x_k) = \|v_0 - x_k\|_{x_k} = k\|w\|_{x_k} = k,$$

and we conclude from this that $\sup_{x \in \mathbf{R}^n} \lambda(f, x) = +\infty$.

For constant functions f, the case $A = 0$, $b = 0$, we have $\|v\|_x = 0$ for all x and v, and consequently $\lambda(f, x) = 0$ for all x.

In summary, we have obtained the following result:

The Newton decrement of downwards unbounded convex quadratic functions (which includes all non-constant affine functions) is infinite at all points. The Newton decrement of downwards bounded convex quadratic functions f is finite at all points, but $\sup_x \lambda(f, x) = \infty$, unless the function is constant. □

We shall give an alternative characterization of the Newton decrement, and for this purpose we need the following useful inequality.

Theorem 15.1.2. *Suppose $\lambda(f, x) < \infty$. Then*

$$|\langle f'(x), v\rangle| \leq \lambda(f, x)\|v\|_x$$

for all $v \in \mathbf{R}^n$.

Proof. Since $\lambda(f, x)$ is assumed to be finite, there exists a Newton direction Δx_{nt} at x, and by definition, $f''(x)\Delta x_{\mathrm{nt}} = -f'(x)$. Using the Cauchy–Schwarz inequality we now obtain:

$$\begin{aligned} |\langle f'(x), v\rangle| &= |\langle f''(x)\Delta x_{\mathrm{nt}}, v\rangle| = |\langle f''(x)^{1/2}\Delta x_{\mathrm{nt}}, f''(x)^{1/2}v\rangle| \\ &\leq \|f''(x)^{1/2}\Delta x_{\mathrm{nt}}\|\|f''(x)^{1/2}v\| = \lambda(f, x)\|v\|_x. \end{aligned}$$

□

Theorem 15.1.3. *Assume as before that x is a point where the second derivative $f''(x)$ is positive semidefinite. Then*

$$\lambda(f, x) = \sup_{\|v\|_x \leq 1} \langle f'(x), v\rangle.$$

Proof. First assume that $\lambda(f,x) < \infty$. Then

$$\langle f'(x), v\rangle \le \lambda(f,x)$$

for all vectors v such that $\|v\|_x \le 1$, according to Theorem 15.1.2. In the case $\lambda(f,x) = 0$ the above inequality holds with equality for $v = 0$, so assume that $\lambda(f,x) > 0$. For $v = -\lambda(f,x)^{-1}\Delta x_{\rm nt}$ we then have $\|v\|_x = 1$ and

$$\langle f'(x), v\rangle = -\lambda(f,x)^{-1}\langle f'(x), \Delta x_{\rm nt}\rangle = \lambda(f,x).$$

This proves that $\lambda(f,x) = \sup_{\|v\|_x \le 1}\langle f'(x), v\rangle$ for finite Newton decrements $\lambda(f,x)$.

Next assume that $\lambda(f,x) = +\infty$, i.e. that no Newton direction exists at x. By the remark after the definition of Newton direction, there exists a vector w such that $f''(x)w = 0$ and $\langle f'(x), w\rangle = 1$. It follows that $\|tw\|_x = t\|w\|_x = t\sqrt{\langle w, f''(x)w\rangle} = 0 \le 1$ and $\langle f'(x), tw\rangle = t$ for all positive numbers t, and this implies that $\sup_{\|v\|_x \le 1}\langle f'(x), v\rangle = +\infty = \lambda(f,x)$. □

We sometimes need to compare $\|\Delta x_{\rm nt}\|$, $\|f'(x)\|$ and $\lambda(f,x)$, and we can do so using the following theorem.

Theorem 15.1.4. *Let $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and the largest eigenvalue of the second derivative $f''(x)$, assumed to be positive semidefinite, and suppose that the Newton decrement $\lambda(f,x)$ is finite. Then*

$$\lambda_{\min}^{1/2}\|\Delta x_{\rm nt}\| \le \lambda(f,x) \le \lambda_{\max}^{1/2}\|\Delta x_{\rm nt}\|$$

and

$$\lambda_{\min}^{1/2}\lambda(f,x) \le \|f'(x)\| \le \lambda_{\max}^{1/2}\lambda(f,x).$$

Proof. Let A be an arbitrary positive semidefinite operator on $\mathbf{R}^n$ with smallest and largest eigenvalue $\mu_{\min}$ and $\mu_{\max}$ respectively. Then

$$\mu_{\min}\|v\| \le \|Av\| \le \mu_{\max}\|v\|$$

for all vectors v.

Since $\lambda_{\min}^{1/2}$ and $\lambda_{\max}^{1/2}$ are the smallest and the largest eigenvalues of the operator $f''(x)^{1/2}$, we obtain the two inequalities of our theorem by applying the general inequality to $A = f''(x)^{1/2}$ and $v = \Delta x_{\rm nt}$, and to $A = f''(x)^{1/2}$ and $v = f''(x)^{1/2}\Delta x_{\rm nt}$, noting that $\|f''(x)^{1/2}\Delta x_{\rm nt}\| = \lambda(f,x)$ and that

$$\|f''(x)^{1/2}(f''(x)^{1/2}\Delta x_{\rm nt})\| = \|f''(x)\Delta x_{\rm nt}\| = \|f'(x)\|.$$ □

Theorem 15.1.4 is a local result, but if the function f is μ-strongly convex, then $\lambda_{\min} \geq \mu$, and if the norm of the second derivative is bounded by some constant M, then $\lambda_{\max} = \|f''(x)\| \leq M$ for all x in the domain of f. Therefore, we get the following corollary to Theorem 15.1.4.

Corollary 15.1.5. *If $f\colon X \to \mathbf{R}$ is a twice differentiable μ-strongly convex function, then*

$$\mu^{1/2}\|\Delta x_{\mathrm{nt}}\| \leq \lambda(f,x) \leq \mu^{-1/2}\|f'(x)\|$$

for all $x \in X$. If moreover $\|f''(x)\| \leq M$, then

$$M^{-1/2}\|f'(x)\| \leq \lambda(f,x) \leq M^{1/2}\|\Delta x_{\mathrm{nt}}\|.$$

The distance from an arbitrary point to the minimum point of a strongly convex function with bounded second derivative can be estimated using the Newton decrement, because we have the following result.

Theorem 15.1.6. *Let $f\colon X \to \mathbf{R}$ be a μ-strongly convex function, and suppose that f has a minimum at the point $\hat{x}$ and that $\|f''(x)\| \leq M$ for all $x \in X$. Then*

$$f(x) - f(\hat{x}) \leq \frac{M}{2\mu}\lambda(f,x)^2$$

and

$$\|x - \hat{x}\| \leq \frac{\sqrt{M}}{\mu}\lambda(f,x).$$

Proof. The theorem follows by combining Theorem 14.1.1 with the estimate $\|f'(x)\| \leq M^{1/2}\lambda(f,x)$ from Corollary 15.1.5. □

The Newton decrement is invariant under surjective affine coordinate transformations. A slightly more general result is the following.

Theorem 15.1.7. *Let f be a twice differentiable function whose domain Ω is a subset of $\mathbf{R}^n$, let $A\colon \mathbf{R}^m \to \mathbf{R}^n$ be an affine map, and let $g = f \circ A$. Let furthermore $x = Ay$ be a point in Ω, and suppose that the second derivative $f''(x)$ is positive semidefinite. The second derivative $g''(y)$ is then positive semidefinite, and the Newton decrements of the two functions g and f satisfy the inequality*

$$\lambda(g,y) \leq \lambda(f,x).$$

Equality holds if the affine map A is surjective.

Proof. The affine map can be written as $Ay = Cy + b$, where C is a linear map and b is a vector, and the chain rule gives us the identities

$$\langle g'(y), w\rangle = \langle f'(x), Cw\rangle \quad \text{and} \quad \langle w, g''(y)w\rangle = \langle Cw, f''(x)Cw\rangle$$

for arbitrary vectors w in $\mathbf{R}^m$. It follows from the latter identity that the second derivative $g''(y)$ is positive semidefinite if $f''(x)$ is so, and that

$$\|w\|_y = \|Cw\|_x.$$

An application of Theorem 15.1.3 now gives

$$\lambda(g,y) = \sup_{\|w\|_y \leq 1} \langle g'(y), w\rangle = \sup_{\|Cw\|_x \leq 1} \langle f'(x), Cw\rangle \leq \sup_{\|v\|_x \leq 1} \langle f'(x), v\rangle = \lambda(f,x).$$

If the affine map A is surjective, then C is a surjective linear map, and hence $v = Cw$ runs through all of $\mathbf{R}^n$ as w runs through $\mathbf{R}^m$. In this case, the only inequality in the above chain of equalities and inequalities becomes an equality, which means that $\lambda(g,y) = \lambda(f,x)$. □

15.2 Newton's method

The algorithm

Newton's method for minimizing a twice differentiable function f is a descent method, in which the search direction in each iteration is given by the Newton direction Δx_{nt} at the current point. The stopping criterion is formulated in terms of the Newton decrement; the algorithm stops when the decrement is sufficiently small. In short, therefore, the algorithm looks like this:

Newton's method

Given a starting point $x \in \operatorname{dom} f$ and a tolerance $\epsilon > 0$.
Repeat

1. Compute a Newton direction Δx_{nt} and the Newton decrement $\lambda(f,x)$ at x.
2. *Stopping criterion:* **stop** if $\lambda(f,x)^2 \leq 2\epsilon$.
3. Determine a step size $h > 0$.
4. *Update:* $x := x + h\Delta x_{\text{nt}}$.

The step size h is set equal to 1 in each iteration in the so-called *pure* Newton method, while it is computed by line search with Armijo's rule or otherwise in *damped* Newton methods.

The stopping criterion is motivated by the fact that $\frac{1}{2}\lambda(f,x)^2$ is an approximation to the decrease $f(x) - f(x + \Delta x_{\text{nt}})$ in function value, and if this decrease is small, it is not worthwhile to continue.

Newton's method generally works well for functions which are convex in a neighborhood of the optimal point, but it breaks down, of course, if it hits

a point where the second derivative is singular and the Newton direction is lacking. We shall show that the pure method, under appropriate conditions on the objective function f, converges to the minimum point if the starting point is sufficiently close to the minimum point. To achieve convergence for arbitrary starting points, it is necessary to use methods with damping.

Example 15.2.1. When applied to a downwards bounded convex quadratic polynomial

$$f(x) = \tfrac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c,$$

Newton's pure method finds the optimal solution after just one iteration, regardless of the choice of starting point x, because $f'(x) = Ax+b$, $f''(x) = A$ and $A\Delta x_{\text{nt}} = -(Ax + b)$, so the update $x^+ = x + \Delta x_{\text{nt}}$ satisfies the equation

$$f'(x^+) = Ax^+ + b = Ax + A\Delta x_{\text{nt}} + b = 0,$$

which means that x^+ is the optimal point. □

Invariance under change of coordinates

Unlike the gradient descent method, Newton's method is invariant under affine coordinate changes.

Theorem 15.2.1. *Let $f\colon X \to \mathbf{R}$ be a twice differentiable function with a positive definite second derivative, and let $(x_k)_0^\infty$ be the sequence generated by Newton's pure algorithm with x_0 as starting point. Let further $A\colon Y \to X$ be an affine coordinate transformation, i.e. the restriction to Y of a bijective affine map. Newton's pure algorithm applied to the function $g = f \circ A$ with $y_0 = A^{-1}x_0$ as the starting point then generates a sequence $(y_k)_0^\infty$ with the property that $Ay_k = x_k$ for each k.*

The two sequences have identical Newton decrements in each iteration, and they therefore satisfy the stopping condition during the same iteration.

Proof. The assertion about the Newton decrements follows from Theorem 15.1.7, and the relationship between the two sequences follows by induction if we show that $Ay = x$ implies that $A(y + \Delta y_{\text{nt}}) = x + \Delta x_{\text{nt}}$, where $\Delta x_{\text{nt}} = -f''(x)^{-1}f'(x)$ and $\Delta y_{\text{nt}} = -g''(y)^{-1}g'(y)$ are the uniquely defined Newton directions at the points x and y of the respective functions.

The affine map A can be written as $Ay = Cy+b$, where C is an invertible linear map and b is a vector. If $x = Ay$, then $g'(y) = C^T f'(x)$ and $g''(y) = C^T f''(x)C$, by the chain rule. It follows that

$$\begin{aligned} C\Delta y_{\text{nt}} &= -Cg''(y)^{-1}g'(y) = -CC^{-1}f''(x)^{-1}(C^T)^{-1}C^T f'(x) \\ &= -f''(x)^{-1}f'(x) = \Delta x_{\text{nt}}, \end{aligned}$$

and hence

$$A(y+\Delta y_{\mathrm{nt}}) = C(y+\Delta y_{\mathrm{nt}})+b = Cy+b+C\Delta y_{\mathrm{nt}} = Ay+\Delta x_{\mathrm{nt}} = x+\Delta x_{\mathrm{nt}}. \ \square$$

Local convergence

We will now study convergence properties for the Newton method, starting with the pure method.

Theorem 15.2.2. *Let $f\colon X \to \mathbf{R}$ be a twice differentiable, μ-strongly convex function with minimum point $\hat{x}$, and suppose that the second derivative f'' is Lipschitz continuous with Lipschitz constant L. Let x be a point in X and set*

$$x^+ = x + \Delta x_{\mathrm{nt}},$$

where Δx_{nt} is the Newton direction at x. Then

$$\|x^+ - \hat{x}\| \leq \frac{L}{2\mu}\|x - \hat{x}\|^2.$$

Moreover, if the point x^+ lies in X then

$$\|f'(x^+)\| \leq \frac{L}{2\mu^2}\|f'(x)\|^2.$$

Proof. The smallest eigenvalue of the second derivative $f''(x)$ is greater than or equal to μ by Theorem 7.3.2. Hence, $f''(x)$ is invertible and the largest eigenvalue of $f''(x)^{-1}$ is less than or equal to μ^{-1}, and it follows that

$$\|f''(x)^{-1}\| \leq \mu^{-1}. \tag{15.3}$$

To estimate the norm of $x^+ - \hat{x}$, we rewrite the difference as

$$\begin{aligned} x^+ - \hat{x} &= x + \Delta x_{\mathrm{nt}} - \hat{x} = x - \hat{x} - f''(x)^{-1}f'(x) \\ &= f''(x)^{-1}\big(f''(x)(x-\hat{x}) - f'(x)\big) = -f''(x)^{-1}w \end{aligned} \tag{15.4}$$

with

$$w = f'(x) - f''(x)(x-\hat{x}).$$

For $0 \leq t \leq 1$ we then define the vektor $w(t)$ as

$$w(t) = f'(\hat{x} + t(x-\hat{x})) - tf''(x)(x-\hat{x}),$$

and note that $w = w(1) - w(0)$, since $f'(\hat{x}) = 0$. By the chain rule,

$$w'(t) = \big(f''(\hat{x} + t(x-\hat{x})) - f''(x)\big)(x-\hat{x}),$$

and by using the Lipschitz continuity of the second derivative, we obtain the estimate

$$\begin{aligned}\|w'(t)\| &\le \|f''(\hat{x} + t(x - \hat{x})) - f''(x)\| \, \|x - \hat{x}\| \\ &\le L\|\hat{x} + t(x - \hat{x}) - x\| \|x - \hat{x}\| = L(1 - t)\|x - \hat{x}\|^2.\end{aligned}$$

Now integrate the above inequality over the interval $[0, 1]$; this results in the inequality

$$\begin{aligned}\|w\| = \Big\| \int_0^1 w'(t)\, dt \Big\| &\le \int_0^1 \|w'(t)\|\, dt \le L\|x - \hat{x}\|^2 \int_0^1 (1 - t)\, dt. \\ &= \frac{1}{2} L\|x - \hat{x}\|^2.\end{aligned} \tag{15.5}$$

By combining equality (15.4) with the inequalities (15.3) and (15.5) we obtain the estimate

$$\|x^+ - \hat{x}\| = \|f''(x)^{-1} w\| \le \|f''(x)^{-1}\| \|w\| \le \frac{L}{2\mu} \|x - \hat{x}\|^2,$$

which is the first claim of the theorem.

To prove the second claim, we assume that x^+ lies in X and consider for $0 \le t \le 1$ the vectors

$$v(t) = f'(x + t\Delta x_{\mathrm{nt}}) - t f''(x) \Delta x_{\mathrm{nt}},$$

noting that

$$v(1) - v(0) = f'(x^+) - f''(x)\Delta x_{\mathrm{nt}} - f'(x) = f'(x^+) + f'(x) - f'(x) = f'(x^+).$$

Since $v'(t) = \big(f''(x + t\Delta x_{\mathrm{nt}}) - f''(x)\big)\Delta x_{\mathrm{nt}}$, it follows from the Lipschitz continuity that

$$\|v'(t)\| \le \|f''(x + t\Delta x_{\mathrm{nt}}) - f''(x)\| \, \|\Delta x_{\mathrm{nt}}\| \le L\|\Delta x_{\mathrm{nt}}\|^2 t,$$

and by integrating this inequality, we obtain the desired estimate

$$\|f'(x^+)\| = \Big\| \int_0^1 v'(t)\, dt \Big\| \le \int_0^1 \|v'(t)\|\, dt \le \frac{L}{2} \|\Delta x_{\mathrm{nt}}\|^2 \le \frac{L}{2\mu^2} \|f'(x)\|^2,$$

where the last inequality follows from Corollary 15.1.5. □

One consequence of the previous theorem is that the pure Newton method converges quadratically when applied to functions with a positive definite second derivative that does not vary too rapidly in a neighborhood of the minimum point, provided that the starting point is chosen sufficiently close to the minimum point. More precisely, the following holds:

Theorem 15.2.3. *Let $f\colon X \to \mathbf{R}$ be a twice differentiable, μ-strongly convex function with minimum point $\hat{x}$, and suppose that the second derivative f'' is Lipschitz continuous with Lipschitz constant L. Let $0 < r \leq 2\mu/L$ and suppose that the open ball $B(\hat{x}; r)$ is included in X.*

Newton's pure method with starting point $x_0 \in B(\hat{x}; r)$ will then generate a sequence $(x_k)_0^\infty$ of points in Ω such that

$$\|x_k - \hat{x}\| \leq \frac{2\mu}{L}\Big(\frac{L}{2\mu}\|x_0 - \hat{x}\|\Big)^{2^k}$$

for all k, and the sequence therefore converges to the minimum point $\hat{x}$ as $k \to \infty$.

The convergence is very rapid. For example,

$$\|x_k - \hat{x}\| \leq \frac{2\mu}{L}\,2^{-2^k}$$

if the initial point is chosen such that $\|x_0 - \hat{x}\| \leq \mu/L$, and this implies that $\|x_k - \hat{x}\| \leq 10^{-19}\mu/L$ already for $k = 6$.

Proof. We keep the notation of Theorem 15.2.2 and then have $x_{k+1} = x_k^+$, so if x_k lies in the ball $B(\hat{x}; r)$, then

$$\|x_{k+1} - \hat{x}\| \leq \frac{L}{2\mu}\|x_k - \hat{x}\|^2, \tag{15.6}$$

and this implies that $\|x_{k+1} - \hat{x}\| < Lr^2/2\mu \leq r$, i.e. the point x_{k+1} lies in the ball $B(\hat{x}; r)$. By induction, all points in the sequence $(x_k)_0^\infty$ lie in $B(\hat{x}; r)$, and we obtain the inequality of the theorem by repeated application of inequality (15.6). □

Global convergence

Newton's damped method converges, under appropriate conditions on the objective function, for arbitrary starting points. The damping is required only during an initial phase, because the step size becomes 1 once the algorithm has produced a point where the gradient is sufficiently small. The convergence is quadratic during this second stage.

The following theorem describes a convergence result for strongly convex functions with Lipschitz continuous second derivative.

Theorem 15.2.4. *Let $f\colon X \to \mathbf{R}$ be a twice differentiable, strongly convex function with a Lipschitz continuous second derivative. Let x_0 be a point in X and suppose that the sublevel set*

$$S = \{x \in X \mid f(x) \le f(x_0)\}$$

is closed.

Then, f has a unique minimum point $\hat{x}$, and Newton's damped algorithm, with x_0 as initial point och with line search according to Armijo's rule with parameters $0 < \alpha < \frac{1}{2}$ and $0 < \beta < 1$, generates a sequence $(x_k)_0^\infty$ of points in S that converges towards the minimum point.

After an initial phase with damping, the algorithm passes into a quadratically convergent phase with step size 1.

Proof. The existence of a unique minimum point is a consequence of Corollary 8.1.7.

Suppose that f is μ-strongly convex and let L be the Lipschitz constant of the second derivative. The sublevel set S is compact since it is bounded according to Theorem 8.1.6. It follows that the distance from the set S to the boundary of the open set X is positive. Fix a positive number r that is less than this distance and also satisfies the inequality

$$r \le \mu/L.$$

Given $x \in S$ we now define the point x^+ by

$$x^+ = x + h\Delta x_{\mathrm{nt}},$$

where h is the step size according to Armijo's rule. In particular, $x_{k+1} = x_k^+$ for all k.

The core of the proof consists in showing that there are two positive constants γ and $\eta \le \mu r$ such that the following two implications hold for all $x \in S$:

(i) $\quad \|f'(x)\| \ge \eta \;\Rightarrow\; f(x^+) - f(x) \le -\gamma;$

(ii) $\quad \|f'(x)\| < \eta \;\Rightarrow\; h = 1 \;\&\; \|f'(x^+)\| < \eta.$

Suppose that we have managed to prove (i) and (ii). If $\|f'(x_k)\| \ge \eta$ for $0 \le k < m$, then

$$f_{\min} - f(x_0) \le f(x_m) - f(x_0) = \sum_{k=0}^{m-1} (f(x_k^+) - f(x_k)) \le -m\gamma,$$

because of property (i). This inequality can not hold for all m, and hence there is a smallest integer k_0 such that $\|f'(x_{k_0})\| < \eta$, and this integer must satisfy the inequality

$$k_0 \le \big(f(x_0) - f_{\min}\big)/\gamma.$$

It now follows by induction from (ii) that the step size h is equal to 1 for all $k \geq k_0$. The damped Newton algorithm is in other words a pure Newton algorithm from iteration k_0 and onwards. Because of Theorem 14.1.1,

$$\|x_{k_0} - \hat{x}\| \leq \mu^{-1}\|f'(x_{k_0})\| < \mu^{-1}\eta \leq r \leq \mu L^{-1},$$

so it follows from Theorem 15.2.3 that the sequence $(x_k)_0^\infty$ converges to $\hat{x}$, and more precisely, that the estimate

$$\|x_{k+k_0} - \hat{x}\| \leq \frac{2\mu}{L}\Big(\frac{L}{2\mu}\|x_{k_0} - \hat{x}\|\Big)^{2^k} \leq \frac{2\mu}{L}\,2^{-2^k}$$

holds for $k \geq 0$.

It thus only remains to prove the existence of numbers η and γ with the properties (i) and (ii). To this end, let

$$S_r = S + \overline{B}(x; r);$$

the set S_r is a convex and compact subset of Ω, and the two continuous functions f' and f'' are therefore bounded on S_r, i.e. there are constants K and M such that

$$\|f'(x)\| \leq K \quad \text{and} \quad \|f''(x)\| \leq M$$

for all $x \in S_r$. It follows from Theorem 7.4.1 that the derivative f' is Lipschitz continuous on the set S_r with Lipschitz constant M, i.e.

$$\|f'(y) - f'(x)\| \leq M\|y - x\|$$

for $x, y \in S_r$.

We now define our numbers η and γ as

$$\eta = \min\Big\{\frac{3(1-2\alpha)\mu^2}{L}, \mu r\Big\} \text{ and } \gamma = \frac{\alpha\beta c\mu}{M}\eta^2, \text{ where } c = \min\Big\{\frac{1}{M}, \frac{r}{K}\Big\}.$$

Let us first estimate the stepsize at a given point $x \in S$. Since

$$\|\Delta x_{\mathrm{nt}}\| \leq \mu^{-1}\|f'(x)\| \leq \mu^{-1}K,$$

the point $x + t\Delta x_{\mathrm{nt}}$ lies in i S_r and especially also in X if $0 \leq t \leq r\mu K^{-1}$. The function

$$g(t) = f(x + t\Delta x_{\mathrm{nt}})$$

is therefore defined for these t-values, and since f is μ-strongly convex and the derivative is Lipschitz continuous with constant M on S_r, it follows from Theorem 1.1.2 and Corollary 15.1.5 that

$$\begin{aligned} f(x + t\Delta x_{\mathrm{nt}}) &\leq f(x) + t\langle f'(x), \Delta x_{\mathrm{nt}}\rangle + \tfrac{1}{2}M\|\Delta x_{\mathrm{nt}}\|^2 t^2 \\ &\leq f(x) + t\langle f'(x), \Delta x_{\mathrm{nt}}\rangle + \tfrac{1}{2}M\mu^{-1}\lambda(f,x)^2 t^2 \\ &= f(x) + t\big(1 - \tfrac{1}{2}M\mu^{-1}t\big)\langle f'(x), \Delta x_{\mathrm{nt}}\rangle. \end{aligned}$$

The number $\hat{t} = c\mu$ lies in the interval $[0, r\mu K^{-1}]$ and is less than or equal to μM^{-1}. Hence, $1 - \frac{1}{2}M\mu^{-1}\hat{t} \geq \frac{1}{2} \geq \alpha$, which inserted in the above inequality gives

$$f(x + \hat{t}\Delta x_{\mathrm{nt}}) \leq f(x) + \alpha\hat{t}\,\langle f'(x), \Delta x_{\mathrm{nt}}\rangle.$$

Now, let $h = \beta^m$ be the step size given by Armijo's rule, which means that the Armijo algorithm terminates in iteration m. Since it does not terminate in iteration $m-1$, we conclude that $\beta^{m-1} > \hat{t}$, i.e.

$$h \geq \beta\hat{t} = \beta c\mu,$$

and this gives us the following estimate for the point $x^+ = x + h\Delta x_{\mathrm{nt}}$:

$$\begin{aligned} f(x^+) - f(x) &\leq \alpha h\langle f'(x), \Delta x_{\mathrm{nt}}\rangle = -\alpha h\,\lambda(f,x)^2 \\ &\leq -\alpha\beta c\mu\,\lambda(f,x)^2 \leq -\alpha\beta c\mu M^{-1}\|f'(x)\|^2 = -\gamma\eta^{-2}\|f'(x)\|^2. \end{aligned}$$

So, if $\|f'(x)\| \geq \eta$ then $f(x^+) - f(x) \leq -\gamma$, which is the content of implication (i).

To prove the remaining implication (ii), we return to the function $g(t) = f(x + t\Delta x_{\mathrm{nt}})$, assuming that $\|f'(x)\| < \eta$. The function is well-defined for $0 \leq t \leq 1$, since

$$\|\Delta x_{\mathrm{nt}}\| \leq \mu^{-1}\|f'(x)\| < \mu^{-1}\eta \leq r.$$

Moreover,

$$g'(t) = \langle f'(x + t\Delta x_{\mathrm{nt}}), \Delta x_{\mathrm{nt}}\rangle \text{ and } g''(t) = \langle \Delta x_{\mathrm{nt}}, f''(x + t\Delta x_{\mathrm{nt}})\Delta x_{\mathrm{nt}}\rangle.$$

By Lipschitz continuity,

$$\begin{aligned} |g''(t) - g''(0)| &= |\langle \Delta x_{\mathrm{nt}}, f''(x + t\Delta x_{\mathrm{nt}})\Delta x_{\mathrm{nt}}\rangle - \langle \Delta x_{\mathrm{nt}}, f''(x)\Delta x_{\mathrm{nt}}\rangle| \\ &\leq \|f''(x + t\Delta x_{\mathrm{nt}}) - f''(x)\|\,\|\Delta x_{\mathrm{nt}}\|^2 \leq tL\|\Delta x_{\mathrm{nt}}\|^3, \end{aligned}$$

and it follows, since $g''(0) = \lambda(f,x)^2$ and $\|\Delta x_{\mathrm{nt}}\| \leq \mu^{-1/2}\lambda(f,x)$, that

$$g''(t) \leq \lambda(f,x)^2 + tL\|\Delta x_{\mathrm{nt}}\|^3 \leq \lambda(f,x)^2 + tL\mu^{-3/2}\,\lambda(f,x)^3.$$

By integrating this inequality over the interval $[0,t]$, we obtain the inequality

$$g'(t) - g'(0) \leq t\lambda(f,x)^2 + \tfrac{1}{2}t^2L\mu^{-3/2}\lambda(f,x)^3.$$

But $g'(0) = \langle f'(x), \Delta x_{\mathrm{nt}}\rangle = -\lambda(f,x)^2$, so it follows that

$$g'(t) \leq -\lambda(f,x)^2 + t\lambda(f,x)^2 + \tfrac{1}{2}t^2L\mu^{-3/2}\lambda(f,x)^3,$$

and further integration results in the inequality

$$g(t) - g(0) \leq -t\lambda(f,x)^2 + \tfrac{1}{2}t^2\lambda(f,x)^2 + \tfrac{1}{6}t^3L\mu^{-3/2}\lambda(f,x)^3.$$

Now, take $t = 1$ to obtain

$$\begin{aligned} f(x + \Delta x_{\mathrm{nt}}) &\leq f(x) - \tfrac{1}{2}\lambda(f,x)^2 + \tfrac{1}{6}L\mu^{-3/2}\lambda(f,x)^3 \\ &= f(x) - \lambda(f,x)^2\big(\tfrac{1}{2} - \tfrac{1}{6}L\mu^{-3/2}\lambda(f,x)\big) \\ &= f(x) + \langle f'(x), \Delta x_{\mathrm{nt}}\rangle\big(\tfrac{1}{2} - \tfrac{1}{6}L\mu^{-3/2}\lambda(f,x)\big). \end{aligned} \tag{15.7}$$

Our assumption $\|f'(x)\| < \eta$ implies that

$$\lambda(f,x) \leq \mu^{-1/2}\|f'(x)\| < \mu^{-1/2}\eta \leq \mu^{-1/2}\cdot 3(1-2\alpha)\mu^2 L^{-1} = 3(1-2\alpha)\mu^{3/2}L^{-1}.$$

We conclude that

$$\tfrac{1}{2} - \tfrac{1}{6}L\mu^{-3/2}\lambda(f,x) > \alpha,$$

which inserted into inequality (15.7) gives us the inequality

$$f(x + \Delta x_{\mathrm{nt}}) \leq f(x) + \alpha\langle f'(x), \Delta x_{\mathrm{nt}}\rangle,$$

which tells us that the step size h is equal to 1.

The iteration leading from x to $x^+ = x + h\Delta x_{\mathrm{nt}}$ is therefore performed according to the pure Newton method. Due to the inequality

$$\|x - \hat{x}\| \leq \mu^{-1}\|f'(x)\| < \mu^{-1}\eta \leq r,$$

which holds by Theorem 14.1.1, x is a point in the ball $B(\hat{x}; r)$, so it follows from the local convergence Theorem 15.2.2 that

$$\|f'(x^+)\| \leq \frac{L}{2\mu^2}\|f'(x)\|^2. \tag{15.8}$$

Since $\eta \leq \mu r \leq \mu^2/L$,

$$\|f'(x^+)\| < \frac{L}{2\mu^2}\eta^2 \leq \frac{\eta}{2} < \eta,$$

and the proof is now complete. □

By iterating inequality (15.8), one obtains in fact the estimate

$$\|f'(x_k)\| \leq \frac{2\mu^2}{L}\Big(\frac{L}{2\mu^2}\|f'(x_{k_0})\|\Big)^{2^{k-k_0}} < \frac{2\mu^2}{L}2^{-2^{k-k_0}}$$

for $k \geq k_0$, and it now follows from Theorem 14.1.1 that

$$f(x_k) - f_{\min} < \frac{2\mu^3}{L^2}2^{-2^{k-k_0+1}}$$

for $k \geq k_0$. Combining this estimate with the previously obtained bound on k_0, one obtains an upper bound on the number of iterations required to estimate the minimum value $f_{\min}$ with a given accuracy. If

$$k > \frac{f(x_0) - f_{\min}}{\gamma} + \log_2 \log_2 \frac{2\mu^3}{L^2\epsilon},$$

then surely $f(x_k) - f_{\min} < \epsilon$. This estimate, however, is of no practical value, because the constants γ, μ and L are rarely known in concrete cases.

Another shortcoming of the classical convergence analysis of Newton's method is that the convergence constants, unlike the algorithm itself, depend on the coordinate system used. For self-concordant functions, it is however possible to carry out the convergence analysis without any unknown constants, as we shall do in Chapter 16.5.

15.3 Equality constraints

With only minor modifications, Newton's algorithm also works well when applied to convex optimization problems with constraints in the form of affine equalities.

Consider the convex optimization problem

$$\text{(P)} \qquad \begin{aligned} &\min && f(x) \\ &\text{s.t.} && Ax = b \end{aligned}$$

where $f\colon \Omega \to \mathbf{R}$ is a twice continuously differentiable convex function, Ω is an open subset of $\mathbf{R}^n$, and A is an $m \times n$-matrix.

The problem's Lagrange function $L\colon \Omega \times \mathbf{R}^m \to \mathbf{R}$ is given by

$$L(x, y) = f(x) + (Ax - b)^T y = f(x) + x^T A^T y - b^T y,$$

and according to the Karush–Kuhn–Tucker theorem (Theorem 11.2.1), a point $\hat{x}$ in Ω is an optimal solution if and only if there is a vector $\hat{y} \in \mathbf{R}^m$ such that

$$\text{(15.9)} \qquad \begin{cases} f'(\hat{x}) + A^T\hat{y} = 0 \\ A\hat{x} \qquad\quad = b. \end{cases}$$

Therefore, the minimization problem (P) is equivalent to the problem of solving the system (15.9) of linear equations.

Example 15.3.1. When f is a convex quadratic function of the form

$$f(x) = \tfrac{1}{2}\langle x, Px\rangle + \langle q, x\rangle + r,$$

the linear system (15.9) becomes

$$\begin{cases} P\hat{x} + A^T\hat{y} = -q \\ A\hat{x} \qquad\quad = \quad b, \end{cases}$$

and this is a quadratic system of linear equations with a symmetric coefficient matrix of order $m+n$. The system has a unique solution if $\operatorname{rank} A = m$ and $\mathcal{N}(A) \cap \mathcal{N}(P) = \{0\}$. See exercise 15.4. In particular, there is a unique solution if the matrix P is positive definite and $\operatorname{rank} A = m$. □

We now return to the general convex minimization problem (P). Let X denote the set of feasible points, so that

$$X = \{x \in \Omega \mid Ax = b\}.$$

In optimization problems without any constraints, the descent direction Δx_{nt} at the point x is a vector which miminizes the Taylor polynom of degree two of the function $f(x+v)$, and the minimization is over all vectors v in $\mathbf{R}^n$. As a new point x^+ with function value less than $f(x)$ we select $x^+ = x + h\Delta x_{\text{nt}}$ with a suitable step size h. In constrained problems, the new point x^+ has to be a feasible point, of course, and this requires that $A\Delta x_{\text{nt}} = 0$. The minimization of the Taylor polynomial is therefore restricted to vectors v that satisfy the condition $Av = 0$, and this means that we have to modify our previous definition of Newton direction and decrement as follows for constrained optimization problems.

Definition. In the equality constrained minimization problem (P), a vector Δx_{nt} is called a *Newton direction* at the point $x \in X$ if there exists a vector $w \in \mathbf{R}^m$ such that

$$\begin{cases} f''(x)\Delta x_{\text{nt}} + A^T w = -f'(x) \\ \quad A\Delta x_{\text{nt}} \qquad\quad = \quad 0. \end{cases} \tag{15.10}$$

The quantity

$$\lambda(f, x) = \sqrt{\langle \Delta x_{\text{nt}}, f''(x)\Delta x_{\text{nt}} \rangle}$$

is called the *Newton decrement*.

It follows from Example 15.3.1 that the Newton direction Δx_{nt} (if it exists) is an optimal solution to the minimization problem

$$\begin{array}{ll} \min & f(x) + \langle f'(x), v \rangle + \frac{1}{2}\langle v, f''(x)v \rangle \\ \text{s.t.} & Av = 0. \end{array}$$

And if $(\Delta x_{\rm nt}, w)$ is a solution to the system (15.10), then

$$\begin{aligned} -\langle f'(x), \Delta x_{\rm nt}\rangle &= \langle f''(x)\Delta x_{\rm nt} + A^T w, \Delta x_{\rm nt}\rangle \\ &= \langle f''(x)\Delta x_{\rm nt}, \Delta x_{\rm nt}\rangle + \langle w, A\Delta x_{\rm nt}\rangle \\ &= \langle f''(x)\Delta x_{\rm nt}, \Delta x_{\rm nt}\rangle + \langle w, 0\rangle = \langle \Delta x_{\rm nt}, f''(x)\Delta x_{\rm nt}\rangle, \end{aligned}$$

so it follows that

$$\lambda(f,x) = \sqrt{-\langle f'(x), \Delta x_{\rm nt}\rangle},$$

just as for unconstrained problems.

The objective function is decreasing in the Newton direction, because

$$\frac{d}{dt} f(x + t\Delta x_{\rm nt})\big|_{t=0} = \langle f'(x), \Delta x_{\rm nt}\rangle = -\lambda(f,x)^2 \leq 0,$$

so $\Delta x_{\rm nt}$ is indeed a descent direction.

Let $P(v)$ denote the Taylor polynomial of degree two of the function $f(x+v)$. Then

$$\begin{aligned} f(x) - f(x + \Delta x_{\rm nt}) &\approx P(0) - P(\Delta x_{\rm nt}) \\ &= -\langle f'(x), \Delta x_{\rm nt}\rangle - \tfrac{1}{2}\langle \Delta x_{\rm nt}, f''(x)\Delta x_{\rm nt}\rangle = \tfrac{1}{2}\lambda(f,x)^2, \end{aligned}$$

just as in the unconstrained case.

With our modified definition of the Newton direction, we can now copy Newton's method verbatim for convex minimization problem of the type

$$\begin{array}{ll} \text{min} & f(x) \\ \text{s.t.} & Ax = b. \end{array}$$

The algorithm looks like this:

Newton's method

Given a starting point $x \in \Omega$ satisfying the constraint $Ax = b$, and a tolerance $\epsilon > 0$.

Repeat

1. Compute the Newton direction $\Delta x_{\rm nt}$ at x by solving the system of equations (15.10), and compute the Newton decrement $\lambda(f,x)$.
2. *Stopping criterion:* **stop** if $\lambda(f,x)^2 \leq 2\epsilon$.
3. Compute a step size $h > 0$.
4. *Update:* $x := x + h\Delta x_{\rm nt}$.

Elimination of constraints

An alternative approach to the optimization problem

$$\text{(P)} \qquad \begin{aligned} &\min && f(x) \\ &\text{s.t.} && Ax = b, \end{aligned}$$

with $x \in \Omega$ as implicit condition and $r = \operatorname{rank} A$, is to solve the system of equations $Ax = b$ and to express r variables as linear combinations of the remaining $n - r$ variables. The former variables can then be eliminated from the objective function, and we obtain in this way an optimization problem in $n - r$ variables without explicit constraints, a problem that can be attacked with Newton's method. We will describe this approach in more detail and compare it with the method above.

Suppose that the set X of feasible points is nonempty, choose a point $a \in X$, and select an affine parametrization

$$x = \xi(z), \quad z \in \tilde{\Omega}$$

of X with $\xi(0) = a$. Since $\{x \in \mathbf{R}^n \mid Ax = b\} = a + \mathcal{N}(A)$, we can write the parametrization as

$$\xi(z) = a + Cz$$

where $C\colon \mathbf{R}^p \to \mathbf{R}^n$ is an injective linear map, whose range $\mathcal{V}(C)$ coincides with the null space $\mathcal{N}(A)$ of the map A, and $p = n - \operatorname{rank} A$. The domain $\tilde{\Omega} = \{z \in \mathbf{R}^p \mid a + Cz \in \Omega\}$ is an open convex subset of $\mathbf{R}^p$.

A practical way to construct the parametrization is of course to solve the system $Ax = b$ by Gaussian elimination.

Let us finally define the function $\tilde{f}\colon \tilde{\Omega} \to \mathbf{R}$ by setting $\tilde{f}(z) = f(\xi(z))$. The problem (P) is then equivalent to the convex optimization problem

$$\text{(}\tilde{\text{P}}\text{)} \qquad \min \tilde{f}(z)$$

which has no explicit constraints.

Let Δx_{nt} be a Newton direction of the function f at the point x, i.e. a vector that satisfies the system (15.10) for a suitably chosen vector w. We will show that the function $\tilde{f}$ has a corresponding Newton direction Δz_{nt} at the point $z = \xi^{-1}(x)$, and that $\Delta x_{\text{nt}} = C\Delta z_{\text{nt}}$.

Since $A\Delta x_{\text{nt}} = 0$ and $\mathcal{N}(A) = \mathcal{V}(C)$, there is a unique vector v such that $\Delta x_{\text{nt}} = Cv$. By the chain rule, $\tilde{f}'(z) = C^T f'(x)$ and $\tilde{f}''(z) = C^T f''(x)C$, so it follows from the first equation in the system (15.10) that

$$\begin{aligned} \tilde{f}''(z)v &= C^T f''(x)Cv = C^T f''(x)\Delta x_{\text{nt}} = -C^T f'(x) - C^T A^T w \\ &= -\tilde{f}'(z) - C^T A^T w. \end{aligned}$$

A general result from linear algebra tells us that $\mathcal{N}(S) = \mathcal{V}(S^T)^\perp$ for arbitrary linear maps S. Applying this result to the maps C^T and A, and using that $\mathcal{V}(C) = \mathcal{N}(A)$, we obtain the equality

$$\mathcal{N}(C^T) = \mathcal{V}(C)^\perp = \mathcal{N}(A)^\perp = \mathcal{V}(A^T)^{\perp\perp} = \mathcal{V}(A^T),$$

which implies that $C^T A^T w = 0$. Hence,

$$\tilde{f}''(z)v = -\tilde{f}'(z),$$

and v is thus a Newton direction of the function $\tilde{f}$ at the point z. So, $\Delta z_{\text{nt}} = v$ is the direction vector we are looking for.

The iteration step $z \to z^+ = z + h\Delta z_{\text{nt}}$ in Newton's method for the unconstrained problem ($\tilde{\text{P}}$) takes us from the point $z = \xi^{-1}(x)$ in $\tilde{\Omega}$ to the point z^+ whose image in X is

$$\begin{aligned}\xi(z^+) &= \xi(z + h\Delta z_{\text{nt}}) = a + C(z + h\Delta z_{\text{nt}}) = a + Cz + hC(\Delta z_{\text{nt}}) \\ &= \xi(z) + h\Delta x_{\text{nt}} = x + h\Delta x_{\text{nt}},\end{aligned}$$

and this is also the point we get by applying Newton's method to the point x in the constrained problem (P).

Also note that the Newton decrements are the same at corresponding points, because

$$\begin{aligned}\lambda(\tilde{f}, z)^2 &= -\langle \tilde{f}'(z), \Delta z_{\text{nt}} \rangle = -\langle C^T f'(x), \Delta z_{\text{nt}} \rangle = -\langle f'(x), C\Delta z_{\text{nt}} \rangle \\ &= -\langle f'(x), \Delta x_{\text{nt}} \rangle = \lambda(f, x)^2.\end{aligned}$$

In summary, we have arrived at the following result.

Theorem 15.3.1. *Let $(x_k)_0^\infty$ be a sequence of points obtained by Newton's method applied to the constrained problem* (P)*. Newton's method applied to the problem* ($\tilde{\text{P}}$)*, obtained by elimination of the constraints and with $\xi^{-1}(x_0)$ as initial point, will then generate a sequence $(z_k)_0^\infty$ with the property that $x_k = \xi(z_k)$ for all k.*

Convergence analysis

No new convergence analysis is needed for the modified version of Newton's method, for we can, because of Theorem 15.3.1, apply the results of Theorem 15.2.4. If the restriction of the function $f\colon \Omega \to \mathbf{R}$ to the set X of feasible points is strongly convex and the second derivative is Lipschitz continuous, then the same also holds for the function $\tilde{f}\colon \tilde{\Omega} \to \mathbf{R}$. (Cf. with exercise 15.5.) Assuming x_0 to be a feasible starting point and the sublevel

set $\{x \in X \mid f(x) \leq f(x_0)\}$ to be closed, the damped Newton algorithm will therefore converge to the minimum point when applied to the constrained problem (P). Close enough to the minimum point, the step size h will also be equal to 1, and the convergence will be quadratic.

Exercises

15.1 Determine the Newton direction, the Newton decrement and the local norm at an arbitrary point $x > 0$ for the function $f(x) = x \ln x - x$.

15.2 Let f be the function $f(x_1, x_2) = -\ln x_1 - \ln x_2 - \ln(4 - x_1 - x_2)$ with $X = \{x \in \mathbf{R}^2 \mid x_1 > 0, x_2 > 0, x_1 + x_2 < 4\}$ as domain. Determine the Newton direction, the Newton decrement and the local norm at the point x when
a) $x = (1, 1)$ b) $x = (1, 2)$.

15.3 Determine a Newton direction, the Newton decrement and the local norm for the function $f(x_1, x_2) = \mathrm{e}^{x_1+x_2} + x_1 + x_2$ at an arbitrary point $x \in \mathbf{R}^2$.

15.4 Assume that P is a symmetric positive semidefinite $n \times n$-matrix and that A is an arbitrary $m \times n$-matrix. Prove that the matrix

$$M = \begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix}$$

is invertible if and only if $\operatorname{rank} A = m$ and $\mathcal{N}(A) \cap \mathcal{N}(P) = \{0\}$.

15.5 Assume that the function $f\colon \Omega \to \mathbf{R}$ is twice differentiable and convex, let $x = \xi(z) = a + Cz$ be an affine parametrization of the set

$$X = \{x \in \Omega \mid Ax = b\},$$

and define the function $\tilde{f}$ by $\tilde{f}(z) = f(\xi(z))$, just as in Section 15.3. Let further σ denote the smallest eigenvalue of the symmetric matrix C^TC.

a) Prove that $\tilde{f}$ is $\mu\sigma$-strongly convex if the restriction of f to X is μ-strongly convex.

b) Assume that the matrix A has full rank and that there are constants K and M such that $Ax = b$ implies

$$\left\| \begin{bmatrix} f''(x) & A^T \\ A & 0 \end{bmatrix}^{-1} \right\| \leq K \quad \text{and} \quad \|f''(x)\| \leq M.$$

Show that $\tilde{f}$ is μ-strongly convex with convexity constant $\mu = \sigma K^{-2} M^{-1}$.

Chapter 16

Self-concordant functions

Self-concordant functions were introduced by Nesterov and Nemirovski in the late 1980s as a product of their analysis of the speed of convergence of Newton's method. Classic convergence results for two times continuously differentiable functions assume that the second derivative is Lipschitz continuous, and the convergence rate depends on the Lipschitz constant. One obvious weakness of these results is that the value of the Lipschitz constant, unlike Newton's method, is not invariant under affine coordinate transformations.

Suppose that a function f, which is defined on an open convex subset X of $\mathbf{R}^n$, has a Lipschitz continuous second derivative with Lipschitz constant L, i.e. that

$$\|f''(y) - f''(x)\| \leq L\|y - x\|$$

for all $x, y \in X$. For the restriction $\phi_{x,v}(t) = f(x + tv)$ of f to a line through x with direction vector v, this means that

$$|\phi''_{x,v}(t) - \phi''_{x,v}(0)| = |\langle v, (f''(x+tv) - f''(x))v\rangle| \leq L\|x+tv-x\|\|v\|^2 = L|t|\|v\|^3.$$

So if the function f is three times differentiable, then consequently

$$|\phi'''_{x,v}(0)| = \lim_{t\to 0}\Big|\frac{\phi''_{x,v}(t) - \phi''_{x,v}(0)}{t}\Big| \leq L\|v\|^3.$$

But

$$\phi'''_{x,v}(0) = \sum_{i,j,k=1}^{n} \frac{\partial^3 f(x)}{\partial x_i \partial x_j \partial x_k} v_i v_j v_k = D^3 f(x)[v, v, v],$$

so a necessary condition for a three times differentiable function f to have a Lipschitz continuous second derivative with Lipschitz constant L is that

$$|D^3 f(x)[v, v, v]| \leq L\|v\|^3 \tag{16.1}$$

for all $x \in X$ and all $v \in \mathbf{R}^n$, and it is easy to show this is also a sufficient condition.

The reason why the value of the Lipschitz constant is not affinely invariant is that there is no natural connection between the Euclidean norm $\|\cdot\|$ and the function f. The analysis of a function's behavior is simplified if we instead use a norm that is adapted to the form of the level surfaces, and for functions with a positive semidefinite second derivative $f''(x)$, such a (semi)norm is the local seminorm $\|\cdot\|_x$, introduced in the previous chapter and defined as $\|v\|_x = \sqrt{\langle v, f''(x)v\rangle}$. Nesterov–Nemirovski's stroke of genius consisted in replacing $\|\cdot\|$ with the local seminorm $\|\cdot\|_x$ in the inequality (16.1). For the function class obtained in this way, it is possible to describe the convergence rate of Newton's method in an affinely independent way and with absolute constants.

16.1 Self-concordant functions

We are now ready for Nesterov–Nemirovski's definition of self-concordance and for a study of the basic properties of self-concordant functions.

Definition. Let $f\colon X \to \mathbf{R}$ be a three times continuously differentiable function with an open convex subset X of $\mathbf{R}^n$ as domain. The function is called *self-concordant* if it is convex, and the inequality

$$\bigl|D^3 f(x)[v,v,v]\bigr| \le 2\bigl(D^2 f(x)[v,v]\bigr)^{3/2} \tag{16.2}$$

holds for all $x \in X$ and all $v \in \mathbf{R}^n$.

Since $D^2 f(x)[v,v] = \|v\|_x^2$, where $\|\cdot\|_x$ is the local seminorm defined by the function f at the point x, we can also write the defining inequality (16.2) as

$$\bigl|D^3 f(x)[v,v,v]\bigr| \le 2\|v\|_x^3,$$

and it is this shorter version that we will prefer, when we work with a single function f.

Remark 1. There is nothing special about the constant 2 in inequality (16.2). If f satisfies the inequality $\bigl|D^3 f(x)[v,v,v]\bigr| \le K\|v\|_x^3$, then the function $F = \frac{1}{4}K^2 f$, obtained from f by scaling, is self-concordant. The choice of 2 as the constant facilitates, however, the wording of a number of results.

Remark 2. For functions f defined on subsets of the real axis and $v \in \mathbf{R}$, $\|v\|_x^2 = f''(x)v^2$ and $D^3 f(x)[v,v,v] = f'''(x)v^3$. Hence, a convex function $f\colon X \to \mathbf{R}$ is self-concordant if and only if

$$|f'''(x)| \leq 2f''(x)^{3/2}$$

for all $x \in X$.

Remark 3. In terms of the restriction $\phi_{x,v}(t) = f(x+tv)$ of the function f to the line through x with direction v, we can equivalently write the inequality

$$\big|D^3 f(x+tv)[v,v,v]\big| \leq 2\|v\|^3_{x+tv}$$

as $|\phi'''_{x,v}(t)| \leq 2\phi''_{x,v}(t)^{3/2}$. A three times continuously differentiable convex function of several variables is therefore self-concordant if and only if all its restrictions to lines are self-concordant.

EXAMPLE 16.1.1. The convex function $f(x) = -\ln x$ is self-concordant on its domain $\mathbf{R}_{++}$. Indeed, inequality (16.2) holds with equality for this function, since $f''(x) = x^{-2}$ and $f'''(x) = -2x^{-3}$. □

EXAMPLE 16.1.2. Convex quadratic functions $f(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c$ are self-concordant since $D^3 f(x)[v,v,v] = 0$ for all x and v.

Hence, affine functions are self-concordant, and the function $x \mapsto \|x\|^2$, where $\|\cdot\|$ is the Euclidean norm, is self-concordant. □

The expression

$$D^3 f(x)[u,v,w] = \sum_{i,k,k=1}^{n} \frac{\partial^3 f(x)}{\partial x_i \partial x_j \partial x_k} u_i v_j w_k$$

is a symmetric trilinear form in the variables u, v, and w, if the function f is three times continuously differentiable in a neighborhood of the point x. For self-concordant functions we have the following generalization of inequality (16.2) in the definition of self-concordance.

Theorem 16.1.1. *Suppose $f\colon X \to \mathbf{R}$ is a self-concordant function. Then,*

$$\big|D^3 f(x)[u,v,w]\big| \leq 2\|u\|_x\|v\|_x\|w\|_x$$

for all $x \in X$ and all vectors u, v, w in $\mathbf{R}^n$.

Proof. The proof is based on a general theorem on norms of symmetric trilinear forms, which is proven in an appendix to this chapter.

Assume first that x is a point where the second derivative $f''(x)$ is positive definite. Then $\|\cdot\|_x$ is a norm with $\langle u, v\rangle_x = \langle u, f''(x)v\rangle$ as the corresponding scalar product. We can therefore apply Theorem 1 of the appendix to the symmetric trilinear form $D^3 f(x)[u,v,w]$ with $\|\cdot\|_x$ as the underlying norm, and it follows that

$$\sup_{u,v,w\neq 0} \frac{\big|D^3 f(x)[u,v,w]\big|}{\|u\|_x\|v\|_x\|w\|_x} = \sup_{v\neq 0} \frac{\big|D^3 f(x)[v,v,v]\big|}{\|v\|^3_x} \leq 2,$$

which is equivalent to the assertion of the theorem.

To cope with points where the second derivative is singular, we consider for $\epsilon > 0$ the scalar product

$$\langle u, v\rangle_{x,\epsilon} = \langle u, f''(x)v\rangle + \epsilon\langle u, v\rangle,$$

where $\langle \cdot\,, \cdot\rangle$ is the usual standard scalar product, and the corresponding norm

$$\|v\|_{x,\epsilon} = \sqrt{\langle v, v\rangle_{x,\epsilon}} = \sqrt{\|v\|_x^2 + \epsilon\|v\|^2}.$$

Obviously, $\|v\|_x \leq \|v\|_{x,\epsilon}$ for all vectors v, and hence

$$|D^3 f(x)[v, v, v]| \leq 2\|v\|_{x,\epsilon}^3$$

for all v, since f is self-concordant. It now follows from Theorem 1 in the appendix that

$$\begin{aligned}|D^3 f(x)[u, v, w]| &\leq 2\|u\|_{x,\epsilon}\|v\|_{x,\epsilon}\|w\|_{x,\epsilon} \\ &= 2\sqrt{(\|u\|_x^2 + \epsilon\|u\|^2)(\|v\|_x^2 + \epsilon\|v\|^2)(\|w\|_x^2 + \epsilon\|u\|w^2)},\end{aligned}$$

and we get the sought-after inequality by letting $\epsilon \to 0$. □

Theorem 16.1.2. *The second derivative $f''(x)$ of a self-concordant function $f\colon X \to \mathbf{R}$ has the same null space $\mathcal{N}(f''(x))$ at all points $x \in X$.*

Proof. We recall that $\mathcal{N}(f''(x)) = \{v \mid \|v\|_x = 0\}$.

Let x and y be two points in X. For reasons of symmetry, we only have to show the inclusion $\mathcal{N}(f''(x)) \subseteq \mathcal{N}(f''(y))$.

Assume therefore that $v \in \mathcal{N}(f''(x))$ and let $x^t = x + t(y - x)$. Since X is an open convex set, there is certainly a number $a > 1$ such that the points x^t lie in X for $0 \leq t \leq a$, and we now define a function $g\colon [0, a] \to \mathbf{R}$ by setting

$$g(t) = D^2 f(x^t)[v, v] = \|v\|_{x^t}^2.$$

Then $g(0) = \|v\|_x^2 = 0$ and $g(t) \geq 0$ for $0 \leq t \leq a$, and since

$$g'(t) = D^3 f(x^t)[v, v, y - x],$$

it follows from Theorem 16.1.1 that

$$|g'(t)| \leq 2\|v\|_{x^t}^2\|y - x\|_{x^t} = 2g(t)\|y - x\|_{x^t}.$$

But the seminorm

$$\|y - x\|_{x^t} = \sqrt{D^2 f(x^t)[y - x, y - x]}$$

depends continuously on t, and it is therefore bounded above by some constant C on the interval $[0, a]$. Hence,

$$|g'(t)| \leq 2Cg(t)$$

for $0 \leq t \leq a$. It now follows from Theorem 2 in the appendix to this chapter that $g(t) = 0$ for all t, and in particular, $g(1) = \|v\|_y^2 = 0$, which proves that $v \in \mathcal{N}(f''(y))$. This proves the inclusion $\mathcal{N}(f''(x)) \subseteq \mathcal{N}(f''(y))$. □

Our next corollary is just a special case of Theorem 16.1.2, because $f''(x)$ is non-singular if and only if $\mathcal{N}(f''(x)) = \{0\}$.

Corollary 16.1.3. *The second derivative of a self-concordant function is either non-singular at all points or singular at all points.*

A self-concordant function will be called *non-degenerate* if its second derivative is positive definite at all points, and by the above corollary, that is the case if the second derivative is positive definite at one single point.

A non-degenerate self-concordant function is in particular strictly convex.

Operations that preserve self-concordance

Theorem 16.1.4. *If f is a self-concordant function and $\alpha \geq 1$, then αf is self-concordant.*

Proof. If $\alpha \geq 1$, then $\alpha \leq \alpha^{3/2}$, and it follows that

$$\begin{aligned}\big|D^3(\alpha f)(x)[v,v,v]\big| &= \alpha\big|D^3 f(x)[v,v,v]\big| \leq 2\alpha\big(D^2 f(x)[v,v]\big)^{3/2} \\ &\leq 2\big(\alpha D^2 f(x)[v,v]\big)^{3/2} = 2\big(D^2(\alpha f)(x)[v,v]\big)^{3/2}.\end{aligned}$$ □

Theorem 16.1.5. *The sum $f + g$ of two self-concordant functions f and g is self-concordant on its domain.*

Proof. We use the elementary inequality

$$a^{3/2} + b^{3/2} \leq (a+b)^{3/2},$$

which holds for all nonnegative numbers a, b (and is easily proven by squaring both sides) and the triangle inequality to obtain

$$\begin{aligned}\big|D^3(f+g)(x)[v,v,v]\big| &= \big|D^3 f(x)[v,v,v] + D^3 g(x)[v,v,v]\big| \\ &\leq 2\big(D^2 f(x)[v,v]\big)^{3/2} + 2\big(D^2 g(x)[v,v]\big)^{3/2} \\ &\leq 2\big(D^2 f(x)[v,v] + D^2 g(x)[v,v]\big)^{3/2} \\ &= 2\big(D^2(f+g)(x)[v,v]\big)^{3/2}.\end{aligned}$$ □

Theorem 16.1.6. *If the function $f\colon X \to \mathbf{R}$ is self-concordant, where X is an open convex subset of $\mathbf{R}^n$, and A is an affine map from $\mathbf{R}^m$ to $\mathbf{R}^n$, then the composition $g = f \circ A$ is a self-concordant function on its domain $A^{-1}(X)$.*

Proof. The affine map A can be written as $Ay = Cy + b$, where C is a linear map and b is a vector. Let y be a point in $A^{-1}(X)$ and let u be a vector in $\mathbf{R}^m$, and write $x = Ay$ och $v = Cu$. According to the chain rule,

$$\begin{aligned} D^2g(y)[u,u] &= D^2f(Ay)[Cu,Cu] = D^2f(x)[v,v] \qquad \text{and} \\ D^3g(y)[u,u,u] &= D^3f(Ay)[Cu,Cu,Cu] = D^3f(x)[v,v,v], \end{aligned}$$

so it follows that

$$\begin{aligned} \big|D^3g(y)[u,u,u]\big| &= \big|D^3f(x)[v,v,v]\big| \le 2\big(D^2f(x)[v,v]\big)^{3/2} \\ &= 2\big(D^2g(y)[u,u]\big)^{3/2}. \end{aligned}$$

□

Example 16.1.3. It follows from Example 16.1.1 and Theorem 16.1.6 that the function $f(x) = -\ln(b - \langle c, x\rangle)$ with domain $\{x \in \mathbf{R}^n \mid \langle c, x\rangle < b\}$ is self-concordant. □

Example 16.1.4. Suppose that the polyhedron

$$X = \bigcap_{j=1}^{p} \{x \in \mathbf{R}^n \mid \langle c_j, x\rangle \le b_j\}$$

has nonempty interior. The function $f(x) = -\sum_{j=1}^{p} \ln(b_j - \langle c_j, x\rangle)$, with $\operatorname{int} X$ as domain, is self-concordant. □

16.2 Closed self-concordant functions

In Section 6.7 we studied the recessive subspace of arbitrary convex functions. The properties of the recessive subspace of a closed self-concordant function is given by the following theorem.

Theorem 16.2.1. *Suppose that $f\colon X \to \mathbf{R}$ is a closed self-concordant function. The function's recessive subspace V_f is then equal to the null space $\mathcal{N}(f''(x)$ of the second derivative $f''(x)$ at an arbitrary point $x \in X$. Moreover,*

(i) $X = X + V_f$.

(ii) $f(x+v) = f(x) + Df(x)[v]$ *for all vectors* $v \in V_f$.

(iii) If $\lambda(f,x) < \infty$, *then* $f(x+v) = f(x)$ *for all* $v \in V_f$.

Proof. Assertions (i) and (ii) are true for the recessive subspace of an arbitrary differentiable convex function according to Theorem 6.7.1, so we only have to prove the remaining assertions.

Let x be an arbitrary point in X and let v be an arbitrary vector in $\mathbf{R}^n$, and consider the restriction $\phi_{x,v}(t) = f(x + tv)$ of f to the line through x with direction v. The domain of $\phi_{x,v}$ is an open interval $I =]\alpha, \beta[$ around 0.

First suppose that $v \in V_f$. Then

$$\phi_{x,v}(t) = f(x) + tDf(x)[v]$$

for all $t \in I$ becuse of property (ii), and it follows that

$$\|v\|_x^2 = D^2 f(x)[v, v] = \phi''_{x,v}(0) = 0,$$

i.e. the vector v belongs to the null space of $f''(x)$. This proves the inclusion $V_f \subseteq \mathcal{N}(f''(x))$. Note that this inclusion holds for arbitrary twice differentiable convex functions without any assumptions concerning self-concordance and closedness.

To prove the converse inclusion $\mathcal{N}(f''(x)) \subseteq V_f$, we instead assume that v is a vector in $\mathcal{N}(f''(x))$. Since $\mathcal{N}(f''(x + tv)) = \mathcal{N}(f''(x))$ for all $t \in I$ due to Theorem 16.1.2, we now have

$$\phi''_{x,v}(t) = D^2 f(x + tv)[v, v] = \|v\|_{x+tv}^2 = 0$$

for all $t \in I$, and it follows that

$$\phi_{x,v}(t) = \phi_{x,v}(0) + \phi'_{x,v}(0)t = f(x) + Df(x)[v]\, t.$$

If $\beta < \infty$, then $x + \beta v$ is a boundary point of X and $\lim_{t\to\beta} \phi_{x,v}(t) < \infty$. However, according to Corollary 8.2.2 this is a contradiction to f being a closed function. Hence, $\beta = \infty$, and similarly, $\alpha = -\infty$. This means that $I =]-\infty, \infty[$, and in particular, I contains the number 1. We conclude that the point $x + v$ lies in X and that $f(x + v) = \phi_{x,v}(1) = f(x) + Df(x)[v]$ for all $x \in X$ and all $v \in \mathcal{N}(f''(x))$, and Theorem 6.7.1 now provides us with the inclusion $\mathcal{N}(f''(x)) \subseteq V_f$. Hence, $V_f = \mathcal{N}(f''(x))$.

Finally, suppose that $\lambda(f, x) < \infty$. Then there exists, by definition, a Newton direction at x, and this implies, according to the remark after the definition of Newton direction, that the implication

$$f''(x)v = 0 \Rightarrow Df(x)[v] = 0$$

holds. Since $V_f = \mathcal{N}(f''(x))$, it now follows from assertion (ii) that $f(x+v) = f(x)$ for all $v \in V_f$. □

The problem of minimizing a degenerate closed self-concordant function $f\colon X \to \mathbf{R}$ with finite Newton decrement $\lambda(f,x)$ at all points $x \in X$ can be reduced to the problem of minimizing a non-degenerate closed self-concordant function as follows.

Assume that the domain X is a subset of $\mathbf{R}^n$, and let V_f denote the recessive subspace of f. Put $m = \dim V_f^\perp$ and let $A\colon \mathbf{R}^m \to \mathbf{R}^n$ be an arbitrary injective linear map onto $V_f^\perp$, and put $X_0 = A^{-1}(X)$. The set X_0 is then an open subset of $\mathbf{R}^m$, and we obtain a function $g\colon X_0 \to \mathbf{R}$ by defining $g(y) = f(Ay)$ for $y \in X_0$.

The function g is self-concordant according to Theorem 16.1.6, and since (y,t) belongs to the epigraph of g if and only if (Ay,t) belongs to the epigraph of f, it follows that g is also a closed function.

Suppose $v \in \mathcal{N}(g''(y))$. Since $g''(y) = A^T f''(Ay)A$,

$$\langle Av, f''(Ay)Av\rangle = \langle v, A^T f''(Ay)Av\rangle = \langle v, g''(y)v\rangle = 0,$$

which means that the vector Av belongs to $\mathcal{N}(f''(Ay))$, i.e. to the recessive subspace V_f. But Av also belongs to $V_f^\perp$, by definition, and $V_f \cap V_f^\perp = \{0\}$, so it follows that $Av = 0$. Hence $v = 0$, since A is an injective map. This proves that $\mathcal{N}(g''(y)) = \{0\}$, which means that g is a non-degenerate function.

Each vector $x \in X$ has a unique decomposition $x = x_1 + x_2$ with $x_1 \in V_f^\perp$ and $x_2 \in V_f$, and x_1 $(= x - x_2)$ lies in X according to Theorem 16.2.1. Consequently, there is a unique point $y \in X_0$ such that $Ay = x_1$. Therefore, $g(y) = f(Ay) = f(x_1) = f(x)$ by the same theorem.

The functions f and g thus have the same ranges, and $\hat{y}$ is a minimum point of g if and only if $A\hat{y}$ is a minimum point of f, and thereby also all points $A\hat{y} + v$ with $v \in V_f$ are minimum points of f.

We also note for future use that

$$\lambda(g,y) \le \lambda(f,Ay) = \lambda(f,Ay+v)$$

for all $y \in X_0$ and all $v \in V_f$, according to Theorem 15.1.7. (In the present case, the two Newton decrements are actually equal, which we leave as an exercise to show.)

Corollary 16.2.2. *A closed self-concordant function $f\colon X \to \mathbf{R}$ is non-degenerate if its domain X does not contain any line.*

Proof. By Theorem 16.2.1, $X = X + V_f$. Hence, if f is degenerate, then X contains all lines through points in X with directions given by nonzero vectors in V_f. So the function must be non-degenerate if its domain does not contain any lines. □

Corollary 16.2.3. *A closed self-concordant function is non-degenerate if and only if it is strictly convex.*

Proof. The second derivative $f''(x)$ of a non-degenerate self-concordant function f is positive definit for all x in its domain, and this implies that f is strictly convex.

The recessive subspace V_f of a degenerate function f is non-trivial, and the restriction $\phi_{x,v}(t) = f(x + tv)$ of f to a line with a direction given by a nonzero vector $v \in V_f$ is affine, according to Theorem 16.2.1. This prevents f from being strictly convex. □

16.3 Basic inequalities for the local seminorm

The graph of a convex function f lies above its tangent planes, and the vertical distance between the point $(y, f(y))$ on the graph and the tangent plane through the point $(x, f(x)$ is greater than or equal to $\frac{1}{2}\mu\|y-x\|^2$ if f is μ-strongly convex. The same distance is also bounded below if the function is self-concordant, but now by an expression that is a function of the local norm $\|y - x\|_x$. The actual function ρ is defined in the following lemma, which also describes all the properties of ρ that we will need.

Lemma 16.3.1. *Let* $\rho\colon]-\infty, 1[\to \mathbf{R}$ *be the function*

$$\rho(t) = -t - \ln(1 - t).$$

(i) The function ρ *is convex, strictly decreasing in the interval* $]-\infty, 0]$*, and strictly increasing in the interval* $[0, 1[$*, and* $\rho(0) = 0$*.*

(ii) For $0 \le t < 1$*,*

$$\rho(t) \le \frac{t^2}{2(1-t)}.$$

In particular, $\rho(t) \le t^2$ *if* $0 \le t \le \frac{1}{2}$*.*

(iii) If $s < 1$ *and* $t < 1$*, then* $\rho(s) + \rho(t) \ge -st$*.*

(iv) If $s \ge 0$*,* $0 \le t < 1$ *and* $\rho(-s) \le \rho(t)$*, then* $s \le \dfrac{t}{1-t}$*.*

Proof. Assertion (i) follows easily by considering the sign of the derivative, and assertion (ii) follows from the Taylor series expansion, which gives

$$\rho(t) = \tfrac{1}{2}t^2 + \tfrac{1}{3}t^3 + \tfrac{1}{4}t^4 + \cdots \le \tfrac{1}{2}t^2(1 + t + t^2 + \cdots) = \tfrac{1}{2}t^2(1-t)^{-1}$$

for $0 \le t < 1$.

To prove (iii), we use the elementary inequality $x - \ln(1 + x) \ge 0$ and take $x = st - s - t$. This gives

$$\begin{aligned} st + \rho(s) + \rho(t) &= st - s - t - \ln(1-s) - \ln(1-t) \\ &= st - s - t - \ln(1 + st - s - t) \ge 0. \end{aligned}$$

Since ρ is strictly decreasing in the interval $]-\infty, 0]$, assertion (iv) will follow once we show that $\rho(-s) \geq \rho(t)$ when $s = t/(1-t)$. To show this inequality, let

$$g(t) = \rho\big(-\frac{t}{1-t}\big) - \rho(t)$$

for $0 \leq t < 1$. We simplify and obtain

$$g(t) = t - 1 + (1-t)^{-1} + 2\ln(1-t).$$

Since $g(0) = 0$ and $g'(t) = 1 + (1-t)^{-2} - 2(1-t)^{-1} = t^2(1-t)^{-2} \geq 0$, we conclude that $g(t) \geq 0$ for all $t \in [0, 1[$, and this completes the proof of assertion (iv). □

The next theorem is used to estimate differences of the form $\|w\|_y - \|w\|_x$, $Df(y)[w] - Df(x)[w]$, and $f(y) - f(x) - Df(x)[y-x]$ in terms of $\|w\|_x$, $\|y-x\|_x$ and the function ρ.

Theorem 16.3.2. *Let $f\colon X \to \mathbf{R}$ be a closed self-concordant function, and suppose that x is a point in X and that $\|y-x\|_x < 1$. Then, y is also a point in X, and the following inequalities hold for the vector $v = y - x$ and arbitrary vectors w:*

$$\frac{\|v\|_x}{1+\|v\|_x} \leq \|v\|_y \leq \frac{\|v\|_x}{1-\|v\|_x} \tag{16.3}$$

$$\frac{\|v\|_x^2}{1+\|v\|_x} \leq Df(y)[v] - Df(x)[v] \leq \frac{\|v\|_x^2}{1-\|v\|_x} \tag{16.4}$$

$$\rho(-\|v\|_x) \leq f(y) - f(x) - Df(x)[v] \leq \rho(\|v\|_x) \tag{16.5}$$

$$(1-\|v\|_x)\|w\|_x \leq \|w\|_y \leq \frac{\|w\|_x}{1-\|v\|_x} \tag{16.6}$$

$$Df(y)[w] - Df(x)[w] \leq D^2 f(x)[v,w] + \frac{\|v\|_x^2\|w\|_x}{1-\|v\|_x} \leq \frac{\|v\|_x\|w\|_x}{1-\|v\|_x}. \tag{16.7}$$

The left parts of the three inequalities (16.3)*,* (16.4) *and* (16.5) *are also satisfied with $v = y - x$ for all $y \in X$.*

Proof. We leave the proof that y belongs to X to the end and start by showing that the inequalities (16.3–16.7) hold under the additional assumption $y \in X$.

I. We begin with inequality (16.6). If $\|w\|_x = 0$, then $\|w\|_z = 0$ for all $z \in X$, according to Theorem 16.1.2. Hence, the inequality holds in this case. Therefore, let w be an arbitrary vector with $\|w\|_x \neq 0$, let $x^t = x + t(y-x)$, and define the function ψ by

$$\psi(t) = \|w\|_{x^t}^{-1} = \big(D^2 f(x^t)[w,w]\big)^{-1/2}.$$

The function ψ is defined on an open interval that contains the interval $[0,1]$, $\psi(0) = \|w\|_x^{-1}$ and $\psi(1) = \|w\|_y^{-1}$. It now follows, using Theorem 16.1.1, that

$$
\begin{aligned}
(16.8) \qquad |\psi'(t)| &= \frac{1}{2}\big|\big(D^2 f(x^t)[w,w]\big)^{-3/2} D^3 f(x^t)[w,w,v]\big| \\
&= \frac{1}{2}\|w\|_{x^t}^{-3}\big|D^3 f(x^t)[w,w,v]\big| \le \frac{1}{2}\|w\|_{x^t}^{-3} \cdot 2\|w\|_{x^t}^2 \|v\|_{x^t} \\
&= \|w\|_{x^t}^{-1}\|v\|_{x^t} = \psi(t)\|v\|_{x^t}.
\end{aligned}
$$

If $\|v\|_x = 0$, then $\|v\|_z = 0$ for all $z \in X$, and hence $\psi'(t) = 0$ for $0 \le t \le 1$. This implies that $\psi(1) = \psi(0)$, i.e. that $\|w\|_y = \|w\|_x$. The inequalities (16.3) and (16.6) are thus satisfied in the case $\|v\|_x = 0$.

Assume henceforth that $\|v\|_x \ne 0$, and first take $w = v$ in the definition of the function ψ. In this special case, inequality (16.8) simplifies to $|\psi'(t)| \le 1$ for $t \in [0,1]$, and hence $\psi(0) - 1 \le \psi(1) \le \psi(0) + 1$, by the mean-value theorem. The right part of this inequality means that $\|v\|_y^{-1} \le \|v\|_x^{-1} + 1$, which after rearrangement gives the left part of inequality (16.3). Note, that this is true even in the case $\|v\|_x \ge 1$.

Correspondingly, the left part of the same inequality gives rise to the right part of inequality (16.3), now under the assumption that $\|v\|_x < 1$.

To prove inequality (16.6), we return to the function ψ with a general w. Since $\|tv\|_x = t\|v\|_x < 1$ for $0 \le t \le 1$, it follows from the already proven inequality (16.3) (with $x^t = x + tv$ instead of y) that

$$
\|v\|_{x^t} = \frac{1}{t}\|tv\|_{x^t} \le \frac{1}{t} \cdot \frac{\|tv\|_x}{1 - \|tv\|_x} = \frac{\|v\|_x}{1 - t\|v\|_x}.
$$

Insert this estimate into (16.8); this gives us the following inequality for the derivative of the function $\ln \psi(t)$:

$$
|(\ln \psi(t))'| = \frac{|\psi'(t)|}{\psi(t)} = \|v\|_{x^t} \le \frac{\|v\|_x}{1 - t\|v\|_x}.
$$

Let us now integrate this inequality over the interval $[0,1]$; this results in the estimate

$$
\begin{aligned}
\big|\ln \frac{\|w\|_y}{\|w\|_x}\big| = \big|\ln \frac{\psi(0)}{\psi(1)}\big| &= \big|\ln \psi(1) - \ln \psi(0)\big| = \Big|\int_0^1 (\ln \psi(t))'\, dt\Big| \\
&\le \int_0^1 \frac{\|v\|_x}{1 - t\|v\|_x}\, dt = -\ln(1 - \|v\|_x),
\end{aligned}
$$

which after exponentiation yields

$$
1 - \|v\|_x \le \frac{\|w\|_y}{\|w\|_x} \le (1 - \|v\|_x)^{-1},
$$

and this is inequality (16.6).

II. To prove the inequality (16.4), we define

$$\phi(t) = Df(x^t)[v],$$

where $x^t = x + t(y - x)$, as before. Then

$$\phi'(t) = D^2 f(x^t)[v, v] = \|v\|_{x^t}^2 = t^{-2}\|tv\|_{x^t}^2,$$

so by using inequality (16.3), we obtain the inequality

$$\frac{\|v\|_x^2}{(1 + t\|v\|_x)^2} = \frac{1}{t^2}\frac{\|tv\|_x^2}{(1 + \|tv\|_x)^2} \leq \phi'(t) \leq \frac{1}{t^2}\frac{\|tv\|_x^2}{(1 - \|tv\|_x)^2} = \frac{\|v\|_x^2}{(1 - t\|v\|_x)^2}$$

for $0 \leq t \leq 1$. The left part of this inequality holds with $v = y - x$ for all $y \in X$, and the right part holds if $\|v\|_x < 1$, and by integrating the inequality over the interval [0,1], we arrive at inequality (16.4).

III. To prove inequality (16.5), we start with the function

$$\Phi(t) = f(x^t) - Df(x)[v]\, t,$$

noting that

$$\Phi(1) - \Phi(0) = f(y) - f(x) - Df(x)[v]$$

and that

$$\Phi'(t) = Df(x^t)[v] - Df(x)[v].$$

By replacing y with x^t in inequality (16.4) , we obtain the following inequality

$$\frac{t\|v\|_x^2}{1 + t\|v\|_x} \leq \Phi'(t) \leq \frac{t\|v\|_x^2}{1 - t\|v\|_x},$$

where the right part holds only if $\|v\|_x < 1$. By integrating the above inequality over the interval $[0, 1]$, we obtain

$$\rho(-\|v\|_x) = \int_0^1 \frac{t\|v\|_x^2}{1 + t\|v\|_x}\, dt \leq \Phi(1) - \Phi(0) \leq \int_0^1 \frac{t\|v\|_x^2}{1 - t\|v\|_x}\, dt = \rho(\|v\|_x),$$

i.e. inequality (16.5).

IV. The proof of inequality (16.7) is analogous to the proof of inequality (16.4), but this time our function ϕ is defined as

$$\phi(t) = Df(x^t)[w].$$

Now, $\phi'(t) = D^2 f(x^t)[w, v]$ and $\phi''(t) = D^3 f(x^t)[w, v, v]$, so it follows from Theorem 16.1.1 and inequality (16.6) that

$$|\phi''(t)| \le 2\|w\|_{x^t}\|v\|_{x^t}^2 \le 2\frac{\|w\|_x\|v\|_x^2}{(1-t\|v\|_x)^3}.$$

By integrating this inequality over the interval $[0, s]$, where $s \le 1$, we get the estimate

$$\begin{aligned}\phi'(s) - \phi'(0) &\le \int_0^s |\phi''(t)|\,dt \le 2\|w\|_x \int_0^s \frac{\|v\|_x^2\,dt}{(1-t\|v\|_x)^3} \\ &= \|w\|_x\Big[\frac{\|v\|_x}{(1-s\|v\|_x)^2} - \|v\|_x\Big],\end{aligned}$$

and another integration over the interval $[0, 1]$ results in the inequality

$$\phi(1) - \phi(0) - \phi'(0) \le \int_0^1 (\phi'(s) - \phi'(0))\,ds \le \frac{\|w\|_x\|v\|_x^2}{1-\|v\|_x},$$

which is the left part of inequality (16.7).

By the Cauchy–Schwarz inequality,

$$\begin{aligned}D^2 f(x)[v, w] &= \langle v, f''(x)w\rangle = \langle f''(x)^{1/2}v, f''(x)^{1/2}w\rangle \\ &\le \|f''(x)^{1/2}v\|\|f''(x)^{1/2}w\| = \|v\|_x\|w\|_x,\end{aligned}$$

and we obtain the right part of inequality (16.7) by replacing $D^2 f(x)[v, w]$ with its majorant $\|v\|_x\|w\|_x$.

V. It now only remains to prove that the condition $\|y-x\|_x < 1$ implies that the point y lies in X.

Assume the contrary. i.e. that there is a point y outside X such that $\|y - x\|_x < 1$. The line segment $[x, y]$ then intersects the boundary of X in a point $x + \bar{t}v$, where $\bar{t}$ is a number in the interval $]0, 1]$. The function ρ is increasing in the interval $[0, 1[$, and hence $\rho(t\|v\|_x) \le \rho(\|v\|_x)$ if $0 \le t < \bar{t}$. It therefore follows from inequality (16.5) that

$$f(x+tv) \le f(x)+tDf(x)[v]+\rho(t\|v\|_x) \le f(x)+|Df(x)[v]|+\rho(\|v\|_x) < +\infty$$

for all t in the interval $[0, \bar{t}[$. However, this is a contradiction, because $\lim_{t\to\bar{t}} f(x + tv) = +\infty$, since f is a closed function and $x + \bar{t}v$ is a boundary point. Thus, y is a point in X. □

16.4 Minimization

This section focuses on minimizing self-concordant functions, and the results are largely based on the following theorem, which also plays a significant role in our study of Newton's algorithm in the next section.

Theorem 16.4.1. *Let $f\colon X \to \mathbf{R}$ be a closed self-concordant function, suppose that $x \in X$ is a point with finite Newton decrement $\lambda = \lambda(f,x)$, let Δx_{nt} be a Newton direction at x, and define*

$$x^+ = x + (1+\lambda)^{-1}\Delta x_{\mathrm{nt}}.$$

The point x^+ is then a point in X and

$$f(x^+) \leq f(x) - \rho(-\lambda).$$

Remark. So a minimum point $\hat{x}$ of f must satisfy the inequality

$$f(\hat{x}) \leq f(x) - \rho(-\lambda)$$

for all $x \in X$ with finite Newton decrement λ.

Proof. The vector $v = (1+\lambda)^{-1}\Delta x_{\mathrm{nt}}$ has local seminorm

$$\|v\|_x = (1+\lambda)^{-1}\|\Delta x_{\mathrm{nt}}\|_x = \lambda(1+\lambda)^{-1} < 1,$$

so it follows from Theorem 16.3.2 that the point $x^+ = x + v$ lies in X and that

$$\begin{aligned} f(x^+) &\leq f(x) + Df(x)[v] + \rho(\|v\|_x) = f(x) + \frac{1}{1+\lambda}\langle f'(x), \Delta x_{\mathrm{nt}}\rangle + \rho(\frac{\lambda}{1+\lambda}) \\ &= f(x) - \frac{\lambda^2}{1+\lambda} - \frac{\lambda}{1+\lambda} - \ln\frac{1}{1+\lambda} = f(x) - \lambda + \ln(1+\lambda) \\ &= f(x) - \rho(-\lambda). \end{aligned}$$

□

Theorem 16.4.2. *The Newton decrement $\lambda(f,x)$ of a downwards bounded closed self-concordant function $f\colon X \to \mathbf{R}$ is finite at each point $x \in X$ and $\inf_{x\in X} \lambda(f,x) = 0$.*

Proof. Let v be an arbitrary vector in the recessive subspace $V_f = \mathcal{N}(f''(x))$. Then

$$f(x+tv) = f(x) + t\langle f'(x), v\rangle$$

for all $t \in \mathbf{R}$ according to Theorem 16.2.1, and since f is supposed to be bounded below, this implies that $\langle f'(x), v\rangle = 0$. This proves the implication

$$f''(x)v = 0 \Rightarrow \langle f'(x), v\rangle = 0,$$

which means that there exists a Newton direction at the point x. Hence, $\lambda(f, x)$ is a finite number.

If there is a positive number δ such that $\lambda(f, x) \geq \delta$ for all $x \in X$, then repeated application of Theorem 16.4.1, with an arbitrary point $x_0 \in X$ as starting point, results in a sequence $(x_k)_0^\infty$ of points in X, defined as $x_{k+1} = x_k^+$ and satisfying the inequality

$$f(x_k) \leq f(x_0) - k\rho(-\delta)$$

for all k. Since $\rho(-\delta) > 0$, this contradicts our assumption that f is bounded below. Thus, $\inf_{x \in X} \lambda(f, x) = 0$. □

Theorem 16.4.3. *All sublevel sets of a non-degenerate closed self-concordant function $f\colon X \to \mathbf{R}$ are compact sets if $\lambda(f, x_0) < 1$ for some point $x_0 \in X$.*

Proof. The sublevel sets are closed since the function is closed, and to prove that they are also bounded it is enough to prove that the particular sublevel set $S = \{x \in X \mid f(x) \leq f(x_0)\}$ is bounded, because of Theorem 6.8.3.

So, let x be an arbitrary point in S, and write $r = \|x - x_0\|_{x_0}$ and $\lambda_0 = \lambda(f, x_0)$ for short. Then

$$f(x) \geq f(x_0) + Df(x_0)[x - x_0] + \rho(-r),$$

according to Theorem 16.3.2, and

$$Df(x_0)[x - x_0] = \langle f'(x_0), x - x_0 \rangle \geq -\lambda(f, x_0)\|x - x_0\|_{x_0} = -\lambda_0 r,$$

by Theorem 15.1.2. Combining these two inequalities we obtain the inequality

$$f(x_0) \geq f(x) \geq f(x_0) - \lambda_0 r + \rho(-r),$$

which simplifies to

$$r - \ln(1 + r) = \rho(-r) \leq \lambda_0 r.$$

Hence,

$$(1 - \lambda_0)r \leq \ln(1 + r)$$

and it follows that $r \leq r_0$, r_0 being the unique positive root of the equation $(1 - \lambda_0)r = \ln(1 + r)$. The sublevel set S is thus included in the ellipsoid $\{x \in \mathbf{R}^n \mid \|x - x_0\|_{x_0} \leq r_0\}$, and it is therefore a bounded set. □

Theorem 16.4.4. *A closed self-concordant function $f\colon X \to \mathbf{R}$ has a minimum point if $\lambda(f, x_0) < 1$ for some point $x_0 \in X$.*

Proof. If in addition f is non-degenerate, then $S = \{x \in X \mid f(x) \leq f(x_0)\}$ is a compact set according to the previous theorem, so the restriction of f to the sublevel set S attains a mininum, and this minimum is clearly a

global minimum of f. The minimum point is furthermore unique, since non-degenerate self-concordant functions are strictly convex.

If f is degenerate, then there is a non-degenerate closed self-concordant function $g\colon X_0 \to \mathbf{R}$ with the same range as f, according to the discussion following Theorem 16.2.1. The relationship between the two functions has the form $g(y) = f(Ay + v)$, where A is an injective linear map and v is an arbitrary vector in the recessive subspace V_f. To the point x_0 there corresponds a point $y_0 \in X_0$ such that $Ay_0 + v = x_0$ for some $v \in V_f$, and $\lambda(g, y_0) \leq \lambda(f, x_0) < 1$. By the already proven part of the theorem, g has a minimum point $\hat{y}$, and this implies that all points in the set $A\hat{y} + V_f$ are minimum points of f. □

Theorem 16.4.5. *Every downwards bounded closed self-concordant function $f\colon X \to \mathbf{R}$ has a minimum point.*

Proof. It follows from Theorem 16.4.2 that there is a point $x_0 \in X$ such that $\lambda(f, x_0) < 1$, so the theorem is a corollary of Theorem 16.4.4. □

Our next theorem describes how well a given point approximates the minimum point of a closed self-concordant function.

Theorem 16.4.6. *Let $f\colon X \to \mathbf{R}$ be a closed self-concordant function with a minimum point $\hat{x}$. If $x \in X$ is an arbitrary point with Newton decrement $\lambda = \lambda(f, x) < 1$, then*

$$\rho(-\lambda) \leq f(x) - f(\hat{x}) \leq \rho(\lambda), \tag{16.9}$$

$$\frac{\lambda}{1+\lambda} \leq \|x - \hat{x}\|_x \leq \frac{\lambda}{1-\lambda}, \tag{16.10}$$

$$\|x - \hat{x}\|_{\hat{x}} \leq \frac{\lambda}{1-\lambda}. \tag{16.11}$$

Remark. Since $\rho(t) \leq t^2$ if $t \leq \frac{1}{2}$, we conclude from inequality (16.9) that

$$f(x) - f_{\min} \leq \lambda(f, x)^2$$

as soon as $\lambda(f, x) \leq \frac{1}{2}$.

Proof. To simplify the notation, let $v = x - \hat{x}$ and $r = \|v\|_x$.

The left part of inequality (16.9) follows directly from the remark after Theorem 16.4.1. To prove the right part of the same inequality, we recall the inequality

$$\langle f'(x), v\rangle \leq \lambda(f, x)\|v\|_x = \lambda r, \tag{16.12}$$

which we combine with the left part of inequality (16.5) in Theorem 16.3.2 and inequality (iii) in Lemma 16.3.1. This results in the following chain of inequalities:

$$\begin{aligned} f(\hat{x}) = f(x - v) &\geq f(x) + \langle f'(x), -v\rangle + \rho(-\|-v\|_x) \\ &= f(x) - \langle f'(x), v\rangle + \rho(-r) \\ &\geq f(x) - \lambda r + \rho(-r) \geq f(x) - \rho(\lambda), \end{aligned}$$

and the proof of inequality (16.9) is now complete.

Since $x - v = \hat{x}$ and $f'(\hat{x}) = 0$, it follows from inequality (16.12) and the left part of inequality (16.4) that

$$\lambda r \geq \langle f'(x), v\rangle = \langle f'(x - v), -v\rangle - \langle f'(x), -v\rangle \geq \frac{\|-v\|_x^2}{1 + \|-v\|_x} = \frac{r^2}{1 + r},$$

and by solving the inequality above with respect to r, we obtain the right part of inequality (16.10).

The left part of the same inequality obviously holds if $r \geq 1$. So assume that $r < 1$. Due to inequality (16.7),

$$\langle f'(x), w\rangle = \langle f'(x - v), -w\rangle - \langle f'(x), -w\rangle \leq \frac{\|-v\|_x \|-w\|_x}{1 - \|-v\|_x} = \frac{r}{1 - r}\|w\|_x,$$

and hence

$$\lambda = \sup_{\|w\|_x \leq 1} \langle f'(x), w\rangle \leq \frac{r}{1 - r},$$

which gives the left part of inequality (16.10).

To prove the remaining inequality (16.11), we use the left part of inequality (16.5) with y replaced by x and x replaced by $\hat{x}$, which results in the inequality

$$\rho(-\|x - \hat{x}\|_{\hat{x}}) \leq f(x) - f(\hat{x}).$$

According to the already proven inequality (16.9), $f(x) - f(\hat{x}) \leq \rho(\lambda)$, so it follows that $\rho(-\|x - \hat{x}\|_{\hat{x}}) \leq \rho(\lambda)$, and by Lemma 16.3.1, this means that $\|x - \hat{x}\|_{\hat{x}} \leq \dfrac{\lambda}{1 - \lambda}$. □

Theorem 16.4.7. *Let f be a closed self-concordant function whose domain X is a subset of $\mathbf{R}^n$, and suppose that*

$$\nu = \sup\{\lambda(f, x) \mid x \in X\} < 1.$$

Then X is equal to the whole space $\mathbf{R}^n$, and f is a constant function.

Proof. It follows from Theorem 16.4.4 that f has a minimum point $\hat{x}$ and from inequality (16.9) in Theorem 16.4.6 that

$$\rho(-\nu) \leq f(x) - f(\hat{x}) \leq \rho(\nu)$$

for all $x \in X$. Thus, f is a bounded function, and since f is closed, this implies that X is a set without boundary points. Hence, $X = \mathbf{R}^n$.

Let v be an arbitrary vector in $\mathbf{R}^n$. By applying inequality (16.11) with $x = \hat{x} + tv$, we obtain the inequality

$$t\|v\|_{\hat{x}} = \|x - \hat{x}\|_{\hat{x}} \leq \frac{\lambda(f,x)}{1-\lambda(f,x)} \leq \frac{\nu}{1-\nu}$$

for all $t > 0$, and this implies that $\|v\|_{\hat{x}} = 0$. The recessive subspace V_f of f is in other words equal to $\mathbf{R}^n$, so f is a constant function according to Theorem 16.2.1. □

16.5 Newton's method for self-concordant functions

In this section we show that Newton's method converges when the objective function $f\colon X \to \mathbf{R}$ is closed, self-concordant and bounded below. We shall also give an estimate of the number of iterations needed to obtain the minimum with a given accuracy ϵ – an estimate that only depends on ϵ and the difference between the minimum value and the function value at the starting point. The algorithm starts with a damped phase, which requires no line search as the step length at the point x can be chosen equal to $1/(1+\lambda(f,x))$, and then enters into a pure phase with quadratic convergence, when the Newton decrement is sufficiently small.

The damped phase

During the damped phase, the points x_k in Newton's algorithm are generated recursively by the equation

$$x_{k+1} = x_k + \frac{1}{1+\lambda_k} v_k,$$

where $\lambda_k = \lambda(f, x_k)$ is the Newton decrement at x_k and v_k is a Newton direction at the same point, i.e

$$f''(x_k)v_k = -f'(x_k).$$

According to Theorem 16.4.1, if the starting point x_0 is a point in X, then all generated points x_k will lie in X and

$$f(x_{k+1}) - f(x_k) \leq \rho(-\lambda_k).$$

If $\delta > 0$ and $\lambda_k \geq \delta$, then $\rho(-\lambda_k) \geq \rho(-\delta)$, because the function $\rho(t)$ is decreasing for för $t < 0$. So if x_N is the first point of the sequence that satisfies the inequality $\lambda_N = \lambda(f, x_N) < \delta$, then

$$\begin{aligned} f_{\min} - f(x_0) \leq f(x_N) - f(x_0) &= \sum_{k=0}^{N-1} (f(x_{k+1}) - f(x_k)) \\ &\leq -\sum_{k=0}^{N-1} \rho(-\lambda_k) \leq -\sum_{k=0}^{N-1} \rho(-\delta) = -N\rho(-\delta), \end{aligned}$$

which implies that att $N \leq (f(x_0) - f_{\min})/\rho(-\delta)$. This proves the following theorem.

Theorem 16.5.1. *Let $f\colon X \to \mathbf{R}$ be a closed, self-concordant and downwards bounded function. Using Newton's damped algorithm with step size as above, we need at most*

$$\left\lfloor \frac{f(x_0) - f_{min}}{\rho(-\delta)} \right\rfloor$$

iterations to generate a point x with Newton decrement $\lambda(f, x) < \delta$ from an arbitrary starting point x_0 in X.

Local convergence

We now turn to the study of Newton's pure method for starting points that are sufficiently close to the minimum point $\hat{x}$. For a corresponding analysis of Newton's damped method we refer to exercise 16.6.

Theorem 16.5.2. *Let $f\colon X \to \mathbf{R}$ be a closed self-concordant function, and suppose that $x \in X$ is a point with Newton decrement $\lambda(f, x) < 1$. Let Δx_{nt} be a Newton direction at x, and let*

$$x^+ = x + \Delta x_{\mathrm{nt}}.$$

Then, x^+ is a point in X and

$$\lambda(f, x^+) \leq \Big(\frac{\lambda(f, x)}{1 - \lambda(f, x)}\Big)^2.$$

Proof. The conclusion that x^+ lies in X follows from Theorem 16.3.2, because $\|\Delta x_{\mathrm{nt}}\|_x = \lambda(f,x) < 1$. To prove the inequality for $\lambda(f,x^+)$, we first use inequality (16.7) of the same theorem with $v = x^+ - x = \Delta x_{\mathrm{nt}}$ and obtain

$$\begin{aligned}\langle f'(x^+), w\rangle &\leq \langle f'(x), w\rangle + \langle f''(x)\Delta x_{\mathrm{nt}}, w\rangle + \frac{\lambda(f,x)^2\|w\|_x}{1-\lambda(f,x)} \\ &= \langle f'(x), w\rangle + \langle -f'(x), w\rangle + \frac{\lambda(f,x)^2\|w\|_x}{1-\lambda(f,x)} = \frac{\lambda(f,x)^2\|w\|_x}{1-\lambda(f,x)}.\end{aligned}$$

But

$$\|w\|_x \leq \frac{\|w\|_{x^+}}{1-\lambda(f,x)},$$

by inequality (16.6), so it follows that

$$\langle f'(x^+), w\rangle \leq \frac{\lambda(f,x)^2\|w\|_{x^+}}{(1-\lambda(f,x))^2},$$

and this implies that

$$\lambda(f,x^+) = \sup_{\|w\|_{x^+}\leq 1} \langle f'(x^+), w\rangle \leq \frac{\lambda(f,x)^2}{(1-\lambda(f,x))^2}. \qquad \square$$

We are now able to prove the following convergence result for Newton's pure method.

Theorem 16.5.3. *Suppose that $f\colon X \to \mathbf{R}$ is a closed self-concordant function and that x_0 is a point in X with Newton decrement*

$$\lambda(f,x_0) \leq \delta < \overline{\lambda} = \tfrac{1}{2}(3-\sqrt{5}) = 0.381966\ldots.$$

Let the sequence $(x_k)_0^\infty$ be recursively defined by

$$x_{k+1} = x_k + v_k,$$

where v_k is a Newton direction at the point x_k.

The sequence $(f(x_k))_0^\infty$ converges to the minimum value $f_{\min}$ of the function f, and if $\epsilon > 0$ then

$$f(x_k) - f_{\min} < \epsilon$$

for $k > A + \log_2(\log_2 B/\epsilon)$, where A and B are constants that only depend on δ.

Moreover, if f is a non-degenerate function, then $(x_k)_0^\infty$ converges to the unique minimum point of f.

Proof. The critical number $\overline{\lambda}$ is a root of the equation $(1-\lambda)^2 = \lambda$, and if $0 \le \lambda < \overline{\lambda}$ then $\lambda < (1-\lambda)^2$.

Let $K(\lambda) = (1-\lambda)^{-2}$; the function K is increasing in the interval $[0, \overline{\lambda}[$ and $K(\lambda)\lambda < 1$. It therefore follows from Theorem 16.5.2 that the following inequality is true for all points $x \in X$ with $\lambda(f,x) \le \delta < \overline{\lambda}$:

$$\lambda(f,x^+) \le K(\lambda(f,x))\,\lambda(f,x)^2 \le K(\delta)\lambda(f,x)^2 \le K(\delta)\delta\lambda(f,x) \le \lambda(f,x) \le \delta.$$

Now, let $\lambda_k = \lambda(f,x_k)$. Due to the inequality above, it follows by induction that $\lambda_k \le \delta$ and that

$$\lambda_{k+1} \le K(\delta)\lambda_k^2$$

for all k, and the latter inequality in turn implies that

$$\lambda_k \le K(\delta)^{-1}\big(K(\delta)\lambda_0\big)^{2^k} \le (1-\delta)^2\big(K(\delta)\delta\big)^{2^k}.$$

Hence, λ_k tends to 0 as $k \to \infty$, because $K(\delta)\delta < 1$. By the remark following Theorem 16.4.6,

$$f(x_k) - f_{\min} \le \lambda_k^2,$$

if $\lambda_k \le \frac{1}{2}$, so we conclude that

$$\lim_{k\to\infty} f(x_k) = f_{\min}.$$

To prove the remaining error estimate, we can without loss of generalization assume that $\epsilon < \delta^2$, because if $\epsilon > \delta^2$ then already

$$f(x_0) - f_{\min} \le \lambda(f,x_0)^2 \le \delta^2 < \epsilon.$$

Let A and B be the constants defined by

$$A = -\log_2\big(-2\log_2(K(\delta)\delta)\big) \quad \text{and} \quad B = (1-\delta)^4.$$

Then $0 < B \le 1$, and $\log_2(\log_2 B/\epsilon)$ is a well-defined number, since $B/\epsilon \ge (1-\delta)^4/\delta^2 = (K(\delta)\delta)^{-2} > 1$. If $k > A + \log_2(\log_2 B/\epsilon)$, then

$$\lambda_k^2 \le (1-\delta)^4\big(K(\delta)\delta\big)^{2^{k+1}} < \epsilon,$$

and consequently $f(x_k) - f_{\min} \le \lambda_k^2 < \epsilon$.

If f is a non-degenerate function, then f has a unique minimum point $\hat{x}$, and it follow from inequality (16.11) in Theorem 16.4.6 that

$$\|x_k - \hat{x}\|_{\hat{x}} \le \frac{\lambda_k}{1-\lambda_k} \to 0, \quad \text{as } k \to \infty.$$

Since $\|\cdot\|_{\hat{x}}$ is a proper norm, this means that $x_k \to \hat{x}$. □

When $\delta = 1/3$, the values of the constants in Theorem 16.5.3 are $A = 0.268\ldots$ and $B = 16/81$, and $A + \log_2(\log_2 B/\epsilon) = 6.87$ for $\epsilon = 10^{-30}$. So with a starting point x_0 satisfying $\lambda(f, x_0) < 1/3$, Newton's algorithm will produce a function value that approximates the minimum value with an error less than 10^{-30} after at most 7 iterations.

Newton's method for self-concordant functions

By combining Newton's damped method with $1/(1+\lambda(f,x))$ as damping factor and Newton's pure method, we arrive at the following variant of Newton's method.

Newton's method

Given a positive number $\delta < \frac{1}{2}(3 - \sqrt{5})$, a starting point $x_0 \in X$, and a tolerance $\epsilon > 0$.

1. *Initiate:* $x := x_0$.
2. Compute the Newton decrement $\lambda = \lambda(f, x)$.
3. Go to line 8 if $\lambda < \delta$ else continue.
4. Compute a Newton direction Δx_{nt} at the point x.
5. *Update:* $x := x + (1 + \lambda)^{-1}\Delta x_{\text{nt}}$.
6. Go to line 2.
7. Compute the Newton decrement $\lambda = \lambda(f, x)$.
8. *Stopping criterion:* **stop** if $\lambda < \sqrt{\epsilon}$. x is an approximate optimal point.
9. Compute a Newton direction Δx_{nt} at the point x.
10. *Update:* $x := x + \Delta x_{\text{nt}}$.
11. Go to line 7.

Assuming that f is closed, self-concordant and downwards bounded, the damped phase of the algorithm, i.e. steps 2–6, continues during at most

$$\lfloor (f(x_0) - f_{\min})/\rho(-\delta) \rfloor$$

iterations, and the pure phase 7–11 ends according to Theorem 16.5.3 after at most $\lceil A + \log_2(\log_2 B/\epsilon) \rceil$ iterations. Therefore, we have the following result.

Theorem 16.5.4. *If the function f is closed, self-concordant and bounded below, then the above Newton method terminates at a point x satisfying $f(x) < f_{\min} + \epsilon$ after at most*

$$\lfloor (f(x_0) - f_{\min}))/\rho(-\delta) \rfloor + \lceil A + \log_2(\log_2 B/\epsilon) \rceil$$

iterations, where A and B are the constants of Theorem 16.5.3.

In particular, $1/\rho(-\delta) = 21.905$ when $\delta = 1/3$, and the second term can be replaced by the number 7 when $\epsilon \geq 10^{-30}$. Thus, at most

$$\lfloor 22(f(x_0) - f_{\min}) \rfloor + 7$$

iterations are required to find an approximation to the minimum value that meets all practical requirements by a wide margin.

Exercises

16.1 Show that the function $f(x) = x \ln x - \ln x$ is self-concordant on $\mathbf{R}_{++}$.

16.2 Suppose $f_i \colon X_i \to \mathbf{R}$ are self-concordant functions for $i = 1, 2, \ldots, m$, and let $X = X_1 \times X_2 \times \cdots \times X_m$. Prove that the function $f \colon X \to \mathbf{R}$, defined by

$$f(x_1, x_2, \ldots, x_m) = f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)$$

for $x = (x_1, x_2, \ldots, x_m) \in X$, is self-concordant.

16.3 Suppose that $f \colon \mathbf{R}_{++} \to \mathbf{R}$ is a three times continuously differentiable, convex function, and that

$$|f'''(x)| \leq 3\frac{f''(x)}{x} \qquad \text{for all } x.$$

a) Prove that the function

$$g(x) = -\ln(-f(x)) - \ln x,$$

with $\{x \in \mathbf{R}_{++} \mid f(x) < 0\}$ as domain, is self-concordant.

[Hint: Use that $3a^2b + 3a^2c + 2b^3 + 2c^3 \leq 2(a^2 + b^2 + c^2)^{3/2}$ if $a, b, c \geq 0$.]

b) Prove that the function

$$F(x, y) = -\ln(y - f(x)) - \ln x$$

is self-concordant on the set $\{(x, y) \in \mathbf{R}^2 \mid x > 0,\ \ y > f(x)\}$.

16.4 Show that the following functions f satisfy the conditions of the previous exercise:

a) $f(x) = -\ln x$ b) $f(x) = x \ln x$ c) $f(x) = -x^p$, where $0 < p \leq 1$.

16.5 Let us write x' for $(x_1, x_2, \ldots, x_{n-1})$ when $x = (x_1, x_2, \ldots, x_n)$, and let $\|\cdot\|$ denote the Euclidean norm in $\mathbf{R}^{n-1}$. Let $X = \{x \in \mathbf{R}^n \mid \|x'\| < x_n\}$, and define the function $f \colon X \to \mathbf{R}$ by $f(x) = -\ln(x_n^2 - \|x'\|^2)$. Prove that the following identity holds for all $v \in \mathbf{R}^n$:

$$D^2 f(x)[v, v] = \frac{1}{2}\big(Df(x)[v]\big)^2 + 2\frac{(x_n^2 - \|x'\|^2)(\|x'\|^2\|v'\|^2 - \langle x', v'\rangle^2) + (v_n\|x'\|^2 - x_n\langle x', v'\rangle)^2}{(x_n^2 - \|x'\|^2)^2\|x'\|^2},$$

and use it to conclude that f is a convex function and that $\lambda(f,x) = 2$ for all $x \in X$.

16.6 *Convergence for Newton's damped method.*
Suppose that the function $f\colon X \to \mathbf{R}$ is closed and self-concordant, and define for points $x \in X$ with finite Newton decrement the point x^+ by

$$x^+ = x + \frac{1}{1+\lambda(f,x)}\Delta x_{\mathrm{nt}},$$

where Δx_{nt} is a Newton direction at x.

a) Then x^+ is a point in X, according to Theorem 16.3.2. Show that

$$\lambda(f,x^+) \leq 2\lambda(f,x)^2,$$

and hence that $\lambda(f,x^+) \leq \lambda(f,x)$ if $\lambda(f,x) \leq \frac{1}{2}$.

b) Suppose x_0 is a point in X with Newton decrement $\lambda(f,x_0) \leq \frac{1}{4}$, and define the sequence $(x_k)_0^\infty$ recursively by $x_{k+1} = x_k^+$. Show that

$$f(x_k) - f_{\min} \leq \tfrac{1}{4} \cdot \left(\tfrac{1}{2}\right)^{2^{k+1}},$$

and hence that $f(x_k)$ converges quadratically to $f_{\min}$.

Appendix

We begin with a result on tri-linear forms which was needed in the proof of the fundamental inequality $\left|D^3 f(x)[u,v,w]\right| \leq 2\|u\|_x\|v\|_x\|w\|_x$ for self-concordant functions.

Fix an arbitrary scalar product $\langle\cdot\,,\cdot\rangle$ on $\mathbf{R}^n$ and let $\|\cdot\|$ denote the corresponding norm, i.e. $\|v\| = \langle v,v\rangle^{1/2}$. If $\phi(u,v,w)$ is a symmetric tri-linear form on $\mathbf{R}^n \times \mathbf{R}^n \times \mathbf{R}^n$, we define its norm $\|\phi\|$ by

$$\|\phi\| = \sup_{u,v,w\neq 0} \frac{|\phi(u,v,w)|}{\|u\|\|v\|\|w\|}.$$

The numerator and the denominator in the expression for $\|\phi\|$ are homogeneous of the same degree 3, hence

$$\|\phi\| = \sup_{(u,v,w)\in S^3} |\phi(u,v,w)|,$$

where S denotes the unit sphere in $\mathbf{R}^n$ with respect to the norm $\|\cdot\|$, i.e.

$$S = \{u \in \mathbf{R}^n \mid \|u\| = 1\}.$$

It follows from the norm definition that

$$|\phi(u,v,w)| \le \|\phi\|\|u\|\|v\|\|w\|$$

for all vectors u, v, w in $\mathbf{R}^n$.

Since tri-linear forms are continuous and the unit sphere is compact, the least upper bound $\|\phi\|$ is attained at some point $(u,v,w) \in S^3$, and we will show that the least upper bound is indeed attained at some point where $u = v = w$. This is the meaning of the following theorem.

Theorem 1. *Suppose that $\phi(u,v,w)$ is a symmetric tri-linear form. Then*

$$\|\phi\| = \sup_{u,v,w\neq 0} \frac{|\phi(u,v,w)|}{\|u\|\|v\|\|w\|} = \sup_{v\neq 0} \frac{|\phi(v,v,v)|}{\|v\|^3}.$$

Remark. The theorem is a special case of the corresponding result for symmetric m-multilinear forms, but we only need the case $m = 3$. The general case is proved by induction.

Proof. Let

$$\|\phi\|' = \sup_{v\neq 0} \frac{|\phi(v,v,v)|}{\|v\|^3} = \sup_{\|v\|=1} |\phi(v,v,v)|.$$

We claim that $\|\phi\| = \|\phi\|'$. Obviously, $\|\phi\|' \le \|\phi\|$, so we only have to prove the converse inequality $\|\phi\| \le \|\phi\|'$.

To prove this inequality, we need the corresponding result for symmetric bilinear forms $\psi(u,v)$. To such a form there is associated a symmetric linear operator (matrix) A such that $\psi(u,v) = \langle Au, v\rangle$, and if $e_1, e_2, \dots, e_n$ is an ON-basis of eigenvectors of A and $\lambda_1, \lambda_2, \dots, \lambda_n$ denote the corresponding eigenvalues with λ_1 as the one with the largest absolute value, and if $u, v \in S$ are vectors with coordinates $u_1, u_2, \dots, u_n$ and $v_1, v_2, \dots, v_n$ with respect to the given ON-basis, then

$$\begin{aligned} |\psi(u,v)| = |\sum_{i=1}^n \lambda_i u_i v_i| &\le \sum_{i=1}^n |\lambda_i||u_i||v_i| \le |\lambda_1| \sum_{i=1}^n |u_i||v_i| \\ &\le |\lambda_1| \Big(\sum_{i=1}^n u_i^2\Big)^{1/2} \Big(\sum_{i=1}^n v_i^2\Big)^{1/2} = |\lambda_1| = |\psi(e_1,e_1)|, \end{aligned}$$

which proves that $\sup_{(u,v)\in S^2} |\psi(u,v)| = \sup_{v\in S} |\psi(v,v)|$.

We now return to the tri-linear form $\phi(u,v,w)$. Let $(\hat{u},\hat{v},\hat{w})$ be a point in S^3 where the least upper bound defining $\|\phi\|$ is attained, i.e.

$$\|\phi\| = \phi(\hat{u},\hat{v},\hat{w}),$$

and consider the function

$$\psi(u,v) = \phi(u,v,\hat{w});$$

this is a symmetric bilinear form on $\mathbf{R}^n \times \mathbf{R}^n$ and

$$\sup_{(u,v)\in S^2} |\psi(u,v)| = \|\phi\|.$$

But as already proven,

$$\sup_{(u,v)\in S^2} |\psi(u,v)| = \sup_{v\in S} |\psi(v,v)|.$$

Therefore, we conclude that we can withour restriction assume that $\hat{u} = \hat{v}$.

We have in other words shown that the set

$$A = \{(v,w) \in S^2 \mid |\phi(v,v,w)| = \|\phi\|\}$$

is nonempty. The set A is a closed subset of S^2, and hence the number

$$\alpha = \max\{\langle v,w\rangle \mid (v,w) \in A\}$$

exists, and obviously $0 \leq \alpha \leq 1$.

Due to tri-linearity,

$$\phi(u+v,u+v,w) - \phi(u-v,u-v,w) = 4\phi(u,v,w).$$

So if u, v, w are arbitrary vectors in S, i.e. vectors with norm 1, then

$$\begin{aligned}
4|\phi(u,v,w)| &\leq |\phi(u+v,u+v,w)| + |\phi(u-v,u-v,w)| \\
&\leq |\phi(u+v,u+v,w)| + \|\phi\|\|u-v\|^2\|w\| \\
&= |\phi(u+v,u+v,w)| - \|\phi\|\|u+v\|^2 + \|\phi\|(\|u+v\|^2 + \|u-v\|^2) \\
&= |\phi(u+v,u+v,w)| - \|\phi\|\|u+v\|^2 + \|\phi\|(2\|u\|^2 + 2\|v\|^2) \\
&= |\phi(u+v,u+v,w)| - \|\phi\|\|u+v\|^2 + 4\|\phi\|.
\end{aligned}$$

Now choose $(\overline{v},\overline{w}) \in A$ such that $\langle \overline{v},\overline{w}\rangle = \alpha$. By the above inequality, we then have

$$\begin{aligned}
4\|\phi\| &= 4|\phi(\overline{v},\overline{v},\overline{w})| = 4|\phi(\overline{v},\overline{w},\overline{v})| \\
&\leq |\phi(\overline{v}+\overline{w},\overline{v}+\overline{w},\overline{v})| - \|\phi\|\|\overline{v}+\overline{w}\|^2 + 4\|\phi\|,
\end{aligned}$$

and it follows that

$$|\phi(\overline{v}+\overline{w},\overline{v}+\overline{w},\overline{v})| \geq \|\phi\|\|\overline{v}+\overline{w}\|^2.$$

Note that $\|\overline{v}+\overline{w}\|^2 = \|\overline{v}\|^2 + \|\overline{w}\|^2 + 2\langle\overline{v},\overline{w}\rangle = 2 + 2\alpha > 0$. Therefore, we can form the vector $\overline{z} = (\overline{v}+\overline{w})/\|\overline{v}+\overline{w}\|$ and write the above inequality as

$$|\phi(\overline{z},\overline{z},\overline{v})| \geq \|\phi\|,$$

which implies that

$$|\phi(\overline{z},\overline{z},\overline{v})| = \|\phi\| \tag{16.13}$$

since $\overline{z}$ and $\overline{v}$ are vectors in S. We conclude that the pair $(\overline{z},\overline{v})$ is an element of the set A, and hence

$$\alpha \geq \langle \overline{z},\overline{v}\rangle = \frac{\langle \overline{v},\overline{v}\rangle + \langle \overline{w},\overline{v}\rangle}{\|\overline{v}+\overline{w}\|} = \frac{1+\alpha}{\sqrt{2+2\alpha}} = \sqrt{\frac{1+\alpha}{2}}.$$

This inequality forces α to be greater than or equal to 1. Hence $\alpha = 1$ and

$$\langle \overline{z},\overline{v}\rangle = 1 = \|\overline{z}\|\|\overline{v}\|.$$

So Cauchy–Schwarz's inequality holds with equality in this case, and this implies that $\overline{z} = \overline{v}$. By inserting this in equality (16.13), we obtain the inequality

$$\|\phi\|' \geq \phi(\overline{v},\overline{v},\overline{v}) = \|\phi\|,$$

and the proof of the theorem is now complete. □

Our second result in this appendix is a uniqueness theorem for functions that satisfy a special differential inequality.

Theorem 2. *Suppose that the function $y(t)$ is continuously differentiable in the interval $I = [0,b[$, that $y(t) \geq 0$, $y(0) = 0$ and $y'(t) \leq Cy(t)^\alpha$ for some given constants $C > 0$ and $\alpha \geq 1$. Then, $y(t) = 0$ in the interval I.*

Proof. Let $a = \sup\{x \in I \mid y(t) = 0 \text{ for } 0 \leq t \leq x\}$. We will prove that $a = b$ by showing that the assumption $a < b$ gives rise to a contradiction.

By continuity, $y(a) = 0$. Choose a point $c \in]a,b[$ and let

$$M = \max\{y(t) \mid a \leq t \leq c\}.$$

Then choose a point d such that $a < d < c$ and $d - a \leq \frac{1}{2}C^{-1}M^{1-\alpha}$. The maximum of the function $y(t)$ on the interval $[a,d]$ is attained at some point e, and by the least upper bound definition of the point a, we have $y(e) > 0$. Of course, we also have $y(e) \leq M$, so it follows that

$$\begin{aligned} y(e) = y(e) - y(a) &= \int_a^e y'(t)\,dt \leq C\int_a^e y(t)^\alpha\,dt \\ &\leq C(d-a)y(e)^\alpha \leq C(d-a)M^{\alpha-1}y(e) \leq \frac{1}{2}y(e), \end{aligned}$$

which is a contradiction. □

Chapter 17

The path-following method

In this chapter, we describe a method for solving the optimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in X \end{array}$$

when X is a closed subset of $\mathbf{R}^n$ with nonempty interior and f is a continuous function which is differentiable in the interior of X. We assume throughout that $X = \mathrm{cl}(\mathrm{int}\, X)$. Pretty soon, we will restrict ourselves to convex problems, i.e. assume that X is a convex set and f is a convex function, in which case, of course, automatically $X = \mathrm{cl}(\mathrm{int}\, X)$ for all sets with nonempty interior.

Descent methods require that the function f is differentiable in a neighborhood of the optimal point, and if the optimal point lies on the boundary of X, then we have a problem. One way to attack this problem is to choose a function $F\colon \mathrm{int}\, X \to \mathbf{R}$ with the property that $F(x) \to +\infty$ as x goes to boundary of X and a parameter $\mu > 0$, and to minimize the function $f(x) + \mu F(x)$ over $\mathrm{int}\, X$. This function's minimum point $\hat{x}(\mu)$ lies in the interior of X, and since $f(x) + \mu F(x) \to f(x)$ as $\mu \to 0$, we can hope that the function value $f(\hat{x}(\mu))$ should be close to the minimum value of f, if the parameter μ is small enough. The function F acts as a barrier that prevents the approximating minimum point from lying on the boundary.

The function $\mu^{-1} f(x) + F(x)$ has of course the same minimum point $\hat{x}(\mu)$ as $f(x) + \mu F(x)$, and for technical reasons it works better to have the parameter in front of the objective function f than in front of the barrier function F. Henceforth, we will therefore instead, with new notation, examine what happens to the minimum point $\hat{x}(t)$ of the function $F_t(x) = tf(x) + F(x)$, when the parameter t tends to $+\infty$.

17.1 Barrier and central path

Barrier

We begin with the formal definition of a barrier.

Definition. Let X be a closed convex set with nonempty interior. A *barrier* to the set X is a differentiable function $F\colon \operatorname{int} X \to \mathbf{R}$ with the property that $\lim_{k\to\infty} F(x_k) = +\infty$ for all sequences $(x_k)_1^\infty$ that converge to a boundary point of X.

If a barrier function has a unique minimum point, then this point is called the *analytic center* of the set X (with respect to the barrier).

Remark 1. A convex function with an open domain goes to ∞ at the boundary if and only if it is a closed function. Hence, if $F\colon \operatorname{int} X \to \mathbf{R}$ is convex and differentiable, then F is a barrier to X if and only if F is closed.

Remark 2. A strictly convex barrier function to a compact convex set has a unique minimum point in the interior of the set. So compact convex sets with nonempty interiors have analytic centers with respect to strictly convex barriers.

Now, let F be a barrier to the closed convex set X, and suppose that we want to minimize a given function $f\colon X \to \mathbf{R}$. For each real number $t \geq 0$ we define the function $F_t\colon \operatorname{int} X \to \mathbf{R}$ by

$$F_t(x) = tf(x) + F(x).$$

In particular, $F_0 = F$. The following theorem is the basis for barrier-based interior-point methods for minimization.

Theorem 17.1.1. *Suppose that $f\colon X \to \mathbf{R}$ is a continuous function, and let F be a downwards bounded barrier to the set X. Suppose that the functions F_t have minimum points $\hat{x}(t)$ in the interior of X for each $t > 0$. Then,*

$$\lim_{t\to+\infty} f(\hat{x}(t)) = \inf_{x\in X} f(x).$$

Proof. Let $v_{\min} = \inf_{x\in X} f(x)$ and $M = \inf_{x\in\operatorname{int} X} F(x)$. (We do not exclude the possibility that $v_{\min} = -\infty$, but M is of course a finite number.)

Choose, given $\eta > v_{\min}$, a point $x^* \in \operatorname{int} X$ such that $f(x^*) < \eta$. Then

$$\begin{aligned} v_{\min} &\leq f(\hat{x}(t)) \leq f(\hat{x}(t)) + t^{-1}(F(\hat{x}(t)) - M) = t^{-1}\big(F_t(\hat{x}(t)) - M\big) \\ &\leq t^{-1}\big(F_t(x^*) - M\big) = f(x^*) + t^{-1}(F(x^*) - M). \end{aligned}$$

Since the right hand side of this inequality tends to $f(x^*)$ as $t \to +\infty$, it follows that $v_{\min} \leq f(\hat{x}(t)) < \eta$ for all sufficiently large numbers t, and this proves the theorem. □

In order to use the barrier method, one needs of course an appropriate barrier to the given set. For sets of the type

$$X = \{x \in \Omega \mid g_i(x) \leq 0, \;\; i = 1, 2, \ldots, m\}$$

we will use the *logarithmic barrier function*

$$F(x) = -\sum_{i=1}^{m} \ln(-g_i(x)). \tag{17.1}$$

Note that the barrier function F is convex if all functions $g_i \colon \Omega \to \mathbf{R}$ are convex. In this case, X is a convex set, and the interior of X is nonempty if Slater's condition is satisfied, i.e. if there is a point $\overline{x} \in \Omega$ such that $g_i(\overline{x}) < 0$ for all i.

Other examples of barriers are the exponential barrier function

$$F(x) = \sum_{i=1}^{m} \mathrm{e}^{-1/g_i(x)}$$

and the power function barriers

$$F(x) = \sum_{i=1}^{m} (-g_i(x))^{-p},$$

where $p > 0$.

Central path

Definition. Let F be a barrier to the set X and suppose that the functions F_t have unique minimum points $\hat{x}(t) \in \operatorname{int} X$ for all $t \geq 0$. The curve $\{\hat{x}(t) \mid t \geq 0\}$ is called the *central path* for the problem $\min_{x \in X} f(x)$.

Note that $\hat{x}(0)$ is the analytic center of X with respect to the barrier F, so the central path starts at the analytic center.

Since the gradient is zero at an optimal point, we have

$$tf'(\hat{x}(t)) + F'(\hat{x}(t)) = 0 \tag{17.2}$$

for all points on the central path. The converse is true if the objective function f and the barrier function F are convex, i.e. $\hat{x}(t)$ is a point on the central path if and only if equation (17.2) is satisfied.

The logarithmic barrier F to $X = \{x \in \Omega \mid g_i(x) \leq 0, \;\; i = 1, 2, \ldots, m\}$ has derivative

$$F'(x) = -\sum_{i=1}^{m} \frac{1}{g_i(x)} g_i'(x),$$

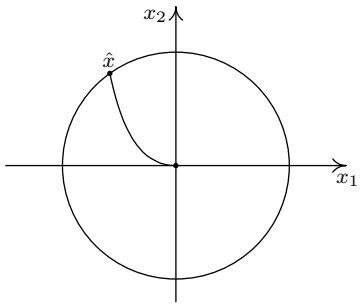

Figure 17.1. The central path associated with the problem of minimizing the function $f(x) = x_1 e^{x_1+x_2}$ over $X = \{x \in \mathbf{R}^2 \mid x_1^2 + x_2^2 \leq 1\}$ with barrier function $F(x) = (1 - x_1^2 - x_2^2)^{-1}$. The minimum point is $\hat{x} = (-0.5825, 0.8128)$.

so the central path equation (17.2) has in this case the following form for $t > 0$:

$$f'(\hat{x}(t)) - \frac{1}{t}\sum_{i=1}^{m} \frac{1}{g_i(\hat{x}(t))} g_i'(\hat{x}(t)) = 0. \tag{17.3}$$

Let us now consider a convex optimization problem of the following type:

$$\text{(P)} \qquad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & g_i(x) \leq 0, \quad i = 1, 2, \ldots, m \end{array}$$

We assume that Slater's condition is satisfied and that the problem has an optimal solution $\hat{x}$.

The corresponding Lagrange function L is given by

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x),$$

and it follows from equation (17.3) that $L_x'(\hat{x}(t), \hat{\lambda}) = 0$, if $\hat{\lambda} \in \mathbf{R}_+^m$ is the vector defined by

$$\hat{\lambda}_i = -\frac{1}{t g_i(\hat{x}(t))}.$$

Since the Lagrange function is convex in the variable x, we conclude that $\hat{x}(t)$ is a minimum point for the function $L(\cdot\,, \hat{\lambda})$. The value at $\hat{\lambda}$ of the dual function $\phi\colon \mathbf{R}_+^m \to \mathbf{R}$ to our minimization problem (P) is therefore by definition

$$\phi(\hat{\lambda}) = L(\hat{x}(t), \hat{\lambda}) = f(\hat{x}(t)) - m/t.$$

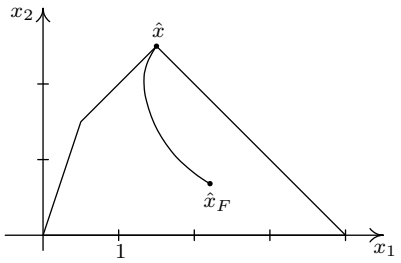

Figure 17.2. The central path for the LP problem $\min_{x\in X} 2x_1 - 3x_2$ with $X = \{x \in \mathbf{R}^2 \mid x_2 \geq 0,\, x_2 \leq 3x_1,\, x_2 \leq x_1 + 1,\, x_1 + x_2 \leq 4\}$ and logarithmic barrier. The point $\hat{x}_F$ is the analytic center of X, and $\hat{x} = (1.5, 2.5)$ is the optimal solution.

By weak duality, $\phi(\hat{\lambda}) \leq f(\hat{x})$, so it follows that

$$f(\hat{x}(t)) - m/t \leq f(\hat{x}).$$

We have thus arrived at the following approximation theorem, which for convex problems with logarithmic barrier provides more precise information than Theorem 17.1.1.

Theorem 17.1.2. *The points $\hat{x}(t)$ on the central path for the convex minimization problem (P) with optimal solution $\hat{x}$ and logarithmic barrier satisfy the inequality*

$$f(\hat{x}(t)) - f(\hat{x}) \leq \frac{m}{t}.$$

Note that the estimate of the theorem depends on the number of constraints but not on the dimension.

17.2 Path-following methods

A strategy for determining the optimal value of the convex optimization problem

$$\text{(P)} \qquad \begin{aligned} &\min \;\; f(x) \\ &\text{s.t.} \;\; g_i(x) \leq 0, \;\; i = 1, 2, \ldots, m \end{aligned}$$

for twice continuously differentiable objective and constraint functions with an error that is less than or equal to ϵ, would in light of Theorem 17.1.2 be to solve the optimization problem $\min F_t(x)$ with logarithmic barrier F for $t = m/\epsilon$, using for example Newton's method. The strategy works for small problems and with moderate demands on accuracy, but better results

are obtained by solving the problems min $F_t(x)$ for an increasing sequence of t-values until $t \geq m/\epsilon$.

A simple version of the barrier method or the *path-following method*, as it is also called, therefore looks like this:

Path-following method

Given a starting point $x = x_0 \in \operatorname{int} X$, a real number $t = t_0 > 0$, an update parameter $\alpha > 1$ and a tolerance $\epsilon > 0$.

Repeat

1. Compute $\hat{x}(t)$ by minimizing $F_t = tf + F$ with x as starting point
2. *Update:* $x := \hat{x}(t)$.
3. *Stopping criterion:* **stop** if $m/t \leq \epsilon$.
4. *Increase* t*:* $t := \alpha t$.

Step 1 is called an *outer iteration* or a *centering step* because it is about finding a point on the central path. To minimize the function F_t, Newton's method is used, and the iterations of Newton's method to compute $\hat{x}(t)$ with x as the starting point are called *inner iterations*.

It is not necessary to compute $\hat{x}(t)$ exactly in the outer iterations; the central path serves no other function than to lead to the optimal point $\hat{x}$, and good approximations to points on the central path will also give rise to a sequence of points which converges to $\hat{x}$.

The computational cost of the method obviously depends on the total number of outer iterations that have to be performed before the stopping criterion is met, and on the number of inner iterations in each outer iteration.

The update parameter α

The parameter α (and the initial value t_0) determines the number of outer iterations required to reach the stopping criterion $t \geq m/\epsilon$. If α is small, i.e. close to 1, then many outer iterations are needed, but on the other hand, each outer iteration requires few inner iterations since the minimum point $x = \hat{x}(t)$ of the function F_t is then a very good starting point in Newton's algorithm for the problem of minimizing the function $F_{\alpha t}$.

For large α values the opposite is true; few outer iterations are needed, but each outer iteration now requires more Newton steps as the starting point $\hat{x}(t)$ is farther from the minimum point $\hat{x}(\alpha t)$.

From experience, it turns out, however, that the above two effects tend to offset each other. The total number of Newton steps is roughly constant over a wide range of α, and values of α between 10 and 20 usually work well.

The initial value t_0

The choice of the starting value t_0 is also significant. A small value requires many outer iterations before the stopping criterion is met. A large value, on the other hand, requires many inner iterations in the first outer iteration before a sufficiently good approximation to the point $\hat{x}(t_0)$ on the central path has been found. Since $f(\hat{x}(t_0)) - f(\hat{x}) \approx m/t_0$, it may be reasonable to choose t_0 so that m/t_0 is of the same magnitude as $f(x_0) - f(\hat{x})$. The problem, of course, is that the optimal value $f(\hat{x})$ is not known a priori, so it is necessary to use a suitable estimate. If, for example, a feasible point λ for the dual problem is known and ϕ is the dual function, then $\phi(\lambda)$ can be used as an approximation of $f(\hat{x})$, and $t_0 = m/(f(x_0) - \phi(\lambda))$ can be taken as initial t-value.

The starting point x_0

The starting point x_0 must lie in the interior of X, i.e. it has to satisfy all constraints with strict inequality. If such a point is not known in advance, one can use the barrier method on an artificial problem to compute such a point, or to conclude that the original problem has no feasible points. The procedure is called *phase 1* of the path-following method and works as follows.

Consider the inequalities

$$g_i(x) \leq 0, \quad i = 1, 2, \ldots, m \tag{17.4}$$

and suppose that the functions $g_i \colon \Omega \to \mathbf{R}$ are convex and twice continuously differentiable. To determine a point that satisfies all inequalities strictly or to determine that there is no such point, we form the optimization problem

$$\begin{array}{ll} \min & s \\ \text{s.t.} & g_i(x) \leq s, \quad i = 1, 2, \ldots, m \end{array} \tag{17.5}$$

in the variables x and s. This problem has strictly feasible points, because we can first choose $x_0 \in \Omega$ arbitrarily and then choose $s_0 > \max_i g_i(x_0)$, and we obtain in this way a point $(x_0, s_0) \in \Omega \times \mathbf{R}$ that satisfies the constraints with strict inequalities. The functions $(x, s) \mapsto g_i(x) - s$ are obviously convex. We can therefore use the path-following method on the problem (17.5), and depending on the sign of the problem's optimal value $v_{\min}$, we get three cases.

$v_{\min} < 0$: The system (17.4) has strictly feasible solutions. Indeed, if (x, s) is a feasible point for the problem (17.5) with $s < 0$, then $g_i(x) < 0$ for all i. This means that it is not necessary to solve the optimization problem (17.5) with great accuracy. The algorithm can be stopped as soon as it has generated a point (x, s) with $s < 0$.

$v_{\min} > 0$: The system (17.4) is infeasible. Also in this case, it is not necessary to solve the problem with great accuracy. We can stop as soon as we have found a feasible point for the dual problem with a positive value of the dual function, since this implies that $v_{\min} > 0$.

$v_{\min} = 0$: If the greatest lower bound $v_{\min} = 0$ is attained, i.e. if there is a point $(\hat{x}, \hat{s})$ with $\hat{s} = 0$, then the system (17.4) is feasible but not strictly feasible. The system (17.4) is infeasible if $v_{\min}$ is not attained. In practice, it is of course impossible to determine exactly that $v_{\min} = 0$; the algorithm terminates with the conclusion that $|v_{\min}| < \epsilon$ for some small positive number ϵ, and we can only be sure that the system $g_i(x) < -\epsilon$ is infeasible and that the system $g_i(x) \leq \epsilon$ is feasible.

Convergence analysis

At the beginning of outer iteration number k, we have $t = \alpha^{k-1} t_0$. The stopping criterion will be triggered as soon as $m/(\alpha^{k-1} t_0) \leq \epsilon$, i.e. when $k - 1 \geq (\log(m/(\epsilon t_0))/\log \alpha$. The number of outer iterations is thus equal to

$$\left\lceil \frac{\log(m/(\epsilon t_0)}{\log \alpha} \right\rceil + 1$$

(for $\epsilon \leq m/t_0$).

The path-following method therefore works, provided that the minimization problems

$$\begin{array}{ll} \min & tf(x) + F(x) \\ \text{s.t.} & x \in \operatorname{int} X \end{array} \tag{17.6}$$

can be solved for $t \geq t_0$. Using Newton's method, this is true, for example, if the objective functions satisfy the conditions of Theorem 15.2.4, i.e. if F_t is strongly convex, has a Lipschitz continuous derivative and the sublevel set corresponding to the starting point is closed.

A question that remains to be resolved is whether the problem (17.6) gets harder and harder, that is requires more innner iterations, when t grows. Practical experience shows that this is not so − in most problems, the number of Newton steps seems to be roughly constant when t grows. For problems with self-concordant objective and barrier functions, it is possible to obtain exact estimates of the total number of iterations needed to solve the optimization problem (P) with a given accuracy, and this will be the theme in Chapter 18.

Chapter 18

The path-following method with self-concordant barrier

18.1 Self-concordant barriers

Definition. Let X be a closed convex subset of $\mathbf{R}^n$ with nonempty interior $\operatorname{int} X$, and let ν be a nonnegative number. A function $f\colon \operatorname{int} X \to \mathbf{R}$ is called a *self-concordant barrier to X with parameter ν*, or shorter a *ν-self-concordant barrier*, if the function is closed, self-concordant and non-constant, and the Newton decrement satisfies the inequality

$$\lambda(f,x) \le \nu^{1/2} \tag{18.1}$$

for all $x \in \operatorname{int} X$.

It follows from Theorem 15.1.2 and Theorem 15.1.3 that inequality (18.1) holds if and only if

$$|\langle f'(x), v\rangle| \le \nu^{1/2}\|v\|_x$$

for all vectors $v \in \mathbf{R}^n$, or equivalently, if and only if

$$\big(Df(x)[v]\big)^2 \le \nu\, D^2 f(x)[v,v]$$

for all $v \in \mathbf{R}^n$.

A closed self-concordant function $f\colon \Omega \to \mathbf{R}$ with the property that $\sup_{x\in\Omega} \lambda(f,x) < 1$ is necessarily constant and the domain Ω is equal to $\mathbf{R}^n$, according to Theorem 16.4.7. The parameter ν of a self-concordant barrier must thus be greater than or equal to 1.

Example 18.1.1. The function $f(x) = -\ln x$ is a 1-self-concordant barrier to the interval $[0,\infty[$, because f is closed and self-concordant and $\lambda(f,x) = 1$ for all $x > 0$. □

EXAMPLE 18.1.2. Convex quadratic functions

$$f(x) = \tfrac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c$$

are self-concordant on $\mathbf{R}^n$, but they do not function as self-concordant barriers, because $\sup \lambda(f, x) = \infty$ for all non-constant convex quadratic functions f, according to Example 15.1.2. □

We will show later that only subsets of halfspaces can have self-concordant barriers, so there is no self-concordant barrier to the whole $\mathbf{R}^n$.

EXAMPLE 18.1.3. Let $g(x)$ be a non-constant convex, quadratic function. The function f, defined by

$$f(x) = -\ln(-g(x)),$$

is a 1-self-concordant barrier to the set $X = \{x \in \mathbf{R}^n \mid g(x) \leq 0\}$.

Proof. Let $g(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c$, let v be an arbitrary vector in $\mathbf{R}^n$, and set

$$\alpha = -\frac{1}{g(x)}Dg(x)[v] \quad \text{and} \quad \beta = -\frac{1}{g(x)}D^2g(x)[v,v] = -\frac{1}{g(x)}\langle v, Av\rangle,$$

where x is an arbitrary point in the interior of X. Note that $\beta \geq 0$ and that $D^3g(x)[v,v,v] = 0$. It therefore follows from the differentiation rules that

$$\begin{aligned}
Df(x)[v] &= -\frac{1}{g(x)}Dg(x)[v] = \alpha,\\
D^2f(x)[v,v] &= \frac{1}{g(x)^2}\big(Dg(x)[v]\big)^2 - \frac{1}{g(x)}D^2g(x)[v,v] = \alpha^2 + \beta \geq 0,\\
D^3f(x)[v,v,v] &= -\frac{2}{g(x)^3}\big(Dg(x)[v]\big)^3 + \frac{3}{g(x)^2}D^2g(x)[v,v]Dg(x)[v]\\
&\quad - \frac{1}{g(x)}D^3g(x)[v,v,v] = 2\alpha^3 + 3\alpha\beta.
\end{aligned}$$

The function f is convex since its second derivative is positive semidefinite, and it is closed since $f(x) \to +\infty$ as $g(x) \to 0$. By squaring it is easy to show that the inequality $|2\alpha^3 + 3\alpha\beta| \leq 2(\alpha^2 + \beta)^{3/2}$ holds for all $\alpha \in \mathbf{R}$ and all $\beta \in \mathbf{R}_+$, and obviously $\alpha^2 \leq \alpha^2 + \beta$. This means that $\big|D^3f(x)[v,v,v]\big| \leq 2\big(D^2f(x)[v,v]\big)^{3/2}$ and that $(Df(x)[v])^2 \leq D^2f(x)[v,v]$. So f is 1-self-concordant. □

The following three theorems show how to build new self-concordant barriers from given ones.

Theorem 18.1.1. *If f is a ν-self-concordant barrier to the set X and $\alpha \geq 1$, then αf is an $\alpha\nu$-self-concordant barrier to X.*

Proof. The proof is left as a simple exercise. □

Theorem 18.1.2. *If f is a μ-self-concordant barrier to the set X and g is a ν-self-concordant barrier to the set Y, then the sum $f+g$ is a self-concordant barrier with parameter $\mu+\nu$ to the intersection $X \cap Y$. And $f+c$ is a μ-self-concordant barrier to X for each constant c.*

Proof. The sum $f+g$ is a closed convex function, and it is self-concordant on the set $\operatorname{int}(X \cap Y)$ according to Theorem 16.1.5. To prove that the sum is a self-concordant barrier with parameter $(\mu+\nu)$, we assume that v is an arbitrary vector in $\mathbf{R}^n$ and write $a = D^2 f(x)[v,v]$ and $b = D^2 g(x)[v,v]$. We then have, by definition,

$$\big(Df(x)[v]\big)^2 \leq \mu a \quad \text{and} \quad \big(Dg(x)[v]\big)^2 \leq \nu b,$$

and using the inequality $2\sqrt{\mu\nu ab} \leq \nu a + \mu b$ between the geometric and the arithmetic mean, we obtain the inequality

$$\begin{aligned}\big(D(f+g)(x)[v]\big)^2 &= \big(Df(x)[v]\big)^2 + \big(Dg(x)[v]\big)^2 + 2Df(x)[v]\cdot Dg(x)[v] \\ &\leq \mu a + \nu b + 2\sqrt{\mu a \nu b} \leq \mu a + \nu b + \nu a + \mu b \\ &= (\mu+\nu)(a+b) = (\mu+\nu)\, D^2(f+g)(x)[v,v],\end{aligned}$$

which means that $\lambda(f+g,x) \leq (\mu+\nu)^{1/2}$.

The assertion about the sum $f+c$ is trivial, since $\lambda(f,x) = \lambda(f+c,x)$ for constants c. □

Theorem 18.1.3. *Suppose that $A\colon \mathbf{R}^m \to \mathbf{R}^n$ is an affine map and that f is a ν-self-concordant barrier to the subset X of $\mathbf{R}^n$. The composition $g = f \circ A$ is then a ν-self-concordant barrier to the inverse image $A^{-1}(X)$.*

Proof. The proof is left as an exercise. □

Example 18.1.4. It follows from Example 18.1.1 and Theorems 18.1.2 and 18.1.3 that the function

$$f(x) = -\sum_{i=1}^{m} \ln(b_i - \langle a_i, x\rangle)$$

is an m-self-concordant barrier to the polyhedron

$$X = \{x \in \mathbf{R}^n \mid \langle a_i, x\rangle \leq b_i,\ \ i = 1,2,\dots,m\}.$$ □

Theorem 18.1.4. *If f is a ν-self-concordant barrier to the set X, then*

$$\langle f'(x), y - x\rangle \leq \nu$$

for all $x \in \operatorname{int} X$ and all $y \in X$.

Remark. It follows that a set with a self-concordant barrier must be a subset of some halfspace. Indeed, a set X with a ν-self-concordant barrier is a subset of the closed halfspace $\{y \in \mathbf{R}^n \mid \langle c, y\rangle \leq \nu + \langle c, x_0\rangle\}$, where $x_0 \in \operatorname{int} X$ is an arbitrary point with $c = f'(x_0) \neq 0$.

Proof. Fix $x \in \operatorname{int} X$ and $y \in X$, let $x^t = x + t(y - x)$ and define the function ϕ by setting $\phi(t) = f(x^t)$. Then ϕ is certainly defined on the open interval $]\alpha, 1[$ for some negative number α, since x is an iterior point. Moreover,

$$\phi'(t) = Df(x^t)[y - x],$$

and especially, $\phi'(0) = Df(x)[y - x] = \langle f'(x), y - x\rangle$. We will prove that $\phi'(0) \leq \nu$.

If $\phi'(0) \leq 0$, then we are done, so assume that $\phi'(0) > 0$. By ν-self-concordance,

$$\phi''(t) = D^2 f(x^t)[y - x, y - x] \geq \nu^{-1}\big(Df(x^t)[y - x]\big)^2 = \nu^{-1}\phi'(t)^2 \geq 0.$$

The derivative ϕ' is thus increasing, and this implies that $\phi'(t) \geq \phi'(0) > 0$ for $t \geq 0$. Furthermore,

$$\frac{d}{dt}\Big(-\frac{1}{\phi'(t)}\Big) = \frac{\phi''(t)}{\phi'(t)^2} \geq \frac{1}{\nu}$$

for all t in the interval $[0, 1[$, so by integrating the last mentioned inequality over the interval $[0, \beta]$, where $\beta < 1$, we obtain the inequality

$$\frac{1}{\phi'(0)} > \frac{1}{\phi'(0)} - \frac{1}{\phi'(\beta)} = \int_0^\beta \frac{d}{dt}\Big(-\frac{1}{\phi'(t)}\Big)\, dt \geq \frac{\beta}{\nu}.$$

Hence, $\phi'(0) < \nu/\beta$ for all $\beta < 1$, which implies that $\phi'(0) \leq \nu$. □

Theorem 18.1.5. *Suppose that f is a ν-self-concordant barrier to the set X. If $x \in \operatorname{int} X$, $y \in X$ and $\langle f'(x), y - x\rangle \geq 0$, then*

$$\|y - x\|_x \leq \nu + 2\sqrt{\nu}.$$

Remark. If $x \in \operatorname{int} X$ is a minimum point, then $\langle f'(x), y - x\rangle = 0$ for all points $y \in X$, since $f'(x) = 0$. Hence, $\|y - x\|_x \leq \nu + 2\sqrt{\nu}$ for all $y \in X$ if x is a minimum point.

Proof. Let $r = \|y-x\|_x$. If $r \le \sqrt{\nu}$, then there is nothing to prove, so assume that $r > \sqrt{\nu}$, and consider for $\alpha = \sqrt{v}/r$ the point $z = x + \alpha(y-x)$, which lies in the interior of X since $\alpha < 1$. By using Theorem 18.1.4 with z instead of x, the assumption $\langle f'(x), y-x\rangle \ge 0$, Theorem 16.3.2 and the equalities $y - z = (1-\alpha)(y-x)$ and $z - x = \alpha(y-x)$, we obtain the following chain of inequalities and equalities:

$$\begin{aligned}\nu &\ge \langle f'(z), y-z\rangle = (1-\alpha)\langle f'(z), y-x\rangle \ge (1-\alpha)\langle f'(z) - f'(x), y-x\rangle \\ &= \frac{1-\alpha}{\alpha}\langle f'(z) - f'(x), z-x\rangle \ge \frac{1-\alpha}{\alpha}\cdot\frac{\|z-x\|_x^2}{1+\|z-x\|_x} \\ &= \frac{(1-\alpha)\alpha\|y-x\|_x^2}{1+\alpha\|y-x\|_x} = \frac{r\sqrt{\nu}-\nu}{1+\sqrt{\nu}}.\end{aligned}$$

The inequality between the extreme ends simplifies to $r \le \nu + 2\sqrt{\nu}$, which is the desired inequality. □

Given a self-concordant funktion f with the corresponding local seminorm $\|\cdot\|_x$, we set

$$\mathcal{E}(x;r) = \{y \in \mathbf{R}^n \mid \|y-x\|_x \le r\}.$$

If f is non-degenerate, then $\|\cdot\|_x$ is a norm at each point $x \in \operatorname{int} X$, and the set $\mathcal{E}(x;r)$ is a closed ellipsoid in $\mathbf{R}^n$ with axis directions determined by the eigenvectors of the second derivative $f''(x)$.

For non-degenerate self-concordant barriers we now have the following corollary to Theorem 18.1.5.

Theorem 18.1.6. *Suppose that f is a non-degenerate ν-self-concordant barrier to the closed convex set X. Then f attains a minimum if and only if X is a bounded set. The minimum point $\hat{x}_f \in \operatorname{int} X$ is unique in that case, and*

$$\mathcal{E}(\hat{x}_f;1) \subseteq X \subseteq \mathcal{E}(\hat{x}_f;\nu+2\sqrt{\nu}).$$

Remark. A closed self-concordant function whose domain does not contain any line, is automatically non-degenerate, so it is not necessary to state explicitly that a self-concordant barrier to a compact set should be non-degenerate.

Proof. The sublevel sets of a closed convex function are closed, so if X is a bounded set, then each sublevel set $\{x \in \operatorname{int} X \mid f(x) \le \alpha\}$ is both closed and bounded, and this implies that f has a minimum, and the minimum point of a non-degenerate convex function is necessarily unique.

Conversely, assume that f has a minimum point $\hat{x}_f$. Then by the remark following Theorem 18.1.5, $\|y - \hat{x}_f\|_{\hat{x}_f} \le \nu + 2\sqrt{\nu}$ for all $y \in X$, and this

amounts to the right inclusion in Theorem 18.1.6, which implies, of course, that X is a bounded set.

The remaining left inclusion follows from Theorem 16.3.2, which implies that the open ellipsoid $\{y \in \mathbf{R}^n \mid \|y - x\|_x < 1\}$ is a subset of $\operatorname{int} X$ for each choice of $x \in \operatorname{int} X$. The closure $\mathcal{E}(x; 1)$ is therefore a subset of X, and we obtain the left inclusion by choosing $x = \hat{x}_f$. □

Given a self-concordant barrier to a set X we will need to compare the local seminorms $\|v\|_x$ and $\|v\|_y$ of a vector at different points x and y, and in order to achieve this we need a measure for the distance from y to x relative the distance from y to the boundary of X along the half-line from x through x. The following definition provides us with the relevant measure.

Definition. Let X be a closed convex subset of $\mathbf{R}^n$ with nonempty interior. For each $y \in \operatorname{int} X$ we define a function $\pi_y \colon \mathbf{R}^n \to \mathbf{R}_+$ by setting

$$\pi_y(x) = \inf\{t > 0 \mid y + t^{-1}(x - y) \in X\}.$$

Obviously, $\pi_y(y) = 0$. To determine $\pi_y(x)$ if $x \neq y$, we consider the half-line from y through x; if the half-line intersects the boundary of X in a point z, then $\pi_y(x) = \|x - y\|/\|z - y\|$ (with respect to arbitrary norms), and if the entire half-line lies in X, then $\pi_y(x) = 0$. We note that $\pi_y(x) < 1$ for interior points x, that $\pi_y(x) = 1$ for boundary points x, and that $\pi_y(x) > 1$ for points outside X.

We could also have defined the function π_y in terms of the Minkowski functional that was introduced in Section 6.10, because

$$\pi_y(x) = \phi_{-y+X}(x - y),$$

where ϕ_{-y+X} is the Minkowski functional of the set $-y + X$.

The following simple estimate of $\pi_y(x)$ will be needed later on.

Theorem 18.1.7. *Let X be a compact convex set, let x and y be points in the interior of X, and suppose that*

$$B(x, r) \subseteq X \subseteq \overline{B}(0; R),$$

where the balls are given with respect to an arbitrary norm $\|\cdot\|$. Then

$$\pi_y(x) \leq \frac{2R}{2R + r}.$$

Proof. The inequality is trivially true if $x = y$, so suppose that $x \neq y$. The half-line from y through x intersects the boundary of X in a point z and $\|z - y\| = \|z - x\| + \|x - y\|$. Furthermore, $\|z - x\| \geq r$ and $\|x - y\| \leq 2R$, so it follows that

$$\pi_y(x) = \frac{\|x - y\|}{\|z - y\|} = \Big(1 + \frac{\|z - x\|}{\|x - y\|}\Big)^{-1} \leq \Big(1 + \frac{r}{2R}\Big)^{-1} = \frac{2R}{2R + r}. \qquad \square$$

The direction derivative $\langle f'(x), v\rangle$ of a ν-self-concordant barrier function f is bounded by $\sqrt{\nu}\|v\|_x$, by definition. Our next theorem shows that the same direction derivative is also bounded by a constant times $\|v\|_y$, if y is an arbitrary point in the domain of f. The two local norms $\|v\|_x$ and $\|v\|_y$ are also compared.

Theorem 18.1.8. *Let f be a ν-self-concordant barrier to X, and let x and y be two points in the interior of X. Then, for all vectors v*

$$|\langle f'(x), v\rangle| \leq \frac{\nu}{1-\pi_y(x)}\,\|v\|_y \tag{18.2}$$

and

$$\|v\|_x \leq \frac{\nu + 2\sqrt{\nu}}{1-\pi_y(x)}\|v\|_y. \tag{18.3}$$

Proof. The two inequalities hold if $y = x$ since

$$|\langle f'(x), v\rangle| \leq \sqrt{\nu}\|v\|_x \leq \nu\|v\|_x$$

and $\pi_x(x) = 0$. They also hold if $\|v\|_y = 0$, i.e. if the vector v belongs to the recessive subspace of f, because then $\|v\|_x = 0$ and $\langle f'(x), v\rangle = 0$. Assume henceforth that $y \neq x$ and that $\|v\|_y \neq 0$.

First consider the case $\|v\|_y = 1$, and let s be an arbitrary number greater than $\nu + 2\sqrt{\nu}$. Then, by Theorems 16.3.2 and 18.1.5, we conclude that

(i) The two points $y \pm v$ lie in X.

(ii) At least one of the two points $x \pm \dfrac{s}{\|v\|_x}v$ lies outside X.

By the definition of $\pi_y(x)$ there is a vector $z \in X$ such that

$$x = y + \pi_y(x)(z - y),$$

and since

$$x \pm (1-\pi_y(x))v = \pi_y(x)z + (1-\pi_y(x))(y \pm v),$$

it follows from convexity that

(iii) The two points $x \pm (1-\pi_y(x))v$ lie in X.

It now follows from (iii) and Theorem 18.1.4 that

$$\langle f'(x), \pm v\rangle = \frac{1}{1-\pi_y(x)}\langle f'(x), x \pm (1-\pi_y(x))v - x\rangle \leq \frac{\nu}{1-\pi_y(x)},$$

which means that

$$|\langle f'(x), v\rangle| \leq \frac{\nu}{1-\pi_y(x)}.$$

This proves inequality (18.2) for vectors v with $\|v\|_y = 1$, and if v is an arbitrary vector with $\|v\|_y \neq 0$, we obtain inequality (18.2) by replacing v in the inequality above with $v/\|v\|_y$.

By combining the two assertions (ii) and (iii) we conclude that

$$1 - \pi_y(x) < \frac{s}{\|v\|_x},$$

i.e. that

$$\|v\|_x < \frac{s}{1 - \pi_y(x)} = \frac{s}{1 - \pi_y(x)}\|v\|_y,$$

and since this holds for all $s > \nu + 2\sqrt{\nu}$, it follows that

$$\|v\|_x \leq \frac{\nu + 2\sqrt{\nu}}{1 - \pi_y(x)}\|v\|_y.$$

This proves inequality (18.3) in the case $\|v\|_y = 1$, and since the inequality is homogeneous, it holds in general. □

Definition. Let $\|\cdot\|_x$ be the local seminorm at x which is associated with the two times differentiable convex function $f\colon X \to \mathbf{R}$, where X is a subset of $\mathbf{R}^n$. The corresponding *dual local norm* is the function $\|\cdot\|_x^*\colon \mathbf{R}^n \to \overline{\mathbf{R}}$, which is defined by

$$\|v\|_x^* = \sup_{\|w\|_x \leq 1} \langle v, w\rangle$$

for all $v \in \mathbf{R}^n$.

The dual norm is easily verified to be subadditive and homogeneous, i.e. $\|v + w\|_x^* \leq \|v\|_x^* + \|w\|_x^*$ and $\|\lambda v\|_x^* = |\lambda|\|v\|_x^*$ for all $v, w \in \mathbf{R}^n$ and all real numbers λ, but $\|\cdot\|_x^*$ is a proper norm on the whole of $\mathbf{R}^n$ only for points x where the second derivative $f''(x)$ is positive definite, because $\|v\|_x^* = \infty$ if v is a nonzero vector in the null space $\mathcal{N}(f''(x))$ since $\|tv\|_x = 0$ for all $t \in \mathbf{R}$ and $\langle v, tv\rangle = t\|v\|^2 \to \infty$ as $t \to \infty$. However, $\|\cdot\|_x^*$ is always a proper norm when restricted to the subspace $\mathcal{N}(f''(x))^\perp$. See exercise 18.2.

By Theorem 15.1.3, we have the following expression for the Newton decrement $\lambda(f, x)$ in terms of the dual local norm:

$$\lambda(f, x) = \|f'(x)\|_x^*.$$

The following variant of the Cauchy–Schwarz inequality holds för the local seminorm.

Theorem 18.1.9. *Assume that* $\|v\|_x^* < \infty$. *Then*

$$|\langle v, w\rangle| \leq \|v\|_x^*\|w\|_x$$

for all vectors w.

Proof. If $\|w\|_x \neq 0$ then $\pm w/\|w\|_x$ are two vectors with local seminorm equal to 1, so it follows from the definition of the dual norm that

$$\pm\frac{1}{\|w\|_x}\langle v, w\rangle = \langle v, \pm w/\|w\|_x\rangle \leq \|v\|_x^*,$$

and we obtain the sought inequality after multiplication by $\|w\|_x$.

If instead $\|w\|_x = 0$, then $\|tw\|_x = 0$ for all real numbers t, and it follows from the supremum definition that $t\langle v, w\rangle = \langle v, tw\rangle \leq \|v\|_x^* < \infty$ for all t. This being possible only if $\langle v, w\rangle = 0$, we conclude that the inequality applies in this case, too. □

Later we will need various estimates of $\|v\|_x^*$. Our first estimate is in terms of the width in different directions of the set X, and this motivates our next definition.

Definition. Given a nonempty subset X of $\mathbf{R}^n$, let $\mathrm{Var}_X\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ be the function defined by

$$\mathrm{Var}_X(v) = \sup_{x\in X}\langle v, x\rangle - \inf_{x\in X}\langle v, x\rangle.$$

$\mathrm{Var}_X(v)$ is obviously a finite number for each $v \in \mathbf{R}^n$ if the set X is bounded, and if v is a unit vector, then $\mathrm{Var}_X(v)$ measures the width of the set X in the direction of v.

Our next theorem shows how to estimate $\|\cdot\|_x^*$ using Var_X.

Theorem 18.1.10. *Suppose that $f\colon X \to \mathbf{R}$ is a closed self-concordant function with a bounded open convex subset X of $\mathbf{R}^n$ as domain, and let $\|\cdot\|_x^*$ be the dual local norm associated with the function f at the point $x \in X$. Then*

$$\|v\|_x^* \leq \mathrm{Var}_X(v)$$

for all $v \in \mathbf{R}^n$.

Proof. It follows from the previous theorem that y is a point in $\mathrm{cl}\, X$ if x is a point in X and $\|y - x\|_x \leq 1$. Hence,

$$\begin{aligned}\|v\|_x^* &= \sup_{\|w\|_x\leq 1}\langle v, w\rangle = \sup_{\|y-x\|_x\leq 1}\langle v, y - x\rangle \leq \sup_{y\in \mathrm{cl}\, X}\langle v, y - x\rangle = \sup_{y\in X}\langle v, y - x\rangle \\ &= \sup_{y\in X}\langle v, y\rangle - \langle v, x\rangle \leq \sup_{y\in X}\langle v, y\rangle - \inf_{y\in X}\langle v, y\rangle = \mathrm{Var}_X(v).\end{aligned}$$ □

We have previously defined the analytic center of a closed convex set X with respect to a given barrier as the unique minimum point of the barrrier,

provided that there is one. According to Theorem 18.1.6, every compact convex set with nonempty interior has an analytic center with respect to any given ν-self-concordant barrier. We can now obtain an upper bound on the dual local norm $\|v\|_x^*$ at an arbitrary point x in terms of the parameter ν and the value of the dual norm at the analytic center.

Theorem 18.1.11. *Let X be a compact convex set, and let $\hat{x}_f$ be the analytic center of the set with respect to a ν-self-concordant barrier f. Then, for each vector $v \in \mathbf{R}^n$ and each $x \in \operatorname{int} X$,*

$$\|v\|_x^* \leq (\nu + 2\sqrt{\nu})\|v\|_{\hat{x}_f}^*.$$

Proof. Let $B_1 = \mathcal{E}(x; 1)$ and $B_2 = \mathcal{E}(\hat{x}_f; \nu + 2\sqrt{\nu})$. Theorems 16.3.2 and 18.1.6 give us the inclusions $B_1 \subseteq X \subseteq B_2$, so it follows from the definition of the dual local norm that

$$\begin{aligned}\|v\|_x^* &= \sup_{\|w\|_x \leq 1} \langle v, w\rangle = \sup_{y \in B_1} \langle v, y - x\rangle \leq \sup_{y \in B_2} \langle v, y - x\rangle \\ &= \langle v, \hat{x}_f - x\rangle + \sup_{y \in B_2} \langle v, y - \hat{x}_f\rangle = \langle v, \hat{x}_f - x\rangle + \sup_{\|w\|_{\hat{x}_f} \leq \nu + 2\sqrt{\nu}} \langle v, w\rangle \\ &= \langle v, \hat{x}_f - x\rangle + (\nu + 2\sqrt{\nu})\|v\|_{\hat{x}_f}^*.\end{aligned}$$

Since $\|-v\|_x^* = \|v\|_x^*$, we may now without loss of generality assume that $\langle v, \hat{x}_f - x\rangle \leq 0$, and this gives us the required inequality. □

18.2 The path-following method

Standard form

Let us say that a convex optimization problem is in *standard form* if it is presented in the form

$$\begin{aligned}&\min && \langle c, x\rangle \\ &\text{s.t.} && x \in X\end{aligned}$$

where X is a compact convex set with nonempty interior and X is equipped with a ν-self-concordant barrier function F.

Remark. One can show that every compact convex set X has a barrier function, but for a barrier function to be useful in a practical optimization problem, it has to be explicitly given so that it is possible to efficiently calculate its partial first and second derivatives.

The assumption that the set X is bounded is not particularly restrictive for problems with finite optimal values, for we can always modify such problems by adding artificial, very big bounds on the variables.

We also recall that an arbitrary convex problem can be transformed into an equivalent convex problem with a linear objective function by an epigraph formulation. (See Chapter 9.3.)

EXAMPLE 18.2.1. Each LP problem with finite optimal value can be written in standard form after suitable transformations. By first identifying the affine hull of the polyhedron of feasible points with $\mathbf{R}^n$ for an appropriate n, we can without restriction assume that the polyhedron has a nonempty interior, and by adding big bounds on the variables, if necessary, we can also assume that our polyhedron X of feasible points is compact. And with X written in the form

$$X = \{x \in \mathbf{R}^n \mid \langle c_i, x\rangle \leq b_i,\ i = 1, 2, \ldots, m\}, \tag{18.4}$$

we get an m-self-concordant barrier F to X, by defining

$$F(x) = -\sum_{i=1}^{m} \ln(b_i - \langle c_i, x\rangle)$$ □

EXAMPLE 18.2.2. Convex quadratic optimization problems, i.e. problems of the type

$$\begin{array}{ll} \min & g(x) \\ \text{s.t.} & x \in X \end{array}$$

where g is a convex quadratic function and X is a bounded polyhedron in $\mathbf{R}^n$ with nonempty interior, can be transformed, using an epigraph formulation and an artificial bound M on the new variable s, to problems of the form

$$\begin{array}{ll} \min & s \\ \text{s.t.} & (x, s) \in Y \end{array}$$

where $Y = \{(x, s) \in \mathbf{R}^n \times \mathbf{R} \mid x \in X,\ g(x) \leq s \leq M\}$ is a compact convex set with nonempty interior. Now assume that the polyhedron X is given by equation (18.4) as an intersection of closed halfspaces. Then the function

$$F(x, s) = -\sum_{i=1}^{m} \ln(b_i - \langle c_i, x\rangle) - \ln(s - g(x)) - \ln(M - s)$$

is an $(m + 2)$-self-concordant barrier to Y according to Example 18.1.3. □

Central path

We will now study the path-following method for the standard problem

$$\text{(SP)} \qquad \begin{array}{ll} \min & \langle c, x\rangle \\ \text{s.t.} & x \in X \end{array}$$

where X is a compact convex subset of $\mathbf{R}^n$ with nonempty interior, and F is a ν-self-concordant barrier to X. The finite optimal value of the problem is denoted by $v_{\min}$.

For $t \geq 0$ we define functions $F_t \colon \operatorname{int} X \to \mathbf{R}$ by

$$F_t(x) = t\langle c, x\rangle + F(x).$$

The functions F_t are closed and self-concordant, and since the set X is compact, each function F_t has a unique minimum point $\hat{x}(t)$. The central path $\{\hat{x}(t) \mid t \geq 0\}$ is in other words well-defined, and its points satisfy the equation

$$tc + F'(\hat{x}(t)) = 0, \tag{18.5}$$

and the starting point $\hat{x}(0)$ is by definition the analytic center $\hat{x}_F$ of X with respect to the given barrier F.

We will use Newton's method to determine the minimum point $\hat{x}(t)$, and for that reason we need to calculate the Newton step and the Newton decrement with respect to the function F_t at points in the interior of X.

Since $F_t''(x) = F''(x)$, the local norm $\|v\|_x$ of a vector v with respect to the function F_t is the same for all $t \geq 0$, namely

$$\|v\|_x = \sqrt{\langle v, F''(x)v\rangle}.$$

In contrast, Newton steps and Newton decrements depend on t; the Newton step at the point x is equal to $-F''(x)^{-1}F_t'(x)$ for the function F_t, and the decrement is given by

$$\lambda(F_t, x) = \sqrt{\langle F_t'(x), F''(x)^{-1}F_t'(x)\rangle} = \|F''(x)^{-1}F_t'(x)\|_x.$$

The following theorem is used to formulate the stopping criterion in the path-following method.

Theorem 18.2.1. *(i) The points $\hat{x}(t)$ on the central path of the optimization problem (SP) satisfy the inequality*

$$\langle c, \hat{x}(t)\rangle - v_{\min} \leq \frac{\nu}{t}.$$

(ii) Moreover, the inequality

$$\langle c, x\rangle - v_{\min} \leq \frac{\nu + \kappa(1-\kappa)^{-1}\sqrt{\nu}}{t}.$$

holds for $t > 0$ *and all point* $x \in \operatorname{int} X$ *satisfying the condition*

$$\lambda(F_t, x) \leq \kappa < 1.$$

Proof. (i) Because of equation (18.5), $c = -t^{-1}F'(\hat{x}(t))$, and it therefore follows from Theorem 18.1.4 that

$$\langle c, \hat{x}(t)\rangle - \langle c, y\rangle = \frac{1}{t}\langle F'(\hat{x}(t)), y - \hat{x}(t)\rangle \leq \frac{\nu}{t}$$

for all $y \in X$. We obtain inequality (i) by choosing y as an optimal solution to the problem (SP).

(ii) Since $\langle c, x\rangle - v_{\min} = (\langle c, x\rangle - \langle c, \hat{x}(t)\rangle) + (\langle c, \hat{x}(t)\rangle - v_{\min})$, it suffices, due to the already proven inequality, to show that

$$\langle c, x\rangle - \langle c, \hat{x}(t)\rangle \leq \frac{\kappa}{1-\kappa} \cdot \frac{\sqrt{\nu}}{t} \tag{18.6}$$

if $x \in \operatorname{int} X$ and $\lambda(F_t, x) \leq \kappa < 1$. But it follows from Theorem 16.4.6 that

$$\|x - \hat{x}(t)\|_{\hat{x}(t)} \leq \frac{\lambda(F_t, x)}{1 - \lambda(F_t, x)} \leq \frac{\kappa}{1-\kappa},$$

so by using that $tc = -F'(\hat{x}(t))$ and that F is ν-self-concordant, we get the following chain of equalities and inequalities:

$$\begin{aligned} t(\langle c, x\rangle - \langle c, \hat{x}(t)\rangle) &= -\langle F'(\hat{x}(t)), x - \hat{x}(t)\rangle \leq \|F'(\hat{x}(t))\|^*_{\hat{x}(t)}\|x - \hat{x}(t)\|_{\hat{x}(t)} \\ &= \lambda(F, \hat{x}(t))\|x - \hat{x}(t)\|_{\hat{x}(t)} \leq \sqrt{\nu}\frac{\kappa}{1-\kappa}, \end{aligned}$$

which proves inequality (18.6). □

Algorithm

The path-following algorithm for solving the standard problem

$$\text{(SP)} \qquad \begin{array}{ll} \min & \langle c, x\rangle \\ \text{s.t.} & x \in X \end{array}$$

works in brief as follows.

We start with a parameter value $t_0 > 0$ and a point $x_0 \in \operatorname{int} X$, which is close enough to the point $\hat{x}(t_0)$ on the central path. "Close enough" is

expressed in terms of the Newton decrement $\lambda(F_{t_0}, x_0)$, which must be sufficiently small.

Then we update the parameter t by defining $t_1 = \alpha t_0$ for a suitable $\alpha > 1$ and minimize the function F_{t_1} using the damped Newton method with x_0 as the starting point. Newton's method is terminated when it has reached a point x_1, which is sufficiently close to the minimum point $\hat{x}(t_1)$ of F_{t_1}.

The procedure is then repeated with $t_2 = \alpha t_1$ as new parameter and with x_1 as starting point in Newton's method for minimization of the function F_{t_2}, etc. As a result we obtain a sequence t_0, x_0, t_1, x_1, t_2, x_2, ... of parameter values and points, and the procedure is terminated when t_k has become sufficiently large with x_k as an approximate optimal point.

From this sketchy description of the algorithm it is clear that we need two parameters, one parameter α to describe the update of t, and one parameter κ to define the stopping criterion in Newton's method. We shall estimate the total number of inner iterations, and the estimate will be the simplest and most obvious if one writes the update parameter α in the form $\alpha = 1+\gamma/\sqrt{\nu}$.

The following precise formulation of the path-following algorithm therefore contains the parameters γ and κ. The addition 'phase 2' is due to the need for an additional phase to generate feasible initial values x_0 and t_0.

Path-following algorithm, phase 2

Given an update parameter $\gamma > 0$, a neighborhood parameter $0 < \kappa < 1$, a tolerance $\epsilon > 0$, a starting point $x_0 \in \operatorname{int} X$, and a starting value $t_0 > 0$ such that $\lambda(F_{t_0}, x_0) \leq \kappa$.

1. *Initiate:* $x := x_0$ and $t := t_0$.
2. *Stopping criterion:* **stop** if $\epsilon t \geq \nu + \kappa(1-\kappa)^{-1}\sqrt{\nu}$.
3. *Increase* t*:* $t := (1 + \gamma/\sqrt{\nu})t$.
4. *Update* x *by using Newton's damped method on the function* F_t *with the current* x *as starting point:*
   (i) Compute the Newton decrement $\lambda = \lambda(F_t, x)$.
   (ii) **quit** Newton's method if $\lambda \leq \kappa$, and go to line 2.
   (iii) Compute the Newtonstep $\Delta x_{\mathrm{nt}} = -F''(x)^{-1}F'_t(x)$.
   (iv) *Uppdate:* $x := x + (1+\lambda)^{-1}\Delta x_{\mathrm{nt}}$
   (v) Go to (i).

We can now show the following convergence result.

Theorem 18.2.2. *Suppose that the above path-following algorithm is applied to the standard problem (SP) with a* ν*-self-concordant barrier* F*. Then the algorithm stops with a point* $x \in \operatorname{int} X$ *which satisfies*

$$\langle c, x\rangle - v_{\min} \leq \epsilon.$$

For each outer iteration, the number of inner iterations in Newton's algorithm is bounded by a constant K, and the total number of inner iterations in the path-following algorithm is bounded by

$$C\sqrt{\nu}\,\ln\Big(\frac{\nu}{t_0\epsilon}+1\Big),$$

where the constants K and C only depend on κ and γ.

Proof. Let us start by examining the inner loop 4 of the algorithm.

Each time the algorithm passes by line 2, it does so with a point x in $\operatorname{int} X$, which belongs to a t-value with Newton decrement $\lambda(F_t,x)\le\kappa$. In step 4, the function F_s, where $s=(1+\gamma/\sqrt{\nu})t$, is then minimized using Newton's damped method with $y_0=x$ as the starting point. The points y_k, $k=1,2,3,\dots$, generated by the method lie in $\operatorname{int} X$ according to Theorem 16.3.2, and the stopping condition $\lambda(F_s,y_k)\le\kappa$ implies, according to Theorem 16.5.1, that the algorithm terminates after at most $\big\lfloor\big(F_s(x)-F_s(\hat{x}(s))\big)/\rho(-\kappa)\big\rfloor$ iterations, where ρ is the function

$$\rho(u)=-u-\ln(1-u).$$

We shall show that there is a constant K, which only depends on the parameters κ and γ, so that

$$\left\lfloor\frac{F_s(x)-F_s(\hat{x}(s))}{\rho(-\kappa)}\right\rfloor\le K,$$

and for that reason we need to estimate the difference $F_s(x)-F_s(\hat{x}(s))$, which we do in the next lemma.

Lemma 18.2.3. *Suppose that $\lambda(F_t,x)\le\kappa<1$. Then, for all $s>0$*

$$F_s(x)-F_s(\hat{x}(s))\le\rho(\kappa)+\frac{\kappa\sqrt{\nu}}{1-\kappa}\cdot\Big|\frac{s}{t}-1\Big|+\nu\,\rho(1-s/t).$$

Proof of the lemma. We start by writing

$$F_s(x)-F_s(\hat{x}(s))=\big(F_s(x)-F_s(\hat{x}(t))\big)+\big(F_s(\hat{x}(t))-F_s(\hat{x}(s))\big). \tag{18.7}$$

By using the equality $tc=-F'(\hat{x}(t))$ and the inequality

$$|\langle F'(\hat{x}(t)),v\rangle|\le\lambda(F,\hat{x}(t))\|v\|_{\hat{x}(t)}\le\sqrt{\nu}\|v\|_{\hat{x}(t)},$$

we obtain the following estimate of the first difference in the right-hand side of (18.7):

$$\begin{aligned}F_s(x)-F_s(\hat{x}(t))&=F_t(x)-F_t(\hat{x}(t))+(s-t)\langle c,x-\hat{x}(t)\rangle\\&=F_t(x)-F_t(\hat{x}(t))-(s/t-1)\langle F'(\hat{x}(t)),x-\hat{x}(t)\rangle\\&\le F_t(x)-F_t(\hat{x}(t))+|s/t-1|\,\sqrt{\nu}\,\|x-\hat{x}(t)\|_{\hat{x}(t)}.\end{aligned} \tag{18.8}$$

By Theorem 16.4.6,

$$F_t(x) - F_t(\hat{x}(t)) \le \rho(\lambda(F_t, x)) \le \rho(\kappa)$$

and

$$\|x - \hat{x}(t)\|_{\hat{x}(t)} \le \frac{\lambda(F_t, x)}{1 - \lambda(F_t, x)} \le \frac{\kappa}{1 - \kappa}.$$

Therefore, it follows from inequality (18.8) that

$$F_s(x) - F_s(\hat{x}(t)) \le \rho(\kappa) + \Big|\frac{s}{t} - 1\Big| \cdot \frac{\kappa\sqrt{\nu}}{1 - \kappa}. \tag{18.9}$$

It remains to estimate the second difference

$$\begin{aligned} \phi(s) &= F_s(\hat{x}(t)) - F_s(\hat{x}(s)) \\ &= s\langle c, \hat{x}(t)\rangle - s\langle c, \hat{x}(s)\rangle + F(\hat{x}(t)) - F(\hat{x}(s)) \end{aligned} \tag{18.10}$$

in the right-hand side of (18.7).

The function $\hat{x}(s)$ is continuously differentiable. This follows from the implicit function theorem, because $\hat{x}(s)$ satisfies the equation

$$sc + F'(\hat{x}(s)) = 0,$$

and the second derivative $F''(x)$ is continuous and non-singular everywhere. By implicit differentiation,

$$c + F''(\hat{x}(s))\hat{x}'(s) = 0,$$

which means that

$$\hat{x}'(s) = -F''(\hat{x}(s))^{-1}c.$$

It now follows from equation (18.10) that the difference $\phi(s)$ is continuously differentiable with derivative

$$\begin{aligned} \phi'(s) &= \langle c, \hat{x}(t)\rangle - \langle c, \hat{x}(s)\rangle - s\langle c, \hat{x}'(s)\rangle - \langle F'(\hat{x}(s), \hat{x}'(s)\rangle \\ &= \langle c, \hat{x}(t) - \hat{x}(s)\rangle - s\langle c, \hat{x}'(s)\rangle + s\langle c, \hat{x}'(s)\rangle \\ &= \langle c, \hat{x}(t) - \hat{x}(s)\rangle, \end{aligned}$$

and a further differentiation gives

$$\begin{aligned} \phi''(s) &= -\langle c, \hat{x}'(s)\rangle = \langle c, F''(\hat{x}(s))^{-1}c\rangle \\ &= \langle s^{-1}F'(\hat{x}(s)), s^{-1}F''(\hat{x}(s))^{-1}F'(\hat{x}(s))\rangle \\ &= s^{-2}\langle F'(\hat{x}(s)), F''(\hat{x}(s))^{-1}F'(\hat{x}(s))\rangle = s^{-2}\lambda(F, \hat{x}(s))^2 \le \nu s^{-2}. \end{aligned}$$

Now note that $\phi(t) = \phi'(t) = 0$. By integrating the inequality for $\phi''(s)$ over the interval $[t, u]$, we therefore obtain the following estimate for $u \geq t$:

$$\phi'(u) = \phi'(u) - \phi'(t) \leq \int_t^u \nu s^{-2}\, ds = \nu(t^{-1} - u^{-1}).$$

Integrating once more over the interval $[t, s]$ results in the inequality

$$\begin{aligned} F_s(\hat{x}(t)) - F_s(\hat{x}(s)) = \phi(s) &= \int_t^s \phi'(u)\, du \leq \nu \int_t^s (t^{-1} - u^{-1})\, du \\ &= \nu\big(\frac{s}{t} - 1 - \ln\frac{s}{t}\big) = \nu\,\rho(1 - s/t) \end{aligned} \tag{18.11}$$

for $s \geq t$. The same conclusion is also reached for $s < t$ by first integrating the inequality for $\phi''(s)$ over the interval $[u, t]$, and then the resulting inequality for $\phi'(u)$ over the interval $[s, t]$.

The inequality in the lemma is now finally a consequence of equation (18.7) and the estimates (18.9) and (18.11). □

Continuation of the proof of Theorem 18.2.2. By using the lemma's estimate of the difference $F_s(x) - F_s(\hat{x}(s))$ when $s = (1 + \gamma/\sqrt{\nu})t$, we obtain the inequality

$$\left\lfloor \frac{F_s(x) - F_s(\hat{x}(s))}{\rho(-\kappa)} \right\rfloor \leq \left\lfloor \frac{\rho(\kappa) + \gamma\kappa(1-\kappa)^{-1} + \nu\,\rho(-\gamma\nu^{-1/2})}{\rho(-\kappa)} \right\rfloor,$$

and $\nu\,\rho(-\gamma\nu^{-1/2}) \leq \frac{1}{2}\gamma^2$, because $\rho(u) = -u - \ln(1-u) \leq \frac{1}{2}u^2$ for $u < 0$. The number of inner iterations in each outer iteration is therefore bounded by the constant

$$K = \left\lfloor \frac{\rho(\kappa) + \gamma\kappa(1-\kappa)^{-1} + \frac{1}{2}\gamma^2}{\rho(-\kappa)} \right\rfloor,$$

which only depends on the parameters κ and γ. For example, $K = 5$ if $\kappa = 0.4$ and $\gamma = 0.32$.

We now turn to the number of outer iterations. Set

$$\beta(\kappa) = \nu + \kappa(1-\kappa)^{-1}\sqrt{\nu}.$$

Suppose that the stopping condition $\epsilon t \geq \beta(\kappa)$ is triggered during iteration number k when $t = (1 + \gamma/\sqrt{\nu})^k t_0$. Because of Theorem 18.2.1, the current point x then satisfies the condition

$$\langle c, x\rangle - v_{\min} \leq \epsilon,$$

which shows that x approximates the minimum point with prescribed accuracy.

Since k is the least integer satisfying the inequality $(1 + \gamma/\sqrt{\nu})^k \geq \beta(\kappa)/t_0\epsilon$, we have

$$k = \left\lceil \frac{\ln(\beta(\kappa)/t_0\epsilon)}{\ln(1+\gamma/\sqrt{\nu})} \right\rceil.$$

To simplify the denominator, we use the fact that $\ln(1 + \gamma x)$ is a concave function. This implies that $\ln(1 + \gamma x) \geq x \ln(1 + \gamma)$ if $0 \leq x \leq 1$, and hence

$$\ln(1 + \gamma/\sqrt{\nu}) \geq \frac{\ln(1+\gamma)}{\sqrt{\nu}}.$$

Furthermore, $\beta(\kappa) = \nu + \kappa(1-\kappa)^{-1}\sqrt{\nu} \leq \nu + \kappa(1-\kappa)^{-1}\nu = (1-\kappa)^{-1}\nu$. This gives us the estimate

$$k \leq \left\lceil \frac{\sqrt{\nu}\ln((1-\kappa)^{-1}\nu/t_0\epsilon)}{\ln(1+\gamma)} \right\rceil \leq K'\sqrt{\nu}\ln\Big(\frac{\nu}{t_0\epsilon} + 1\Big)$$

for the number of outer iterations with a constant K' that only depends on κ and γ, and by multiplying this with the constant K we obtain the corresponding estimate for the total number of inner iterations. □

Phase 1

In order to use the path-following algorithm, we need a $t_0 > 0$ and a point $x_0 \in \operatorname{int} X$ with Newton decrement $\lambda(F_{t_0}, x_0) \leq \kappa$ to start from. Since the central path begins in the analytic center $\hat{x}_F$ of X and $\lambda(F, \hat{x}_F) = 0$, it can be expected that (x_0, t_0) is good enough as a starting pair if only x_0 is close enough to $\hat{x}_F$ and $t_0 > 0$ is sufficiently small. Indeed, this is true, and we shall show that one can generate such a pair by solving an artificial problem, given that one knows a point $\overline{x} \in \operatorname{int} X$.

Therefore, let $G_t\colon \operatorname{int} X \to \mathbf{R}$, where $0 \leq t \leq 1$, be the functions defined by

$$G_t(x) = -t\langle F'(\overline{x}), x\rangle + F(x).$$

The functions G_t are closed and self-concordant, and they have unique minimum points $\overline{x}(t)$.

Note that $G_0 = F$, and hence $\overline{x}(0) = \hat{x}_F$. Since $G_t'(x) = -tF'(\overline{x}) + F'(x)$, $G_1'(\overline{x}) = 0$, and this means that $\overline{x}$ is the minimum point of the function G_1. Hence, $\overline{x}(1) = \overline{x}$. The curve $\{\overline{x}(t) \mid 0 \leq t \leq 1\}$ thus starts in the analytic center $\hat{x}_F$ and ends in the given point $\overline{x}$. By using the path-following method, now following the curve *backwards*, we will therefore obtain a suitable starting point for phase 2 of the algorithm.

We use Newton's damped method to minimize G_t and note that $G_t'' = F''$ for all t, so the local norm with respect to the function G_t coincides with the local norm with respect to the function F, and we can thus unambiguously use the symbol $\|\cdot\|_x$ for the local norm at the point x.

The algorithm for determining a starting pair (x_0, t_0) now looks like this.

Path-following algorithm, phase 1

Given $\overline{x} \in \operatorname{int} X$, and parameters $0 < \gamma < \frac{1}{2}\sqrt{\nu}$ and $0 < \kappa < 1$.

1. *Initiate:* $x := \overline{x}$ and $t := 1$.
2. *Stopping criterion:* **stop** if $\lambda(F, x) < \frac{3}{4}\kappa$ and set $x_0 = x$.
3. *Decrease* t*:* $t := (1 - \gamma/\sqrt{\nu})t$.
4. *Update* x *by using Newton's damped method on the function* G_t *with the current* x *as starting point:*
   (i) Compute $\lambda = \lambda(G_t, x)$.
   (ii) **quit** Newton's method if $\lambda \le \kappa/2$, and go to line 2.
   (iii) Compute the Newton step $\Delta x_{\mathrm{nt}} = -F''(x)^{-1} G_t'(x)$.
   (iv) *Update:* $x := x + (1 + \lambda)^{-1} \Delta x_{\mathrm{nt}}$.
   (v) Go to (i).

When the algorithm has stopped with a point x_0, we define t_0 by setting

$$t_0 = \max\{t \mid \lambda(F_t, x_0) \le \kappa\}.$$

The number of iterations in phase 1 is given by the following theorem.

Theorem 18.2.4. *Phase 1 of the path-following algorithm stops with a point* $x_0 \in \operatorname{int} X$ *after at most*

$$C\sqrt{\nu} \ln\Big(\frac{\nu}{1 - \pi_{\hat{x}_F}(\overline{x})} + 1\Big)$$

inner iterations, where the constant C *only depends on* κ *and* γ*, the number* t_0 *satisfies the conditions* $\lambda(F_{t_0}, x_0) \le \kappa$ *and* $t_0 \ge \kappa/4 \operatorname{Var}_X(c)$.

Proof. We start by estimating the number of inner iterations in each outer iteration; this number is bounded by the quotient

$$\frac{G_s(x) - G_s(\overline{x}(s))}{\rho(-\kappa/2)},$$

where $s = (1 - \gamma/\sqrt{\nu})t$, and Lemma 18.2.3 gives us the majorant

$$\rho(\kappa/2) + \frac{\kappa\sqrt{\nu}}{2 - \kappa} \cdot \frac{\gamma}{\sqrt{\nu}} + \nu\, \rho(\gamma/\sqrt{\nu})$$

for the numerator of the quotient. By Lemma 16.3.1, $\nu\rho(\gamma/\sqrt{\nu}) \le \gamma^2$, so the number of inner iterations in each outer iteration is bounded by the constant

$$\frac{\rho(\kappa/2) + \kappa(2-\kappa)^{-1}\gamma + \gamma^2}{\rho(-\kappa/2)}.$$

We now consider the outer iterations. Since $F' = G'_t + tF'(\overline{x})$,

$$\begin{aligned}(18.12)\quad \lambda(F,x) &= \|F'(x)\|_x^* = \|G'_t(x) + tF'(\overline{x})\|_x^* \le \|G'_t(x)\|_x^* + t\|F'(\overline{x})\|_x^* \\ &= \lambda(G_t,x) + t\|F'(\overline{x})\|_x^*.\end{aligned}$$

It follows from Theorem 18.1.11 that

$$\|F'(\overline{x})\|_x^* \le (\nu + 2\sqrt{\nu})\|F'(\overline{x})\|_{\hat{x}_F}^* \le 3\nu\|F'(\overline{x})\|_{\hat{x}_F}^*,$$

and from Theorem 18.1.8 that

$$\|F'(\overline{x})\|_{\hat{x}_F}^* = \sup_{\|v\|_{\hat{x}_F} \le 1} \langle F'(\overline{x}), v\rangle \le \frac{\nu}{1 - \pi_{\hat{x}_F}(\overline{x})}.$$

Hence

$$(18.13)\qquad \|F'(\overline{x})\|_x^* \le \frac{3\nu^2}{1 - \pi_{\hat{x}_F}(\overline{x})}.$$

During outer interation number k, we have $t = (1 - \gamma/\sqrt{\nu})^k$ and the point x satisfies the condition $\lambda(G_t, x) \le \kappa/2$ when Newton's method stops. So it follows from inequality (18.12) and the estimate (18.13) that the stopping condition $\lambda(F,x) < \frac{3}{4}\kappa$ in line 2 of the algorithm is fulfilled if

$$\frac{1}{2}\kappa + \frac{3\nu^2}{1 - \pi_{\hat{x}_F}(\overline{x})}(1 - \gamma/\sqrt{\nu})^k \le \frac{3}{4}\kappa,$$

i.e. if

$$k\ln(1 - \gamma/\sqrt{\nu}) < -\ln\Big(\frac{12\kappa^{-1}\nu^2}{1 - \pi_{\hat{x}_F}(\overline{x}}\Big).$$

By using the inequality $\ln(1-x) \le -x$, which holds for $0 < x < 1$, we see that the stopping condition is fulfilled for

$$k > \frac{\sqrt{\nu}}{\gamma}\ln\Big(\frac{12\kappa^{-1}\nu^2}{1 - \pi_{\hat{x}_F}(\overline{x}}\Big).$$

So the number of outer iterations is less than

$$K\sqrt{\nu}\ln\Big(\frac{\nu}{1 - \pi_{\hat{x}_F}(\overline{x})} + 1\Big),$$

where the constant K only depends on κ and γ, and this proves the estimate of the theorem, since the number of inner iterations in each outer iteration is bounded by a constant, which only depends on κ and γ.

The definition of t_0 implies that $\kappa = \lambda(F_{t_0}, x_0)$, so we get the following inequalities with the aid of Theorem 18.1.10:

$$\begin{aligned}\kappa = \lambda(F_t, x_0) &= \|F_t'(x_0)\|_{x_0}^* = \|t_0 c + F'(x_0)\|_{x_0}^* \le t_0\|c\|_{x_0}^* + \|F'(x_0)\|_{x_0}^* \\ &= t_0\|c\|_{x_0}^* + \lambda(F, x_0) \le t_0 \operatorname{Var}_X(c) + \frac{3}{4}\kappa.\end{aligned}$$

It follows that

$$t_0 \ge \frac{\kappa}{4 \operatorname{Var}_X c}. \qquad \square$$

The following complexity result is now obtained by combining the two phases of the path-following algorithm.

Theorem 18.2.5. *A standard problem (SP) with ν-self-concordant barrier, tolerance level $\epsilon > 0$ and starting point $\overline{x} \in \operatorname{int} X$ can be solved with at most*

$$C\sqrt{\nu}\ln(\nu\Phi/\epsilon + 1)$$

Newton steps, where

$$\Phi = \frac{\operatorname{Var}_X(c)}{1 - \pi_{\hat{x}_F}(\overline{x})}$$

and the constant C only depends on γ and κ.

Proof. Phase 1 provides a starting point x_0 and an initial value t_0 for phase 2, satisfying the condition $t_0 \ge \kappa/(4 \operatorname{Var}_X(c))$. The number of inner iterations in phase 2 is therefore bounded by

$$O(1)\sqrt{\nu}\ln\Big(\frac{4\nu \operatorname{Var}_X(c)}{\kappa\epsilon} + 1\Big) = O(1)\sqrt{\nu}\ln\Big(\frac{\nu \operatorname{Var}_X(c)}{\epsilon} + 1\Big).$$

So the total number of inner iterations in the two phases is

$$\begin{aligned}&O(1)\sqrt{\nu}\ln\Big(\frac{\nu}{1 - \pi_{\hat{x}_F}(\overline{x})} + 1\Big) + O(1)\sqrt{\nu}\ln\Big(\frac{\nu \operatorname{Var}_X(c)}{\epsilon} + 1\Big) \\ &\qquad = O(1)\sqrt{\nu}\ln(\nu\Phi/\epsilon + 1). \qquad \square\end{aligned}$$

Remark. The algorithms in this section provide nice theoretical complexity results, but they are not suitable for practical use. The main limitation is the low updating factor $(1 + O(1)\nu^{-1/2})$ of the penalty parameter t, which implies that the total number of Newton steps will be proportional to $\sqrt{\nu}$.

For an LP problem with $n = 1000$ variables and $m = 10000$ inequalities, one would need to solve hundreds of linear equations with 1000 variables, which requires far more time than what is needed by the simplex algorithm. In the majority of outer iterations, one can, however, in practice increase the penalty parameter much faster than what is needed for the theoretical worst case analysis, without necessarily having to increase the number of Newton steps to maintain proximity to the central path. There are good practical implementations of the algorithm that use various dynamic strategies to control the penalty parameter t, and as a result only a moderate total number of Newton steps is needed, regardless of the size of the problem.

18.3 LP problems

We now apply the algorithm in the previous section on LP problems. Consider a problem of the type

$$\begin{array}{ll} \min & \langle c, x\rangle \\ \text{s.t.} & Ax \leq b \end{array} \tag{18.14}$$

where $A = [a_{ij}]$ is an $m \times n$-matrix. We assume that the polyhedron

$$X = \{x \in \mathbf{R}^n \mid Ax \leq b\}$$

of feasible points is bounded and has a nonempty interior. The boundedness assumption implies that $m > n$.

The ith row of the matrix A is denoted by a_i, that is $a_i = [a_{i1}\ a_{i2}\ \ldots\ a_{in}]$. The matrix product $a_i x$ is thus well-defined.

As a barrier to the set X we use the m-self-concordant function

$$F(x) = -\sum_{i=1}^{m} \ln(b_i - a_i x).$$

The path-following algorithm started from an arbitrary point $\overline{x} \in \operatorname{int} X$ results in an ϵ-solution, i.e. a point with a value of the objective function that approximates the optimal value with an error less than ϵ, after at most

$$O(1)\sqrt{m}\ln(m\Phi/\epsilon + 1)$$

inner iterations, where

$$\Phi = \frac{\operatorname{Var}_X(c)}{1 - \pi_{\hat{x}_F}(\overline{x})}.$$

We now estimate the number of arithmetic operations (additions, subtractions, multiplications and divisions) that are required during phase 2 of the algorithm to obtain this ϵ-solution.

For each inner iteration of the Newton algorithm, we first have to compute the gradient and the hessian of the barrier function at the current point x, i.e.

$$F'(x) = \sum_{i=1}^{m} \frac{a_i^T}{b_i - a_i x} \quad \text{och} \quad F''(x) = \sum_{i=1}^{m} \frac{a_i^T a_i}{(b_i - a_i x)^2}.$$

This can be done with $O(mn^2)$ arithmetic operations. The Newton direction Δx_{nt} at x is obtained as solution to the quadratic system

$$F''(x)\Delta x_{\text{nt}} = -(tc + F'(x))$$

of linear equations, and using Gaussian elimination, we find the solution after $O(n^3)$ arithmetic operations. Finally, $O(n)$ additional arithmetic operations, including one square root extraction, are needed to compute the Newton decrement $\lambda = \lambda(F_t, x)$ and the new point $x^+ = x + (1 + \lambda)^{-1}\Delta x_{\text{nt}}$.

The corresponding estimate of the number of operations is also true for phase 1 of the algorithm.

The gradient and hessian computation is the most costly of the above computations since $m > n$. The total number of arithmetic operations in each iteration is therefore $O(mn^2)$, and by multiplying with the number of inner iterations, the overall arithmetic cost of the path-following algorithm is estimated to be no more than $O(m^{3/2}n^2)\ln(m\Phi/\epsilon + 1)$ operations.

The resulting approximate minimum point $\hat{x}(\epsilon)$ is an interior point of the polyhedron X, but the minimum is of course attained at an extreme point on the the boundary of X. However, there is a simple procedure, called *purification* and described below, which starting from $\hat{x}(\epsilon)$ finds an extreme point $\hat{x}$ of X after no more than $O(mn^2)$ arithmetic operations and with an objective function value that does not exceed the value at $\hat{x}(\epsilon)$. This means that we have the following result.

Theorem 18.3.1. *For the LP problem* (18.14) *at most*

$$O(m^{3/2}n^2)\ln(m\Phi/\epsilon + 1)$$

arithmetic operations are needed to find an extreme point $\hat{x}$ of the polyhedron of feasible points that approximates the minimum value with an error less than ϵ.

Purification

The proof of the following theorem describes an algorithm for how to generate an extreme point with a value of the objective function that does not exceed the value at a given interior point of the polyhedron of feasible points.

Theorem 18.3.2. *Let*

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

be an LP problem with n variables and m constraints, and suppose that the polyhedron X of feasible points is line-free and that the objective function is bounded below on X. For each point of X we can generate an extreme point of X with a value of the objective function that does not exceed the value at the given point with an algorithm using at most $O(mn^2)$ arithmetic operations.

Proof. The idea is very simple: Follow a half-line from the given point $x^{(0)}$ with non-increasing function values until hitting upon a point $x^{(1)}$ in a face F_1 of the polyhedron X. Then follow a half-line in the face F_1 with non-increasing function values until hitting upon a point $x^{(2)}$ in the intersection $F_1 \cap F_2$ of two faces, etc. After n steps, one has reached a point $x^{(n)}$ in the intersection of n (independent) faces, i.e. an extreme point, with a function value that is less than or equal to the value at the starting point.

To estimate the number of arithmetic operation we have to study the above procedure in a little more detail.

We start by defining $v^{(1)} = \mathbf{e}_1$ if $c_1 < 0$, $v^{(1)} = -\mathbf{e}_1$ if $c_1 > 0$, and $v^{(1)} = \pm\mathbf{e}_1$ if $c_1 = 0$, where the sign in the latter case should be chosen so that the half-line $x^{(0)}+tv^{(1)}$, $t \geq 0$, intersects the boundary of the polyhedron; this is possible since the polyhedron is assumed to be line-free. In the first two cases, the half-line also intersects the boundary of the polyhedron, because $\langle c, x^{(0)} + tv^{(1)}\rangle = \langle c, x^{(0)}\rangle - t|c_1|$ tends to $-\infty$ as t tends to ∞ and the objective function is assumed to be bounded below on X. The intersection point $x^{(1)} = x^{(0)} + t_1 v^{(1)}$ between the half-line and the boundary of X can be computed with $O(mn)$ arithmetic operations, since we only have to compute the vectors $b - Ax^{(0)}$ and $Av^{(1)}$, and quotients between their coordinates in order to find the nonnegative parameter value t_1.

After renumbering the equations, we may assume that the point $x^{(1)}$ lies in the hyperplane $a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1$. We now eliminate the variable x_1 from the constraints and the objective function, which results in

a system of the form

$$\begin{cases} x_1 + a'_{12}x_2 + \cdots + a'_{1n}x_n = b'_1 \\ A' \begin{bmatrix} x_2 \\ \vdots \\ x_n \end{bmatrix} \le b' \end{cases} \tag{18.15}$$

where A' is an $(m-1) \times (n-1)$-matrix, and in a new objective function

$$c'_2x_2 + \cdots + c'_nx_n + d',$$

which is the restriction of the original objective function to the current face. The number of operations required to perform the eliminations is $O(mn)$.

After $O(mn)$ operations we have thus managed to find a point $x^{(1)}$ in a face F_1 of X with an objectiv function value $\langle c, x^{(1)} \rangle = \langle c, x^{(0)} \rangle - t_1|c_1|$ not exceeding $\langle c, x^{(0)} \rangle$, and determined the equation of the face and the restriction of the objective function to the face. We now have a problem of lower dimension $n-1$ and with $m-1$ constraints.

We continue by choosing a descent vector $v^{(2)}$ for the objective function that is parallel to the face F_1, and we achieve this by defining $v^{(2)}$ so that $v_2^{(2)} = \pm 1$, $v_3^{(2)} = \cdots = v_n^{(2)} = 0$ (and $v_1^{(2)} = -a'_{12}v_2^{(2)}$), where the sign of $v_2^{(2)}$ should be chosen so that the objective function is non-decreasing along the half-line $x^{(1)} + tv^{(2)}$, $t \ge 0$, and the half-line instersects the relative boundary of F_1. This means that $v_2^{(2)} = 1$ if $c'_2 < 0$ and $v_2^{(2)} = -1$ if $c'_2 > 0$, while the sign of $v_2^{(2)}$ is determined by the requirement that the half-line should intersect the boundary in the case $c'_2 = 0$.

We then determine the intersection between the half-line $x^{(1)} + tv^{(2)}$, $t \ge 0$, and the relative boundary of F_1, which occurs in one of the remaining hyperplanes. If this hyperplane is the hyperplane $a'_{21}x_2 + \cdots + a'_{2n}x_n = b'_2$, say, we eliminate the variable x_2 from the remaining constraints and the objective function. All this can be done with at most $O(mn)$ operations and results in a point $x^{(2)}$ in the intersection of two faces, and the new value of the objective function is $\langle c, x^{(2)} \rangle = \langle c, x^{(1)} \rangle - t_2|c'_2| \le \langle c, x^{(1)} \rangle$.

After n iterations, which together require at most $nO(mn) = O(mn^2)$ arithmetic operations, we have reached an extreme point $\hat{x} = x^{(n)}$ with a function value that does not exceed the value at the starting point $x^{(0)}$. The coordinates of the extreme point are obtained by solving a triangular system of equations, which only requires $O(n^2)$ operations. The total number of operations is thus $O(mn^2)$. □

EXAMPLE 18.3.1. We exemplify the purification algorithm with the LP problem

$$\begin{array}{ll} \min & -2x_1 + x_2 + 3x_3 \\ \text{s.t.} & \begin{cases} -x_1 + 2x_2 + x_3 \leq 4 \\ -x_1 + x_2 + x_3 \leq 2 \\ x_1 - 2x_2 \leq 1 \\ x_1 - x_2 - 2x_3 \leq 1 \end{cases} \end{array}$$

Starting from the interior point $x^{(0)} = (1, 1, 1)$ with objectiv function value $c^T x^{(0)} = 2$, we shall find an extreme point with a lower value.

Since $c_1 = -2 < 0$, we begin by choosing $v^{(1)} = (1, 0, 0)$ and by determining the point of intersection between the half-line $x = x^{(0)} + tv^{(1)} = (1+t, 1, 1)$, $t \geq 0$, and the boundary of the polyhedron of feasible points. We find that the point $x^{(1)} = (3, 1, 1)$, corresponding to $t = 2$, satisfies all constraints and the third constraint with equality. So $x^{(1)}$ lies in the face obtained by intersecting the polyhedron X with the supporting hyperplane $x_1 - 2x_2 = 1$. We eliminate x_1 from the objectiv function and from the remaining constraints using the equation of this hyperplane, and consider the restriction of the objective function to the corresponding face, i.e. the function $f(x) = -3x_2 + 3x_3 - 2$ restricted to the polyhedron given by the system

$$\begin{cases} x_1 - 2x_2 = 1 \\ x_3 \leq 5 \\ -x_2 + x_3 \leq 3 \\ x_2 - 2x_3 \leq 0 \end{cases}$$

The x_2-coefficient of our new objective function is negative, so we follow the half-line $x_2 = 1 + t$, $x_3 = 1$, $t \geq 0$, in the hyperplane $x_1 - 2x_2 = 1$ until it hits a new supporting hyperplane, which occurs for $t = 1$, when it intersects the hyperplane $x_2 - 2x_3 = 0$ in the point $x^{(2)} = (5, 2, 1)$. Elimination of x_2 results in the objective function $f(x) = -3x_3 - 2$ and the system

$$\begin{cases} x_1 - 2x_2 = 1 \\ x_2 - 2x_3 = 0 \\ x_3 \leq 5 \\ -x_3 \leq 3 \end{cases}$$

Our new half-line in the face $F_1 \cap F_2$ is given by the equation $x_3 = 1+t$, $t \geq 0$, and the halfline intersects the third hyperplane $x_3 = 5$ when $t = 4$, i.e. in a point with x_3-coordinate equal to 5. Back substitution gives $x^{(3)} = (21, 10, 5)$, which is an extreme point with objective function value equal to -17. □

18.4 Complexity

By the *complexity* of a problem we here mean the number of arithmetic operations needed to solve it, and in this section we will study the complexity of LP problems with rational coefficients. The *solution* of an LP problem consists by definition of the problem's optimal value and, provided the value is finite, of an optimal point. All known estimates of the complexity depend not only on the number of variables and constraints, but also on the size of the coefficients, and an appropriate measure of the size of a problem is given by the number of binary bits needed to represent all its coefficients.

Definition. The *input length* of a vector $x = (x_1, x_2, \ldots, x_n)$ in $\mathbf{R}^n$ is the integer $\ell(x)$ defined as

$$\ell(x) = \sum_{j=1}^{n} \lceil \log_2(|x_j| + 1) \rceil.$$

The number of digits in the binary expansion of a positive integer z is equal to $\lceil \log_2(|z| + 1) \rceil$. The binary representation of a negative integer z requires one bit more in order to take care of the sign, and so does the representation of $z = 0$. The number of bits to represent an arbitrary vector x in $\mathbf{R}^n$ with integer coordinates is therefore at most $\ell(x) + n$.

The norm of a vector can be estimated using the input length, and we shall need the following simple estimate in the two cases $p = 1$ and $p = 2$.

Lemma 18.4.1. *$\|x\|_p \le 2^{\ell(x)}$ for all $x \in \mathbf{R}^n$ and all $p \ge 1$.*

Proof. The inequality is a consequence of the following trivial inequalities $\sum_{j=1}^{n} a_j \le \prod_{j=1}^{n} (a_j + 1)$, $a^p + 1 \le (a + 1)^p$ and $\log_2(a + 1) \le \lceil \log_2(a + 1) \rceil$, which hold for nonnegative numbers a, a_j, and imply that

$$\|x\|_p^p = \sum_{j=1}^{n} |x_j|^p \le \prod_{j=1}^{n} (|x_j|^p + 1) \le \prod_{j=1}^{n} (|x_j| + 1)^p \le 2^{p\,\ell(x)}. \qquad \square$$

We will now study LP problems of the type

$$\text{(LP)} \qquad \begin{aligned} &\min \ \langle c, x \rangle \\ &\text{s.t.} \ \ Ax \le b \end{aligned}$$

where all coefficients of the $m \times n$-matrix $A = [a_{ij}]$ and of the vectors b and c are integers. Every LP problem with rational coefficients can obviously be replaced by an equivalent problem of this type after multiplication with a

suitable least common denominator. The polyhedron of feasible points will be denoted by X so that

$$X = \{x \in \mathbf{R}^n \mid Ax \leq b\}.$$

Definition. The two integers

$$\ell(X) = \ell(A) + \ell(b) \quad \text{and} \quad L = \ell(X) + \ell(c) + m + n,$$

where $\ell(A)$ denotes the input length of the matrix A, considered as a vector in $\mathbf{R}^{mn}$, are called the *input length of the polyhedron* X and the *input length of the given LP problem* (LP), respectively.[†]

The main result of this section is the following theorem, which implies that there is a solution algorithm that is polynomial in the input length of the LP problem.

Theorem 18.4.2. *There is an algorithm which solves the LP problem* (LP) *with at most* $O((m+n)^{7/2}L)$ *arithmetic operations.*

Proof. **I.** We begin by noting that we can without restriction assume that the polyhedron X of feasible points is line-free. Indeed, we can, if necessary replace the problem (LP) with the equivalent and line-free LP problem

$$\begin{aligned} \min \quad & \langle c, x^+ \rangle - \langle c, x^- \rangle \\ \text{s.t.} \quad & \begin{cases} Ax^+ - Ax^- \leq b \\ -x^+ \leq 0 \\ -x^- \leq 0. \end{cases} \end{aligned}$$

This LP problem in $n' = 2n$ variables and with $m' = m + 2n$ constraints has input length

$$\begin{aligned} L' &= 2\ell(A) + 2n + \ell(b) + 2\ell(c) + m' + n' \\ &\leq 2(\ell(A) + \ell(b) + \ell(c) + m + n) + 4n = 2L + 4n \leq 6L, \end{aligned}$$

so any algorithm that solves this problem with $O((m'+n')^{7/2}L')$ operations also solves problem (LP) with $O((m+n)^{7/2}L)$ operations since $m' + n' \leq 4(m+n)$ and $L' \leq 6L$.

From now on, we therefore assume that X *is a line-free polyhedron,* and for nonempty polyhedra X this implies that $m \geq n$ and that X has at least one extreme point.

[†]Since $\ell(X) + mn + m$ bits are needed to represent all coefficients of the polyhedron X and $L + mn$ bits are needed to represent all coefficients of the given LP problem, it would be more logical to call these numbers the input length of the polyhedron and of the LP problem, respectively. However, the forthcoming calculations will be simpler with our conventions.

The assertion of the theorem is also trivially true for LP problems with only one variable, so we assume that $m \geq n \geq 2$. Finally, we can naturally assume that all the rows of the matrix A are nonzero, for if the kth row is identically zero, then the corresponding constraint can be deleted if $b_k \geq 0$, while the polyhedron X of feasible point is empty if $b_k < 0$. In the future, we can thus make use of the inequalities

$$\ell(X) \geq \ell(A) \geq m \geq n \geq 2 \text{ and } L \geq \ell(X) + m + n \geq \ell(X) + 4.$$

II. Under the above assumptions, we will prove the theorem by showing:

1. With $O(m^{7/2}L)$ operations, one can determine whether the optimal value of the problem is $+\infty$, $-\infty$ or finite, i.e. whether there are any feasible points or not, and if there are feasible points whether the objective functions is bounded below or not.
2. Given that the optimal value is finite, one can then determine an optimal solution with $O(m^{3/2}n^2L)$ operations.

Since the proof of statement 1 uses the solution of an appropriate auxiliary LP problem with finite value, we begin by showing statement 2.

III. As a first building block we need a lemma that provides information about the extreme points of the polyhedron X in terms of its input length.

Lemma 18.4.3. *(i) Let $\hat{x}$ be an extreme point of the polyhedron X. Then, the following inequality holds for all nonzero coordinates $\hat{x}_j$:*

$$2^{-\ell(X)} \leq |\hat{x}_j| \leq 2^{\ell(X)}.$$

Thus, all extreme points of X lie in the cube $\{x \in \mathbf{R}^n \mid \|x\|_\infty \leq 2^{\ell(X)}\}$.

(ii) If $\hat{x}$ and $\tilde{x}$ are two extreme points of X and $\langle c, \hat{x}\rangle \neq \langle c, \tilde{x}\rangle$, then

$$|\langle c, \hat{x}\rangle - \langle c, \tilde{x}\rangle| \geq 4^{-\ell(X)}.$$

Proof. To prove the lemma, we begin by recalling *Hadamard's inequality* for $k \times k$-matrices $C = [c_{ij}]$ with columns $C_{*1}, C_{*2}, \ldots, C_{*k}$, and which reads as follows:

$$|\det C| \leq \prod_{j=1}^{k} \|C_{*j}\|_2 = \prod_{j=1}^{k} \Big(\sum_{i=1}^{k} c_{ij}^2\Big)^{1/2}.$$

The inequality is geometrically obvious − the left-hand side $|\det C|$ is the volume of a (hyper)parallelepiped, spanned by the matrix columns, while the right-hand side is the volume of a (hyper)cuboid whose edges are of the same length as the edges of the parallelepiped.

By combining Hadamard's inequality with Lemma 18.4.1, we obtain the inequality

$$|\det C| \leq \prod_{j=1}^{k} 2^{\ell(C_{*j})} = 2^{\ell(C)}.$$

If C is a quadratic submatrix of the matrix $\begin{bmatrix} A & b \end{bmatrix}$, then obviously $\ell(C) \leq \ell(A) + \ell(b) = \ell(X)$, and it follows from the above inequality that

$$|\det C| \leq 2^{\ell(X)}. \tag{18.16}$$

Now let $\hat{x}$ be an extreme point of the polyhedron X. According to Theorem 5.1.1, there is a set $\{i_1, i_2, \ldots, i_n\}$ of row indices such that the extreme point $\hat{x}$ is obtained as the unique solution to the equation system

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = i_1, i_2, \ldots, i_n.$$

By Cramer's rule, we can write the solution in the form

$$\hat{x}_j = \frac{\Delta_j}{\Delta},$$

where Δ is the determinant of the coefficient matrix and Δ_j is the determinant obtained by replacing column number j in Δ with the right-hand side of the equation system. The determinants Δ and Δ_j are integers, and their absolute values are at most equal to $2^{\ell(X)}$, because of inequality (18.16). This leads to the following estimates for all nonzero coordinates $\hat{x}_j$, i.e. for all j with $\Delta_j \neq 0$:

$$|\hat{x}_j| = |\Delta_j|/|\Delta| \leq 2^{\ell(X)}/1 = 2^{\ell(X)} \text{ and } |\hat{x}_j| = |\Delta_j|/|\Delta| \geq 1/2^{\ell(X)} = 2^{-\ell(X)},$$

which is assertion (i) of the lemma.

(ii) The value of the objective function at the extreme point $\hat{x}$ is

$$\langle c, \hat{x} \rangle = \Big(\sum_{j=1}^{n} c_j \Delta_j \Big) / \Delta = T/\Delta,$$

where the numerator T is an integer. If $\tilde{x}$ is another extreme point, then of course we also have $\langle c, \tilde{x} \rangle = T'/\Delta'$ for some integer T' and determinant Δ' with $|\Delta'| \leq 2^{\ell(X)}$. It follows that the difference

$$\langle c, \tilde{x} \rangle - \langle c, \hat{x} \rangle = (T\Delta' - T'\Delta)/\Delta\Delta'$$

is either equal to zero or, if the numerator is nonzero, an integer with absolute value $\geq 1/|\Delta\Delta'| \geq 4^{-\ell(X)}$. □

IV. We shall use the path-following method, but this assumes that the polyhedron of feasible points is bounded and that there is an inner point from which to start phase 1. To get around this difficulty, we consider the following auxiliary problems in $n+1$ variables and $m+2$ linear constraints:

$$(\mathrm{LP}_M) \qquad \begin{array}{ll} \min & \langle c, x\rangle + Mx_{n+1} \\ \text{s.t.} & \begin{cases} Ax + (b - \mathbf{1})x_{n+1} \le b \\ x_{n+1} \le 2 \\ -x_{n+1} \le 0. \end{cases} \end{array}$$

Here, M is a positive integer, $\mathbf{1}$ denotes the vector $(1, 1, \dots, 1)$ in $\mathbf{R}^m$, and x is as before the n-tuple $(x_1, x_2, \dots, x_n)$.

Let X' denote the polyhedron of feasible points for the problem (LP_M). Since $(x, x_{n+1}) = (0, 1)$ satisfies all constraints with strict inequality, $(0, 1)$ is an inner point in X'.

We obtain the following estimates for the input length $\ell(X')$ of the polyhedron X' and the input lenght $L(M)$ of problem (LP_M):

$$\begin{aligned} (18.17) \qquad \ell(X') &= \ell(A) + \sum_{i=1}^{m} \big\lceil \log_2\big(|b_i - 1| + 1\big)\big\rceil + 1 + 1 + \ell(b) + 2 \\ &\le \ell(X) + 4 + \sum_{i=1}^{m} \big(1 + \big\lceil \log_2\big(1 + |b_i\big)\big\rceil\big) \\ &= \ell(X) + 4 + m + \ell(b) \le 2\ell(X) + 4 \le 2L - 4, \end{aligned}$$

$$\begin{aligned} (18.18) \qquad L(M) &= \ell(X') + \ell(c) + \lceil \log_2(M + 1)\rceil + m + n + 3 \\ &\le 2\ell(X) + 2\ell(c) + \lceil \log_2 M\rceil + m + n + 8 \\ &= 2L + \lceil \log_2 M\rceil - (m + n) + 8 \le 2L + \lceil \log_2 M\rceil + 4. \end{aligned}$$

The reason for studying our auxiliary problem (LP_M) is given in the following lemma.

Lemma 18.4.4. *Assume that problem* (LP) *has a finite value. Then:*

(i) Problem (LP_M) *has a finite value for each integer* $M > 0$.

(ii) If $(\hat{x}, 0)$ *is an optimal solution to problem* (LP_M), *then* $\hat{x}$ *is an optimal solution to the original problem* (LP).

(iii) Assume that $M \ge 2^{4L}$ *and that the extreme point* $(\hat{x}, \hat{x}_{n+1})$ *of* X' *is an optimal solution to problem* (LP_M). *Then,* $\hat{x}_{n+1} = 0$, *so* $\hat{x}$ *is an optimal solution to problem* (LP).

Proof. (i) The assumption of finite value means that the polyhedron X is nonempty and that the objective function $\langle c, x\rangle$ is bounded below on X, and

by Theorem 12.1.1, this implies that the vector c lies in the dual cone of the recession cone $\operatorname{recc} X$. Since

$$\begin{aligned}\operatorname{recc} X' &= \{(x, x_{n+1}) \mid Ax + (b - \mathbf{1})x_{n+1} \leq 0,\ \ x_{n+1} = 0\} \\ &= \operatorname{recc} X \times \{0\},\end{aligned}$$

the dual cone of $\operatorname{recc} X'$ is equal to $(\operatorname{recc} X)^+ \times \mathbf{R}$. We conclude that the vector (c, M) lies in the dual cone $(\operatorname{recc} X')^+$, which means that the objective function of problem (LP_M) is bounded below on the nonempty set X'. Hence, our auxiliary problem has a finite value.

The polyhedron X' is line-free, since

$$\begin{aligned}\operatorname{lin} X' &= \{(x, x_{n+1}) \mid Ax + (b - \mathbf{1})x_{n+1} = 0,\ \ x_{n+1} = 0\} \\ &= \operatorname{lin} X \times \{0\} = \{(0, 0)\}.\end{aligned}$$

(ii) The point $(x, 0)$ is feasible for problem (LP_M) if and only if x belongs to X, i.e. is feasible for our original problem (LP). So if $(\hat{x}, 0)$ is an optimal solution to the auxiliary problem, then in particular

$$\langle c, \hat{x}\rangle = \langle c, \hat{x}\rangle + M \cdot 0 \leq \langle c, x\rangle + M \cdot 0 = \langle c, x\rangle$$

for all $x \in X$, which shows that $\hat{x}$ is an optimal solution to problem (LP).

(iii) Assume that $(\hat{x}, \hat{x}_{n+1})$ is an extreme point of the polyhedron X' and an optimal solution to problem (LP_M). By Lemma 18.4.3, applied to the polyhedron X', and the estimate (18.17), we then have the inequality

$$\|\hat{x}\|_\infty \leq 2^{\ell(X')} \leq 2^{2\ell(X)+4} \leq 2^{2L-4}, \tag{18.19}$$

so it follows by using Lemma 18.4.1 that

$$\begin{aligned}|\langle c, \hat{x}\rangle| &\leq \sum_{j=1}^{n} |c_j||\hat{x}_j| \leq \|c\|_1 \|\hat{x}\|_\infty \leq 2^{\ell(c)} \cdot 2^{2\ell(X)+4} \leq 2^{2\ell(X)+2\ell(c)+4} \\ &\leq 2^{2L-2m-2n+4} \leq 2^{2L-4}.\end{aligned}$$

Assume that $\hat{x}_{n+1} \neq 0$. Then $\hat{x}_{n+1} \geq 2^{-\ell(X')} \geq 2^{-2L}$, according to Lemma 18.4.3. The optimal value $\hat{v}_M$ of the auxiliary problem (LP_M) therefore satisfies the inequality

$$\hat{v}_M = \langle c, \hat{x}\rangle + M\hat{x}_{n+1} \geq M\hat{x}_{n+1} - |\langle c, \hat{x}\rangle| \geq M \cdot 2^{-2L} - 2^{2L-4}.$$

Let now x be an arbitrary extreme point of X. Since $(x, 0)$ is a feasible point for problem (LP_M) and since $\|x\|_\infty \leq 2^{\ell(X)}$ by lemma 18.4.3, the optimal value $\hat{v}_M$ must also satisfy the inequality

$$\hat{v}_M \leq \langle c, x\rangle + M \cdot 0 \leq |\langle c, x\rangle| \leq \|c\|_1 \cdot \|x\|_\infty \leq 2^{\ell(c)+\ell(X)} = 2^{L-m-n} \leq 2^{L-4}.$$

By combining the two inequalities for $\hat{v}_M$, we obtain the inequality

$$2^{L-4} \geq M \cdot 2^{-2L} - 2^{2L-4},$$

which implies that

$$M \leq 2^{3L-4} + 2^{4L-4} < 2^{4L}.$$

So if $M \geq 2^{4L}$, then $\hat{x}_{n+1} = 0$. □

V. We are now ready for the main step in the proof of Theorem 18.4.2.

Lemma 18.4.5. *Suppose that problem* (LP) *has a finite value. The path-following algorithm, applied to the problem* (LP_M) *with* $\|x\|_\infty \leq 2^{2L}$ *as an additional constraint,* $M = 2^{4L}$, $\epsilon = 2^{-4L}$, *and* $(0, 1)$ *as starting point for phase 1, and complemented with a subsequent purification operation, generates an optimal solution to problem* (LP) *after at most* $O(m^{3/2}n^2L)$ *aritmetic operations.*

Proof. It follows from the previous lemma and the estimate (18.19) that the LP problem (LP_M) has an optimal solution $(\hat{x}, 0)$ which satisfies the additional constraint $\|\hat{x}\|_\infty \leq 2^{2L}$ if $M = 2^{4L}$. The LP problem obtained from (LP_M) by adding the $2n$ constraints

$$x_j \leq 2^{2L} \quad \text{and} \quad -x_j \leq 2^{2L}, \qquad j = 1, 2, \ldots, n,$$

therefore has the same optimal value as (LP_M).

The extended problem has $m+2n+2 = O(m)$ linear constraints, and the point $\overline{z} = (\overline{x}, \overline{x}_{n+1}) = (0, 1)$ is an interior point of the compact polyhedron of feasible points, which we denote by Z. By Theorem 18.3.1, the path-following algorithm with $\epsilon = 2^{-4L}$ and $\overline{z}$ as the starting point therefore stops after $O((m+2n+2)^{3/2}n^2)\ln((m+2n+2)\Phi/\epsilon+1) = O(m^{3/2}n^2)\ln(m2^{4L}\Phi+1)$ arithmetic operations at a point in the polyhedron X' and with a value of the objective function that approximates the optimal value $\hat{v}_M$ with an error less than 2^{-4L}.

Purification according to the method in Theorem 18.3.2 leads to an extreme point $(\hat{x}, \hat{x}_{n+1})$ of X' with a value of the objective function less than $\hat{v}_M + 2^{-4L}$, and since $2^{-4L} = 4^{-2L} < 4^{-\ell(X')}$, it follows from Lemma 18.4.3 that $(\hat{x}, \hat{x}_{n+1})$ is an optimal solution to (LP_M). By Lemma 18.4.4, this implies that $\hat{x}$ is an optimal solution to the original problem (LP).

The purification process requires $O(mn^2)$ arithmetic operations, so the total arithmetic cost is

$$O(mn^2) + O(m^{3/2}n^2)\ln(m2^{4L}\Phi + 1) = O(m^{3/2}n^2)\ln(m2^{4L}\Phi + 1)$$

operations. It thus only remains to prove that $\ln(m2^{4L}\Phi + 1) = O(L)$, and since $m \leq L$, this will follow if we show that $\ln \Phi = O(L)$.

By definition,

$$\Phi = \mathrm{Var}_Z(c, M) \cdot \frac{1}{1 - \pi_{\hat{z}_F}(\overline{z})},$$

where $\hat{z}_F$ is the analytic center of Z with respect to the relevant logarithmic barrier F. The absolute value of the objective function at an arbitrary point $(x, x_{n+1}) \in Z$ can be estimated by

$$|\langle c, x\rangle + M x_{n+1}| \le \|c\|_1 \|x\|_\infty + 2M \le 2^{\ell(c)+2L} + 2 \cdot 2^{4L} \le 2^{4L+2},$$

and the maximal variation of the function is at most twice this value. Hence,

$$\mathrm{Var}_Z(c, M) \le 2^{4L+3}.$$

The second component of Φ is estimated using Theorem 18.1.7. Let $\overline{B}_\infty(a, a_{n+1}; r)$ denote the closed ball of radius r in $\mathbf{R}^{n+1} = \mathbf{R}^n \times \mathbf{R}$ with center at the point (a, a_{n+1}) and with distance given by the maximum norm, i.e.

$$\overline{B}_\infty(a, a_{n+1}; r) = \{(x, x_{n+1}) \in \mathbf{R}^n \times \mathbf{R} \mid \|x - a\|_\infty \le r,\ \ |x_{n+1} - a_{n+1}| \le r\}.$$

The polyhedron Z is by definition included in the ball $\overline{B}_\infty(0, 0; 2^{2L})$. On the other hand, the tiny ball $\overline{B}_\infty(\overline{z}; 2^{-L})$ is included in Z, for if $\|x\|_\infty \le 2^{-L}$ and $|x_{n+1} - 1| \le 2^{-L}$, then

$$\begin{aligned}
&\sum_{j=1}^{n} a_{ij}x_j + (b_i - 1)x_{n+1} - b_i = \sum_{j=1}^{n} a_{ij}x_j + b_i(x_{n+1} - 1) - x_{n+1} \\
&\quad \le \sum_{j=1}^{n} |a_{ij}||x_j| + |b_i||x_{n+1} - 1| - x_{n+1} \le 2^{-L}\Big(\sum_{j=1}^{n} |a_{ij}| + |b_i|\Big) - (1 - 2^{-L}) \\
&\quad \le 2^{-L+\ell(X)} + 2^{-L} - 1 \le 2^{-4} + 2^{-L} - 1 < 0,
\end{aligned}$$

which proves that the ith inequality of the system $Ax + (b-1)x_{n+1} \le b$ holds with strict inequality for $i = 1, 2, \ldots, m$, and the remaining inequalities that define the polyhedron Z are obviously strictly satisfied.

It therefore follows from Theorem 18.1.7 that

$$\pi_{\hat{z}_F}(\overline{z}) \le \frac{2 \cdot 2^{2L}}{2 \cdot 2^{2L} + 2^{-L}},$$

and that consequently

$$\frac{1}{1 - \pi_{\hat{z}_F}(\overline{z})} \le 2 \cdot 2^{3L} + 1 < 2^{3L+2}.$$

This implies that $\Phi \le 2^{4L+3} \cdot 2^{3L+2} = 2^{7L+5}$. Hence, $\ln \Phi = O(L)$, which completes the proof of the lemma. □

VI. It remains to show that $O(m^{7/2}L)$ operations are sufficient to decide whether the optimal value of the original problem (LP) is $+\infty$, $-\infty$ or finite.

To decide whether the value is $+\infty$ or not, i.e. whether the polyhedron X is empty or not, we consider the artificial LP problem

$$\begin{array}{ll} \min & x_{n+1} \\ \text{s.t.} & \begin{cases} Ax - \mathbf{1}x_{n+1} \leq b \\ \qquad -x_{n+1} \leq 0 \end{cases} \end{array}$$

This problem has feasible points since $(0, t)$ satisfies all constraints for sufficiently large positive numbers t. The optimal value of the problem is apparently greater than or equal to zero, and it is equal to zero if and only if $X \neq \emptyset$.

So we can decide whether the polyhedron X is empty or not by determining an optimal solution to the artificial problem. The input length of this problem is $\ell(X) + 2m + n + 4$, and since this number is $\leq 2L$, it follows from Lemma 18.4.5 that we can decide whether X is empty or not with $O(m^{3/2}n^2L)$ aritmethic operations.

Note that we do not need to solve the artificial problem exactly. If the value is greater than zero, then, because of Lemma 18.4.3, it is namely greater than or equal to 2^{-2L}. It is therefore sufficient to determine a point that approximates the value with an error of less than 2^{-2L} to know if the value is zero or not.

VII. If the polyhedron X is nonempty, we have as the next step to decide whether the objective function is bounded below. This is the case if and only if the dual problem to problem (LP) has feasible points, and this dual maximization problem is equivalent to the minimization problem

$$\begin{array}{ll} \min & \langle -b, y \rangle \\ \text{s.t.} & \begin{cases} A^T y \leq c \\ -A^T y \leq -c \\ \qquad -y \leq 0, \end{cases} \end{array}$$

which is a problem with m variables, $2n+m$ $(= O(m))$ constraints and input length

$$2\ell(A) + m + 2\ell(c) + \ell(b) + m + (2n + m) \leq 2L + m \leq 3L.$$

So it follows from step VI that we can decide whether the dual problem has any feasible points with $O(m^{7/2}L)$ operations.

The proof of Theorem 18.4.2 is now complete. □

Exercises

18.1 Show that if the functions f_i are ν_i-self-concordant barriers to the subsets X_i of $\mathbf{R}^{n_i}$, then $f(x_1, \ldots, x_m) = f_1(x_1) + \cdots + f_m(x_m)$ is a $(\nu_1 + \cdots + \nu_m)$-self-concordant barrier to the product set $X_1 \times \cdots \times X_m$.

18.2 Prove that the dual local norm $\|v\|_x^*$ that is associated with the function f is finite if and only if v belongs to $\mathcal{N}(f''(x))^\perp$, and that the restriction of $\|\cdot\|_x^*$ to $\mathcal{N}(f''(x))^\perp$ is a proper norm.

18.3 Let X be a closed proper convex cone with nonempty interior, let $\nu \geq 1$ be a real number, and suppose that the function $f\colon \operatorname{int} X \to \mathbf{R}$ is closed and self-concordant and that $f(tx) = f(x) - \nu \ln t$ for all $x \in \operatorname{int} X$ and all $t > 0$. Prove that

a) $f'(tx) = t^{-1} f'(x)$ b) $f'(x) = -f''(x)x$ c) $\lambda(f, x) = \sqrt{\nu}$.

The function f is in other words a ν-self-concordant barrier to X.

18.4 Show that the nonnegative orthant $X = \mathbf{R}^n_+$, $\nu = n$ and the logarithmic barrier $f(x) = -\sum_{i=1}^n \ln x_i$ fulfill the assumptions of the previous exercise.

18.5 Let $X = \{(x, x_{n+1}) \in \mathbf{R}^n \times \mathbf{R} \mid x_{n+1} \geq \|x\|_2\}$.

a) Show that the function $f(x) = -\ln(x_{n+1}^2 - (x_1^2 + \cdots + x_n^2))$ is self-concordant on $\operatorname{int} X$.

b) Show that X, $\nu = 2$ and f fulfill the assumptions of exercise 18.3. The function f is thus a 2-self-concordant barrier to X.

18.6 Suppose that the function $f\colon \mathbf{R}_{++} \to \mathbf{R}$ is convex, three times continuously differentiable and that

$$|f'''(x)| \leq 3\frac{f''(x)}{x}$$

for all $x > 0$. The function

$$F(x, y) = -\ln(y - f(x)) - \ln x$$

with $X = \{(x, y) \in \mathbf{R}^2 \mid x > 0,\, y > f(x)\}$ as domain is self-concordant according to exercise 16.3. Show that F is a 2-self-concordant barrier to the closure $\operatorname{cl} X$.

18.7 Prove that the function

$$F(x, y) = -\ln(y - x \ln x) - \ln x$$

is a 2-self-concordant barrier to the epigraph

$$\{(x, y) \in \mathbf{R}^2 \mid y \geq x \ln x,\, x \geq 0\}.$$

18.8 Prove that the function

$$G(x, y) = -\ln(\ln y - x) - \ln y$$

is a 2-self-concordant barrier to the epigraph $\{(x, y) \in \mathbf{R}^2 \mid y \geq \mathrm{e}^x\}$.

Bibliografical and historical notices

Basic references in convex analysis are the books by Rockafellar [1] from 1970 and Hiriart-Urutty–Lemarechal [1] from 1993. Almost all results from Chapters 1–10 of this book can be found in one form or another in Rockafellar's book, which also contains a historical overview with references to the original works in the field.

A more accessible book on the same subject is Webster [1]. Among textbooks in convexity with an emphasis on polyhedra, one should mention Stoer–Witzgall [1] and the more combinatorially oriented Grünbaum [1].

A modern textbook on convex optimization is Boyd–Vandenberghe [1], which in addition to theory and algorithms also contains lots of interesting applications from a variety of fields.

Part 1

The general convexity theory was founded around the turn of the century 1900 by Hermann Minkowski [1, 2] as a byproduct of his number theoretic studies. Among other things, Minkowski introduced the concepts of separation and extreme point, and he showed that every compact convex set is equal to the convex hull of its extreme points and that every polyhedron is finitely generated, i.e. one direction of Theorem 5.3.1 − the converse was noted later by Weyl [1].

The concept of dual cone was introduced by Steinitz [1], who also showed basic results about the recession cone.

The theory of linear inequalities is surprisingly young − a special case of Theorem 3.3.7 (Exercise 3.11a) was proved by Gordan [1], the algebraic version of Farkas's lemma, i.e. Corollary 3.3.3, can be found in Farkas [1], and a closely related result (Exercise 3.11b) is given by Stiemke [1]. The first systematic treatment of the theory is given by Weyl [1] and Motzkin [1]. Significant contributions have also been provided by Tucker [1]. The proof

in Chapter 3 of Farkas's lemma has a geometrical character; an alternative algebraic induction proof of the lemma has been given by Kuhn [1].

Extreme points and faces are treated in detail in Klee [1,2].

Jensen [1] studied convex functions of one real variable and showed that convex functions with $\mathbf{R}$ as domain are continuous and have one-sided derivatives everywhere. Jensen's inequality, however, was shown earlier for functions with positive second derivative by Hölder [1].

The conjugate function was introduced by Fenchel [1], and a modern treatment of the theory of convex cones, sets and functions can be found in Fenchel [2], which among other things contains original results about the closure of convex functions and about the subdifferential.

Part II

The earliest known example of linear programming can be found in Fourier's works from the 1820s (Fourier [1]) and deals with the problem of determining the best, with respect to the maximum norm, fit to an overdetermined system of linear equations. Fourier reduced this problem to minimizing a linear form over a polyhedron, and he also hinted a method, equivalent to the simplex algorithm, to compute the minimum.

It was to take until the 1940s before practical problems on a larger scale began to be formulated as linear programming. The transportation problem was formulated by Hitchcock [1], who also gave a constructive solution method, and the diet problem was studied by Stigler [1], who, however, failed to compute the exact solution. The Russian mathematician and economist Kantorovich [1] had some years before formulated and solved LP problems in production planning, but his work was not noticed outside the USSR and would therefore not influence the subsequent development.

The need for mathematical methods for solving military planning problems had become apparent during the Second World War, and in 1947 a group of mathematicians led by George Dantzig and Marshall Wood worked at the U.S. Department of the Air Force with such problems. The group's work resulted in the realization of the importance of linear programming, and the first version of the simplex algorithm was described by Dantzig [1] and Wood–Dantzig [1].

The simplex algorithm is contemporary with the first computers, and this suddenly made it possible to treat large problems numerically and contributed to a breakthrough for linear programming. A conference on linear programming, arranged by Tjalling Koopmans 1949 in Chicago, was also an important step in the popularization of linear programming. During this conference, papers on linear programming were presented by economists,

mathematicians, and statisticians. The papers were later published in Koopmans [1], and this book became the start for a rapidly growing literature on linear programming.

The theory of convex programs has its roots in a paper by Kuhn–Tucker which deals with necessary and sufficient conditions for optimality in non-linear problems. Kuhn–Tucker [1] noted the connection between Lagrange multipliers and saddle points, and they focused on the role of convexity. A related result with Lagrange multiplier conditions had otherwise been shown before by John [1] for general differentiable constraints, and KKT conditions are present for the first time in an unpublished master's thesis by Karush [1]. Theorem 11.2.1 can be found in Uzawa [1].

The duality theorem in linear programming was known as a result of game theory by John von Neumann, but the first published proof of this theorem appears in Gale–Kuhn–Tucker [1].

There are numerous textbooks on linear programming. Two early such books, written by pioneers in the field, are Dantzig [4], which in addition to the mathematical material also contains a thorough historical overview, many applications and an extensive bibliography, and Gale [1], which provides a concise but mathematically rigorous presentation of linear programming with an emphasis on economic applications. More recent books are Chvatal [1] and Luenberger [1].

Part III

Dantzig's [2] basic article 1951 treated the non-degenerate case of the simplex algorithm, and the possibility of cycling in the degenerate case caused initially some concern. The first example with cycling was constructed by Hoffman [1], but even before this discovery Charnes [1] had proposed a method for avoiding cycling. Other such methods were then given by Dantzig–Orden–Wolfe [1] and Wolfe [2]. Bland's [1] simple pivoting rule is relatively recent.

It is easy to modify the simplex algorithm so that it is directly applicable to LP problems with bounded variables, which was first noted by Charnes–Lemke [1] and Dantzig [3].

The dual simplex algorithm was developed by Beale [1] and Lemke [1]. The currently most efficient variants of the simplex algorithm are primal-dual algorithms.

Convex quadratic programs can be solved by a variant of the simplex algorithm, formulated by Wolfe [1].

Khachiyan's [1] complexity results was based on the ellipsoid algorithm, which was first proposed by Shor [1] as a method in general convex optimization. See Bland–Goldfarb–Todd [1] for an overview of the ellipsoid method.

Many variants of Karmarkar's [1] algorithm were developed after his publication in 1984. Algorithms for LP problems with $O(n^3L)$ as complexity bound are described by Gonzaga [1] and Ye [1].

Part IV

Newton's method is a classic iterative algorithm for finding critical points of differentiable functions. Local quadratic convergence for functions with Lipschitz continuous, positive definite second derivatives in a neighborhood of the critical point was shown by Kantorovich [2].

Barrier methods for solving nonlinear optimization problems were first used during the 1950s. The central path with logarithmic barriers was studied by Fiacco and McCormick, and their book on sequential minimization techniques − Fiacco–McCormick [1], first published in 1968 − is the standard work in the field. The methods worked well in practice, for the most part, but there were no theoretical complexity results. They lost in popularity in the 1970s and then experienced a renaissance in the wake of Karmarkar's discovery.

Karmarkar's [1] polynomial algorithm for linear programming proceeds by mapping the polyhedron of feasible points and the current approximate solution x_k onto a new polyhedron and a new point x'_k which is located near the center of the new polyhedron, using a projective scaling transformation. Thereafter, a step is taken in the transformed space which results in a point x_{k+1} with a lower objective function value. The progress is measured by means of a logarithmic potential function.

It was soon noted that Karmarkar's potential-reducing algorithm was akin to previously studied path-following methods, and Renegar [1] and Gonzaga [1] managed to show that the path-following method with logarithmic barrier is polynomial for LP problems.

A general introduction to linear programming and the algorithm development in the area until the late 1980s (the ellipsoid method, Karmarkar's algorithm, etc.) is given by Goldfarb–Todd [1]. An overview of potential-reducing algorithms is given by Todd [1], while Gonzaga [2] describes the evolution of path-following algorithms until 1992.

A breakthrough in convex optimization occurred in the late 1980s, when Yurii Nesterov discovered that Gonzaga's and Renegar's proof only used two properties of the logarithmic barrier function, namely, that it satisfies the two differential inequalities, which with Nesterov's terminology means that the barrier is self-concordant with finite parameter ν. Since explicit computable self-concordant barriers exist for a number of important types of convex sets, the theoretical complexity results for linear programming could now be

extended to a large class of convex optimization problems, and Nemirovskii together with Nesterov developed algorithms for convex optimization based on self-concordant barriers. See Nesterov–Nesterovski [1].

References

Adler, I. & Megiddo, N.

[1] A simplex algorithm whose average number of steps is bounded between two quadratic functions of the smaller dimension. *J. ACM* 32 (1985), 871–895.

Beale, E.M.L

[1] An alternative method of linear programming, *Proc. Cambridge Philos. Soc.* 50 (1954), 513–523.

Bland, R.G.

[1] New Finite Pivoting Rules for the Simplex Method, *Math. Oper. Res.* 2 (1977), 103–107.

Bland, R.G., Goldfarb, D. & Todd, M.J.

[1] The ellipsoid method: A survey, *Oper. Res.* 29 (1981), 1039–1091.

Borgwardt, K.H.

[1] *The Simplex Method − a probabilistic analysis,* Springer-Verlag, 1987.

Boyd, S. & Vandenberghe, L.

[1] *Convex Optimization*, Cambridge Univ. Press, Cambridge, UK, 2004.

Charnes, A.

[1] Optimality and Degeneracy in Linear Programming, *Econometrica* 20 (1952), 160–170.

Charnes, A. & Lemke, C.E.

[1] *The bounded variable problem.* ONR memorandum 10, Carnegie Institute of Technology, 1954.

Chvátal, V.

[1] *Linear Programming.* W.H. Freeman, 1983.

Dantzig, G.B.

[1] Programming of Interdependent Activities. II. Mathematical Model, *Econometrica* 17 (1949), 200–211.

[2] Maximization of Linear Functions of Variables Subject to Linear Inequalities. Pages 339–347 in T.C. Koopmans (ed.), *Activity Analysis of Production and Allocation*, John Wiley, 1951.

Dantzig, G.B.

[3] Upper Bounds, Secondary Constraints and Block Triangularity in Linear Programming, *Econometrica* 23 (1955), 174–183.

[4] *Linear Programming and Extensions.* Princeton University Press, 1963.

Dantzig, G.B., Orden, A. & Wolfe, P.

[1] The genereralized simplex method for minimizing a linear form under linear inequality constraints, *Pacific J. Math.* 5 (1955), 183–195.

Farkas, J.

[1] Theorie der einfachen Ungleichungen, *J. reine angew. Math.* 124 (1902), 1–27.

Fenchel, W.

[1] On conjugate convex functions. *Canad. J. Math.* 1 (1949), 73-77.

[2] *Convex Cones, Sets and Functions.* Lecture Notes, Princeton University, 1951.

Fiacco, A.V. & McCormick, G.P.

[1] *Nonlinear Programming: Sequential Unconstrained Minimization Techniques.* Society for Industrial and Applied Mathematics, 1990. (First published in 1968 by Research Analysis Corporation.)

Fourier, J.-B.

[1] Solution d'une question particulière du calcul des inégalités. Sid 317–328 i *Oeuvres de Fourier II*, 1890.

Gale, D.

[1] *The Theory of Linear Economic Models.* McGraw–Hill, 1960.

Gale, D., Kuhn, H.W. & Tucker, A.W.

[1] Linear programming and the theory of games. Pages 317–329 in Koopmans, T.C. (ed.), *Activity Analysis of Production and Allocation*, John Wiley & Sons, 1951.

Goldfarb, D.G. & Todd, M.J.

[1] Linear programming. Chapter 2 in Nemhauser, G.L. et al. (eds.), *Handbooks in Operations Research and Management Science, vol. 1: Optimization*, North-Holland, 1989.

Gonzaga, C.C.

[1] An algorithm for solving linear programming problems in $O(n^3L)$ operations. Pages 1–28 in Megiddo, N. (ed.), *Progress in Mathematical Programming: Interior-Point and Related Methods,* Springer-Verlag, 1988.

[2] Path-Following Methods for Linear Programming, *SIAM Rev.* 34 (1992), 167–224.

Gordan, P.

[1] Über die Auflösung linearer Gleichungen mit reellen Coefficienten, *Math. Ann.* 6 (1873), 23–28.

Grünbaum, B.

[1] *Convex Polytopes.* Interscience publishers, New York, 1967.

Hiriart-Urruty, J.-B. & Lemaréchal, C.

[1] *Convex Analysis and Minimization Algorithms.* Springer, 1993.

Hitchcock, F.L.

[1] The distribution of a product from several socurces to numerous localities, *J. Math. Phys.* 20 (1941), 224–230.

Hoffman, A. J.

[1] *Cycling in the Simplex Algorithm.* Report No. 2974, National Bureau of Standards, Gaithersburg, MD, USA, 1953.

Hölder, O.

[1] Über einen Mittelwertsatz, *Nachr. Ges. Wiss. Göttingen*, 38–47, 1989.

Jensen, J.L.W.V.

[1] Sur les fonctions convexes et les inégalités entre les valeur moyennes, *Acta Math.* 30 (1906), 175–193.

John, F.

[1] Extremum problems with inequalities as subsidiary conditions, 1948. Sid 543–560 i Moser J. (ed.), *Fritz John, Collected Papers*, Birkhäuser Verlag, 1985.

Kantorovich, L.V.

[1] Mathematical methods of organizing and planning production, Leningrad State Univ. (in Russian), 1939. Engl. translation in *Management Sci.* 6 (1960), 366-422.

[2] *Functional Analysis and Applied Mathematics.* National Bureau of Standards, 1952. (First published in Russian in 1948.)

Karmarkar, N.

[1] A new polynomial-time algorithm for linear programming, *Combinatorica* 4 (1984), 373–395.

Karush, W.

[1] *Minima of Functions of Several Variables with Inequalities as Side Constraints.* M. Sc. Disseratation. Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois, 1939.

Khachiyan, L.G.

[1] A polynomial algorithm in linear programming, *Dokl. Akad. Nauk SSSR* 244 (1979), 1093–1096. Engl. translation in *Soviet Math. Dokl.* 20 (1979), 191-194.

Klee, V.
[1] Extremal structure of convex sets, *Arch. Math.* 8 (1957), 234–240.
[2] Extremal structure of convex sets, II, *Math. Z.* 69 (1958), 90–104.

Klee, V. & Minty, G.J.
[1] How Good is the Simplex Algorithm? Pages 159–175 in Shisha, O. (ed.), *Inequalities, III,* Academic Press, 1972.

Koopmans, T.C., ed.
[1] *Activity Analysis of Production and Allocation.* John Wiley & Sons, 1951.

Kuhn, H.W.
[1] Solvability and Consistency for Linear Equations and Inequalities, *Amer. Math. Monthly* 63 (1956), 217–232.

Kuhn, H.W. & Tucker, A.W.
[1] Nonlinear programming. Pages 481–492 in *Proc. of the second Berkeley Symposium on Mathematical Statistics and Probability.* Univ. of California Press, 1951.

Lemke, C.E.
[1] The dual method of solving the linear programming problem, *Naval Res. Logist. Quart.* 1 (1954), 36–47.

Luenberger, D.G.
[1] *Linear and Nonlinear Programming.* Addison–Wesley, 1984

Minkowski, H.
[1] *Geometrie der Zahlen.* Teubner, Leipzig, 1910.
[2] *Gesammelte Abhandlungen von Hermann Minkowski, Vol. 1, 2.* Teubner, Leibzig, 1911

Motzkin, T.
[1] *Beiträge zur Theorie der linearen Ungleichungen.* Azviel, Jerusalem, 1936.

Nesterov, Y. & Nemirovskii, A.
[1] *Interior-Point Polynomial Algorithms in Convex Programming.* Society for Industrial and Applied Mathematics, 1994.

Renegar, J.
[1] A polynomial-time algorithm based on Newton's method for linear programming, *Math. Programm.* 40 (1988), 59–94.

Rockafellar, R.T.
[1] *Convex Analysis.* Princeton Univ. Press., 1970

Shor, N.Z.
[1] Utilization of the operation of space dilation in the minimization of convex functions, *Cybernet. System Anal.* 6 (1970), 7–15.

Smale, S.
[1] On the average number of steps in the simplex method of linear programming, *Math. Program.* 27 (1983), 241–262.

Steinitz, E.
[1] Bedingt konvergente Reihen und konvexe Systeme, I, II, III, *J. Reine Angew. Math.* 143 (1913), 128–175; 144 (1914), 1–40; 146 (1916), 1–52.

Stiemke, E.
[1] Über positive Lösungen homogener linearer Gleichungen, *Math. Ann.* 76 (1915), 340–342.

Stigler, G.J.
[1] The Cost of Subsistence, *J. Farm Econ.* 27 (1945), 303–314.

Stoer, J. & Witzgall, C.
[1] *Convexity and Optimization in Finite Dimensions I.* Springer-Verlag, 1970.

Todd, M.
[1] Potential-reduction methods in mathematical programming, *Math. Program.* 76 (1997), 3–45.

Tucker, A.W.
[1] Dual Systems of Homogeneous Linear Relations. Pages 3–18 in Kuhn, H.W. & Tucker, A.W. (eds.), *Linear Inequalities and Related Systems*, Princeton Univ. Press, 1956.

Uzawa, H.
[1] The Kuhn–Tucker theorem in concave programming. In Arrow, K.J., Hurwicz, L. & H. Uzawa, H. (eds.), *Studies in Linear and Non-linear Programming,* Stanford Univ. Press, 1958.

Webster, R.
[1] *Convexity.* Oxford University Press, 1954.

Weyl, H.
[1] Elementare Theorie der konvexen Polyeder, *Comment. Math. Helv.* 7 (1935), 290–306.

Wolfe, P.
[1] The Simplex Method for Quadratic Programming, *Econometrica* 27 (1959), 382–398.
[2] A Technique for Resolving Degeneracy in Linear Programming, *J. of the Soc. for Industrial and Applied Mathematics* 11 (1963), 205–211.

Wood, M.K. & Dantzig, G.B.
[1] Programming of Interdependent Activities. I. General discussion, *Econometrica* 17 (1949), 193–199.

Wright, S.

[1] *Primal-dual interior-point methods.* SIAM Publications, 1997.

Ye, Y.

[1] An $O(n^3L)$ potential reduction algorithm for linear programming. *Math. Program.* 50 (1991), 239–258.

[2] *Interior point algorithms.* John Wiley and Sons, 1997.

Answers and solutions to the exercises

Chapter 2

2.2 a) $\{x \in \mathbf{R}^2 \mid 0 \leq x_1 + x_2 \leq 1,\ x_1, x_2 \geq 0\}$
b) $\{x \in \mathbf{R}^2 \mid \|x\| \leq 1\}$
c) $\mathbf{R}^2_{++} \cup \{(0,0)\}$

2.3 E.g. $\{(0,1)\} \cup (\mathbf{R} \times \{0\})$ in $\mathbf{R}^2$.

2.4 $\{x \in \mathbf{R}^3_{++} \mid x_3^2 \leq x_1 x_2\}$

2.5 Use the triangle inequality

$$\left(\sum_1^n (x_j + y_j)^2\right)^{1/2} \leq \left(\sum_1^n x_j^2\right)^{1/2} + \left(\sum_1^n y_j^2\right)^{1/2}$$

to show that the set is closed under addition of vectors. Or use the perspective map; see example 2.3.4.

2.6 Follows from the fact that $-\mathbf{e}_k$ is a conic combination of the vectors $\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_n$.

2.7 Let X be the halfspace $\{x \in \mathbf{R}^n \mid \langle c, x\rangle \geq 0\}$. Each vector $x \in X$ is a conic combination of c and the vector $y = x - \langle c, x\rangle \|c\|^{-2} c$, and y lies in the $(n-1)$-dimensional subspace $Y = \{x \in \mathbf{R}^n \mid \langle c, x\rangle = 0\}$, which is generated by n vectors as a cone according to the previous exercise. Hence, x is a conic combination of these n vectors and c.

2.8 The intersection between the cone X and the unit circle is a closed circular arc with endpoints x and y, say. The length of the arc is either less than π, equal to π, or equal to 2π. The cone X is proper and generated by the two vectors x and y in the first case. It is equal to a halfspace in the second case and equal to $\mathbf{R}^2$ in the third case, and it is generated by three vectors in both these cases.

2.9 Use exercise 2.8.

2.10 a) $\operatorname{recc} X = \{x \in \mathbf{R}^2 \mid x_1 \geq x_2 \geq 0\}$, $\lin X = \{(0,0)\}$
b) $\operatorname{recc} X = \lin X = \{(0,0)\}$

2.10 c) $\operatorname{recc} X = \{x \in \mathbf{R}^3 \mid 2x_1 + x_2 + x_3 \le 0,\ x_1 + 2x_2 + x_3 \le 0\}$,
$\operatorname{lin} X = \{(t, t, -3t) \mid t \in \mathbf{R}\}$

d) $\operatorname{recc} X = \{x \in \mathbf{R}^3 \mid x_1 \ge |x_2|\}$,
$\operatorname{lin} X = \{x \in \mathbf{R}^3 \mid x_1 = x_2 = 0\}$.

2.12 b) (i) $c(X) = \{x \in \mathbf{R}^2 \mid 0 \le \frac{1}{3}x_1 \le x_2 \le \frac{1}{2}x_1\} = \operatorname{cl}(c(X))$

(ii) $c(X) = \{x \in \mathbf{R}^2 \mid 0 < x_2 \le \frac{1}{2}x_1\} \cup \{(0,0)\}$,
$\operatorname{cl}(c(X) = \{x \in \mathbf{R}^2 \mid 0 \le x_2 \le \frac{1}{2}x_1\}$,

(iii) $c(X) = \{x \in \mathbf{R}^3 \mid x_1 x_3 \ge x_2^2,\ x_3 > 0\} \cup \{(0,0,0)\}$,
$\operatorname{cl}(c(X)) = c(X) \cup \{(x_1, 0, 0) \mid x_1 \ge 0\}$.

c) $c(X) = \{(x, x_{n+1}) \in \mathbf{R}^n \times \mathbf{R} \mid \|x\| \le x_{n+1}\}$.

2.14 Let $z_n = x_n + y_n$, $n = 1, 2, \ldots$ be a convergent sequence of points in $X + Y$ with $x_n \in X$ and $y_n \in Y$ for all n and limit z_0. The sequence $(y_n)_1^\infty$ ha a convergent subsequence $(y_{n_k})_{k=1}^\infty$ with limit $y_0 \in Y$, since Y is compact. The corresponding subsequence $(z_{n_k} - y_{n_k})_{k=1}^\infty$ of points in X converges to $z_0 - y_0$, and the limit point belongs to X since X is a closed set. Hence, $z_0 = (z_0 - y_0) + y_0$ lies in $X + Y$, and this means that $X + Y$ is a closed set.

Chapter 3

3.1 E.g. $\{x \in \mathbf{R}^2 \mid x_2 \le 0\}$ and $\{x \in \mathbf{R}^2 \mid x_2 \ge \mathrm{e}^{x_1}\}$.

3.2 Follows from Theorem 3.1.3 for closed sets and from Theorem 3.1.5 for open sets.

3.4 a) $\mathbf{R}_+ \times \mathbf{R}$ b) $\{0\} \times \mathbf{R}$ c) $\{0\} \times \mathbf{R}_+$ d) $\mathbf{R}_+ \times \mathbf{R}_+$
e) $\{x \in \mathbf{R}^2 \mid x_2 \ge x_1 \ge 0\}$

3.6 a) $X = X^{++} = \{x \in \mathbf{R}^2 \mid x_1 + x_2 \ge 0,\ x_2 \ge 0\}$,
$X^+ = \{x \in \mathbf{R}^2 \mid x_2 \ge x_1 \ge 0\}$

b) $X = X^{++} = \mathbf{R}^2$, $X^+ = \{(0,0)\}$

c) $X = \mathbf{R}^2_{++} \cup \{(0,0)\}$, $X^+ = X^{++} = \mathbf{R}^2_+$

3.7 (i) $\Rightarrow$ (iii): Since $-a_j \notin \operatorname{con} A$, there is, for each j, a vector c_j such that $-\langle c_j, a_j\rangle < 0$ and $\langle c_j, x\rangle \ge 0$ for all $x \in \operatorname{con} A$, which implies that $\langle c_j, a_j\rangle > 0$ and $\langle c_j, a_k\rangle \ge 0$ for all k. It follows that $c = c_1 + c_2 + \cdots + c_m$ works.

(iii) $\Rightarrow$ (ii): Suppose that $\langle c, a_j\rangle > 0$ for all j. Then $\sum_1^m \lambda_j a_j = 0$ implies $0 = \langle c, 0\rangle = \sum_1^m \lambda_j \langle c, a_j\rangle$, so if $\lambda_j \ge 0$ for all j then $\lambda_j \langle c, a_j\rangle = 0$ for all j, with the conclusion that $\lambda_j = 0$ for all j.

(ii) $\Rightarrow$ (i): If there is a vector x such that $x = \sum_1^m \lambda_j a_j$ and $-x = \sum_1^m \mu_j a_j$ with nonnegative scalars λ_j, μ_j, then by addition we obtan

the equality $\sum_1^m(\lambda_j+\mu_j)a_j = 0$ with the conclusions $\lambda_j+\mu_j = 0$, $\lambda_j = \mu_j = 0$ and $x = 0$.

3.8 No solution.

3.10 Solvable for $\alpha \le -2$, $-1 < \alpha < 1$ and $\alpha > 1$.

3.11 a) The systems (S) and (S*) are equivalent to the systems

$$\begin{cases} Ax \ge 0 \\ -Ax \ge 0 \\ Ex \ge 0 \\ \mathbf{1}^T x > 0 \end{cases} \quad \text{and} \quad \begin{cases} A^T(y'-y'') + Ez + \mathbf{1}t = 0 \\ y', y'', z \ge 0,\ t > 0, \end{cases}$$

respectively (with y corresponding to $y''-y'$). The assertion therefore follows from Theorem 3.3.7.

b) The systems (S) and (S*) are equivalent to the systems

$$\begin{cases} Ax \ge 0 \\ -Ax \ge 0 \\ Ex > 0 \end{cases} \quad \text{and} \quad \begin{cases} A^T(y'-y'') + Ez = 0 \\ y', y'', z \ge 0,\ z \ne 0, \end{cases}$$

respectively. Now apply Theorem 3.3.7.

3.12 By Theorem 3.3.7, the system is solvable if and only if the dual system

$$\begin{cases} A^T(y'-y'') + z + u = 0 \\ A(w+u) = 0 \\ y', y'', z, w, u \ge 0,\ u \ne 0 \end{cases}$$

has no solution. It follows from the two equations of the dual system that:

$$\begin{aligned} 0 &= -(w+u)^T A^T = -(w+u)^T A^T(y'-y'') = (w+u)^T(z+u) \\ &= w^T z + w^T u + u^T z + u^T u, \end{aligned}$$

and all the four terms in the last sum are nonnegative. We conclude that $u^T u = 0$, and hence $u = 0$. So the dual system has no solution.

Chapter 4

4.1 a) $\operatorname{ext} X = \{(1,0),(0,1)\}$ b) $\operatorname{ext} X = \{(0,0),(1,0),(0,1),(\frac{1}{2},1)\}$
c) $\operatorname{ext} X = \{(0,0,1),(0,0,-1)\} \cup \{(x_1,x_2,0) \mid (x_1-1)^2 + x_2^2 = 1\} \setminus \{(0,0,0)\}$

4.2 Suppose $x \in \operatorname{cvx} A \setminus A$; then $x = \lambda a + (1-\lambda)y$ where $a \in A$, $y \in \operatorname{cvx} A$ and $0 < \lambda < 1$. It follows that $x \notin \operatorname{ext}(\operatorname{cvx} A)$.

4.3 We have $\operatorname{ext} X \subseteq A$, according to the previous exercise. Suppose that $a \in A \setminus \operatorname{ext} X$. Then $a = \lambda x_1 + (1-\lambda)x_2$, where $x_1, x_2 \in X$, $x_1 \neq x_2$ and $0 < \lambda < 1$. We have $x_i = \mu_i a + (1-\mu_i)y_i$, where $0 \leq \mu_i < 1$ and $y_i \in \operatorname{cvx}(A \setminus \{a\})$. It now follows from the equality

$$a = (1 - \lambda\mu_1 - (1-\lambda)\mu_2)^{-1}(\lambda(1-\mu_1)y_1 + (1-\lambda)(1-\mu_2)y_2),$$

that a lies in $\operatorname{cvx}(A \setminus \{a\})$.Therefore, $\operatorname{cvx}(A \setminus \{a\}) = \operatorname{cvx} A = X$, which contradicts the minimality of A. Hence, $\operatorname{ext} X = A$.

4.4 The set $X \setminus \{x_0\}$ is convex if and only if $]a,b[\subseteq X \setminus \{x_0\}$ for all $a, b \in X \setminus \{x_0\}$, i.e. if and only if $x_0 \notin]a,b[$ for all $a, b \in X \setminus \{x_0\}$, i.e. if and only if $x_0 \in \operatorname{ext} X$.

4.5 E.g. the set in exercise 4.1 c).

4.6 a) Follows directly from Theorem 4.1.3.

b) The extreme point $(1,0)$ of $\left\{x \in \mathbf{R}^2 \mid x_2 \leq \sqrt{1-x_1^2},\ |x_1| \leq 1\right\}$ is not exposed.

4.7 b) A zero-dimensional general face is an extreme point, and a zero-dimensional exposed face is an exposed point. Hence, exercise 4.6 b) contains an example of a general face which is not an exposed face.

c) Suppose that $a, b \in X$ and that the open line segment $]a,b[$ intersects F'. Since $F' \subseteq F$, the same line segment also intersects F, so it follows that $a, b \in F$. But since F' is a general face of F, it follows that $a, b \in F'$. So F' is indeed a general face of X.
The set X in exercise 4.6 b) has $F = \{1\} \times]-\infty, 0]$ as an exposed face, and $F' = \{(1,0)\}$ is an exposed face of F but not of X.

d) Fix a point $x_0 \in F \cap \operatorname{rint} C$. To each $x \in C$ there is a point $y \in C$ such that x_0 lies on the open line segment $]x,y[$, and it now follows from the definition of a general face that $x \in F$.

e) Use the result in d) on the set $C = X \cap \operatorname{cl} F$. Since $\operatorname{rint} C$ contains $\operatorname{rint} F$ as a subset, $F \cap \operatorname{rint} C \neq \emptyset$, so it follows that $C \subseteq F$. The converse inclusion is of course trivial.

f) Use the result in d) with $F = F_1$ och $C = F_2$, which gives us the inclusion $F_2 \subseteq F_1$. The converse inclusion is obtained analogously.

g) If F is a general face and $F \cap \operatorname{rint} X \neq \emptyset$, then $X \subseteq F$ by d) above. For faces $F \neq X$ we therefore have $F \cap \operatorname{rint} X = \emptyset$, which means that $F \subseteq \operatorname{rbdry} X$.

Chapter 5

5.1 a) $(-\frac{2}{3}, \frac{4}{3})$, and $(4,-1)$ b) $(-\frac{2}{3}, \frac{4}{3})$, $(4,-1)$, and$(-3,-1)$
c) $(0,0,0)$, $(2,0,0)$, $(0,2,0)$, $(0,0,4)$, and $(\frac{4}{3}, \frac{4}{3}, 0)$

5.1 d) $(0,4,0,0)$, $(0,\frac{5}{2},0,0)$, $(\frac{3}{2},\frac{5}{2},0,0)$, $(0,1,1,0)$, and $(0,\frac{5}{2},0,\frac{3}{2})$

5.2 The extreme rays are generated by $(-2,4,3)$, $(1,1,0)$, $(4,-1,1)$, and $(1,0,0)$.

5.3 $C = \begin{bmatrix} 1 & -2 & 1 \\ -1 & 2 & 3 \\ -3 & 2 & 5 \end{bmatrix}$

5.4 a) $A = \{(1,0),\, (0,1)\},\ B = \{(-\frac{2}{3},\frac{4}{3}),\, (4,-1)\}$

b) $A = \emptyset,\ B = \{(-\frac{2}{3},\frac{4}{3}),\, (4,-1),\, (-3,-1)\}$

c) $A = \{(1,1,-3),\, (-1,-1,3),\, (4,-7,-1),\, (-7,4,-1)\}$,
$B = \{(0,0,0),\, (2,0,0),\, (0,2,0),\, (0,0,4),\, (\frac{4}{3},\frac{4}{3},0)\}$

d) $A = \emptyset$,
$B = \{(0,4,0,0),\, (0,\frac{5}{2},0,0),\, (\frac{3}{2},\frac{5}{2},0,0),\, (0,1,1,0),\, (0,\frac{5}{2},0,\frac{3}{2})\}$.

5.5 The inclusion $X = \operatorname{cvx} A + \operatorname{con} B \subseteq \operatorname{con} A + \operatorname{con} B = \operatorname{con}(A \cup B)$ implies that $\operatorname{con} X \subseteq \operatorname{con}(A \cup B)$. Obviously, $A \subseteq \operatorname{cvx} A \subseteq X$. Since $\operatorname{cvx} A$ is a compact set, $\operatorname{recc} X = \operatorname{con} B$, so using the assumption $0 \in X$, we obtain the inclusion $B \subseteq \operatorname{con} B \subseteq X$. Thus, $A \cup B \subseteq X$, and it follows that $\operatorname{con}(A \cup B) \subseteq \operatorname{con} X$.

Chapter 6

6.1 E.g. $f_1(x) = x - |x|$ and $f_2(x) = -x - |x|$.

6.3 $a \geq 5$ and $a > 5$, respectively.

6.4 Use the result of exercise 2.1.

6.5 Follows from $f(x) = \max(x_{i_1} + x_{i_2} + \cdots + x_{i_k})$, where the maximum is taken over all subsets $\{i_1, i_2, \ldots, i_k\}$ of $\{1, 2, \ldots, n\}$ consisting of k elements.

6.6 The inequality is trivial if $x_1 + x_2 + \cdots + x_n = 0$, and it is obtained by adding the n inequalities

$$f(x_i) \leq \frac{x_i}{x_1 + \cdots + x_n} f(x_1 + \cdots + x_n) + \big(1 - \frac{x_i}{x_1 + \cdots + x_n}\big) f(0)$$

if $x_1 + \cdots + x_n > 0$.

6.7 Choose

$$c = \frac{f(x_2) - f(x_1)}{\|x_2 - x_1\|^2}\,(x_1 - x_2),$$

to obtain $f(x_1) + \langle c, x_1\rangle = f(x_2) + \langle c, x_2\rangle$. By quasiconvexity,

$$f(\lambda x_1 + (1-\lambda)x_2) + \langle c, \lambda x_1 + (1-\lambda)x_2\rangle \leq f(x_1) + \langle c, x_1\rangle,$$

which simplifies to

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2).$$

6.8 Let $f\colon \mathbf{R}^n \times \mathbf{R} \to \overline{\mathbf{R}}$ be the function defined by

$$f(x,t) = \begin{cases} t & \text{if } (x,t) \in C, \\ +\infty & \text{if } (x,t) \notin C. \end{cases}$$

Then $\inf\{t \in \mathbf{R} \mid (x,t) \in C\} = \inf\{f(x,t) \mid t \in \mathbf{R}\}$, and Theorem 6.2.6 now follows from Corollary 6.2.7.

6.9 Choose, given $x, y \in X$, sequences $(x_k)_1^\infty$, $(y_k)_1^\infty$ of points $x_k, y_k \in \operatorname{int} X$ such that $x_k \to x$ and $y_k \to y$ as $k \to \infty$. Since the points $\lambda x_k + (1-\lambda)y_k$ belong to $\operatorname{int} X$,

$$f(\lambda x_k + (1-\lambda)y_k) \le \lambda f(x_k) + (1-\lambda) f(y_k),$$

and since f is continuous on X, we now obtain the desired inequality $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y)$ by passing to the limit.

6.10 Let $m = \inf\{f(x) \mid x \in \operatorname{rint}(\operatorname{dom} f)\}$ and fix a relative interior point x_0 of $\operatorname{dom} f$. If $x \in \operatorname{dom} f$ is arbitrary and $0 < \lambda < 1$, then $\lambda x + (1-\lambda)x_0$ is a relative interior point of $\operatorname{dom} f$, and it follows that

$$m \le f(\lambda x + (1-\lambda)x_0) \le \lambda f(x) + (1-\lambda) f(x_0).$$

The inequality $f(x) \ge m$ now follows by letting $\lambda \to 1$.

6.11 Minimum 8 at $x = (\frac{1}{8}, 2)$.

6.12 a) $\|x\|_p$ b) $\max(x_1, 0)$.

Chapter 7

7.2 Yes.

7.5 Let J be a subinterval of I. If $f'_+(x) \ge 0$ for all $x \in J$, then

$$f(y) - f(x) \ge f'_+(x)(y-x) \ge 0$$

for all $y > x$ in the interval J, i.e. f is increasing on J. If instead $f'_+(x) \le 0$ for all $x \in J$, then $f(y) - f(x) \ge f'_+(x)(y-x) \ge 0$ for all $y < x$, i.e. f is decreasing on J.

Since the right derivative $f'_+(x)$ is increasing on I, there are three different cases to consider. Either $f'_+(x) \ge 0$ for all $x \in I$, and f is then increasing on I, or $f'_+(x) \le 0$ for all $x \in I$, and f is then decreasing on I, or there is a point $c \in I$ such that $f'_+(x) \le 0$ to the left of c and $f'_+(x) > 0$ to the right of c, and f is in this case decreasing to the left of c and increasing to the right of c.

7.6 a) The existence of the limits is a consequence of the results of the previous exercise.

b) Consider the epigraph of the extended function.

7.7 Follows directly from exercise 7.6 b).

7.8 Suppose that $f \in \mathcal{F}$. Let $x_0 \in \mathbf{R}^n$ be an arbitrary point, and consider the function $g(x) = f(x) - \langle f'(x_0), x - x_0\rangle$. The function g belongs to $\mathcal{F}$ and $g'(x_0) = 0$. It follows that $g(x) \geq g(x_0)$ for all x, which means that $f(x) \geq f(x_0) + \langle f'(x_0), x - x_0\rangle$ for all x. Hence, f is convex by Theorem 7.2.1.

7.9 $\phi(t) = f(x + tv) = f(x) + t\langle f'(x), v\rangle$ for $v \in V_f$ by Theorem 6.7.1. Differentiate two times to obtain $D^2 f(x)[v, v] = \phi''(0) = 0$, with the conlusion that $f''(x)v = 0$.

7.13 By combining Theorem 7.3.1 (i) with x replaced by $\hat{x}$ and $v = x - \hat{x}$ with the Cauchy-Schwarz inequality, we obtain the inequality $\mu\|x - \hat{x}\|^2 \leq \langle f'(x), x - \hat{x}\rangle \leq \|f'(x)\|\|x - \hat{x}\|$.

Chapter 8

8.1 Suppose that f is μ-strongly convex, where $\mu > 0$, and let c be a subgradient at 0 of the convex function $g(x) = f(x) - \frac{1}{2}\mu\|x\|^2$. Then $f(x) \geq f(0) + \langle c, x\rangle + \frac{1}{2}\mu\|x\|^2$ for all x, and the right-hand side tends to ∞ as $\|x\| \to \infty$. Alternatively, one could use Theorem 8.1.6.

8.2 The line segment $[-\mathbf{e}_1, \mathbf{e}_2]$, where $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$.

8.3 a) $\overline{B}_2(0; 1) = \{x \mid \|x\|_2 \leq 1\}$ b) $\overline{B}_1(0; 1) = \{x \mid \|x\|_1 \leq 1\}$
c) $\overline{B}_\infty(0; 1) = \{x \mid \|x\|_\infty \leq 1\}$.

8.4 a) $\operatorname{dom} f^* = \{a\}$, $f^*(a) = b$
b) $\operatorname{dom} f^* = \{x \mid x < 0\}$, $f^*(x) = -1 - \ln(-x)$
c) $\operatorname{dom} f^* = \mathbf{R}_+$, $f^*(x) = x \ln x - x$, $f^*(0) = 0$
d) $\operatorname{dom} f^* = \mathbf{R}$, $f^*(x) = \mathrm{e}^{x-1}$
e) $\operatorname{dom} f^* = \mathbf{R}_-$, $f^*(x) = -2\sqrt{-x}$.

Chapter 9

9.1
$$\min \; 5000x_1 + 4000x_2 + 3000x_3 + 4000x_4$$
$$\text{s.t.} \begin{cases} -x_1 + 2x_2 + 2x_3 + x_4 \geq 16 \\ 4x_1 + x_2 + 2x_4 \geq 40 \\ 3x_1 + x_2 + 2x_3 + x_4 \geq 24, \; x \geq 0 \end{cases}$$

9.2
$$\max \; v$$
$$\text{s.t.} \begin{cases} 2x_1 + x_2 - 4x_3 \geq v \\ x_1 + 2x_2 - 2x_3 \geq v \\ -2x_1 - x_2 + 2x_3 \geq v \\ x_1 + x_2 + x_3 = 1, \; x \geq 0 \end{cases}$$

9.3 The row player should choose row number 2 and the column player column number 1.

9.4 Payoff matrix:

	Sp E	Ru E	Ru 2
Sp E	-1	1	-1
Ru E	1	-1	-2
Sp 2	-1	2	2

The column players problem can be formulated as

$$\min\ u \quad \text{s.t.} \begin{cases} -y_1 + y_2 + y_3 \le u \\ y_1 - y_2 - 2y_3 \le u \\ -y_1 + 2y_2 + 2y_3 \le u \\ y_1 + y_2 + y_3 = 1,\ y \ge 0 \end{cases}$$

9.5 a) $(\frac{4}{5}, \frac{13}{15})$

9.6 a) $\max\ r$ s.t. $\begin{cases} -x_1 + x_2 + r\sqrt{2} \le 0 \\ x_1 - 2x_2 + r\sqrt{5} \le 0 \\ x_1 + x_2 + r\sqrt{2} \le 1 \end{cases}$

b) $\max\ r$ s.t. $\begin{cases} -x_1 + x_2 + 2r \le 0 \\ x_1 - 2x_2 + 3r \le 0 \\ x_1 + x_2 + 2r \le 1 \end{cases}$

Chapter 10

10.1 $\phi(\lambda) = 2\lambda - \frac{1}{2}\lambda^2$

10.2 The dual functions ϕ_a and ϕ_b of the two problems are given by:

$$\phi_a(\lambda) = 0 \text{ for all } \lambda \ge 0 \text{ and } \phi_b(\lambda) = \begin{cases} 0 & \text{if } \lambda = 0, \\ \lambda - \lambda \ln \lambda & \text{if } 0 < \lambda < 1, \\ 1 & \text{if } \lambda \ge 1. \end{cases}$$

10.5 The inequality $g_i(x_0) \ge g_i(\hat{x}) + \langle g_i'(\hat{x}), x_0 - \hat{x} \rangle = \langle g_i'(\hat{x}), x_0 - \hat{x} \rangle$ holds for all $i \in I(\hat{x})$. It follows that $\langle g_i'(\hat{x}), \hat{x} - x_0 \rangle \ge -g_i(x_0) > 0$ for $i \in I_{\text{oth}}(\hat{x})$, and $\langle g_i'(\hat{x}), \hat{x} - x_0 \rangle \ge -g_i(x_0) \ge 0$ for $i \in I_{\text{aff}}(\hat{x})$.

10.6 a) $v_{\min} = -1$ for $x = (-1, 0)$ b) $v_{\max} = 2 + \frac{\pi}{4}$ for $x = (1, 1)$
c) $v_{\min} = -\frac{1}{3}$ for $x = \pm(\frac{2}{\sqrt{6}}, -\frac{1}{\sqrt{6}})$ d) $v_{\max} = \frac{1}{54}$ for $x = (\frac{1}{6}, 2, \frac{1}{3})$

Chapter 11

11.1 $\hat{\lambda} = 2b$

11.3 b) Let $L\colon \Omega \times \Lambda \to \mathbf{R}$ and $L_1\colon (\mathbf{R} \times \Omega) \times (\mathbf{R}_+ \times \Lambda) \to \mathbf{R}$ be the Lagrange functions of the problems (P) and (P′), respectively, and

let ϕ and ϕ_1 be the corresponding dual functions. The two Lagrange functions are related as follows:

$$L_1(t, x, \lambda_0, \lambda) = (1 - \lambda_0)(t - f(x)) + L(x, \lambda).$$

The Lagrange function L_1 is for fixed $(\lambda_0, \lambda) \in \mathbf{R}_+ \times \Lambda$ bounded below if and only if $\lambda_0 = 1$ and $\lambda \in \operatorname{dom}\phi$. Hence, $\operatorname{dom}\phi_1 = \{1\} \times \operatorname{dom}\phi$. Moreover, $\phi_1(1, \lambda) = \phi(\lambda)$ for all $\lambda \in \operatorname{dom}\phi$.

11.4 Let I be the index set of all non-affine constraints, and let k be the number of elements of I. Slater's condition is satisfied by the point $\overline{x} = k^{-1} \sum_{i \in I} \overline{x}_i$.

11.5 Let $b^{(1)}$ and $b^{(2)}$ be two points in U, and let $0 < \lambda < 1$. Choose, given $\epsilon > 0$, feasible points $x^{(i)}$ for the problems $(\mathrm{P}_{b^{(i)}})$ so that $f(x^{(i)}) < v_{\min}(b^{(i)}) + \epsilon$. The point $x = \lambda x^{(1)} + (1 - \lambda)x^{(2)}$ is feasible for the problem (P_b), where $b = \lambda b^{(1)} + (1 - \lambda)b^{(2)}$. Therefore,

$$\begin{aligned} v_{\min}(\lambda b^{(1)} + (1 - \lambda)b^{(2)}) &\le f(x) \le \lambda f(x^{(1)}) + (1 - \lambda) f(x^{(2)}) \\ &< \lambda v_{\min}(b^{(1)}) + (1 - \lambda) v_{\min}(b^{(2)}) + \epsilon, \end{aligned}$$

and since $\epsilon > 0$ is arbitrary, this shows that the function $v_{\min}$ is convex on U.

11.6 a) $v_{\min} = 2$ for $x = (0, 0)$ b) $v_{\min} = 2$ for $x = (0, 0)$
c) $v_{\min} = \ln 2 - 1$ for $x = (-\ln 2, \frac{1}{2})$ d) $v_{\min} = -5$ for $x = (-1, -2)$
e) $v_{\min} = 1$ for $x = (1, 0)$ f) $v_{\min} = 2\,\mathrm{e}^{1/2} + \frac{1}{4}$ for $x = (\frac{1}{2}, \frac{1}{2})$

11.7 $v_{\min} = 2 - \ln 2$ for $x = (1, 1)$

11.9 $\min \ 50x_1^2 + 80x_1x_2 + 40x_2^2 + 10x_3^2$

$$\text{s.t.} \begin{cases} 0.2x_1 + 0.12x_2 + 0.04x_3 \ge 0.12 \\ x_1 + \quad x_2 + \quad x_3 = 1, \quad x \ge 0 \end{cases}$$

Optimum for $x_1 = x_3 = 0.5$ miljon dollars.

Chapter 12

12.1 All nonempty sets $X(b) = \{x \mid Ax \ge b\}$ of feasible points have the same recession cone, since $\operatorname{recc} X(b) = \{x \mid Ax \ge 0\}$ if $X(b) \ne \emptyset$. Therefore, it follows from Theorem 12.1.1 that the optimal value $v(b)$ is finite if $X(b) \ne \emptyset$. The convexity of the optimal value function v is a consequence of the same theorem, because

$$v(b) = \min\{\langle -b, y\rangle \mid A^T y \le c,\ y \ge 0\},$$

according to the duality theorem.

12.2 E.g. $\min\ x_1 - x_2$ s.t. $\begin{cases} -x_1 \geq 1 \\ x_2 \geq 1,\ x \geq 0 \end{cases}$ and $\max\ y_1 + y_2$ s.t. $\begin{cases} -y_1 \leq 1 \\ y_2 \leq -1,\ y \geq 0 \end{cases}$

12.5 $v_{\max} = \begin{cases} \dfrac{t-3}{t+1} & \text{for } x = \left(-\dfrac{2}{t+1}, \dfrac{t-1}{t+1}\right) \text{ if } t < -2, \\ 5 & \text{for } x = (2,3) \text{ if } t \geq -2. \end{cases}$

Chapter 13

13.1 a) $\min\ 2x_1 - 2x_2 + x_3$ s.t. $\begin{cases} x_1 + x_2 - x_3 - s_1 = 3 \\ x_1 + x_2 - x_3 + s_2 = 2 \\ x_1, x_2, x_3, s_1, s_2 \geq 0 \end{cases}$

b) $\min\ x_1 + 2x_2' - 2x_2''$ s.t. $\begin{cases} x_1 + x_2' - x_2'' - s_1 = 1 \\ x_2' - x_2'' - s_2 = -2 \\ x_1, x_2', x_2'', s_1, s_2 \geq 0 \end{cases}$

13.2 a) $(5,5,0)$ and $(7\frac{1}{2}, 0, 2\frac{1}{2})$ b) $(3,0,0,0)$ and $(0,0,0,3)$

13.3 $\max\ y_1 + 7y_2$ s.t. $\begin{cases} y_1 + y_2 \leq 1 \\ 2y_2 \leq 1 \\ -y_1 + 7y_2 \leq 4 \end{cases}$

13.4 a) $v_{\min} = -1$ for $x = (0,0,4,1)$ b) $v_{\max} = 56$ for $x = (24,0,0,1,11)$
c) $v_{\max} = 30\frac{6}{7}$ for $x = (1\frac{5}{7}, \frac{3}{7}, 0)$ d) $v_{\max} = 23$ for $x = (2,0,3,0,5)$
e) $v_{\min} = -\infty$ f) $v_{\min} = -1\frac{13}{15}$ for $x = (0, \frac{2}{3}, 0, \frac{2}{5})$

13.5 $v_{\min} = -2$ is attained at all points on the line segment between the points $(0,3,1,1,0)$ and $(0,2,2,0,1)$.

13.6 $v_{\max} = 15$ for $x = (2\frac{1}{2}, 2\frac{1}{2}, 2\frac{1}{2}, 0)$

13.8 $v_{\min} = 9$ for $x = (\frac{2}{3}, 1\frac{2}{3}, 1\frac{2}{3})$

13.9 $v_{\min} = -40\frac{3}{5}$ for $x = (-3\frac{3}{5}, 11\frac{4}{5})$

13.10 a) $v_{\min} = 4\frac{1}{4}$ for $x = (\frac{3}{4}, \frac{1}{2}, \frac{3}{4})$ b) $v_{\min} = \frac{4}{5}$ for $x = (0, \frac{2}{5}, 0)$
c) $v_{\min} = 5\frac{7}{12}$ for $x = (1\frac{1}{4}, \frac{11}{12}, 0)$

13.12 $v_{\max} = \begin{cases} 7 & \text{for } x = (3\frac{1}{2}, 0) \text{ if } t \leq 1, \\ 4 + 3t & \text{for } x = (2,3) \text{ if } 1 < t < 2, \\ 5t & \text{for } x = (0,5) \text{ if } t \geq 2. \end{cases}$

13.13 500 pairs of model A and 700 pairs of model B.

13.14 4 liters of milk and 1 loaf. The milk price could rise to 10 SEK/l.

13.17 First, use the algorithm $\mathcal{A}$ on the system consisting of the linear inequalities $Ax \geq b$, $x \geq 0$, $A^T y \leq c$, $y \geq 0$, $\langle c, x\rangle \leq \langle b, y\rangle$. If the algorithm delivers a solution $(\overline{x}, \overline{y})$, then $\overline{x}$ is an optimal solution to the minimization problem because of the complementarity theorem.
If the algorithm instead shows that the system has no solution, then we use the algorithm on the system $Ax \geq b$, $x \geq 0$ to determine whether the minimization problem has feasible points or not. If this latter system has feasible points, then it follows from our first investigation that the dual problem has no feasible points, and we conclude that the objective function is unbounded below, because of the duality theorem.

Chapter 14

14.1 $x_1 = (\frac{4}{9}, -\frac{1}{9})$, $x_2 = (\frac{2}{27}, \frac{2}{27})$, $x_3 = (\frac{8}{243}, -\frac{2}{243})$.

14.3 $hf'(x_k) = f(x_k) - f(x_{k+1}) \to f(\hat{x}) - f(\hat{x}) = 0$ and $hf'(x_k) \to hf'(\hat{x})$. Hence, $f'(\hat{x}) = 0$.

Chapter 15

15.1 $\Delta x_{\mathrm{nt}} = -x \ln x$, $\lambda(f, x) = \sqrt{x} \ln x$, $\|v\|_x = |v|/\sqrt{x}$.

15.2 a) $\Delta x_{\mathrm{nt}} = (\frac{1}{3}, \frac{1}{3})$, $\lambda(f, x) = \sqrt{\frac{1}{3}}$, $\|v\|_x = \frac{1}{2}\sqrt{5v_1^2 + 2v_1v_2 + 5v_2^2}$

b) $\Delta x_{\mathrm{nt}} = (\frac{1}{3}, -\frac{2}{3})$, $\lambda(f, x) = \sqrt{\frac{1}{3}}$, $\|v\|_x = \frac{1}{2}\sqrt{8v_1^2 + 8v_1v_2 + 5v_2^2}$.

15.3 $\Delta x_{\mathrm{nt}} = (v_1, v_2)$, where $v_1 + v_2 = -1 - \mathrm{e}^{-(x_1+x_2)}$,
$\lambda(f, x) = \mathrm{e}^{(x_1+x_2)/2} + \mathrm{e}^{-(x_1+x_2)/2}$, $\|v\|_x = \mathrm{e}^{(x_1+x_2)/2}|v_1 + v_2|$.

15.4 If $\operatorname{rank} A < m$, then $\operatorname{rank} M < m + n$, and if $\mathcal{N}(A) \cap \mathcal{N}(P)$ contains a nonzero vector x, then $M \begin{bmatrix} x \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Hence, the matrix M has no inverse in these cases.
Conversely, suppose that $\operatorname{rank} A = m$, i.e. that $\mathcal{N}(A^T) = \{0\}$, and that $\mathcal{N}(A) \cap \mathcal{N}(P) = \{0\}$. We show that the coefficient matrix M is invertible by showing that the homogeneous system

$$\begin{cases} Px + A^T y = 0 \\ Ax = 0 \end{cases}$$

has no other solutions than the trivial one, $x = 0$ and $y = 0$.
By multiplying the first equation from the left by x^T we obtain

$$0 = x^T Px + x^T A^T y = x^T Px + (Ax)^T y = x^T Px,$$

and since P is positive semidefinite, it follows that $Px = 0$. The first equation now gives $A^T y = 0$. Hence, $x \in \mathcal{N}(A) \cap \mathcal{N}(P)$ and $y \in \mathcal{N}(A^T)$, which means that $x = 0$ and $y = 0$.

15.5 a) By assumption, $\langle v, f''(x)v\rangle \geq \mu\|v\|^2$ if $Av = 0$. Since $AC = 0$, we conclude that

$$\begin{aligned}\langle w, \tilde{f}''(z)w\rangle &= \langle w, C^T f''(x)Cw\rangle = \langle Cw, f''(x)Cw\rangle \geq \mu\|Cw\|^2 \\ &= \mu\langle w, C^T Cw\rangle \geq \mu\sigma\|w\|^2\end{aligned}$$

for all $w \in \mathbf{R}^p$, which shows that the function $\tilde{f}$ is $\mu\sigma$-strongly convex.

b) The assertion follows from a) if we show that the restriction of f to X is a $K^{-2}M^{-1}$-strongly convex function. So assume that $x \in X$ and that $Av = 0$. Then

$$\begin{bmatrix} f''(x) & A^T \\ A & 0 \end{bmatrix}\begin{bmatrix} v \\ 0 \end{bmatrix} = \begin{bmatrix} f''(x)v \\ 0 \end{bmatrix}$$

and due to the bound on the norm of the inverse matrix, we conclude that

$$\|v\| \leq K\|f''(x)v\|.$$

The positive semidefinite second derivative $f''(x)$ has a positive semidefinite square root $f''(x)^{1/2}$ and $\|f''(x)^{1/2}\| = \|f''(x)\|^{1/2} \leq M^{1/2}$. It follows that

$$\begin{aligned}\|f''(x)v\|^2 &= \|f''(x)^{1/2}f''(x)^{1/2}v\|^2 \leq \|f''(x)^{1/2}\|^2\|f''(x)^{1/2}v\|^2 \\ &\leq M\|f''(x)^{1/2}v\|^2 = M\langle v, f''(x)v\rangle,\end{aligned}$$

which inserted in the above inequality results in the inequality

$$\langle v, f''(x)v\rangle \geq K^{-2}M^{-1}\|v\|^2.$$

Chapter 16

16.2 Let P_i denote the projection of $\mathbf{R}^{n_1} \times \cdots \times \mathbf{R}^{n_m}$ onto then ith factor $\mathbf{R}^{n_i}$. Then $f(x) = \sum_{i=1}^m f_i(P_i x)$, so it follows from Theorems 16.1.5 and 16.1.6 that f is self-concordant.

16.3 a) The function g is convex, since $g''(x) = \dfrac{f'(x)^2}{f(x)^2} - \dfrac{f''(x)}{f(x)} + \dfrac{1}{x^2} \geq 0$.

$g'''(x) = -\dfrac{f'''(x)}{f(x)} + 3\dfrac{f'(x)f''(x)}{f(x)^2} - 2\dfrac{f'(x)^3}{f(x)^3} - \dfrac{2}{x^3}$ implies that

$|g'''(x)| \leq 3\dfrac{f''(x)}{x|f(x)|} + 3\dfrac{|f'(x)|f''(x)}{f(x)^2} + 2\dfrac{|f'(x)|^3}{|f(x)|^3} + 2\dfrac{1}{x^3}$.

The inequality $|g'''(x)| \leq 2g''(x)^{3/2}$, which proves that the function g is self-concordant, is now obtained by choosing $a = \sqrt{f''(x)/|f(x)|}$, $b = |f'(x)|/|f(x)|$ and $c = 1/x$ in the equality

$$3a^2b + 3a^2c + 2b^3 + 2c^3 \leq 2(a^2 + b^2 + c^2)^{3/2}.$$

To prove this inequality, we can due to homogeneity assume that

$$a^2 + b^2 + c^2 = 1.$$

Inserting $a^2 = 1 - b^2 - c^2$ into the inequality, we can rewrite it as $(b + c)(3 - (b + c)^2) \leq 2$, which holds since $x(3 - x^2) \leq 2$ for $x \geq 0$.

16.3 b) Let $\phi(t) = F(x_0 + \alpha t, y_0 + \beta t)$ be the restriction of F to an arbitrary line through the point (x_0, y_0) in $\operatorname{dom} F$. We will prove that ϕ is self-concordant, and we have to treat the cases $\alpha = 0$ and $\alpha \neq 0$ separately. If $\alpha = 0$, then $\phi(t) = -\ln(\beta t + a) + b$, where $a = y_0 - f(x_0)$ and $b = -\ln x_0$, so ϕ is self-concordant in this case.
To prove the case $\alpha \neq 0$, we note that $f(x) - Ax - B$ satisfies the assumptions of the exercise for each choice of the constants A and B, and hence $h(x) = -\ln(Ax + B - f(x)) - \ln x$ is self-concordant according to the result in a). But $\phi(t) = h(\alpha t + x_0)$, where $A = \beta/\alpha$ and $B = y_0 - \beta x_0/\alpha$. Thus, ϕ is self-concordant.

16.6 a) Set $\lambda = \lambda(f, x)$ and use the inequalities (16.7) and (16.6) in Theorem 16.3.2 with $y = x^+$ and $v = x^+ - x = (1 + \lambda)^{-1}\Delta x_{\mathrm{nt}}$. This results in the inequality

$$\begin{aligned}
\langle f'(x^+), w\rangle &\leq \langle f'(x), w\rangle + \frac{1}{1+\lambda}\langle f''(x)\Delta x_{\mathrm{nt}}, w\rangle + \frac{\lambda^2\|w\|_x}{(1+\lambda)^2(1-\lambda/(1+\lambda))} \\
&= \langle f'(x), w\rangle - \frac{1}{1+\lambda}\langle f'(x), w\rangle + \frac{\lambda^2}{1+\lambda}\|w\|_x \\
&= \frac{\lambda}{1+\lambda}\langle f'(x), w\rangle + \frac{\lambda^2}{1+\lambda}\|w\|_x \\
&\leq \frac{\lambda}{1+\lambda}\lambda\|w\|_x + \frac{\lambda^2}{1+\lambda}\|w\|_x = \frac{2\lambda^2}{1+\lambda}\|w\|_x \\
&\leq \frac{2\lambda^2\|w\|_{x^+}}{(1+\lambda)(1-\lambda/(1+\lambda))} = 2\lambda^2\|w\|_{x^+}
\end{aligned}$$

with $\lambda(f, x^+) \leq 2\lambda^2$ as conclusion.

Chapter 18

18.1 Follows from Theorems 18.1.3 and 18.1.2.

18.2 To prove the implication $\|v\|_x^* < \infty \Rightarrow v \in \mathcal{N}(f''(x))^\perp$ we write v as $v = v_1 + v_2$ with $v_1 \in \mathcal{N}(f''(x))$ and $v_2 \in \mathcal{N}(f''(c))^\perp$, noting that $\|v_1\|_x = 0$. Hence $\|v\|_1^2 = \langle v_1, v_1\rangle = \langle v, v_1\rangle \leq \|v\|_x^*\|v_1\|_x = 0$, and we conclude that $v_1 = 0$. This proves that v belongs to $\mathcal{N}(f''(x))^\perp$.
Given $v \in \mathcal{N}(f''(x))^\perp$ there exists a vector u such that $v = f''(x)u$. We

shall prove that $\|v\|_x^* = \|u\|_x$. From this follows that $\|v\|_x^* < \infty$ and that $\|\cdot\|_x^*$ is a norm on the subspace $\mathcal{N}(f''(x))^\perp$ of $\mathbf{R}^n$.
Let $w \in \mathbf{R}^n$ be arbitrary. By Cauchy–Schwarz's inequality,

$$\begin{aligned}\langle v, w\rangle &= \langle f''(x)u, w\rangle = \langle f''(x)^{1/2}u, f''(x)^{1/2}w\rangle \\ &\le \|f''(x)^{1/2}u\|\|f''(x)^{1/2}w\| = \|u\|_x\|v\|_x,\end{aligned}$$

and this implies that $\|v\|_x^* \le \|u\|_x$. Suppose $v \neq 0$. Then u does not belong to $\mathcal{N}(f''(x))$, which means that $\|u\|_x \neq 0$, and for $w = u/\|u\|_x$ we get the identity

$$\langle v, w\rangle = \|u\|_x^{-1}\langle f''(x)^{1/2}u, f''(x)^{1/2}u\rangle = \|u\|_x^{-1}\|f''(x)^{1/2}u\|^2 = \|u\|_x,$$

which proves that $\|v\|_x^* = \|u\|_x$. If on the other hand $v = 0$, then u is a vector in $\mathcal{N}(f''(x))$ so we have $\|v\|_x^* = \|u\|_x$ in this case, too.

18.3 a) Differentiate the equality $f(tx) = f(x) - \nu \ln t$ with respect to x.

b) Differentiate the equality obtained in a) with respect to t and then take $t = 1$.

c) Since X does not contain any line, f is a non-degenerate self-concordant function, and it follows from the result in b) that x is the unique Newton direction of f at the point x. By differentiating the equality $f(tx) = f(x) - \nu \ln t$ with respect to t and then putting $t = 1$, we obtain $\langle f'(x), x\rangle = -\nu$. Hence

$$\nu = -\langle f'(x), x\rangle = -\langle f'(x), \Delta x_{\mathrm{nt}}\rangle = \lambda(f, x)^2.$$

18.5 Define $g(x, x_{n+1}) = (x_1^2 + \cdots + x_n^2) - x_{n+1}^2 = \|x\|^2 - x_{n+1}^2$, so that

$$f(x) = -\ln(-g(x, x_{n+1})),$$

and let $w = (v, v_{n+1})$. Then

$$\begin{aligned}Dg &= Dg(x, x_{n+1})[w] = 2(\langle v, x\rangle - x_{n+1}v_{n+1}),\\ D^2g &= D^2g(x, x_{n+1})[w, w] = 2(\|v\|^2 - v_{n+1}^2),\\ D^3g &= D^3g(x, x_{n+1})[w, w, w] = 0,\\ Df &= Df(x, x_{n+1})[w] = -\frac{1}{g}Dg\\ D^2f &= D^2f(x, x_{n+1})[w, w] = \frac{1}{g^2}\big((Dg)^2 - gD^2g\big),\\ D^3f &= D^3f(x, x_{n+1})[w, w, w] = \frac{1}{g^3}\big(-2(Dg)^3 + 3gDgD^2g\big).\end{aligned}$$

Consider the difference

$$\Delta = (Dg)^2 - gD^2g = 4(\langle x, v\rangle - x_{n+1}v_{n+1})^2 + 2(x_{n+1}^2 - \|x\|^2)(\|v\|^2 - v_{n+1}^2).$$

Since $x_{n+1} > \|x\|$, we have $\Delta \geq 0$ if $|v_{n+1}| \leq \|v\|$. So suppose that $|v_{n+1}| > \|v\|$. Then

$$\begin{aligned} |x_{n+1}v_{n+1} - \langle x, v\rangle| &\geq x_{n+1}|v_{n+1}| - |\langle x, v\rangle| \\ &\geq x_{n+1}|v_{n+1}| - \|x\|\|v\| \geq 0, \end{aligned}$$

and it follows that

$$\begin{aligned} \Delta &\geq 4(x_{n+1}|v_{n+1}| - \|x\|\|v\|)^2 + 2(x_{n+1}^2 - \|x\|^2)(\|v\|^2 - v_{n+1}^2) \\ &= 2(x_{n+1}|v_{n+1}| - \|x\|\|v\|)^2 + 2(x_{n+1}\|v\| - \|x\||v_{n+1}|)^2 \geq 0. \end{aligned}$$

This shows that $D^2 f = \Delta/g^2 \geq 0$, so f is a convex function.

To prove that the function is self-concordant, we shall show that

$$4(D^2 f)^3 - (D^3 f)^2 \geq 0.$$

After simplification we obtain

$$4(D^2 f)^3 - (D^3 f)^2 = g^{-4}(D^2 g)^2(3(Dg)^2 - 4gD^2 g),$$

and the problem has now been reduced to showing that the difference

$$\begin{aligned} \Delta' &= 3(Dg)^2 - 4gD^2 g \\ &= 12(\langle x, v\rangle - x_{n+1}v_{n+1})^2 + 8(x_{n+1}^2 - \|x\|^2)(\|v\|^2 - v_{n+1}^2) \end{aligned}$$

is nonnegative. This is obvious if $|v_{n+1}| \leq \|v\|$, and if $|v_{n+1}| > \|v\|$ then we get in a similar way as above

$$\begin{aligned} \Delta' &\geq 12(x_{n+1}|v_{n+1}| - \|x\|\|v\|)^2 + 8(x_{n+1}^2 - \|x\|^2)(\|v\|^2 - v_{n+1}^2) \\ &= 4(x_{n+1}|v_{n+1}| - \|x\|\|v\|)^2 + 8(x_{n+1}\|v\| - \|x\||v_{n+1}|)^2 \geq 0. \end{aligned}$$

18.6 Let $w = (u, v)$ be an arbitrary vector in $\mathbf{R}^2$. Writing $a = 1/(y - f(x))$, $b = -1/x$, $A = f'(x)$ and $B = f''(x)$ for short, where $a > 0$ and $B \geq 0$, we obtain

$$\begin{aligned} DF(x, y)[w] &= (aA + b)u - av \\ D^2F(x, y)[w, w] &= (aB + a^2A^2 + b^2)u^2 - 2a^2Auv + a^2v^2, \end{aligned}$$

and

$$\begin{aligned} &2D^2F(x, y)[w, w] - \big(DF(x, y)[w]\big)^2 \\ &\quad = a^2A^2u^2 + b^2u^2 + a^2v^2 + 2abuv - 2a^2Auv - 2abAu^2 + 2aBu^2 \\ &\quad = (aAu - bu - av)^2 + 2aBu^2 \geq 0. \end{aligned}$$

So F is a 2-self-concordant function.

18.7 Use the previous exercise with $f(x) = x \ln x$.

18.8 Taking $f(x) = -\ln x$ in exercise 18.5, we see that

$$F(x, y) = -\ln(\ln x + y) - \ln x$$

is a 2-self-concordant barrier to the closure of the region $-y < \ln x$. Since $G(x, y) = F(y, -x)$, it then follows from Theorem 18.1.3 that G is a 2-self-concordant barrier to the region $y \geq \mathrm{e}^x$.

Index