Optimisation

Embryonic notes for the course A4B33OPT
This text is incomplete and may be added to and improved during the semester.
This version: **25th February 2015**

Tomáš Werner
Translated from Czech to English by Libor Špaček

Czech Technical University
Faculty of Electrical Engineering

Contents

Introduction

Optimisation (more precisely mathematical optimisation) attempts to solve the minimisation (or maximisation) of functions of many variables in the presence of possible constraint conditions. This formulation covers many practical problems in engineering and in the natural sciences; often we want to do something 'in the best possible way' in the 'given circumstances'. It is very useful for an engineer to be able to recognise optimisation problems in various situations. Optimisation, also called *mathematical programming*, is a branch of applied mathematics, combining aspects of mathematical analysis, linear algebra and computer science. It is a modern, fast developing subject.

Examples of some tasks leading to optimisation problems:

- Approximate some observed functional dependence by a function of a given class (of functions, e.g. a polynomial).
- Choose some shares to invest in, so that the expected return is large and the expected risk is small.
- Build a given number of shops around a town so that every inhabitant lives near one.
- Determine the sequence of control signals to a robot, so that its hand moves from place A to place B along the shortest path (or in the shortest time, using the minimum energy, etc.) and without a collision.
- Regulate the intake of gas to a boiler so that the temperature in the house remains nearly optimal.
- Design a printed circuit board in such a way that the length of the connections is minimal.
- Find the shortest path through a computer network.
- Find the best connection from place A to place B using bus/train timetables.
- Design the best school timetable.
- Build a bridge of a given carrying capacity using the least amount of building materials.
- Train a neural network.

Apart form the engineering practice, optimisation is also important in natural sciences. Most physical laws can be formulated in terms of some variable attaining an extreme value. Living organisms are, at any given moment, consciously or unconsciously, solving a number of optimisation problems – e.g. they are choosing the best possible behaviours.

You will not learn on this course how to solve all these problems but you will learn how to recognise the type of a problem and its difficulty. You will gain the foundations for solving the easiest problems and for an approximate solution of the more difficult ones. The spectrum of problems that you will be able to solve will be further significantly enhanced after the completion of the follow-up course *Combinatorial Optimisation*.

Goal: To achieve a thorough understanding of vector calculus, including both problem solving and theoretical aspects. The orientation of the course is toward the problem aspects, though we go into great depth concerning the theory behind the computational skills that are developed.

This goal shows itself in that we present no 'hard' proofs, though we do present 'hard' theorems. This means that you are expected to understand these theorems as to their hypotheses and conclusions but not to understand or even see their proofs. However, 'easy' theorems are discussed throughout the course, and you are expected to understand their proofs completely. For example, it is a hard theorem that a continuous real-valued function defined on a closed interval of the real numbers attains its maximum value. But it is an easy theorem that if that maximum value is taken at an interior point of the interval and if the function is differentiable there, then its derivative equals to zero at that point.

You will also learn to grasp quite a large number of important definitions.

Chapter 1

Formalising Optimisation Tasks

1.1 Mathematical notation

Bold **font** in these notes indicates a newly defined concept, which you should strive to comprehend and memorise. Words in *italics* mean either emphasis or a newly introduced concept that is generally known. Paragraphs, sentences, proofs, examples and exercises marked by a star ($\star$) are elaborations (and thus more difficult) and not essential for the examination.

We now review mathematical notation used in these notes. The reader ought to become thoroughly familiar with it.

1.1.1 Sets

We will be using the standard sets notation:

$\{a_1, \ldots, a_n\}$	a set with elements $a_1, \ldots, a_n$
$a \in A$	element a belongs to set A (or a is an element of A)
$A \subseteq B$	A is a subset of set B, i.e., every element of A belongs to B
$A = B$	set A equals to set B, then $A \subseteq B$ and also $B \subseteq A$
$\{\, a \in A \mid \varphi(a) \,\}$	set of elements of A with property φ. Sometimes we abbreviate this as $\{\, a \mid \varphi(a) \,\}$
$A \cup B$	union of sets, set $\{\, a \mid a \in A$ or $a \in B \,\}$
$A \cap B$	intersection of sets, set $\{\, a \mid a \in A$ and also zároveň $a \in B \,\}$
$(a_1, \ldots, a_n)$	ordered n-tuple of elements $a_1, \ldots, a_n$
$A \times B$	cartesian product of sets, set of all pairs $\{\, (a, b) \mid a \in A,\ b \in B \,\}$
A^n	cartesian product of n identical sets, $A^n = A \times \cdots \times A$ (n-krát)
$\emptyset$	empty set
iff	if and only if ($\iff$)

Names of sets will be denoted by inclined capital letters, e.g. A or X. Numerical sets will be written as follows:

$\mathbb{N}$	set of natural numbers
$\mathbb{Z}$	set of integers
$\mathbb{Q}$	set rational numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}_+$	set of non-negative real numbers
$\mathbb{R}_{++}$	set of positive real numbers
$[x_1, x_2]$	closed real interval (set $\{ x \in \mathbb{R} \mid x_1 \leq x \leq x_2 \}$)
(x_1, x_2)	open real interval (set $\{ x \in \mathbb{R} \mid x_1 < x < x_2 \}$)
$\mathbb{C}$	set of complex numbers

1.1.2 Mappings

A mapping from set A to set B is written as

$$f\colon A \to B. \tag{1.1}$$

Mapping can be imagined as a 'black box' which associates each element $a \in A$ (in domain A) with exactly one element $b = f(a) \in B$ (in codomain B). The formal definition is as follows: subset f of the cartesian product $A \times B$ (i.e. *relation*) is called *mapping*, if $(a, b) \in f$ and $(a, b') \in f$ implies $b = b'$. Even though *mapping* (*map*) means exactly the same as *function*, the word 'function' is normally used only for mapping into numerical sets (e.g. $B = \mathbb{R}$, $\mathbb{Z}$, $\mathbb{C}$ etc.).

The set of all images (codomain elements b) of all arguments (domain elements a) with property φ, is abbreviated as:

$$\{ f(a) \mid a \in A,\, \varphi(a) \} = \{ b \in B \mid b = f(a),\, a \in A,\, \varphi(a) \}$$

or just $\{ f(a) \mid \varphi(a) \}$, when A is clear from the context. Here $\varphi(a)$ is a logical expression which can be true or false. For example, set $\{ x^2 \mid -1 < x < 1 \}$ is half-closed interval $[0, 1)$. The domain set A in the mapping f is written $f(A) = \{ f(a) \mid a \in A \}$.

1.1.3 Functions and mappings of several real variables

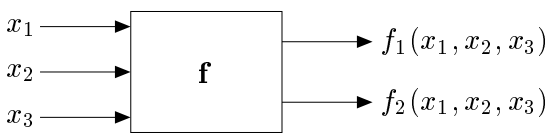
An ordered n-tuple $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ of real numbers is called (n-dimensional) **vector**.

$$\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m \tag{1.2}$$

denotes a mapping, which associates with vector $\mathbf{x} \in \mathbb{R}^n$ vector

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(x_1, \ldots, x_n) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})) = (f_1(x_1, \ldots, x_n), \ldots, f_m(x_1, \ldots, x_n)) \in \mathbb{R}^m,$$

where $f_1, \ldots, f_m\colon \mathbb{R}^n \to \mathbb{R}$ are the *components* of the mapping. We can write also $\mathbf{f} = (f_1, \ldots, f_m)$. The following figure illustrates the mapping $\mathbf{f}\colon \mathbb{R}^3 \to \mathbb{R}^2$:



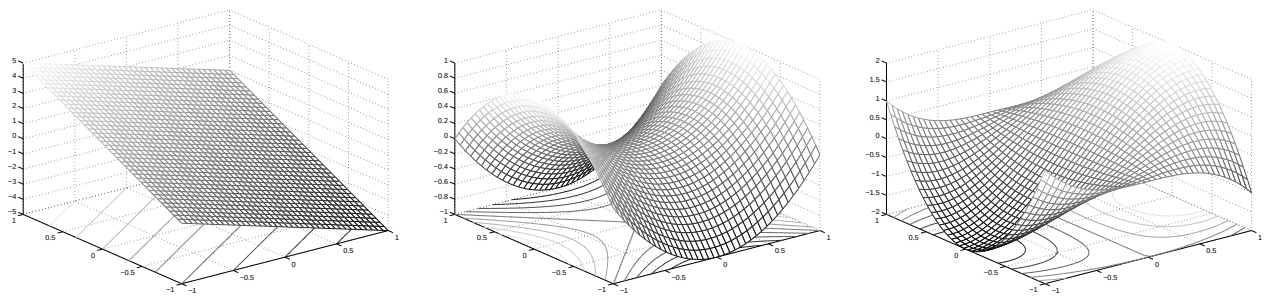
When $m = 1$ then the codomain values are scalars, written in italics, as follows: f. When $m > 1$ then the codomain values are vectors, written in bold font, $\mathbf{f}$. Even though strictly speaking the words 'function' and 'mapping' mean one and the same thing, it is common to talk about a *function* when $m = 1$ and a *mapping* when $m > 1$.

Definitions and statements in this text will be formulated so as to apply to functions and mappings whose definition domain is the entire $\mathbb{R}^n$. However, this need not always be the case, e.g. the definition domain of the function $f(x) = \sqrt{1 - x^2}$ is the interval $[-1, 1] \subset \mathbb{R}$. Nonetheless, the above default domain simplifies the notation and the reader should find it easy to generalise any given statement to a different definition domain.

We use the following terms for functions $f\colon \mathbb{R}^n \to \mathbb{R}$:

- **Graph** of the function is the set $\{\, (\mathbf{x}, y) \in \mathbb{R}^{n+1} \mid \mathbf{x} \in \mathbb{R}^n,\ y = f(\mathbf{x}) \,\}$.
- **Contour** of the value y of the function is the set $\{\, \mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) = y \,\}$.

The following figure shows examples of the graph and the contours of functions of two variables on the rectangle $[-1, 1]^2$ (created by the matlab command `meshc`):



$f(x, y) = -2x + 3y$ $\qquad$ $f(x, y) = x^2 - y^2$ $\qquad$ $f(x, y) = 3x - x^3 - 3xy^2$

1.2 Minimum of a function over a set

Given set $Y \subseteq \mathbb{R}$, we call $y \in Y$ its *smallest element* (or *minimum*), iff $y \leq y'$ for all $y' \in Y$. Not all subsets of $\mathbb{R}$ have the smallest element (e.g. interval $(0, 1]$ does not).

Take function $f\colon X \to \mathbb{R}$, where X is an arbitrary set. Denote codomain Y of X by function f

$$Y = f(X) = \{\, f(x) \mid x \in X \,\} \subseteq \mathbb{R}$$

When set Y has the smallest element, we define

$$\min_{x \in X} f(x) = \min Y$$

called *minimum of the function f over the set X*. In this case there exists at least one element $x \in X$, so that $f(x) = \min Y$. We say that the function *attains minimum* at the point x. The subset of set X, at which the minimum is reached, is denoted by the symbol 'argument of the minimum'

$$\operatorname*{argmin}_{x \in X} f(x) = \{\, x \in X \mid f(x) = \min Y \,\}.$$

We define the maximum of function over a set similarly. Minima and maxima of a function are generically called its *extrema* or *optima*.

Example 1.1.

- $\min\limits_{x\in\mathbb{R}} |x-1| = \min\{\, |x-1| \mid x \in \mathbb{R}\,\} = \min \mathbb{R}_+ = 0, \ \operatorname*{argmin}\limits_{x\in\mathbb{R}} |x-1| = \{1\}$
- Let $f(x) = \max\{|x|, 1\}$. Then $\operatorname*{argmin}\limits_{x\in\mathbb{R}} f(x) = [-1, 1]$.
- Let $(a_1, a_2, \ldots, a_5) = (1, 2, 3, 2, 3)$. Then[1] $\max\limits_{i=1}^{5} a_i = 3, \ \operatorname*{argmax}\limits_{i=1}^{5} a_i = \{3, 5\}$. □

1.3 The general problem of continuous optimisation

Optimisation problems are formulated as searching for the minimum of a given real function $f\colon X \to \mathbb{R}$ over a given set X. This formalisation is very general, as the set X is quite arbitrary. There are three broad categories of problems:

- *Combinatorial optimisation*, when set X is finite (even though possibly very large). Its elements can be, for example, paths in a graph, configuration of the Rubik's cube or text strings of finite lengths. Examples are finding the shortest path through the graph or the problem of the travelling salesman.
- *Continuous optimisation* when set X contains real numbers or real vectors. An example is linear programming.
- *Variational calculus* when set X contains real functions. An example is to find the planar curve of given length which encloses the maximum area.

This course addresses continuous optimisation. The general problem of continuous optimisation is usually formulated as follows: we seek the minimum of function $f\colon \mathbb{R}^n \to \mathbb{R}$ on set $X \subseteq \mathbb{R}^n$, which contains all solutions $(x_1, \ldots, x_n)$ of a set of m inequalities and ℓ equations.

$$g_i(x_1, \ldots, x_n) \le 0, \quad i = 1, \ldots, m \tag{1.3a}$$
$$h_i(x_1, \ldots, x_n) = 0, \quad i = 1, \ldots, \ell \tag{1.3b}$$

for given functions $g_1, \ldots, g_m, h_1, \ldots, h_\ell\colon \mathbb{R}^n \to \mathbb{R}$. In vector notation we write:

$$X = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{g}(\mathbf{x}) \le \mathbf{0},\ \mathbf{h}(\mathbf{x}) = \mathbf{0}\,\},$$

where $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{h}\colon \mathbb{R}^n \to \mathbb{R}^\ell$ and $\mathbf{0}$ denote null vectors of an appropriate dimension. We seek the minimum of given function $f\colon \mathbb{R}^n \to \mathbb{R}$ on set X. That is written also as

$$\begin{aligned} \min \quad & f(x_1, \ldots, x_n) \\ \text{condition to} \quad & g_i(x_1, \ldots, x_n) \le 0, \quad i = 1, \ldots, m \\ & h_i(x_1, \ldots, x_n) = 0, \quad i = 1, \ldots, \ell. \end{aligned} \tag{1.4}$$

Example 1.2. A shepherd has 100 metres of fencing. He wants to make a sheep paddock that is as large (in area) as possible. It is to be a rectangle whose three sides will be formed by the fence and the remaining side by a river, as sheep cannot swim.

[1] $\max\limits_{i=1}^{5} a_i$ is more often written as $\max\limits_{i=1,\ldots,5} a_i$. We use the first method, following an analogy with the standard notation $\sum\limits_{i=1}^{5} a_i$.

Let's call the sides of the rectangle x, y. We are solving the problem

$$\begin{aligned} \max \quad & xy \\ \text{condition to} \quad & 2x + y = 100 \end{aligned}$$

or

$$\max\{\, xy \mid x \in \mathbb{R},\ y \in \mathbb{R},\ 2x + y = 100 \,\}.$$

Here we have $n = 2$, $m = 0$, $\ell = 1$.

We know how to solve this problem easily. From the constraint $2x + y = 100$ we have $y = 100 - 2x$, therefore instead of the original problem, we can solve the equivalent problem without constraints:

$$\min_{x \in \mathbb{R}} x(100 - 2x).$$

The minimum of the quadratic function $x(100 - 2x)$ is easily found by means of analysis of functions of a single variable. □

Example 1.3. Find the pair of nearest points in the plane. One point lies on the circle of unit radius with the centre at the origin and the second point lies in the square with the centre at point $(2, 2)$ and the side of one unit. This problem can, of course, be solved easily by some thought. However, let's write it in the form (1.4).

Point (x_1, x_2) on the circle satisfies $x_1^2 + x_2^2 = 1$. Point (x_3, x_4) in the square satisfies $-\frac{1}{2} \le x_3 - 2 \le \frac{1}{2}$, $-\frac{1}{2} \le x_4 - 2 \le \frac{1}{2}$. We have $n = 4$, $m = 4$, $\ell = 1$ and

$$X = \{\, (x_1, x_2, x_3, x_4) \mid x_1^2 + x_2^2 - 1 = 0,\ \tfrac{3}{2} - x_3 \le 0,\ x_3 - \tfrac{5}{2} \le 0,\ \tfrac{3}{2} - x_4 \le 0,\ x_4 - \tfrac{5}{2} \le 0 \,\}.$$

We are solving the problem

$$\begin{aligned} \min \quad & \sqrt{(x_1 - x_3)^2 + (x_2 - x_4)^2} \\ \text{subject to} \quad & x_1^2 + x_2^2 - 1 = 0 \\ & \tfrac{3}{2} - x_3 \le 0 \\ & x_3 - \tfrac{5}{2} \le 0 \\ & \tfrac{3}{2} - x_4 \le 0 \\ & x_4 - \tfrac{5}{2} \le 0 \end{aligned}$$

□

In mathematical analysis, the solution of problem (1.4) is called *extrema of function f, subject to constraints* (1.3). When the constraints are missing, we talk about *free extrema* of function f. Mathematical optimisation is commonly using somewhat different terminology:

- Function f is called the *objective* (also penalty, cost, criteria) function.
- elements of the set X are called *admissible solutions*, which is somewhat contradictory, as they need not be the solutions of the problem (1.4). elements of the set $\operatorname{argmin}_{\mathbf{x} \in X} f(\mathbf{x})$ are then called *optimal solutions*.
- Equations and inequalities (1.3) are called *constraining conditions*, in short *constraints*.
- Constraints (1.3a), respectively (1.3b), are called constraints of *inequality type*, respectively *equality type*. When the constraints are missing ($m = \ell = 0$), then we talk about *unconstrained* optimisation.
- When the set X of admissible solutions is empty (constraints are in a conflict with each other), then the problem is called *inadmissible*.
- When the objective function can grow above any bounds while fulfilling the constraints, then the problem is called *unbounded*.

1.4 Exercises

1.1. Solve the following problems. Express the textual problem descriptions in the form of (1.4). All that is necessary is some common sense and the derivatives of functions of a single variable.

a) $\min\{\, x^2 + y^2 \mid x > 0,\ y > 0,\ xy \geq 1 \,\}$

b) $\min\{\, (x-2)^2 + (y-1)^2 \mid x^2 \leq 1,\ y^2 \leq 1 \,\}$

c) You are to make a cardboard box with the volume of 72 litres, whose length is twice its width. What will be its dimensions using the minimum amount of cardboard? The thickness of the sides is negligible.

d) What are the dimensions of a cylinder with the unit volume and the minimum surface area?

e) Find the dimensions of a half-litre beer glass that requires the minimum amount of glass. The thickness of the glass is uniform.

f) Find the area of the largest rectangle inscribed inside a semi-circle of radius 1.

g) A rectangle in a plane has one corner at the origin and another corner lies on the curve $y = x^2 + x^{-2}$. For what value of x will its area be minimal? Can its area be arbitrarily large?

h) Find the point in the plane, nearest to the point $(3, 0)$ and lying on the parabola given by the equation $y = x^2$.

i) One hectare plot (10K square metres) of rectangular shape is to be surrounded on three sides by a hedge that costs 1000 crowns per metre and on the remaining side by an ordinary fence that costs 500 crowns per metre. What will be the cheapest dimensions for the plot?

j) x, y are numbers in the interval $[1, 5]$, such that their sum is 6. Find such numbers so that xy^2 is (a) minimal and (b) maximal.

k) We seek the n-tuple of numbers $x_1, \ldots, x_n \in \{-1, +1\}$, such that their product is positive and their sum is minimal. As your result, write down a formula (as simple as possible), giving the value of this sum for any n.

l) *Rat biathlon.* A rat stands on the bank of a circular pond with unit radius. The rat wants to reach the opposite point on the bank of the pond. It swims with velocity v_1 and runs with velocity v_2. It wants to get there as quickly as possible by swimming, running or a combination of both. What path will it choose? The rat's strategy can change depending on different relative values of v_1 and v_2. Solve this problem for all combinations of these two values.

Chapter 2

Matrix Algebra

Real **matrix** of dimensions $m \times n$ is the table

$$\mathbf{A} = [a_{ij}] = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix},$$

where a_{ij} are the **elements** of the matrix. Matrix can also be understood as the mapping $\{1, \ldots, m\} \times \{1, \ldots, n\} \to \mathbb{R}$. The set of all real matrices of dimensions $m \times n$ (i.e., s m rows and n columns) is written as $\mathbb{R}^{m \times n}$.

We will use the following terminology:

- When $m = n$ the matrix is called **square** and for $m \neq n$ **rectangular**, while for $m < n$ it is **wide** and for $m > n$ it is **narrow**.
- **Diagonal elements** of the matrix are elements $a_{11}, \ldots, a_{pp}$, where $p = \min\{m, n\}$. A matrix is **diagonal**, when all non-diagonal elements are zero (this applies to both square and rectangular matrices). When $\mathbf{A}$ is square diagonal ($m = n$), we write $\mathbf{A} = \mathrm{diag}(a_{11}, a_{22}, \ldots, a_{nn})$.
- **Zero matrix** has all elements zero, written $\mathbf{0}_{m \times n}$ (when the dimensions are clear from the context, then simply $\mathbf{0}$).
- **identity matrix** is square diagonal and its diagonal elements are all 1s, written $\mathbf{I}_n$ (when the dimensions are clear from the context, then simply $\mathbf{I}$).
- A matrix can be composed of several **sub-matrices** (sometimes also called **blocks**), e.g.:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \end{bmatrix}, \quad \begin{bmatrix} sub - matrices \mathbf{A} \\ \mathbf{B} \end{bmatrix}, \quad \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}, \quad \begin{bmatrix} \mathbf{A} & \mathbf{I} \\ \mathbf{0} & \mathbf{D} \end{bmatrix}. \tag{2.1}$$

  The dimensions of the individual blocks must be compatible. The dimensions of the identity matrix $\mathbf{I}$ and the zero matrix $\mathbf{0}$ in the fourth example are determined by the dimensions of the matrices $\mathbf{A}$ and $\mathbf{D}$.

2.1 Matrix operations

The following operations are defined on the matrices:

- The product of scalar[1] $\alpha \in \mathbb{R}$ and matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is the matrix $\alpha\mathbf{A} = [\alpha a_{ij}] \in \mathbb{R}^{m\times n}$.
- Addition of matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$ is the matrix $\mathbf{A} + \mathbf{B} = [a_{ij} + b_{ij}] \in \mathbb{R}^{m\times n}$.
- **Matrix product** of $\mathbf{A} \in \mathbb{R}^{m\times p}$ and $\mathbf{B} \in \mathbb{R}^{p\times n}$ is the matrix $\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m\times n}$ with elements
$$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}. \tag{2.2}$$

Properties of the matrix product:

- $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
- $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$ a $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
- $\mathbf{AI}_n = \mathbf{A} = \mathbf{I}_m\mathbf{A}$
- $(\alpha\mathbf{A})\mathbf{B} = \mathbf{A}(\alpha\mathbf{B}) = \alpha(\mathbf{AB})$ (We might be tempted to think that the expression $\alpha\mathbf{A}$ is also a matrix product, where the scalar $\alpha \in \mathbb{R}$ is considered to be a matrix of dimension 1×1. However, this is not the case because the the inner dimensions of matrices would be generally different.)

Generally it is not true that $\mathbf{AB} = \mathbf{BA}$ (square matrices are generally non-commutitative).

It is useful to remember the following rule for the multipication of matrices composed of blocks
$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{AX} + \mathbf{BY} \\ \mathbf{CX} + \mathbf{DY} \end{bmatrix}.$$

2.2 Transposition and symmetry

The transpose of matrix $\mathbf{A} = [a_{ij}] \in \mathbb{R}^{m\times n}$ is written as $\mathbf{A}^T = [a_{ji}] \in \mathbb{R}^{n\times m}$. The properties of transposition are:

- $(\alpha\mathbf{A})^T = \alpha\mathbf{A}^T$
- $(\mathbf{A}^T)^T = \mathbf{A}$
- $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$
- $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$

A square matrix is called

- **symmetric**, when $\mathbf{A}^T = \mathbf{A}$, i.e., $a_{ij} = a_{ji}$,
- **skew-symmetric**, when $\mathbf{A}^T = -\mathbf{A}$, i.e., $a_{ij} = -a_{ji}$ (from which it necessarily follows that $a_{ii} = 0$)

2.3 Rank and inversion

Rank of a matrix is the size of the largest subset of its linearly independent columns. In other words, it is the dimension of the linear envelope of the matrix columns. Rank is written as $\operatorname{rank}\mathbf{A}$. The following holds (but it is not easy to prove)
$$\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{A}^T, \tag{2.3}$$

[1] The term *scalar* in the real matrix algebra denotes a real number. More precisely, considering the set of all matrices of dimensions $m \times n$ as a linear space, then it is the scalar of this linear space.

thus instead of using the columns, it is equivalently possible to define the rank using the rows. It follows that for any matrix

$$\operatorname{rank} \mathbf{A} \leq \min\{m, n\}. \tag{2.4}$$

When $\operatorname{rank} \mathbf{A} = \min\{m, n\}$, we say that the matrix is of **full rank**. A square matrix of full rank is called **regular**, otherwise it is said to be **singular**.

When matrices $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{B} \in \mathbb{R}^{m \times n}$ satisfy

$$\mathbf{AB} = \mathbf{I}, \tag{2.5}$$

then matrix $\mathbf{B}$ is the **right inverse** of matrix $\mathbf{A}$ and matrix $\mathbf{A}$ is the **left inverse** of matrix $\mathbf{B}$. The right or the left inverse need not exist or they need not be unique. For example, when $m < n$, then the equality (2.5) never holds (why?). The right inverse of matrix $\mathbf{A}$ exists iff the rows of $\mathbf{A}$ are linearly independent. The left inverse of matrix $\mathbf{B}$ exists iff the columns of $\mathbf{B}$ are linearly independent.

For $m = n$ (square matrix $\mathbf{A}$), its right inverse exists iff $\mathbf{A}$ is regular (this is why a regular matrix is also called **invertible**). In this case it is unique and equal to the left inverse of the matrix $\mathbf{A}$. Then we talk only about an **inverse** of matrix $\mathbf{A}$ and denote it as $\mathbf{A}^{-1}$. Properties of an inverse:

- $\mathbf{AA}^{-1} = \mathbf{I} = \mathbf{A}^{-1}\mathbf{A}$
- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
- $(\alpha\mathbf{A})^{-1} = \alpha^{-1}\mathbf{A}^{-1}$
- $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$, which is abbreviated to $\mathbf{A}^{-T}$.

2.4 Determinants

Determinant is the function $\mathbb{R}^{n \times n} \to \mathbb{R}$ (i.e. it associates a scalar with a square matrix) defined as

$$\det \mathbf{A} = \sum_{\sigma} \operatorname{sgn} \sigma \prod_{i=1}^{n} a_{i\,\sigma(i)}, \tag{2.6}$$

where we are adding over all permutations n of elements σ: $\{1, \ldots, n\} \to \{1, \ldots, n\}$, where $\operatorname{sgn} \sigma$ denotes the sign of each permutation. Some properties of determinants:

- Determinant is a multilinear function of the matrix columns, i.e., it is a linear function of an arbitrary column when all the other columns are constant.
- Determinant is an alternating function of the matrix columns, i.e., swapping two neighbouring columns swaps the sign of the determinant.
- $\det \mathbf{I} = 1$
- $\det \mathbf{A} = 0$ iff $\mathbf{A}$ is singular
- $\det \mathbf{A}^T = \det \mathbf{A}$
- $\det(\mathbf{AB}) = (\det \mathbf{A})(\det \mathbf{B})$
- $\det \mathbf{A}^{-1} = (\det \mathbf{A})^{-1}$ (it follows from the above for $\mathbf{B} = \mathbf{A}^{-1}$)

2.5 Matrix of a single column or a single row

A matrix with just one column (i.e. a element of $\mathbb{R}^{n\times 1}$) is also called a **column vector**[2]. A matrix with just one row (i.e. an element of $\mathbb{R}^{1\times m}$) is also called a **row vector**.

The linear space $\mathbb{R}^{n\times 1}$ of all matrices with one column is 'almost the same' as the linear space $\mathbb{R}^n$ of all ordered n-tuples $(x_1, \dots, x_n)$. Therefore it is customary not to distinguish between the two spaces and to move between their two meanings without a warning. We will call the elements

$$\mathbf{x} = \underbrace{(x_1, \dots, x_n)}_{\text{uspořádaná } n\text{-tice}} = \underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}}_{\text{matrix } n\times 1} \in \mathbb{R}^n$$

of this space simply **vectors**. In other words, by the unqualified term *vector* will be meant a *column vector* or equally, an ordered n-tuple of numbers[3].

The cases where vectors occur in the matrix products are important:

- Given matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and vector $\mathbf{x} \in \mathbb{R}^n$, the expression $\mathbf{y} = \mathbf{A}\mathbf{x}$ is the matrix product of matrix $m \times n$ with matrix $n \times 1$, therefore according to (2.2), it is

  $$y_i = \sum_{j=1}^{n} a_{ij} x_j.$$

  The vector $\mathbf{y} \in \mathbb{R}^m$ is the linear combination (with the coefficients $x_1, \dots, x_n$) of the columns of the matrix $\mathbf{A}$.
- For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{x}^T\mathbf{y} = x_1y_1 + \cdots + x_ny_n$ is the matrix product of the row vector $\mathbf{x}^T$ and the column vector $\mathbf{y}$, the result of which is a scalar.
- For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{x}\mathbf{y}^T$ is $m \times n$ matrix of rank 1, sometimes called the *outer product* of the vectors $\mathbf{x}$ a $\mathbf{y}$.

Symbol $\mathbf{1}_n = (1, \dots, 1) \in \mathbb{R}^n$ will denote the *column* vector with all its elements equal to one. When n is clear from the context, we will write just $\mathbf{1}$. For example, for $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{1}^T\mathbf{x} = x_1 + \cdots + x_n$.

2.6 Matrix sins

When manipulating matrix expressions and equations, you should aim to gain the same profficiency as with manipulating scalar expressions and equations. Students sometimes make gross errors when manipulating matrix expressions; errors which it is possible to avoid by paying some minimal attention. Next, we give some typical examples of these 'sins'.

2.6.1 An expression is nonsensical because of the matrices dimensions

As the first example we note blunders where an expression lacks meaning due to the dimensions of the matrices and vectors. The first type of these errors involves breaking the syntax rules,

[2] The term *vector* has a more general meaning in the general linear algebra than in the matrix algebra; there it means an element of a general linear space.

[3] Of course, we could do the same with rows (and some do, e.g. in computer graphics).

e.g.:

- When $\mathbf{A} \in \mathbb{R}^{2\times 3}$ and $\mathbf{B} \in \mathbb{R}^{3\times 3}$, then the following expressions are wrong:

$$\mathbf{A}+\mathbf{B}, \quad \mathbf{A}=\mathbf{B}, \quad [\mathbf{A}\ \mathbf{B}], \quad \mathbf{A}^T\mathbf{B}, \quad \mathbf{A}^{-1}, \quad \det \mathbf{A}, \quad \mathbf{A}^2.$$

- A frightful example is the use of a 'fraction' for a matrix, e.g. $\dfrac{\mathbf{A}}{\mathbf{B}}$.

In the second type of errors, the culprit produces an expression or a conclusion which does not contradict the syntax rules but does not make sense semantically, e.g.:

- Inversion of an evidently singular square matrix. For example $(\mathbf{w}\mathbf{w}^T)^{-1}$, where $\mathbf{w} \in \mathbb{R}^3$.
- Assuming the existence of the left inverse of a fat matrix or the right inverse of a slim matrix. For example writing $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}$, where $\mathbf{Q} \in \mathbb{R}^{5\times 3}$.
- The assertion that $\operatorname{rank} \mathbf{A} = 5$, where $\mathbf{A} \in \mathbb{R}^{3\times 5}$, is wrong because every quintuple of vectors from $\mathbb{R}^3$ is linearly dependent.

Example 2.1. When we see the expression $(\mathbf{A}^T\mathbf{B})^{-1}$, we must immediately realise the following about the dimensions of the matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{k\times p}$:

- In order to avoid a syntactical error in the multiplication, it must be the case that $m = k$.
- As the product $\mathbf{A}^T\mathbf{B}$ has the dimensions $n \times p$, we must have $n = p$ in order to avoid a syntax error in the inversion. So, now we know that both matrices must have the same dimensions.
- Since $\operatorname{rank}(\mathbf{AB}) \leq \min\{\operatorname{rank} \mathbf{A}, \operatorname{rank} \mathbf{B}\}$, then should $\mathbf{A}^T$ be narrow or $\mathbf{B}$ wide, it would follow that $\mathbf{A}^T\mathbf{B}$ would certainly be singular and we would have a semantic error. In order to avoid the error, both matrices must be either square or narrow, $m \geq n$.

Conclusion: in order for expression $(\mathbf{A}^T\mathbf{B})^{-1}$ to make sense, both matrices must have the same dimensions and must be square or narrow. You may well object that, even so, the matrix $\mathbf{A}^T\mathbf{B}$ still need not have an inverse – however, our goal was to find only *the necessary conditions for the dimensions of the matrices* to make sense. □

2.6.2 The use of non-existent matrix identities

Matrix manipulation skills can be improved by memorising a stock of matrix identities. Though, of course, they must not be wrong. Typical examples:

- $(\mathbf{AB})^T = \mathbf{A}^T\mathbf{B}^T$ (when the inner dimensions in the matrix product $\mathbf{A}^T\mathbf{B}^T$ differ, then it is also a syntax error)
- $(\mathbf{AB})^{-1} = \mathbf{A}^{-1}\mathbf{B}^{-1}$ (for non-square matrices it is also a syntax error, for square but singular matrices it is also a semantic error)
- $(\mathbf{A}+\mathbf{B})^2 = \mathbf{A}^2 + 2\mathbf{AB} + \mathbf{B}^2$. This identity is based on a very 'useful' but non-existent identity $\mathbf{AB} = \mathbf{BA}$. Correctly it should be $(\mathbf{A}+\mathbf{B})^2 = \mathbf{A}^2 + \mathbf{AB} + \mathbf{BA} + \mathbf{B}^2$.

2.6.3 Non equivalent manipulations of equations and inequalities

Here the culprit takes a wrong step with *nonequivalent manipulation* of an equation or an inequality. We are all familiar with equivalent and nonequivalent manipulations of scalar equations from school. For example, the operation 'take a square root of an equation' is nonequivalent, since, though $a = b$ implies $a^2 = b^2$, $a^2 = b^2$ does not imply $a = b$. Examples:

- The assumption that $\mathbf{a}^T\mathbf{x} = \mathbf{a}^T\mathbf{y}$ implies $\mathbf{x} = \mathbf{y}$ (not true even when the vector $\mathbf{a}$ is non zero).
- The assumption that when $\mathbf{A} \in \mathbb{R}^{3\times 5}$ and $\mathbf{AX} = \mathbf{AY}$, then $\mathbf{X} = \mathbf{Y}$ (not true because $\mathbf{A}$ does not have a left inverse, i.e. linearly independent columns).
- The assumption that $\mathbf{A}^T\mathbf{A} = \mathbf{B}^T\mathbf{B}$ implies $\mathbf{A} = \mathbf{B}$ (not true even for scalars).

2.6.4 Further ideas for working with matrices

- Draw rectangles (with dimensions) under matrix expressions to clarify their dimensions.
- When encountering a matrix equation or a system of equations, count the scalar equations and the unknowns.
- Work with Matlab as well as with the paper. Matrix expression manipulations can often be verified on random matrices. For example, if we want to verify the equality of $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$, we can try e.g. `A=randn(5,3); B=randn(3,6); (A*B)'-B'*A'`. Of course, it is not a proof.

2.7 Exercises

2.1. Solve these equations for the unknown matrix $\mathbf{X}$ (assume that, if needed, its inverse exists):

a) $\mathbf{AX} + \mathbf{B} = \mathbf{A}^2\mathbf{X}$
b) $\mathbf{X} - \mathbf{A} = \mathbf{XB}$
c) $2\mathbf{X} - \mathbf{AX} = 2\mathbf{A} = \mathbf{0}$

2.2. Solve the system of equations $\{\, \mathbf{b}_i = \mathbf{X}\mathbf{a}_i,\ i = 1, \ldots, k \,\}$ for the unknown matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$. What must be the value of k, so that the system will have the same number of equations as unknowns? On what condition does the system have a single solution?

2.3. Solve the system of equations $\{\, \mathbf{Ax} = \mathbf{b},\ \mathbf{x} = \mathbf{A}^T\mathbf{y} \,\}$, where $\mathbf{x}, \mathbf{y}$ are unknown vectors and the matrix $\mathbf{A}$ is wide with full rank. Find only the solution for $\mathbf{x}$, we are not interested in $\mathbf{y}$. Verify in Matlab on a random example obtained by commands `A=randn(m,n); b=randn(n,1)`.

2.4. Consider the system of equations in unknowns $\mathbf{x}$ and $\mathbf{y}$:

$$\begin{aligned} \mathbf{Ax} + \mathbf{By} &= \mathbf{a} \\ \mathbf{Cx} + \mathbf{Dy} &= \mathbf{b} \end{aligned}$$

a) Express this system in the form $\mathbf{Pu} = \mathbf{q}$.
b) Suppose that $\mathbf{a}, \mathbf{x} \in \mathbb{R}^m$, $\mathbf{b}, \mathbf{y} \in \mathbb{R}^n$. Show that $\mathbf{x} = (\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1}(\mathbf{a} - \mathbf{BD}^{-1}\mathbf{b})$. What is its computational advantage over computing $\mathbf{u}$ directly from the system $\mathbf{Pu} = \mathbf{q}$?

2.5. Which of these equation systems are linear? Lower case denotes vectors, upper case matrices. Assume the most general dimensions of the matrices and vectors. What is the number of equations and unknowns in each system?

a) $\mathbf{Ax} = \mathbf{b}$, unknown $\mathbf{x}$

b) $\mathbf{x}^T\mathbf{A}\mathbf{x} = 1$, unknown $\mathbf{x}$
c) $\mathbf{a}^T\mathbf{X}\mathbf{b} = 0$, unknown $\mathbf{X}$
d) $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^T = \mathbf{C}$, unknown $\mathbf{X}$
e) $\{\, \mathbf{X}^T\mathbf{Y} = \mathbf{A},\ \mathbf{Y}^T\mathbf{X} = \mathbf{B} \,\}$, unknown $\mathbf{X}, \mathbf{Y}$

2.6. Mapping $\operatorname{vec}\colon \mathbb{R}^{m\times n} \to \mathbb{R}^{mn}$ ('vectorisation' matrix, in Matlab written as `A(:)`), is defined so that $\operatorname{vec}\mathbf{A}$ is the matrix $\mathbf{A}$ rearranged by columns into a single vector. The *Kronecker matrix product* (in Matlab `kron(A,B)`) is defined as

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}.$$

For arbitrary matrices (with compatible dimensions), we have:

$$\operatorname{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A}) \operatorname{vec}\mathbf{B}. \tag{2.7}$$

Use this formula to transform the following systems of equations in the unknown matrix $\mathbf{X}$ into the form $\mathbf{P}\mathbf{u} = \mathbf{q}$ in the unknown vector $\mathbf{u}$. Assume that the number of equations is equal to the number of unknowns. Assume that the matrices and vectors have the most general dimensions that make sense.

a) $\{\, \mathbf{b}_i^T\mathbf{X}\mathbf{a}_i = 0,\ i = 1, \ldots, k \,\}$
b) $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^T = \mathbf{C}$

2.7. The sum of the diagonal elements of a square matrix is called its *trace*.

a) Prove that the matrices $\mathbf{AB}$ and $\mathbf{BA}$ have the same trace.
b) Prove that the equation $\mathbf{AB} - \mathbf{BA} = \mathbf{I}$ has no solution for any $\mathbf{A}, \mathbf{B}$.

2.8. *The commutator* of two matrices is the matrix $[\mathbf{A}, \mathbf{B}] = \mathbf{AB} - \mathbf{BA}$. Prove the *Jacobi's identity* $[\mathbf{A}, [\mathbf{B}, \mathbf{C}]] + [\mathbf{B}, [\mathbf{C}, \mathbf{A}]] + [\mathbf{C}, [\mathbf{A}, \mathbf{B}]] = \mathbf{0}$.

2.9. Prove the *Sherman-Morrison formula* ($\mathbf{A}$ is square regular and $\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u} \neq 1$):

$$(\mathbf{A} - \mathbf{u}\mathbf{v}^T)^{-1} = \mathbf{A}^{-1}\left(\mathbf{I} + \frac{\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}}{1 - \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}}\right).$$

2.10. Prove that $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

2.11. Prove that for every square matrix $\mathbf{A}$

a) $\mathbf{A} + \mathbf{A}^T$ is symmetric
b) $\mathbf{A} - \mathbf{A}^T$ is skew-symmetric
c) there exists symmetric $\mathbf{B}$ and skew-symmetric $\mathbf{C}$, such that $\mathbf{A} = \mathbf{B} + \mathbf{C}$, where $\mathbf{B}, \mathbf{C}$ are uniquely determined.

2.12. Prove that for each $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times m}$, the matrix

$$\mathbf{L} = \begin{bmatrix} \mathbf{I} - \mathbf{BA} & \mathbf{B} \\ 2\mathbf{A} - \mathbf{ABA} & \mathbf{AB} - \mathbf{I} \end{bmatrix}$$

has the property $\mathbf{L}^2 = \mathbf{I}$ (where $\mathbf{L}^2$ is the abbreviation for $\mathbf{LL}$). A matrix with this property is called the *involution*.

2.13. When is a diagonal matrix regular? What is the inverse of a diagonal matrix?

2.14. ($\star$) Prove that when $\mathbf{I} - \mathbf{A}$ is regular, then $\mathbf{A}(\mathbf{I} - \mathbf{A})^{-1} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{A}$.

2.15. ($\star$) Prove that when $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{A} + \mathbf{B}$ are regular, then

$$\mathbf{A}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B} = \mathbf{B}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{A} = (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}.$$

2.16. ($\star$) Let square matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ be such that $\mathbf{A}\mathbf{B}^T$ a $\mathbf{C}\mathbf{D}^T$ are symmetric and it holds that $\mathbf{A}\mathbf{D}^T - \mathbf{B}\mathbf{C}^T = \mathbf{I}$. Prove that $\mathbf{A}^T\mathbf{D} - \mathbf{C}^T\mathbf{B} = \mathbf{I}$.

Chapter 3

Linearity

Set $\mathbb{R}^{m\times n}$ of matrices of fixed dimensions $m \times n$, together with the operations $+$ (adding matrices) and $\cdot$ (multiplying matrices by a scalar), form a *linear space* over the field of real numbers. A special case is the linear space $\mathbb{R}^{n\times 1}$ of one column matrices or, applying the identity §2.5, the linear space $\mathbb{R}^n$ of all n-tuples of real numbers.

Let's review the notion of the linear space from linear algebra:

3.1 Linear subspaces

Linear combination of vectors $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^n$ is the vector

$$\alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k$$

for some scalars $\alpha_1, \dots, \alpha_k \in \mathbb{R}$.
Vectors are **linearly independent**, when the following implication holds

$$\alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k = \mathbf{0} \quad \Longrightarrow \quad \alpha_1 = \cdots = \alpha_k = 0. \tag{3.1}$$

Otherwise they are **linearly dependent**.
Linear span of a set of vectors $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ is the set

$$\operatorname{span}\{\mathbf{x}_1, \dots, \mathbf{x}_k\} = \{\, \alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k \mid \alpha_1, \dots, \alpha_k \in \mathbb{R} \,\}$$

of all their linear combinations (here we are assuming that the number of the vectors is finite).

The set $X \subseteq \mathbb{R}^n$ is called the **linear subspace** (briefly **subspace**) of the linear space $\mathbb{R}^n$, when an arbitrary linear combination of arbitrary vectors from X is contained in X (we say that the set X is closed with respect to the linear combinations).
A **basis** of the linear subspace $X \subseteq \mathbb{R}^n$ is a linearly independent set of vectors, whose linear envelope is X. A nontrivial subspace of $\mathbb{R}^n$ has an infinite number of bases, where each basis has the same number of vectors. This number is the **dimension** of the linear subspace, written $\dim X$.

3.2 Linear mapping

The mapping $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ is called **linear**, when for each $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^n$ and $\alpha_1, \dots, \alpha_k \in \mathbb{R}$,

$$\mathbf{f}(\alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k) = \alpha_1\mathbf{f}(x_1) + \cdots + \alpha_k\mathbf{f}(x_k), \tag{3.2}$$

in other words, when 'the mapping of a linear combination is equal to the linear combination of the mappings'.

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, then the mapping

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x} \tag{3.3}$$

is clearly linear, since

$$\mathbf{f}(\alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k) = \mathbf{A}(\alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k) = \alpha_1\mathbf{A}\mathbf{x}_1 + \cdots + \alpha_k\mathbf{A}\mathbf{x}_k = \alpha_1\mathbf{f}(x_1) + \cdots + \alpha_k\mathbf{f}(x_k).$$

Conversely, it is possible to prove (we omit the detailed proof), that for each linear mapping $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$, there exists precisely one matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, such that $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$. We say that the matrix $\mathbf{A}$ *represents* the linear mapping.

For $m = 1$ the linear mapping is a function $f\colon \mathbb{R}^n \to \mathbb{R}$ of the form

$$f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} = a_1x_1 + \cdots + a_nx_n, \tag{3.4}$$

where $\mathbf{a} \in \mathbb{R}^n$. This function is also known as a *linear form*.

A composition of linear mappings is another linear mapping. When $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ and $\mathbf{g}(\mathbf{y}) = \mathbf{B}\mathbf{y}$, then

$$\mathbf{g}(\mathbf{f}(\mathbf{x})) = (\mathbf{g} \circ \mathbf{f})(\mathbf{x}) = \mathbf{B}(\mathbf{A}\mathbf{x}) = (\mathbf{B}\mathbf{A})\mathbf{x},$$

i.e. $\mathbf{B}\mathbf{A}$ is the matrix of the composed mappings $\mathbf{g} \circ \mathbf{f}$. Therefore the matrix of composed mappings is the product of the matrices of the individual mappings. This is the main reason why it makes sense to define the matrix multiplication as in (2.2): the matrix multiplication corresponds to the composition of the linear mappings.

3.2.1 The range and the null space

There are two linear subspaces closely associated with linear mappings: the range and the null space (or kernel). When the mapping is represented by a matrix, as in $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, we talk about the range and the null space of the matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$.

The **range** of matrix $\mathbf{A}$ is the set

$$\operatorname{rng}\mathbf{A} = \mathbf{f}(\mathbb{R}^n) = \{\, \mathbf{A}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n \,\} = \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} \subseteq \mathbb{R}^m, \tag{3.5}$$

where $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$ are the columns of the matrix $\mathbf{A}$. Therefore the range is the linear envelope of the columns of the matrix, as $\mathbf{A}\mathbf{x} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n$ is the linear combination of the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$ with the coefficients $x_1, \ldots, x_n$. It is the set of all possible values of the mapping $\mathbf{f}$, i.e. the set of all $\mathbf{y}$, for which the system $\mathbf{y} = \mathbf{A}\mathbf{x}$ has a solution. The range is a linear subspace of $\mathbb{R}^m$. From the definition of the rank of a matrix it is clear that

$$\dim \operatorname{rng}\mathbf{A} = \operatorname{rank}\mathbf{A}. \tag{3.6}$$

The **null space** of matrix $\mathbf{A}$ is the set

$$\operatorname{null}\mathbf{A} = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{0} \,\} \subseteq \mathbb{R}^n \tag{3.7}$$

of all vectors which map into the null vector. Sometimes it is also called the *kernel* of the mapping. It is a linear subspace of $\mathbb{R}^n$. The null space is trivial (contains only the vector $\mathbf{0}$)

iff matrix $\mathbf{A}$ has linearly independent columns. That is, every fat matrix has a nontrivial null space.

The dimensions of the range and of the null space are related by:

$$\underbrace{\dim \operatorname{rng} \mathbf{A}}_{\operatorname{rank} \mathbf{A}} + \dim \operatorname{null} \mathbf{A} = n. \tag{3.8}$$

You will find the proof of this important relationship in every textbook of linear algebra.

3.3 Affine subspace and mapping

Affine combination of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n$ is such linear combination

$$\alpha_1 \mathbf{x}_1 + \cdots + \alpha_k \mathbf{x}_k,$$

for which $\alpha_1 + \cdots + \alpha_k = 1$.
Affine envelope of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ is the set of all their affine combinations. **Affine subspace**[1] of linear space $\mathbb{R}^n$ is such set $A \subseteq \mathbb{R}^n$ which is closed with respect to affine combinations (i.e. every affine combination of vectors from A is in A).

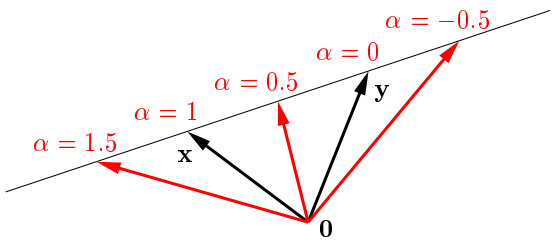
Example 3.1. Consider two linearly independent vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$. Their linear envelope is the set

$$\operatorname{span}\{\mathbf{x}, \mathbf{y}\} = \{\, \alpha \mathbf{x} + \beta \mathbf{y} \mid \alpha, \beta \in \mathbb{R} \,\},$$

i.e. the plane passing through these two points and through the origin $\mathbf{0}$, that is the entire $\mathbb{R}^2$. Their affine envelope is the set

$$\operatorname{aff}\{\mathbf{x}, \mathbf{y}\} = \{\, \alpha \mathbf{x} + \beta \mathbf{y} \mid \alpha, \beta \in \mathbb{R},\ \alpha + \beta = 1 \,\} = \{\, \alpha \mathbf{x} + (1 - \alpha) \mathbf{y} \mid \alpha \in \mathbb{R} \,\},$$

which is the line passing through the points $\mathbf{x}, \mathbf{y}$. The following figure shows the vectors $\alpha \mathbf{x} + (1 - \alpha) \mathbf{y}$ for various values of α:



Similarly, the linear envelope of two linearly independent vectors in $\mathbb{R}^3$ is the plane passing through these two points and the origin $\mathbf{0}$ and their affine envelope is the line passing through these two points. The affine envelope of three linearly independent points in $\mathbb{R}^3$ is the plane passing through these three points. □

Theorem 3.1.

- *Let A be an affine subspace of $\mathbb{R}^n$ and $\mathbf{x}_0 \in A$. Then the set $A - \mathbf{x}_0 = \{\, \mathbf{x} - \mathbf{x}_0 \mid \mathbf{x} \in A \,\}$ is a linear subspace of $\mathbb{R}^n$.*

[1] Here we define the affine *subspace* of a linear space rather than the affine *space* itself. The definition of affine space not referring to some linear space exists but it is not needed here, so it is omitted.

- *Let X be a linear subspace of $\mathbb{R}^n$ and $\mathbf{x}_0 \in \mathbb{R}^n$. Then the set $X + \mathbf{x}_0 = \{\, \mathbf{x} + \mathbf{x}_0 \mid \mathbf{x} \in X \,\}$ is an affine subspace of $\mathbb{R}^n$.*

Proof. We prove only the first part, as the proof of the second part is similar. We want to prove that an arbitrary linear combination of vectors from the set $A - \mathbf{x}_0$ is in $A - \mathbf{x}_0$. That means $\mathbf{x}_1, \dots, \mathbf{x}_k \in A$ and $\alpha_1, \dots, \alpha_k \in \mathbb{R}$ must satisfy $\alpha_1(\mathbf{x}_1 - \mathbf{x}_0) + \cdots + \alpha_k(\mathbf{x}_k - \mathbf{x}_0) \in A - \mathbf{x}_0$ or

$$\alpha_1(\mathbf{x}_1 - \mathbf{x}_0) + \cdots + \alpha_k(\mathbf{x}_k - \mathbf{x}_0) + \mathbf{x}_0 = \alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k + (1 - \alpha_1 - \cdots - \alpha_k)\mathbf{x}_0 \in A.$$

This holds because $\alpha_1 + \cdots + \alpha_k + (1 - \alpha_1 - \cdots - \alpha_k) = 1$ and therefore the last term is an affine combination of vectors from A, which by the assumption was in A. $\square$

This theorem shows that an affine subspace is just a 'shifted' linear subspace (i.e. it need not pass through the origin like the linear subspace). The **dimension of an affine subspace** is the dimension of that linear subspace. Affine subspaces of $\mathbb{R}^n$ with dimensions 0, 1, 2 and $n-1$ are called respectively the **point**, **line**, **plane** and **superplane**.

Mapping $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ is called an **affine** mapping, when (3.2) holds for all $\alpha_1 + \cdots + \alpha_k = 1$, i.e. the mapping of an affine combination is the same as the affine combination of the mappings. It is possible to show (do it!), that the mapping $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ is affine iff there exists matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and vector $\mathbf{b} \in \mathbb{R}^m$, such that

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}. \tag{3.9}$$

For $m = 1$ the mapping (3.9) is also called an **affine function** [2] $f\colon \mathbb{R}^n \to \mathbb{R}$ and has the form

$$f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b = a_1x_1 + \cdots + a_nx_n + b, \tag{3.10}$$

where $\mathbf{a} \in \mathbb{R}^n$ a $b \in \mathbb{R}$.

3.4 Exercises

3.1. Decide whether the following sets form linear or affine subspaces of $\mathbb{R}^n$ and determine their dimensions:

a) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} = 0 \,\}$ for given $\mathbf{a} \in \mathbb{R}^n$
b) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} = \alpha \,\}$ for given $\mathbf{a} \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$
c) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x}^T\mathbf{x} = 1 \,\}$
d) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}\mathbf{x}^T = \mathbf{I} \,\}$ for given $\mathbf{a} \in \mathbb{R}^n$

3.2. Given the mapping $\mathbf{f}(\mathbf{x}) = \mathbf{x} \times \mathbf{y}$, where $\mathbf{y} \in \mathbb{R}^3$ is a fixed (constant) vector and $\times$ denotes vector product (therefore this is a mapping from $\mathbb{R}^3$ to $\mathbb{R}^3$), is this mapping linear? If so, find the matrix $\mathbf{A}$, so that $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$. What is $\mathbf{A}^T$ equal to? What is the rank of $\mathbf{A}$?

3.3. Given mapping $\mathbf{f}\colon \mathbb{R}^2 \to \mathbb{R}^3$ determined by the rule $\mathbf{f}(x, y) = (\, x+y,\ 2x-1,\ x-y \,)$, is this mapping linear? Is this mapping affine? Prove both of your answers.

[2] The word 'linear' means something different in linear algebra and in mathematical analysis. For example, you called the function of a single variable $f(x) = ax + b$ linear at school. However, in linear algebra, it is not linear – it is affine. Although the equation system $\mathbf{A}\mathbf{x} = \mathbf{b}$ is called 'linear' even in linear algebra.

3.4. Prove that (a) the set of solutions of a homogeneous linear system $\mathbf{Ax} = \mathbf{0}$ is a linear subspace and (b) the set of solutions of a non-homogeneous linear system $\mathbf{Ax} = \mathbf{b}$ (for $\mathbf{b} \neq \mathbf{0}$) is an affine subspace.

3.5. Find the space of the range and the null space for each of the following linearch mappings:

a) $\mathbf{f}(x_1, x_2, x_3) = (\, x_1 - x_2,\ x_2 - x_3 + 2x_1 \,)$

b) $\mathbf{f}(x_1, x_2) = (\, 2x_1 + x_2,\ x_1 - x_2,\ x_1 + 2x_2 \,)$

3.6. Write the shortest possible matlab code to determine whether the spaces of the ranges for two given matrices are the same. What are the most general dimensions of the matrices required for the task to make sense?

3.7. Which of the following assertions are true? Prove each one or find a counter-example. Some of the assertions may be valid only for certain matrices dimensions – in those cases, find the most general conditions for the matrices dimensions for the assertion to be true.

a) When $\mathbf{AB}$ is of full rank, then $\mathbf{A}$ and $\mathbf{B}$ are of full ranks.

b) When $\mathbf{A}$ and $\mathbf{B}$ are of full ranks, then $\mathbf{AB}$ is of full rank.

c) When $\mathbf{A}$ and $\mathbf{B}$ have trivial null spaces, then $\mathbf{AB}$ has the trivial null space.

d) ($\star$) When $\mathbf{A}$ and $\mathbf{B}$ are both slim with full rank and $\mathbf{A}^T\mathbf{B} = \mathbf{0}$, then matrix $[\mathbf{A}\ \mathbf{B}]$ is slim with full rank.

e) ($\star$) When matrix $\begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}$ is of full rank, then $\mathbf{A}$ and $\mathbf{B}$ are both of full ranks.

Chapter 4

Orthogonality

4.1 Scalar product

The space $\mathbb{R}^n$ is naturally equipped with the **standard scalar product**

$$\mathbf{x}^T\mathbf{y} = x_1y_1 + \cdots + x_ny_n.$$

Scalar product obeys the **Cauchy-Schwarz inequality** $(\mathbf{x}^T\mathbf{y})^2 \leq (\mathbf{x}^T\mathbf{x})(\mathbf{y}^T\mathbf{y})$.

Standard scalar product induces the **euclidian norm**[1]

$$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T\mathbf{x}} = (x_1^2 + \cdots + x_n^2)^{1/2},$$

The norm fulfills the **triangle inequality** $\|\mathbf{x}+\mathbf{y}\|_2 \leq \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2$, which follows easily from the Cauchy-Schwarz inequality (square it and multiply out). The norm measures the *length* (more commonly called the magnitude) of the vector $\mathbf{x}$. The *angle* φ between a pair of vectors is given as

$$\cos\varphi = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|_2\,\|\mathbf{y}\|_2}.$$

The euclidian norm induces the **euclidian metric**

$$d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2,$$

which measures the *distance* between points $\mathbf{x}$ and $\mathbf{y}$.

4.2 Orthogonal vectors

A pair of vectors is called **orthogonal** (perpendicular), when $\mathbf{x}^T\mathbf{y} = 0$. It is written as $\mathbf{x} \perp \mathbf{y}$.

A vector is called **normalised**, when it has a unit magnitude ($\|\mathbf{x}\|_2 = 1 = \mathbf{x}^T\mathbf{x}$). The set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is called **orthonormal**, when each vector in this set is normalised and each pair of vectors from this set is orthogonal, that is:

$$\mathbf{x}_i^T\mathbf{x}_j = \begin{cases} 0 & \text{when } i \neq j, \\ 1 & \text{when } i = j. \end{cases} \tag{4.1}$$

[1] We use the symbol $\|\cdot\|_2$ for the euclidian norm instead of just $\|\cdot\|$ because later we will introduce other norms.

An othonormal set of vectors is linearly independent. To prove this take the scalar product of the left hand side of the implication (3.1) with vector $\mathbf{x}_i$, which gives

$$0 = \mathbf{x}_i^T \mathbf{0} = \alpha_1 \mathbf{x}_i^T \mathbf{x}_1 + \cdots + \alpha_k \mathbf{x}_i^T \mathbf{x}_k = \alpha_i \mathbf{x}_i^T \mathbf{x}_i = \alpha_i.$$

therefore $\alpha_i = 0$. Repeating this for each i, we get $\alpha_1 = \cdots = \alpha_k = 0$.

Orthonormal sets of vectors are in some sense 'the most linearly independent' sets of vectors.

4.3 Orthogonal subspaces

Subspaces X and Y of space $\mathbb{R}^n$ are called **orthogonal**, when $\mathbf{x} \perp \mathbf{y}$ for each $\mathbf{x} \in X$ and $\mathbf{y} \in Y$. Written as $X \perp Y$. The testing for orthogonality of subspaces nonetheless does not require the testing of an infinite number of pairs of vectors. It is sufficient (prove it!) to check that for two arbitrary bases of X and Y, each base vector of X is orthogonal to each base vector of Y.

Orthogonal complement of subspace X in space $\mathbb{R}^n$ is the set

$$X^\perp = \{\, \mathbf{y} \in \mathbb{R}^n \mid \mathbf{x}^T \mathbf{y} = 0 \text{ for all } \mathbf{x} \in X \,\}. \tag{4.2}$$

Thus it is the set of all vectors in $\mathbb{R}^n$, such that each is orthogonal to each vector in X. In other words, $X^\perp$ is the 'largest' subspace of $\mathbb{R}^n$, orthogonal to X. Properties of the orthogonal complement:

- $(X^\perp)^\perp = X$.
- $\dim X + \dim(X^\perp) = n$
- For each vector $\mathbf{z} \in \mathbb{R}^n$, there exists exactly one $\mathbf{x} \in X$ and exactly one $\mathbf{y} \in X^\perp$, such that $\mathbf{z} = \mathbf{x} + \mathbf{y}$.

Example 4.1. Two perpendicular lines in $\mathbb{R}^3$ passing through the origin are orthogonal subspaces. However, they are not orthogonal complements of each other. Orthogonal complement of a line in $\mathbb{R}^3$ passing through the origin is the plane through the origin which is perpendicular to the line. □

4.4 The four fundamental subspaces of a matrix

Every matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ generates four **fundamental subspaces**:

- $\operatorname{rng} \mathbf{A} = \{\, \mathbf{A}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n \,\}$ is the set of all linear combinations of the columns of $\mathbf{A}$,
- $\operatorname{null} \mathbf{A} = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{0} \,\}$ is the set of all vectors orthogonal to the rows of $\mathbf{A}$,
- $\operatorname{rng}(\mathbf{A}^T) = \{\, \mathbf{A}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^m \,\}$ is the set of all linear combinations of the rows of $\mathbf{A}$,
- $\operatorname{null}(\mathbf{A}^T) = \{\, \mathbf{x} \in \mathbb{R}^m \mid \mathbf{A}^T\mathbf{x} = \mathbf{0} \,\}$ is the set of all vectors orthogonal to the columns of $\mathbf{A}$.

It follows from the definition of the orthogonal complement (think about it!), that these subspaces are related as follows:

$$(\operatorname{null} \mathbf{A})^\perp = \operatorname{rng}(\mathbf{A}^T), \tag{4.3a}$$

$$(\operatorname{rng} \mathbf{A})^\perp = \operatorname{null}(\mathbf{A}^T). \tag{4.3b}$$

4.5 Matrix with orthonormal columns

Let columns of matrix $\mathbf{U} \in \mathbb{R}^{m\times n}$ form an orthonormal set of vectors. Since orthonormal vectors are linearly independent, then necessarily $m \geq n$. The condition of orthonormality (4.1) of the columns then can be expressed concisely as:

$$\mathbf{U}^T\mathbf{U} = \mathbf{I}_n. \tag{4.4}$$

linear mapping $\mathbf{f}(\mathbf{x}) = \mathbf{U}\mathbf{x}$ (i.e. mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$) preserves the scalar product, as

$$\mathbf{f}(\mathbf{x})^T\mathbf{f}(\mathbf{y}) = (\mathbf{U}\mathbf{x})^T(\mathbf{U}\mathbf{y}) = \mathbf{x}^T\mathbf{U}^T\mathbf{U}\mathbf{y} = \mathbf{x}^T\mathbf{y}.$$

When $\mathbf{x} = \mathbf{y}$ then it preserves also the euclidian norm, $\|\mathbf{f}(\mathbf{x})\|_2 = \|\mathbf{U}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$. That is the mapping preserves distances and angles. Such mappings are called **isometric**.

When the matrix $\mathbf{U}$ is square ($m = n$), the following relationships are mutually equivalent:

$$\mathbf{U}^T\mathbf{U} = \mathbf{I} \quad \Longleftrightarrow \quad \mathbf{U}^T = \mathbf{U}^{-1} \quad \Longleftrightarrow \quad \mathbf{U}\mathbf{U}^T = \mathbf{I}. \tag{4.5}$$

The proof is not difficult. Since the columns of $\mathbf{U}$ are orthonormal, they are linearly independent and $\mathbf{U}$ is regular. Multiplying the leftmost equation on the right by $\mathbf{U}^{-1}$ we get the middle equation. Multiplying the middle equation on the left by $\mathbf{U}$ we get the rightmost equation. The remaining implications are proven analogously.

Equivalence (4.5) tells us that when a square matrix has orthonormal columns, then its columns are orthonormal, too. Moreover, the inversion of such a matrix is easily computed by a trivial transposition. A square matrix obeying the conditions (4.5) is called **orthogonal matrix**.

It is worth emphasising that when $\mathbf{U}$ is rectangular with orthonormal columns, then it is not true that $\mathbf{U}\mathbf{U}^T = \mathbf{I}$. Further, when $\mathbf{U}$ has orthogonal (but not orthonormal) columns, it need not have orthogonal rows[2].

Let $\mathbf{U}$ be an orthogonal matrix. Computing the determinant of both sides of the equation $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, we get $\det(\mathbf{U}^T\mathbf{U}) = \det(\mathbf{U}^T)\det\mathbf{U} = (\det\mathbf{U})^2 = 1$. That is $\det\mathbf{U}$ can take on two values:

- When $\det\mathbf{U} = 1$, the matrix is called **special orthogonal** or also **rotational**, as the mapping $\mathbf{f}(\mathbf{x}) = \mathbf{U}\mathbf{x}$ (mapping from $\mathbb{R}^n$ to itself) means a *rotation* of vector $\mathbf{x}$ around the origin. Every rotation in the $\mathbb{R}^n$ space can be uniquely represented by a rotation matrix.
- When $\det\mathbf{U} = -1$, then the mapping $\mathbf{f}$ is the composition of a rotation and a *reflection* around a superplane passing through the origin.

Example 4.2. All 2×2 rotational matrices can be written as

$$\mathbf{U} = \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix}$$

for some value of φ. Multiplying a vector by this matrix corresponds to the rotation of the vector in the plane by the angle φ. Check that $\mathbf{U}^T\mathbf{U} = \mathbf{I} = \mathbf{U}\mathbf{U}^T$ and $\det\mathbf{U} = 1$. □

[2] this is perhaps the reason why a square matrix with orthonormal columns (therefore also rows) is not called 'orthonormal' but 'orthogonal'. Rectangular matrix with orthonormal columns and square matrix with orthogonal (but not orthonormal) columns do not have special names.

Example 4.3. Permutation matrix is a square matrix, the columns of which are permuted vectors of the standard basis, e.g.

$$\begin{bmatrix}\mathbf{e}_3 & \mathbf{e}_1 & \mathbf{e}_2\end{bmatrix} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 1 & 0 & 0\end{bmatrix}.$$

Permutation matrices are orthogonal (prove it!) and their determinants are equal to the sign of the permutation.

Remember: multiplying an arbitrary matrix $\mathbf{A}$ by a permutation matrix on the left permutes the rows of matrix $\mathbf{A}$. Multiplying matrix $\mathbf{A}$ by a permutation matrix on the right permutes the columns of matrix $\mathbf{A}$. □

4.6 QR decomposition

Matrix $\mathbf{A}$ is **upper triangular** when $a_{ij} = 0$ for each $i > j$ (there are only zeroes under the main diagonal). It is **lower triangular** when $a_{ij} = 0$ for each $i < j$ (there are only zeroes above the main diagonal).

Every matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with $m \geq n$ can be decomposed into the product

$$\mathbf{A} = \mathbf{QR}, \tag{4.6}$$

where $\mathbf{Q} \in \mathbb{R}^{m\times n}$ has orthonormal columns ($\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$) and $\mathbf{R} \in \mathbb{R}^{n\times n}$ is upper triangular. When $\mathbf{A}$ is of full rank (i.e. n) and the condition that the diagonal elements $\mathbb{R}$ be positive ($r_{ii} > 0$) is satisfied, then matrices $\mathbf{Q}$ and $\mathbf{R}$ are unique. The QR decomposition is implemented in Matlab by the command[3] `[Q,R]=qr(A,0)`.

Since columns of $\mathbf{Q}$ are linearly independent, $\mathbf{Ax} = \mathbf{QRx} = \mathbf{0}$ precisely when $\mathbf{Rx} = \mathbf{0}$. That means $\operatorname{null}\mathbf{A} = \operatorname{null}\mathbf{R}$. Then, using identity (3.8), we have: $\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{R}$.

When $\mathbf{A}$ if of full rank, matrix $\mathbf{R}$ is regular and therefore (think carefully!) $\operatorname{rng}\mathbf{A} = \operatorname{rng}\mathbf{Q}$. This demonstrates that when $\mathbf{A}$ is of full rank, then the QR decomposition can be understood as finding the orthonormal basis of the subspace $\operatorname{rng}\mathbf{A}$, where the basis is formed by the columns of the matrix $\mathbf{Q}$.

QR decomposition has many applications. It is typically used for solving linear systems. For example, let us solve the system $\mathbf{Ax} = \mathbf{b}$ with regular square matrix $\mathbf{A}$. Decompose $\mathbf{A} = \mathbf{QR}$ and left-multiply the system by $\mathbf{Q}^T$, which gives

$$\mathbf{Rx} = \mathbf{Q}^T\mathbf{b}. \tag{4.7}$$

This is ?ekvivalentní úprava?, since $\mathbf{Q}$ is regular. However, as $\mathbf{R}$ is triangular, this system can be solved easily by back-substitution.

4.6.1 (⋆) Gramm-Schmidt orthonormalisation

Gramm-Schmidtova orthonormalisation is an algorithm, which for given linearly independent vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$ finds vectors $\mathbf{q}_1, \ldots, \mathbf{q}_n \in \mathbb{R}^m$, such that

[3] Note that the command `[Q,R]=qr(A)` computes so called *full QR decomposition*, in which $\mathbf{R}$ is upper triangular and of the same size as $\mathbf{A}$, and $\mathbf{Q}$ is orthogonal of the size $m \times m$. Find out about this command using `help qr`!

- $\mathbf{q}_1, \ldots, \mathbf{q}_n$ are orthonormal,
- For each $k = 1, \ldots, n$ $\operatorname{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_k\} = \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$.

The idea of the algorithm is simple. Suppose that we already have vectors $\mathbf{q}_1, \ldots, \mathbf{q}_{k-1}$ with the described properties. We add to the vector $\mathbf{a}_k$ such linear combination of vectors $\mathbf{q}_1, \ldots, \mathbf{q}_{k-1}$, so that it becomes orthogonal to them all. Then we normalise this vector, i.e.

$$\mathbf{q}_k := \mathbf{a}_k - \sum_{j=1}^{k-1} r_{jk}\mathbf{q}_j, \quad \mathbf{q}_k := \frac{\mathbf{q}_k}{\|\mathbf{q}_k\|_2}. \tag{4.8}$$

The algorithm iterates step by step for $k = 1, \ldots, n$.

How to find the coefficients r_{jk}? From (4.8) it follows that

$$\mathbf{a}_k = \sum_{j=1}^{k} r_{jk}\mathbf{q}_j. \tag{4.9}$$

here we have an extra coefficient r_{kk}, which represents the change of the vector $\mathbf{q}_k$ by normalisation. Relation (4.9) enables us to compute the coefficients r_{jk} from the requirement of ortonormality of the vectors $\mathbf{q}_1, \ldots, \mathbf{q}_k$. Multiplying it by vector $\mathbf{q}_j$, we get $r_{jk} = \mathbf{q}_j^T\mathbf{a}_k$.

An improved version of Gramm-Schmidt orthonormalisation can be used for computing the QR decomposition. Equation (4.9) can be written in the matrix form as $\mathbf{A} = \mathbf{QR}$, where vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$ are columns of matrix $\mathbf{A}$, vectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$ are columns of $\mathbf{Q}$, and $\mathbf{R}$ is upper triangular with elements $r_{jk} = \mathbf{q}_i^T\mathbf{a}_k$. QR decomposition is then achieved by improvements to this algorithm, which reduce the rounding errors and also allow for the linear dependence of the columns of $\mathbf{A}$.

4.7 Exercises

4.1. Find the orthogonal complement of the space $\operatorname{span}\{(0,1,1),(1,2,3)\}$.

4.2. Find two orthonormal vectors $\mathbf{x}, \mathbf{y}$, such that $\operatorname{span}\{\mathbf{x},\mathbf{y}\} = \operatorname{span}\{(0,1,1),(1,2,3)\}$.

4.3. Find the orthonormal basis of the subspace $\operatorname{span}\{\,(1,1,1,-1),\ (2,-1,-1,1),\ (-1,2,2,1)\,\}$ using QR decomposition.

4.4. Prove that the product of orthogonal matrices is an orthogonal matrix.

4.5. For which n is the matrix $\operatorname{diag}(-\mathbf{1}_n)$ (i.e. diagonal matrix with minus ones along the diagonal) rotational?

4.6. What are the conditions on numbers a, b so that the matrix $\begin{bmatrix} a+b & b-a \\ a-b & b+a \end{bmatrix}$ is orthogonal?

4.7. The number of independent parameters (degrees of freedom) of an orthogonal matrix $n \times n$ is determined by the difference of the number of matrix elements (n^2) and the number of independent equations in the condition $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. Informally speaking, it is the number of 'dials' you can independently 'twiddle' during a rotation in the n-dimensional space. What is this number for $n = 2, 3, 4$? Find the general formula for any n.

4.8. ($\star$) Consider the mapping $\mathbf{F}\colon \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ given by the formula $\mathbf{F}(\mathbf{A}) = (\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A})^{-1}$. Prove that:

a) For each $\mathbf{A}$, such that $\mathbf{I}+\mathbf{A}$ is regular, $(\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A})^{-1} = (\mathbf{I}+\mathbf{A})^{-1}(\mathbf{I}-\mathbf{A})$.
b) The matrix $\mathbf{F}(\mathbf{A})$ is orthogonal for a skew-symmetric matrix $\mathbf{A}$.
c) The matrix $\mathbf{F}(\mathbf{A})$ is skew-symmetric for an orthogonal matrix $\mathbf{A}$ such that $\mathbf{I}+\mathbf{A}$ is regular.
d) The mapping $\mathbf{F}$ is a self-inversion, i.e. $\mathbf{F}(\mathbf{F}(\mathbf{A})) = \mathbf{A}$ for each $\mathbf{A}$. This applies for any matrix $\mathbf{A}$, not just for an orthogonal or a skew-symmetric $\mathbf{A}$.

Before working out your proofs, check in Matlab that the above statements are valid for a random matrix.

4.9. Let X, Y be subspaces of $\mathbb{R}^n$. We define $X+Y = \{\, \mathbf{x}+\mathbf{y} \mid \mathbf{x} \in X,\ \mathbf{y} \in Y \,\}$. Prove that:

a) $X \subseteq Y \implies X^\perp \supseteq Y^\perp$
b) $(\star)$ $(X+Y)^\perp = X^\perp \cap Y^\perp$
c) $(X \cap Y)^\perp = X^\perp + Y^\perp$ Hint: prove this from the previous point by using $(X^\perp)^\perp = X$.

Chapter 5

Spectral Decomposition and Quadratic Functions

5.1 Eigenvalues and eigenvectors

When

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}. \tag{5.1}$$

for square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, vector $\mathbf{v} \in \mathbb{C}^n$, $\mathbf{v} \neq \mathbf{0}$ and scalar $\lambda \in \mathbb{C}$,

then λ is called an **eigenvalue** of the matrix and $\mathbf{v}$ is the **eigenvector** associated with the eigenvalue λ. Eigenvalues and eigenvectors can be in general complex-valued.

Equation (5.1) can be re-written as

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}. \tag{5.2}$$

This is a system of homogeneous linear equations in $\mathbf{v}$, which has a non-trivial solution iff the matrix $\mathbf{A} - \lambda\mathbf{I}$ is singular. That is eigenvalues are the roots of the polynomial

$$p_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}), \tag{5.3}$$

which is called the **characteristic polynomial**. Eigenvectors associated with eigenvalues λ can then be found from the equations system (5.2).

Example 5.1. Find the eigenvalues of the matrix $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$. The characteristic equation is

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \det \begin{bmatrix} 1-\lambda & 2 \\ 3 & 4-\lambda \end{bmatrix} = (1-\lambda)(4-\lambda) - 3\cdot 2 = \lambda^2 - 5\lambda - 2 = 0.$$

This quadratic equation has two roots $\lambda = (5 \pm \sqrt{33})/2$. These are then the eigenvalues of matrix A. Eigenvectors belonging to each λ will be found by solving the homogeneous linear system:

$$\begin{bmatrix} 1-\lambda & 2 \\ 3 & 4-\lambda \end{bmatrix} \mathbf{v} = \mathbf{0}.$$

□

It follows from the definition of the determinant (2.6) (think about it!), that the characteristic polynomial is of degree n, therefore it has n (in general complex) roots. Labeling the roots $\lambda_1, \ldots, \lambda_n$, it follows that:

$$p_{\mathbf{A}}(\lambda) = \prod_{i=1}^{n} (\lambda - \lambda_i).$$

There may be some multiple roots. From this perspective, the matrix has exactly n eigenvalues, of which some may be the same. This list of eigenvalues is sometimes called the **spectrum** matrix.

Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of matrix A and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be their associated eigenvectors. Equation (5.1) for them can be written as the single matrix equation (think!)

$$\mathbf{AV} = \mathbf{VD}, \tag{5.4}$$

where the diagonal matrix $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ has the eigenvalues on the diagonal and the columns of the square matrix $\mathbf{V} = [\mathbf{v}_1 \ \cdots \ \mathbf{v}_n]$ are the eigenvectors.

The eigenvectors are not uniquely determined by their eigenvalues. All eigenvectors associated with one particular eigenvalue form the subspace $\mathbb{R}^n$, since when $\mathbf{Au} = \lambda\mathbf{u}$ and $\mathbf{Av} = \lambda\mathbf{v}$, then $\mathbf{A}(\alpha\mathbf{u}) = \lambda(\alpha\mathbf{u})$ and $\mathbf{A}(\mathbf{u}+\mathbf{v}) = \lambda(\mathbf{u}+\mathbf{v})$. Eigenvectors can be in general linearly dependent. This is not a simple question and we will not discuss it here in detail. Let us just say that there is a good reason to choose such eigenvectors, so that the rank of matrix $\mathbf{V}$ is as large as possible.

How are the eigenvalues and eigenvectors calculated? The characteristic polynomial is mostly a theoretical tool and a direct solution for its roots is not suited to numerical computation. Numerical iteration algorithms are used for larger matrices. Different types of algorithms are best suited to different types of matrices. The matlab function `[V,D]=eig(A)` computes matrices $\mathbf{V}$ and $\mathbf{D}$ fulfilling (5.4).

5.1.1 Spectral decomposition

When $\mathbf{V}$ is regular (i.e., there exist n linearly independent eigenvectors), then it is invertible and (5.4) can be written as

$$\mathbf{A} = \mathbf{VDV}^{-1}. \tag{5.5}$$

This identity (5.5) is then called **eigenvalues decomposition of a matrix** or **spectral decomposition**. In this case matrix $\mathbf{A}$ is similar to a diagonal matrix (it is **diagonalisable**), since (5.5) implies $\mathbf{V}^{-1}\mathbf{AV} = \mathbf{D}$.

Many properties of matrices are known to guarantee diagonalisability. The most important one is symmetry.

Theorem 5.1. *Let matrix $\mathbf{A}$ of dimensions $n \times n$ be symmetric. Then all its eigenvectors are real and there exists an orthonormal set n of its eigenvectors.*

This is sometimes called the **spectral theorem**. It says that for any symmetric $\mathbf{A}$ in the identity (5.4), matrix $\mathbf{D}$ is real and $\mathbf{V}$ can be chosen as orthogonal, $\mathbf{V}^{-1} = \mathbf{V}^T$. Therefore

$$\mathbf{A} = \mathbf{VDV}^T. \tag{5.6}$$

Eigenvalues and eigenvectors are an extensive subject which we have by no means exhausted here. From now on we will need only the spectral decomposition of a symmetric matrix.

5.2 Quadratic form

Quadratic form over $\mathbb{R}^n$ is the function $f\colon \mathbb{R}^n \to \mathbb{R}$ given by the formula

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}, \tag{5.7}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$.

Every square matrix can be written as the sum of a symmetric and a skew-symmetric matrix:

$$\mathbf{A} = \underbrace{\tfrac{1}{2}(\mathbf{A}+\mathbf{A}^T)}_{\text{symetrická}} + \underbrace{\tfrac{1}{2}(\mathbf{A}-\mathbf{A}^T)}_{\text{antisymetrická}}$$

(see Exercise 2.11). However,

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \tfrac{1}{2}\mathbf{x}^T(\mathbf{A}+\mathbf{A}^T)\mathbf{x} + \tfrac{1}{2}\underbrace{\mathbf{x}^T(\mathbf{A}-\mathbf{A}^T)\mathbf{x}}_{0},$$

since $\mathbf{x}^T(\mathbf{A}-\mathbf{A}^T)\mathbf{x} = \mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T\mathbf{A}^T\mathbf{x} = \mathbf{x}^T\mathbf{A}\mathbf{x} - (\mathbf{x}^T\mathbf{A}\mathbf{x})^T = 0$, where we used the fact that a transposition of a scalar is the same scalar.

Therefore when $\mathbf{A}$ is not symmetric, we can substitute for it its symmetric part $\frac{1}{2}(\mathbf{A}+\mathbf{A}^T)$, leaving the quadratic form unchanged. Thus in what follows we will safely assume that $\mathbf{A}$ is symmetric.

Definition 5.1. *Symmetric matrix* $\mathbf{A}$ *is*

- **positive [negative] semidefinite**, *when for each* $\mathbf{x}$, $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ *[*$\mathbf{x}^T\mathbf{A}\mathbf{x} \leq 0$*]*
- **positive [negative] definite**, *when for each* $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ *[*$\mathbf{x}^T\mathbf{A}\mathbf{x} < 0$*]*
- **indefinite**, *when there exist* $\mathbf{x}$ *and* $\mathbf{y}$, *such that* $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ *and* $\mathbf{y}^T\mathbf{A}\mathbf{y} < 0$.

A matrix may have several of these properties. For example, a positive definite matrix is also positive semidefinite. A null matrix is both positive and negative semidefinite.

Even though the definition makes sense for an arbitrary square matrix, it is customary to talk about these properties only for symmetric matrices. Sometimes these properties are defined not for a matrix but more generally for the quadratic form.

It is clear from definition 5.1 whether a quadratic form has an extremum and of what kind:

- When $\mathbf{A}$ is positive [negative] semidefinite, then the quadratic form has a minimum [maximum] at the origin.
- When $\mathbf{A}$ is positive [negative] definite, then the quadratic form has a sharp minimum [maximum] at the origin.
- When $\mathbf{A}$ is indefinite, then the quadratic form does not have an extremum.

This statement is easy to prove. When $\mathbf{A}$ is positive semidefinite, then the quadratic form can not be negative and at $\mathbf{x} = \mathbf{0}$ must be zero, therefore it has a minimum at $\mathbf{x} = \mathbf{0}$ (and possibly elsewhere, too). When $\mathbf{A}$ is indefinite and e.g. $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$, then a point $\mathbf{x}$ can not be a maximum because $(2\mathbf{x})^T\mathbf{A}(2\mathbf{x}) > \mathbf{x}^T\mathbf{A}\mathbf{x}$. It can not be a minimum either because for some $\mathbf{y}$, $\mathbf{y}^T\mathbf{A}\mathbf{y} < 0$.

Theorem 5.2. *A symmetric matrix is*

- *positive [negative] semidefinite, iff all its eigenvalues are non-negative [non-positive]*

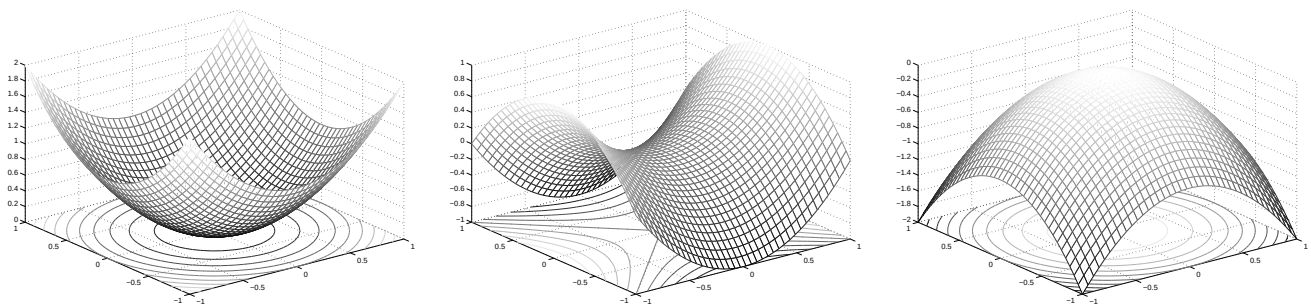
- *positive [negative] definite, iff all its eigenvalues are positive [negative]*
- *indefinite, iff it has at least one positive and one negative eigenvalues.*

Proof. By the eigenvalues decomposition (5.6):

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{V}\mathbf{D}\mathbf{V}^T\mathbf{x} = \mathbf{y}^T\mathbf{D}\mathbf{y} = \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2, \tag{5.8}$$

where $\mathbf{y} = \mathbf{V}^T\mathbf{x}$. The substitution $\mathbf{x} = \mathbf{V}\mathbf{y}$ thus diagonalised the matrix of the quadratic form. As $\mathbf{V}$ is regular, the definity of matrix $\mathbf{A}$ is the same as the definity of $\mathbf{D}$. However, as $\mathbf{D}$ is diagonal, its definity is immediately clear from the signs of λ_i. For example, expression (5.8) is non-negative for each $\mathbf{y}$ iff all λ_i are non-negative. □

When each λ_i is positive ($\mathbf{A}$ is positive definite), then the shape of the function $g(\mathbf{y}) = \mathbf{y}^T\mathbf{D}\mathbf{y}$ 'looks like a pit'. When each λ_i is negative ($\mathbf{A}$ is negative definite), then the function 'looks like a peak'. When some λ_i are positive and some negative ($\mathbf{A}$ is indefinite), then the shape of the function is a 'saddle':



$g(y_1, y_2) = y_1^2 + y_2^2$ $\qquad$ $g(y_1, y_2) = y_1^2 - y_2^2$ $\qquad$ $g(y_1, y_2) = -y_1^2 - y_2^2$

However, as $\mathbf{V}$ is orthogonal, the transformation $\mathbf{x} = \mathbf{V}\mathbf{y}$ is just an isometry, thus the form of f will differ from the diagonal form g only by rotation/reflection in the domain space.

5.3 Quadratic function

General **quadratic function** (or *second degree polynomial*) of n variables has the form:

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c, \tag{5.9}$$

where $\mathbf{A}^T = \mathbf{A} \neq \mathbf{0}$. Compared to the quadratic form it has additional linear and constant terms. Conversely, the quadratic form is the same as qadratic function without linear and constant terms. Note that for $n = 1$ (5.9) is the well known quadratic function of a single variable $f(x) = ax^2 + bx + c$.

How to find extrema of quadratic functions? Find the natural extrema using derivatives, see later. Another method is to transform the quadratic function into a quadratic form by translation of the coordinates.

Sometimes we can find vector $\mathbf{x}_0 \in \mathbb{R}^n$ and scalar y_0, such that

$$\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c = (\mathbf{x} - \mathbf{x}_0)^T\mathbf{A}(\mathbf{x} - \mathbf{x}_0) + y_0. \tag{5.10}$$

The expression on the right hand side is a quadratic form with the origin moved to the point $\mathbf{x}_0$, plus a constant. This transformation is called **completing the square**. You know it for the

case of $n = 1$, as the school method for deriving the formula for the roots of the quadratic equation of a single variable. We determine $\mathbf{x}_0, y_0$ from the given $\mathbf{A}, \mathbf{b}, c$ as follows. Multiplying out the right hand side, we get

$$\begin{aligned}(\mathbf{x}-\mathbf{x}_0)^T\mathbf{A}(\mathbf{x}-\mathbf{x}_0)+y_0 &= \mathbf{x}^T\mathbf{A}\mathbf{x}-\mathbf{x}^T\mathbf{A}\mathbf{x}_0-\mathbf{x}_0^T\mathbf{A}\mathbf{x}+\mathbf{x}_0^T\mathbf{A}\mathbf{x}_0+y_0\\ &= \mathbf{x}^T\mathbf{A}\mathbf{x}-2\mathbf{x}_0^T\mathbf{A}\mathbf{x}+\mathbf{x}_0^T\mathbf{A}\mathbf{x}_0+y_0.\end{aligned}$$

Comparing the terms of the same degree we obtain

$$\mathbf{b} = -2\mathbf{A}\mathbf{x}_0, \tag{5.11a}$$

$$c = \mathbf{x}_0^T\mathbf{A}\mathbf{x}_0 + y_0, \tag{5.11b}$$

from which we find $\mathbf{x}_0$ and y_0. When the sytem (5.11a) has no solution, then the completion of the square is not possible.

When the completion of the square is possible, then the solution of the extrema of a quadratic function is no different to the solution of the extrema of a quadratic form because the only difference is the translation by $\mathbf{x}_0$. When the completion of the square is not possible, then the quadratic function does not have any extrema.

Example 5.2. Given the quadratic function

$$f(x,y) = 2x^2-2xy+y^2-2y+3 = \begin{bmatrix}x\\y\end{bmatrix}^T\begin{bmatrix}2 & -1\\-1 & 1\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix}+\begin{bmatrix}0\\-2\end{bmatrix}^T\begin{bmatrix}x\\y\end{bmatrix}+3.$$

Its completion of the square is

$$f(x,y) = 2(x-1)^2-2(x-1)(y-2)+(y-2)^2-3 = \begin{bmatrix}x-1\\y-2\end{bmatrix}^T\begin{bmatrix}2 & -1\\-1 & 1\end{bmatrix}\begin{bmatrix}x-1\\y-2\end{bmatrix}-3,$$

thus we have $\mathbf{x}_0 = (1,2)$, $y_0 = -3$. Since matrix $\mathbf{A}$ is positive definite (verify!), the quadratic function has an etremum at the point $\mathbf{x}_0$. □

Example 5.3. For the quadratic function

$$f(x,y) = x^2-y = \begin{bmatrix}x\\y\end{bmatrix}^T\begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix}+\begin{bmatrix}0\\-1\end{bmatrix}^T\begin{bmatrix}x\\y\end{bmatrix}$$

the square can not be completed. □

Contour of a quadratic function is called **quadric**, (or *quadric surface*). E.g. quadric is the set

$$\{\,\mathbf{x}\in\mathbb{R}^n \mid \mathbf{x}^T\mathbf{A}\mathbf{x}+\mathbf{b}^T\mathbf{x}+c=0\,\}. \tag{5.12}$$

For $n = 2$ the quadric is called **conic**. An important special case of a quadric is **ellipsoid surface**[1], which is the set $\{\,\mathbf{x}\in\mathbb{R}^n \mid \mathbf{x}^T\mathbf{A}\mathbf{x}=1\,\}$ for a positive definite $\mathbf{A}$.

[1] Sometimes it is also called *ellipsoid* but the terminology is ambiguous and some authors mean by an ellipsoid the set $\{\,\mathbf{x}\in\mathbb{R}^n \mid \mathbf{x}^T\mathbf{A}\mathbf{x}\le 1\,\}$. The difference is between the surface of a solid and the whole solid.

5.4 Exercises

5.1. Compute the eigenvectors and eigenvalues of the matrices $\begin{bmatrix} 1 & 2 \\ -4 & -2 \end{bmatrix}$, $\begin{bmatrix} 1 & 2 \\ 2 & -3 \end{bmatrix}$.

5.2. Write down the equation whose roots are the eigenvectors of the matrix $\begin{bmatrix} 2 & 0 & 3 \\ 0 & -2 & -1 \\ 3 & -1 & 2 \end{bmatrix}$.

5.3. Find the eigenvalues and eigenvectors of (a) null, (b)unit, (c) diagonal, matrices. Find the eigenvalues of a triangular matrix.

5.4. Show that $\lambda_1 + \cdots + \lambda_n = \operatorname{trace} \mathbf{A}$ and $\lambda_1 \times \cdots \times \lambda_n = \det \mathbf{A}$.

5.5. Suppose you know the eigenvalues and eigenvectors of matrix $\mathbf{A}$. What are the eigenvalues and eigenvectors of the matrix $\mathbf{A} + \alpha \mathbf{I}$?

5.6. ($\star$) We said that finding the roots of the characteristic polynomial (5.3) is not a suitable method for finding the eigenvalues. On the contrary, finding the roots of an arbitrary polynomial can be changed into finding the eigenvalues of a matrix, called the *accompanying matrix* of the polynomial. Derive the shape of this matrix. Verify in Matlab for various polynomials.

5.7. It is well known that an arbitrary rotation in the 3D space can be performed as a rotation around some line (passing through the origin) by some angle. Using geometrical reasoning only (i.e., without any calculations), deduce as much as you can about the eigenvalues and eigenvectors of a rotational matrix of dimensions 3×3.

5.8. In §6.1.3 we defined projection as matrix $\mathbf{P}$ satisfying $\mathbf{P}^2 = \mathbf{P}$. Using geometrical reasoning, find at least one eigenvalue and an associated eigenvector of projection.

5.9. ($\star$) Using geometrical reasoning, find at least two eigenvectors and associated eigenvalues of the Householder's matrix from Exercise 6.16.

5.10. What is $\mathbf{A}^n$ equal to, when $\mathbf{A}$ is a symmetric matrix?

5.11. Show that the eigenvalues of a skew-symmetric matrix are either zero or purely imaginary.

5.12. ($\star$) Show that two square matrices commute iff they have identical eigenvectors.

5.13. ($\star$) Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times m}$. Show that all non-zero eigenvalues of the matrices $\mathbf{AB}$ and $\mathbf{BA}$ are identical.

5.14. For each following matrix determine whether it is positive/negative (semi)definite or indefinite:

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}, \quad \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

5.15. Determine whether the following quadratic functions have a minimum, maximum, and at which point. Use the completion of the square.

a) $f(x, y) = x^2 + 4xy - 2y^2 + 3x - 6y + 5$

b) $f(\mathbf{x}) = \mathbf{x}^T \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 2 & -1 \end{bmatrix} \mathbf{x}$

5.16. Consider the matrix $\mathbf{A} = \begin{bmatrix} 1 & -3 \\ 2 & -4 \end{bmatrix}$. Which of the following statements are true?

a) $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is non-negative for each $\mathbf{x} \in \mathbb{R}^2$.
b) $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is non-positive for each $\mathbf{x} \in \mathbb{R}^2$.
c) The function $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$ has an extremum at the point $\mathbf{x} = \mathbf{0}$.

Hint: is the matrix symmetric?

5.17. ($\star$) Implement a Matlab function `ellipse(Q)` that draws an ellipse given by the equation $\mathbf{x}^T\mathbf{A}\mathbf{x} = 1$, for positive definite $\mathbf{A}$. Think how to proceed when designing a function `conic(Q)`, that draws the conic section $\mathbf{x}^T\mathbf{A}\mathbf{x} = 1$ for $\mathbf{A}$ of an arbitrary definity (recall that a general conic section can be unbounded,therefore it is necessary to cut it off at the boundary of a given rectangle).

5.18. Prove that the matrix $\mathbf{A}^T\mathbf{A}$ is positive semidefinite for any matrix $\mathbf{A}$.

5.19. Prove that the matrix $\mathbf{A}^T\mathbf{A} + \mu\mathbf{I}$ is positive definite for any matrix $\mathbf{A}$ and for any $\mu > 0$.

5.20. Prove that a (square symmetric) matrix is positive definite iff its inverse is positive definite.

5.21. Must a positive semidefinite matrix have non-negative elements along its diagonal? Prove your answer, weather it was positive or negative.

5.22. ($\star$) Positive semidefinite matrix can be understood as a generalisation of non-negative numbers. This is why positive semidefinity is sometimes denoted as $\mathbf{A} \succeq 0$ and positive definity as $\mathbf{A} \succ 0$. The notation $\mathbf{A} \succeq \mathbf{B}$ is an abbreviation of $\mathbf{A} - \mathbf{B} \succeq 0$. Based on this analogy, we might expect that:

a) If $\mathbf{A} \succeq \mathbf{B}$ and $\mathbf{C} \succeq \mathbf{D}$, then $\mathbf{A} + \mathbf{C} \succeq \mathbf{B} + \mathbf{D}$.
b) If $\mathbf{A} \succeq 0$ and $\alpha \geq 0$, then $\alpha\mathbf{A} \succeq 0$.
c) If $\mathbf{A} \succeq 0$, then $\mathbf{A}^2 \succeq 0$.
d) If $\mathbf{A} \succeq 0$ and $\mathbf{B} \succeq 0$, then $\mathbf{AB} \succeq 0$.
e) If $\mathbf{A} \succ 0$, then $\mathbf{A}^{-1} \succ 0$.
f) If $\mathbf{A} \succeq 0$ and $\mathbf{B} \succeq 0$, then $\mathbf{ABA} \succeq 0$.

Which of these statements are really true? Prove or find counter-examples.

5.23. ($\star$) Consider a random square matrix whose elements are independent random numbers drawn from the normal distribution with zero mean and unit variance. Such matrix is obtained in Matlab by the command `A=randn(n)`. Suppose we generate in this way a large number of matrices. What proportions of them will be positive definite, positive semidefinite, and indefinite? Justify your answer. Try it in Matlab for finite samples of matrices.

Chapter 6

Nonhomogeneous Linear Systems

consider the system of m linear equations in n unknowns

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{6.1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$. This system has (at least one) solution iff $\mathbf{b} \in \operatorname{rng} \mathbf{A}$ (i.e. $\mathbf{b}$ is a linear combination of the columns of $\mathbf{A}$), which can also be written as $\operatorname{rank}[\mathbf{A}\ \mathbf{b}] = \operatorname{rank} \mathbf{A}$ (the Frobenius theorem). The set of solutions of the system is an affine subspace of $\mathbb{R}^n$ (see Exercise 3.4).

The system is **homogeneous** when $\mathbf{b} = \mathbf{0}$ and **nonhomogeneous** when $\mathbf{b} \neq \mathbf{0}$. In this chapter we will concentrate solely on nonhomogeneous systems. We distinguish three cases:

- The system has no solution (this arises typically for $m > n$, though this condition is neither necessary nor sufficient). In this case we may wish to solve the system approximately, which is the subject of section §6.1.
- The system has exactly one solution.
- The system has infinitely many solutions (this arises typically for $m < n$, this condition again being neither necessary nor sufficient). In this case we may wish to choose a single solution from the infinite set of solutions, which is the subject of section §6.2.

6.1 An approximate solution of the system in the least squares sense

When the system (6.1) does not have a solution, solve it approximately. Consider the vector $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$ of the *residuals*) and seek such $\mathbf{x}$, so that its euclidian norm $\|\mathbf{r}\|_2$ is as small as possible. The problem does not change (why?), when instead of the euclidian norm we minimise its square

$$\|\mathbf{r}\|_2^2 = \mathbf{r}^T\mathbf{r} = \sum_{i=1}^m r_i^2 = \sum_{i=1}^m (\mathbf{a}_i^T\mathbf{x} - b_i)^2,$$

where $\mathbf{a}_i^T$ denotes the rows of matrix $\mathbf{A}$. Therefore we are solving the following problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2. \tag{6.2}$$

As we are minimising the sum of squares of the residuals, it is called an approximate solution of the system **in the least squares sense** or the *least squares solution*.

Example 6.1. The system of three equations in two unknowns

$$\begin{aligned} x + 2y &= 6 \\ -x + y &= 3 \\ x + y &= 4 \end{aligned}$$

is over-determined. Its least squares solution means finding such numbers x, y, which minimise $(x + 2y - 6)^2 + (-x + y - 3)^2 + (x + y - 4)^2$. □

It is possible to express many useful problems in the form of (6.2). Sometimes this is not easy to see at the first glance and this can cause some difficulties to students.

Example 6.2. We seek the shortest connecting line between two nonintersecting lines (skew-lines) in the $\mathbb{R}^n$ space. Let i-th line be defined by two points, denoted $\mathbf{p}_i, \mathbf{q}_i \in \mathbb{R}^n$, for $i = 1, 2$. We wish to formulate this problem in the form of (6.2). We are solving the required system

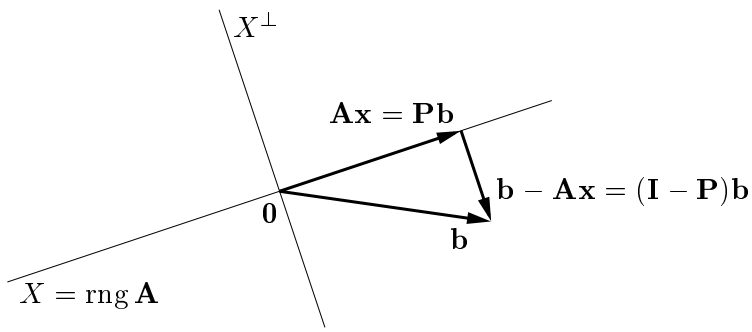
$$\mathbf{p}_1 + t_1(\mathbf{q}_1 - \mathbf{p}_1) \approx \mathbf{p}_2 + t_2(\mathbf{q}_2 - \mathbf{p}_2).$$

This system has n equations in 2 unknowns t_1, t_2. It can be written as $\mathbf{Ax} \approx \mathbf{b}$ where

$$\mathbf{A} = \begin{bmatrix} \mathbf{q}_1 - \mathbf{p}_1 & \mathbf{p}_2 - \mathbf{q}_2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}, \quad \mathbf{b} = \mathbf{p}_2 - \mathbf{p}_1.$$

□

We solve Example (6.2) using the following analysis. In order that $\|\mathbf{Ax} - \mathbf{b}\|_2$ (i.e. the distance between the points $\mathbf{Ax}$ and $\mathbf{b}$) is minimal, then the vector $\mathbf{b} - \mathbf{Ax}$ must be orthogonal to the space $\operatorname{rng} \mathbf{A}$, i.e. to every column of matrix $\mathbf{A}$. The following figure shows the situation:



This condition can be written as $\mathbf{A}^T(\mathbf{Ax} - \mathbf{b}) = \mathbf{0}$, or

$$\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}. \tag{6.3}$$

System (6.3) is therefore called the **normal equation**. It is a system of n equations in n unknowns.

Equation (6.3) can be derived in other ways, too. Example (6.2) seeks the minimum of the quadratic function

$$\begin{aligned} \|\mathbf{Ax} - \mathbf{b}\|_2^2 &= (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b}) \\ &= (\mathbf{x}^T\mathbf{A}^T - \mathbf{b}^T)(\mathbf{Ax} - \mathbf{b}) \\ &= \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - \mathbf{x}^T\mathbf{A}^T\mathbf{b} - \mathbf{b}^T\mathbf{Ax} + \mathbf{b}^T\mathbf{b} \\ &= \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - 2\mathbf{b}^T\mathbf{Ax} + \mathbf{b}^T\mathbf{b}, \end{aligned} \tag{6.4}$$

where we used the fast that a scalar is equal to its transpose and thus $\mathbf{b}^T\mathbf{A}\mathbf{x} = (\mathbf{b}^T\mathbf{A}\mathbf{x})^T = \mathbf{x}^T\mathbf{A}^T\mathbf{b}$. Let us attempt to find an extremum of this quadratic function by completing the square (see §5.3). System (5.11a) will have the form $\mathbf{A}^T\mathbf{A}\mathbf{x}_0 = \mathbf{A}^T\mathbf{b}$ (warning: $\mathbf{A}, \mathbf{b}$ means something different in (6.4) and in (5.11a)), i.e. we obtained the normal equations. At the same time it is clear that the matrix $\mathbf{A}^T\mathbf{A}$ is positive semidefinite, as for every $\mathbf{x} \in \mathbb{R}^n$ we have

$$\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = (\mathbf{A}\mathbf{x})^T\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|_2^2 \geq 0. \tag{6.5}$$

Therefore the point $\mathbf{x}_0$ will be minimum.

When matrix $\mathbf{A}$ is of full rank (i.e. n), then by (6.8) the matrix $\mathbf{A}^T\mathbf{A}$ is regular and the system can be solved by inversion:

$$\mathbf{x} = \mathbf{A}^+\mathbf{b}, \quad \text{kde} \quad \mathbf{A}^+ = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T. \tag{6.6}$$

Matrix $\mathbf{A}^+$ is called the **pseudoinverse** of the (slim) matrix $\mathbf{A}$. It is one of the left inverses of matrix $\mathbf{A}$, since $\mathbf{A}^+\mathbf{A} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{A} = \mathbf{I}$.

When $\mathbf{A}$ is not of full rank, then the matrix $\mathbf{A}^T\mathbf{A}$ is singular and the solution (6.6) can not be used. In that case the system (6.3) and thus also Example (6.2) have an infinite number (an affine subspace) of solutions (warning: this does not mean that the system (6.1) has an infinite number of solutions!).

6.1.1 ($\star$) Solvability of the normal equations

Let us prove that the system (6.3) always has a solution, which is not immediately obvious.

Theorem 6.1.

$$\operatorname{rng}(\mathbf{A}^T\mathbf{A}) = \operatorname{rng}(\mathbf{A}^T) \tag{6.7}$$

where $\mathbf{A}$ *is an arbitrary matrix.*

Proof. First we prove these two statements:

- $\operatorname{null}\mathbf{A} \subseteq \operatorname{null}(\mathbf{A}^T\mathbf{A})$, since $\mathbf{A}\mathbf{x} = \mathbf{0} \;\Rightarrow\; \mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{0}$.
- $\operatorname{null}(\mathbf{A}^T\mathbf{A}) \subseteq \operatorname{null}\mathbf{A}$, since $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{0} \;\Rightarrow\; \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|_2^2 = \mathbf{0} \;\Rightarrow\; \mathbf{A}\mathbf{x} = \mathbf{0}$.

Putting these two statements together, we obtain $\operatorname{null}\mathbf{A} = \operatorname{null}(\mathbf{A}^T\mathbf{A})$. Now applying identity (3.8) to matrices $\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$, it follows that

$$\dim\operatorname{rng}(\mathbf{A}^T\mathbf{A}) = \operatorname{rank}(\mathbf{A}^T\mathbf{A}) = \operatorname{rank}\mathbf{A}^T = \dim\operatorname{rng}(\mathbf{A}^T). \tag{6.8}$$

It follows from definition (3.5) (think about it!), that $\operatorname{rng}(\mathbf{A}^T\mathbf{A}) \subseteq \operatorname{rng}(\mathbf{A}^T)$. However, when a subspace is a subset of another subspace and both subspaces have the same dimension, then they must be the same. This much is clear: an arbitrary basis $\operatorname{rng}(\mathbf{A}^T\mathbf{A})$ is also in $\operatorname{rng}(\mathbf{A}^T)$ and as both subspaces have the same dimension, it is also the basis of $\operatorname{rng}(\mathbf{A}^T)$. □

Corollary 6.2. *System (6.3) has a solution for any* $\mathbf{A}$ *and* $\mathbf{b}$.

Proof. According to (6.7): $\mathbf{A}^T\mathbf{b} \in \operatorname{rng}(\mathbf{A}^T) = \operatorname{rng}(\mathbf{A}^T\mathbf{A})$. □

6.1.2 Solution using QR decomposition

Formula (6.6) is not always best suited to numerical computation (where we necessarily use limited precision arithmetic with finite length representation of numbers), even when matrix $\mathbf{A}$ is of full rank.

Example 6.3. Solve the system $\mathbf{Ax} = \mathbf{b}$ with

$$\mathbf{A} = \begin{bmatrix} 3 & 6 \\ 1 & 2.01 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 9 \\ 3.01 \end{bmatrix}.$$

Matrix $\mathbf{A}$ is regular. Suppose we use floating point arithmetic with precision of three digits. Gaussian elimination will find the exact solution of the system $\mathbf{x} = (1, 1)$. Whereas if we express the normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ in this arithmetic, we get:

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 10 & 20 \\ 20 & 40 \end{bmatrix}, \quad \mathbf{A}^T\mathbf{b} = \begin{bmatrix} 30 \\ 60.1 \end{bmatrix}.$$

The matrix of this system is now singular, since rounding occured in the product $\mathbf{A}^T\mathbf{A}$. □

Numerically more suitable method is to solve the normal equations *without* an explicit evaluation of the $\mathbf{A}^T\mathbf{A}$ product. That can be done using QR decomposition $\mathbf{A} = \mathbf{QR}$. Substituting this into the normal equations we get $\mathbf{R}^T\mathbf{Q}^T\mathbf{QRx} = \mathbf{R}^T\mathbf{Q}^T\mathbf{b}$. Simplifying using $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ and left-multipying by the matrix $\mathbf{R}^{-T}$ (which is an equivalence operation), we have

$$\mathbf{Rx} = \mathbf{Q}^T\mathbf{b}. \tag{6.9}$$

This is the same formula as (4.7), the only difference being that $\mathbf{Q}$ in (4.7) is a square matrix, whereas here it is rectangular.

Matlab implements the solution of the nonhomogeneous linear system by the operator \ (*backslash*). When a system is over-determined, then the result is an approximate solution in the least squares sense and the algorithm uses QR decomposition. Learn to understand how the operators *slash* and *backslash* work by studying the output of the commands `help mrdivide` and `help mldivide`.

6.1.3 More about orthogonal projection

It is instructive to develop further the geometrical reasoning we used to derive the normal equations. Suppose $\mathbf{x}$ is the solution of the normal equations, then vector $\mathbf{Ax}$ is an orthogonal projection of vectoru $\mathbf{b}$ into the subspace $X = \operatorname{rng} \mathbf{A}$ (see figure above). When $\mathbf{A}$ is of full rank, then (6.6) gives

$$\mathbf{Ax} = \mathbf{Pb}, \quad \text{where} \quad \mathbf{P} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T. \tag{6.10}$$

This is an important result: an orthogonal projection of a vector into the subspace X is a linear mapping represented by the matrix $\mathbf{P}$. Therefore this matrix if often called the **projektor**.

Subspace X, which we are projecting into, is represented by the basis (columns of matrix $\mathbf{A}$). Projektor $\mathbf{P}$ should not change when we use a different basis of the subspace. Various basis of the subspace X are represented by the columns of the matrix $\tilde{\mathbf{A}} = \mathbf{AC}$, for various regular matrices $\mathbf{C} \in \mathbb{R}^{n\times n}$ (i.e. $\mathbf{C}$ is the transfer matrix to a different basis). It is easy to verify that, indeed

$$\tilde{\mathbf{A}}(\tilde{\mathbf{A}}^T\tilde{\mathbf{A}})^{-1}\tilde{\mathbf{A}}^T = \mathbf{AC}(\mathbf{C}^T\mathbf{A}^T\mathbf{AC})^{-1}\mathbf{C}^T\mathbf{A}^T = \mathbf{ACC}^{-1}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{C}^{-T}\mathbf{C}^T\mathbf{A}^T = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T.$$

When X is represented by an orthonormal basis, then $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ and the expression (6.10) simplifies to[1] $\mathbf{P} = \mathbf{A}\mathbf{A}^T$. A special case of orthogonal projection is $\dim X = 1$, i.e. the projection onto a line. Let $X = \operatorname{span}\{\mathbf{a}\}$, where we assume $\|\mathbf{a}\|_2 = 1$. Then $\mathbf{P} = \mathbf{a}\mathbf{a}^T$. The formula $(\mathbf{a}^T\mathbf{b})\mathbf{a} = \mathbf{a}\mathbf{a}^T\mathbf{b} = \mathbf{P}\mathbf{b}$ for the projection of vector $\mathbf{b}$ onto a normalised vector $\mathbf{a}$ ought to be familiar to you from the secondary school.

By purely geometrical reasoning we can see what is the range and the null space of the projector. An arbitrary vector from $\mathbb{R}^m$ is projected into subspace X. An arbitrary vector orthogonal to X is projected to the null vector $\mathbf{0}$. Therefore

$$\operatorname{rng}\mathbf{P} = X, \tag{6.11a}$$

$$\operatorname{null}\mathbf{P} = X^\perp. \tag{6.11b}$$

The figure shows that the vector $\mathbf{b} - \mathbf{A}\mathbf{x} = \mathbf{b} - \mathbf{P}\mathbf{b} = (\mathbf{I} - \mathbf{P})\mathbf{b}$ is an orthogonal projection of vectoru $\mathbf{b}$ into $X^\perp$. Therefore the projector into $X^\perp$ is the matrix $\mathbf{I} - \mathbf{P}$. Note that the projector into $X^\perp$ has a natural role in an approximate solution of a system: the value of the minimum in problem (6.2) is $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 = \|\mathbf{b} - \mathbf{P}\mathbf{b}\|_2^2 = \|(\mathbf{I} - \mathbf{P})\mathbf{b}\|_2^2$.

Note about general projection. *Projection* in linear algebra means such linear mapping $\mathbf{f}(\mathbf{y}) = \mathbf{P}\mathbf{y}$, which satisfies $\mathbf{f}(\mathbf{f}(\mathbf{y})) = \mathbf{f}(\mathbf{y})$, i.e. $\mathbf{P}\mathbf{P} = \mathbf{P}^2 = \mathbf{P}$. This expresses an understandable requirement that, once a vector is projected, further projection should leave it unchanged. Projection does not have to be orthogonal in general; it can also be skewed – then the projection is along subspace $\operatorname{null}\mathbf{P}$ into subspace $\operatorname{rng}\mathbf{P}$. Projection is orthogonal, when $\operatorname{null}\mathbf{P} \perp \operatorname{rng}\mathbf{P}$. This[2] occurs exactly when, in addition to $\mathbf{P}^2 = \mathbf{P}$, also $\mathbf{P}^T = \mathbf{P}$ (we leave out the proof of this assertion). Verify that the projector defined by formula (6.10) satisfies $\mathbf{P}^2 = \mathbf{P} = \mathbf{P}^T$.

6.1.4 Using the least squares for regression

Regression is the modelling of the dependency of variable $y \in \mathbb{R}$ on variable $t \in T$ by the regression function

$$y = f(t, \mathbf{x}).$$

The function is known, except for the parameters $\mathbf{x} \in \mathbb{R}^n$. Given a list of pairs (t_i, y_i), $i = 1, \ldots, m$, where measurements of $y_i \in \mathbb{R}$ are subject to errors, the goal is to find parameters $\mathbf{x}$, so that $y_i \approx f(t_i, \mathbf{x})$ for all i. We are minimising the sum of squares of the residuals, i.e. solving the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \sum_{i=1}^{m} [y_i - f(t_i, \mathbf{x})]^2. \tag{6.12}$$

Let us choose the regression function as a linear combination

$$f(t, \mathbf{x}) = x_1\varphi_1(t) + \cdots + x_n\varphi_n(t) = \boldsymbol{\varphi}(t)^T\mathbf{x}$$

[1] Remember (see §4.5), that matrix $\mathbf{A}$ with orthonormal columns need not have orthonormal rows, in other words, $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ does not imply $\mathbf{A}\mathbf{A}^T = \mathbf{I}$. Then the question arises: what is matrix $\mathbf{A}\mathbf{A}^T$? Here you got the answer.

[2] Of course, it is not true that the null space and the range space of a general square matrix are mutually orthogonal. They are even less likely to be orthogonal complements. Do not confuse with relations (4.3)!

of the given functions $\varphi_i\colon T \to \mathbb{R}$. Then

$$\sum_{i=1}^{m} [y_i - f(t_i, \mathbf{x})]^2 = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2,$$

where $\mathbf{y} = (y_1, \ldots, y_m)$ and the elements of matrix $\mathbf{A}$ are $a_{ij} = \varphi_j(t_i)$ (think about it!). Thus we expressed problem (6.12) in the form of (6.2).

Example 6.4. *Polynomial regression.* Let $T = \mathbb{R}$ a $\varphi_i(t) = t^{i-1}$. Then the regression function is the polynomial of degree $n-1$,

$$f(t, \mathbf{x}) = x_1 + x_2 t + x_3 t^2 + \cdots + x_n t^{n-1}.$$

Specifically, for $n = 1$ problem (6.12) becomes $\min_x \sum_i (y_i - x)^2$. The solution is the arithmetic mean (average): $x = \frac{1}{m}\sum_{i=1}^{m} y_i$ (verify!). □

6.2 Least norm solution of a system

Suppose now that the system (6.1) is underdetermined, in other words it has infinitely many solutions. Let $\mathbf{x}'$ be an arbitrary vector satisfying $\mathbf{A}\mathbf{x}' = \mathbf{b}$ (so called **particular solution** of the system). Since for each $\mathbf{x} \in \operatorname{null} \mathbf{A}$, $\mathbf{A}(\mathbf{x}' + \mathbf{x}) = \mathbf{A}\mathbf{x}' = \mathbf{b}$, it is possible to express the set of the solutions of the system parametrically, as

$$\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{b} \,\} = \mathbf{x}' + \operatorname{null} \mathbf{A}. \tag{6.13}$$

It is often useful to pick just one solution from this set of solutions, according to some criteria. A natural criterion is to minimise the euclidian norm of the solution, which results in the problem

$$\min\{\, \|\mathbf{x}\|_2^2 \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{A}\mathbf{x} = \mathbf{b} \,\}. \tag{6.14}$$

Instead of minimising the norm $\|\mathbf{x}\|_2$, we are again minimising its square. This problem is known as solving the nonhomogeneous linear system with the **least norm** (*least norm solution*). Note that sometimes it is appropriate to use other criteria than the least euclidian norm, see e.g. Exercise 10.25.

Example 6.5. The system of two equations in three unknowns

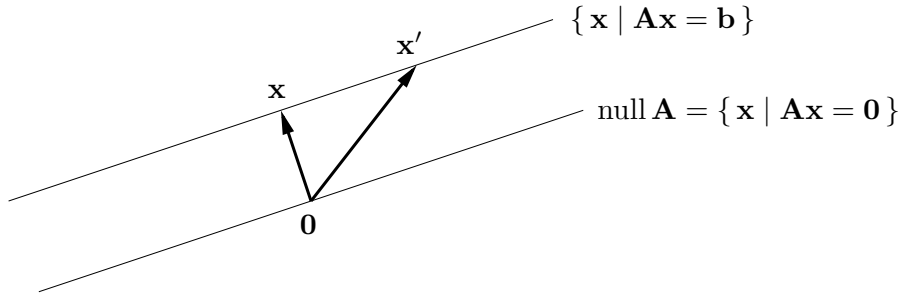
$$\begin{aligned} x + 2y + z &= 1 \\ -x + y + 2z &= 2 \end{aligned}$$

is underdetermined, i.e., it has infinitely many solutions. The solutions set is

$$(x_0, y_0, z_0) + \operatorname{null} \mathbf{A} = (1, -1, 2) + \operatorname{span}\{(1, -1, 1)\} = \{\, (1+\alpha, -1-\alpha, 2+\alpha) \mid \alpha \in \mathbb{R} \,\}.$$

Its least norm solution is the solution which minimises the number $x^2 + y^2 + z^2$. □

Problem (6.14) is easy to solve by the method of Lagrange multipliers. This will be covered in a later chapter. For now we sove it by inspection.



vectors $\mathbf{x}$ and $\mathbf{x}'$ are two different solutions of the system but only $\mathbf{x}$ has the least norm. It is clear that an arbitrary solution $\mathbf{x}$ has the least norm (i.e., is the nearest to the origin $\mathbf{0}$) iff vector $\mathbf{x}$ is orthogonal to the null space of matrix $\mathbf{A}$. According to (4.3a), this means that $\mathbf{x} \in \operatorname{rng}(\mathbf{A}^T)$, i.e., $\mathbf{x}$ must be a linear combination of rows of $\mathbf{A}$. In other words, there must exist some vector $\boldsymbol{\lambda} \in \mathbb{R}^m$, such that $\mathbf{x} = \mathbf{A}^T\boldsymbol{\lambda}$. Thus in order to solve problem (6.14), we must solve the system of equations

$$\mathbf{A}^T\boldsymbol{\lambda} = \mathbf{x}, \tag{6.15a}$$

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{6.15b}$$

This is a system of $m + n$ equations in $m + n$ unknowns $(\mathbf{x}, \boldsymbol{\lambda})$.

Let us solve this system. Substituting $\mathbf{x}$ into the second equation, $\mathbf{A}\mathbf{A}^T\boldsymbol{\lambda} = \mathbf{b}$. Assume that matrix $\mathbf{A}$ is of full rank (i.e. m). Then $\boldsymbol{\lambda} = (\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b}$. Substituting into the first equation, we get

$$\mathbf{x} = \mathbf{A}^+\mathbf{b}, \quad \text{where} \quad \mathbf{A}^+ = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}. \tag{6.16}$$

Matrix $\mathbf{A}^+$ is called the **pseudoinverse** of the (fat) matrix $\mathbf{A}$. It is a right-inverse of matrix $\mathbf{A}$ (verify!).

6.2.1 Pseudoinverse of a general matrix of full rank

Pseudoinverse of a slim matrix was defined earlier by formula (6.6). Summary: when matrix $\mathbf{A}$ is of full rank (i.e. $\max\{m, n\}$), its pseudoinverse is defined as

$$\mathbf{A}^+ = \begin{cases} (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T & \text{when } m \geq n, \\ \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1} & \text{when } m \leq n. \end{cases} \tag{6.17}$$

Vector $\mathbf{x} = \mathbf{A}^+\mathbf{b}$ is in the first case the least squares solution of the system $\mathbf{A}\mathbf{x} = \mathbf{b}$, in the second case it is the least norm solution. When $m = n$, then in both cases $\mathbf{A}^+ = \mathbf{A}^{-1}$ (verify!).

In case $\mathbf{A}$ is not of full rank, then it is not possible to use formula (6.17) and the pseudoinverse has to be defined differently. We will return to this question later, in §7.6.

6.3 Exercises

6.1. Given the system $\mathbf{A}\mathbf{x} = \mathbf{b}$, where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \neq \mathbf{0}$, are the following statements true? Prove your answers, whether positive or negative.

a) When $m < n$, then the system always has a solution.

b) When $m > n$, then the system never has a solution.

c) When $m < n$ and $\mathbf{A}$ is of full rank, then the system always has an infinite number of solutions.

6.2. Solve approximately in the least squares sense the following system, using (a) pseudoinverse, (b) QR decomposition:

$$\begin{bmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & -3 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

6.3. We seek the point $\mathbf{x} \in \mathbb{R}^m$, which minimises the sum of squares of the distances to the given points $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$, i.e., it minimises the expression $\sum_{i=1}^n \|\mathbf{a}_i - \mathbf{x}\|_2^2$. Express the problem in the form of (6.2) (analogously to Example 6.2). Prove that the minimum is attained at the 'center of gravity' $\mathbf{x} = \frac{1}{n}\sum_{i=1}^n \mathbf{a}_i$.

6.4. Given vectors $\mathbf{a}, \mathbf{s}, \mathbf{y} \in \mathbb{R}^n$, find the point $\mathbf{y}$ which is the nearest to the line $\{\, \mathbf{a}+t\mathbf{s} \mid t \in \mathbb{R} \,\}$. Express this problem in the form of (6.2).

6.5. Given vectors $\mathbf{a}, \mathbf{y} \in \mathbb{R}^n$ and scalar $b \in \mathbb{R}$, find the point $\mathbf{y}$ which is the nearest to the superplane $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} = b \,\}$. Express this problem in the form of (6.2).

6.6. Given set m of lines (affine subspaces of dimension 1) in space $\mathbb{R}^n$, where i-th line is the set $\{\, \mathbf{a}_i + t_i\mathbf{s}_i \mid t_i \in \mathbb{R} \,\}$; find the point $\mathbf{y}$ whose sum of squares of distances to the lines is minimal. Express this problem in the form of (6.2). Hint: Minimise over the variables $\mathbf{y}$ and $\mathbf{t} = (t_1, \ldots, t_m)$.

6.7. Expand on Exercise 6.6 for the case where instead of m lines we have m affine subspaces of dimensions $d_1, \ldots, d_m$.

6.8. Given m lines in a plane, where i-th line's equation is $\mathbf{a}_i^T\mathbf{x} = b_i$ for given $\mathbf{a}_i \in \mathbb{R}^2$ and $b_i \in \mathbb{R}$; find the point minimising the sum of squares of distances to each of the lines. Express this problem in the form of (6.2).

6.9. A plank of wood has n holes in it with coordinates $x_1, \ldots, x_n \in \mathbb{R}$, all in one line. We measure distances $d_{ij} = x_j - x_i$ between selected pairs of points $(i, j) \in E$, where set $E \subseteq \{1, \ldots, n\} \times \{1, \ldots, n\}$ is given. The pairs are chosen so that $x_j > x_i$. Use the distances d_{ij} to estimate the coordinates $x_1, \ldots, x_n$. Express this problem in the form of (6.2), i.e., find the matrix $\mathbf{A}$ and the vector $\mathbf{b}$. Is it possible to achieve that $\mathbf{A}$ is of full rank? If not, how would you change the problem so that it is of full rank?

6.10. In the problem of *weighted least squares*, we want to find $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ minimising the function

$$f(\mathbf{x}) = \sum_{i=1}^m w_i \left(\sum_{j=1}^n a_{ij}x_j - b_i \right)^2$$

whre w_i are non-negative weights. Express the function in matrix form (hint: collect the scalars w_i into the diagonal matrix $\mathbf{W} = \operatorname{diag}(\mathbf{w})$). Write down the matrix expression for the optimal $\mathbf{x}$. Under what conditions does this problem have a solution?

6.11. Given vectors $\mathbf{u} = (2, 1, -3)$ and $\mathbf{v} = (1, -1, 1)$; find the orthogonal projections of vector $(2, 0, 1)$ into subspaces (a) $\operatorname{span}\{\mathbf{u}\}$, (b) $(\operatorname{span}\{\mathbf{u}\})^\perp$, (c) $\operatorname{span}\{\mathbf{u}, \mathbf{v}\}$, (d) $(\operatorname{span}\{\mathbf{u}, \mathbf{v}\})^\perp$.

6.12. Let $X = \operatorname{span}\{\,(-\frac{3}{5}, 0, \frac{4}{5}, 0),\ (0, 0, 0, 1),\ (\frac{4}{5}, 0, \frac{3}{5}, 0)\,\}$. Find the projectors into the subspace X and the subspace $X^{\perp}$. Hint: are the vectors orthonormal, per chance?

6.13. Given $\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 4 & 1 \\ 1 & 2 & 0 \end{bmatrix}$, find the orthogonal projections of vector $(1, 1, 1)$ into the subspaes $\operatorname{rng}\mathbf{A}$, $\operatorname{null}\mathbf{A}$, $\operatorname{rng}(\mathbf{A}^T)$, $\operatorname{null}(\mathbf{A}^T)$.

6.14. The null space of a projector is typically non-trivial, i.e. projector $\mathbf{P}$ is a singular matrix. When is $\mathbf{P}$ regular? In that case what are the matrix $\mathbf{A}$ in formula (6.10) and the subspace $X = \operatorname{rng}\mathbf{A}$? What is the geometrical meaning of this situation?

6.15. ($\star$) We have shown in §6.1.3 that the matrix $\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ can be interpreted as a projection into the subspace $\operatorname{rng}\mathbf{A}$. Based on the analysis of §6.2, it is natural to construct a similar matrix $\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}$. What is the geometrical interpretation of this matrix?

6.16. ($\star$) For $\|\mathbf{a}\|_2 = 1$, $\mathbf{H} = \mathbf{I} - 2\mathbf{a}\mathbf{a}^T$ is known as the *Householder's Matrix*. Transformation $\mathbf{H}\mathbf{x}$ is the reflection of vector $\mathbf{x}$ in the superplane with the normal vector $\mathbf{a}$. This is why $\mathbf{H}$ is sometimes also called an *elementary reflector*.

a) Derive the matrix $\mathbf{H}$ using similar reasoning as we used to derive the projector.
b) Show that $\mathbf{H} = \mathbf{H}^T$ and $\mathbf{H}^T\mathbf{H} = \mathbf{I}$ (i.e., matrix H is symmetric and orthogonal).
c) It follows from the above two properties that $\mathbf{H}\mathbf{H} = \mathbf{I}$. What does that say about the transformation $\mathbf{H}\mathbf{x}$?
d) Show that $\det\mathbf{H} = -1$.
e) What is $\mathbf{H}\mathbf{a}$? What is $\mathbf{H}\mathbf{x}$, when $\mathbf{a}^T\mathbf{x} = 0$? Demonstrate your answers algebraically and justify (explain) them geometrically.

6.17. ($\star$) *RQ decomposition* decomposes matrix $\mathbf{A} = \mathbf{R}\mathbf{Q}$, where $\mathbf{R}$ is upper triangular and $\mathbf{Q}$ is orthogonal. How would you calculate the RQ decomposition from the QR decomposition?

6.18. ($\star$) Matrix $\mathbf{A}$ is *normal*, when $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T$. An example is a symmetric matrix (but not all normal matrices are symmetric). Prove that $\operatorname{rng}\mathbf{A} \perp \operatorname{null}\mathbf{A}$ for normal matrix A. Hint: start with (6.7).

6.19. Given an arbitrary matrix A of full rank, prove the following properties of its pseudoinverse from (6.17):

a) $\mathbf{A}^+ = \mathbf{A}^{-1}$ when $\mathbf{A}$ is square
b) $(\mathbf{A}^+)^+ = \mathbf{A}$
c) $(\mathbf{A}^T)^+ = (\mathbf{A}^+)^T$
d) $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$, $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$, $(\mathbf{A}\mathbf{A}^+)^T = \mathbf{A}\mathbf{A}^+$, $(\mathbf{A}^+\mathbf{A})^T = \mathbf{A}^+\mathbf{A}$
e) $\mathbf{A}^T = \mathbf{A}^T\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+\mathbf{A}\mathbf{A}^T$
f) $(\mathbf{A}^T\mathbf{A})^+ = \mathbf{A}^+(\mathbf{A}^T)^+$, $(\mathbf{A}\mathbf{A}^T)^+ = (\mathbf{A}^T)^+\mathbf{A}^+$

Chapter 7

Singular Values Decomposition (SVD)

Every matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ can be decomposed as

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{7.1}$$

where

- $\mathbf{S} \in \mathbb{R}^{m\times n}$ is diagonal. Its diagonal elements $\sigma_1, \ldots, \sigma_p$, where $p = \min\{m, n\}$, are the **singular numbers** of matrix $\mathbf{A}$. put them in descending order $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$. When this condition is satisfied, then the singular numbers are uniquely determined by the matrix.
- $\mathbf{U} \in \mathbb{R}^{m\times m}$, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. The columns of matrix $\mathbf{U}$ are **left singular vectors** of matrix $\mathbf{A}$.
- $\mathbf{V} \in \mathbb{R}^{n\times n}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. The columns of matrix $\mathbf{V}$ are **right singular vectors** of matrix $\mathbf{A}$.

Decomposition (7.1) is called **SVD** (*Singular Value Decomposition*).

The number of non-zero singular numbers is equal to the rank of the matrix $\mathbf{A}$. Let $r = \operatorname{rank}\mathbf{A} \leq p$ be the number of non-zero singular numbers. Then (7.1) can be written as

$$\mathbf{A} = \underbrace{\begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix}}_{\mathbf{U}} \underbrace{\begin{bmatrix}\mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}}_{\mathbf{S}} \underbrace{\begin{bmatrix}\mathbf{V}_1^T \\ \mathbf{V}_2^T\end{bmatrix}}_{\mathbf{V}^T} = \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^T \tag{7.2}$$

where $\mathbf{S}_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$ is square diagonal matrix whose diagonal consists of all non-zero singular numbers. The sizes of the blocks $\mathbf{U}_1, \mathbf{U}_2, \mathbf{V}_1, \mathbf{V}_2$ and of the zero blocks are determined by the size of the matrix $\mathbf{S}_1$ (when some block has one zero dimension, it is considered to be empty). The decomposition $\mathbf{A} = \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^T$ is called **Reduced SVD**.

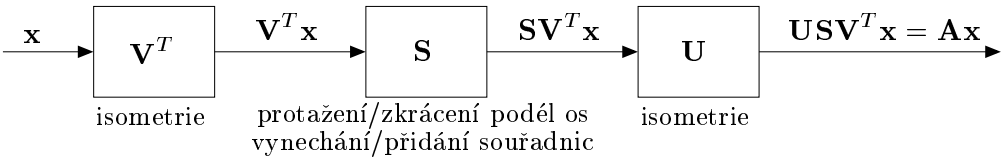
Reduced SVD is obtained from full SVD (7.1) by cutting matrix $\mathbf{S}$ to make it square $r \times r$, leaving out the last $m-r$ columns from matrix $\mathbf{U}$ and leaving out the last $n-r$ columns from matrix $\mathbf{V}$. Full SVD is obtained from reduced SVD by adding columns to slim matrices $\mathbf{U}_1$ and $\mathbf{V}_1$ to make them square orthogonal, and adding zeros to the square matrix $\mathbf{S}$ to make it rectangular of the same dimensions as $\mathbf{A}$.

Example 7.1. Here is an example of the full and reduced SVDs of a 2×3 matrix:

$$\mathbf{A} = \begin{bmatrix}3 & 2 & 2 \\ 2 & 3 & -2\end{bmatrix} = \begin{bmatrix}1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2}\end{bmatrix}\begin{bmatrix}5 & 0 & 0 \\ 0 & 3 & 0\end{bmatrix}\begin{bmatrix}1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{18} & -1/\sqrt{18} & 4/\sqrt{18} \\ 2/3 & -2/3 & 1/3\end{bmatrix} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

$$= \begin{bmatrix}1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2}\end{bmatrix}\begin{bmatrix}5 & 0 \\ 0 & 3\end{bmatrix}\begin{bmatrix}1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{18} & -1/\sqrt{18} & 4/\sqrt{18}\end{bmatrix} = \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^T \qquad \square$$

SVD is a powerful tool for analysing linear mapping represented by matrix $\mathbf{A}$. Formula (7.1) reveals that every linear mapping is a composition of three simpler linear mappings, specifically of isometry $\mathbf{V}^T$, diagonal mapping $\mathbf{S}$ and isometry $\mathbf{U}$. Linear mapping represented by a diagonal matrix is simply stretching or shrinking along the coordinate axes. Possibly, when the matrix is fat, it means leaving out some coordinate axes or, when the matrix is slim, adding zero coordinates.



In the language of the basis it means that for any linear mapping it is possible to find orthonormal bases of the domain space and of the co-domain space, such that with respect to these bases, the mapping is diagonal.

Matlab command `[U,S,V]=svd(A)` calculates the full SVD. The reduced SVD is not directly implemented but can be easily obtained by using the command `[U,S,V]=svd(A,'econ')`, which returns $\mathbf{U} \in \mathbb{R}^{m\times p}$, $\mathbf{S} \in \mathbb{R}^{p\times p}$ and $\mathbf{V} \in \mathbb{R}^{n\times p}$.

Note on numerical linear algebra. We introduced already three different matrix decompositions: QR, spectral decomposition, and SVD. There are many more. The design of numerical algorithms for matrix operations, solutions of systems of linear equations and decompositions of matrices by vectors is the subject of the *numerical linear algebra*. Freely accessible software packages for numerical linear algebra do exist, for example LAPACK and BLAS. Matlab is built on top of the LAPACK package.

7.1 SVD from spectral decomposition

Let (7.1) be satisfied. Then

$$\mathbf{A}^T\mathbf{A} = \mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S}\mathbf{V}^T = \mathbf{V}\mathbf{S}^T\mathbf{S}\mathbf{V}^T, \qquad \text{where} \quad \mathbf{S}^T\mathbf{S} = \operatorname{diag}(\sigma_1^2, \dots, \sigma_r^2, \underbrace{0, \dots, 0}_{(n-r)}), \tag{7.3a}$$

$$\mathbf{A}\mathbf{A}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{V}\mathbf{S}^T\mathbf{U}^T = \mathbf{U}\mathbf{S}\mathbf{S}^T\mathbf{U}^T, \qquad \text{where} \quad \mathbf{S}\mathbf{S}^T = \operatorname{diag}(\sigma_1^2, \dots, \sigma_r^2, \underbrace{0, \dots, 0}_{(m-r)}). \tag{7.3b}$$

Note that (7.3a) is the spectral decomposition of the symmetric matrix $\mathbf{A}^T\mathbf{A}$ (see §5.1.1). The diagonal elements of the matrix $\mathbf{S}^T\mathbf{S}$ are the eigenvalues of the matrix $\mathbf{A}^T\mathbf{A}$. They are non-negative, which is in accord with $\mathbf{A}^T\mathbf{A}$ being positive semidefinite (see (6.5)). The columns of the orthogonal matrix $\mathbf{V}$ are eigenvectors of the matrix $\mathbf{A}^T\mathbf{A}$.

Similarly, (7.3b) is the spectral decomposition of the symmetric positive definite matrix $\mathbf{A}\mathbf{A}^T$.

So we see that the right singular vectors of matrix $\mathbf{A}$ are eigenvectors of the matrix $\mathbf{A}^T\mathbf{A}$, the left singular vectors of matrix $\mathbf{A}$ are eigenvectors of the matrix $\mathbf{A}\mathbf{A}^T$, and that non-zero singular numbers of matrix $\mathbf{A}$ are square roots of the non-zero eigenvalues of the $\mathbf{A}^T\mathbf{A}$ (and also of $\mathbf{A}\mathbf{A}^T$).

Thus we demonstrated that decomposition (7.1) exists and can be found using the spectral decomposition. This computation is not numerically satisfactory, since an explicit computation

of the matrices $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ can lead to rounding errors (see §6.1.2). Therefore SVD is typically computed by algorithms which manage to avoid the computation of these matrices. On the other hand, when we do not mind the loss of precision, then the computation of the SVD by spectral decomposition can be faster. For instance, when $m \ll n$ and we need to compute only the matrices $\mathbf{U}$ and $\mathbf{S}$ (and do not need $\mathbf{V}$), then the spectral decomposition of the matrix $\mathbf{A}\mathbf{A}^T$ will be typically faster, as the size of this matrix is small ($m \times m$).

7.2 Orthonormal basis of the fundamental subspaces of a matrix

SVD reveals orthonormal basis of all four fundamental subspaces generated by matrix $\mathbf{A}$ (see §4.4), as

$$\operatorname{rng} \mathbf{U}_1 = \operatorname{rng} \mathbf{A}, \tag{7.4a}$$
$$\operatorname{rng} \mathbf{V}_1 = \operatorname{rng}(\mathbf{A}^T), \tag{7.4b}$$
$$\operatorname{rng} \mathbf{U}_2 = \operatorname{null}(\mathbf{A}^T), \tag{7.4c}$$
$$\operatorname{rng} \mathbf{V}_2 = \operatorname{null} \mathbf{A}. \tag{7.4d}$$

Identity (7.4a) can be proven as follows:

$$\operatorname{rng} \mathbf{A} = \{\, \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n \,\} \overset{\text{(a)}}{=} \{\, \mathbf{U}_1\mathbf{S}_1\mathbf{y} \mid \mathbf{y} \in \mathbb{R}^r \,\} \overset{\text{(b)}}{=} \{\, \mathbf{U}_1\mathbf{z} \mid \mathbf{z} \in \mathbb{R}^r \,\} = \operatorname{rng} \mathbf{U}_1.$$

Here the identity marked (a) is valid because $\mathbf{V}_1$ is of full rank and thus (by the Frobenius Theorem) for each $\mathbf{y} \in \mathbb{R}^r$ there exists $\mathbf{x} \in \mathbb{R}^n$ satisfying $\mathbf{y} = \mathbf{V}_1^T\mathbf{x}$. In other words, $\operatorname{rng}(\mathbf{V}_1^T) = \mathbb{R}^r$. The identity marked (b) is valid for the similar reason: $\mathbf{S}_1$ is square regular, thus $\operatorname{rng} \mathbf{S}_1 = \mathbb{R}^r$.

Identity (7.4b) follows from (7.4a), as $\mathbf{A}^T = (\mathbf{U}_1\mathbf{S}_1\mathbf{V}_1)^T = \mathbf{V}_1\mathbf{S}_1^T\mathbf{U}_1^T = \mathbf{V}_1\mathbf{S}_1\mathbf{U}_1^T$.

Matrices $\mathbf{U}$ and $\mathbf{V}$ are orthogonal. Thus from the definition of orthogonal complement it is clear that $(\operatorname{rng} \mathbf{U}_1)^\perp = \operatorname{rng} \mathbf{U}_2$ and $(\operatorname{rng} \mathbf{V}_1)^\perp = \operatorname{rng} \mathbf{V}_2$. Identities (7.4c) and (7.4d) now follow from (4.3).

7.3 The nearest matrix of a lower rank

Frobenius norm of matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is the number

$$\|\mathbf{A}\|_{\mathrm{F}} = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \right)^{1/2} = \left(\sum_{j=1}^{n} \|\mathbf{a}_j\|_2^2 \right)^{1/2} \tag{7.5}$$

where $\mathbf{a}_1, \ldots, \mathbf{a}_n$ are the columns of matrix $\mathbf{A}$. Since clearly $\|\mathbf{A}\|_{\mathrm{F}} = \|\mathbf{A}^T\|_{\mathrm{F}}$, we could also write rows instead of columns in (7.5). Similarly to the euclidian norm, the Frobenius norm does not change under an isometric transformation of rows or columns of a matrix, or

$$\mathbf{U}^T\mathbf{U} = \mathbf{I}, \;\; \mathbf{V}^T\mathbf{V} = \mathbf{I} \quad \Longrightarrow \quad \|\mathbf{A}\|_{\mathrm{F}} = \|\mathbf{U}\mathbf{A}\|_{\mathrm{F}} = \|\mathbf{A}\mathbf{V}^T\|_{\mathrm{F}} = \|\mathbf{U}\mathbf{A}\mathbf{V}^T\|_{\mathrm{F}}. \tag{7.6}$$

This follows easily (think about it!) from (7.5).

Consider the problem where given matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ of rank r, we wish to find the nearest (in the Frobenius norm sense) matrix $\mathbf{A}'$ of a given lower rank $r' \le r$. So, we are solving the problem:

$$\min\{\, \|\mathbf{A} - \mathbf{A}'\|_{\mathrm{F}} \mid \mathbf{A}' \in \mathbb{R}^{m\times n},\ \operatorname{rank} \mathbf{A}' = r' \,\}. \tag{7.7}$$

Theorem 7.1 (Eckart-Young). *Let SVD matrix* $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$, *where* $\mathbf{S} = \operatorname{diag}(\sigma_1, \dots, \sigma_r)$. *Let* $\mathbf{S}' = \operatorname{diag}(\sigma'_1, \dots, \sigma'_r)$, *where*

$$\sigma'_i = \begin{cases} \sigma_i & \text{when } i \le r', \\ 0 & \text{when } i > r'. \end{cases}$$

Then the solution of problem (7.7) is matrix $\mathbf{A}' = \mathbf{U}\mathbf{S}'\mathbf{V}^T$ *and*

$$\|\mathbf{A} - \mathbf{A}'\|_{\mathrm{F}} = (\sigma_{r'+1}^2 + \cdots + \sigma_r^2)^{1/2}. \tag{7.8}$$

We present the main part of this theorem without a proof. We will prove only the assertion (7.8). Using (7.6), we have:

$$\|\mathbf{A} - \mathbf{A}'\|_{\mathrm{F}} = \|\mathbf{U}\mathbf{S}\mathbf{V}^T - \mathbf{U}\mathbf{S}'\mathbf{V}^T\|_{\mathrm{F}} = \|\mathbf{U}(\mathbf{S} - \mathbf{S}')\mathbf{V}^T\|_{\mathrm{F}} = \|\mathbf{S} - \mathbf{S}'\|_{\mathrm{F}} = (\sigma_{r'+1}^2 + \cdots + \sigma_r^2)^{1/2}.$$

In this sense the *singular numbers give the distance of matrix A to the matrix of a given lower rank*.

The theorem says that we can find the nearest matrix of the given lower rank r' by setting to zero $r - r'$ smallest singular numbers in the SVD of the original matrix (so that the number of the remaining singular numbers is r'). Putting it another way, SVD decomposition (7.1) can be written as the sum

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \tag{7.9}$$

where $\mathbf{u}_1, \dots, \mathbf{u}_m$ are the columns of matrix $\mathbf{U}$ and $\mathbf{v}_1, \dots, \mathbf{v}_n$ are the columns of matrix $\mathbf{V}$. Note that $\mathbf{u}_i\mathbf{v}_i^T \in \mathbb{R}^{m\times n}$ is matrix of rank 1 (see §2.5). The matrix of a lower rank is obtained by taking only the first r' terms of this sum:

$$\mathbf{A}' = \mathbf{U}\mathbf{S}'\mathbf{V}^T = \sum_{i=1}^{r} \sigma'_i \mathbf{u}_i \mathbf{v}_i^T = \sum_{i=1}^{r'} \sigma_i \mathbf{u}_i \mathbf{v}_i^T.$$

We see that the singular numbers give not just the rank of a matrix but by (7.8) also tell us how 'far' the matrix is from the matrix of a given lower rank. Singular vectors not only define the orthonormal bases of all the fundamental subspaces of a matrix by (7.4) but in addition they show how these subspaces would change, should the matrix be substituted by one of a given lower rank.

7.4 Fitting a subspace to given points

We seek the (linear) subspace $X \subseteq \mathbb{R}^m$ of a given dimension that minimises the sum of squares of the distances to the given points[1] $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^m$. This task can not be turned into the least squares problem of §6.1. However, it can be solved by using Theorem 7.1:

$$r = \operatorname{rank} \mathbf{A} = \dim \operatorname{span}\{\mathbf{a}_1, \dots, \mathbf{a}_n\}, \tag{7.10}$$

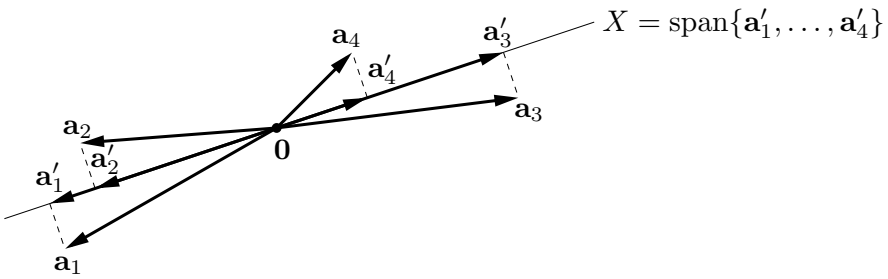
$$r' = \operatorname{rank} \mathbf{A}' = \dim \operatorname{span}\{\mathbf{a}'_1, \dots, \mathbf{a}'_n\}, \tag{7.11}$$

[1] This problem is called the *principal component analysis (PCA)* or *Karhunen-Loewe transform* in statistics.

where $\mathbf{a}_j$ are the columns of matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{a}'_j$ are the columns of matrix $\mathbf{A}'$. Further,

$$\|\mathbf{A} - \mathbf{A}'\|_{\mathrm{F}}^2 = \sum_{j=1}^{n} \|\mathbf{a}_j - \mathbf{a}'_j\|_2^2.$$

I.e. $X = \operatorname{span}\{\mathbf{a}'_1, \dots, \mathbf{a}'_n\} = \operatorname{rng} \mathbf{A}'$ is such subspace of dimension r', that the sum of squares of the perpendicular distances of points $\mathbf{a}_1, \dots, \mathbf{a}_n$ from this subspace is minimal:



Usually we need not find the points $\mathbf{a}'_j$ but only the subspace X. We can easily find its orthonormal basis from the relationships (7.4). Since only the first r' singular numbers of matrix $\mathbf{A}'$ are non-zero, the basis of the subspace $X = \operatorname{rng} \mathbf{A}'$ is the set of the first r' columns of matrix $\mathbf{U}$ in the decomposition $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$. Sometimes it can be more advantageous to seek the orthogonal complement $X^\top = \operatorname{null}(\mathbf{A}')^T$ of the desired subspace. Its basis is the last $m - r'$ columns of matrix $\mathbf{U}$.

Example 7.2. Given n points $\mathbf{a}_1, \dots, \mathbf{a}_n$ in the space $\mathbb{R}^3$. Let the full SVD of the matrix, whose columns are the given points, be $\begin{bmatrix}\mathbf{a}_1 \cdots \mathbf{a}_n\end{bmatrix} = \mathbf{U}\mathbf{S}\mathbf{V}^T$. Denote the columns of matrix $\mathbf{U} \in \mathbb{R}^{3\times 3}$ as $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$.

Find the line passing through the origin, such that the sum of squares of the perpendicular distances of these points from the line is minimal. Such line is the set

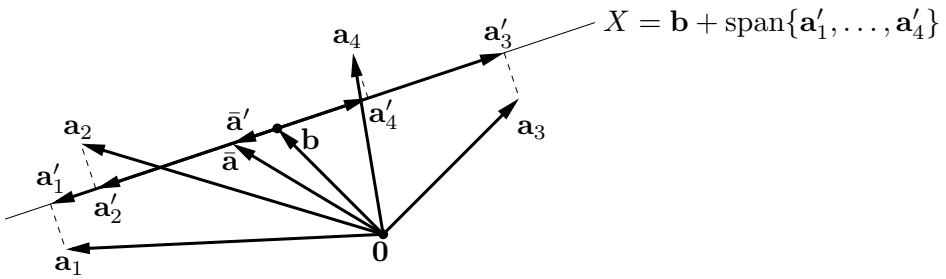
$$\operatorname{span}\{\mathbf{u}_1\} = \{\, \alpha\mathbf{u}_1 \mid \alpha \in \mathbb{R} \,\} = \{\, \mathbf{x} \in \mathbb{R}^3 \mid \mathbf{u}_2^T\mathbf{x} = \mathbf{u}_3^T\mathbf{x} = 0 \,\} = \operatorname{span}\{\mathbf{u}_2, \mathbf{u}_3\}^\perp.$$

Find the plane passing through the origin, such that the sum of squares of the perpendicular distances of these points from the plane is minimal. Such plane is the set

$$\operatorname{span}\{\mathbf{u}_1, \mathbf{u}_2\} = \{\, \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 \mid \alpha_1, \alpha_2 \in \mathbb{R} \,\} = \{\, \mathbf{x} \in \mathbb{R}^3 \mid \mathbf{u}_3^T\mathbf{x} = 0 \,\} = \operatorname{span}\{\mathbf{u}_3\}^\perp. \qquad \square$$

7.4.1 Generalisation to affine subspace

Genealising the previous problem, now instead of the linear subspace we seek the *affine* subspace of dimension r' that minimises the sum of squares of perpendicular distances from the points $\mathbf{a}_1, \dots, \mathbf{a}_n$. This affine subspace can be written as $X = \mathbf{b} + \operatorname{span}\{\mathbf{a}'_1, \dots, \mathbf{a}'_n\}$ for some translation $\mathbf{b} \in \mathbb{R}^m$ (see §3.3):



The sum of squares of perpendicular distances from X is (consult the figure)

$$\sum_{j=1}^{n} \|\mathbf{a}_j - \mathbf{a}'_j - \mathbf{b}\|_2^2 = \|\mathbf{A} - \mathbf{A}' - \mathbf{b}\mathbf{1}^T\|_{\mathrm{F}}^2. \tag{7.12}$$

We seek $\mathbf{b} \in \mathbb{R}^m$ and $\mathbf{A}' \in \mathbb{R}^{m\times n}$, which minimise (7.12) given the condition that $\operatorname{rank} \mathbf{A}' = r'$.

When $\mathbf{A}'$ is fixed, the minimisation of (7.12) with respect to variable $\mathbf{b}$ can be easily solved explicitly (see Exercise 6.3): the minimum is achieved at the point

$$\mathbf{b} = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{a}_j - \mathbf{a}'_j) = \bar{\mathbf{a}} - \bar{\mathbf{a}}',$$

where $\bar{\mathbf{a}} = \frac{1}{n}(\mathbf{a}_1 + \cdots + \mathbf{a}_n)$ and $\bar{\mathbf{a}}' = \frac{1}{n}(\mathbf{a}'_1 + \cdots + \mathbf{a}'_n)$. As $\bar{\mathbf{a}}' \in \operatorname{span}\{\mathbf{a}'_1, \ldots, \mathbf{a}'_n\}$, then $\bar{\mathbf{a}} = \mathbf{b} + \bar{\mathbf{a}}' \in X$ (see the figure). Thus we proved that the optimal affine subspace X passes through the 'center of gravity' $\bar{\mathbf{a}}$ of points $\mathbf{a}_1, \ldots, \mathbf{a}_n$.

Now the solution is clear. We seek the affine subspace passing through the point $\mathbf{b}$, which minimises the sum of squares of the distances to points $\mathbf{a}_1, \ldots, \mathbf{a}_n$. Therefore it is sufficient to first translate all the points so as to place their center at the origin and then to find the linear subspace that minimises the sum of squares of the distances to the translated points.

7.5 Approximate solution of homogeneous systems

Let us solve the homogeneous linear system

$$\mathbf{A}\mathbf{x} = \mathbf{0} \tag{7.13}$$

for $\mathbf{A} \in \mathbb{R}^{m\times n}$. The set of solutions is the set $\operatorname{null} \mathbf{A}$, which is a linear subspace of $\mathbb{R}^n$ of dimension $d = n - \operatorname{rank} \mathbf{A}$ (see §3.2.1). One of the solutions is always $\mathbf{x} = \mathbf{0}$ (so called trivial solution).

Can a homogeneous system be over-determined? Over-determined can be defined as dimension d of the solutions space being less than some given dimension $d' > d$. A special case is when the system has only the trivial solution ($d = 0$) but we would like a non-trivial solution. Let us solve the system approximately, so that matrix $\mathbf{A}$ is changed as little as possible, while the solution space gains the desired dimension d'. In other words we first find the matrix $\mathbf{A}'$ of rank $n - d'$ which is the nearest to matrix $\mathbf{A}$ (by Theorem 7.1) and then solve the system $\mathbf{A}'\mathbf{x} = \mathbf{0}$.

Relationship to the nonhomogeneous case. In §6.1 we formulated an approximate solution of nonhomogeneous ($\mathbf{b} \neq \mathbf{0}$) system $\mathbf{A}\mathbf{x} = \mathbf{b}$ as the problem $\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$. It may appear that this formulation is totally different from the formulation of an approximate solution of a homogeneous system given here. However, this is not the case. Let us formulate an approximate solution of a nonhomogeneous system as follows: when the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ has no solution, change vector $\mathbf{b}$ as little as possible, such that the system has a solution. More precisely, we seek the vector $\mathbf{b}'$ such that for some $\mathbf{x}$, $\mathbf{A}\mathbf{x} = \mathbf{b}'$ and the number $\|\mathbf{b} - \mathbf{b}'\|_2$ is as small as possible. This problem can be written as

$$\min\{\, \|\mathbf{b} - \mathbf{b}'\|_2 \mid \mathbf{A}\mathbf{x} = \mathbf{b}',\ \mathbf{x} \in \mathbb{R}^n,\ \mathbf{b}' \in \mathbb{R}^m \,\}.$$

Here we minimise with respect to the variables $\mathbf{x}$ and $\mathbf{b}'$ (it does not matter that $\mathbf{x}$ does not occur in the criterion). It is possible to simplify this problem (think about it!): substituting $\mathbf{b}' = \mathbf{A}\mathbf{x}$ into the criterion $\|\mathbf{b} - \mathbf{b}'\|_2$, gives $\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$. To sum up:

- In an approximate solution of nonhomogeneous system $\mathbf{A}\mathbf{x} = \mathbf{b}$, change vector $\mathbf{b}$ as little as possible, so that the system has a solution.
- In an approximate solution of homogeneous system $\mathbf{A}\mathbf{x} = \mathbf{0}$, change matrix $\mathbf{A}$ as little as possible, so that the system has the solutions space of a given dimension.

7.6 ($\star$) Pseudoinverse of a general matrix

Let us now return to the solution of nonhomogeneous linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ for $\mathbf{A} \in \mathbb{R}^{m\times n}$. In §6 we separately discussed the cases where the system had none, one, or infinitely many solutions. Now we merge all these cases into just one general formulation

$$\min \left\{ \|\mathbf{x}\|_2^2 \,\middle|\, \mathbf{x} \in \operatorname*{argmin}_{\mathbf{x}'\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x}' - \mathbf{b}\|_2^2 \right\}. \tag{7.14}$$

That means we seek vector $\mathbf{x}$ for which the number $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ is minimal; should there be several such vectors, we select the one with the smallest norm $\|\mathbf{x}\|_2$.

Let SVD of matrix $\mathbf{A}$ be given by formula (7.2). Then:

$$\begin{aligned}
\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 &= \|\mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{x} - \mathbf{b}\|_2^2 \\
&= \|\mathbf{U}^T(\mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{x} - \mathbf{b})\|_2^2 && \text{as } \|\mathbf{U}^T\mathbf{z}\|_2 = \|\mathbf{z}\|_2 \text{ for each } \mathbf{z} \\
&= \|\mathbf{S}\mathbf{V}^T\mathbf{x} - \mathbf{U}^T\mathbf{b}\|_2^2 && \text{as } \mathbf{U}^T\mathbf{U} = \mathbf{I} \\
&= \|\mathbf{S}\mathbf{y} - \mathbf{c}\|_2^2 \\
&= \left\| \begin{bmatrix} \mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} - \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix} \right\|_2^2 \\
&= \left\| \begin{bmatrix} \mathbf{S}_1\mathbf{y}_1 - \mathbf{c}_1 \\ -\mathbf{c}_2 \end{bmatrix} \right\|_2^2 \\
&= \|\mathbf{S}_1\mathbf{y}_1 - \mathbf{c}_1\|_2^2 + \|\mathbf{c}_2\|_2^2,
\end{aligned} \tag{7.15}$$

where

$$\mathbf{V}^T\mathbf{x} = \begin{bmatrix} \mathbf{V}_1^T\mathbf{x} \\ \mathbf{V}_2^T\mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \mathbf{y}, \qquad \mathbf{U}^T\mathbf{b} = \begin{bmatrix} \mathbf{U}_1^T\mathbf{b} \\ \mathbf{U}_2^T\mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix} = \mathbf{c}. \tag{7.16}$$

What have we achieved here? We have shown that the exression $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ is equal to the expression (7.15), which is much easier to minimise, since matrix $\mathbf{S}_1$ is diagonal and regular. The minimum of (7.15) is thus achieved for $\mathbf{y}_1 = \mathbf{S}_1^{-1}\mathbf{c}_1$, as then $\mathbf{S}_1\mathbf{y}_1 = \mathbf{c}_1$. Since $\mathbf{S}_1$ is diagonal, its inverse is simply $\mathbf{S}_1^{-1} = \operatorname{diag}(\sigma_1^{-1}, \ldots, \sigma_r^{-1})$.

Expression (7.15) does not depend on vector $\mathbf{y}_2$, which can thus be chosen arbitrarily. Let us choose it such that vector $\mathbf{y}$ has the smallest norm. This will evidently occur when $\mathbf{y}_2 = \mathbf{0}$. Additionally, $\mathbf{x}$ will also have the smallest norm because $\|\mathbf{y}\|_2 = \|\mathbf{V}^T\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ (follows from orthogonality of $\mathbf{V}$).

The solution of problem (7.14) is obtained by back-substitution from (7.16):

$$\mathbf{x} = \mathbf{V}\mathbf{y} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \end{bmatrix} \begin{bmatrix} \mathbf{S}_1^{-1}\mathbf{c}_1 \\ \mathbf{0} \end{bmatrix} = \mathbf{V}_1\mathbf{S}_1^{-1}\mathbf{c}_1 = \mathbf{V}_1\mathbf{S}_1^{-1}\mathbf{U}_1^T\mathbf{b}.$$

The matrix

$$\mathbf{A}^+ = \mathbf{V}_1\mathbf{S}_1^{-1}\mathbf{U}_1^T \tag{7.17}$$

is the **pseudoinverse** (of the general) matrix $\mathbf{A}$. It is also called the **Moore-Penrose pseudoinverse**. When $\mathbf{A}$ is of full rank, then this definition agrees with formulae (6.6) and (6.16) (verify!).

Note that while we needed the full SVD for the derivation of formula (7.17), only reduced SVD occurs in the formula itself. Matrices $\mathbf{U}_2$ and $\mathbf{V}_2$ were needed only for the derivation of the formula.

7.7 Exercises

7.1. Given the matrices

$$\mathbf{A} = \begin{bmatrix} 0.528 & 0.896 & -0.72 \\ -1.204 & -0.528 & 0.96 \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} 0.6 & -0.8 \\ -0.8 & 0.6 \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} 0.64 & 0.6 & -0.48 \\ 0.48 & -0.8 & -0.36 \\ -0.6 & 0 & -0.8 \end{bmatrix}.$$

Calculate the matrix $\mathbf{B}$ of rank one, such that $\|\mathbf{A} - \mathbf{B}\|_{\mathrm{F}}$ is minimal (where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm). Find the value of $\|\mathbf{A} - \mathbf{B}\|_{\mathrm{F}}$ for the matrix $\mathbf{B}$.

Answer: $\|\mathbf{A} - \mathbf{B}\|_{\mathrm{F}} = 0.5$.

7.2. Find the orthonormal basis of the subspace $\operatorname{span}\{\,(1,1,1,-1),\ (2,-1,-1,1),\ (-1,2,2,1)\,\}$ using SVD.

7.3. ($\star$) Solve the system of Exercise 6.2 approximately, in the least squares sense, using SVD.

7.4. Given $\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 4 & 1 \\ 1 & 2 & 0 \end{bmatrix}$, find the orthonormal bases of subspaces $\operatorname{rng}\mathbf{A}$, $\operatorname{null}\mathbf{A}$, $\operatorname{rng}(\mathbf{A}^T)$, $\operatorname{null}(\mathbf{A}^T)$. You may use a computer.

7.5. ($\star$) Prove the properties of the pseudoinverse in Exercise 6.19, using (7.17) for arbitrary (square or rectangular) matrices of any rank.

Chapter 8

Nonlinear Functions and Mappings

In previous chapters we encountered the linear and affine mappings and quadratic functions. In this chapter we will consider in more detail the nonlinear functions $f\colon \mathbb{R}^n \to \mathbb{R}$ and the mappings $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$. Let us revise the functions and mappings notation from §1.1.3:

Example 8.1. Examples of functions and mappings of several variables:

1. $f\colon \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 - y^2$
2. $f\colon \mathbb{R}^n \to \mathbb{R}$, $f(\mathbf{x}) = x_1$ (even when $x_2, \ldots, x_n$ is missing, f is still understood to be a function of n variables)
3. $f\colon \mathbb{R}^n \to \mathbb{R}$, $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$ (linear function)
4. $f\colon \mathbb{R}^n \to \mathbb{R}$, $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$ (affine function)
5. $f\colon \mathbb{R}^n \to \mathbb{R}$, $f(\mathbf{x}) = \mathrm{e}^{-\|\mathbf{x}\|_2^2}$
6. $f\colon \mathbb{R}^n \to \mathbb{R}$, $f(\mathbf{x}) = \max\limits_{i=1}^{n} x_i$
7. $\mathbf{f}\colon \mathbb{R} \to \mathbb{R}^2$, $\mathbf{f}(t) = (\cos t, \sin t)$ (parametrisation of a circle, set $\mathbf{f}([0, 2\pi))$ represents a circle)
8. $\mathbf{f}\colon \mathbb{R} \to \mathbb{R}^3$, $\mathbf{f}(t) = (\cos t, \sin t, at)$ (parametrisation of a helix)
9. $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{f}(\mathbf{x}) = \mathbf{x}$ (identity mapping)
10. $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ (linear mapping)
11. $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ (affine mapping)
12. $\mathbf{f}\colon \mathbb{R}^2 \to \mathbb{R}^3$, $\mathbf{f}(u, v) = (\,(R + r\cos v)\cos u,\ (R + r\cos v)\sin u,\ r\sin v\,)$ (parametrisation of a torus or annuloid, set $\mathbf{f}([0, 2\pi) \times [0, 2\pi))$ represents a torus)
13. The *image morphing* technique deforms an image (e.g. of a face) to another image (face). Morphing is represented by the mapping $\mathbb{R}^2 \to \mathbb{R}^2$.
14. An electric field associates with every point in $\mathbb{R}^3$ a vector in $\mathbb{R}^3$. □

8.1 Continuity

Definition 8.1. *Mapping* $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ *is* **continuous at point** $\mathbf{x} \in \mathbb{R}^n$, *iff*

$$\forall \varepsilon > 0 \ \ \exists \delta > 0 \ \ \forall \mathbf{y} \in \mathbb{R}^n : \quad \|\mathbf{x} - \mathbf{y}\| < \delta \implies \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| < \varepsilon.$$

A mapping is **continuous over set** $X \subseteq \mathbb{R}^n$ *iff it is continuous at every point* $\mathbf{x} \in X$.

Informally speaking, a mapping is continuous if it associates a pair of near points with a pair of near points. However, definition 8.1 is not convenient for checking continuity. We give a sufficient (but not necessary) condition that is more practical. We assume that the reader can verify the continuity of functions of one variable. We leave out the proof.

Theorem 8.1.

(a) *Let function $f\colon \mathbb{R} \to \mathbb{R}$ be continuous at point x. Let $k \in \{1, \dots, n\}$ and let function $g\colon \mathbb{R}^n \to \mathbb{R}$ be given by $g(x_1, \dots, x_n) = f(x_k)$ (i.e. g depends solely on variable x_k). Then function g is continuous at every point $(x_1, \dots, x_n)$ where $x_k = x$.*

(b) *Let functions $f, g\colon \mathbb{R}^n \to \mathbb{R}$ be continuous at point $\mathbf{x}$. Then the functions $f + g$, $f - g$ and fg are continuous at point $\mathbf{x}$. When $g(\mathbf{x}) \neq 0$, then the function f/g is also continuous at point $\mathbf{x}$.*

(c) *Let $g\colon \mathbb{R}^n \to \mathbb{R}$ be continuous at point $\mathbf{x}$ and $f\colon \mathbb{R} \to \mathbb{R}$ be continuous at point $y = g(\mathbf{x})$. Then the composite function $f \circ g\colon \mathbb{R}^n \to \mathbb{R}$ is continuous at point $\mathbf{x}$.*

(d) *Let functions $f_1, \dots, f_m\colon \mathbb{R}^n \to \mathbb{R}$ be continuous at point $\mathbf{x}$. Then the mapping $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ defined by $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))$ is continuous at point $\mathbf{x}$.*

Example 8.2. Using the above theorem we can easily show that, for example, the (frightfully looking) function

$$f(x, y) = \sqrt{\sin(x^3 y - y^4) + |x^2 + y^3 \mathrm{e}^x|}$$

is continuous. E.g. by (a), x^3 is a continuous function of two variables (x, y). Similarly, y is a continuous function of variables (x, y). Then, by (b), the function $x^3 y$ is continuous. The continuity of the whole function can be proved in this 'recursive' manner. □

8.2 Partial differentiation

The partial derivative of function $f\colon \mathbb{R}^n \to \mathbb{R}$ with respect to x_i is denoted in the following ways:

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = f_{x_i}(\mathbf{x}) = \frac{\partial y}{\partial x_i},$$

where the last notation assumes that $y = f(\mathbf{x})$. The partial derivative is evaluated by treating all the variables x_j, $j \neq i$ as constants and differentiating the function with respect to the single variable x_i.

Example 8.3. Consider the function $f(x, y) = x^2 y + \sin(x - y^3)$. Its partial derivatives are

$$\begin{aligned}
\frac{\partial f(x, y)}{\partial x} &= f_x(x, y) = 2xy + \cos(x - y^3), \\
\frac{\partial f(x, y)}{\partial y} &= f_y(x, y) = x^2 - 3y^2 \cos(x - y^3).
\end{aligned}$$

□

8.3 The total derivative

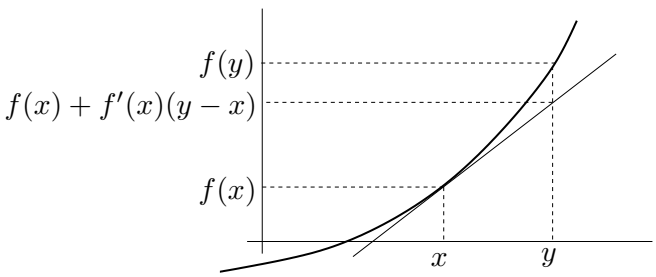
Let us review the definition of the derivative of function $f\colon \mathbb{R} \to \mathbb{R}$ of a single variable at point x. When the limit

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = f'(x) = \lim_{y \to x} \frac{f(y) - f(x)}{y - x}, \tag{8.1}$$

exists, then function f is *differentiable* at point x and the value of the limit is its *derivative* at point x. Differentiability means that the function can be 'well approximated' in the nighbourhood of point x by the affine function

$$f(y) \approx f(x) + f'(x)(y - x). \tag{8.2}$$

As shown in this figure:



How to generalise the concepts of differentiability and derivative to the mapping $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$? It appears that it is not easy to do so by a generalisation of the limit concept (8.1). It is better to use formula (8.2). Let us approximate the mapping in the neighbourhood of point $\mathbf{x}$ by:

$$\mathbf{f}(\mathbf{y}) \approx \mathbf{f}(\mathbf{x}) + \mathbf{f}'(\mathbf{x})(\mathbf{y} - \mathbf{x}). \tag{8.3}$$

When $\mathbf{x}$ is fixed, then the right hand side of (8.3) is an affine mapping in the variable $\mathbf{y}$. Since $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^m$, then $\mathbf{f}'(\mathbf{x})$ must be a matrix of size $m \times n$. A mapping is **differentiable** at point $\mathbf{x}$ if it is 'similar' to an affine mapping in the neighbourhood of $\mathbf{x}$. E.g. there exists matrix $\mathbf{f}'(\mathbf{x})$ such that the approximation error $\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})(\mathbf{y} - \mathbf{x})$ is 'small' for a 'small' $\mathbf{y} - \mathbf{x}$. In order to express this condition precisely we would need to use the limit of a function of several variables, the knowledge of which we do not expect of the reader. We therefore leave the concept of 'differentiable mapping' undefined and instead define a somewhat stronger property which is in practice sufficient:

Definition 8.2. *mapping* $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ *at point* $\mathbf{x}$ *is* **continuously differentiable**, *iff at point* $\mathbf{x}$ *all the partial derivatives* $\partial f_i(\mathbf{x})/\partial x_j$ *exist and are continuous.*

It is possible to prove that when a mapping is at some point continuously differentiable, then it is at that point also differentiable.

Example 8.4. Consider the function of Exercise 8.3; both its partial derivatives are continuous functions over the entire $\mathbb{R}^2$, therefore the function is differentiable at each point $(x, y) \in \mathbb{R}^2$.□

Note that the mere existence of all the partial derivatives is not sufficient for differentiability.

Example 8.5. Let the function $f\colon \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$f(x,y) = \begin{cases} 1 & \text{když } xy = 0, \\ 0 & \text{když } x \neq 0 \text{ a } y \neq 0. \end{cases}$$

At point $(0,0)$ both partial derivatives exist (both are equal to zero) but the function $\partial x/\partial f(x,y)$ is not continuous function of (x,y) at $(0,0)$. It is possible to show that f is not differentiable at point $(0,0)$. This is not surprising as the function is not at all like an affine function in the neighbourhood of this point. □

When mapping $\mathbf{f}$ is differentiable at point $\mathbf{x}$, then in this case the matrix $\mathbf{f}'(\mathbf{x})$ has a natural shape: its elements are the partial derivatives of all the mapping elements with respect to all the variables:

$$\frac{\mathrm{d}\mathbf{f}(\mathbf{x})}{\mathrm{d}\mathbf{x}} = \mathbf{f}'(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix}. \tag{8.4}$$

matrix (8.4) is called **the total derivative**[1] (or shortly just **the derivative**) of the mapping $\mathbf{f}$ at point $\mathbf{x}$. For historical reasons it is also called the **Jacobi's matrix**. Special cases:

- For $f\colon \mathbb{R} \to \mathbb{R}$, $f'(x)$ is a *scalar* and is the same as ordinary derivative (8.1).
- For $\mathbf{f}\colon \mathbb{R} \to \mathbb{R}^m$, $\mathbf{f}'(x)$ is a *column vector*.
- For $f\colon \mathbb{R}^n \to \mathbb{R}$, $f'(\mathbf{x})$ is a *row vector*.

8.3.1 Derivative of mapping composition

The 'chain rule' for differentiation of function compositions can be naturally extended to mappings. The proof of the following theorem is long and so we will not give it here.

Theorem 8.2. *Let* $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^m$ *and* $\mathbf{f}\colon \mathbb{R}^m \to \mathbb{R}^\ell$ *be differentiable mappings. The derivative of the mappings composition* $\mathbf{f} \circ \mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^\ell$ *is*

$$(\mathbf{f} \circ \mathbf{g})'(\mathbf{x}) = \frac{\mathrm{d}\mathbf{f}(\mathbf{g}(\mathbf{x}))}{\mathrm{d}\mathbf{x}} = \mathbf{f}'(\mathbf{g}(\mathbf{x}))\,\mathbf{g}'(\mathbf{x}). \tag{8.5}$$

The dimensions of the relevant spaces can be succinctly expressed by the following diagram:

$$\mathbb{R}^n \xrightarrow{\mathbf{g}} \mathbb{R}^m \xrightarrow{\mathbf{f}} \mathbb{R}^\ell. \tag{8.6}$$

If we put $\mathbf{u} = \mathbf{g}(\mathbf{x})$ and $\mathbf{y} = \mathbf{f}(\mathbf{u})$, the rule can also be written in the Leibnitz notation as:

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\mathbf{x}} = \frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\mathbf{u}}\,\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}\mathbf{x}}, \tag{8.7}$$

which is easy to remember, as $\mathrm{d}\mathbf{u}$ can be 'as if eliminated' (however, this is not a proof!). Let us emphasise that this equality is *matrix multiplication*. The left hand side expression is matrix $\ell \times n$, the first expression on the right hand side is matrix $\ell \times m$ and the second one is matrix

[1] The term 'differential' is sometimes used instead of 'the total derivative'. These terms are similar but not identical: the total derivative is a *matrix*, whereas the total differential is a *linear mapping* represented by the matrix. This difference is exactly the same as saying, in linear algebra, just 'matrix' instead of 'linear mapping'.

$m \times n$. When $\ell = m = n = 1$ we get the well known chain rule for differentiating compositions of functions of a single variable. The rule can be evidently extended to the compositions of more than two mappings: *The Jacobi's matrix of the composed mapping is the product of the Jacobi's matrices of the individual mappings.*

Example 8.6. Let $f(u, v)$ be a differentiable function of two variables. Determine the (total) derivative of the function $z = f(x + y, xy)$ with respect to (w.r.t.) the vector (x, y), i.e. its partial derivatives w.r.t. x and y.

Given the diagram $\mathbb{R}^2 \xrightarrow{\mathbf{g}} \mathbb{R}^2 \xrightarrow{f} \mathbb{R}$, where mapping $\mathbf{g}$ is given by

$$\mathbf{g}(x, y) = (u, v) = (x + y, xy).$$

The derivative of mapping f w.r.t. the vector (u, v) is the 1×2 matrix (row vector):

$$f'(u, v) = \begin{bmatrix} \frac{\partial f(u,v)}{\partial u} & \frac{\partial f(u,v)}{\partial v} \end{bmatrix} = \begin{bmatrix} f_u(u, v) & f_v(u, v) \end{bmatrix}.$$

The derivative of mapping $\mathbf{g}$ w.r.t. the vector (x, y) is the 2×2 matrix:

$$\mathbf{g}'(x, y) = \frac{\mathrm{d}\mathbf{g}(x, y)}{\mathrm{d}(x, y)} = \begin{bmatrix} \frac{\partial(x+y)}{\partial x} & \frac{\partial(x+y)}{\partial y} \\ \frac{\partial(xy)}{\partial x} & \frac{\partial(xy)}{\partial y} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ y & x \end{bmatrix}.$$

The derivative of the mapping $f \circ \mathbf{g}$: $\mathbb{R}^2 \to \mathbb{R}$ w.r.t. the vector (x, y) is the 1×2 matrix (row vector)

$$\begin{aligned} \frac{\mathrm{d}z}{\mathrm{d}(x, y)} = \frac{\mathrm{d}f(\mathbf{g}(x, y))}{\mathrm{d}(x, y)} &= f'(u, v)\mathbf{g}'(x, y) \\ &= \begin{bmatrix} f_u(u, v) & f_v(u, v) \end{bmatrix} \begin{bmatrix} 1 & 1 \\ y & x \end{bmatrix} \\ &= \begin{bmatrix} f_u(u, v) + y f_v(u, v) & f_u(u, v) + x f_v(u, v) \end{bmatrix}. \end{aligned}$$

□

Example 8.7. Show two methods how to determine the partial derivative f_x of the function $f(x, y) = \mathrm{e}^{(x+y)^2+(xy)^2}$:

- Treat y as a constant and differentiate f as a function of single variable x:

$$f_x = [2(x + y) + 2(xy)y]\mathrm{e}^{(x+y)^2+(xy)^2} = 2(x + y + xy^2)\mathrm{e}^{(x+y)^2+(xy)^2}.$$

- Put $u = x + y$, $v = xy$, $f(u, v) = \mathrm{e}^{u^2+v^2}$. From Example 8.6, we have $f_x = f_u + y f_v$. Since

$$f_u = 2u\mathrm{e}^{u^2+v^2}, \quad f_v = 2v\mathrm{e}^{u^2+v^2},$$

  we have $f_x = f_u + y f_v = 2u\mathrm{e}^{u^2+v^2} + y(2v)\mathrm{e}^{u^2+v^2} = 2(x + y + xy^2)\mathrm{e}^{(x+y)^2+(xy)^2}$. □

Example 8.8. Given differentiable function f: $\mathbb{R}^2 \to \mathbb{R}$, determine the derivative of the function $z = f(t + t^2, \sin t)$ w.r.t. t.

Consider the diagram $\mathbb{R} \xrightarrow{\mathbf{g}} \mathbb{R}^2 \xrightarrow{f} \mathbb{R}$, where $\mathbf{g}(t) = (u, v) = (t + t^2, \sin t)$. Then

$$\frac{\mathrm{d}z}{\mathrm{d}t} = f'(u, v)\mathbf{g}'(t) = \begin{bmatrix} f_u(u, v) & f_v(u, v) \end{bmatrix} \begin{bmatrix} 1 + 2t \\ \cos t \end{bmatrix} = f_u(u, v)(1 + 2t) + f_v(u, v)\cos t.$$

□

8.3.2 Differentiation of expressions with matrices

When a function or a mapping are given by an expression containing vectors and matrices, then the derivatives can always be computed by 'brute force', i.e., by expanding the expression into its individual elemements and computing the partial derivatives of each element w.r.t. each variable. Strictly speaking this solves the problem. Nonetheless, it is advantageous to simplify the result and turn it into a matrix expression.

Example 8.9. Let us find the (total) derivative of the qudratic form $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is an arbitrary (not necessarily symmetric) matrix of size $n \times n$. Writing out function f in detail:

$$\begin{aligned}\mathbf{x}^T\mathbf{A}\mathbf{x} = a_{11}x_1^2 &+ a_{21}x_2x_1 + \cdots + a_{n1}x_nx_1 + \\ a_{12}x_1x_2 &+ a_{22}x_2^2 + \cdots + a_{n2}x_nx_1 + \\ &\vdots \\ a_{1n}x_1x_n &+ a_{2n}x_2x_n + \cdots + a_{nn}x_n^2.\end{aligned}$$

With a bit of effort we can see from this expression that

$$\frac{\partial f(\mathbf{x})}{\partial x_1} = 2a_{11}x_1 + (a_{21} + a_{12})x_2 + \cdots + (a_{n1} + a_{1n})x_n$$

and similarly for the derivatives w.r.t. the remaining variables. Note that these partial derivatives can be arranged in a row vector

$$f'(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} = \mathbf{x}^T(\mathbf{A} + \mathbf{A}^T).$$

□

The following table lists other often seen derivatives. Derive them all as an exercise! The chain rule is often useful for this.

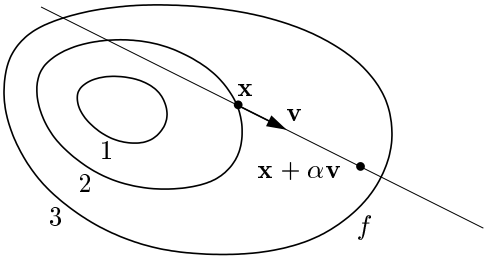
$\mathbf{f}(\mathbf{x})$	$\mathbf{f}'(\mathbf{x})$	notes
$\mathbf{x}$	$\mathbf{I}$	$\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$
$\mathbf{A}\mathbf{x}$	$\mathbf{A}$	$\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$
$\mathbf{x}^T\mathbf{x}$	$2\mathbf{x}^T$	$\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}$
$\mathbf{x}^T\mathbf{A}\mathbf{x}$	$\mathbf{x}^T(\mathbf{A} + \mathbf{A}^T)$	$\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}$
$\mathbf{x}^T\mathbf{a} = \mathbf{a}^T\mathbf{x}$	$\mathbf{a}^T$	$\mathbf{a} \in \mathbb{R}^n$, $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}$
$\lVert\mathbf{x}\rVert_2$	$\mathbf{x}^T/\lVert\mathbf{x}\rVert_2$	$\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}$
$\mathbf{g}(\mathbf{A}\mathbf{x})$	$\mathbf{g}'(\mathbf{A}\mathbf{x})\mathbf{A}$	$\mathbf{A} \in \mathbb{R}^{\ell\times n}$, $\mathbf{g}\colon \mathbb{R}^\ell \to \mathbb{R}^m$, $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$
$\mathbf{g}(\mathbf{x})^T\mathbf{g}(\mathbf{x})$	$2\mathbf{g}(\mathbf{x})^T\mathbf{g}'(\mathbf{x})$	$\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}$
$\mathbf{g}(\mathbf{x})^T\mathbf{h}(\mathbf{x})$	$\mathbf{g}(\mathbf{x})^T\mathbf{h}'(\mathbf{x}) + \mathbf{h}(\mathbf{x})^T\mathbf{g}'(\mathbf{x})$	$\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{h}\colon \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}$

8.4 Directional derivative

The **cut** of function $f\colon \mathbb{R}^n \to \mathbb{R}$ at point $\mathbf{x} \in \mathbb{R}^n$ in direction $\mathbf{v} \in \mathbb{R}^n$ is the function $\varphi\colon \mathbb{R} \to \mathbb{R}$:

$$\varphi(\alpha) = f(\mathbf{x} + \alpha\mathbf{v}). \tag{8.8}$$

The following figure illustrates a cut for the case $n = 2$:



The **directional derivative** of function f at point $\mathbf{x}$ in direction $\mathbf{v}$ is the scalar

$$\varphi'(0) = \left.\frac{\mathrm{d}\varphi(\alpha)}{\mathrm{d}\alpha}\right|_{\alpha=0} = \lim_{\alpha\to 0}\frac{f(\mathbf{x}+\alpha\mathbf{v}) - f(\mathbf{x})}{\alpha}. \tag{8.9}$$

The directional derivative in the direction of the ith standard basis vector $(0,\ldots,0,1,0,\ldots,0)$ (1 in the ith position) is just the partial derivative w.r.t. the variable x_i.

The directional derivative of a mapping is obtained by computing the directional derivatives of each component. I.e. the directional derivative of mapping $\mathbf{f} = (f_1,\ldots,f_m)\colon \mathbb{R}^n \to \mathbb{R}^m$ at point $\mathbf{x} \in \mathbb{R}^n$ in direction $\mathbf{v} \in \mathbb{R}^n$ is the vector $(\varphi_1'(0),\ldots,\varphi_m'(0)) \in \mathbb{R}^m$, where $\varphi_i(\alpha) = f_i(\mathbf{x}+\alpha\mathbf{v})$.

Theorem 8.3. *Let mapping* $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ *be differentiable at point* $\mathbf{x}$. *Then its directional derivative at point* $\mathbf{x}$ *in direction* $\mathbf{v}$ *is* $\mathbf{f}'(\mathbf{x})\mathbf{v}$.

Proof. Mapping $\mathbf{y} = \boldsymbol{\varphi}(\alpha) = \mathbf{f}(\mathbf{x}+\alpha\mathbf{v})$ is a composition of two mappings $\mathbf{y} = \mathbf{f}(\mathbf{u})$ and $\mathbf{u} = \mathbf{x}+\alpha\mathbf{v}$. We have diagram $\mathbb{R} \xrightarrow{\mathbf{u}=\mathbf{x}+\alpha\mathbf{v}} \mathbb{R}^n \xrightarrow{\mathbf{y}=\mathbf{f}(\mathbf{x})} \mathbb{R}^m$ and $\mathrm{d}\mathbf{u}/\mathrm{d}\alpha = \mathbf{v}$. By the chain rule

$$\boldsymbol{\varphi}'(\alpha) = \frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\alpha} = \frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\mathbf{u}}\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}\alpha} = \frac{\mathrm{d}\mathbf{f}(\mathbf{u})}{\mathrm{d}\mathbf{u}}\mathbf{v}.$$

Putting $\alpha = 0$ gives $\mathbf{u} = \mathbf{x}$, which proves the theorem. □

8.5 Gradient

The transpose of the total derivative of function $f\colon \mathbb{R}^n \to \mathbb{R}$ is called the **gradient** and is written as

$$f'(\mathbf{x})^T = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} = \nabla f(\mathbf{x})$$

(∇ is read as 'nabla'). Whereas $f'(\mathbf{x})$ is a row vector, the gradient is a column vector[2].

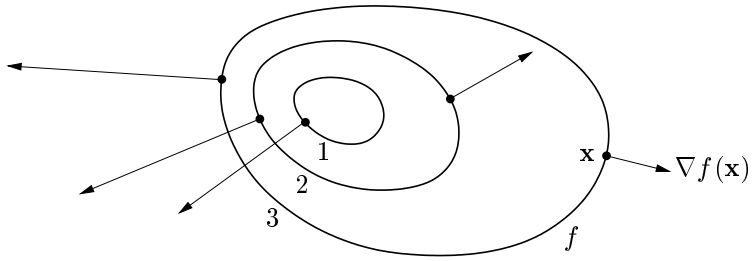
Consider the directional derivatives at a fixed point $\mathbf{x}$ in various directions given by a normalised vector $\mathbf{v}$ (i.e. $\|\mathbf{v}\|_2 = 1$). Such derivative is $f'(\mathbf{x})\mathbf{v}$, that is the scalar product of the gradient at point $\mathbf{x}$ and the vector $\mathbf{v}$. It is clear (but think about it), that:

[2] Introducing a new term for the transpose of the derivative seems superfluous – nonetheless the justification is that the total derivative is a *linear function*, whereas the gradient is a *vector*. Unfortunately, the literature is inconsistent in drawing a distinction between the gradient and the (total) derivative function. Sometimes they are treated as identical, both denoted as $\nabla f(\mathbf{x})$. However, this leads to an inconsistency with the notation used in linear algebra, as the derivative of function $\mathbb{R}^n \to \mathbb{R}$ is then no longer a special case of the derivative of mapping $\mathbb{R}^n \to \mathbb{R}^m$ (i.e. Jacobi's matrix), which is a row vector when $m = 1$.

- The directional derivative is maximal in the direction $\mathbf{v} = \nabla f(\mathbf{x})/\|\nabla f(\mathbf{x})\|_2$, i.e. when $\mathbf{v}$ is parallel with the gradient and has the same orientation. That means the gradient direction is the *direction of the steepest ascent* of a function.
- The gradient magnitude $\|\nabla f(\mathbf{x})\|_2$ expresses the steepness of the slope of a function in the direction of the steepest ascent.
- The directional derivative in the direction perpendicular to the gradient is zero.

Further, it can be shown (see §10.2.1) that the gradient is always *perpendicular to the contour*.

The following figure shows three contours of function $f\colon \mathbb{R}^2 \to \mathbb{R}$ and its gradients at several points:



8.6 Second order partial derivatives

Differentiating function $f\colon \mathbb{R}^n \to \mathbb{R}$ first w.r.t. x_i and then w.r.t. x_j produces the partial derivative of the second order, denoted

$$\frac{\partial}{\partial x_j}\frac{\partial f(\mathbf{x})}{\partial x_i} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i\,\partial x_j}.$$

When $i = j$, we write in the condensed form

$$\frac{\partial}{\partial x_i}\frac{\partial f(\mathbf{x})}{\partial x_i} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i^2}.$$

It can be proved that when the mixed partial derivatives

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_i\,\partial x_j}, \qquad \frac{\partial^2 f(\mathbf{x})}{\partial x_j\,\partial x_i}$$

are continuous at point $\mathbf{x}$, then they are equal, i.e. the order of the differentiation w.r.t. the individual variables can be changed.

Example 8.10. Determine all the second derivatives of the function $f(x, y) = x^2y + \sin(x - y^3)$ from Example 8.3. The first derivatives are already given in that example. Now follow the second derivatives:

$$\begin{aligned}
\frac{\partial^2 f(x,y)}{\partial x^2} &= \frac{\partial}{\partial x}[2xy + \cos(x - y^3)] &&= 2y - \sin(x - y^3)\\
\frac{\partial^2 f(x,y)}{\partial x\,\partial y} &= \frac{\partial}{\partial y}[2xy + \cos(x - y^3)] &&= 2x + 3y^2\sin(x - y^3)\\
\frac{\partial^2 f(x,y)}{\partial y\,\partial x} &= \frac{\partial}{\partial x}[x^2 - 3y^2\cos(x - y^3)] &&= 2x + 3y^2\sin(x - y^3)\\
\frac{\partial^2 f(x,y)}{\partial y^2} &= \frac{\partial}{\partial y}[x^2 - 3y^2\cos(x - y^3)] &&= -6y\cos(x - y^3) - 9y^4\sin(x - y^3).
\end{aligned}$$

Note that the order of differentiation w.r.t. x and y is indeed immaterial. □

We write the matrix of all the second partial derivatives of function $f\colon \mathbb{R}^n \to \mathbb{R}$ as follows

$$f''(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_n} \end{bmatrix}.$$

It is a symmetric matrix of dimensions $n \times n$, often called the **Hess matrix**.

What might be the second derivative of mapping $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$? It will no longer be just a two dimensional table (i.e. matrix) of dimensions $n \times n$ but rather a three dimensional table of dimensions $m \times n \times n$.

8.7 Taylor's polynomial

Let function of one variable $f\colon \mathbb{R} \to \mathbb{R}$ have at point x derivatives up to order k. Its **Taylor's polynomial** of degree k at point x is the polynomial $T_k\colon \mathbb{R} \to \mathbb{R}$ of degree k such that at point x all its derivatives up to order k are the same as those of function f. In this sense polynomial T_k is an approximation function of f in the neighbourhood of point x.

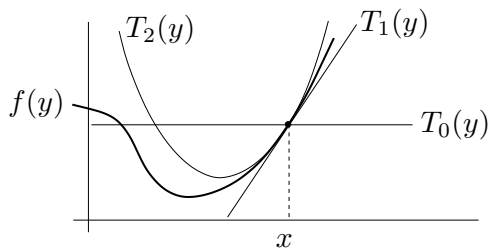
Taylor's polynomial is uniquely defined by this requirement and its form is (derive!):

$$T_k(y) = \sum_{i=0}^{k} \frac{1}{i!} f^{(i)}(x)\,(y-x)^i, \tag{8.10}$$

where $f^{(i)}$ denotes i-th derivative of function f (here zero-th derivative is the function itself $f^{(0)} = f$) and where we put $0! = 1$. The form of the polynomial up to degree two:

$$\begin{aligned} T_0(y) &= f(x), \\ T_1(y) &= f(x) + f'(x)\,(y-x), \\ T_2(y) &= f(x) + f'(x)\,(y-x) + \tfrac{1}{2} f''(x)\,(y-x)^2. \end{aligned}$$

Taylor's polynomial of zero-th degree T_0 is a very poor approximation, equal simply to constant function. We already know the polynomial of the first degree $T_1(x)$ from the formula (8.2). The polynomial of the second degree T_2 is a parabola which has the same value and the same first two derivatives with function f at point x. See the following figure:



Can we generalise Taylor's polynomial to functions of several variables $f\colon \mathbb{R}^n \to \mathbb{R}$? Taylor's polynomial of kth degree (function f in the neighbourhood of point $\mathbf{x}$) is the polynomial $T_k\colon \mathbb{R}^n \to \mathbb{R}$ of degree k, which has at point $\mathbf{x}$ all its partial derivative up to order k the same

as function f. We list now the polynomials only up to degree two, without giving the general formula for any degree:

$$T_0(\mathbf{y}) = f(\mathbf{x}), \tag{8.11a}$$
$$T_1(\mathbf{y}) = f(\mathbf{x}) + f'(\mathbf{x})\,(\mathbf{y}-\mathbf{x}), \tag{8.11b}$$
$$T_2(\mathbf{y}) = f(\mathbf{x}) + f'(\mathbf{x})\,(\mathbf{y}-\mathbf{x}) + \tfrac{1}{2}(\mathbf{y}-\mathbf{x})^T f''(\mathbf{x})\,(\mathbf{y}-\mathbf{x}). \tag{8.11c}$$

Here $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $f'(\mathbf{x}) \in \mathbb{R}^{1\times n}$ is Jacobi matrix (row vector) and $f''(\mathbf{x}) \in \mathbb{R}^{n\times n}$ is Hess matrix. Function (8.11b) is affine and function (8.11c) is quadratic.

Taylor's polynomial can be generalised to the mapping $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ in such a way that we take Taylor's polynomials of all components $f_1, \ldots, f_m$. The polynomial of the first degree thus results in the mapping

$$\mathbf{T}_1(\mathbf{y}) = \mathbf{f}(\mathbf{x}) + \mathbf{f}'(\mathbf{x})\,(\mathbf{y}-\mathbf{x}), \tag{8.12}$$

which is the same as (8.3). The polynomial of the second degree leads to the mapping $\mathbf{T}_2$, whose components are functions (8.11c). This can not be written in a matrix form, as all $m \times n \times n$ second partial derivatives do not 'fit' into a matrix.

Example 8.11. Find Taylor's polynomial of second degree at point $(x_0, y_0) = (2, 1)$ of function $f(x, y) = \sin(x + y^2)$ of Example 8.3.

We have

$$\begin{aligned}
f(x_0, y_0) &= \sin 3,\\
f'(x_0, y_0) &= \cos(x+y^2)\begin{bmatrix}1 & 2y\end{bmatrix}\Big|_{(x,y)=(2,1)} = (\cos 3)\begin{bmatrix}1 & 2\end{bmatrix},\\
f''(x_0, y_0) &= -\sin(x+y^2)\begin{bmatrix}1 & 2y\\ 2y & 4y^2-2\end{bmatrix}\Bigg|_{(x,y)=(2,1)} = -(\sin 3)\begin{bmatrix}1 & 2\\ 2 & 2\end{bmatrix}.
\end{aligned}$$

Thus by (8.11c) we have (watch out, our variables are denoted differently than in (8.11c))

$$\begin{aligned}
T_2(x, y) &= f(x_0, y_0) + f'(x_0, y_0)\begin{bmatrix}x-x_0\\ y-y_0\end{bmatrix} + \frac{1}{2}\begin{bmatrix}x-x_0\\ y-y_0\end{bmatrix}^T f''(x_0, y_0)\begin{bmatrix}x-x_0\\ y-y_0\end{bmatrix}\\
&= \sin 3 + (\cos 3)\begin{bmatrix}1 & 2\end{bmatrix}\begin{bmatrix}x-2\\ y-1\end{bmatrix} - \frac{\sin 3}{2}\begin{bmatrix}x-2\\ y-1\end{bmatrix}^T\begin{bmatrix}1 & 2\\ 2 & 2\end{bmatrix}\begin{bmatrix}x-2\\ y-1\end{bmatrix}\\
&= \sin 3 + (\cos 3)(x-2+2(y-1)) - \frac{\sin 3}{2}((x-2)^2 + 4(x-2)(y-1) + 2(y-1)^2)\\
&= (\cos 3)(x+2y-4) + (\sin 3)(-x^2/2 - y^2 - 2xy + 3x + 6y + 6). \qquad \square
\end{aligned}$$

8.8 Exercises

8.1. Given sets $X = [-1, 1] \times \{0\} = \{\,(x, 0) \mid -1 \le x \le 1\,\} \subseteq \mathbb{R}^2$ and $Y = [-1, 1] \times [-1, 1]$. Sketch the following sets:

a) $\left\{\,\mathbf{y} \in \mathbb{R}^2 \;\middle|\; \min\limits_{\mathbf{x}\in X} \|\mathbf{x}-\mathbf{y}\|_2 \le 1\,\right\}$

b) $\left\{\,\mathbf{y} \in \mathbb{R}^2 \;\middle|\; \max\limits_{\mathbf{x}\in X} \|\mathbf{x}-\mathbf{y}\|_2 \le 2\,\right\}$

c) contours of height 1 in function $f(\mathbf{x}) = \min\limits_{\mathbf{x}\in Y} \|\mathbf{x}-\mathbf{y}\|_2$

d) contours of height $\sqrt{2}$ in function $f(\mathbf{x}) = \max_{\mathbf{x} \in Y} \|\mathbf{x} - \mathbf{y}\|_2$

8.2. Given function of two variables $f(x, y)$.

a) Calculate the derivative of f in polar coordinates (φ, r), where $x = r\cos\varphi$, $y = r\sin\varphi$.

b) Point (x, y) is moving in time t along the curve given by equation $(x, y) = (t^2 + 2t, \ln(t^2 + 1))$. Find the time derivative of f.

8.3. Calculate the derivative of function $g(\mathbf{u}) = f(\mathbf{a}^T\mathbf{u}, \|\mathbf{u}\|_2)$ with respect to the vector $\mathbf{u}$.

8.4. An altitude (height above the sea level) of some lanscape is given by the formula $h(d, s) = 2s^2 + 3sd - d^2 + 5$, where d is geographical longitute (increasing from the West to the East) and s is geographical latitude (increasing from the South to the North). At point $(s, d) = (1, -1)$, determine:

a) the direction of the steepest ascent of the terrain

b) the steepness of the terrain in the South-Easterly direction.

8.5. Find the second derivative $f''(x, y)$ (i.e. Hess matrix) of the functions

a) $f(x, y) = \mathrm{e}^{-x^2-y^2}$

b) $f(x, y) = \ln(\mathrm{e}^x + \mathrm{e}^y)$

8.6. Hess matrix of the quadratic form $f(\mathbf{x}) = \mathbf{x}^T\mathbf{a}\mathbf{x}$ is $f''(\mathbf{x}) = \mathbf{a} + \mathbf{a}^T$. Derive.

8.7. Given function $f(x, y) = 6xy^2 - 2x^3 - 3y^3$, find Taylor's polynomials of first and second degree at point $(x_0, y_0) = (1, -2)$.

8.8. *The finite difference method* calculates function derivatives approximately as

$$f'(x) \approx \frac{f(x + h) - f(x)}{h},$$

where h is a small number (a good choice is $h = \sqrt{\varepsilon}$, where ε is machine accuracy). This is applicable also to partial derivatives. Make up two mappings $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{f}\colon \mathbb{R}^m \to \mathbb{R}^\ell$ for some unequal dimensions $n, m, \ell > 1$. Choose a point $\mathbf{x} \in \mathbb{R}^n$. Calculate approximately the total derivative (Jacobi matrix) $\mathbf{g}'(\mathbf{x})$ and $\mathbf{f}'(\mathbf{g}(\mathbf{x}))$ in Matlab, using the method of finite differences. Then calculate the derivative of the composite mapping $(\mathbf{f} \circ \mathbf{g})'(\mathbf{x})$ by the method of the finite differences and also by multiplying matrices $\mathbf{g}'(\mathbf{x})$ and $\mathbf{f}'(\mathbf{g}(\mathbf{x}))$. Compare the results.

Chapter 9

Extrema of a Function over a Set

9.1 Minimum and infimum

We begin by defining the following concepts for a set $Y \subseteq \mathbb{R}$:

- **Lower bound** of set Y is any number $a \in \mathbb{R}$, such that $a \leq y$ for all $y \in Y$.
- **Infimum** of set Y is its greatest lower bound. It is denoted as $\inf Y$.
- **The smallest element** (or **minimum**) of set Y is its lower bound that is also a member of Y. When the minimum exists, it is unique. It is denoted as $\min Y$.

Upper bound, the greatest element (maximum, $\max Y$) and supremum ($\sup Y$) are defined correspondingly.

Minimum or maximum of a subset of real numbers need not exist. Nonetheless, it is a deep property of real numbers, that there exists an infimum [supremum] for every lower [upper] bound subset. This property is called **competeness**.

Let us introduce two elements $-\infty$ and $+\infty$ (which do not belong to the set $\mathbb{R}$) and let us define $-\infty < a < +\infty$ for each $a \in \mathbb{R}$. When set Y is not bound below [above], we define $\inf Y = -\infty$ [$\sup Y = +\infty$]. For an empty set, we define $\inf \emptyset = +\infty$ and $\sup \emptyset = -\infty$.

Example 9.1.

1. The set of all upper bounds of the real interval $[0, 1)$ is $[1, +\infty)$.
2. The set of all upper bounds of set $\mathbb{R}$ is $\emptyset$.
3. The set of all upper bounds of $\emptyset$ is $\mathbb{R}$.
4. Interval $[0, 1)$ does not have the maximum but it does have the supremum 1.
5. The minimum of set $\{\, 1/n \mid n \in \mathbb{N} \,\}$ does not exist but the infimum is 0.
6. The maximum of set $\{\, x \in \mathbb{Q} \mid x^2 \leq 2 \,\} = [-\sqrt{2}, \sqrt{2}] \cap \mathbb{Q}$ does not exist but the supremum is $\sqrt{2}$.
7. $\max\{1, 2, 3\} = \sup\{1, 2, 3\} = 3$ (the minimum and maximum of each finite set exist and are equal to infimum and supremum respectively).
8. Set $\mathbb{R}$ does not have the smallest element, i.e. $\min \mathbb{R}$ does not exist.
9. Empty set $\emptyset$ does not have the smallest element, i.e. $\min \emptyset$ does not exist. $\square$

9.2 Properties of subsets of $\mathbb{R}^n$

The set

$$U_\varepsilon(\mathbf{x}) = \{\, \mathbf{y} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{y}\| < \varepsilon \,\}, \tag{9.1}$$

where $\varepsilon > 0$ and $\mathbf{x} \in \mathbb{R}^n$, is called[1] **$\varepsilon$-neighbourhood** of point $\mathbf{x}$. It is a sphere (without a boundary) with center at $\mathbf{x}$ and non-zero radius ε.

Definition 9.1. *Consider set $X \subseteq \mathbb{R}^n$. Point $\mathbf{x} \in \mathbb{R}^n$ is called its*

- **interior point**, *when there exists $\varepsilon > 0$ such that $U_\varepsilon(\mathbf{x}) \subseteq X$*
- **boundary point**, *when for each $\varepsilon > 0$, $U_\varepsilon(\mathbf{x}) \cap X \neq \emptyset$ and $U_\varepsilon(\mathbf{x}) \cap (\mathbb{R}^n \setminus X) \neq \emptyset$*
- **exterior point**, *when there exists $\varepsilon > 0$, such that $U_\varepsilon(\mathbf{x}) \cap X = \emptyset$*
- **accumulation point**, *when for each $\varepsilon > 0$, $(U_\varepsilon(\mathbf{x}) \setminus \{\mathbf{x}\}) \cap X \neq \emptyset$*
- **isolated point**, *when there exists $\varepsilon > 0$, such that $U_\varepsilon(\mathbf{x}) \cap X = \{\mathbf{x}\}$.*

Note that the boundary and the accumulation points of a set need not belong to that set. When a point does belong to a set, it is either an interior or a boundary point but not both (prove!). **Interior** [**boundary**] of a set is the set of all its interior [boundary] points.

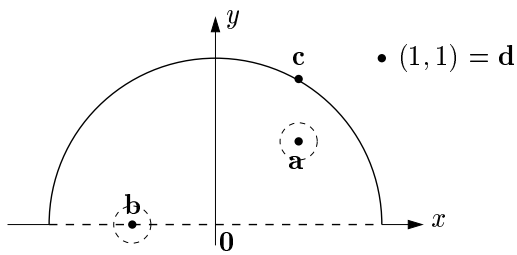
A set is called

- **open**, iff all its points are interior points;
- **closed**, iff it contains all its boundary points.

It can be proved that set X is closed [open], iff its complement $\mathbb{R}^n \setminus X$ is open [closed]. Openness and closeness are not mutually exclusive: sets $\emptyset$ and $\mathbb{R}^n$ are open and closed. Furthermore, some sets are neither open nor closed, e.g., the interval $(0, 1]$.

Set X is **bounded**, iff there is $r \in \mathbb{R}$ such that $\|\mathbf{x} - \mathbf{y}\|_2 < r$ for all $\mathbf{x}, \mathbf{y} \in X$. In other words, the set 'fits' into a sphere of finite radius.

Example 9.2. Consider set $\{\, (x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1,\ y > 0 \,\} \cup \{(1, 1)\} \subseteq \mathbb{R}^2$ shown in this figure:



Point $\mathbf{a}$ is an interior point of the set because there exists $\varepsilon > 0$ such that neighbourhood $U_\varepsilon(\mathbf{a})$ is wholly contained in the set.
Point $\mathbf{b}$ is a boundary point because a neighbourhood $U_\varepsilon(\mathbf{b})$ has for each $\varepsilon > 0$ a non-empty intersection with the set and also with its complement. Note that $\mathbf{b}$ does not belong to the set.
Point $\mathbf{a}$ is not a boundary point and point $\mathbf{b}$ is not an interior point.
Point $\mathbf{c}$ is not interior, is a boundary point, and belongs to the set.
Points $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are accumulation points, point $\mathbf{d}$ is isolated. Point $(1, 1)$ is exterior and therefore

[1] The norm in (9.1) can be euclidian or any other vector p-norm (see §12.3.1). The interior and the boundary of the set are independent of the choice of the norm.

does not belong to the set.

This set is not open because, for example, point $\mathbf{c}$ is not interior. Nor is it closed because, for example, point $\mathbf{b}$ is a boundary point but does not belong to the set. The set is bounded.□

Example 9.3. Point $1/2$ is an interior point of the interval $(0,1] \subseteq \mathbb{R}$ and points 0 and 1 are boundary points. □

Example 9.4. Set $[0,1] \times \{1\} = \{\,(x,y) \mid 0 \le x \le 1,\ y = 1\,\} \subseteq \mathbb{R}^2$ (line segment in a plane) has no interior points. All its points are boundary and accumulation. It is thus its own boundary. It is not open, it is closed, and is bounded. □

9.3 Existence of extrema

Let us recall (see §1.2) the concept of extrema of a function over a set. Function $f\colon \mathbb{R}^n \to \mathbb{R}$ over set $X \subseteq \mathbb{R}^n$ achieves its minimum at point $\mathbf{x} \in X$, iff for all $\mathbf{x}' \in X$, $f(\mathbf{x}) \le f(\mathbf{x}')$. The value of this minimum is $f(\mathbf{x})$. In other words, the value of the minimum is the smallest element of set

$$f(X) = \{\, f(\mathbf{x}) \mid \mathbf{x} \in X \,\} \subseteq \mathbb{R},$$

which is the image of set X under the mapping f.
The minimum need not always exist, as set $f(X)$ need not have the smallest element.

Example 9.5. function $f(x) = x$ over set $(0,1) \subseteq \mathbb{R}$ (open interval) has no minimum, as set $f((0,1)) = (0,1)$ does not have the smallest element. □

Example 9.6. function $f(x,y) = x + y$ over set $X = \{\,(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\,\}$ has no minimum, as set $f(X) = (-\sqrt{2}, \sqrt{2})$ does not have the smallest element. □

Theorem 9.2 gives sufficient condition for the existence of extrema of a function over a set. First though, we give without proof the following more general fact.

Theorem 9.1. *Continuous mapping image of a closed bounded set is a closed bounded set.*

Thus for closed and bounded set $X \subseteq \mathbb{R}^n$ and continuous mapping $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$, set $\mathbf{f}(X) = \{\,\mathbf{f}(\mathbf{x}) \mid \mathbf{x} \in X\,\} \subseteq \mathbb{R}^m$ is also closed and bounded[2].

It might appear that continuous mapping would preserve closeness even without compactness. However, it is easy to find counterexamples.

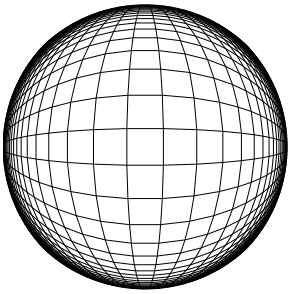
Example 9.7. let X be defined by interval $[1, +\infty) \subseteq \mathbb{R}$. This set is closed but not bounded. Continuous mapping $f(x) = 1/x$ produces image set $f(X) = (0,1]$, which is not closed and is bounded. □

Example 9.8. consider mapping $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^n$ given by

$$\mathbf{f}(\mathbf{x}) = (1 + \mathbf{x}^T\mathbf{x})^{-1/2}\,\mathbf{x}.$$

The image of unbounded set $\mathbb{R}^n$ under mapping $\mathbf{f}$ is open bounded set $\mathbf{f}(\mathbb{R}^n) = \{\,\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x}^T\mathbf{x} < 1\,\}$ (unit sphere without a boundary). For illustration we display in the figure set $\mathbf{f}(X) \subseteq \mathbb{R}^2$ for $X = (\mathbb{R} \times \mathbb{Z}) \cup (\mathbb{Z} \times \mathbb{R}) \subseteq \mathbb{R}^2$ (i.e. X is a planar grid. Think about it!):

[2] Sets which are both closed and bounded are called *compact*.



□

Theorem 9.1 has important consequences for optimisation. It is known as the *extreme value theorem* or *Weierstrass Theorem*.

Theorem 9.2. *Continuous function $f\colon \mathbb{R}^n \to \mathbb{R}$ over a closed bounded set $X \subseteq \mathbb{R}^n$ achieves its minimum.*

Proof. For function $f\colon \mathbb{R}^n \to \mathbb{R}$, the image of closed bounded set $X \subseteq \mathbb{R}^n$ is closed bounded set $f(X) \subseteq \mathbb{R}$. That can be none other than closed finite interval or a union of such intervals. Such set certainly has the smallest element. □

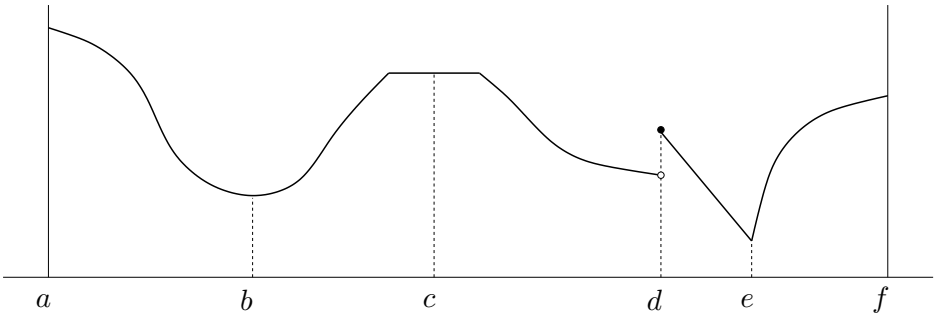
We emphasise that Theorem 9.2 gives only sufficient but not necessary conditions for existence of minimum of a function over a set. E.g. function $f(x) = x^2$ has a minimum over set $\mathbb{R}$, even though set $\mathbb{R}$ is unbounded.

9.4 Local extrema

Definition 9.2. *For function $f\colon \mathbb{R}^n \to \mathbb{R}$ and set $X \subseteq \mathbb{R}^n$. Point $\mathbf{x} \in X$ is called* **local minimum** *function f na množině X, iff there exists $\varepsilon > 0$ such that $\mathbf{x}$ is minimum of function f over the set $U_\varepsilon(\mathbf{x}) \cap X$.*

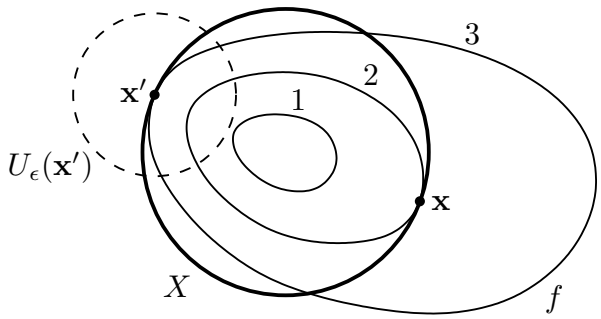
Local maximum is defined similarly. Every minimum of function f over set X is also a local minimum of function f over set X (but, in general, the converse is not true). In the context of discussing local extrema, we can sometimes use the term **global extrema** to emphasise the 'ordinary' extrema (in the sense of §1.2). When a reference to set X is missing, then the full domain of function f is meant.

Example 9.9. Function of one variable in the figure has in the closed interval $[a, f]$ at point a a local and also the global maximum, at point b a local minimum, at point c a local maximum and also a local minimum, at point d a local maximum, at point e a local and also the global minimum, at point f a local maximum.



□

Example 9.10. Let $X \subseteq \mathbb{R}^2$ je kružnice a funkce $f\colon \mathbb{R}^2 \to \mathbb{R}$ má vrstevnice jako na obrázku:



V bodě $\mathbf{x}$ nabývá funkce f na množině X globálního (a tedy i lokálního) minima, protože v žádném bodě na kružnici X nemá funkce menší hodnotu než $f(\mathbf{x}) = 2$. V bodě $\mathbf{x}'$ nabývá funkce f na množině X lokálního minima, protože existuje $\varepsilon > 0$ tak, že v bodě $\mathbf{x}'$ nabývá funkce f na části kružnice $U_\varepsilon(\mathbf{x}') \cap X$ svého (globálního) minima. □

Example 9.11. Function $f(\mathbf{x}) = x_1$ has over set $\{\, \mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 \leq 1 \,\}$ the global (and therefore also a local) minimum at point $(-1, 0, \ldots, 0)$. □

Example 9.12. Any function $f\colon \mathbb{R} \to \mathbb{R}$ has over $\mathbb{Z}$ (the set of integers) at an arbitrary point $x \in \mathbb{Z}$ a local minimum and also a local maximum. □

9.5 Exercises

9.1. Sketch the following subsets of $\mathbb{R}^2$:

a) $[-1, 0] \times \{1\}$
b) $\{\, (x, y) \mid x > 0,\ y > 0,\ xy = 1 \,\}$
c) $\{\, (x, y) \mid \min\{x, y\} = 1 \,\}$

9.2. Which of the following sets is the union of a finite number of (open, closed, half-closed) intervals? Find these intervals. Example: $\{\, x^2 \mid x \in \mathbb{R} \,\} = [0, +\infty)$.

a) $\{\, 1/x \mid x \geq 1 \,\}$
b) $\{\, 1/x \mid |x| \geq 1 \,\}$
c) $\{\, \mathrm{e}^{-x^2} \mid x \in \mathbb{R} \,\}$
d) $\{\, x + y \mid x^2 + y^2 < 1 \,\}$
e) $\{\, x + y \mid x^2 + y^2 = 1 \,\}$
f) $\{\, x - y \mid x^2 + y^2 = 1 \,\}$
g) $\{\, |x| + |y| \mid x^2 + y^2 = 1 \,\}$
h) $\{\, x_1 + \cdots + x_n \mid \mathbf{x} \in \mathbb{R}^n,\ x_1^2 + \cdots + x_n^2 = 1 \,\}$
i) $\{\, |x - y| \mid x \in [0, 1],\ y \in (1, 2] \,\}$
j) $\{\, x + y \mid |x| \geq 1,\ |y| \geq 1 \,\}$

9.3. What is the interior and the boundary of these sets?

a) $\{\, (x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1,\ y \geq 0 \,\}$
b) $\{\, (x, y) \in \mathbb{R}^2 \mid y = x^2,\ -1 < x \leq 1 \,\}$

c) $\{ (x, y) \in \mathbb{R}^2 \mid xy < 1,\ x > 0,\ y > 0 \}$
d) $\{ \mathbf{x} \in \mathbb{R}^n \mid \max_{i=1}^n x_i \leq 1 \}$
e) $\{ \mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} = b \}$, where $\mathbf{a} \in \mathbb{R}^n$, $b \in \mathbb{R}$ (superplane)
f) $\{ \mathbf{x} \in \mathbb{R}^n \mid b \leq \mathbf{a}^T\mathbf{x} \leq c \}$, where $\mathbf{a} \in \mathbb{R}^n$, $b, c \in \mathbb{R}$ (panel)
g) $\{ \mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}\mathbf{x} = \mathbf{b} \}$, where $\mathbf{a}$ is wide (affine subspace of $\mathbb{R}^n$)

9.4. Given function $f\colon \mathbb{R}^n \to \mathbb{R}$, sets $Y \subseteq X \subseteq \mathbb{R}^n$, and point $\mathbf{x} \in Y$. Consider these two assertions:

a) function f has at point $\mathbf{x}$ local minimum over set X.
b) function f has at point $\mathbf{x}$ local minimum over set Y.

Does the second assertion follow from the first one? Does the first one follow from the second? Prove from the definition of local extrema or disprove by giving a counter example.

9.5. Can it happen that a function over a set has a local minimum but does not have global minimum? Prove your answer.

Chapter 10

Analytical Conditions for Local Extrema

10.1 Free local extrema

Theorem 10.1. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ a $\mathbf{x} \in X \subseteq \mathbb{R}^n$. Let*

- *funkce f je v bodě $\mathbf{x}$ diferencovatelná,*
- *$\mathbf{x}$ je vnitřní bod množiny X,*
- *$\mathbf{x}$ je local extrém funkce f na množině X.*

Pak $f'(\mathbf{x}) = \mathbf{0}$, neboli všechny parciální derivace funkce f v bodě $\mathbf{x}$ jsou nulové.

Proof. Z Definice 9.2 plyne, že existuje $\varepsilon > 0$ tak, že funkce f má v bodě $\mathbf{x}$ (globální) extrém na okolí $U_\varepsilon(\mathbf{x})$. Z toho ovšem plyne, že řez $\varphi(\alpha) = f(\mathbf{x} + \alpha\mathbf{v})$ funkce f (viz §8.4) v libovolném směru $\mathbf{v} \neq \mathbf{0}$ má (globální) extrém v bodě $\alpha = 0$ na množině $\{\, \alpha \in \mathbb{R} \mid |\alpha| \leq \varepsilon/\|\mathbf{v}\| \,\}$. Tedy funkce φ má v bodě $\alpha = 0$ local extrém. Tedy její derivace je v tomto bodě nulová (to víme z analýzy funkcí jedné proměnné). Ale tato derivace je směrová derivace funkce f v bodě $\mathbf{x}$ ve směru $\mathbf{v}$. Parciální derivace jsou speciálním případem směrové derivace. □

Bod, ve kterém má funkce všechny parciální derivace nulové, se nazývá její **stacionární bod**. Věta 10.1 svádí k tomu, aby se použila v situacích, kdy nejsou splněny její předpoklady. Uveďme příklady tohoto chybého použití.

Example 10.1. V Příkladu 9.9 jsou předpoklady Věty 10.1 splněny pouze pro body b, c, které jsou stacionární a vnitřní. Body a, f jsou hraniční (tedy ne vnitřní) body intervalu $[a, f]$ a v bodech d, e není funkce diferencovatelná. □

Example 10.2. Funkce $f(x) = x^3$ má na $\mathbb{R}$ v bodě 0 stacionární bod, ale nemá tam local extrém. To není v rozporu s Větou 10.1. □

Example 10.3. Funkce $f(\mathbf{x}) = \|\mathbf{x}\|_2$ má na hyperkrychli $X = \{\, \mathbf{x} \in \mathbb{R}^n \mid -\mathbf{1} \leq \mathbf{x} \leq \mathbf{1} \,\}$ v bodě $\mathbf{0}$ volné local minimum (nakreslete si množinu a vrstevnice funkce pro $n = 1$ a pro $n = 2$!). Nemá tam ale stacionární bod, protože tam není diferencovatelná. Dále má funkce na množině X local maxima ve všech rozích hyperkrychle, např. v bodě $\mathbf{1}$. V bodě $\mathbf{1}$ ale není stacionární bod, což není v rozporu s Větou 10.1, protože $\mathbf{1}$ není vnitřní bod X. □

Věta 10.1 říká, že stacionární body jsou body 'podezřelé' z volného lokálního extrému. Udává podmínku *prvního řádu* na volné extrema, protože obsahuje první derivace. Následující podmínka *druhého řádu* pomůže zjistit, zda je stacionární bod lokálním extrémem, případně jakým.

Theorem 10.2. *Let* $f\colon \mathbb{R}^n \to \mathbb{R}$ *a* $\mathbf{x} \in X \subseteq \mathbb{R}^n$. *Let*

- *funkce* f *je v bodě* $\mathbf{x}$ *dvakrát diferencovatelná,*
- $\mathbf{x}$ *je vnitřní bod množiny* X,
- $f'(\mathbf{x}) = \mathbf{0}$.

Pak platí:

- *Je-li Hessova matice* $f''(\mathbf{x})$ *pozitivně [negativně] definitní, pak* $\mathbf{x}$ *je local minimum [maximum] funkce* f *na množině* X.
- *Je-li* $f''(\mathbf{x})$ *indefinitní, pak* $\mathbf{x}$ *není local extrém funkce* f *na množině* X.

I když Větu 10.2 nebudeme dokazovat, základní myšlenka důkazu není překvapující. Místo funkce f vyšetřujeme v blízkosti bodu $\mathbf{x}$ její Taylorův polynom druhého stupně (8.11c),

$$T_2(\mathbf{y}) = f(\mathbf{x}) + \underbrace{f'(\mathbf{x})(\mathbf{y}-\mathbf{x})}_{\mathbf{0}} + \tfrac{1}{2}(\mathbf{y}-\mathbf{x})^T f''(\mathbf{x})\,(\mathbf{y}-\mathbf{x}).$$

Protože $f'(\mathbf{x}) = \mathbf{0}$, lineární člen je nulový a polynom je tedy kvadratická forma posunutá do bodu $\mathbf{x}$. Rozdíl je ale v tom, že pokud je kvadratická forma (pozitivně či negativně) semidefinitní, má v počátku extrém, zatímco Věta 10.2 o případu, kdy je $f''(\mathbf{x})$ semidefinitní, nic nepraví. V tom případě v bodě $\mathbf{x}$ local extrém být může nebo nemusí (příkladem jsou funkce $f(x) = x^3$ a $f(x) = x^4$ v bodě $x = 0$). Bod $\mathbf{x}$, ve kterém je $f'(\mathbf{x}) = \mathbf{0}$ a matice $f''(\mathbf{x})$ je indefinitní, se nazývá **sedlový bod**.

Example 10.4. extrema kvadratické funkce (5.9) umíme hledat pomocí rozkladu na čtverec. Ovšem je to také možné pomocí derivací. Podmínka stacionarity je

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c) = 2\mathbf{x}^T\mathbf{A}^T + \mathbf{b}^T = \mathbf{0}.$$

Po transpozici dostaneme rovnici (5.11a). Druh extrému určíme podle druhé derivace (Hessiánu), který je roven $2\mathbf{A}$ (předpokádáme symetrii $\mathbf{A}$). To souhlasí s klasifikací extrémů kvadratické formy z §5. □

10.2 local extrema vázané rovnostmi

Hledejme minimum funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ na množině

$$X = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{g}(\mathbf{x}) = \mathbf{0} \,\}, \tag{10.1}$$

kde $\mathbf{g} = (g_1, \ldots, g_m)\colon \mathbb{R}^n \to \mathbb{R}^m$. To odpovídá úloze (1.4) s omezeními typu rovnosti:

$$\begin{array}{rl} \min & f(x_1, \ldots, x_n) \\ \text{za podmínek} & g_i(x_1, \ldots, x_n) = 0, \quad i = 1, \ldots, m. \end{array} \tag{10.2}$$

Mluvíme o minimu funkce f *vázaném rovnostmi* $\mathbf{g}(\mathbf{x}) = \mathbf{0}$.

Množina X obsahuje všechna řešení soustavy $\mathbf{g}(\mathbf{x}) = \mathbf{0}$, což je soustava m rovnic o n neznámých. Množina X obvykle nemá žádné vnitřní body, proto nelze použít Větu 10.1. V některých případech ale lze vyjádřit všechna řešení soustavy $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ parametricky a úlohu tak převést na úlohu bez omezení. Toto jsme použili v Příkladu 1.2, uveďme další příklady.

Example 10.5. Hledejme obdélník s jednotkovým obsahem a minimálním obvodem. Tedy minimalizujeme funkci $f(x, y) = x + y$ za podmínky $xy = 1$, neboli hledáme minima f na množině $X = \{ (x, y) \in \mathbb{R}^2 \mid g(x, y) = 1 - xy = 0 \}$.

Množina X nemá žádné vnitřní body (dokažte!), proto nelze použít Větu 10.1. Z podmínky ale máme $y = 1/x$, což dosazeno do účelové funkce dá $f(x, 1/x) = x + 1/x$. Dle Věty 10.1 má tato funkce na svém definičním oboru dva stacionární body $x = \pm 1$. Tedy body podezřelé z lokálního extrému jsou $(x, y) = \pm(1, 1)$. □

Example 10.6. Řešme úlohu

$$\begin{aligned} \min \quad & x + y \\ \text{za podmínky} \quad & x^2 + y^2 = 1, \end{aligned} \tag{10.3}$$

tedy hledáme minimum funkce $f(x, y) = x + y$ na množině $X = \{ (x, y) \in \mathbb{R}^2 \mid g(x, y) = 0 \}$ kde $g(x, y) = 1 - x^2 - y^2$.

Množina X nemá žádné vnitřní body. Ale lze ji parametrizovat jako $X = \{ (\cos z, \sin z) \mid z \in \mathbb{R} \}$. Úlohu tak převedeme na hledání lokálních extrémů funkce jedné proměnné $f(\cos z, \sin z) = \cos z + \sin z$. Podmínka stacionarity $\mathrm{d}f(\cos z, \sin z)/\mathrm{d}z = -\sin z + \cos z = 0$ má dvě řešení $z = \pm\frac{\pi}{2}$. Tedy body podezřelé z lokálního extrému jsou $(x, y) = \pm\frac{\sqrt{2}}{2}(1, 1)$. □

Někdy ovšem množinu (10.1) parametrizovat nejde nebo je to složité. Nyní proto odvodíme obecnější postup, *metodu Lagrangeových multiplikátorů*.

10.2.1 Tečný a ortogonální prostor k povrchu

Zapomeňme nejprve na účelovou funkci f a zkoumejme jen množinu (10.1). Předpokládejme, že zobrazení $\mathbf{g}$ je v okolí nějakého bodu $\mathbf{x} \in X$ spojitě diferencovatelné. V tom případě je množina X v okolí bodu $\mathbf{x}$ 'zakřivený povrch'[1] v $\mathbb{R}^n$. Pak existuje **tečný prostor** (množina všech tečných vektorů) a **ortogonální prostor** (množina všech kolmých vektorů) k povrchu X v bodě $\mathbf{x}$. Tyto dva prostory jsou ortogonální doplněk jeden druhého (viz §4.3). Zde přesné definice pojmů 'vektor tečný k povrchu' a 'vektor kolmý k povrchu' neuvádíme a spoléháme na geometrickou intuici. Následující lema uvádíme bez důkazu[2].

Lemma 10.3. *Let zobrazení* $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^m$ *je v bodě* $\mathbf{x} \in X$ *spojitě diferencovatelné. Let*

$$\operatorname{rank} \mathbf{g}'(\mathbf{x}) = m. \tag{10.4}$$

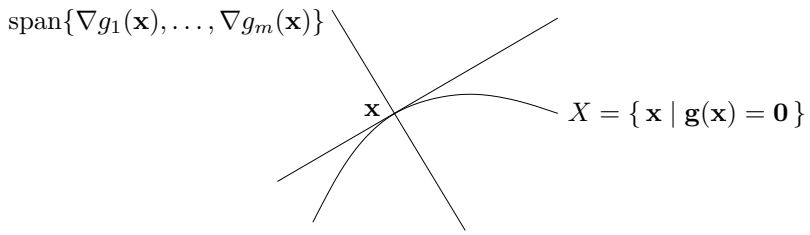
Pak ortogonální prostor k množině X *v bodě* $\mathbf{x}$ *je množina*

$$\operatorname{rng} \mathbf{g}'(\mathbf{x})^T = \operatorname{span}\{\nabla g_1(\mathbf{x}), \ldots, \nabla g_m(\mathbf{x})\}. \tag{10.5}$$

[1] Přesněji, množina X je příkladem objektu, který se nazývá *diferencovatelný manifold*. Studiem takových objektů se zabývá *diferenciální geometrie*.

[2] Lema lze dokázat např. pomocí *věty o implicitní funkci*, která se standardně vyučuje v kursech vícerozměrné analýzy, ale ke které jste se nedostali.

Viz obrázek:



Jelikož řádky Jacobiho matice $\mathbf{g}'(\mathbf{x})$ jsou gradienty $\nabla g_i(\mathbf{x})$, podmínka (10.4) vlastně říká, že gradienty $\nabla g_1(\mathbf{x}), \dots, \nabla g_m(\mathbf{x})$ musí být lineárně nezávislé. Bodu $\mathbf{x} \in X$ splňující podmínku (10.4) se někdy říká **regulární bod** povrchu.

Pro $m = 1$ podmínka (10.4) zní $\nabla g(\mathbf{x}) \neq \mathbf{0}$ a lema zobecňuje skutečnost, kterou jsme bez důkazu uvedli v §8.5, totiž že gradient funkce je v každém bodě kolmý k její vrstevnici. Lema ale navíc říká, že *každý* vektor kolmý k vrstevnici musí být násobek gradientu.

Example 10.7. Let $g\colon \mathbb{R}^2 \to \mathbb{R}$ je funkce $g(x, y) = x^2 + y^2 - 1$. Množina X je jednotková kružnice v $\mathbb{R}^2$. Máme $\nabla g(x, y) = (2x, 2y)$. Protože pro každé $(x, y) \in X$ je $\nabla g(x, y) \neq (0, 0)$, předpoklady Lematu 10.3 jsou splněny a ortogonální prostor k X v bodě (x, y) je množina $\mathrm{span}\{\nabla g(x, y)\} = \{\, (\alpha x, \alpha y) \mid \alpha \in \mathbb{R} \,\}$, což je přímka kolmá ke kružnici. Tečný prostor v bodě (x, y) je ortogonální doplněk této přímky, tedy přímka tečná ke kružnici. □

Example 10.8. Let $g\colon \mathbb{R}^3 \to \mathbb{R}$ je funkce $g(x, y, z) = x^2 + y^2 + z^2 - 1$. Množina X je jednotková sféra v $\mathbb{R}^3$. Máme $\nabla g(x, y, z) = (2x, 2y, 2z)$. Ortogonální prostor k X v bodě (x, y, z) je množina $\mathrm{span}\{\nabla g(x, y, z)\} = \{\, (\alpha x, \alpha y, \alpha z) \mid \alpha \in \mathbb{R} \,\}$, což je přímka kolmá ke sféře. Tečný prostor v bodě (x, y, z) je ortogonální doplněk této přímky, tedy rovina tečná ke sféře. □

Example 10.9. Let $\mathbf{g} = (g_1, g_2)\colon \mathbb{R}^3 \to \mathbb{R}^2$ je zobrazení

$$\mathbf{g}(x, y, z) = (\, x^2 + y^2 + z^2 - 1,\ (x - 1)^2 + y^2 + z^2 - 1 \,).$$

Nulová vrstevnice funkce g_1 je jednotková sféra se středem v bodě $(0, 0, 0)$, nulová vrstevnice funkce g_2 je jednotková sféra se středem v bodě $(1, 0, 0)$. Množina X je průnik těchto dvou sfér, je to tedy kružnice v $\mathbb{R}^3$. Máme $\nabla g_1(x, y, z) = 2(x, y, z)$ a $\nabla g_2(x, y, z) = 2(x - 1, y, z)$. Ortogonální prostor k množině X v bodě (x, y, z) je množina $\mathrm{span}\{\nabla g_1(x, y, z), \nabla g_2(x, y, z)\} = \{\, \alpha_1(x, y, z) + \alpha_2(x - 1, y, z) \mid \alpha_1, \alpha_2 \in \mathbb{R} \,\}$, což je rovina kolmá ke kružnici v bodě (x, y, z). Tečný prostor je ortogonální doplněk této množiny, tedy přímka tečná ke kružnici. □

Example 10.10. Let $g\colon \mathbb{R}^2 \to \mathbb{R}$ je funkce $g(x, y) = (x^2 + y^2 - 1)^2$. Množina X je stejná kružnice jako v Příkladě 10.7. Máme $\nabla g(x, y) = 4(x^2 + y^2 - 1)(x, y)$. Pro každý bod $(x, y) \in X$ je $\nabla g(x, y) = (0, 0)$, tedy předpoklady Lematu 10.3 nejsou splněny. Ortogonální prostor není množina $\mathrm{span}\{\nabla g(x, y)\} = \{(0, 0)\}$. □

10.2.2 Podmínky prvního řádu

Nyní přidáme do našich úvah i účelovou funkci f. Je intuitivně zřejmé (důkaz neuvádíme), že pokud $\mathbf{x}$ má být local extrém funkce f na množině X, směrová derivace $f'(\mathbf{x})\mathbf{v} = \nabla f(\mathbf{x})^T \mathbf{v}$ funkce f v bodě $\mathbf{x}$ v každém směru $\mathbf{v}$ tečném k povrchu X musí být nulová. To znamená, že

gradient $\nabla f(\mathbf{x})$ musí být kolmý k tečnému prostoru v bodě $\mathbf{x}$, neboli musí patřit do ortogonálního prostoru (10.5), neboli musí být lineární kombinací gradientů $\nabla g_1(\mathbf{x}), \dots, \nabla g_m(\mathbf{x})$. Tedy existují čísla $\lambda_1, \dots, \lambda_m \in \mathbb{R}$ tak, že

$$\nabla f(\mathbf{x}) + \lambda_1 \nabla g_1(\mathbf{x}) + \cdots + \lambda_m \nabla g_m(\mathbf{x}) = \mathbf{0}. \tag{10.6}$$

Výsledek těchto úvah se obvykle formuluje následujícím způsobem.

Theorem 10.4. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$, $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{x} \in X$. Let*

- *f a $\mathbf{g}$ jsou v bodě $\mathbf{x}$ spojitě diferencovatelné,*
- $\operatorname{rank} \mathbf{g}'(\mathbf{x}) = m$,
- *bod $\mathbf{x}$ je local extrém funkce f na množině X.*

Pak existují čísla $(\lambda_1, \dots, \lambda_m) = \boldsymbol{\lambda} \in \mathbb{R}^m$ tak, že $L'(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{0}$, kde funkce $L\colon \mathbb{R}^{n+m} \to \mathbb{R}$ je dána jako

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x}) = f(\mathbf{x}) + \lambda_1 g_1(\mathbf{x}) + \cdots + \lambda_m g_m(\mathbf{x}). \tag{10.7}$$

Zápis $L'(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{0}$ označuje, že parciální derivace funkce L podle $x_1, \dots, x_n, \lambda_1, \dots, \lambda_m$ jsou nulové, neboli bod $(\mathbf{x}, \boldsymbol{\lambda}) \in \mathbb{R}^{m+n}$ je stacionární bod funkce L. Rovnost $\partial L(\mathbf{x}, \boldsymbol{\lambda})/\partial \mathbf{x} = \mathbf{0}$ je ekvivalentní rovnosti (10.6). Rovnost $\partial L(\mathbf{x}, \boldsymbol{\lambda})/\partial \boldsymbol{\lambda} = \mathbf{g}(\mathbf{x}) = \mathbf{0}$ je ekvivalentní omezením. Číslům λ_i se říká **Lagrangeovy multiplikátory** a funkci (10.7) **Lagrangeova funkce**.

Example 10.11. Řešme znovu úlohu (10.3). Lagrangeova funkce je

$$L(x, y, \lambda) = x + y + \lambda(1 - x^2 - y^2).$$

Její stacionární body (x, y, λ) jsou řešeními soustavy tří rovnic o třech neznámých

$$\begin{aligned} \partial L(x, y, \lambda)/\partial x &= 1 - 2\lambda x &&= 0 \\ \partial L(x, y, \lambda)/\partial y &= 1 - 2\lambda y &&= 0 \\ \partial L(x, y, \lambda)/\partial \lambda &= 1 - x^2 - y^2 &&= 0. \end{aligned}$$

První dvě rovnice dají $x = y = 1/(2\lambda)$. Dosazením do třetí máme $2/(2\lambda)^2 = 1$, což dá dva kořeny $\lambda = \pm 1/\sqrt{2}$. Stacionární body funkce L jsou dva, $(x, y, \lambda) = \pm(1, 1, 1)/\sqrt{2}$. Tedy máme dva kandidáty na local extrema, $(x, y) = \pm(1, 1)/\sqrt{2}$.

Tuto jednoduchou úlohu je samozřejmě snadné vyřešit úvahou. Nakreslete si kružnici $X = \{\, (x, y) \mid x^2 + y^2 = 1 \,\}$ a několik vrstevnic funkce f a najděte kýžené extrema! □

Example 10.12. Řešme úlohu (10.3), kde ale omezení změníme na $g(x, y) = (1 - x^2 - y^2)^2 = 0$. Podle Příkladu 10.10 máme $g'(x, y) = (0, 0)$ pro každé $(x, y) \in X$, čekáme tedy problém.

Stacionární body Lagrangeovy funkce $L(x, y, \lambda) = x + y + \lambda(1 - x^2 - y^2)^2$ musí splňovat

$$\begin{aligned} \partial L(x, y, \lambda)/\partial x &= 1 - 4\lambda x(1 - x^2 - y^2) &&= 0 \\ \partial L(x, y, \lambda)/\partial y &= 1 - 4\lambda y(1 - x^2 - y^2) &&= 0 \\ \partial L(x, y, \lambda)/\partial \lambda &= (1 - x^2 - y^2)^2 &&= 0. \end{aligned}$$

Tyto rovnice si odporují. Jelikož $1 - x^2 - y^2 = 0$, tak např. první rovnice říká $1 - 4\lambda x \cdot 0 = 0$, což neplatí pro žádné (x, λ). Závěr je, že local extrema $(x, y) = \pm(1, 1)/\sqrt{2}$ jsme nenašli. □

Example 10.13. Vrať me se k úloze (6.14), tedy k hledání řešení nehomogenní lineární soustavy s nejmenší normou. Lagrangeova funkce je

$$L(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{x}^T\mathbf{x} + 2\boldsymbol{\lambda}^T(\mathbf{b} - \mathbf{A}\mathbf{x}),$$

kde přidaná dvojka nemění situaci. Je $\partial L(\mathbf{x}, \boldsymbol{\lambda})/\partial \mathbf{x} = 2\mathbf{x}^T - 2\boldsymbol{\lambda}^T\mathbf{A}$ (odvoďte!). Stacionární body funkce L tedy získáme řešením soustavy (6.15), kterou jsme v 6.2 odvodili úvahou. □

Předchozí příklad vyžaduje od studenta nejen znalost metody Lagrangeových multiplikátorů, ale i jistou zručnost v manipulaci s maticovými výrazy. Cvičte tuto zručnost ve Cvičeních 10.22–10.25!

Věta 10.4 udává podmínky prvního řádu na extrema vázané rovnostmi. Říká, že pokud $(\mathbf{x}, \boldsymbol{\lambda})$ je stacionární bod Lagrangeovy funkce, pak bod $\mathbf{x}$ je 'podezřelý' z lokálního extrému funkce f na množině X. Jak poznáme, zda tento bod je local extrém, případně jaký? Podmínky druhého řádu pro vázané extrema uvádíme nepovinně v §10.2.3. Zde pouze zdůrazníme, že druh lokálního extrému *nelze* zjistit podle definitnosti Hessovy matice $L''(\mathbf{x}, \boldsymbol{\lambda})$, tedy je chybou použít Větu 10.2 na funkci L.

10.2.3 (⋆) Podmínky druhého řádu

Theorem 10.5. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$, $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{x} \in \mathbb{R}^n$ a $\boldsymbol{\lambda} \in \mathbb{R}^m$. Let*

- *$(\mathbf{x}, \boldsymbol{\lambda})$ je stacionární bod Lagrangeovy funkce, neboli $\partial L(\mathbf{x}, \boldsymbol{\lambda})/\partial \mathbf{x} = \mathbf{0}$ a $\partial L(\mathbf{x}, \boldsymbol{\lambda})/\partial \boldsymbol{\lambda} = \mathbf{0}$,*
- *f a $\mathbf{g}$ jsou dvakrát diferencovatelné v bodě $\mathbf{x}$.*

Pak platí:

- *Je-li $\partial^2 L(\mathbf{x}, \boldsymbol{\lambda})/\partial \mathbf{x}^2$ pozitivně [negativně] definitní na nulovém prostoru matice $\mathbf{g}'(\mathbf{x})$, má f v bodě $\mathbf{x}$ ostré local minimum [maximum] vázané podmínkou $\mathbf{g}(\mathbf{x}) = \mathbf{0}$.*
- *Je-li $\partial^2 L(\mathbf{x}, \boldsymbol{\lambda})/\partial \mathbf{x}^2$ indefinitní na nulovém prostoru matice $\mathbf{g}'(\mathbf{x})$, nemá f v bodě $\mathbf{x}$ local minimum ani local maximum vázané podmínkou $\mathbf{g}(\mathbf{x}) = \mathbf{0}$.*

Zde výraz

$$\frac{\partial^2 L(\mathbf{x}, \boldsymbol{\lambda})}{\partial \mathbf{x}^2} = f''(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i\, g_i''(\mathbf{x})$$

značí druhou derivaci (Hessovu matici) funkce $L(\mathbf{x}, \boldsymbol{\lambda})$ podle $\mathbf{x}$ v bodě $(\mathbf{x}, \boldsymbol{\lambda})$. Tvrzení, že matice $\mathbf{A}$ je pozitivně definitní na množině T znamená, že $\mathbf{y}^T\mathbf{A}\mathbf{y} > 0$ pro každé $\mathbf{y} \in T \setminus \{\mathbf{0}\}$.

Jak zjistíme definitnost dané matice $\mathbf{A}$ na nulovém prostoru Jacobiho matice $\mathbf{g}'(\mathbf{x})$? Najdeme-li bázi $\mathbf{B}$ tohoto nulového prostoru, pak každý prvek množiny T lze parametrizovat jako $\mathbf{y} = \mathbf{B}\mathbf{z}$. Protože $\mathbf{y}^T\mathbf{A}\mathbf{y} = \mathbf{z}^T\mathbf{B}^T\mathbf{A}\mathbf{B}\mathbf{z}$, převedli jsme problém na zjišťování definitnosti matice $\mathbf{B}^T\mathbf{A}\mathbf{B}$.

Example 10.14. Najděme strany kvádru s jednotkovým objem a minimálním povrchem. Tedy minimalizujeme $xy + xz + yz$ za podmínky $xyz = 1$. Lagrangeova funkce je

$$L(x, y, z, \lambda) = xy + xz + yz + \lambda(1 - xyz).$$

Položením derivací L rovným nule máme soustavu

$$\begin{aligned}
L'_x(x, y, z, \lambda) &= y + z - \lambda yz &= 0\\
L'_y(x, y, z, \lambda) &= x + z - \lambda xz &= 0\\
L'_z(x, y, z, \lambda) &= x + y - \lambda xy &= 0\\
L'_\lambda(x, y, z, \lambda) &= xyz - 1 &= 0.
\end{aligned}$$

Soustava je zjevně splněna pro $(x, y, z, \lambda) = (1, 1, 1, 2)$. Máme ukázat, že tento bod odpovídá lokálnímu minimu. Máme

$$\frac{\partial^2 L(x, y, z, \lambda)}{\partial (x, y, z)^2} = \begin{bmatrix} 0 & 1-\lambda z & 1-\lambda y \\ 1-\lambda z & 0 & 1-\lambda x \\ 1-\lambda y & 1-\lambda x & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix}. \tag{10.8}$$

Ukážeme, že tato matice je pozitivně definitní na nulovém prostoru Jacobiho matice

$$g'(x, y, z) = \begin{bmatrix} -yz & -xz & -xy \end{bmatrix} = \begin{bmatrix} -1 & -1 & -1 \end{bmatrix}.$$

Nejdříve zkusme štěstí, zda matice (10.8) není pozitivně definitní již na $\mathbb{R}^3$ – v tom případě by zjevně byla pozitivně definitní i na nulovém prostoru $g'(x, y, z)$ (promyslete, proč to tak je!). Není tomu tak, protože její vlastní čísla jsou $\{-2, 1, 1\}$, tedy je indefinitní.

Nějakou bázi nulového prostoru matice $g'(x, y, z)$ snadno najdeme ručně, např.

$$\mathbf{B} = \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Snadno zjistíme, že matice

$$\mathbf{B}^T \frac{\partial^2 L(x, y, z, \lambda)}{\partial (x, y, z)^2} \mathbf{B} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

má vlastní čísla $\{2, 1\}$, tedy je pozitivně definitní. □

10.3 Cvičení

10.1. Co je vnitřek a hranice těchto množin? Výsledek napište v množinovém zápisu.

a) $\{ (x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1,\ y \geq 0 \}$
b) $\{ (x, y) \in \mathbb{R}^2 \mid y = x^2,\ -1 < x \leq 1 \}$
c) $\{ (x, y) \in \mathbb{R}^2 \mid xy < 1,\ x > 0,\ y > 0 \}$
d) $\{ \mathbf{x} \in \mathbb{R}^n \mid \max_{i=1}^n x_i \leq 1 \}$
e) $\{ \mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} = b \}$, kde $\mathbf{a} \in \mathbb{R}^n$, $b \in \mathbb{R}$ (nadrovina)
f) $\{ \mathbf{x} \in \mathbb{R}^n \mid b < \mathbf{a}^T\mathbf{x} \leq c \}$, kde $\mathbf{a} \in \mathbb{R}^n$, $b, c \in \mathbb{R}$

10.2. Je dána funkce $f\colon \mathbb{R}^n \to \mathbb{R}$, množiny $Y \subseteq X \subseteq \mathbb{R}^n$, a bod $\mathbf{x} \in Y$. Uvažujme dva výroky:

a) Funkce f má v bodě $\mathbf{x}$ local minimum na množině X.
b) Funkce f má v bodě $\mathbf{x}$ local minimum na množině Y.

Vyplývá (b) z (a)? Vyplývá (a) z (b)? Dokažte z definice lokálního extrému nebo vyvraťte nalezením protipříkadu.

10.3. Může nastat případ, kdy funkce na množině má local minimum ale nemá na ní globální minimum? Odpověď dokažte.

10.4. Funkce $f\colon \mathbb{R}^3 \to \mathbb{R}$ má stacionární bod $(2, 1, 5)$. Co se dá o tomto stacionárním bodě říci, když Hessova matice $f''(2, 1, 5)$ v něm má vlastní čísla

a) $\{2, 3, -1\}$
b) $\{2, 3, 0\}$
c) $\{0, -1, 1\}$

10.5. Pro následující funkce spočítejte (na papíře) stacionární body. Pro každý stacionární bod určete, zda je to local minimum, local maximum, či sedlový bod. Pokud to určit nedokážete, odůvodněte.

a) $f(x, y) = x(1 - \frac{2}{3}x^2 - y^2)$
b) $f(x, y) = 1/x + 1/y + xy$
c) $f(x, y) = \mathrm{e}^y(y^2 - x^2)$
d) $f(x, y) = 3x - x^3 - 3xy^2$
e) $f(x, y) = 6xy^2 - 2x^3 - 3y^4$
f) $f(x, y) = x^4/3 + y^4/2 - 4xy^2 + 2x^2 + 2y^2 + 3$
g) $f(x, y, z) = x^3 + y^3 + 2xyz + z^2$

10.6. Dokažte, že funkce $f(x, y) = x$ nabývá za podmínky $x^3 = y^2$ minima pouze v počátku. Ukažte, že metoda Lagrangeových multiplikátorů toto minimum nenajde.

Následující úlohy se pokuste vyřešit parametrizací podmínek (analogicky k Příkladu 10.6) a pak metodou Lagrangeových multiplikátorů. Pokud jedna z těchto metod není použitelná, vynechte ji. Při použití metody Lagrangeových multiplikátorů stačí pouze najít stacionární body Lagrangeovy funkce – nemusíte určovat, jde-li o local extrema a případně jaké.

10.7. Najděte local extrema funkcí

a) $f(x, y) = 2x - y$
b) $f(x, y) = x(y - 1)$
c) $f(x, y) = x^2 + 2y^2$
d) $f(x, y) = x^2y$
e) $f(x, y) = x^4 + y^2$
f) $f(x, y) = \sin(xy)$
g) $f(x, y) = \mathrm{e}^{xy}$

na kružnici $x^2 + y^2 = 1$. Nápověda: Někdy je dobré účelovou funkci zjednodušit, pokud to nezmění řešení.

10.8. Najděte extrema funkce

a) $f(x, y, z) = x + yz$ za podmínek $x^2 + y^2 + z^2 = 1$ a $z^2 = x^2 + y^2$
b) $f(x, y, z) = xyz$ za podmínek $x^2 + y^2 + z^2 = 1$ a $xy + yz + zx = 1$

10.9. Najděte extrema funkce

a) $f(x, y, z) = (x + y)(y + z)$
b) $f(x, y, z) = a/x + b/y + c/z$, kde $a, b, c > 0$ jsou dány
c) $f(x, y, z) = x^3 + y^2 + z$
d) $f(x, y, z) = x^3 + y^3 + z^3 + 2xyz$
e) $(\star)$ $f(x, y, z) = x^3 + y^3 + z^3 - 3xyz$
f) $(\star)$ $f(x, y, z) = x^3 + 2xyz - z^3$

na sféře $x^2 + y^2 + z^2 = 1$.

10.10. Rozložte dané kladné reálné číslo na součin n kladných reálných čísel tak, aby jejich součet byl co nejmenší.

10.11. Spočítejte rozměry tělesa tak, aby mělo při daném objemu nejmenší povrch:

a) kvádr
b) kvádr bez víka (má jednu dolní stěnu a čtyři boční, horní stěna chybí)
c) válec
d) půllitr (válec bez víka)
e) ($\star$) kelímek (komolý kužel bez víka). Objem komolého kužele je $V = \frac{\pi}{3}h(R^2+Rr+r^2)$ a povrch pláště (bez podstav) je $S = \pi(R+r)\sqrt{(R-r)^2+h^2}$. Můžete použít vhodný numerický software na řešení vzniklé soustavy rovnic.

10.12. Najděte bod nejblíže počátku na křivce

a) $x + y = 1$
b) $x + 2y = 5$
c) $y = x^3 + 1$
d) $x^2 + 2y^2 = 1$

10.13. Let $\mathbf{x}^*$ je bod nejblíže počátku na nadploše $h(\mathbf{x}) = 0$. Ukažte metodou Lagrangeových multiplikátorů, že vektor $\mathbf{x}^*$ je kolmý k tečné nadrovině plochy v bodě $\mathbf{x}^*$.

10.14. Máme kouli o poloměru r a středu $\mathbf{x}_0$, tj. množinu $\{\, \mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{x}_0\|_2 \leq r \,\}$. Máme nadrovinu $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} = b \,\}$.

10.15. Do elipsy o daných délkách os vepište obdélník s maximálním obsahem. Předpokládejte přitom, že strany obdélníku jsou rovnoběžné s osami elipsy.

10.16. *Fermatův princip* v paprskové optice říká, že cesta mezi libovolnými dvěma body na paprsku má takový tvar, aby ji světlo proběhlo za čas kratší než jí blízké dráhy. Později se zjistilo, že správným kritériem není *nejkratší* ale *extrémní* čas. Tedy skutečná dráha paprsku musí mít čas větší nebo menší než jí blízké dráhy. Z tohoto principu odvoďte:

a) Zákon odrazu od zrcadla: úhel dopadu se rovná úhlu odrazu.
b) Snellův zákon lomu: na rozhraní dvou prostředí se světlo lomí tak, že

$$\frac{c_1}{c_2} = \frac{\sin\alpha_1}{\sin\alpha_2},$$

kde α_i je úhel paprsku od normály rozhraní a c_i je rychlost světla v prostředí i.

Odvození udělejte

a) pro rovinné zrcadlo a rovinné rozhraní (což vede na minimalizaci bez omezení),
b) pro zrcadlo a rozhraní tvaru obecné plochy s rovnicí $g(\mathbf{x}) = 0$. Dokážete najít situaci, kdy skutečná dráha paprsku má čas *větší* než jí blízké dráhy?

10.17. Rozdělení pravděpodobnosti diskrétní náhodné proměnné je funkce $p\colon \{1, \ldots, n\} \to \mathbb{R}_+$ (tj. soubor nezáporných čísel $p(1), \ldots, p(n)$) splňující $\sum_{x=1}^n p(x) = 1$.

a) *Entropie* náhodné proměnné s rozdělením p je rovna $-\sum_{x=1}^n p(x)\log p(x)$, kde log je přirozený logaritmus. Najděte rozdělení s maximální entropií.

b) Dokažte *Gibbsovu nerovnost* (též zvanou *informační nerovnost*): pro každé dvě rozdělení p, q platí

$$\sum_{x=1}^{n} p(x) \log q(x) \geq \sum_{x=1}^{n} p(x) \log p(x),$$

přičemž rovnost nastává jen tehdy, když $p = q$.

10.18. ($\star$) Máme trojúhelník se stranami délek a, b, c. Uvažujme bod, který má takovou polohu, že součet čtverců jeho vzdáleností od stran trojúhelníku je nejmenší možný. Jaké budou vzdálenosti x, y, z tohoto bodu od stran trojúhelníku?

10.19. ($\star$) Máme krychli s délkou hrany 2. Do stěny krychle je vepsána kružnice (která má tedy poloměr 1) a okolo sousední stěny je opsána kružnice (která má tedy poloměr $\sqrt{2}$). Najděte nejmenší a největší vzdálenost mezi body na kružnicích.

10.20. ($\star$) Najděte extrema funkce

$$f(x, y, z, u, v, w) = (1 + x + u)^{-1} + (1 + y + v)^{-1} + (1 + z + w)^{-1}$$

za podmínek $xyz = a^3$, $uvw = b^3$ a $x, y, z, u, v, w > 0$.

10.21. Popište množinu řešení soustavy

$$\begin{aligned} x + 2y + z &= 1 \\ 2x - y - 2z &= 2. \end{aligned}$$

Najděte takové řešení soustavy, aby výraz $\sqrt{x^2 + y^2 + z^2}$ byl co nejmenší. Najděte co nejvíce způsobů řešení.

10.22. Minimalizujte $\mathbf{x}^T\mathbf{x}$ za podmínky $\mathbf{a}^T\mathbf{x} = 1$. Jaký je geometrický význam úlohy?

10.23. Maximalizujte $\mathbf{a}^T\mathbf{x}$ za podmínky $\mathbf{x}^T\mathbf{x} = 1$. Jaký je geometrický význam úlohy?

10.24. Minimalizujte $\mathbf{x}^T\mathbf{A}\mathbf{x}$ za podmínky $\mathbf{b}^T\mathbf{x} = 1$, kde $\mathbf{A}$ je pozitivně definitní.

10.25. Minimalizujte $\|\mathbf{C}\mathbf{x}\|_2$ za podmínky $\mathbf{A}\mathbf{x} = \mathbf{b}$, kde $\mathbf{A}$ má lineárně nezávislé řádky a $\mathbf{C}$ má lineárně nezávislé sloupce.

10.26. ($\star$) Minimalizujte $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ za podmínky $\mathbf{C}\mathbf{x} = \mathbf{0}$, kde $\mathbf{A}$ má lineárně nezávislé sloupce a $\mathbf{C}$ má lineárně nezávislé řádky.

10.27. ($\star$) Minimalizujte $\|\mathbf{C}\mathbf{x}\|_2$ za podmínek $\mathbf{A}\mathbf{x} = \mathbf{0}$ a $\mathbf{x}^T\mathbf{x} = 1$.

10.28. ($\star$) Minimalizujte $\|\mathbf{A}\mathbf{x}\|_2$ za podmínky $\mathbf{x}^T\mathbf{C}\mathbf{x} = 1$, kde $\mathbf{C}$ je pozitivně definitní.

10.29. ($\star$) Minimalizujte $\mathbf{a}^T\mathbf{x}$ za podmínky $\mathbf{x}^T\mathbf{C}\mathbf{x} = 1$, kde $\mathbf{C}$ je pozitivně definitní.

10.30. ($\star$) Jaké musí být vlastnosti matice $\mathbf{A}$ a vektoru $\mathbf{b}$, aby $\max\{\, \|\mathbf{A}\mathbf{x}\|_2 \mid \mathbf{b}^T\mathbf{x} = 0 \,\} = 0$?

Hints and Solutions

10.1.a) vnitřek $\emptyset$, hranice původní množina

10.1.b) vnitřek $\emptyset$, hranice $\{\, (x, y) \in \mathbb{R}^2 \mid y = x^2,\ -1 \leq x \leq 1 \,\}$

10.1.c) vnitřek původní množina, hranice $\{\, (x, 0) \mid x \geq 0 \,\} \cup \{\, (0, y) \mid y \geq 0 \,\} \cup \{\, (x, y) \mid xy = 1 \,\}$

10.1.d) $\max_{i=1}^{n} x_i \leq 1$ je totéž co $x_i \leq 1$ pro všechna i, tedy množina jde napsat také jako $(-\infty, 1]^n$ (kartézský součin n stejných polootevřených intervalů). Vnitřek je $(-\infty, 1)^n$, hranice (těžko se popíše krátčeji) je $(-\infty, 1]^n \setminus (-\infty, 1)^n$

10.1.e) vnitřek $\emptyset$, hranice původní množina

10.1.f) vnitřek $\{\,\mathbf{x} \mid b < \mathbf{a}^T\mathbf{x} < c\,\}$, hranice $\{\,\mathbf{x} \mid \mathbf{a}^T\mathbf{x} = b\,\} \cup \{\,\mathbf{x} \mid \mathbf{a}^T\mathbf{x} = c\,\}$

10.3. může

10.4.a) funkce nemá v tomto bodě local extrém na $\mathbb{R}^3$

10.4.b) funkce má v tomto bodě local minimum na $\mathbb{R}^3$

10.5.d) Stacionární body jsou 4.

10.5.e) Stacionární body jsou 3.

10.5.f) Stacionárních bodů je 5.

10.5.g) Stacionární body jsou 3, a to $(0,0,0)$, $(3/2,3/2,-9/4)$, $(3/2,3/2,-9/4)$.

10.25. $\mathbf{x} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{A}^T(\mathbf{A}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{A}^T)^{-1}\mathbf{b}$.

10.26. $\mathbf{x} = [\mathbf{I} - (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{C}^T(\mathbf{C}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{C}^T)^{-1}\mathbf{C}^T]\,(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$

Chapter 11

Iterační algoritmy na volné local extrema

Zde se budeme věnovat numerickým iteračním algoritmům na nalezení volného lokálního minima diferencovatelných funkcí na množině $\mathbb{R}^n$.

11.1 Sestupné metody

Iterační algoritmy na hledání lokálního minima spojité funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ mají tvar

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \, \mathbf{v}_k, \tag{11.1}$$

kde vektor $\mathbf{v}_k \in \mathbb{R}^n$ je **směr hledání** a skalár $\alpha_k > 0$ je **délka kroku**. Ve třídě algoritmů zvaných **sestupné metody** (*descent methods*) hodnota účelové funkce monotonně klesá[1], $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$.

Let je funkce f diferencovatelná. Směr $\mathbf{v}_k$ se nazývá **sestupný**, jestliže

$$f'(\mathbf{x}_k)\, \mathbf{v}_k < 0, \tag{11.2}$$

tedy směrová derivace ve směru $\mathbf{v}_k$ je záporná. Pokud v bodě $\mathbf{x}_k$ existuje sestupný směr, existuje délka kroku $\alpha_k > 0$ tak, že $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$. Pokud v bodě $\mathbf{x}_k$ sestupný směr neexistuje, vektor $f'(\mathbf{x}_k)$ je nutně nulový (proč?) a tedy $\mathbf{x}_k$ je stacionární bod.

Máme-li sestupný směr, optimální délku kroku α_k najdeme minimalizací funkce f na polopřímce z bodu $\mathbf{x}_k$ ve směru $\mathbf{v}_k$. Tedy minimalizujeme funkci jedné proměnné

$$\varphi(\alpha_k) = f(\mathbf{x}_k + \alpha_k \mathbf{v}_k) \tag{11.3}$$

přes všechny $\alpha_k \geq 0$. Tato úloha je v kontextu vícerozměrné optimalizace nazývána *line search*. Úlohu stačí řešit přibližně. Takovou přibližnou metodu není obtížné vymyslet a proto se jí dále nebudeme zabývat.

Dále uvedeme nejznámější zástupce sestupných metod.

[1] Existují totiž i algoritmy, ve kterých hodnota $f(\mathbf{x}_k)$ neklesá monotonně (tj. někdy stoupne a někdy klesne) a přesto konvergují k optimu (např. *subgradientní metody*).

11.2 Gradientní metoda

Tato nejjednodušší metoda volí směr sestupu jako záporný gradient funkce f v bodě $\mathbf{x}_k$:

$$\mathbf{v}_k = -f'(\mathbf{x}_k)^T = -\nabla f(\mathbf{x}_k). \tag{11.4}$$

Tento směr je sestupný, což je okamžitě vidět dosazením do (11.2).

Nevýhodou gradientní metody je to, že konvergence může být pomalá kvůli 'cik-cak' chování. To se může stát tehdy, když funkce v okolí lokálního optima je v některých směrech mnohem protaženější než v jiných (přesněji, když vlastní čísla Hessiánu $f''(\mathbf{x})$ mají velmi různé velikosti). Výhodou metody je spolehlivost, protože směr je vždy sestupný.

11.2.1 ($\star$) Závislost na lineární transformaci souřadnic

Transformujme vektor proměnných $\mathbf{x}$ lineární transformací $\tilde{\mathbf{x}} = \mathbf{A}\mathbf{x}$, kde $\mathbf{A}$ je čtvercová regulární matice. Je jasné, že úloha v nových proměnných bude mít stejné optimum jako v původních proměnných. Tedy

$$\min_{\mathbf{x}} f(\mathbf{x}) = \min_{\tilde{\mathbf{x}}} \tilde{f}(\tilde{\mathbf{x}}), \qquad \text{kde} \quad \tilde{f}(\tilde{\mathbf{x}}) = \tilde{f}(\mathbf{A}\mathbf{x}) = f(\mathbf{x}) = f(\mathbf{A}^{-1}\tilde{\mathbf{x}}).$$

Iterace gradientní metody v nových proměnných je

$$\tilde{\mathbf{x}}_{k+1} = \tilde{\mathbf{x}}_k - \alpha_k \, \tilde{f}'(\tilde{\mathbf{x}}_k)^T. \tag{11.5}$$

Zkoumejme, jaké iteraci to odpovídá v původních proměnných. K tomu potřebujeme vyjádřit (11.5) v proměnných $\mathbf{x}$. Použitím řetězového pravidla odvodíme

$$\tilde{f}'(\tilde{\mathbf{x}}) = \frac{\mathrm{d}\tilde{f}(\tilde{\mathbf{x}})}{\mathrm{d}\tilde{\mathbf{x}}} = \frac{\mathrm{d}\tilde{f}(\tilde{\mathbf{x}})}{\mathrm{d}\mathbf{x}} \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\tilde{\mathbf{x}}} = \frac{\mathrm{d}f(\mathbf{x})}{\mathrm{d}\mathbf{x}} \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\tilde{\mathbf{x}}} = f'(\mathbf{x})\mathbf{A}^{-1}.$$

Dosazením za $\tilde{\mathbf{x}}$ a $\tilde{f}'(\tilde{\mathbf{x}})$ do (11.5) a úpravou dostaneme

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \, (\mathbf{A}^T\mathbf{A})^{-1} f'(\mathbf{x}_k)^T. \tag{11.6}$$

To lze napsat ve tvaru (11.1) se směrem hledání

$$\mathbf{v}_k = -(\mathbf{A}^T\mathbf{A})^{-1} f'(\mathbf{x}_k)^T. \tag{11.7}$$

Tento směr se liší od původního směru (11.4) vynásobením maticí $(\mathbf{A}^T\mathbf{A})^{-1}$. Vidíme tedy, že gradientní metoda *není invariantní* vůči lineární transformaci souřadnic.

Ovšem lze ukázat, že nový směr (11.7) je také sestupný. Dosazením (11.4) do (11.2) to znamená, že $-f'(\mathbf{x}_k)(\mathbf{A}^T\mathbf{A})^{-1} f'(\mathbf{x}_k)^T < 0$. To je ale pravda, neboť matice $\mathbf{A}^T\mathbf{A}$ a tedy i její inverze je pozitivně definitní, viz Cvičení 5.18.

Na vzorec (11.7) se lze dívat ještě obecněji. Je jasné, že směr $\mathbf{v}_k = -\mathbf{C}_k^{-1} f'(\mathbf{x}_k)^T$ je sestupný, je-li matice $\mathbf{C}_k$ pozitivně definitní. Dá se ukázat i opak, totiž že každý sestupný směr lze napsat takto. Matice $\mathbf{C}_k$ může být jiná v každém kroku. Uvidíme, že algoritmy uvedené dále budou mít vždy tento tvar.

11.3 Newtonova metoda

Newtonova metoda (přesněji Newton-Raphsonova) je slavný iterační algoritmus na řešení soustav nelineárních rovnic. Lze ho použít i na minimalizaci funkce tak, že hledáme nulový gradient. Oba způsoby použití popíšeme.

11.3.1 Použití na soustavy nelineárních rovnic

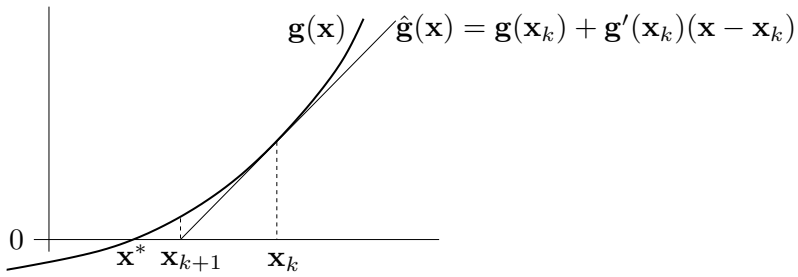
Řešme rovnici $\mathbf{g}(\mathbf{x}) = \mathbf{0}$, kde $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^n$ je diferencovatelné zobrazení. Jedná se tedy o soustavu n rovnic s n neznámými. Zobrazení $\mathbf{g}$ aproximujeme v okolí bodu $\mathbf{x}_k$ Taylorovým polynomem prvního stupně

$$\mathbf{g}(\mathbf{x}) \approx \hat{\mathbf{g}}(\mathbf{x}) = \mathbf{g}(\mathbf{x}_k) + \mathbf{g}'(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k), \tag{11.8}$$

kde Jacobiho matice $\mathbf{g}'(\mathbf{x}_k) \in \mathbb{R}^{n\times n}$ je derivace zobrazení v bodě $\mathbf{x}_k$. Další iteraci $\mathbf{x}_{k+1}$ najdeme řešením nehomogenní lineární soustavy $\hat{\mathbf{g}}(\mathbf{x}_{k+1}) = \mathbf{0}$. Pokud je Jacobiho matice regulární, řešením je

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{g}'(\mathbf{x}_k)^{-1}\mathbf{g}(\mathbf{x}_k). \tag{11.9}$$

Viz obrázek:



Hlavní výhodou Newtonovy metody je, že v blízkém okolí řešení obvykle konverguje velmi rychle (mnohem rychleji než gradientní metoda). Nevýhodou je, že je nutno začít s poměrně přesnou aproximací $\mathbf{x}_0$ skutečného řešení, jinak algoritmus snadno diverguje.

Example 11.1. *Babylónská metoda* na výpočet druhé odmocniny čísla $a \geq 0$ je dána iterací

$$x_{k+1} = \frac{1}{2}\Big(x_k + \frac{a}{x_k}\Big).$$

To není nic jiného než Newtonova metoda pro řešení rovnice $0 = g(x) = x^2 - a$. Opravdu,

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)} = x_k - \frac{x_k^2 - a}{2x_k} = x_k - \frac{1}{2}\Big(x_k - \frac{a}{x_k}\Big) = \frac{1}{2}\Big(x_k + \frac{a}{x_k}\Big). \qquad \square$$

Example 11.2. Hledejme průsečík křivek $(x-1)^2 + y^2 = 1$ a $x^4 + y^4 = 1$. Máme $n = 2$ a

$$\mathbf{x} = (x, y) = \begin{bmatrix} x \\ y \end{bmatrix}, \quad \mathbf{g}(\mathbf{x}) = \mathbf{g}(x, y) = \begin{bmatrix} (x-1)^2 + y^2 - 1 \\ x^4 + y^4 - 1 \end{bmatrix}, \quad \mathbf{g}'(\mathbf{x}) = \mathbf{g}'(x, y) = \begin{bmatrix} 2(x-1) & 2y \\ 4x^3 & 4y^3 \end{bmatrix}.$$

Iterace (11.9) je

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \begin{bmatrix} 2(x_k - 1) & 2y_k \\ 4x_k^3 & 4y_k^3 \end{bmatrix}^{-1} \begin{bmatrix} (x_k - 1)^2 + y_k^2 - 1 \\ x_k^4 + y_k^4 - 1 \end{bmatrix}.$$

Načrtneme-li si obě křivky, vidíme, že mají dva průsečíky. Zvolme počáteční odhad pro horní průsečík $(x_0, y_0) = (1, 1)$. První iterace bude

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 & 2 \\ 4 & 4 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.75 \\ 1 \end{bmatrix}.$$

Šestá iterace $(x_6, y_6) = (\,0.671859751039018,\ 0.944629015546222\,)$ je taková, že rovnice jsou splněny se strojovou přesností. □

Example 11.3. Funkce $g(x) = x^2 - 1$ má dva nulové body $x = \pm 1$. Pokud v nějaké iteraci bude $x_k = 0$, nastane dělení nulou. Pokud bude x_k velmi malé, dělení nulou nenastane, ale iterace x_{k+1} se ocitne velmi daleko od kořene. □

Example 11.4. Pro funkci $g(x) = x^3 - 2x + 2$ zvolme $x_0 = 0$. Další iterace bude $x_1 = 1$ a další $x_2 = 0$. Algoritmus bude oscilovat mezi hodnotami 0 a 1, tedy bude divergovat. □

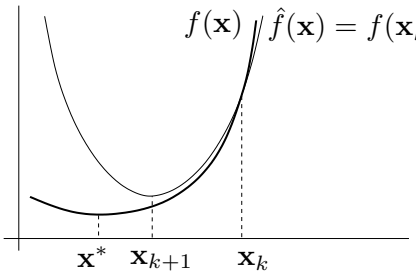
11.3.2 Použití na minimalizaci funkce

Newtonovu metodu lze použít pro hledání lokálního extrému dvakrát diferencovatelné funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ tak, že v algoritmu (11.9) položíme $\mathbf{g}(\mathbf{x}) = f'(\mathbf{x})^T$. Tím dostaneme iteraci

$$\mathbf{x}_{k+1} = \mathbf{x}_k - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k)^T, \tag{11.10}$$

kde $f''(\mathbf{x}_k)$ je Hessova matice funkce f v bodě $\mathbf{x}_k$.

Význam iterace (11.9) byl takový, že se zobrazení $\mathbf{g}$ aproximovalo Taylorovým polynomem prvního stupně (tedy afinním zobrazením) a pak se našel kořen tohoto polynomu. Význam iterace (11.10) je takový, že se funkce f aproximuje Taylorovým polynomem druhého stupně (tedy kvadratickou funkcí) a pak se najde minimum této kvadratické funkce. Odvoďte podrobně, že tomu tak je!



Iteraci (11.10) lze napsat v obecnějším tvaru (11.1), kde

$$\mathbf{v}_k = -f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k)^T. \tag{11.11}$$

Výhodou tohoto zobecnění je možnost zvolit optimální (ne nutně jednotkovou) délku kroku pomocí jednorozměrné minimalizace (11.3). Algoritmu (11.10) s jednotkovou délkou kroku se pak říká **čistá** Newtonova metoda.

Vektoru (11.11) říkáme **Newtonův směr**. Vidíme, že se od gradientního směru (11.4) liší násobením hessovou maticí $f''(\mathbf{x}_k)$. Aby to byl sestupný směr, musí být

$$f'(\mathbf{x}_k)\,\mathbf{v}_k = -f'(\mathbf{x}_k) f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k)^T < 0.$$

Toto platí, když $f'(\mathbf{x}_k) \neq \mathbf{0}$ (tj. $\mathbf{x}_k$ není stacionární bod) a matice $f''(\mathbf{x}_k)$ je pozitivně definitní (neboť pak bude pozitivně definitní i její inverze, viz Cvičení 5.20).

11.4 Nelineární metoda nejmenších čtverců

Řešme přeurčenou soustavu rovnic $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ pro $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^m$ (tedy soustavu m rovnic s n neznámými) ve smyslu nejmenších čtverců. To vede na minimalizaci funkce

$$f(\mathbf{x}) = \|\mathbf{g}(\mathbf{x})\|_2^2 = \mathbf{g}(\mathbf{x})^T\mathbf{g}(\mathbf{x}) = \sum_{i=1}^{m} g_i(\mathbf{x})^2, \tag{11.12}$$

kde g_i jsou složky zobrazení $\mathbf{g}$. Speciálním případem je přibližné řešení lineární nehomogenní soustavy $\mathbf{Ax} = \mathbf{b}$, kde $\mathbf{g}(\mathbf{x}) = \mathbf{b} - \mathbf{Ax}$ (viz §6.1). Zde ovšem předpokládáme obecně nelineární zobrazení $\mathbf{g}$.

Zatímco v §11.2 a §11.3.2 bylo cílem minimalizovat *obecnou* funkci, zde chceme minimalizovat funkci ve speciálním tvaru (11.12). Nyní máme dvě možnosti. Buď můžeme nasadit na funkci (11.12) jednu z metod pro minimalizaci obecné funkce, k čemuž se vrátíme v §11.4.2. Nebo můžeme být chytřejší a využít speciálního tvaru funkce (11.12), což popíšeme v §11.4.1.

11.4.1 Gauss-Newtonova metoda

Aproximujme opět zobrazení $\mathbf{g}$ Taylorovým polynomem prvního stupně (11.8). Úloha (11.12) pak vyžaduje minimalizovat $\|\hat{\mathbf{g}}(\mathbf{x})\|_2^2$. To je úloha lineárních nejmenších čtverců, kterou již známe z §6.1. Vede na normální rovnici

$$\mathbf{g}'(\mathbf{x}_k)^T\mathbf{g}'(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k) = -\mathbf{g}'(\mathbf{x}_k)^T\mathbf{g}(\mathbf{x}_k).$$

Pokud má Jacobiho matice $\mathbf{g}'(\mathbf{x}_k)$ lineárně nezávislé sloupce (tedy hodnost n, viz §6.1), tuto rovnici můžeme vyřešit pseudoinverzí:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \underbrace{(\mathbf{g}'(\mathbf{x}_k)^T\mathbf{g}'(\mathbf{x}_k))^{-1}\mathbf{g}'(\mathbf{x}_k)^T}_{\mathbf{g}'(\mathbf{x}_k)^+}\mathbf{g}(\mathbf{x}_k) \tag{11.13}$$

Algoritmus (11.13) je znám jako **Gauss-Newtonova metoda**. Můžeme jej opět napsat obecněji ve tvaru (11.1) se směrem hledání

$$\begin{aligned}
\mathbf{v}_k &= -(\mathbf{g}'(\mathbf{x}_k)^T\mathbf{g}'(\mathbf{x}_k))^{-1}\mathbf{g}'(\mathbf{x}_k)^T\mathbf{g}(\mathbf{x}_k) && \text{(11.14a)}\\
&= -\mathbf{g}'(\mathbf{x}_k)^+\mathbf{g}(\mathbf{x}_k) && \text{(11.14b)}\\
&= -\tfrac{1}{2}(\mathbf{g}'(\mathbf{x}_k)^T\mathbf{g}'(\mathbf{x}_k))^{-1}f'(\mathbf{x}_k)^T. && \text{(11.14c)}
\end{aligned}$$

Pro $m = n$ máme $\mathbf{g}'(\mathbf{x}_k)^+ = \mathbf{g}'(\mathbf{x}_k)^{-1}$, tedy Gauss-Newtonova metoda se redukuje na Newtonovu metodu (11.9) na řešení soustavy n rovnic s n neznámými.

Tvar (11.14c) dostaneme z (11.14a) dosazením derivace účelové funkce $f'(\mathbf{x}) = 2\mathbf{g}(\mathbf{x})^T\mathbf{g}'(\mathbf{x})$ (viz §8.3.2). Vidíme, že Gauss-Newtonův směr (11.14c) se liší od gradientního směru (11.4) pouze násobením maticí $\frac{1}{2}(\mathbf{g}'(\mathbf{x}_k)^T\mathbf{g}'(\mathbf{x}_k))^{-1}$. Aby byl tento směr sestupný, musí být

$$f'(\mathbf{x}_k)\,\mathbf{v}_k = -\tfrac{1}{2}f'(\mathbf{x}_k)(\mathbf{g}'(\mathbf{x}_k)^T\mathbf{g}'(\mathbf{x}_k))^{-1}f'(\mathbf{x}_k)^T < 0.$$

Toto platí, když $f'(\mathbf{x}_k) \neq \mathbf{0}$ a matice $\mathbf{g}'(\mathbf{x}_k)^T\mathbf{g}'(\mathbf{x}_k)$ je pozitivně definitní (viz Cvičení 5.20). Matice $\mathbf{g}'(\mathbf{x}_k)^T\mathbf{g}'(\mathbf{x}_k)$ je pozitivně definitní právě tehdy, když $\mathbf{g}'(\mathbf{x}_k)$ má lineárně nezávislé sloupce (dokažte!), což ovšem již předpokládáme kvůli existenci inverze. Tedy vidíme, že za přirozených podmínek je Gauss-Newtonův směr vždy sestupný.

Čistá Gauss-Newtonova metoda (tj. s jednotkovou délkou kroku) může divergovat, a to i když je počáteční odhad $\mathbf{x}_0$ libovolně blízko lokálnímu minimu funkce (11.12). Protože ale Gauss-Newtonův směr je vždy sestupný, vhodnou volbou délky kroku α_k lze vždy zajistit konvergenci.

Example 11.5. V systému GPS máme m satelitů se známými souřadnicemi $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$ a chceme spočítat souřadnice pozorovatele $\mathbf{x} \in \mathbb{R}^n$ z naměřených vzdáleností $y_i = \|\mathbf{a}_i - \mathbf{x}\|_2$ pozorovatele od satelitů. Měření jsou zatížena chybou, proto obecně tato soustava rovnic nebude mít žádné řešení. Řešme tuto přeurčenou nelineární soustavu ve smyslu nejmenších čtverců, tedy minimalizujme funkci

$$f(\mathbf{x}) = \sum_{i=1}^{m} \big(\|\mathbf{x} - \mathbf{a}_i\|_2 - y_i\big)^2.$$

Máme tedy $\mathbf{g} = (g_1, \ldots, g_m)\colon \mathbb{R}^n \to \mathbb{R}^m$, kde $g_i(\mathbf{x}) = \|\mathbf{x} - \mathbf{a}_i\|_2 - y_i$. Derivace složek $\mathbf{g}$ je (pomůže nám §8.3.2, ale udělejte sami!) $g_i'(\mathbf{x}) = (\mathbf{x} - \mathbf{a}_i)^T / \|\mathbf{x} - \mathbf{a}_i\|_2$. Tedy

$$\mathbf{g}'(\mathbf{x}) = \begin{bmatrix} (\mathbf{x} - \mathbf{a}_1)^T / \|\mathbf{x} - \mathbf{a}_1\|_2 \\ \vdots \\ (\mathbf{x} - \mathbf{a}_m)^T / \|\mathbf{x} - \mathbf{a}_m\|_2 \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

Pak dosadíme do vzorečku (11.13). □

11.4.2 Rozdíl proti Newtonově metodě

Předpokládejme, že bychom optimalizovali naší účelovou funkci (11.12) přímo Newtonovou metodou z §11.3.2. Spočítejme (proveďte sami!) Hessián funkce (11.12):

$$f''(\mathbf{x}) = 2\mathbf{g}'(\mathbf{x})^T \mathbf{g}'(\mathbf{x}) + 2 \sum_{i=1}^{m} g_i(\mathbf{x}) g_i''(\mathbf{x}). \tag{11.15}$$

Hessián je součtem členu obsahujícího derivace prvního rádu a členu obsahujícího derivace druhého řádu. Vidíme, že Gauss-Newtonův směr (11.14c) se liší od Newtonova směru (11.11) zanedbáním členu druhého řádu v Hessiánu (11.15). Jinými slovy, Gauss-Newtonovu metodu je možno vnímat jako aproximaci Newtonovy metody na minimalizaci funkce (11.12) spočívající v tom, že skutečný Hessián (11.15) se aproximuje výrazem $2\mathbf{g}'(\mathbf{x})^T \mathbf{g}'(\mathbf{x})$.

To se projevuje tím, že Gauss-Newtonova metoda obvykle konverguje pomaleji než plná Newtonova metoda. Ovšem vyhnuli jsme se počítání druhých derivací funkce $\mathbf{g}$, což je hlavní výhoda Gauss-Newtonovy metody.

11.4.3 Levenberg-Marquardtova metoda

Levenberg-Marquardtova metoda je široce používané vylepšení Gauss-Newtonovy metody, které matici $\mathbf{g}'(\mathbf{x})^T \mathbf{g}'(\mathbf{x})$ v iteraci (11.13) nahrazuje maticí

$$\mathbf{g}'(\mathbf{x}_k)^T \mathbf{g}'(\mathbf{x}_k) + \mu_k \mathbf{I} \tag{11.16}$$

pro nějaké zvolené $\mu_k > 0$. Vidíme, že:

- Pro malé μ_k se Levenberg-Marquardtova iterace blíží Gauss-Newtonově iteraci.

- Pro velké μ_k je inverze matice (11.16) blízká $\mu_k^{-1}\mathbf{I}$, tedy Levenberg-Marquardtova iterace je blízká $\mathbf{x}_{k+1} = \mathbf{x}_k - \mu_k^{-1} f'(\mathbf{x}_k)^T$. Ale to je iterace gradientní metody s délkou kroku μ_k^{-1}.

Tím jsou spojeny výhody Gauss-Newtonovy metody (typicky rychlá konvergence v okolí optima) a gradientní metody (spolehlivost i daleko od optima). Volbou parametru μ_k spojitě přecházíme mezi oběma metodami.

Parametr μ_k měníme během algoritmu. Začneme např. s $\mu_0 = 10^3$ a pak v každé iteraci:

- Pokud iterace snížila účelovou funkci, iteraci přijmeme a μ_k zmenšíme.
- Pokud iterace nesnížila účelovou funkci, iteraci odmítneme a μ_k zvětšíme.

Zvětšování a zmenšování μ_k děláme násobením a dělením konstantou, např. 10. Všimněte si, toto nahrazuje optimalizaci délky kroku α_k (*line search*).

Na algoritmus lze pohlížet i jinak. V iteraci (11.13) se počítá inverze matice $\mathbf{g}'(\mathbf{x}_k)^T\mathbf{g}'(\mathbf{x}_k)$. Tato matice je sice vždy pozitivně semidefinitní, ale může být blízká singulární (kdy se to stane?). To neblaze ovlivní stabilitu algoritmu. Matice (11.16) je ale vždy pozitivně definitní (viz Cvičení 5.19), a tedy regulární.

11.4.4 Statistické odůvodnění kritéria nejmenších čtverců

Zde a dříve v §6.1 jsme ukázali metody na přibližné řešení přeurčených soustav rovnic ve smyslu nejmenších čtverců. Nyní podáme statistický důvod, odkud se kritérium nejmenších čtverců vzalo.

Odhadujme skryté parametry $\mathbf{x}$ nějakého systému z měření $\mathbf{y}$ na systému. Budiž vázány známou závislostí $\mathbf{y} = \mathbf{f}(\mathbf{x})$. Měření jsou zatížena chybami, které jsou způsobeny šumem senzorů, nepřesnostmi měření, nedokonalou znalostí modelu, apod. Tedy

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) + \mathbf{r}, \tag{11.17}$$

kde $\mathbf{r} = (r_1, \ldots, r_m)$ jsou náhodné proměnné modelující chyby měření $\mathbf{y} = (y_1, \ldots, y_m)$. Metoda nejmenších čtverců říká, že máme minimalizovat $\|\mathbf{r}\|_2^2 = \sum_{i=1}^m r_i^2$, ale neříká proč.

Důvod odvodíme statistickou úvahou. Metoda činí dva předpoklady:

- Náhodné proměnné r_i mají normální (neboli Gaussovo) rozdělení s nulovou střední hodnotou a směrodatnou odchylkou σ, s hustotou pravděpodobnosti

  $$p(r_i) = c\,\mathrm{e}^{-r_i^2/(2\sigma^2)},$$

  kde $c = \left(\sigma\sqrt{2\pi}\right)^{-1}$ je normalizační konstanta.
- Náhodné proměnné $r_1, \ldots, r_m$ jsou na sobě nezávislé. Tedy sdružená hustota pravděpodobnosti je rovna součinu

  $$p(\mathbf{r}) = p(r_1, \ldots, r_m) = \prod_{i=1}^m p(r_i) = \prod_{i=1}^m c\,\mathrm{e}^{-r_i^2/(2\sigma^2)}. \tag{11.18}$$

Dále použijeme *princip maxima věrohodnosti*. Ten říká, že parametry $\mathbf{x}$ se mají najít tak, aby $p(\mathbf{r}) = p(\mathbf{y} - \mathbf{f}(\mathbf{x}))$ bylo maximální. Je pohodlnější minimalizovat záporný logaritmus

$$-\log p(r_1, \ldots, r_m) = -\sum_{i=1}^m \log p(r_i) = \sum_{i=1}^m \left(\frac{r_i^2}{2\sigma^2} - \log c\right).$$

Jelikož σ je konstanta, je to totéž jako minimalizovat $\sum_i r_i^2$.

11.5 Cvičení

11.1. Najděte local extrém funkce $f(x, y) = x^2 - y + \sin(y^2 - 2x)$ čistou Newtonovou metodou. Počáteční odhad zvolte $(x_0, y_0) = (1, 1)$.

11.2. Máme m bodů v rovině o souřadnicích (x_i, y_i), $i = 1, \ldots, m$. Tyto body chceme proložit kružnicí ve smyslu nejmenších čtverců – tj. hledáme kružnici se středem (u, v) a poloměrem r takovou, aby součet čtverců kolmých vzdáleností bodů ke kružnici byl minimální. Zformulujte příslušnou optimalizační úlohu. Odvoďte iteraci Gauss-Newtonovy a Levenberg-Marquardtovy metody.

Chapter 12

Lineární programování

Lineární rovnicí rozumíme výrok $a_1x_1+\cdots+a_nx_n = b$, neboli $h(\mathbf{x}) = 0$ kde h je afinní funkce. **Lineární nerovnicí** rozumíme výrok $a_1x_1 + \cdots + a_nx_n \leq b$ či $a_1x_1 + \cdots + a_nx_n \geq b$, neboli $g(\mathbf{x}) \leq 0$ či $g(\mathbf{x}) \geq 0$ kde g je afinní funkce. Úloha **lineárního programování** (LP, také zvané lineární optimalizace) znamená minimalizaci lineární funkce za podmínek ve tvaru lineárních rovnic a nerovnic. Neboli v obecné formulaci (1.4) je funkce f lineární (tj. tvaru (3.4)) a funkce g_i, h_i jsou afinní (tj. tvaru (3.10)).

Stejně jako pro obecnou úlohu spojité optinalizace (viz §1.3), pro řešitelnost LP mohou nastat tři případy:

- úloha má (alespoň jedno) optimální řešení,
- úloha je *nepřípustná* (množina přípustných řešení je prázdná, omezení si odporují),
- úloha je *neomezená* (účelovou funkci lze za daných omezení libovolně zlepšovat).

Jednoduché úlohy lineárního programování lze řešit graficky.

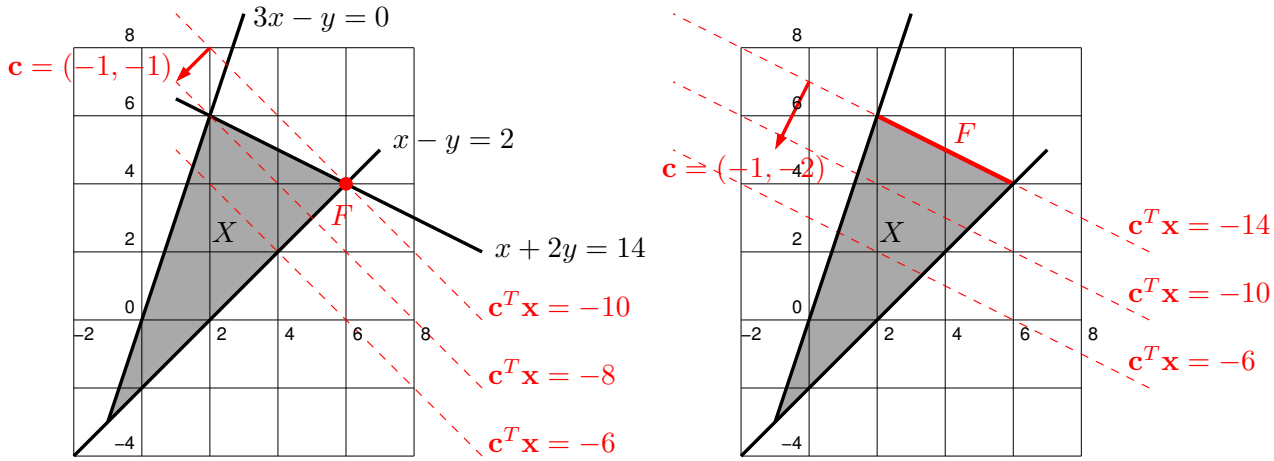
Example 12.1. Mějme lineární program

$$\begin{array}{rrcl} \min & -x - & y & \\ \text{za podmínek} & x + & 2y \leq & 14 \\ & 3x - & y \geq & 0 \\ & x - & y \leq & 2 \end{array} \tag{12.1}$$

Množina přípustných řešení této úlohy

$$X = \{\, (x, y) \in \mathbb{R}^2 \mid x + 2y \leq 14,\ 3x - y \geq 0,\ x - y \leq 2 \,\} \tag{12.2}$$

je průnik tří polorovin $\{\, (x, y) \mid x + 2y \leq 14 \,\}$, $\{\, (x, y) \mid 3x - y \geq 0 \,\}$ a $\{\, (x, y) \mid x - y \leq 2 \,\}$. Tuto množinu snadno nakreslíme:



Účelová funkce $-x - y$, neboli $\mathbf{c}^T\mathbf{x}$ pro $\mathbf{x} = (x, y)$ a $\mathbf{c} = (-1, -1)$, má vrstevnice kolmé k vektoru $\mathbf{c}$ a roste ve směru $\mathbf{c}$. Proto (viz levý obrázek) účelová funkce na množině X nabývá (globálního) minima v bodě $(x, y) = (6, 4)$. Úloha má tedy jediné optimální řešení.

Pokud bychom účelovou funkci úlohy (12.1) změnili na $-x - 2y$, bude tato funkce na množině X nabývat minima ve všech bodech úsečky spojující body $(2, 6)$ a $(6, 4)$ (viz pravý obrázek). Úloha má tedy nekonečně mnoho optimálních řešení. $\square$

12.1 Různé tvary úloh LP

Při zápisu úlohy LP je zvykem odděleně zapisovat obecná lineární omezení a omezení na znaménka jednotlivých proměnných. Obecnou úlohu LP tedy zapíšeme jako

$$\begin{array}{rrl}
\min & c_1x_1 + \cdots + c_nx_n & \\
\text{za podmínek} & a_{i1}x_1 + \cdots + a_{in}x_n \geq b_i, & i \in I_+ \\
& a_{i1}x_1 + \cdots + a_{in}x_n \leq b_i, & i \in I_- \\
& a_{i1}x_1 + \cdots + a_{in}x_n = b_i, & i \in I_0 \\
& x_j \geq 0\,, & j \in J_+ \\
& x_j \leq 0\,, & j \in J_- \\
& x_j \in \mathbb{R}, & j \in J_0
\end{array}$$

kde

$$\begin{aligned}
I &= \{1, \ldots, m\} = I_0 \cup I_+ \cup I_- \\
J &= \{1, \ldots, n\} \;= J_0 \cup J_+ \cup J_-
\end{aligned}$$

jsou rozklady indexových množin. Zápis $x_j \geq 0$ značí, že proměnná x_j může nabývat pouze nezáporných hodnot, zatímco $x_j \in \mathbb{R}$ značí, že x_j může nabývat libovolných hodnot.

Počítačové algoritmy na řešení LP často předpokládají úlohu v nějakém speciálním tvaru, kdy jsou dovoleny pouze jisté typy omezení. Nejčastěji užívané speciální tvary jsou:

- Dovolíme pouze omezení typu '=' a nezáporné proměnné ($I_+ = I_- = J_- = J_0 = \emptyset$), tj.

$$\begin{array}{rrl}
\min & c_1x_1 + \cdots + c_nx_n & \\
\text{za podmínek} & a_{i1}x_1 + \cdots + a_{in}x_n = b_i, & i = 1, \ldots, m \\
& x_j \geq 0\,, & j = 1, \ldots, n
\end{array}$$

To[1] lze psát maticově jako $\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \,\}$, kde $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{c} \in \mathbb{R}^n$.

- Tvar $\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{A}\mathbf{x} \geq \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \,\}$.
- Tvar $\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{A}\mathbf{x} \geq \mathbf{b} \,\}$.

Tyto speciální tvary nemají menší vyjadřovací schopnost než obecný tvar, neboť obecný tvar se dá převést na libovolný speciální tvar některou z následujících úprav:

- Rovnost $\mathbf{a}_i^T\mathbf{x} = b_i$ nahradíme dvěma nerovnostmi $\mathbf{a}_i^T\mathbf{x} \geq b_i$, $-\mathbf{a}_i^T\mathbf{x} \geq -b_i$.
- Nerovnost $a_{i1}x_1 + \cdots + a_{in}x_n \leq b_i$ převedeme na rovnost přidáním pomocné **slackové proměnné**[2] $u_i \geq 0$ jako $a_{i1}x_1 + \cdots + a_{in}x_n + u_i = b_i$.

  Podobně převedeme nerovnost $a_{i1}x_1 + \cdots + a_{in}x_n \geq b_i$ na rovnost (jak?).
- Proměnnou bez omezení $x_i \in \mathbb{R}$ rozdělíme na dvě nezáporné proměnné $x_i^+ \geq 0$, $x_i^- \geq 0$ přidáním podmínky $x_i = x_i^+ - x_i^-$.

Úloha získaná z původní úlohy pomocí těchto úprav je ekvivalentní původní úloze v tom smyslu, že hodnota jejich optima je stejná a argument optima původní úlohy lze 'snadno' získat z argumentu optima nové úlohy.

Example 12.2. V úloze (12.1) chceme první podmínku převést na rovnost. To uděláme zavedením slackové proměnné $u \geq 0$. Transformovaná úloha je

$$\begin{array}{rrcrcrcr}
\min & -x & - & y & & & & \\
\text{za podmínek} & x & + & 2y & + & u & = & 14 \\
 & 3x & - & y & & & \geq & 0 \\
 & x & - & y & & & \leq & 2 \\
 & & & & & u & \geq & 0
\end{array}$$

Je-li (x, y, u) optimum této úlohy, optimum úlohy (12.1) je (x, y). □

Example 12.3. V úloze (12.1) obě proměnné mohou mít libovolné znaménko. Chceme převést úlohu na tvar, kde všechny proměnné jsou nezáporné. Dosadíme $x = x_+ - x_-$ a $y = y_+ - y_-$, kde $x_+, x_-, y_+, y_- \geq 0$. Výsledná úloha je

$$\begin{array}{rrcrcrcrcr}
\min & -x_+ & + & x_- & - & y_+ & + & y_- & & \\
\text{za podmínek} & x_+ & - & x_- & + & 2y_+ & - & 2y_- & \leq & 14 \\
 & 3x_+ & - & 3x_- & - & y_+ & + & y_- & \geq & 0 \\
 & x_+ & - & x_- & - & y_+ & + & y_- & \leq & 2 \\
 & & & & & & & x_+, x_-, y_+, y_- & \geq & 0
\end{array}$$

□

[1] Tomuto tvaru se někdy říká *standardní*. Bohužel názvosloví různých tvarů LP není jednotné, názvy jako 'standardní tvar', 'základní tvar' či 'kanonický tvar' tedy mohou znamenat v různých knihách něco jiného.

[2] *Slack* znamená anglicky např. mezeru mezi zdí a skříní, která není zcela přiražená ke zdi. Termín *slack variable* nemá ustálený český ekvivalent, někdy se překládá jako *skluzová proměnná*.

12.1.1 Po částech afinní funkce

Někdy je možné převést na LP některé úlohy, které jako LP na první pohled nevypadají. K tomu uvedeme dvě jednoduché skutečnosti.

Za prvé, pro každou množinu X a funkci $f\colon X \to \mathbb{R}$ platí

$$\min_{x \in X} f(x) = \min\{\, z \mid f(x) \le z,\ x \in X,\ z \in \mathbb{R}\,\}, \tag{12.3}$$

Důkaz: v optimu pravé úlohy je $f(x) = z$, protože kdyby bylo $f(x) < z$, mohli bychom z zmenšit bez porušení omezení a tedy (x, z) by nebylo optimální.

Za druhé, pro libovolná čísla $a_1, \ldots, a_k, b$ platí ekvivalence

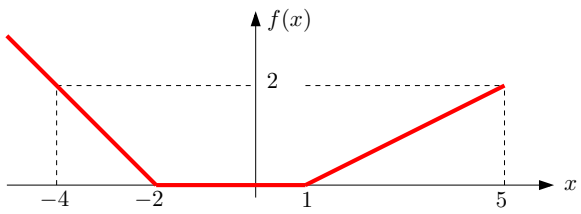
$$(\max_{i=1}^k a_i \le b) \iff (\forall i = 1, \ldots, k)(a_i \le b) \tag{12.4a}$$

$$(\min_{i=1}^k a_i \ge b) \iff (\forall i = 1, \ldots, k)(a_i \ge b). \tag{12.4b}$$

Mějme nyní funkci $f\colon \mathbb{R}^n \to \mathbb{R}$ danou vzorcem

$$f(\mathbf{x}) = \max_{i=1}^{k} (\mathbf{c}_i^T \mathbf{x} + d_i), \tag{12.5}$$

kde $\mathbf{c}_i \in \mathbb{R}^n$ a $d_i \in \mathbb{R}$ jsou dány. Tato funkce není lineární ani afinní, je po částech afinní (viz Cvičení 12.5). Příkladem pro $n = 1$ a $k = 3$ je funkce $f(x) = \max\{-x - 1,\ 0,\ \frac{1}{2}(x - 1)\}$, jejíž graf je na obrázku:



Řešme úlohu

$$\min\{\, f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{A}\mathbf{x} \ge \mathbf{b}\,\}. \tag{12.6}$$

To není úloha LP, neboť její účelová funkce není lineární. Ale s použitím (12.3) a (12.4) máme

$$\begin{aligned}
\min\{\, f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{A}\mathbf{x} \ge \mathbf{b}\,\} &= \min\{\, z \mid (\mathbf{x}, z) \in \mathbb{R}^{n+1},\ f(\mathbf{x}) \le z,\ \mathbf{A}\mathbf{x} \ge \mathbf{b}\,\} \\
&= \min\{\, z \mid (\mathbf{x}, z) \in \mathbb{R}^{n+1},\ \max_i (\mathbf{c}_i^T \mathbf{x} + d_i) \le z,\ \mathbf{A}\mathbf{x} \ge \mathbf{b}\,\} \\
&= \min\{\, z \mid (\mathbf{x}, z) \in \mathbb{R}^{n+1},\ \mathbf{c}_i^T \mathbf{x} + d_i \le z\ (\forall i),\ \mathbf{A}\mathbf{x} \ge \mathbf{b}\,\}.
\end{aligned}$$

Tedy jsme úlohu převedli na LP.

Example 12.4. Úloha

$$\begin{array}{rl}
\min & \max\{\, x_1 + 2x_2,\ 2x_1 + x_2\,\} \\
\text{za podm.} & x_1 + 2x_2 \le 14 \\
& 3x_1 - x_2 \ge 0 \\
& x_1 - x_2 \le 2
\end{array}$$

není LP, protože účelová funkce $f(x_1, x_2) = \max\{\, x_1 + 2x_2,\ 2x_1 + x_2 \,\}$ není lineární ani afinní (nakreslete si na papír její vrstevnice!). Úlohu lze ale přeformulovat na LP

$$\begin{aligned} \min \quad & z \\ \text{za podm.} \quad & 3x_1 + 4x_2 \le z \\ & 2x_1 - 3x_2 \le z \\ & x_1 + 2x_2 \le 14 \\ & 3x_1 - x_2 \ge 0 \\ & x_1 - x_2 \le 2 \end{aligned}$$

□

Skutečnost (12.4) lze také použít na některé úlohy, které mají maximum či minimum v omezeních. Např. je

$$\min\{\, \mathbf{c}^T\mathbf{x} \mid \max_i \mathbf{d}_i^T\mathbf{x} \le e,\ \mathbf{A}\mathbf{x} \ge \mathbf{b} \,\} = \min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{d}_i^T\mathbf{x} \le e\ (\forall i),\ \mathbf{A}\mathbf{x} \ge \mathbf{b} \,\}.$$

Example 12.5. Platí rovnost

$$\min\{\, x - y \mid x \ge 0,\ y \ge 0,\ \max\{x, y\} \le 1 \,\} = \min\{\, x - y \mid x \ge 0,\ y \ge 0,\ x \le 1,\ y \le 1 \,\}$$

protože $\max\{x, y\} \le 1$ je ekvivalentní $x \le 1$, $y \le 1$. Úloha vlevo není LP, úloha vpravo ano. □

Tento převod lze užít i pro funkce obsahující absolutní hodnoty, neboť $|x| = \max\{-x, x\}$. Lze ho dále použít i na minima, podle druhé formy (12.4). Je ale nutná opatrnost: neplatí nic takového jako $(\min_i a_i \ge b) \iff (\forall i)(a_i \ge b)$, tedy máme-li špatnou kombinaci minim/maxim a nerovností, převod na LP není možný.

12.2 Některé aplikace LP

12.2.1 Optimální výrobní program

Z m druhů surovin vyrábíme n druhů výrobků.

- a_{ij} = množství suroviny druhu i potřebné na výrobu výrobku druhu j
- b_i = množství suroviny druhu i, které máme k dispozici
- c_j = zisk z vyrobení jednoho výrobku druhu j
- x_j = počet vyrobených výrobků druhu j

Úkolem je zjistit, kolik jakých výrobků máme vyrobit, abychom dosáhli největšího zisku. Řešení:

$$\max\left\{ \sum_{j=1}^{n} c_j x_j \;\middle|\; \sum_{j=1}^{n} a_{ij} x_j \le b_i,\ x_j \ge 0 \right\}. \tag{12.7}$$

Example 12.6. Pán u stánku prodává lupínky za 120 Kč/kg a hranolky za 76 Kč/kg. Na výrobu 1 kg lupínků se spotřebuje 2 kg brambor a 0.4 kg oleje. Na výrobu 1 kg hranolku se spotřebuje 1.5 kg brambor a 0.2 kg oleje. Je nakoupeno 100 kg brambor a 16 kg oleje. Brambory

stály 12 Kč/kg, olej 40 Kč/kg. Kolik má pán vyrobit lupínků a kolik hranolků, aby co nejvíce vydělal? To lze vyjádřit jako LP

$$\begin{aligned} \max \quad & 120l + 76h \\ \text{za podmínek} \quad & 2l + 1.5h \leq 100 \\ & 0.4l + 0.2h \leq 16 \\ & l, h \geq 0 \end{aligned}$$

Přitom předpokládáme, že zbytky surovin se po pracovní době vyhodí. Pokud se zbytky využijí, tak maximalizujeme $(120 - 24 - 16)l + (76 - 18 - 8)h = 80l + 50h$.

V obou případech je optimální řešení $l = 20$ kg lupínků a $h = 40$ kg hranolků. □

12.2.2 Směšovací (dietní) problém

Z n druhů surovin, z nichž každá je směsí m druhů látek, máme namíchat konečný produkt o požadovaném složení tak, aby cena surovin byla minimální.

- a_{ij} = množství látky druhu i obsažené v jednotkovém množství suroviny druhu j
- b_i = nejmenší požadované množství látky druhu i v konečném produktu
- c_j = jednotková cena suroviny druhu j
- x_j = množství suroviny druhu j

Řešení:

$$\min \left\{ \sum_{j=1}^{n} c_j x_j \;\middle|\; \sum_{j=1}^{n} a_{ij} x_j \geq b_i,\ x_j \geq 0 \right\}. \tag{12.8}$$

Example 12.7. Jste kuchařka v menze a chcete uvařit pro studenty co nejlevnější oběd, ve kterém ovšem kvůli předpisům musí být dané minimální množství živin (cukrů, bílkovin a vitamínů). Oběd vaříte ze tří surovin: brambor, masa a zeleniny. Jsou dány hodnoty v tabulce:

	na jednotku brambor	na jednotku masa	na jednotku zeleniny	min. požadavek na jeden oběd
obsah cukrů	2	1	1	8
obsah bílkovin	2	6	1	16
obsah vitamínů	1	3	6	8
cena	25	50	80	

Kolik je třeba každé suroviny na jeden oběd?

Minimalizujeme $25b+50m+80z$ za podmínek $2b+m+z \geq 8$, $2b+6m+z \geq 16$, $b+3m+6z \geq 8$ a $b, m, z \geq 0$. Optimální řešení je $b = 3.2$, $m = 1.6$, $z = 0$ s hodnotou 160. □

12.2.3 Dopravní problém

Máme m výrobců a n spotřebitelů.

- a_i = množství zboží vyráběné výrobcem i
- b_j = množství zboží požadované spotřebitelem j
- c_{ij} = cena dopravy jednotky zboží od výrobce i ke spotřebiteli j
- x_{ij} = množství zboží vezené od výrobce i ke spotřebiteli j

Chceme co nejlevněji rozvézt zboží od výrobců ke spotřebitelům. Řešení:

$$\min\left\{ \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} \;\middle|\; \sum_{j=1}^{n} x_{ij} = a_i,\ \sum_{i=1}^{m} x_{ij} = b_j,\ x_{ij} \geq 0 \right\}. \tag{12.9}$$

Zadání musí splňovat $\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j$ (nabídka musí být rovna poptávce), jinak bude úloha nepřípustná. Úloha jde modifikovat tak, že dovolíme $\sum_{i=1}^{m} a_i \geq \sum_{j=1}^{n} b_j$ (proveďte!).

12.2.4 Distribuční problém

Máme m strojů a n druhů výrobků.

- a_i = počet hodin, který je k dispozici na stroji i
- b_j = požadované množství výrobku druhu j
- c_{ij} = cena jedné hodiny práce stroje i na výrobku typu j
- k_{ij} = hodinový výkon stroje i při výrobě výrobku druhu j
- x_{ij} = počet hodin, po který bude stroj i vyrábět výrobek druhu j

Pro každý ze strojů máme určit, kolik výrobků se na něm bude vyrábět. Řešení:

$$\min\left\{ \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} \;\middle|\; \sum_{j=1}^{n} x_{ij} \leq a_i,\ \sum_{i=1}^{m} k_{ij}x_{ij} = b_j,\ x_{ij} \geq 0 \right\}. \tag{12.10}$$

12.3 Použití na nehomogenní lineární soustavy

12.3.1 Vektorové normy

Norma formalizuje pojem 'délky' vektoru $\mathbf{x}$.

Definition 12.1. *Funkce* $\|\cdot\|\colon \mathbb{R}^n \to \mathbb{R}$ *se nazývá vektorová*[3] **norma**, *jestliže splňuje tyto axiomy:*

1. *Jestliže* $\|\mathbf{x}\| = 0$ *pak* $\mathbf{x} = \mathbf{0}$.
2. $\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|$ *pro každé* $\alpha \in \mathbb{R}$ *a* $\mathbf{x} \in \mathbb{R}^n$ *(norma je kladně homogenní).*
3. $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ *pro každé* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ *(trojúhelníková nerovnost).*

Z axiomů plynou tyto další vlastnosti normy:

- $\|\mathbf{0}\| = 0$, což plyne z homogenity pro $\alpha = 0$
- $\|\mathbf{x}\| \geq 0$ pro každé $\mathbf{x} \in \mathbb{R}^n$. To jde odvodit tak, že v trojúhelníkové nerovnosti položíme $\mathbf{y} = -\mathbf{x}$, což dá
$$\|\mathbf{x}-\mathbf{x}\| = \|\mathbf{0}\| = 0 \leq \|\mathbf{x}\| + \|-\mathbf{x}\| = 2\|\mathbf{x}\|,$$
kde na pravé straně jsme použili homogenitu.

[3] Existují i *maticové* normy, což jsou funkce $\mathbb{R}^{m\times n} \to \mathbb{R}$.

Jednotková sféra normy je množina $\{\, \mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\| = 1 \,\}$, tedy vrstevnice normy jednotkové výšky. Díky homogenitě je jednotková sféra středově symetrická a její tvar zcela určuje normu.

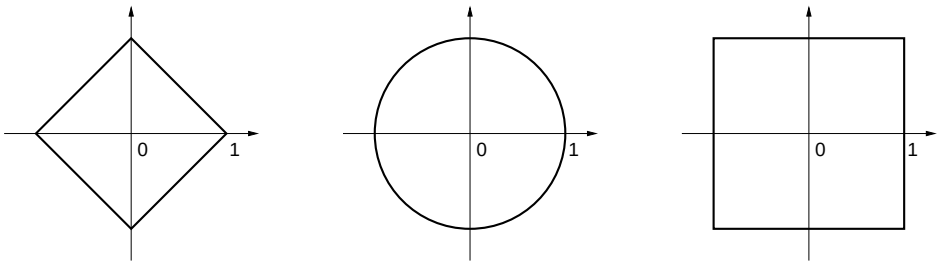
Uveďme příklady norem. Základním příkladem je p**-norma**

$$\|\mathbf{x}\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}.$$

Musí být $p \geq 1$, jinak neplatí trojúhelníková nerovnost. Nejčastěji narazíte na:

- $\|\mathbf{x}\|_1 = |x_1| + \cdots + |x_n|$. Někdy se jí říká *manhattanská norma*, protože v systému pravoúhlých ulic je vzdálenost mezi body $\mathbf{x}$ a $\mathbf{y}$ rovna $\|\mathbf{x} - \mathbf{y}\|_1$.
- $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2} = \sqrt{\mathbf{x}^T\mathbf{x}}$. Je to známá *eukleidovská norma*.
- $\|\mathbf{x}\|_\infty = \lim_{p\to\infty} \|\mathbf{x}\|_p = \max\{|x_1|, \ldots, |x_n|\}$ (dokažte rovnost výpočtem limity!). Někdy se jí říká *Čebyševova norma* nebo *max-norma*.

Jednotkové sféry těchto norem v $\mathbb{R}^2$ vypadají takto:



$\|\mathbf{x}\|_1 = 1$ $\qquad$ $\|\mathbf{x}\|_2 = 1$ $\qquad$ $\|\mathbf{x}\|_\infty = 1$

Existují ale i normy, které nejsou p-normy, např.

- $\|\mathbf{x}\| = 2|x_1| + \sqrt{x_2^2 + x_3^2} + \max\{|x_4|, |x_5|\}$ je norma na $\mathbb{R}^5$.
- Je-li $\|\mathbf{x}\|$ norma a $\mathbf{A}$ je čtvercová nebo úzká matice s plnou hodností, je také $\|\mathbf{A}\mathbf{x}\|$ norma.

12.3.2 Přibližné řešení přeurčených soustav

Mějme přeurčenou lineární soustavu $\mathbf{A}\mathbf{x} = \mathbf{b}$, kde $\mathbf{A} \in \mathbb{R}^{m\times n}$ a $\mathbf{0} \neq \mathbf{b} \in \mathbb{R}^m$. Nalezení jejího přibližného řešení formulujme jako úlohu

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p. \tag{12.11}$$

Uvažujme tři případy:

- Pro $p = \infty$ hledáme takové $\mathbf{x}$, které minimalizuje výraz

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_\infty = \max_{i=1}^{m} |\mathbf{a}_i^T\mathbf{x} - b_i|, \tag{12.12}$$

  tedy minimalizuje maximální residuum. Toto řešení je známé pod názvem *minimaxní* nebo *Čebyševovo*. Úloha je ekvivalentní lineárnímu programu

$$\begin{array}{rll} \min & z & \\ \text{za podm.} & \mathbf{a}_i^T\mathbf{x} - b_i \leq z, & i = 1, \ldots, m \\ & -\mathbf{a}_i^T\mathbf{x} + b_i \leq z, & i = 1, \ldots, m \end{array}$$

  který lze zapsat elegantněji jako

$$\min\{\, z \in \mathbb{R} \mid \mathbf{x} \in \mathbb{R}^n,\ -z\mathbf{1} \leq \mathbf{A}\mathbf{x} - \mathbf{b} \leq z\mathbf{1} \,\}. \tag{12.13}$$

- Pro $p = 2$ dostaneme řešení ve smyslu nejmenších čtverců, které jsme odvodili v §6.1.
- Pro $p = 1$ hledáme takové $\mathbf{x}$, které minimalizuje výraz

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1 = \sum_{i=1}^{m} |\mathbf{a}_i^T \mathbf{x} - b_i|, \tag{12.14}$$

kde $\mathbf{a}_1, \ldots, \mathbf{a}_m$ jsou řádky matice $\mathbf{A}$. Úloha je ekvivalentní lineárnímu programu

$$\begin{array}{rll} \min & \displaystyle\sum_{i=1}^{m} z_i & \\ \text{za podm.} & \mathbf{a}_i^T \mathbf{x} - b_i \le z_i, & i = 1, \ldots, m \\ & -\mathbf{a}_i^T \mathbf{x} + b_i \le z_i, & i = 1, \ldots, m \end{array}$$

který lze zapsat elegantněji v maticovém tvaru jako

$$\min\{\, \mathbf{1}^T \mathbf{z} \mid \mathbf{z} \in \mathbb{R}^m,\ \mathbf{x} \in \mathbb{R}^n,\ -\mathbf{z} \le \mathbf{A}\mathbf{x} - \mathbf{b} \le \mathbf{z} \,\}. \tag{12.15}$$

12.3.3 Lineární regrese

Vraťme se k lineární regresi z §6.1.4 (znovu přečtěte!). Funkční závislost přibližně popsanou naměřenými dvojicemi (t_i, y_i), $i = 1, \ldots, m$, jsme aproximovali regresní funkcí

$$f(t, \mathbf{x}) = x_1 \varphi_1(t) + \cdots + x_n \varphi_n(t) = \boldsymbol{\varphi}(t)^T \mathbf{x},$$

kde parametry $\mathbf{x}$ jsou takové, aby $y_i \approx f(t_i, \mathbf{x})$ pro všechna i. Přibližné rovnosti $\approx$ jsme chápali ve smyslu nejmenších čtverců, tedy hledali jsme takové $\mathbf{x}$ které minimalizovalo funkci

$$\sum_{i=1}^{m} (y_i - f(t_i, \mathbf{x}))^2 = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2, \tag{12.16}$$

kde $\mathbf{y} = (y_1, \ldots, y_m)$ a prvky matice $\mathbf{A}$ jsou $a_{ij} = \varphi_j(t_i)$. Tedy řešíme úlohu (12.11) pro $p = 2$. Můžeme ale použít i jiné normy než eukleidovskou. Pro $p = 1$ minimalizujeme

$$\sum_{i=1}^{m} |y_i - f(t_i, \mathbf{x})| = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_1 \tag{12.17}$$

a pro $p = \infty$ minimalizujeme

$$\max_{i=1}^{m} |y_i - f(t_i, \mathbf{x})| = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_\infty. \tag{12.18}$$

Dále ukážeme, k čemu to může být dobré.

Regrese ve smyslu ∞-normy je vhodná např. při aproximaci funkcí.

Example 12.8. Na počítači bez matematického koprocesoru potřebujeme mnohokrát vyhodnocovat funkci sinus na intervalu $[0, \frac{\pi}{2}]$. Protože přesné vyhodnocení této funkce by trvalo příliš dlouho, chceme ji aproximovat polynomem třetího stupně $x_1 + x_2 t + x_3 t^2 + x_4 t^3$, který se vyhodnotí mnohem rychleji. Spočítejme hodnoty $y_i = \sin t_i$ funkce v dostatečném počtu bodů $t_i = \frac{\pi i}{2n}$ pro $i = 1, \ldots, m$. Koeficienty polynomu je vhodné hledat minimalizací Čebyševova kritéria (12.18), neboť to nám dá záruku, že chyba aproximace nikde nepřesáhne hodnotu, která je nejmenší možná pro daný stupeň polynomu. □

Regrese ve smyslu 1-normy je užitečná tehdy, když je malá část hodnot y_i naměřená úplně špatně (např. se někdo při zapisování čísel spletl v desetinné čárce). Takovým hodnotám se říká **vychýlené hodnoty** (*outliers*). Disciplína zabývající se modelováním funkčních závislostí za přítomnosti vychýlených hodnot se nazývá **robustní regrese**. V tomto případě řešení ve smyslu nejmenších čtverců není vhodné (není 'robustní'), protože i jediný vychýlený bod velmi ovlivní řešení. Regrese ve smyslu 1-normy je proti vychýleným bodům odolnější.

Ukážeme to na nejjednodušším možném případu regrese: odhad hodnoty jediného čísla ze souboru jeho nepřesných měření. Pro daná čísla $\mathbf{y} = (y_1, \dots, y_m) \in \mathbb{R}^m$ hledáme $x \in \mathbb{R}$ minimalizující funkci

$$f(x) = \|(x - y_1, \dots, x - y_m)\|_p = \|\mathbf{1}x - \mathbf{y}\|_p. \tag{12.19}$$

- Pro $p = \infty$ je $f(x) = \max_{i=1}^m |x - y_i|$. Řešením je $x = \frac{1}{2}\big(\min_{i=1}^m y_i + \max_{i=1}^m y_i\big)$, tedy bod v polovině mezi krajními body.
- Pro $p = 2$ je $f(x) = \sqrt{\sum_{i=1}^m (x - y_i)^2}$. Řešením je aritmetický průměr, $x = \frac{1}{m}\sum_{i=1}^m y_i$ (viz Příklad 6.4).
- Pro $p = 1$ je $f(x) = \sum_{i=1}^m |x - y_i|$. Řešením je *medián* z čísel y_i (dokažte!). Medián se vypočte tak, že seřadíme čísla y_i podle velikosti a vezmeme prostřední z nich. Pokud je m sudé, máme dva 'prostřední prvky' a v tom případě funkce f nabývá minima v jejich libovolné konvexní kombinaci. Je pak úzus definovat medián jako aritmetický průměr prostředních prvků.

Předpokládejme nyní, že jedno z čísel, např. y_1, se zvětšuje. V tom případě se řešení x pro různá p budou chovat různě. Např. aritmetický průměr se bude zvětšovat, a to tak, že zvětšováním hodnoty y_1 dosáhneme *libovolné* hodnoty x. Pro medián to ovšem neplatí – zvětšováním jediného bodu y_1 ovlivníme x jen natolik, nakolik to změní pořadí bodů. Jeho libovolným zvětšováním nedosáhneme libovolné hodnoty x.

Example 12.9. Šuplérou změříme průměr ocelové kuličky v několika místech, dostaneme hodnoty $\mathbf{y} = (1.02,\ 1.04,\ 0.99,\ 2.03)$ (cm). Při posledním měření jsme se na stupnici přehlédli, proto je poslední hodnota úplně špatně. Z těchto měření chceme odhadnout skutečný průměr. Máme

$$\frac{1}{2}\big(\min_{i=1}^m y_i + \max_{i=1}^m y_i\big) = 1.51, \qquad \frac{1}{m}\sum_{i=1}^m y_i = 1.27, \qquad \operatorname{median}_{i=1}^m y_i = 1.03.$$

Je zjevné, že medián je neovlivněn vychýleným bodem, zatímco ostatní odhady ano. □

Ve složitějším případě, např. prokládání dat polynomem jako v Příkladu 6.4, se nedá robustnost rešení ve smyslu 1-normy takto jednoduše formálně ukázat a analýza může být mnohem těžší. Ale intuitivně bude situace obdobná: řešení ve smyslu 1-normy bude méně citlivé na vychýlené body než řešení ve smyslu 2-normy.

12.4 Cvičení

12.1. Najděte graficky množinu optimálních řešení úlohy

$$\begin{aligned} \min \quad & c_1x_1 + c_2x_2 + c_3x_3 \\ \text{za podm.} \quad & x_1 + x_2 \geq 1 \\ & x_1 + 2x_2 \leq 3 \\ & x_1 + x_2 \leq 10 \\ & x_1, x_2, x_3 \geq 0 \end{aligned}$$

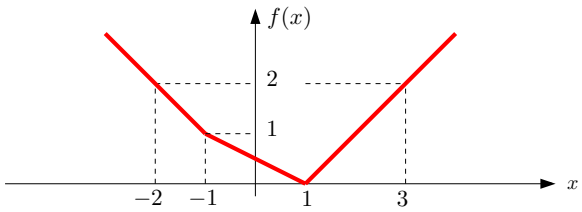
pro následující případy: (a) $\mathbf{c} = (-1, 0, 1)$, (b) $\mathbf{c} = (0, 1, 0)$, (c) $\mathbf{c} = (0, 0, -1)$.

12.2. Vyřešte úvahou tyto jednoduché úlohy LP a napište co nejjednodušší vzorec pro optimální hodnotu. Vektor $\mathbf{c} \in \mathbb{R}^n$ a číslo $k \in \mathbb{N}$, $1 \le k \le n$, jsou dány.

a) $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{0} \le \mathbf{x} \le \mathbf{1} \,\}$
b) $\max\{\, \mathbf{c}^T\mathbf{x} \mid -\mathbf{1} \le \mathbf{x} \le \mathbf{1} \,\}$
c) $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \ge \mathbf{0},\ \mathbf{1}^T\mathbf{x} = 1 \,\}$
d) $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \ge \mathbf{0},\ \mathbf{1}^T\mathbf{x} \le 1 \,\}$
e) $\max\{\, \mathbf{c}^T\mathbf{x} \mid -1 \le \mathbf{1}^T\mathbf{x} \le 1 \,\}$
f) $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \ge \mathbf{0},\ \mathbf{1}^T\mathbf{x} = k \,\}$
g) $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{0} \le \mathbf{x} \le \mathbf{1},\ \mathbf{1}^T\mathbf{x} = k \,\}$
h) $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{0} \le \mathbf{x} \le \mathbf{1},\ \mathbf{1}^T\mathbf{x} \le k \,\}$
i) $\max\{\, \mathbf{c}^T\mathbf{x} \mid 0 \le x_1 \le x_2 \le \cdots \le x_n \le 1 \,\}$
j) $(\star)$ $\max\{\, \mathbf{c}^T\mathbf{x} \mid -\mathbf{y} \le \mathbf{x} \le \mathbf{y},\ \mathbf{1}^T\mathbf{y} = k,\ \mathbf{y} \le \mathbf{1} \,\}$

12.3. Převeďte na LP nebo odůvodněte, proč to nejde. Proměnné jsou vždy $\mathbf{x}$ či x_i.

a) $\max\{\, |x_1 - c_1| + \cdots + |x_n - c_n| \mid a_1x_1 + \cdots + a_nx_n \ge b \,\}$
b) $\min\{\, |x_1| + |x_2| \mid 2x_1 - x_2 \ge 1,\ -x_1 + 2x_2 \ge 1 \,\}$
c) $\max\{\, |\mathbf{c}^T\mathbf{x}| \mid \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \ge \mathbf{0} \,\}$
d) $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b},\ |\mathbf{d}^T\mathbf{x}| \le 1,\ \mathbf{x} \ge \mathbf{0} \,\}$
e) $\min_{\mathbf{x}\in\mathbb{R}^n} \sum_{l=1}^{L} \max_{k=1}^{K} (\mathbf{c}_{kl}^T\mathbf{x} + d_{kl})$
f) $\min_{\mathbf{x}\in\mathbb{R}^n} \sum_{i=1}^{m} f(\mathbf{a}_i^T\mathbf{x} - b_i)$, kde funkce f je definována obrázkem



g) $\min\{\, \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1 \mid \mathbf{x} \in \mathbb{R}^n,\ \|\mathbf{x}\|_\infty \le 1 \,\}$
h) $\min\{\, \|\mathbf{x}\|_1 \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{A}\mathbf{x} = \mathbf{b} \,\}$
i) $\min\{\, \|\mathbf{x}\|_1 \mid \mathbf{x} \in \mathbb{R}^n,\ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_\infty \le 1 \,\}$
j) $\min_{\mathbf{x}\in\mathbb{R}^n} (\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1 + \|\mathbf{x}\|_\infty)$

12.4. Dokažte nebo vyvraťte následující rovnosti. Zde $\mathbf{c} \in \mathbb{R}^n$ a $\mathbf{A} \in \mathbb{R}^{m\times n}$ jsou dány, $\|\cdot\|$ je libovolná norma, a optimalizuje se přes proměnné $\mathbf{x} \in \mathbb{R}^n$.

a) $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \|\mathbf{x}\| = 1 \,\} = \max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \|\mathbf{x}\| \le 1 \,\}$
b) $\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \|\mathbf{x}\| = 1 \,\} = \min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \|\mathbf{x}\| \le 1 \,\}$
c) $\max\{\, \|\mathbf{A}\mathbf{x}\| \mid \mathbf{x} \in \mathbb{R}^n,\ \|\mathbf{x}\| = 1 \,\} = \max\{\, \|\mathbf{A}\mathbf{x}\| \mid \mathbf{x} \in \mathbb{R}^n,\ \|\mathbf{x}\| \le 1 \,\}$

Nápověda: Inspirujte se úvahou v §12.1.1.

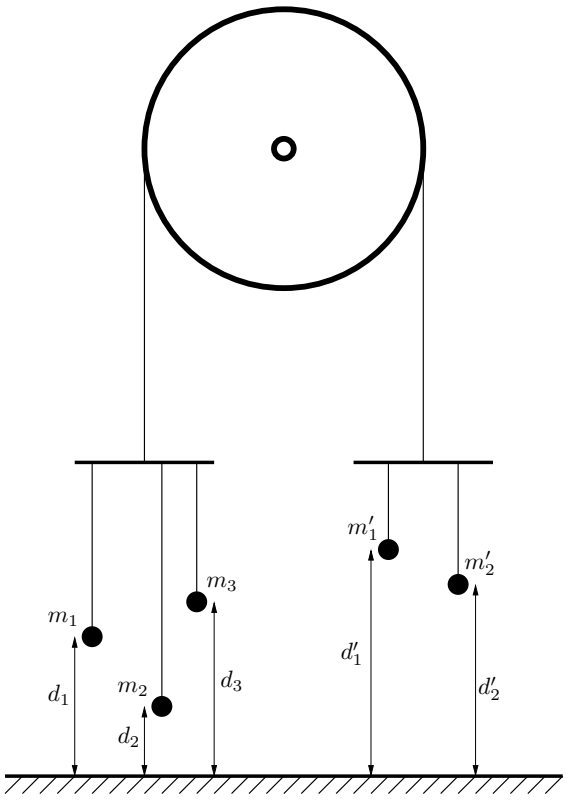
12.5. Pochopte kód v Matlabu, který nakreslí graf funkce $f(\mathbf{x}) = \max_{i=1}^{k} (\mathbf{c}_i^T\mathbf{x} + d_i)$ pro $\mathbf{x} \in \mathbb{R}^2$:

```
k = 200; N = 40;
cd = randn(3,k);
```

```
x1 = ones(N,1)*linspace(-1,1,N); x2 = linspace(-1,1,N)'*ones(1,N);
x = [x1(:)'; x2(:)']; x(3,:) = 1;
meshc(x1,x2,reshape(max(cd'*x,[],1),[N N])); axis vis3d
```

12.6. Hledáme největší hyperkouli $B(\mathbf{a}, r) = \{\, \mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}-\mathbf{a}\|_2 \leq r \,\}$, která se vejde do polyedru $P = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \leq \mathbf{b} \,\}$. Tedy hledáme maximální r za podmínky $B(\mathbf{a}, r) \subseteq P$, kde optimalizujeme přes proměnné $(\mathbf{a}, r)$. Vyjádřete jako LP.

12.7. Máme kladku s provazem, jehož oba konce končí hákem. Na levém háku visí n závaží na provázcích, přičemž i-té závaží má tíhu m_i a jeho výška nad zemí je d_i, pro $i = 1, \ldots, n$. Na pravém háku visí n' závaží na provázcích, přičemž i-té závaží má tíhu m'_i a jeho výška nad zemí je d'_i, pro $i = 1, \ldots, n'$. Výšky d_i a d'_i se měří v poloze, kdy jsou oba háky ve stejné výšce nad zemí. Kladka se pohybuje bez tření, provaz a provázky jsou nekonečně ohebné, provázky a háky mají nulovou hmotnost. Obrázek ukazuje příklad pro $n = 3$, $n' = 2$.



Soustava má jediný stupeň volnosti daný otáčením kladky. Označme jako x výšku levého háku nad bodem, kdy jsou oba háky ve stejné výšce – tedy pro $x = 0$ jsou oba háky ve stejné výšce a pro $x > 0$ bude levý hák o $2x$ výše než pravý hák. V závislosti na x každé závaží buď visí nad zemí (pak je jeho potenciální energie rovna m_i krát výška nad zemí) nebo leží na zemi (pak je jeho potenciální energie nulová). Soustava bude v rovnováze při minimální celkové potenciální energii.

a) Napište vzorec pro celkovou potenciální energii soustavy jako funkci x.
b) Napište lineární program, jehož optimum je rovno minimální potenciální energii soustavy. Není-li to možné, vysvětlete.

12.8. Veverka před zimou potřebuje přerovnat zásoby oříšků. Stávající zásoby má v m jamkách, přičemž i-tá jamka má souřadnice $\mathbf{p}_i \in \mathbb{R}^2$ a je v ní a_i oříšků. Potřebuje je přenosit do n nových připravených jamek, přičemž j-tá jamka má souřadnice $\mathbf{q}_j \in \mathbb{R}^2$ a na konci v ní bude y_j oříšků. Veverka unese najednou jen jeden oříšek. Let x_{ij} označuje celkový počet oříšků přenesených ze staré jamky i do nové jamky j. Uvažujte dvě úlohy:

a) Čísla y_j jsou dána. Hledají se taková čísla x_{ij}, aby se veverka vykonala co nejméně práce, kde práce na přenesení jednoho oříšku je přímo úměrná vzdálenosti (vzdušnou čarou). Běh bez oříšku se za práci nepovažuje.
b) Hledají se čísla x_{ij} a y_j tak, aby veverka vykonala co nejméně práce a navíc byly v nových jamkách oříšky rozloženy co nejrovnoměrněji, čímž minimalizuje škodu způ-

sobenou případnou krádeží. Přesněji, aby rozdíl mezi největším a největším z čísel y_j byl menší než dané číslo t .

Formulujte obě úlohy jako LP. Předpokládejte, že počty oříšků jsou nezáporná reálná čísla, ač ve skutečnosti mohou být pouze nezáporná celá čísla.

Hints and Solutions

12.2.a) $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{0} \le \mathbf{x} \le \mathbf{1} \,\} = \sum_{i=1}^n \max\{0, c_i\}$, tedy optimální hodnota je součet kladných čísel c_i. Důkaz: Ukažme, že optimum se nabývá pro takové $\mathbf{x}$, že $x_i = 0$ pro $c_i < 0$ a $x_i = 1$ pro $c_i > 0$ (pro $c_i = 0$ je x_i libovolné). Kdyby to tak totiž nebylo, mohli bychom číslo $\mathbf{c}^T\mathbf{x}$ zvětšit zmenšením nějakého x_i pro $c_i < 0$ nebo zvětšením pro $c_i > 0$. Tedy $\mathbf{x}$ by nebyl optimální argument.

12.2.b) $\sum_{i=1}^n |c_i|$. Dokáže se podobně.

12.2.c) $\max_{i=1}^n c_i$

12.2.d) $\max_{i=1}^n \max\{0, c_i\} = \max\bigl\{0, \max_{i=1}^n c_i\bigr\}$

12.2.i) Nápověda: substituujte $y_i = x_i - x_{i-1}$

12.3.a) nejde

12.3.b) $\min\{\, z_1 + z_2 \mid x_1, x_2, z_1, z_2 \in \mathbb{R},\ 2x_1 - x_2 \ge 1,\ -x_1 + 2x_2 \ge 1,\ x_1 \le z_1,\ x_2 \le z_2,\ -x_1 \le z_1,\ -x_2 \le z_2 \,\}$

12.3.c) nejde

12.3.d) $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{d}^T\mathbf{x} \le 1,\ -\mathbf{d}^T\mathbf{x} \le 1,\ \mathbf{x} \ge \mathbf{0} \,\}$

12.3.e) $\min\{\, \mathbf{1}^T\mathbf{z} \mid \mathbf{c}_{kl}^T\mathbf{x} + d_{kl} \le z_l\ (\forall k, l),\ -\mathbf{c}_{kl}^T\mathbf{x} - d_{kl} \le z_l\ (\forall k, l),\ \mathbf{x} \in \mathbb{R}^n,\ \mathbf{z} \in \mathbb{R}^L \,\}$ (analogické §12.1.1)

12.3.g) $\min\{\, \mathbf{1}^T\mathbf{z} \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{z} \in \mathbb{R}^m,\ -\mathbf{z} \le \mathbf{A}\mathbf{x} - \mathbf{b} \le \mathbf{z},\ -\mathbf{1} \le \mathbf{x} \le \mathbf{1} \,\}$

12.3.j) $\min\{\, \mathbf{1}^T\mathbf{y} + z \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{y} \in \mathbb{R}^m,\ z \in \mathbb{R},\ -\mathbf{z} \le \mathbf{A}\mathbf{x} - \mathbf{b} \le \mathbf{z},\ -z\mathbf{1} \le \mathbf{x} \le z\mathbf{1} \,\}$

12.7.a) $E(x) = \sum_{i=1}^n m_i \max(d_i + x, 0) + \sum_{i=1}^{n'} m'_i \max(d'_i - x, 0)$

12.7.b) $\min\{\, \sum_{i=1}^n m_i z_i + \sum_{i=1}^{n'} m'_i z'_i \mid x, z_i, z'_i \in \mathbb{R},\ z_i \ge d_i + x,\ z'_i \ge d'_i - x,\ z_i \ge 0,\ z'_i \ge 0 \,\}$

Chapter 13

Konvexní množiny a polyedry

Definition 13.1. *Množina $X \subseteq \mathbb{R}^n$ se nazývá* **konvexní**, *jestliže*

$$\mathbf{x} \in X,\ \mathbf{y} \in X,\ 0 \le \alpha \le 1 \quad \Longrightarrow \quad \alpha\mathbf{x} + (1-\alpha)\mathbf{y} \in X. \tag{13.1}$$

Množina $\{\, \alpha\mathbf{x} + (1-\alpha)\mathbf{y} \mid 0 \le \alpha \le 1 \,\}$ je úsečka spojující body $\mathbf{x}$ a $\mathbf{y}$ (viz Příklad 3.1). Definice tedy říká, že množina je konvexní, jestliže s každými dvěma body obsahuje i úsečku, která je spojuje. Obrázek ukazuje příklad konvexní a nekonvexní množiny v $\mathbb{R}^2$:

Konvexní množinu lze definovat i abstraktněji. **Konvexní kombinace** vektorů $\mathbf{x}_1, \dots, \mathbf{x}_k$ je jejich lineární kombinace $\alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k$ taková, že $\alpha_1 + \cdots + \alpha_k = 1$ a $\alpha_1, \dots, \alpha_k \ge 0$. Množina je konvexní právě tehdy, když je uzavřená vůči konvexním kombinacím (neboli každá konvexní kombinace vektorů z množiny leží v množině). Lze dokázat indukcí, že tato definice je ekvivalentní Definici 13.1. Všimněte si, že $\alpha\mathbf{x} + (1-\alpha)\mathbf{y}$ pro $0 \le \alpha \le 1$ je konvexní kombinací dvou vektorů $\mathbf{x}, \mathbf{y}$.

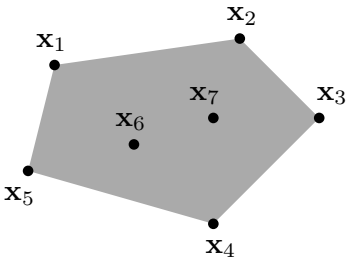
Konvexní obal vektorů $\mathbf{x}_1, \dots, \mathbf{x}_k$ je množina všech jejich konvexních kombinací. Tuto k-tici vektorů můžeme vnímat jako množinu $X = \{\mathbf{x}_1, \dots, \mathbf{x}_k\}$, konvexní obal pak značíme

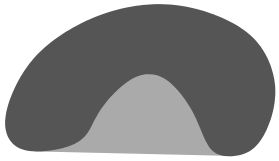
$$\operatorname{conv} X = \operatorname{conv}\{\mathbf{x}_1, \dots, \mathbf{x}_k\} = \{\, \alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k \mid \alpha_1 + \cdots + \alpha_k = 1,\ \alpha_1, \dots, \alpha_k \ge 0 \,\}. \tag{13.2}$$

Jak se definuje konvexní obal množiny s *nekonečným* počtem prvků, např. pravém obrázku výše? Nelze použít definice (13.2), neboť není jasné, co znamená součet $\alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k$ pro nekonečný počet vektorů (množina X může být i nespočetná). Konvexní obal libovolné (konečné či nekonečné) množiny $X \subseteq \mathbb{R}^n$ se definuje jako průnik všech konvexních množin, které množinu obsahují, tedy

$$\operatorname{conv} X = \bigcap \{\, Y \mid Y \supseteq X,\ Y \text{ konvexní} \,\}.$$

Obrázek ukazuje konvexní obal konečné (vlevo) a nekonečné (vpravo) množiny pro $n = 2$:





13.1 Čtyři kombinace a čtyři obaly

Konvexní kombinace je lineární kombinace, jejíž koeficienty splňují omezení $\alpha_1 + \cdots + \alpha_k = 1$ a $\alpha_1, \ldots, \alpha_k \geq 0$. Všimněte si, že když vynecháme druhé omezení, dostaneme afinní kombinaci (viz §3.3). Podle toho, které ze dvou omezení vyžadujeme, dostaneme čtyři druhy kombinací. Udělejme si v nich nyní pořádek.

Vážený součet $\alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k$ vektorů $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n$ se nazývá jejich

lineární kombinace,	jestliže	$\alpha_1, \ldots, \alpha_k \in \mathbb{R}$.		
afinní kombinace,	jestliže	$\alpha_1, \ldots, \alpha_k \in \mathbb{R}$,	$\alpha_1 + \cdots + \alpha_k = 1$.	
nezáporná kombinace,	jestliže	$\alpha_1, \ldots, \alpha_k \in \mathbb{R}$,		$\alpha_1, \ldots, \alpha_k \geq 0$.
konvexní kombinace,	jestliže	$\alpha_1, \ldots, \alpha_k \in \mathbb{R}$,	$\alpha_1 + \cdots + \alpha_k = 1$,	$\alpha_1, \ldots, \alpha_k \geq 0$.

Množina, která je uzavřená vůči

lineárním kombinacím,	se nazývá	**lineární podprostor**.
afinním kombinacím,	se nazývá	**afinní podprostor**.
nezáporným kombinacím,	se nazývá	**konvexní kužel**.
konvexním kombinacím,	se nazývá	**konvexní množina**.

K tomu, co již znáte, přibyl pojem nezáporné kombinace a konvexního kuželu.

Lineární [afinní, nezáporný, konvexní] **obal** vektorů $\mathbf{x}_1, \ldots, \mathbf{x}_k$ je množina všech jejich lineárních [afinních, nezáporných, konvexních] kombinací. Obecněji, lineární [afinní, nezáporný, konvexní] obal množiny $X \subseteq \mathbb{R}^n$ je průnik všech lineárních podprostorů [afinních podprostorů, konvexních kuželů, konvexních množin] obsahující množinu X.

Example 13.1. Mějme tři body v $\mathbb{R}^3$, které neleží v jedné rovině s počátkem. Jejich lineární obal je celé $\mathbb{R}^3$. Jejich afinní obal je rovina jimi procházející. Jejich nezáporný obal je nekonečný trojboký hranol, jehož vrchol je v počátku a jehož hrany jsou tři polopřímky určené počátkem a danými body. Jejich konvexní obal je trojúhelník jimi určený. □

Jako cvičení si nakreslete lineární, afinní, nezáporný a konvexní obal náhodně zvolených k vektorů v $\mathbb{R}^n$ pro všech devět případů $k, n \in \{1, 2, 3\}$.

13.2 Operace zachovávající konvexitu množin

Jaké operace s konvexními množinami mají za výsledek opět konvexní množinu? Zdaleka nejdůležitější taková operace je průnik. Následující větu je snadné dokázat.

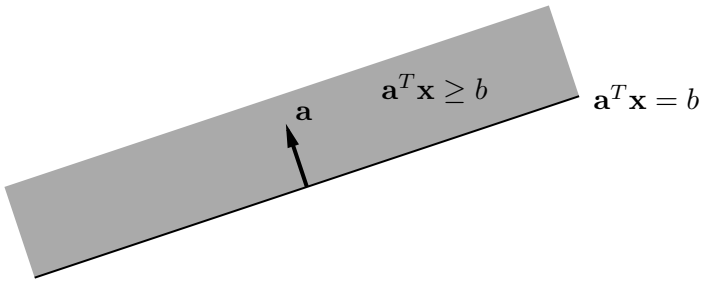
Theorem 13.1. *Průnik (konečně či nekonečně mnoha) konvexních množin je konvexní množina.*

Proof. Uděláme jen pro dvě množiny, zobecnění na libovolná počet množin je očividné. Let $X, Y \subseteq \mathbb{R}^n$ jsou konvexní. Let $\mathbf{x}, \mathbf{y} \in X \cap Y$, tedy každý z bodů $\mathbf{x}, \mathbf{y}$ je současně v množině X i v množině Y. Proto pro $0 \leq \alpha \leq 1$ bude bod $\alpha\mathbf{x} + (1-\alpha)\mathbf{y}$ také v množině X i Y, tedy bude v množině $X \cap Y$. □

Sjednocení konvexních množin ale *nemusí* být konvexní množina.

13.3 Konvexní polyedry

Poloprostor je množina $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} \geq b \,\}$ pro nějaké $\mathbf{a} \in \mathbb{R}^n$, $b \in \mathbb{R}$. Jeho hranice je nadrovina $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} = b \,\}$. Vektor $\mathbf{a} \in \mathbb{R}^n$ je normála této nadroviny. Obrázek znázorňuje tyto pojmy pro $n = 2$:

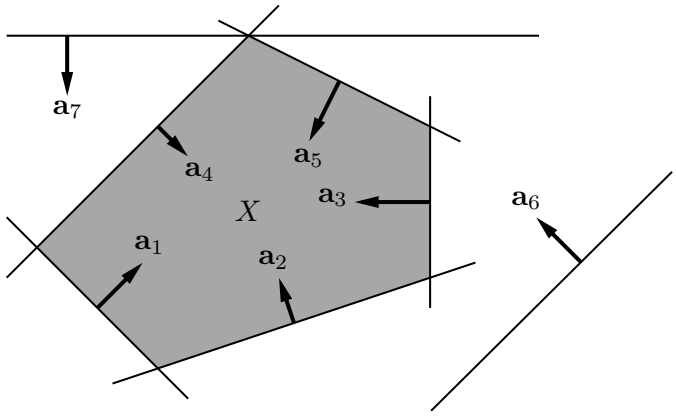


Definition 13.2. Konvexní polyedr *je průnik konečně mnoha poloprostorů.*

Konvexní polyedr je tedy množina

$$X = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}_i^T\mathbf{x} \geq b_i,\ i = 1, \ldots, m \,\} = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \geq \mathbf{b} \,\}, \tag{13.3}$$

kde $\mathbf{a}_1^T, \ldots, \mathbf{a}_m^T$ jsou řádky matice $\mathbf{A} \in \mathbb{R}^{m \times n}$ a $b_1, \ldots, b_m \in \mathbb{R}$ jsou složky vektoru $\mathbf{b} \in \mathbb{R}^m$. Všimněte si, že definice dovoluje i omezení typu rovnosti $\mathbf{a}_i^T\mathbf{x} = b_i$, protože to je ekvivalentní $\mathbf{a}_i^T\mathbf{x} \leq b_i$, $\mathbf{a}_i^T\mathbf{x} \geq b_i$. Obrázek ukazuje příklad pro $n = 2$, $m = 7$:



Všimněte si, že omezení 6 a 7 jsou redundantní – jejich odebráním by se polyedr nezměnil.

Jelikož poloprostor je očividně konvexní množina, plyne konvexita konvexního polyedru z Věty 13.1. Všimněte si, že konvexní polyedr nemusí být omezený.

Example 13.2. Množina (12.2) je konvexní polyedr, který vidíme na obrázku v Příklad 12.1.□

Example 13.3. Příklady konvexních polyedrů v $\mathbb{R}^n$:

- prázdná množina $\emptyset$
- celý prostor $\mathbb{R}^n$
- každý afinní podprostor (např. bod, přímka, rovina, nadrovina)
- polopřímka $\{\, \mathbf{x} + \alpha \mathbf{v} \mid \alpha \geq 0 \,\}$
- poloprostor
- panel $\{\, \mathbf{x} \in \mathbb{R}^n \mid b_1 \leq \mathbf{a}^T \mathbf{x} \leq b_2 \,\}$
- hyperkrychle $\{\, \mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_\infty \leq 1 \,\} = \{\, \mathbf{x} \in \mathbb{R}^n \mid -1 \leq x_i \leq 1,\ i = 1, \ldots, n \,\}$
- simplex, to jest konvexní obal $n+1$ afinně nezávislých bodů
- standardní simplex $\{\, \mathbf{x} \in \mathbb{R}^n \mid x_i \geq 0,\ \sum_{i=1}^n x_i \leq 1 \,\}$
- pravděpodobnostní simplex $\{\, \mathbf{x} \in \mathbb{R}^n \mid x_i \geq 0,\ \sum_{i=1}^n x_i = 1 \,\}$ (množina všech rozdělení pravděpodobnosti diskrétní náhodné proměnné)
- zobecněný osmistěn $\{\, \mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i| \leq 1 \,\}$. □

Example 13.4. Koule v $\mathbb{R}^n$ je průnikem nekonečně mnoha poloprostorů $\mathbf{a}^T \mathbf{x} \leq 1$ pro všechna $\|\mathbf{a}\|_2 = 1$. Je to konvexní množina, ale není to konvexní polyedr (protože počet poloprostorů není konečný). □

13.3.1 Stěny konvexního polyedru

Definition 13.3. **Stěna** *konvexního polyedru* $X \subseteq \mathbb{R}^n$ *je množina* $\operatorname*{argmin}_{\mathbf{x} \in X} \mathbf{c}^T \mathbf{x}$ *pro nějaké* $\mathbf{c} \in \mathbb{R}^n$.

Dimenze stěny je dimenze jejího afinního obalu (zopakujte si pojem afinního obalu z §13.1 a dimenze afinního podprostoru z §3.3). Stěny některých dimenzí mají jméno:

- stěna dimenze 0 se nazývá **vrchol**,
- stěna dimenze 1 se nazývá **hrana**,
- stěna dimenze $n-1$ se nazývá **faceta** (angl. *facet*, zatímco *face* znamená stěna).

Z Definice 13.3 a Věty 13.1 plyne, že každá stěna konvexního polyedru je sama o sobě konvexní polyedr.

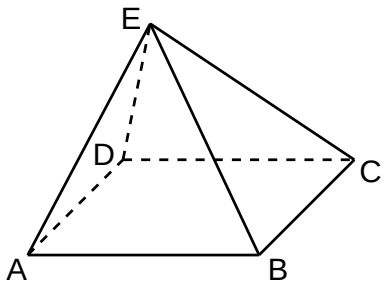
Definice 13.3 definuje stěny konvexního polyedru geometricky. Bez důkazu uvedeme algebraickou definici stěny, která předpokládá, že polyedr je ve tvaru (13.3).

Theorem 13.2. *$F \subseteq X$ je stěna konvexního polyedru (13.3) právě tehdy, když*

$$F = \{\, \mathbf{x} \in X \mid \mathbf{a}_i^T \mathbf{x} = \mathbf{b}_i,\ i \in I \,\} \tag{13.4}$$

pro nějakou podmnožinu indexů $I \subseteq \{1, \ldots, m\}$.

Example 13.5. Let polyedr (13.3) je pyramida v $\mathbb{R}^3$:



Tento polyedr je průnikem pěti poloprostorů, tedy $m = 5$ a $n = 3$. Let omezení $\mathbf{a}_i^T\mathbf{x} \geq b_i$ pro $i = 1, 2, 3, 4, 5$ je poloprostor, jehož hranicí je polorovina určená po řadě body ABCD, ABE, BCE, CDE, ADE. Pro $I = \{1\}$ je množina (13.4) faceta ABCD. Pro $I = \{1, 2\}$ je množina (13.4) hrana AB. Pro $I = \{1, 2, 3\}$ je množina (13.4) vrchol B. □

13.3.2 Dvě reprezentace konvexního polyedru

Následující věta je hluboká a uvádíme ji bez důkazu. (Pro neomezené konvexní polyedry platí podobná věta, trochu složitější, kterou neuvádíme.)

Theorem 13.3. *Konvexní obal konečně mnoha bodů je omezený konvexní polyedr. Obráceně, omezený konvexní polyedr je konvexním obalem svých vrcholů.*

Máme tedy dvě reprezentace omezeného polyedru:

- **H-reprezentace**: průnik konečně mnoha poloprostorů ('H' jako *half-space*)
- **V-reprezentace**: konvexní obal konečně mnoha bodů ('V' jako *vertex*)

Přechod od jedné reprezentace ke druhé může být výpočetně těžký či prakticky nemožný. Důvodem je to, že polyedr definovaný jako průnik malého počtu (přesněji, tento počet je polynomiální funkcí n) poloprostorů může mít vlemi velký (exponenciální funkce n) počet vrcholů. Naopak, polyedr s malým počtem vrcholů může mít exponenciální počet facet. V tom případě by algoritmus, který převádí H-reprezentaci na V-reprezentaci nebo naopak, by při polynomiálně dlouhém vstupu musel vydat exponenciálně dlouhý výstup.

Example 13.6. Uvažujme následující konvexní polyedry v $\mathbb{R}^n$ (viz Příklad 13.3):

- Simplex má $n + 1$ vrcholů a $n + 1$ facet.
- Hyperkrychle má $2n$ facet a 2^n vrcholů.
- Zobecněný osmistěn má $2n$ vrcholů a 2^n facet. □

13.4 Cvičení

13.1. Odpovězte, zda následující množiny jsou konvexní a odpověď dokažte z definice konvexní množiny:

a) interval $[a, b] \subseteq \mathbb{R}$, kde $a \leq b$
b) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \leq \mathbf{b},\ \mathbf{C}\mathbf{x} = \mathbf{d} \,\}$
c) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x}^T\mathbf{A}\mathbf{x} \leq 1 \,\}$, kde $\mathbf{A}$ je pozitivně semidefinitní
d) $\mathbb{Z}$ (množina celých čísel)

e) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \max\{x_1, , \ldots, , x_n\} \geq 0 \,\}$

13.2. Které z následujících množin jsou konvexní? Nemusíte dokazovat z definice, stačí uvést přesvědčivý argument. Množinu si nakreslete pro případ $n = 1$ a $n = 2$.

a) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \sum_{i=1}^n x_i = 1 \,\}$
b) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \sum_{i=1}^n x_i \geq 1 \,\}$
c) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} \geq \mathbf{0},\ \sum_{i=1}^n x_i = 1 \,\}$
d) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} \geq \mathbf{0},\ \sum_{i=1}^n x_i \leq 1 \,\}$
e) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 = 1 \,\}$
f) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 < 1 \,\}$
g) $\{\, (x, y) \in \mathbb{R}^2 \mid x \geq 0,\ y \geq 0,\ xy = 1 \,\}$
h) $\{\, (x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 2 \,\} \cap \{\, (x, y) \in \mathbb{R}^2 \mid (x-1)^2 + y^2 \leq 2 \,\}$

13.3. Které z následujících množin jsou konvexní polyedry? Pokud je množina konvexní polyedr, dokážete ji vyjádřit ve tvaru $\{\, \mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{b} \,\}$ (tj. jako průnik poloprostorů)?

a) $\left\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} \geq \mathbf{0},\ \sum_i x_i a_i = b,\ \sum_i x_i a_i^2 = c \,\right\}$, kde a_i, b, c jsou dané skaláry
b) $\{\, \mathbf{C}\mathbf{x} \mid \mathbf{x} \geq \mathbf{0},\ \mathbf{1}^T\mathbf{x} = 1 \,\}$, kde matice $\mathbf{C}$ je dána
c) $\{\, \mathbf{C}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \|\mathbf{x}\|_2 \leq 1 \,\}$, kde matice $\mathbf{C}$ je dána
d) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{a}\|_2 \leq \|\mathbf{x} - \mathbf{b}\|_2 \,\}$, kde $\mathbf{a}, \mathbf{b}$ jsou dány

13.4. Mějme vektory $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$. Pro každé $i = 1, \ldots, m$ definujeme množinu

$$X_i = \{\, \mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{a}_i\|_2 \leq \|\mathbf{x} - \mathbf{a}_j\|_2,\ j \neq i \,\}.$$

Ukažte, že množiny $X_1, \ldots, X_m$ jsou konvexní polyedry. Ukažte, že tyto množiny tvoří rozklad (zopakujte si, co je to rozklad množiny) množiny $\mathbb{R}^n$. Sjednocení hranic těchto množin se nazývá *Voronoiův diagram*. Nakreslete si ho pro $n = 2$ a $m = 4$ pro různé konfigurace bodů $\mathbf{a}_1, \ldots, \mathbf{a}_4$.

13.5. Bude Věta 13.1 platit, pokud v ní sousloví 'konvexní množina' nahradíme souslovím 'lineární podprostor' (příp. 'afinní podprostor', 'konvexní kužel')? Kladnou i zápornou odpověď dokažte.

Hints and Solutions

13.1.a) Konvexní, protože pro libovolné $\alpha \in [0, 1]$ je $\alpha a + (1 - \alpha)b \in [a, b]$.

13.1.b) Konvexní.

13.1.c) Konvexní.

13.1.d) Nekonvexní. Např. pro $x = 1$, $y = 2$, $\alpha = \frac{1}{2}$ číslo $\alpha x + (1 - \alpha)y = 1.5$ není celé.

13.1.e) Nekonvexní.

13.2.a) nadrovina, konvexní

13.2.b) poloprostor, konvexní

13.2.c) průnik poloprostorů a nadroviny, konvexní polyedr

13.2.d) průnik poloprostorů, konvexní

13.2.e) sféra, není konvexní

13.2.f) koule bez hranice, konvexní

13.2.g) graf jedné větve hyperboly, není konvexní

13.2.h) průnik dvou koulí, konvexní

Chapter 14

Simplexová metoda

Zde popíšeme algoritmus na řešení úloh lineárního programování zvaný **simplexová metoda**.

Zapomeňme prozatím na účelovou funkci a zkoumejme množinu přípustných řešení LP ve tvaru

$$X = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \,\}, \tag{14.1}$$

kde $\mathbf{A} \in \mathbb{R}^{m \times n}$ je široká ($m < n$) matice s hodností m, tedy její řádky jsou lineárně nezávislé.

Soustava $\mathbf{A}\mathbf{x} = \mathbf{b}$ má nekonečně mnoho řešení. Položíme-li však $n - m$ složek vektoru $\mathbf{x}$ rovno nule (tedy učiníme-li $n - m$ z podmínek $\mathbf{x} \geq \mathbf{0}$ aktivních), soustava má nejvýše jedno řešení. Tato úvaha vede k následujícím definicím:

- Množina $J \subseteq \{1, 2, \ldots, n\}$ se nazývá **báze** úlohy, pokud $|J| = m$ a sloupce matice $\mathbf{A}$ s indexy J jsou lineárně nezávislé. Tedy sloupce J tvoří regulární matici $m \times m$.
- Vektor $\mathbf{x}$ je **bázové řešení** příslušné bázi J, pokud $\mathbf{A}\mathbf{x} = \mathbf{b}$ a $x_j = 0$ pro $j \notin J$.
- Bázové řešení $\mathbf{x}$ je **přípustné**, pokud $\mathbf{x} \geq \mathbf{0}$.
- Bázové řešení $\mathbf{x}$ je **degenerované**, pokud má méně než m nenulových složek.
- Dvě báze jsou **sousední**, pokud mají $m - 1$ společných prvků.

Protože matice $\mathbf{A}$ má hodnost m, existuje aspoň jedna báze a každé bázi přísluší právě jedno bázové řešení. Bázové řešení však může příslušet více než jedné bázi, což se stane právě tehdy, když je toto bázové řešení degenerované.

Example 14.1. Let je soustava $\mathbf{A}\mathbf{x} = \mathbf{b}$ dána tabulkou (blokovou maticí)

$$\begin{bmatrix} \mathbf{A} & \mathbf{b} \end{bmatrix} = \left[\begin{array}{rrrrrr|r} -1 & 1 & 3 & 1 & 0 & 2 & 1 \\ 1 & 0 & 4 & 0 & 1 & 4 & 4 \\ -1 & 0 & 4 & 1 & 1 & 4 & 2 \end{array}\right]. \tag{14.2}$$

- $J = \{2, 3, 5\}$ není báze, protože sloupce $2, 3, 5$ matice $\mathbf{A}$ jsou lineárně závislé.
- $J = \{1, 4, 5\}$ je báze, protože tyto sloupce jsou lineárně nezávislé. Bázové řešení $\mathbf{x} = (x_1, x_2, \ldots, x_6)$ příslušné bázi J se najde řešením soustavy

$$\begin{bmatrix} -1 & 1 & 0 \\ 1 & 0 & 1 \\ -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} \tag{14.3}$$

  a položením $x_2 = x_3 = x_6 = 0$. Soustava (14.3) má právě jedno řešení, neboť její matice je regulární. Dostaneme $\mathbf{x} = (3, 0, 0, 4, 1, 0)$. Toto bázové řešení je přípustné. Není degenerované, protože má $m = 3$ nenulových složek.

- $J = \{1, 2, 4\}$ je báze. Bázové řešení je $\mathbf{x} = (4, -1, 0, 6, 0, 0)$. Je nepřípustné, protože $x_2 < 0$.
- $J = \{3, 4, 5\}$ je báze. Bázové řešení $\mathbf{x} = (0, 0, 1, -2, 0, 0)$ je degenerované, protože má méně než $m = 3$ nenulových složek.
- Stejné bázové řešení $\mathbf{x} = (0, 0, 1, -2, 0, 0)$ dostaneme volbou báze $J = \{3, 4, 6\}$. Vidíme, že degenerované bázové řešení odpovídá více než jedné bázi.
- Báze $\{1, 4, 5\}$ a $\{2, 4, 5\}$ jsou sousední, protože mají společné dva prvky $\{4, 5\}$. Báze $\{1, 4, 5\}$ a $\{2, 4, 6\}$ nejsou sousední. □

14.1 Geometrie simplexové metody

Následující věta udává spojitost mezi algebraickým a geometrickým popisem konvexního polyedru (14.1). Důkaz vynecháme.

Theorem 14.1. *Přípustná bázová řešení jsou vrcholy polyedru (14.1). Dvojice sousedních bází odpovídají buď jedinému (degenerovanému) vrcholu nebo dvojici vrcholů spojených hranou.*

Dle Definice 13.3 se všechna minima lineární funkce na konvexním polyedru X nabývají na nějaké stěně polyedru. Předpokládejme, že tato optimální stěna obsahuje alespoň jeden vrchol (to bude platit téměř vždy). Pak se alespoň jedno minimum nabývá i v tomto vrcholu. To nám dovoluje navrhnout naivní algoritmus na řešení LP: uděláme výčet všech přípustných bázových řešení a nalezneme to s nejlepší hodnotou účelové funkce. Tento algoritmus není praktický, protože bázových řešení je exponenciálně mnoho.

Simplexová metoda je efektivnější obměna tohoto přístupu: přechází mezi sousedními bázemi tak, že bázová řešení jsou stále přípustná (tedy přechází po hranách polyedru X) a účelová funkce se zlepšuje (nebo aspoň nezhoršuje).

14.2 Stavební kameny algoritmu

Zde vysvětlíme jednotlivé stavební kameny simplexové metody, které nakonec v §14.3 spojíme v celý algoritmus.

14.2.1 Přechod k sousední standardní bázi

Simplexová metoda pracuje pouze se *standardními* bázemi, tj. sloupce J jsou (permutované) vektory standardní báze. To má výhodu v tom, že (i) nemusíme kontrolovat, zda jsou sloupce J lineárně nezávislé a (ii) nenulové složky bázového řešení $\mathbf{x}$ jsou rovny přímo složkám vektoru $\mathbf{b}$. Na počátku algoritmu se předpokládá, že matice $\mathbf{A}$ obsahuje standardní bázi.

Z lineární algebry známe *ekvivalentní řádkové úpravy* soustavy $\mathbf{Ax} = \mathbf{b}$: libovolný řádek tabulky $\begin{bmatrix}\mathbf{A} & \mathbf{b}\end{bmatrix}$ můžeme vynásobit nenulovým číslem a můžeme k němu přičíst lineární kombinaci ostatních řádků. Tyto úpravy nemění množinu řešení soustavy.

Ukážeme, jak přejít od aktuální standardní báze J k sousední standardní bázi, tedy nějaký sloupec $j' \in J$ bázi opustí a nějaký sloupec $j \notin J$ do báze vstoupí. Let i je řádek, ve kterém je $a_{ij'} = 1$. Prvek (i, j) matice se nazývá **pivot** (angl. znamená *čep*). Let $a_{ij} \neq 0$. Chceme nastavit pivot a_{ij} na jedničku, vynulovat prvky nad i pod pivotem, a nezměnit přitom sloupce $J \setminus \{j'\}$. Toho se dosáhne těmito ekvivalentními řádkovými úpravami:

1. Vyděl řádek i číslem a_{ij}.
2. Pro každé $i' \neq i$ odečti $a_{i'j}$-násobek řádku i od řádku i'.

Říkáme, že jsme provedli *ekvivalentní úpravu kolem pivotu* s indexy (i, j).

Example 14.2. Mějme soustavu

$$\begin{bmatrix}\mathbf{A} & \mathbf{b}\end{bmatrix} = \left[\begin{array}{cccccc|c} 0 & 2 & 6 & 1 & 0 & 4 & 4 \\ 1 & \boxed{1} & 3 & 0 & 0 & 2 & 3 \\ 0 & -1 & 1 & 0 & 1 & 2 & 1 \end{array}\right] \tag{14.4}$$

se (standardní) bází $J = \{1, 4, 5\}$. Vidíme ihned odpovídající bázové řešení, $\mathbf{x} = (3, 0, 0, 4, 1, 0)$.

Chceme nahradit bázový sloupec $j' = 1$ nebázovým sloupcem $j = 2$, tedy přejít k sousední bázi $\{2, 4, 5\}$. Máme $i = 2$, tedy pivot je prvek a_{22} (v tabulce orámován). Ekvivalentními řádkovými úpravami musíme docílit, aby pivot byl roven jedné a prvky nad ním a pod ním byly nulové. Při tom smíme změnit sloupec 1, ale sloupce 4 a 5 se změnit nesmějí. Toho se docílí vydělením řádku 2 číslem a_{22} (což zde nemá žádný efekt, protože náhodou $a_{22} = 1$) a pak přičtením vhodných násobků řádku 2 k ostatním řádkům. Výsledek:

$$\begin{bmatrix}\mathbf{A} & \mathbf{b}\end{bmatrix} = \left[\begin{array}{cccccc|c} -2 & 0 & 0 & 1 & 0 & 0 & -2 \\ 1 & 1 & 3 & 0 & 0 & 2 & 3 \\ 1 & 0 & 4 & 0 & 1 & 4 & 4 \end{array}\right].$$

Nyní sloupce $\{2, 4, 5\}$ tvoří standardní bázi. □

14.2.2 Kdy je sousední bázové řešení přípustné?

Uvedeným způsobem můžeme od aktuální báze přejít k libovolné sousední bázi. Přitom nové bázové řešení může nebo nemusí být přípustné. Je-li aktuální bázové řešení přípustné, jak poznáme, zda i nové bázové řešení bude přípustné?

Protože nenulové složky bázového řešení $\mathbf{x}$ jsou rovny složkám vektoru $\mathbf{b}$, bázové řešení je přípustné právě tehdy, když $\mathbf{b} \geq \mathbf{0}$. Let v aktuální tabulce je $\mathbf{b} \geq \mathbf{0}$. Proveďme ekvivalentní úpravu kolem pivotu (i, j). Hledáme podmínky na (i, j), za kterých bude i po úpravě $\mathbf{b} \geq \mathbf{0}$.

Po ekvivalentní úpravě kolem pivotu (i, j) se vektor $\mathbf{b}$ změní takto (viz §14.2.1):

- b_i se změní na b_i/a_{ij},
- pro každé $i' \neq i$ se $b_{i'}$ změní na $b_{i'} - a_{i'j}b_i/a_{ij}$.

Tato čísla musejí být nezáporná. To nastane právě tehdy, když platí následující podmínky:

$$a_{ij} > 0, \tag{14.5a}$$

$$\text{pro každé } i' \neq i \text{ platí } a_{i'j} \leq 0 \text{ nebo } \frac{b_i}{a_{ij}} \leq \frac{b_{i'}}{a_{i'j}}, \tag{14.5b}$$

kde 'nebo' je užito v nevylučovacím smyslu. Podmínka (14.5a) je zřejmá. Podmínka (14.5b) je ekvivalentní podmínce $b_{i'} - a_{i'j}b_i/a_{ij} \geq 0$, neboť $a_{ij} > 0$, $b_i \geq 0$, $b_{i'} \geq 0$ (rozmyslete!).

Example 14.3. Uvažujme opět tabulku (14.4).

- Ekvivalentní úprava okolo pivotu $(i, j) = (3, 2)$ nepovede k přípustnému bázovému řešení, neboť $a_{ij} = -1 < 0$, což porušuje podmínku (14.5a).

- Ekvivalentní úprava okolo pivotu $(i,j) = (2,2)$ nepovede k přípustnému bázovému řešení, neboť pro $i' = 1$ je $a_{i'j} > 0$ a $\frac{3}{1} > \frac{4}{2}$, tedy podmínka (14.5b) je porušena.
- Ekvivalentní úprava okolo pivotu $(i,j) = (3,6)$ povede k přípustnému bázovému řešení. Podmínky (14.5) jsou splněny, neboť $a_{ij} = 2 > 0$ a $\frac{1}{2} \leq \frac{4}{4}$, $\frac{1}{2} \leq \frac{3}{2}$. □

14.2.3 Co když je celý sloupec nekladný?

Jestliže jsou všechny prvky v nějakém nebázovém sloupci j nekladné, víme z podmínky (14.5a), že tento sloupec se nemůže stát bázovým. Platí ale navíc, že souřadnice x_j bodu $\mathbf{x}$ se může libovolně zvětšovat a bod $\mathbf{x}$ přesto zůstane v polyedru X. Tedy existuje polopřímka s počátkem v $\mathbf{x}$ ležící celá v polyedru X. Tedy polyedr X je neomezený.

Example 14.4. Let $\begin{bmatrix}\mathbf{A} & \mathbf{b}\end{bmatrix}$ je tabulka

$$\begin{array}{rrrrrrr|r} & 0 & -2 & 6 & 1 & 0 & 4 & 4 \\ & 1 & -1 & 3 & 0 & 0 & 2 & 3 \\ & 0 & -1 & 1 & 0 & 1 & 2 & 1 \\ \hline \mathbf{x} = & 3 & 0 & 0 & 4 & 1 & 0 & \end{array}$$

s bází $\{1,4,5\}$. Pod tabulkou je napsáno bázové řešení $\mathbf{x}$. Když se x_2 bude libovolně zvětšovat, změnu lze kompenzovat současným zvětšováním bázových proměnných x_1, x_4, x_5 tak, že vektor $\mathbf{Ax}$ zůstane nezměněn a tedy roven $\mathbf{b}$. Konkrétně, vektor pro každé $\alpha \geq 0$ bude vektor $\mathbf{x} = (3,0,0,4,1,0) + \alpha(1,1,0,2,1,0)$ splňovat $\mathbf{Ax} = \mathbf{b}$ a $\mathbf{x} \geq \mathbf{0}$. □

14.2.4 Ekvivalentní úpravy účelového řádku

Dosud jsme prováděli ekvivalentní řádkové úpravy pouze na soustavě $\mathbf{Ax} = \mathbf{b}$ a účelové funkce si nevšímali. Tyto úpravy lze rozšířit na celou úlohu LP včetně účelové funkce. Nebudeme účelovou funkci uvažovat ve tvaru $\mathbf{c}^T\mathbf{x}$, ale v mírně obecnějším tvaru $\mathbf{c}^T\mathbf{x} - d$. Tedy řešíme LP

$$\min\{\, \mathbf{c}^T\mathbf{x} - d \mid \mathbf{Ax} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \,\}. \tag{14.6}$$

Úlohu budeme reprezentovat **simplexovou tabulkou**

$$\begin{bmatrix} \mathbf{c}^T & d \\ \mathbf{A} & \mathbf{b} \end{bmatrix}. \tag{14.7}$$

Přičtěme k účelovému řádku $\begin{bmatrix}\mathbf{c}^T & \mathrm{d}\end{bmatrix}$ libovolnou lineární kombinaci $\mathbf{y}^T\begin{bmatrix}\mathbf{A} & \mathbf{b}\end{bmatrix}$ ostatních řádků $\begin{bmatrix}\mathbf{A} & \mathbf{b}\end{bmatrix}$, kde $\mathbf{y}$ jsou koeficienty lineární kombinace. Ukážeme, že tato úprava zachová hodnotu účelové funkce $\mathbf{c}^T\mathbf{x} - d$ pro každé $\mathbf{x}$ splňující $\mathbf{Ax} = \mathbf{b}$. Nový účelový řádek bude

$$\begin{bmatrix}\mathbf{c}^T & d\end{bmatrix} + \mathbf{y}^T\begin{bmatrix}\mathbf{A} & \mathbf{b}\end{bmatrix} = \begin{bmatrix}\mathbf{c}^T + \mathbf{y}^T\mathbf{A} & d + \mathbf{y}^T\mathbf{b}\end{bmatrix}.$$

Nová účelová funkce bude tedy

$$(\mathbf{c}^T + \mathbf{y}^T\mathbf{A})\mathbf{x} - (d + \mathbf{y}^T\mathbf{b}) = \mathbf{c}^T\mathbf{x} - d + \mathbf{y}^T(\mathbf{Ax} - \mathbf{b}).$$

Ale to je rovno $\mathbf{c}^T\mathbf{x} - d$ pro každé $\mathbf{x}$ splňující $\mathbf{Ax} = \mathbf{b}$.

14.2.5 Co udělá přechod k sousední bázi s účelovou funkcí?

Let J je standardní báze. Přičtěme k účelovému řádku takovou lineární kombinaci ostatních řádků, aby pro všechna $j \in J$ bylo $c_j = 0$ (novému vektoru $\mathbf{c}$ se pak říká *redukované ceny*). Protože bázové řešení $\mathbf{x}$ je v nebázových sloupcích nulové, znamená to $\mathbf{c}^T\mathbf{x} = 0$. Tedy hodnota účelové funkce $\mathbf{c}^T\mathbf{x} - d$ v bázovém řešení $\mathbf{x}$ je rovna jednoduše $-d$.

Navíc je snadno vidět, co udělá přechod k nové bázi s účelovou funkcí. Let j' je sloupec opouštějící bázi a j je sloupec vstupující do báze. Při přechodu k nové bázi se číslo $x_{j'}$ stane nulovým a číslo x_j se zvětší z nuly na kladné (nebo se nezmění). Protože $c_{j'} = 0$, číslo $\mathbf{c}^T\mathbf{x} - d$ při $c_j \geq 0$ stoupne (nebo se nezmění) a při $c_j \leq 0$ klesne (nebo se nezmění).

Example 14.5. Mějme úlohu se simplexovou tabulkou

$$\begin{bmatrix} \mathbf{c}^T & d \\ \mathbf{A} & \mathbf{b} \end{bmatrix} = \left[\begin{array}{rrrrrr|r} 1 & -2 & -3 & -1 & 2 & 1 & 4 \\ \hline 0 & 2 & 6 & 1 & 0 & 4 & 4 \\ 1 & 1 & 3 & 0 & 0 & 2 & 3 \\ 0 & -1 & 1 & 0 & 1 & 2 & 1 \end{array}\right],$$

kde $J = \{1, 4, 5\}$. Složky vektoru $\mathbf{c}$ v bázových sloupcích vynulujeme tak, že k účelovému řádku přičteme první řádek, odečteme druhý řádek, a odečteme dvojnásobek třetího řádku:

$$\begin{array}{r|rrrrrr|r} & 0 & 1 & -2 & 0 & 0 & -1 & 3 \\ \hline & 0 & 2 & 6 & 1 & 0 & 4 & 4 \\ & 1 & 1 & 3 & 0 & 0 & 2 & 3 \\ & 0 & -1 & 1 & 0 & 1 & 2 & 1 \\ \hline \mathbf{x} = & 3 & 0 & 0 & 4 & 1 & 0 & \end{array}$$

Pod tabulku jsme napsali bázové řešení $\mathbf{x}$. Nyní je $\mathbf{c}^T\mathbf{x} = 0$, a tedy hodnota účelové funkce v bázovém řešení je $\mathbf{c}^T\mathbf{x} - d = -d = -3$.

Dejme tomu, že chceme přidat do báze nebázový sloupec 2 a vyloučit z ní některý z bázových sloupců $\{1, 4, 5\}$. Po tomto přechodu se x_2 stane kladné nebo zůstane nulové a jedna ze složek x_1, x_4, x_5 se vynuluje. Protože $c_1 = c_4 = c_5 = 0$, změna x_1, x_4, x_5 se na účelové funkci neprojeví a ta se změní o c_2x_2. Kritérium tedy stoupne nebo zůstane stejné, protože $c_2 = 1 > 0$. □

Pokud v některém sloupci j je $c_j \leq 0$ a $a_{ij} \leq 0$ pro všechna i, pak můžeme proměnnou x_j libovolně zvětšovat (viz §14.2.3) a účelovou funkci libovolně zmenšovat. Úloha je tedy neomezená.

14.3 Základní algoritmus

Spojením popsaných stavebních kamenů dostaneme iteraci simplexového algoritmu na řešení úlohy (14.6). Iterace přejde k sousední standardní bázi takové, že bázové řešení zůstane přípustné a účelová funkce se zmenší nebo alespoň nezmění. Vstupem i výstupem iterace je simplexová tabulka (14.7) s těmito vlastnostmi:

- podmnožina sloupců $\mathbf{A}$ tvoří standardní bázi J,
- bázové řešení odpovídající této bázi je přípustné, $\mathbf{b} \geq \mathbf{0}$,
- složky vektoru $\mathbf{c}$ v bázových sloupcích jsou nulové, $c_j = 0$ pro $j \in J$.

Iteraci se provede v těchto krocích:

1. Vyber index j pivotu tak, aby $c_j < 0$ (viz §14.2.5).
2. Vyber index i pivotu podle podmínek (14.5). Z těchto podmínek plyne (promyslete!)

$$i \in \operatorname*{argmin}_{i' \mid a_{i'j} > 0} \frac{b_{i'}}{a_{i'j}}, \tag{14.8}$$

   kde $\operatorname{argmin}_{i' \mid a_{i'j} > 0}$ označuje, že se minimalizuje přes všechna i' splňující $a_{i'j} > 0$.
3. Udělej ekvivalentní úpravu okolo pivotu (i, j) (viz §14.2.1).
4. Udělej ekvivalentní úpravu účelového řádku, která vynuluje c_j v novém bázovém sloupci j (viz §14.2.5).

Algoritmus, který opakuje uvedenou iteraci, nazveme **základní simplexový algoritmus**. Algoritmus končí, když už nelze iteraci provést. To nastane z jednoho z těchto důvodů:

- Všechny koeficienty c_j jsou nezáporné. Účelovou funkci nelze zlepšit a jsme v optimu.
- V některém sloupci j je $c_j < 0$ a $a_{ij} \leq 0$ pro všechna i. Úloha je neomezená.

Výběr indexů (i, j) pivotu v krocích 1 a 2 nemusí být jednoznačný, tedy může být více sloupců j s vhodným znaménkem c_j a více řádků i může splňovat podmínky (14.5) (tedy může být více argumentů minima v podmínce (14.8)). Algoritmus, který vybírá jediný pivot z těchto možností, se nazývá **pivotové pravidlo**.

Zřídka se algoritmus může dostat do stavu, kdy cyklicky prochází stále stejnou množinu bází, které odpovídají jedinému degenerovanému bázovému řešení a tedy účelová funkce se nemění. Tomuto problému **cyklení** se dá zabránit použitím vhodného pivotového pravidla (nejznámější je *Blandovo anticyklící pravidlo*), které ale popisovat nebudeme.

Example 14.6. Vyřešte lineární program (14.6) simplexovou metodou, když výchozí simplexová tabulka (14.7) je

$$\begin{array}{rrrrrr|r} 0 & -2 & 1 & 0 & 0 & -3 & 0 \\ \hline 0 & 2 & 6 & 1 & 0 & 4 & 4 \\ 1 & 1 & 3 & 0 & 0 & 2 & 3 \\ 0 & -1 & 1 & 0 & 1 & \boxed{2} & 1 \end{array}.$$

Báze je $J = \{1, 4, 5\}$ a bázové řešení $\mathbf{x} = (3, 0, 0, 4, 1, 0)$.

Účelový řádek budeme nazývat nultý, ostatní pak prvý, druhý atd. První iterace simplexového algoritmu se provede v těchto krocích:

1. Vybereme sloupec j, který vstoupí do báze. To může být libovolný sloupec, který má v nultém řádku záporné číslo. Je rozumné vzít nejmenší takové číslo, zde -3, tedy $j = 6$.
2. Vybereme řádek i pivotu dle (14.8) nalezením argumentu minima z čísel $\frac{4}{4}, \frac{3}{2}, \frac{1}{2}$. Bude tedy $i = 3$. Výsledný pivot je označen rámečkem. Všimněte si, že řádek $i = 3$ má v aktuální bázi jedničku ve sloupci 5, sloupec 5 tedy bázi opustí.

3, 4. Uděláme ekvivalentní úpravu okolo pivotu (i, j) a zároveň vynulujeme číslo c_j. Neboli chceme, aby se z pivotu a_{ij} stala jednička a nad i pod pivotem byly nuly, a to včetně nultého řádku. Tedy nejprve třetí řádek vydělíme dvěma a potom k nultému řádku přičteme trojnásobek třetího řádku, od prvního řádku odečteme čtyřnásobek třetího řádku, a od

druhého řádku odečteme dvojnásobek třetího řádku. Všimněte si: k žádnému řádku nikdy nepřičítáme násobky jiného řádku než pivotového. Výsledek:

$$\begin{array}{rrrrrr|r}
0 & -3.5 & 2.5 & 0 & 1.5 & 0 & 1.5 \\
\hline
0 & \boxed{4} & 4 & 1 & -2 & 0 & 2 \\
1 & 2 & 2 & 0 & -1 & 0 & 2 \\
0 & -0.5 & 0.5 & 0 & 0.5 & 1 & 0.5
\end{array}$$

Na konci první iterace máme bázi $J = \{1, 4, 6\}$, bázové řešení $\mathbf{x} = (2, 0, 0, 2, 0, 0.5)$, a hodnotu účelové funkce $-d = -1.5$.

Druhá iterace: pivot je ve sloupci $j = 2$. Jeho řádek najdeme dle (14.8) porovnáním čísel $\frac{2}{4}, \frac{2}{2}$, tedy $i = 1$. Výsledek druhé iterace:

$$\begin{array}{rrrrrr|r}
0 & 0 & 6 & 0.875 & -0.25 & 0 & 3.25 \\
\hline
0 & 1 & 1 & 0.25 & -0.5 & 0 & 0.5 \\
1 & 0 & 0 & -0.5 & 0 & 0 & 1 \\
0 & 0 & 1 & 0.125 & \boxed{0.25} & 1 & 0.75
\end{array}$$

Výsledek třetí iterace:

$$\begin{array}{rrrrrr|r}
0 & 0 & 7 & 1 & 0 & 1 & 4 \\
\hline
0 & 1 & 3 & 0.5 & 0 & 2 & 2 \\
1 & 0 & 0 & -0.5 & 0 & 0 & 1 \\
0 & 0 & 4 & 0.5 & 1 & 4 & 3
\end{array}$$

Protože všechna čísla v účelovém řádku jsou nezáporná, algoritmus končí. Úloha má optimální řešení s hodnotou -4 v bodě $(x_1, x_2, x_3, x_4, x_5, x_6) = (1, 2, 0, 0, 3, 0)$. □

Example 14.7. Let simplexová tabulka (14.7) je

$$\begin{array}{rrrrr|r}
-2 & 6 & 1 & 0 & 0 & 0 \\
\hline
-1 & -1 & -1 & 1 & 0 & 2 \\
\boxed{2} & -1 & -2 & 0 & 1 & 1
\end{array}$$

Tabulka po první iteraci je

$$\begin{array}{rrrrr|r}
0 & 5 & -1 & 0 & 1 & 1 \\
\hline
0 & -1.5 & -2 & 1 & 0.5 & 2.5 \\
1 & -0.5 & -1 & 0 & 0.5 & 0.5
\end{array}$$

Podle nultého řádku by další pivot měl být ve třetím sloupci. Ale čísla a_{i3} jsou všechna záporná (viz §14.2.3). Tedy úloha je neomezená. V nové tabulce je vidět, že můžeme zvětšovat x_3 libovolně a kompenzovat to vhodným nárůstem x_1 a x_4. Jelikož $c_1 = c_4 = 0$, změny x_1 a x_4 se na účelové funkci neprojeví a jediný vliv na ní bude mít x_3, které ho bude libovolně zmenšovat. □

14.4 Inicializace algoritmu

Na začátku základního simplexového algoritmu musí být úloha zadána ve tvaru

$$\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \,\}, \tag{14.9}$$

kde matice $\mathbf{A}$ obsahuje standardní bázi a $\mathbf{b} \geq \mathbf{0}$. Ukážeme, jak lze obecnou úlohu LP převést na tento tvar.

Někdy je převod snadný. Pokud má úloha tvar $\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{A}\mathbf{x} \leq \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \,\}$ a platí $\mathbf{b} \geq \mathbf{0}$, přidáme slackové proměnné $\mathbf{u} \geq \mathbf{0}$ a omezení převedeme na $\mathbf{A}\mathbf{x} + \mathbf{u} = \mathbf{b}$. Úloha tedy bude mít simplexovou tabulku

$$\begin{bmatrix} \mathbf{c}^T & \mathbf{0} & 0 \\ \mathbf{A} & \mathbf{I} & \mathbf{b} \end{bmatrix},$$

ve které sloupce příslušné proměnným $\mathbf{u}$ tvoří standardní bázi.

Example 14.8. Vyřešte simplexovým algoritmem:

$$\begin{aligned} \min \quad & -3x_1 - x_2 - 3x_3 \\ \text{za podmínek} \quad & 2x_1 + x_2 + x_3 \leq 2 \\ & x_1 + 2x_2 + 3x_3 \leq 5 \\ & 2x_1 + 2x_2 + x_3 \leq 6 \\ & x_1, x_2, x_3 \geq 0 \end{aligned}$$

Přidáme slackové proměnné $u_1, u_2, u_3 \geq 0$, abychom omezení uvedli do tvaru rovností:

$$\begin{aligned} \min \quad & -3x_1 - x_2 - 3x_3 \\ \text{za podmínek} \quad & 2x_1 + x_2 + x_3 + u_1 = 2 \\ & x_1 + 2x_2 + 3x_3 + u_2 = 5 \\ & 2x_1 + 2x_2 + x_3 + u_3 = 6 \\ & x_1, x_2, x_3, u_1, u_2, u_3 \geq 0 \end{aligned}$$

Zde je výchozí simplexová tabulka:

$$\begin{array}{cccccc|c} -3 & -1 & -3 & 0 & 0 & 0 & 0 \\ \hline 2 & 1 & 1 & 1 & 0 & 0 & 2 \\ 1 & 2 & 3 & 0 & 1 & 0 & 5 \\ 2 & 2 & 1 & 0 & 0 & 1 & 6 \end{array}$$

□

14.4.1 Dvoufázová simplexová metoda

Pokud je úloha zadána v obecném tvaru, operacemi z §12.1 ji lze vždy převést do tvaru (14.9). Vynásobením vhodných řádků záporným číslem vždy zajistíme $\mathbf{b} \geq \mathbf{0}$, matice $\mathbf{A}$ ale nemusí obsahovat standardní bázi. Máme dokonce vážnější problém: není vůbec jasné, zda úloha (14.9) je přípustná. V tomto případě nejdříve vyřešíme *pomocnou úlohu* LP, která najde *nějaké* (ne nutně optimální) přípustné řešení. Z něj pak získáme standardní bázi. Pomocná úloha je

$$\min\{\, \mathbf{1}^T\mathbf{u} \mid \mathbf{A}\mathbf{x} + \mathbf{u} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0},\ \mathbf{u} \geq \mathbf{0} \,\} \tag{14.10}$$

a má simplexovou tabulku

$$\begin{bmatrix} \mathbf{0} & \mathbf{1}^T & 0 \\ \mathbf{A} & \mathbf{I} & \mathbf{b} \end{bmatrix}.$$

Pro libovolné $\mathbf{u} \geq \mathbf{0}$ je $\mathbf{1}^T\mathbf{u} \geq 0$, přičemž $\mathbf{1}^T\mathbf{u} = 0$ právě tehdy, když $\mathbf{u} = \mathbf{0}$. Tedy úloha (14.9) je přípustná právě tehdy, je-li optimální hodnota úlohy (14.10) rovna 0. Na počátku tvoří sloupce příslušné proměnným $\mathbf{u}$ standardní bázi, lze tedy na ní pustit základní simplexový algoritmus. Ten může skončit dvěma způsoby:

- Pokud je optimum větší než 0, pak úloha (14.9) je nepřípustná.
- Pokud je optimum rovno 0, pak úloha (14.9) je přípustná. Pokud není optimální řešení $(\mathbf{x}, \mathbf{u})$ úlohy (14.10) degenerované, po skončení simplexového algoritmu jsou všechny bázové proměnné kladné. Protože $\mathbf{u} = \mathbf{0}$, proměnné $\mathbf{u}$ budou tedy nebázové. Proto mezi sloupci příslušnými proměnným $\mathbf{x}$ bude existovat standardní báze.

  Pokud je optimální řešení $(\mathbf{x}, \mathbf{u})$ úlohy (14.10) degenerované, některé proměnné $\mathbf{u}$ mohou být na konci algoritmu bázové. Pak je nutno udělat dodatečné úpravy kolem pivotů ve sloupcích příslušných bázovým proměnným $\mathbf{u}$, abychom tyto sloupce dostali z báze ven.

Nalezení nějakého přípustného řešení v pomocné úloze (14.10) se nazývá **první fáze** a řešení původní úlohy pak **druhá fáze** algoritmu, mluvíme tedy o **dvoufázové simplexové metodě**.

Example 14.9. Řešte

$$\begin{array}{rrcl} \min & -20x_1 - 30x_2 - 40x_3 & & \\ \text{za podmínek} & 3x_1 + 2x_2 + x_3 & = & 10 \\ & x_1 + 2x_2 + 2x_3 & = & 15 \\ & x_1, x_2, x_3 & \geq & 0 \end{array}$$

Máme sice $\mathbf{b} \geq \mathbf{0}$, ale není jasné, zda existuje přípustné $\mathbf{x}$, tím méně není vidět standardní báze. Provedeme první fázi algoritmu. Pomocná úloha bude

$$\begin{array}{rrcl} \min & u_1 + u_2 & & \\ \text{za podmínek} & 3x_1 + 2x_2 + x_3 + u_1 & = & 10 \\ & x_1 + 2x_2 + 2x_3 + u_2 & = & 15 \\ & x_1, x_2, x_3, u_1, u_2 & \geq & 0 \end{array}$$

s tabulkou

$$\begin{array}{ccccc|c} 0 & 0 & 0 & 1 & 1 & 0 \\ \hline 3 & 2 & 1 & 1 & 0 & 10 \\ 1 & 2 & 2 & 0 & 1 & 15 \end{array}$$

Sloupce nad přidanými proměnnými tvoří standardní bázi, můžeme tedy na pomocnou úlohu pustit základní simplexový algoritmus. Po vynulování ceny nad bázovými proměnnými budou kroky algoritmu vypadat takto:

$$\begin{array}{rrrrr|r} -4 & -4 & -3 & 0 & 0 & -25 \\ \hline 3 & \boxed{2} & 1 & 1 & 0 & 10 \\ 1 & 2 & 2 & 0 & 1 & 15 \end{array}$$

$$\begin{array}{rrrrr|r} 2 & 0 & -1 & 2 & 0 & -5 \\ \hline 1.5 & 1 & 0.5 & 0.5 & 0 & 5 \\ -2 & 0 & \boxed{1} & -1 & 1 & 5 \end{array}$$

$$\begin{array}{rrrrr|r} 0 & 0 & 0 & 1 & 1 & 0 \\ \hline 2.5 & 1 & 0 & 1 & -0.5 & 2.5 \\ -2 & 0 & 1 & -1 & 1 & 5 \end{array}$$

Optimum je rovno 0, tedy původní úloha je přípustná. Proměnné u_1, u_2 jsou nebázové a tedy rovny nule, bázové proměnné jsou x_2, x_3. Teď tedy můžeme začít druhou fázi (řešení původní

úlohy) s počáteční tabulkou

$$\begin{array}{rrr|r} -20 & -30 & -40 & 0 \\ \hline 2.5 & 1 & 0 & 2.5 \\ -2 & 0 & 1 & 5 \end{array}$$

□

14.5 Cvičení

14.1. V tabulce označte všechny pivoty takové, že ekvivalentní úprava kolem nich povede k přípustnému bázovému řešení:

$$\begin{bmatrix} \mathbf{A} & \mathbf{b} \end{bmatrix} = \left[\begin{array}{rrrrrrrr|r} 0 & 2 & 6 & 1 & 0 & -4 & 3 & 0 & 4 \\ 1 & 1 & -3 & 0 & 0 & 2 & 3 & 0 & 3 \\ 0 & -1 & 1 & 0 & 1 & -2 & -3 & 0 & 1 \\ 0 & -2 & 2 & 0 & 0 & 2 & -1 & 1 & 1 \end{array}\right]$$

14.2. Zapište lineární program

$$\begin{array}{rl} \min & -x_1 \qquad\qquad - \ x_4 - 3x_5 \\ \text{za podmínek} & 2x_1 \qquad\qquad + \ x_4 + \ x_5 + x_6 = 2 \\ & -x_1 + x_2 \qquad + 2x_4 + 3x_5 \qquad = 4 \\ & 2x_1 \qquad + x_3 + 2x_4 - \ x_5 \qquad = 6 \\ & \qquad\qquad\qquad x_1, x_2, x_3, x_4, x_5, x_6 \geq 0 \end{array}$$

do simplexové tabulky. Předpokládejte, že aktuální báze je $\{2, 3, 6\}$. Jaké je aktuální bázové řešení? Je toto bázové řešení přípustné. Je degenerované? Pokud je to možné, udělejte jeden krok simplexového algoritmu. Pokud to možné není, vysvětlete proč.

14.3. Vyřešte simplexovou metodou:

$$\begin{array}{rl} \max & 2x_1 - \ x_2 - 3x_3 \\ \text{za podmínek} & -2x_1 - \ x_2 + \ x_3 \leq 2 \\ & -x_1 + 2x_2 - 3x_3 \leq 5 \\ & -2x_1 - 4x_2 + \ x_3 \leq 6 \\ & x_1, x_2, x_3 \geq 0 \end{array}$$

14.4. Vyřešte simplexovou metodou (navzdory tomu, že lze řešit úvahou):

$$\begin{array}{rl} \max & 6x_1 + 9x_2 + 5x_3 + 9x_4 \\ \text{za podmínek} & x_1 + x_2 + x_3 + x_4 = 1 \\ & x_1, x_2, x_3, x_4 \geq 0 \end{array}$$

14.5. Let úloha (14.6) má více než jedno optimální řešení. Jak se to pozná v simplexové tabulce? Navrhněte algoritmus, jehož výstupem bude výčet všech optimálních bázových řešení.

14.6. Upravte do vhodného tvaru a vyřešte dvoufázovou simplexovou metodou:

$$\begin{array}{rrcl} \max & 3x_1 - 4x_2 & & \\ \text{za podmínek} & -2x_1 - 5x_2 & \leq & 10 \\ & 3x_1 + x_2 & \leq & 3 \\ & -2x_1 + x_2 & \leq & -2 \\ & x_1 & \geq & 0 \\ & x_2 & \leq & -1 \end{array}$$

Hints and Solutions

14.6. Optimum je $(x_1, x_2) = (25, -36)/13$.

Chapter 15

Dualita v lineárním programování

Ke každé úloze LP lze sestavit podle dále popsaného postupu jinou úlohu LP. Novou úlohu nazýváme **duální**, původní úlohu nazýváme **primární** či **přímou**. Konstrukce je symetrická: duální úloha k duální úloze je původní úloha. Tedy má smysl říkat, že primární a duální úloha jsou *navzájem* duální. Dvojice duálních úloh je svázána zajímavými vztahy.

15.1 Konstrukce duální úlohy

K úloze LP v obecném tvaru (viz §12.1) se duální úloha získá dle tohoto postupu:

$$
\begin{array}{rlrll}
\min & \sum_{j\in J} c_j x_j & \max & \sum_{i\in I} b_i y_i & \\
\text{za podm.} & \sum_{j\in J} a_{ij} x_j = b_i & \text{za podm.} & y_i \in \mathbb{R}, & i \in I_0 \\
 & \sum_{j\in J} a_{ij} x_j \geq b_i & & y_i \geq 0\ , & i \in I_+ \\
 & \sum_{j\in J} a_{ij} x_j \leq b_i & & y_i \leq 0\ , & i \in I_- \\
 & x_j \in \mathbb{R} & & \sum_{i\in I} a_{ij} y_i = c_j, & j \in J_0 \\
 & x_j \geq 0 & & \sum_{i\in I} a_{ij} y_i \leq c_j, & j \in J_+ \\
 & x_j \leq 0 & & \sum_{i\in I} a_{ij} y_i \geq c_j, & j \in J_-
\end{array}
$$

V levém sloupci je primární úloha, v prostředním sloupci je z ní vytvořená duální úloha. V pravém sloupci jsou množiny indexů pro obě úlohy: $I = \{1, \ldots, m\} = I_0 \cup I_+ \cup I_-$ je indexová množina primárních omezení a duálních proměnných, $J = \{1, \ldots, n\} = J_0 \cup J_+ \cup J_-$ je indexová množina primárních proměnných a duálních omezení.

Všimněte si následující symetrie: i-tému primárnímu omezení $\sum_j a_{ij} x_j \geq b_i$ odpovídá duální proměnná $y_i \geq 0$. Opačně, j-tá primární proměnná $x_j \geq 0$ odpovídá j-tému duálnímu omezení $\sum_i a_{ij} y_i \leq c_j$. Podobně pro ostatní řádky.

Pro speciální tvary LP se dvojice duálních úloh přehledněji napíše v maticové formě. Např. pro $I_0 = I_- = J_0 = J_- = \emptyset$ obdržíme

$$
\begin{array}{rlrl}
\min & \mathbf{c}^T\mathbf{x} & \max & \mathbf{b}^T\mathbf{y} \\
\text{za podm.} & \mathbf{A}\mathbf{x} \geq \mathbf{b} & \text{za podm.} & \mathbf{y} \geq \mathbf{0} \\
 & \mathbf{x} \geq \mathbf{0} & & \mathbf{A}^T\mathbf{y} \leq \mathbf{c}
\end{array}
\tag{15.1}
$$

15.2 Věty o dualitě

Následující věty platí pro obecný tvar LP, ale důkazy uděláme pouze pro speciální tvar (15.1). V důkazech si všimněte, že $\mathbf{b}^T\mathbf{y} = \mathbf{y}^T\mathbf{b}$ a $\mathbf{A}^T\mathbf{y} \leq \mathbf{c}$ je totéž jako $\mathbf{y}^T\mathbf{A} \leq \mathbf{c}^T$.

Theorem 15.1 (o slabé dualitě). *Let* $\mathbf{x}$ *je přípustné primární řešení a* $\mathbf{y}$ *přípustné duální řešení. Pak* $\mathbf{c}^T\mathbf{x} \geq \mathbf{b}^T\mathbf{y}$.

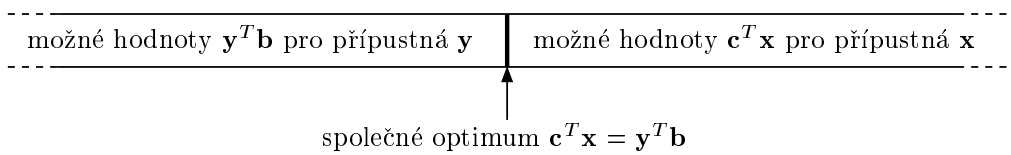
Proof. Díky přípustnosti $\mathbf{x}$ a $\mathbf{y}$ platí $\mathbf{y}^T\mathbf{A} \leq \mathbf{c}^T$ a $\mathbf{x} \geq \mathbf{0}$, z čehož plyne (proč?) $\mathbf{y}^T\mathbf{A}\mathbf{x} \leq \mathbf{c}^T\mathbf{x}$. Podobně, díky přípustnosti $\mathbf{x}$ a $\mathbf{y}$ platí $\mathbf{A}\mathbf{x} \geq \mathbf{b}$ a $\mathbf{y} \geq \mathbf{0}$, z čehož plyne $\mathbf{y}^T\mathbf{A}\mathbf{x} \geq \mathbf{y}^T\mathbf{b}$. Z toho

$$\mathbf{c}^T\mathbf{x} \geq \mathbf{y}^T\mathbf{A}\mathbf{x} \geq \mathbf{y}^T\mathbf{b}. \tag{15.2}$$

□

Theorem 15.2 (o silné dualitě). *Primární úloha má optimální řešení právě tehdy, když má duální úloha optimální řešení. Má-li primární úloha optimální řešení* $\mathbf{x}$ *a duální úloha optimální řešení* $\mathbf{y}$*, platí* $\mathbf{c}^T\mathbf{x} = \mathbf{b}^T\mathbf{y}$.

Důkaz věty o silné dualitě není jednoduchý a vynecháme jej. Věty o slabé a silné dualitě mají jasnou interpretaci: pro přípustná $\mathbf{x}$ a $\mathbf{y}$ není hodnota duální účelové funkce nikdy větší než hodnota primární účelové funkce a tyto hodnoty se potkají ve společném optimu:



Theorem 15.3 (o komplementaritě). *Let* $\mathbf{x}$ *je přípustné primární řešení a* $\mathbf{y}$ *přípustné duální řešení. Pak* $\mathbf{c}^T\mathbf{x} = \mathbf{b}^T\mathbf{y}$ *právě tehdy, když zároveň platí tyto dvě podmínky:*

$$y_i\left(\sum_{j\in J} a_{ij}x_j - b_i\right) = 0 \qquad \forall i \in I, \tag{15.3a}$$

$$x_j\left(\sum_{i\in I} a_{ij}y_j - c_j\right) = 0 \qquad \forall j \in J. \tag{15.3b}$$

Všimněte si, co podmínky (15.3) říkají: na každém řádku ve dvojici duálních úloh je vždy alespoň jedno omezení aktivní, buď primární nebo duální (přičemž omezení typu rovnosti bereme vždy za aktivní).

Proof. Pro libovolné vektory $\mathbf{u}, \mathbf{v} \geq \mathbf{0}$ platí

$$\forall i\,(u_i = 0 \text{ nebo } v_i = 0) \iff \forall i\,(u_i v_i = 0) \iff \mathbf{u}^T\mathbf{v} = 0.$$

Protože $\mathbf{x}$ a $\mathbf{y}$ jsou přípustné, podmínky (15.3) je tedy možno psát jako

$$\mathbf{y}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) = 0, \tag{15.4a}$$

$$(\mathbf{c}^T - \mathbf{y}^T\mathbf{A})\mathbf{x} = 0, \tag{15.4b}$$

neboli

$$\mathbf{c}^T\mathbf{x} = \mathbf{y}^T\mathbf{A}\mathbf{x} = \mathbf{y}^T\mathbf{b}. \tag{15.5}$$

Vztah (15.5) zjevně implikuje $\mathbf{c}^T\mathbf{x} = \mathbf{y}^T\mathbf{b}$. Obráceně, $\mathbf{c}^T\mathbf{x} = \mathbf{y}^T\mathbf{b}$ implikuje (15.5), neboť jsme dříve ukázali, že pro přípustné $\mathbf{x}, \mathbf{y}$ platí (15.2). □

Example 15.1. Mějme dvojici navzájem duálních úloh LP:

$$
\begin{array}{rrcl}
\min & 2x_1 + 5x_2 + 6x_3 & = & \mathbf{5.4} \\
\mathbf{3} = & 2x_1 + x_2 + 2x_3 & \geq & 3 \\
\mathbf{2.4} = & x_1 + 2x_2 + 2x_3 & \geq & 1 \\
\mathbf{3} = & x_1 + 3x_2 + x_3 & \geq & 3 \\
\mathbf{-0.6} = & -x_1 + x_2 - 2x_3 & \geq & -1 \\
\mathbf{1.2} = & x_1 & \geq & 0 \\
\mathbf{0.6} = & x_2 & \geq & 0 \\
\mathbf{0} = & x_3 & \geq & 0
\end{array}
\qquad
\begin{array}{rrcl}
\max & 3y_1 + y_2 + 3y_3 - y_4 & = & \mathbf{5.4} \\
\mathbf{0.2} = & y_1 & \geq & 0 \\
\mathbf{0} = & y_2 & \geq & 0 \\
\mathbf{1.6} = & y_3 & \geq & 0 \\
\mathbf{0} = & y_4 & \geq & 0 \\
\mathbf{2} = & 2y_1 + y_2 + y_3 - y_4 & \leq & 2 \\
\mathbf{5} = & y_1 + 2y_2 + 3y_3 + y_4 & \leq & 5 \\
\mathbf{2} = & 2y_1 + 2y_2 + y_3 - 2y_4 & \leq & 6
\end{array}
$$

Spočetli jsme optimální řešení obou úloh a dosadili tato řešení do účelových funkcí a do omezení. Hodnoty optimálních řešení $\mathbf{x}^* = (1.2, 0.6, 0)$ a $\mathbf{y}^* = (0.2, 0, 1.6)$ a hodnoty omezení a účelových funkcí v optimech jsou napsané tučně před/za rovnítky. Dle věty o silné dualitě se optima rovnají. Vezmeme-li libovolný řádek (kromě účelového), je na něm alespoň jedno z obou omezení aktivní. Např. ve druhém řádku je primární omezení $2x_1+x_2+2x_3 \geq 3$ aktivní a duální omezení $y_1 \geq 0$ je neaktivní. Podle věty o komplementaritě se nemůže stát, že by na některém řádku byly obě omezení zároveň neaktivní (mohou být obě ale zároveň aktivní, což zde nenastává, ale může to nastat v případě degenerace). □

Předložíme-li přípustná primární a duální řešení taková, že se účelové funkce rovnají, dokázali jsme optimalitu obou úloh. Pro velké úlohy to může být nejsnadnější důkaz optimality.

Máme-li duální optimální řešení, jak z něj co nejlevněji spočítat primární optimální řešení? Obecně je k tomu nutno vyřešit soustavu lineárních nerovnic (což není o moc snadnější než vyřešit lineární program). Někdy ale postačí vyřešit soustavu rovnic.

Example 15.2. Je dána primární úloha z Příkladu 15.1. Zkuste dokázat bez použití algoritmu na řešení LP, že $\mathbf{x} = (x_1, x_2, x_3) = (1.2, 0.6, 0)$ je optimální řešení primární úlohy (přičemž optimální duální řešení $\mathbf{y}$ není zadáno).

Pomocí věty o komplementaritě zkusíme z daného optimálního $\mathbf{x}$ zkusíme spočítat optimální $\mathbf{y}$. Protože jsou druhé a čtvrté primární omezení neaktivní, z komplementarity plyne $y_2 = y_4 = 0$. Protože $x_1 > 0$ a $x_2 > 0$, z komplementarity musí být první a druhé duální omezení aktivní. Máme tedy soustavu lineárních rovnic

$$
\begin{aligned}
2y_1 + y_3 &= 2 \\
y_1 + 3y_3 &= 5
\end{aligned}
\tag{15.6}
$$

která má jediné řešení $(y_1, y_3) = (0.2, 1.6)$. Tedy $\mathbf{y} = (0.2, 0, 1.6, 0)$. Toto duální řešení je přípustné (tj. splňuje všechna duální omezení). Protože se hodnota primární účelové funkce v bodě $\mathbf{x}$ rovná hodnotě duální účelové funkce v bodě $\mathbf{y}$, musejí být $\mathbf{x}$ a $\mathbf{y}$ optimální řešení.

Tento postup nemusí vést vždy k cíli. Pokud by duální úloha by měla nekonečně mnoho optimálních řešení, soustava (15.6) by měla nekonečně mnoho řešení (měla by např. více proměnných než neznámých). Z nich by bylo nutno vybrat přípustná duální řešení, tedy $\mathbf{y} \geq \mathbf{0}$. Museli bychom tedy řešit soustavu rovnic *a nerovnic*. □

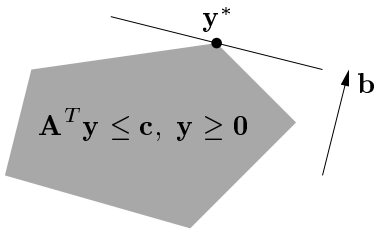
Zkoumejme, jak se změní optimální hodnota úlohy $\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \,\}$, jestliže nepatrně změníme pravé strany omezení $\mathbf{b}$. Odpověď je snadno vidět v duálu.

Theorem 15.4 (o stínových cenách). *Let funkce $f\colon \mathbb{R}^m \to \mathbb{R}$ je definována jako*

$$f(\mathbf{b}) = \min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{b},\ \mathbf{x} \geq \mathbf{0}\,\} = \max\{\, \mathbf{b}^T\mathbf{y} \mid \mathbf{A}^T\mathbf{y} \leq \mathbf{c},\ \mathbf{y} \geq \mathbf{0}\,\}.$$

Jestliže má duální úloha pro dané $\mathbf{b}$ jediné optimální řešení $\mathbf{y}^*$, pak je funkce f v bodě $\mathbf{b}$ diferencovatelná a platí $f'(\mathbf{b}) = \mathbf{y}^{*T}$, neboli $\partial f(\mathbf{b})/\partial b_i = \mathbf{y}_i^*$.

Proof. Je-li $\mathbf{y}^*$ duální optimální řešení pro dané $\mathbf{b}$, je $f(\mathbf{b}) = \mathbf{b}^T\mathbf{y}^*$. Jelikož je toto optimální řešení jediné, nabývá se ve vrcholu polyedru přípustných řešení $\{\, \mathbf{y} \in \mathbb{R}^m \mid \mathbf{A}^T\mathbf{y} \leq \mathbf{c},\ \mathbf{y} \geq \mathbf{0}\,\}$, viz obrázek:



Změníme-li nepatrně $\mathbf{b}$, optimální duální řešení $\mathbf{y}^*$ se nezmění a zůstane jediné (toto odůvodnění není zcela rigorózní, ale geometricky je dostatečně názorné). Tedy při malé změně vektoru $\mathbf{b}$ je hodnota optima stále rovna $f(\mathbf{b}) = \mathbf{b}^T\mathbf{y}^*$. Derivací získáme $f'(\mathbf{b}) = \mathbf{y}^{*T}$. □

Zdůrazněme předpoklad jednoznačnosti optimálního řešení. Kdyby množina duálních optimálních řešení byla ne jediný vrchol, ale stěna vyšší dimenze, po malé změně účelového vektoru $\mathbf{b}$ by se optimální stěna mohla stát vrcholem a funkce f by tedy v bodě $\mathbf{b}$ nebyla diferencovatelná.

Protože $\mathbf{b}$ je zároveň vektor pravých stran primární úlohy, optimální duální proměnné $\mathbf{y}^*$ vyjadřují *citlivost* optima primární úlohy na změnu pravých stran primárních omezení $\mathbf{A}\mathbf{x} \geq \mathbf{b}$. Interpretujeme-li naše LP jako optimální výrobní plán (12.7) (pozor, liší se obrácenou nerovností v omezení), pak hodnota y_i^* říká, jak by se náš výdělek zvětšil, kdybychom trochu uvolnili omezení na výrobní zdroje $\mathbf{a}_i^T\mathbf{x} \leq b_i$. V ekonomii se proto duálním proměnným říká **stínové ceny** primárních omezení.

Všimněte si, že věta o stínových cenách je ve shodě s větou o komplementaritě. Pokud $y_i^* = 0$, je $\mathbf{a}_i^T\mathbf{x} < b_i$, tedy malá změna b_i nemá na optimum vliv.

Example 15.3. Let je známo, že duální úloha v Příkladu 15.1 má jediné optimální řešení. Stínová cena prvního primárního omezení $2x_1+x_2+2x_3 \geq 3$ je $y_1 = 0.2$. Změňme pravou stranu $b_1 = 3$ tohoto omezení o malou hodnotu $h = 0.01$ a zkoumejme, jak se změní optimum. Tato změna nezmění argument $\mathbf{y}^*$ duálního optima, pouze jeho hodnotu $\mathbf{b}^T\mathbf{y}^*$. Podle silné duality hodnota primárního optima musí být rovna hodnotě duálního optima (argument $\mathbf{x}^*$ primárního optima se může nějak změnit, to nás ale nezajímá). Dvojice úloh tedy bude vypadat takto:

$$\begin{array}{rrcr}
\min & 2x_1 + 5x_2 + 6x_3 & = & \mathbf{5.402} \\
 & 2x_1 + x_2 + 2x_3 & \geq & 3.01 \\
 & x_1 + 2x_2 + 2x_3 & \geq & 1 \\
 & x_1 + 3x_2 + x_3 & \geq & 3 \\
 & -x_1 + x_2 - 2x_3 & \geq & -1 \\
 & x_1 & \geq & 0 \\
 & x_2 & \geq & 0 \\
 & x_3 & \geq & 0
\end{array}
\qquad
\begin{array}{rrcr}
\max & 3.01y_1 + y_2 + 3y_3 - y_4 & = & \mathbf{5.402} \\
\mathbf{0.2} = & y_1 & \geq & 0 \\
\mathbf{0} = & y_2 & \geq & 0 \\
\mathbf{1.6} = & y_3 & \geq & 0 \\
\mathbf{0} = & y_4 & \geq & 0 \\
\mathbf{2} = & 2y_1 + y_2 + y_3 - y_4 & \leq & 2 \\
\mathbf{5} = & y_1 + 2y_2 + 3y_3 + y_4 & \leq & 5 \\
\mathbf{2} = & 2y_1 + 2y_2 + y_3 - 2y_4 & \leq & 6
\end{array}$$

V okolí bodu $\mathbf{b} = (3, 1, 3, -1)$, ve kterém se nemění optimální $\mathbf{y}^*$, bude $f(\mathbf{b}) = \mathbf{b}^T\mathbf{y}^*$ a tedy hodnota společného optima se změní o $h \cdot \partial f(\mathbf{b})/\partial b_1 = h \cdot y_1 = 0.2 \cdot 0.01$ na 5.402. □

15.3 Příklady na konstrukci a interpretaci duálních úloh

Dualita umožňuje *vhled* do řešeného problému, často velmi netriviální. Abychom danou úlohu (fyzikální, ekonomickou či jinou) popsanou lineárním programem porozuměli do hloubky, je často třeba pochopit význam nejen primární úlohy, ale i duální úlohy a vět o dualitě.

Example 15.4 (Ekonomická interpretace duality). Vrať me se k Příkladu 12.6 o výrobci lupínků a hranolků z brambor a oleje. Napišme k této úloze duální úlohu:

$$\begin{array}{rrcl} \max & 120l + 76h & & \\ \text{za podm.} & 2l + 1.5h & \leq & 100 \\ & 0.4l + 0.2h & \leq & 16 \\ & l & \geq & 0 \\ & h & \geq & 0 \end{array} \qquad \begin{array}{rrcl} \min & 100a + 16b & & \\ \text{za podm.} & a & \geq & 0 \\ & b & \geq & 0 \\ & 2a + 0.4b & \geq & 120 \\ & 1.5a + 0.2b & \geq & 76 \end{array}$$

Přijde překupník a chce koupit od výrobce jeho zásoby brambor a oleje. Překupník řeší tuto otázku: Jaké nejnižší ceny musím nabídnout, aby mi výrobce své zásoby prodal? Tvrdíme, že toto je význam duální úlohy.

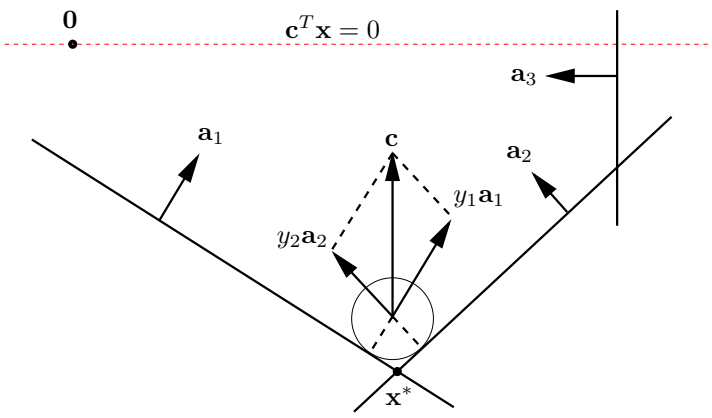
Vskutku, Let a, b označují nabízenou cenu za jednotku brambor a oleje. Překupník chce minimalizovat celkovou cenu za suroviny $100a + 16b$. Musí být $2a + 0.4b \geq 120$, neboť jinak by výrobci více vyplatilo vyrobit ze všech brambor a oleje lupínky a prodat je, než prodat suroviny. Ze stejného důvodu musí být $1.5a + 0.2b \geq 76$. Optimální duální řešení je $a = 32$ a $b = 140$.

Toto je další důvod (kromě Věty 15.4), proč se optimálním duálním proměnným někdy říká *stínové ceny* odpovídajících primárních omezení. Např. stínová cena brambor je 32 Kč/kg. □

Example 15.5 (Fyzikální interpretace duality). Uvažujme dvojici duálních úloh

$$\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{b},\ \mathbf{x} \in \mathbb{R}^n \,\} = \max\{\, \mathbf{b}^T\mathbf{y} \mid \mathbf{A}^T\mathbf{y} = \mathbf{c},\ \mathbf{y} \geq \mathbf{0} \,\}.$$

Uvažujme následující 'analogový počítač'. Mějme polyedr tvořený třemi poloprostory $\mathbf{a}_i^T\mathbf{x} \geq b_i$ a vektor $\mathbf{c}$ míříci svisle vzhůru:



Hoďme do polyedru malý míček, na který působí tíhová síla $-\mathbf{c}$. Míček s pozicí $\mathbf{x}$ má potenciální energii $\mathbf{c}^T\mathbf{x}$. Míček se bude pohybovat do té doby, než nalezne místo s nejmenší potenciální energií, což je nejnižší vrchol $\mathbf{x}^*$. Tedy $\mathbf{x}^*$ je řešením primární úlohy.

V bodě $\mathbf{x}^*$ je míček v klidu a proto pro něj platí rovnováha sil: tíha $-\mathbf{c}$ se vyrovnává silami stěn. Tedy existují skaláry $y_i^* \geq 0$ tak, že $\mathbf{c} = \sum_i y_i^* \mathbf{a}_i = \mathbf{A}^T\mathbf{y}^*$. Skaláry y_i^* jsou nezáporné, protože stěny působí silou jen dovnitř polyedru, ne ven.

Pokud $\mathbf{a}_i^T\mathbf{x}^* > b_i$, míček se i-té stěny nedotýká. V tom případě je síla stěny na míček nutně nulová, $y_i^* = 0$. Proto pro každé i platí $y_i^*(\mathbf{a}_i^T\mathbf{x}^* - b_i) = 0$, což jsou podmínky (15.3). Dle věty o komplementaritě tedy je $\mathbf{c}^T\mathbf{x}^* = \mathbf{b}^T\mathbf{y}^*$.

Zdůrazněme, že tato úvaha *nedokazuje* žádnou z vět o dualitě. Předpokládá totiž platnost fyzikálních zákonů, které nelze matematicky dokázat ale pouze experimentálně pozorovat. □

Example 15.6. Mějme úlohu

$$\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{1}^T\mathbf{x} = 1,\ \mathbf{x} \geq \mathbf{0} \,\} = \min\left\{ \sum_{i=1}^n c_i x_i \,\middle|\, \sum_{i=1}^n x_i = 1,\ x_i \geq 0 \right\},$$

kde $\mathbf{c} = (c_1, \ldots, c_n)$ je dáno a optimalizuje se přes $\mathbf{x} = (x_1, \ldots, x_n)$. Napište duální úlohu a iterpretujte věty o silné dualitě a komplementaritě.

Úvahou snadno vidíme (viz Příklad 12.2), že optimální hodnota je $\min_{i=1}^n c_i$ a nabývá se ve vektoru $\mathbf{x}$ jehož všechny složky jsou nulové kromě složek příslušných minimálnímu c_i. Pokud je více minimálních prvků c_i, optimální $\mathbf{x}$ není dáno jednoznačně. Např. pro $\mathbf{c} = (1, 3, 1, 2)$ bude optimálním řešením každé $\mathbf{x} = (x_1, 0, x_3, 0)$ pro $x_1, x_3 \geq 0$ splňující $x_1 + x_3 = 1$.

Podle návodu na konstrukci duální úlohy dostaneme duál

$$\max\{\, y \in \mathbb{R} \mid y\mathbf{1} \leq \mathbf{c} \,\} = \max\{\, y \in \mathbb{R} \mid y \leq c_i,\ i = 1, \ldots, n \,\}.$$

Neboli hledá se největší číslo y, které je menší než všechna čísla c_i. Takové číslo y se rovná minimu z čísel c_i. Tedy platí silná dualita.

Podmínky komplementarity říkají, že v optimech bude alespoň jedno z odpovídající dvojice primární-duální omezení aktivní. Dvojice omezení $\sum_i x_i = 1$, $y \in \mathbb{R}$ splňuje podmínky komplementarity triviálně. Dvojice omezení $x_i \geq 0$, $y \leq c_i$ je splňuje právě tehdy, když je splněna aspoň jedna z rovností $x_i = 0$, $y = c_i$. To znamená:

- Pokud je v duálu $y < c_i$, v primáru musí být $x_i = 0$. To je ale jasné, protože $y < c_i$ znamená, že c_i není nejmenší ze složek vektoru $\mathbf{c}$ a tudíž (dle úvahy výše) mu v primáru nemůžeme přiřadit nenulovou váhu x_i.
- Obráceně, pokud je v primáru $x_i > 0$, musí být v duálu $y = c_i$. To je jasné, protože pokud jsme v primáru přiřadili číslu c_i nenulovou váhu, musí být nejmenší. □

Example 15.7. Z §12.3 víme, že optimální argument úlohy

$$\begin{aligned}\min_{x\in\mathbb{R}} \sum_{i=1}^n |x - a_i| &= \min\left\{ \sum_{i=1}^n z_i \,\middle|\, z_i \in \mathbb{R},\ x \in \mathbb{R},\ -z_i \leq x - a_i \leq z_i \right\} \\ &= \min\{\, \mathbf{1}^T\mathbf{z} \mid \mathbf{z} \in \mathbb{R}^n,\ x \in \mathbb{R},\ -\mathbf{z} \leq \mathbf{1}x - \mathbf{a} \leq \mathbf{z} \,\} \qquad (15.7)\end{aligned}$$

je medián z čísel $a_1, \ldots, a_n$. Napište duální úlohu a co nejvíce ji zjednodušte. Úvahou nalezněte optimální hodnotu primární a duální úlohy a ověřte, že se (dle silné duality) rovnají.

Ryclý způsob jak vytvořit duální úlohu je podle předpisu v §15.1, to ovšem vyžaduje zkušenost a snadno se v tom udělá chyba. Zdlouhavější avšak bezpečnější způsob je přes maticovou formu. Primární a duální úlohu napíšeme v maticovém tvaru, kde zvolíme názvy matic tak, aby nekolidovaly s (15.7):

$$\begin{array}{rl} \min & \mathbf{h}^T\mathbf{u} \\ \text{za podm.} & \mathbf{F}\mathbf{u} \geq \mathbf{g} \\ & \mathbf{u} \in \mathbb{R}^{1+n} \end{array} \qquad \begin{array}{rl} \max & \mathbf{g}^T\mathbf{v} \\ \text{za podm.} & \mathbf{v} \geq \mathbf{0} \\ & \mathbf{F}^T\mathbf{v} = \mathbf{h} \end{array}$$

Matice zvolíme tak, aby primární úloha odpovídala úloze (15.7):

$$\mathbf{F} = \begin{bmatrix} \mathbf{1} & \mathbf{I} \\ -\mathbf{1} & \mathbf{I} \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} \mathbf{a} \\ -\mathbf{a} \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} 0 \\ \mathbf{1} \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} x \\ \mathbf{z} \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} \mathbf{p} \\ \mathbf{q} \end{bmatrix}.$$

Vektor duálních proměnných $\mathbf{v}$ jsme zároveň rozdělili na dva bloky $\mathbf{p}, \mathbf{q}$, odpovídající blokům matic $\mathbf{F}$ a $\mathbf{g}$. Vynásobním matic přepíšeme duální úlohu do tvaru (ověřte!)

$$\begin{aligned} &\max\{\, \mathbf{a}^T(\mathbf{p}-\mathbf{q}) \mid \mathbf{1}^T(\mathbf{p}-\mathbf{q}) = 0,\ \mathbf{p}+\mathbf{q} = \mathbf{1},\ \mathbf{p} \geq \mathbf{0},\ \mathbf{q} \geq \mathbf{0} \,\} \\ &= \max\left\{ \sum_{i=1}^n a_i(p_i - q_i) \;\middle|\; \sum_{i=1}^n (p_i - q_i) = 0,\ p_i + q_i = 1,\ p_i \geq 0,\ q_i \geq 0 \right\} \end{aligned} \tag{15.8}$$

Úlohu (15.8) lze dále zjednodušit substitucí

$$2p_i = 1 + t_i, \quad 2q_i = 1 - t_i.$$

Po této substituci je $p_i - q_i = t_i$ a podmínka $p_i + q_i = 1$ je splněna automaticky. Podmínka $\sum_i (p_i - q_i) = 0$ odpovídá podmínce $\sum_i t_i = 0$. Podmínka $p_i \geq 0$ odpovídá $t_i \geq -1$ a podmínka $q_i \geq 0$ odpovídá $t_i \leq 1$. Duální úloha s novými proměnnými $\mathbf{t} \in \mathbb{R}^n$ je tedy

$$\max\left\{ \sum_{i=1}^n a_i t_i \;\middle|\; \sum_{i=1}^n t_i = 0,\ -1 \leq t_i \leq 1 \right\} \;=\; \max\{\, \mathbf{a}^T\mathbf{t} \mid \mathbf{1}^T\mathbf{t} = 0,\ -\mathbf{1} \leq \mathbf{t} \leq \mathbf{1} \,\}. \tag{15.9}$$

Primární úloha (15.7) a duální úloha (15.9) spolu zdánlivě vůbec nesouvisejí – avšak podle silné duality jejich optimální hodnoty musí být stejné! Zkusme pochopit, proč tomu tak je.

Nejprve si všimneme, že optimální hodnota primární úlohy (15.7) se nezmění, posuneme-li čísla $a_1, \ldots, a_n$ o libovolnou konstantu $b \in \mathbb{R}$. To je jasné, neboť medián x se posune o stejnou konstantu a je $|(x-b) - (a_i - b)| = |x - a_i|$. Totéž platí pro duální úlohu (15.9), neboť díky podmínce $\sum_i t_i = 0$ je $\sum_i (a_i - b)t_i = \sum_i a_i t_i$. Proto bez ztráty obecnosti můžeme zvolit $b = \operatorname{median}_i a_i$, neboli posunout body tak, že jejich medián bude $x = 0$.

Nyní je primární optimální hodnota rovna jednoduše $\sum_i |x - a_i| = \sum_i |a_i|$. Protože kladných a záporných čísel a_i je stejný počet, duální úloha nabývá optima v takovém vektoru $\mathbf{t}$, kde $t_i = -1$ pro $a_i < 0$ a $t_i = 1$ pro $a_i > 0$ (což splňuje podmínku $\sum_i t_i = 0$). Tedy duální optimální hodnota je také $\sum_i a_i t_i = \sum_i |a_i|$. □

15.4 Cvičení

15.1. Ukažte pro dvojici úloh LP v §15.1, že duál duálu se rovná původní úloze. Musíte nejdříve duální úlohu (prostřední sloupec) vpravo převést do tvaru primární úlohy (první sloupec), tj. např. musíte převést maximalizaci na minimalizaci.

15.2. Napište duální úlohu a podmínky komplementarity k následujícím úlohám. Pokud úloha není LP, nejdříve převeďte na LP (dle §12.1). Výslednou duální úlohu co nejvíce zjednodušte, příp. převeďte do skalární formy, je-li skalární forma výstižnější.

a) $\min_{x\in\mathbb{R}} \max_{i=1}^{n} |a_i - x|$ (střed intervalu)
b) úloha (12.13)
c) úloha (12.15)
d) dopravní problém (12.9)
e) všechny úlohy ze Cvičení 12.2
f) úloha vzniklá ve Cvičení 12.7
g) $\min_{\mathbf{x}\in\mathbb{R}^n} \max_{i=1}^{m} (\mathbf{a}_i^T\mathbf{x} + b_i)$ (viz §12.1.1)

($\star$) Pro každou úlohu se pokuste nalézt význam duální úlohy, podobně jako v Příkladu 15.6.

15.3. Dokažte bez užití algoritmu na řešení LP, že $\mathbf{x} = (1, 1, 1, 1)$ je optimální řešení úlohy

$$\begin{array}{ll} \min & \begin{bmatrix} 47 & 93 & 17 & -93 \end{bmatrix} \mathbf{x} \\ \text{za podm.} & \begin{bmatrix} -1 & -6 & 1 & 3 \\ -1 & -2 & 7 & 1 \\ 0 & 3 & -10 & -1 \\ -6 & -11 & -2 & 12 \\ 1 & 6 & -1 & -3 \end{bmatrix} \mathbf{x} \le \begin{bmatrix} -3 \\ 5 \\ -8 \\ -7 \\ 4 \end{bmatrix} \end{array}$$

Hints and Solutions

15.2.d) $\max\left\{ \sum_{i=1}^{m} a_i p_i + \sum_{j=1}^{n} b_j q_j \mid p_i \in \mathbb{R},\ q_i \in \mathbb{R},\ p_i + q_j \le c_{ij} \right\}$

15.2.f) Duál: $\max\{ \sum_{i=1}^{n} y_i d_i + \sum_{i=1}^{n'} y'_i d'_i \mid \sum_{i=1}^{n} y_i = \sum_{i=1}^{n'} y'_i,\ y_i \le m_i,\ \ y'_i \le m'_i,\ y_i, y'_i \ge 0 \}$.
Podmínky komplementarity: $z_i(y_i - m_i) = 0$, $z'_i(y'_i - m'_i) = 0$, $(z_i - d_i - x)y_i = 0$, $(z'_i - d'_i + x)y'_i = 0$.

Chapter 16

Konvexní funkce

Definition 16.1. *Funkce* $f\colon \mathbb{R}^n \to \mathbb{R}$ *je* **konvexní** *na konvexní množině* $X \subseteq \mathbb{R}^n$*, jestliže*

$$\mathbf{x} \in X,\ \mathbf{y} \in X,\ 0 \le \alpha \le 1 \quad \Longrightarrow \quad f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \le \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}). \tag{16.1}$$

Funkce f *je* **konkávní** *na množině* X*, jestliže je funkce* $-f$ *konvexní na množině* X*.*

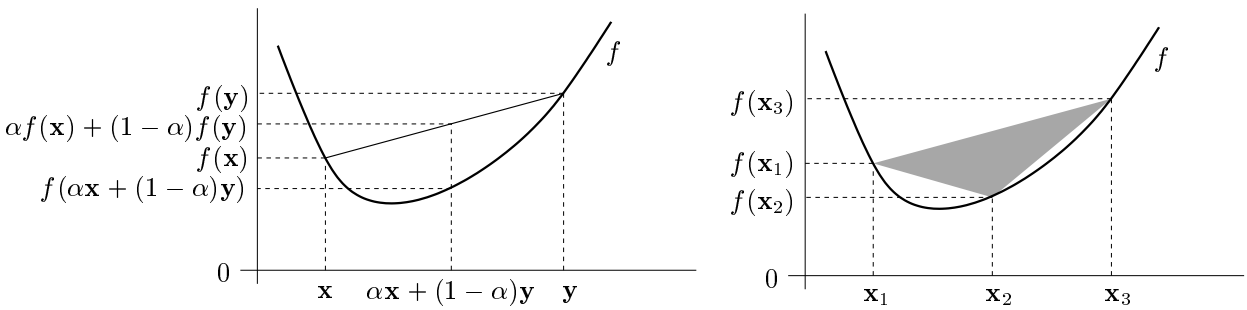
Rozlišujte pojem *konvexní množina* a *konvexní funkce*, jde o různé věci. Dále si všimněte, že X musí být konvexní množina – pojem konvexní funkce na nekonvexní množině nemá smysl. Pokud X je celý definiční obor funkce f, odkaz na X můžeme vynechat a říkáme pouze, že funkce f je konvexní.

Podmínku (16.1) lze zobecnit pro více než dva body: funkce f je konvexní právě tehdy, když

$$\mathbf{x}_1, \dots, \mathbf{x}_k \in X,\ \ \alpha_1, \dots, \alpha_k \ge 0,\ \ \alpha_1 + \dots + \alpha_k = 1 \quad \Longrightarrow \\ f(\alpha_1\mathbf{x}_1 + \dots + \alpha_k\mathbf{x}_k) \le \alpha_1 f(\mathbf{x}_1) + \dots + \alpha_k f(\mathbf{x}_k). \tag{16.2}$$

Podmínka (16.2) zjevně implikuje (16.1) a indukcí lze dokázat, že to platí i naopak. Podmínce (16.2) se někdy říká **Jensenova nerovnost**. Porovnejte s definicí lineárního zobrazení (3.2)!

Geometrický význam podmínky (16.1) je ten, že úsečka spojující body $(\mathbf{x}, f(\mathbf{x}))$ a $(\mathbf{y}, f(\mathbf{y}))$ leží nad grafem funkce (viz levý obrázek). Geometrický význam podmínky (16.2) je ten, že konvexní polyedr vybarvený šedě (viz pravý obrázek) leží nad grafem funkce. Podrobně rozmyslete, jak tyto geometrické interpretace odpovídají výrazům (16.1) a (16.2)!



Důkaz konvexity funkce z Definice 16.1 vyžaduje někdy kreativitu, neexistuje na to mechanický postup. Chceme-li dokázat, že funkce není konvexní, stačí nalézt jedinou trojici $(\mathbf{x}, \mathbf{y}, \alpha)$ porušující (16.1) – její 'uhodnutí' však také vyžaduje, abychom měli o funkci představu.

Example 16.1. Dokažme z Definice 16.1, že funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ definovaná jako $f(\mathbf{x}) = \max_{i=1}^n x_i$ je konvexní. Máme dokázat, že pro každé $\mathbf{x}$, $\mathbf{y}$ a $0 \le \alpha \le 1$ platí

$$\begin{aligned} f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) &= \max_i(\alpha x_i + (1-\alpha)y_i) && \text{(16.3a)}\\ &\le \max_i \alpha x_i + \max_i (1-\alpha)y_i && \text{(16.3b)}\\ &= \alpha \max_i x_i + (1-\alpha)\max_i y_i = \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) && \text{(16.3c)} \end{aligned}$$

kde rovnost (16.3c) plyne z nezápornosti čísel α a $1-\alpha$.

Nerovnost (16.3b) plyne z toho, že pro každé $a_1, \dots, a_n, b_1, \dots, b_n \in \mathbb{R}$ platí

$$\max_i(a_i + b_i) \le \max_i a_i + \max_i b_i. \tag{16.4}$$

Nerovnost (16.4) dokážeme takto. Let i^*, j^*, k^* jsou indexy, ve kterých se nabývají maxima, tedy $a_{i^*} + b_{i^*} = \max_i(a_i + b_i)$, $a_{j^*} = \max_i a_i$, $b_{k^*} = \max_i b_i$. Proto $a_{i^*} \le a_{j^*}$ a $b_{i^*} \le b_{k^*}$. Tedy $\max_i(a_i + b_i) = a_{i^*} + b_{i^*} \le a_{j^*} + b_{k^*} = \max_i a_i + \max_i b_i$. □

Example 16.2. Dokažme z Definice 16.1, že funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ definovaná jako $f(\mathbf{x}) = \min_{i=1}^n x_i$ není konvexní. Např. volba $n = 2$, $\mathbf{x} = (0, 2)$, $\mathbf{y} = (2, 0)$, $\alpha = \frac{1}{2}$ nesplňuje (16.1), nebot'

$$f((\mathbf{x} + \mathbf{y})/2) = f(1, 1) = 1 > (f(\mathbf{x}) + f(\mathbf{y}))/2 = (0 + 0)/2 = 0.$$ □

Použitím Jensenovy nerovnosti na vhodnou konvexní funkci lze získat mnoho užitečných nerovností.

Example 16.3. Funkce log je konkávní na $\mathbb{R}_{++}$. Napišme pro tuto funkci Jensenovu nerovnost (16.2) (jelikož funkce je konkávní a ne konvexní, musíme v Jensenově nerovnosti obrátit znaménko nerovnosti), ve které položíme $\alpha_1 = \dots = \alpha_n = \frac{1}{n}$:

$$\log \frac{x_1 + \dots + x_n}{n} \ge \frac{\log x_1 + \dots + \log x_n}{n}$$

kde $x_1, \dots, x_n$ jsou kladné. Vezmeme-li exponenciálu každé strany, dostaneme

$$\frac{x_1 + \dots + x_n}{n} \ge (x_1 \cdots x_n)^{1/n}.$$

Tato známá nerovnost říká, že aritmetický průměr není menší než geometrický. □

Example 16.4. Uveďme často potkávané jednoduché konvexní či konkávní funkce:

1. Exponenciála $f(x) = \mathrm{e}^{ax}$ je konvexní na $\mathbb{R}$, pro libovolné $a \in \mathbb{R}$.
2. Mocnina $f(x) = x^a$ je na $\mathbb{R}_{++}$ konvexní pro $a \ge 1$ nebo $a \le 0$ a konkávní pro $0 \le a \le 1$.
3. Mocnina absolutní hodnoty $f(x) = |x|^a$ je pro $a \ge 1$ konvexní na $\mathbb{R}$ (speciálně: absolutní hodnota $|x|$ je konvexní).
4. Logaritmus $f(x) = \log x$ je konkávní na $\mathbb{R}_{++}$.
5. Záporná entropie $f(x) = x \log x$ je konvexní na $\mathbb{R}_{++}$ (nebo i na $\mathbb{R}_+$, pokud dodefinujeme $0 \log 0 = 0$, což se často dělá, protože $\lim_{x\to 0+} x \log x = 0$).
6. Afinní funkce $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$ je zároveň konvexní i konkávní.

7. Kvadratická forma $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$ je konvexní pro $\mathbf{A}$ pozitivně semidefinitní, konkávní pro $\mathbf{A}$ negativně semidefinitní, a nekonvexní a nekonkávní pro $\mathbf{A}$ indefinitní (viz Příklad 16.5).
8. Maximum složek $f(\mathbf{x}) = \max_{i=1}^n x_i = \max\{x_1, \dots, x_n\}$ je konvexní na $\mathbb{R}^n$.
9. Log-sum-exp funkce $f(\mathbf{x}) = \log(\mathrm{e}^{x_1} + \cdots + \mathrm{e}^{x_n})$ je konvexní. Tato funkce se někdy nazývá *měkké maximum*, neboť funkce
$$f_t(\mathbf{x}) = f(t\mathbf{x})/t = \log(\mathrm{e}^{tx_1} + \cdots + \mathrm{e}^{tx_n})/t$$
se pro $t \to +\infty$ blíží funkci $\max_{i=1}^n x_i$ (dokažte výpočtem limity!).
10. Geometrický průměr $f(\mathbf{x}) = (x_1 \cdots x_n)^{1/n}$ je konkávní na $\mathbb{R}^n_+$.
11. Každá norma (viz Definice 12.1) je konvexní funkce, neboť pro každé $0 \le \alpha \le 1$ máme
$$\|\alpha\mathbf{x} + (1-\alpha)\mathbf{y}\| \le \|\alpha\mathbf{x}\| + \|(1-\alpha)\mathbf{y}\| = \alpha\|\mathbf{x}\| + (1-\alpha)\|\mathbf{y}\|,$$
kde nerovnost plyne z trojúhelníkové nerovnosti a rovnost z homogenity.

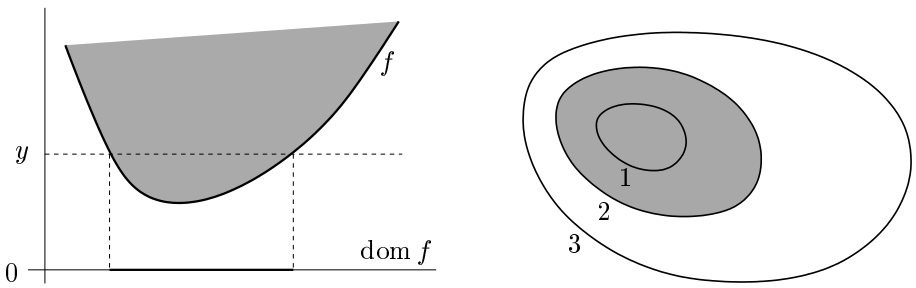
Nakreslete či představte si vrstevnice a grafy těchto funkcí (v případě více proměnných pro $n = 1$ a $n = 2$)! □

16.1 Vztah konvexní funkce a konvexní množiny

Zopakujte si pojmy vrstevnice a graf funkce z §1.1.3! Zavedeme dva podobné pojmy, které se liší pouze nahrazením rovnosti nerovností. Pro funkci $f\colon \mathbb{R}^n \to \mathbb{R}$ definujeme:

- **Subkontura**[1] výšky y je množina $\{\, \mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) \le y \,\}$.
- **Epigraf** funkce je množina $\{\, (\mathbf{x}, y) \in \mathbb{R}^{n+1} \mid \mathbf{x} \in \mathbb{R}^n,\ f(\mathbf{x}) \le y \,\}$.

Levý obrázek znázorňuje subkonturu výšky y a epigraf funkce $\mathbb{R} \to \mathbb{R}$, pravý obrázek subkonturu výšky 2 funkce $\mathbb{R}^2 \to \mathbb{R}$:



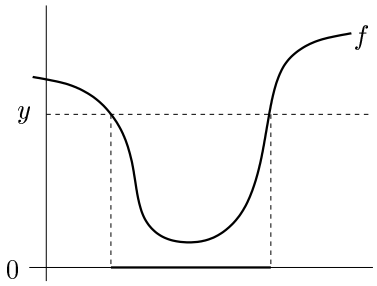
Existují těsné vztahy mezi konvexitou funkce a konvexitou jejího epigrafu a subkontur (což jsou množiny), dané následujícími větami.

Theorem 16.1. *Je-li f konvexní funkce, pak je každá subkontura této funkce konvexní množina.*

Proof. Předpokládejme, že body $\mathbf{x}_1$ a $\mathbf{x}_2$ patří do subkontury, tedy $f(\mathbf{x}_1) \le y$ a $f(\mathbf{x}_2) \le y$. Pro každé $0 \le \alpha \le 1$ platí
$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \le \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2) \le \alpha y + (1-\alpha) y = y,$$
kde první nerovnost plyne z konvexity funkce a druhá z nerovností $f(\mathbf{x}_1) \le y$, $f(\mathbf{x}_2) \le y$. Tedy bod $\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2$ patří do subkontury, která je proto konvexní množina. □

[1] Slovo 'subkontura' je pokus o český překlad anglického 'sublevel set'.

Obrácená implikace ve Větě 16.1 neplatí: snadno najdeme funkci, která není konvexní a jejíž každá subkontura je konvexní množina[2]. Příklad je na obrázku:



Theorem 16.2. *Funkce f je konvexní právě tehdy, když její epigraf je konvexní množina.*

Proof. Předpokládejme, že funkce f je konvexní. Vezměme dva body $(\mathbf{x}_1, y_1)$ a $(\mathbf{x}_2, y_2)$ z epigrafu, tedy $f(\mathbf{x}_1) \leq y_1$ a $f(\mathbf{x}_2) \leq y_2$. Pro každé $0 \leq \alpha \leq 1$ platí

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha)f(\mathbf{x}_2) \leq \alpha y_1 + (1-\alpha)y_2,$$

kde první nerovnost plyne z konvexity funkce a druhá nerovnost z $f(\mathbf{x}_1) \leq y_1$ a $f(\mathbf{x}_2) \leq y_2$. Tedy bod $\alpha(\mathbf{x}_1, y_1) + (1-\alpha)(\mathbf{x}_2, y_2)$ patří do epigrafu, který je proto konvexní množina.

Předpokládejme, že epigraf je konvexní množina. Tedy pokud body $(\mathbf{x}_1, y_1)$ a $(\mathbf{x}_2, y_2)$ patří do epigrafu, pak také bod $\alpha(\mathbf{x}_1, y_1) + (1-\alpha)(\mathbf{x}_2, y_2)$ patří do epigrafu pro každé $0 \leq \alpha \leq 1$. Volbou $y_1 = f(\mathbf{x}_1)$ a $y_2 = f(\mathbf{x}_2)$ máme

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha y_1 + (1-\alpha)y_2 = \alpha f(\mathbf{x}_1) + (1-\alpha)f(\mathbf{x}_2),$$

proto je funkce f konvexní. □

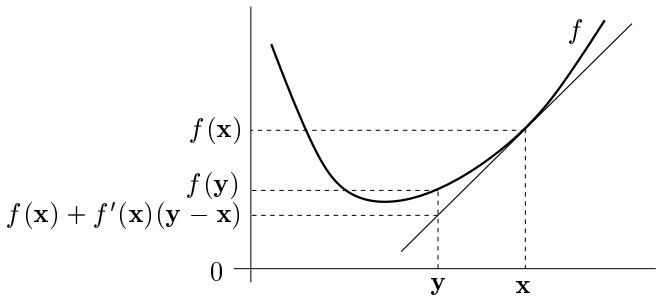
16.2 Konvexita diferencovatelných funkcí

Konvexní funkce nemusí být v každém bodě diferencovatelná (uvažte např. funkci $f(x) = |x|$). Pokud je ale funkce jednou či dvakrát diferencovatelná, její konvexitu lze snadněji než pomocí Definice 16.1 charakterizovat pomocí derivací. Následující dvě věty uvedeme bez důkazů.

Theorem 16.3 (Podmínka prvního řádu). *Let funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ je diferencovatelná na konvexní množině $X \subseteq \mathbb{R}^n$. Funkce f je konvexní na množině X právě tehdy, když*

$$\mathbf{x} \in X,\ \mathbf{y} \in X \quad \Longrightarrow \quad f(\mathbf{y}) \geq f(\mathbf{x}) + f'(\mathbf{x})\,(\mathbf{y} - \mathbf{x}).$$

To znamená, že Taylorův polynom prvního řádu funkce f v každém bodě $\mathbf{x} \in X$ (viz (8.11b)) je všude (tj. pro každé $\mathbf{y}$) menší nebo roven funkci f:



[2] Funkce, jejíž každá subkontura je konvexní množina, se nazývá *kvazikonvexní*. Kvazikonvexní funkce nejsou zdaleka tak hezké jako konvexní funkce.

Theorem 16.4 (Podmínka druhého řádu). *Let $X \subseteq \mathbb{R}^n$ je konvexní množina, která má pouze vnitřní body. Let funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ je dvakrát diferencovatelná na X. Funkce f je konvexní na množině X právě tehdy, když v každém bodě $\mathbf{x} \in X$ je Hessova matice $f''(\mathbf{x})$ pozitivně semidefinitní.*

Example 16.5. Let $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$, kde $\mathbf{A}$ je symetrická pozitivně semidefinitní. Ukažme konvexitu této funkce třemi způsoby:

- Dokažme konvexitu z Věty 16.4. To je triviální, protože Hessián je $f''(\mathbf{x}) = 2\mathbf{A}$ a tedy je pozitivně semidefinitní.
- Dokažme konvexitu z Věty 16.3. Protože $f'(\mathbf{x}) = 2\mathbf{x}^T\mathbf{A}$, máme dokázat, že

  $$\mathbf{y}^T\mathbf{A}\mathbf{y} \geq \mathbf{x}^T\mathbf{A}\mathbf{x} + 2\mathbf{x}^T\mathbf{A}(\mathbf{y} - \mathbf{x}).$$

  To jde upravit na $\mathbf{x}^T\mathbf{A}\mathbf{x} - 2\mathbf{x}^T\mathbf{A}\mathbf{y} + \mathbf{y}^T\mathbf{A}\mathbf{y} \geq 0$. Platí[3]

  $$\mathbf{x}^T\mathbf{A}\mathbf{x} - 2\mathbf{x}^T\mathbf{A}\mathbf{y} + \mathbf{y}^T\mathbf{A}\mathbf{y} = (\mathbf{x} - \mathbf{y})^T\mathbf{A}(\mathbf{x} - \mathbf{y}), \tag{16.5}$$

  což je nezáporné pro každé $\mathbf{x}, \mathbf{y}$, protože $\mathbf{A}$ je pozitivně semidefinitní.
- Dokážme konvexitu z Definice 16.1. Musíme dokázat, že pro každé $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ a $0 \leq \alpha \leq 1$ platí (16.1), tedy

  $$[\alpha\mathbf{x} + (1-\alpha)\mathbf{y}]^T\mathbf{A}[\alpha\mathbf{x} + (1-\alpha)\mathbf{y}] \leq \alpha\mathbf{x}^T\mathbf{A}\mathbf{x} + (1-\alpha)\mathbf{y}^T\mathbf{A}\mathbf{y}$$

  Po roznásobení a převedení všech členů na jednu stranu upravujeme:

  $$\begin{aligned}(\alpha - \alpha^2)\mathbf{x}^T\mathbf{A}\mathbf{x} - 2\alpha(1-\alpha)\mathbf{x}^T\mathbf{A}\mathbf{y} + ((1-\alpha) - (1-\alpha)^2)\mathbf{y}^T\mathbf{A}\mathbf{y} &\geq 0\\ \alpha(1-\alpha)(\mathbf{x}^T\mathbf{A}\mathbf{x} - 2\mathbf{x}^T\mathbf{A}\mathbf{y} + \mathbf{y}^T\mathbf{A}\mathbf{y}) &\geq 0.\end{aligned}$$

  Výraz $\alpha(1-\alpha)$ je pro každé $0 \leq \alpha \leq 1$ nezáporný. Nezápornost výrazu (16.5) jsme již ukázali. □

16.3 Operace zachovávající konvexitu funkcí

Operace zachovávající konvexitu funkcí umožňují z jednoduchých konvexních funkcí získat složitější. Konvexitu složitější funkce je často snadnější dokázat pohodlněji pomocí těchto operací než z Definice 16.1.

Jsou-li $g_1, \ldots, g_k\colon \mathbb{R}^n \to \mathbb{R}$ konvexní funkce a $\alpha_1, \ldots, \alpha_k \geq 0$, je snadné dokázat z Definice 16.1 (proveďte!), že také funkce

$$f = \alpha_1 g_1 + \cdots + \alpha_k g_k$$

je konvexní. Speciálně, jsou-li f a g konvexní funkce, pak $f + g$ je konvexní.

Zkoumejme nyní složenou funkci $f(\mathbf{x}) = (h \circ \mathbf{g})(\mathbf{x}) = h(\mathbf{g}(\mathbf{x}))$, kde $\mathbb{R}^n \xrightarrow{\mathbf{g}} \mathbb{R}^m \xrightarrow{h} \mathbb{R}$. Obecně *neplatí* ani v případě $m = n = 1$, že konvexita funkcí g a h zaručuje konvexitu funkce f. Nutné a postačující podmínky pro konvexitu složené funkce jsou obecně dosti komplikované a nebudeme je uvádět. Uvedeme jen nejdůležitější případ, kdy $\mathbf{g}$ je afinní zobrazení.

[3] Všimněte si, že pro $n = 1$ a $\mathbf{A} = 1$ se rovnost (16.5) zjednoduší na známé $x^2 - 2xy + y^2 = (x - y)^2$.

Theorem 16.5. *Let funkce $h\colon \mathbb{R}^m \to \mathbb{R}$ je konvexní. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ a $\mathbf{b} \in \mathbb{R}^m$. Pak funkce $f(\mathbf{x}) = h(\mathbf{A}\mathbf{x} + \mathbf{b})$ je konvexní.*

Proof. Pro každé $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ a $0 \le \alpha \le 1$ platí

$$\begin{aligned} f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) &= h(\mathbf{A}[\alpha\mathbf{x} + (1-\alpha)\mathbf{y}] + \mathbf{b}) \\ &= h(\alpha(\mathbf{A}\mathbf{x} + \mathbf{b}) + (1-\alpha)(\mathbf{A}\mathbf{y} + \mathbf{b})) \\ &\le \alpha h(\mathbf{A}\mathbf{x} + \mathbf{b}) + (1-\alpha)h(\mathbf{A}\mathbf{y} + \mathbf{b}) \\ &= \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y}). \end{aligned}$$

□

Nejzajímavější operace zachovávající konvexitu funkcí je maximum.

Theorem 16.6. *Let I je libovolná množina a $g_i\colon \mathbb{R}^n \to \mathbb{R}$, $i \in I$, jsou konvexní funkce. Pak funkce*

$$f(\mathbf{x}) = \max_{i\in I} g_i(\mathbf{x}) \tag{16.6}$$

je konvexní, kde předpokládáme, že pro každé $\mathbf{x}$ maximum existuje[4].

Proof. Postupujeme podobně jako v Příkladu 16.1. Máme

$$\begin{aligned} f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) &= \max_{i\in I} g_i(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) && \text{(16.7a)} \\ &\le \max_{i\in I} \alpha g_i(\mathbf{x}) + (1-\alpha)g_i(\mathbf{y}) && \text{(16.7b)} \\ &= \max_{i\in I} \alpha g_i(\mathbf{x}) + \max_{i\in I}(1-\alpha)g_i(\mathbf{y}) && \text{(16.7c)} \\ &\le \alpha \max_{i\in I} g_i(\mathbf{x}) + (1-\alpha)\max_{i\in I} g_i(\mathbf{y}) = \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y}), && \text{(16.7d)} \end{aligned}$$

kde nerovnost (16.7b) plyne z konvexity funkcí g_i a rovnost (16.7c) z nezápornosti α a $1-\alpha$. Nerovnost (16.7d) plyne z nerovnosti (16.4), kterou jsme sice dokázali jen pro konečnou množinu I ale zřejmě platí i pro nekonečnou I. □

Uveďme ještě jiný, méně podrobný ale jednodušší, důkaz Věty 16.6.

Proof. Protože funkce g_i jsou konvexní, dle Věty 16.2 jsou jejich epigrafy konvexní množiny. Snadno ověříme (podrobný důkaz vynecháme), že epigraf funkce (16.6) je průnik epigrafů funkcí g_i. Dle Věty 13.1 je průnik konvexních množin konvexní množina. Tedy epigraf funkce (16.6) je konvexní množina. Dle Věty 16.2 je tedy funkce f konvexní. □

Example 16.6. Let $f(\mathbf{x}) = \max_{i=1}^n x_i$ je maximum ze složek $\mathbf{x}$. Konvexitu této funkce jsme dokázali z Definice 16.1, nicméně dokažme ji z Věty 16.6. Máme $g_i(\mathbf{x}) = x_i$. Funkce g_i jsou lineární, tedy konvexní. Tedy funkce $f(\mathbf{x}) = \max_{i=1}^n g_i(\mathbf{x})$ je konvexní. □

Example 16.7. Funkce

$$f(\mathbf{x}) = \max_{i=1}^{k}(\mathbf{a}_i^T\mathbf{x} + b_i)$$

je maximem afinních funkcí. Tuto funkci jsme již potkali v §12.1.1. Protože afinní funkce jsou konvexní, je i jejich maximum konvexní. □

[4] Pokud pro nějaké $\mathbf{x}$ množina $\{\, g_i(\mathbf{x}) \mid i \in I \,\}$ nemá největší prvek (což se může stát jen tehdy, je-li množina I nekonečná), můžeme maximum v (16.6) nahradit supremem a věta stále platí.

Example 16.8. Let $C \subseteq \mathbb{R}^n$ je libovolná (ne nutně konvexní) množina. Funkce

$$f(\mathbf{x}) = \max_{\mathbf{y} \in C} \|\mathbf{x} - \mathbf{y}\|$$

udává vzdálenost bodu $\mathbf{x}$ od nejvzdálenějšího bodu množiny C (zde předpokládáme, že maximum existuje). Dle Věty 16.5 je pro každé pevné $\mathbf{y}$ výraz $\|\mathbf{x} - \mathbf{y}\|$ konvexní funkcí $\mathbf{x}$. Tedy výraz $\|\mathbf{x} - \mathbf{y}\|$ lze chápat jako množinu konvexních funkcí $\mathbf{x}$ indexovaných indexem $\mathbf{y}$ (můžeme označit $\|\mathbf{x} - \mathbf{y}\| = g_{\mathbf{y}}(\mathbf{x})$). Jelikož f je maximem těchto funkcí, je i funkce f konvexní. □

Example 16.9. Mějme funkci

$$f(\mathbf{c}) = \max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{A}\mathbf{x} \geq \mathbf{b} \,\},$$

která vyjadřuje závislost optimální hodnoty daného lineárního programu na vektoru $\mathbf{c}$ (viz §12). Máme $f(\mathbf{c}) = \max_{\mathbf{x} \in X} \mathbf{c}^T\mathbf{x}$ a $X = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \geq \mathbf{b} \,\}$ (zde předpokládáme, že pro každé $\mathbf{c}$ maximum existuje, neboli množina X je neprázdná a omezená). Je-li $\mathbf{x}$ pevné, je $\mathbf{c}^T\mathbf{x}$ lineární funkce vektoru $\mathbf{c}$. Funkce f je tedy maximum nekonečného množství lineárních funkcí, tedy je konvexní. □

Example 16.10. Let $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$, $b_1, \ldots, b_n \in \mathbb{R}$ a $\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$ je vektor nezáporných vah. Přibližné řešení soustavy $\mathbf{a}_i^T\mathbf{x} = b_i$, $i = 1, \ldots, n$, ve smyslu *vážených nejmenších čtverců* (viz §6.10) znamená vypočítat

$$f(\mathbf{w}) = \min_{\mathbf{x} \in \mathbb{R}^m} \sum_{i=1}^{n} w_i(\mathbf{a}_i^T\mathbf{x} - b_i)^2,$$

kde jsme označili hodnotu výsledného minima jako funkci vektoru vah. Funkce f je konkávní, protože je minimem lineárních funkcí. □

16.4 Cvičení

16.1. Pro každou funkci $f\colon \mathbb{R}^n \to \mathbb{R}$ dokažte z Definice 16.1, které z těchto čtyř tvrzení platí: funkce je konvexní, konkávní, konvexní i konkávní, ani konvexní ani konkávní.

a) $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$
b) $f(\mathbf{x}) = \mathbf{x}^T\mathbf{x}$
c) $f(\mathbf{x}) =$ aritmetický průměr čísel $x_1, \ldots, x_n$
d) $f(\mathbf{x}) = \operatorname{median}_{i=1}^n x_i$ (medián čísel $x_1, \ldots, x_n$)

16.2. Pro každou funkci dokažte, které z těchto čtyřech tvrzení platí: funkce je konvexní, konkávní, konvexní i konkávní, ani konvexní ani konkávní. Můžete to dokázat buď z Definice 16.1, pomocí derivací, nebo pomocí operací zachovávajících konvexitu.

a) $f(x) = \mathrm{e}^{x^2}$
b) $f(x) = \mathrm{e}^{-x^2}$
c) $f(x, y) = |x - y|$
d) $f(x, y) = -y$
e) $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$

f) $f(\mathbf{x}) = \sum_{i=1}^{n} x_i \log x_i$ na množině $\mathbb{R}^n_{++}$

g) $f(\mathbf{x}) = \sum_{i=1}^{k} \log(b_i - \mathbf{x}^T\mathbf{a}_i)$ na množině $X = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x}^T\mathbf{a}_i < b_i \,, i = 1, \ldots, k \,\}$

h) $f(\mathbf{x}) = \min_{i=1}^{n} |x_i|$

i) $f(\mathbf{x}) = \max_{i=1}^{n} x_i + \min_{i=1}^{n} x_i$

j) $f(\mathbf{x}) = \max_{i=1}^{n} x_i - \min_{i=1}^{n} x_i$

k) $(\star)$ $f(\mathbf{x})$ = součet k největších čísel $x_1, \ldots, x_n$ (kde $k \leq n$ je dáno)

16.3. Robustní prokládání přímky množinou bodů $(\mathbf{x}_i, y_i) \in (\mathbb{R}^n \times \mathbb{R})$, $i = 1, \ldots, m$ vyžaduje minimalizaci funkce

$$f(\mathbf{a}, b) = \sum_{i=1}^{m} \max\{-\mathbf{a}^T\mathbf{x}_i + b + y_i - \varepsilon,\ 0\ ,\ \mathbf{a}^T\mathbf{x}_i + b - y_i - \varepsilon\},$$

kde $\mathbf{a} \in \mathbb{R}^n$ a $b \in \mathbb{R}$. Dokažte, že $f(\mathbf{a}, b)$ je konvexní funkce.

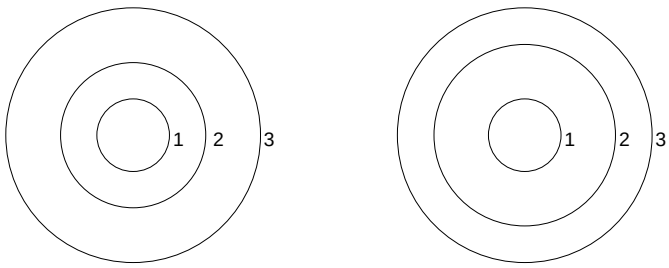
16.4. Je dána funkce $f(x) = -\cos x$ a množina $X = [-\pi, +\pi]$ (kde $[\cdot]$ značí uzavřený interval). Zakroužkujte pravdivá tvrzení (může jich být i více):

a) Funkce f je na množině X konvexní.

b) Funkce f je na množině X konkávní.

c) Funkce f není na množině X ani konvexní ani konkávní.

16.5. Každý z obrázků zobrazuje některé vrstevnice funkce dvou proměnných a jejich výšky. Je možné, aby funkce, která má tyto vrstevnice, byla konvexní? Dokažte z Definice 16.1.



16.6. Co je subkontura výšky 2 funkce jedné proměnné $f(x) = x^2 - x$?

16.7. Zkuste dokázat z Definice 16.1 konvexitu či konkavitu funkcí z Příkladu 16.4. Jestliže to nesvedete, dokažte jejich konvexitu či konkavitu pomocí Vět 16.3 a 16.4.

Hints and Solutions

16.1.a) Konvexní i konkávní, nerovnost (16.1) platí s rovností.

16.1.b) Je konvexní, není koknávní.

16.1.c) Konvexní i konkávní, nerovnost (16.1) platí s rovností.

16.1.d) Ani konvexní ani konkávní.

16.5. V Definici 16.1 zvolte $\mathbf{x}, \mathbf{y}$ na vrstevnicích výšky 1 a 3. Zvolte chytře α. Odpověď: ne, ano.

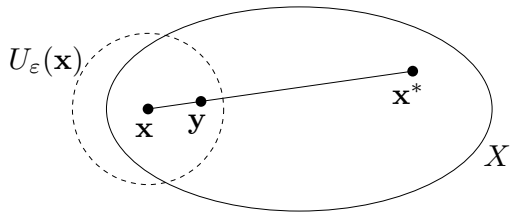
16.7. Interval $[-1, 2]$.

Chapter 17

Konvexní optimalizační úlohy

Theorem 17.1. *Let funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ je konvexní na konvexní množině $X \subseteq \mathbb{R}^n$. Pak každé local minimum funkce f na množině X je zároveň globální.*

Proof. Let $\mathbf{x}$ je lokálním minimem f na X, viz obrázek:



Dle Definice 9.2 tedy existuje $\varepsilon > 0$ tak, že $f(\mathbf{x}) \le f(\mathbf{y})$ pro všechna $\mathbf{y} \in U_\varepsilon(\mathbf{x}) \cap X$. Let ale $\mathbf{x}$ není globální minimum, tedy existuje $\mathbf{x}^* \in X$ takové, že $f(\mathbf{x}^*) < f(\mathbf{x})$. Ukážeme, že to vede ke sporu. Pro každé ε totiž můžeme zvolit $0 < \alpha < 1$ tak, že bod $\mathbf{y} = \alpha\mathbf{x} + (1-\alpha)\mathbf{x}^*$ leží v okolí $U_\varepsilon(\mathbf{x})$. Protože je množina X konvexní, leží bod $\mathbf{y}$ zároveň i v X. Máme

$$f(\mathbf{y}) \;=\; f(\alpha\mathbf{x} + (1-\alpha)\mathbf{x}^*) \;\le\; \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{x}^*) \;<\; \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{x}) \;=\; f(\mathbf{x}).$$

Ale tvrzení $f(\mathbf{y}) < f(\mathbf{x})$ je ve sporu s předpokladem, že $\mathbf{x}$ je local minimum. □

Význam Věty 17.1 je v tom, že najít local minimum funkce na množině je obvykle mnohem snadnější než najít globální minimum. Úloze, ve které minimalizujeme konvexní funkci na konvexní množině, se říká **konvexní optimalizační úloha**.

Zopakujme nyní obecnou úlohu spojité optimalizace ve standarním tvaru (1.4):

$$\begin{aligned} \min \quad & f(x_1, \dots, x_n) \\ \text{za podmínek} \quad & g_i(x_1, \dots, x_n) \le 0, \quad i = 1, \dots, m \\ & h_i(x_1, \dots, x_n) = 0, \quad i = 1, \dots, l \end{aligned} \tag{17.1}$$

kde $f\colon \mathbb{R}^n \to \mathbb{R}$, $(g_1, \dots, g_m) = \mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^m$, $(h_1, \dots, h_l) = \mathbf{h}\colon \mathbb{R}^n \to \mathbb{R}^l$. Její množina přípustných řešení je konvexní, jestliže funkce $f, g_1, \dots, g_m$ jsou konvexní a funkce $h_1, \dots, h_l$ jsou afinní (tedy zobrazení $\mathbf{h}$ je afinní). Tuto množinu totiž můžeme psát jako

$$X = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{g}(\mathbf{x}) \le \mathbf{0},\ \mathbf{h}(\mathbf{x}) = \mathbf{0} \,\} = \bigcap_{i=1}^{m} \{\, \mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) \le 0 \,\} \cap \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{h}(\mathbf{x}) = \mathbf{0} \,\}.$$

Zde každá množina $\{\, \mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) \le 0 \,\}$ je konvexní, neboť je to subkontura konvexní funkce g_i (Věta 16.1). Množina $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{h}(\mathbf{x}) = \mathbf{0} \,\}$ je afinní podprostor, tedy také konvexní. Množina X je průnik konvexních množin, tedy (dle Věty 13.1) je konvexní.

Podmínka, že funkce $f, g_1, \ldots, g_m$ jsou konvexní a zobrazení $\mathbf{h}$ je afinní, je postačující ale nikoliv nutná pro konvexitu množiny X.

Example 17.1. Uvažujme dvě ekvivalentní definice téže množiny

$$X = \{\, \mathbf{x} \in \mathbb{R}^2 \mid x_1/(1+x_2^2) \le 0,\ (x_1+x_2)^2 = 0 \,\} = \{\, \mathbf{x} \in \mathbb{R}^2 \mid x_1 \le 0,\ x_1 + x_2 = 0 \,\}.$$

Oba tvary jsou ekvivalentní (proč?). V prvním tvaru funkce $g(\mathbf{x}) = x_1/(1+x_2^2)$ není konvexní (dokažte z Definice 16.1!) a funkce $h(\mathbf{x}) = (x_1+x_2)^2$ není afinní. Přesto je množina X konvexní, což je vidět ze druhého tvaru. □

Úloze tvaru (17.1), ve které jsou funkce $f, g_1, \ldots, g_m$ konvexní a zobrazení $\mathbf{h}$ afinní, říkáme **konvexní optimalizační úloha ve standardním tvaru**. Tato podmínka je nutná ale nikoliv postačující pro konvexitu

17.1 Třídy optimalizačních úloh

Optimalizační úlohy ve tvaru (17.1) se taxonomizují podle druhu funkcí f, g_i, h_i. Pro každou třídu existují specializované algoritmy schopné najít local minimum[1].

17.1.1 Lineární programování (LP)

V *lineárním programování* jsou všechny funkce f, g_i, h_i afinní. Jde tedy v jistém smyslu o nejjednoduší případ konvexní optimalizační úlohy. Přesto jsme viděli v Kapitole 12, že již tento jednoduchý případ má velmi mnoho aplikací.

17.1.2 Kvadratické programování (QP)

V *kvadratickém programování* jsou funkce g_i, h_i afinní a funkce f je kvadratická konvexní, tedy $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$, kde matice $\mathbf{A}$ je pozitivně semidefinitní.

Example 17.2. Při řešení soustavy ve smyslu nejmenších čtverců počítáme konvexní QP bez omezení $\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$.

Tuto úlohu lze všelijak modifikovat, např. můžeme přidat omezení $\mathbf{c} \le \mathbf{x} \le \mathbf{d}$, tj. každá proměnná x_j musí být v intervalu $[c_j, d_j]$. To vede na konvexní QP s omezeními. □

Example 17.3. Hledání řešení přeurčené lineární soustavy s nejmenší normou vede na úlohu $\min\{\, \mathbf{x}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b} \,\}$, což je konvexní QP s omezeními. □

Example 17.4. Chceme spočítat vzdálenost polyedrů

$$P_1 = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}_1\mathbf{x} \le \mathbf{b}_1 \,\}, \qquad P_2 = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}_2\mathbf{x} \le \mathbf{b}_2 \,\}$$

[1] Viz např. http://www.neos-guide.org.

danou jako $d(P_1, P_2) = \inf\{\, \|\mathbf{x}_1 - \mathbf{x}_2\|_2 \mid \mathbf{x}_1 \in P_1,\ \mathbf{x}_2 \in P_2 \,\}$. Úloha vede na QP

$$\min\{\, \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 \mid \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n,\ \mathbf{A}_1\mathbf{x}_1 \le \mathbf{b}_1,\ \mathbf{A}_2\mathbf{x}_2 \le \mathbf{b}_2 \,\}.$$

Pokud se polyedry protínají, jejich vzdálenost je nula. Pokud je aspoň jeden polyedr prázdný, úloha je nepřípustná □

Example 17.5. Je dáno m bodů v $\mathbb{R}^n$, z nichž každý patří do jedné ze dvou tříd, označených -1 a 1. Jinými slovy, je dána množina dvojic $(\mathbf{x}_i, y_i) \in \mathbb{R}^n \times \{-1, 1\}$ pro $i = 1, \dots, m$. V úloze *lineární klasifikace* hledáme nadrovinu, která odděluje body z obou tříd, tedy hledáme $\mathbf{a} \in \mathbb{R}^n$ a $b \in \mathbb{R}$ takové, že

$$\begin{aligned} \mathbf{a}^T\mathbf{x}_i - b &< 0 \quad \text{pro } y_i = -1, \\ \mathbf{a}^T\mathbf{x}_i - b &> 0 \quad \text{pro } y_i = 1. \end{aligned}$$

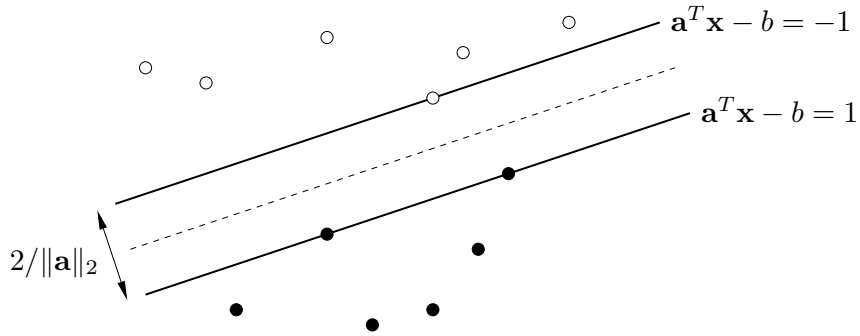
Tyto nerovnice se nezmění vynásobením $(\mathbf{a}, b)$ libovolným kladným číslem, tedy ekvivalentně hledáme $\mathbf{a}, b$ tak, že

$$\begin{aligned} \mathbf{a}^T\mathbf{x}_i - b &\le -1 \quad \text{pro } y_i = -1, \\ \mathbf{a}^T\mathbf{x}_i - b &\ge 1 \quad \text{pro } y_i = 1, \end{aligned}$$

což můžeme napsat také jako

$$y_i(\mathbf{a}^T\mathbf{x}_i - b) \ge 1 \quad \forall i = 1, \dots, m. \tag{17.2}$$

Body jsou vlastně odděleny pásem $\{\, \mathbf{x} \in \mathbb{R}^n \mid -1 \ge \mathbf{a}^T\mathbf{x} - b \ge 1 \,\}$, viz obrázek:



V úloze *support vector machine* (SVM) chceme najít nejen rozdělující nadrovinu, ale navíc maximalizovat šířku tohoto pásu. Snadno spočítáme (proveďte!), že šířka pásu je $2/\|\mathbf{a}\|_2$. Maximalizace této funkce je stejná jako minimalizace $\|\mathbf{a}\|_2^2 = \mathbf{a}^T\mathbf{a}$. Tedy chceme minimalizovat $\mathbf{a}^T\mathbf{a}$ za podmínek (17.2). To je úloha QP. □

17.1.3 Kvadratické programování s kvadratickými omezeními (QCQP)

Obecnější variantou je *kvadratické programování s kvadratickými omezeními* (QCQP, *quadratically constrained quadratic programming*), kde všechny funkce f, g_i, h_i jsou kvadratické. Úloha je konvexní jen tehdy, když kvadratické funkce f, g_i jsou konvexní (tj. s pozitivně semidefinitní maticí) a funkce h_i jsou afinní.

17.1.4 Semidefinitní programování (SDP)

Theorem 17.2. *Množina všech pozitivně semidefinitních matic rozměru $n \times n$ je konvexní kužel.*

Proof. Let pro $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ a $\mathbf{x} \in \mathbb{R}^n$ je $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ a $\mathbf{x}^T\mathbf{B}\mathbf{x} \geq 0$. Pak pro každé $\alpha, \beta \geq 0$

$$\mathbf{x}^T(\alpha\mathbf{A} + \beta\mathbf{B})\mathbf{x} = \alpha\mathbf{x}^T\mathbf{A}\mathbf{x} + \beta\mathbf{x}^T\mathbf{B}\mathbf{x} \geq 0. \qquad \square$$

Konvexní kužel je konvexní množina. To umožňuje formulovat třídu konvexních úloh známou jako *semidefinitní programování* (SDP). Jednou z možných formulací je

$$\min\{\, \mathbf{C} \cdot \mathbf{X} \mid \mathbf{X} \in \mathbb{R}^{n\times n} \text{ pozitivně semidefinitní}, \ \mathbf{A}_i \cdot \mathbf{X} = b_i \ \forall i = 1, \ldots, m \,\}, \tag{17.3}$$

kde matice $\mathbf{C}, \mathbf{A}_1, \ldots, \mathbf{A}_m \in \mathbb{R}^{n\times n}$ jsou dány a optimalizujeme přes pozitivně semidefinitní matice $\mathbf{X} \in \mathbb{R}^{n\times n}$. Operace

$$\mathbf{A} \cdot \mathbf{X} = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}X_{ij}$$

označuje skalární součin matic. Na řešení úlohy (17.3) existují efektivní algoritmy.

SDP je velmi obecná třída konvexních úloh. Lze ukázat, že LP, QP i QCQP lze formulovat jako speciální případy SDP. Pro ilustraci ukážeme, že pokud matice $\mathbf{C}, \mathbf{A}_1, \ldots, \mathbf{A}_m$ jsou diagonální, úloha (17.3) se redukuje na LP. V tom případě v součinech $\mathbf{C} \cdot \mathbf{X}$ a $\mathbf{A}_i \cdot \mathbf{X}$ nediagonální prvky matice $\mathbf{X}$ nehrají žádnou roli. Diagonální matice je pozitivně semidefinitní právě tehdy, když všechny její prvky jsou nezáporné (viz Cvičení 5.21). Tedy úloha (17.3) se redukuje na

$$\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n, \ \mathbf{x} \geq \mathbf{0}, \ \mathbf{a}_i^T\mathbf{x} = b_i \ \forall i = 1, \ldots, m \,\},$$

kde vektory $\mathbf{c}, \mathbf{a}_i \in \mathbb{R}^n$ jsou diagonály matic $\mathbf{C}, \mathbf{A}_i$.

Dále uvedeme příklady konvexních úloh, které na první pohled nespadají do žádné z uvedených tříd.

Example 17.6. Jsou dány body $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$ a chceme minimalizovat funkci

$$f(\mathbf{x}) = \sum_{i=1}^{m} \|\mathbf{a}_i - \mathbf{x}\|_2 \tag{17.4}$$

přes $\mathbf{x} \in \mathbb{R}^n$. Řešení této úlohy je známo jako *geometrický medián*. Pro $n = 1$ se funkce redukuje na $f(x) = \sum_{i=1}^{m} |x - a_i|$, jejímž minimem je obyčejný medián.

Úloha není úlohou LP, QP, ani QCQP. Ovšem lze ji přeformulovat jako SDP (podrobnosti vynecháme).

Pro případ $n = 2$ má úloha jednoduchý mechanický model. Do vodorovného prkna vyvrtáme díry o souřadnicích $\mathbf{a}_i$. Každou dírou provlečeme provázek. Provázky jsou nahoře svázané uzlem do jednoho bodu a dole mají závaží o stejné hmotnosti. Poloha uzlu je $\mathbf{x}$. Hodnota $f(\mathbf{x})$ je potenciální energie soustavy a ustálený stav odpovídá minimu $f(\mathbf{x})$. $\square$

Example 17.7. *Analytický střed polyedru* $X = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \geq \mathbf{b} \,\}$ je bod $\mathbf{x} \in X$, který maximalizuje funkci

$$f(\mathbf{x}) = \sum_{i=1}^{m} \log(\mathbf{a}_i^T \mathbf{x} - b_i), \tag{17.5}$$

kde $\mathbf{a}_i^T$ jsou řádky matice $\mathbf{A}$. Všimněte si, že definiční obor funkce f je X. Lze ukázat, že f má na množině X minimum a toto minimum je jediné právě tehdy, když polyedr X je neprázdný a omezený. □

17.2 Cvičení

17.1. Významnou vlastností konvexních funkcí je to, že každé local minimum funkce je zároveň globální (Věta 17.1). Ne každá funkce s touto vlastností je ovšem konvexní. Člověk by si mohl myslet, že součet dvou funkcí (ne nutně konvexních) s touto vlastností bude mít tuto vlastnost také. Je toto tvrzení pravdivé? Odpověď dokažte.

17.2. Dokažte, že množina optimálních řešení konvexní optimalizační úlohy je konvexní.

17.3. Mějme úlohu

$$\min\{\, f(x,y) \mid x, y \geq 0,\ 2x + y \geq 1,\ x + 3y \geq 1 \,\}.$$

Nakreslete množinu přípustných řešení. Pro každou z následujících účelových funkcí najděte úvahou množinu optimálních řešení a optimální hodnotu:

a) $f(x,y) = x + y$
b) $f(x,y) = x$
c) $f(x,y) = \min\{x, y\}$
d) $f(x,y) = \max\{x, y\}$
e) $f(x,y) = |x + y|$
f) $f(x,y) = x^2 + 9y^2$

V kterých případech se jedná o konvexní optimalizační úlohu?

17.4. Najděte explicitní řešení pro následující úlohy QCQP ($\mathbf{A}, \mathbf{B}$ jsou pozitivně definitní):

a) $\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{x}^T\mathbf{A}\mathbf{x} \leq 1 \,\}$
Nápověda: Viz Cvičení 12.4.
b) $\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ (\mathbf{x} - \mathbf{b})^T\mathbf{A}(\mathbf{x} - \mathbf{b}) \leq 1 \,\}$
Nápověda: substituujte $\mathbf{y} = \mathbf{x} - \mathbf{b}$.
c) $\min\{\, \mathbf{x}^T\mathbf{B}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{x}^T\mathbf{A}\mathbf{x} \leq 1 \,\}$

17.5. Formulujte úlohu $\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_4$ jako konvexní QCQP.

17.6. Máme konvexní funkci jedné proměnné $f\colon \mathbb{R} \to \mathbb{R}$. Dáme do grafu funkce žebřík o délce 1 tak, aby oba konce ležely na grafu. Předpokládáme-li, že tření mezi žebříkem a grafem je nulové, zaujme žebřík stav lokálního minima potenciální energie (která je přímo úměrná výšce středu žebříku). Zformulujte jako optimalizační úlohu. Bude tato úloha konvexní? Pokud ne, najděte situaci, kdy potenciální energie bude mít více než jedno local minimum.

Chapter 18

Příklady nekonvexních úloh

Example 18.1. Řešení homogenní lineární soustavy ve smyslu nejmenších čtverců vede na úlohu

$$\min\{\, \|\mathbf{A}\mathbf{x}\|_2^2 \mid \mathbf{x}^T\mathbf{x} = 1 \,\}. \tag{18.1}$$

To je instance QCQP, ale není to konvexní úloha kvůli omezení $\mathbf{x}^T\mathbf{x} = 1$. Dokonce ani nejde na konvexní úlohu transformovat. Je jasné, že množina $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x}^T\mathbf{x} = 1 \,\}$ není konvexní. Někdo by si mohl myslet, že omezení $\mathbf{x}^T\mathbf{x} = 1$ lze nahradit konvexním omezením $\mathbf{x}^T\mathbf{x} \leq 1$, podobně jako ve Cvičení 12.4. To ale nelze, neboť

$$\min\{\, \|\mathbf{A}\mathbf{x}\|_2^2 \mid \mathbf{x}^T\mathbf{x} = 1 \,\} \neq \min\{\, \|\mathbf{A}\mathbf{x}\|_2^2 \mid \mathbf{x}^T\mathbf{x} \leq 1 \,\} = 0.$$

My ale víme, že úlohu (18.1) lze řešit pomocí SVD, protože hledáme nadrovinu s normálovým vektorem $\mathbf{x}$, která minimalizuje součet čtverců kolmých vzdáleností řádků $\mathbf{a}_1, \ldots, \mathbf{a}_m$ matice $\mathbf{A}$ k nadrovině. □

Example 18.2. Obecněji, úloha (7.7) je nekonvexní, neboť její množina přípustných řešení je nekonvexní.[1]. □

V posledních dvou příkladech bylo snadné najít globální optimum nekonvexní úlohy. To je ale řídká výjimka – častěji je nalezení globálního minima nekonvexní úlohy velmi těžké, neboť úloha má velmi mnoho lokálních minim.

Example 18.3. Uveďme příklad, na kterém bude na první pohled vidět, že nekonvexní úloha může mít velmi mnoho lokálních minim. Řešme úlohu

$$\min\{\, -\mathbf{x}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ -\mathbf{1} \leq \mathbf{x} \leq \mathbf{1} \,\}. \tag{18.2}$$

Množina přípustných řešení je hyperkrychle, $X = [-1, 1]^n$. Účelová funkce $f(\mathbf{x}) = -\mathbf{x}^T\mathbf{x}$ je konkávní. Je očividné, že funkce f má na množině X local minimum v každém vrcholu hyperkrychle X (nakreslete si obrázek pro $n = 2$, tedy pro obyčejný čtverec!). Pro n proměnných má úloha 2^n lokálních minim. Připomeňme, že konvexní polyedr popsaný polynomiálním počtem lineárních nerovnic může mít exponenciální počet vrcholů (viz §13.3.2).

[1] Namítnete, že vůbec nemůžeme mluvit o konvexitě úlohy (7.7), protože v této úloze optimalizujeme přes množinu matic a konvexitu jsme v Definicích 13.1 a 16.1 definovali pro množiny a funkce vektorů. Definice konvexity lze ovšem snadno zobecnit na množiny a funkce matic: buď matici $\mathbb{R}^{m\times n}$ v úloze (7.7) můžeme přerovnat do vektoru $\mathbb{R}^{mn}$, nebo (lépe) můžeme konvexitu definovat místo na prostoru $\mathbb{R}^n$ na obecném vektorového prostoru (viz učebnice lineární algebry).

V tomto případě jsou local minima všechna stejná, tedy úlohu snadno vyřešíme. Ale již mírnou modifikací úlohy se stane nalezení globálního optima prakticky nemožné. Uvažujme úlohu

$$\min\{\, \mathbf{x}^T\mathbf{A}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ -\mathbf{1} \le \mathbf{x} \le \mathbf{1}\,\}. \tag{18.3}$$

Je jasné, že pro $\mathbf{A} = -\mathbf{I}$ dostaneme úlohu (18.2). Je známo, že neexistuje algoritmus, který by pro libovolnou (tedy ne nutně pozitivně semidefinitní) matici $\mathbf{A} \in \mathbb{R}^{n\times n}$ vyřešil úlohu (18.3) v čase, který je shora omezen polynomiální funkcí čísla n. □

Uveďme dále praktičtější příklady.

Example 18.4. Mějme m bodů v rovině $\mathbf{a}_1, \dots, \mathbf{a}_m \in \mathbb{R}^d$. Úkolem je rozmístit dalších n bodů $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ tak, aby nejdelší vzdálenost bodu $\mathbf{a}_i$ k nejbližšímu bodu $\mathbf{x}_j$ byla nejmenší. Tedy minimalizujeme účelovou funkci

$$f(\mathbf{x}_1, \dots, \mathbf{x}_n) = \sum_{i=1}^{m} \min_{j=1}^{n} \|\mathbf{a}_i - \mathbf{x}_j\| \tag{18.4}$$

přes vektory $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$. Máme $f\colon \mathbb{R}^{dn} \to \mathbb{R}$, tedy přesněji můžeme říci, že minimalizujeme funkci f přes jediný vektor $(\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{dn}$.

Úloha je známá jako *shlukování*. Jako motivaci si představme optimální rozmístění cisteren ve vesnici, kde občas neteče voda. Zde máme $n = 2$, $\mathbf{a}_i$ jsou souřadnice domů a $\mathbf{x}_j$ jsou souřadnice cisteren. Chceme, aby průměrná vzdálenost obyvatele k nejbližší cisterně byla co nejmenší.

Je funkce (18.4) konvexní? Vezměme jednoduchý případ $d = 1$, $m = 1$, $n = 2$, $a_1 = 0$. Pak (18.4) má tvar $f(x_1, x_2) = \min\{|x_1|, |x_2|\}$. Snadno dokážeme (proveďte!), že toto není konvexní funkce. Bez důkazu uveďme, že (nepřekvapivě) funkce není konvexní ani pro větší d, m, n.

Je známo, že neexistuje algoritmus, který by našel optimální řešení úlohy (18.4) v čase, který je polynomiální funkcí čísel d, m, n. V praktické situaci tedy nezbývá nic jiného, než použít algoritmus, který najde pouze přibližné optimum. Takovým algoritmem je např. *k-means*.

Funkci (18.5) lze modifikovat takto:

$$f(\mathbf{x}_1, \dots, \mathbf{x}_n) = \max_{i=1}^{m} \min_{j=1}^{n} \|\mathbf{a}_i - \mathbf{x}_j\|. \tag{18.5}$$

Jaký význam má tato formulace? □

18.1 Celočíselné programování

Významnou skupinou nekonvexních úloh jsou úlohy, ve kterých přípustná řešení nabývají pouze celočíselných hodnot. Z nich nejvýznamější je **celočíselné lineární programování** (ILP, *integer linear programming*)

$$\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{Z}^n,\ \mathbf{A}\mathbf{x} \ge \mathbf{b}\,\}. \tag{18.6}$$

Rozdíl oproti obyčejnému LP je v tom, že místo $\mathbf{x} \in \mathbb{R}^n$ je $\mathbf{x} \in \mathbb{Z}^n$. Často proměnné nabývají dokonce pouze dvou stavů, tedy $\mathbf{x} \in \{0, 1\}^n$, pak mluvíme o *binárním LP* nebo *0-1 LP*. Množina přípustných řešení této úlohy je nekonvexní, obsahuje konečný počet izolovaných bodů.

Množinu přípustných řešení můžeme napsat dvěma způsoby:

$$X = \{\, \mathbf{x} \in \mathbb{Z}^n \mid \mathbf{A}\mathbf{x} \geq \mathbf{b} \,\} = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \geq \mathbf{b} \,\} \cap \mathbb{Z}^n.$$

Druhý způsob říká, že X jsou body celočíselné mřížky $\mathbb{Z}^n$ ležící uvnitř konvexního polyedru $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \geq \mathbf{b} \,\}$. Tento polyedr je množinou přípustných řešení obyčejného LP.

Zatímco LP je řešitelné v polynomiálním čase, ILP je NP-těžké. Mnoho úloh kombinatorické optimalizace (např. úloh na grafech) se pohodlně formuluje jako ILP.

Example 18.5. V *úloze o pokrytí množiny* je dána konečná množina S a množiny $S_1, \ldots, S_n \subseteq S$. Úkolem je vybrat co nejmenší počet těchto množin takových, že jejich sjednocení je stejné jako sjednocení původních množin. Tedy najít co nejmenší množinu indexů $I \subseteq \{1, \ldots, n\}$ takovou, že $\bigcup_{i \in I} S_i = \bigcup_{i=1}^n S_i$. Je známo, že tato úloha je NP-těžká.

Formulujme ji jako ILP. Proměnné budou $x_1, \ldots, x_n \in \{0, 1\}$, kde $x_i = 1$ indikuje $i \in I$.

$$\begin{aligned} \min \quad & \sum_{i=1}^n x_i \\ \text{za podmínek} \quad & \sum_{i \,|\, e \in S_i} x_i \geq 1, \qquad \forall e \in S_1 \cup \cdots \cup S_n \\ & x_1, \ldots, x_n \in \{0, 1\} \end{aligned}$$

Let např. $F = \{\{a, b\}, \{b, c\}, \{a, c\}\}$. Existují tři optimální pokrytí, každé obsahuje dvě z daných tří množin: $\mathbf{x} = (1, 1, 0)$, $\mathbf{x} = (1, 0, 1)$ a $\mathbf{x} = (0, 1, 1)$. Každé z nich má optimální hodnotu ILP rovnu 2. □

18.2 Konvexní relaxace nekonvexních úloh

Relaxace je technika, kterou lze někdy získat přibližná řešení obtížných úloh. Spočívá na očividné skutečnosti (promyslete!), že pro každou množinu $X \subseteq \mathbb{R}^n$ a funkci $f\colon X \to \mathbb{R}$ platí

$$Y \supseteq X \quad \Longrightarrow \quad \min_{\mathbf{x} \in Y} f(\mathbf{x}) \leq \min_{\mathbf{x} \in X} f(\mathbf{x}). \tag{18.7}$$

Jak toho použít? Let úloha $\min_{\mathbf{x} \in X} f(\mathbf{x})$ je obtížná její obtížnost pramení ze složitosti množiny přípustných řešení X. Nahradíme množinu X 'jednodušší' množinou $Y \supseteq X$ a řešíme snadnější úlohu $\min_{\mathbf{x} \in Y} f(\mathbf{x})$. Když budeme mít štěstí, bude nerovnost v (18.7) platit s rovností, tedy obě optima budou stejná. Když ne, získáme alespoň dolní mez na optimální řešení.

Typická situace je, že množina X je nekonvexní a my ji nahradíme vhodnou konvexní množinou $Y \supseteq X$. Pokud funkce f je konvexní, získáme konvexní úlohu. Mluvíme o **konvexní relaxaci**. Tak například konvexní relaxace úlohy ILP (18.6) je úloha LP

$$\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{A}\mathbf{x} \geq \mathbf{b} \,\}. \tag{18.8}$$

Zde máme $X = \mathbb{Z}^n$, $Y = \mathbb{R}^n \supset X$, a $f(\mathbf{x}) = \mathbf{c}^T\mathbf{x}$.

Example 18.6. V úloze 18.5 je optimální hodnota LP relaxace rovna $\frac{3}{2}$, odpovídající $\mathbf{x} = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$. □

18.3 Cvičení

18.1. Dokažte, že účelové funkce vystupující v následujících úlohách jsou nekonvexní:

a) úloha (18.5)
b) Příklad 11.5
c) Cvičení 11.2

Chapter 19

Vícekriteriální optimalizace

19.1 Uspořádání na množině

Binární relace na množině Y je množina $R \subseteq Y \times Y$. Binární relace je

- **reflexivní**, když $(x, x) \in R$ pro každé $x \in Y$,
- **transitivní**, když $(x, y) \in R$ a $(y, z) \in R$ implikuje $(x, z) \in R$,
- **antisymetrická**, když $(x, y) \in R$ a $(y, x) \in R$ implikuje $x = y$.

Částečné uspořádání (krátce jen **uspořádání**) na množině Y je binární relace na Y, která je reflexivní, transitivní a antisymetrická.

Relace uspořádání se obvykle značí infixově symbolem $\preceq$, tedy místo $(x, y) \in R$ píšeme $x \preceq y$. Pokud potřebujeme rozlišit více různých kvasi-uspořádání, používáme symboly jako $\leq_1, \leq_2, \leq', \preceq'$ atd.

Prvky $x, y \in Y$ jsou **srovnatelné** v uspořádání $\preceq$, když $x \preceq y$ nebo $y \preceq x$ nebo obojí. (Kvasi-)uspořádání je **úplné** (neboli **totální**), když každé dva prvky z Y jsou srovnatelné.

Example 19.1.

- $Y = \mathbb{R}$ a $\preceq$ je přirozené uspořádání reálných čísel. Tato relace je úplné uspořádání.
- $Y \subseteq 2^U$ a $\preceq$ je relace inkluze na množině 2^U, tedy $x \preceq y$ právě když $x \subseteq y$. Zde U je libovolná množina a 2^U značí množinu všech jejích podmnožin. Tato relace je uspořádání, ale není úplné.
- $Y = \mathbb{N}$ a $\preceq$ je relace dělitelnosti, tj. $x \preceq y$ právě když x dělí y. Tato relace je uspořádání, není úplné.
- $\mathbf{x} \preceq \mathbf{y}$ právě když $\sum_{i=1}^n x_i \leq \sum_{i=1}^n y_i$. Tato relace není antisymetrická, tedy není uspořádání. □

Nás ovšem nejvíce zajímá případ $Y = \mathbb{R}^n$.

Example 19.2. Příklady uspořádání na množině $\mathbb{R}^n$:

- Uspořádání 'po složkách': $\mathbf{x} \preceq \mathbf{y}$ právě když $x_i \leq y_i$ pro všechna $i = 1, \ldots, n$. Toto uspořádání není úplné: např. pro $n = 2$ jsou vektory $\mathbf{x} = (0, 1)$ a $\mathbf{y} = (1, 0)$ nesrovnatelné.
- Lexikografické uspořádání: $\mathbf{x} \preceq \mathbf{y}$ právě když

$$(\mathbf{x} = \mathbf{y}) \text{ nebo } (\exists m)(x_m < y_m)(\forall i < m)(x_i = y_i).$$

  Toto uspořádání je úplné.

- Definujme $\mathbf{x} \preceq \mathbf{y}$ právě když $\sum_{i=1}^{n} x_i \leq \sum_{i=1}^{n} y_i$. Tato relace není uspořádání, protože není antisymetrická. $\square$

Definition 19.1. *Prvek $x \in Y$ se nazývá (vzhledem ke kvasi-uspořádání $\preceq$)*

- **minimální prvek** *množiny Y, když $y \preceq x$ implikuje $x \preceq y$, pro všechna $y \in Y$.*
- **nejmenší prvek** *množiny Y, když $x \preceq y$, pro všechna $y \in Y$.*

Pro totální uspořádání oba pojmy splývají.

19.2 Úlohy vícekriteriální optimalizace

V klasické optimalizaci jsme se zabývali úlohami typu

$$\min_{x \in X} f(x), \tag{19.1}$$

kde X je množina přípustných řešení a $f\colon X \to \mathbb{R}$ je účelová (neboli kriteriální) funkce. Optimální hodnoty této úlohy jsou minimální prvky množiny $f(X) = \{\, f(x) \mid x \in X \,\} \subseteq \mathbb{R}$. Zde pojem 'minimální prvek' se myslí vzhledem k přirozenému uspořádání na $\mathbb{R}$.

Zobecněme tuto úlohu. Let $f\colon X \to Y$ a Let $\preceq$ je (kvasi-)uspořádání na množině Y. Pak úlohou (19.1) budeme rozumět nalezení minimálních prvků množiny $f(X) \subseteq Y$ vzhledem k uspořádání $\preceq$. Případně pro každý minimální prvek y množiny $f(X)$ chceme najít argument $x \in X$, ve kterém se nabývá, tedy splňujícím $y = f(x)$.

Nejčastěji v aplikacích potkáme případ $Y = \mathbb{R}^n$. Pak mluvíme o *vícekriteriální optimalizaci*[1], nebot vlastně chceme minimalizovat více skalárních kritérií (složek zobrazení $\mathbf{f}\colon X \to \mathbb{R}^n$, hodnoty zobrazení jsou vektory a tedy ho píšeme tučně) najednou. Dále se omezíme pouze na tento případ.

Example 19.3. V obchodě nabízejí čtyři druhy aut s těmito vlastnostmi:

		VW Golf	Opel Astra	Ford Focus	Toyota Corolla
cena	[tis. euro]	16	15	14	15
spotřeba	[l/100km]	7.2	7.0	7.5	8.2

Chceme levné auto s malou spotřebou. Která auta je dobré si koupit a která naopak nekoupit?

Máme $X = \{\,$VW Golf, Opel Astra, Ford Focus, Toyota Corolla$\,\}$ a $Y = \mathbb{R}^2$. Tabulka definuje zobrazení $\mathbf{f}$. Není ovšem jasné, jaké (kvasi-)uspořádání na množině $\mathbb{R}^2$ použít pro rozhodování.

Rozhodujme se vzhledem k uspořádání 'po složkách' $\leq^2$. Vzhledem k tomuto uspořádání nemá množina $\mathbf{f}(X)$ nejmenší prvek, neboli kritéria jsou v konfliktu. Její minimální prvky jsou $\mathbf{f}(\text{Opel Astra}) = (15, 7.0)$ a $\mathbf{f}(\text{Ford Focus}) = (14, 7.5)$.

Rozhodujme se vzhledem k lexikografickému uspořádání, přesněji nejprve se rozhodujme dle ceny a pak dle spotřeby. Nyní minimální prvek množiny $\mathbf{f}(X)$ je $\mathbf{f}(\text{Ford Focus}) = (14, 7.5)$. $\square$

[1] Angl. *multicriteria optimization*. Názvosloví ovšem není jednotné, jindy se používají názvy *multiobjective optimization* nebo *vector optimization* (neboť hodnoty zobrazení f jsou vektory).

Example 19.4. Chceme řešit (přesně či přibližně) nehomogenní lineární soustavu $\mathbf{Ax} \approx \mathbf{b}$, ale zároveň chceme, aby velikost vektoru $\mathbf{x}$ byla malá. Řešíme tedy úlohu

$$\min_{\mathbf{x}\in\mathbb{R}^n} (\, \|\mathbf{Ax} - \mathbf{b}\|_2,\ \|\mathbf{x}\|_2 \,).$$

Jaké jsou minimální prvky množiny $\mathbf{f}(\mathbb{R}^2)$ vzhledem k uspořádání po složkách? Je jich nekonečně mnoho, obrázek. Vezmeme-li lexikografické uspořádání a je-li soustava přeurčená, dostaneme metodu nejmenší normy. □

Example 19.5. V okrese je n vesnic se souřadnicemi $\mathbf{a}_1, \ldots, a_n \in \mathbb{R}^2$. Do jakého místa $\mathbf{x} \in \mathbb{R}^2$ máme umístit heliport, aby byl ke všem vesnicím blízko?

Máme $X = \mathbb{R}^2$, $Y = \mathbb{R}^n$, a $f_i(\mathbf{x}) = \|\mathbf{a}_i - \mathbf{x}\|_2$. Řešíme úlohu

$$\min_{\mathbf{x}\in\mathbb{R}^2} (\, \|\mathbf{a}_1 - \mathbf{x}\|_2, \cdots, \|\mathbf{a}_n - \mathbf{x}\|_2 \,).$$

Součástí úlohy není (kvasi-)uspořádání na množině $\mathbb{R}^n$. Jaká (kvasi-)uspořádání jsou vhodná?

Uspořádání po složkách: množina minimálních prvků množiny $\mathbf{f}(\mathbb{R}^2)$ je konvexní obal bodů $\mathbf{a}_1, \ldots, \mathbf{a}_n$. To je intuitivně jasné (i když přesný důkaz vynecháme): neleží-li $\mathbf{x}$ v tomto konvexním obalu, můžeme změnit $\mathbf{x}$ tak, že se vzdálenost k některým bodům $\mathbf{a}_i$ zmenší a k ostatním se nezvětší. Tedy je hloupost umístit heliport mimo tento konvexní obal.

Max-uspořádání: vede na úlohu

$$\min_{\mathbf{x}\in\mathbb{R}^2} \max_{i=1}^{n} \|\mathbf{a}_i - \mathbf{x}\|_2.$$

Toto je úloha klasické (skalární) optimalizace. Této formulaci se někdy říká *minimaxní* formulace. Minimalizujeme vzdálenost heliportu od nejvzdálenější vesnice. □