OPPA European Social Fund
Prague & EU: We invest in your future.

Optimisation

Embryonic notes for the course A4B33OPT
This text is incomplete and may be added to and improved during the semester.
This version: **17th October 2013**

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.

15 14 13 12 11.2 11.1 10 9 8 7 6 5 4 3 2 **??**

[**?**].

We also recommend to study exercises §3.2 in the notes [**?**]. [**?**] [**?**] [**?**] [**?**]

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

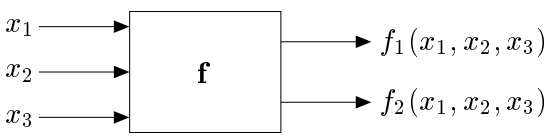
An ordered n-tuple $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ of real numbers is called (n-dimensional) **vector**.

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

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 $\mathbf{f}\colon \mathbb{R}^3 \to \mathbb{R}^2$:



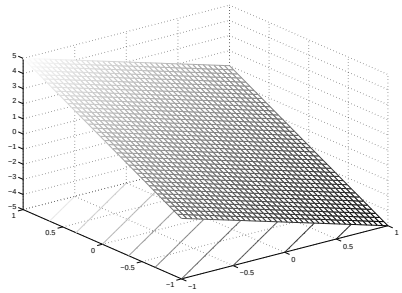
When $m = 1$ then the codomain values are scalars, written in italics, as follows: f. When $m > 1$ then the codomain values are vectors, written in bold font, $\mathbf{f}$. Even though strictly speaking the words 'function' and 'mapping' mean one and the same thing, it is common to talk about a *function*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

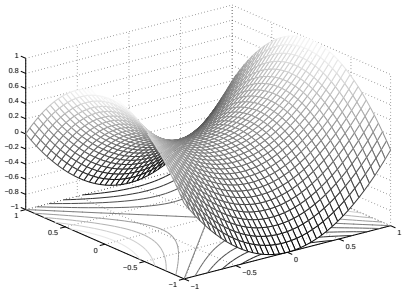
We use the following terms for functions $f\colon \mathbb{R}^n \to \mathbb{R}$:

- **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 $\{\, \mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) = y \,\}$.

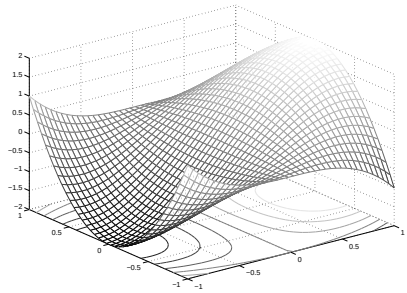
The following figure shows examples of the graph and the contours of functions of two variables on the rectangle $[-1, 1]^2$ (created by the matlab command `meshc`):



$f(x, y) = -2x + 3y$



$f(x, y) = x^2 - y^2$



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

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

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

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:

- $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I} = \mathbf{A}^{-1}\mathbf{A}$
- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
- $(\alpha\mathbf{A})^{-1} = \alpha^{-1}\mathbf{A}^{-1}$
- $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$, 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 $\sigma\colon \{1, \ldots, n\} \to \{1, \ldots, n\}$, 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 $(x_1,\ldots,x_n)$. 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

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

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 $x_1,\ldots,x_n$) 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 $\mathbf{1}_n = (1,\ldots,1)\in\mathbb{R}^n$ 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 $\operatorname{rank}(\mathbf{AB}) \le \min\{\operatorname{rank}\mathbf{A}, \operatorname{rank}\mathbf{B}\}$, 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, $m \ge n$.

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 $\{\, \mathbf{b}_i = \mathbf{Xa}_i,\ i = 1, \ldots, k \,\}$ 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:

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

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) $\{\, \mathbf{b}_i^T\mathbf{X}\mathbf{a}_i = 0,\ i = 1,\dots,k \,\}$
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 $\mathbf{A}\mathbf{B}$ and $\mathbf{B}\mathbf{A}$ have the same trace.
b) Prove that the equation $\mathbf{A}\mathbf{B} - \mathbf{B}\mathbf{A} = \mathbf{I}$ has no solution for any $\mathbf{A}, \mathbf{B}$.

2.8. *The commutator* of two matrices is the matrix $[\mathbf{A},\mathbf{B}] = \mathbf{A}\mathbf{B} - \mathbf{B}\mathbf{A}$. 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 $(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

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

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

has the property $\mathbf{L}^2 = \mathbf{I}$ (where $\mathbf{L}^2$ is the abbreviation for $\mathbf{L}\mathbf{L}$). 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 $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n$ is the vector

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

for some scalars $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$.
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 $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is the set

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

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 $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n$ and $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$,

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

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.

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

$$\operatorname{rng}\mathbf{A} = \mathbf{f}(\mathbb{R}^n) = \{\,\mathbf{A}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n\,\} = \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} \subseteq \mathbb{R}^m, \tag{3.4}$$

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

$$\dim\operatorname{rng}\mathbf{A} = \operatorname{rank}\mathbf{A}. \tag{3.5}$$

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

$$\operatorname{null}\mathbf{A} = \{\,\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{0}\,\} \subseteq \mathbb{R}^n \tag{3.6}$$

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:

$$\underbrace{\dim\operatorname{rng}\mathbf{A}}_{\operatorname{rank}\mathbf{A}} + \dim\operatorname{null}\mathbf{A} = n. \tag{3.7}$$

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 $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^n$ 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 $\mathbf{x}_1, \dots, \mathbf{x}_k$ 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).

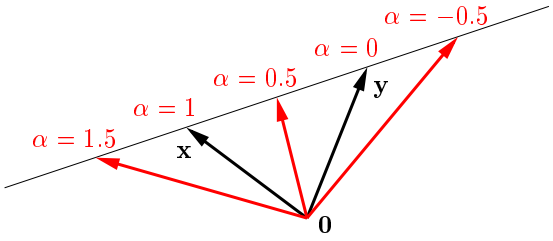
Example 3.1. Consider two linearly independent vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$. Their linear envelope is the set

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

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 $\mathbb{R}^n$.*
- *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. □

[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.

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

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

A note on terminology. 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 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.

3.4. Prove that (a) the set of solutions of a homogeneous linear system $\mathbf{A}\mathbf{x} = \mathbf{0}$ is a linear subspace and (b) the set of solutions of a non-homogeneous linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ (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 (3.7), 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 $\mathbf{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ 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]*

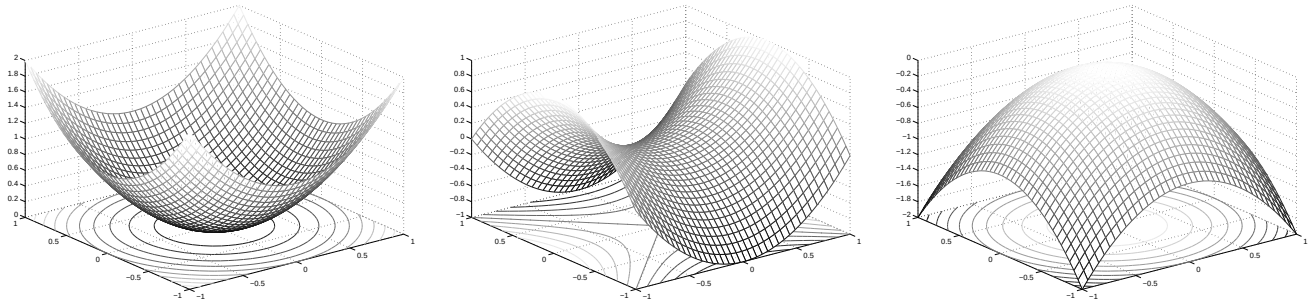
- *positive [negative] definite, iff all its eigenvalues are positive [negative]*
- *indefinite, iff it has at least one positive and one negative eigenvalues.*

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':



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

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. (⋆) 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. (⋆) 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. (⋆) Show that two square matrices commute iff they have identical eigenvectors.

5.13. (⋆) 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 $\mathbf{A}\mathbf{B} \succeq 0$.
e) If $\mathbf{A} \succ 0$, then $\mathbf{A}^{-1} \succ 0$.
f) If $\mathbf{A} \succeq 0$ and $\mathbf{B} \succeq 0$, then $\mathbf{A}\mathbf{B}\mathbf{A} \succeq 0$.

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 (skewlines) 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

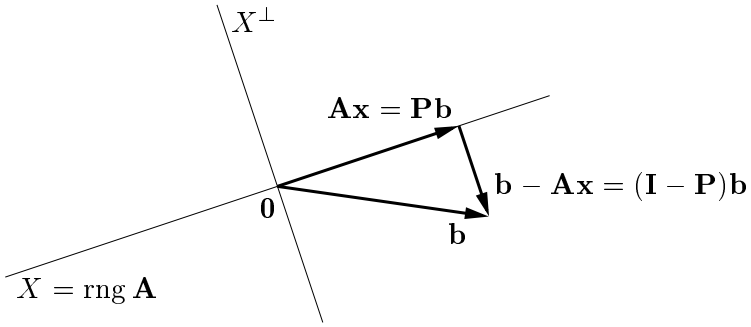
$$\mathbf{p}_1 + t_1(\mathbf{q}_1 - \mathbf{p}_1) \approx \mathbf{p}_2 + t_2(\mathbf{q}_2 - \mathbf{p}_2).$$

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 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 (⋆) 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 (3.7) 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 (3.4) (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 9.24.

Example 6.5. The system of two equations in three unknowns

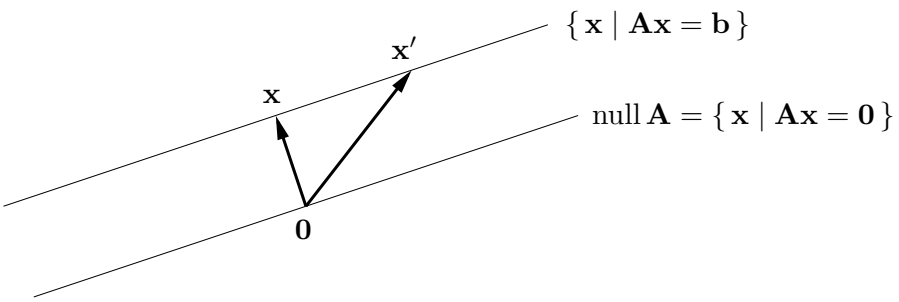
$$\begin{aligned} x + 2y + z &= 1 \\ -x + y + 2z &= 2 \end{aligned}$$

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 $\mathbf{x} \in \mathrm{rng}(\mathbf{A}^T)$, 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 $\mathbf{W} = \mathrm{diag}(\mathbf{w})$). 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) $\mathrm{span}\{\mathbf{u}\}$, (b) $(\mathrm{span}\{\mathbf{u}\})^\perp$, (c) $\mathrm{span}\{\mathbf{u}, \mathbf{v}\}$, (d) $(\mathrm{span}\{\mathbf{u}, \mathbf{v}\})^\perp$.

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**.

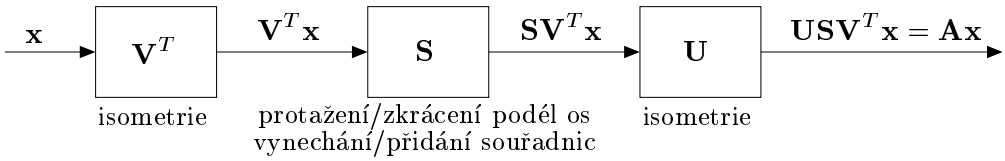
Reduced SVD is obtained from full SVD (7.1) by cutting matrix $\mathbf{S}$ to make it square $r \times r$, leaving out the last $m-r$ columns from matrix $\mathbf{U}$ and leaving out the last $n-r$ columns from matrix $\mathbf{V}$. Full SVD is obtained from reduced SVD by adding columns to slim matrices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

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

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

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* $\mathbf{S} = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$. *Let* $\mathbf{S}' = \operatorname{diag}(\sigma'_1, \ldots, \sigma'_r)$, *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 $\mathbf{u}_1, \ldots, \mathbf{u}_m$ are the columns of matrix $\mathbf{U}$ and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ 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] $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$. This task can not be turned into the least squares problem of §6.1. However, it can be solved by using Theorem 7.1:

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

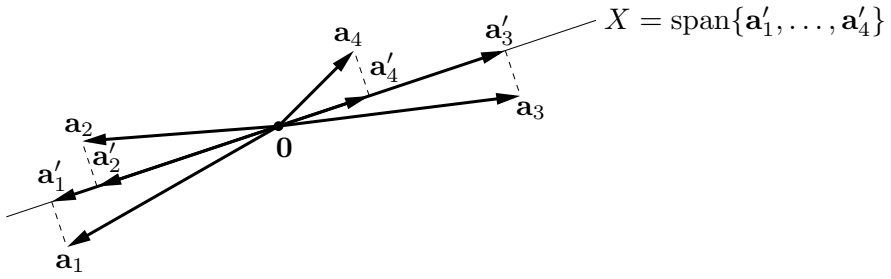
$$r' = \operatorname{rank}\mathbf{A}' = \dim \operatorname{span}\{\mathbf{a}'_1, \ldots, \mathbf{a}'_n\}, \tag{7.11}$$

[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 $\mathbf{A} = \mathbf{USV}^T$. 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 $\begin{bmatrix}\mathbf{a}_1 \cdots \mathbf{a}_n\end{bmatrix} = \mathbf{USV}^T$. 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

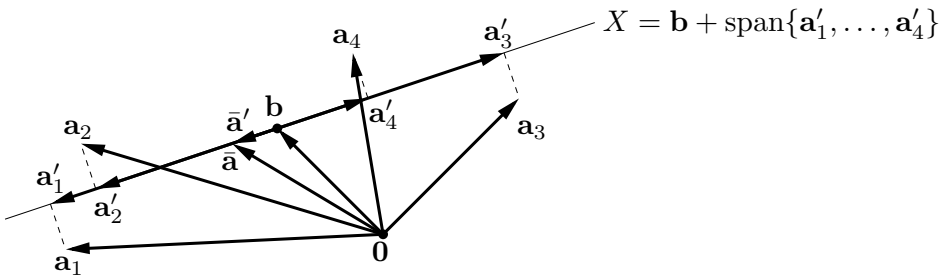
$$\operatorname{span}\{\mathbf{u}_1\} = \{\, \alpha\mathbf{u}_1 \mid \alpha \in \mathbb{R} \,\} = \{\, \mathbf{x} \in \mathbb{R}^3 \mid \mathbf{u}_2^T\mathbf{x} = \mathbf{u}_3^T\mathbf{x} = 0 \,\} = \operatorname{span}\{\mathbf{u}_2, \mathbf{u}_3\}^\perp.$$

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. $f\colon \mathbb{R}^n \to \mathbb{R}$, $f(\mathbf{x}) = \max_{i=1}^{n} x_i$
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 $k \in \{1, \ldots, n\}$ and let function $g\colon \mathbb{R}^n \to \mathbb{R}$ be given by $g(x_1, \ldots, x_n) = f(x_k)$ (i.e. g depends solely on variable x_k). Then function g is continuous at every point $(x_1, \ldots, x_n)$ 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 $f_1, \ldots, f_m\colon \mathbb{R}^n \to \mathbb{R}$ be continuous at point $\mathbf{x}$. Then the mapping $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ defined by $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$ 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 funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ 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

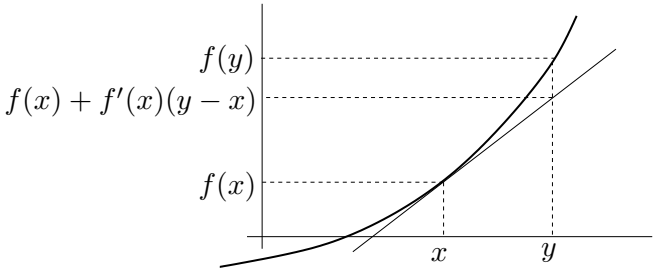
Let us review the definition of the derivative of function $f\colon \mathbb{R} \to \mathbb{R}$ of a single variable at point x. When the limit

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

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:

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

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.

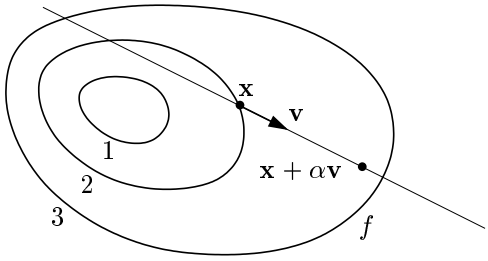
$\mathbf{f}(\mathbf{x})$	$\mathbf{f}'(\mathbf{x})$	notes
$\mathbf{x}$	$\mathbf{I}$	$\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$
$\mathbf{A}\mathbf{x}$	$\mathbf{A}$	$\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$
$\mathbf{x}^T\mathbf{x}$	$2\mathbf{x}^T$	$\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}$
$\mathbf{x}^T\mathbf{A}\mathbf{x}$	$\mathbf{x}^T(\mathbf{A} + \mathbf{A}^T)$	$\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}$
$\mathbf{x}^T\mathbf{a} = \mathbf{a}^T\mathbf{x}$	$\mathbf{a}^T$	$\mathbf{a} \in \mathbb{R}^n$, $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}$
$\|\mathbf{x}\|_2$	$\mathbf{x}^T/\|\mathbf{x}\|_2$	$\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}$
$\mathbf{g}(\mathbf{A}\mathbf{x})$	$\mathbf{g}'(\mathbf{A}\mathbf{x})\mathbf{A}$	$\mathbf{A} \in \mathbb{R}^{\ell\times n}$, $\mathbf{g}\colon \mathbb{R}^\ell \to \mathbb{R}^m$, $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$
$\mathbf{g}(\mathbf{x})^T\mathbf{g}(\mathbf{x})$	$2\mathbf{g}(\mathbf{x})^T\mathbf{g}'(\mathbf{x})$	$\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}$
$\mathbf{g}(\mathbf{x})^T\mathbf{h}(\mathbf{x})$	$\mathbf{g}(\mathbf{x})^T\mathbf{h}'(\mathbf{x}) + \mathbf{h}(\mathbf{x})^T\mathbf{g}'(\mathbf{x})$	$\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{h}\colon \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}$

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 i^{th} standard basis vector $(0,\ldots,0,1,0,\ldots,0)$ (1 in the i^{th} 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].

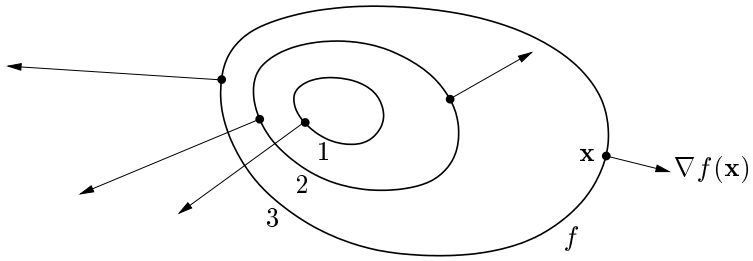
Consider the directional derivatives at a fixed point $\mathbf{x}$ in various directions given by a normalised vector $\mathbf{v}$ (i.e. $\|\mathbf{v}\|_2 = 1$). Such derivative is $f'(\mathbf{x})\mathbf{v}$, that is the scalar product of the gradient at point $\mathbf{x}$ and the vector $\mathbf{v}$. 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 §9.2.1) that the gradient is always *perpendicular to the contour*.

The following figure shows three contours of funkce $f\colon \mathbb{R}^2 \to \mathbb{R}$ 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 funkce jedné proměnné $f\colon \mathbb{R} \to \mathbb{R}$ má v bodě x derivace až do řádu k. Její **Taylorův polynom** stupně k v bodě x is funkce $T_k\colon \mathbb{R} \to \mathbb{R}$ daná předpisem

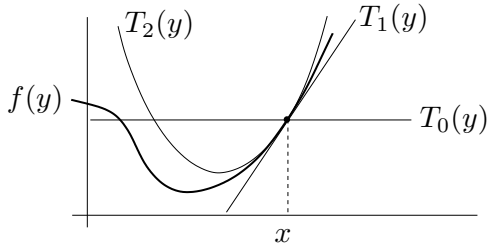
$$T_k(y) = \sum_{i=0}^{k} \frac{1}{i!} f^{(i)}(x)\,(y-x)^i, \tag{8.10}$$

kde $f^{(i)}$ označuje i-tou derivaci funkce f (kde nultá derivace je funkce sama, $f^{(0)} = f$) a kde klademe $0! = 1$. Polynom T_k je definován vlastností, že v bodě x má všechny derivace až do řádu k stejné jako funkce f (dokažte!). V tomto smyslu is polynom T_k aproximací funkce f v okolí bodu x.

Tvary polynomu až do řádu 2:

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

Taylorův polynom nultého řádu T_0 is hodně špatná aproximace, rovná jednoduše konstantní funkci. Polynom prvního řádu $T_1(x)$ už známe ze vzorce (8.2). Polynom druhého řádu T_2 is parabola, která má s funkcí f v bodě x společnou hodnotu a první dvě derivace. Viz obrázek:



Jak zobecnit Taylorův polynom pro funkci více proměnných $f\colon \mathbb{R}^n \to \mathbb{R}$? Nebudeme uvádět vzorec pro polynom libovolného stupně, napíšeme jen polynomy do stupně dva:

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

Zde $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $f'(\mathbf{x}) \in \mathbb{R}^{1\times n}$ is Jacobiho matrix (řádkový vector) a $f''(\mathbf{x}) \in \mathbb{R}^{n\times n}$ je Hessova matrix. Funkce (8.11b) is affine a funkce (8.11c) je kvadratická.

Taylorův polynom lze zobecnit na mapping $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ tak, že vezmeme Taylorovy polynomy všech složek $f_1, \ldots, f_m$. Polynom prvního stupně tak vede na mapping

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

což není nic jiného než (8.3). Polynom druhého stupně vede na mapping $\mathbf{T}_2$, jehož složky jsou funkce (8.11c). To nejde napsat v maticové formě, protože všech $m \times n \times n$ druhých partialch derivací se 'nevejde' do matrix.

8.8 Vlastnosti podmnožin $\mathbb{R}^n$

Pro $\varepsilon > 0$ a $\mathbf{x} \in \mathbb{R}^n$ se set

$$U_\varepsilon(\mathbf{x}) = \{\, \mathbf{y} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{y}\| < \varepsilon \,\} \tag{8.13}$$

nazývá[3] **ε-okolí** bodu $\mathbf{x}$. is to koule (bez hranice) se středem $\mathbf{x}$ a nenulovým poloměrem ε.

Definition 8.3. *consider množinu $X \subseteq \mathbb{R}^n$. Bod $\mathbf{x} \in \mathbb{R}^n$ se nazývá její*

- **vnitřní bod**, *jestliže existuje $\varepsilon > 0$ tak, že $U_\varepsilon(\mathbf{x}) \subseteq X$*
- **hraniční bod**, *jestliže pro každé $\varepsilon > 0$ platí $U_\varepsilon(\mathbf{x}) \cap X \neq \emptyset$ a $U_\varepsilon(\mathbf{x}) \cap (\mathbb{R}^n \setminus X) \neq \emptyset$*
- **vnější bod**, *jestliže existuje $\varepsilon > 0$ tak, že $U_\varepsilon(\mathbf{x}) \cap X = \emptyset$*
- **hromadný bod**, *jestliže pro každé $\varepsilon > 0$ platí $(U_\varepsilon(\mathbf{x}) \setminus \{\mathbf{x}\}) \cap X \neq \emptyset$*
- **izolovaný bod**, *jestliže existuje $\varepsilon > 0$ tak, že $U_\varepsilon(\mathbf{x}) \cap X = \{\mathbf{x}\}$.*

Všimněte si, že hraniční a hromadný bod set nemusí patřit do této set. Pokud leží bod v množině, is buď vnitřní nebo hraniční, ale ne obojí najednou (dokažte!). **Vnitřek** [**hranice**] set je set všech jejích vnitřních [hraničních] bodů.

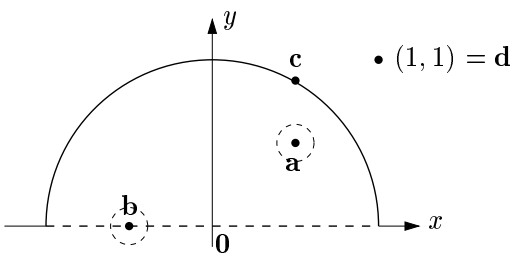
set se nazývá

- **otevřená**, jestliže všechny její body jsou vnitřní;
- **uzavřená**, jestliže obsahuje každý svůj hraniční bod.

Lze dokázat, že set X is uzavřená [otevřená], právě když její complement $\mathbb{R}^n \setminus X$ is otevřený [uzavřený]. Otevřenost a uzavřenost se nevylučují: set $\emptyset$ a $\mathbb{R}^n$ jsou zároveň otevřené i uzavřené. Naopak, některé set nejsou ani otevřené ani uzavřené, např. interval $(0, 1]$.

set X is **omezená**, jestliže existuje $r \in \mathbb{R}$ takové, že $\|\mathbf{x} - \mathbf{y}\|_2 < r$ pro všechna $\mathbf{x}, \mathbf{y} \in X$. Jinými slovy, set se 'vejde' do koule konečného průměru.

Example 8.11. Given množinu $\{\, (x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1,\ y > 0 \,\} \cup \{(1, 1)\} \subseteq \mathbb{R}^2$ na obrázku:



[3] Norma v (8.13) může být eukleidovská, ale i libovolná jiná vectorová p-norma (viz §11.2.1). Vnitřek a hranice set na výběru normy nezávisí.

Bod $\mathbf{a}$ is vnitřní bod set, protože existuje $\varepsilon > 0$ takové, že okolí $U_\varepsilon(\mathbf{a})$ celé leží v množině. Bod $\mathbf{b}$ je hraniční, protože okolí $U_\varepsilon(\mathbf{b})$ má pro každé $\varepsilon > 0$ neprázdný průnik s set i s jejím doplňkem. Všimněte si, že $\mathbf{b}$ nepatří do set. Bod $\mathbf{a}$ není hraniční a bod $\mathbf{b}$ není vnitřní. Bod $\mathbf{c}$ není vnitřní, is hraniční a patří do set. Bod $\mathbf{d}$ není vnitřní, is hraniční a patří do set. Body $\mathbf{a}, \mathbf{b}, \mathbf{c}$ jsou hromadné, bod $\mathbf{d}$ is izolovaný.

set není otevřená, protože např. bod $\mathbf{c}$ není vnitřní. Není ani uzavřená, protože např. bod $\mathbf{b}$ is hraniční ale nepatří do set. set is omezená. □

Example 8.12. Bod $1/2$ is vnitřní bod intervalu $(0, 1] \subseteq \mathbb{R}$ a body 0 a 1 jsou hraniční. □

Example 8.13. Given množinu $[0, 1] \times \{1\} = \{\, (x, y) \mid 0 \leq x \leq 1,\ y = 1 \,\} \subseteq \mathbb{R}^2$ (úsečka v rovině). Nemá žádné vnitřní body. Všechny její body jsou hraniční a hromadné, is tedy sama svou vlastní hranicí. Není otevřená, is uzavřená, není omezená. □

8.9 Věta o extrémní hodnotě

Uvažujme obraz

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

set $X \subseteq \mathbb{R}^n$ v mapping $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$. Položme si otázku, zda některá mapping zachovávají vlastnosti jako otevřenost, uzavřenost či omezenost. Tedy např. je-li X uzavřená, jestli is také $\mathbf{f}(X)$ uzavřená. Následující větu uvádíme bez důkazu.

Theorem 8.4. *Spojité mapping uzavřené omezené set is uzavřená omezená set.*

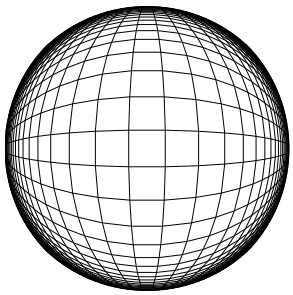
Mohlo by se zdát, že spojité mapping bude zachovávat např. uzavřenost bez omezenosti. Uveďme protipříklad.

Example 8.14. Let X is interval $[1, +\infty) \subseteq \mathbb{R}$. Tato set je uzavřená a není omezená. Spojité mapping $f(x) = 1/x$ zobrazí tuto množinu na interval $f(X) = (0, 1]$, který není uzavřený a is omezený. □

Example 8.15. consider spojité mapping $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^n$ dané vzorcem

$$\mathbf{f}(\mathbf{x}) = (1 + \mathbf{x}^T\mathbf{x})^{-1/2}\,\mathbf{x}.$$

Obraz neomezené set $\mathbb{R}^n$ v mapping $\mathbf{f}$ is otevřená omezená set $\mathbf{f}(\mathbb{R}^n) = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x}^T\mathbf{x} < 1 \,\}$ (jednotková koule bez hranice). Pro ilustraci na obrázku ukažme množinu $\mathbf{f}(X) \subseteq \mathbb{R}^2$ pro $X = (\mathbb{R} \times \mathbb{Z}) \cup (\mathbb{Z} \times \mathbb{R}) \subseteq \mathbb{R}^2$ (tedy X is pravidelná mřížka v rovině):



□

Věta 8.4 má důležitý důsledek pro optimalizaci, který is znám jako *věta o extrémní hodnotě* nebo *Weierstrassova věta*.

Theorem 8.5. *Spojitá funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ nabývá na uzavřené omezené množině $X \subseteq \mathbb{R}^n$ svého minima. Neboli existuje prvek $\mathbf{x}^* \in X$ takový, že $f(\mathbf{x}^*) = \min f(X) = \min_{\mathbf{x}\in X} f(\mathbf{x})$.*

Proof. Pro funkci $f\colon \mathbb{R}^n \to \mathbb{R}$ is obraz uzavřené omezené set $X \subseteq \mathbb{R}^n$ uzavřená omezená set $f(X) \subseteq \mathbb{R}$. To ale nemůže být nic jiného než uzavřený konečný interval nebo sjednocení takových intervalů. Taková set jistě má nejmenší prvek. □

8.10 Exercises

8.1. is dána funkce dvou proměnných $f(x, y)$.

a) Spočítejte derivace f podle polárních souřadnic (φ, r), kde $x = r\cos\varphi$, $y = r\sin\varphi$.
b) Bod (x, y) se v čase t pohybuje po křivce dané rovnicí $(x, y) = (t^2 + 2t, \ln(t^2 + 1))$. Najděte derivaci f podle času.

8.2. Spočítejte derivaci funkce $g(\mathbf{u}) = f(\mathbf{a}^T\mathbf{u}, \|\mathbf{u}\|_2)$ podle vectoru $\mathbf{u}$.

8.3. Nadmořská výška krajiny is dána vzorcem $h(d, s) = 2s^2 + 3sd - d^2 + 5$, kde d is zeměpisná délka (zvětšuje se od západu k východu) a s is zeměpisná šířka (zvětšuje se od jihu k severu). V bodě $(s, d) = (1, -1)$ určete

a) směr nejstrmějšího stoupání terénu
b) strmost terénu v jihovýchodním směru.

8.4. Spočítejte druhou derivaci $f''(x, y)$ (i.e., Hessovu matici) funkcí

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

8.5. Hessova matrix kvadratické formy $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$ je $f''(\mathbf{x}) = \mathbf{A} + \mathbf{A}^T$. Odvoďte.

8.6. is dána funkce $f(x, y) = 6xy^2 - 2x^3 - 3y^3$. V bodě $(x_0, y_0) = (1, -2)$ najděte Taylorův polynom prvního a druhého stupně.

8.7. *Metoda konečných diferencí* počítá derivaci funkce přibližně jako

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

kde h is malé číslo (dobrá volba is $h = \sqrt{\varepsilon}$, kde ε is strojová přesnost). Toto jde použít i na partial derivace. Vymyslete si dvě mapping $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^m$ a $\mathbf{f}\colon \mathbb{R}^m \to \mathbb{R}^\ell$ pro nějaké navzájem různé dimenze $n, m, \ell > 1$. Zvolte bod $\mathbf{x} \in \mathbb{R}^n$. Spočítejte přibližně totální derivace (Jacobiho matrix) $\mathbf{g}'(\mathbf{x})$ a $\mathbf{f}'(\mathbf{g}(\mathbf{x}))$ v Matlab metodou konečných diferencí. Potom spočítejte derivaci složeného mapping $(\mathbf{f} \circ \mathbf{g})'(\mathbf{x})$ jednak metodou konečných diferencí a jednak vynásobením matic $\mathbf{g}'(\mathbf{x})$ a $\mathbf{f}'(\mathbf{g}(\mathbf{x}))$. Porovnejte.

8.8. Načrtněte následující podset $\mathbb{R}^2$:

a) $[-1, 0] \times \{1\}$
b) $\{\, (x, y) \mid x > 0,\ y > 0,\ xy = 1 \,\}$

c) $\{\,(x,y) \mid \min\{x,y\} = 1\,\}$

8.9. Každá z následujících množin is sjednocením konečného počtu (otevřených, uzavřených či polouzavřených) intervalů. Najděte tyto intervaly. Příklad: $\{\,x^2 \mid x \in \mathbb{R}\,\} = [0, +\infty)$.

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) $\{\,x + y \mid |x| \geq 1,\ |y| \geq 1\,\}$

8.10. Given set $X = [-1,1] \times \{0\} = \{\,(x,0) \mid -1 \leq x \leq 1\,\} \subseteq \mathbb{R}^2$ a $Y = [-1,1] \times [-1,1]$. Načrtněte následující set:

a) $\left\{\,\mathbf{y} \in \mathbb{R}^2 \,\middle|\, \min_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{y}\|_2 \leq 1\,\right\}$
b) $\left\{\,\mathbf{y} \in \mathbb{R}^2 \,\middle|\, \max_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{y}\|_2 \leq 2\,\right\}$
c) vrstevnice výšky 1 funkce $f(\mathbf{x}) = \min_{\mathbf{x} \in Y} \|\mathbf{x} - \mathbf{y}\|_2$
d) vrstevnice výšky $\sqrt{2}$ funkce $f(\mathbf{x}) = \max_{\mathbf{x} \in Y} \|\mathbf{x} - \mathbf{y}\|_2$

8.11. Co is vnitřek a hranice těchto množin?

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) $\{\,\mathbf{x} \in \mathbb{R}^n \mid b \leq \mathbf{a}^T\mathbf{x} \leq c\,\}$, kde $\mathbf{a} \in \mathbb{R}^n$, $b, c \in \mathbb{R}$ (panel)
g) $\{\,\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{b}\,\}$, kde $\mathbf{A}$ is široká (afinní subspace $\mathbb{R}^n$)

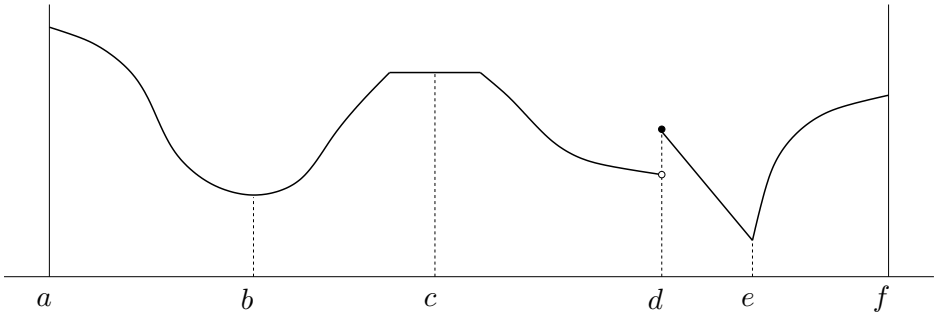
Chapter 9

Analytické Conditions on Local Extrema

Definition 9.1. *consider funkci* $f\colon \mathbb{R}^n \to \mathbb{R}$ *a množinu* $X \subseteq \mathbb{R}^n$. *Bod* $\mathbf{x} \in X$ *se nazývá* **lokální minimum** *funkce* f *na množině* X, *existuje-li* $\varepsilon > 0$ *takové, že* $\mathbf{x}$ *is minimum funkce* f *na množině* $U_\varepsilon(\mathbf{x}) \cap X$ *(viz §1.2), neboli* $f(\mathbf{x}) \leq f(\mathbf{y})$ *pro všechna* $\mathbf{y} \in U_\varepsilon(\mathbf{x}) \cap X$. *Lokální maximum se definuje obdobně.*

Každé minimum funkce f na množině X is zároveň lokální minimum funkce f na množině X (naoThen obecně neplatí). Mluvíme-li o lokálních extrémech, pro zdůraznění někdy 'obyčejné' extrémy nazýváme **globální extrémy**. Pokud odkaz na množinu X chybí, myslí se celý definiční obor $X = \mathbb{R}^n$.

Example 9.1. Funkce jedné proměnné na obrázku má na uzavřeném intervalu $[a, f]$ v bodě a lokální a zároveň globální maximum, v bodě b lokální minimum, v bodě c lokální maximum, v bodě d lokální maximum, v bodě e lokální a zároveň globální minimum, v bodě f lokální maximum.



□

Example 9.2.

- Funkce $f(x) = \sin x$ má v každém bodě $\pi/2 + 2k\pi$ lokální maximum a globální maximum.
- Funkce $f(x, y) = x^2 + y^2$ má v bodě $(0, 0)$ globální minimum.
- Funkce $f(x, y) = x^2$ má v bodech $(0, y)$ pro $y \in \mathbb{R}$ globální minimum.
- Konstantní funkce má v každém bodě globální i lokální minimum i maximum.
- Funkce $f(\mathbf{x}) = x_1$ má na množině $\{\, \mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 \leq 1 \,\}$ globální minimum v bodě $(-1, 0, \ldots, 0)$ a globální maximum v bodě $(1, 0, \ldots, 0)$. □

9.1 Volné lokální extrémy

Theorem 9.1. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ a $\mathbf{x} \in X \subseteq \mathbb{R}^n$. Necht'*

- *funkce f is v bodě $\mathbf{x}$ diferencovatelná,*
- *$\mathbf{x}$ je vnitřní bod set X,*
- *$\mathbf{x}$ je lokální extrém funkce f na množině X.*

Pak $f'(\mathbf{x}) = \mathbf{0}$, neboli všechny partial derivace funkce f v bodě $\mathbf{x}$ jsou nulové.

Proof. Z Definice 9.1 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 $\alpha = 0$ lokální extrém. Tedy její derivace is v tomto bodě nulová (to víme z analýzy funkcí jedné proměnné). Ale tato derivace is směrová derivace funkce f v bodě $\mathbf{x}$ ve směru $\mathbf{v}$. partial derivace jsou speciálním případem směrové derivace. □

Bod, ve kterém má funkce všechny partial derivace nulové, se nazývá její **stacionární bod**. Věta 9.1 svádí k tomu, aby se použila v situacích, kdy nejsou splněny její předpoklady. Uveďme příklady tohoto chybého použití.

Example 9.3. V Příkladu 9.1 jsou předpoklady Věty 9.1 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 9.4. Funkce $f(x) = x^3$ má na $\mathbb{R}$ v bodě 0 stacionární bod, ale nemá tam lokální extrém. To není v rozporu s Větou 9.1. □

Example 9.5. Funkce $f(\mathbf{x}) = \|\mathbf{x}\|_2$ má na hyperkrychli $\{\, \mathbf{x} \in \mathbb{R}^n \mid -\mathbf{1} \leq \mathbf{x} \leq \mathbf{1} \,\}$ v bodě $\mathbf{0}$ volné lokální 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 vázaná lokální maxima ve všech rozích hyperkrychle, např. v bodě $\mathbf{1}$. V bodě $\mathbf{1}$ ale není stacionární bod, což není v rozporu s Větou 9.1, protože bod $\mathbf{1}$ není vnitřní bod hyperkrychle. □

Věta 9.1 říká, že stacionární body jsou body 'podezřelé' z volného lokálního extrému. Udává podmínku *prvního řádu* na volné extrémy, protože obsahuje první derivace. Následující podmínka *druhého řádu* pomůže zjistit, zda is stacionární bod lokálním extrémem, případně jakým.

Theorem 9.2. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ a $\mathbf{x} \in X \subseteq \mathbb{R}^n$. Necht'*

- *funkce f is v bodě $\mathbf{x}$ dvakrát diferencovatelná,*
- *$\mathbf{x}$ je vnitřní bod set X,*
- *$f'(\mathbf{x}) = \mathbf{0}$.*

Then platí:

- *Je-li Hessova matrix $f''(\mathbf{x})$ pozitivně [negativně] definitní, pak $\mathbf{x}$ je lokální minimum [maximum] funkce f na množině X.*
- *Je-li $f''(\mathbf{x})$ indefinitní, pak $\mathbf{x}$ není lokální extrém funkce f na množině X.*

I když Větu 9.2 nebudeme dokazovat, základní myšlenka důkazu není překvapující. Místo funkce f vyšetřujeme v blízkosti bodu $\mathbf{x}$ její Taylorův polynom druhého řádu (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 $f'(\mathbf{x}) = \mathbf{0}$, linear člen is nulový a polynom is tedy pouhá kvadratická forma posunutá do bodu $\mathbf{x}$. Rozdíl is ale v tom, že pokud is kvadratická forma (pozitivně či negativně) semidefinitní, má v počátku extrém, zatímco Věta 9.2 o případu, kdy je $f''(\mathbf{x})$ semidefinitní, nic nepraví. V tom případě v bodě $\mathbf{x}$ lokální 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 is $f'(\mathbf{x}) = \mathbf{0}$ a matrix $f''(\mathbf{x})$ is indefinitní, se nazývá **sedlový bod**.

Example 9.6. Extrémy kvadratické funkce (5.9) umíme hledat pomocí decomposition na čtverec. Ovšem is 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 is roven $2\mathbf{A}$ (předpokádáme symetrii $\mathbf{A}$). To souhlasí s klasifikací extrémů kvadratické formy z §5. □

9.2 Lokální extrémy vázané rovnostmi

Hledejme minumum funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ na množině

$$X = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{g}(\mathbf{x}) = \mathbf{0} \,\}, \tag{9.1}$$

kde $\mathbf{g} = (g_1, \ldots, g_m)\colon \mathbb{R}^n \to \mathbb{R}^m$. Tedy řešíme úlohu (1.4) s omezeními typu rovnosti:

$$\begin{array}{rl} \min & f(x_1, \ldots, x_n) \\ \text{za podmínek} & g_i(x_1, \ldots, x_n) = 0, \quad i = 1, \ldots, m. \end{array} \tag{9.2}$$

Mluvíme o minimu funkce f *vázaném rovnostmi* $\mathbf{g}(\mathbf{x}) = \mathbf{0}$.

set (9.1) obsahuje všechna řešení soustavy $\mathbf{g}(\mathbf{x}) = \mathbf{0}$, což is soustava m (obecně nelinárních) rovnic o n neznámých. Tato set obvykle nemá žádné vnitřní body, proto nelze použít Věty 9.1. V některých případech ale lze vyjádřit všechna řešení soustavy v parametrické formě, to jest najít mapping $\boldsymbol{\varphi}\colon \mathbb{R}^\ell \to \mathbb{R}^n$ takové, že $X = \{\, \boldsymbol{\varphi}(\mathbf{y}) \mid \mathbf{y} \in \mathbb{R}^\ell \,\}$. Then lze úlohu převést na úlohu bez omezení.

Example 9.7. 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 $X = \{\, (x, y) \in \mathbb{R}^2 \mid g(x, y) = 1 - xy = 0 \,\}$. set X nemá žádné vnitřní body (rozmyslete!), proto nelze použít Větu 9.1. Z podmínky ale we have $y = 1/x$, což dosazeno do účelové funkce dá $f(x, 1/x) = x + 1/x$. Najdeme lokální extrémy této funkce na množině $\mathbb{R}$. Dostaneme dva stacionární body $(x, y) = \pm(1, 1)$. □

Obvykle ale množinu (9.1) parametrizovat nejde. Nyní proto odvodíme obecnější postup, *metodu Lagrangeových multiplikátorů*.

9.2.1 Tečný subspace

Zkoumejme množinu X. Pokud is mapping $\mathbf{g}$ spojitě differentiable, set X is 'zakřivený hladký povrch' v $\mathbb{R}^n$. Zvolíme-li bod $\mathbf{x} \in X$, za jistých předpokladů existuje *tečný subspace* k množině X v bodě $\mathbf{x}$. Definice tečného subspaceu is dosti složitá a nebudeme ji uvádět[1]. Bez důkazu ale uvedeme, že pokud má Jacobiho matrix $\mathbf{g}'(\mathbf{x})$ hodnost m, Then tento tečný subspace je set

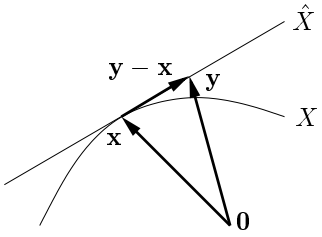
$$\hat{X} = \{\, \mathbf{y} \in \mathbb{R}^n \mid \hat{\mathbf{g}}(\mathbf{y}) = \mathbf{0} \,\}, \tag{9.3}$$

kde

$$\hat{\mathbf{g}}(\mathbf{y}) = \mathbf{g}(\mathbf{x}) + \mathbf{g}'(\mathbf{x})(\mathbf{y} - \mathbf{x}) = \mathbf{g}'(\mathbf{x})(\mathbf{y} - \mathbf{x})$$

je affine aproximace mapping $\mathbf{g}$ podle vztahu (8.3). Zde jsme využili, že $\mathbf{x} \in X$ a tedy $\mathbf{g}(\mathbf{x}) = \mathbf{0}$. set $\hat{X}$ is affine subspace $\mathbb{R}^n$.

Rovnice $\mathbf{g}'(\mathbf{x})(\mathbf{y} - \mathbf{x}) = \mathbf{0}$ říká, že pro každé $\mathbf{y} \in \hat{X}$ je vector $\mathbf{y} - \mathbf{x}$ kolmý na rows matrix $\mathbf{g}'(\mathbf{x})$, neboli platí $(\mathbf{y} - \mathbf{x}) \perp \operatorname{span}\{\nabla g_1(\mathbf{x}), \ldots, \nabla g_m(\mathbf{x})\}$. Všimněte si, že pro $m = 1$ is set X vrstevnice funkce g nulové výšky. Dostali jsme tedy tvrzení, které jsme bez důkazu uvedli v §8.5: gradient is vždy kolmý k vrstevnici.



Example 9.8. Let $g\colon \mathbb{R}^n \to \mathbb{R}$ is funkce $g(\mathbf{x}) = \mathbf{x}^T\mathbf{x} - 1$. set X is tedy jednotková n-rozměrná sféra. affine aproximace funkce g v bodě $\mathbf{x} \in X$ je

$$\hat{g}(\mathbf{y}) = g'(\mathbf{x})(\mathbf{y} - \mathbf{x}) = 2\mathbf{x}^T(\mathbf{y} - \mathbf{x}) = 2\mathbf{x}^T\mathbf{y} - 2,$$

kde jsme využili, že $\mathbf{x}^T\mathbf{x} = 1$. Tečný subspace $\hat{X} = \{\, \mathbf{y} \in \mathbb{R}^n \mid \mathbf{x}^T\mathbf{y} = 1 \,\}$ is tečná nadrovina ke sféře v bodě $\mathbf{x}$.

Speciálně pro $n = 2$ is set X kružnice a $\hat{X}$ is tečna k této kružnici. □

Example 9.9. Let $\mathbf{a} = (1, 0, 0) \in \mathbb{R}^3$ a $\mathbf{g} = (g_1, g_2)\colon \mathbb{R}^3 \to \mathbb{R}^2$ je mapping se složkami

$$g_1(\mathbf{x}) = \mathbf{x}^T\mathbf{x} - 1, \quad g_2(\mathbf{x}) = (\mathbf{x} - \mathbf{a})^T(\mathbf{x} - \mathbf{a}) - 1.$$

Nulová vrstevnice funkce g_1 is jednotková sféra se středem v bodě $\mathbf{0}$, nulová vrstevnice funkce g_2 is jednotková sféra se středem v bodě $\mathbf{a}$. set X is průnik těchto dvou sfér, tedy kružnice v $\mathbb{R}^3$. Tečný subspace $\hat{X}$ is tečna k této kružnici, tedy přímka v prostoru. □

Proč is nutný předpoklad, že Jacobiho matrix $\mathbf{g}'(\mathbf{x})$ má hodnost m? Pokud má hodnost menší než m, může se stát, že tečný subspace v bodě $\mathbf{x}$ buď neexistuje, nebo existuje ale nerovná se množině (9.3).

[1] Tato definice se formuluje v rámci *diferenciální geometrie*, která se zabývá studiem zakřivených prostorů. V diferenciální geometrii is naše set X příkladem objektu, který se nazývá *diferencovatelný manifold*.

Example 9.10. Let $g\colon \mathbb{R}^n \to \mathbb{R}$ is funkce $g(\mathbf{x}) = (\mathbf{x}^T\mathbf{x} - 1)^2$. Je jasné (proč?), že set

$$X = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x}^T\mathbf{x} - 1 = 0 \,\} = \{\, \mathbf{x} \in \mathbb{R}^n \mid (\mathbf{x}^T\mathbf{x} - 1)^2 = 0 \,\}.$$

je stejná sféra jako v Příkladu 9.8. Jacobiho matrix is (ověřte!)

$$g'(\mathbf{x}) = 4(\mathbf{x}^T\mathbf{x} - 1)\mathbf{x}^T.$$

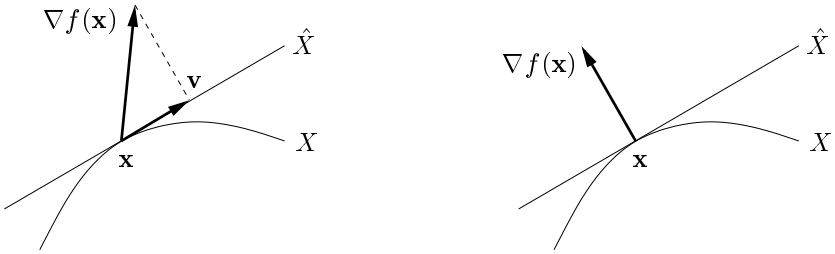
Ovšem pro každé $\mathbf{x} \in X$ we have $\mathbf{x}^T\mathbf{x} - 1 = 0$ a tedy $g'(\mathbf{x}) = \mathbf{0}$. Tedy gradient funkce g is nulový, neboli hodnost matrix $g'(\mathbf{x})$ je rovna nule. Tedy $\hat{X} = \{\, \mathbf{y} \in \mathbb{R}^n \mid g'(\mathbf{x})(\mathbf{y} - \mathbf{x}) = 0 \,\} = \mathbb{R}^n$.

Závěr: i když tečný prostor k množině X v každém jejím bodě existuje (našli jsme ho v Příkladu 9.8), tak zde není roven množině $\hat{X}$. □

9.2.2 Lokální minimum na tečném subspaceu

Lze ukázat, že pokud bod $\mathbf{x}$ is lokální extrém funkce f na množině X, Then bod $\mathbf{x}$ is lokální extrém funkce f také na tečném subspaceu $\hat{X}$. To naši úlohu výrazně zjednodušuje, protože místo na množině X ověřujeme existenci lokálního minima na affinem subspaceu $\hat{X}$.

Let vector $\mathbf{v}$ označuje průmět gradientu $\nabla f(\mathbf{x})$ do tečného subspaceu $\hat{X}$. Směrová derivace funkce f v bodě $\mathbf{x}$ ve směru $\mathbf{v}$ is číslo $f'(\mathbf{x})\mathbf{v}$. Pokud tato směrová derivace je non-zero (obrázek dole vlevo), bod $\mathbf{x}$ není lokální extrém funkce f na množině $\hat{X}$. Aby byla nulová, musí být $\mathbf{v} = \mathbf{0}$, neboli gradient $\nabla f(\mathbf{x})$ musí být kolmý na tečný subspace $\hat{X}$ (obrázek vpravo).



we have tedy tento výsledek: pokud $\mathbf{x}$ is lokální extrém funkce f na množině X, is vector $\nabla f(\mathbf{x})$ kolmý na tečný subspace $\hat{X}$. To znamená, že vector $\nabla f(\mathbf{x})$ is kolmý na nulový subspace Jacobiho matrix $\mathbf{g}'(\mathbf{x})$. Dle rovnosti (4.3) tedy platí

$$\nabla f(\mathbf{x}) \in [\operatorname{null} \mathbf{g}'(\mathbf{x})]^\perp = \operatorname{rng}[\mathbf{g}'(\mathbf{x})^T] = \operatorname{span}\{\nabla g_1(\mathbf{x}), \ldots, \nabla g_m(\mathbf{x})\}, \tag{9.4}$$

neboli vector $\nabla f(\mathbf{x})$ is linear kombinací vectors $\nabla g_i(\mathbf{x})$. Tedy existuje vector $\boldsymbol{\lambda} \in \mathbb{R}^m$ tak, že

$$f'(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{g}'(\mathbf{x}) = \mathbf{0}. \tag{9.5}$$

9.2.3 Podmínky prvního řádu

Výsledek úvah z §9.2.1 a §9.2.2 se obvykle formuluje následujícím způsobem.

Theorem 9.3. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$, $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^m$ a $\mathbf{x} \in X$. Necht'*

- *f a $\mathbf{g}$ jsou v bodě $\mathbf{x}$ spojitě differentiable,*
- *matrix $\mathbf{g}'(\mathbf{x})$ má hodnost m,*
- *bod $\mathbf{x}$ is lokální extrém funkce f na množině X.*

Then existují numbers $(\lambda_1, \ldots, \lambda_m) = \boldsymbol{\lambda} \in \mathbb{R}^m$ *tak, že* $L'(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{0}$, *kde funkce* $L\colon \mathbb{R}^{n+m} \to \mathbb{R}$ *je dána jako*

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x}) = f(\mathbf{x}) + \lambda_1 g_1(\mathbf{x}) + \cdots + \lambda_m g_m(\mathbf{x}). \tag{9.6}$$

Zápis $L'(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{0}$ označuje, že partial derivace funkce L podle $x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m$ jsou nulové, neboli bod $(\mathbf{x}, \boldsymbol{\lambda}) \in \mathbb{R}^{m+n}$ is stacionární bod funkce L. Rovnost $\partial L(\mathbf{x}, \boldsymbol{\lambda})/\partial \mathbf{x} = \mathbf{0}$ is ekvivalentní rovnosti (9.5). Rovnost $\partial L(\mathbf{x}, \boldsymbol{\lambda})/\partial \boldsymbol{\lambda} = \mathbf{g}(\mathbf{x}) = \mathbf{0}$ is ekvivalentní omezením. Číslům λ_i se říká **Lagrangeovy multiplikátory** a funkci (9.6) **Lagrangeova funkce**.

Example 9.11. Řešme znovu Příklad 9.7. is $L(x, y, \lambda) = x + y + \lambda(1 - xy)$ a řešíme soustavu

$$\begin{aligned} \partial L(x, y, \lambda)/\partial x &= 1 - \lambda y = 0 \\ \partial L(x, y, \lambda)/\partial y &= 1 - \lambda x = 0 \\ \partial L(x, y, \lambda)/\partial \lambda &= xy - 1 = 0. \end{aligned}$$

Soustava is zjevně splněna pro $(x, y, \lambda) = \pm(1, 1, 1)$. □

Example 9.12. Hledejme extrémy funkce $f(x, y) = x + y$ za podmínky $g(x, y) = 1 - x^2 - y^2 = 0$ we have $n = 2$, $m = 1$. Lagrangeova funkce is $L(x, y, \lambda) = x + y + \lambda(1 - x^2 - y^2)$. 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 we have dva kandidáty na lokální extrémy, $(x, y) = \pm(1, 1)/\sqrt{2}$.

Tuto jednoduchou úlohu is samozřejmě snadné vyřešit úvahou. Nakreslete si kružnici $X = \{\, (x, y) \mid x^2 + y^2 = 1 \,\}$ a několik vrstevnic funkce f a najděte kýžené extrémy! □

Example 9.13. Vrat'me se k úloze (6.14), tedy k hledání řešení nehomogení linear 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. is $\partial L(\mathbf{x}, \boldsymbol{\lambda})/\partial \mathbf{x} = 2\mathbf{x}^T - 2\boldsymbol{\lambda}^T \mathbf{A}$ (odvoď te!). Stacionární body funkce L tedy získáme řešením soustavy (6.15), kterou jsme v 6.2 odvodili úvahou. □

Příklad 9.14 vyžaduje od studenta nejen znalost metody Lagrangeových multiplikátorů, ale i zručnost v manipulaci s maticovými výrazy. is typické, že student správně napíše Lagrangeovu funkci a někdy i derivaci (9.7a), ale Then už nedokáže vyřešit soustavu (9.7). Trénujte tyto dovednosti ve Cvičeních 9.21–9.24! Pokud vám to nejde, zopakujte si §2!

Example 9.14. Najděte takové $\mathbf{x} \in \mathbb{R}^n$, které minimalizuje $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ za podmínky $\mathbf{c}^T \mathbf{x} = 0$. Předpokládejte, že $\mathbf{A}$ is obdélníková úzká s plnou hodností.

Místo $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ minimalizujme $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$. Lagrangeova funkce je

$$L(\mathbf{x}, \lambda) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + 2\lambda \mathbf{c}^T \mathbf{x} = \mathbf{x}^T \mathbf{A}^T \mathbf{A}\mathbf{x} - 2\mathbf{b}^T \mathbf{A}\mathbf{x} + \mathbf{b}^T \mathbf{b} + 2\lambda \mathbf{c}^T \mathbf{x},$$

kde jsme Lagrangeův multiplikátor nazvali 2λ. Řešíme soustavu rovnic

$$\partial L(\mathbf{x},\lambda)/\partial \mathbf{x} = \mathbf{A}^T\mathbf{A}\mathbf{x} - \mathbf{A}^T\mathbf{b} + \lambda\mathbf{c} = \mathbf{0} \tag{9.7a}$$
$$\partial L(\mathbf{x},\lambda)/\partial \lambda = 2\mathbf{c}^T\mathbf{x} = 0 \tag{9.7b}$$

s proměnnými $\mathbf{x}$ a λ. Z rovnice (9.7a) dostaneme

$$\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{b} - \lambda\mathbf{c}) = \mathbf{A}^+\mathbf{b} - \lambda(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{c}. \tag{9.8}$$

Dosazení do (9.7b) dá $\mathbf{c}^T\mathbf{A}^+\mathbf{b} = \lambda\mathbf{c}^T(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{c}$, z toho

$$\lambda = \frac{\mathbf{c}^T\mathbf{A}^+\mathbf{b}}{\mathbf{c}^T(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{c}}.$$

Dosazení do (9.8) dá hledané optimální řešení počáteční úlohy

$$\mathbf{x} = \mathbf{A}^+\mathbf{b} - \frac{\mathbf{c}^T\mathbf{A}^+\mathbf{b}}{\mathbf{c}^T(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{c}}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{c}.$$

□

Example 9.15. Řešme Příklad 9.12, kde ale omezení změníme na $g(x,y) = (1-x^2-y^2)^2 = 0$. Podle Příkladu 9.10 we have $g'(x,y) = (0,0)$ 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 lokální extrémy $(x,y) = \pm(1,1)/\sqrt{2}$ jsme nenašli. □

Věta 9.3 udává podmínky prvního řádu na extrémy vázané rovnostmi. Říká, že pokud $(\mathbf{x},\boldsymbol{\lambda})$ is stacionární bod Lagrangeovy funkce, Then bod $\mathbf{x}$ is 'podezřelý' z lokálního extrému funkce f na množině X. Jak poznáme, zda tento bod is lokální extrém, případně jaký? Podmínky druhého řádu pro vázané extrémy existují, jsou ale dosti složité a uvádíme je v §9.2.4. Zde pouze zdůrazníme, že druh lokálního extrému *nelze* zjistit podle definitnosti Hessovy matrix $L''(\mathbf{x},\boldsymbol{\lambda})$, tedy is chybou použít Větu 9.2 na funkci L. Důvodem je, že pokud $\mathbf{x}$ is vázaný lokální extrém funkce f a $(\mathbf{x},\boldsymbol{\lambda})$ is stacionární bod funkce L, pak $(\mathbf{x},\boldsymbol{\lambda})$ není lokální extrém funkce L. Naopak, lze ukázat, že *vždy* is to její sedlový bod.

9.2.4 (⋆) Podmínky druhého řádu

Theorem 9.4. *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* $\boldsymbol{\lambda} \in \mathbb{R}^m$. *Necht'*

- $(\mathbf{x},\boldsymbol{\lambda})$ *is 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 differentiable v bodě* $\mathbf{x}$.

Then platí:

- *Je-li* $\partial^2 L(\mathbf{x},\boldsymbol{\lambda})/\partial\mathbf{x}^2$ *pozitivně [negativně] definitní na nulovém prostoru matrix* $\mathbf{g}'(\mathbf{x})$, *má* f *v bodě* $\mathbf{x}$ *ostré lokální 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 matrix* $\mathbf{g}'(\mathbf{x})$*, nemá* f *v bodě* $\mathbf{x}$ *lokální minimum ani lokální 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 matrix $\mathbf{A}$ 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 matrix $\mathbf{A}$ na nulovém prostoru Jacobiho matrix $\mathbf{g}'(\mathbf{x})$? Najdeme-li bázi $\mathbf{B}$ tohoto nulového prostoru, Then každý prvek set 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 matrix $\mathbf{B}^T\mathbf{A}\mathbf{B}$.

Example 9.16. 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 we have 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 is zjevně splněna pro $(x,y,z,\lambda) = (1,1,1,2)$. Máme ukázat, že tento bod odpovídá lokálnímu minimu. Máme

$$\frac{\partial^2 L(x,y,z,\lambda)}{\partial (x,y,z)^2} = \begin{bmatrix} 0 & 1-\lambda z & 1-\lambda y \\ 1-\lambda z & 0 & 1-\lambda x \\ 1-\lambda y & 1-\lambda x & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix}. \tag{9.9}$$

Ukážeme, že tato matrix is pozitivně definitní na nulovém prostoru Jacobiho matrix

$$g'(x,y,z) = \begin{bmatrix} -yz & -xz & -xy \end{bmatrix} = \begin{bmatrix} -1 & -1 & -1 \end{bmatrix}.$$

Nejdříve zkusme štěstí, zda matrix (9.9) není pozitivně definitní již na $\mathbb{R}^3$ – v tom případě by zjevně byla pozitivně definitní i na nulovém prostoru $g'(x,y,z)$ (chvíli zamyslete, proč to tak je). Není tomu tak, protože její vlastní numbers jsou $\{-2,1,1\}$, tedy is indefinitní.

Nějakou bázi nulového prostoru matrix $g'(x,y,z)$ snadno najdeme ručně, např.

$$\mathbf{B} = \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Snadno zjistíme, že matrix

$$\mathbf{B}^T \frac{\partial^2 L(x,y,z,\lambda)}{\partial (x,y,z)^2} \mathbf{B} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

má vlastní numbers $\{2,1\}$, tedy is pozitivně definitní. □

9.3 Lokální extrémy vázané nerovnostmi ‘hrubou silou’

Změňme nyní úlohu (9.2) tak, že podmínky budou nerovnosti. we have tedy úlohu

$$\begin{array}{rl} \min & f(\mathbf{x}) \\ \text{za podmínek} & g_i(\mathbf{x}) \le 0, \quad i = 1, \dots, m. \end{array} \tag{9.10}$$

Omezení $g_i(\mathbf{x}) \le 0$ nazveme **aktivní** v bodě $\mathbf{x}$ když $g_i(\mathbf{x}) = 0$, a **neaktivní** když $g_i(\mathbf{x}) < 0$. Pokud $\mathbf{x}$ je optimální řešení úlohy (9.10), neaktivní omezení v bodě $\mathbf{x}$ nehrají žádnou roli, můžeme is odstranit a úloha se nezmění. Kdybychom věděli předem (což bohužel nevíme), která omezení budou v optimu neaktivní, mohli bychom is vypustit a úlohu tak zjednodušit.

Tato úvaha nám dovolí navrhnout algoritmus na vyřešení úlohy (9.10) ‘hrubou silou’. Pro každou podmnožinu $I \subseteq \{1, \dots, m\}$ najdeme všechna lokální minima funkce f za podmínek $g_i(\mathbf{x}) = 0$, $i \in I$. Pro každé takto získané lokální minimum ověříme, zda is přípustné, tedy zda splňuje $g_i(\mathbf{x}) \le 0$, $i = 1, \dots, m$.

Algoritmus lze snadno zobecnit na úlohu (1.4), obsahující omezení typu rovnosti i nerovnosti.

Tento algoritmus má nevýhodu v tom, že musíme vyzkoušet všech 2^m podmnožin I. Proto jej lze použít jen pro velmi malé m.

Example 9.17. Hledejme všechny lokální extrémy funkce $f(x, y, z)$ za podmínek

$$\begin{aligned} x^2 + y^2 + z^2 &\le 1 \\ z &\ge 0 \end{aligned}$$

we have $g_1(x, y, z) = x^2 + y^2 + z^2 - 1$ a $g_2(x, y, z) = -z$. set přípustných řešení is polokoule. Provedeme postupně tyto kroky:

- $I = \emptyset$ (*obě podmínky neaktivní*): Najdeme všechny lokální extrémy funkce f na celém $\mathbb{R}^3$. Pro každý nalezený lokální extrém ověříme, zda splňuje podmínky $x^2+y^2+z^2 \le 1$ a $z \ge 0$ (tedy leží v půlkouli).
- $I = \{1\}$ (*první podmínka aktivní, druhá neaktivní*): Najdeme všechny lokální extrémy funkce f na sféře $x^2 + y^2 + z^2 = 1$. Pro každý z nich ověříme, zda splňuje podmínku $z \ge 0$ (tedy leží na správné polovině sféry).
- $I = \{2\}$ (*první podmínka neaktivní, druhá aktivní*): Najdeme všechny lokální extrémy funkce f na rovině $z = 0$. Pro každý z nich ověříme, zda splňuje podmínku $x^2+y^2+z^2 \le 1$ (tedy leží v kruhu, jenž je průnikem koule a roviny $z = 0$).
- $I = \{1, 2\}$ (*obě podmínky aktivní*): Najdeme všechny lokální extrémy funkce f za podmínek $x^2 + y^2 = 1$ a $z = 0$ (tedy na kružnici). Nemusíme ověřovat nic. □

Example 9.18. Najděme všechny lokální extrémy funkce $f(x, y) = x^2y + y^2 + x$ na množině $\{\, (x, y) \in \mathbb{R}^2 \mid -1 \le x - y \le 1 \,\}$. Given dvě omezení, $-1 \le x - y$ a $x - y \le 1$. Protože obě najednou nemohou být aktivní, we have tři možnosti: žádné ativní, první aktivní, druhé aktivní.

Extrém bez omezení vyjde $(x, y) = (1, -\frac{1}{2})$, což ale není přípustné řešení. Vázané extrémy můžeme najít dosazením. Extrémy za omezení $x-y = -1$ jsou komplexní. Extrémy za omezení $x - y = 1$ jsou $x = \pm 1/\sqrt{3}$, $y = x - 1$. Extrém se záporným x is lokální minimum, extrém s kladným x is lokální maximum. Shrnuto, lokální extrémy na množině jsou celkově dva. □

9.4 Exercises

9.1. Co is vnitřek a hranice těchto množin?

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) $\{\,\mathbf{x} \in \mathbb{R}^n \mid b \leq \mathbf{a}^T\mathbf{x} \leq c\,\}$, kde $\mathbf{a} \in \mathbb{R}^n$, $b, c \in \mathbb{R}$ (panel)
g) $\{\,\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{b}\,\}$, kde $\mathbf{A}$ is široká (affine subspace $\mathbb{R}^n$)

9.2. is dána funkce $f\colon \mathbb{R}^n \to \mathbb{R}$, set $Y \subseteq X \subseteq \mathbb{R}^n$, a bod $\mathbf{x} \in Y$. Uvažujme dva výroky:

a) Funkce f má v bodě $\mathbf{x}$ lokální minimum na množině X.
b) Funkce f má v bodě $\mathbf{x}$ lokální minimum na množině Y.

Vyplývá druhý výrok z prvního? Vyplývá první výrok z druhého? Dokažte z definice lokálního extrému nebo vyvraťe nalezením protipříkadu.

9.3. (velmi snadné) Funkce $f(x,y,z)$ má stacionární bod $(2,1,5)$. Co se dá o tomto stacionárním bodě říci, když Hessova matrix $f''(2,1,5)$ v něm má vlastní numbers

a) $\{2,3,-1\}$
b) $\{2,3,0\}$
c) $\{0,-1,1\}$

9.4. Pro následující funkce spočítejte (na papíře) stacionární body. Pro každý stacionární bod určete, zda is to lokální minimum, lokální 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) $f(x,y) = 3x - x^3 - 3xy^2$ (jsou 4)
e) $f(x,y) = 6xy^2 - 2x^3 - 3y^4$ (jsou 3)
f) $f(x,y) = x^4/3 + y^4/2 - 4xy^2 + 2x^2 + 2y^2 + 3$ (je jich 5)
g) $f(x,y,z) = x^3 + y^3 + 2xyz + z^2$ (jsou 3: $(0,0,0)$, $(3/2,3/2,-9/4)$, $(3/2,3/2,-9/4)$)

Nápověda: Dávejte dobrý pozor při řešení soustav rovnic vzniklých z podmínky na stacionární bod. Snadno se totiž stane, že vám nějaké řešení unikne.

9.5. 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 9.7) a Then 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 lokální extrémy a případně jaké.

9.6. Najděte lokální extrémy funkcí

a) $f(x,y) = 2x - y$
b) $f(x,y) = x(y-1)$
c) $f(x,y) = x^2 + 2y^2$
d) $f(x,y) = x^2 y$
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 is dobré účelovou funkci zjednodušit, pokud to nezmění řešení.

9.7. Najděte extrémy 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 $xy + yz + zx = 1$

9.8. Najděte extrémy 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 $x^2 + y^2 + z^2 = 1$.

9.9. Rozložte dané kladné reálné číslo na součin n kladných reálných čísel tak, aby jejich součet byl co nejmenší.

9.10. 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.

9.11. Najděte bod nejblíže počátku na křivce

a) $x + y = 1$
b) $x + 2y = 5$
c) $y = x^3 + 1$
d) $x^2 + 2y^2 = 1$

9.12. Let $\mathbf{x}^*$ is bod nejblíže počátku na nadploše $h(\mathbf{x}) = 0$. Ukažte metodou Lagrangeových multiplikátorů, že vector $\mathbf{x}^*$ je kolmý k tečné nadrovině plochy v bodě $\mathbf{x}^*$.

9.13. Given kouli o poloměru r a středu $\mathbf{x}_0$, i.e., množinu $\{\, \mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{x}_0\|_2 \leq r \,\}$. Máme nadrovinu $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} = b \,\}$.

9.14. Do elipsy o daných délkách os vepište obdélník s maximálním obsahem. Předpokládejte přitom, že strany obdélníku jsou rovnoběžné s osami elipsy.

9.15. *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 is úhel paprsku od normály rozhraní a c_i is 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?

9.16. Rozdělení pravděpodobnosti diskrétní náhodné proměnné is funkce $p\colon \{1,\dots,n\} \to \mathbb{R}_+$ (i.e., soubor nezáporných čísel $p(1),\dots,p(n)$) splňující $\sum_{x=1}^{n} p(x) = 1$.

a) *Entropie* náhodné proměnné s rozdělením p is rovna $-\sum_{x=1}^{n} p(x)\log p(x)$, kde log is přirozený logaritmus. Najděte rozdělení s maximální entropií. Udělejte totéž za omezení, že is předepsána střední hodnota $\mu = \sum_{i=1}^{m} x\,p(x)$.
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 $p = q$.

9.17. (⋆) Given trojúhelník se stranami délek a, b, c. Uvažujme bod, který má takovou polohu, že součet čtverců jeho vzdáleností od stran trojúhelníku is nejmenší možný. Jaké budou vzdálenosti x, y, z tohoto bodu od stran trojúhelníku?

9.18. (⋆) Given krychli s délkou hrany 2. Do stěny krychle is vepsána kružnice (která má tedy poloměr 1) a okolo sousední stěny is opsána kružnice (která má tedy poloměr $\sqrt{2}$). Najděte nejmenší a největší vzdálenost mezi body na kružnicích.

9.19. (⋆) Najděte extrémy 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 $x, y, z, u, v, w > 0$.

9.20. 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.

9.21. Minimalizujte $\mathbf{x}^T\mathbf{x}$ za podmínky $\mathbf{a}^T\mathbf{x} = 1$. Jaký je geometrický význam úlohy?

9.22. Maximalizujte $\mathbf{a}^T\mathbf{x}$ za podmínky $\mathbf{x}^T\mathbf{x} = 1$. Jaký je geometrický význam úlohy?

9.23. Minimalizujte $\mathbf{x}^T\mathbf{A}\mathbf{x}$ za podmínky $\mathbf{b}^T\mathbf{x} = 1$, kde $\mathbf{A}$ je pozitivně definitní.

9.24. Minimalizujte $\|\mathbf{C}\mathbf{x}\|_2$ za podmínky $\mathbf{A}\mathbf{x} = \mathbf{b}$, kde $\mathbf{C}$ is square nebo úzká s linearly nezávislými sloupci.

9.25. ($\star$) Minimalizujte $\|\mathbf{C}\mathbf{x}\|_2$ za podmínek $\mathbf{A}\mathbf{x} = \mathbf{0}$ a $\mathbf{x}^T\mathbf{x} = 1$.

9.26. ($\star$) Minimalizujte $\|\mathbf{A}\mathbf{x}\|_2$ za podmínky $\mathbf{x}^T\mathbf{C}\mathbf{x} = 1$, kde $\mathbf{C}$ is positivně definitní.

9.27. ($\star$) Minimalizujte $\mathbf{a}^T\mathbf{x}$ za podmínky $\mathbf{x}^T\mathbf{C}\mathbf{x} = 1$, kde $\mathbf{C}$ je positivně definitní.

9.28. ($\star$) Jaké musí být vlastnosti matrix $\mathbf{A}$ a vectoru $\mathbf{b}$, aby $\max\{\, \|\mathbf{A}\mathbf{x}\|_2 \mid \mathbf{b}^T\mathbf{x} = 0 \,\} = 0$?

Chapter 10

Numerical Algorithms for Free Local Extrema

Zde se budeme věnovat numerickým iteračním algoritmům na nalezení volného lokálního minima diferencovatelných funkcí na množině $\mathbb{R}^n$.

10.1 Rychlost konvergence iteračních algoritmů

Numerický iterační algoritmus na řešení nějaké úlohy konstruuje posloupnost bodů $\mathbf{x}_k$, která konverguje k řešení úlohy $\mathbf{x}$. Posloupnost zbytků $a_k = \|\mathbf{x}_k - \mathbf{x}\|$ is nezáporná, $a_k \geq 0$, a konverguje k nule, $\lim_{k\to\infty} a_k = 0$. Zkoumejme rychlost konvergence této posloupnosti.

Pokud existuje limita

$$\lim_{k\to\infty} \frac{a_{k+1}}{a_k} = r, \tag{10.1}$$

řekneme, že posloupnost $\{a_k\}$ konverguje

- **sublinearly**, pokud $r = 1$
- **linearly**, pokud $0 < r < 1$
- **superlinearly**, pokud $r = 0$.

Je jasné, že čím is r menší, tím posloupnost konverguje 'rychleji'. Sublinear konvergence znamená velmi (často nepoužitelně) pomalý algoritmus. linear konvergence znamená přijatelnou rychlost, přibližně rovnou rychlosti konvergence geometrické řady. Většina numerických algoritmů konverguje linearly. Superlinear konvergence znamená výtečný algoritmus.

Example 10.1.

1. Posloupnost $\{a_k\} = \{2^{-k}\} = \left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \ldots\right\}$ konverguje linearly, protože $a_{k+1}/a_k = 1/2$, což is independent na k. Posloupnost is obyčejná geometrická řada.
2. Posloupnost $\{a_k\} = \{1/k\} = \left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\right\}$ konverguje sublinearly, protože $a_{k+1}/a_k = k/(k+1)$, což pro $k \to \infty$ se blíží 1.
3. Posloupnost $\{a_k\} = \{2^{-2^k}\} = \left\{\frac{1}{4}, \frac{1}{16}, \frac{1}{256}, \ldots\right\}$ konverguje superlinearly, protože

$$\frac{a_{k+1}}{a_k} = \frac{2^{-2^{k+1}}}{2^{-2^k}} = 2^{-2^{k+1}+2^k} = 2^{-2^k}$$

a tedy limita (10.1) is rovna 0.

Uvědomte si, jak fantasticky rychlá is to konvergence. Znamená to, že $a_{k+1} = a_k^2$, i.e., s každou iterací se zhruba *zdvojnásobí počet platných cifer*. Strojové přesnosti dosáhneme za několik málo iterací.

4. Posloupnost $\{a_k\} = \{k^{-k}\}$ konverguje superlinearly (limitu (10.1) spočtěte!).
5. Pro posloupnost $\{a_k\} = \{\frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}, \dots\}$, i.e., 'koktavou' verzi posloupnosti $\{2^{-k}\}$, limita (10.1) neexistuje, protože výraz a_{k+1}/a_k is jiný pro sudé a pro liché k.
6. Pro posloupnost $\{a_k\} = \{1, \frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{3}, \frac{1}{8}, \frac{1}{4}, \frac{1}{16}, \dots\}$, i.e., proložené posloupnosti $\{1/k\}$ a $\{2^{-k}\}$, limita (10.1) neexistuje z podobného důvodu. □

Poslední dva příklady ukazují nedostatečnost stávající definice: limita (10.1) neexistuje, přestože posloupnost is jinak 'rozumná'. Proto se zavádí obecnější definice. Řekneme, že posloupnost $\{a_k\}$ konverguje **alespoň** sublinearly [linearly, superlinárně], existuje-li posloupnost $\{a'_k\}$, která konverguje sublinearly [linearly, superlinárně] a $a'_k \geq a_k$ pro každé k. Např. posloupnost z příkladu 5 výše konverguje alespoň linearly, protože můžeme zvolit $a'_k = 2^{-k/2}$.

10.2 (⋆) Metoda zlatého řezu

Půlení intervalu is známá iterační metoda na hledání nulové hodnoty spojité funkce $g\colon \mathbb{R} \to \mathbb{R}$ (i.e., hledání kořene rovnice $g(x) = 0$). Metoda nevyžaduje počítání derivací funkce, které ani nemusejí existovat. Na začátku k-té iterace we have dva body $x_1 < x_2$ takové, že

$$g(x_1)g(x_2) < 0. \tag{10.2}$$

To zaručuje, že v intervalu $[x_1, x_2]$ leží aspoň jeden kořen. V $(k+1)$-ní iteraci přídáme bod $x_3 = (x_1 + x_2)/2$. Nezbytně bude bud' $g(x_1)g(x_3) < 0$ nebo $g(x_3)g(x_2) < 0$ nebo $g(x_3) = 0$. V prvním případě interval $[x_1, x_2]$ nahradíme intervalem $[x_1, x_3]$, ve druhém případě intervalem $[x_3, x_2]$. Pokračujeme stejně. Protože v každé iteraci se interval neučitosti zúží na polovinu, metoda konverguje linearly ($r = 1$).

Hledejme nyní nikoliv nulovou hodnotu, ale minimum spojité funkce $f\colon \mathbb{R} \to \mathbb{R}$. Funkci nazveme **unimodální** na intervalu $[x_1, x_2]$, pokud existuje bod x takový, že $x_1 < x < x_2$ a na intervalu $[x_1, x]$ funkce striktně klesá a na intervalu $[x, x_2]$ striktně roste. V tom případě má funkce na intervalu právě jedno minimum x, které se nabývá v jeho vnitřním bodě.

Zatímco pro zachycení kořene stačila dvojice bodů splňující (10.2), pro zachycení minima potřebujeme *trojici* bodů. Let v k-té iteraci Given tři body $x_1 < x_3 < x_2$ tak, že funkce is na intervalu $[x_1, x_2]$ unimodální a platí

$$[f(x_3) - f(x_1)][f(x_2) - f(x_3)] < 0. \tag{10.3}$$

Trojici (x_1, x_3, x_2) říkáme **závorka** (*bracket*). V $(k+1)$-ní přidáme bod x_4, dejme tomu mezi body x_3 a x_2. Musí nastat jeden z těchto případů:

- $f(x_3) \leq f(x_4)$: funkce is unimodální na intervalu $[x_1, x_3]$ a minimum is zachyceno závorkou (x_1, x_3, x_4).
- $f(x_3) > f(x_4)$: funkce is unimodální na intervalu $[x_3, x_2]$ a minimum is zachyceno závorkou (x_3, x_4, x_2).

Zůstává otázka, jak volit pozici bodů, aby bylo zaručeno největší možné zmenšení intervalu neurčitosti, a to při obou možnostech 1 a 2. Chceme, aby závorky (x_1, x_3, x_2), (x_1, x_3, x_4) a (x_3, x_4, x_2) byly rozděleny ve stejném poměru. Tedy

$$\frac{b}{a} = \frac{a}{c} = \frac{b-c}{c},$$

kde $a = x_3 - x_1$, $b = x_2 - x_3$, $c = x_4 - x_3$. Odtud dostaneme $\varphi - \varphi^{-1} = 1$, kde jsme označili $b/a = \varphi$. Kladný kořen této rovnice is číslo $\varphi = (1+\sqrt{5})/2 \approx 1.618$, známé z antiky jako **zlatý řez**. we have zaručeno, že v další iteraci bude interval neurčitosti φ-krát kratší.

Protože se v každé iteraci interval neurčitosti zmenší φ-krát, algoritmus konverguje linearly $(r = \varphi^{-1})$.

Kdy algoritmus půlení intervalu a algoritmus zlatého řezu ukončit? Lze ukázat, že kvůli zaokrouhlovacím chybám nejde interval neurčitosti zmenšit na méně než asi $\sqrt{\varepsilon}$, kde ε is strojová přesnost.

10.3 Sestupné metody

Iterační algoritmy na hledání lokálního minima spojité funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ mají tvar

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \, \mathbf{v}_k, \tag{10.4}$$

kde vector $\mathbf{v}_k \in \mathbb{R}^n$ is **směr hledání** a skalár $\alpha_k > 0$ is **délka kroku**. Ve třídě algoritmů zvaných **sestupné metody** (*descent methods*) hodnota funkce monotonně klesá[1], $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$.

Let is funkce f diferencovatelná. Směr $\mathbf{v}_k$ se nazývá **sestupný**, jestliže

$$f'(\mathbf{x}_k)\, \mathbf{v}_k < 0, \tag{10.5}$$

tedy směrová derivace ve směru $\mathbf{v}_k$ is 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, vector $f'(\mathbf{x}_k)$ is nutně nulový (proč?). Tedy x_k je stacionární bod. V tom případě $\mathbf{x}_k$ může (a skoro vždy je) ale také nemusí být lokální minimum.

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

$$\varphi(\alpha_k) = f(\mathbf{x}_k + \alpha_k \mathbf{v}_k) \tag{10.6}$$

přes všechny $\alpha_k \geq 0$. Tato úloha is v kontextu vícerozměrné optimisation nazývána *line search*. Úlohu stačí řešit přibližně. Takovou přibližnou metodu není obtížné vymyslet a proto se jí dále nebudeme zabývat. Poznamenejme ale, že metodu zlatého řezu nelze beze změn použít, protože funkce φ nemusí být unimodální.

Dále uvedeme nejznámější zástupce sestupných metod.

[1] Existují totiž i algoritmy, ve kterých hodnota $f(\mathbf{x}_k)$ neklesá monotonně (i.e., někdy stoupne a někdy klesne) a přesto konvergují k optimu (např. *subgradientní metody*).

10.4 Gradientní metoda

Tato nejjednodušší metoda volí zvolit směr sestupu jako záporný gradient funkce f v bodě $\mathbf{x}_k$:

$$\mathbf{v}_k = -f'(\mathbf{x}_k)^T = -\nabla f(\mathbf{x}_k). \tag{10.7}$$

Tento směr is sestupný, což is okamžitě vidět dosazením do (10.5).

Rychlost konvergence gradientní metody is linear. Konvergence je často pomalá kvůli 'cik-cak' chování. Výhodou metody is její spolehlivost, protože směr is vždy sestupný.

10.4.1 (⋆) Závislost na linear transformaci souřadnic

Transformujme vector proměnných $\mathbf{x}$ linear transformací $\tilde{\mathbf{x}} = \mathbf{A}\mathbf{x}$, kde $\mathbf{A}$ is square regular matrix. 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

$$\tilde{\mathbf{x}}_{k+1} = \tilde{\mathbf{x}}_k - \alpha_k \, \tilde{f}'(\tilde{\mathbf{x}}_k)^T. \tag{10.8}$$

Zkoumejme, jaké iteraci to odpovídá v původních proměnných. K tomu potřebujeme vyjádřit (10.8) 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 (10.8) a úpravou dostaneme

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \, (\mathbf{A}^T\mathbf{A})^{-1} f'(\mathbf{x}_k)^T. \tag{10.9}$$

To lze napsat ve tvaru (10.4) se směrem hledání

$$\mathbf{v}_k = -(\mathbf{A}^T\mathbf{A})^{-1} f'(\mathbf{x}_k)^T. \tag{10.10}$$

Tento směr se liší od původního směru (10.7) vynásobením maticí $(\mathbf{A}^T\mathbf{A})^{-1}$. Vidíme tedy, že gradientní metoda *není invariantní* vůči linear transformaci souřadnic.

Ovšem lze ukázat, že nový směr (10.10) is také sestupný. Dosazením (10.7) do (10.5) to znamená, že $-f'(\mathbf{x}_k)(\mathbf{A}^T\mathbf{A})^{-1}f'(\mathbf{x}_k)^T < 0$. To is ale pravda, neboť matrix $\mathbf{A}^T\mathbf{A}$ a tedy i její inverze is pozitivně definitní, viz Cvičení 5.18.

Na vzorec (10.10) se lze dívat ještě obecněji. is jasné, že směr $\mathbf{v}_k = -\mathbf{C}_k^{-1} f'(\mathbf{x}_k)^T$ is sestupný, je-li matrix $\mathbf{C}_k$ pozitivně definitní. Dá se ukázat i opak, totiž že každý sestupný směr lze napsat takto. matrix $\mathbf{C}_k$ může být jiná v každém kroku. Uvidíme, že algoritmy uvedené dále budou mít vždy tento tvar.

10.5 Newtonova metoda

Newtonova metoda (přesněji Newton-Raphsonova) is slavný iterační algoritmus na řešení soustav Nonlinearch rovnic. Lze ho použít i na minimalizaci funkce tak, že hledáme nulový gradient. Oba způsoby použití popíšeme.

10.5.1 Použití na soustavy Nonlinearch rovnic

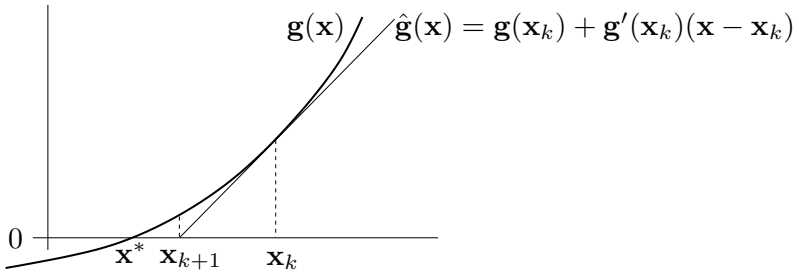
Řešme rovnici $\mathbf{g}(\mathbf{x}) = \mathbf{0}$, kde $\mathbf{g}$: $\mathbb{R}^n \to \mathbb{R}^n$ je differentiable mapping. Jedná se tedy o soustavu n rovnic s n neznámými, které obecně mohou být Nonlinear. mapping $\mathbf{g}$ aproximujeme v okolí bodu $\mathbf{x}_k$ Taylorovým polynomem prvního řádu

$$\mathbf{g}(\mathbf{x}) \approx \hat{\mathbf{g}}(\mathbf{x}) = \mathbf{g}(\mathbf{x}_k) + \mathbf{g}'(\mathbf{x}_k)\,(\mathbf{x} - \mathbf{x}_k), \tag{10.11}$$

kde $\mathbf{g}'(\mathbf{x}_k)$ is derivace mapping v bodě $\mathbf{x}_k$, tedy (Jacobiho) matrix rozměru $n \times n$. Další iteraci $\mathbf{x}_{k+1}$ najdeme řešením nonhomogeneous linear soustavy $\hat{\mathbf{g}}(\mathbf{x}_{k+1}) = \mathbf{0}$. Pokud is Jacobiho matrix regular, řešením je

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{g}'(\mathbf{x}_k)^{-1}\,\mathbf{g}(\mathbf{x}_k). \tag{10.12}$$

Viz obrázek:



Newtonova metoda konverguje obvykle (i když ne vždy) superlinearly, tedy velmi rychle. Její nevýhodou je, že is nutno začít poměrně přesnou aproximací $\mathbf{x}_0$ skutečného řešení, jinak algoritmus snadno diverguje.

Example 10.2. *Babylónská metoda* na výpočet druhé odmocniny numbers $a \geq 0$ is 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,

$$x_{k+1} = x_k - \frac{g(x)}{g'(x)} = x_k - \frac{x^2 - a}{2x} = x_k - \frac{1}{2}\Big(x_k - \frac{a}{x_k}\Big) = \frac{1}{2}\Big(x_k + \frac{a}{x_k}\Big). \qquad \square$$

Example 10.3. Hledejme průsečík křivek $(x-1)^2 + y^2 = 1$ a $x^4 + y^4 = 1$. Given $n = 2$ 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 (10.12) 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. $\square$

Example 10.4. Funkce $f(x) = x^2 - 1$ 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 pravděpodobně ocitne velmi daleko od kořene. □

Example 10.5. Pro funkci $f(x) = x^3 - 2x + 2$ 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. □

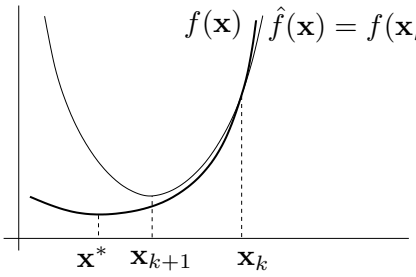
10.5.2 Použití na minimalizaci funkce

Newtonovu metodu lze použít pro hledání lokálního extrému dvakrát differentiable funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ tak, že v algoritmu (10.12) položíme $\mathbf{g}(\mathbf{x}) = f'(\mathbf{x})^T$. Tím dostaneme iteraci

$$\mathbf{x}_{k+1} = \mathbf{x}_k - f''(\mathbf{x}_k)^{-1}\, f'(\mathbf{x}_k)^T, \tag{10.13}$$

kde $f''(\mathbf{x}_k)$ is Hessova matrix funkce f v bodě $\mathbf{x}_k$.

Význam iterace (10.12) byl takový, že se mapping $\mathbf{g}$ aproximovalo Taylorovým polynomem prvního řádu (tedy affinem mappingm) a Then se našel kořen tohoto polynomu. Význam iterace (10.13) is takový, že se funkce f aproximuje Taylorovým polynomem druhého řádu (tedy kvadratickou funkcí) a Then se najde minimum této kvadratické funkce. Odvoďte podrobně, že tomu tak je!



Iteraci (10.13) lze napsat v obecnějším tvaru (10.4), kde

$$\mathbf{v}_k = -f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k)^T. \tag{10.14}$$

Výhodou tohoto zobecnění is možnost zvolit optimální (ne nutně jednotkovou) délku kroku pomocí jednorozměrné minimalizace (10.6). Algoritmu (10.13) s jednotkovou délkou kroku se Then říká **čistá** Newtonova metoda.

vectoru (10.14) říkáme **Newtonův směr**. Aby to byl sestupný směr, musí být

$$f'(\mathbf{x}_k)\, \mathbf{v}_k = -f'(\mathbf{x}_k)\, f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k)^T < 0.$$

Postačující podmínkou pro to je, aby matrix $f''(\mathbf{x}_k)$ byla pozitivně definitní (nebot' Then bude pozitivně definitní i její inverze, viz Cvičení 5.20). To znamená, že funkce f v bodě $\mathbf{x}_k$ se musí lokálně podobat 'údolí'.

10.6 Gauss-Newtonova metoda

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

$$f(\mathbf{x}) = \|\mathbf{g}(\mathbf{x})\|_2^2 = \mathbf{g}(\mathbf{x})^T\, \mathbf{g}(\mathbf{x}) = \sum_{i=1}^{m} g_i(\mathbf{x})^2, \tag{10.15}$$

kde g_i jsou složky mapping $\mathbf{g}$. Speciálním případem je přibližné řešení linear nonhomogeneous soustavy $\mathbf{A}\mathbf{x} = \mathbf{b}$, kde $\mathbf{g}(\mathbf{x}) = \mathbf{b} - \mathbf{A}\mathbf{x}$ (viz §6.1). Zde ovšem předpokládáme obecně Nonlinear mapping $\mathbf{g}$.

Všimněte si, že zatímco v §10.4 a §10.5.2 bylo cílem minimalizovat *obecnou* funkci, zde chceme minimalizovat funkci ve speciálním tvaru (10.15). Nyní we have dvě možnosti. Buď můžeme nasadit na funkci (10.15) jednu z metod pro minimalizaci obecné funkce, k čemuž se vrátíme v §10.6.1. Nebo můžeme být chytřejší a využít speciálního tvaru funkce (10.15), což uděláme teď.

Aproximujme opět mapping $\mathbf{g}$ Taylorovým polynomem prvního řádu (10.11). Úloha (10.15) Then vyžaduje minimalizovat $\|\hat{\mathbf{g}}(\mathbf{x})\|_2^2$. To is úloha linearch nejmenších čtverců, kterou již známe z §6.1. Vede na normální rovnice

$$\mathbf{g}'(\mathbf{x}_k)^T\, \mathbf{g}'(\mathbf{x}_k)\, (\mathbf{x} - \mathbf{x}_k) = -\mathbf{g}'(\mathbf{x}_k)^T\, \mathbf{g}(\mathbf{x}_k).$$

Pokud má Jacobiho matrix $\mathbf{g}'(\mathbf{x}_k)$ plnou hodnost, řešíme pomocí pseudoinverze:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \underbrace{[\mathbf{g}'(\mathbf{x}_k)^T\, \mathbf{g}'(\mathbf{x}_k)]^{-1}\, \mathbf{g}'(\mathbf{x}_k)^T}_{\mathbf{g}'(\mathbf{x}_k)^+}\, \mathbf{g}(\mathbf{x}_k) \tag{10.16}$$

Algoritmus (10.16) is znám jako **Gauss-Newtonova metoda**. Můžeme jej opět napsat obecněji ve tvaru (10.4) se směrem hledání

$$\mathbf{v}_k = -[\mathbf{g}'(\mathbf{x}_k)^T\, \mathbf{g}'(\mathbf{x}_k)]^{-1}\, \mathbf{g}'(\mathbf{x}_k)^T\, \mathbf{g}(\mathbf{x}_k). \tag{10.17}$$

Pro $m = n$ we have $\mathbf{g}'(\mathbf{x}_k)^+ = \mathbf{g}'(\mathbf{x}_k)^{-1}$, tedy Gauss-Newtonova metoda se redukuje na Newtonovu metodu (10.12) na řešení soustavy n rovnic s n neznámými.

Snadno spočítáme (viz §8.3.2) derivaci účelové funkce (10.15), is rovna $f'(\mathbf{x}) = 2\mathbf{g}(\mathbf{x})^T\mathbf{g}'(\mathbf{x})$. Z toho vidíme, že Gauss-Newtonův směr (10.17) lze psát ekvivalentně jako

$$\mathbf{v}_k = -\tfrac{1}{2}[\mathbf{g}'(\mathbf{x}_k)^T\, \mathbf{g}'(\mathbf{x}_k)]^{-1}\, f'(\mathbf{x}_k)^T. \tag{10.18}$$

Tento směr se liší od gradientního směru (10.7) pouze násobením maticí $\frac{1}{2}[\mathbf{g}'(\mathbf{x}_k)^T\, \mathbf{g}'(\mathbf{x}_k)]^{-1}$. Pokud Jacobián $\mathbf{g}'(\mathbf{x}_k)$ má plnou hodnost (tedy n), tato matrix je pozitivně definitní. Podobnou úvahou jako v §10.5.2 dostaneme, že směr (10.17) is *vždy* sestupný.

Čistá Gauss-Newtonova metoda (i.e., s jednotkovou délkou kroku) může divergovat, a to i když is počáteční odhad $\mathbf{x}_0$ libovolně blízko lokálnímu minimu funkce (10.15). Protože ale Gauss-Newtonův směr je vždy sestupný, vhodnou volbou délky kroku α_k lze vždy zajistit konvergenci.

Example 10.6. V systému GPS we have 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 Nonlinear soustavu ve smyslu nejmenších čtverců, tedy minimalizujme funkci

$$f(\mathbf{x}) = \sum_{i=1}^{m} \big(\|\mathbf{x} - \mathbf{a}_i\|_2 - y_i\big)^2.$$

we have tedy $\mathbf{g} = (g_1, \ldots, g_m)$: $\mathbb{R}^n \to \mathbb{R}^m$, kde $g_i(\mathbf{x}) = \|\mathbf{x} - \mathbf{a}_i\|_2 - y_i$. Derivace složek $\mathbf{g}$ is (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

$$\mathbf{g}'(\mathbf{x}) = \begin{bmatrix} (\mathbf{x} - \mathbf{a}_1)^T / \|\mathbf{x} - \mathbf{a}_1\|_2 \\ \vdots \\ (\mathbf{x} - \mathbf{a}_m)^T / \|\mathbf{x} - \mathbf{a}_m\|_2 \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

Then dosadíme do vzorečku (10.16). □

10.6.1 Rozdíl proti Newtonově metodě

Předpokládejme, že bychom optimalizovali naší účelovou funkci (10.15) přímo Newtonovou metodou z §10.5.2. Spočítejme (proveďte sami!) Hessián funkce (10.15):

$$f''(\mathbf{x}) = 2\mathbf{g}'(\mathbf{x})^T \mathbf{g}'(\mathbf{x}) + 2 \sum_{i=1}^{m} g_i(\mathbf{x})\, g_i''(\mathbf{x}).$$

Hessián is součtem členu obsahujícího derivace prvního rádu a členu obsahujícího derivace druhého řádu. Vidíme, že směr (10.18) se liší od Newtonova směru (10.14) zanedbáním členu druhého řádu v Hessiánu $f''(\mathbf{x}_k)$. To se projevuje tím, že Gauss-Newtonova metoda má horší lokální konvergenční chování než plná Newtonova metoda – ani v blízkém okolí řešení nemusí konvergovat superlinearly. Na druhou stranu, vyhnuli jsme se počítání druhých derivací funkce $\mathbf{g}$, což is velké zjednodušení.

10.6.2 Levenberg-Marquardtova metoda

Levenberg-Marquardtova metoda is široce používané vylepšení Gauss-Newtonovy metody, které matici $\mathbf{g}'(\mathbf{x})^T \mathbf{g}'(\mathbf{x})$ v iteraci (10.16) nahrazuje maticí

$$\mathbf{g}'(\mathbf{x}_k)^T \mathbf{g}'(\mathbf{x}_k) + \mu_k \mathbf{I} \tag{10.19}$$

pro nějaké zvolené $\mu_k > 0$. Vidíme, že:

- Pro malé μ_k se Levenberg-Marquardtova iterace blíží Gauss-Newtonově iteraci.
- Pro velké μ_k is inverze matrix (10.19) blízká $\mu_k^{-1}\mathbf{I}$, tedy Levenberg-Marquardtova iterace is blízká $\mathbf{x}_{k+1} = \mathbf{x}_k - \mu_k^{-1} f'(\mathbf{x}_k)^T$. Ale to is 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 Then 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 (10.16) se počítá inverze matrix $\mathbf{g}'(\mathbf{x}_k)^T \mathbf{g}'(\mathbf{x}_k)$. Tato matrix is sice vždy pozitivně semidefinitní, ale může být blízká singular (kdy se to stane?). To neblaze ovlivní stabilitu algoritmu. matrix (10.19) is ale vždy pozitivně definitní (viz Cvičení 5.19), a tedy regular.

10.7 Statistické odůvodnění kritéria nejmenších čtverců

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

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) + \boldsymbol{\varepsilon}, \tag{10.20}$$

kde $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)$ jsou náhodné proměnné modelující chyby měření $\mathbf{y} = (y_1, \ldots, y_m)$. Metoda nejmenších čtverců říká, že we have minimalizovat $\|\boldsymbol{\varepsilon}\|_2^2 = \sum_{i=1}^m \varepsilon_i^2$, ale neříká proč.

Důuvod odvodíme statistickou úvahou. Metoda činí dva předpoklady:

- Náhodné proměnné ε_i mají normální (neboli Gaussovo) rozdělení s nulovou střední hodnotou a směrodatnou odchylkou σ,

  $$p(\varepsilon_i) = c\,\mathrm{e}^{-\varepsilon_i^2/(2\sigma^2)},$$

  kde $c = \big(\sigma\sqrt{2\pi}\big)^{-1}$ is normalizační konstanta.
- Náhodné proměnné $\varepsilon_1, \ldots, \varepsilon_m$ jsou na sobě independent. Tedy sdružená pravděpodobnost is rovna součinu

  $$p(\boldsymbol{\varepsilon}) = p(\varepsilon_1, \ldots, \varepsilon_m) = \prod_{i=1}^m p(\varepsilon_i) = \prod_{i=1}^m c\,\mathrm{e}^{-\varepsilon_i^2/(2\sigma^2)}. \tag{10.21}$$

Dále použijeme *princip maxima věrohodnosti*. Ten říká, že parametry $\mathbf{x}$ se mají najít tak, aby $p(\boldsymbol{\varepsilon}) = p(\mathbf{y} - \mathbf{f}(\mathbf{x}))$ bylo maximální. is pohodlnější minimalizovat záporný logaritmus

$$-\log p(\varepsilon_1, \ldots, \varepsilon_m) = -\sum_{i=1}^m \log p(\varepsilon_i) = \sum_{i=1}^m \Big(\frac{\varepsilon_i^2}{2\sigma^2} - \log c\Big).$$

Jelikož σ is konstanta, is to totéž jako minimalizovat $\sum_i \varepsilon_i^2$.

10.8 Exercises

10.1. Najděte lokální extrém funkce $f(x, y) = x^2 - y + \sin(y^2 - 2x)$ čistou Newtonovou metodou. Počáteční odhad zvolte $(x_0, y_0) = (1, 1)$.

10.2. Given m bodů v rovině o souřadnicích (x_i, y_i), $i = 1, \ldots, m$. Tyto body chceme proložit kružnicí ve smyslu nejmenších čtverců – i.e., 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 11

Convexity

11.1 Konvexní set

Definition 11.1. *set $X \subseteq \mathbb{R}^n$ se nazývá* **konvexní**, *jestliže*

$$\mathbf{x} \in X,\ \mathbf{y} \in X,\ 0 \leq \alpha \leq 1 \quad \Longrightarrow \quad \alpha\mathbf{x} + (1-\alpha)\mathbf{y} \in X. \tag{11.1}$$

set $\{\, \alpha\mathbf{x} + (1-\alpha)\mathbf{y} \mid 0 \leq \alpha \leq 1 \,\}$ je úsečka spojující body $\mathbf{x}$ a $\mathbf{y}$ (zopakujte si Příklad 3.1). Definice tedy říká, že set is konvexní, jestliže s každými dvěma body obsahuje i úsečku, která is spojuje. Obrázek ukazuje příklad konvexní a nekonvexní set v $\mathbb{R}^2$:

Konvexní množinu lze definovat i abstraktněji. **Konvexní kombinace** vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ is jejich linear kombinace $\alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k$ taková, že $\alpha_1 + \cdots + \alpha_k = 1$ a $\alpha_1, \ldots, \alpha_k \geq 0$. set je konvexní právě tehdy, když is uzavřená vůči konvexním kombinacím (neboli každá konvexní kombinace vectors z set leží v množině). Lze dokázat indukcí, že tato definice is ekvivalentní Definici 11.1. Všimněte si, že $\alpha\mathbf{x} + (1-\alpha)\mathbf{y}$ pro $0 \leq \alpha \leq 1$ is konvexní kombinací dvou vectors $\mathbf{x}, \mathbf{y}$.

Konvexní obal vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ is set všech jejich konvexních kombinací. Tuto k-tici vectors můžeme vnímat jako množinu $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, konvexní obal Then značíme

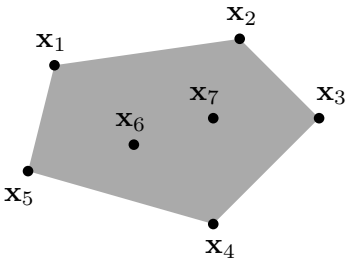
$$\operatorname{conv} X = \operatorname{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_k\} = \{\, \alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k \mid \alpha_1 + \cdots + \alpha_k = 1,\ \alpha_1, \ldots, \alpha_k \geq 0 \,\}. \tag{11.2}$$

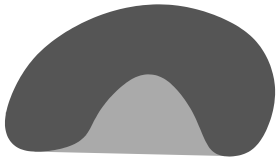
Jak se definuje konvexní obal set s *nekonečným* počtem prvků, např. pravém obrázku výše? Nelze použít definice (11.2), neboť není jasné, co znamená součet $\alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k$ pro nekonečný počet vectors (uvědomme si, že set X může být i nespočetná). Konvexní obal libovolné (konečné či nekonečné) set $X \subseteq \mathbb{R}^n$ se definuje jako set všech konvexních kombinací všech konečných podmnožin X, tedy

$$\operatorname{conv} X = \{\, \alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k \mid \mathbf{x}_1, \ldots, \mathbf{x}_k \in X,\ \alpha_1 + \cdots + \alpha_k = 1,\ \alpha_1, \ldots, \alpha_k \geq 0,\ k \in \mathbb{N} \,\}.$$

Ekvivalentně se dá konvexní obal set definovat jako nejmenší konvexní set, která množinu obsahuje (přesněji, jako průnik všech konvexních množin, které množinu obsahují).

Obrázek ukazuje konvexní obal konečné (vlevo) a nekonečné (vpravo) set pro $n = 2$:





11.1.1 Čtyři kombinace a čtyři obaly

Konvexní kombinace is linear kombinace, jejíž coefficienty splňují omezení $\alpha_1 + \cdots + \alpha_k = 1$ a $\alpha_1, \ldots, \alpha_k \geq 0$. Všimněte si, že když vynecháme druhé omezení, dostaneme affine kombinaci (viz §3.3). Podle toho, které ze dvou omezení vyžadujeme, dostaneme čtyři druhy kombinací. Udělejme si v nich nyní pořádek.

Vážený součet $\alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k$ vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n$ se nazývá jejich

linear kombinace,	jestliže	$\alpha_1, \ldots, \alpha_k \in \mathbb{R}$.		
affine 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	$\alpha_1, \ldots, \alpha_k \in \mathbb{R}$,	$\alpha_1 + \cdots + \alpha_k = 1$,	$\alpha_1, \ldots, \alpha_k \geq 0$.

set, která is uzavřená vůči

linearm kombinacím,	se nazývá	**linear subspace**.
affinem kombinacím,	se nazývá	**affine subspace**.
nezáporným kombinacím,	se nazývá	**konvexní kužel**.
konvexním kombinacím,	se nazývá	**konvexní set**.

K tomu, co již znáte, přibyl pojem nezáporné kombinace a konvexního kuželu.

linear [affine, nezáporný, konvexní] **obal** vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ is set všech jejich linearch [affinech, nezáporných, konvexních] kombinací. Obecněji, obal (konečné či nekonečné) set $X \subseteq \mathbb{R}^n$ is set kombinací všech konečných podmnožin X. Ekvivalentně, linear [affine, nezáporný, konvexní] obal set $X \subseteq \mathbb{R}^n$ is nejmenší linear subspace [affine subspace, konvexní kužel, konvexní set] obsahující množinu X.

Jako cvičení si nakreslete linear, affine, nezáporný a konvexní obal náhodně zvolených k vectors v $\mathbb{R}^n$ pro devět případů $k, n \in \{1, 2, 3\}$.

11.1.2 Operace zachovávající konvexitu množin

Jaké operace s konvexními setmi mají za výsledek opět konvexní množinu? Zdaleka nejdůležitější taková operace is průnik. Následující větu is snadné dokázat.

Theorem 11.1. *Průnik (konečně či nekonečně mnoha) konvexních množin is konvexní set.*

Sjednocení konvexních množin ale *nemusí* být konvexní set.

11.2 Konvexní funkce

Definition 11.2. *Funkce* $f\colon \mathbb{R}^n \to \mathbb{R}$ *is* **konvexní** *na množině* $X \subseteq \mathbb{R}^n$, *jestliže set* X *is konvexní a platí*

$$\mathbf{x} \in X,\ \mathbf{y} \in X,\ 0 \le \alpha \le 1 \quad \Longrightarrow \quad f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \le \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}). \tag{11.3}$$

Funkce f *is* **konkávní** *na množině* X, *jestliže is funkce* $-f$ *konvexní na množině* X.

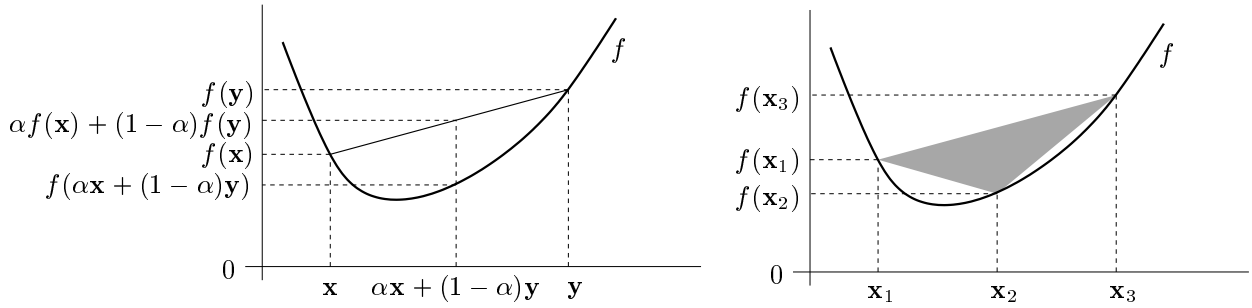
Rozlišujte pojem *konvexní set* a *konvexní funkce*, jde o různé věci. Dále si všimněte, že X musí být konvexní set – pojem konvexní funkce na nekonvexní množině nemá smysl. Pokud $X = \mathbb{R}^n$, odkaz na X můžeme vynechat a říkáme pouze, že funkce f is konvexní.

Podmínku (11.3) lze zobecnit pro více než dva body: funkce f is konvexní právě tehdy, když

$$\mathbf{x}_1, \ldots, \mathbf{x}_k \in X,\ \ \alpha_1, \ldots, \alpha_k \ge 0,\ \ \alpha_1 + \cdots + \alpha_k = 1 \quad \Longrightarrow \\ f(\alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k) \le \alpha_1 f(\mathbf{x}_1) + \cdots + \alpha_k f(\mathbf{x}_k), \tag{11.4}$$

neboli 'funkční hodnota konvexní kombinace není větší než konvexní kombinace funkčních hodnot'. Podmínka (11.4) zjevně implikuje podmínku (11.3) a indukcí lze dokázat, že to platí i naopak. Podmínce (11.4) se někdy říká **Jensenova nerovnost**. Porovnejte ji s definicí linearho mapping (3.2)!

Geometrický význam podmínky (11.3) is 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 (11.4) is ten, že konvexní polyedr vybarvený šedě (viz pravý obrázek) leží nad grafem funkce. Podrobně rozmyslete, jak tyto geometrické interpretace odpovídají výrazům (11.3) a (11.4)!



Důkaz konvexity funkce z Definice 11.2 vyžaduje někdy kreativitu, neexistuje na to mechanický postup. Naopak, chceme-li dokázat, že funkce *není* konvexní, stačí nám (samozřejmě!) jediný protipříklad – jeho nalezení však také může vyžadovat intuici.

Example 11.1. Dokažme z Definice 11.2, že funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ daná 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 (11.3), neboť

$$f((\mathbf{x}+\mathbf{y})/2) = f(1,1) = 1 > (f(\mathbf{x}) + f(\mathbf{y}))/2 = (0+0)/2 = 0. \qquad \square$$

Poznamenejme, že použitím Jensenovy nerovnosti na vhodnou konvexní funkci lze dokázat mnoho známých nerovností.

Example 11.2. Funkce log is konkávní na $\mathbb{R}_{++}$. Napišme pro tuto funkci Jensenovu nerovnost (11.4) (jelikož funkce is konkávní a ne konvexní, musíme v Jensenově nerovnosti obrátit znaménko nerovnosti), ve které položíme $\alpha_1 = \cdots = \alpha_n = \frac{1}{n}$:

$$\log \frac{x_1 + \cdots + x_n}{n} \geq \frac{\log x_1 + \cdots + \log x_n}{n}$$

kde $x_1, \ldots, x_n$ jsou kladné. Vezmeme-li exponenciálu každé strany, dostaneme

$$\frac{x_1 + \cdots + x_n}{n} \geq (x_1 \cdots x_n)^{1/n}$$

Tato známá nerovnost říká, že aritmetický průměr není nikdy menší než geometrický. □

Example 11.3. Uveďme často potkávané jednoduché konvexní či konkávní funkce:

1. Exponenciála $f(x) = \mathrm{e}^{ax}$ is konvexní na $\mathbb{R}$, pro libovolné $a \in \mathbb{R}$.
2. Mocnina $f(x) = x^a$ is na $\mathbb{R}_{++}$ konvexní pro $a \geq 1$ nebo $a \leq 0$ a konkávní pro $0 \leq a \leq 1$.
3. Mocnina absolutní hodnoty $f(x) = |x|^a$ is pro $a \geq 1$ konvexní na $\mathbb{R}$ (speciálně: absolutní hodnota $|x|$ is konvexní).
4. Logaritmus $f(x) = \log x$ is konkávní na $\mathbb{R}_{++}$.
5. Záporná entropie $f(x) = x \log x$ is 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. affine funkce $f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ is zároveň konvexní i konkávní.
7. Kvadratická forma $f(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x}$ is konvexní pro $\mathbf{A}$ pozitivně semidefinitní, konkávní pro $\mathbf{A}$ negativně semidefinitní, a nekonvexní a nekonkávní pro $\mathbf{A}$ indefinitní (viz Příklad 11.4).
8. Maximum složek $f(\mathbf{x}) = \max_{i=1}^n x_i = \max\{x_1, \ldots, x_n\}$ is 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

   $$f_a(\mathbf{x}) = f(a\mathbf{x})/a = \log(\mathrm{e}^{ax_1} + \cdots + \mathrm{e}^{ax_n})/a$$

   se pro $a \to +\infty$ blíží funkci $\max_{i=1}^n x_i$ (dokažte výpočtem limity!).
10. Geometrický průměr $f(\mathbf{x}) = (x_1 \cdots x_n)^{1/n}$ is konkávní na $\mathbb{R}^n_+$.

Nakreslete či představte si vrstevnice a grafy těchto funkcí! □

11.2.1 vectorové normy

Norma formalizuje pojem 'délky' vectoru $\mathbf{x}$.

Definition 11.3. *Funkce* $\|\cdot\|: \mathbb{R}^n \to \mathbb{R}$ *se nazývá* **norma**, *jestliže splňuje tyto axiomy:*

1. *Jestliže* $\|\mathbf{x}\| = 0$ *Then* $\mathbf{x} = \mathbf{0}$.
2. $\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$ *pro každé* $\alpha \in \mathbb{R}$ *a* $\mathbf{x} \in \mathbb{R}^n$ *(norma is kladně homogeneous).*
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 $\alpha = 0$

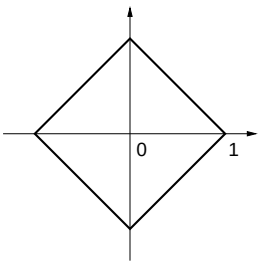
- $\|\mathbf{x}\| \geq 0$ pro každé $\mathbf{x} \in \mathbb{R}^n$. To jde odvodit tak, že v trojúhelníkové nerovnosti položíme $\mathbf{y} = -\mathbf{x}$, což dá
$$\|\mathbf{x} - \mathbf{x}\| = \|\mathbf{0}\| = 0 \leq \|\mathbf{x}\| + \|-\mathbf{x}\| = 2\|\mathbf{x}\|,$$
kde na pravé straně jsme použili homogenitu.
- Norma is konvexní funkce, neboť pro každé $0 \leq \alpha \leq 1$ máme
$$\|\alpha\mathbf{x} + (1-\alpha)\mathbf{y}\| \leq \|\alpha\mathbf{x}\| + \|(1-\alpha)\mathbf{y}\| = \alpha\|\mathbf{x}\| + (1-\alpha)\|\mathbf{y}\|,$$
kde nerovnost plyne z trojúhelníkové nerovnosti a rovnost z homogenity.

Jednotková sféra normy is set $\{\,\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\| = 1\,\}$ všech vectors s jednotkovou normou. Díky homogenitě is jednotková sféra středově symetrická a její tvar zcela určuje normu.

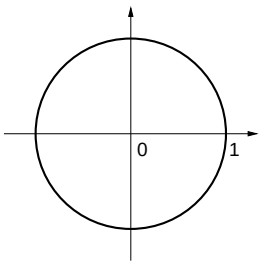
Uveďme příklady norem. Základním příkladem is **p-norma**
$$\|\mathbf{x}\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}.$$
Musí být $p \geq 1$, jinak neplatí trojúhelníková nerovnost. Nejčastěji narazíte na:

- $\|\mathbf{x}\|_1 = |x_1| + \cdots + |x_n|$. Někdy se jí říká *manhattanská norma*, protože v systému pravoúhlých ulic is vzdálenost mezi body $\mathbf{x}$ a $\mathbf{y}$ rovna $\|\mathbf{x} - \mathbf{y}\|_1$.
- $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2} = \sqrt{\mathbf{x}^T\mathbf{x}}$. is 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*.

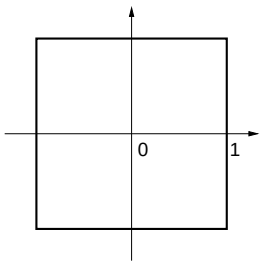
Jednotkové sféry těchto norem v $\mathbb{R}^2$ vypadají takto:



$\|\mathbf{x}\|_1 = 1$



$\|\mathbf{x}\|_2 = 1$



$\|\mathbf{x}\|_\infty = 1$

Existují ale i normy, které nejsou p-normy, např.

- $\|\mathbf{x}\| = 2|x_1| + \sqrt{x_2^2 + x_3^2} + \max\{|x_4|, |x_5|\}$ is norma na $\mathbb{R}^5$.
- Je-li $\|\mathbf{x}\|$ norma a $\mathbf{A}$ is square nebo úzká matrix s plnou hodností, is také $\|\mathbf{A}\mathbf{x}\|$ norma.

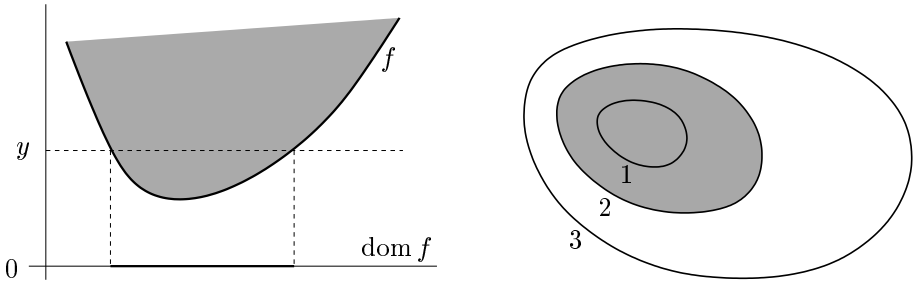
11.2.2 Epigraf a subkontura

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 is set $\{\,\mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) \leq y\,\}$.
- **Epigraf** funkce is set $\{\,(\mathbf{x}, y) \in \mathbb{R}^{n+1} \mid \mathbf{x} \in \mathbb{R}^n,\ f(\mathbf{x}) \leq y\,\}$.

Levý obrázek znázorňuje subkonturu výšky y a epigraf funkce $\mathbb{R} \to \mathbb{R}$, pravý obrázek subkonturu výšky 2 funkce $\mathbb{R}^2 \to \mathbb{R}$:

[1] Slovo 'subkontura' is pokus o český překlad anglického 'sublevel set'.



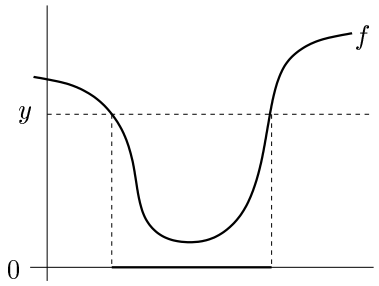
Theorem 11.2. *Je-li f konvexní funkce, Then is každá subkontura této funkce konvexní set.*

Proof. Předpokládejme, že body $\mathbf{x}_1$ a $\mathbf{x}_2$ patří do subkontury, tedy $f(\mathbf{x}_1) \leq y$ a $f(\mathbf{x}_2) \leq y$. Pro každé $0 \leq \alpha \leq 1$ platí

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha)f(\mathbf{x}_2) \leq \alpha y + (1-\alpha)y = y,$$

kde první nerovnost plyne z konvexity funkce a druhou nerovnost dostaneme sečtením nerovnice $f(\mathbf{x}_1) \leq y$ vynásobené α a nerovnice $f(\mathbf{x}_2) \leq y$ vynásobené $1-\alpha$. Tedy bod $\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2$ patří do subkontury, která is proto konvexní set. □

Obrácená implikace ve Větě 11.2 neplatí: snadno najdeme funkci, která není konvexní a jejíž každá subkontura is konvexní set[2]. Příklad is na obrázku:



Theorem 11.3. *Funkce f is konvexní právě tehdy, když její epigraf je konvexní set.*

Proof. Předpokládejme, že funkce f is 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 set.

Předpokládejme, že epigraf is konvexní set. Tedy pokud body $(\mathbf{x}_1, y_1)$ a $(\mathbf{x}_2, y_2)$ patří do epigrafu, Then 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 is funkce f konvexní. □

[2] Funkce, jejíž každá subkontura is konvexní set, se nazývá *kvazikonvexní* ('quasi' znamená latinsky 'jako když', 'skoro'). Kvazikonvexní funkce nejsou zdaleka tak hezké jako konvexní funkce.

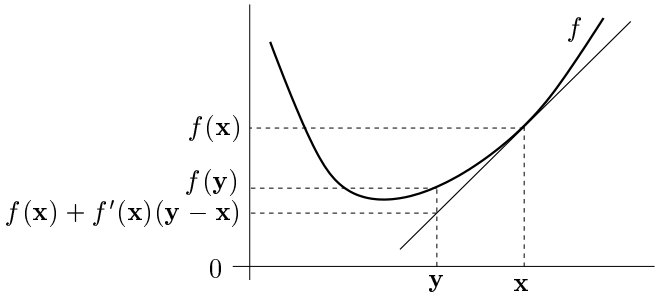
11.2.3 convexity diferencovatelných funkcí

Konvexní funkce nemusí být v každém bodě diferencovatelná (uvažte např. funkci $f(x) = |x|$). Pokud is ale funkce jednou či dvakrát diferencovatelná, její konvexitu lze snadněji než pomocí Definice 11.2 charakterizovat pomocí derivací.

Theorem 11.4. *Let funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ is diferencovatelná na konvexní množině $X \subseteq \operatorname{dom} f$. Funkce f is 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 tečna ke grafu funkce v každém bodě $\mathbf{x} \in X$ leží celá (i.e., pro každé $\mathbf{y}$) pod grafem (promyslete a porovnejte s definicí derivace v §8.3!):



Theorem 11.5. *Let funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ is dvakrát diferencovatelná na konvexní množině $X \subseteq \mathbb{R}^n$. Funkce f is konvexní na množině X právě tehdy, když v každém bodě $\mathbf{x} \in X$ is Hessova matrix $f''(\mathbf{x})$ pozitivně semidefinitní.*

Example 11.4. Let $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$, kde $\mathbf{A}$ is symetrická pozitivně semidefinitní. Ukažme konvexitu této funkce třemi způsoby:

- Dokažme konvexitu z Věty 11.5. To je triviální, protože Hessián is $f''(\mathbf{x}) = 2\mathbf{A}$ a tedy is pozitivně semidefinitní.
- Dokažme konvexitu z Věty 11.4. Protože $f'(\mathbf{x}) = 2\mathbf{x}^T\mathbf{A}$, we have 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 $\mathbf{x}^T\mathbf{A}\mathbf{x} - 2\mathbf{x}^T\mathbf{A}\mathbf{y} + \mathbf{y}^T\mathbf{A}\mathbf{y} \geq 0$. Ale zjevně platí[3]

  $$\mathbf{x}^T\mathbf{A}\mathbf{x} - 2\mathbf{x}^T\mathbf{A}\mathbf{y} + \mathbf{y}^T\mathbf{A}\mathbf{y} = (\mathbf{x} - \mathbf{y})^T\mathbf{A}(\mathbf{x} - \mathbf{y}), \tag{11.5}$$

  což is nezáporné pro každé $\mathbf{x}, \mathbf{y}$, protože $\mathbf{A}$ is pozitivně semidefinitní.
- Dokážme konvexitu z Definice 11.2. Musíme dokázat, že pro každé $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ a $0 \leq \alpha \leq 1$ platí (11.3), 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 $\alpha(1-\alpha)$ is pro každé $0 \leq \alpha \leq 1$ nezáporný. Nezápornost výrazu (11.5) jsme již ukázali. □

[3] Všimněte si, že pro $n = 1$ a $\mathbf{A} = 1$ se rovnost (11.5) zjednoduší na známé $x^2 - 2xy + y^2 = (x-y)^2$.

11.2.4 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 11.2.

Jsou-li $g_1, \ldots, g_k \colon \mathbb{R}^n \to \mathbb{R}$ konvexní funkce a $\alpha_1, \ldots, \alpha_k \geq 0$, is snadné dokázat z Definice 11.2 (proveďte!), že také funkce

$$f = \alpha_1 g_1 + \cdots + \alpha_k g_k$$

je konvexní. Speciálně, jsou-li f a g konvexní funkce, Then $f + g$ je konvexní.

Zkoumejme nyní složenou funkci $f(\mathbf{x}) = (g \circ \mathbf{h})(\mathbf{x}) = g(\mathbf{h}(\mathbf{x}))$, kde $\mathbb{R}^n \xrightarrow{\mathbf{h}} \mathbb{R}^m \xrightarrow{g} \mathbb{R}$. Obecně *neplatí* ani v případě $m = n = 1$, že convexity funkcí g a h zaručuje konvexitu funkce f. Nutné a postačující podmínky pro konvexitu složené funkce jsou obecně dosti komplikované a nebudeme is uvádět. Uvedeme jen nejdůležitější případ, kdy $\mathbf{h}$ is affine mapping.

Theorem 11.6. *Let funkce $g \colon \mathbb{R}^m \to \mathbb{R}$ is konvexní. Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ a $\mathbf{b} \in \mathbb{R}^m$. Then funkce $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x} + \mathbf{b})$ je konvexní.*

Proof. Pro každé $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ a $0 \leq \alpha \leq 1$ platí

$$\begin{aligned} f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) &= g(\mathbf{A}[\alpha\mathbf{x} + (1-\alpha)\mathbf{y}] + \mathbf{b}) \\ &= g(\alpha(\mathbf{A}\mathbf{x} + \mathbf{b}) + (1-\alpha)(\mathbf{A}\mathbf{y} + \mathbf{b})) \\ &\leq \alpha g(\mathbf{A}\mathbf{x} + \mathbf{b}) + (1-\alpha) g(\mathbf{A}\mathbf{y} + \mathbf{b}) \\ &= \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}). \end{aligned}$$

□

Example 11.5. Let funkce $f \colon \mathbb{R}^{2n} \to \mathbb{R}$ is dána předpisem $f(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$, kde argument funkce f is vector $(\mathbf{x}, \mathbf{y}) = (x_1, \ldots, x_n, y_1, \ldots, y_n) \in \mathbb{R}^{2n}$. Ukažme, že funkce f is konvexní. Položme $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ a $\mathbf{A} = \begin{bmatrix} \mathbf{I} & -\mathbf{I} \end{bmatrix}$. Máme

$$\mathbf{x} - \mathbf{y} = \begin{bmatrix} \mathbf{I} & -\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \mathbf{A}\mathbf{z}.$$

Tedy $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{z}) = \|\mathbf{A}\mathbf{z}\|$. Jelikož norma is konvexní, plyne convexity f z Věty 11.6. □

Zdaleka nejzajímavější operace zachovávající konvexitu funkcí is ovšem maximum.

Theorem 11.7. *Necht' I is libovolná set a $g_i \colon \mathbb{R}^n \to \mathbb{R}$, $i \in I$, jsou konvexní funkce. Then funkce*

$$f(\mathbf{x}) = \max_{i \in I} g_i(\mathbf{x}) \tag{11.6}$$

je konvexní, kde předpokládáme, že pro každé $\mathbf{x}$ maximum existuje[4].

Proof. Podle Věty 11.3 is funkce konvexní právě tehdy, když její epigraf is konvexní set. is jasné, že epigraf funkce (11.6) is průnik epigrafů funkcí g_i. Ale průnik (konečného i nekonečného počtu) konvexních množin is konvexní set. Tedy epigraf funkce (11.6) is konvexní set. □

[4] Pokud pro nějaké $\mathbf{x}$ set $\{\, g_i(\mathbf{x}) \mid i \in I \,\}$ nemá největší prvek (což se může stát jen tehdy, je-li set I nekonečná), můžeme maximum v (11.6) nahradit supremem a věta stále platí.

Velká obecnost věty plyne z toho, že indexová set I může být konečná i nekonečná (a to spočetná i nespočetná). Uveďme nejprve příklady pro konečnou množinu I.

Example 11.6. Funkce

$$f(\mathbf{x}) = \max_{i=1}^{k}(\mathbf{a}_i^T\mathbf{x} + b_i)$$

je maximem affinech funkcí. Protože affine funkce jsou konvexní, is i jejich maximum konvexní. Tuto funkci jsme již potkali ve vzorci (12.3). □

Example 11.7. Let $f(\mathbf{x}) = \max_{i=1}^{n} x_i$ is maximum ze složek $\mathbf{x}$. Konvexitu této funkce lze poměrně snadno dokázat z Definice 11.2, nicméně dokažme ji z Věty 11.7. Given $g_i(\mathbf{x}) = x_i$. Funkce g_i jsou linear, tedy konvexní. Tedy funkce $f(\mathbf{x}) = \max_{i=1}^{n} g_i(\mathbf{x})$ je konvexní. □

Dále uveďme příklady pro nekonečnou množinu I.

Example 11.8. Let $C \subseteq \mathbb{R}^n$ is libovolná (ne nutně konvexní) set. Funkce

$$f(\mathbf{x}) = \max_{\mathbf{y}\in C} \|\mathbf{x} - \mathbf{y}\|$$

udává vzdálenost bodu $\mathbf{x}$ od nejvzdálenějšího bodu set C (zde předpokládáme, že maximum existuje). Pro každé pevné $\mathbf{y}$ je $\|\mathbf{x} - \mathbf{y}\|$ konvexní funkcí vectoru $\mathbf{x}$. Tedy výraz $\|\mathbf{x} - \mathbf{y}\|$ lze chápat jako množinu konvexních funkcí $\mathbf{x}$ indexovaných indexem $\mathbf{y}$ – pro zdůraznění této skutečnosti můžeme psát $\|\mathbf{x} - \mathbf{y}\| = g_{\mathbf{y}}(\mathbf{x})$. Jelikož f is maximem těchto funkcí, je i funkce f konvexní. □

Example 11.9. consider 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 linearho programu na vectoru $\mathbf{c}$. we have $f(\mathbf{c}) = \max_{\mathbf{x}\in X} \mathbf{c}^T\mathbf{x}$ 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 set X is neprázdná a omezená). Je-li $\mathbf{x}$ pevné, is $\mathbf{c}^T\mathbf{x}$ linear funkce vectoru $\mathbf{c}$. Funkce f is tedy maximum nekonečného množství linearch funkcí, tedy is konvexní. □

Example 11.10. Let $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$, $b_1, \ldots, b_n \in \mathbb{R}$ a $\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$ is vector 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

$$f(\mathbf{w}) = \min_{\mathbf{x}\in\mathbb{R}^m} \sum_{i=1}^{n} w_i(\mathbf{a}_i^T\mathbf{x} - b_i),$$

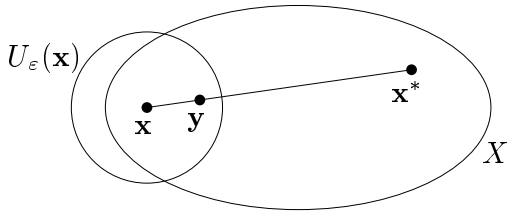
kde jsme označili hodnotu výsledného minima jako funkci vectoru vah. Funkce f is konkávní, protože is minimem linearch funkcí. □

11.3 Minima konvexní funkce na konvexní množině

Pro optimalizaci is klíčové, že každé lokální minimum konvexní funkce na konvexní množině is nutně globální. Proto jsou konvexní funkce a konvexní set v optimalizaci tak důležité.

Theorem 11.8. *Let funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ is konvexní na konvexní množině $X \subseteq \mathbb{R}^n$. Then každé lokální minimum funkce f na množině X is zároveň globální minimum.*

Proof. Let $\mathbf{x}$ is lokálním minimem f na X, viz obrázek:



Dle Definice 9.1 tedy existuje $\varepsilon > 0$ tak, že $f(\mathbf{x}) \le f(\mathbf{y})$ pro všechna $\mathbf{y} \in U_\varepsilon(\mathbf{x}) \cap X$. Necht' 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 set 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 $f(\mathbf{y}) < f(\mathbf{x})$ is ve sporu s předpokladem, že $\mathbf{x}$ je lokální minimum. □

11.3.1 Konvexní optimalizační úlohy

Zopakujme obecnou úlohu spojité optimisation (1.4) (viz §1.3)

$$\begin{aligned} \min \quad & f(x_1, \dots, x_n) \\ \text{za podmínek} \quad & g_i(x_1, \dots, x_n) \le 0, \quad i = 1, \dots, m \\ & h_i(x_1, \dots, x_n) = 0, \quad i = 1, \dots, \ell \end{aligned} \tag{11.7}$$

kde $f\colon \mathbb{R}^n \to \mathbb{R}$, $(g_1, \dots, g_m) = \mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^m$, $(h_1, \dots, h_\ell) = \mathbf{h}\colon \mathbb{R}^n \to \mathbb{R}^\ell$.

Definition 11.4. Konvexní optimalizační úloha *je úloha (11.7), kde funkce $f, g_1, \dots, g_m$ jsou konvexní a funkce $h_1, \dots, h_\ell$ jsou affine (tedy mapping $\mathbf{h}$ is affine).*

set přípustných řešení konvexní úlohy is konvexní. Můžeme ji totiž psát jako

$$\begin{aligned} X &= \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{g}(\mathbf{x}) \le \mathbf{0},\ \mathbf{h}(\mathbf{x}) = \mathbf{0} \,\} \\ &= \{\, \mathbf{x} \in \mathbb{R}^n \mid g_1(\mathbf{x}) \le 0 \,\} \cap \dots \cap \{\, \mathbf{x} \in \mathbb{R}^n \mid g_m(\mathbf{x}) \le 0 \,\} \cap \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{h}(\mathbf{x}) = \mathbf{0} \,\}. \end{aligned}$$

Zde každá set $\{\, \mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) \le 0 \,\}$ je konvexní, neboť is to subkontura konvexní funkce g_i (Věta 11.2). set $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{h}(\mathbf{x}) = \mathbf{0} \,\}$ is affine subspace, tedy také konvexní. set X is průnik konvexních množin, tedy konvexní (Věta 11.1).

Mohli bychom si myslet, že is přirozenější definovat konvexní optimalizační úlohu jednoduše jako minimalizaci konvexní funkce na konvexní množině. Tato definice is obecnější, protože set přípustných řešení X může být konvexní i tehdy, když funkce g_i nejsou konvexní nebo funkce h_i nejsou affine. Výhoda Definice 11.4 is v tom, že zatímco konvexitu set X nemusí být snadné dokázat, obvykle is snadno vidět, zda jsou funkce g_i konvexní a funkce h_i affine.

Example 11.11. Uvažujme dvě ekvivalentní definice téže set

$$X = \{\, \mathbf{x} \in \mathbb{R}^2 \mid x_1/(1+x_2^2) \leq 0,\ (x_1+x_2)^2 = 0 \,\} = \{\, \mathbf{x} \in \mathbb{R}^2 \mid x_1 \leq 0,\ x_1 + x_2 = 0 \,\}.$$

Oba tvary jsou ekvivalentní (proč?). V prvním tvaru funkce $g(\mathbf{x}) = x_1/(1+x_2^2)$ není konvexní (dokažte z Definice 11.2!) a funkce $h(\mathbf{x}) = (x_1+x_2)^2$ není affine. Přesto is set X konvexní, což is vidět ze druhého tvaru. □

Je obvykle relativně snadné najít nějaké lokální minimum optimalizační úlohy (at' konvexní či nekonvexní). Pro úlohy bez omezení s diferencovatelnou účelovou funkcí to lze udělat např. gradientní metodou (viz §10.4). Numerické algoritmy pro úlohy s omezeními existují, ale neuváděli jsme je, protože jsou poměrně složité. Konvexní optimalizační úlohy se těší výsadě dané Větou 11.8, totiž že každé lokální minimum is zároveň globální. Pokud is úloha nekonvexní, obvykle (avšak ne vždy, viz Příklad 15.5) má mnoho lokálních minim a kvůli tomu is těžké najít globální optimum. Nekonvexní úloha s větším množstvím proměnných je tedy velmi často prakticky neřešitelná.

Zformuluje-li tedy inženýr svůj problém jako optimalizační úlohu, musí si ihned položit otázku, zda is možné ji formulovat jako konvexní optimalizační úlohu. Na to neexistuje žádný mechanický postup – pomůže pouze intuice získaná zkušeností. Jelikož nekonvexních úloh je v jistém smyslu 'mnohem' více než konvexních, většina praktických problémů is nekonvexních. Avšak pro překvapivě mnoho užitečných problémů se dá konvexní tvar najít.

11.4 Exercises

11.1. Dokažte z definice konvexní set, že následucící set jsou konvexní:

a) interval $[a,b] \subseteq \mathbb{R}$
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 $\mathbf{A}$ is pozitivně semidefinitní

11.2. Které z následujících množin jsou konvexní? Nemusíte dokazovat z definice, stačí uvést přesvědčivý argument. Pokud to jde, zkuste množinu načrtnout v prostoru malé dimenze.

a) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \sum_{i=1}^n x_i = 1 \,\}$ (nadrovina, konvexní)
b) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \sum_{i=1}^n x_i \geq 1 \,\}$ (poloprostor, konvexní)
c) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} \geq \mathbf{0},\ \sum_{i=1}^n x_i = 1 \,\}$ (průnik poloprostorů a nadroviny, tedy konvexní polyedr)
d) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} \geq \mathbf{0},\ \sum_{i=1}^n x_i \leq 1 \,\}$ (průnik poloprostorů, konvexní)
e) $\mathbb{R}^n \setminus \{\, \mathbf{x} \in \mathbb{R}^n \mid \sum_{i=1}^n x_i \geq 1 \,\}$ (complement uzavřeného poloprostoru, tedy otevřený poloprostor, konvexní)
f) $\mathbb{R}^n \setminus \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} \geq \mathbf{0},\ \sum_{i=1}^n x_i \leq 1 \,\}$ (complement simplexu, není konvexní)
g) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 = 1 \,\}$ (sféra, není konvexní)
h) $\{\, \mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 < 1 \,\}$ (koule bez hranice, konvexní)
i) $\{\, (x,y) \in \mathbb{R}^2 \mid x \geq 0,\ y \geq 0,\ xy = 1 \,\}$ (graf jedné větve hyperboly, není konvexní)
j) $\{\, (x,y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 2 \,\} \cap \{\, (x,y) \in \mathbb{R}^2 \mid (x-1)^2 + y^2 \leq 2 \,\}$ (průnik dvou koulí, konvexní)

11.3. Pro každou funkci dokažte z Definice 11.2, které z těchto čtyřech tvrzení platí: funkce is 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) $f(\mathbf{x}) = \max_{i=1}^n x_i$
d) $f(\mathbf{x}) =$ aritmetický průměr čísel $x_1, \ldots, x_n$

11.4. Pro každou funkci zjistěte, které z těchto čtyřech tvrzení platí: funkce is konvexní, konkávní, konvexní i konkávní, ani konvexní ani konkávní. Můžete to udělat buď z Definice 11.2, 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) $f(\mathbf{b}) = \min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{A}\mathbf{x} \geq \mathbf{b} \,\}$ na množině všech vectors $\mathbf{b}$, pro které je polyedr $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \geq \mathbf{b} \,\}$ neprázdný a omezený ($\mathbf{A} \in \mathbb{R}^{m\times n}$ a $\mathbf{c} \in \mathbb{R}^n$ jsou dány). Nápověda: Použijte LP dualitu.
i) $f(\mathbf{x}) = \max_{i=1}^n |x_i|$
j) $f(\mathbf{x}) = \min_{i=1}^n |x_i|$
k) $f(\mathbf{x}) = \max_{i=1}^n x_i + \min_{i=1}^n x_i$
l) $f(\mathbf{x}) = \max_{i=1}^n x_i - \min_{i=1}^n x_i$
m) ($\star$) $f(\mathbf{x}) = \mathrm{median}_{i=1}^n x_i$ (medián čísel $x_1, \ldots, x_n$)
n) ($\star$) $f(\mathbf{x}) =$ součet k největších čísel $x_1, \ldots, x_n$ (kde $k \leq n$ is dáno)

11.5. Robustní prokládání přímky set bodů $(\mathbf{x}_i, y_i) \in (\mathbb{R}^n \times \mathbb{R})$, $i = 1, \ldots, m$ vyžaduje minimalizaci kritéria

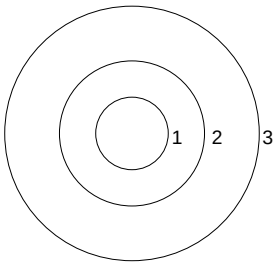
$$f(\mathbf{a}, b) = \sum_{i=1}^m \max\{-\mathbf{a}^T\mathbf{x}_i + b + y_i - \varepsilon,\ 0\ ,\ \mathbf{a}^T\mathbf{x}_i + b - y_i - \varepsilon\},$$

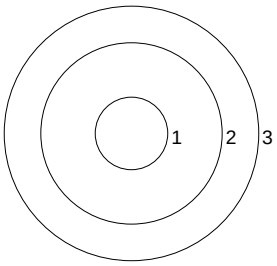
kde $\mathbf{a} \in \mathbb{R}^n$ a $b \in \mathbb{R}$. Dokažte, že $f(\mathbf{a}, b)$ je konvexní funkce.

11.6. is dána funkce $f(x) = -\cos x$ a set $X = [-\pi, +\pi]$ (kde $[\cdot]$ značí uzavřený interval). Zakroužkujte pravdivá tvrzení (může jich být i více):

a) Funkce f is na množině X konvexní.
b) Funkce f is na množině X konkávní.
c) Funkce f není na množině X ani konvexní ani konkávní.

11.7. Každý z obrázků zobrazuje vrstevnice funkce dvou proměnných a jejich výšky. Mohou být funkce konvexní? Dokažte z Definice 11.2. (Odpověď: ne, ano)





11.8. Významnou vlastností konvexních funkcí is to, že každé lokální minimum funkce is zároveň globální (Věta 11.8). Ne každá funkce s touto vlastností is ovšem konvexní. Člověk by si mohl myslet, že součet dvou funkcí (ne nutně konvexních) s touto vlastností bude mít tuto vlastnost také. Dokažte nebo najděte protipříklad.

11.9. Dokažte, že set optimálních řešení konvexní optimalizační úlohy is konvexní.

11.10. consider 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) $f(x,y) = x^2+9y^2$

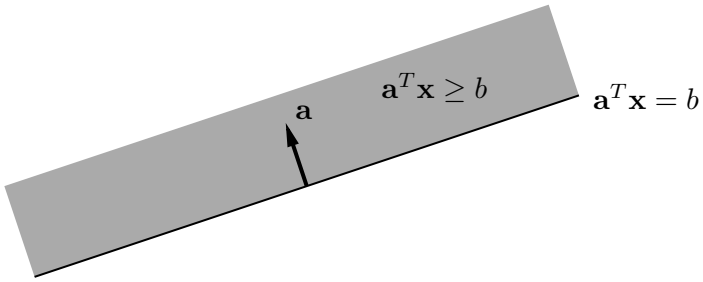
V kterých případech se jedná o konvexní optimalizační úlohu?

Chapter 12

Linear Programming

12.1 Konvexní polyedry

Poloprostor is set $\{\,\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} \geq b\,\}$ pro nějaké $\mathbf{a} \in \mathbb{R}^n$, $b \in \mathbb{R}$. Jeho hranice je nadrovina $\{\,\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} = b\,\}$. Obrázek znázorňuje tyto pojmy pro $n = 2$:



Poloprostor is očividně konvexní set.

Definition 12.1. Konvexní polyedr *is průnik konečně mnoha poloprostorů.*

Konvexní polyedr is tedy set

$$X = \{\,\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}_i^T\mathbf{x} \geq b_i,\ i = 1, \ldots, m\,\} = \{\,\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \geq \mathbf{b}\,\},$$

kde $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$ jsou rows matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ a $b_1, \ldots, b_m \in \mathbb{R}$ jsou složky vectoru $\mathbf{b} \in \mathbb{R}^m$. Je jasné, že definice dovoluje i omezení typu $\mathbf{a}_i^T\mathbf{x} \leq b_i$, protože to is ekvivalentní $-\mathbf{a}_i^T\mathbf{x} \geq -b_i$. Dovoluje i omezení typu rovnosti $\mathbf{a}_i^T\mathbf{x} = b_i$, které is ekvivalentní $\mathbf{a}_i^T\mathbf{x} \leq b_i$, $\mathbf{a}_i^T\mathbf{x} \geq b_i$.

Jelikož poloprostor is konvexní set, plyne convexity konvexního polyedru z Věty 11.1. Všimněte si, že konvexní polyedr nemusí být omezený.

Example 12.1. Příklady konvexních polyedrů v $\mathbb{R}^n$:

- každý affine subspace (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 $\{\,\mathbf{x} \in \mathbb{R}^n \mid -1 \leq x_i \leq 1,\ i = 1, \ldots, n\,\}$
- 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 $\{\,\mathbf{x} \in \mathbb{R}^n \mid x_i \geq 0,\ \sum_{i=1}^n x_i = 1\,\}$ (set všech rozdělení pravděpodobnosti diskrétní náhodné proměnné)
- zobecněný osmistěn $\{\,\mathbf{x} \in \mathbb{R}^n \mid \sum_{i=1}^n |x_i| \leq 1\,\}$. □

12.1.1 Stěny konvexního polyedru

Definition 12.2. Stěna *konvexního polyedru* $X \subseteq \mathbb{R}^n$ *is set*

$$F = \operatorname*{argmin}_{\mathbf{x} \in X} \mathbf{c}^T\mathbf{x}$$

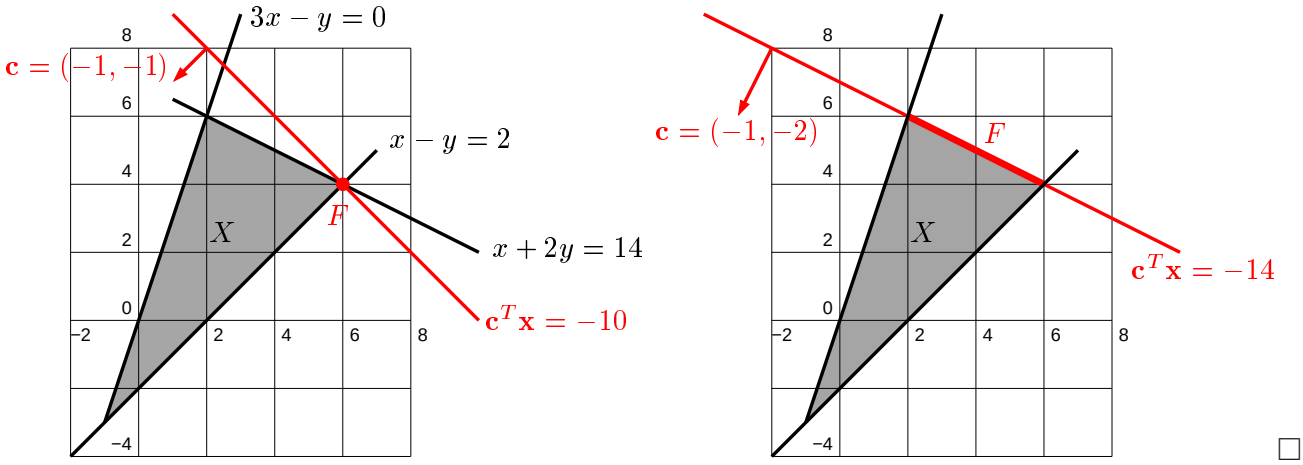
pro nějaké $\mathbf{c} \in \mathbb{R}^n$.

Tedy stěna konvexního polyedru is set všech jeho bodů, ve kterých nějaká linear funkce nabývá minima. Volbou $\mathbf{c} = \mathbf{0}$ dostaneme, že jednou ze stěn is i celý polyedr X.

Každá stěna konvexního polyedru is sama o sobě konvexní polyedr, neboť je průnikem polyedru X a nadroviny $\mathbf{c}^T\mathbf{x} = d$, kde $d = \min_{\mathbf{x}\in X} \mathbf{c}^T\mathbf{x}$. **Dimenze stěny** is dimenze jejího affineho obalu (zopakujte si pojem affineho obalu z §11.1.1 a dimenze affineho subspaceu 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).

Example 12.2. Polyedr $X = \{\,(x,y) \in \mathbb{R}^2 \mid x + 2y \leq 14,\ 3x - y \geq 0,\ x - y \leq 2\,\}$ snadno nakreslíte tužkou a pravítkem. Červená přímka je vrstevnice funkce $\mathbf{c}^T\mathbf{x}$ výšky $\min_{\mathbf{x}\in X} \mathbf{c}^T\mathbf{x}$. Zvolíme-li $\mathbf{c} = (-1,-1)$, minimum funkce $\mathbf{c}^T\mathbf{x}$ se na polyedru X nabývá ve vrcholu $F = \{(6,4)\}$, což is stěna dimenze 0 (levý obrázek). Zvolíme-li $\mathbf{c} = (-1,-2)$, minimum se nabývá na hraně $F = \operatorname{conv}\{(6,4),(2,6)\}$, což is stěna dimenze 1 (pravý obrázek).



□

12.1.2 Dvě reprezentace polyedru

Následující věta is hluboká a její důkaz není snadný.

Theorem 12.1. *Konvexní obal konečně mnoha bodů is omezený konvexní polyedr. Obráceně, omezený*[1] *konvexní polyedr is konvexním obalem svých vrcholů.*

we have 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 velmi těžký nebo i nemožný. Důvodem is to, že polyedr definovaný jako průnik malého počtu (přesněji, tento počet is polynomiální funkcí dimenze n) poloprostorů může mít obrovský (exponenciální v dimenzi 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 12.3.

- Simplex (tedy konvexní obal $n+1$ bodů) is konvexní polyedr, který má $n+1$ vrcholů a $n+1$ facet.
- Hyperkrychle má $2n$ facet a 2^n vrcholů.
- Zobecněný osmistěn $\{\, \mathbf{x} \in \mathbb{R}^n \mid \sum_i |x_i| \leq 1 \,\}$ má $2n$ vrcholů a 2^n facet. □

12.2 Úloha linearho programování

Úloha **linearho programování** (LP, také zvané linear optimisation) znamená minimalizaci linear funkce za podmínek linearch rovností a nerovností. Přesněji, v obecné formulaci (1.4) is funkce f linear a funkce g_i, h_i jsou affine.

Example 12.4. Příkladem linearho programu is úloha z Příkladu 12.2,

$$\begin{aligned} \min \quad & -x - y \\ \text{za podmínek} \quad & x + 2y \leq 14 \\ & 3x - y \geq 0 \\ & x - y \leq 2 \end{aligned} \tag{12.1}$$

□

Podle §12.1 is set přípustných řešení úlohy LP konvexní polyedr a úlohu LP lze tedy vždy psát jako

$$\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{A}\mathbf{x} \geq \mathbf{b} \,\}.$$

Můžeme samozřejmě hledat maximum místo minima, stačí otočit znaménko vectoru $\mathbf{c}$.

Zopakujme (viz §1.3), že pro řešitelnost LP mohou nastat tři případy: úloha má (alespoň jedno) optimální řešení, úloha je *nepřípustná* (i.e., set přípustných řešení is prázdná, neboli podmínky si navzájem odporují), úloha is *neomezená* (i.e., účelovou funkci lze za splněných podmínek zlepšovat nade všechny meze).

[1] Pro neomezené konvexní polyedry platí podobná věta, trochu složitější, kterou zde nebudeme uvádět.

12.3 Různé tvary úloh LP

Při zápisu úlohy LP is zvykem odděleně zapisovat obecná linear omezení a omezení na znaménka jednotlivých proměnných. Obecnou úlohu LP tedy zapíšeme jako

$$\begin{aligned}
\min \quad & c_1x_1 + \cdots + c_nx_n \\
\text{za podmínek} \quad & a_{i1}x_1 + \cdots + a_{in}x_n \geq b_i, \quad i \in I_+ \\
& a_{i1}x_1 + \cdots + a_{in}x_n \leq b_i, \quad i \in I_- \\
& a_{i1}x_1 + \cdots + a_{in}x_n = b_i, \quad i \in I_0 \\
& x_j \geq 0\,, \quad j \in J_+ \\
& x_j \leq 0\,, \quad j \in J_- \\
& x_j \in \mathbb{R}, \quad j \in J_0
\end{aligned}$$

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 decompositiony 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 ($I_+ = I_- = J_- = J_0 = \emptyset$), i.e.,

  $$\begin{aligned}
  \min \quad & c_1x_1 + \cdots + c_nx_n \\
  \text{za podmínek} \quad & a_{i1}x_1 + \cdots + a_{in}x_n = b_i, \quad i = 1, \ldots, m \\
  & x_j \geq 0\,, \quad j = 1, \ldots, n
  \end{aligned}$$

  To[2] 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 $\mathbf{x}, \mathbf{c} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m \times n}$.
- 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 pomocí vhodných operací. Některé takové operace jsou jasné, např. nahrazení omezení $\mathbf{a}_i^T\mathbf{x} = b_i$ dvěma omezeními $\mathbf{a}_i^T\mathbf{x} \geq b_i$, $-\mathbf{a}_i^T\mathbf{x} \geq -b_i$. Uvedeme dvě méně zřejmé operace:

- Nerovnost $a_{i1}x_1 + \cdots + a_{in}x_n \leq b_i$ převedeme na rovnost přidáním pomocné **slackové proměnné**[3] $u_i \geq 0$ 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 $x_i = x_i^+ - x_i^-$.

[2] Tomuto tvaru se někdy říká *standardní*. Podotkněme ovšem, že názvosloví různých tvarů LP není jednotné, názvy jako 'standardní tvar', 'základní tvar' či 'kanonický tvar' tedy mohou znamenat v různých knihách něco jiného.

[3] *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*.

Úloha získaná z původní úlohy pomocí těchto operací is ekvivalentní původní úloze v tom smyslu, že hodnota jejich optima is stejná a argument optima původní úlohy lze 'snadno' získat z argumentu optima nové úlohy.

Example 12.5. V úloze (12.1) chceme první podmínku převést na rovnost. To uděláme zavedením slackové proměnné $u \geq 0$. Transformovaná úloha je

$$\begin{array}{rrrrrl} \min & -x & - & y & & \\ \text{za podmínek} & x & + & 2y & + u & = 14 \\ & 3x & - & y & & \geq 0 \\ & x & - & y & & \leq 2 \\ & & & & u & \geq 0 \end{array}$$

Je-li (x, y, u) optimum této úlohy, optimum úlohy (12.1) je (x, y). □

Example 12.6. 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

$$\begin{array}{rrrrrrrrl} \min & -x_+ & + & x_- & - & y_+ & + & y_- & \\ \text{za podmínek} & x_+ & - & x_- & + 2y_+ & - & 2y_- & \leq & 14 \\ & 3x_+ & - & 3x_- & - & y_+ & + & y_- & \geq 0 \\ & x_+ & - & x_- & - & y_+ & + & y_- & \leq 2 \\ & & & & & x_+, x_-, y_+, y_- & & \geq & 0 \end{array}$$

□

12.3.1 Po částech affine funkce

consider účelovou funkci

$$f(\mathbf{x}) = \max_{k=1}^{K} (\mathbf{c}_k^T \mathbf{x} + d_k), \tag{12.2}$$

kde $\mathbf{c}_k \in \mathbb{R}^n$ a $d_k \in \mathbb{R}$ jsou dány (viz Cvičení 12.5). Řešme úlohu

$$\begin{array}{rl} \min & f(\mathbf{x}) \\ \text{za podmínek} & \mathbf{A}\mathbf{x} \geq \mathbf{b} \end{array} \tag{12.3}$$

Toto není úloha LP, neboť funkce f není linear nebo affine, je pouze po částech affine. Ovšem podle Věty 11.7 is tato funkce *konvexní*, proto úloha (12.3) je minimalizace konvexní funkce na konvexním polyedru a tedy se na ní vztahuje Věta 11.8.

Úloha jde převést na LP zavedením pomocné proměnné:

$$\begin{array}{rll} \min & z & \\ \text{za podmínek} & \mathbf{c}_k^T \mathbf{x} + d_k \leq z, & k = 1, \ldots, K \\ & \mathbf{A}\mathbf{x} \geq \mathbf{b} & \end{array} \tag{12.4}$$

kde minimalizujeme přes proměnné $(x_1, \ldots, x_n, z) = (\mathbf{x}, z) \in \mathbb{R}^{n+1}$. Ekvivalence úloh (12.3) a (12.4) se dokáže takto. Předpokládejme, že $(\mathbf{x}, z)$ is optimum úlohy (12.4). Then musí být alespoň jedno z omezení $\mathbf{c}_k^T \mathbf{x} + d_k \leq z$ aktivní (tedy musí platit s rovností), protože jinak bychom mohli z zmenšit a neporušit přitom žádné omezení. Z toho plyne $z = \max_{k=1}^{K} (\mathbf{c}_k^T \mathbf{x} + d_k)$.

Tento trik lze užít i na funkce obsahující absolutní hodnoty, neboť $|x| = \max\{-x, x\}$.

Při těchto převodech is nutná opatrnost: pokud bychom v úloze (12.3) maximalizovali místo minimalizovali, převod na LP by nebyl možný. Vodítkem je, že pokud úloha nejde jednoduše převést na minimalizaci konvexní funkce na konvexním polyedru, nepůjde převést na LP.

Example 12.7. Úloha

$$\begin{aligned} \min \quad & \max\{\, 3x_1 + 4x_2,\ 2x_1 - 3x_2 \,\} \\ \text{za podm.} \quad & x_1 + 2x_2 \le 14 \\ & 3x_1 - x_2 \ge 0 \\ & x_1 - x_2 \le 2 \end{aligned}$$

není LP, protože účelová funkce $f(x_1, x_2) = \max\{\, 3x_1 + 4x_2,\ 2x_1 - 3x_2 \,\}$ není linear ani affine. Lze ji ale přeformulovat na

$$\begin{aligned} \min \quad & z \\ \text{za podm.} \quad & 3x_1 + 4x_2 \le z \\ & 2x_1 - 3x_2 \le z \\ & x_1 + 2x_2 \le 14 \\ & 3x_1 - x_2 \ge 0 \\ & x_1 - x_2 \le 2 \end{aligned}$$

což is LP, neboť účelová funkce $f(x_1, x_2, z) = z$ is linear a omezení jsou také linear. □

Podobně lze často na LP převést úlohy, které obsahují minima a maxima v omezeních.

Example 12.8. 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 $\max\{x, y\} \le 1$ is ekvivalentní $x \le 1$, $y \le 1$. Úloha vlevo není LP, úloha vpravo ano. □

12.4 Některé aplikace LP

Zde uvedeme typické aplikace LP. Zdaleka to ale není výčet všech aplikací, ten is totiž nepřeberný.

12.4.1 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 we have k dispozici
- c_j = zisk z vyrobení jednoho výrobku druhu j
- x_j = počet vyrobených výrobků druhu j

Úkolem is zjistit, kolik jakých výrobků we have vyrobit, abychom dosáhli největšího zisku. Řešení:

$$\max\left\{ \sum_{j=1}^{n} c_j x_j \;\middle|\; \sum_{j=1}^{n} a_{ij} x_j \le b_i,\ x_j \ge 0 \right\}. \tag{12.5}$$

Example 12.9. 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. is 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

$$\begin{aligned} \max \quad & 120\ell + 76h \\ \text{za podmínek} \quad & 2\ell + 1.5h \leq 100 \\ & 0.4\ell + 0.2h \leq 16 \\ & \ell, h \geq 0 \end{aligned}$$

Přitom předpokládáme, že zbytky surovin se po pracovní době vyhodí. Pokud se zbytky využijí, tak maximalizujeme $(120 - 24 - 16)\ell + (76 - 18 - 8)h = 80\ell + 50h$. □

12.4.2 Směšovací (dietní) problém

Z n druhů surovin, z nichž každá is 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:

$$\min \left\{ \sum_{j=1}^{n} c_j x_j \;\middle|\; \sum_{j=1}^{n} a_{ij} x_j \geq b_i,\ x_j \geq 0 \right\}$$

Example 12.10. Jste kuchařka v menze a máte uvařit pro studenty co nejlevnější oběd, ve kterém ovšem musí být dané minimální množství živin (cukrů, bílkovin a vitamínů). Oběd budete vařit ze tří surovin: brambor, masa a zeleniny. Viz tabulka:

	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 is třeba každé suroviny na jeden oběd?

Minimalizujeme $25b + 50m + 80z$ za podmínek $2b + 1m + 1z \geq 8$, $2b + 6m + 1z \geq 16$, $1b + 3m + 6z \geq 8$. Optimální řešení is $b = 3.2$, $m = 1.6$, $z = 0$ s hodnotou 160. □

12.4.3 Dopravní problém

Given 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:

$$\min\left\{ \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} \;\middle|\; \sum_{j=1}^{n} x_{ij} = a_i,\ \sum_{i=1}^{m} x_{ij} = b_j,\ x_{ij} \geq 0 \right\}.$$

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!).

12.4.4 Distribuční problém

Given m strojů a n druhů výrobků.

- a_i = počet hodin, který is 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 we have určit, kolik výrobků se na něm bude vyrábět. Řešení:

$$\min\left\{ \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} \;\middle|\; \sum_{j=1}^{n} x_{ij} \leq a_i,\ \sum_{i=1}^{m} k_{ij}x_{ij} = b_j,\ x_{ij} \geq 0 \right\}.$$

12.5 Řešení přeurčených linearch soustav

consider přeurčenou linear soustavu $\mathbf{Ax} = \mathbf{b}$, 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

$$\min\{\, \|\mathbf{Ax} - \mathbf{b}\|_p \mid \mathbf{x} \in \mathbb{R}^n \,\}. \tag{12.6}$$

Zaměřme se podrobněji na tři případy:

- Pro $p = \infty$ hledáme takové $\mathbf{x}$, které minimalizuje výraz

$$\|\mathbf{Ax} - \mathbf{b}\|_\infty = \max_{i=1}^{m} |\mathbf{a}_i^T\mathbf{x} - b_i|, \tag{12.7}$$

  tedy minimalizuje maximální residuum. Toto řešení is známé pod názvem *minimaxní* nebo *Čebyševovo*. Úloha is ekvivalentní linearmu programu

$$\begin{array}{rl} \min & z \\ \text{za podm.} & \mathbf{a}_i^T\mathbf{x} - b_i \leq z, \quad i = 1,\ldots,m \\ & -\mathbf{a}_i^T\mathbf{x} + b_i \leq z, \quad i = 1,\ldots,m \end{array}$$

  který lze zapsat elegantněji jako

$$\min\{\, z \in \mathbb{R} \mid \mathbf{x} \in \mathbb{R}^n,\ -z\mathbf{1} \leq \mathbf{Ax} - \mathbf{b} \leq z\mathbf{1} \,\}. \tag{12.8}$$

- Pro $p = 2$ dostaneme řešení ve smyslu nejmenších čtverců, kterým jsme se zabývali v §6.1.
- Pro $p = 1$ hledáme takové $\mathbf{x}$, které minimalizuje výraz

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1 = \sum_{i=1}^{m} |\mathbf{a}_i^T \mathbf{x} - b_i|, \tag{12.9}$$

kde $\mathbf{a}_1, \ldots, \mathbf{a}_m$ jsou rows matrix $\mathbf{A}$. Úloha is ekvivalentní linearmu programu

$$\begin{aligned} \min \quad & \sum_{i=1}^{m} z_i \\ \text{za podm.} \quad & \mathbf{a}_i^T \mathbf{x} - b_i \le z_i, \quad i = 1, \ldots, m \\ & -\mathbf{a}_i^T \mathbf{x} + b_i \le z_i, \quad i = 1, \ldots, m \end{aligned}$$

který lze zapsat elegantněji v maticovém tvaru jako

$$\min\{\, \mathbf{1}^T \mathbf{z} \mid \mathbf{z} \in \mathbb{R}^m,\ \mathbf{x} \in \mathbb{R}^n,\ -\mathbf{z} \le \mathbf{A}\mathbf{x} - \mathbf{b} \le \mathbf{z} \,\}. \tag{12.10}$$

12.5.1 Použití na robustní regresi

Řešení ve smyslu 1-normy se používá tehdy, když potřebujeme modelovat funkční závislost naměřených dat (tedy děláme regresi, viz §6.1.4) a malá část dat is naměřená úplně špatně (např. se někdo při zapisování čísel spletl v desetinné čárce). Takovým datovým bodům s hrubou chybou se říká **vychýlené body** (*outliers*). Disciplína zabývající se modelováním funkčních závislostí za přítomnosti vychýlených bodů se nazývá **robustní regrese**.

V tomto případě řešení ve smyslu nejmenších čtverců není vhodné (není 'robustní'), protože i jediný vychýlený bod velmi ovlivní řešení. Řešení ve smyslu 1-normy tuto neblahou vlastnost nemá, přesněji, má ji v menší míře.

Ukážeme to na nejjednodušším možném případu regrese: odhad hodnoty jediného reálného numbers ze souboru jeho nepřesných měření. consider numbers $\mathbf{b} = (b_1, \ldots, b_m) \in \mathbb{R}^m$ a řešme úlohu (12.6) ve tvaru

$$\min\{\, \|(x - b_1, \ldots, x - b_m)\|_p \mid x \in \mathbb{R} \,\} = \min\{\, \|\mathbf{1}x - \mathbf{b}\|_p \mid x \in \mathbb{R} \,\}. \tag{12.11}$$

- Pro $p = \infty$ minimalizujeme funkci $f(x) = \max_{i=1}^m |x - b_i|$. Řešením is střed intervalu krajních bodů, $x = \frac{1}{2}\big(\min_{i=1}^m b_i + \max_{i=1}^m b_i\big)$.
- Pro $p = 2$ minimalizujeme funkci $f(x) = \sqrt{\sum_{i=1}^m (x - b_i)^2}$. Řešením is aritmetický průměr, $x = \frac{1}{m}\sum_{i=1}^m b_i$ (viz Příklad 6.4).
- Pro $p = 1$ minimalizujeme funkci $f(x) = \sum_{i=1}^m |x - b_i|$. Řešením x is *medián* z čísel b_i (dokažte!). Medián se vypočte tak, že seřadíme numbers b_i podle velikosti a vezmeme prostřední z nich. Pokud is m sudé, we have dva 'prostřední prvky' a v tom případě funkce f nabývá minima v jejich libovolné konvexní kombinaci. is Then úzus definovat medián jako aritmetický průměr prostředních prvků.

Předpokládejme nyní, že jeden libovolný bod (např. b_1) se bude zvětšovat. V tom případě se řešení x pro různá p budou chovat různě. Např. aritmetický průměr se bude zvětšovat, a to tak, že zvětšováním hodnoty b_1 dosáhneme *libovolné* hodnoty x. Pro medián to ovšem neplatí – zvětšováním jediného bodu b_i ovlivníme x jen natolik, nakolik to změní pořadí bodů. Jeho libovolným zvětšováním nedosáhneme libovolné hodnoty x.

Example 12.11. Šuplérou změříme průměr ocelové kuličky v několika místech, dostaneme hodnoty $\mathbf{b} = (1.02, 1.04, 0.99, 2.03)$ (cm). Při posledním měření jsme se na stupnici přehlédli, proto is poslední hodnota úplně špatně. Z těchto měření chceme odhadnout skutečný průměr. Máme

$$\frac{1}{2}\big(\min_{i=1}^{m} b_i + \max_{i=1}^{m} b_i\big) = 1.51, \qquad \frac{1}{m}\sum_{i=1}^{m} b_i = 1.27, \qquad \operatorname*{median}_{i=1}^{m} b_i = 1.03.$$

Je zjevné, že medián is 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.

12.6 Exercises

12.1. Které z následujících množin jsou konvexní polyedry? Pokud je set konvexní polyedr, dokážete ji vyjádřit ve tvaru $\{\,\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{b}\,\}$ (i.e., jako průnik poloprostorů)?

a) $\{\,2y_1 + 3y_2 \mid -1 \leq y_1 \leq 1,\ -1 \leq y_2 \leq 1\,\}$
b) $\big\{\,\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} \geq \mathbf{0},\ \mathbf{1}^T\mathbf{x} = 1,\ \sum_i x_i a_i = b_1,\ \sum_i x_i a_i^2 = b_2\,\big\}$, kde a_i, b_1, b_2 jsou dané skaláry
c) $\{\,\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{a}\|_2 \leq \|\mathbf{x} - \mathbf{b}\|_2\,\}$, kde $\mathbf{a}, \mathbf{b}$ jsou dány
d) $\{\,\mathbf{C}\mathbf{y} \mid \mathbf{y} \geq \mathbf{0},\ \mathbf{1}^T\mathbf{y} = 1\,\}$, kde matrix $\mathbf{C}$ is dána
e) $\{\,\mathbf{C}\mathbf{y} \mid \|\mathbf{y}\|_2 \leq 1\,\}$, kde matrix $\mathbf{C}$ is dána

12.2. consider vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$. 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 set $X_1, \ldots, X_m$ jsou konvexní polyedry. Ukažte, že tyto set tvoří decomposition (zopakujte si, co is to decomposition set) set $\mathbb{R}^n$. Sjednocení hranic těchto množin se nazývá *Voronoiův diagram*. Nakreslete si ho pro $n = 2$ a $m = 4$ pro různé konfigurace bodů $\mathbf{a}_1, \ldots, \mathbf{a}_4$.

12.3. 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: $\mathbf{c} = (-1, 0, 1)$, $\mathbf{c} = (0, 1, 0)$, $\mathbf{c} = (0, 0, -1)$.

12.4. Vyřešte úvahou tyto úlohy a napište vzorec pro optimální hodnotu. Ve všech úlohách optimalizujeme přes proměnné $\mathbf{x} \in \mathbb{R}^n$ (příp. také $\mathbf{y} \in \mathbb{R}^n$). Parametry $\mathbf{c} \in \mathbb{R}^n$ a $k \in \mathbb{N}$, $1 \leq k \leq n$, jsou dány.

a) $\min\{\,\mathbf{c}^T\mathbf{x} \mid \mathbf{0} \leq \mathbf{x} \leq \mathbf{1}\,\}$ (výsledek: $\sum_{i \mid c_i < 0} c_i$, tedy záporných čísel c_i)

b) $\min\{\, \mathbf{c}^T\mathbf{x} \mid -\mathbf{1} \le \mathbf{x} \le \mathbf{1} \,\}$ (výsledek: $-\sum_i |c_i|$)

c) ($\star$) $\min\{\, \mathbf{c}^T\mathbf{x} \mid 0 \le x_1 \le x_2 \le \cdots \le x_n \le 1 \,\}$
Nápověda: Proveďte substituci $y_i = x_i - x_{i-1}$.

d) $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \ge \mathbf{0},\ \mathbf{1}^T\mathbf{x} = 1 \,\}$

e) $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \ge \mathbf{0},\ \mathbf{1}^T\mathbf{x} \le 1 \,\}$

f) $\max\{\, \mathbf{c}^T\mathbf{x} \mid -1 \le \mathbf{1}^T\mathbf{x} \le 1 \,\}$

g) $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \ge \mathbf{0},\ \mathbf{1}^T\mathbf{x} = k \,\}$

h) $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{0} \le \mathbf{x} \le \mathbf{1},\ \mathbf{1}^T\mathbf{x} = k \,\}$

i) $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{0} \le \mathbf{x} \le \mathbf{1},\ \mathbf{1}^T\mathbf{x} \le k \,\}$

j) ($\star$) $\max\{\, \mathbf{c}^T\mathbf{x} \mid -\mathbf{y} \le \mathbf{x} \le \mathbf{y},\ \mathbf{1}^T\mathbf{y} = k,\ \mathbf{y} \le \mathbf{1} \,\}$ kde předpokládáme $\mathbf{a} \ge \mathbf{b}$

12.5. Pochopte kód v Matlab, který vizualizuje funkci $f(\mathbf{x}) = \max_{k=1}^{K}(\mathbf{c}_k^T\mathbf{x} + d_k)$ pro $n = 2$:

```
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. Převeďte na LP nebo odůvodněte, proč to nejde:

a) $\max\{\, |\mathbf{c}^T\mathbf{x}| \mid \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \ge \mathbf{0} \,\}$

b) $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b},\ |\mathbf{d}^T\mathbf{x} + e| \le f,\ \mathbf{x} \ge \mathbf{0} \,\}$

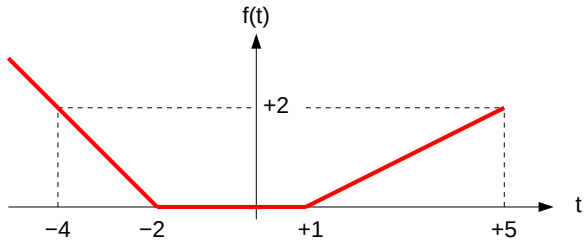
c) $\min\left\{ \sum_{\ell=1}^{L} \max_{k=1}^{K}(\mathbf{c}_{k\ell}^T\mathbf{x} + d_{k\ell}) \,\middle|\, \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \ge \mathbf{0} \right\}$

d) $\max\{\, |x - 1| + 2|y + 1| \mid x, y \in \mathbb{R},\ x + y \le 2 \,\}$

e) $\min\{\, |x_1| + |x_2| + |x_3| \mid 2x_1 - x_2 - x_3 \ge 1,\ -x_1 + 2x_2 - x_3 \ge 1,\ -x_1 - x_2 + 2x_3 \ge 1 \,\}$

f) $\max\{\, \min\{\mathbf{p}^T\mathbf{x}, \mathbf{q}^T\mathbf{x}\} - |\mathbf{r}^T\mathbf{x}| \ \mid\ \mathbf{x} \in \mathbb{R}^n,\ \mathbf{A}\mathbf{x} \ge \mathbf{b} \,\}$, kde $\mathbf{p}, \mathbf{q}, \mathbf{r} \in \mathbb{R}^n$

g) $\min_{\mathbf{x}\in\mathbb{R}^n} \sum_{i=1}^{m} f(\mathbf{a}_i^T\mathbf{x} - b_i)$, kde funkce f is dána obrázkem



h) $\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1$

i) $\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_\infty$

j) $\min\{\, \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1 \mid \mathbf{x} \in \mathbb{R}^n,\ \|\mathbf{x}\|_\infty \le 1 \,\}$

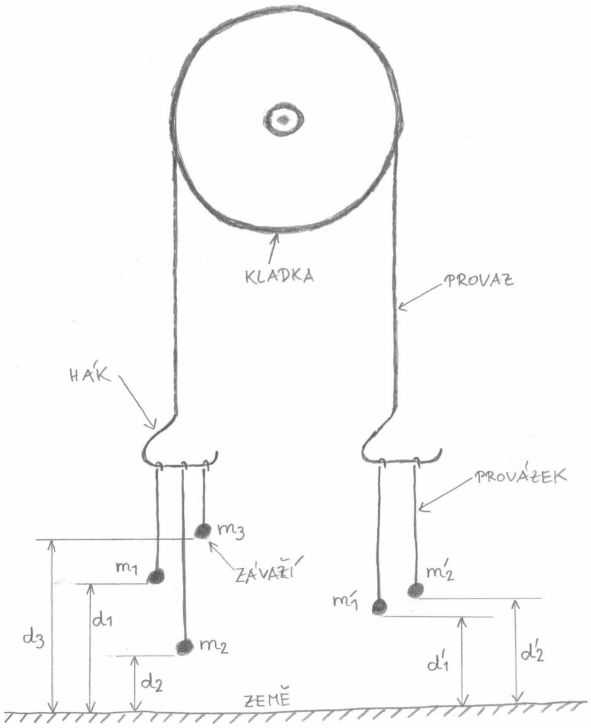
k) $\min\{\, \|\mathbf{x}\|_1 \mid \mathbf{x} \in \mathbb{R}^n,\ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_\infty \le 1 \,\}$

l) $\min_{\mathbf{x}\in\mathbb{R}^n}(\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1 + \|\mathbf{x}\|_\infty)$

m) $\min_{\mathbf{x}\in\mathbb{R}^n} \sum_{k=1}^{K} \|\mathbf{A}_k\mathbf{x} - \mathbf{b}_k\|_\infty$, kde $\mathbf{A}_1, \ldots, \mathbf{A}_K$ jsou dané matrix a $\mathbf{b}_1, \ldots, \mathbf{b}_K$ jsou dané vectors.

12.7. 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.

12.8. Given 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 is 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 is 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 (Then is jeho potenciální energie rovna m_i krát výška nad zemí) nebo leží na zemi (Then is jeho potenciální energie nulová). Soustava bude v rovnováze při minimální celkové potenciální energii.

Napište vzorec pro celkovou potenciální energii soustavy jako funkci x. Je-li to možné, napište linear program, jehož optimum is rovno minimální potenciální energii soustavy. Není-li to možné, vysvětlete.

12.9. Dokažte nebo vyvrat'te následující rovnosti:

- $\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \|\mathbf{x}\| = 1 \,\} = \max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \|\mathbf{x}\| \leq 1 \,\}$
- $\min\{\, \mathbf{c}^T\mathbf{x} \mid \|\mathbf{x}\| = 1 \,\} = \min\{\, \mathbf{c}^T\mathbf{x} \mid \|\mathbf{x}\| \leq 1 \,\}$
- $\max\{\, \|\mathbf{A}\mathbf{x}\| \mid \|\mathbf{x}\| = 1 \,\} = \max\{\, \|\mathbf{A}\mathbf{x}\| \mid \|\mathbf{x}\| \leq 1 \,\}$

Zde $\|\cdot\|$ is libovolná norma a $\mathbf{c} \in \mathbb{R}^n$ a $\mathbf{A} \in \mathbb{R}^{m\times n}$ jsou dány. Nápověda: Inspirujte se úvahou v §12.3.1.

12.10. Veverka před zimou potřebuje přerovnat zásoby oříšků. Stávající zásoby má v m jamkách, přičemž i-tá jamka má souřadnice $\mathbf{p}_i$ a je v ní a_i oříšků. Potřebuje is přenosit do n nových připravených jamek, přičemž j-tá jamka má souřadnice $\mathbf{q}_j$ a na konci v ní bude y_j oříšků. Veverka za prvé chce vykonat co nejméně práce, kde práce na přenesení jednoho oříšku is přímo úměrná vzdálenosti (vzdušnou čarou) nesení (běh bez oříšku se za práci nepovažuje). Za druhé chce, aby v nových jamkách byly oříšky rozloženy co nejrovnoměrněji, přesněji, aby rozdíl mezi největším a největším z čísel y_j byl menší než dané číslo t (tím minimalizuje škodu způsobenou případnou krádeží). Spočítejte x_{ij} a y_j, kde x_{ij} is počet oříšků

přenesených ze staré jamky i do nové jamky j. Zanedbejte skutečnost, že počty oříšků mohou být celá (a ne pouze nezáporná) numbers.

Chapter 13

Simplex Method

Zde popíšeme algoritmus na řešení úloh linearho programování zvaný **simplexová metoda**.

Zapomeňme prozatím na účelovou funkci a zkoumejme množinu přípustných řešení LP ve tvaru

$$X = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \,\}, \tag{13.1}$$

kde $\mathbf{A} \in \mathbb{R}^{m \times n}$ is široká ($m < n$) matrix s hodností m, tedy její rows jsou linearly independent.

Soustava $\mathbf{A}\mathbf{x} = \mathbf{b}$ má nekonečně mnoho řešení. Položíme-li však $n - m$ složek vectoru $\mathbf{x}$ rovno nule, soustava může mít jediné řešení. Tato úvaha vede k následujícím definicím:

- set $L \subseteq \{1, 2, \ldots, n\}$ se nazývá **báze** úlohy, pokud $|L| = m$ a columns matrix $\mathbf{A}$ s indexy L jsou linearly independent. Tedy columns L tvoří regular matici $m \times m$.
- vector $\mathbf{x}$ se nazývá **bázové řešení** příslušné bázi L, pokud $\mathbf{A}\mathbf{x} = \mathbf{b}$ a $x_j = 0$ pro $j \notin L$.
- Bázové řešení $\mathbf{x}$ se nazývá **přípustné**, pokud $\mathbf{x} \geq \mathbf{0}$.
- Bázové řešení $\mathbf{x}$ se nazývá **degenerované**, pokud má méně než m nenulových složek.
- Dvě báze se nazývají **sousední**, pokud mají $m - 1$ společných prvků.

Protože $\mathbf{A}$ má hodnost m, existuje aspoň jedna báze. is jasné, že báze určuje jednoznačně bázové řešení. Bázové řešení však může odpovídat více než jedné bázi, což se stane právě tehdy, když is toto bázové řešení degenerované.

Example 13.1. Let is soustava $\mathbf{A}\mathbf{x} = \mathbf{b}$ dána tabulkou

$$[\,\mathbf{A} \,|\, \mathbf{b}\,] = \left[\begin{array}{cccccc|c} -1 & 1 & 3 & 1 & 0 & 2 & 1 \\ 1 & 0 & 4 & 0 & 1 & 4 & 4 \\ -1 & 0 & 4 & 1 & 1 & 4 & 2 \end{array}\right].$$

- $L = \{2, 3, 5\}$ není báze, protože columns $2, 3, 5$ matrix $\mathbf{A}$ jsou linearly závislé.
- $L = \{1, 2, 4\}$ is báze, protože tyto columns jsou linearly independent. Bázové řešení $\mathbf{x} = (x_1, x_2, \ldots, x_6)$ příslušné bázi L se najde řešením soustavy

$$\begin{bmatrix} -1 & 1 & 1 \\ 1 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix}$$

  a položením $x_3 = x_5 = x_6 = 0$. Dostaneme $\mathbf{x} = (4, -1, 0, 6, 0, 0)$. Toto bázové řešení is nepřípustné, protože $x_2 < 0$. Není degenerované, protože má $m = 3$ nenulových složek.

- $L = \{1, 2, 6\}$ is báze. Bázové řešení je $\mathbf{x} = (1, \frac{1}{2}, 0, 0, 0, \frac{3}{4})$. is přípustné, protože $\mathbf{x} \geq \mathbf{0}$.
- $L = \{3, 4, 5\}$ is báze. Bázové řešení je $\mathbf{x} = (0, 0, 1, -2, 0, 0)$. is nepřípustné. Navíc is degenerované, protože má méně než $m = 3$ nenulových složek.
- Stejné bázové řešení $\mathbf{x} = (0, 0, 1, -2, 0, 0)$ dostaneme volbou báze $L = \{3, 4, 6\}$. Vidíme, že bázové řešení odpovídá více než jedné bázi, protože is degenerované.
- Báze $\{2, 3, 5\}$ a $\{3, 4, 5\}$ jsou sousední, protože mají společné dva prvky $\{3, 5\}$. Báze $\{2, 3, 5\}$ a $\{1, 2, 4\}$ nejsou sousední, protože mají společný jen jeden prvek. □

Následující věta udává spojitost mezi algebraickým a geometrickým popisem konvexního polyedru (13.1). Důkaz vynecháme, není obtížný.

Theorem 13.1. *Přípustná bázová řešení jsou vrcholy polyedru (13.1), přičemž dvojice sousedních bází odpovídají dvojici vrcholů spojených hranou.*

Víme, že optimum linear funkce na polyedru X se nabývá na jeho stěně, tedy alespoň v jednom vrcholu. To nám dovoluje navrhnout naivní algoritmus na řešení LP: uděláme výčet všech přípustných bázových řešení a nalezneme to s nejlepší hodnotou účelové funkce. Tato metoda samozřejmě nelze prakticky použít, protože přípustných bázových řešení je exponenciálně mnoho. Vlastně is jen o málo chytřejší než metoda popsaná v §9.3.

Simplexová metoda is 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).

13.1 Stavební kameny algoritmu

Zde vysvětlíme jednotlivé stavební kameny simplexové metody, které nakonec v §13.2 spojíme v celý algoritmus.

13.1.1 Přechod k sousední standardní bázi

Simplexový algoritmus udržuje pouze *standardní báze*. V tom případě jsou nenulové složky bázového řešení $\mathbf{x}$ rovny jednoduše složkám vectoru $\mathbf{b}$.

Z linear algebry známe *ekvivalentní řádkové úpravy* soustavy $\mathbf{Ax} = \mathbf{b}$: libovolný řádek tabulky $[\,\mathbf{A} \,|\, \mathbf{b}\,]$ můžeme vynásobit nenulovým číslem a můžeme k němu přičíst libovolnou linear kombinaci ostatních řádků. Tyto úpravy nemění množinu řešení soustavy.

Ukážeme, jak přejít od aktuální standardní báze L k sousední standardní bázi, tedy nahradit jeden bázový sloupec $j' \in L$ nebázovým columnsm $j \notin L$. Let i is takový řádek, ve kterém má sloupec j' jedničku, $a_{ij'} = 1$. Prvek a_{ij} 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 columns $L \setminus \{j'\}$. 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 13.2. consider soustavu

$$[\,\mathbf{A}\,|\,\mathbf{b}\,] = \left[\begin{array}{cccccc|c} 0 & 2 & 6 & 1 & 0 & 4 & 4 \\ 1 & \boxed{1} & 3 & 0 & 0 & 2 & 3 \\ 0 & -1 & 1 & 0 & 1 & 2 & 1 \end{array}\right] \tag{13.2}$$

se standardní bází $L = \{1, 4, 5\}$. Vidíme ihned odpovídající bázové řešení, $\mathbf{x} = (3, 0, 0, 4, 1, 0)$.

Nahradíme bázový sloupec $j' = 1$ nebázovým columnsm $j = 2$, tedy přejdeme k sousední bázi $\{2, 4, 5\}$. we have $i = 2$, tedy pivot is prvek a_{22} (v tabulce zvýrazněn). Řádkovými úpravami musíme docílit, aby pivot byl roven jedné a prvky nad ním a pod ním byly nulové. Při tom smíme 'zničit' sloupec 1, ale columns 4 a 5 se změnit nesmějí. Toho se docílí vydělením řádku 2 číslem a_{22} (což zde lze vynechat, protože náhodou we have $a_{22} = 1$) a Then přičtením vhodných násobků řádku 2 k ostatním řádkům. Výsledek:

$$[\,\mathbf{A}\,|\,\mathbf{b}\,] = \left[\begin{array}{cccccc|c} -2 & 0 & 0 & 1 & 0 & 0 & -2 \\ 1 & 1 & 3 & 0 & 0 & 2 & 3 \\ 1 & 0 & 4 & 0 & 1 & 4 & 4 \end{array}\right].$$

Nyní columns $\{2, 4, 5\}$ tvoří standardní bázi. □

13.1.2 Kdy is sousední bázové řešení přípustné?

Uvedeným způsobem můžeme od aktuální standardní báze přejít k libovolné sousední standardní 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 vectoru $\mathbf{b}$, bázové řešení is přípustné právě tehdy, když $\mathbf{b} \geq \mathbf{0}$. Let v aktuální tabulce is $\mathbf{b} \geq \mathbf{0}$. 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 vector $\mathbf{b}$ změní takto (viz §13.1.1):

- b_i se změní na b_i/a_{ij},
- pro $i' \neq i$ se $b_{i'}$ změní na $b_{i'} - a_{i'j}(b_i/a_{ij})$.

Tato numbers musejí být nezáporná. To nastane právě tehdy, když platí následující podmínky:

$$a_{ij} > 0 \tag{13.3a}$$

$$\text{Pro každé } i' \neq i \text{ platí } a_{i'j} \leq 0 \text{ nebo } \frac{b_i}{a_{ij}} \leq \frac{b_{i'}}{a_{i'j}} \tag{13.3b}$$

kde 'nebo' is užito v nevylučovacím smyslu. Podmínka (13.3a) is zřejmá. Podmínka (13.3b) je ekvivalentní podmínce $b_{i'} - a_{i'j}(b_i/a_{ij}) \geq 0$, uvědomíme-li si, že $a_{ij} > 0$, $b_i > 0$, $b_{i'} > 0$.

Example 13.3. Uvažujme opět soustavu (13.2).

- Povede ekvivalentní úprava okolo pivotu $(i, j) = (3, 2)$ k přípustnému řešení? Ne, protože $a_{ij} = -1 < 0$, což porušuje podmínku (13.3a).
- Povede ekvivalentní úprava okolo pivotu $(i, j) = (2, 2)$ k přípustnému řešení (tuto úpravu jsme již provedli v Příkladu 13.2)? Ne, protože (i, j) nesplňuje podmínku (13.3b). Pro $i' = 3$ we have $a_{i'j} = -1 \leq 0$, tedy podmínka (13.3b) is splněna. Ale pro $i' = 1$ is $a_{i'j} > 0$, tedy musí být $\frac{3}{1} \leq \frac{4}{2}$, což není pravda.

- Povede ekvivalentní úprava okolo pivotu $(i,j) = (3,6)$ k přípustnému řešení? Podmínka (13.3a) is splněna, protože $a_{ij} = 2 > 0$. Podmínka (13.3b) vyžaduje $\frac{1}{2} \leq \frac{4}{4}$ a $\frac{1}{2} \leq \frac{3}{2}$, což platí. □

13.1.3 Co znamená nekladný sloupec?

Jestliže jsou všechny prvky v nebázovém sloupci nekladné, tento sloupec se nemůže stát bázovým, nebot' v něm nelze vybrat pivot splňující podmínku (13.3a). V tom případě se některé složky vectoru $\mathbf{x}$ mohou zvětšovat nade všechny meze. Tedy existuje polopřímka s počátkem v $\mathbf{x}$ ležící celá v polyedru X. To znamená, že polyedr X is neomezený.

Example 13.4. Let tabulka $[\,\mathbf{A}\,|\,\mathbf{b}\,]$ vypadá takto:

$$\begin{array}{rrrrrrr|l} & 0 & -2 & 6 & 1 & 0 & 4 & 4 \\ & 1 & -1 & 3 & 0 & 0 & 2 & 3 \\ & 0 & -1 & 1 & 0 & 1 & 2 & 1 \\ \hline \mathbf{x} = & 3 & 0 & 0 & 4 & 1 & 0 & \end{array}$$

Báze is $\{1,4,5\}$. Pod tabulkou is napsáno odpovídající bázové řešení $\mathbf{x}$. Když se x_2 bude libovolně zvětšovat, změnu lze kompenzovat současným zvětšováním x_1, x_4, x_5 tak, že vector $\mathbf{Ax}$ zůstane nezměněn a tedy roven $\mathbf{b}$. Konkrétně, vector $\mathbf{x} = (\,3+x_2,\ x_2,\ 0,\ 4+2x_2,\ 1+x_2,\ 0\,)$ bude pro každé $x_2 \geq 0$ splňovat $\mathbf{Ax} = \mathbf{b}$ a $\mathbf{x} \geq \mathbf{0}$. □

13.1.4 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

$$\min\{\,\mathbf{c}^T\mathbf{x} - d \mid \mathbf{Ax} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0}\,\}. \tag{13.4}$$

Úlohu budeme reprezentovat **simplexovou tabulkou**

$$\left[\begin{array}{c|c} \mathbf{c}^T & d \\ \hline \mathbf{A} & \mathbf{b} \end{array}\right]. \tag{13.5}$$

Přičtěme k účelovému řádku $[\,\mathbf{c}^T\,|\,d\,]$ libovolnou linear kombinaci $\mathbf{y}^T[\,\mathbf{A}\,|\,\mathbf{b}\,]$ ostatních řádků $[\,\mathbf{A}\,|\,\mathbf{b}\,]$, kde $\mathbf{y}$ jsou coefficienty linear 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

$$[\,\mathbf{c}^T\,|\,d\,] + \mathbf{y}^T[\,\mathbf{A}\,|\,\mathbf{b}\,] = [\,\mathbf{c}^T + \mathbf{y}^T\mathbf{A}\,|\,d + \mathbf{y}^T\mathbf{b}\,].$$

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 is rovno $\mathbf{c}^T\mathbf{x} - d$ pro každé $\mathbf{x}$ splňující $\mathbf{Ax} = \mathbf{b}$.

13.1.5 Co udělá přechod k sousední bázi s účelovou funkcí?

Let columns L tvoří standardní bázi. Přičtěme k účelovému řádku takovou linear kombinaci ostatních řádků, aby coefficienty c_j v bázových sloupcích $j \in L$ byly nulové. To lze vždy udělat. Protože bázové řešení $\mathbf{x}$ is v nebázových sloupcích nulové, znamená to $\mathbf{c}^T\mathbf{x} = 0$. Tedy hodnota kritéria $\mathbf{c}^T\mathbf{x} - d$ v bázovém řešení $\mathbf{x}$ is rovna jednoduše $-d$. Navíc is na první pohled vidět, co udělá s kritériem vložení nebázového columns $j \notin L$ do báze: při $c_j \geq 0$ kritérium stoupne nebo se nezmění, při $c_j \leq 0$ kritérium klesne nebo se nezmění.

Example 13.5. consider úlohu se standardní bází $\{1, 4, 5\}$:

$$\left[\begin{array}{c|c} \mathbf{c}^T & d \\ \hline \mathbf{A} & \mathbf{b} \end{array}\right] = \left[\begin{array}{cccccc|c} 1 & -2 & -3 & -1 & 2 & 1 & 4 \\ \hline 0 & 2 & 6 & 1 & 0 & 4 & 4 \\ 1 & 1 & 3 & 0 & 0 & 2 & 3 \\ 0 & -1 & 1 & 0 & 1 & 2 & 1 \end{array}\right].$$

Vynulujme hodnoty vectoru $\mathbf{c}$ v bázových sloupcích. To uděláme tak, že ke kriteriálnímu řádku přičteme první řádek, odečteme druhý řádek, a odečteme dvojnásobek třetího řádku:

	0	1	−2	0	0	−1	3
	0	2	6	1	0	4	4
	1	1	3	0	0	2	3
	0	−1	1	0	1	2	1
$\mathbf{x} =$	3	0	0	4	1	0	

Do tabulky jsme úplně dolů navíc napsali odpovídající bázové řešení $\mathbf{x}$. is vidět, že $\mathbf{c}^T\mathbf{x} = 0$ a tedy aktuální hodnota kritéria 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 kritériu neprojeví a kritérium se změní o c_2x_2. Kritérium tedy stoupne nebo zůstane stejné, protože $c_2 = 1 > 0$. □

Pokud v některém sloupci j platí $c_j \leq 0$ a $a_{ij} \leq 0$ pro všechna i, Then můžeme proměnnou x_j libovolně zvětšovat (viz 13.1.3) a úloha is tedy neomezená (její optimum se blíží $-\infty$).

13.2 Základní algoritmus

Spojením popsaných stavebních kamenů dostaneme iteraci simplexové metody. Iterace přejde k sousední standardní bázi takové, že bázové řešení zůstane přípustné a účelová funkce se zmenší nebo alespoň nezmění. Vstupem i výstupem iterace is simplexová tabulka (13.4) s těmito vlastnostmi:

- podset sloupců $\mathbf{A}$ tvoří standardní bázi,
- bázové řešení odpovídající této bázi is přípustné, i.e., $\mathbf{b} \geq \mathbf{0}$,
- složky vectoru $\mathbf{c}$ v bázových sloupcích jsou nulové.

Iteraci se provede v těchto krocích:

1. Vyber index j pivotu podle znamének čísel $c_1, \ldots, c_n$.

2. Vyber index i pivotu podle podmínek (13.3). Z těchto podmínek plyne (promyslete!)

$$i \in \operatorname*{argmin}_{i' \mid a_{i'j}>0} \frac{b_{i'}}{a_{i'j}}, \tag{13.6}$$

kde tento zápis znamená, že minimalizujeme přes všechna i' splňující $a_{i'j} > 0$.

3. Udělej ekvivalentní úpravu tabulky $[\,\mathbf{A} \mid \mathbf{b}\,]$ okolo pivotu (i, j).
4. Udělej ekvivalentní úpravu účelového řádku, která vynuluje c_j v novém bázovém sloupci j.

Algoritmus, který opakuje uvedenou iteraci, nazveme **základní simplexový algoritmus**. Algoritmus končí, když už nelze iteraci provést. To nastane z jednoho z těchto důvodů:

- Všechny coefficienty c_j jsou nezáporné (kritérium nelze zlepšit a jsme v optimu).
- V některém sloupci is $c_j < 0$ a $a_{ij} \leq 0$ pro všechna i (úloha is neomezená).

Výběr indexů (i, j) pivotu v kroku 1 nemusí být jednoznačný, tedy může být více sloupců j s vhodným znaménkem c_j a více řádků i může splňovat podmínky (13.3) (tedy může být více argumentů minima v podmínce (13.6)). Algoritmus, který vybírá jediný pivot z několika 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 is *Blandovo anticyklící pravidlo*), které ale popisovat nebudeme[1].

Example 13.6. Vyřešte simplexovou metodou:

$$\begin{aligned} \min \quad & -6x_1 - 8x_2 - 5x_3 - 9x_4 \\ \text{za podmínek} \quad & 2x_1 + x_2 + x_3 + 3x_4 + x_5 = 5 \\ & x_1 + 3x_2 + x_3 + 2x_4 + x_6 = 3 \\ & x_1, x_2, x_3, x_4, x_5, x_6 \geq 0 \end{aligned}$$

Výchozí tabulka je

-6	-8	-5	-9	0	0	0
2	1	1	3	1	0	5
1	3	1	$\boxed{2}$	0	1	3

Účelový řádek budeme nazývat nultý, ostatní Then prvý, druhý atd. Počáteční báze is $L = \{5, 6\}$.

Vybereme sloupec, který nově vstoupí do báze. To může být libovolný sloupec, který má v nultém řádku záporné číslo. is rozumné vzít nejmenší takové číslo, zde -9, tedy sloupec 4.

Protože do báze chceme přidat sloupec 4, musí některý ze sloupců 5 a 6 z báze ven. Jeho index získáme porovnáním čísel $\frac{5}{3}$ a $\frac{3}{2}$ (čitatel is vždy vpravo, jmenovatel is vždy ve sloupci, který má přijít do báze; uvažujeme ale jen podíly s kladným jmenovatelem): vybereme to nejmenší z nich. Protože $\frac{5}{3} > \frac{3}{2}$, pivot bude v řádku 2. Všimněte si, že přes stávající standardní bázi řádek 2 odpovídá sloupci 2 – tento sloupec tedy půjde z báze ven.

Výsledný pivot (který is v takto nalezeném řádku a sloupci) is označen rámečkem. Na základě něj spočítáme novou tabulku. To uděláme ekvivalentními řádkovými úpravami, kterými musíme dosáhnout toho, že:

[1] Anticyklící pivotové pravidlo tedy zaručí skončení simplexového algoritmu za konečný počet iterací. is hlubokým otevřeným problémem, zda existuje pivotové pravidlo, které zaručuje ukončení simplexového algoritmu v *polynomiálním* počtu iterací.

- z pivotu se stane jednička,
- nad i pod pivotem budou nuly, a to včetně nultého řádku.

Jediný způsob, jak toho dosáhnout, is pomocí těchto dvou úprav:

- přičítat vhodné násobky pivotového řádku k ostatním řádkům (tedy k ničemu nikdy nepřičítáme násobky jiného řádku než pivotového)
- samotný pivotový řádek dělit kladným číslem

Tedy k nultému řádku přičteme $\frac{9}{2}$ druhého řádku, k prvnímu řádku přičteme $-\frac{3}{2}$ druhého řádku, a druhý řádek vydělíme dvěma:

$$\begin{array}{cccccc|c} -1.5 & 5.5 & -0.5 & 0 & 0 & 4.5 & 13.5 \\ \hline \boxed{0.5} & -3.5 & -0.5 & 0 & 1 & -1.5 & 0.5 \\ 0.5 & 1.5 & 0.5 & 1 & 0 & 0.5 & 1.5 \end{array}$$

Všimněme si, že vše is v pořádku: v nové tabulce we have opět standardní bázi (columns 5 a 4), nad ní we have v nultém řádku nuly, a numbers v nejvíce pravém sloupci jsou nezáporná. V políčku vpravo nahoře máme aktuální hodnotu kritéria, -13.5.

Další krok is zde:

$$\begin{array}{cccccc|c} 0 & -5 & -2 & 0 & 3 & 0 & 15 \\ \hline 1 & -7 & -1 & 0 & 2 & -3 & 1 \\ 0 & \boxed{5} & 1 & 1 & -1 & 2 & 1 \end{array}$$

Další krok:

$$\begin{array}{cccccc|c} 0 & 0 & -1 & 1 & 2 & 2 & 16 \\ \hline 1 & 0 & 0.4 & 1.4 & 0.6 & -0.2 & 2.4 \\ 0 & 1 & \boxed{0.2} & 0.2 & -0.2 & 0.4 & 0.2 \end{array}$$

Další krok:

$$\begin{array}{cccccc|c} 0 & 5 & 0 & 2 & 1 & 4 & 17 \\ \hline 1 & -2 & 0 & 1 & 1 & -1 & 2 \\ 0 & 5 & 1 & 1 & -1 & 2 & 1 \end{array}$$

Protože všechna numbers v účelovém řádku jsou nezáporná, končíme. Původní LP má optimální řešení, které má hodnotu -17 a nastává v bodě $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6) = (2, 0, 1, 0, 0, 0)$.

Pro kontrolu si dosaďte řešení do původního zadání a zkontrolujte, že řešení is přípustné a že hodnota kritéria is -17. □

Example 13.7. Vyřešte simplexovou metodou:

$$\begin{aligned} \min \quad & -2x_1 + 6x_2 + x_3 \\ \text{za podmínek} \quad & -x_1 - x_2 - x_3 + x_4 = 2 \\ & 2x_1 - x_2 - 2x_3 + x_5 = 1 \\ & x_1, x_2, x_3, x_4, x_5 \geq 0 \end{aligned}$$

Výchozí tabulka je

$$\begin{array}{ccccc|c} -2 & 6 & 1 & 0 & 0 & 0 \\ \hline -1 & -1 & -1 & 1 & 0 & 2 \\ \boxed{2} & -1 & -2 & 0 & 1 & 1 \end{array}$$

Druhá tabulka is zde:

0	5	−1	0	1	1
0	−1.5	−2	1	0.5	2.5
1	−0.5	−1	0	0.5	0.5

Podle nultého řádku by další pivot měl být ve třetím sloupci. Ale v něm jsou numbers v prvém a druhém řádku záporná. Tedy úloha is neomezená (to bylo ostatně patrné hned ze zadání). V nové tabulce je vidět, že můžeme zvětšovat x_3 libovolně, což bude kompenzováno příslušným nárůstem x_1 a x_4. Jelikož x_1 a x_4 nejsou v kritériu, jejich změny se na něm neprojeví a jediný vliv na kritérium bude mít x_3, které ho bude libovolně zmenšovat. □

13.3 Inicializace algoritmu

Na začátku základního simplexového algoritmu musí být úloha zadána ve tvaru

$$\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \,\}, \tag{13.7}$$

kde matrix $\mathbf{A}$ obsahuje standardní bázi a $\mathbf{b} \geq \mathbf{0}$. Pokud toto není splněno, nemůžeme základní algoritmus spustit. Ukážeme, jak lze každou úlohu převést na tento tvar.

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 $\mathbf{b} \geq \mathbf{0}$, převod is snadný: přidáme slackové proměnné $\mathbf{u} \geq \mathbf{0}$ a omezení převedeme na tvar $\mathbf{A}\mathbf{x} + \mathbf{u} = \mathbf{b}$. Úloha tedy bude mít simplexovou tabulku

$$\left[\begin{array}{cc|c} \mathbf{c}^T & \mathbf{0} & 0 \\ \hline \mathbf{A} & \mathbf{I} & \mathbf{b} \end{array}\right],$$

ve které columns příslušné proměnným $\mathbf{u}$ tvoří standardní bázi.

Example 13.8. 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 jsou kroky algoritmu:

-3	-1	-3	0	0	0	0
$\boxed{2}$	1	1	1	0	0	2
1	2	3	0	1	0	5
2	2	1	0	0	1	6

0	0.5	-1.5	1.5	0	0	3
1	0.5	0.5	0.5	0	0	1
0	1.5	$\boxed{2.5}$	-0.5	1	0	4
0	1	0	-1	0	1	4

0	1.4	0	1.2	0.6	0	5.4
1	0.2	0	0.6	-0.2	0	0.2
0	0.6	1	-0.2	0.4	0	1.6
0	1	0	-1	0	1	4

Úloha má optimální řešení s hodnotou -5.4 v bodě $\mathbf{x} = (x_1, x_2, x_3) = (0.2, 0, 1.6)$ (ověřte v původním zadání!). Hodnota slackových proměnných is $(u_1, u_2, u_3) = (0, 0, 4)$. □

Pokud jsou naše omezení zadána v obecném tvaru, operacemi z §12.3 is lze vždy převést do tvaru (13.7). Vynásobením některých řádků záporným číslem vždy zajistíme $\mathbf{b} \geq \mathbf{0}$. matrix $\mathbf{A}$ ale nemusí obsahovat standardní bázi. we have dokonce vážnější problém: není vůbec jasné, zda úloha (13.7) is přípustná. V tomto případě nejdříve vyřešíme *pomocnou úlohu* LP, která najde *nějaké* (obecně ne optimální) přípustné řešení. Z něj Then lze získat kýženou standardní bázi. Pomocná úloha je

$$\min\{\, \mathbf{1}^T\mathbf{u} \mid \mathbf{A}\mathbf{x} + \mathbf{u} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0},\ \mathbf{u} \geq \mathbf{0} \,\} \tag{13.8}$$

a má simplexovou tabulku

$$\left[\begin{array}{cc|c} \mathbf{0} & \mathbf{1}^T & 0 \\ \hline \mathbf{A} & \mathbf{I} & \mathbf{b} \end{array}\right].$$

Pro libovolné $\mathbf{u} \geq \mathbf{0}$ is $\mathbf{1}^T\mathbf{u} \geq 0$, přičemž $\mathbf{1}^T\mathbf{u} = 0$ právě tehdy, když $\mathbf{u} = \mathbf{0}$. Tedy (promyslete!) úloha (13.7) is přípustná právě tehdy, je-li optimální hodnota úlohy (13.8) rovna 0. Na počátku tvoří columns příslušné proměnným $\mathbf{u}$ standardní bázi, lze tedy na ní pustit základní simplexový algoritmus. Ten může skončit dvěma způsoby:

- Pokud is optimum větší než 0, Then úloha (13.7) je nepřípustná.
- Pokud is optimum rovno 0, Then úloha (13.7) je přípustná. Mohou dále nastat dva případy:
  - Pokud není optimální řešení $(\mathbf{x}, \mathbf{u})$ úlohy (13.8) degenerované, po skončení simplexového algoritmu nemůže být žádná bázová proměnná nulová. Protože $\mathbf{u} = \mathbf{0}$, proměnné $\mathbf{u}$ budou nutně nebázové. Tedy mezi sloupci příslušnými proměnným $\mathbf{x}$ bude existovat standardní báze.
  - Pokud is optimální řešení $(\mathbf{x}, \mathbf{u})$ úlohy (13.8) degenerované, některé z proměnných $\mathbf{u}$ mohou být na konci algoritmu bázové. Then is nutno udělat dodatečné úpravy kolem pivotů ve sloupcích příslušných bázovým proměnným $\mathbf{u}$, abychom tyto proměnné dostali z báze ven.

Nalezení nějakého přípustného řešení v pomocné úloze (13.8) se nazývá **první fáze** a řešení původní úlohy Then **druhá fáze** algoritmu, mluvíme tedy o **dvoufázové simplexové metodě**.

Example 13.9. Řešte

$$\begin{aligned} \min \quad & -20x_1 - 30x_2 - 40x_3 \\ \text{za podmínek} \quad & 3x_1 + 2x_2 + x_3 = 10 \\ & x_1 + 2x_2 + 2x_3 = 15 \\ & x_1, x_2, x_3 \geq 0 \end{aligned}$$

we have 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

$$\begin{aligned} \min \quad & u_1 + u_2 \\ \text{za podmínek} \quad & 3x_1 + 2x_2 + x_3 + u_1 = 10 \\ & x_1 + 2x_2 + 2x_3 + u_2 = 15 \\ & x_1, x_2, x_3, u_1, u_2 \geq 0 \end{aligned}$$

s tabulkou

0	0	0	1	1	0
3	2	1	1	0	10
1	2	2	0	1	15

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:

−4	−4	−3	0	0	−25
3	2	1	1	0	10
1	2	2	0	1	15

2	0	−1	2	0	−5
1.5	1	0.5	0.5	0	5
−2	0	1	−1	1	5

0	0	0	1	1	0
2.5	1	0	1	−0.5	2.5
−2	0	1	−1	1	5

Optimum is rovno 0, tedy původní úloha is 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

−20	−30	−40	0
2.5	1	0	2.5
−2	0	1	5

□

13.4 Exercises

13.1. V tabulce

$$[\,\mathbf{A}\,|\,\mathbf{b}\,] = \left[\begin{array}{cccccccc|c} 0 & 2 & 6 & 1 & 0 & -4 & 3 & 0 & 4 \\ 1 & 1 & -3 & 0 & 0 & 2 & 3 & 0 & 3 \\ 0 & -1 & 1 & 0 & 1 & -2 & -3 & 0 & 1 \\ 0 & -2 & 2 & 0 & 0 & 2 & -1 & 1 & 1 \end{array}\right]$$

označte všechny takové pivoty, že ekvivalentní úprava kolem nich povede k přípustnému bázovému řešení.

13.2. Zapište linear program

$$\begin{array}{rrrrrrrrrl} \min & -x_1 & & & & - & x_4 & - 3x_5 & & \\ \text{za podmínek} & 2x_1 & & & & + & x_4 + & x_5 + x_6 & = & 2 \\ & -x_1 + x_2 & & & & + 2x_4 + & 3x_5 & & = & 4 \\ & 2x_1 & & + x_3 + 2x_4 - & & x_5 & & & = & 6 \\ & & & & & & x_1, x_2, x_3, x_4, x_5, x_6 & & \geq & 0 \end{array}$$

do simplexové tabulky. Předpokládejte, že aktuální báze je tvořena sloupci $2, 3, 6$ a hodnota kritéria v aktuálním bázovém řešení je nula.

a) Jaké is aktuální bázové řešení?
b) is toto bázové řešení přípustné či degenerované?
c) Pokud is to možné, udělejte jeden krok simplexového algoritmu. Pokud to možné není, vysvětlete proč.

13.3. Vyřešte simplexovou metodou:

$$\begin{array}{rrrrrl} \max & 2x_1 - & x_2 - 3x_3 & & & \\ \text{za podmínek} & -2x_1 - & x_2 + & x_3 & \leq & 2 \\ & -x_1 + & 2x_2 - 3x_3 & & \leq & 5 \\ & -2x_1 - & 4x_2 + & x_3 & \leq & 6 \\ & & & x_1, x_2, x_3 & \geq & 0 \end{array}$$

13.4. Vyřešte simplexovou metodou (navzdory tomu, že lze řešit úvahou):

$$\begin{array}{rl} \max & 6x_1 + 9x_2 + 5x_3 + 9x_4 \\ \text{za podmínek} & x_1 + x_2 + x_3 + x_4 = 1 \\ & x_1, x_2, x_3, x_4 \geq 0 \end{array}$$

13.5. Úloha (13.4) má více než jedno optimální řešení. Jak se to projeví v simplexové tabulce? Můžeme udělat výčet všech optimálních bázových řešení?

13.6. Následující úlohu vyřešte nejdříve graficky, Then ji upravte do podoby vhodné pro simplexovou metodu a vyřešte simplexovou metodou. Použijte doufázovou metodu.

$$\begin{array}{rrrrl} \max & 3x_1 - 4x_2 & & & \\ \text{za podmínek} & -2x_1 - 5x_2 & \leq & 10 \\ & 3x_1 + x_2 & \leq & 3 \\ & -2x_1 + x_2 & \leq & -2 \\ & x_1 & \geq & 0 \\ & & x_2 & \leq & -1 \end{array}$$

Řešení: Optimum is $(x_1, x_2) = (25, -36)/13$.

Chapter 14

Duality in Linear Programing

Ke každé úloze LP lze sestavit podle dále popsaného postupu jinou úlohu LP. Novou úlohu nazýváme **duální**, původní úlohu nazýváme **primární** či **přímou**. Konstrukce is symetrická: duál duálu is původní úloha. Tedy má smysl říkat, že primární a duální úloha jsou *navzájem* duální. Dvojice duálních úloh is svázána zajímavými vztahy.

14.1 Konstrukce duální úlohy

K úloze LP v obecném tvaru (viz §12.3) se duální úloha získá dle tohoto postupu:

$$
\begin{array}{rlrll}
\min & \sum_{j\in J} c_j x_j & \max & \sum_{i\in I} y_i b_i & \\
\text{za podm.} & \sum_{j\in J} a_{ij}x_j = b_i & \text{za podm.} & y_i \in \mathbb{R}, & i \in I_0 \\
& \sum_{j\in J} a_{ij}x_j \geq b_i & & y_i \geq 0\,, & i \in I_+ \\
& \sum_{j\in J} a_{ij}x_j \leq b_i & & y_i \leq 0\,, & i \in I_- \\
& x_j \in \mathbb{R} & & \sum_{i\in I} y_i a_{ij} = c_j, & j \in J_0 \\
& x_j \geq 0 & & \sum_{i\in I} y_i a_{ij} \leq c_j, & j \in J_+ \\
& x_j \leq 0 & & \sum_{i\in I} y_i a_{ij} \geq c_j, & j \in J_-
\end{array}
$$

V levém sloupci is primární úloha, v prostředním sloupci is z ní vytvořená duální úloha. V pravém sloupci jsou set indexů pro obě úlohy: $I = \{1,\dots,m\} = I_0 \cup I_+ \cup I_-$ is indexová set primárních omezení a duálních proměnných, $J = \{1,\dots,n\} = J_0 \cup J_+ \cup J_-$ is indexová set primárních proměnných a duálních omezení.

Všimněte si, že 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 odpovídá j-tému duálnímu omezení $\sum_i a_{ij}x_j \leq c_j$.

Pro speciální tvary LP se dvojice duálních úloh přehledněji napíše v maticové formě. Např. pro $I_0 = I_- = J_0 = J_- = \emptyset$ obdržíme

$$
\begin{array}{rlrll}
\min & \mathbf{c}^T\mathbf{x} & \max & \mathbf{y}^T\mathbf{b} & \\
\text{za podm.} & \mathbf{A}\mathbf{x} \geq \mathbf{b} & \text{za podm.} & \mathbf{y} \geq \mathbf{0} & \\
& \mathbf{x} \geq \mathbf{0} & & \mathbf{y}^T\mathbf{A} \leq \mathbf{c}^T &
\end{array}
\tag{14.1}
$$

14.2 Věty o dualitě

Následující věty platí pro obecný tvar LP, ale důkazy uděláme pouze pro speciální tvar (14.1).

Theorem 14.1 (o slabé dualitě). *Let* $\mathbf{x}$ *is přípustné primární řešení a* $\mathbf{y}$ *přípustné duální řešení. Then* $\mathbf{c}^T\mathbf{x} \geq \mathbf{y}^T\mathbf{b}$.

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 $\mathbf{y}^T\mathbf{A}\mathbf{x} \geq \mathbf{y}^T\mathbf{b}$. Napíšeme-li tyto dvě nerovnosti za sebe, máme

$$\mathbf{c}^T\mathbf{x} \geq \mathbf{y}^T\mathbf{A}\mathbf{x} \geq \mathbf{y}^T\mathbf{b}. \tag{14.2}$$

□

Theorem 14.2 (o komplementaritě). *Let* $\mathbf{x}$ *is přípustné primární řešení a* $\mathbf{y}$ *přípustné duální řešení. Then* $\mathbf{c}^T\mathbf{x} = \mathbf{y}^T\mathbf{b}$ *právě tehdy, když zároveň platí tyto dvě podmínky komplementarity:*

$$\textit{Pro každé } i \in I \textit{ platí } y_i = 0 \textit{ nebo } \sum_{j \in J} a_{ij}x_j = b_i. \tag{14.3a}$$

$$\textit{Pro každé } j \in J \textit{ platí } x_j = 0 \textit{ nebo } \sum_{i \in I} y_i a_{ij} = c_j. \tag{14.3b}$$

'Nebo' is zde užito v nevylučovacím smyslu, i.e., mohou nastat obě možnosti současně.

Proof. Klíčové is si uvědomit (rozmyslete!), že pro libovolné vectors $\mathbf{u}, \mathbf{v} \geq \mathbf{0}$ platí

$$\forall i\,(u_i = 0 \text{ nebo } v_i = 0) \quad \Longleftrightarrow \quad \forall i\,(u_i v_i = 0) \quad \Longleftrightarrow \quad \mathbf{u}^T\mathbf{v} = 0.$$

Tedy podmínky (14.3) is možno psát jako

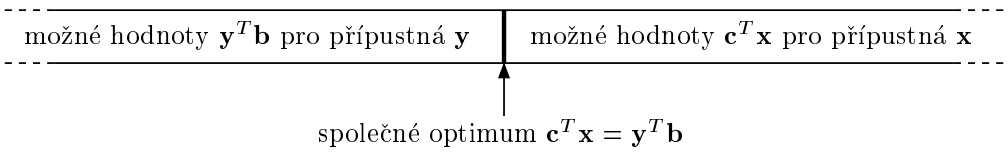
$$\mathbf{y}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) = 0 \tag{14.4a}$$

$$(\mathbf{c}^T - \mathbf{y}^T\mathbf{A})\mathbf{x} = 0. \tag{14.4b}$$

Tvrzení $\mathbf{c}^T\mathbf{x} = \mathbf{y}^T\mathbf{b}$ is ekvivalentní tomu, že obě nerovnosti ve vztahu (14.2) jsou rovnostmi. Then $\mathbf{c}^T\mathbf{x} = \mathbf{y}^T\mathbf{A}\mathbf{x}$ is ekvivalentní (14.4b) a $\mathbf{y}^T\mathbf{A}\mathbf{x} = \mathbf{y}^T\mathbf{b}$ je ekvivalentní (14.4a). □

Theorem 14.3 (o silné dualitě). *Primární úloha má optimální řešení právě tehdy, když má duální úloha optimální řešení. Mají-li obě úlohy optimální řešení, platí* $\mathbf{c}^T\mathbf{x} = \mathbf{y}^T\mathbf{b}$, *kde* $\mathbf{x}$ *a* $\mathbf{y}$ *značí tato optimální řešení.*

Důkaz věty o silné dualitě is obtížný (a vynecháme jej). To není překvapivé, neboť tato věta is jedním z nejhlubších výsledků v linearm programování. Věty o slabé a silné dualitě mají jasnou interpretaci: pro přípustná $\mathbf{x}$ a $\mathbf{y}$ není hodnota duálního kritéria nikdy větší než hodnota primárního kritéria, a tyto hodnoty se potkají ve společném optimu. Viz obrázek:



Dobře si uvědomte, že věta o komplementaritě is slabší než věta o silné dualitě, protože věta o komplementaritě neříká, že rovnost $\mathbf{c}^T\mathbf{x} = \mathbf{y}^T\mathbf{b}$ vůbec někdy nastane. Uveďme ještě jeden jednoduchý důsledek slabé duality, který is opět slabší než silná dualita.

Corollary 14.4. *Let* $\mathbf{x}$ *is přípustné primární řešení a* $\mathbf{y}$ *is přípustné duální řešení. Let* $\mathbf{c}^T\mathbf{x} = \mathbf{y}^T\mathbf{b}$. *Potom* $\mathbf{x}$ *a* $\mathbf{y}$ *jsou zároveň optimální řešení.*

Proof. Pro libovolné primární přípustné řešení $\mathbf{x}'$ plyne z věty o slabé dualitě $\mathbf{y}^T\mathbf{b} \leq \mathbf{c}^T\mathbf{x}'$. Z předpokladu máme $\mathbf{c}^T\mathbf{x} = \mathbf{y}^T\mathbf{b}$. Z toho plyne $\mathbf{c}^T\mathbf{x} \leq \mathbf{c}^T\mathbf{x}'$. Jelikož toto platí pro *každé* přípustné $\mathbf{x}'$, řešení $\mathbf{x}$ musí být optimální.

Optimalita $\mathbf{y}$ se dokáže symetricky. □

Example 14.1. consider dvojici navzájem duálních úloh LP:

$$\begin{array}{rrrrrrrrrrr}
\min & 2x_1 & + 5x_2 & + 6x_3 & = \mathbf{5.4} & \qquad \max & 3y_1 & + y_2 & + 3y_3 & - y_4 & = \mathbf{5.4} \\
\mathbf{3} = & 2x_1 & + x_2 & + 2x_3 & \geq 3 & \mathbf{0.2} = & y_1 & & & & \geq 0 \\
\mathbf{2.4} = & x_1 & + 2x_2 & + 2x_3 & \geq 1 & \mathbf{0} = & & y_2 & & & \geq 0 \\
\mathbf{3} = & x_1 & + 3x_2 & + x_3 & \geq 3 & \mathbf{1.6} = & & & y_3 & & \geq 0 \\
\mathbf{-0.6} = & -x_1 & + x_2 & - 2x_3 & \geq -1 & \mathbf{0} = & & & & y_4 & \geq 0 \\
\mathbf{1.2} = & x_1 & & & \geq 0 & \mathbf{2} = & 2y_1 & + y_2 & + y_3 & - y_4 & \leq 2 \\
\mathbf{0.6} = & & x_2 & & \geq 0 & \mathbf{5} = & y_1 & + 2y_2 & + 3y_3 & + y_4 & \leq 5 \\
\mathbf{0} = & & & x_3 & \geq 0 & \mathbf{2} = & 2y_1 & + 2y_2 & + y_3 & - 2y_4 & \leq 6
\end{array}$$

Spočetli jsme optimální řešení obou úloh a dosadili tato řešení do kritérií 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 kritérií v optimech jsou napsané tučně před/za rovnítky. Vidíme, že obě optima se sobě rovnají, jak to musí být podle věty o silné dualitě. Vezmeme-li libovolný řádek (kromě účelového), is na něm alespoň jedno z obou omezení aktivní (i.e., platí s rovností). Např. ve druhém řádku is 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 rovnosti zároveň neaktivní (mohou být obě ale zároveň aktivní, což zde nenastává, ale může to nastat v případě degenerace). □

Zopakujme (viz §12), že pro každou úlohu LP mohou nastat 3 možnosti: úloha má optimální řešení, úloha is neomezená, úloha je nepřípustná.

Theorem 14.5. *Z devíti možností pro dvojici duálních úloh se realizují tyto:*

primární/duální	*má optimum*	*neomezená*	*nepřípustná*
má optimum	*ano*	*ne*	*ne*
neomezená	*ne*	*ne*	*ano*
nepřípustná	*ne*	*ano*	*ano*

Proof. Snadno najdeme příklady dvojic duálních úloh, které realizují povolené kombinace. Zbývá dokázat, že zakázané kombinace nemohou nastat.

Čtyři zakázané kombinace v prvním řádku a prvním sloupci plynou z první části věty o silné dualitě (primární úloha má optimum *právě tehdy*, když duální úloha má optimum).

Pokud is primární [duální] úloha neomezená, její optimum (přesněji infimum [supremum]) is $-\infty$ $[+\infty]$. Věta o slabé dualitě zakazuje, aby úlohy byly zároveň neomezené, protože Then bychom měli $-\infty \geq +\infty$. □

Předložíme-li přípustná primární a duální řešení taková, že se kritéria rovnají, dokázali jsme optimalitu obou úloh. Takové dvojici řešení se říká *certifikát optimality*. Pro velké úlohy to může být nejsnadnější důkaz optimality.

Někdy lze spočítat optimální duální řešení levně z optimálního primárního řešení, jak ukazuje následující příklad.

Example 14.2. Je dána primární úloha z Příkladu 14.1. Dokažte 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 není zadáno duální řešení $\mathbf{y}$)

Optimalitu daného $\mathbf{x}$ zkusíme dokázat pomocí věty o komplementaritě. Předpokládejme, že $\mathbf{y}$ (které zatím neznáme) is optimální řešení duální úlohy. Protože jsou druhé a čtvrté primární omezení neaktivní (neplatí v nich rovnost ale pouze nerovnost), z komplementarity musí být $y_2 = y_4 = 0$. Protože $x_1 > 0$ a $x_2 > 0$, z komplementarity musí být první a druhé duální omezení aktivní. we have tedy soustavu linearch rovnic

$$\begin{aligned} 2y_1 + \ \ y_3 &= 2 \\ y_1 + 3y_3 &= 5 \end{aligned} \tag{14.5}$$

která má jediné řešení $(y_1, y_3) = (0.2, 1.6)$. Tedy $\mathbf{y} = (0.2, 0, 1.6, 0)$. Toto duální řešení is přípustné (i.e., splňuje všechna duální omezení). Protože se hodnota primárního kritéria v bodě $\mathbf{x}$ rovná hodnotě duálního kritéria v bodě $\mathbf{y}$, musejí být $\mathbf{x}$ a $\mathbf{y}$ optimální řešení.

Zdůrazněme, že tento postup nemusí vést vždy k cíli. Když bude mít duální úloha více než jedno řešení, bude mít soustava (14.5) nekonečně mnoho (affine subspace) řešení. Mezi nimi sice budou přípustná duální řešení, ale k jejich nalezení budeme potřebovat řešit soustavu rovnic *a nerovnic* (což už není snadné). □

14.3 Stínové ceny

Theorem 14.6 (o stínových cenách). *Označme*

$$f(\mathbf{b}) = \min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \,\} = \max\{\, \mathbf{y}^T\mathbf{b} \mid \mathbf{y}^T\mathbf{A} \leq \mathbf{c}^T,\ \mathbf{y} \geq \mathbf{0} \,\}$$

optimální hodnotu dvojice duálních úloh jako funkci vectoru $\mathbf{b}$. *Jestliže má duální úloha pro dané* $\mathbf{b}$ *jediné optimální řešení* $\mathbf{y}^*$, *Then is funkce* f *v bodě* $\mathbf{b}$ *diferencovatelná a platí* $f'(\mathbf{b}) = \mathbf{y}^{*T}$.

Proof. Jelikož is optimální řešení $\mathbf{y}^*$ jediné, nabývá se ve vrcholu polyedru přípustných řešení $\{\, \mathbf{y} \in \mathbb{R}^m \mid \mathbf{y}^T\mathbf{A} \leq \mathbf{c}^T,\ \mathbf{y} \geq \mathbf{0} \,\}$. Změníme-li nepatrně $\mathbf{b}$, set duálních optimálních řešení se nezmění, neboli budeme mít stále jediné optimální řešení ve stejném vrcholu $\mathbf{y}^*$ (tento argument není zcela rigorózní, ale geometricky is dosti názorný). Tedy v malém okolí bodu $\mathbf{b}$ is hodnota optima jednoduše rovna $\mathbf{y}^{*T}\mathbf{b}$. Derivací toho získáme $f'(\mathbf{b}) = \mathbf{y}^{*T}$. □

Uvědomte si nutnost předpokladu o jednoznačnosti optimálního řešení. Kdyby set duálních optimálních řešení byla ne jediný vrchol, ale stěna vyšší dimenze, po infinitezimální změně účelového vectoru $\mathbf{b}$ by se optimální stěna mohla stát vrcholem a funkce f by tedy v bodě $\mathbf{b}$ nebyla diferencovatelná. Předpoklad o jednoznačnosti řešení lze vypustit, ale Then by věta byla složitější.

Protože $\mathbf{b}$ is zároveň vector 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 (12.5) (pozor, liší se obrácenou nerovností v omezení), Then 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 is ve shodě s větou o komplementaritě. Pokud $y_i^* = 0$, is $\mathbf{a}_i^T\mathbf{x} < b_i$, tedy malá změna b_i nemá na optimum vliv.

Example 14.3. V Příkladu 14.1 is $y_1 = 0.2$ is stínová cena prvního primárního omezení $2x_1 + x_2 + 2x_3 \geq 3$. 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 změní jeho *hodnotu* $\mathbf{y}^{*T}\mathbf{b}$. Podle silné duality hodnota primárního optima musí být rovna hodnotě duálního optima (argument $\mathbf{x}^*$ primárního optima se nějak změní, my ale nepotřebujeme vědět, jak). Dvojice úloh tedy bude vypadat takto:

$$
\begin{array}{lrcl}
\min & 2x_1 + 5x_2 + 6x_3 = \mathbf{5.402} & & \\
& 2x_1 + x_2 + 2x_3 \geq & 3.01 & \\
& x_1 + 2x_2 + 2x_3 \geq & 1 & \\
& x_1 + 3x_2 + x_3 \geq & 3 & \\
& -x_1 + x_2 - 2x_3 \geq & -1 & \\
& x_1 \geq & 0 & \\
& x_2 \geq & 0 & \\
& x_3 \geq & 0 &
\end{array}
\qquad
\begin{array}{lrcl}
\max & 3.01y_1 + y_2 + 3y_3 - y_4 = \mathbf{5.402} & & \\
\mathbf{0.2} = & y_1 \geq & 0 & \\
\mathbf{0} = & y_2 \geq & 0 & \\
\mathbf{1.6} = & y_3 \geq & 0 & \\
\mathbf{0} = & y_4 \geq & 0 & \\
\mathbf{2} = & 2y_1 + y_2 + y_3 - y_4 \leq & 2 & \\
\mathbf{5} = & y_1 + 2y_2 + 3y_3 + y_4 \leq & 5 & \\
\mathbf{2} = & 2y_1 + 2y_2 + y_3 - 2y_4 \leq & 6 &
\end{array}
$$

Věta o stínových cenách říká, že v malém okolí bodu $\mathbf{b} = (3, 1, 3, -1)$, ve kterém se nemění optimální $\mathbf{y}^*$, bude $f(\mathbf{b}) = \mathbf{y}^{*T}\mathbf{b}$ a tedy

$$5.402 - 5.4 = \frac{\partial f(\mathbf{b})}{\partial b_1} h = y_1 h = 0.2 \cdot 0.01.$$

□

14.4 Příklady na konstrukci a interpretaci duálních úloh

Dualita umožňuje *vhled* do řešeného problému, často velmi netriviální. Dá se říci, že abychom jakoukoli úlohu (s fyzikální, ekonomickou či jinou interpretací) popsanou linearm programem porozuměli do hloubky, is třeba pochopit význam nejen primární úlohy, ale i duální úlohy a vět o dualitě.

Example 14.4. consider 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 $\mathbf{c} = (c_1, \ldots, c_n)$ is dáno a optimalizuje se přes $\mathbf{x} = (x_1, \ldots, x_n)$. Najděte elementární úvahou hodnotu optima, napište duální úlohu. Vysvětlete, co v dané úloze znamenají věty o silné dualitě a komplementaritě.

Optimální hodnota is $\min_{i=1}^n c_i$, tedy nejmenší z čísel c_i. Dosahuje se ve vectoru $\mathbf{x}$ jehož všechny složky jsou nulové kromě složek příslušných minimálnímu c_i. To is jasné, protože je nejvýhodnější soustředit všechnu 'váhu' rozdělení $\mathbf{x}$ do nejmenšího prvku. Pokud is více

minimálních prvků c_i, optimální vector $\mathbf{x}$ není dán jednoznačně. Např. pro $\mathbf{c} = (1, 3, 1, 2)$ budou optimálními řešeními vectors $\mathbf{x} = (x_1, 0, x_3, 0)$ pro všechna $x_1, x_3 \geq 0$ splňující $x_1 + x_3 = 1$.

Podle návodu na konstrukci duální úlohy dostaneme duál

$$\max\{\, y \in \mathbb{R} \mid y\mathbf{1} \leq \mathbf{c} \,\} = \max\{\, y \in \mathbb{R} \mid y \leq c_i,\ i = 1, \ldots, n \,\}.$$

Tato úloha má jasný význam: hledá se největší číslo y, které je menší než všechna numbers c_i. Takové číslo y se rovná minimu z čísel c_i.

Význam silné duality is jasný: hodnoty primárního i duálního optima jsou si rovny.

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 $x_i \geq 0$, $y \leq c_i$ is splňuje právě tehdy, když is splněna aspoň jedna z rovností $x_i = 0$, $y = c_i$. To znamená:

- Pokud is v duálu $y < c_i$, v primáru musí být $x_i = 0$. To is ale jasné, protože $y < c_i$ znamená, že c_i není nejmenší ze složek vectoru $\mathbf{c}$ a tudíž v primáru by byla hloupost mu přiřadit nenulovou váhu x_i.
- Obráceně, pokud is v primáru $x_i > 0$, musí být v duálu $y = c_i$. To is jasné, protože pokud jsme v primáru přiřadili číslu c_i nenulovou váhu, musí být nejmenší. □

Example 14.5. Medián čísel $\mathbf{b} = (b_1, \ldots, b_m) \in \mathbb{R}^m$ is řešením úlohy (viz §12.5)

$$\min_{x \in \mathbb{R}} \sum_{i=1}^{m} |x - b_i| \;=\; \min\{\, \mathbf{1}^T\mathbf{z} \mid \mathbf{z} \in \mathbb{R}^m,\ x \in \mathbb{R},\ -\mathbf{z} \leq \mathbf{1}x - \mathbf{b} \leq \mathbf{z} \,\}. \tag{14.6}$$

Najděme k této úloze duál a výsledek co možná nejvíce zjednodušme.

Primární a duální úlohu napíšeme ve tvaru (14.1), kde ale přejmenujeme názvy matic, aby nekolidovaly s (14.6):

$$\begin{array}{rl} \min & \mathbf{h}^T\mathbf{u} \\ \text{za podm.} & \mathbf{F}\mathbf{u} \geq \mathbf{g} \\ & \mathbf{u} \in \mathbb{R}^{1+m} \end{array} \qquad\qquad \begin{array}{rl} \max & \mathbf{v}^T\mathbf{g} \\ \text{za podm.} & \mathbf{v} \geq \mathbf{0} \\ & \mathbf{v}^T\mathbf{F} = \mathbf{h}^T \end{array}$$

matrix zvolíme tak, aby primární úloha odpovídala úloze (14.6), tedy (promyslete!)

$$\mathbf{F} = \begin{bmatrix} 1 & \mathbf{I} \\ -1 & \mathbf{I} \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} \mathbf{b} \\ -\mathbf{b} \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} 0 \\ \mathbf{1} \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} x \\ \mathbf{z} \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} \mathbf{p} \\ \mathbf{q} \end{bmatrix}.$$

vector duálních proměnných $\mathbf{v}$ jsme zároveň rozdělili na dva blocks $\mathbf{p}, \mathbf{q}$, odpovídající blockům matic $\mathbf{F}$ a $\mathbf{g}$. Vynásobním blockových matic přepíšeme duální úlohu do tvaru (ověřte na papíře!)

$$\max\{\, \mathbf{b}^T(\mathbf{p} - \mathbf{q}) \mid \mathbf{1}^T(\mathbf{p} - \mathbf{q}) = 0,\ \mathbf{p} + \mathbf{q} = \mathbf{1},\ \mathbf{p} \geq \mathbf{0},\ \mathbf{q} \geq \mathbf{0} \,\},$$

což ve skalárním tvaru lze psát jako

$$\max\left\{ \sum_{i=1}^{m} b_i(p_i - q_i) \;\middle|\; \sum_{i=1}^{m} (p_i - q_i) = 0,\ p_i + q_i = 1,\ p_i \geq 0,\ q_i \geq 0 \right\}. \tag{14.7}$$

Překvapivě, úlohu (14.7) lze dále zjednodušit chytrou substitucí

$$2p_i = 1 + t_i, \qquad 2q_i = 1 - t_i.$$

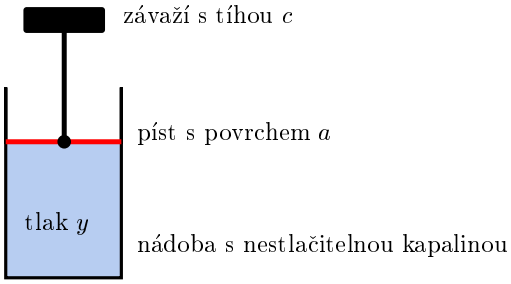
Po této substituci is $p_i - q_i = t_i$, podmínka $p_i + q_i = 1$ is splněna automaticky, podmínka $p_i \geq 0$ odpovídá $t_i \geq -1$, a podmínka $q_i \geq 0$ odpovídá $t_i \leq 1$ (vše promyslete!). Úloha (14.7) s novými proměnnými $\mathbf{t}$ má tedy tvar

$$\min\left\{ \sum_{i=1}^{m} b_i t_i \;\middle|\; \sum_{i=1}^{m} t_i = 0,\ -1 \leq t_i \leq 1 \right\} \;=\; \min\{\, \mathbf{b}^T\mathbf{t} \mid \mathbf{1}^T\mathbf{t} = 0,\ -\mathbf{1} \leq \mathbf{t} \leq \mathbf{1} \,\}. \tag{14.8}$$

Zde význam vět o dualitě není vůbec očividný! Např. na první pohled není ani trochu vidět, že optimální hodnoty úloh (14.6) a (14.8) jsou stejné (silná dualita). Jako nepovinné cvičení to chytrý student může zkusit odůvodnit. □

Example 14.6. (⋆) Uvažujme fyzikální systém ('analogový počítač'), který sestává z nádob s nestlačitelnou kapalinou uzavřených písty a ze závaží. K pochopení jeho činnosti budeme potřebovat tyto známé poučky z hydrostatiky:

- Objem kapaliny v uzavřené nádobě is při libovolném tlaku stejný.
- Tlak v kapalině is při rovnováze všude stejný.
- Let y is tlak v nádobě s kapalinou uzavřené pístem. Nechť píst má povrch a a působí na něj síla c (viz obrázek). Then $c = ay$.



Obrázek 14.1 ukazuje celý stroj, v němž

- a_{ij} = povrch svislého pístu v nádobě i spojeného se závažím j (pro $a_{ij} > 0$ is píst nahoře, pro $a_{ij} < 0$ is píst dole).
- x_j = výška závaží j (měřeno směrem dolů od roviny nulové výšky)
- b'_i = šířka mezery mezi tyčí a stěnou. Při $\mathbf{x} = \mathbf{0}$ platí $b'_i = b_i$.
- c_j = tíha závaží j
- Vodorovné písty mají povrch jednotkový.

Ze zachování objemu kapaliny v nádobě i plyne $b'_i = b_i - a_{i1}x_1 - \cdots - a_{in}x_n$. Protože vodorovné tyče nemohou projít stěnou, is vždy $b'_i \geq 0$. Tedy $\mathbf{b}' = \mathbf{b} - \mathbf{A}\mathbf{x} \geq \mathbf{0}$, tedy $\mathbf{A}\mathbf{x} \leq \mathbf{b}$.

Potenciální energie závaží j is $-c_j x_j$. Libovolný statický systém v rovnováze zaujme stav s nejnižší potenciální energií. Proto se závaží ustálí v takových výškách, že jejich celková potenciální energie bude minimální, neboli $c_1x_1 + \cdots + c_nx_n = \mathbf{c}^T\mathbf{x}$ bude maximální. Tedy stroj řeší linear program

$$\max\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b},\ \mathbf{x} \in \mathbb{R}^n \,\}.$$

Jeho duál je

$$\min\{\, \mathbf{y}^T\mathbf{b} \mid \mathbf{y}^T\mathbf{A} = \mathbf{c}^T,\ \mathbf{y} \geq \mathbf{0} \,\}.$$

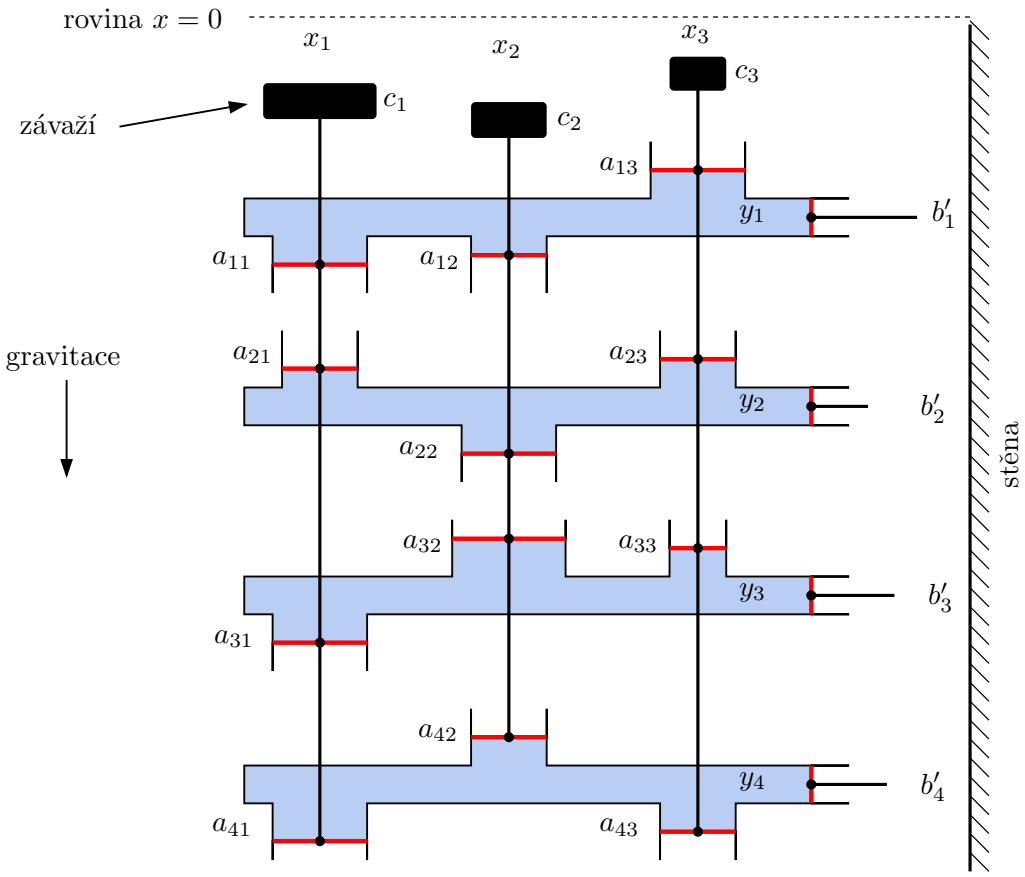


Figure 14.1: Hydraulický počítač řešící primární i duální úlohu LP.

Tento duální LP jsme našli *čistě formálně* podle postupu o konstrukci duální úlohy. To nám ale vůbec neříká, jaký má duál vztah k našemu stroji. Pokusme se tento vztah odhalit. Klíčové pro objevení tohoto vztahu is přiřadit duální proměnné y_i význam *tlaku* v nádobě i (všimněte si, že v primární úloze tlak vůbec nevystupuje). Teď dokážeme interpretovat duální LP a jeho vztah k primárnímu LP daný větami o dualitě (vynecháváme větu o slabé dualitě):

- Jelikož stěna působí silou vždy od sebe, musí být tlak v y_i v nádrži i nezáporný. To dá duální omezení $\mathbf{y} \geq \mathbf{0}$.
- Protože povrch vodorovných pístů is jednotkový, tlak y_i se rovná síle vodorovné tyče i na stěnu. Rovnováha sil pro svislou tyč j bude $a_{1j}y_1 + \cdots + a_{mj}y_m = c_j$, což is duální omezení $\mathbf{y}^T\mathbf{A} = \mathbf{c}^T$.
- Dle *věty o komplementaritě* v ustáleném stavu platí buď $b'_i = b_i - a_{i1}x_1 - \cdots - a_{in}x_n = 0$ nebo $y_i = 0$, pro každé i. Ale to is jasné, protože když se některá vodorovná tyč nedotýká stěny, musí být tlak v příslušné nádobě nulový.
- Dle *věty o silné dualitě* is v ustáleném stavu duální kritérium $\mathbf{y}^T\mathbf{b} = y_1b_1 + \cdots + y_mb_m$ minimální. Proč to tak je? Potenciální energie všech závaží is rovna práci, nutné na jejich vyzdvižení do roviny $x = 0$. Tato práce se dá vykonat buď přímo zdvihnutím závaží (což odpovídá primárnímu kritériu $\mathbf{c}^T\mathbf{x}$) nebo odtlačením vodorovných tyčí od stěny do vzdáleností b_i. Ukážeme, že druhý způsob odpovídá duálnímu kritériu. Zafixujeme-li všechny vodorovné tyče kromě jediné tyče i, při odtlačování tyče i se síla, kterou tlačíme na tyč, nemění (promyslete!). Tedy vykonáme práci y_ib_i. Když takto odtlačíme od stěny

postupně všechny tyče, vykonáme práci $\mathbf{y}^T\mathbf{b}$.

- Věta o *stínových cenách* říká, že se změnou b_i se optimum mění tím více, čím is větší tlak y_i. To is ale jasné, protože čím is větší tlak y_i, tím větší práce is třeba na odtlačení tyče od stěny do vzdálenosti b_i.

Zdůrazněme, že tyto úvahy *nedokazují* žádnou ze tří vět o dualitě. Předpokládáme totiž platnost fyzikálních zákonů, které ale nelze matematicky dokázat, lze is pouze experimentálně pozorovat. Skutečnost, že z chování stroje 'vyplývá' např. věta o silné dualitě, není matematický důkaz – ten is totiž čistou logickou dedukcí a žádné fyzikální zákony nepředpokládá.

Tím, že se nám podařilo pochopit význam duální úlohy ve stroji, jsme se o fyzice našeho stroje dozvěděli něco nového – tedy, že se dá podmínka rovnováhy formulovat pomocí tlaků v nádobách. Toho bychom si nejspíše nevšimli, kdybychom se nezabývali duální úlohou. □

14.5 Exercises

14.1. Ukažte pro dvojici úloh LP v §14.1, že duál duálu se rovná původní úloze. Musíte nejdříve duální úlohu (prostřední sloupec) vpravo převést do tvaru primární úlohy (první sloupec), i.e., např. musíte převést maximalizaci na minimalizaci.

14.2. Napište duální úlohy a podmínky komplementarity k následujícím úlohám. Výsledek 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 (12.8)
c) úloha (12.10)
d) všechny úlohy ze Cvičení 12.4
e) úloha LP vzniklá ve Cvičení 12.8
f) Příklad 12.9.

(⋆) Dále pro každou úlohu zkuste interpretovat větu o silné dualitě (i.e., úvahou odvoďte, jaká is optimální hodnota duální úlohy, a tato musí být stejná jako optimální hodnota primární úlohy) a podmínky komplementarity, podobně jako v Příkladu 14.4. Interpretace vět o dualitě is obecně velmi netriviální, takže většinou se vám to nepodaří – ale aspoň to zkuste.

14.3. Dokažte bez užití algoritmu na řešení LP, že $\mathbf{x} = (1, 1, 1, 1)$ je optimální řešení úlohy

$$\begin{array}{rl} \min & \begin{bmatrix} 47 & 93 & 17 & -93 \end{bmatrix} \mathbf{x} \\ \text{za podm.} & \begin{bmatrix} -1 & -6 & 1 & 3 \\ -1 & -2 & 7 & 1 \\ 0 & 3 & -10 & -1 \\ -6 & -11 & -2 & 12 \\ 1 & 6 & -1 & -3 \end{bmatrix} \mathbf{x} \leq \begin{bmatrix} -3 \\ 5 \\ -8 \\ -7 \\ 4 \end{bmatrix} \end{array}$$

Chapter 15

Convex Optimisation Problems

15.1 Třídy optimalizačních úloh

Optimalizační úlohy ve tvaru (11.7) se taxonomizují podle druhu funkcí f, g_i, h_i. Pro každou třídu existují specializované algoritmy schopné najít lokální minimum[1].

linear programování (LP)

V *linearm programování* jsou všechny funkce f, g_i, h_i affine. 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.

Kvadratické programování (QP)

V *kvadratickém programování* jsou funkce g_i, h_i affine a funkce f is kvadratická konvexní, tedy $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$, kde matrix $\mathbf{A}$ is pozitivně semidefinitní.

Example 15.1. 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 $\mathbf{c} \leq \mathbf{x} \leq \mathbf{d}$, i.e., každá proměnná x_j musí být v intervalu $[c_j, d_j]$. To vede na konvexní QP s omezeními. □

Example 15.2. Hledání řešení linear soustavy s nejmenší normou vede na úlohu $\min\{\, \mathbf{x}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b} \,\}$, což is konvexní QP s omezeními. □

Example 15.3. Chceme spočítat vzdálenost polyedrů

$$P_1 = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}_1\mathbf{x} \leq \mathbf{b}_1 \,\}, \qquad P_2 = \{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}_2\mathbf{x} \leq \mathbf{b}_2 \,\}$$

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

$$\min\{\, \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 \mid \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n,\ \mathbf{A}_1\mathbf{x}_1 \leq \mathbf{b}_1,\ \mathbf{A}_2\mathbf{x}_2 \leq \mathbf{b}_2 \,\}.$$

Pokud se polyedry protínají, jejich vzdálenost is nula. Pokud je aspoň jeden polyedr prázdný, úloha is nepřípustná □

[1] Viz např. http://www.neos-guide.org.

Kvadratické programování s kvadratickými omezeními (QCQP)

Obecnější variantou is *kvadratické programování s kvadratickými omezeními* (QCQP, *quadratically constrained quadratic programming*), kde všechny funkce f, g_i, h_i jsou kvadratické. Úloha is konvexní jen tehdy, když kvadratické funkce f, g_i jsou konvexní (i.e., s pozitivně semidefinitní maticí) a funkce h_i jsou affine.

Example 15.4. Na závěr ukažme jednoduchou konvexní úlohu, ktera na první pohled nespadá do žádné z uvedených tříd. Jsou dány body $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$ a chceme minimalizovat funkci

$$f(\mathbf{x}) = \sum_{i=1}^{m} \|\mathbf{a}_i - \mathbf{x}\|_2 \tag{15.1}$$

přes $\mathbf{x} \in \mathbb{R}^n$. Řešení této úlohy is známo jako *geometrický medián*. Pro $n = 1$ se funkce redukuje na $f(x) = \sum_{i=1}^{m} |x - a_i|$, jejímž minimem is obyčejný medián.

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 is $\mathbf{x}$. Hodnota $f(\mathbf{x})$ je potenciální energie soustavy a ustálený stav odpovídá minimu $f(\mathbf{x})$. □

15.2 Příklady nekonvexních úloh

Example 15.5. Řešení homogeneous linear soustavy ve smyslu nejmenších čtverců vede na úlohu

$$\min\{\, \|\mathbf{A}\mathbf{x}\|_2^2 \mid \mathbf{x}^T\mathbf{x} = 1 \,\}. \tag{15.2}$$

To is 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 set $\{\, \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x}^T\mathbf{x} = 1 \,\}$ 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 12.9. To ale nelze, nebot'

$$\min\{\, \|\mathbf{A}\mathbf{x}\|_2^2 \mid \mathbf{x}^T\mathbf{x} = 1 \,\} \neq \min\{\, \|\mathbf{A}\mathbf{x}\|_2^2 \mid \mathbf{x}^T\mathbf{x} \leq 1 \,\} = 0.$$

My ale víme, že úlohu (15.2) lze řešit pomocí SVD, protože hledáme nadrovinu s normálovým vectorem $\mathbf{x}$, která minimalizuje součet čtverců kolmých vzdáleností řádků $\mathbf{a}_1, \ldots, \mathbf{a}_m$ matrix $\mathbf{A}$ k nadrovině. □

V tomto příkladě měla nekonvexní úloha jediné lokální minimum. To je ale řídká výjimka – v naprosté většině mají nekonvexní úlohy mnoho lokálních extrémů.

Example 15.6. Uveďme příklad, na kterém bude na první pohled vidět, že nekonvexní úloha může mít velmi mnoho lokálních minim. Řešme úlohu

$$\min\{\, -\mathbf{x}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ -\mathbf{1} \leq \mathbf{x} \leq \mathbf{1} \,\}. \tag{15.3}$$

set přípustných řešení is hyperkrychle, $X = [-1, 1]^n$. Účelová funkce $f(\mathbf{x}) = -\mathbf{x}^T\mathbf{x}$ is konkávní. is očividné, že funkce f má na množině X lokální 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 linearch nerovnic může mít exponenciální počet vrcholů (viz §12.1.2).

V tomto případě jsou lokální minima všechna stejná, tedy úlohu snadno vyřešíme. Ale již mírnou modifikací úlohy se stane nalezení globálního optima prakticky nemožné. Uvažujme úlohu

$$\min\{\, \mathbf{x}^T\mathbf{Q}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ -\mathbf{1} \leq \mathbf{x} \leq \mathbf{1} \,\}. \tag{15.4}$$

Je jasné, že pro $\mathbf{Q} = -\mathbf{I}$ dostaneme úlohu (15.3). Je známo, že neexistuje algoritmus, který by pro libovolnou (tedy také negativně semidefinitní) matici $\mathbf{Q} \in \mathbb{R}^{n\times n}$ vyřešil úlohu (15.4) v čase, který is shora omezen polynomiální funkcí numbers n. □

Uveďme dále praktičtější příklady.

Example 15.7. consider m bodů v rovině $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^2$. Úkolem je rozmístit dalších n bodů $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^2$ tak, aby nejdelší vzdálenost bodu $\mathbf{a}_i$ k nejbližšímu bodu $\mathbf{x}_j$ byla nejmenší. Tedy minimalizujeme účelovou funkci

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \max_{i=1}^{m} \min_{j=1}^{n} \|\mathbf{a}_i - \mathbf{x}_j\| \tag{15.5}$$

přes vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^2$. we have $f\colon \mathbb{R}^{2n} \to \mathbb{R}$, tedy přesněji můžeme říci, že minimalizujeme funkci f přes jediný vector $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{2n}$.

Úloha is známá jako *shlukování*. Jako motivaci si představme optimální rozmístění cisteren ve vesnici, kde občas neteče voda. Zde $\mathbf{a}_i$ jsou souřadnice domů a $\mathbf{x}_j$ jsou souřadnice cisteren. Chceme, aby obyvatel každého domu měl k nejbližší cisterně co nejblíže.

Je funkce f konvexní? Funkce $g_i(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \|\mathbf{a}_i - \mathbf{x}_j\|$ jsou konvexní pro každé i. Ale funkce $h_i(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \min_{j=1}^{n} \|\mathbf{a}_i - \mathbf{x}_j\|$ již konvexní být nemusí (Větu 11.7 nelze použít, ta hovoří o *maximu* konvexních funkcí). Tedy ani funkce $f(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \max_{i=1}^{m} h_i(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ nemusí být konvexní. Dobře si ujasněte význam funkcí g_i, h_i, f a jejich definiční obory!

Tím, že se nám nepodařilo dokázat konvexitu funkce f, jsme samozřejmě nedokázali její nekonvexitu. A už vůbec jsme nedokázali, že funkce f má více než jedno lokální minimum. Nicméně is známo, že neexistuje algoritmus, který by našel optimální řešení úlohy (15.5) pro libovolný soubor bodů $\mathbf{a}_i$ v čase, který is polynomiální funkcí čísel m a n. V praktické situaci tedy nezbývá nic jiného, než použít algoritmus, který najde pouze přibližné optimum. Takovým algoritmem is např. *k-means*.

Úlohu (15.5) lze modifikovat nahrazením maxima součtem,

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{m} \min_{j=1}^{n} \|\mathbf{a}_i - \mathbf{x}_j\|. \tag{15.6}$$

Opět se jedná o nekonvexní úlohu. Jaký význam má tato formulace? □

Example 15.8. consider školní třídu tvaru konvexní set $T \subseteq \mathbb{R}^2$. Před písemkou chceme rozmístit n studentů tak, aby se jim co nejhůře opisovalo, tedy aby nejmenší vzdálenost mezi každými dvěma studenty byla co největší. Maximalizujeme tedy funkci

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \min_{i,j=1}^{n} \|\mathbf{x}_i - \mathbf{x}_j\|$$

přes vectors $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in T^n$. set přípustných řešení T^n is sice konvexní, ale funkce f není konkávní. Jedná se tedy o nekonvexní úlohu. □

15.2.1 Celočíselné programování

Významnou skupinou nekonvexních úloh jsou úlohy, ve kterých přípustná řešení nabývají pouze celočíselných hodnot. Z nich nejvýznamější je **celočíselné linear programování** (ILP, *integer linear programming*)

$$\min\{\, \mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in \mathbb{Z}^n,\ \mathbf{A}\mathbf{x} \geq \mathbf{b} \,\}. \tag{15.7}$$

Rozdíl oproti obyčejnému LP is 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 $\mathbf{x} \in \{0,1\}^n$. set přípustných řešení této úlohy je nekonvexní, obsahuje konečný počet izolovaných bodů. Neformálně se dá říci, že v jistém smyslu žádná set není 'méně konvexní' než set 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 is set přípustných řešení obyčejného LP.

Zatímco LP is snadné vyřešit (LP is řešitelné v polynomiálním čase), ILP is NP-úplné. ILP má obrovskou aplikovatelnost. Mnoho úloh kombinatorické optimisation (např. úloh na grafech) se dá formulovat jako ILP.

Example 15.9. V *úloze o pokrytí set* Given dán systém množin $F = \{S_1, \ldots, S_n\}$ (tedy S_i jsou set a F je set množin) a úkolem is vybrat z něj co nejmenší podmnožinu, která má stejné sjednocení jako původní systém. Jedná se o jednu z klasických NP-úplných úloh.

Formulujme ji jako ILP. Proměnné budou $x_1, \ldots, x_n \in \{0,1\}$, kde $x_i = 1$ indikuje, že set S_i byla vybrána.

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} x_i \\ \text{za podmínek} \quad & \sum_{i \,|\, e \in S_i} x_i \geq 1, \qquad \forall e \in S_1 \cup \cdots \cup S_n \\ & x_1, \ldots, x_n \in \{0,1\} \end{aligned}$$

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. □

15.3 Exercises

15.1. 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 12.9.

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) $\min\{\, \mathbf{x}^T\mathbf{B}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{x}^T\mathbf{A}\mathbf{x} \leq 1 \,\}$

15.2. Formulujte úlohu $\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_4$ jako konvexní QCQP.

15.3. Given 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 is nulové, zaujme žebřík stav lokálního minima potenciální energie (která is 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 lokální minimum.

15.4. Dokažte, že účelové funkce vystupující v následujících úlohách jsou nekonvexní:

a) úloha (15.5)
b) úloha (15.6)
c) Příklad 10.6
d) Cvičení 10.2

Možný postup is metoda Monte Carlo: V Matlab generujte náhodně (commandem `randn`) potřebné vectors a numbers tak dlouho, dokud neporuší podmínku (11.3).

OPPA European Social Fund
Prague & EU: We invest in your future.