TU WIEN SWM ORCOS
Operations Research and Control Systems

Introduction to Mathematical Optimization

Vladimir M. Veliov

Institute of Statistics and Mathematical Methods in Economics
Vienna University of Technology

Vienna, 2019

Contents

Chapter 1

Introduction, Examples, General Remarks

1.1 The role of optimization

An abstract formulation of any optimization problem reads as follows:

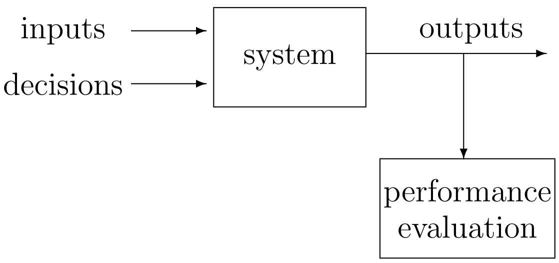
$$\min_{x\in K} f(x),$$

where K is a set (of any kind) and $f : K \to \mathbf{R}$ is a real-valued function defined in K.

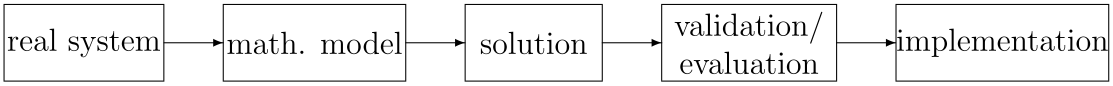
Where do optimization problems arise?

1. Design and maintenance of real systems (engineering, technics, management)

   Real system:



   Approach:



Optimization problems are involved in the block called *math. model.*The influence of various feasible decision variables on the system is modeled with the purpose to optimize the systems performance. Methods for optimization are involved in the block *solution.*

2. Description and analyses of real phenomena:

   - extremal principles in physics
   - qualitative studies in economics.

3. Auxiliary tools for other mathematical problems:

   - for solving equations
   - in approximation theory
   - in methods for identification and estimation.

What do we want to be able to:

- to formulate different optimization problems
- to judge about existence and uniqueness of a solution
- to derive necessary/sufficient optimality conditions
- to obtain qualitative properties of the solutions
- to (approximately) solve the problem numerically.

1.2 A classification of the optimization problems

The classification given below is neither ”canonical” nor needed for the understanding of the exposition below. Moreover, it is neither strict nor comprehensive. Its role is to give a broad view to the problems of mathematical optimization, to acquaint the reader with the used terminology, and to place the material of the present course in the overall optimization context.

Generally speaking, an optimization problem cam be formulated as follows: let K be a set (of any kind) and let $f : K \to \mathbf{R}$ be a real-valued function defined in K. The problem is to find those points in K (if any) at which f attains its minimum on K (more precise definitions will be given in Sections 1.3). Formally this is written as

$$\min_{x \in K} f(x). \tag{1.1}$$

Essentially, different classes of optimization problems are determined by specific features of the set K and the function f.

1. *Continuous optimization problem.* ("Kontinuierliches Optimierungsproblem). Here K is a subset of a normed linear space. For this kind of problems the classical tools of the mathematical analyses are applicable and play an important role (continuity, derivatives, etc.)

1.1. *Finite-dimensional problems.* Here K is a subset of a finite-dimensional space, essentially $\mathbf{R}^n$.

1.1.1. *Linear optimization problems* (*Linear Optimization/Programming*) ("Lineare Optimierung"). Here the objective function is linear and the set K is defined by a finite system of linear equalities and inequalities, as in Example 1.5 below. This topic will be studied in Chapters 4 and 5.

1.1.2. *Non-linear optimization problems.* Here the objective function may be non-linear and the set K may be defined by a finite system of possibly non-linear equalities and inequalities, as in Example 1.7 below. This topic will be studied in Chapters 2 and 3.

1.1.3. *Semi-infinite optimization problems.* Here K is a subset of $\mathbf{R}^n$ which is defined by infinitely many equalities and inequalities.

1.1.4. Several other subclasses of this topic can be identified, in which specific techniques are employed: one-dimensional problems, linear-quadratic problems, optimization problems involving rational functions, problems with specific structures (*structural optimization* – a direction that is currently intensively developing), etc. Partially, such types of problems are presented in a master course at TU Wien.

1.2. *Infinite-dimensional problems.* Here K is a subset of an infinite-dimensional (functional) space. For problems of this type the solution is a function, rather than a finite-dimensional vector. Different branches of this topic are:

1.2.1. *Calculus of variations* ("Variationsrechnung"). Example: find a closed curve with a fixed length passing through given points in the plane which surrounds a maximal surface.

1.2.2. *Optimal control* ("Optimale Steuerung/Kontrolle") of systems described by ordinary differential equations (ODE) (such a course is available for master students at TU Wien, under the name *Dynamic Optimization* ("Dynamische Optimierung")). Example: find the optimal (time-dependent) policy of a firm that maximizes the total revenue.

1.2.3. Optimal control of distributed systems (a course on that is available at TU Wien). Examples: Find the optimal (time-dependent) heating regime of a body, aimed to achieve a certain temperature distribution in a minimal time or with minimal energy.

2. *Discrete optimization*("Discrete Optimierung"). Here the constraining set K has a discrete nature, hence combinatorial methods are more relevant than methods based on the classical mathematical analyses. Such type of problems are included in the course of *Operations Research* within the bachelor study at TU Wien.

2.1. *Integer optimization problems* ("Ganzzahlige Optimierung"). This type of problems has the specific feature that the set K consists of a finite or countable number of elements, therefore is numerable by integer numbers. Even if K contains only finitely many elements, their number can be large, therefore the topic involves hard mathematical considerations of combinatorial and analytic nature, in particular continuous optimization problems as auxiliary tools. As an example we may take the same model as in Example 1.5, with the additional condition that the variables x_j may take as values only natural numbers.

2.2. *Optimization problems on graphs* ("Netzwerkoptimierungsprobleme"). Here the description of the set K involves a graph. Example: for a given graph with fixed and given "lengths" of the edges, find the shortest path connecting two given nodes.

Stochastic Optimization. All of the above types of problems have versions in which some random variables are involved. Additional technical tools from the probability theory and the theory of stochastic processes are needed to study such problems.

1.3 Notions of optimality and terminology

Let $K \subset \mathbf{R}^n$ and $f : K \to \mathbf{R}$. Consider the problem

$$\min_{x \in K} f(x). \tag{1.2}$$

The next definition gives the meaning of the above formulation.

Definition 1.1 The point $x^* \in \mathbf{R}^n$ is called *local solution* ("lokale Lösung") of problem (1.2) if $x^* \in K$ and there exists an open set $\mathcal{O} \subset \mathbf{R}^n$ such that $x^* \in \mathcal{O}$ and

$$f(x^*) \le f(x) \quad \text{for every } \; x \in K \cap \mathcal{O}. \tag{1.3}$$

If $f(x^*) < f(x)$ for every $x \in K \cap \mathcal{O}$, $x \neq x^*$, then x^* is called *strict local solution.* If (1.3) is fulfilled with $\mathcal{O} = \mathbf{R}^n$ then x^* is called *global solution.* □

Terminology. Problem (1.2) should be understood as "find a solution in the sense of the above definitions". The function f is often called *objective function* ("Zielfunktion", "Nutzenfunktion"). The set K is called *constraining set* or simply *constraint* ("Nebenbedingung"). The elements of K are called *admissible (*or *feasible) points* ("zulässige Lösungen"). When we say "minimize f subject to the constraints ..." we mean the problem $\min_{x \in K}$, where K consists of those x for which the constraints ... are satisfied.

We mention that considering a minimization problem (1.2) is not a restriction, since the problem $\max_{x \in K} f(x)$ can be equivalently reformulated as $\min_{x \in K}(-f(x))$.

Sometimes we shall use the term "minimizer", respectively "maximizer", instead of "solution".

Problem (1.2) does not necessarily have a solution. Possible reasons for which a solution may fail to exist may be that the set K is not compact or the objective function f is not continuous.

Theorem 1.2 *Let $f : \mathbf{R}^n \to \mathbf{R}$ be continuous, and let $K \subset \mathbf{R}^n$. Assume that there exists a number b such that the set $K_b := \{x \in K : \ f(x) \leq b\}$ is non-empty and compact. Then problem (1.2) has a global solution.*

Proof. Since f is continuous and K_b is non-empty and compact, f attains its minimum on K_b at some point $x^* \in K_b$. Then $f(x) \geq f(x^*)$ for every $x \in K_b$ and $f(x) > b \geq f(x^*)$ for every $x \in K \setminus K_b$. Thus x^* is a global minimizer of f on K. Q.E.D.

Example 1.3 Consider the problem

$$\min_{(x_1,x_2) \in K} \{-x_1 + \alpha x_2\},$$

where K is the set of those $x = (x_1, x_2) \in \mathbf{R}^2$ which satisfy the inequalities

$$x_1 + x_2 \leq 1,$$

$$x_2 \geq 0.$$

Here α is a parameter and we take $\alpha = 1$. The set K is not compact, but the set K_b is compact for $b = 0$. Indeed, $(0, 0) \in K_0$, hence the set is non-empty. From the constraints we have $x_1 \leq 1$ for every $(x_1, x_2) \in K$. Moreover, $-x_1 + x_2 \leq 0$ implies that $0 \leq x_2 \leq x_1 \leq 1$, hence K_0 is bounded. The closedness of K_0 is evident. Then the problem has a solution according to Theorem 1.2.

Exercise 1.4 Find the set of all values of the parameter α for which the problem in Example 1.3 has a solution.

Neither the local nor the global solution of (1.2) (if such exist) is necessarily unique. The issue of uniqueness will be discussed later on.

1.4 Examples of optimization problems

Example 1.5 (*Optimal diet*) The first row of the next table represents the prices of different nutritious foods, c_j, where the index j indicates a specific food. The main part of the table represents the nutritious contents of different foods, a_{ij}, where i indicates a specific content. The two right-hand columns represent the lower and the upper bound of the tolerable quantities per week of different nutritious components.

	Milk	Potatoes	Bred	......	Wine	Lower b.	Upper b.
Price	c_1	c_2	c_3	......	c_n		
Calories	a_{11}	a_{12}	a_{13}	......	a_{1n}	b'_1	b''_1
Fat	a_{21}	a_{22}	a_{23}	......	a_{2n}	b'_2	b''_2
Cholesterol	a_{31}	a_{32}	a_{33}	......	a_{3n}	b'_2	b''_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
Vitamin C	a_{m1}	a_{m2}	a_{m3}	......	a_{mn}	b'_m	b''_m

Table 2: Foods, nutritious components, prices c_1, contents a_{ij}, bounds b'_i, b''_i.

A diet $x = (x_1, x_2, \dots, x_n) \in \mathbf{R}^n$ is called *admissible* if

$$b'_1 \le \sum_{j=1}^{n} a_{1j} x_j \le b''_1,$$

$$b'_2 \le \sum_{j=1}^{n} a_{2j} x_j \le b''_2,$$

$$\dots \qquad \dots \qquad \dots \tag{1.4}$$

$$b'_m \le \sum_{j=1}^{n} a_{mj} x_j \le b''_m,$$

$$x_1 \ge 0, \;\; x_2 \ge 0, \;\; \dots, \;\; x_n \ge 0. \tag{1.5}$$

Then an admissible diet of minimal cost could be found by solving the minimization problem

$$\min_x \{c_1x_1 + c_2x_2 + \ldots + c_nx_n\}$$

subject to the constraints (1.4), (1.5).

This optimization problem belongs to the group 1 in Section 1.1 as it is a problem of optimal design. The input is a set of foods, the decision variables x_i determine the appropriate quantities, so that the combined foot (the "diet") obtained as an output satisfies the medical requirements and is of highest performance (lowest possible price).

The above problem is a special case of (1.2) with

$$f(x) = c_1x_1 + c_2x_2 + \ldots + c_nx_n, \quad K = \{x \in \mathbf{R}^n : \text{ (1.4), (1.5) are satisfied}\}.$$

Since f is linear and K is defined only by linear inequalities, the problem is called *linear*. Problems of this type are called *linear optimization (programming) problems* ("lineare Optimierungsprobleme").

Example 1.6 *Production maximization under a budget constraint.* A production involves n factors. If $x_1, \ldots, x_n$ are the used quantities of each of the factors, then the produced quantity will be

$$f(x_1, \ldots, x_n) = \left(\sum_{k=1}^{n} c_k (x_k)^\sigma \right)^{1/\sigma},$$

where $\sigma \in (0, 1)$, $c_i > 0$. The above production function is widely used in economics (the so-called production function with constant elasticity of substitution). The producer wants to maximize f facing a budget constraint:

$$\sum_{k=1}^{n} p_k x_k = b,$$

where p_k is the price per unit of the k-th factor and b is the available budget. The used quantities should be non-negative, therefore we should add the constraints

$$x_k \geq 0, \quad k = 1, \ldots, n.$$

The problem of the producer is to maximize $f(x)$ subject to the budget constraint and the non-negativity constraints. This problem also belongs to the group 1 in Section 1.1. In contrast to the previous example, here the objective function is not linear. Problems involving a non-linear objective function or non-linear functions describing the constraints are called *non-linear optimization problems.*

Example 1.7 *Identification of unknown parameters.* Let a real system with input I and output Z be described by the input-output function $I \mapsto F(\xi, I) =: Z$, where $\xi \in \mathbf{R}^n$ is a vector of parameters which is not known for the particular system at hand. However, it may be known that the vector ξ satisfies additional relations (say, due to physical reasons or other additional a priori knowledge):

$$h_j(\xi) \le 0, \quad j = 1, \dots, r. \tag{1.6}$$

In order to identify the values of the parameters ξ one can make use of measurements: for inputs I_k, $k = 1, \dots, s$, the corresponding measured outputs $\tilde{Z}_k$ are assumed to be known, so that

$$F(\xi, I_k) = \tilde{Z}_k + \varepsilon_k,$$

where ε_k are (unknown) measurement errors. A reasonable way to identify the parameters ξ is to solve with respect to ξ the optimization problem

$$\min \sum_{k=1}^{s} \Big(F(\xi, I_k) - \tilde{Z}_k \Big)^2$$

subject to the constraints (1.6), that is to find ξ which explains the available measurements with minimal *mean squared error* $\sum_{k=1}^{s} (\varepsilon_k)^2$. This method was proposed and grounded by Gauss (about 1795) and is known as *method of least squares* ("Methode der kleinsten Quadrate").

The above optimization problem belongs to group 3 in Section 1.1. Notice that here both the objective function and the functions h_i in the constraints may be non-linear and that the constraints have the form of inequalities.

Example 1.8 *Production planning under resource constraints.* To some extent this example is similar to Example 1.6. We consider a firm that has facilities for producing n types of products with market prices p_1, ..., p_n. The manager should decide what amount, x_k, to produce of each product $k \in \{1, \dots, n\}$. Production of quantities $x = (x_1, \dots, x_n)'$ costs $C(x)$. Thus the net revenue of production x will be $\sum_{k=1}^{n} p_k x_k - C(x)$. The production requires r types of resources (labor, various machines or other facilities) and the firm has quantity R_j of the j-th resource, $j = 1, \dots, r$. Production x requires resource $h_j(x)$ of type j. Thus the firm faces the following optimization problem:

$$\max \left\{ \sum_{k=1}^{n} p_k x_k \; - \; C(x) \right\}$$

subject to the constraints

$$h_j(x) \le R_j, \quad j = 1, \dots, r,$$

$$x_k \ge 0, \quad k = 1, \dots, n.$$

The non-negativity constraints are needed in order to avoid negative production, which may formally appear as a solution if this is not prevented by constraints.

Chapter 2

Smooth Constrained Optimization Problems

In this chapter we shall investigate optimization problems in which the objective function is differentiable and the constraining set is defined by a finite number of equalities and inequalities involving differentiable functions. In Section 2.1 we recall some basic notions from the Course of Analysis 1 and in Section 2.2 we give geometric interpretations that help the intuitive understanding of the formal material. Section 2.3 recalls the classical Lagrange theorem for optimization problems with equality constraints.

In the next lines we shortly describe the logics behind the organization of the material in Sections 2.4–2.6.

Let us consider the problem

$$\max_{x\in[a,b]} f(x),$$

where $[a,b] \subset \mathbf{R}$ is a compact interval. (Notice that the constraint $x \in [a,b]$ can be rewritten as a system of two inequality constraints: $-x+a \le 0$ and $x-b \le 0$.) As we know from school, if x^* is a (local or global) solution of the above problem, and if the function f is differentiable, then

$$\frac{\mathrm{d}f}{\mathrm{d}x}(x^*) \begin{cases} = 0 & \text{if } x^* \in (a,b), \\ \le 0 & \text{if } x^* = a, \\ \ge 0 & \text{if } x^* = b. \end{cases}$$

This fact can be interpreted in the following way: whatever is the solution x^*, the gradient (the derivative) of f at x^* is not directed inwards with respect to the interval $[a,b]$: if $x^* = a$ is a maximizer, then the gradient at $x = x^*$ must point to the left (be ≤ 0); if $x^* = b$ is an optimal solution, then the gradient at $x = x^*$ must point to the right (≥ 0); if x^* belongs to the interior of $[a,b]$, then any non-zero vector is directed inward (as far as we can move left and right from x^* still remaining in $[a, b]$), hence the gradient must be zero.

Below in this chapter we shall elaborate the above finding for optimization problems in multi-dimensional spaces. This requires to develop the concept of "inward/outward" directions, that leads to the notion of *normal cone*. This will be done in Section 2.5. In doing this we will need some basic facts from the *convex analysis*, which are presented in Section 2.4. Equipped with all this knowledge we come to Section 2.6, which is the core of this chapter and gives a necessary optimality condition for problems with equality and inequality constraints. In Section 2.8 we briefly discuss second order (sufficient) optimality conditions. Section 2.9 is devoted to examples and exercises.

2.1 Derivatives and an implicit function theorem

We view the linear space $\mathbf{R}^n$ as consisting of n-dimensional column-vectors $x = (x_1, \ldots, x_n)'$, where "$'$" means transposition. For x, $y \in \mathbf{R}^n$ we define the scalar product $\langle x, y\rangle := \sum_{i=1}^n x_i y_i$ and the norm $|x| = \sqrt{\langle x, x\rangle}$. The symbol 0 will be used for the zero vector (the origin) of $\mathbf{R}^n$ and also for the zero-matrix of any dimension; the dimension of 0 should be clear from the context.

Let $U \subset \mathbf{R}^n$ be an open set and let $F : U \to \mathbf{R}^m$. The function F is called *Fréchet differentiable* ("Fréchet-differenzierbar") at $x \in U$ if there exists a linear mapping ("Abbildung") (denoted by) $\partial F(x) : \mathbf{R}^n \to \mathbf{R}^m$ such that

$$\lim_{\xi \to 0,\, \xi \neq 0} \frac{F(x+\xi) - F(x) - \partial F(x)\xi}{|\xi|} = 0.$$

Notice that the above definition can be written as follows: there exists a function $\mathbf{R}^n \ni \xi \mapsto \eta(\xi) \in \mathbf{R}^m$ defined in some neighborhood of $0 \in \mathbf{R}^n$, such that $\eta(\xi) \longrightarrow 0$ with $\xi \longrightarrow 0$ and

$$F(x+\xi) - F(x) = \partial F(x)\xi + |\xi|\eta(\xi). \tag{2.1}$$

The function $\partial F(x)$ is called *Fréchet derivative* ("Fréchet-Ableitung") of F at x. In the text below "differentiable" or "derivative" will always mean "Fréchet differentiable" or "Fréchet derivative".

If $F(x) = (F_1(x), \ldots, F_m(x))'$ and all partial derivatives $\frac{\partial F_i}{\partial x_j}$ exist and are continuous in a neighborhood of x, then F is Fréchet differentiable at x and $\partial F(x)$ can be identified with the matrix

$$\partial F(x) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(x) & \ldots & \frac{\partial F_1}{\partial x_n}(x) \\ \ldots & \ldots & \ldots \\ \frac{\partial F_m}{\partial x_1}(x) & \ldots & \frac{\partial F_m}{\partial x_n}(x) \end{pmatrix}.$$

If $m = 1$ then

$$\nabla F(x) := (\partial F(x))' = \left(\frac{\partial F}{\partial x_1}(x),\ \ldots,\ \frac{\partial F}{\partial x_n}(x)\right)'$$

is the gradient F at x.

Obviously, for $\lambda \in \mathbf{R}^m$ we have

$$\lambda' \, \partial F(x) = \begin{pmatrix} \left\langle \lambda, \frac{\partial F}{\partial x_1}(x) \right\rangle \\ \vdots \\ \left\langle \lambda, \frac{\partial F}{\partial x_n}(x) \right\rangle \end{pmatrix}' = \begin{pmatrix} \sum_{i=1}^m \lambda_i \frac{\partial F_i}{\partial x_1}(x) \\ \vdots \\ \sum_{i=1}^m \lambda_i \frac{\partial F_i}{\partial x_n}(x). \end{pmatrix}' .$$

If the function $F : U \to \mathbf{R}^m$ is Fréchet differentiable at every point in some neighborhood of $x_0 \in U$ and the derivative $\partial F(x)$ is continuous in this neighborhood, then F is called *continuously differentiable* ("stetig differenzierbar") around x_0. Continuous differentiability in an open set is defined as continuous differentiability around every point of this set.

Partial derivatives. If $F : \mathbf{R}^{n_1} \times \mathbf{R}^{n_2} \to \mathbf{R}^m$, then $\partial_{x_1} F(x_1, x_2)$ (with $x_1 \in \mathbf{R}^{n_1}$ and $x_2 \in \mathbf{R}^{n_2}$) denotes the derivative of the function $x_1 \mapsto F(x_1, x_2)$ (if it exists) with respect to $x_1 \in \mathbf{R}^{n_1}$ for a fixed $x_2 \in \mathbf{R}^{n_2}$. Similarly, $\partial_{x_2} F(x_1, x_2)$ is the derivative with respect to x_2 for a fixed $x_1 \in \mathbf{R}^{n_1}$.

Foe example for the function $F : \mathbf{R}^3 \to \mathbf{R}^3$ given by

$$F(x) = F(x_1, x_2, x_3) = \begin{pmatrix} x_1 + (x_2)^2 + (x_3)^3 \\ x_1 x_2 \\ x_2 x_3 \end{pmatrix}$$

we have

$$\partial F(x) = \begin{pmatrix} 1 & 2x_2 & 3x_3 \\ x_2 & x_1 & 0 \\ 0 & x_3 & x_2 \end{pmatrix}$$

$$\partial_{x_1} F(x) = \begin{pmatrix} 1 \\ x_2 \\ 0 \end{pmatrix} \quad \partial_{x_2} F(x) = \begin{pmatrix} 2x_2 \\ x_1 \\ x_3 \end{pmatrix} \quad \partial_{x_3} F(x) = \begin{pmatrix} 3x_3 \\ 0 \\ x_2 \end{pmatrix}.$$

A few particular cases. If A is an $(m \times n)$-matrix, then the derivative of the linear function $\mathbf{R}^n \ni x \mapsto F(x) = Ax \in \mathbf{R}^m$ is $\partial F(x) = A$ for every x. If $G : \mathbf{R}^m \to \mathbf{R}^r$ is a function which is differentiable at $y = Ax$, then $H(x) = G(Ax)$ is differentiable at x and (according to the chain rule) $\partial H(x) = \partial G(Ax)A$.

Now we shall recall the formulation of the classical *implicit function theorem* ("Satz über implizite Funktionen").

Theorem 2.1 *Let $(x_0, y_0) \in \mathbf{R}^p \times \mathbf{R}^q$ and let $F : U \to \mathbf{R}^q$, where U is a neighborhood of (x_0, y_0). Assume that $F(x_0, y_0) = 0$, F is continuously differentiable around (x_0, y_0) and (the $(q \times q)$-matrix) $\partial_y F(x_0, y_0)$ is invertible.*

Then there exists a neighborhood V of x_0 and a function $x \mapsto y(x) \in \mathbf{R}^q$ defined in V such that $(x, y(x)) \in U$ for every $x \in V$,

$$F(x, y(x)) = 0 \quad \text{for every } x \in V,$$

$y(\cdot)$ is continuously differentiable in V, and

$$\partial y(x) = -\left[\partial_y F(x, y(x))\right]^{-1} \partial_x F(x, y(x)).$$

2.2 Geometric interpretations

Here we recall some geometric representations relating functions with sets through their graphs.

If $f : \mathbf{R}^n \to \mathbf{R}$ is differentiable at x then according to (2.1)

$$f(x + \xi) = f(x) + \langle \nabla f(x), \xi \rangle + |\xi| \eta(\xi), \quad \eta(\xi) \to 0 \text{ with } \xi \to 0. \tag{2.2}$$

As a consequence we have the following lemma.

Lemma 2.2 *If f is differentiable at x and for some $\xi \in \mathbf{R}^n$ it holds that $\langle \nabla f(x), \xi \rangle > 0$, then there exists $\bar{\delta} > 0$ such that*

$$f(x + \delta\xi) > f(x) \qquad \forall\, \delta \in (0, \bar{\delta}).$$

Indeed, applying (2.2) with $\xi := \delta\xi$ we obtain that inequality in Lemma 2.2 will be fulfilled if $|\xi|\eta(\delta\xi) < \langle \nabla f(x), \xi \rangle$, which holds for all sufficiently small $\delta > 0$.

Geometrically, Lemma 2.2 says that the "angle" between $\nabla f(x)$ and ξ is acute ("spitzer Winkel")), then $f(x + \delta\xi) > f(x)$ for all sufficiently small positive numbers δ. This means that we may increase the value of f by making a sufficiently small step in the direction of ξ. Conversely, if the angle between $\nabla f(x)$ and ξ is obtuse ("stumpf") (that is, $\langle \nabla f(x), \xi \rangle < 0$), then $f(x + \delta\xi) < f(x)$ for all sufficiently small positive numbers δ. This is illustrated in Fig. 2.1.

Now consider a differentiable function $g : \mathbf{R}^n \to \mathbf{R}$ and the set

$$K := \{x \in \mathbf{R}^n : \; g(x) = 0\}.$$

The set K can have a rather complicated structure if the gradient of g can vanish, even in the case $n = 2$ that we consider below. However, if the gradient does not vanish, then according to the implicit function theorem one can solve (locally) the equation $g(x) = 0$ with respect to one of the two components of x and obtain a piece of curve on the plane. The gradient of g at each point of this curve is perpendicular ("senkrecht") to the curve at this point. This is visualized in Fig. 2.2.

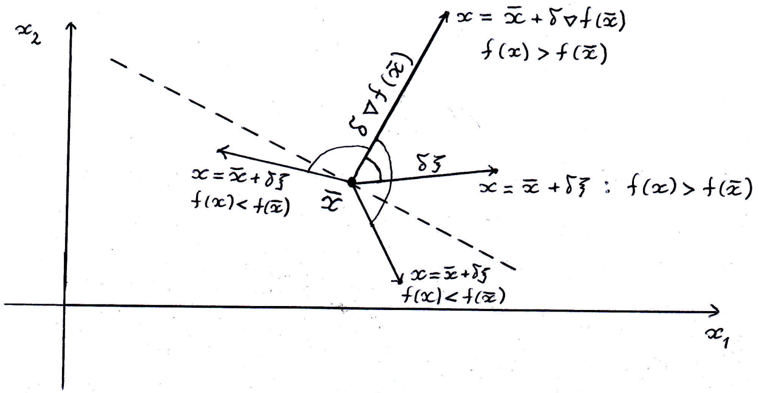

x_2
$x = \bar{x} + \delta\nabla f(\bar{x})$
$f(x) > f(\bar{x})$
$\delta\nabla f(\bar{x})$
$\delta\xi$
$x = \bar{x} + \delta\xi$
$f(x) < f(\bar{x})$
$\bar{x}$
$x = \bar{x} + \delta\xi \; : \; f(x) > f(\bar{x})$
$x = \bar{x} + \delta\xi$
$f(x) < f(\bar{x})$
x_1

Figure 2.1: The gradient $\nabla f(\bar{x})$, directions of increase and descent. The dashed line is perpendicular to $\nabla f(\bar{x})$.

Remark 2.3 Notice that on this plot the drawn arrows are the vectors that begin at x and end at $x + \nabla g(x)$, that is, they are the gradients $\nabla g(x)$ attached at (shifted to) the point x.

Now consider a set defined by an inequality:

$$K := \{x \in \mathbf{R}^2 : \; h(x) \leq 0\},$$

where $h : \mathbf{R}^2 \to \mathbf{R}$ is differentiable. This set can be bounded or unbounded, connected or disconnected. If $h(x) < 0$, then the point x belongs to the interior ("Inneres" oder "offener Kern") of K: $x \in \operatorname{int} K$. If $h(x) = 0$ and $\nabla h(x) \neq 0$, then x belongs to the boundary ("Rand") of K. The gradient is directed towards the exterior of K. This is visualized in Fig. 2.3.

2.3 Problems with equality constraints; Lagrange theorem

Consider first the problem

$$\min f(x), \tag{2.3}$$

where $f : \mathbf{R}^n \to \mathbf{R}$. Here the constraining set K is the whole $\mathbf{R}^n$, therefore it is not included in the formulation of the problem. The following theorem is attributed to Pierre de Fermat, who formulated it between 1629 and 1638 for the case of a polynomial function f. The very notion of derivative was introduced much later and, in fact, the theorem was first proved by Leibniz in 1736 (although apparently it had been known earlier to Newton).

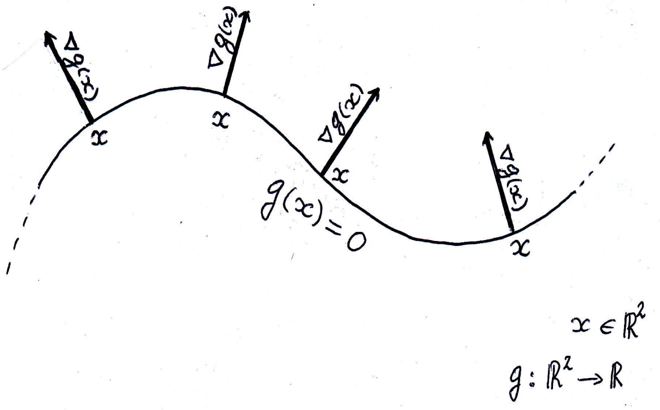


Figure 2.2: The curve $g(x) = 0$ and gradient vectors (normal to the curve).

Theorem 2.4 *Let x^* be a local solution of problem (2.3). Assume that all partial derivatives $\frac{\partial f}{\partial x_j}$ exist at x^*. Then*

$$\frac{\partial f}{\partial x_j}(x^*) = 0, \qquad j = 1, \ldots, n.$$

The claim of the theorem is true also for problems with constraints $x \in K$, provided that the solution x^* belongs to the interior of K: $x^* \in \operatorname{int} K$.

Now consider the minimization problem (2.3) subject to the constraints

$$g(x) = 0, \tag{2.4}$$

where $f : \mathbf{R}^n \to \mathbf{R}$, $g : \mathbf{R}^n \to \mathbf{R}^m$. This formulation should be understood as equivalent to

$$\min_{x \in K} f(x),$$

where $K = \{x \in \mathbf{R}^n : \; g(x) = 0\}$.

Theorem 2.5 *Let x^* be a local solution of problem (2.3), (2.4). Assume that the functions f and g are continuously differentiable around x^*.*
(i) Then there exists a non-zero vector $(\lambda_0, \lambda) \in \mathbf{R} \times \mathbf{R}^m$ such that

$$\lambda_0 \, \partial f(x^*) + \lambda' \, \partial g(x^*) = 0. \tag{2.5}$$

(ii) If, in addition, $\operatorname{rank}(\partial g(x^*)) = m$, *then claim (i) is true with $\lambda_0 = 1$ and the vector λ for which (2.5) holds is unique.*

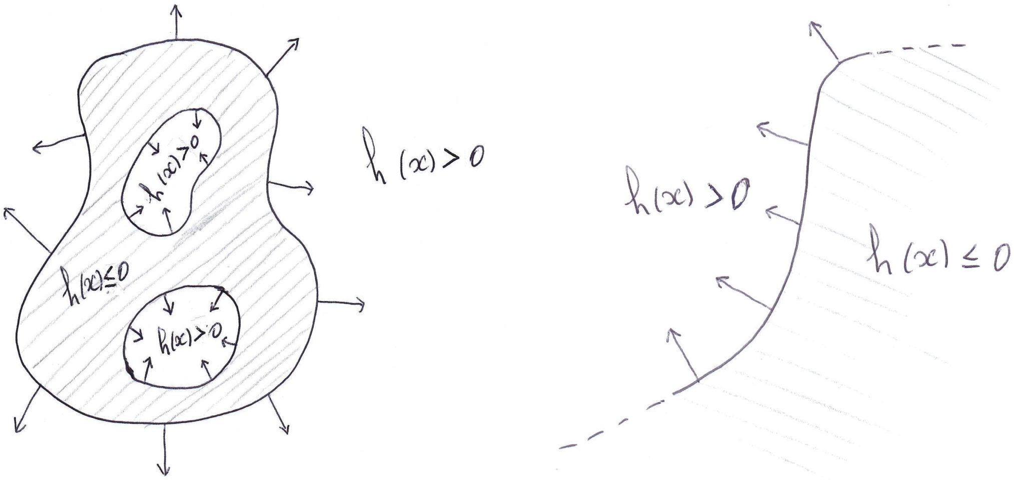


Figure 2.3: Two sets defined by inequalities $h(x) \leq 0$ and gradient vectors at some of their boundary points.

In a detailed form equation (2.5) reads as

$$\lambda_0 \frac{\partial f}{\partial x_j}(x^*) + \sum_{i=1}^{m} \lambda_i \frac{\partial g_i}{\partial x_j}(x^*) = 0, \qquad j = 1, \ldots, n,$$

where $\lambda = (\lambda_1, \ldots, \lambda_m)'$.

The proof of the theorem is given in the course of Analysis 2, but it will also follow from the more general result in the next section.

The above theorem is a version of a general principle that Joseph Louis Lagrange formulated in the year 1788. In words close to those of Lagrange the principle can be formulated as follows: "*The following general principle can be stated. If a maximum or a minimum of a function of several variables is being sought under the condition that there is a constraint connecting the variables which is determined by one or several equations, one must add the functions determining the equations of the constraint multiplied by indeterminate multipliers to the function to be minimized and then seek the maximum or the minimum of the constructed sum as if the variables were independent. The equations obtained, when added to the constraint equations, will serve for determining all the unknowns.*"

Several remarks follow.

Remark 2.6 So-formulated, the Lagrange principle is not a theorem. It is only a suggestion. For many classes of problems this suggestion can take the form of a strict theorem, as in Theorem 2.5, but one has to be careful, especially for problems in infinite-dimensional spaces.

Remark 2.7 The Lagrange principle has played and still plays a key role in the theory of extremal problems, as well as in mechanics and physics.

Remark 2.8 Exercise 2.39 in Section 2.9 shows that claim (ii) of Theorem 2.5 may fail to be true without the additional rank-condition involved.

Remark 2.9 Let us introduce the *Lagrange function* or shortly *Lagrangian* $L : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$

$$L(x,\lambda) = f(x) + \lambda' g(x) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x).$$

Then the Lagrange principle can be formulated as follows: If we want to maximize or minimize the function f under the constraints $g(x) = 0$, we should maximize or minimize the Lagrange function L with respect to x disregarding the constraints. The equations $\partial_x L = 0$ (resulting from the Fermat theorem) together with the equations $g(x) = 0$ can be used for finding the solution of the original optimization problem. Notice that the number of unknowns $(n+m)$ coincides with the number of the equations.

Remark 2.10 The Lagrangian L defined above is in the so-called *normal form.* The first part of Theorem 2.5 suggests to define also the general form Lagrangian

$$\hat{L}(x,\lambda_0,\lambda) = \lambda_0 f(x) + \lambda' g(x)$$

Then Theorem 2.5 can be formulated as follows: for any solution x^* there exists a non-zero vector $(\lambda_0,\lambda) \in \mathbf{R} \times \mathbf{R}^n$ such that

$$\partial_x \hat{L}(x^*,\lambda_0,\lambda) = 0, \qquad g(x^*) = 0.$$

Notice that here the number of unknowns is $n+m+1$ while the number of equations is $n+m$. However, due to the homogeneity of the equation $\partial_x \hat{L}(x^*,\lambda_0,\lambda) = 0$ the vector (λ_0,λ) may be sought either in the form $(0,\lambda)$ (abnormal case) or in the form $(1,\lambda)$ (normal case). The first possibility is eliminated if the rank condition in Theorem 2.5 is fulfilled.

2.4 Convex sets and a separation theorem

Convexity of sets and functions is a very convenient property that enables fruitful analysis in the optimization and in many other contexts. Here we give some preliminary material on convex sets, while Chapter 3 will be entirely devoted to optimization problems formulated in terms of convex sets and convex functions.

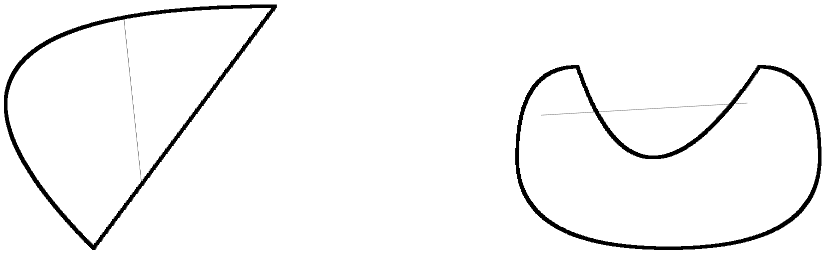


Figure 2.4: A convex and a non-convex set

Definition 2.11 The set $C \subset \mathbf{R}^n$ is called *convex* ("konvex") if for every x_1, $x_2 \in C$ and for every number $\alpha \in (0,1)$ it holds that

$$\alpha x_1 + (1-\alpha)x_2 \in C.$$

Lemma 2.12 *Let $C \subset \mathbf{R}^n$ be convex. Then for every finite number of points $x_1, \ldots, x_k \in C$ and for every $\alpha_1, \ldots, \alpha_k \in [0,1]$ with $\sum_{i=1}^k \alpha_i = 1$, also the point $\sum_{i=1}^k \alpha_i x_i$ belongs to C.*

Proof. For $k = 1$ the claim is trivial. For $k = 2$ the claim repeats the definition of convexity. Let $k = 3$ and let x_i, α_i, $i = 1, 2, 3$, be as in the formulation of the lemma. At least one of the numbers α_i should be strictly smaller than one, and let α_1 be such. Then

$$\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 = \alpha_1 x_1 + (1-\alpha_1)\left(\frac{\alpha_2}{1-\alpha_1}x_2 + \frac{\alpha_3}{1-\alpha_1}x_3\right) = \alpha_1 x_1 + (1-\alpha_1)x \in C,$$

since $x := \frac{\alpha_2}{1-\alpha_1}x_2 + \frac{\alpha_3}{1-\alpha_1}x_3 \in C$ due to

$$\frac{\alpha_2}{1-\alpha_1} + \frac{\alpha_3}{1-\alpha_1} = 1, \qquad \frac{\alpha_i}{1-\alpha_1} \geq 0,\ i = 2, 3.$$

For $k > 3$ one can proceed in the same manner by induction. Q.E.D.

Any point $\sum_{i=1}^k \alpha_i x_i$ with $\alpha_i \geq 0$, $\sum_{i=1}^k \alpha_i = 1$ is called *convex combination* ("Konvex-Kombination") of $x_1, \ldots, x_k$.

Definition 2.13 The set of all convex combinations of the points $x_1, \ldots, x_k \in \mathbf{R}^n$ is called *convex hull* ("konvexe Hülle") of these points and is denoted by $\operatorname{co}\{x_1, \ldots, x_k\}$.

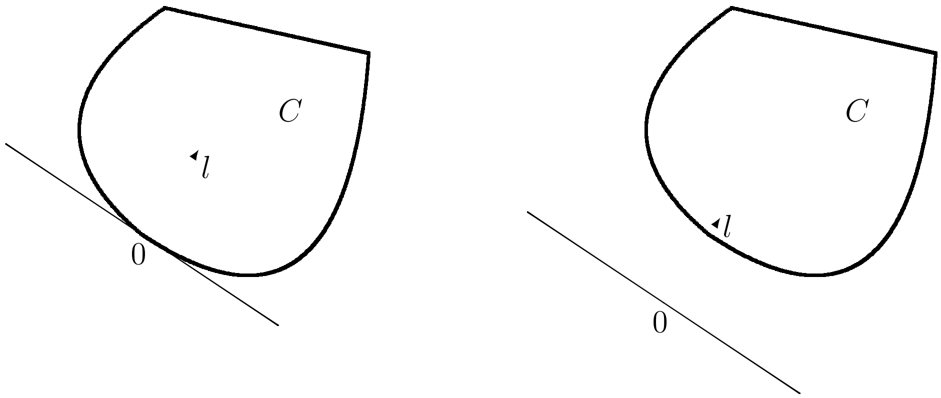


Figure 2.5: Separation and strict separation of C from the origin by a hyperplane.

Lemma 2.14 *The convex hull of a finite number of points is a closed convex set.*

Exercise 2.15 Prove Lemma 2.14.

Lemma 2.16 *The intersection of any family of convex sets is a convex set.*

Exercise 2.17 Prove Lemma 2.16.

Below we give a simple version of the Hahn-Banach separation theorem ("Trennungssatz"), which will be presented in a more general form in the lectures on *functional analysis.*

Theorem 2.18 *Let $C \subset \mathbf{R}^n$ be a nonempty convex set and let $0 \notin C$.*
(i) Then there exists a nonzero vector $l \in \mathbf{R}^n$ such that

$$\langle l, x \rangle \geq 0 \quad \textit{for every } x \in C.$$

(ii) If, in addition, C is closed, then there exists a non-zero $l \in \mathbf{R}^n$ and a real number $\varepsilon > 0$ such that

$$\langle l, x \rangle \geq \varepsilon \quad \textit{for every } x \in C.$$

(This property is called strict separation.*)*

The meaning is that the set C can be separated from the origin by a hyperplane (see Fig. 2.5): the origin belongs to one of the two closed half-spaces defined by l, while C is entirely contained in the other. We give a proof of this theorem for completeness.

Proof of Theorem 2.18. First we prove part (ii) of the theorem. Assume that C is closed and $0 \notin C$. Then we consider the optimization problem

$$\min_{x \in C} |x|. \tag{2.6}$$

Since $C \neq \emptyset$, then there exists some $x_0 \in C$. Then with $b = |x_0|$ we have that the set $C \cap \{x : \ |x| \leq b\}$ is nonempty. It is obviously compact as an intersection of a closed and a compact set. Since $x \longrightarrow |x|$ is a continuous function, Theorem 1.2 implies that the above problem has a global solution. Denote this solution by l. Obviously $l \neq 0$ since $0 \notin C$. We shall prove that the inequality in the second claim of the theorem is true with the chosen l.

Take an arbitrary $x \in C$. For every $\alpha \in (0,1)$ we have $\alpha x + (1-\alpha) l \in C$ due to the convexity of C. Then according to the definition of l as a solution of (2.6) we have

$$|\alpha x + (1-\alpha) l| \ \geq \ |l|.$$

From here

$$|l|^2 \leq |\alpha(x-l) + l|^2 \ = \ \langle \alpha(x-l)+l, \alpha(x-l)+l \rangle \ = \ \alpha^2 |x-l|^2 + 2\alpha \langle x-l, l \rangle + |l|^2,$$

hence

$$\langle l, x \rangle \ \geq \ |l|^2 - \frac{\alpha}{2} |x-l|^2.$$

Since this inequality is fulfilled for any $\alpha \in (0,1)$, it is fulfilled also for $\alpha = 0$, which implies the second claim of the theorem with $\varepsilon = |l|^2 > 0$.

Now we pass to the proof of part (i). Assume that this claim is false. Denote by $\mathcal{S}$ the unit circle in $\mathbf{R}^n$: $\mathcal{S} := \{x \in \mathbf{R}^n : \ |x| = 1\}$. For any $x \in C$ define the set

$$A_x := \{l \in \mathbf{R}^n : \ \langle l, x \rangle < 0\}.$$

We shall prove by contradiction that $\mathcal{S} \not\subset \cup_{x \in C} A_x$. To do this, let us assume that $\mathcal{S} \subset \cup_{x \in C} A_x$. Notice that each of the sets A_x is open (an open half-space) and the union of these open sets covers the compact set $\mathcal{S}$. The compactness of $\mathcal{S}$ implies that there exists a finite sub-covering $A_{x_1}, \ldots, A_{x_k}$. That is, $\mathcal{S} \subset \cup_{i=1,\ldots,k} A_{x_i}$. We know that the set $D := \operatorname{co}\{x_1, \ldots, x_k\}$ is convex and compact and since $x_i \in C$, we have $D \subset C$ according to Lemma 2.12. In particular, $0 \notin D$, since $0 \notin C$. Then we can apply to D the already proved second claim of the theorem: there exists $l \neq 0$ such that $\langle l, x \rangle \geq 0$ for every $x \in D$. Replacing, if necessary, l with $l/|l|$ we may assume that $|l| = 1$, hence $l \in \mathcal{S}$. In particular, $\langle l, x_i \rangle \geq 0$, $i = 1, \ldots, k$, which means that $l \notin A_{x_i}$. This contradicts the assumption that $\mathcal{S} \subset \cup_{x \in C} A_x$, which is then false.

So we have $\mathcal{S} \not\subset \cup_{x \in C} A_x$, that is, there exists $l^* \in \mathcal{S}$ such that $l^* \notin \cup_{x \in C} A_x$. Then for every $x \in C$ we have $l^* \notin A_x$, hence $\langle l^*, x \rangle \geq 0$. Thus the first claim of the theorem is fulfilled with $l = l^*$. Q.E.D.

2.5 Tangent and normal cones to sets

Let $K \subset \mathbf{R}^n$ be a nonempty closed set.

Definition 2.19 A vector $l \in \mathbf{R}^n$ is called *tangent vector* to K at the point $x \in K$ if there exists a sequence of positive numbers $\delta_i \longrightarrow 0$ and a sequence of vectors $\xi_i \longrightarrow 0$ in $\mathbf{R}^n$ such that

$$x + \delta_i l + \delta_i \xi_i \in K \quad \text{for every } \; i = 1,\, 2,\, \ldots\,.$$

The set of all tangent vectors to K at $x \in K$ will be denoted by $T_K(x)$ and called *tangent cone* ("Tangentialkegel") to K at x.

Definition 2.20 A vector $\nu \in \mathbf{R}^n$ is called *normal vector* to K at the point $x \in K$ if

$$\langle \nu, l \rangle \leq 0 \quad \text{for every } \; l \in T_K(x).$$

The set of all normal vectors to K at $x \in K$ will be denoted by $N_K(x)$ and called *normal cone* ("Normalkegel") of K at $x \in K$. It is convenient to define $N_K(x) = \emptyset$ if $x \notin K$.

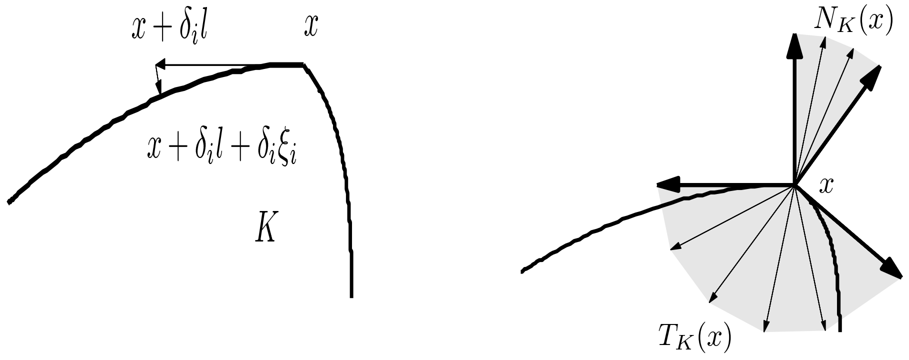


Figure 2.6: A tangent vector to K (left), the tangent and the normal cones (right).

The fact that both the tangent and the normal cones are really cones follows directly from the definitions. We remind that a set $M \subset \mathbf{R}^n$ is a cone if $x \in M$ implies that $\alpha x \in M$ for all $\alpha \geq 0$. Moreover, it easily follows from the definitions that both cones are closed.

For a cone $P \subset \mathbf{R}^n$ the set

$$P^\circ := \{\nu \in \mathbf{R}^n : \ \langle \nu, p \rangle \leq 0 \ \forall p \in P\}$$

is also a cone which is called *polar cone* ("Polarkegel") of P. Thus one can say that the normal cone to a set at a given point is the polar cone to the tangent cone at this point.

We also mention that for a point x from the interior of $K \subset \mathbf{R}^n$ we have $T_K(x) = \mathbf{R}^n$, hence $N_K(x) = \{0\}$. If the set K is defined by an inequality constraint

$$K = \{x \in \mathbf{R}^n : \ h(x) \leq 0\},$$

where $h : \mathbf{R}^n \to \mathbf{R}$ is a differentiable function, then

$$T_K(x) \subset \left\{ \begin{array}{cl} \mathbf{R}^n & \text{if } h(x) < 0, \\ \{l \in \mathbf{R}^n : \ \langle \nabla h(x), l \rangle \leq 0\} & \text{if } h(x) = 0. \end{array} \right.$$

$$N_K(x) \supset \left\{ \begin{array}{cl} \emptyset & \text{if } h(x) > 0, \\ \{0\} & \text{if } h(x) < 0, \\ \{\alpha \nabla h(x) : \ \alpha \geq 0\} & \text{if } h(x) = 0. \end{array} \right.$$

These relations are easy to verify by the definition. In fact, from the results below it will follow that the inclusions are actually equalities, provided that $\nabla h(x) \neq 0$.

The relevance of the above notions in the optimization context is revealed by the following simple proposition, which in fact, provides the ground of the smooth optimization, as presented below.

Proposition 2.21 *Let $K \subset \mathbf{R}^n$ be a nonempty closed set and let $f : \mathbf{R}^n \to \mathbf{R}$. Let x^* be a local solution of the problem*

$$\min_{x \in K} f(x)$$

and let f be differentiable at x^*. Then

$$-\nabla f(x^*) \in N_K(x^*).$$

Proof. Let $l \in T_K(x^*)$ be arbitrarily chosen. According to Definition 2.19, there exist sequences $\delta_i \longrightarrow 0$ and $\xi_i \longrightarrow 0$ such that $x^* + \delta_i l + \delta_i \xi_i \in K$.

Since x^* is a local minimizer, there exists an open set $\mathcal{O}$ containing x^* such that the inequality $f(x) \geq f(x^*)$ holds for every $x \in K \cap \mathcal{O}$. Since $x^* + \delta_i l + \delta_i \xi_i$ belongs to K for all i, and belongs to $\mathcal{O}$ for all sufficiently large i, we have that

$$f(x^* + \delta_i l + \delta_i \xi_i) \geq f(x^*)$$

for all sufficiently large i. Moreover, $\delta_i l + \delta_i \xi_i$ converges to zero. Due to the (Fréchet) differentiability at x^* we have (see (2.1))

$$0 \le f(x^* + \delta_i l + \delta_i \xi_i) - f(x^*) = \partial f(x^*)(\delta_i l + \delta_i \xi_i) + |\delta_i l + \delta_i \xi_i| \, \eta_i,$$

where $\eta_i \longrightarrow 0$. Dividing by δ_i and passing to the limit we obtain that

$$\partial f(x^*) \, l \ge 0, \quad \text{that is, } \langle -\nabla f(x^*), l \rangle \le 0.$$

Since $l \in T_K(x^*) \setminus \{0\}$ was arbitrarily chosen, we conclude from Definition 2.20 that $-\nabla f(x^*) \in N_K(x^*)$. Q.E.D.

The geometric meaning of the proposition is clear: a differentiable function f strictly decreases from $f(x^*)$ in all directions l for which $\langle \nabla f(x), \, l \rangle < 0$. Since x^* is a minimizer of f on K, such directions should be outward looking from K. In particular, (informally) $\langle \nabla f(x), \, l \rangle \ge 0$ must hold for every $l \in T_K(x^*)$, since the tangent directions are (almost) inward looking. This means that $-\nabla f(x^*) \in N_K(x^*)$.

The above proposition shows that it is important to characterize the tangent and the normal cones of sets defined by a system of equalities and inequalities. Namely, consider

$$K := \{x \in \mathbf{R}^n : \; g(x) = 0, \; h(x) \le 0\}, \tag{2.7}$$

where $g : \mathbf{R}^n \to \mathbf{R}^m$, $h : \mathbf{R}^n \to \mathbf{R}^r$. If g and h are continuous, then K is a closed (but not necessarily compact) set. By convention, the relation $h \le 0$ with $h = (h_1, \, \ldots, \, h_r)' \in \mathbf{R}^r$ means that $h_j \le 0$ for every $j = 1, \, \ldots, \, r$.

For a given $x \in K$, denote $J(x) = \{j : \; h_j(x) = 0\}$. That is, $J(x)$ consists of those indexes $j_1, \, \ldots, \, j_p$, for which $h_{j_s}(x) = 0$ (the indexes of the *active constraints* "aktive Nebenbedingungen"). We use the shortening

$$\partial h(x)_{|J(x)} := \begin{pmatrix} \partial h_{j_1}(x) \\ \vdots \\ \partial h_{j_p}(x) \end{pmatrix}. \tag{2.8}$$

Lemma 2.22 *Let the functions g and h_j, $j \in J(x)$, be differentiable at $x \in K$. Then*

$$T_K(x) \subset \left\{ l \in \mathbf{R}^n : \; \partial g(x) \, l = 0, \; \partial h(x)_{|J(x)} \, l \le 0 \right\}. \tag{2.9}$$

The right-hand side of (2.9) reads in detail as

$$\{ l \in \mathbf{R}^n : \; \langle \nabla g_i(x), l \rangle = 0, \; i = 1, \, \ldots, \, m, \; \langle \nabla h_j(x), l \rangle \le 0, \; j \in J(x) \}.$$

Proof. Take an arbitrary $l \in T_K(x)$. Then there are sequences $\delta_k \longrightarrow 0$, $\delta_k > 0$, and $\xi_k \longrightarrow 0$ such that $x + \delta_k l + \delta_k \xi_k \in K$, that is,

$$g(x + \delta_k l + \delta_k \xi_k) = 0, \quad h(x + \delta_k l + \delta_k \xi_k) \le 0.$$

Since for $j \in J(x)$ we have that $h_j(x) = 0$, from the definition of (Fréchet) derivative (see (2.1)) we have that

$$0 \ge h_j(x + \delta_k l + \delta_k \xi_k) - h_j(x) = \partial h_j(x)(\delta_k l + \delta_k \xi_k) + |\delta_k l + \delta_k \xi_k| \eta_k$$

for some sequence $\eta_k \longrightarrow 0$. Since $\delta_k > 0$, we may divide by δ_k and pass to the limit. This gives $\langle \nabla h_j(x), l \rangle \le 0$. In a similar way one can show that $\langle \nabla g_i(x), l \rangle = 0$, which completes the proof. Q.E.D.

The inverse inclusion in (2.9) is not true, in general. A simple example is

$$K = \{(x_1, x_2)' \in \mathbf{R}^2 : \; x_2 - x_1^2 = 0, \; 2x_2 - x_1^2 = 0\}.$$

Here

$$K = \{0\}, \quad T_K(0) = \{0\}, \qquad \text{but } \; \partial g(0) = \begin{pmatrix} 0 & 1 \\ 0 & 2 \end{pmatrix},$$

thus the right-hand side of (2.9) is $\mathbf{R} \times \{0\}$.

However, equality in (2.9) holds under an additional (*regularity*) assumption, as in the next crucial theorem, proved (in a more general framework) by Ljusternik, 1934.

Theorem 2.23 *Let the set $K \subset \mathbf{R}^n$ be defined as in (2.7). Assume that the functions g and h are continuously differentiable around the point $x \in K$. Assume also that the linear mapping $\partial g(x) : \mathbf{R}^n \to \mathbf{R}^m$ is surjective*[1] *(that is,* $\operatorname{rank} \partial g(x) = m$*) and that there exists $\bar{l} \in \mathbf{R}^n$ such that*

$$\partial g(x)\, \bar{l} = 0, \quad \partial h_j(x)\, \bar{l} < 0 \;\; \textit{for} \;\; j \in J(x).$$

Then

$$T_K(x) \; = \; \left\{ l \in \mathbf{R}^n : \; \partial g(x)\, l = 0, \; \partial h(x)_{|J(x)}\, l \le 0 \right\}. \tag{2.10}$$

Proof. Due to Lemma 2.22 we have to prove only the inclusion $L \subset T_K(x)$, where L denotes the set in the right-hand side of (2.10). Let us fix an arbitrary $l \in L$. For any $\alpha > 0$ we consider the vector $p_\alpha := l + \alpha \bar{l}$. We will prove that $p_\alpha \in T_K(x)$ for every $\alpha > 0$. Then also $l \in T_K(x)$ since $p_\alpha \longrightarrow l$ with $\alpha \longrightarrow 0$ and $T_K(x)$ is closed.

[1] We remind that the *image* of a linear mapping $G : \mathbf{R}^n \to \mathbf{R}^m$ is defined as $\operatorname{Im} G := G\mathbf{R}^n := \{Gx : x \in \mathbf{R}^n\}$. The mapping G is surjective if $\operatorname{Im} G = \mathbf{R}^m$, that is, if the rows of the matrix G are linearly independent.

So, it remains to prove that $p_\alpha \in T_K(x)$ for all $\alpha > 0$. Let us fix an $\alpha > 0$ and further skip the subscript α, setting $p := p_\alpha$. Due to the definition of the set L and the properties of $\bar{l}$ we have

$$\partial g(x)\, p = \partial g(x)\, l + \alpha \partial g(x)\, \bar{l} = 0 + 0 = 0,$$

$$\partial h_j(x)\, p = \partial h_j(x)\, l + \alpha \partial h_j(x)\, \bar{l} \le \alpha \partial h_j(x)\, \bar{l} < 0, \quad j \in J(x).$$

Since $\partial g(x)$ is surjective, there is an $(n \times m)$-matrix Y such that the $(m \times m)$-matrix $\partial g(x)\, Y$ is invertible. Indeed, one may define the j-th column of Y as a vector $y_j \in \mathbf{R}^n$ for which $\partial g(x)\, y_j = e_j$, where e_j is the j-th canonical coordinate vector in $\mathbf{R}^m$ (such y_j exists due to the surjectivity of $\partial g(x)$).

Consider the function

$$F(t, s) := g(x + tp + Ys), \quad t \in \mathbf{R},\ s \in \mathbf{R}^m.$$

We want to find $s = s(t)$ so that $F(t, s(t)) = 0$ for all t close to zero. To do this we shall use the implicit function Theorem 2.1. Due to the assumption for g, the function F is continuously differentiable around $(0, 0)$ and

$$\partial_t F(0, 0) = \partial g(x)\, p = 0, \quad \partial_s F(0, 0) = \partial g(x)\, Y \ \text{ – invertible.}$$

Moreover, $F(0, 0) = g(x) = 0$. Then we can apply the implicit function Theorem 2.1: there exists a neighborhood $(-\delta, \delta)$ of $t = 0$ and a function $s(\cdot) : (-\delta, \delta) \to \mathbf{R}^m$ such that $s(\cdot)$ is differentiable at $t = 0$ and

$$s(0) = 0, \quad F(t, s(t)) = 0, \quad t \in (-\delta, \delta).$$

Moreover, due to the last claim of Theorem 2.1 we have

$$\partial s(0) = -[\partial_s F(0, s(0))]^{-1}\, \partial_t F(0, s(0)) = -[\partial g(x)\, Y]^{-1}\, 0 = 0.$$

Now consider the function

$$x(t) = x + tp + Ys(t), \quad t \in [0, \delta),$$

which can be written in the form

$$x(t) = x + tp + t\xi(t) \quad \text{with} \quad \xi(t) = \frac{1}{t} Y s(t). \tag{2.11}$$

Due to the differentiability of $s(\cdot)$, we have

$$s(t) = s(0) + t\partial s(0) + t\eta(t) = t\eta(t),$$

where $\eta(t) \longrightarrow 0$ with $t \longrightarrow 0$ (see (2.1)). Then $\xi(t) = Y\eta(t) \longrightarrow 0$ with $t \longrightarrow 0$. In order to prove that $p \in T_K(x)$ (by taking $t = \delta_i$, $\xi_i = \xi(\delta_i)$ in the definition of tangent vector) we need to show that $x(t) \in K$ for all sufficiently small $t \in (0, \delta)$.

We have

$$g(x(t)) = F(t, s(t)) = 0.$$

For $j \in J(x)$ we have $\partial h_j(x)\, p < 0$, thus

$$\begin{aligned} h_j(x(t)) &= h_j(x(0)) + t\,\partial h_j(x(0))\,\partial x(0) + o(t) \\ &= t\,\partial h_j(x)\,(p + Y\partial s(0)) + o(t) = t\,\partial h_j(x)\,p + o(t) < 0. \end{aligned}$$

Here $o(t)/t \to 0$, hence $t\,\partial h_j(x)\,p + o(t) < 0$ for all sufficiently small t.

For $j \notin J(x)$ we have

$$h_j(x(t)) = h_j(x(0)) + [h_j(x(t)) - h_j(x(0))] < 0$$

for all sufficiently small $t \in (0, \delta)$ (the last inequality uses the continuity of the function $t \longrightarrow h_j(x(t))$ at $t = 0$).

Thus we obtain that $x(t) \in K$. The proof is complete. Q.E.D.

The two conditions in the Ljusternik theorem (the surjectivity and the existence of $\bar{l}$) are known as *Mangasarian-Fromowitz constraint qualification* (also in German).

Notice that the tangent cone $T_K(x)$ is represented in (2.10) by a system of linear equalities and inequalities. Then in order to apply Proposition 2.21 we need to describe the normal cone to a set defined by a system of linear equalities and inequalities. The following lemma is related to a result obtained by Farkas, 1902 (known as lemma of Farkas) and we call it also Farkas lemma, although the original formulation is different and will be given in the next chapter.

Lemma 2.24 *Let*

$$P = \{x \in \mathbf{R}^n : \; Gx = 0, \; Hx \le 0\},$$

where G is an $(m \times n)$-matrix and H is an $(r \times n)$-matrix. Then the polar cone is

$$P^\circ = \{G'\lambda + H'\mu : \; \lambda \in \mathbf{R}^m, \; \mu \in \mathbf{R}^r, \; \mu \ge 0\}. \tag{2.12}$$

Proof. Denote by Q the right-hand side of (2.12).

1. For arbitrary $x \in P$ and $q \in Q$ we have (with some λ and μ as in (2.12))

$$\langle q, x\rangle = \langle G'\lambda + H'\mu, x\rangle = \langle \lambda, Gx\rangle + \langle \mu, Hx\rangle = \langle \mu, Hx\rangle \le 0,$$

where in the last inequality we use that $\mu \geq 0$. Thus $Q \subset P^\circ$.

2. Now assume that there is $\nu \in P^\circ \setminus Q$.

Applying directly the definition of convexity we verify that Q is convex. Obviously Q is also a cone. Moreover, Q is a closed set. This fact is not obvious. It follows from Lemma 4.5 in Chapter 4 and for now it will be taken as known.

Consider the set $C = Q - \nu := \{q - \nu : \ q \in Q\}$. Since Q is a closed convex set, so is C. Moreover, $0 \notin C$ since $\nu \notin Q$. Then according to the separation Theorem 2.18 there exist $l \in \mathbf{R}^n$ and $\varepsilon > 0$ such that

$$\langle l, c\rangle \geq \varepsilon \ \ \forall\, c \in C, \qquad \text{hence} \ \ \langle l, q\rangle \geq \langle l, \nu\rangle + \varepsilon \ \ \forall\, q \in Q. \tag{2.13}$$

Since Q is a cone, the boundedness from below of $\langle l,\, q\rangle$ implies that $\langle l, q\rangle \geq 0$ for every $q \in Q$. (Indeed, if $\langle l,\, q_0\rangle < 0$ for some $q_0 \in Q$, then for a sufficiently large number $b > 0$ it will hold that $\langle l,\, bq_0\rangle < \langle l, \nu\rangle + \varepsilon$, while $bq_0 \in Q$.) So we have

$$\langle l,\, G'\lambda + H'\mu\rangle \geq 0. \tag{2.14}$$

for every $\lambda \in \mathbf{R}^m$ and $\mu \in \mathbf{R}^r$, $\mu \geq 0$. Taking $\mu = 0$ we obtain $\langle Gl,\, \lambda\rangle \geq 0$ for all $\lambda \in \mathbf{R}^m$, which is only possible if $Gl = 0$. Taking $\lambda = 0$ in (2.14) we obtain $\langle Hl,\, \mu\rangle \geq 0$ for all $\mu \geq 0$, which is only possible if $Hl \geq 0$. As a result we obtain that $-l \in P$. Since $\nu \in P^\circ$ it must hold that

$$\langle \nu, -l\rangle \leq 0.$$

On the other hand, taking $q = 0$ in (2.13) we obtain $0 \geq \langle l, \nu\rangle + \varepsilon$, which contradicts the last inequality since $\varepsilon > 0$. This completes the proof. Q.E.D.

2.6 Problems with equality and inequality constraints

This is the core of the present chapter. Here we consider a problem with equality and inequality constraints:

$$\min f(x)$$

subject to the constraints

$$g_i(x) = 0, \qquad i = 1,\ \ldots,\ m, \tag{2.15}$$

$$h_j(x) \leq 0, \qquad j = 1,\ \ldots,\ r, \tag{2.16}$$

where f, g_i, $h_j : \mathbf{R}^n \to \mathbf{R}$. In other words, we deal with the problem

$$\min_{x \in K} f(x), \tag{2.17}$$

where

$$K = \{x \in \mathbf{R}^n : \ g(x) = 0, \ h(x) \leq 0\} \tag{2.18}$$

and $g(x) = (g_1(x), \ldots, g_m(x))'$, $h(x) = (h_1(x), \ldots, h_r(x))'$, so that $g : \mathbf{R}^n \to \mathbf{R}^m$, $h : \mathbf{R}^n \to \mathbf{R}^r$. As before, the vector inequality $h \leq 0$ is understood componentwise: $h_j \leq 0$ for every $j = 1, \ldots, r$.

The following theorem combines results obtained by Karush (1939), and Kuhn and Tucker (1951) and will be further referred to as *Karush-Kuhn-Tucker (KKT) theorem.* It provides *necessary optimality conditions* ("notwendige Optimalitätsbedingungen")

As before, we denote $J(x) = \{j : \ h_j(x) = 0\}$ (remember also the notation (2.8)).

Theorem 2.25 *Let $x^* \in K$ be a local solution of problem (2.17), (2.18). Assume that the functions f, g and h are continuously differentiable around x^*.*

Then

(i) there exists a non-zero vector $(\lambda_0, \lambda, \mu) \in \{0, 1\} \times \mathbf{R}^m \times \mathbf{R}^r$ such that

$$\begin{aligned} \lambda_0 \partial f(x^*) + \lambda' \partial g(x^*) + \mu' \partial h(x^*) &= 0, & (2.19) \\ \mu' h(x^*) &= 0, & (2.20) \\ g(x^*) &= 0, & (2.21) \\ h(x^*) &\leq 0, & (2.22) \\ \mu &\geq 0. & (2.23) \end{aligned}$$

(ii) If, in addition, the matrix $\partial g(x^*)$ is surjective (that is, rank$\partial g(x^*) = m$) and there exists $\bar{l} \in \mathbf{R}^n$ such that $\partial g(x^*) \, \bar{l} = 0$ and $\partial h(x^*)_{|J(x^*)} \, \bar{l} < 0$, then claim (i) of the theorem is true with $\lambda_0 = 1$.

Before proving the theorem we make some remarks.

Remark 2.26 First we mention that in view of (2.23) and (2.22), condition (2.20) can be written as

$$\mu_j h_j(x^*) = 0, \quad j = 1, \ldots, r. \tag{2.24}$$

Indeed, if $\mu' h(x^*) = \sum_{j=1}^r \mu_j h_j(x^*) = 0$, then each summand must be equal to zero, since each summand is non-positive.

Remark 2.27 In a detailed form, the KKT conditions (2.19), (2.20) read as

$$\lambda_0 \partial f(x^*) + \sum_{i=1}^m \lambda_i \, \partial g_i(x^*) + \sum_{j=1}^r \mu_j \, \partial h_j(x^*) = 0,$$

$$\mu_j = 0 \quad \text{for those } j \text{ for which } \ h_j(x^*) < 0.$$

The second group of conditions is known as *complementary slackness condition* (this term is accepted also in German; also the term "Komplementäre Schlupfbedingung" is used).

System (2.19)–(2.23) (also with (2.24) instead of the equivalent (2.20)) of equations and inequalities for the optimal solutions and the corresponding Lagrange multipliers is called KKT system and its solutions $(x, \lambda_0, \lambda, \mu)$ are called KKT points.

Remark 2.28 Formally counted, the number of equations in the KKT system is $m + n + r$. The unknowns x, λ, μ have (in the regular case $\lambda_0 = 1$) in total the same dimension. One can hope that this system of equations may have a unique or finite number of solutions. The inequalities in the KKT system may help to rule out some of them.

In general, there is no easy way to solve the KKT system. Nevertheless, it may be used as a tool for finding explicit solutions in simple cases (see the examples in Section 2.9.3) and as a base for numerical methods or qualitative investigations.

Remark 2.29 According to Remark 2.27, if $(x, 0, \lambda, \mu)$ is an abnormal KKT point of problem (2.17), (2.18), then

$$\sum_{i=1}^{m} \lambda_i \, \partial g_i(x) + \sum_{j \in J(x)}^{r} \mu_j \, \partial h_j(x) = 0, \tag{2.25}$$

which means that the vectors $\partial g_i(x), \partial h_j(x)$, $i = 1, \ldots, m$, $j \in J(x)$ are linearly dependent. Then

$$\operatorname{rank} \begin{pmatrix} \partial g(x) \\ \partial h(x)_{|J(x)} \end{pmatrix} < m + |J(x)|,$$

or equivalently, the above matrix is not surjective. Conversely, if

$$\operatorname{rank} \begin{pmatrix} \partial g(x) \\ \partial h(x)_{|J(x)} \end{pmatrix} = m + |J(x)|, \tag{2.26}$$

then the Mangasarian-Fromowitz constraint qualification is fulfilled. Indeed, the surjectivity implies that $\partial g(x)$ is also surjective and moreover, that the system

$$\begin{pmatrix} \partial g(x) \\ \partial h(x)_{|J(x)} \end{pmatrix} l = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -1 \\ \vdots \\ -1 \end{pmatrix},$$

(where the first m components of the vector in the right-hand side equal zero, the remaining components equal -1) has a solution $\bar{l}$.

Condition (2.26) is stronger than the Mangasarian-Fromowitz constraint qualification. This can be seen, for example, in Example 3.22 in the next chapter.

We mention that if the surjectivity condition (2.26) does not hold, then (2.25) has a nonzero solution (λ, μ), but this does not necessarily mean that $(x, 0, \lambda, \mu)$ is an abnormal KKT point, since some of the components of μ may be negative. Exercise 2.53 provides an example.

Remark 2.30 If we define the (extended) Lagrange function

$$L(x, \lambda, \mu) := f(x) + \lambda' g(x) + \mu' h(x) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{r} \mu_j h_j(x),$$

then in the normal case $\lambda_0 = 1$ condition (2.19) can be written as

$$\partial_x L(x^*, \lambda, \mu) = 0.$$

Remark 2.31 If there are no inequality constraints (that is, $r = 0$), then Theorem 2.25 reduces to the Lagrange Theorem 2.5.

Example 2.46 in Section 2.9 shows that claim (ii) of Theorem 2.25 may not be true without the additional surjectivity condition.

Proof of Theorem 2.25. *Part 1: The normal case.* We start with the proof of claim (ii) of the theorem. According to Proposition 2.21

$$-\nabla f(x^*) \in N_K(x^*) = (T_K(x^*))^{\circ}.$$

Due to Theorem 2.23 we have

$$T_K(x^*) \;=\; \left\{ l \in \mathbf{R}^n : \; \partial g(x^*)\, l = 0, \;\; \partial h(x^*)_{|J(x^*)}\, l \leq 0 \right\}.$$

Then from Lemma 2.24 applied with $G = \partial g(x^*)$ and $H = \partial h(x^*)_{|J(x^*)}$ we obtain that there exist $\lambda \in \mathbf{R}^m$ and $\mu_j \geq 0$, $j \in J(x^*)$, such that

$$-\nabla f(x^*) = (\partial g(x^*))'\lambda + \sum_{j \in J(x^*)} (\partial h_j(x^*))'\, \mu_j.$$

Transposing we obtain

$$-\partial f(x^*) = \lambda' \partial g(x^*) + \sum_{j \in J(x^*)} \mu_j\, \partial h_j(x^*).$$

Defining $\mu_j = 0$ for $j \notin J(x^*)$ we obtain (2.19) with $\lambda_0 = 1$. Equalities (2.23) and (2.20) are obviously also fulfilled.

Part 2: The abnormal case. Now we assume that the additional assumption (the constraint qualification) in claim (ii) of the theorem is not fulfilled. That is, either $\partial g(x^*)$ is not surjective, or $\bar{l}$ with the required properties does not exist.

First we notice that if $\partial g(x^*)$ is not surjective, then its rows are linearly dependent, which means that $\lambda' \, \partial g(x^*) = 0$ for some $\lambda \neq 0$. Therefore claim (i) holds with $\lambda_0 = 0$ and $\mu = 0$.

Now we consider the second possibility, namely, that $\bar{l}$ does not exist. We shall prove by contradiction that claim (i) of the theorem holds with $\lambda_0 = 0$. Assume that the last statement is false. Then for every non-zero (λ, μ) with $\mu \geq 0$ and $\mu_j = 0$ for $j \notin J(x^*)$ it holds that $\lambda' \, \partial g(x^*) + \mu' \, \partial h(x^*) \neq 0$, hence

$$\partial g(x^*)' \, \lambda + \partial h(x^*)' \, \mu \neq 0. \tag{2.27}$$

Define the set

$$C = \left\{ \partial g(x^*)' \, \lambda + \partial h(x^*)' \, \mu : \; \lambda \in \mathbf{R}^m, \; \mu \in \mathbf{R}^r, \; \mu \geq 0, \; \sum_{j=1}^{r} \mu_j = 1, \; \mu_j = 0 \text{ for } j \notin J(x^*) \right\}.$$

Obviously C is a convex set and from Lemma 4.5 in Chapter 4 it follows that C is closed. Moreover, $0 \notin C$ due to (2.27). Then according to the second claim in the Hahn-Banach Theorem 2.18 there exists $l \neq 0$ and $\varepsilon > 0$ such that

$$\langle l, c \rangle \geq \varepsilon \quad \forall \, c \in C.$$

Thus for $\bar{l} := -l$ it holds that

$$\langle \bar{l}, \partial g(x^*)' \, \lambda + \partial h(x^*)' \, \mu \rangle \leq -\varepsilon,$$

hence

$$\langle \partial g(x^*) \, \bar{l}, \lambda \rangle + \langle \partial h(x^*) \, \bar{l}, \mu \rangle \leq -\varepsilon \quad \forall (\lambda, \mu) \text{ as in the definition of } C.$$

Since the inequality has to be satisfied for every $\lambda \in \mathbf{R}^m$, we conclude that $g(x^*) \, \bar{l} = 0$. Moreover, if we take $\lambda = 0$, $\mu_{j_0} = 1$ for some $j_0 \in J(x^*)$ and $\mu_j = 0$ for $j \neq j_0$, we obtain that

$$\partial h(x^*)_{j_0} \, \bar{l} \leq -\varepsilon < 0.$$

Thus $\bar{l}$ satisfies the requirements in claim (ii) of the theorem, which is a contradiction. The proof is complete. Q.E.D.

We give a short recapitulation of the logics that drives the exposition in the last two sections.

(i) The fundamental observation is that the anti-gradient (that is, the gradient with minus sign) of the objective function at a local minimizer belongs to the normal cone to the constraining set.

(ii) The normal cone is polar to the tangent one, and the latter is described by the Ljusternik theorem as a cone defined by **linear** equalities and inequalities (by a local linearization of the constraints).

(iii) The Farkas lemma gives a description of the normal cone to sets defined by linear equalities and inequalities.

Combining these facts we come up with the KKT theorem.

We shall conclude this section with a discussion of Example 1.8 in Chapter 1. In the normal case $\lambda_0 = 1$ the Lagrange function (after reformulation of the problem as a minimization problem) reads as

$$L(x, \mu, \mu^0) = -\langle p, x\rangle + C(x) + \mu'(h(x) - R) - (\mu^0)'x,$$

where $p = (p_1, \ldots, p_n)'$, $h(x) = (h_1(x), \ldots, h_r(x))'$, $R = (R_1, \ldots, R_r)'$, $\mu = (\mu_1, \ldots, \mu_r)'$, $\mu^0 = (\mu^0_1, \ldots, \mu^0_n)'$. The KKT conditions become

$$\begin{aligned} \partial_x L &= -p + \partial C(x) + \mu'\partial h(x) - \mu^0 = 0, \\ &\mu_j(h_j(x) - R_j) = 0, \quad j = 1, \ldots, r, \qquad \mu^0_k\, x_k = 0, \quad k = 1, \ldots, n, \\ &\mu \geq 0, \qquad \mu^0 \geq 0. \end{aligned}$$

The above system consists of $n + r + n$ equations for the $n + r + n$ unknowns x, μ, μ^0. It may still have more than one solution and the additional inequality constraints may help to eliminate some of them. This will be demonstrated in another example in Section 2.9.

2.7 Interpretation of the Lagrange multipliers

The Lagrange multipliers (λ, μ) have a striking interpretation that plays an important role in economics. Let us consider again the problem

$$\min f(x)$$

subject to the constraints

$$\begin{aligned} g_i(x) &= 0, \qquad i = 1, \ \ldots, \ m, \\ h_i(x) &\leq 0, \qquad j = 1, \ \ldots, \ r. \end{aligned}$$

Let us assume that the conditions of claim (ii) of the KKT Theorem 2.25 are fulfilled. We embed this problem into a family of problems parameterized by the vectors $b \in \mathbf{R}^m$, $c \in \mathbf{R}^r$:

$$g_i(x) = b_i, \qquad i = 1, \ldots, m, \tag{2.28}$$

$$h_j(x) \le c_j, \qquad j = 1, \ldots, r. \tag{2.29}$$

The vectors $b \in \mathbf{R}^m$ and $c \in \mathbf{R}^r$ will take values in a presumably small neighborhood $\mathcal{O}$ of the origin in $\mathbf{R}^m \times \mathbf{R}^r$, so that constraints (2.28) and (2.29) can be viewed as resulting from small disturbances in the original constraints (the latter correspond to $b = 0$ and $c = 0$). Assume that for each $(b, c) \in \mathcal{O}$ the corresponding problem has a unique solution $x^*(b, c)$.

Our aim will be to find out how the optimal objective value $f(x^*(b, c))$ depends on the parameters b and c. To do this we make three additional assumptions:

(i) the set of active constraints is the same for all $(b, c) \in \mathcal{O}$: $J(x^*(b, c)) = J^*$;

(ii) the Mangasarian-Fromowitz condition in part (ii) of Theorem 2.25 is fulfilled at $x^*(b, c)$ for every $(b, c) \in \mathcal{O}$;

(iii) the function $(b, c) \longrightarrow x^*(b, c)$ is differentiable in $\mathcal{O}$.

Let us consider the derivative of the optimal value $f(x^*(b, c))$ with respect to b, making use of the KKT Theorem 2.25. Denoting by $\lambda(b, c)$ and $\mu(b, c)$ the corresponding Lagrange multipliers we obtain

$$\begin{aligned}
\partial_b \left(f(x^*(b,c))\right) &= \partial f(x^*(b,c))\, \partial_b x^*(b,c) \\
&= [-\lambda(b,c)'\partial g(x^*(b,c)) - \mu(b,c)'\partial h(x^*(b,c))]\, \partial_b x^*(b,c) \\
&= -\lambda(b,c)'\, [\partial g(x^*(b,c))\partial_b x^*(b,c)] - \mu(b,c)'\, [\partial h(x^*(b,c))\partial_b x^*(b,c)] \\
&= -\lambda(b,c)'\partial_b(g(x^*(b,c))) - \mu(b,c)'\partial_b(h(x^*(b,c))) \\
&= -\lambda(b,c)'\partial_b(b) - \mu(b,c)'\partial_b(h(x^*(b,c)) - c) \\
&= -\lambda(b,c)'I - \partial_b(\mu(b,c)'(h(x^*(b,c)) - c)) = -\lambda(b,c)',
\end{aligned}$$

where I is the unit matrix of dimension $(m \times m)$ and we have used the complementary slackness condition $\mu(b,c)'(h(x^*(b,c)) - c) = 0$.

In a similar way we may prove that $\partial_c \left(f(x^*(b,c))\right) = -\mu(b,c)$. Thus we obtained the following relations for the optimal objective value $f^*(b,c) := f(x^*(b,c))$:

$$\nabla_b f^*(b,c) = -\lambda(b,c), \qquad \nabla_c f^*(b,c) = -\mu(b,c).$$

One can say that the Lagrange multipliers measure the marginal change of the optimal objective value when the right-hand side of the constraints changes. This will be given an economic meaning below in this section. Before this, we stress that the above relations were obtained under demanding conditions. Assumptions (i)–(iii) need not be fulfilled, in general. Conditions that imply (i)–(iii) are known, but will not be discussed in this course.

As an illustration we return to Example 1.8 in Chapter 1. Here we address the question of how can the firm evaluate the marginal value of the different resources, $R_1, \dots, R_r$. This means the following: at what price would it be profitable for the firm to buy a (small) additional amount of resource j, or equivalently, at what price would the firm agree to sell a (small) amount of resource j? In other words, the question is how valuable is resource j for the firm. Of course, this value depends on the entire specification of the firms problem, including the available quantities $R_1, \dots, R_r$.

To answer the above question we evaluate the change of the optimal objective value if instead of quantity R_j of resource j the firm has a quantity $R_j + \varepsilon$. The resulting solution is denoted by $x^*(\varepsilon)$. According to the considerations in this section,

$$\frac{f^*(\varepsilon) - f^*(0)}{\varepsilon} \approx \partial f^*(\varepsilon)_{|\varepsilon=0} = \mu_j,$$

where μ_j is the Lagrange multiplier corresponding to resource j in the KKT theorem. Here μ_j is taken with positive sign since our original problem is for maximization (thus for minimization of $(-f)$).

It would be profitable for the firm to buy a (small) quantity of resource j at any price lower than μ_j. Therefore, the Lagrange multiplier may be interpreted as "shadow price" ("Schattenpreis") of the corresponding resource. If the constraint $h_j(x) \le 0$ is not active, then $h_j(x^*) < 0$ and the complementary slackness condition implies that $\mu_j = 0$. That is, some amount of resource j is useless for the firm and can be sold at any positive price. If for some equality constraint $g_i(x) = 0$ it happens that the corresponding Lagrange multiplier is negative, then the firm should sell some quantity of the respective resource even at negative price (to get rid of it).

We stress that the shadow price has nothing to do with the market price of the resource. The shadow price is specific to each firm and changes if the parameters of the firm change.

2.8 Second order optimality conditions

We begin with a reminder from calculus. Let $F : \mathbf{R}^n \to \mathbf{R}$ be Fréchet differentiable around a point x. Then $x \mapsto (\partial F(x))'$ is a mapping from $\mathbf{R}^n$ to $\mathbf{R}^n$, which may happen to be Fréchet differentiable itself. In this case its derivative is denoted by $\partial_{xx} F(x) : \mathbf{R}^n \to \mathbf{R}^n$. If F is twice continuously differentiable around x, then its Fréchet second derivative coincides with the Hessian ([“Hesse-Matrix”])

$$\partial_{xx}F(x) = \begin{pmatrix} \frac{\partial^2 F}{\partial x_1 \partial x_1}(x) & \dots & \frac{\partial^2 F}{\partial x_1 \partial x_n}(x) \\ \dots & \dots & \dots \\ \frac{\partial^2 F}{\partial x_n \partial x_1}(x) & \dots & \frac{\partial^2 F}{\partial x_n \partial x_n}(x) \end{pmatrix}.$$

Similarly, if $F : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$, then we denote by $\partial_{xx}F(x,y)$ the second (Fréchet) derivative of F with respect to x. If F is twice continuously differentiable around x, then

$$\partial_{xx}F(x,y) = \begin{pmatrix} \frac{\partial^2 F}{\partial x_1 \partial x_1}(x,y) & \dots & \frac{\partial^2 F}{\partial x_1 \partial x_n}(x,y) \\ \dots & \dots & \dots \\ \frac{\partial^2 F}{\partial x_n \partial x_1}(x,y) & \dots & \frac{\partial^2 F}{\partial x_n \partial x_n}(x,y) \end{pmatrix}.$$

The following lemma related to the Taylor expansion will be used later.

Lemma 2.32 *Let $F : \mathbf{R}^n \to \mathbf{R}$ be twice continuously differentiable around the point x, and let $\{x_k\}$ be a sequence such that*

$$x_k \neq x, \quad x_k \longrightarrow x, \quad \frac{x_k - x}{|x_k - x|} \longrightarrow l \in \mathbf{R}^n \quad \text{as } k \to \infty.$$

Then

$$\lim_{k\to\infty} \frac{F(x_k) - F(x) - \partial F(x)(x_k - x)}{|x_k - x|^2} = \frac{1}{2}\, l'\, \partial_{xx}F(x)\, l.$$

Now we consider again problem (2.17), (2.18), this time assuming that the functions f, g and h are twice continuously differentiable around a local solution x^*.

With a triplet $x \in K$, $\lambda \in \mathbf{R}^m$, $\mu \in \mathbf{R}^r_+ := \{\mu \in \mathbf{R}^r : \ \mu \geq 0\}$ we associate the set

$$\begin{aligned} C(x,\lambda,\mu) := \{l \in \mathbf{R}^n : \quad & \partial g(x)\, l = 0, \\ & \partial h_j(x)\, l = 0 \ \text{ if } \ h_j(x) = 0 \text{ and } \ \mu_j > 0, \\ & \partial h_j(x)\, l \leq 0 \ \text{ if } \ h_j(x) = 0 \text{ and } \ \mu_j = 0\}. \end{aligned}$$

The above set is a convex cone called *critical cone* (“kritischer Kegel”). It is very closely related to the tangential cone $T_K(x)$, since obviously

$$C(x,\lambda,\mu) \subset \left\{ l \in \mathbf{R}^n : \ \partial g(x)\, l = 0, \ \partial h(x)_{|J(x)}\, l \leq 0 \right\},$$

and the right-hand side coincides with $T_K(x)$ if the assumptions in the Ljusternik Theorem 2.23 are fulfilled. The right-hand side may fail to be a subset of $C(x,\lambda,\mu)$ since the definition of the latter requires that $\partial h_j(x) l = 0$ if $h_j(x) = 0$ and $\mu_j > 0$.

Theorem 2.33 *Let x^* be a local solution of problem (2.17), (2.18), and let the constraint qualification (2.26) be fulfilled. Let (x^*, λ, μ) be a KKT point (the normal form of the KKT Theorem 2.25 holds due to the constraint qualification – see Remark 2.29). Then*

$$l' \partial_{xx} L(x^*, \lambda, \mu)\, l \geq 0 \quad \textit{for all} \;\; l \in C(x^*, \lambda, \mu).$$

We shall not present the proof of this theorem, which involves arguments similar to those in the proof of the Ljusternik theorem.

Along with the KKT necessary optimality condition in Theorem 2.25 the above theorem containes an additional inequality involving second derivatives. Thus we have a *second order necessary condition* ("Notwendige Bedingung 2. Ordnung"). If there are no constraints, then $C(x^*) = \mathbf{R}^n$ the KKT conditions reduce to $\partial f(x) = 0$, and the inequality in Theorem 2.33 becomes $l' \partial_{xx} f(x^*)\, l \geq 0$ for every $l \in \mathbf{R}^n$, that is, the $(n \times n)$-matrix $\partial_{xx} f(x^*)$ must be non-negative definite.

It is remarkable that the necessary optimality conditions in Theorem 2.33 are rather close to the *sufficient optimality conditions* ("hinreichende Optimalitätsbedingungen") presented in the next theorem.

Theorem 2.34 *Let (x^*, λ, μ) be a KKT point for problem (2.17), (2.18), and let*

$$l' \partial_{xx} L(x^*, \lambda, \mu)\, l > 0 \quad \textit{for all} \;\; l \in C(x^*, \lambda, \mu) \setminus \{0\}.$$

Then x^* is a strict local minimizer.

Proof. Assume that the claim of the theorem is false, that is, for every natural number k there exists $x_k \in K$ such that

$$|x_k - x^*| \leq \frac{1}{k}, \quad x_k \neq x^*, \quad \text{and} \quad f(x_k) \leq f(x^*).$$

The vectors $\frac{x_k - x^*}{|x_k - x^*|}$ belong to the unit ball, therefore we can extract a subsequence (we use the same enumeration) that converges to some unit vector l. Then $x_k = x^* + \beta_k l + o(\beta_k)$, where $\beta_k := |x_k - x^*|$, thus $l \in T_K(x^*)$. From Proposition 2.21 we know that $-\nabla f(x^*) \in N_K(x^*)$, hence

$$\partial f(x^*)\, l \geq 0. \tag{2.30}$$

Moreover, according to Lemma 2.22, we obtain that

$$l \in \left\{ p \in \mathbf{R}^n : \; \partial g(x^*)\, p = 0, \;\; \partial h(x^*)_{|J(x^*)}\, p \leq 0 \right\}. \tag{2.31}$$

Then either we have that $l \in C(x^*, \lambda, \mu)$, or there exists some $j_0 \in J(x^*)$ for which $\mu_{j_0} > 0$ but $\partial h_{j_0}(x^*)\, l < 0$.

Let us assume that $l \notin C(x^*, \lambda, \mu)$. From the KKT conditions we have that

$$\partial f(x^*) + \sum_{i=1}^{m} \lambda_i \partial g_i(x^*) + \sum_{j=1}^{r} \mu_j \partial h_j(x^*) = 0.$$

Multiplying from the right by l and using the complementary slackness condition and (2.31) we obtain that

$$\partial f(x^*)\, l + \sum_{j \in J(x^*)}^{r} \mu_j \partial h_j(x^*)\, l = 0.$$

For every $j \in J(x^*)$ it holds that $\mu_j \geq 0$ and $\partial h_j(x^*)\, l \leq 0$. Thus every summand is non-positive. However, $\mu_{j_0} \partial h_{j_0}(x^*)\, l < 0$, hence

$$\partial f(x^*)\, l > 0.$$

On the other hand, $f(x_k) - f(x^*) \leq 0$, which implies that

$$\partial f(x^*)\, l = \lim_{\beta \to 0+} \frac{f(x^* + \beta l) - f(x^*)}{t} = \lim_{k \to \infty} \frac{f(x^* + \beta_k l + o(\beta_k)) - f(x^*)}{\beta_k} \leq 0. \quad (2.32)$$

This contradiction proves that $l \in C(x^*, \lambda, \mu)$.

From (2.30) and (2.32) we obtain that $\partial f(x^*)\, l = 0$. Since (x^*, λ, μ) is a KKT point and $x_k \in K$, we have

$$F(x_k) = f(x_k) + \lambda' \underbrace{g(x_k)}_{=0} + \underbrace{\mu'}_{\geq 0} \underbrace{h(x_k)}_{\leq 0} \leq f(x_k)$$

and

$$F(x*) = f(x^*) + \lambda' \underbrace{g(x^*)}_{=0} + \underbrace{\mu' h(x^*)}_{=0} = f(x^*).$$

Then, using that $\partial F(x^*) = 0$ we obtain

$$F(x_k) - F(x^*) - \partial F(x^*)(x_k - x^*) \leq f(x_k) - f(x^*).$$

According to Lemma 2.32 applied for $F(x) := L(x, \lambda, \mu)$, we have

$$\frac{1}{2}\, l'\, \partial_{xx} F(x^*)\, l = \lim_{k \to \infty} \frac{F(x_k) - F(x^*) - \partial F(x^*)(x_k - x^*)}{|x_k - x^*|^2} = \lim_{k \to \infty} \frac{f(x_k) - f(x^*)}{|x_k - x^*|^2} \leq 0.$$

This contradicts the inequality in the formulation of the theorem and completes the proof. Q.E.D.

Remark 2.35 In the next lines we comment the advantages and the drawbacks of the necessary versus sufficient optimality conditions, in general. A necessary optimality condition must be satisfied by every solution of the optimization problem if such exists. Therefore, the points that satisfy a necessary optimality condition can be viewed as "candidates" for optimal solutions. All solutions are "candidates", but not all "candidates" need be really optimal. Even if a unique "candidate" exists, it does not need to be optimal. An easy example is provided by the problem $\min_{x\in\mathbf{R}} x^3$, where $x = 0$ is the only point satisfying the necessary optimality condition $\partial f(x) = 0$ (that is, the only "candidate"), but nevertheless, it is not an optimal solution. An optimal solution does not exist in this problem. Here we see the importance of existence theorems, such as Theorem 1.2. If it is known that an optimal solution exists and a necessary optimality condition is satisfied by a unique point, then this point is optimal, that is, the necessary condition turns into a sufficient one.

In contrast, a sufficient optimality condition is very useful if for a given problem one can find a point that satisfies this condition. This point is automatically optimal. However, in general, the sufficient conditions are more demanding than the necessary ones. It may happen that the optimal point does not satisfy a particular sufficient condition. For example, $f(x) = x^4$ has a minimum at $x = 0$, but the sufficient condition (known from school) $f'(x) = 0$, $f''(x) > 0$, is not fulfilled.

A reasonable approach is to try to find all "candidates" and then to show that some of them satisfy a sufficient condition, or at least that some of them do not satisfy a second order necessary conditions and can be eliminated.

2.9 Additional examples and exercises

We start with a few exercises on geometric representations.

Exercise 2.36 Draw the geometric representation of the following sets and the gradients of the involved non-linear functions at 2-3 points, plotted as attached to the respective points (see Remark 2.3):

(i) $K := \{(x_1, x_2) \in \mathbf{R}^2 :\ 0 \le x_1 + x_2 \le 1,\ -1 \le x_1 \le 1\}$;

(ii) $K := \{(x_1, x_2) \in \mathbf{R}^2 :\ x_1 - (x_2)^2 \le 1,\ -x_1 + (x_2)^2 \le 0,\ x_1 \le 3\}$;

(iii) $K := \{(x_1, x_2) \in \mathbf{R}^2 :\ (x_1)^2 + (x_2)^2 \le 1,\ (x_1 - 1)^2 + (x_2 - 1)^2 = 1\}$.

2.9.1 Lagrange theorem

Example 2.37 *Production maximization under budget constraint.* Here we solve the problem in Example 1.6 (Section 1.4). The problem is

$$\max \left(\sum_{k=1}^{n} c_k (x_k)^\sigma \right)^{1/\sigma}$$

subject to

$$\sum_{k=1}^{n} p_k x_k = b, \tag{2.33}$$

$$x_k \geq 0, \quad k = 1, \dots, n,$$

where $\sigma \in (0,1)$, $c_k > 0$, $p_k > 0$ and $b > 0$ are given data. Notice that due to the monotonicity of the function $y \mapsto y^\sigma$ the above problem is equivalent to the one of maximization of

$$f(x) := \sum_{k=1}^{n} c_k (x_k)^\sigma$$

under the same constraints. The problem has a solution. Indeed, from (2.33) we have that $x_k \leq b/p_k$, so the constraining set is compact and f is continuous on it. Moreover, every solution x satisfies $x_i > 0$. Indeed, assume that for example $x_1 = 0$. There is some k such that $x_k > 0$. Then consider another point $\tilde{x}$, where $\tilde{x}_1 = \varepsilon$, $\tilde{x}_k = x_k - \varepsilon p_1/p_k$ and the rest of the components are unchanged. If $\varepsilon > 0$ is sufficiently small, then $\tilde{x}_k \geq 0$ and obviously (2.33) is also satisfied. On the other hand

$$f(\tilde{x}) - f(x) = c_1 \varepsilon^\sigma + c_k (x_k - \varepsilon p_1/p_k)^\sigma - c_k (x_k)^\sigma$$

The right-hand side is zero for $\varepsilon = 0$ and its derivative is strictly positive for all sufficiently small positive ε. Then the right-hand side is positive for sufficiently small positive ε, which contradicts the optimality of x.[2]

So, the problem has a solution and every solution satisfies $x_k > 0$. Then any solution is also a local solution of the problem with only the equality constraint (2.33). Since $\operatorname{rank} \partial g(x) = \operatorname{rank}(p_1, \dots, p_n) = 1 = m$, the Lagrange theorem can be applied in its stronger version (with $\lambda_0 = 1$). Taking into account that we have to replace f by $-f$ in order to pass to a minimization problem, the Lagrangian is

$$L(x, \lambda) = -\sum_{j=1}^{n} c_j (x_j)^\sigma + \lambda \left(\sum_{j=1}^{n} p_j x_j - b \right).$$

[2] This argumentation is often used when one wants to prove that some vector x is not an optimal solution. Namely, find an appropriate small variation of x which is admissible and gives a better value of the objective function.

The Lagrange equation $\partial_x L(x,\lambda) = 0$ takes the form

$$-\sigma c_k (x_k)^{\sigma-1} + \lambda p_k = 0, \quad k = 1, \ldots, n.$$

From here we obtain that

$$x_k = \left(\frac{\lambda p_k}{\sigma c_k}\right)^{\frac{1}{\sigma-1}}, \quad k = 1, \ldots, n.$$

Substituting this in equation (2.33) we determine the Lagrange multiplier λ:

$$\lambda^{\frac{1}{\sigma-1}} = \frac{b}{\sum_{j=1}^{n} p_k \left(\frac{p_k}{\sigma c_k}\right)^{\frac{1}{\sigma-1}}}.$$

Hence,

$$x_k = \frac{b}{\sum_{j=1}^{n} p_j \left(\frac{p_j}{\sigma c_j}\right)^{\frac{1}{\sigma-1}}} \left(\frac{p_k}{\sigma c_k}\right)^{\frac{1}{\sigma-1}}, \quad k = 1, \ldots, n.$$

Exercise 2.38 Find an explicit formula for the solution of the problem in Example 2.37.

Exercise 2.39 *(The rank condition is essential for the normal form of the Lagrange Theorem 2.5).* Show that the Lagrange principle is not fulfilled with $\lambda_0 = 1$ for the following (trivial) problem

$$\min x$$

subject to

$$x^2 = 0, \quad x \in \mathbf{R}^1.$$

2.9.2 Tangents and normals

Exercise 2.40 What are $T_K(x)$ and $N_K(x)$ for

$$K = [0,2] \subset \mathbf{R}, \qquad \text{and } x = 0,\ x = 1,\ x = 2.$$

Example 2.41 Let us describe analytically the tangent cone and the normal cone to the set

$$K = \{(x_1, x_2) \in \mathbf{R}^2 :\ x_1 + x_2 \leq 1,\ x_1 \geq 0,\ x_2 \geq 0\}.$$

at $\bar{x} = (1,0)' \in K$. (Of course, this can be directly done by using the Ljusternik theorem and the lemma of Farkas from Section 2.5, but here we want to do this just by applying the definitions.)

If $l = (l_1, l_2)' \in T_K(x)$, then we have

$$\bar{x} + \delta_i l + \delta_i \xi_i \in K, \quad \text{with } 0 < \delta_i \to 0, \;\; \xi_i = (\xi_i^1, \xi_i^2) \to 0.$$

This means that

$$(1 + \delta_i l_1 + \delta_i \xi_i^1) + (0 + \delta_i l_2 + \delta_i \xi_i^2)) \leq 1, \tag{2.34}$$

$$1 + \delta_i l_1 + \delta_i \xi_i^1 \geq 0, \quad 0 + \delta_i l_2 + \delta_i \xi_i^2 \geq 0. \tag{2.35}$$

Since $\xi_i \to 0$ and $\delta_i > 0$, we obtain from the last inequality that $l_2 \geq 0$. The first inequality is

$$\delta_i(l_1 + l_2) + \delta_i(\xi_i^1 + \xi_i^2) \leq 0,$$

which similarly implies $l_1 + l_2 \leq 0$. Then we have

$$T_K(\bar{x}) \subset \{(l_1, l_2) : \; l_1 + l_2 \leq 0, \; l_2 \geq 0\}.$$

To prove the inverse inclusion we take (l_1, l_2) from the right-side and check that (2.34), (2.35) are fulfilled with $\xi_i = 0$. Thus we have described the tangent cone. (See it geometrically!)

For the normal cone we have that $(\nu_1, \nu_2) \in N_K(\bar{x})$ if and only if $\nu_1 l_1 + \nu_2 l_2 \leq 0$ for every $(l_1, l_2) \in T_K(x)$, that is, for which $l_1 + l_2 \leq 0, \; l_2 \geq 0$. In particular, for $l = (-1, 0)'$ we obtain that $-\nu_1 \leq 0$ and for $l = (-1, 1)$ we obtain $-\nu_1 + \nu_2 \leq 0$. Then

$$N_K(\bar{x}) \subset \{(\nu_1, \nu_2) : \; -\nu_1 + \nu_2 \leq 0, \; \nu_1 \geq 0\}.$$

To prove the inverse inclusion we fix (ν_1, ν_2) from the right-hand side, take an arbitrary $(l_1, l_2)' \in T_K(\bar{x})$ and represent

$$(l_1, l_2) = l_2(-1, 1) - (l_1 + l_2)(-1, 0).$$

Then we have

$$\langle \nu, l \rangle = l_2 \langle \nu, (-1, 1)' \rangle - (l_1 + l_2) \langle \nu, (-1, 0)' \rangle.$$

Since we already ensured that the two scalar products are non-positive and since $l_2 \geq 0$ and $l_1 + l_2 \leq 0$, we obtain that $\langle \nu, l \rangle \leq 0$, hence $\nu \in N_K(\bar{x})$.

Example 2.42 Describe analytically the tangent and the normal cone to the set $K \subset \mathbf{R}^3$ defined by the constraints

$$(x_1)^2 + 2(x_3)^2 \leq 3,$$

$$(x_1)^3 + (x_2)^3 = 9$$

at the point $x^* = (1, 2, 1)' \in K$.

Solution. Here

$$g(x) = (x_1)^3 + (x_2)^3 - 9, \qquad h(x) = (x_1)^2 + 2(x_3)^2 - 3.$$

Then

$$\begin{aligned} \partial g(x^*) &= \big(3(x_1^*)^2,\, 3(x_2^*)^2,\, 0\big) = (3, 12, 0), \\ \partial h(x^*) &= (2x_1^*,\, 0,\, 4x_3^*) = (2, 0, 4). \end{aligned}$$

For $\bar{l} := (-4, 1, 0)'$ (this is just one of many possible choices) we have

$$\partial g(x^*)\,\bar{l} = (3, 12, 0) \begin{pmatrix} -4 \\ 1 \\ 0 \end{pmatrix} = 0, \qquad \partial h(x^*)\,\bar{l} = (2, 0, 4) \begin{pmatrix} -4 \\ 1 \\ 0 \end{pmatrix} = -8 < 0,$$

thus the assumptions of the Ljusternik theorem are fulfilled.

Then

$$\begin{aligned} T_K(x^*) &= \left\{ l \in \mathbf{R}^3 :\ (3, 12, 0) \begin{pmatrix} l_1 \\ l_2 \\ l_3 \end{pmatrix} = 0, \quad (2, 0, 4) \begin{pmatrix} l_1 \\ l_2 \\ l_3 \end{pmatrix} \le 0 \right\} \\ &= \{ l \in \mathbf{R}^3 :\ 3l_1 + 12l_2 = 0,\ 2l_1 + 4l_3 \le 0 \} \\ &= \{ l \in \mathbf{R}^3 :\ l_1 + 4l_2 = 0,\ l_1 + 2l_3 \le 0 \}. \end{aligned}$$

Then we apply the Farkas lemma with $G = (1, 4, 0)$ and $H = (1, 0, 2)$ and obtain

$$\begin{aligned} N_K(x^*) &= \left\{ \begin{pmatrix} 1 \\ 4 \\ 0 \end{pmatrix} \lambda \ + \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} \mu :\ \lambda \in \mathbf{R},\ \mu \ge 0 \right\} \\ &= \left\{ \begin{pmatrix} \lambda + \mu \\ 4\lambda \\ 2\mu \end{pmatrix} :\ \lambda \in \mathbf{R},\ \mu \ge 0 \right\}. \end{aligned}$$

Exercise 2.43 Describe analytically and graphically the tangential and the normal cone to the set

$$K = \{ (x_1, x_2)' \in \mathbf{R}^2 :\ (x_1)^2 - x_1 + x_2 \le 0,\ x_2 \ge 0 \}$$

at the point $\bar{x} = (1, 0)' \in K$.

Answer:

$$T_K(\bar{x}) = \{ (l_1, l_2)' = \beta(-1, \alpha)' \in \mathbf{R}^2 :\ \beta \ge 0,\ \alpha \in [0, 1] \}$$

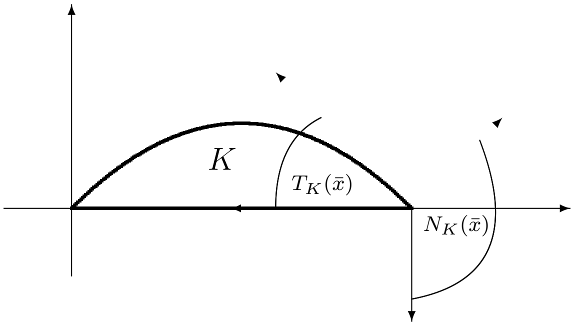


Figure 2.7: The tangent and the normal cones from Example 2.43 translated at the point $\bar{x} = (1.0)'$.

$$N_K(\bar{x}) = \{(\nu_1, \nu_2)' = \beta(\alpha, 2\alpha - 1)' \in \mathbf{R}^2 : \ \beta \geq 0, \ \alpha \in [0, 1]\}.$$

The two sets are visualized in Figure 2.7. Notice that both cones in this figure are shifted from the origin to the point $\bar{x}$.

Exercise 2.44 Describe analytically $T_K(x)$ and $N_K(x)$ for

$$K = \left\{x \in \mathbf{R}^3 : \ (x_1)^2 + (x_2)^2 + (x_3)^2 = 1, \ x_3 \geq 0\right\} \quad \text{and } x = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0\right)'.$$

Exercise 2.45 What are $T_K(x)$ and $N_K(x)$ for

$$K = \left\{(x_1, x_2)' \in \mathbf{R}^2 : \ x_1 = t \cos\frac{1}{t}, \ x_2 = t \sin\frac{1}{t}, \ t > 0\right\} \cup \{0\} \quad \text{and } x = (0, 0)'.$$

2.9.3 The KKT theorem

Example 2.46 *Abnormal case.* Consider the problem

$$\min\{x_1\}$$

subject to

$$-(x_1)^3 + x_2 \leq 0$$

$$-x_2 \leq 0.$$

The point $(0,0)'$ satisfies the constraints and the corresponding objective value is 0. For any other admissible point $(x_1,x_2)'$ we have $0 \leq x_2 \leq (x_1)^3$, hence $x_1 \geq 0$. The corresponding objective value is non-negative. Then $x^* = (0,0)'$ is a solution of the problem. The Lagrange equation (2.19) reads in this case as

$$\lambda_0(1 \quad 0) + \mu_1(0 \quad 1) + \mu_2(0 \quad -1) = (0 \quad 0),$$

which is satisfied only if $\lambda_0 = 0$. Thus the normal form of the KKT does not hold in this case. Notice that here there are two active constraints and for every $l = (l_1,l_2)' \in \mathbf{R}2$

$$\partial h(x^*)l = \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} l_1 \\ l_2 \end{pmatrix} = \begin{pmatrix} l_2 \\ -l_2 \end{pmatrix}.$$

This cannot be a vector with strictly negative components, therefore the Mangasarian-Fromowitz constraint qualification fails.

Example 2.47 *([4, p. 34])* Consider the problem

$$\min\{(x_1)^2 - 4x_1 + (x_2)^2 - 6x_2\}$$

subject to

$$\begin{aligned} x_1 + x_2 &\leq 3 \\ -2x_1 + x_2 &\leq 2. \end{aligned}$$

First of all we can apply Theorem 1.2 to prove existence of a global solution (here K is not compact, but one can appropriately define a compact set K_b; with what number b?).

The Lagrange function reads as

$$L(x_1,x_2,\mu_1,\mu_2) = (x_1)^2 - 4x_1 + (x_2)^2 - 6x_2 + \mu_1(x_1 + x_2 - 3) + \mu_2(-2x_1 + x_2 - 2).$$

We can apply the strong claim of the KKT theorem (with $\lambda_0 = 1$) since

$$\partial h(x) = \begin{pmatrix} 1 & 1 \\ -2 & 1 \end{pmatrix}$$

and the surjectivity condition (2.26) in Remark 2.29 holds no matter whether $J(x) = \emptyset$, $\{1\}$, $\{2\}$, or $\{1,2\}$.

The KKT system consists of the conditions

$$\begin{aligned} h_1 &= x_1 + x_2 - 3 \le 0, \\ h_2 &= -2x_1 + x_2 - 2 \le 0, \\ \partial_{x_1} L &= 2x_1 - 4 + \mu_1 - 2\mu_2 = 0, \\ \partial_{x_2} L &= 2x_2 - 6 + \mu_1 + \mu_2 = 0, \\ \mu_1 h_1 &= \mu_1(x_1 + x_2 - 3) = 0, \\ \mu_2 h_2 &= \mu_2(-2x_1 + x_2 - 2) = 0, \\ & \mu_1 \ge 0, \ \mu_2 \ge 0. \end{aligned}$$

We shall consider the following four cases.

Case 1: $\mu_1 = \mu_2 = 0$. Here the second group of two equations determines $x_1 = 2$, $x_2 = 3$. However, the point $(2, 3)'$ does not satisfy the first inequality, thus we can disregard it.

Case 2: $\mu_1 \neq 0$, $\mu_2 = 0$. In this case the solution of the respective equations for x_1, x_2, μ_1 (the third, the fourth, and the fifth equations) is $x = (1, 2)'$, $\mu_1 = 2$. All the inequalities are fulfilled in this case, thus $(x, \mu) = (1, 2, 2, 0)'$ is a KKT point and $x = (1, 2)'$ is candidate for a local solution (that is, it satisfies the necessary optimality conditions in the KKT theorem).

Case 3: $\mu_1 = 0$, $\mu_2 \neq 0$. Here the solution of the third, the fourth, and the sixth equations is $x = (4/5, 18/5)'$, $\mu_2 = -6/5$. The Lagrange multiplier μ_2 is negative, therefore we can disregard also this case.

Case 4: $\mu_1 \neq 0$, $\mu_2 \neq 0$. In this case the system of equations for x and μ consists of the third, the fourth, the fifth, and the sixth equations. The unique solution has a negative $\mu_2 = -8/9$, therefore it also does not bring a KKT point.

The above considerations show that there is only a single KKT point appearing in Case 2. Since we know that a solution exists, we conclude that it is $x^* = (1, 2)'$.

We mention that sometimes more than one KKT point may exist. Not all of them determine an (even locally) optimal solution. In this case we can either use second order necessary or sufficient conditions or compare the corresponding objective values in order to judge which "candidate" for a solution is really optimal.

Example 2.48 *Solve the problem*

$$\min\{(x_1)^2+2(x_2)^2+(x_3)^2\}$$

$$\begin{aligned} x_1+x_2-\ln x_3 &\geq 1, \\ x_3 &\geq 1. \end{aligned}$$

by using the KKT theorem.

Solution. First we reformulate the problem in the form as in the KKT theorem:

$$\min\{(x_1)^2+2(x_2)^2+(x_3)^2\}$$

$$-x_1-x_2+\ln x_3+1 \leq 0 \tag{2.36}$$

$$-x_3+1 \leq 0. \tag{2.37}$$

This problem has a solution since the level sets $\{x\in\mathbf{R}^3: \ f(x)\leq c\}$ of the objective function are compact and the existence theorem from Chapter 1 is applicable. Every optimal solution is a part of a KKT point.

First we search for KKT points with $\lambda_0=1$. The Lagrange function is

$$L(x,\mu)=(x_1)^2+2(x_2)^2+(x_3)^2+\mu_1(-x_1-x_2+\ln x_3+1)+\mu_2(-x_3+1).$$

The KKT conditions consist of the equations

$$\partial_{x_1}L=2x_1-\mu_1=0, \tag{2.38}$$

$$\partial_{x_2}L=4x_2-\mu_1=0, \tag{2.39}$$

$$\partial_{x_3}L=2x_3+\frac{\mu_1}{x_3}-\mu_2=0, \tag{2.40}$$

$$\mu_1(-x_1-x_2+\ln x_3+1)=0, \tag{2.41}$$

$$\mu_2(-x_3+1)=0, \tag{2.42}$$

the inequalities (2.36), (2.37), and $\mu_1\geq 0$, $\mu_2\geq 0$.

We consider the following four cases.

(i) $\mu_1=\mu_2=0$. Then from (2.38)–(2.40) $x_1=x_2=x_3=0$, which is not a feasible point due to (2.37).

(ii) $\mu_1 = 0$, $\mu_2 > 0$. Then from (2.42) $x_3 = 1$ and from (2.38), (2.39) $x_1 = x_2 = 0$. This is not a feasible point due to (2.36).

(iii) $\mu_1 > 0$, $\mu_2 = 0$. Then (2.38)–(2.41) give

$$2x_1 - \mu_1 = 0,$$

$$4x_2 - \mu_1 = 0,$$

$$2x_3 + \frac{\mu_1}{x_3} = 0,$$

$$x_1 + x_2 - \ln x_3 = 1.$$

The third equation gives $\mu_1 = -2(x_3)^2$, which contradicts the inequality $\mu_1 > 0$. Thus we do not obtain a KKT point also in this case.

(iv) $\mu_1 > 0$, $\mu_2 > 0$. From (2.42) we get $x_3 = 1$. Equations (2.38), (2.39) implay $x_1 = 2x_2$ and (2.41) implies $x_1 + x_2 = 1$. Then $x_1 = \frac{2}{3}$, $x_2 = \frac{1}{3}$. Then $\mu_1 = \frac{4}{3} > 0$ and $\mu_2 = \frac{10}{3} > 0$ and we obtain a KKT point with $x = (\frac{2}{3}, \frac{1}{3}, 1)'$.

Now we try to find abnormal KKT points (with $\lambda_0 = 0$). According to Remark 2.29, x can be a part of such a point only if (2.26) fails. We have

$$\partial h(x) = \begin{pmatrix} -1 & -1 & \frac{1}{x_3} \\ 0 & 0 & -1 \end{pmatrix},$$

which has rank $= 2$ for every $x_3 \geq 1$. Thus the surjectivity condition is fulfilled and there are no abnormal KKT points.

Since the only KKT point is $x = (\frac{2}{3}, \frac{1}{3}, 1)'$ and since the problem has a solution, this point is the unique optimal solution.

Exercise 2.49 *([4, p. 34]) The KKT conditions are not sufficient, in general!* Consider the problem

$$\max\{(x-1)^3\}$$

subject to the constraints

$$-x \leq 0, \quad x - 2 \leq 0.$$

Prove that KKT system has a solution with $x = 1$, but this is not a locally optimal solution.

Exercise 2.50 Find all KKT points (normal and abnormal) of the problem

$$\max\left\{(x_1+1)^2+(x_2+1)^2\right\}$$

subject to the constraints

$$(x_1)^2+(x_2)^2\leq 5, \quad x_1\leq 1.$$

Which of the KKT points correspond to global solutions of the problem and why?

Exercise 2.51 Solve the following problem using the KKT theorem:

$$\max\left\{\ln(5-x_1+5x_2)\right\}$$

subject to the constraints

$$(x_1-1)^2+4(x_2)^2\leq 1, \quad x_1+x_2\leq 2.$$

Exercise 2.52 *([3, p. 372])* Solve the problem

$$\min\left\{x_1\right\}$$

subject to the constraints

$$(x_1-4)^2+(x_2)^2\leq 16,$$

$$(x_1-3)^2+(x_2-2)^2=13$$

using the KKT theorem. Illustrate the set K and the solution graphically.

Exercise 2.53 Find all KKT points (normal and abnormal) of the problem

$$\min\left\{(x_1)^2+(x_2)^2\right\}$$

subject to the constraints

$$x_1\geq 1,$$

$$(x_1)^2+4(x_2)^2\geq 1.$$

Does the problem have a solution, and which of the KKT points correspond to global solutions?

Chapter 3

Convex Optimizations Problems

Convex optimization problems are those of minimizing a convex function (or maximizing a concave function) on a convex set. These problems have several features that allow to obtain stronger results than in the general optimization problem considered in the previous chapter. Among these features we mention the following ones:

(i) for convex functions the local minimizers are global;

(ii) the tangent and the normal cones to convex sets can be conveniently characterized;

(iii) the characterization of the minimizers in terms of Lagrange functions can be done without assuming differentiability;

(iv) The elegant "duality theory" becomes really meaningful and applicable for convex problems.

We begin with a short introduction to convex functions. Then in Section 3.2 we prove some stronger versions of the KKT theorem. In Sections 3.3 and 3.4 we present the important concepts of *saddle point* ("Sattelpunkt") and *duality* ("Dualität"). Additional examples and exercises are given in Section 3.5.

3.1 Convex functions

Let $K \subset \mathbf{R}^n$ be a convex set.

Definition 3.1 The function $f : K \to \mathbf{R}$ is called *convex* if for every two different points x_1, $x_2 \in K$ and for every number $\alpha \in (0,1)$ it holds that

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2).$$

If the last inequality is strict, then f is called *strictly convex.* The function f is called *concave* if $(-f)$ is convex (that is, f satisfies the reverse inequality in Definition 3.1).

Obviously all affine functions $f(x) = \langle l, x \rangle + b$, with $l \in \mathbf{R}^n$, $b \in \mathbf{R}$ are convex. It is also obvious that if $f_1, f_2 : \mathbf{R}^n \to \mathbf{R}$ are two convex functions, then $\beta f_1 + \gamma f_2$ with $\beta, \gamma \geq 0$ is also convex (**check this using the definition!**).

A quadratic function $f(x) := \langle Ax, x \rangle$ (where A is a symmetric $(n \times n)$-matrix) is convex if and only if the matrix A is non-negative definite, that is $\langle Ax, x \rangle \geq 0$ for every $x \in \mathbf{R}^n$. A necessary and sufficient condition for non-negative definiteness of A is known from linear algebra (the Sylvester criterion): all principle minors are non-negative.

Even more, if the function $f : \mathbf{R}^n \to \mathbf{R}$ has continuous second partial derivatives $\frac{\partial^2 f}{\partial x_i \partial x_j}$, then f is convex if and only if the second total derivative of f,

$$\partial_{xx} f(x) := \frac{\partial^2 f(x)}{\partial x^2} := \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \cdots & \cdots & \cdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{pmatrix},$$

is non-negative definite for every $x \in \mathbf{R}^n$. The last two assertions need proofs, which we skip.

The next lemma claims that the supremum of convex functions is also convex.

Lemma 3.2 *Let A be an arbitrary set and let $f : A \times \mathbf{R}^n \to \mathbf{R}$ be a given function such that for every $a \in A$ the function $f(a, \cdot) : \mathbf{R}^n \to \mathbf{R}$ is convex. Define*

$$\hat{f}(x) := \sup_{a \in A} f(a, x), \qquad K := \{x \in \mathbf{R}^n : \; \hat{f}(x) < +\infty\}.$$

Then the set K is convex and $\hat{f} : K \to \mathbf{R}$ is a convex function.

Proof. Let x_1, $x_2 \in K$, $\alpha \in (0,1)$ and $x = \alpha x_1 + (1-\alpha) x_2$. Then for every $a \in A$ we have

$$f(a, x) \leq \alpha f(a, x_1) + (1-\alpha) f(a, x_2) \leq \alpha \hat{f}(x_1) + (1-\alpha) \hat{f}(x_2).$$

Taking the supremum in $a \in A$ we obtain

$$\hat{f}(x) \leq \alpha \hat{f}(x_1) + (1-\alpha) \hat{f}(x_2),$$

which proves both claims of the lemma. Q.E.D.

Notice that the maximum of differentiable functions does not need to be differentiable. For example, $f(x) := |x|$, $x \in \mathbf{R}$, is not differentiable at $x = 0$ although $|x| = \max\{x, -x\}$ and both $f_1(x) = x$ and $f_2(x) = -x$ are linear, hence differentiable and convex.

The convex functions have the following nice property in the context of the optimization problem

$$\min_{x \in K} f(x). \tag{3.1}$$

Theorem 3.3 *Assume that the set K is convex and the objective function $f : K \to \mathbf{R}$ is convex. Then every local solution of (3.1) is a global solution. If f is strictly convex, then there is at most one local solution and if it exists, it is the unique global solution.*

Proof. Let $x^* \in K$ be a local solution. Then there exists a neighborhood $\mathcal{O}$ of x^* such that

$$f(x) \geq f(x^*) \quad \forall\, x \in K \cap \mathcal{O}.$$

Take an arbitrary point $x \in K$. Then $x_\alpha := \alpha x + (1-\alpha)x^* \in K$ for every $\alpha \in (0,1)$. Even more, $x_\alpha \in \mathcal{O}$ for all sufficiently small $\alpha > 0$. Then using the local optimality of x^* and the convexity of f we have

$$f(x^*) \leq f(x_\alpha) \leq \alpha f(x) + (1-\alpha) f(x^*).$$

Hence $\alpha f(x^*) \leq \alpha f(x)$ and since $\alpha > 0$ and $x \in K$ is arbitrary, we obtain that x^* is globally optimal.

Now let f be strictly convex and let x_1^* and x_2^* be two different local solutions. According to the already proved claim, both x_1^* and x_2^* are global solutions. In particular, $f(x_1^*) = f(x_2^*)$. For the point $x = \frac{1}{2}(x_1^* + x_2^*)$ we have (due to the strict convexity of f)

$$f(x) < \frac{1}{2} f(x_1^*) + \frac{1}{2} f(x_2^*) = f(x_1^*),$$

which contradicts the global optimality of x_1^*. Q.E.D.

In the sequel we need the following two facts.

Lemma 3.4 *Let $f : \mathbf{R}^n \to \mathbf{R}$ be convex. If f is differentiable at a point x, then*

$$f(y) \geq f(x) + \partial f(x)(y - x) \quad \forall\, y \in \mathbf{R}^n.$$

Proof. The definition of convexity can be reformulated as

$$f(x + \alpha(y-x)) \leq f(x) + \alpha(f(y) - f(x)) \quad \forall\, y \in \mathbf{R}^n, \ \ \forall\, \alpha \in (0,1).$$

Then

$$\frac{f(x + \alpha(y-x)) - f(x)}{\alpha} \leq f(y) - f(x).$$

Passing to the limit with $\alpha \to 0$ we obtain $\partial f(x)(y-x) \le f(y) - f(x)$, which had to be proved. Q.E.D.

As a consequence of this lemma we obtain the following.

Corollary 3.5 *If the function $f : \mathbf{R}^n \to \mathbf{R}$ is convex and differentiable, then x^* is a global minimizer of f in $\mathbf{R}^n$ if and only if $\nabla f(x^*) = 0$.*

3.2 The Kuhn-Tucker theorem for convex problems

Consider again the problem

$$\min_{x \in K} f(x), \tag{3.2}$$

where

$$K := \{x \in \mathbf{R}^n : \; g(x) = 0, \; h(x) \le 0\} \tag{3.3}$$

and $f : \mathbf{R}^n \to \mathbf{R}$, $g : \mathbf{R}^n \to \mathbf{R}^m$, $h : \mathbf{R}^n \to \mathbf{R}^r$ are given functions. As everywhere in this script, the vector inequality $h \le 0$ is understood componentwise.

If the functions g_i are affine and the functions h_j are convex, then the set K is convex. The affinity of $g_i(x)$ means that $g_i(x) = G_i x - b_i$, where G_i is an $(1 \times n)$-dimensional row-vector and $b_i \in \mathbf{R}$. In a matrix form, $g(x) = Gx - b$, where G is the $(m \times n)$-matrix with rows G_i, and $b = (b_1, \ldots, b_m)'$.

Indeed, if x^1, $x^2 \in K$ and $\alpha \in [0,1]$, then

$$g(\alpha x^1 + (1-\alpha)x^2) = G(\alpha x^1 + (1-\alpha)x^2) - b = \alpha(Gx^1 - b) + (1-\alpha)(Gx^2 - b) = 0 + 0 = 0$$

and

$$h(\alpha x^1 + (1-\alpha)x^2) \le \alpha h(x^1) + (1-\alpha)h(x^2) \le 0 + 0 = 0.$$

Hence, $\alpha x^1 + (1-\alpha)x^2 \in K$ and K is convex.

Let us look at the claims and the assumptions of the KKT Theorem 2.25 in the case of affine g and convex f and h. In this case for $\mu \ge 0$ the Lagrange function $L(x, \lambda, \mu)$ is convex with respect to x. Then according to Corollary 3.5 equation (2.19) is equivalent to the relation

$$L(x^*, \lambda, \mu) \le L(x, \lambda, \mu) \quad \forall\, x \in \mathbf{R}^n.$$

Notice that derivatives do not appear in this relation.

Moreover, the existence of $\bar{l}$ such that $\partial g(x^*)\,\bar{l} = 0$ and $\partial h_j(x^*)\,\bar{l} < 0$ for $j \in J(x^*)$ (that appears as a part of the assumptions in the second claim of the KKT theorem) is implied by the existence of $\bar{x} \in K$ such that $h_j(\bar{x}) < 0$ for $j \in J(x^*)$. (Notice that the latter condition does not involve derivatives.) Indeed, we may define $\bar{l} = \bar{x} - x^*$ and then due to the affinity of g and due to the convexity of h we have

$$\partial g(x^*)\,\bar{l} = G\,(\bar{x} - x^*) = b - b = 0,$$

$$\partial h_j(x^*)\,\bar{l} = \partial h_j(x^*)\,(\bar{x} - x^*) \leq h_j(\bar{x}) - h_j(x^*) = h_j(\bar{x}) < 0,$$

where in the last inequality we use Lemma 3.4 and the equality $h_j(x^*) = 0$ for $j \in J(x^*)$.

The above considerations raise the question whether the claim of the KKT theorem can be reformulated for convex problems without assuming differentiability at all. We shall elaborate this issue in the rest of this section.

To bring some more flexibility of our results we consider an a bit more general problem. First we give the following definition.

Definition 3.6 A set $K_0 \subset \mathbf{R}^n$ is called *polyhedral* if there is an $(s \times n)$-matrix A^0 (s is an arbitrary natural number) and $d^0 \in \mathbf{R}^s$ such that $K_0 = \{x \in \mathbf{R}^n : \ A^0 x \ \leq d^0\}$.

Consider the problem

$$\min_{x \in K \cap K_0} f(x), \tag{3.4}$$

where $K_0 \subset \mathbf{R}^n$ is a polyhedral set and $K \subset \mathbf{R}^n$ is defined by the system of equalities and inequalities

$$G_i x - b_i = 0, \quad i = 1, \ldots, m, \tag{3.5}$$

$$H_j x - c_j \leq 0, \quad j \in J^a := \{1, \ldots, r'\}, \tag{3.6}$$

$$h_j(x) \leq 0, \qquad j \in J^c := \{r' + 1, \ldots, r\}, \tag{3.7}$$

where $0 \leq r' \leq r$ and the functions f and h_j, $j \in J^c$, are convex. Thus the equality constraints $g(x) = 0$ become affine ($g(x) = Gx - b$) and some of the inequality constraints (with indexes $j \in J^a$) are also affine ($h_j(x) = H_j x - c_j$).

The assumptions that the set K_0 is polyhedral, and that f and h_j, $j \in J^c$, are convex will be *standing* throughout this chapter.

Now we introduce an additional condition that can be viewed as a replacement of the Mangasarian-Fromowitz condition in the KKT Theorem 2.25.

Assumption S (Slater condition): For every $j \in J^c$ there exists $\tilde{x}^j \in K \cap K_0$ such that $h_j(\tilde{x}^j) < 0$.

Remark 3.7 It is important to notice that Assumption S is equivalent to the following one:

$$\exists\, \tilde{x} \in K \cap K_0 \;\text{ such that } h_j(\tilde{x}) < 0 \;\; \forall\, j \in J^c. \tag{3.8}$$

Indeed, if the Slater condition holds, then one may define $\tilde{x} := \frac{1}{|J^c|} \sum_{j \in J^c} \tilde{x}^j$ and show that (3.8) holds true thanks to the convexity of h_j and of the sets K_0 and K. On the other hand, if (3.8) is fulfilled, then one may take $\tilde{x}^j = \tilde{x}$ and the Slater condition holds.

The main contributor to the next theorem (which we call *KKT Theorem for convex problems*) is Fritz John (1948).

Theorem 3.8 *Consider problem (3.4)–(3.7) under the standing assumptions and Assumption S. Then x^* is an optimal solution if and only if there exist vectors $\lambda \in \mathbf{R}^m$ and $\mu \in \mathbf{R}^r$ such that the following relations are satisfied:*

$$L(x^*, \lambda, \mu) \le L(x, \lambda, \mu) \quad \forall\, x \in K_0, \tag{3.9}$$

$$\mu' \, h(x^*) = 0, \tag{3.10}$$

$$x^* \in K \cap K_0, \tag{3.11}$$

$$\mu \ge 0. \tag{3.12}$$

Remark 3.9 As in the previous chapter, the equality $\mu' h(x^*) = 0$ is equivalent to $\mu_j h_j(x^*) = 0$ for every $j = 1, \ldots, r$, due to the inequalities $h_j(x^*) \le 0$ and $\mu_j \ge 0$.

Proof. The sufficiency of (3.9)–(3.12) for optimality of x^* is almost evident. First of all $x^* \in K \cap K_0$. Moreover, for every $x \in K \cap K_0$

$$\begin{aligned} f(x^*) &= f(x^*) + \lambda' \underbrace{g(x^*)}_{=0} + \underbrace{\mu' h(x^*)}_{=0} = L(x^*, \lambda, \mu) \le L(x, \lambda, \mu) \\ &= f(x) + \lambda' \underbrace{g(x)}_{=0} + \underbrace{\mu' h(x)}_{\le 0} \le f(x). \end{aligned}$$

The proof of necessity will be split into several parts. In the first part we shall prove the necessity under a simplifying condition and with $K_0 = \mathbf{R}^n$. Using this, in Part 2 we prove the necessity without the simplifying conditions, but still with $K_0 = \mathbf{R}^n$. Finally, in Part 3 also the assumption $K_0 = \mathbf{R}^n$ is eliminated.

Part 1. In this part we will prove the theorem in the case $K_0 = \mathbf{R}^n$ under the following simplifying condition:

Simplifying constraint qualification. If the equality

$$\sum_{i=1}^{m} \lambda_i G_i + \sum_{j=1}^{r'} \mu_j H_j = 0$$

is fulfilled for some $\lambda \in \mathbf{R}^m$ and some $\mu \in \mathbf{R}^r_+$, then $\lambda = 0$ and $\mu = 0$.

Let x^* be a solution of $\min_{x \in K} f(x)$. We shall use the abbreviations $I := \{1, \dots, m\}$, $J := J^a \cup J^c = \{1, \dots, r\}$. As before, $J(x^*)$ is the set of all indexes $j \in J$ for which $h_j(x^*) = 0$ (the active constraints).

Consider the set

$$C := \{(\alpha, \beta_i, \gamma_j)_{i \in I, j \in J} : \; \alpha > f(x) - f(x^*), \; \beta_i = g_i(x), \; \gamma_j \geq h_j(x) \; \text{ for some } x \in \mathbf{R}^m\}.$$

Clearly (by the same argument as in the beginning of the present section) the set $C \subset \mathbf{R} \times \mathbf{R}^m \times \mathbf{R}^r$ is convex. Moreover, $0 \notin C$, since the opposite would imply that for some x

$$0 > f(x) - f(x^*), \quad 0 = g_i(x), \quad 0 \geq h_j(x),$$

which contradicts the optimality of x^*.

According to the Hahn-Banach Theorem 2.18, there is a non-zero triplet $(\lambda_0, \lambda, \mu)$ with $\lambda_0 \in \mathbf{R}$, $\lambda \in \mathbf{R}^m$, $\mu \in \mathbf{R}^r$ such that

$$\lambda_0 \alpha + \sum_{i \in I} \lambda_i \beta_i + \sum_{j \in J} \mu_j \beta_j \; \geq 0 \qquad \forall \, (\alpha, \beta, \gamma) \in C. \tag{3.13}$$

Due to the unboundedness from above of α and γ_j, it must hold that $\lambda_0 \geq 0$ and $\mu_j \geq 0$, that is, (3.12) holds. Moreover, taking $x = x^*$ in the definition of the set C we have that for every $\varepsilon > 0$

$$(\varepsilon, g_i(x^*), h_j(x^*)) = (\varepsilon, 0, h_j(x^*)) \in C,$$

hence,

$$\varepsilon \lambda_0 + \sum_{j \in J} \mu_j h_j(x^*) \geq 0.$$

This holds for an arbitrary $\varepsilon > 0$, therefore it holds also with $\varepsilon = 0$. Then

$$\sum_{j \in J} \mu_j h_j(x^*) \geq 0.$$

Since $\mu_j \geq 0$ and $h_j(x^*) \leq 0$, each summand in the above sum must be equal to zero. Thus (3.10) is satisfied.

The next step is to prove that $\lambda_0 > 0$. Assume that $\lambda_0 = 0$. Then due to the definition of C and (3.13) we have for every $x \in \mathbf{R}^n$ that

$$\sum_{i \in I} \lambda_i g_i(x) + \sum_{j \in J} \mu_j h_j(x) \geq 0. \tag{3.14}$$

In particular, for the vector $\bar{x}$ from the Slater condition we have $g_i(\bar{x}) = 0$, $h_j(\bar{x}) \leq 0$ for $j \in J$ and $h_j(\bar{x}) < 0$ for $j \in J^c$, hence,

$$\sum_{j \in J^a} \mu_j h_j(\bar{x}) + \sum_{j \in J^c} \mu_j h_j(\bar{x}) \geq 0.$$

Since the first sum is non-positive, and since $h_j(\bar{x}) < 0$, $j \in J^c$, the above inequality implies that $\mu_j = 0$ for every $j \in J^c$. Then (3.14) becomes

$$\sum_{i \in I} \lambda_i (G_i x - b_i) + \sum_{j \in J^a} \mu_j (H_j x - c_j) \geq 0.$$

Rearranging the terms we have that

$$\left(\sum_{i \in I} \lambda_i G_i + \sum_{j \in J^a} \mu_j H_j \right) x \;\geq\; \sum_{i \in I} \lambda_i b_i + \sum_{j \in J^a} \mu_j c_j.$$

Since the above inequality holds for every $x \in \mathbf{R}^n$ we must have that

$$\sum_{i \in I} \lambda_i G_i + \sum_{j \in J^a} \mu_j H_j = 0.$$

According to the simplifying constraint qualification the above equality implies $\lambda_i = 0$, $\mu_j = 0$. Thus we obtained that if $\lambda_0 = 0$, then $\lambda = 0$ and $\mu = 0$, which is a contradiction with $(\lambda_0, \lambda, \mu) \neq 0$. Hence $\lambda_0 > 0$ and we may assume that $\lambda_0 = 1$, since we can divide λ and μ by λ_0 and still have (3.13) fulfilled.

It remains to prove (3.9). For every $x \in \mathbf{R}^n$ and $\varepsilon > 0$ we have that

$$(\varepsilon + f(x) - f(x^*),\; g_i(x),\; h_j(x)) \in C.$$

From (3.13) (with $\lambda_0 = 1$) we obtain

$$\varepsilon + f(x) - f(x^*) + \sum_{i \in I} \lambda_i g_i(x) + \sum_{j \in J} \mu_j h_j(x) \;\geq\; 0.$$

Hence, taking into account that $\varepsilon > 0$ is arbitrary, we have

$$f(x) + \sum_{i \in I} \lambda_i g_i(x) + \sum_{j \in J} \mu_j h_j(x) \;\geq\; f(x^*) = f(x^*) + \sum_{i \in I} \lambda_i \underbrace{g_i(x^*)}_{=0} + \sum_{j \in J} \underbrace{\mu_j h_j(x^*)}_{=0}$$

and (3.9) is proved.

Part 2. Now we shall prove the theorem again for $K_0 = \mathbf{R}^n$, but without the simplifying condition. The proof consists of four sub-parts

Part 2.1: Reduction of the constraints. Let x^* be a solution of $\min_{x\in K} f(x)$. In this sub-part we shall pass to a new problem in which a part of the linear constrains are eliminated while the solution remains the same, x^*.

Let us fix a set of indexes $\tilde{I} \in I$ such that the rows of G with indexes $i \in \tilde{I}$ are linearly independent and have maximal rank ($= \operatorname{rank} G$). Denote the resulting matrix by $G_{\tilde{I}}$. Thus the rows of $G_{\tilde{I}}$ are linearly independent and each row of G is a linear combination of some rows of $G_{\tilde{I}}$.

Now we shall select a set of indexes $\tilde{J} \subset J^a \cap J(x^*)$ and denote by $H_{\tilde{J}}$ the matrix consisting of the rows H_j with indexes $j \in \tilde{J}$. We do this step by step considering one after the other the active rows of H. At the beginning we set $\tilde{J} := \emptyset$. Take the first active row of H and include its index in the set $\tilde{J}$ if and only if it is not linearly dependent from the rows of $G_{\tilde{I}}$. Proceeding further, we include the index $s \in J^a \cap J(x^*)$ into $\tilde{J}$ if and only if H_s cannot be represented as a linear combination

$$H_s = -\sum_{i\in\tilde{I}} \lambda_i G_i - \sum \mu_j H_j \quad \text{with } \mu_j \geq 0, \tag{3.15}$$

where the second sum is taken only for indexes j that are already included in the set $\tilde{J}$. Proceed in this way till the last active row of H. According to this selection procedure, the obtained matrices $G_{\tilde{I}}$ and $H_{\tilde{J}}$ have the following properties:

(i) if $k \in I \setminus \tilde{I}$, then there exist real numbers λ_i^k, $i \in \tilde{I}$, such that $G_k = \sum_{i\in\tilde{I}} \lambda_i^k G_i$;

(ii) if $s \in (J^a \cap J(x^*)) \setminus \tilde{J}$, then there exist real numbers λ_i^s, $i \in \tilde{I}$, and $\mu_j^s \geq 0$, $j \in \tilde{J}$, such that $H_s = -\sum_{i\in\tilde{I}} \lambda_i^s G_i - \sum_{j\in\tilde{J}} \mu_j^s H_i$;

(iii) if the equality $\sum_{i\in\tilde{I}} \lambda_i G_i + \sum_{j\in\tilde{J}} \mu_j H_j = 0$ is fulfilled with $\lambda_i \in \mathbf{R}$ and $\mu_j \geq 0$, then $\lambda_i = 0$ and $\mu_j = 0$ for every $i \in \tilde{I}$ and $j \in \tilde{J}$.

The first two properties are directly granted by the selection procedure. To show the third one we notice that if all μ_j are zero, then we obtain a contradiction with the linear independence of G_i, $i \in \tilde{I}$, unless all the λ-s are equal to zero. If not all μ_j are zero, then we take the non-zero one with the maximal index, say s, divide the equality in (iii) by $\mu_s > 0$ and obtain a contradiction with the selection procedure (see (3.15) in its context).

Now we consider the following (reduced) problem:

$$\min f(x) \tag{3.16}$$

subject to

$$G_{\tilde{I}}x - b_{\tilde{I}} = 0, \quad H_{\tilde{J}}x - c_{\tilde{J}} \leq 0, \quad h_j(x) \leq 0,\ j \in J^c. \tag{3.17}$$

Part 2.2: The equivalence of the reduced and the original problem.

Auxiliary ("helfend") claim: x^* is an optimal solution of problem (3.16), (3.17).

Obviously x^* satisfies (3.17). Notice that due to the property (i) we have for $k \in I \setminus \tilde{I}$

$$b_k = G_k x^* = \sum_{i\in\tilde{I}} \lambda_i^k G_i x^* = \sum_{i\in\tilde{I}} \lambda_i^k b_i. \tag{3.18}$$

Similarly, according to (ii), for $s \in (J^a \cap J(x^*)) \setminus \tilde{J}$ we have

$$c_k = H_k x^* = \sum_{i\in\tilde{I}} \lambda_i^s G_i x^* + \sum_{j\in\tilde{J}} \mu_j^s H_j x^* = \sum_{i\in\tilde{I}} \lambda_i^s b_i + \sum_{j\in\tilde{J}} \mu_j^s c_j \tag{3.19}$$

with $\mu_j^s \geq 0$.

Now we shall prove that for every x that satisfies (3.17) it holds that $f(x) \geq f(x^*)$. Let x satisfy (3.17). For $k \in I \setminus \tilde{I}$ we have using (3.18)

$$G_k x = \sum_{i \in \tilde{I}} \lambda_i^k G_i x = \sum_{i \in \tilde{I}} \lambda_i^k b_i = b_k,$$

and for $s \in (J^a \cap J(x^*)) \setminus \tilde{J}$ we have using (3.19)

$$H_s x = \sum_{i \in \tilde{I}} \lambda_i^s G_i x + \sum_{j \in \tilde{J}} \mu_j^s H_i x \leq \sum_{i \in \tilde{I}} \lambda_i^s b_i + \sum_{j \in \tilde{J}} \mu_j^s c_j = c_k.$$

Thus x satisfies all active constraints of the original problem (3.2), (3.3). Hence

$$f(x) \geq f(x^*),$$

which implies that x^* is an optimal solution of problem (3.4)–(3.7) (with $K_0 = \mathbf{R}^n$). The auxiliary claim is proved.

Part 2.3: The KKT theorem for the reduced problem. Now we shall prove that the claim of the theorem holds for the auxiliary problem (3.16), (3.17). That is, that there exist $\lambda \in \mathbf{R}^{|\tilde{I}|}$ and $\mu \in \mathbf{R}^{|\tilde{J}|+|J^c|}$ such that

$$f(x^*) + \sum_{i \in \tilde{I}} \lambda_i g_i(x^*) + \sum_{j \in \tilde{J} \cup J^c} \mu_j h_j(x^*) \tag{3.20}$$

$$\leq f(x) + \sum_{i \in \tilde{I}} \lambda_i g_i(x) + \sum_{j \in \tilde{J} \cup J^c} \mu_j h_j(x) \qquad \forall x \in \mathbf{R}^n,$$

$$\mu_j \, h_j(x^*) = 0, \quad \forall j \in \tilde{J} \cup J^c, \tag{3.21}$$

$$\mu \geq 0. \tag{3.22}$$

However, this follows from the result in Part 1, since property (iii) of the reduced problem is just the simplifying condition for this problem.

Part 2.4: Back to the original problem. Now we return to the problem $\min_{x \in K} f(x)$. Define $\lambda_i = 0$ for all $i \in \{1, \dots, m\} \setminus \tilde{I}$ and $\mu_j = 0$ for all $j \in J^a \setminus \tilde{J}$. Then all relations in (3.9)–(3.12) (with $K_0 = \mathbf{R}^n$) follow from (3.20)–(3.22).

Part 3. Now let K_0 be an arbitrary closed polyhedral set. We have to prove that any solution x^* of (3.4)–(3.7) satisfies (3.9)–(3.12) together with some $\lambda \in \mathbf{R}^m$ and $\mu \in \mathbf{R}^r$. Since K_0 is a polyhedral set there are vectors $A_j^0 \in \mathbf{R}^n$, $j \in J_0$ (a set of indexes) and $d^0 \in \mathbf{R}^{|J_0|}$ such that

$$K_0 = \{x \in \mathbf{R}^n : \; A_j^0 x - d_j^0 \leq 0, \; j \in J_0\}.$$

Then we may add all the inequality constraints in the definition of K_0 to the constraints (3.6), and pass to a problem with $K_0 = \mathbf{R}^n$ also having x^* as a solution. The Slater condition remains fulfilled independently on whether the polyhedral constraints are included in K_0 or in the system of affine inequalities. Due to the already proved result there are $\lambda \in \mathbf{R}^m$, $\mu \in \mathbf{R}^r$, $\mu^0 \in \mathbf{R}^{|J_0|}$ such that

$$\begin{aligned}
&f(x^*) + \sum_{i=1}^m \lambda_i g_i(x^*) + \sum_{j=1}^r \mu_j h_j(x^*) + \sum_{j \in J_0} \mu_j^0 (A_j^0 x^* - d_j^0) \\
&\le\ f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^r \mu_j h_j(x) + \sum_{j \in J_0} \mu_j^0 (A_j^0 x - d_j^0) \qquad \forall\, x \in \mathbf{R}^n, \\
&\mu_j\, h_j(x^*) = 0 \ \ \forall j \in \{1, \ldots, r\}, \quad \mu_j^0 (A_j^0 x^* - d_j^0) = 0 \ \ \forall\, j \in J_0, \\
&\mu \ge 0, \quad \mu_0 \ge 0.
\end{aligned}$$

The last two group of relations imply, in particular, (3.10) and (3.12). It remains to prove (3.9). Take an arbitrary $x \in K_0$. We have

$$\begin{aligned}
L(x^*, \lambda, \mu) &= f(x^*) + \sum_{i=1}^m \lambda_i g_i(x^*) + \sum_{j=1}^r \mu_j h_j(x^*) \\
&= f(x^*) + \sum_{i=1}^m \lambda_i g_i(x^*) + \sum_{j=1}^r \mu_j h_j(x^*) + \sum_{j \in J_0} \underbrace{\mu_j^0 (A_j^0 x^* - d_j^0)}_{=0} \\
&\le f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^r \mu_j h_j(x) + \sum_{j \in J_0} \underbrace{\mu_j^0}_{\ge 0} \underbrace{(A_j^0 x - d_j^0)}_{\le 0} \\
&\le L(x, \lambda, \mu).
\end{aligned}$$

This completes the proof of the theorem. Q.E.D.

Remark 3.10 As Part 3 of the proof of the above theorem shows, one can include in the description of the set K_0 any part of the linear equality and inequality constraints of the problem. Usually one places in K_0 the "easy" constraints, such as constraints of the sort $a_i \le x_i \le b_i$, which do not create problems in the minimization of the Lagrange function involved in (3.9).

Remark 3.11 We stress three remarkable features of Theorem 3.8, which are missing in the KKT Theorem 2.25 in the preceding chapter. First, no surjectivity-type assumptions are needed, which is essential for the material in the next chapter. Second, no differentiability is required. In Example 3.25 in the next section we demonstrate how the theorem can be applied for a non-differentiable function f. Third, Theorem 3.8

provides optimality conditions that are both necessary and sufficient. This is one of the advantages that the linear-convex structure of the problem brings (compare with Remark 2.35).

Theorem 3.8 claims that x^* is a minimizer in the constrained problem (3.2), (3.3) if and only if it minimizes the Lagrange function in $\mathbf{R}^n$, provided that the multipliers λ and μ are appropriately chosen. In the next two sections we shall say more on the "appropriate" choice of the multipliers.

3.3 Saddle points

In this section we present the general concept of saddle points (related to the celebrated Nash equilibria) and show its connection to optimization.

Definition 3.12 Let $K_0 \subset \mathbf{R}^n$, $Y_0 \subset \mathbf{R}^k$ and $F : K_0 \times Y_0 \to \mathbf{R}$. A point $(\bar{x}, \bar{y}) \in K_0 \times Y_0$ is called *saddle point* ("Sattelpunkt") of F in $K_0 \times Y_0$ if

$$F(\bar{x}, y) \le F(\bar{x}, \bar{y}) \le F(x, \bar{y}) \quad \forall\, x \in K_0,\ \forall\, y \in Y_0.$$

Equivalently,

$$\sup_{y \in Y_0} F(\bar{x}, y) \le F(\bar{x}, \bar{y}) \le \inf_{x \in K_0} F(x, \bar{y}).$$

Since the "sup" and "inf" are achieved ("erreicht, angenommen") at $\bar{y}$ and $\bar{x}$, respectively, we may rewrite the above inequalities as

$$\max_{y \in Y_0} F(\bar{x}, y) = F(\bar{x}, \bar{y}) = \min_{x \in K_0} F(x, \bar{y}). \tag{3.23}$$

(As usual, writing "max" instead of "sup" means that the supremum is achieved. Similarly for "min".)

Now consider the problem

$$\min_{x \in K \cap K_0} f(x), \tag{3.24}$$

where

$$K = \{x \in \mathbf{R}^n :\ g(x) = 0,\ h(x) \le 0\}, \tag{3.25}$$

$f : \mathbf{R}^n \to \mathbf{R}$, $g : \mathbf{R}^n \to \mathbf{R}^m$, $h : \mathbf{R}^n \to \mathbf{R}^r$ are given functions, and $K_0 \subset \mathbf{R}^n$ is a closed set.

Theorem 3.8 suggests that it is reasonable to consider the restriction of the Lagrange function to the set of those $(x, (\lambda, \mu))$ for which $x \in K_0$, $(\lambda, \mu) \in Y_0$, where

$$Y_0 := \mathbf{R}^m \times \mathbf{R}^r_+, \qquad \mathbf{R}^r_+ := \{\mu \in \mathbf{R}^r :\ \mu \ge 0\}.$$

Applying Definition 3.12 to this particular case we say that $(\bar{x}, (\bar{\lambda}, \bar{\mu})) \in K_0 \times Y_0$ is a saddle point of the Lagrange function $L(x, \lambda, \mu) = f(x) + \lambda' g(x) + \mu' h(x)$ in $K_0 \times Y_0$ if

$$L(\bar{x}, \lambda, \mu) \;\le\; L(\bar{x}, \bar{\lambda}, \bar{\mu}) \;\le\; L(x, \bar{\lambda}, \bar{\mu}) \quad \forall\, x \in K_0,\ \forall\, (\lambda, \mu) \in Y_0. \tag{3.26}$$

The next lemma shows that the saddle points of the Lagrange function are exactly the KKT points for problem (3.24). We remind (following Section 3.2) that by definition, $(\bar{x}, \bar{\lambda}, \bar{\mu})$ is a KKT point for (3.24) if

$$L(\bar{x}, \bar{\lambda}, \bar{\mu}) \le L(x, \bar{\lambda}, \bar{\mu}), \quad \forall x \in K_0, \tag{3.27}$$

$$\bar{x} \in K \cap K_0, \tag{3.28}$$

$$(\bar{\lambda}, \bar{\mu}) \in Y_0, \tag{3.29}$$

$$\bar{\mu}'\, h(\bar{x}) = 0. \tag{3.30}$$

Lemma 3.13 *The point $(\bar{x}, (\bar{\lambda}, \bar{\mu})) \in K_0 \times Y_0$ is a saddle point of L in $K_0 \times Y_0$ if and only if it is a KKT point for (3.24).*

Proof. Let $(\bar{x}, (\bar{\lambda}, \bar{\mu})) \in K_0 \times Y_0$ be a saddle point of L in $K_0 \times Y_0$. Inequality (3.27) coincides with the second inequality in (3.26). Inclusion (3.29) and $\bar{x} \in K_0$ are included in the definition of saddle point. To prove that $\bar{x} \in K$ and equality (3.30) we rewrite the left inequality in (3.26) as

$$(\lambda - \bar{\lambda})' g(\bar{x}) + (\mu - \bar{\mu})' h(\bar{x}) = L(\bar{x}, \lambda, \mu) - L(\bar{x}, \bar{\lambda}, \bar{\mu}) \le 0 \quad \forall\, (\lambda, \mu) \in \mathbf{R}^m \times \mathbf{R}^r_+.$$

Taking $\mu = \bar{\mu}$ we have $(\lambda - \bar{\lambda})' g(\bar{x}) \le 0$ for all $\lambda \in \mathbf{R}^m$, which is only possible if $g(\bar{x}) = 0$. Taking $\lambda = \bar{\lambda}$ we have that $(\mu - \bar{\mu})' h(\bar{x}) \le 0$ for every $\mu \in \mathbf{R}^r_+$, which is only possible if $h(\bar{x}) \le 0$. In addition, if $h_j(\bar{x}) < 0$ and $\bar{\mu}_j > 0$, then the inequality $(\mu - \bar{\mu})' h(\bar{x}) \le 0$ fails with $\mu = 0$. Thus $(\bar{\mu})' h(\bar{x}) = 0$. We obtained (3.28) and (3.30).

Now, let $(\bar{x}, (\bar{\lambda}, \bar{\mu}))$ be a KKT point for (3.24). Then $(\bar{\lambda}, \bar{\mu}) \in Y_0$ and $\bar{x} \in K_0$. Moreover, the second inequality in (3.26) coincides with (3.27). Finally, for every $(\lambda, \mu) \in Y_0$, we have

$$L(\bar{x}, \lambda, \mu) \;=\; f(\bar{x}) + \lambda' \underbrace{g(\bar{x})}_{=0} + \underbrace{\mu' h(\bar{x})}_{\le 0} \;\le\; f(\bar{x}) \;=\; f(\bar{x}) + \bar{\lambda}' \underbrace{g(\bar{x})}_{=0} + \underbrace{\bar{\mu}' h(\bar{x})}_{=0} \;=\; L(\bar{x}, \bar{\lambda}, \bar{\mu}),$$

thus the left inequality in (3.26) holds. Q.E.D.

We stress that at this level of generality (nothing is assumed in this section about the functions f, g, h) the KKT conditions are no longer necessary for optimality of $\bar{x}$. The next theorem shows, however, that the KKT conditions (3.27)–(3.30) are *sufficient optimality conditions* ("hinreichende Optimalitätsbedingungen").

Theorem 3.14 *Let $(x^*, (\lambda^*, \mu^*))$ be a saddle point of the Lagrange function $L(x, \lambda, \mu)$ in $K_0 \times Y_0$ (or equivalently, a KKT point). Then x^* is a globally optimal solution of the optimization problem (3.24), (3.25).*

Proof. As it is remarked after the proof of the sufficiency in Theorem 3.8, this proof is valid for arbitrary functions f, g and h, thus the present theorem holds true. Below, we repeat this simple proof for convenience. Since $(x^*, (\lambda^*, \mu^*))$ is a KKT point, we have $x^* \in K \cap K_0$. To prove the optimality of x^* we take an arbitrary $x \in K \cap K_0$. Due to (3.30) and (3.27) we have

$$\begin{aligned} f(x^*) &= f(x^*) + (\lambda^*)' \underbrace{g(x^*)}_{=0} + \underbrace{(\mu^*)'h(x^*)}_{=0} = L(x^*, \lambda^*, \mu^*) \\ &\le L(x, \lambda^*, \mu^*) = f(x) + (\lambda^*)' \underbrace{g(x)}_{=0} + \underbrace{(\mu^*)'}_{\ge 0} \underbrace{h(x)}_{\le 0} \le f(x). \end{aligned}$$

This proves the global optimality of x^*. Q.E.D.

The above theorem raises the following question: is it true that for any global solution x^* of problem (3.24), (3.25) there exists a pair $(\lambda^*, \mu^*) \in Y_0$ such that $(x^*, (\lambda^*, \mu^*))$ is a saddle point of the Lagrange function? For a problem for which the answer is positive, Theorem 3.14 provides not only a sufficient, but also a necessary optimality condition.

The answer of the above question is not positive, in general. An example is given in Exercise 3.27 in Section 3.5.

However, the answer is positive for problems with linear-convex structure as in Theorem 3.8.

Theorem 3.15 *Consider the problem*

$$\min_{K \cap K_0} f(x)$$

where $K \subset \mathbf{R}^n$ is defined by the constraints

$$G_i x - b_i = 0, \quad i = 1, \dots, m, \tag{3.31}$$

$$H_j x - c_j \le 0, \quad j = 1, \dots, r', \tag{3.32}$$

$$h_j(x) \le 0, \quad j = r' + 1, \dots, r, \tag{3.33}$$

$0 \le r' \le r$, and K_0 is a polyhedral set. Assume that the functions f and h_j are convex and that there exists $\tilde{x} \in K_0$ satisfying (3.31) and (3.32), and also satisfying (3.33) as strict inequalities (the Slater condition).

Then x^* is an optimal solution if and only if there exists $(\lambda^*, \mu^*) \in Y_0$ such that $(x^*, (\lambda^*, \mu^*))$ is a saddle point of the Lagrange function $L(x, \lambda, \mu)$ in $K_0 \times Y_0$.

Proof. If x^* is an optimal solution, then Theorem 3.8 claims existence of (λ^*, μ^*) such that $(x^*, (\lambda^*, \mu^*))$ is a KKT point for the problem. Then we apply Lemma 3.13. Q.E.D.

3.4 Duality

Let us consider again the problem (that we call further "Problem $\mathcal{P}_0$")

Problem $\mathcal{P}_0$: $$\min_{x \in K \cap K_0} f(x),$$

where
$$K = \{x \in \mathbf{R}^n : \ g(x) = 0, \ h(x) \le 0\}, \tag{3.34}$$
$f : \mathbf{R}^n \to \mathbf{R}$, $g : \mathbf{R}^n \to \mathbf{R}^m$, $h : \mathbf{R}^n \to \mathbf{R}^r$ are given functions and $K_0 \subset \mathbf{R}^n$ is a closed set.

In the previous section we showed the importance of the saddle points of the Lagrange function in the context of the above problem. We recall the definition of a saddle point of $L(x, \lambda, \mu)$, using as before the notation $Y_0 := \mathbf{R}^m \times \mathbf{R}^r_+$. By definition, the point $(\bar{x}, (\bar{\lambda}, \bar{\mu}))$ is a saddle point of L in $K_0 \times Y_0$ if $(\bar{x}, (\bar{\lambda}, \bar{\mu})) \in K_0 \times Y_0$ and
$$\sup_{(\lambda,\mu) \in Y_0} L(\bar{x}, \lambda, \mu) \ \le \ L(\bar{x}, \bar{\lambda}, \bar{\mu}) \ \le \ \inf_{x \in K_0} L(x, \bar{\lambda}, \bar{\mu}).$$

These inequalities suggest to search for a saddle point by minimizing with respect to $\bar{x}$ the leftmost quantity and maximizing with respect to $(\bar{\lambda}, \bar{\mu})$ the rightmost quantity. Therefore, we define the functions
$$P(x) := \sup_{(\lambda,\mu) \in Y_0} L(x, \lambda, \mu), \qquad D(\lambda, \mu) := \inf_{x \in K_0} L(x, \lambda, \mu).$$

and consider the two problems

Problem $\mathcal{P}$ (primal problem): $$\min_{x \in K_0} P(x),$$

Problem $\mathcal{D}$ (dual problem): $$\max_{(\lambda,\mu) \in Y_0} D(\lambda, \mu).$$

Notice that $P(x)$ may take the "value" $+\infty$ and $D(\lambda,\mu)$ may take the "value" $-\infty$. This will not lead to any confusion if we interpret $-\infty$ and $+\infty$ as symbols and extend the relations $<$, $\leq$, $>$ and $\geq$ in the usual way. For example, $-\infty < a < \infty$ for every $a \in \mathbf{R}$.

Moreover, the "sup" and the "inf" in the definitions of the functions P and D need not be achieved. As it was mentioned before, if the supremum is achieved, then we write "max" instead of "sup". Similar convention applies to the infimum.

We also mention that for every $x \in K_0$ and $(\lambda,\mu) \in Y_0$

$$D(\lambda,\mu) \leq L(x,\lambda,\mu) \leq P(x),$$

hence

$$D(\lambda,\mu) \leq P(x) \quad \forall\, x \in K_0,\ \forall\,(\lambda,\mu) \in Y_0. \tag{3.35}$$

Remark 3.16 It is also important to notice that according to Lemma 3.2 the function D is concave on the set where it is not $-\infty$, hence every local maximizer is a global maximizer. In particular, if D is differentiable at some point $(\lambda,\mu) \in Y_0$ and $\partial D(\lambda,\mu) = 0$ then this point is a global maximizer of D.

First of all we shall prove that problems $\mathcal{P}_0$ and $\mathcal{P}$ are equivalent.

Lemma 3.17 *It holds that*

$$P(x) = \begin{cases} f(x) & \text{if } x \in K, \\ +\infty & \text{if } x \notin K. \end{cases}$$

In particular, the point $x^* \in K_0$ is a solution of Problem $\mathcal{P}$ if and only if it is a solution of Problem $\mathcal{P}_0$.

Proof. If $x \notin K$, then either there is some index i_0 for which $g_{i_0}(x) \neq 0$, or some index j_0 for which $h_{j_0}(x) > 0$. Then from the definition of $P(x)$ we have, respectively (by taking μ or λ equal to zero)

$$P(x) \geq \sup_{\lambda_i \in \mathbf{R}} \{f(x) + \lambda_i \underbrace{g_i(x)}_{\neq 0}\} = +\infty \qquad (\text{take } \mu = 0 \text{ and } \lambda_i = 0 \text{ for } i \neq i_0),$$

or

$$P(x) \geq \sup_{\mu_j \geq 0} \{f(x) + \mu_j \underbrace{h_j(x)}_{>0}\} = +\infty \qquad (\text{take } \lambda = 0 \text{ and } \mu_j = 0 \text{ for } j \neq j_0).$$

If $x \in K$ then

$$P(x) = \sup_{(\lambda,\mu)\in Y_0} \{f(x) + \lambda' \underbrace{g(x)}_{=0} + \underbrace{\mu' h(x)}_{\leq 0}\} \leq f(x) = L(x,(0,0)) \leq P(x).$$

Thus the above inequalities are satisfied as equalities. Q.E.D.

The next theorem reveals how Problems $\mathcal{P}$ and $\mathcal{D}$ are related to the saddle points of the Lagrange function associated with Problem $\mathcal{P}_0$.

Theorem 3.18 *Let $f : \mathbf{R}^n \to \mathbf{R}$, $g : \mathbf{R}^n \to \mathbf{R}^m$, $h : \mathbf{R}^n \to \mathbf{R}^r$ be arbitrary functions, $K_0 \subset \mathbf{R}^n$ be a closed set, and $Y_0 := \mathbf{R}^m \times \mathbf{R}^r_+$.*

(i) If $(x^*, (\lambda^*, \mu^*))$ is a saddle point of L in $K_0 \times Y_0$, then x^* is a solution of Problem $\mathcal{P}$, (λ^*, μ^*) is a solution of Problem $\mathcal{D}$, and $P(x^*) = D(\lambda^*, \mu^*)$.

(ii) If for $(x^*, (\lambda^*, \mu^*)) \in K_0 \times Y_0$ it holds that $P(x^*) = D(\lambda^*, \mu^*)$, then $(x^*, (\lambda^*, \mu^*))$ is a saddle point of L in $K_0 \times Y_0$.

Proof. (i) Let $(x^*, (\lambda^*, \mu^*))$ be a saddle point of L in $K_0 \times Y_0$. From Theorem 3.14 x^* is a solution of Problem $\mathcal{P}_0$, thus from Lemma 3.17 it is also a solution of Problem $\mathcal{P}$.

According to the definition of a saddle point

$$P(x^*) = \sup_{(\lambda,\mu)\in Y_0} L(x^*, \lambda, \mu) = \inf_{x \in K_0} L(x, \lambda^*, \mu^*) = D(\lambda^*, \mu^*).$$

According to (3.35) we have $D(\lambda, \mu) \leq P(x^*)$ for every $(\lambda, \mu) \in Y_0$. Since $P(x^*) = D(\lambda^*, \mu^*)$, we conclude that (λ^*, μ^*) is a solution of Problem $\mathcal{D}$.

(ii) Now let $P(x^*) = D(\lambda^*, \mu^*)$ be fulfilled for $(x^*, (\lambda^*, \mu^*)) \in K_0 \times Y_0$. Then for every $(\lambda, \mu) \in Y_0$ and $x \in K_0$ we have

$$L(x^*, \lambda, \mu) \leq P(x^*) = D(\lambda^*, \mu^*) \leq L(x, \lambda^*, \mu^*).$$

Thus $(x^*, (\lambda^*, \mu^*))$ is a saddle point of L. Q.E.D.

In general, it may happen that problems $\mathcal{P}$ and $\mathcal{D}$ have solutions x^* and (λ^*, μ^*), respectively, for which $P(x^*) > D(\lambda^*, \mu^*)$ (compare with (3.35)). This effect is known as "duality gap". Theorems 3.15 and 3.18 together show that for linear-convex optimization problems there is no duality gap.

The possibility to include some of the constraints in the set K_0 implies that the pair of a primal and a dual problem associated with a given optimization problem is not unique, in general. It depends on which affine constraints are included in K_0 instead in the system of inequalities (3.34). This is illustrated by the following example.

Example 3.19 Consider the problem

$$\min \{x_1 + 4x_2\}$$

subject to the constraints

$$(x_1)^2 + 2(x_2)^2 \le 8, \qquad -x_2 - 1 \le 0.$$

Obviously the problem is convex. Moreover, it has a solution since the domain defined by the constraints is a subset of an ellipse centered at zero. In addition, the Slater condition holds, for example with $\bar{x} = (0,0)'$.

Here we may define $K_0 = \mathbf{R}^2$ and the Lagrange function

$$L(x_1, x_2, \mu_1, \mu_2) = x_1 + 4x_2 + \mu_1 \left((x_1)^2 + 2(x_2)^2 - 8\right) + \mu_2(-x_2 - 1).$$

Then Theorem 3.15 guarantees the existence of a saddle point of the Lagrange function, and it can be found by solving the dual problem as claimed by Theorem 3.18.

Alternatively, we may define the polyhedral set $K_0 := \{(x_1, x_2) : \;\; x_2 \ge -1\}$ and consider the Lagrange function

$$L(x_1, x_2, \mu) = x_1 + 4x_2 + \mu \left((x_1)^2 + 2(x_2)^2 - 8\right)$$

on the set $K_0 \times \mathbf{R}_+$. Here we shall demonstrate the second possibility, while the first will be given as an exercise.

The dual objective function is

$$D(\mu) = \inf_{x_1 \in \mathbf{R},\, x_2 \ge -1} \{x_1 + 4x_2 + \mu((x_1)^2 + 2(x_2)^2 - 8)\}, \qquad \mu \ge 0. \tag{3.36}$$

We know how to solve this problem from school. Taking the derivative with respect to each of the variables we obtain

$$x_1 = -\frac{1}{2\mu}, \qquad x_2 = \begin{cases} -1 & \text{if } \mu \in (0,1], \\ -\frac{1}{\mu} & \text{if } \mu > 1. \end{cases} \tag{3.37}$$

In the case $\mu = 0$ the infimum of F is $-\infty$. Then we substitute these expressions in (3.36) and obtain that

$$D(\mu) \;=\; \begin{cases} -\infty & \text{for } \mu = 0, \\ -4 - \frac{1}{4\mu} - 6\mu & \text{for } \mu \in (0,1], \\ -\frac{9}{4\mu} - 8\mu & \text{for } \mu > 1. \end{cases}$$

The dual problem is

$$\max_{\mu \ge 0} D(\mu).$$

According to Remark 3.16 it is enough to find a point where D is differentiable and the derivative is zero. Differentiating for $\mu \in (0,1)$, we obtain such a point $\mu^* = \frac{1}{2\sqrt{6}}$. Then we calculate $D(\mu^*) = -4 - \sqrt{6}$. Thus, the dual problem is solved.

Let us evaluate x_1 and x_2 from (3.37). We obtain $x_1^* = -\sqrt{6}$, $x_2^* = -1$. The point $x^* := (x_1^*, x_2^*)'$ belongs to K. Then $P(x^*) = f(x^*) = -\sqrt{6} - 4$, and $P(x^*) = D(\mu^*)$. Thus (x^*, μ^*) is a saddle point of L in $K_0 \times \mathbf{R}_+$ and x^* is a solution of our original problem.

Notice that the main results in the last section and in the present section apply to arbitrary functions f, g, h and any set K_0. The role of the convexity is only to provide a class of problems for which the general statements are meaningful: there is no gap between the necessary and the sufficient conditions and there is no duality gap.

3.5 Additional examples and exercises

Exercise 3.20 Show that the function $f(x_1, x_2) = x_1 x_2$ is neither convex nor concave in $\mathbf{R}_+^2$.

Hint: Consider the following two pairs of points: $(0,0)$, $(1,1)$ and $(0,1)$, $(1,0)$.

Exercise 3.21 Prove that the function $f : \mathbf{R}^n \to \mathbf{R}$ is convex if and only if the set

$$\mathrm{Epi}(f) := \{(x,y) \in \mathbf{R}^n \times \mathbf{R} : \ x \in \mathbf{R}^n \text{ and } y \geq f(x)\}$$

is convex. (The notation Epi(f) is an abbreviation for the epigraph of f – the area in $\mathbf{R}^n \times \mathbf{R}$ “above” the graph of f.)

Hint: Just apply the definitions of convex set and of convex function.

Example 3.22 Consider the problem

$$\min\{-x_1\}$$

subject to

$$(x_1)^2 + (x_2)^2 \leq 4,$$

$$(x_1 - 1)^2 + (x_2)^2 \leq 1.$$

Here the solution is easy to find geometrically: it is the point $x^* := (2,0)'$. Notice that for the gradients of the two constraints at x^* we have

$$\partial h_1(2,0) = (4,0), \qquad \partial h_2(2,0) = (2,0).$$

The two gradients are linearly dependent, hence the condition (2.26) in Remark 2.29 is not fulfilled. However, the Mangasarian-Fromowitz constraint qualification is fulfilled with $\bar{l} = (-1, 0)'$. The Slater condition is also fulfilled with $\bar{x} := (1, 0)'$. Thus x^* must satisfy the normal for of the KKT conditions.

Exercise 3.23 Solve the problem in Example 3.22 using Theorem 3.8.

Exercise 3.24 Find the global maximizers in the problem

$$\max\{3x_1 + x_2\}$$

subject to

$$(x_1 - 1)^2 + 4(x_2 + 1)^2 \le 4$$

$$e^{x_1} + x_2 \le 2.$$

Apply the KKT theorem for convex problems. (Is the problem convex?)

Example 3.25 Consider the problem

$$\min\{|x_1 - x_2| - x_3\}$$

subject to

$$x_1 + x_2 \le 1$$

$$x_1^2 + x_3^2 \le 1.$$

A solution of this problem exists due to Theorem 1.2 ((x_1, x_3) is bounded in the unit ball, hence, if $|x_2|$ takes too large values the objective value will be worse than that at $(0, 0, 0)$).

Here the objective function is not differentiable, but is convex, as well as the inequalities. The Slater condition is fulfilled with $\bar{x} = (0, 0, 0)$. So we can apply Theorem 3.8. The KKT conditions read as

$$L = |x_1 - x_2| - x_3 + \mu_1(x_1 + x_2 - 1) + \mu_2(x_1^2 + x_3^2 - 1) \longrightarrow \min_{(x_1,x_2,x_3)\in\mathbf{R}^3}, \tag{3.38}$$

$$\mu_1(x_1 + x_2 - 1) = 0, \tag{3.39}$$

$$\mu_2(x_1^2 + x_3^2 - 1) = 0, \tag{3.40}$$

$$x_1 + x_2 \le 1 \tag{3.41}$$

$$x_1^2 + x_3^2 \le 1, \tag{3.42}$$

$$\mu_1 \ge 0, \quad \mu_2 \ge 0. \tag{3.43}$$

Case 1. Let the minimum be achieved at a point x at which the objective function is differentiable (that is, $x_1 \neq x_2$), hence L is differentiable, too. According to Theorem 2.4, the minimizers of (3.38) satisfy $\partial_x L = 0$. Thus we have

$$\partial_{x_1} L = \sigma + \mu_1 + 2\mu_2 x_1 = 0, \tag{3.44}$$

$$\partial_{x_2} L = -\sigma + \mu_1 = 0,$$

$$\partial_{x_3} L = -1 + 2\mu_2 x_3 = 0,$$

$$\tag{3.45}$$

where $\sigma = 1$ if $x_1 > x_2$ and $\sigma = -1$ if $x_1 < x_2$. Due to (3.43) we conclude from the second equation that $\sigma = 1$, hence $x_1 > x_2$, $\mu_1 = 1$. Then (3.44) becomes $2\mu_2 x_1 = -2$, which implies $x_1 < 0$. Since $x_1 > x_2$, we have also $x_2 < 0$. Then (3.39) cannot be fulfilled. There are no KKT points in Case 1.

Case 2. Let the minimum be achieved at a point x at which $x_1 = x_2$ (a point of non-differentiability of the objective function). A solution (x, μ) of (3.38)–(3.43) must exist in this case since our optimization problem has a solution! Since $x_1 = x_2$, the following relations should be satisfied:

$$\tilde{L} \;=\; -x_3 + \mu_1(2x_1 - 1) + \mu_2(x_1^2 + x_3^2 - 1) \longrightarrow \min_{(x_1,x_3)\in\mathbf{R}^2},$$

$$\mu_1(2x_1 - 1) = 0,$$

$$\mu_2(x_1^2 + x_3^2 - 1) = 0,$$

$$2x_1 \leq 1$$

$$x_1^2 + x_3^2 \leq 1,$$

$$\mu_1 \geq 0, \quad \mu_2 \geq 0.$$

The first relation can now be replaced with $\partial_{x_1}\tilde{L} = 0$, $\partial_{x_2}\tilde{L} = 0$ and the solution continues as in the examples for illustration of the KKT Theorem 2.25 by considering four sub-cases for (3.39), (3.40). The solution is left as an exercise.

Exercise 3.26 Complete the solution in Example 3.25.

Exercise 3.27 Consider the problem

$$\min\{-x\}$$

subject to the constraints

$$x \in \mathbf{R}^1, \quad x^2 \leq 0.$$

Prove that the Lagrange function $L(x, \mu)$ does not have a saddle point in $K_0 \times Y_0 := \mathbf{R}^1 \times \mathbf{R}_+$, although the problem has an optimal solution. What is the reason? (Notice that the problem is convex!)

Example 3.28 Consider the problem

$$\min_{x \in K} \{f(x) := x_1 - 2x_2 + 3x_3 - 4x_4\}$$

with

$$K := \left\{x \in \mathbf{R}^4 : \ (x_1)^2 + (x_2)^2 + (x_3)^2 + (x_4)^2 \leq 2\right\}$$

as a primal problem and find the function $D(\mu)$ of the corresponding dual problem

$$\max_{\mu \geq 0} D(\mu).$$

Then find the solution μ^* of the dual problem and evaluate the corresponding x^* (resulting from the definition of $D(\mu^*)$). Is x^* a solution of the primal problem?

Solution. Here

$$L(x, \mu) = x_1 - 2x_2 + 3x_3 - 4x_4 + \mu \left((x_1)^2 + (x_2)^2 + (x_3)^2 + (x_4)^2 - 2\right). \tag{3.46}$$

Since by definition $D(\mu) = \min_{x \in \mathbf{R}^4} L(x, \mu)$ and L is convex with respect to x, the minimizing x is determined by the condition $\partial_x L(x, \mu) = 0$:

$$x_1 = -\frac{1}{2\mu}, \quad x_2 = \frac{1}{\mu}, \quad x_3 = -\frac{3}{2\mu}, \quad x_4 = \frac{2}{\mu}. \tag{3.47}$$

Then substituting in (3.46) we calculate

$$D(\mu) = -\frac{15}{2\mu} - 2\mu.$$

Since $D(\mu)$ is concave, its maximum on the set $\mu \geq 0$ is attained either for $\mu = 0$ (which gives the "bad" value $D = -\infty$, which cannot be maximal), or at a point where $\partial D(\mu) = 0$:

$$\frac{15}{2\mu^2} - 2 = 0,$$

hence, taking into account that $\mu \geq 0$,

$$\mu^* = \frac{\sqrt{15}}{2}.$$

We evaluate $D(\mu^*) = -2\sqrt{15}$.

From (3.47) we calculate a "candidate" for a solution of the primal problem:

$$x_1^* = -\frac{1}{\sqrt{15}}, \quad x_2^* = \frac{2}{\sqrt{15}}, \quad x_3^* = -\frac{3}{\sqrt{15}}, \quad x_4^* = \frac{4}{\sqrt{15}}.$$

Then we evaluate

$$P(x^+) = f(x^*) = \frac{1}{\sqrt{15}}(-1-4-9-16) = -2\sqrt{15} = D(\mu^*),$$

which means that x^* is an optimal solution due to Theorem 3.18.

Exercise 3.29 Consider again the problem in Example 3.19. Take there $K_0 := \mathbf{R}^2$ and

$$L(x_1, x_2, \mu_1, \mu_2) = x_1 + 4x_2 + \mu_1\left((x_1)^2 + 2(x_2)^2 - 8\right) + \mu_2(-x_2 - 1).$$

1. Determine explicitly the dual objective function $D(\mu_1, \mu_2)$, μ_1, $\mu_2 \geq 0$.
2. Find a solution of the dual problem.
3. Find the solution of the original problem using the solution of the dual one.

The solution is essentially similar to the one in Example 3.19, but the calculations are different.

Exercise 3.30 a) Solve geometrically the problem

$$\max\{(x_1)^2 + (x_2)^2\}$$

subject to

$$2x_1 + x_2 \leq 5, \qquad x_1 \geq 0, \ \ x_2 \geq 0.$$

b) Find explicitly the dual objective function D with $K_0 = \mathbf{R}^2_+$.
c) Is there a duality gap for this problem?

Exercise 3.31 Consider the problem

$$\min\{(x_1)^2 + 2(x_2)^2 + 3(x_3)^2 + 4(x_4)^2)\}$$

subject to

$$-x_1 - x_2 + 2 \leq 0$$

$$-x_3 - x_4 + 3 \leq 0.$$

1. Find explicitly the dual function $D(\mu_1, \mu_2)$ (taking $K_0 = \mathbf{R}^4$).
2. Find a solution (μ_1^*, μ_2^*) of the dual problem.
3. Find a solution x^* of the problem $\min_{x\in\mathbf{R}^4} L(x, \mu_1^*, \mu_2^*)$. Is x^* a solution of the original problem and why?

Chapter 4

Linear Optimization Problems (Linear Programming)

Dealing with linear problems of any kind has the advantage that the instruments (theoretical and numerical) of linear algebra are applicable. Many problems in physics and economics can be formulated as (or approximated by) linear ones.

In the context of optimization, the pioneering works of the Soviet mathematician L. Kantorovich (Nobel prize winner, 1975, together with T. Koopmans) in the late 30-ties and of the consequent substantial contributions by the American mathematician G. Dantzig in the 1950-ties turned the linear optimization to a powerful instrument for economic decision making. Combined with the emergence and the progress of computers, this opened a new era of the implementation of the mathematical optimization in practice.

In this chapter we take advantage of the previously obtained results, apply them in the case of linear optimization problems, and develop them further by making use of the linear structure. In the first section we give various formulations of linear optimization problems and some basic geometric interpretations. In Section 4.3 we continue the duality analysis from the end of the last chapter to obtain the astounding symmetry between every linear optimization problem and its dual one. Then in Section 4.4 we give an economic interpretation of the dual problem and its solutions. Section 4.5 introduces the concept of basis (extreme) points that plays a crucial role for the widely used "simplex method" for numerical solution (to be presented in the next chapter). We conclude this chapter with some additional exercises.

4.1 A model of production/transportation planning

Linear optimization problems often arise in production/transportation planning. Below we give an example. Numerous applications of linear programming in diverse areas are given e.g. in [6].

A firm is planning to build new facilities for producing a single good in an area where the demand for this good is concentrated at n locations, D_1, $D_2, \ldots, D_n$. There are m possible places S_1, $S_2, \ldots, S_m$ where production facilities can be build. The production costs of one unit of good produced at S_i is p_i. Due to resource limitations, no more than q_i units can be produced per day at this place. The demanded quantity at place D_j is know to be d_j units per day. The transportation cost from S_i to D_j is c_{ij} per unit. The maximal quantity that can be transported from S_i to D_j is b_{ij} units per day. (Note that $b_{ij} = 0$ if there is no road from S_i to D_j.)

The problem of the firm is to decide what production capacities, y_1, $y_2, \ldots, y_m$, to install at each of the possible locations S_1, $S_2, \ldots, S_m$. (Note that $y_i = 0$ means that no facility will be build at place S_i.) Let x_{ij} denotes the quantity that will be transported every day from S_i to D_j. Then the variables (unknowns) y_1, $y_2, \ldots, y_m$, $\{x_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n\}$ have to satisfy the following constraints:

$$\sum_{j=1}^{n} x_{ij} = y_i, \quad i = 1, \ldots, m,$$

$$y_i \le q_i, \quad i = 1, \ldots, m,$$

$$0 \le x_{ij} \le b_{ij}, \quad i = 1, \ldots, m, \ \ j = 1, \ldots, n,$$

$$\sum_{i=1}^{m} x_{ij} = d_j, \quad j = 1, \ldots, n.$$

The firm wants to minimize the total expenditures per day:

$$\min \left\{ \sum_{i=1}^{m} p_i \, y_i + \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} \, x_{ij} \right\}.$$

4.2 Formulations and geometric interpretation

The general problem of linear optimization (programming) ["Lineare Optimierung (Programmierung)"] reads as

Problem GLP:

$$\min_{x \in K} \langle c, x \rangle,$$

where $K \subset \mathbf{R}^n$ is defined by the linear constraints

$$A_i x = b_i, \;\; i = 1, \ldots, m,$$

$$A_i x \geq b_i, \;\; i = m+1, \ldots, m+r,$$

$$x_k \geq 0, \;\; k = 1, \ldots, l.$$

Here $c \in \mathbf{R}^n$, A_i are n-dimensional row-vectors, and b_i are real numbers. The abbreviation *GLP* stays for "General Linear Primal" problem, where the meaning of "primal" will become clear in the next section. In the present section the abbreviation can be interpreted as "General Linear Problem". The unknowns x_k will be sometimes referred to as *decision variables*. They determine a "program" (in an economic context, e.g. a production plan), where the term "linear programming" originates from (having originally nothing to do with "computer programming").

A few remarks about the above formulation follow.

The inequality constraints are split into two parts: general linear inequality constraints $A_i x \geq b_i$ and non-negativity constraints $x_k \geq 0$.[1] This redundancy is motivated by the simplicity of the non-negativity constraints, which allows to treat them in an easier way in the theoretical and numerical considerations below.

The numbers m and/or r could be zero, so that equality/inequality constraints can be missing, at all. Also l can be equal to zero if non-negativity constraints are not present. Notice that the constraints are assumed to be sorted in such a way that the first m constraints are of equality type, the rest r constraints are of inequality type. Similarly, the variables x_k are enumerated so that the first l ones are restricted to be non-negative.

Each of the equality constraints $A_i x = b_i$ can be replaced by two inequality constraints: $A_i x \geq b_i$ and $-A_i x \geq -b_i$, so that considering a problem with inequality constraints only ($m = 0$) is not a restriction of the generality (although this "shortcut" is not recommended when the problem is to be solved numerically).

On the other hand, one can reformulate Problem GLP in such a way that only equality constraints are present ($r = 0$) and there are non-negativity constraints for all decision variables x_k. Let us make this clear. An inequality constraint $A_i x \geq b_i$ can be replaced with the equality constraint $A_i x - s_i = b_i$, where s_i is a new decision variable (called in German "Schlupfvariable") which has to satisfy the constraint $s_i \geq 0$. Moreover, if there is no non-negativity constraint for some x_k (that is, $k > l$), then one can represent $x_k = x_k^+ - x_k^-$ with x_k^+, $x_k^- \geq 0$ and substitute it in all other constraints. In this way we may transform Problem GLP to a problem of the following form (called further *Standard Linear (optimization) Problem*):

[1] Notice that the inequalities are reversed, compared with the formulations in the preceding chapters. This is due to a historically established "standard" and is not essential, of course.

Problem SLP:

$$\min_{x \in \mathbf{R}^n} \langle c, x \rangle \tag{4.1}$$

subject to

$$Ax = b, \tag{4.2}$$

$$x \geq 0. \tag{4.3}$$

Of course, here the space dimension n, the matrix A and the vector b are not the same as in Problem GLP. In the next chapter we shall present a numerical algorithm for solving SLP.

Example 4.1 Consider the linear optimization problem

$$\max\{x_1 + 2x_2\}$$

subject to the constraints

$$\begin{aligned} x_1 + x_2 &\leq 1, \\ x_1 - x_2 &\geq -1, \\ x_2 &\geq 0. \end{aligned}$$

We introduce the additional variables s_1, s_2, substitute $x_1 = x_1^+ - x_1^-$, and obtain the equivalent problem

$$\min\{-x_1^+ + x_1^- - 2x_2\}$$

$$\begin{aligned} -x_1^+ + x_1^- - x_2 - s_1 &= -1, \\ x_1^+ - x_1^- - x_2 - s_2 &= -1, \\ x_1^+,\, x_1^-,\, x_2,\, s_1,\, s_2 &\geq 0. \end{aligned}$$

This problem has the form of (4.1)–(4.3) with $x = (x_1^+, x_1^-, x_2, s_1, s_2)'$,

$$A = \begin{pmatrix} -1 & 1 & -1 & -1 & 0 \\ 1 & -1 & -1 & 0 & -1 \end{pmatrix}, \qquad b = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \qquad c = (-1, 1, -2, 0, 0)'.$$

Now we make some geometric interpretations.

Any linear equality

$$\langle g, x \rangle = \beta$$

with $g \in \mathbf{R}^n$ defines an $(n-1)$-dimensional (affine) hyperplane ("Hyperebene") Γ in $\mathbf{R}^n$, which has a visual geometric meaning if $n = 2$ ("Gerade") or $n = 3$ ("Ebene"). The vector g is orthogonal to Γ.

The inequality $\langle g, x\rangle \geq \beta$ defines an (affine) half-space with a boundary Γ, namely the half-space defined by Γ to which the vector g points.

Then a system of linear equalities and inequalities defines a set K (in Definition 3.6 we called such sets polyhedral) which is an intersection of a finite number of half-spaces and hyperplanes.

Figure 4.1 shows the set defined by the original inequalities (before the reformulation to an SLP) in Example 4.1. (After the reformulation the resulting set K lies in a 5-dimensional space and is hard to be represented geometrically.) The point from K where the maximum is achieved is the one with the greatest projection (where also the sign is regarded) on the vector c. This is because $\langle c, x\rangle/|c|$ is exactly the length of this projection with plus or minus sign depending on the sign of $\langle c, x\rangle$.

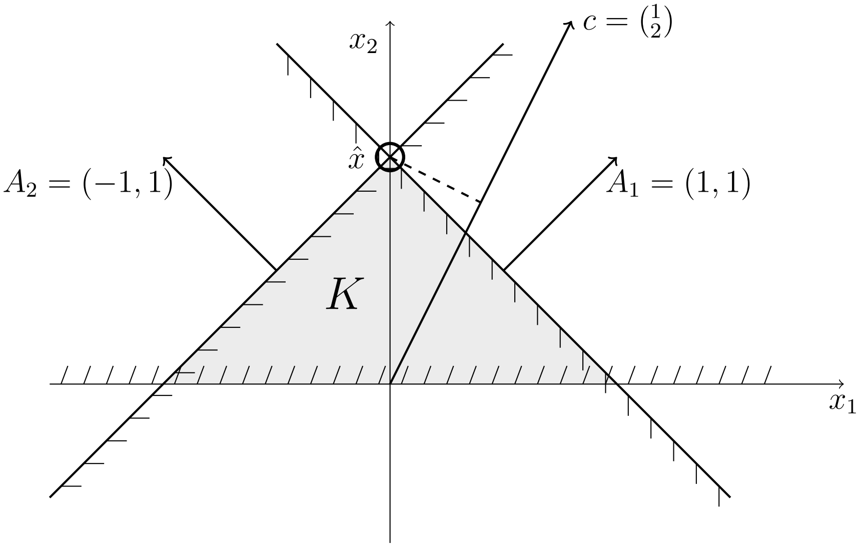


Figure 4.1: The polygon K from Example 4.1 and the solution x^*.

4.3 Duality theory for linear problems

In this section we apply the duality theory from Section 3.4 in the case of linear optimization problems and establish a striking symmetry between the primal and the dual problem. It shows that solving Problem GLP is equivalent to solving its dual problem, which brings the advantage to have the choice which one to solve or to analyze. Some numerical methods involve simultaneous solving of both problems.

Consider again the general linear (primal) problem

Problem GLP:

$$\min \langle c, x\rangle,$$

subject to

$$\begin{aligned} A_i x &= b_i, \quad i = 1, \dots, m, \\ A_i x &\geq b_i, \quad i = m+1, \dots, m+r, \\ x_k &\geq 0, \quad k = 1, \dots, l, \end{aligned}$$

where $c \in \mathbf{R}^n$, A_i are n-dimensional row-vectors and b_i are real numbers.

As in the last three sections of Chapter 3 we separate the non-negativity constraints from the rest, including them in the set

$$K_0 := \{x \in \mathbf{R}^n : \; x_k \geq 0 \text{ for } k = 1, \dots, l\}.$$

Then Problem GLP can be rewritten as

$$\min_{x \in K \cap K_0} \{c' x\}$$

where

$$K := \{x \in \mathbf{R}^n : \; b_i - A_i x = 0, \; i = 1, \dots, m, \; b_j - A_j x \leq 0, \; j = m+1, \dots, m+r\}.$$

The Lagrange function reads as

$$L(x, \lambda, \mu) = c'x + \sum_{i=1}^{m} \lambda_i (b_i - A_i x) + \sum_{j=m+1}^{m+r} \mu_{j-m}(b_j - A_j x), \qquad \lambda \in \mathbf{R}^m, \; \mu \in \mathbf{R}^r_+. \quad (4.4)$$

According to Lemma 3.17 the GLP is equivalent to the problem

$$\min_{x \in K_0} P(x), \qquad \text{where } P(x) := \sup_{(\lambda,\mu) \in \mathbf{R}^m \times \mathbf{R}^r_+} L(x, \lambda, \mu),$$

and if x^* is an optimal solution of these problems, then $c'x^* = P(x^*)$. Even more,

$$P(x) = \begin{cases} c'x & \text{if } x \in K, \\ +\infty & \text{if } x \notin K. \end{cases} \quad (4.5)$$

Let us denote $y := (\lambda, \mu)$. In this notation

$$L(x, \lambda, \mu) = L(x, y) = c'x + \langle y, b - Ax\rangle.$$

Then the dual problem of GLP has the objective function

$$\begin{aligned} D(y) &= \inf_{x\in K_0} L(x,y) = \inf_{x\in K_0} \{\langle c,x\rangle + \langle y, b-Ax\rangle\} \\ &= \inf_{x\in K_0} \{\langle b,y\rangle + \langle c-A'y, x\rangle\} \\ &= \inf_{x\in K_0} \left\{\langle b,y\rangle + \sum_{k=1}^{l} x_k[c-A'y]_k + \sum_{k=l+1}^{n} x_k[c-A'y]_k\right\} \\ &= \left\{ \begin{array}{ll} b'y & \text{if } [A'y-c]_k \le 0 \ \forall\, k\in\{1,\dots,l\} \text{ and} \\ & \quad [A'y-c]_k = 0 \ \forall\, k\in\{l+1,\dots,n\}, \\ -\infty & \text{else.} \end{array}\right. \end{aligned}$$

This is because the linear functions $x_k \mapsto -[A'y-c]_k\, x_k =: -\alpha x_k$ is unbounded from below in $(-\infty,\infty)$ if $\alpha \neq 0$, also it is unbounded from below in $[0,\infty)$ if $\alpha > 0$, and achieves its minimum at $x_k = 0$ if $\alpha \le 0$. As usual, $[v]_k$ means the k-th component of the vector v if v is a vector and the k-th row of the matrix v if v is a matrix. Then the dual problem (GLD is an abbreviation of General Linear Dual) reads as

Problem GLD:

$$\max\{b'y\}$$

subject to the constraints

$$\begin{aligned} [A']_k\, y &\le c_k, \qquad k\in\{1,\dots,l\}, && (4.6) \\ [A']_k\, y &= c_k, \qquad k\in\{l+1,\dots,n\}, && (4.7) \\ y_j &\ge 0, \qquad j\in\{m+1,\dots,m+r\}. \end{aligned}$$

Obviously this problem can be formulated as

$$\max_{y\in Y\cap Y_0} \{b'y\},$$

where Y is the set of those $y \in \mathbf{R}^{m+r}$ for which the constraints (4.6) and (4.7) are satisfied, and

$$Y_0 := \left\{y\in\mathbf{R}^{m+r} :\ y_j \ge 0,\ j = m+1,\dots,m+r\right\} = \mathbf{R}^m \times \mathbf{R}^r_+.$$

From the representation of the dual objective function $D(y)$ and the introduced notations we have

$$D(y) = \left\{ \begin{array}{ll} b'y & \text{if } y\in Y, \\ -\infty & \text{if } y\notin Y. \end{array}\right. \tag{4.8}$$

Thus we established that the dual problem $\max_{y \in Y_0} D(y)$ and the problem GLD are equivalent and if y^* is a solution, then $b'y^* = D(y^*)$.

One can formulate the following rules for obtaining the dual problem to the GLP:
– it is a maximization problem;
– the number of variables equals the number of constraints in the GLP:
each variable corresponds to one constraint in GLP;
– the number of constraints equals the number of variables in GLP:
each constraint corresponds to one variable in GLP;
– the matrix of the constraints is A';
– the right-hand sides of the constraints are the coefficients of the objective function of GLP;
– the constraints corresponding to free variables in GLP are equalities,
the rest are of "$\leq$" type;
– the variables corresponding to equality constraints of GLP are free,
the rest are non-negative;
– the coefficients of the objective function are the right-hand sides of the constraints in GLP.

The above rules are summarized in the following diagram:

$$\begin{array}{lcl}
\min \langle c, x \rangle & \longrightarrow & \max \langle b, y \rangle \\
 & & \\
A_1 x = b_1 & \longrightarrow & y_1 \text{ – free} \\
\dots \dots & & \dots \dots \\
A_m x = b_m & \longrightarrow & y_m \text{ – free} \\
A_{m+1} x \geq b_{m+1} & \longrightarrow & y_{m+1} \geq 0 \\
\dots \dots & & \dots \dots \\
A_{m+r} x \geq b_{m+r} & \longrightarrow & y_{m+r} \geq 0 \\
 & & \\
x_1 \geq 0 & \longrightarrow & [A']_1 y \leq c_1 \\
\dots \dots & & \dots \dots \\
x_l \geq 0 & \longrightarrow & [A']_l y \leq c_l \\
x_{l+1} \text{ – free} & \longrightarrow & [A']_{l+1} y = c_{l+1} \\
\dots \dots & & \dots \dots \\
x_n \text{ – free} & \longrightarrow & [A']_n y = c_n
\end{array}$$

Let us write down the dual problem to GLD. To do this we represent the GLD in the GLP format:

$$\min \{-b'y\}$$

subject to the constraints

$$\begin{aligned} -[A']_k\, y &= -c_k, \qquad k \in \{l+1, \ldots, n\}, \\ -[A']_k\, y &\geq -c_k, \qquad k \in \{1, \ldots, l\}, \\ y_j &\geq 0, \qquad j \in \{m+1, \ldots, m+r\}. \end{aligned}$$

Then writing down the dual problem of the last one (using the above rules) we obtain the original problem GLP (*check it!*). Thus we have the following remarkable fact.

Lemma 4.2 *The dual problem to the dual of a given linear optimization problem coincides with the given problem.*

The lemma points out the symmetry between a linear optimization problem and its dual. In particular, the above diagram for the definition of the dual problem can be read from right to left if the optimization problem has the form as on the right column of the diagram.

Remark 4.3 Notice also that GLP and GLD have the same Lagrange function $L(x, y)$ (up to the sign), where x appears as a Lagrange multiplier for GLD, where the variable is y.

The next theorem summarizes the connections between the two problems in a duality pair, which almost readily follow from the results obtained in the previous chapter.

Theorem 4.4 *The following three statements are equivalent:*
(A1) Both GLP and GLD have feasible solutions (that is, $K \cap K_0 \neq \emptyset$, $Y \cap Y_0 \neq \emptyset$);
(A2) Either GLP or GLD has an optimal solution;
(A3) Both GLP and GLD have optimal solutions.

Also the following six statements are equivalent:
(B1) x^* and y^* are optimal solutions of GLP and GLD, respectively;
(B2) $x^* \in K_0$, $y^* \in Y_0$, and $P(x^*) = D(y^*)$ (that is, $c'x^* = b'y^*$);
(B3) (x^*, y^*) is a saddle point of the Lagrange function (4.4) in $K_0 \times Y_0$;
(B4) x^* and y^* are feasible points for GLP and GLD, respectively, and the complementary slackness conditions hold for the inequality constraints:

$$y_j^*\,(A_j x^* - b_j) = 0,\; j = m+1, \ldots, m+r, \qquad x_k^*\,([A']_k y^* - c_k) = 0,\; k = 1, \ldots, l.$$

(B5) x^* is an optimal solution of GLP and y^* is a Lagrange multiplier associated with x^* in the KKT conditions for GLP;
(B6) y^* is an optimal solution of GLD and x^* is a Lagrange multiplier associated with y^* in the KKT conditions for GLD.

Proof. **Part 1.** The implication (A3) $\Longrightarrow$ (A2) is obvious. Let us prove that (A2) implies (A3). Assume that GLP has an optimal solution x^*. According to Theorem 3.15 there exists $y^* = (\lambda^*, \mu^*) \in Y_0$ such that (x^*, y^*) is a saddle point of the Lagrange function in $K_0 \times Y_0$. Then Theorem 3.18 (i) implies that y^* is a solution of GLD, thus (A3).

To prove that (A1) implies (A3) we need the following lemma (which was used also in Chapter 2).

Lemma 4.5 *Let $M \subset \mathbf{R}^n$ be a polyhedral set (see Definition 3.6) and let P be an $(m \times n)$-matrix. Then the set $PM := \{Px : \; x \in M\}$ is convex and closed.*

The convexity of PM is obvious since M is convex and P is linear. The proof of the closedness requires some work (in the case of unbounded M) which we skip in this version of the script.

Let us continue with the proof of the implication (A1) $\Longrightarrow$ (A3). Take some $\tilde{y} \in Y \cap Y_0$. Due to (4.5) and (4.8) for every $x \in K \cap K_0$ we have

$$c'x = P(x) = \sup_{y \in Y_0} L(x, y) \geq L(x, \tilde{y}) \geq D(\tilde{y}) = b'\tilde{y}.$$

Then the linear function $c'x$ is bounded from below in $K \cap K_0$ by the number $\langle b, \tilde{y} \rangle$. On the other hand, according to Lemma 4.5 the set $\{c'x : \; x \in K \cap K_0\}$ is convex and closed. Due to the boundedness from below this set is an interval of the form $[p, q]$ or $[p, +\infty)$. If $x^* \in K \cap K_0$ is such that $c'x^* = p$, then $c'x^* \leq c'x$ for every $x \in K \cap K_0$, thus x^* is a solution of GLP.

The fact that GLD also has a solution is analogous.

Part 2. Now we consider claims (B1)–(B6). Claims (B2) and (B3) are equivalent according to Theorem 3.18. Also (B3) implies (B1) due to the same theorem. Claims (B3) and (B5) are equivalent due to Lemma 3.13. Also (B3) and (B6) are equivalent due to the same lemma and Remark 4.3. Below we shall prove in separate steps the implications (B1) $\Longrightarrow$ (B2); (B3) $\Longrightarrow$ (B4); (B4) $\Longrightarrow$ (B3).

2.1. Let (B1) be fulfilled. Since x^* is a solution of GLP, Theorem 3.15 claims that there exists $\bar{y} \in Y_0$ such that $(x^*, \bar{y})$ is a saddle point of L in $K_0 \times Y_0$. Then $P(x^*) = D(\bar{y})$ according to claim (i) of Theorem 3.18. Since y^* is a solution of GLD, we have $D(\bar{y}) \leq D(y^*)$. This gives $P(x^*) \leq D(y^*)$, which implies $P(x^*) = D(y^*)$ (here we use that $P(x) \geq D(y)$ for every $x \in K_0$ and $y \in Y_0$). This proves (B2).

2.2. Let (B3) hold. According to Lemma 3.13 (x^*, y^*) is a KKT point for GLP, hence, the feasibility of x^* and the first group of equalities in (B4) (these are the complementary slackness conditions (3.30)). But (x^*, y^*) is also a KKT point for GLD (see Remark 4.3), hence, the feasibility of y^* and the second group of equalities in (B4).

2.3. It remains to prove that (B4) implies (B3). Using the first set of equalities in (B4) and $x^* \in K$ we have for any $y \in Y_0$

$$\begin{aligned} L(x^*, y^*) - L(x^*, y) &= (y^* - y)'(b - Ax^*) = \sum_{j=m+1}^{m+r} (y_j^* - y_j)(b_j - A_j x^*) \\ &= -\sum_{j=m+1}^{m+r} \underbrace{y_j}_{\geq 0} \underbrace{(b_j - A_j x^*)}_{\leq 0} \geq 0. \end{aligned}$$

Similarly, using the second set of equalities in (B4) and $y^* \in Y$ we obtain that for any $x \in K_0$

$$L(x, y^*) - L(x^*, y^*) \geq 0.$$

Hence $L(x^*, y) \leq L(x^*, y^*) \leq L(x, y^*)$ for every $x \in K_0$ and every $y \in Y_0$. This means that (x^*, y^*) is a saddle point of L in $K_0 \times Y_0$. Q.E.D.

The above theorem provides the following possibilities: (i) to prove existence of an optimal solution by "double feasibility" (that is by proving that both the primal and the dual problem have feasible points); (ii) to prove existence of a solution by proving that the dual problem has a solution; (iii) to solve the dual problem and then to find a solution of the primal problem using claim (B4) of the theorem.

The next examples demonstrate some "hand-solvable" applications of the duality Theorem 4.4.

Example 4.6 Consider the linear optimization problem

$$\max\{-x_1 + 2x_2 + x_3\}$$

subject to

$$x_1 + x_2 + x_3 = 5$$

$$x_1 - x_3 \leq 2$$

$$3x_1 + x_2 + 2x_3 \geq 7$$

$$x_1,\ x_3 \geq 0.$$

In order to obtain the dual problem we first rewrite the above problem in the GLP format:

$$\min\{x_1 - 2x_2 - x_3\}$$

subject to

$$\begin{aligned} x_1 + x_2 + x_3 &= 5 \\ -x_1 + x_3 &\geq -2 \\ 3x_1 + x_2 + 2x_3 &\geq 7 \\ x_1,\, x_3 &\geq 0. \end{aligned}$$

According to the definition, Problem GLD reads as

$$\max\{5y_1 - 2y_2 + 7y_3\}$$

subject to

$$\begin{aligned} y_1 - y_2 + 3y_3 &\leq 1 \\ y_1 + y_3 &= -2 \\ y_1 + y_2 + 2y_3 &\leq -1 \\ y_2,\, y_3 &\geq 0. \end{aligned}$$

The second constraint is of equality type because x_2 is free in GLP. Since the first constraint in GLP is an equality, the variable y_1 in GLD is free.

The point $x^* = (0, 3, 2)' \in K \cap K_0$ gives a value $P(x^*) = -8$ of the objective function of GLP, and the point $y^* = (-3, 0, 1)' \in Y \cap Y_0$ gives $D(y^*) = -8$. Then part (B2) in Theorem 4.4 implies that x^* is a solution of the considered problem and y^* is a solution of its dual.

Example 4.7 Consider the linear optimization problem

$$\max\{-3x_1 - x_2\}$$

subject to

$$x_1 - x_2 \leq 2$$

$$2x_1 + x_2 \geq 2$$

$$x_1,\, x_2 \geq 0.$$

In order to obtain the dual problem we first rewrite the above problem in the GLP format:

$$\min\{3x_1 + x_2\}$$

subject to

$$\begin{aligned} -x_1 + x_2 &\geq -2 \\ 2x_1 + x_2 &\geq 2 \\ x_1,\, x_2 &\geq 0. \end{aligned}$$

Then the dual problem is

$$\max\{-2y_1 + 2y_2\}$$

subject to

$$\begin{aligned} -y_1 + 2y_2 &\leq 3 \\ y_1 + y_2 &\leq 1 \\ y_1,\, y_2 &\geq 0. \end{aligned}$$

Let us solve the dual problem graphically. From Figure 4.2 we see that the solution is $y^* = (0,1)'$, which gives the optimal value $D(y^*) = 2$. Due to part (A2) of Theorem 4.4 the primal problem has a solution x^*, and according to part (B4) of the same theorem it satisfies the second inequality $2x_1 + x_2 \geq 2$ as an equality, due to the complementary slackness condition and $y_2 > 0$. Moreover, the first inequality, $-y_1 + 2y_2 \leq 3$, is satisfied by y^* as a strict inequality ("<"), hence $x_1^* = 0$. Thus we obtain that $x^* = (0,2)'$ is a solution of our problem. The minimal objective value is $P(x^*) = 2 = D(y^*)$, as it should be, according to part (B2) of Theorem 4.4.

4.4 Economic interpretation of the dual problem

Let us consider again the problem

$$\max_x c'x,$$

under constraints

$$Ax \leq b, \qquad x \geq 0.$$

Similarly as in Example 1.8 one can interpret $x \in \mathbf{R}^n$ as a production plan (x_k is the quantity of product k that is to be produced), $b \in \mathbf{R}^r$ is the vector of available resources, the component a_{ij} of the matrix A is the amount of resource of type i needed for the

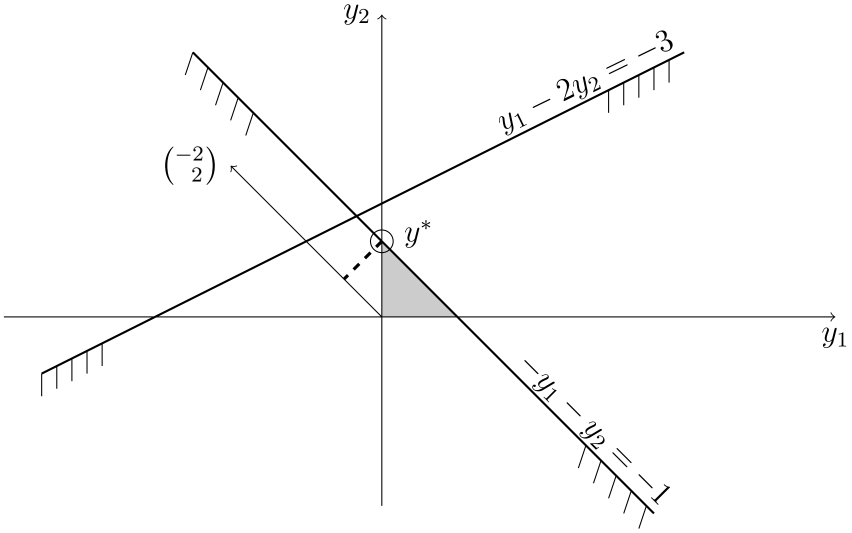


Figure 4.2: Graphical solution of the dual problem in Example 4.7

production of one unit of product j, c is the price vector. In Section 2.7 we interpreted the Lagrange multiplier μ_i resulting from the KKT theorem, corresponding to an active constraint as the shadow price of resource i.[2]

Now we shall interpret the dual problem

$$\min_x b'y,$$

under constraints

$$A'y \geq c, \qquad y \geq 0.$$

Any feasible y is a reasonable evaluation of the resource prices: the constraints

$$\sum_{i=1}^{m} a_{ik} y_i \geq c_k$$

mean that with a feasible evaluation of the resource prices the cost of production of any good is not smaller than the market price c_j of this product. That is, no production gives a positive profit, which should be the case in a competitive market: if a producer would make a profit from some product, then the price of the resources will increase as long as the profit is positive. A solution of the dual problem gives such a resource price-vector y^*, for which the total value of the resource, $y'b$, is minimal, under the condition that with this price system no positive profit can be made.

[2] In this section we use the notations b and c in a way which is inconsistent with that in Section 2.7. This should not lead to a confusion.

According to (B4) in Theorem 4.4, we have for the optimal solutions of the primal and of the dual problem that

$$x_k^* \left(\sum_{i=1}^{m} a_{ik} y_i^* - c_k \right) = 0.$$

This means that a product k will be produced ($x_k^* > 0$) only if this does not lead to losses ($\sum_{i=1}^m a_{ik} y_i^* - c_k = 0$).

In short, one can say that the solution of the dual problem provides an efficient evaluation of the value of the resources in a competitive economy.

4.5 Extreme points and basis solutions

From the geometric interpretation of the linear optimization problems we may guess that if an optimal solution exists, then the optimal value is achieved also at a point which is a vertex of the set of feasible points. In this section we make this observation precise and supply it with a proof. The underlying idea will be used in the next chapter to develop a powerful numerical solver – the celebrated *simplex method.*

Let $K \subset \mathbf{R}^n$ be an arbitrary closed convex set.

Definition 4.8 A point $x \in K$ is called *extreme point* of K if the equality $x = \alpha x^1 + (1-\alpha)x^2$ with x^1, $x^2 \in K$ and $\alpha \in (0,1)$ implies that $x^1 = x^2$.

In other words, extreme points are such points that cannot be represented as a proper convex combinations of other points of K.

In what follows we focus on the special case of a polyhedral set in the standard format

$$K = \{x \in \mathbf{R}^n : \ Ax = b, \ x \geq 0\},$$

where A is an $(m \times n)$-matrix and $b \in \mathbf{R}^m$. Further on, we denote the columns of A by A^i, $i = 1, \ldots, n$, so that $A = [A^1 \ \ldots \ A^n]$.

Proposition 4.9 *Let* $r := \operatorname{rank}(A) \geq 1$*. The point* $x = (x_1, \ldots, x_n)' \in \mathbf{R}^n$ *is an extreme point of* K *if and only if there exist indexes* $i_1, \ldots, i_r$ *such that the columns* $A^{i_1}, \ldots, A^{i_r}$ *are linearly independent and*

$$x \geq 0, \quad x_i = 0 \text{ for } i \notin \{i_1, \ldots, i_r\}, \quad \sum_{s=1}^{r} x_{i_s} A^{i_s} = b. \tag{4.9}$$

Proof. **1.** Let x be an extreme point of K. The case $x = 0$ is trivial. Indeed, from $Ax = b$ we obtain $b = 0$. Then one can take $A^{i_1}, \ldots, A^{i_r}$ to be any system of r linearly independent columns of A. The equalities (4.9) are fulfilled by $x = 0$.

1.1. Let $x \neq 0$ and let $x_{i_1}, \ldots, x_{i_k}$ be all the positive components of x. Obviously we have

$$\sum_{s=1}^{k} x_{i_s} A^{i_s} = b, \qquad x_i = 0 \text{ for } i \neq i_s,\ s = 1, \ldots, k. \tag{4.10}$$

Let us prove that the columns $A^{i_1}, \ldots, A^{i_k}$ are linearly independent. Assume that for some real numbers $\alpha_1, \ldots, \alpha_k$

$$\sum_{s=1}^{k} \alpha_s A^{i_s} = 0. \tag{4.11}$$

For a real number ε we denote by x^ε the vector with components $x^\varepsilon_{i_s} = x_{i_s} + \varepsilon\alpha_s$ and the remaining components equal to zero. Since $x_{i_s} > 0$, we have that also $x^\varepsilon_{i_s} \geq 0$, $s = 1, \ldots, k$, provided that $|\varepsilon|$ is sufficiently small, namely, for ε satisfying $|\varepsilon||\alpha_s| \leq x_{i_s}$ for all $s = 1, \ldots, k$. Let us fix such an $\varepsilon \neq 0$. From (4.10) and (4.11) we have that $Ax^{\pm\varepsilon} = b$, thus $x^{\pm\varepsilon} \in K$. Obviously $x = 0.5x^{-\varepsilon} + 0.5x^{\varepsilon}$. Since x is an extreme point of K this equality implies that $x^{-\varepsilon} = x^{\varepsilon}$, which means that $\alpha_1 = \ldots = \alpha_k = 0$. Thus the vectors $A^{i_1}, \ldots, A^{i_k}$ are linearly independent.

1.2. If $k = r$, then (4.10) coincides with (4.9). If $k < r$ ($k > r$ is not possible due to the linear independence of $A^{i_1}, \ldots, A^{i_k}$), then one can complement the linearly independent vectors $A^{i_1}, \ldots, A^{i_k}$ with additional $r - k$ columns of A, preserving the linear independence. Then (4.9) will be satisfied, since the components of x corresponding to the additional columns of A are all zero.

2. Now let relations (4.9) hold for a set of indexes $i_1, \ldots, i_r$ for which the columns $A^{i_1}, \ldots, A^{i_r}$ are linearly independent. Assume that $x = \alpha x^1 + (1 - \alpha)x^2$ with some $\alpha \in (0, 1)$ and some x^1, $x^2 \in K$. For $i \notin \{i_1, \ldots, i_r\}$ we have $x_i = 0$, hence $0 = \alpha x^1_i + (1 - \alpha)x^2_i$. Since x^1_i, $x^2_i \geq 0$ this implies $x^1_i = x^2_i = 0$. Thus

$$\sum_{s=1}^{r} x^1_{i_s} A^{i_s} = b, \qquad \sum_{s=1}^{r} x^2_{i_s} A^{i_s} = b.$$

Subtracting the second equality from the first we obtain that

$$\sum_{s=1}^{r} (x^1_{i_s} - x^2_{i_s}) A^{i_s} = 0.$$

Since $A^{i_1}, \ldots, A^{i_r}$ are linearly independent, this implies $x^1 = x^2$, thus x is an extreme point of K. Q.E.D.

Definition 4.10 A point $x \in \mathbf{R}^n_+$ is called *basis point* (also "basis solution") of K if there exist indexes $i_1, \ldots, i_r$ (where $r := \operatorname{rank}(A) \geq 1$) such that the columns $A^{i_1}, \ldots, A^{i_r}$ are linearly independent and equalities (4.9) are fulfilled.

Clearly, every base point is feasible, due to (4.9). The above proposition says that the notions of basis point and extreme point coincide, except in the degenerate case $A = 0$.

Obviously every basis point is also feasible. The components x_{i_s} of a basis point x are called *basis components.* If all basis components are positive, the basis point is called *non-degenerate.* If a basis point has some basis components equal to zero, it is called *degenerate.*

The next theorem shows that in searching for an optimal solution of a linear optimization problem one can restrict the search to basis points only.

Theorem 4.11 *Let A be a non-zero matrix. If the linear problem*

$$\min_{x \in K} \{c'x\}, \qquad K = \{x \in \mathbf{R}^n : \ Ax = b, \ x \geq 0\},$$

has a feasible point (that is, $K \neq \emptyset$), then it has a basis point.

If the problem has a solution, then it has a solution which is a basis point.

Proof. We shall prove both statements simultaneously. Let x be a feasible point. If $x = 0$, then x is a basis point by the argument in the first paragraph of the proof of Proposition 4.9. Let $x \neq 0$ and let $x_{i_1}, \ldots, x_{i_k}$ be all the strictly positive components of x. Then (4.10) is satisfied. If the vectors $A^{i_1}, \ldots, A^{i_k}$ are linearly independent, then one can prove that x is a basis point exactly as in part 1.2 of the proof of Proposition 4.9. In particular, if x is an optimal solution, then it is an optimal basis solution.

Now consider the case of linearly dependent vectors $A^{i_1}, \ldots, A^{i_k}$. There exist numbers $\alpha_1, \ldots, \alpha_k$, not all of them equal to zero, such that

$$\sum_{s=1}^{k} \alpha_s A^{i_s} = 0.$$

Without any restriction we may assume that at least one of the numbers α_i is positive (otherwise we can set $\alpha := -\alpha$). Consider the point x^ε with components $x^\varepsilon_{i_s} = x_{i_s} - \varepsilon \alpha_s$ for $s = 1, \ldots, k$, and $x^\varepsilon_i = x_i = 0$ for all other components. As in part 1.1 of the proof of Proposition 4.9 we have that $Ax^\varepsilon = b$. If we define

$$\varepsilon = \min_{s:\, \alpha_s > 0} \left\{ \frac{x_{i_s}}{\alpha_s} \right\}. \tag{4.12}$$

then also $x^\varepsilon \geq 0$ is fulfilled, thus $x^\varepsilon \in K$. Moreover, $x^\varepsilon_{i_s} = 0$ for those s for which the minimum in (4.12) is attained. Thus x^ε is feasible and has less than k non-zero components.

Let x be an optimal solution. Since $x^\varepsilon \in K$ for all sufficiently small $|\varepsilon|$, we have

$$0 \leq c'x^\varepsilon - c'x = \varepsilon \sum_{s=1}^{k} c_{i_s} \alpha_s.$$

Since ε can be a positive or a negative number (with $|\varepsilon|$ sufficiently small) the above inequality implies that $\sum_{s=1}^{k} c_{i_s} \alpha_s = 0$, hence $c'x^\varepsilon = c'x$ and x^ε is also optimal. Then defining ε as in (4.12) we obtain as above an optimal solution x^ε with less than k non-zero components.

In both cases we may repeat the same procedure as long as the currently obtained point $x := x^\varepsilon$ is non-zero and the columns of A corresponding to the non-zero components of x are linearly dependent. Since at every step the number of non-zero components strictly decreases, after a finite number of steps the procedure must terminate, which means, according to the first part of the proof, that we have arrived at a basis point. Q.E.D.

The above theorem suggests the following approach: (i) calculate $r := \text{rank}(A)$; (ii) consider all possible combinations of r columns of A, and for those which are of rank r solve the system of linear equations in (4.9); (iii) among all "solutions" obtained in this way and having only non-negative components choose one for which the objective function has a minimal value.

Then according to the above theorem one will obtain an optimal solution, if such exists. If an optimal solution does not exist, the set of non-negative solutions in point (iii) would be empty.

The above numerical approach is not efficient, since it requires solving C_n^r systems of linear equations, where C_n^r is the number of combinations of r elements out of n elements. This could be a too heavy task even for the existing high-performance computers since C_n^r may increase exponentially with n (say, with $r \approx n/2$). In the next chapter, we present a much more efficient realization of the idea provided by the above theorem. Here we only mention that nowadays computers solve linear optimization problems with hundreds of millions of unknowns x_i.

Example 4.12 Consider the problem

$$\max\{x_1 - x_2 + x_3 - x_4\},$$

subject to

$$x_1 + x_2 + x_3 + x_4 = 10,$$

$$x_1 - 2x_2 + 3x_3 - x_4 = 0,$$

$$x_1,\ x_2,\ x_3,\ x_4 \geq 0.$$

Here

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -2 & 3 & -1 \end{pmatrix}$$

and rank$(A) = 2$. All possible combinations of basis components are

$$(1,2),\ (1,3),\ (1,4),\ (2,3),\ (2,4),\ (3,4).$$

Solving the system

$$\begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 10 \\ 0 \end{pmatrix},$$

we obtain one basis point $x^1 := (20/3, 10/3, 0, 0)'$. Then we solve the system

$$\begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_3 \end{pmatrix} = \begin{pmatrix} 10 \\ 0 \end{pmatrix}$$

The solution is $(15, -5)$, which cannot be a part of a basis solution since it contains a negative component. The system

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_4 \end{pmatrix} = \begin{pmatrix} 10 \\ 0 \end{pmatrix}$$

gives a basis point $x^2 := (5, 0, 0, 5)'$. In the same way we obtain two more basis points $x^3 := (0, 6, 4, 0)'$ and $x^4 := (0, 0, 2.5, 7.5)'$, while the solution corresponding to the pair of columns $(2, 4)$ has a negative component.

The objective values corresponding to the four basis solutions are 10/3, 0, -2, -5, therefore there is a single optimal basis solutions: $x^* = x^1 = (20/3, 10/3, 0, 0)'$.

4.6 Additional examples and exercises

Example 4.13 A firm has quantities $R_1, \ldots, R_n$ of n different resources. This firm is able to produce m different types of products, where production of one unit of the j-th product ($j = 1, \ldots, m$) requires quantity $q_{1,j}$ of resource 1, $q_{2,j}$ of resource 2, ..., quantity $q_{n,j}$ of resource n. Any unit of product j can be sold at price p_j ($j = 1, \ldots, m$). Moreover, the firm may sell or buy resource i from the market at price c_i, but due to transport limitations the firm cannot sell more than quantity a_i and buy more than a quantity b_i of resource i ($i = 1, \dots, n$). The firm has to decide how much of any resource to buy or sell and how to allocate the resources to products, so that its total revenue (regarding bought or sold resources and sold production) is maximal.

1. Formulate a linear optimization problem that represents this firm's problem.
2. Reformulate the obtained linear optimization problem as a problem in a standard format (SLP), see Section 4.2.

Example 4.14 Write down the dual problem to the following one:

$$\min\{2x_1 + x_2 + 4x_3\},$$

$$x_1 - x_2 + x_3 \geq 2,$$

$$-2x_1 + x_2 + 2x_3 \geq 1,$$

$$x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0.$$

Solve the dual problem geometrically and use the duality theorem to determine an optimal solution of the primal problem.

Solution. The dual problem reads as

$$\max\{2y_1 + y_2\},$$

$$y_1 - 2y_2 \leq 2,$$

$$-y + y_2 \leq 1,$$

$$y_1 + 2y_2 \leq 4,$$

$$y_1 \geq 0,\ y_2 \geq 0.$$

The graphical solution (**make it!**) gives $y_1^* = 3$, $y_2^* = \frac{1}{2}$. The second inequality is non-active. The complementary slackness condition in the duality theorem implies that $x_2^* = 0$. Since y_1^* and y_2^* are both positive, the complementary slackness condition implies also that the two inequality constraints in the primal problem are active. Taking into account that $x_2^* = 0$ we come up with the equations

$$\begin{aligned} x_1 + x_3 &= 2, \\ -2x_1 + 2x_3 &= 1. \end{aligned}$$

Solving them we obtain the following solution of the primal problem: $x^* = \left(\frac{3}{4}, 0, \frac{5}{4}\right)'$.

Exercise 4.15 Write down the dual problem to the following one:

$$\max\{2x_1 - x_2 + x_3\},$$

subject to

$$\begin{gathered} x_1 - x_2 + x_3 = 4, \\ x_2 + x_3 \geq 1, \\ 2x_1 + 3x_2 \leq 6, \\ x_2 \geq 0. \end{gathered}$$

Rewrite the dual problem in the GLP format and find its dual. (It should coincide with the original problem due to the duality Lemma 4.2.)

Exercise 4.16 Write down the dual problem to the following one:

$$\max\{2x_1 + 6x_2 + x_3\},$$

subject to

$$\begin{gathered} x_1 - 2x_2 - x_3 \leq -3, \\ 2x_1 + 3x_2 - x_3 \geq -3, \\ x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0. \end{gathered}$$

Solve the dual problem graphically, then use Theorem 4.4 to find all solutions of the primal problem (if any).

Exercise 4.17 Let a polyhedral set be defined by the (standard format) constraints

$$\begin{aligned} 2x_1 + x_3 &= 2, \\ x_1 + x_2 - 4x_4 &= 3, \\ x_2 - 2x_3 &= 1, \\ x_1,\, x_2,\, x_3,\, x_4 &\geq 0. \end{aligned}$$

Find all basis points.

Exercise 4.18 Consider the problem

$$\max\{x_1 + x_2\},$$

subject to

$$-x_1 + 3x_2 \geq 1,$$

$$2x_1 + 5x_2 \leq 10,$$

$$x_1,\, x_2 \geq 0.$$

1. Solve this problem geometrically.
2. Rewrite the problem in the standard LP format (by introducing two “Schlupfvariable”).
3. Find all basis points of the SLP formulation and the optimal one among them.
4. What is the relation between the basis points and the vertices of the set $K \subset \mathbf{R}^2$ of the original problem?

Chapter 5

Numerical Methods for Continuous Optimization

5.1 Linear optimization problems

Solving numerically large linear optimization problems was made possible by the invention of the so-called *simplex method* by George Dantzig in 1947. Although numerical methods based on diverse approaches have been developed—some of which theoretically more efficient—the simplex method remains the most used one. It will be at the focus of the present section.

5.1.1 The simplex method

The idea of the simplex method is to start from a basis point (extreme point of the polyhedral set defined by the constraints) and to move in each step to a neighboring basis point that gives a better objective value. This can be done by a modest number of calculations and, moreover, it turns out that if at a certain step a better neighboring basis point does not exist, then we have reached a solution of the problem.

We consider the standard format linear optimization problem

$$\min_{x \in \mathbf{R}^n} \langle c, x \rangle, \tag{5.1}$$

subject to

$$Ax = b, \tag{5.2}$$

$$x \geq 0. \tag{5.3}$$

Here $c \in \mathbf{R}^n$, A is an $(m \times n)$-dimensional matrix with column-vectors $A^1, \dots, A^n$, and $b \in \mathbf{R}^m$. We assume that $\operatorname{rank}(A) = m$, which is not a restriction of the generality, as we shall see in the next subsection.

We remind that under the condition $\operatorname{rank}(A) = m$ a point x that satisfies (5.2) and (5.3) is called *basis point* if there exists a set of m indexes $I^b(x) = \{i_1, \dots, i_m\}$ such that the columns $A^{i_1}, \dots, A^{i_m}$ are linearly independent and $x_j = 0$ for $j \notin I^b(x)$ (see Definition 4.10). The set of indexes $I^b(x)$ of a basis point x is called *index set* of the basis point and its elements are called *basis indexes*.

1. We start the derivation of the simplex method with some preliminary considerations.

Let us assume that a basis point $\bar{x}$ with an index set $I^b(\bar{x})$ is given. This means that $\operatorname{rank}\{A^i\}_{i \in I^b(\bar{x})} = m$ and $\bar{x}_j = 0$ for $j \notin I^b(\bar{x})$. The system of equations (5.2) can be rewritten as

$$\sum_{i \in I^b(\bar{x})} x_i A^i + \sum_{j \notin I^b(\bar{x})} x_j A^j = b. \tag{5.4}$$

Since $\operatorname{rank}\{A^i\}_{i \in I^b(\bar{x})} = m = |I^b(\bar{x})|$, one can solve equation (5.4) with respect to x_i, $i \in I^b(\bar{x})$. Thus, we obtain a system of equalities that is equivalent to (5.2) and has the form

$$x_i + \sum_{j \notin I^b(\bar{x})} \alpha_{i,j} x_j = \beta_i, \quad i \in I^b(\bar{x}), \tag{5.5}$$

with appropriate numbers $\alpha_{i,j}$ and β_i. To do this formally we denote by $\bar{A}$ the matrix with columns $A^{i_1}, \dots, A^{i_m}$. Since rank $\bar{A} = m$ this matrix is invertible. Multiplying (5.2) by $\bar{A}^{-1}$ from the left, we obtain $\bar{A}^{-1} A x = \bar{A}^{-1} b$. Then in the representation (5.4) we have $\beta = \bar{A}^{-1} b$ and $\alpha_{*,j}$ is the j-column of $\bar{A}^{-1} A$ for $j \notin I^b(\bar{x})$ (the columns with indexes $j_1, \dots, j_m$ form the unit $(m \times m)$-matrix).

Notice that system (5.5) is satisfied by $\bar{x}$, and since $\bar{x}_j = 0$ for $j \notin I^b(\bar{x})$, we have $\beta_i = \bar{x}_i \geq 0$ (the inequality is even strict if the basis point $\bar{x}$ is non-degenerate).

The representation (5.5) of the equality constraints is called *adapted* to the basis point $\bar{x}$.

2. Let $\bar{x}$ be a basis point (for shortness we denote $\bar{I} := I^b(\bar{x})$) and let the constraints (5.2) be written in the form (5.5) adapted to $\bar{x}$. Now we shall investigate the following question: given an index $\nu \notin \bar{I}$, can we find an index $\kappa \in \bar{I}$ so that if we remove κ from the set $\bar{I}$ and replace it with ν, there will exist a basis point $\tilde{x}$ with the new index set $I^b(\tilde{x}) = \tilde{I} := \big(\bar{I} \setminus \{\kappa\}\big) \cup \{\nu\}$?

Let $\kappa \in \bar{I}$ be an index for which $\alpha_{\kappa,\nu} > 0$ (if such exists). If $\tilde{x}$ is a basis point with index set $\tilde{I}$ then, in particular, it is feasible and must satisfy (5.5). Moreover, $\tilde{x}_j$ must equal zero for $j \notin \tilde{I}$, while $\tilde{x}_j \geq 0$ for $j \in \tilde{I}$. Thus, we obtain from (5.5) that

$$\tilde{x}_i + \alpha_{i,\nu} \tilde{x}_\nu = \beta_i, \quad i \in \bar{I}.$$

Since $\tilde{x}_\kappa = 0$ (due to $\kappa \notin \tilde{I}$) we have $\alpha_{\kappa,\nu}\tilde{x}_\nu = \beta_\kappa$. Then

$$\tilde{x}_\nu = \frac{\beta_\kappa}{\alpha_{\kappa,\nu}} \qquad \tilde{x}_i = \beta_i - \frac{\alpha_{i,\nu}}{\alpha_{\kappa,\nu}}\beta_\kappa, \quad i \in \bar{I}. \tag{5.6}$$

For $\tilde{x}$ to be a basis point we have to ensure that $\tilde{x} \geq 0$. For $j \notin \tilde{I}$ we have $\tilde{x}_j = 0$ by definition. For $j = \nu$ we have $\tilde{x}_\nu = \frac{\beta_\kappa}{\alpha_{\kappa,\nu}} \geq 0$ since $\beta_\kappa \geq 0$. For $i \in \tilde{I} \setminus \{\nu\}$ we have

$$\tilde{x}_i \geq 0 \quad \Longleftrightarrow \quad \beta_i - \frac{\alpha_{i,\nu}}{\alpha_{\kappa,\nu}}\beta_\kappa \geq 0.$$

Thus we obtain that for the fixed κ and ν with $\alpha_{\kappa,\nu}$, a basis point $\tilde{x}$ with the basis index set $\tilde{I}$ exists if and only if

$$\alpha_{\kappa,\nu} > 0 \quad \text{and} \quad \alpha_{\kappa,\nu}\,\beta_i - \alpha_{i,\nu}\,\beta_\kappa \geq 0, \quad i \in \bar{I} \setminus \{\kappa\}, \tag{5.7}$$

and its non-zero elements are determined by the formulas (5.6). (Notice that for $i = \kappa$ the second inequality in (5.7) is automatically satisfied as an equality.)

We have chosen κ such that $\alpha_{\kappa,\nu} > 0$ (if such exists). For an index $i \in \bar{I} \setminus \{\kappa\}$ for which $\alpha_{i,\nu} \leq 0$, the second inequality in (5.7) is satisfied, since $\beta_i \geq 0$ and $\alpha_{\kappa,\nu} > 0$. For the indexes $i \in \bar{I} \setminus \{\kappa\}$ for which $\alpha_{i,\nu} > 0$ the second inequality in (5.7) is equivalent to

$$\frac{\beta_i}{\alpha_{i,\nu}} \geq \frac{\beta_\kappa}{\alpha_{\kappa,\nu}}.$$

Summarizing, if for a given index $\nu \in \{1, \dots, n\}$ the index κ is chosen so that

$$\frac{\beta_\kappa}{\alpha_{\kappa,\nu}} = \min_{i \in \bar{I}^+} \frac{\beta_i}{\alpha_{i,\nu}}, \qquad \text{where } \bar{I}^+ := \{i \in \bar{I} : \ \alpha_{i,\nu} > 0\}, \tag{5.8}$$

then the point $\tilde{x}$ defined by (5.6) (with the rest of the components equal to zero) is a basis point. Notice that we have assumed that an index $\kappa \in \bar{I}$ with $\alpha_{\kappa,\nu} > 0$ exists, so that the set $\bar{I}^+$ is non-empty. The case in which such κ does not exist for the chosen ν will be investigated in Lemma 5.2 below.

3. Now assume that for a given $\nu \notin \bar{I}$ the set $\bar{I}^+$ is nonempty and κ is chosen so that (5.8) holds. In order to check whether the new basis point $\tilde{x}$ gives a better objective value than the basis point $\bar{x}$ we consider the difference $\Delta := c'\bar{x} - c'\tilde{x}$. Taking into account (5.6) and also that $\bar{x}_i = \beta_i$ for $i \in \bar{I}$ we obtain that

$$\Delta \,=\, \sum_{i \in \bar{I}} c_i \beta_i - \sum_{i \in \bar{I}} c_i \left(\beta_i - \frac{\alpha_{i,\nu}}{\alpha_{\kappa,\nu}}\beta_\kappa\right) - c_\nu \frac{\beta_\kappa}{\alpha_{\kappa,\nu}} \,=\, \sum_{i \in \bar{I}} c_i \frac{\alpha_{i,\nu}}{\alpha_{\kappa,\nu}}\beta_\kappa - c_\nu \frac{\beta_\kappa}{\alpha_{\kappa,\nu}}.$$

Hence

$$\Delta = \frac{\beta_\kappa}{\alpha_{\kappa,\nu}} \left(\sum_{i \in \bar{I}} c_i \alpha_{i,\nu} - c_\nu \right) = -\frac{\beta_\kappa}{\alpha_{\kappa,\nu}} \Delta_\nu, \tag{5.9}$$

where

$$\Delta_\nu := c_\nu - \sum_{i \in \bar{I}} c_i \, \alpha_{i,\nu}. \tag{5.10}$$

Since $\frac{\beta_\kappa}{\alpha_{\kappa,\nu}} \geq 0$, the inequality $c'\bar{x} \geq c'\tilde{x}$ holds if and only if $\Delta_\nu \leq 0$.

4. If the index ν (which enters the basis index set) and the index κ (which leaves the basis index set) are fixed so that $\tilde{x}$ is a new basis point, we want to represent the equality constraints $Ax = b$ in a form adapted to $\tilde{x}$. To do this, we just solve the equation with index κ in (5.5) with respect to x_ν and substitute the resulting expression in the rest of the equations. Thus

$$x_\nu + \frac{1}{\alpha_{\kappa,\nu}} x_\kappa + \sum_{j \notin (\bar{I} \cup \{\nu\})} \frac{\alpha_{\kappa,j}}{\alpha_{\kappa,\nu}} x_j = \frac{\beta_\kappa}{\alpha_{\kappa,\nu}}. \tag{5.11}$$

From (5.5) we have for $i \in \bar{I} \setminus \{\kappa\} = \tilde{I} \setminus \{\nu\}$ that

$$x_i + \alpha_{i,\nu} x_\nu + \sum_{j \notin (\bar{I} \cup \{\nu)\}} \alpha_{i,j} \, x_j = \beta_i$$

and substituting x_ν from (5.11) we obtain

$$x_i + \alpha_{i,\nu} \left(\frac{\beta_\kappa}{\alpha_{\kappa,\nu}} - \frac{1}{\alpha_{\kappa,\nu}} x_\kappa - \sum_{j \notin (\bar{I} \cup \{\nu\})} \frac{\alpha_{\kappa,j}}{\alpha_{\kappa,\nu}} x_j \right) + \sum_{j \notin (\bar{I} \cup \{\nu\})} \alpha_{i,j} \, x_j = \beta_i.$$

From the above relations we obtain a representation of the equality constraints $Ax = b$ in a form adapted to the new basis point $\tilde{x}$:

$$x_i + \sum_{j \notin \tilde{I}} \tilde{\alpha}_{i,j} \, x_j = \tilde{\beta}_i, \quad i \in \tilde{I}, \tag{5.12}$$

where for $j \in \{1, \ldots, n\} \setminus \tilde{I}$

$$\tilde{\alpha}_{\nu,j} = \begin{cases} \frac{1}{\alpha_{\kappa,\nu}} & \text{if } j = \kappa \\ \frac{\alpha_{\kappa,j}}{\alpha_{\kappa,\nu}} & \text{if } j \neq \kappa \end{cases}, \quad \tilde{\beta}_\nu = \frac{\beta_\kappa}{\alpha_{\kappa,\nu}}, \tag{5.13}$$

$$\tilde{\alpha}_{i,j} = \begin{cases} -\frac{\alpha_{i,\nu}}{\alpha_{\kappa,\nu}} & \text{if } j = \kappa \\ \alpha_{i,j} - \alpha_{i,\nu} \frac{\alpha_{\kappa,j}}{\alpha_{\kappa,\nu}} & \text{if } j \neq \kappa \end{cases}, \quad \tilde{\beta}_i = \beta_i - \frac{\alpha_{i,\nu}}{\alpha_{\kappa,\nu}} \beta_\kappa, \quad i \in \tilde{I} \setminus \{\nu\}. \tag{5.14}$$

The above considerations lead us to the following algorithm.

Step 0. Assume that the equality constraints have already been written in the form (5.5) adapted to a given basis point $\bar{x}$ with basis index set $\bar{I} = I^b(\bar{x})$. The issue of how to find an initial basis solution $\bar{x}$, if such exists at all, will be elaborated in the next subsection.

Step 1. Find an index $\nu \notin \bar{I}$ (to be eventually included in the new basis index set) such that $\Delta_\nu := c_\nu - \sum_{i\in\bar{I}} c_i\alpha_{i,\nu} < 0$. If $\Delta_\nu \geq 0$ for all $\nu \notin \bar{I}$, then the algorithm terminates. We shall prove in Lemma 5.1 below that in this case $\bar{x}$ is an optimal solution. If $\Delta_\nu < 0$ continue with the next step.

Step 2. If $\alpha_{i,\nu} \leq 0$ for every $i \in \bar{I}$ (that is, $\bar{I}^+ = \emptyset$, see (5.8)) then the algorithm terminates. We shall prove in Lemma 5.2 that in this case the problem does not have an optimal solution since the objective function is unbounded from below on the set of feasible points.

If there exists $\kappa \in \bar{I}$ such that $\alpha_{\kappa,\nu} > 0$, then we define a new basis point $\tilde{x}$ by (5.6), where the index κ that is removed from the basis index set $\bar{I}$ is determined from (5.8). According to (5.9) $c'\tilde{x} \leq c'\bar{x}$. Then continue with the next step.

Step 3. Using (5.13) and (5.14) calculate the coefficients $\tilde{\alpha}$ and $\tilde{\beta}$ of the adapted representation (5.12) of the equality constraints corresponding to the basis point $\tilde{x}$, then go back to Step 1 with the new basis point $\bar{x} := \tilde{x}$. This completes the loop of the simplex method.

It remains to prove the claims in the termination cases at Step 1 and Step 2. The convergence and the numerical complexity of the algorithm will be discussed in Subsection 5.1.3.

Lemma 5.1 *If for a basis point $\bar{x}$ and for every $\nu \notin I^b(\bar{x}) =: \bar{I}$ it holds that $\Delta_\nu \geq 0$ (see (5.10)), then $\bar{x}$ is an optimal solution.*

Proof. For an arbitrary feasible point x we have

$$c'x = \sum_{j=1}^{n} c_j x_j = \sum_{i\in\bar{I}} c_i x_i + \sum_{j\notin\bar{I}} c_j x_j$$

and using the representation (5.5) of the equality constraints adapted to $\bar{x}$ we obtain

$$\begin{aligned} c'x &= \sum_{i\in\bar{I}} c_i \left(\beta_i - \sum_{j\notin\bar{I}} \alpha_{i,j}\, x_j\right) + \sum_{j\notin\bar{I}} c_j x_j \\ &= \sum_{i\in\bar{I}} c_i\beta_i + \sum_{j\notin\bar{I}} \left(c_j - \sum_{i\in\bar{I}} c_i\alpha_{i,j}\right) x_j = \sum_{i\in\bar{I}} c_i\beta_i + \sum_{j\notin\bar{I}} \Delta_j x_j. \end{aligned} \tag{5.15}$$

Since $x_j \geq 0$ and $\Delta_j \geq 0$, we have $c'x \geq \sum_{i\in\bar{I}} c_i\beta_i = c'\bar{x}$, which completes the proof. Q.E.D.

Lemma 5.2 *If ν is fixed as in Step 1 of the simplex algorithm and $\alpha_{i,\nu} \leq 0$ for every $i \in I^b(\bar{x}) =: \bar{I}$, then the problem does not have an optimal solution since the objective function is unbounded from below on the set of feasible points.*

Proof. Assume that $\alpha_{i,\nu} \leq 0$ for every $i \in \bar{I}$. We shall define a new feasible point x in the following way. Define $x_\nu = N$ with an arbitrarily large number N, and $x_j = 0$ for $j \notin \bar{I}$ and $j \neq \nu$. For $i \in \bar{I}$ we define x_i from (5.5):

$$x_i = \beta_i - \sum_{j\notin\bar{I}} \alpha_{i,j}\, x_j = \beta_i - \alpha_{i,\nu} N.$$

The non-negativity conditions are satisfied since $\alpha_{i,\nu} \leq 0$. Thus x is a feasible point for any $N \geq 0$. On the other hand, from the representation (5.5) of the equality constraints we may express the objective value corresponding to x as in (5.15):

$$c'x = \sum_{i\in\bar{I}} c_i\beta_i + \sum_{j\notin\bar{I}} \Delta_j x_j.$$

According to the choice of $\nu \notin \bar{I}$ we have $\Delta_\nu < 0$. Thus we obtain that

$$c'x = \sum_{i\in\bar{I}} c_i\beta_i + \Delta_\nu N \longrightarrow -\infty \text{ with } N \to +\infty.$$

This completes the proof of the lemma. Q.E.D.

Example 5.3 Consider the problem

$$\min\{-3x_1 - x_4 - 3x_6\}$$

$$\begin{array}{rrrrrrcr} x_1 & +x_2 & & -x_4 & -x_5 & -2x_6 & = & -3 \\ -2x_1 & -2x_2 & +x_3 & & & -x_6 & = & 2 \\ 3x_1 & +2x_2 & -x_3 & +2x_4 & +x_5 & +4x_6 & = & 3 \end{array}$$

$$x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6 \geq 0.$$

Step 0. First we illustrate the preliminary Step 0 of the simplex algorithm (which will be elaborated in the next subsection). Let us take $\bar{I} = \{2, 3, 5\}$ (for now this choice of $\bar{I}$ can be viewed just as an appropriate guess). Then

$$\bar{A} = \begin{pmatrix} 1 & 0 & -1 \\ -2 & 1 & 0 \\ 2 & -1 & 1 \end{pmatrix}, \qquad \bar{A}^{-1} = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 3 & 2 \\ 0 & 1 & 1 \end{pmatrix}.$$

Then the matrix $\mathcal{A} := \bar{A}^{-1}A$ and the vector $\beta := \bar{A}^{-1}b$ are

$$\mathcal{A} = \begin{pmatrix} 2 & 1 & 0 & 1 & 0 & 1 \\ 2 & 0 & 1 & 2 & 0 & 1 \\ 1 & 0 & 0 & 2 & 1 & 3 \end{pmatrix}, \qquad \beta = \begin{pmatrix} 2 \\ 6 \\ 5 \end{pmatrix}.$$

Thus the equality constraints take the following form adapted to the basis (x_2, x_3, x_5):

$$\begin{array}{ccccccccc} x_2 & & & + & (2x_1 & +x_4 & +x_6) & = & 2 \\ & x_3 & & + & (2x_1 & +2x_4 & +x_6) & = & 6 \\ & & x_5 & + & (x_1 & +2x_4 & +3x_6) & = & 5. \end{array} \tag{5.16}$$

The basis solution is then $(\bar{x}_2, \bar{x}_3, \bar{x}_5) = (2, 6, 5)$, the rest of the components are zero. This completes the preliminary Step 0 of the simplex method. As mentioned above this part assumes that a basis index set is given and involves the computationally hard problem of matrix inversion. However, this preliminary step will be avoided by the technique presented in the next subsection.

Step 1. With $\alpha_{i,j}$ obtained in (5.16) we calculate (see (5.10))

$$\Delta_1 = c_1 - (c_2\alpha_{2,1} + c_3\alpha_{3,1} + c_5\alpha_{5,1}) = -3 - (0 + 0 + 0) = -3,$$

$$\Delta_4 = c_4 - (c_2\alpha_{2,4} + c_3\alpha_{3,4} + c_5\alpha_{5,4}) = -1 - (0 + 0 + 0) = -1,$$

$$\Delta_6 = c_6 - (c_2\alpha_{2,6} + c_3\alpha_{3,6} + c_5\alpha_{5,6}) = -3 - (0 + 0 + 0) = -3.$$

we can take any of the indexes 1, 4 and 6 as a new basis index. Let us take $\nu = 1$.

Step 2. We have $\bar{I}^+ := \{i : \ \alpha_{i,1} > 0\} = \{2, 3, 5\}$ and calculate

$$\min_{i \in \bar{I}^+} \frac{\beta_i}{\alpha_{i,1}} = \min \left\{ \frac{\beta_2}{\alpha_{2,1}}, \frac{\beta_3}{\alpha_{3,1}}, \frac{\beta_5}{\alpha_{5,1}} \right\} = \min \left\{ \frac{2}{2}, \frac{6}{2}, \frac{5}{1} \right\} = \frac{2}{2},$$

thus we choose the index $\kappa = 2$ to be removed from the basis.

Step 3. We calculate the representation of the equality constraints adapted to the new basis (x_1, x_3, x_5). This can be done by using the formulas (5.13) and (5.14) or directly, as below. First we express the new basis variable x_1 from the equation for $x_\kappa = x_2$, which leaves the basis:

$$x_1 + \left(\frac{1}{2} x_2 + \frac{1}{2} x_4 + \frac{1}{2} x_6 \right) = 1.$$

Then we exclude x_1 from the other two equations:

$$x_3 + \left[2 \left(1 - \left(\frac{1}{2} x_2 + \frac{1}{2} x_4 + \frac{1}{2} x_6 \right) \right) + 2x_4 + x_6 \right] = 6,$$

$$x_5 + \left[1 - \left(\frac{1}{2} x_2 + \frac{1}{2} x_4 + \frac{1}{2} x_6 \right) + 2x_4 + 3x_6 \right] = 5.$$

So we obtain the following representation of the equality constraints adapted to the basis (x_1, x_3, x_5):

$$\begin{array}{ccccccccc}
x_1 & & & + & \left(\frac{1}{2}x_2\right. & +\frac{1}{2}x_4 & + & \left.\frac{1}{2}x_6\right) & = 1 \\
& x_3 & & + & (-x_2 & +x_4 & &) & = 4 \\
& & x_5 & + & \left(-\frac{1}{2}x_2\right. & +\frac{3}{2}x_4 & + & \left.\frac{5}{2}x_6\right) & = 4.
\end{array}$$

Now we turn back again to Step 1. This time we have $\bar{I} = \{1, 3, 5\}$, calculate

$$\Delta_2 = c_2 - (c_1\alpha_{1,2} + c_3\alpha_{3,2} + c_5\alpha_{5,2}) = 0 - \left((-3)\frac{1}{2} + 0.(-1) + 0.(-\frac{1}{2}) \right) = \frac{3}{2},$$

$$\Delta_4 = c_4 - (c_1\alpha_{1,4} + c_3\alpha_{3,4} + c_5\alpha_{5,4}) = -1 - \left((-3)\frac{1}{2} + 0.(1) + 0.(\frac{3}{2}) \right) = \frac{1}{2},$$

$$\Delta_6 = c_6 - (c_1\alpha_{1,6} + c_3\alpha_{3,6} + c_5\alpha_{5,6}) = -3 - \left((-3)\frac{1}{2} + 0.(0) + 0.(\frac{5}{2}) \right) = -\frac{3}{2}$$

and find $\nu = 6$ according to the rule formulated in Step 1. Since $\Delta_6 < 0$ we continue with Step 2. From the coefficients $\alpha_{1,6} = \frac{1}{2}$, $\alpha_{3,6} = 0$ and $\alpha_{5,6} = \frac{5}{2}$ only the first and the third are positive, thus $\bar{I}^+ = \{1, 5\}$. Then

$$\min_{i \in \bar{I}^+} \frac{\beta_i}{\alpha_{i,6}} = \min \left\{ \frac{\beta_1}{\alpha_{1,6}}, \frac{\beta_5}{\alpha_{5,6}} \right\} = \min \left\{ \frac{1}{1/2}, \frac{4}{5/2} \right\} = \frac{8}{5},$$

thus we choose the index $\kappa = 5$ to be removed from the basis. In Step 3 we solve the third equation with respect to x_6 and substitute in the other two equations:

$$\begin{aligned} x_6 + \frac{2}{5}\left(-\frac{1}{2}x_2 + \frac{3}{2}x_4 + x_5\right) &= \frac{8}{5}, \\ x_1 + \left[\frac{1}{2}x_2 + \frac{1}{2}x_4 + \frac{1}{2}\left(\frac{8}{5} - \frac{2}{5}\left(-\frac{1}{2}x_2 + \frac{3}{2}x_4 + x_5\right)\right)\right] &= 1 \\ x_3 + (-x_2 + x_4) &= 4. \end{aligned}$$

From here we obtain the representation of the equality constraints that is adapted to the new basis (x_1, x_3, x_6):

$$\begin{array}{ccccccccc} x_1 & & & + & \left(\frac{3}{5}x_2\right. & +\frac{1}{5}x_4 & - \frac{1}{5}x_5\left.\right) & = & \frac{1}{5} \\ & x_3 & & + & (-x_2 & +x_4 &) & = & 4 \\ & & x_6 & + & \left(-\frac{1}{5}x_2\right. & +\frac{3}{5}x_4 & + \frac{2}{5}x_5\left.\right) & = & \frac{8}{5}. \end{array}$$

We return again to Step 1. We calculate

$$\Delta_2 = c_2 - (c_1\alpha_{1,2} + c_3\alpha_{3,2} + c_6\alpha_{6,2}) = 0 - \left((-3)\frac{3}{5} + 0.(-1) + (-3)\left(-\frac{1}{5}\right)\right) = \frac{6}{5},$$

$$\Delta_4 = c_4 - (c_1\alpha_{1,4} + c_3\alpha_{3,4} + c_6\alpha_{6,4}) = -1 - \left((-3)\frac{1}{5} + 0.(1) + (-3)\frac{3}{5}\right) = \frac{7}{5},$$

$$\Delta_5 = c_5 - (c_1\alpha_{1,5} + c_3\alpha_{3,5} + c_6\alpha_{6,5}) = 0 - \left((-3)\left(-\frac{1}{5}\right) + 0.(0) + (-3)\frac{2}{5}\right) = \frac{3}{5}.$$

Since $\Delta_j \geq 0$ for every $j \notin \bar{I}$, that is $j \in \{2, 4, 5\}$, the algorithm terminates and $x^* = \left(\frac{1}{5}, 0, 4, 0, 0, \frac{8}{5}\right)'$ is an optimal solution.

Exercise 5.4 Solve by the simplex method the following problem

$$\min\{-3x_2 - 3x_3 + 2x_5\}$$

$$\begin{array}{ccccccc} -x_1 & -2x_2 & -x_3 & & +2x_5 & & = -3 \\ & 2x_2 & +x_3 & +x_4 & +4x_5 & & = \;\;6 \\ & x_2 & +3x_3 & & +4x_5 & +x_6 & = \;\;3 \end{array}$$

$$x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6 \geq 0.$$

Hint: Follow Example 5.3, using that the obvious initial basis point with index set $\{1, 4, 6\}$ (so that Step 0 is not needed).

5.1.2 Finding an initial basis point

As above, we consider the standard format linear optimization problem

$$\min_{x \in \mathbf{R}^n} \; c'x, \tag{5.17}$$

subject to

$$Ax = b, \tag{5.18}$$

$$x \geq 0, \tag{5.19}$$

where $x \in \mathbf{R}^n$ and the number of constraints in (5.18) is m. Obviously, one can assume without any restriction that $b \geq 0$ (multiplying by -1 the equation with negative right-hand sides).

In order to obtain an initial basis point and the corresponding adapted representation of (5.18) we consider the following auxiliary problem with m additional variables $z \in \mathbf{R}^m$, indexed by the indexes $n+1, \ldots, n+m$:

$$\min_{(x,z) \in \mathbf{R}^n \times \mathbf{R}^m} \sum_{j=n+1}^{n+m} z_j \tag{5.20}$$

subject to

$$z + Ax = b, \tag{5.21}$$

$$x \geq 0, \quad z \geq 0. \tag{5.22}$$

Notice that for this problem the rank of the matrix of the constraints, $(I \; A)$ (where I is the identity matrix), equals m, thus our assumption about the rank made in the previous subsection is fulfilled.

For the problem (5.20) we readily have a basis point $(x, z) := (0, b)$ (remember that $b \geq 0$) with the corresponding basis representation adapted to the basis variables $z_1, \ldots, z_m$. (Strictly speaking we should write $\begin{pmatrix} x \\ z \end{pmatrix}$ and $\begin{pmatrix} 0 \\ b \end{pmatrix}$ instead of (x, z) and $(0, b)$, but this slight abuse of dimensions will not lead to a confusion.) Then we can apply the simplex algorithm for solving this problem. Since the feasible set is nonempty and the objective value is bounded from below by 0 due to the constraints $z \geq 0$, this problem has a solution (the proof uses Lemma 4.5 like in the proof of Theorem 4.4, part A). Let (x^*, z^*) be a basis optimal solution obtained by the simplex algorithm (with the anti-cycling modification of the simplex method presented in the next subsection such a solution will be found in a finite number of steps).

The next propositions and the consequent analysis explain how the obtained solution (x^*, z^*) helps to find a basis point and the corresponding adapted representation in our original problem (5.17)–(5.19).

Proposition 5.5

(i) If $z^* \neq 0$, then problem (5.17)–(5.19) does not have a solution due to emptiness of the feasible set.

(ii) If $z^* = 0$, then x^* is an extreme point of the feasible set (5.18), (5.19).

Proof. (i) If $z^* \neq 0$, then the optimal objective value in problem (5.20)–(5.22) is strictly positive. On the other hand, if we assume that problem (5.17)–(5.19) has a feasible point x, then $(x, 0)$ is a feasible point for (5.20)–(5.22) and the corresponding objective value is 0, which is a contradiction.

(ii) Now let $z^* = 0$. Then x^* is a feasible point of (5.17)–(5.19). It remains to show that it is a extreme point.

Assume that $x^* = \alpha x^1 + (1-\alpha)x^2$ with $\alpha \in (0,1)$ and some feasible points x^1 and x^2. Then $(x^*, 0) = \alpha(x^1, 0) + (1-\alpha)(x^2, 0)$. According to Definition 4.10 and Proposition 4.9, $(x^*, 0)$ is an extreme point of the polyhedral set (5.21), (5.22), hence $(x^1, 0) = (x^2, 0)$. Thus $x^1 = x^2$. This implies that x^* is an extreme point of the feasible set of (5.18), (5.19). Q.E.D.

Notice, that in the case (ii) we cannot claim that x^* is a basis point for the problem (5.17)–(5.19), since it is not assumed that $\text{rank}(A) = m$.

The above proposition solves our original problem (5.17)–(5.19) if $z^* \neq 0$: its feasible set is empty in this case.

If $z^* = 0$ the proposition does not solve completely the problem of finding an initial basis point of the original problem (5.17)–(5.19). The reason is that although $z^* = 0$, some of the z-variables can still belong the basis (which means that the basis point $(x^*, 0)$ is degenerate) and we do not have a basis representation adapted to x^*.

Let us consider the case $z^* = 0$ in detail. If the basis index set $I^* := I^b(x^*, 0)$ of the basis point $(x^*, 0)$ contains only indexes from $\{1, \ldots, n\}$, then we have found a basis point of our original problem (5.17)–(5.19) and I^* is its basis index set (in this case necessarily $\text{rank}(A) = m$).

If I^* contains some indexes bigger than n then we may present $I^* = \bar{I} \cup I^z$ with $\bar{I} \subset \{1, \ldots, n\}$, $I^z \subset \{n{+}1, \ldots, n{+}m\}$. Then the basis representation of the constraints (5.21) (adapted to I^*) has the form

$$x_i \ + \sum_{j \in \{1,\ldots,n\} \setminus \bar{I}} \alpha_{i,j}\, x_j \ + \sum_{j \in \{n+1,\ldots,n+m\} \setminus I^z} \alpha_{i,j}\, z_j = \beta_i, \quad i \in \bar{I}, \tag{5.23}$$

$$z_i \ + \sum_{j \in \{1,\ldots,n\} \setminus \bar{I}} \alpha_{i,j}\, x_j \ + \sum_{j \in \{n+1,\ldots,n+m\} \setminus I^z} \alpha_{i,j}\, z_j = 0, \quad i \in I^z. \tag{5.24}$$

If it happens that all coefficients $\alpha_{i,j}$ with $i \in I^z$ and $j \in \{1, \dots, n\} \setminus \bar{I}$ are equal to zero, then the second group of equations is independent of x, hence $z = 0$ is a solution of (5.24) for every x. Then $(x, 0)$ is a solution of (5.23), (5.24) if and only if x solves the system

$$x_i + \sum_{j \in \{1,\dots,n\} \setminus \bar{I}} \alpha_{i,j} \, x_j = \beta_i, \quad i \in \bar{I}. \tag{5.25}$$

On the other hand (5.23), (5.24) is equivalent to (5.21) and $(x, 0)$ is a solution of the latter if and only if x is a solution of (5.18). Thus we obtain that equation (5.25) is equivalent to (5.18). It is readily in a form adapted to x^* with basis index set $\bar{I}$. (Notice that the number of basis components is smaller than m now, but still equals the number of equations in (5.25).)

In the alternative case, there is some $\kappa \in I^z$ and $\nu \in \{1, \dots, n\} \setminus \bar{I}$ such that $\alpha_{\kappa,\nu} \neq 0$. Then we can solve the equation with index κ with respect to x_ν and substitute it in the remaining equations in (5.23) and (5.24). Notice that the right-hand sides do not change, thus remain non-negative. Thus we obtain a new basis index set and adapted basis representation for the basis point $(x^*, 0)$ having one more component in $\{1, \dots, n\}$ than $\bar{I}$. Therefore, after a finite number of steps we shall obtain a basis and basis representation of the constraints.

After having found a basis point and the corresponded adapted representation of the system $Ax = b$, we may start solving problem by the simplex method as described in the previous subsection.

In fact, the two stages of the simplex method (finding an initial basis representation of the constraints and the consequent iterations described in Subsection 5.1.1) can be joined together by considering the so-called *M-problem* ("Big-M-Methode"):

$$\min_{(x,z) \in \mathbf{R}^n \times \mathbf{R}^m} \left\{ c'x \, + \, M \sum_{j=n+1}^{n+m} z_j \right\} \tag{5.26}$$

subject to

$$z + Ax = b, \tag{5.27}$$

$$x \geq 0, \quad z \geq 0, \tag{5.28}$$

where M is a sufficiently large number. Here $(0, b)$ is again an initial basis point. Theoretically, M should be so large, that the term $c'x$ in the objective function plays no role in the implementation of the simplex method as long as there are some non-zero z-variables. Practically, M could be treated as a algebraic symbol (and kept as a symbol in the implementation of the simplex method) such that $aM > b$ for every numbers $a > 0$ and b.

Example 5.6 Consider the problem

$$\min\{4x_1 + 2x_2 - x_3\}$$

$$\begin{array}{rrrrl} x_1 & +5x_2 & -2x_3 & +2x_4 & = 4 \\ & -3x_2 & +2x_3 & -x_4 & = 2 \\ x_1 & -x_2 & +2x_3 & & = 8 \end{array}$$

$$x_1,\ x_2,\ x_3,\ x_4\ \geq\ 0.$$

In order to obtain an initial basis point we consider the auxiliary problem

$$\min\{z_5 + z_6 + z_7\}$$

$$\begin{array}{cccrrrrl} z_5 & & & +x_1 & +5x_2 & -2x_3 & +2x_4 & = 4 \\ & z_6 & & & -3x_2 & +2x_3 & -x_4 & = 2 \\ & & z_7 & +x_1 & -x_2 & +2x_3 & & = 8 \end{array}$$

$$x_1,\ x_2,\ x_3,\ x_4\, z_5,\ z_6,\ z_7\ \geq\ 0.$$

We have an initial basis representation for the basis index set $I^* = \{5,\ 6,\ 7\}$. The coefficients of the basis representation are

$$\begin{pmatrix} \alpha_{5,1} = 1 & \alpha_{5,2} = 5 & \alpha_{5,3} = -2 & \alpha_{5,4} = 2 \\ \alpha_{6,1} = 0 & \alpha_{6,2} = -3 & \alpha_{6,3} = 2 & \alpha_{6,4} = -1 \\ \alpha_{7,1} = 1 & \alpha_{7,2} = -1 & \alpha_{7,3} = 2 & \alpha_{7,4} = 0 \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_5 = 4 \\ \beta_6 = 2 \\ \beta_7 = 8 \end{pmatrix}.$$

Then, following the simplex algorithm, we calculate

$$\Delta_1 = c_1 - (c_5\alpha_{5,1} + c_6\alpha_{6,1} + c_7\alpha_{7,1}) = -2$$

$$\Delta_2 = c_2 - (c_5\alpha_{5,2} + c_6\alpha_{6,2} + c_7\alpha_{7,2}) = -1$$

$$\Delta_3 = c_3 - (c_5\alpha_{5,3} + c_6\alpha_{6,3} + c_7\alpha_{7,3}) = -2$$

$$\Delta_4 = c_4 - (c_5\alpha_{5,4} + c_6\alpha_{6,4} + c_7\alpha_{7,4}) = -1.$$

According to Step 1 of the simplex algorithm we can any of the indexes $\nu = 1,\ 2,\ 3,\ 4$ in the new basis. Let us choose $\nu = 1$. Then, according to Step 2, we compare

$$\frac{\beta_5}{\alpha_{5,1}} = \frac{4}{1} \quad \text{and} \quad \frac{\beta_7}{\alpha_{7,1}} = \frac{8}{1}$$

(since $\alpha_{6,1}$ is not positive) and choose $\kappa = 5$, so that z_5 will be excluded from the basis. We obtain the corresponding basis representation as described in Step 3 of the simplex method:

$$\begin{array}{cccrrrrl} x_1 & & & +5x_2 & -2x_3 & +2x_4 & +z_5 & = 4 \\ & z_6 & & -3x_2 & +2x_3 & -x_4 & & = 2 \\ & & z_7\ +\ (4 - 5x_2 + 2x_3 - 2x_4 - z_5) & -x_2 & +2x_3 & & & = 8. \end{array}$$

Thus the coefficient in the representation are

$$\begin{pmatrix} \alpha_{1,2}=5 & \alpha_{1,3}=-2 & \alpha_{1,4}=2 & \alpha_{1,5}=1 \\ \alpha_{6,2}=-3 & \alpha_{6,3}=2 & \alpha_{6,4}=-1 & \alpha_{6,5}=0 \\ \alpha_{7,2}=-6 & \alpha_{7,3}=4 & \alpha_{7,4}=-2 & \alpha_{7,5}=-1 \end{pmatrix}, \qquad \beta=\begin{pmatrix} \beta_1=4 \\ \beta_6=2 \\ \beta_7=4 \end{pmatrix}.$$

Then we calculate the corresponding

$$\Delta_2 = c_2-(c_1\alpha_{1,2}+c_6\alpha_{6,2}+c_7\alpha_{7,2}) = 9$$

$$\Delta_3 = c_3-(c_1\alpha_{1,3}+c_6\alpha_{6,3}+c_7\alpha_{7,3}) = -6$$

$$\Delta_4 = c_4-(c_1\alpha_{1,4}+c_6\alpha_{6,4}+c_7\alpha_{7,4}) = 3$$

$$\Delta_5 = c_5-(c_1\alpha_{1,5}+c_6\alpha_{6,5}+c_7\alpha_{7,5}) = 2.$$

The only possible choice of a new basis variable is $\nu = 3$. In the column of x_3 we have two positive coefficients and

$$\frac{\beta_6}{\alpha_{6,3}} = \frac{2}{2} \text{ and } \quad \frac{\beta_7}{\alpha_{7,3}} = \frac{4}{4}$$

are equal, so that we can choose either z_6 or z_7 to leave the basis. Let us choose z_6. As prescribed by the simplex method we solve for x_3 the equation where z_6 appears and obtain the equation

$$x_3 - \frac{3}{2}x_2 - \frac{1}{2}x_4 + \frac{1}{2}z_6 = 1$$

and excluding x_3 from the rest of the equations we obtain altogether the following system adapted to the basis $I^* = \{1,\, 3,\, 7\}$:

$$\begin{array}{ccccccccl} x_1 & & & +2x_2 & +x_4 & +z_5 & +z_6 & = & 6 \\ & x_3 & & -\frac{3}{2}x_2 & -\frac{1}{2}x_4 & & +\frac{1}{2}z_6 & = & 1 \\ & & z_7 & & & -z_5 & -2z_6 & = & 0. \end{array}$$

Calculating the next decrements we obtain $\Delta_2 = 1$, $\Delta_4 = 0$, $\Delta_5 = 2$, $\Delta_6 = 3$. Since there is no negative among them we have found an optimal basis solution $(6, 0, 1, 0, 0, 0)$ of the auxiliary problem. Observe that the variable z_7 belongs to the basis. Here we have $\bar{I} = \{1, 3\}$, $I^z = \{7\}$. However, all coefficients $\alpha_{i,j}$, with $i \in I^z$ and $j \in \{1, \dots, n\} \setminus \bar{I}$ are zero. According to algorithm we ignore the last equation and delete the columns corresponding to z_5 and z_6. Thus we obtain the representation

$$\begin{array}{ccccccl} x_1 & & +2x_2 & +x_4 & = & 6 \\ & x_3 & -\frac{3}{2}x_2 & -\frac{1}{2}x_4 & = & 1. \end{array}$$

Starting from this basis representation of the constraints in the original problem (with basis index set $\{1, 3\}$) we continue by applying the simplex method with the original objective function $c'x$. After two iterations we obtain the optimal solution $(0, 0, 4, 6)$ (please, do this!).

Notice that in this problem the rank r of the matrix A equals $2 < m$. This is the reason why one auxiliary variable remained in the basis at the solution of the auxiliary problem and the corresponding constraint was eliminated.

5.1.3 Anti-cycle enhancements

Theoretically, the simplex method, as presented in Subsection 5.1.1, may happen to be infinitely cycling, if the objective value is not strictly decreasing at every iteration. According to (5.9), a non-strict decrease may happen at a certain step of the algorithm only if $\beta_\kappa = 0$ (see Step 2 of the algorithm), in which case we have a degenerate basis. Thus cycling cannot happen if all basis points in our problem are non-degenerate. Of course, this property cannot be checked in advance. However, it turns out that the degeneracy is an "unstable" property: an arbitrarily small appropriate perturbation in the data of a degenerate problem leads to a non-degenerate problem. In contrast, the non-degenerate problems are stable. This explains why cycling does not happen in practice.

Nevertheless, several modifications of the simplex method that avoid cycles are known. We shall present a rather easy one, introduced by R. Bland in 1977. We remind that at each step of the simplex algorithm we first choose an index ν which to enter the basis. We know from the consideration in the previous section that ν had to satisfy $\Delta_\nu < 0$. The Bland rule requires at every step to choose the minimal index ν for which $\Delta_\nu < 0$.

Then a basis index κ has to be determined from (5.8) to leave the basis. Bland's rule requires to choose the minimal index κ for which (5.8) is satisfied (which makes sense only if the minimum in (5.8) is achieved at more than one index).

It can be proved that with these modifications cycles do not appear, hence the simplex method terminates after a finite number of steps.

5.1.4 Complexity of the simplex method and alternative methods

Although the simplex method finds a solution (or establishes non-existence of a solution) after a finite number of (exact) calculations, the question of how large this number is, is of theoretical and practical interest. Here, we provide some information related to this issue.

The standard format linear optimization problem is defined by $(n + 1) \times (m + 1)$

numbers – the components of the matrix A and the vectors b and c. Given a particular version of the simplex method (including the procedure for finding an initial basis point and the implemented anti-cycling rule), the amount of calculations needed to solve the problem depends on the size of the data, but also on the particular values of the coefficients. Let $N(n, m; P)$ be the number of calculations needed to solve a particular problem P of size $n \times m$.

One point of view on the complexity is the *worst-case criterion*, in which

$$\mathbf{N}(n, m) := \sup_P N(n, m; P)$$

is considered, with the maximum taken over all problems P in which the matrix A of size $m \times n$. Then the asymptotics of $\mathbf{N}(n, m)$ when n and m (or only n) tend to infinity can be used as a measure of the complexity of the simplex method.

A simplified and reasonable version of the above worst-case complexity criterion is the of number of calculations as a function of the size of the data (roughly $s = n\,m$):

$$\hat{\mathbf{N}}(s) := \max_{n\,m=s} \sup_P N(n, m; P).$$

A desired property would be that $\hat{\mathbf{N}}(s)$ increases with s in a polynomial way. However, for most of the used versions of the simplex method it has been shown that $\hat{\mathbf{N}}(s)$ grows exponentially with s. Said shortly, the simplex method has an *exponential (worst-case) complexity.*

However, there is a huge practical and statistical evidence (published in many hundreds of papers) that the simplex method performs much better than it could be expected from the worst-case analysis. This brings into consideration a different type of complexity criteria called *averaged* or *statistical.* In these criteria, the average number of calculations is taken into account, where the particular problems of a given size are generated randomly with a given probability distributions (also structural specificity or the density of the non-zero elements of the matrix A can be taken into account). For several probability distributions for the entries of A (normal, uniform, etc.) it has been proved that the expected computational complexity is polynomial. These results provide one explanation for the practically evident efficiency of the simplex method even for rather large problems. Even more, the empirically estimated statistical number of iterations of the simplex method is proportional to m. (The number of calculations at each iteration is proportional to $n\,m$.)

Since, in general, the simplex method has an exponential complexity, one could ask if alternative methods do exist that may provide a polynomial worst-case complexity. Such a method was first proposed by the Soviet (American since 1989 till his death in 2005) mathematician L. Khachiyan in 1979. He introduced the so-called *ellipsoidal methods* (to be presented in a lecture course for master students), which are applicable also for non-linear problems, and proved polynomial complexity (with a rather high degree) in the case of linear optimization problems. Various numerical methods with polynomial complexity appeared later on. However, the simplex method still remains one of the most efficient and the most used in practice.

Nowadays extra-large problems (with billions of unknowns) arise in the design of internet searching algorithms, in signal or image recognition, etc., which have rather specific structure and require special methods. This is the subject of the so-called *structural optimization* (also for non-linear problems), which is a hot topic in the mathematical optimization.

5.1.5 Additional examples and exercises

Example 5.7 Solve the problem

$$\min\{2x_1 + x_2 + 2x_3 + x_4\}$$

subject to

$$\begin{array}{ccccccccc} x_1 & + & x_2 & + & 5x_3 & + & x_4 & = & 7 \\ & & 2x_2 & + & 7x_3 & + & x_4 & = & 11, \end{array}$$

$$x_1,\, x_2,\, x_3,\, x_4 \geq 0.$$

by using the simplex method.

Hint: Use x_1 and one additional variable z_5 in the second equation as initial basis variables in the auxiliary problem for finding an initial basis for the given problem.

Solution. The auxiliary problem for finding an initial basis point is

$$\min\{z_5\}$$

subject to

$$\begin{array}{ccccccccccc} x_1 & & & + & x_2 & + & 5x_3 & + & x_4 & = & 7 \\ & & z_5 & + & 2x_2 & + & 7x_3 & + & x_4 & = & 11, \end{array}$$

$$x_1,\, x_2,\, x_3,\, x_4,\, z_5 \geq 0.$$

We have (with $c = (0, 0, 0, 0, 1)'$)

$$\Delta_2 = c_2 - (1c_1 + 2c_5) = -2,$$

$$\Delta_3 = c_3 - (5c_1 + 7c_5) = -7,$$

$$\Delta_4 = c_4 - (1c_1 + 1c_5) = -1.$$

We can chose any of the non-basis variables as a new basis variable. Let us take $\nu = 3$ (since Δ_2 is the minimal). Then according to the rule for choosing which basis variable to be removed we have to choose $\kappa = 1$, since

$$\frac{7}{5} < \frac{11}{7}.$$

Then we have

$$\begin{array}{ccccccccc} x_3 & & + & \frac{1}{5}x_1 & + & \frac{1}{5}x_2 & + & \frac{1}{5}x_4 & = & \frac{7}{5} \\ & z_5 & + & 2x_2 & + & 7\left(\frac{7}{5} - \frac{1}{5}x_1 - \frac{1}{5}x_2 - \frac{1}{5}x_4\right) & + & x_4 & = & 11, \end{array}$$

Hence, the adapted to the basis $(x_{,} z_5)$ representation of the equality constraints is

$$\begin{array}{ccccccccc} x_3 & & + & \frac{1}{5}x_1 & + & \frac{1}{5}x_2 & + & \frac{1}{5}x_4 & = & \frac{7}{5} \\ & z_5 & - & \frac{7}{5}x_1 & + & \frac{3}{5}x_2 & - & \frac{2}{5}x_4 & = & \frac{6}{5}, \end{array}$$

Then we calculate (still with $c = (0,0,0,0,1)'$)

$$\Delta_1 = c_1 - (c_3\frac{1}{5} + c_5\frac{-7}{5}) = \frac{7}{5},$$

$$\Delta_2 = c_2 - (c_3\frac{1}{5} + c_5\frac{3}{5}) = -\frac{3}{5},$$

$$\Delta_4 = c_4 - (c_3\frac{1}{5} + c_5\frac{-2}{5}) = \frac{2}{5}.$$

According to the rule of the simplex algorithm we have to choose $\nu = 2$, and since

$$\frac{7/5}{1/5} > \frac{6/5}{3/5}$$

we have to choose $\kappa = 5$. That is, z_5 leaves the basis and x_2 enters the basis.

We express

$$\begin{array}{ccccccccccc} x_2 & & - & \frac{7}{3}x_1 & - & \frac{2}{3}x_4 & + & \frac{5}{3}z_5 & = & 2 \\ & x_3 & + & \frac{1}{5}x_1 & + & \frac{1}{5}\left(2 - \frac{7}{3}x_1 - \frac{2}{3}x_4 - \frac{5}{3}z_5\right) & + & \frac{1}{5}x_4 & = & \frac{7}{5}, \end{array}$$

Since all the auxiliary variables have left the basis (this only z_5 in our problem and we ignore it further) we have found an initial basis for the original problem, $((x_2,\, x_3))$, and the adapted representation, found from the above equations, is

$$\begin{array}{ccccccc} x_2 & & - & \frac{7}{3}x_1 & - & \frac{2}{3}x_4 & = & 2 \\ & x_3 & + & \frac{2}{3}x_1 & + & \frac{1}{3}x_4 & = & 1. \end{array}$$

Now we calculate (this time with $c = (2, 1, 2, 1)'$)

$$\Delta_1 = c_1 - (c_2 \frac{-7}{3} + c_3 \frac{2}{3}) = 3,$$

$$\Delta_4 = c_4 - (c_2 \frac{-2}{3} + c_3 \frac{1}{3}) = 1.$$

Since Δ_1 and Δ_4 are both positive, we have reached an optimal solution, namely, $x^* = (0, 2, 1, 0)'$.

Exercise 5.8 Solve by the simplex method the following problem

$$\min\{2x_1 + x_2 + 2x_3 + x_4\}$$

subject to

$$4x_1 + 2x_2 + 13x_3 + 3x_4 = 17$$

$$x_1 + x_2 + 5x_3 + x_4 = 7$$

$$x_1,\ x_2,\ x_3,\ x_4 \geq 0.$$

Introduce first two additional variables in order to find an initial basis solution.

5.2 Non-linear optimization problems

5.2.1 Minimization of functions of a scalar variable

We start with the simplest case of minimization of $f : [a, b] \to \mathbf{R}$, where $[a, b]$ is a compact interval of real numbers. As it will be seen later, minimization of a scalar function is often an ingredient of optimization methods of functions in $\mathbf{R}^n$ without or with constraints.

Here, we present only a brief account of several methods, some of them applicable for non-differentiable functions, other making use of derivatives.

Let us assume that the function f is *unimodular*, that is, f is continuous and there exist $\alpha, \beta \in [a, b]$, $\alpha \leq \beta$ such that f is strictly monotone decreasing in $[a, \alpha]$, strictly monotone increasing in $[\beta, b]$ and is constant on $[\alpha, \beta]$. Clearly, the value of f in $[\alpha, \beta]$ equals $\inf_{x \in [a,b]} f(x)$.

1. Method of dividing the interval in half. This is a rather simple method that assumes unimodularity of f but does not assume differentiability. Let us fix a (presumably small) number $\delta \in (0, b - a)$. We set $a_0 = a$, $b_0 = b$ and carry the following iterations. Define $u' = (a_0 + b_0 - \delta)/2$ and $u'' = (a_0 + b_0 + \delta)/2$ and calculate $f(u')$ and $f(u'')$. Then we define a new interval $[a_1, b_1]$ as

$$[a_1, b_1] = [a_0, u''] \text{ if } f(u') \leq f(u'')$$

$$[a_1, b_1] = [u', b_0] \text{ if } f(u') > f(u'').$$

If f is unimodular on $[a_0, b_0]$ then obviously it is unimodular also on $[a_1, b_1]$ and contains a minimizer of f on $[a, b]$. Moreover, $b_1 - a_1 = (b - a)/2 + \delta/2$, hence $\delta \in (0, b_1 - a_1)$. Then we can continue the same procedure, generating a sequence of intervals $[a_k, b_k]$, each of which contains a minimizer of f on $[a, b]$. We can easily calculate that

$$b_k - a_k = \frac{b-a}{2^k} + \left(1 - \frac{1}{2^k}\right)\delta,$$

which implies, in particular, that $\delta \in (0, b_k - a_k)$ and the procedure can be continued. After $k \geq \log_2((b-a)/\delta)$ such iterations, the point $x_k = (a_k + b_k)/2$ will be at distance at most δ to a minimizer of f in $[a, b]$. Thus we have to chose in advance δ so small as the accuracy with which we want to solve the minimization problem.

2. Method of golden section. In the above method, the length of the interval $[a_k, b_k]$ containing a minimizer shortens almost by half at each iteration (assuming that δ is very small), but each iteration requires two computations of the function f, so that roughly the interval shortens by 1/4 (25%) per evaluation of f. The golden section method described below has a better performance in this sense: every calculation of f shortens the interval $[a_k, b_k]$ containing a minimizer by about 38.2%.

Golden section ("Goldener Schnitt") of the interval $[a_0, b_0] = [a, b]$ is a point $u \in [a_0, b_0]$ such that the ratio of the shorter subinterval to the longer one equals the ratio of the longer subinterval to the whole interval $[a_0, b_0]$. That is, if $u - a_0 < b_0 - u$, then for u to be a golden section it must hold that

$$\frac{u - a_0}{b_0 - u} = \frac{b_0 - u}{b_0 - a_0}.$$

Solving the above equation with respect to u we obtain a unique solution in $[0, 1]$:

$$u_0' = a_0 + (3 - \sqrt{5})(b_0 - a_0)/2 = a + 0.381966011...(b_0 - a_0).$$

In the case $u - a_0 > b_0 - u$ (that is, $[a_0, u]$ is the larger subinterval) we similarly obtain the point

$$u_0'' = a_0 + (\sqrt{5} - 1)(b_0 - a_0)/2 = a + 0.618033989...(b_0 - a_0).$$

A remarkable property of the golden section is that each of the two golden sections provides a golden section of the interval determined by the other one and containing it. That is, u_0' is (the bigger) golden section of $[a_0, u_0'']$ and u_0'' is (the smaller) golden section of $[u_0', b_0]$.

Exercise 5.9 Prove the above fact.

The golden section method begins with calculation of $f(u_0')$ and $f(u_0'')$ and setting

$$[a_1, b_1] = [a_0, u_0''] \text{ if } f(u_0') \le f(u_0'')$$

$$[a_1, b_1] = [u_0', b_0] \text{ if } f(u_0') > f(u_0'').$$

In the first case $u_1'' := u_0'$ is the larger golden section of $[a_1, b_1]$ and $u_1' = a_1 + (b_1 - u_1'') = a_1 + b_1 - u_1''$ is the smaller one. In the second case $u_1' = u_0''$, $u_1'' = b_1 - (u_1' - a_1)a_1 + b_1 - u_1'$. Then the process continues in the same way, generating a sequence of intervals $[a_k, b_k]$ each of which containing a minimizer of f.

It is important to notice that at each step, the value of f at one of the two new golden section points u_k' and u_k'' has already been calculated at the previous step, therefore, each step requires only one calculation of f. At each step the length of the interval shortens by a fraction of $(3 - \sqrt{5})/2 = 0.381966011...$, which is considerably larger than 0.25 that we obtained for the method of dividing by half.

3. Method using the Fibonacci numbers. The Fibonacci sequence is defined as

$$F_{k+2} = F_{k+1} + F_k, \quad k = 1, 2, \dots, \quad F_0 = F_1 = 1.$$

Let us fix a number n – the maximal number of iterations we are going to make. Then the method starts with

$$u_0' = a_0 + (b_0 - a_0)F_n/F_{n+2}, \quad u_0'' = a_0 + (b_0 - a_0)F_{n+1}/F_{n+2}.$$

and the next interval is chosen exactly as in the golden section method. Then at step $k \in \{1, \dots, n\}$ we define

$$u_k' = a_k + (b_0 - a_0)F_{n-k}/F_{n+2}, \quad u_k'' = a_k + (b_0 - a_0)F_{n-k+1}/F_{n+2}.$$

As in the golden section method it is important to notice that at each step one of the new section points has already appeared at the previous step, therefore at each step the function f has to be evaluated only ones.

The method of Fibonacci numbers is in a certain sense optimal among methods in which the number of calculations of f is given in advance, but the points at which f has to be calculated are not given and their choice may depend on the results obtained in the previous iterations. We shall not discuss this point. Despite its optimality, the method of the Fibonacci numbers is of little practical value.

4. Method of Tangents. Now we assume that $f : [a, b] \to \mathbf{R}$ is convex and differentiable. As before we set $a_0 = a$, $b_0 = b$ and generate a sequence of shrinking intervals $[a_k, b_k]$ containing a minimizer of f. The idea is at every step to approximate from below f on $[a_k, b_k]$ by a function φ which consists of two linear pieces, and to use the minimizer of this function to obtain a section that defines the next interval $[a_{k+1}, b_{k+1}]$. Namely, we define

$$\varphi_1(x) = \max_{x \in [a_0, b_0]} \{f(a_0) + f'(a_0)(x - a_0),\ f(b_0) + f'(b_0)(x - b_0)\}$$

and take the point u_0 which minimizes this function (see Figure 5.1), that is (due to the convexity of f), the solution of the equation

$$f(a_0) + f'(a_0)(x - a_0) = f(b_0) + f'(b_0)(x - b_0).$$

If $f'(u_0) = 0$ then we stop, since u_0 is a solution. Otherwise, we set

$$[a_1, b_1] = [a_0, u_0] \ \text{ if } \ f'(u_0) > 0,$$

$$[a_1, b_1] = [u_0, b_0] \ \text{ if } \ f'(u_0) < 0.$$

end continue in the same way.

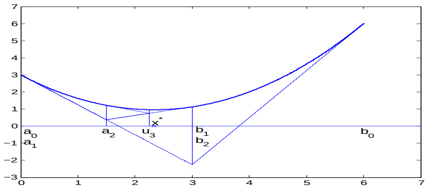


Figure 5.1: Three successive iterations of the method of tangents.

Theorem 5.10 *Let f be two times continuously differentiable, let $\inf_{x\in[a,b]} f''(x) > 0$, and let x^* be (the unique) minimizer of f. If $f'(u_k) = 0$ at some step k, then $u_k = x^*$. Otherwise for every $\varepsilon > 0$ there exists a constant $C = C_\varepsilon$ such that*

$$|u_k - x^*| \le C_\varepsilon \left(\frac{1+\varepsilon}{2}\right)^k, \quad k = 1, 2, \dots .$$

5. Interpolation methods. In the method of tangents the minimization of f on the current interval $[a_k, b_k]$ was replaced with the trivial problem of minimization of a function that consists of two linear pieces. Alternatively, one can use other classes of simple functions to approximate f on $[a_k, b_k]$ and then to minimize the approximation. One can use polynomials of degree 2 and 3, for example, or simple rational functions. The approximation is usually obtained by interpolation using values of f (or of f and f'). We shall not present the details.

5.2.2 Gradient methods

We consider the minimization problem

$$\min_{x\in\mathbf{R}^n} f(x), \tag{5.29}$$

where $f : \mathbf{R}^n \to \mathbf{R}$ is a differentiable function. We remind Lemma 2.2, which claims that if $\langle \nabla f(x), l\rangle < 0$ for some $l \in \mathbf{R}^n$, then $f(x + \sigma l) < f(x)$ for all sufficiently small $\sigma > 0$. This simple lemma provides the base of the following *Conceptual Algorithm* (CA):

Step 1. Choose some initial $x_0 \in \mathbf{R}^n$. Set $k = 0$.
Step 2. If $\nabla f(x_k) = 0$ then stop.
Step 3. Find $l_k \in \mathbf{R}^n$ such that $\langle \nabla f(x_k), l_k\rangle < 0$.
Step 4. Find σ_k such that $f(x_k + \sigma_k l_k) < f(x_k)$.
Step 5. Set $x_{k+1} = x_k + \sigma_k l_k$. Set $k := k + 1$, and go to Step 2.

Observe that l_k in Step 3 and σ_k in Step 4 always exist. Various versions of the CA differ from each other by the way the particular l_k and σ_k are chosen at each stage k of the algorithm. We mention also that in order to ensure that the algorithm will stop in finite time, we should replace the stopping criterion $\nabla f(x_k) = 0$ with $|\nabla f(x_k)| < \varepsilon$ (there are other possibilities, too). However, this modification of Step 2 is not enough for finiteness of (CA), since the vectors $\sigma_k l_k$ may become too small, so that $\sum_0^\infty \sigma_k |l_k|$ is even smaller than the distance from x_0 to any point x where $|\nabla f(x)| < \varepsilon$. Additional conditions for σ_k have to be posed.

In the next subsections we present a few particular realizations of the conceptual algorithm.

The steepest descent method

Let at a point $x \in \mathbf{R}^n$ we have $\nabla f(x) \neq 0$. Among all unit vectors $l \in \mathbf{R}^n$, the vector $-\nabla f(x)/|\nabla f(x)|$ gives the minimal value of the scalar product $\langle \nabla f(x), l\rangle$. Indeed, for every unit vector $l \in \mathbf{R}^n$

$$\left\langle \nabla f(x), -\frac{\nabla f(x)}{|\nabla f(x)|} \right\rangle = -|\nabla f(x)| = -|\nabla f(x)|\,|l| \leq -\langle \nabla f(x), l\rangle.$$

This makes it reasonable to chose

$$l_k = -\nabla f(x_k)$$

at Step 3 of (CA). That is, we choose the direction of *steepest descent* ("steilste Abstieg").

One way to realize Step 4 of (CA) is to take σ_k which minimizes f in the direction l_k. This means to find σ_k as a minimizer in the problem

$$\min_{\sigma \geq 0} \left\{ \varphi(\sigma) := f(x_k + \sigma l_k) \right\}. \tag{5.30}$$

This can be approximately done numerically by one of the methods for scalar minimization, but this may be too costly, since f has to be evaluated many times. Other methods for choosing the step size will be considered later on. Here we mention that if σ_k is chosen according to (5.30), then $\varphi'(\sigma_k) = 0$. Hence,

$$0 = \varphi'(\sigma_k) = \partial f(x_k + \sigma_k l_k)\, l_k = \langle \nabla f(x_k + \sigma_k l_k), l_k\rangle.$$

Since $x_{k+1} = x_k + \sigma_k l_k$ and $l_{k+1} = -\nabla f(x_{k+1})$, we obtain that

$$\langle l_k, l_{k+1}\rangle = 0.$$

That is, each two successive directions are orthogonal to each other. Of course, this is not true, in general, if the step size σ is chosen in a different way. On the other hand, if σ_k is determined from (5.30), then

$$\langle \nabla f(x_{k+1}), l_k\rangle = 0, \tag{5.31}$$

no matter how is l_k defined.

The case of a quadratic function f

Now we consider the particular case where f is a quadratic function:

$$f(x) = \frac{1}{2}\langle Qx, x\rangle + \langle q, x\rangle,$$

where Q is a symmetric (strictly) positive definite $(n \times n)$-matrix, and $q \in \mathbf{R}^n$. Since f is strictly convex, it has at most one minimizer, and if it exists, it is uniquely determined by the equation $f'(x) = 0$ (see Corollary 3.5).

The steepest descent direction l_k is in this case

$$l_k = -(Qx_k + q).$$

The derivative of the function $\varphi(\sigma)$ in (5.30) takes the form

$$\varphi'(\sigma) = \sigma l_k^\top Q l_k + (Qx_k + q)^\top l_k,$$

and has a single zero at

$$\sigma_k = -\frac{(Qx_k + q)^\top l_k}{l_k^\top Q l_k} = -\frac{\nabla f(x_k)^\top l_k}{l_k^\top Q l_k} = \frac{l_k^\top l_k}{l_k^\top Q l_k} = \frac{|l_k|^2}{l_k^\top Q l_k}.$$

Thus the steepest descent method takes the explicit form

$$x_{k+1} := x_k + \frac{|l_k|^2}{l_k^\top Q l_k} l_k \tag{5.32}$$

with $l_k = -(Qx_k + q) = -\nabla f(x_k)$.

Theorem 5.11 *Let x^* be the unique minimizer of f. Then there exist constants $\alpha \in (0,1)$ and C such that the sequence $\{x_k\}$ produced by (5.32) (with any initial $x_0 \in \mathbf{R}^n$) satisfies the inequality*

$$|x_k - x^*| \le C\alpha^k.$$

A proof of this theorem can be found e.g. in [1, Chapter 4].

As we know from Corollary 3.5, the minimization of a positive definite quadratic function f is equivalent to solving the equation $f'(x) = 0$. It reads as $Qx = -q$, so that the solution is $x^* = Q^{-1}q$. If the dimension n is large, solving this equation may be a problem. The steepest descent method that we presented above is, in fact, an iterative approximation scheme for solving the equation $Qx = -q$. The convergence, as stated in Theorem 5.11, is like a geometric progression. This type of convergence is known in the literature as *linear convergence.*

A convergence result

As mentioned above, one (theoretical) way to choose the step size in the conceptual algorithm in the beginning of this subsection is by scalar minimization:

$$\min_{\sigma \ge 0} \{\varphi(\sigma) := f(x_k + \sigma l_k)\}.$$

Exact solution of the above problem is not always possible, therefore we consider an approximate version, where at step k of the algorithm the size σ_k satisfies

$$\varphi(\sigma_k) \leq \inf_{\sigma \geq 0} \varphi(\sigma) + \delta_k, \tag{5.33}$$

where $\delta_k > 0$.

Theorem 5.12 *Consider the steepest descent method ($l_k = -\nabla f(x_k)$) with the choice (5.33) of σ_k. Assume that f is bounded form below, has Lipschitz continuous derivative and that $\sum_{k=1}^{\infty} \delta_k < +\infty$. Then for any $x_0 \in \mathbf{R}^n$ the sequence $\{x_k\}$ generated by the method satisfies $\lim_{k\to\infty} \nabla f(x_k) = 0$.*

Proof. First of all we remind the (easy to prove) fact that

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq \frac{L}{2}|y - x|^2 \tag{5.34}$$

for every x, $y \in \mathbf{R}^n$.

According to (5.33) we have for every $\alpha > 0$

$$f(x_{k+1}) \leq \varphi(\alpha) + \delta_k = f(x_k - \alpha\nabla f(x_k)) + \delta_k,$$

hence, using (5.34),

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq f(x_k) - f(x_k - \alpha\nabla f(x_k)) - \delta_k \\ &\geq \alpha|\nabla f(x_k)|^2 - \frac{\alpha^2 L}{2}|\nabla f(x_k)|^2 - \delta_k = \alpha\left(1 - \frac{L}{2}h\right)|\nabla f(x_k)|^2 - \delta_k. \end{aligned}$$

Since this inequality holds for every $\alpha > 0$, we have also

$$f(x_k) - f(x_{k+1}) \geq \sup_{\alpha>0} \alpha\left(1 - \frac{L}{2}h\right)|\nabla f(x_k)|^2 - \delta_k = \frac{1}{2L}|\nabla f(x_k)|^2 - \delta_k. \tag{5.35}$$

In particular, $f(x_{k+1}) \leq f(x_k)+\delta_k$ and $\sum_{k=1}^{\infty} \delta_k < +\infty$. Moreover, the sequence $\{f(x_k)\}$ is bounded from below. It is an easy exercise to prove that this implies convergence of the sequence $\{f(x_k)\}$. Then passing to the limit in (5.35) we obtain that $0 \geq \lim_{k\to+\infty} \frac{1}{2L}|\nabla f(x_k)|^2$, which completes the proof. Q.E.D.

The above choice of σ_k is also not very practical, since the residual $\varphi(\sigma_k)-\inf_{\sigma\geq 0}\varphi(\sigma)$ is difficult to estimate, and since there is also freedom in the choice of δ_k.

Some choices used in the literature are:

$$\sigma_k = -\frac{\langle \nabla f(x_k), l_k\rangle}{L|l_k|^2},$$

where L is the Lipschitz constant of $\nabla f(x)$ in the set $\{x \in \mathbf{R}^n :\ f(x) \leq f(x_k)\}$.

Gradient projection methods

Here we consider a problem with constraints:

$$\min_{x \in K} f(x),$$

where $K \subset \mathbf{R}^n$ is closed.

Denote by $\mathcal{P}_K(x) : \{y \in K : \ |x - y| = \operatorname{dist}(x, K)\}$ the projection (set) of s on K. Then for every $\sigma > 0$ we define $x_{k+1}(\sigma)$ so that it satisfies

$$x_{k+1}(\sigma) \in \mathcal{P}_K(x_k - \sigma \nabla f(x_k)).$$

Then we choose σ_k by scalar minimization:

$$\min_{\sigma > 0} f(x_{k+1}(\sigma)).$$

Here we mention, that $-\nabla f(x) \in N_K(x)$ (the normal cone to K at x) is a necessary optimality condition for x to be a minimizer (see Proposition 2.21). If it is not satisfied at x_k, that is, if $-\nabla f(x_k) \not\in N_K(x_k)$, then it is easy to prove that $f(x_{k+1}(\sigma)) < f(x_{k+1}(0)) = f(x_k)$ and the method is descending.

Another version a gradient projection method is applicable in the case of a convex K. Define $\bar{x}_k$ by solving the problem

$$\min_{x \in K} \langle \nabla f(x_k), x - x_k \rangle.$$

This is a problem of minimization of a *linear function* on K and could be much easier to solve (if, say, K is a polyhedral set). Obtain σ_k by solving the scalar problem

$$\min_{\sigma \in [0,1]} f(x_k + \sigma(\bar{x}_k - x_k))$$

and define

$$x_{k+1} = x_k + \sigma_k(\bar{x}_k - x_k).$$

Due to the convexity of K we have $x_{k+1} \in K$.

Method of conjugate gradients for quadratic problems

Above we considered the gradient method for solving the problem of optimization of quadratic functions. The method had linear convergence. Below we shall present a method for minimization of a quadratic function $f(x) = \frac{1}{2}\langle Qx, x \rangle + \langle q, x \rangle$ (with a symmetric positive definite matrix Q) that reaches the exact solution in finite number of steps, assuming that all calculations are done exactly. We shall present the idea considering in detail the first two steps. In the presentation we shall often use the obvious relation $\nabla f(x+y) = \nabla f(x) + Qy$.

Step 0. choose an arbitrary $x_0 \in \mathbf{R}^n$. If $\nabla f(x_0) = 0$ then we have reached the solution. Else define

$$l_0 := -\nabla f(x_0), \quad x_1 := x_0 + \sigma_0 l_0,$$

where σ_0 is chosen by scalar minimization:

$$0 = \frac{\mathrm{d}}{\mathrm{d}\sigma} f(x_0 + \sigma l_0) = \langle \nabla f(x_0 + \sigma l_0), l_0 \rangle = \langle \nabla f(x_0) + \sigma Q l_0, l_0 \rangle,$$

hence

$$\sigma_0 = -\frac{\langle \nabla f(x_0), l_0 \rangle}{\langle Q l_0, l_0 \rangle} = \frac{|\nabla f(x_0)|^2}{\langle Q l_0, l_0 \rangle}.$$

According to (5.31) we have

$$\langle \nabla f(x_1), l_0 \rangle = 0, \quad \text{hence also } \langle \nabla f(x_1), \nabla f(x_0) \rangle = 0. \tag{5.36}$$

An important observation is, that if we define the $(n-1)$-dimensional affine space (remember that $l_0 \neq 0$)

$$\Gamma_1 = \{x \in \mathbf{R}^n : \langle Q l_0, x - x_1 \},$$

then we have that the solution $x^* = -Q^{-1} q$ belongs to Γ_1. Indeed (see (5.36)),

$$\langle Q l_0, -Q^{-1} q - x_1 \rangle = -\langle Q l_0, Q^{-1} q \rangle - \langle Q l_0, x_1 \rangle = -\langle l_0, q + Q x_1 \rangle = -\langle l_0, \nabla f(x_1) \rangle = 0.$$

Step 1. If $\nabla f(x_1) = 0$ then we have reached the solution. Else define $l_1 := -\nabla f(x_1) + \beta_1 l_0$, where β_1 will be chosen in such a way that

$$\langle Q l_1, l_0 \rangle = 0, \tag{5.37}$$

which gives

$$\beta_1 = \frac{\langle Q \nabla f(x_1), l_0 \rangle}{\langle Q l_0, l_0 \rangle}.$$

Then, using that $l_0 = (x_1 - x_0)/\sigma_0$, the identity $Q(x_1 - x_0) = \nabla f(x_1) - \nabla f(x_0)$ and the second equality in (5.36), we obtain that

$$\beta_1 = \frac{\langle \nabla f(x_1), Q(x_1 - x_0) \rangle}{\sigma_0 \langle Q l_0, l_0 \rangle} = \frac{\langle \nabla f(x_1), \nabla f(x_1) - \nabla f(x_0) \rangle}{|\nabla f(x_0)|^2} = \frac{|\nabla f(x_1)|^2}{|\nabla f(x_0)|^2}.$$

Define $x_2 = x_1 + \sigma_1 l_1$, where σ_1 is obtain by scalar minimization:

$$\sigma_1 = -\frac{\langle \nabla f(x_1), l_1 \rangle}{\langle Q l_1, l_1 \rangle} = \frac{\langle \nabla f(x_1), \nabla f(x_1) - \beta_1 l_0 \rangle}{\langle Q l_1, l_1 \rangle} = \frac{|\nabla f(x_1)|^2}{\langle Q l_1, l_1 \rangle}.$$

Some more work is needed to verify that the following relations are satisfied:

$$\langle \nabla f(x_2), \nabla f(x_1) \rangle = \langle \nabla f(x_2), \nabla f(x_0) \rangle = 0.$$

Now consider the $(n-2)$-dimensional affine subspace

$$\Gamma_2 = \{x \in \mathbf{R}^n : \langle Q l_0, x - x_1 \rangle = \langle Q l_1, x - x_2 \rangle = 0\}.$$

The dimension of Γ_2 is $n-2$ since $\nabla f(x_1)$ is non-zero and orthogonal to l_0 (see (5.36)), thus l_0 and l_1 are linearly independent. Since

$$\langle Q l_1, x^* - x_2 \rangle = \langle Q l_1, -Q^{-1} q - x_2 \rangle = -\langle l_1, q + Q x_2 \rangle = -\langle l_1, \nabla f(x_2) \rangle = 0,$$

we obtain that $x^* \in \Gamma_2$.

Step k. If $\nabla f(x_k) = 0$ then we have reached the solution. Else calculate

$$\beta_k := \frac{|\nabla f(x_k)|^2}{|\nabla f(x_{k-1})|^2}, \qquad l_k := -\nabla f(x_k) + \beta_k l_{k-1},$$

$$\sigma_k := -\frac{\nabla f(x_k)^\top l_k}{\langle Q l_k, l_k \rangle}, \qquad x_{k+1} = x_k + \sigma_k l_k.$$

By induction one can prove as before that

$$\langle Q l_j, l_i \rangle = 0, \quad 0 \le i < j \le k, \tag{5.38}$$

$$\langle \nabla f(x_k), \nabla f(x_i) \rangle = 0, \quad \langle \nabla f(x_k), l_i \rangle = 0, \;\; 0 \le i < k.$$

From these relations one can prove as above that the affine space

$$\Gamma_{k+1} = \{x \in \mathbf{R}^n : \langle Q l_0, x - x_1 \rangle = \ldots = \langle Q l_k, x - x_{k+1} \rangle = 0\}$$

is $(n-k-1)$-dimensional and $x^* \in \Gamma_{k+1}$. The last implies that x^* will be determined at latest at step number $n-1$.

Vectors $l_0, l_1, \ldots, l_k$ that satisfy (5.38) are called *conjugate* ("konjugiert") to each other (orthogonal with respect to the scalar product $\langle Qx, y \rangle$).

In practice, when applied for solving large systems of linear equations, the method is not run $n-1$ iterations. Clearly, if we stop at iteration number k, only the components of x^* in the space Γ_{n-k-1} will remain unknown. Therefore, it is desirable to preliminary manipulate the matrix Q in such a way that the projections of x^* on the vectors $Q l_k$ become small when k is large enough. The technique of this manipulation is known as *preconditioning* ("Präkonditionierung").

5.2.3 Methods involving the Lagrange function

let us consider the probelm

$$\min f(x)$$

$$h_i(x) \le 0, \quad i = 1, \dots, r,$$

$$x \in K_0,$$

where $x \in \mathbf{R}^n$, f and h_i are continuously differentiable, $K_0 \subset \mathbf{R}^n$ is a "simple" set. As in the context of the KKT theorem we define $Y_0 := \mathbf{R}^r_+ = \{\mu \in \mathbf{R}^r : \ \mu_i \ge 0\}$ and the Lagrange function

$$L(x, \mu) = f(x) + \langle \mu, h(x) \rangle.$$

With the purpose to obtain a KKT point (hence, eventually, a solution of the problem) we consider the following iterative procedure:

$$x_{k+1} = \mathcal{P}_{K_0}(x_k - \alpha_k \partial_x L(x_k, \mu_k)^\top),$$

$$\mu_{k+1} = \mathcal{P}_{Y_0}(\mu_k + \alpha_k \partial_\mu L(x_k, \mu_k)^\top),$$

where the step length could be chosen by the same ways as in the gradient methods. Here $\mathcal{P}_{K_0}$ and $\mathcal{P}_{Y_0}$ are the projection operators on K_0 and Y_0, respectively. Since the sets are "simple" these operators are also simple. For example,

$$\mathcal{P}_{Y_0}(\mu) = \nu \quad \text{with} \ \ \nu_i = \max\{\mu_i, 0\}.$$

Convergence results are available.

5.2.4 Higher order methods

The classical Newton method. Consider the unconstraint problem

$$\min_{x \in \mathbf{R}^n} f(x),$$

with a twice continuously differentiable function f. Let x_k be already determined. Then find x_{k+1} as a solution of the problem

$$\min_{x \in \mathbf{R}^n} F(x), \tag{5.39}$$

where F is the quadratic approximation of f around x_k:

$$F(x) = f(x_k) + \partial f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^\top \partial_{xx} f(x_k)(x - x_k).$$

The necessary optimality condition for F becomes

$$0 = \partial F(x) = \partial f(x_k) + \partial_{xx} f(x_k)(x - x_k),$$

which gives

$$x_{k+1} = x_k + (\partial_{xx} f(x_k))^{-1} \nabla f(x_k),$$

provided that $\partial_{xx}(x_k)$ is invertible. This is the case if f is a strongly convex function, since in this case $\partial_{xx}(x)$ is invertible for every x. Moreover the problem has a unique solution x^* in this case.

Theorem 5.13 *Assume that f is twice differentiable and the second derivative is Lipschitz continuous. Moreover, let a number $\rho > 0$ exist such that $y^\top \partial_{xx} f(x) y \geq \rho |y|^2$ for every $y \in \mathbf{R}^n$. Then if the initial point x_0 is sufficiently close to the solution x^*, there exists $q \in (0, 1)$ such that the sequence $\{x_k\}$ generated by the Newton method satisfies the error estimation*

$$|x_k - x^*| \leq \frac{\rho}{L} q^{2^k}.$$

We mention that one does not necessarily need to invert the matrix $\partial_{xx} f(x_k)$ in order to calculate the next iteration x_{k+1}. One can (approximately) solve the quadratic problem (5.39) by the method of conjugate gradients or other method.

Convergence for which the above inequality holds is called *quadratic*. It is substantially faster than the linear convergence. For example, the linear convergence like $(1/2)^k$ produces the sequence 0.5, 0.25, 0.125, 0.0625, 0.03125, 0.01562, 0.007812, 0.003906, while the quadratic convergence $(1/2)^{2^k}$ produces the sequence 2.5×10^{-1}, 6.25×10^{-2}, 3.906×10^{-3}, 1.526×10^{-5}, 2.328×10^{-10}, 5.421×10^{-20}, 2.939×10^{-39}, ...

Sequential Quadratic Programming (SQP). The Newton method has a large number of extensions for various kinds of problems. One of the most widely used methods for numerical solving of mathematical programming problems is also a version of the Newton method known as Sequential Quadratic Programming (SQP).

Consider the mathematical programming problem

$$\min f(x) \tag{5.40}$$

$$g_i(x) = 0, \qquad i = 1, \ldots, m, \tag{5.41}$$

$$h_j(x) \leq 0, \qquad j = 1, \ldots, r, \tag{5.42}$$

where f, g_i, $h_j : \mathbf{R}^n \to \mathbf{R}$ are two times continuously differentiable functions. If x_k is a vector obtained at the k-th iteration, one approach suggests to obtain x_{k+1} by solving a problem which is obtained by replacing f, g_i and h_j by their quadratic approximations at x_k. This means to obtain x_{k+1} as a solution of the problem

$$\min F(x) \tag{5.43}$$

$$G_i(x) = 0, \qquad i = 1,\ \ldots,\ m, \tag{5.44}$$

$$H_j(x) \le 0, \qquad j = 1,\ \ldots,\ r, \tag{5.45}$$

where

$$\begin{aligned} F(x) &= f(x_k) + \partial f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^\top \partial_{xx} f(x_k)(x - x_k), \\ G_i(x) &= g_i(x_k) + \partial g_i(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^\top \partial_{xx} g_i(x_k)(x - x_k), \\ H_j(x) &= h_j(x_k) + \partial h_j(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^\top \partial_{xx} h_j(x_k)(x - x_k). \end{aligned}$$

This approach has the disadvantage that the constraint are non-linear (quadratic). Having only linear constraints is a substantial advantage although the objective functional is quadratic. There are fast methods for solving problems of minimization of a quadratic function subject to linear constraints.

For this reason, an alternative approach is to approximate the constraints by linear functions, that is, to define

$$\begin{aligned} F(x) &= f(x_k) + \partial f(x_k)(x - x_k), \\ G_i(x) &= g_i(x_k) + \partial g_i(x_k)(x - x_k), \\ H(x) &= h_j(x_k) + \partial h_j(x_k)(x - x_k). \end{aligned}$$

In this case problem (5.43)–(5.45) is efficiently solvable, but the speed of convergence drops down compared with the Newton method.

An alternative approach to use second order information about the constraints is to pass it to the objective function by using the Lagrange function. We remind that the Lagrange function, associated with problem (5.40)–(5.42) is $L(x, \lambda, \mu) := f(x) + \langle \lambda, g(x) \rangle + \langle \mu, h(x) \rangle = f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^r \mu_j h_j(x)$.

Given the k-th iterate x_k, and the vectors λ^k and μ^k the SQP method consists of the following two steps at every iteration:

0. Choose $x_0 \in \mathbf{R}^n$, $\lambda^0 \in \mathbf{R}^m$ and $\mu^0 \in \mathbf{R}^r_+$; set $k = 0$.
1. If (x_k, λ^k, μ^k) is a KKT point of problem (5.40)–(5.42), that is, if

$$\partial_x L(x_k, \lambda^k, \mu^k) = 0, \quad g(x_k) = 0, \quad h(x_k) \le 0, \quad \langle \mu^k, h(x_k) \rangle = 0, \quad \mu \ge 0,$$

then STOP.

2. Find a KKT point $(x_{k+1}, \lambda^{k+1}, \mu^{k+1})$ of the quadratic problem with linear constraints

$$\min\{f(x_k) + \partial f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^\top \partial_{xx} L(x_k, \lambda^k, \mu^k)(x - x_k)\}$$

$$g_i(x_k) + \partial g_i(x_k)(x - x_k) = 0, \qquad i = 1, \ldots, m,$$

$$h_j(x_k) + \partial h_j(x_k)(x - x_k) \leq 0, \qquad j = 1, \ldots, r.$$

Set $k := k + 1$ and go to 1.

Notice that if there are no inequality constraints, then the KKT system for the linear-quadratic problem that has to be solved at every step consists of linear equations only. If inequality constraints are present, the KKT system for the linear-quadratic problem consists of linear equations and inequalities, and in addition, the complementarity condition, which creates a trouble. Special methods have been developed for such linear complementarity problems.

We do not formulate the precise convergence theorem for the SQP method, mentioning only that the convergence is quadratic, provided that the functions f, g and h have Lipschitz continuous second derivatives, the starting point (x_0, λ^0, μ^0) is sufficiently close to a point (x^*, λ^*, μ^*), where x^* is a local minimizer in problem (5.40)–(5.42) and (λ^*, μ^*) are corresponding Lagrange multipliers, and in addition, a condition somewhat stronger than the Mangasarian-Fromovitz constraint qualification holds, and the second order sufficient optimality condition in Theorem 2.34 is fulfilled.

More details on the SQP method will be given in the lectures on Methods of Mathematical Programming for master students.

Bibliography

[1] W. Alt. *Nichtlineare Optimierung.* Vieweg+Teubner Verlag, 2011 (in German).

[2] K.H. Borgwardt. *Optimierung, Operations Research, Spieltheorie.* Birghäuser, 2001 (in German).

[3] M. Gredts and F. Lempio. *Mathematische Optimierungsverfahren des Operations research.* De Gruyter, 2011 (in German).

[4] M. Luptáčik. *Mathematical optimization and economic analysis.* Springer, 2010.

[5] J. Nocedal and S. Wright. *Numerical optimization.* Springer, 2006.

[6] R. Pler, J. Mula, and M. Diaz-Madroñero. *Operations research problems.*Springer, 2014.

[7] R.T. Rockafellar. *Fundamentals of optimization.* Available at: http://www.math.washington.edu/˜rtr/mypage.html.