Nonlinear Optimization

Prof. Dr. Martin Schmidt
Trier University

Summer term 2022

Contents

List of Algorithms

List of Figures

Glossary

$(x^k)_k$	Sequence of vectors x^k with running index k
U	Usually an open subset of $\mathbb{R}^n$
$X \subseteq \mathbb{R}^n$	Feasible set
$B_\varepsilon(x)$	Open ε-ball around x
$\text{lev}^\alpha(f, X)$	Lower level set of a function $f : X \to \mathbb{R}$ for the level $\alpha \in \mathbb{R}$
$\nabla^2 f(x) \in \mathbb{R}^{n \times n}$	Hessian of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$
$\nabla f(x) \in \mathbb{R}^n$	Gradient of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$
$\|\cdot\|$	Euclidean norm in $\mathbb{R}^n$
d	Usually a direction in $\mathbb{R}^n$
$f : \mathbb{R}^n \to \mathbb{R}$	Objective function
$g_i : \mathbb{R}^n \to \mathbb{R}$	Inequality constraints
$h_j : \mathbb{R}^n \to \mathbb{R}$	Equality constraints
x^*	A local/global solution of an optimization problem

Some Words of Caution

The lecture in its current form is given for the fourth time at Trier University. This means that—although the author tries as hard as possible and although there have been already some rounds of corrections—it is most likely that there are some mistakes in these lecture notes. Thus, whenever you are struggling with an index ("Is this really i? Shouldn't it be j?") or with a proof ("Isn't an argument missing here?"), or in any other comparable situation, please do not hesitate to ask me. Either I can explain the correct content again or there will be an improved version of these lecture notes online the next day!

These notes are mainly based on three great textbooks on nonlinear optimization:

- The second part on unconstrained optimization problems is mainly taken from Ulbrich and Ulbrich (2012).
- The third part on constrained optimization problems is mainly taken from Geiger and Kanzow (2002).
- The fourth part on algorithms is mainly taken from Nocedal and Wright (2006).

These three books are excellent sources. If you really want to study nonlinear optimization, get them and read them! Whenever a certain topic in these lecture notes are based on another source, this will be explicitly stated.

I want to thank Kevin Goergen, Ioana Molan, and Andreas Horländer for their help in creating many of the pictures in these lecture notes as well as Maximilian Lämmerer for pointing out some bugs in previous versions of these lecture notes. Finally, I thank Anna-Sophia Leuck and Madleen-Maria Popa for their help in compiling the short CVs of the mathematicians at the end of these lecture notes.

Martin Schmidt

Trier, summer term 2022

Part I

Introduction

1

Problem Statement and Basic Notations

In this lecture we consider finite-dimensional optimization problems. These problems (in their minimization version) can be stated as

$$\min_{x \in \mathbb{R}^n} \quad f(x) \tag{1.1a}$$

$$\text{s.t.} \quad x \in X \subseteq \mathbb{R}^n, \tag{1.1b}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ denotes the so-called *objective function*, $x \in \mathbb{R}^n$ are the *variables* of the optimization problem, and X denotes its *feasible set*. The latter is usually defined by a finite number of equality and inequality constraints:

$$X := \{x \in \mathbb{R}^n \colon g_i(x) \geq 0, i \in I,\, h_j(x) = 0, j \in J\}.$$

Here, I and J are finite (and maybe empty) index sets of inequality and equality constraints.

In a short form, we can thus also write

$$\min_{x \in X} \; f(x)$$

or

$$\min \{f(x) \colon x \in X\}.$$

If we do not want to consider the minimization of an objective function (like, e.g., a cost function), we can also consider maximization problems that are given as

$$\max_{x \in X} \; f(x).$$

Indeed, there is no real distinction between these two versions of an optimization problem because it is easy to see that

$$\min_{x \in X} \; f(x) = - \max_{x \in X} \; -f(x)$$

holds. In what follows, we typically consider minimization problems such as (1.1).

Optimization problems appear in many different fields of applications like in transportation, energy, telecommunication, economics, physics, and many more; see, e.g., the book by Schewe and Schmidt (2019). We will later also see and discuss some specific examples for optimization problems.

In optimization, we are interested in characterizing and computing minimizers of the objective function. Let us formalize this notion with some definitions.

Definition 1.1 (Feasible point)**.** A vector (or point) $x \in \mathbb{R}^n$ is called *feasible* for Problem (1.1) if $x \in X$ holds.

Definition 1.2 (Local minimizer)**.** A point $x^* \in \mathbb{R}^n$ is called a *local minimizer* of Problem (1.1) if x^* is feasible and if an $\varepsilon > 0$ exists such that $f(x) \geq f(x^*)$ for all $x \in X \cap B_\varepsilon(x^*)$.

Here and in what follows we denote by

$$B_\varepsilon(x^*) := \{x \in \mathbb{R}^n \colon \|x - x^*\| < \varepsilon\}$$

the open ε-ball at x^* and $\|x\| = \sqrt{x^\top x}$ denotes the Euclidean norm in $\mathbb{R}^n$. See also Figure 1.1 for an illustration of the definition of local minima.

Definition 1.3 (Strict local minimizer)**.** A point $x^* \in \mathbb{R}^n$ is called a *strict local minimizer* of Problem (1.1) if x^* is feasible and if an $\varepsilon > 0$ exists such that $f(x) > f(x^*)$ for all $x \in (X \cap B_\varepsilon(x^*)) \setminus \{x^*\}$.

Besides local minimizers we will also consider global minimizers.

Definition 1.4 ((Strict) global minimizers)**.** A point $x^* \in \mathbb{R}^n$ is called a *global minimizer* of Problem (1.1) if x^* is feasible and if $f(x) \geq f(x^*)$ holds for all $x \in X$. The point is called a *strict global minimizer* if $f(x) > f(x^*)$ holds for all $x \in X \setminus \{x^*\}$.

Example 1.5. Consider the function $f : \mathbb{R} \to \mathbb{R}$ given in Figure 1.2. The point x^1 is a local but not a global maximizer, whereas x^3 is a local and a

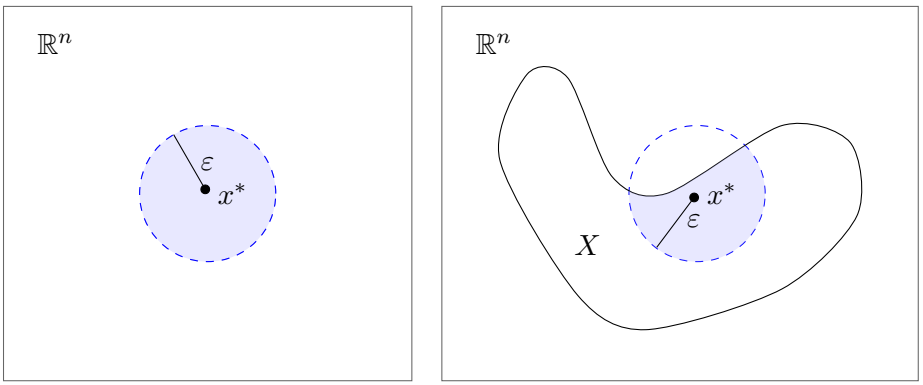


Figure 1.1: Illustration of the definition of local minima. For an unconstrained problem (left) you have to compare $f(x^*)$ with all other objective function values $f(x)$ with $x \in B_\varepsilon(x^*)$. In the constrained case, this ball first needs to be intersected with the feasible set X before the objective function values of the remaining points are compared with $f(x^*)$ (right).

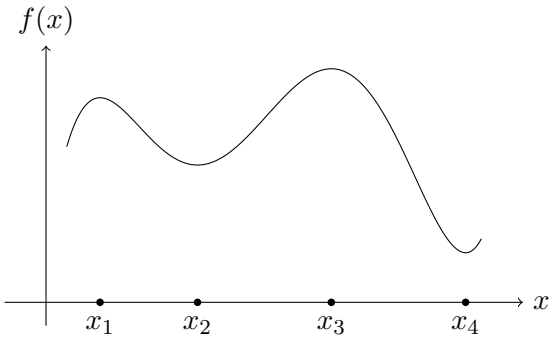


Figure 1.2: Local vs. global minimizers and maximizers

global maximizer. The point x^2 is a local but not a global minimizer, whereas x^4 is a local and a global minimizer.

Throughout the lecture we will use both the gradient as well as the Hessian matrix of a sufficiently smooth function $f : \mathbb{R}^n \to \mathbb{R}$. We define the *gradient* of f as the column vector

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right)^\top \in \mathbb{R}^n.$$

The *Hessian* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as the matrix of the second partial derivatives, i.e.,

$$\nabla^2 f(x) := \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \ldots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \ldots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Note that the Hessian is symmetric (by the Theorem of Schwarz) if the function f is twice continuously differentiable.

With the gradient at hand we can already define so-called stationary points.

Definition 1.6 (Stationary point)**.** Let $f : X \subseteq \mathbb{R}^n \to \mathbb{R}$ be differentiable in an open neighborhood of $x^* \in X$. Then, x^* is called a *stationary point* of f if $\nabla f(x^*) = 0$ holds.

Later on we will discuss the relationship between stationary points and minimizers or maximizers of a function.

2

Classification of Optimization Problems

Optimization problems are typically characterized by their key properties and these properties also determine the types of algorithms that can be used for solving these problems.

Unconstrained vs. constrained optimization If the feasible set X of Problem (1.1) is $\mathbb{R}^n$, then we call the optimization problem an *unconstrained optimization problem*; see, e.g., Geiger and Kanzow (2013). This type of problems will be discussed in the second part of this lecture. If $X \subset \mathbb{R}^n$ holds, i.e., if the variable vector needs to satisfy some constraints in order to be feasible, we call the problem a *constrained optimization problem*. These problems will be discussed in the third part of the lecture. Some good textbooks on the latter topic are, e.g., Ulbrich and Ulbrich (2012), Nocedal and Wright (2006), Bertsekas (2016), and Geiger and Kanzow (2002).

Continuous vs. (mixed-)integer optimization In many applications it is required to restrict certain or all variables to be integer valued (or even to be binary, i.e., in $\{0, 1\}$). In these cases, the problem is called a *(mixed-)integer optimization problem*; see, e.g., Schrijver (1998). If no variable needs to satisfy an integrality condition, the problem is called a *continuous optimization problem*. In this lecture, we will mainly consider continuous optimization problems.

Nonlinear vs. linear optimization If the objective function f is linear and if the feasible set X is polyhedral (i.e., it can be described by linear equality and inequality constraints), then the problem is called a *linear optimization problem*; see, e.g., Chvátal (1983). If the objective or at

least one of the constraints used to describe X is nonlinear, the problem is called a *nonlinear optimization problem*, which is, e.g., discussed in Ulbrich and Ulbrich (2012), Nocedal and Wright (2006), Bertsekas (2016), and Geiger and Kanzow (2002).

Smooth vs. nonsmooth optimization For nonlinear optimization problems, we additionally distinguish between the case that the objective function and all constraint functions are smooth (i.e., at least in $\mathcal{C}^2$) and the case where this is not true. The latter case is the topic of *nonsmooth optimization*; see, e.g., Lemarechal and Mifflin (1978). In this lecture we will only consider smooth optimization problems.

Finite- vs. infinite-dimensional optimization In practice it might also be required to consider an infinite number of constraints or variables. These are instances of *infinite optimization problems*, which are discussed, e.g., in Liberzon (2011). However, in this lecture we only consider finite-dimensional problems.

Global vs. local optimization We will learn that if an optimization problem is convex, it is as hard to compute global optima as it is to compute local ones. For nonconvex optimization problems, finding global optima is much harder. Finding global optima of nonconvex problems is dealt with in the field of *global optimization* (see, e.g., Stein (2018) and Horst and Tuy (2013)), whereas "only" finding local optima is done in the field of *local optimization*. In this lecture, we are only interested in finding local optima.

Deterministic optimization vs. optimization under uncertainties Usually, one assumes that the data, i.e., the parameters of the optimization problem to be solved are given and not subject to uncertainties. This is then called *deterministic optimization*. However, in practice, many optimization problems are subject to uncertainty since several parameters of the problem such as objective function coefficients or right-hand sides of inequality constraints are not exactly known. Dealing with these kinds of problems is done in the field of *optimization under uncertainties*; see, e.g., Birge and Louveaux (2011) and Ben-Tal et al. (2009).

3

Solvability

Before we turn to the discussion of optimality conditions and algorithms, we need to clarify what we understand to be a "solvable optimization problem". This is easy to clarify: We say that an optimization problem is *solvable* if there exists at least one (locally/globally) optimal point. However, to check whether this is the case or not is not always easy to do. In this section, we discuss which reasons for unsolvability exist as well as state and prove theorems on solvability.

Without any additional assumptions we can assign a *generalized minimum value* to every optimization problem

$$\min_{x \in \mathbb{R}^n} \; f(x) \quad \text{s.t.} \quad x \in X \subseteq \mathbb{R}^n$$

and this generalized value is the infimum of f on X. To define this formally, let us first define a value $\alpha \in \mathbb{R}$ to be a *lower bound* for f on X if it satisfies

$$\alpha \leq f(x) \quad \text{for all} \quad x \in X.$$

The *infimum* of f on X then is the largest lower bound for f on X. This means,

$$v = \inf_{x \in X} \; f(x)$$

if

(a) $v \leq f(x)$ for all $x \in X$, i.e., v itself is a lower bound for f on X, and

(b) $\alpha \leq v$ for all other lower bounds α for f on X.

If f is not bounded below on X, we set

$$\inf_{x \in X} \; f(x) = -\infty$$

and we define the infimum over the empty set as

$$\inf_{x \in \emptyset} \; f(x) = +\infty.$$

In the latter case, the specific choice of f does not play any role. As specified above, the infimum (i.e., the generalized minimum value) is defined as an element of the extended real numbers $\mathbb{R} \cup \{-\infty, +\infty\}$. In lectures on analysis it is shown that the infimum exists and that it is unique (without any assumptions on f and X); see, e.g., Heuser (1990) or Abbott (2015).

We can now define *solvability* for an optimization problem more formally.

Definition 3.1 (Solvable optimization problem)**.** We call the minimization problem (1.1) *solvable* if there exists an $x^* \in X$ with

$$f(x^*) = \inf_{x \in X} \; f(x).$$

This means a problem is solvable if its infimum is attained on X. It is not hard to see that the following theorem is correct.

Theorem 3.2. The minimization problem (1.1) is solvable if and only if it has a global minimizer.

Proof. First, let Problem (1.1) be solvable. This means there exists an $x^* \in X$ with

$$f(x^*) = \min_{x \in X} f(x).$$

Since $f(x^*)$ is also an infimum, it is a lower bound of f on X. Thus, it holds $f(x^*) \leq f(x)$ for all $x \in X$, which implies, by definition, that x^* is a global minimizer.

On the other hand, let x^* be a global minimizer of f on X. Thus, $x^* \in X$ holds and $f(x^*)$ is a lower bound of f on X. Assume now that there exists a larger lower bound α of f on X. This implies

$$f(x^*) < \alpha \leq f(x) \quad \text{for all} \quad x \in X,$$

which immediately leads to a contradiction for $x = x^*$. Thus, x^* is the largest

lower bound of f on X and we obtain

$$f(x^*) = \inf_{x \in X} f(x). \qquad \square$$

In Stein (2018) it is also shown that there are exactly three reasons for a problem being unsolvable:

(a) It holds

$$\inf_{x \in X} f(x) = +\infty.$$

This case corresponds to $X = \emptyset$. This means that the optimization problem (1.1) does not have any feasible point, i.e., it is *infeasible*. Although this case seems to be trivial at a first glance, it is often hard to decide infeasibility for real-world and, in particular, large optimization problems in practice.[1]

(b) It holds

$$\inf_{x \in X} f(x) = -\infty.$$

If the objective function f is continuous, this implies that X needs to be unbounded. However, just because the feasible set is unbounded does not mean that $\inf_{x \in X} f(x) = -\infty$ needs to hold.

(c) The last possibility is that a finite infimum is not attained on the feasible set X. One reason might again be an unbounded feasible set X. For instance, the minimization problem with the continuous objective function $f(x) = e^x$ is not solvable on $X = \mathbb{R}$ because the infimum 0 is not attained. On the other hand, an open feasible set might also render an optimization problem unsolvable. Consider, for instance, the problem

$$\min_x \; f(x) = x \quad \text{s.t.} \quad x \in X = (0, 1]$$

sketched in Figure 3.1. Obviously, the infimum is 0 but it is not attained on X. A remedy for this issue is to assume that the feasible set is closed. If, as usual, the feasible set X is given by finitely many inequality and equality constraints, i.e.,

$$X = \{x \in \mathbb{R}^n \colon g_i(x) \geq 0, i \in I, \, h_j(x) = 0, j \in J\}$$

with finite index sets I, J, closedness of the set X is implied by the continuity of the constraint functions g_i, $i \in I$, and h_j, $j \in J$.

Finally, it is also possible that a minimization problem is not solvable

[1] If we say that an optimization problem is *large* we mean that it has a large number of variables and/or a large number of constraints.

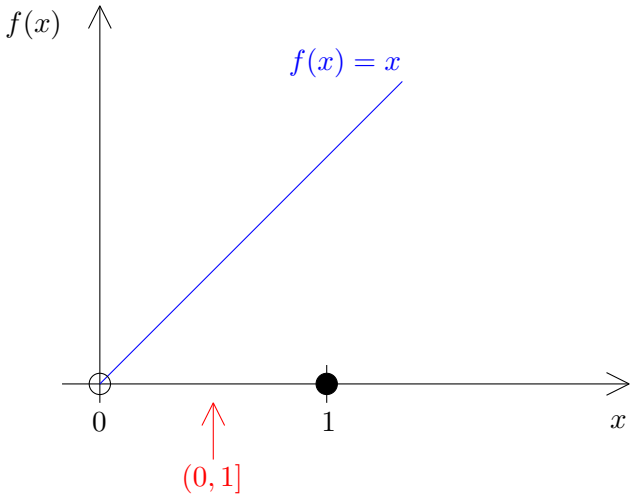


Figure 3.1: An open feasible set

although the feasible set is closed (even if it is compact[2]). This might be the case if the objective function f is not continuous, i.e., if it has jumps. For instance, the objective function

$$f(x) = \begin{cases} 1, & x \leq 0, \\ x, & x > 0, \end{cases}$$

has no global minimizer on the non-empty and compact feasible set $X = [-1, 1]$; see Figure 3.2.

This discussion motivates the classic existence theorem of Weierstraß.

Theorem 3.3 (Theorem of Weierstraß)**.** Suppose that the set X is non-empty and compact and that the function $f : X \to \mathbb{R}$ is continuous. Then, f has at least one global minimizer and at least one global maximizer.

Proof. Let $v = \inf_{x \in X} f(x)$. Since $X \neq \emptyset$ we know that $v < +\infty$ holds. We need to show that there exists an $x^* \in X$ with $v = f(x^*)$. Since v is an infimum, there exists a sequence $(x^k)_k \subseteq X$ with $\lim_{k\to\infty} f(x^k) = v$. Moreover, as X is compact, the theorem of Bolzano–Weierstraß ensures the existence of a sub-sequence that converges in X. Let this sub-sequence be denoted by $(x^{k_i})_i$. Thus, there exists an $\bar{x} \in X$ with $\lim_{i\to\infty} x^{k_i} = \bar{x}$. Because f is continuous on X we obtain

$$f(\bar{x}) = f\left(\lim_{i\to\infty} x^{k_i}\right) = \lim_{i\to\infty} f(x^{k_i}) = v.$$

[2] A set in $\mathbb{R}^n$ is called *compact* if it is bounded and closed.

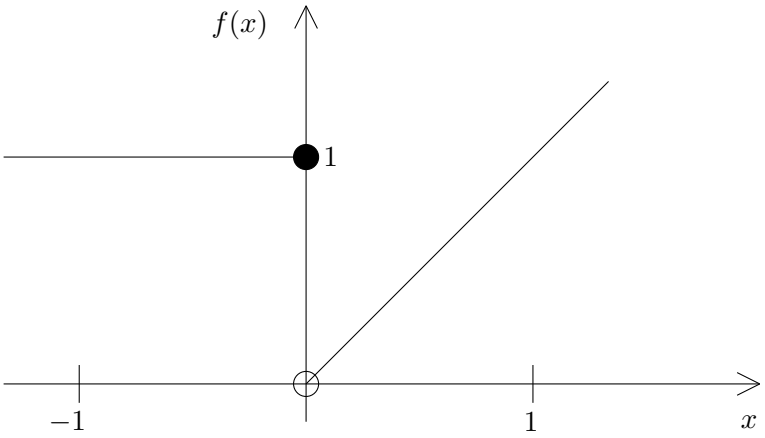


Figure 3.2: A discontinuous objective function

Thus, the proof is completed by choosing $x^* = \bar{x}$. □

The assumptions of the theorem of Weierstraß can be weakened. To this end, we first need to define so-called *lower level sets*.

Definition 3.4 (Lower level set)**.** Let $X \subseteq \mathbb{R}^n$, $f : X \to \mathbb{R}$, and $\alpha \in \mathbb{R} \cup \{-\infty, +\infty\}$ be given. The set

$$\operatorname{lev}^{\alpha}(f, X) := \{x \in X \colon f(x) \leq \alpha\}$$

is called the *lower level set* of f on X w.r.t. the level α.

See Figure 3.3 for an exemplary illustration of the level sets of a quadratic function.

Moreover, we need the set of global minimizers

$$S := \{x^* \in X \colon f(x) \geq f(x^*) \text{ for all } x \in X\}.$$

Obviously, solvability of Problem (1.1) is equivalent to $S \neq \emptyset$.

Lemma 3.5. Let $v = \inf\{f(x) \colon x \in X\}$. Then, $S = \operatorname{lev}^{v}(f, X)$ holds.

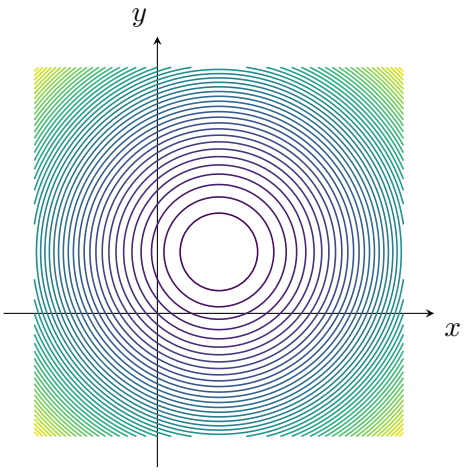


Figure 3.3: Lower level sets of the function $f(x_1, x_2) = (x_1 - 1)^2 + (x_2 - 1)^2$

Proof. It holds

$$\begin{aligned} x^* \in S &\iff x^* \text{ is a global minimizer} \\ &\iff x^* \in X \text{ and } f(x^*) = v \\ &\iff x^* \in X \text{ and } f(x^*) \leq v \text{ (since } \{x \in X \colon f(x) < v\} = \emptyset) \\ &\iff x^* \in \operatorname{lev}^v(f, X). \end{aligned}$$

□

Lemma 3.6. Let $\alpha \in \mathbb{R}$ be given such that $\operatorname{lev}^\alpha(f, X) \neq \emptyset$. Then, $S \subseteq \operatorname{lev}^\alpha(f, X)$ holds.

Proof. The assumption $\operatorname{lev}^\alpha(f, X) \neq \emptyset$ implies the existence of a point $\tilde{x} \in X$ with $f(\tilde{x}) \leq \alpha$. Let now x^* be an arbitrary global minimizer. Then, $x^* \in X$ and $f(x^*) \leq f(\tilde{x}) \leq \alpha$ holds. Thus, $x^* \in \operatorname{lev}^\alpha(f, X)$. □

With these results we can now state and prove a stronger version of the theorem of Weierstraß.

Theorem 3.7. Let a set $X \subseteq \mathbb{R}^n$ be given and suppose that the function $f : X \to \mathbb{R}$ is continuous. Moreover, assume that there exists a value $\alpha \in \mathbb{R}$ with $\operatorname{lev}^\alpha(f, X)$ being non-empty and compact. Then, S is also non-empty and compact.

Proof. Lemma 3.6 implies that Problem (1.1) and the problem

$$\min_x \quad f(x) \quad \text{s.t.} \quad x \in \operatorname{lev}^\alpha(f, X) \tag{3.1}$$

have the same optimal solutions and the same optimal objective function value v. Moreover, Problem (3.1) satisfies the assumptions of Theorem 3.3 (Weierstraß), which implies $S \neq \emptyset$. We also have

$$S = \text{lev}^v(f, X) \subseteq \text{lev}^\alpha(f, X),$$

which implies the boundedness of S. It remains to prove that S is closed. To this end, we consider the converging sequence $(x^k)_k \subseteq S$ with limit x^*. Since S is contained in the closed set $\text{lev}^\alpha(f, X)$, we obtain $x^* \in \text{lev}^\alpha(f, X) \subseteq X$. Moreover, the continuity of f on X and $f(x^k) \leq v$ for all $k \in \mathbb{N}$ implies $f(x^*) \leq v$. Thus, $x^* \in \text{lev}^v(f, X) = S$ and S is compact. □

This theorem also holds for unconstrained optimization problems, i.e., the case $X = \mathbb{R}^n$.

Corollary 3.8. Suppose that the function $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and that $\text{lev}^\alpha(f, \mathbb{R}^n)$ is non-empty and compact. Then, S is non-empty and compact as well.

Proof. Simply apply Theorem 3.7 with $X = \mathbb{R}^n$. □

For the moment, this should be enough regarding the existence of optimizers, i.e., regarding the solvability of optimization problems. Even more existence theorems can be found, e.g., in Stein (2018) or in Ulbrich and Ulbrich (2012).

Part II

Unconstrained Optimization Problems

4

Problem Statement and a First Example

In this part of the lecture we study unconstrained optimization problems of the form

$$\min_{x \in \mathbb{R}^n} \; f(x),$$

i.e., we have $X = \mathbb{R}^n$.

Before we establish optimality conditions and derive algorithms, we first consider the probably most prominent example of an unconstrained optimization problem: the least-squares problem.

To this end, we assume a physical, technical, or economical process that delivers a system's response $y \in \mathbb{R}^s$ for a given input $u \in \mathbb{R}^r$; see Figure 4.1. The goal now is to approximate the system's behavior by a parameterized approach

$$u \mapsto g(u; x)$$

with parameters $x \in \mathbb{R}^n$. For this, we use measurements y^i obtained for inputs u^i, $i = 1, \dots, N$, and try to find parameters x so that $g(u^i; x)$ fits as good as possible to the measurements y^i. Here, "fits good" can be measured using a norm of $\mathbb{R}^s$ and we choose the Euclidean norm. This problem can be

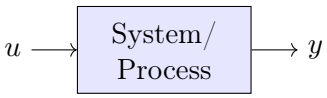


Figure 4.1: Input-output system

modeled as the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{N} \|y^i - g(u^i; x)\|_2^2.$$

This problem is called the *least-squares* or *nonlinear regression problem*. If g is chosen to be a linear ansatz function this problem leads to a *linear regression problem*, which is often already discussed in introductory lectures on linear algebra.

5

Optimality Conditions

We start with a first-order necessary condition for a local minimizer.

Theorem 5.1 (Necessary first-order optimality condition)**.** Let $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$ be differentiable on the open set $U \subseteq \mathbb{R}^n$ and let x^* be a local minimizer of f. Then, $\nabla f(x^*) = 0$ holds, i.e., x^* is a stationary point.

Proof. Consider the difference quotient

$$\frac{f(x^* + td) - f(x^*)}{t}.$$

For an arbitrary $d \in \mathbb{R}^n$ and a sufficiently small $t \in \mathbb{R}$, it holds that the above difference quotient is non-negative. (Otherwise, x^* would not be a local minimizer.) Taking $t \searrow 0$ yields the directional derivative $\nabla f(x^*)^\top d \geq 0$ and the choice $d = -\nabla f(x^*)$ implies $\nabla f(x^*) = 0$. □

The above optimality condition is called a "first-order" optimality condition since it uses a condition on the first derivative of the function.

Note that the stationarity condition of Theorem 5.1 is necessary but not sufficient. For instance, it holds $\nabla(-f) = -\nabla f$ and thus every stationary point of f is a stationary point of $-f$ as well. Hence, this stationarity condition cannot distinguish between minimizers and maximizers. Moreover, a stationary point does not even need to be a minimizer or a maximizer.

Definition 5.2 (Saddle point)**.** A stationary point x^* of f, which is neither a local minimizer nor a local maximizer is called a *saddle point* of f.

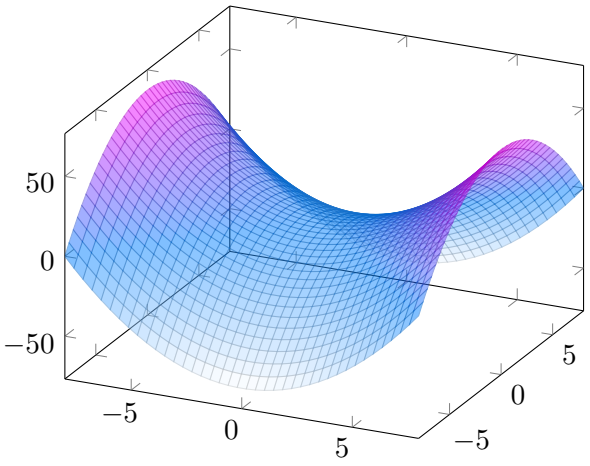


Figure 5.1: $f(x) = x_1^2 - x_2^2$

Example 5.3. Consider the function $f(x) = x_1^2 - x_2^2$; see Figure 5.1. The gradient of f is $\nabla f(x) = (2x_1, -2x_2)^\top \in \mathbb{R}^2$. Thus, the point $x^* = 0 \in \mathbb{R}^2$ is the only stationary point of f. However, since f has a positive curvature in x_1-direction and a negative curvature in x_2-direction, x^* is a saddle point.

For being able to distinguish between local minimizers, local maximizers, and saddle points we need to take the curvature of f (i.e., the second-order information about f) into account as well. For this, we also need the small-o notation.

Definition 5.4 (Small-o notation)**.** Given two non-negative infinite sequences of scalars $(\eta_k)_k$ and $(\nu_k)_k$, we write

$$\eta_k = o(\nu_k)$$

if

$$\lim_{k\to\infty} \frac{\eta_k}{\nu_k} = 0$$

holds. This notation can also be defined for functions. Let $\eta : \mathbb{R} \to \mathbb{R}$ be a function. We say that

$$\eta(\nu) = o(\nu)$$

to indicate that $\eta(\nu)/\nu$ approaches zero either as $\nu \to 0$ or $\nu \to \infty$.[1]

Theorem 5.5 (Necessary second-order optimality condition)**.** Let $f : U \subseteq$

[1] The specific limit should always be clear from the context.

$\mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable on the open set $U \subseteq \mathbb{R}^n$ and let $x^* \in U$ be a local minimum of f. Then, it holds

(a) $\nabla f(x^*) = 0$, i.e., x^* is a stationary point of f, and

(b) the Hessian matrix $\nabla^2 f(x^*)$ is positive semi-definite, i.e.,

$$d^\top \nabla^2 f(x^*) d \geq 0 \quad \text{for all} \quad d \in \mathbb{R}^n.$$

Proof. Condition (a) directly follows from Theorem 5.1. Thus, we need to prove the second condition. To this end, let $d \in \mathbb{R}^n \setminus \{0\}$ be arbitrary.[2] Let now $\tau = \tau(d) > 0$ be sufficiently small so that the Taylor expansion[3] of f leads to

$$0 \leq f(x^* + td) - f(x^*) = t\nabla f(x^*)^\top d + \frac{t^2}{2} d^\top \nabla^2 f(x^*) d + \rho(t)$$

for all $t \in (0, \tau]$ and with $\rho(t) = o(t^2)$. Here, we used that x^* is a local minimizer for the first inequality. Due to Condition (a), this implies

$$d^\top \nabla^2 f(x^*) d \geq -\frac{2\rho(t)}{t^2}.$$

The right-hand side tends to 0 for $t \to 0$, which completes the proof. □

Example 5.6. The conditions of the last theorem are necessary but not sufficient. Consider, e.g., the saddle point $x^* = 0$ of the function $f(x) = x^3$.

This motivates the following theorem.

Theorem 5.7 (Sufficient second-order optimality condition)**.** Let $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable on the open set $U \subseteq \mathbb{R}^n$ and let $x^* \in U$ be a point that satisfies

(a) $\nabla f(x^*) = 0$, i.e., x^* is a stationary point of f, and

(b) the Hessian matrix $\nabla^2 f(x^*)$ is positive definite, i.e.,

$$d^\top \nabla^2 f(x^*) d > 0 \quad \text{for all} \quad d \in \mathbb{R}^n \setminus \{0\}$$

holds.

Then, the point x^* is a strict local minimizer of f.

[2] Note that for $d = 0$ the condition is trivially satisfied.

[3] See, e.g., Nocedal and Wright (2006) or Heuser (1990) and Heuser (1993) for Taylor's theorem.

Proof. Suppose that both conditions hold. Then, there exists a constant $\mu > 0$ with

$$d^\top \nabla^2 f(x^*) d \geq \mu \|d\|^2 \quad \text{for all} \quad d \in \mathbb{R}^n.$$

Using the Taylor expansion again as well as the stationarity of x^*, we can find an $\varepsilon > 0$ such that

$$f(x^* + d) - f(x^*) = \frac{1}{2} d^\top \nabla^2 f(x^*) d + o(\|d\|^2) \geq \frac{\mu}{2} \|d\|^2 + o(\|d\|^2)$$

holds for all $d \in B_\varepsilon(0)$. Thus, x^* is a strict local minimizer. □

Example 5.8. The global minimizer $x^* = 0$ of the function $f(x) = x^4$ shows that the conditions of the last theorem are sufficient but not necessary.

6

Convexity

We now consider the class of functions that have the property that every local minimizer also is a global minimizer. It will turn out that this is the case for convex functions.

Definition 6.1 (Convex set)**.** A set $X \subseteq \mathbb{R}^n$ is called *convex* if for all $x, y \in X$ and for all $\lambda \in [0, 1]$ it holds

$$(1 - \lambda)x + \lambda y \in X.$$

In words: If two points are elements of a convex set, then the entire connecting line is in the set as well; cf. Figure 6.1 (left).

Example 6.2. See Figure 6.1 for a convex set (top) and a nonconvex set (bottom).

Definition 6.3 (Convex function)**.** Let $f : X \to \mathbb{R}$ be a function defined on a convex set $X \subseteq \mathbb{R}^n$. Then, f is called

(a) *convex* if for all $x, y \in X$ and all $\lambda \in [0, 1]$ it holds

$$f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y);$$

(b) *strictly convex* if for all $x, y \in X$ with $x \neq y$ and all $\lambda \in (0, 1)$ it holds

$$f((1 - \lambda)x + \lambda y) < (1 - \lambda)f(x) + \lambda f(y);$$

(c) *uniformly convex* if there exists a constant $\mu > 0$ so that for all $x, y \in X$

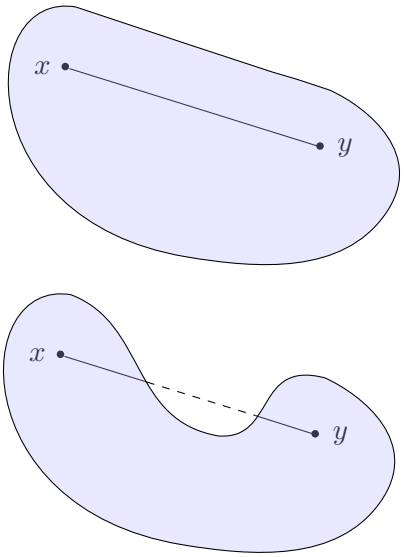


Figure 6.1: A convex (top) and a nonconvex set (bottom)

and all $\lambda \in [0,1]$ it holds

$$f((1-\lambda)x + \lambda y) + \mu\lambda(1-\lambda)\|y-x\|^2 \leq (1-\lambda)f(x) + \lambda f(y).$$

Figure 6.2 shows a convex (but not strictly convex), a strictly convex, and a nonconvex function.

Remark 6.4. Concave functions, strictly concave functions, and uniformly concave functions are defined in analogy to Definition 6.3 but with "$\geq$" in (a) and (c) as well as "$>$" in (b) instead of "$\leq$" and "$<$", respectively.

For the case of f being differentiable we can also give another characterization of convexity.

Theorem 6.5. Let $f : X \to \mathbb{R}$ be continuously differentiable on an open and convex set $X \subseteq \mathbb{R}^n$. Then, the following holds:

(a) The function f is convex if and only if for all $x, y \in X$ it holds

$$\nabla f(x)^\top (y-x) \leq f(y) - f(x). \tag{6.1}$$

(b) The function f is strictly convex if and only if for all $x, y \in X$ with $x \neq y$ it holds

$$\nabla f(x)^\top (y-x) < f(y) - f(x).$$

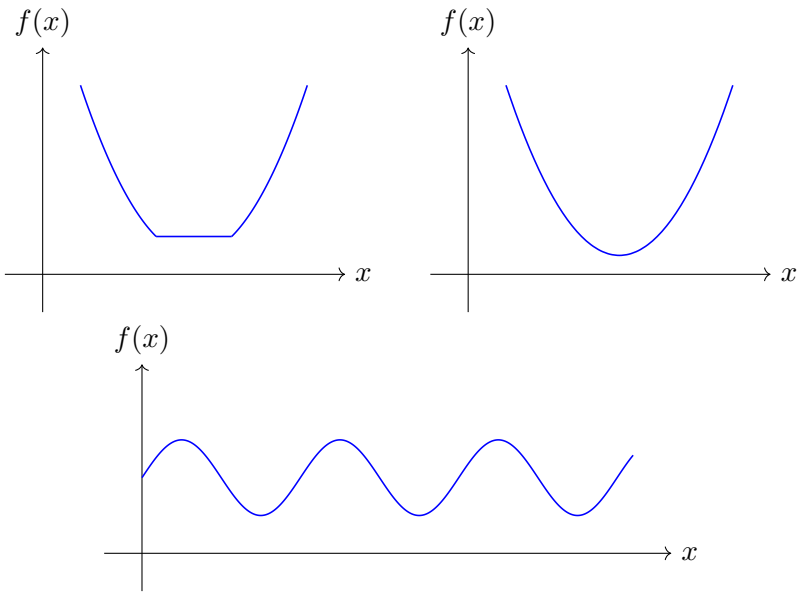


Figure 6.2: A convex (but not strictly convex; top left), a strictly convex (top right), and a nonconvex function (bottom)

(c) The function f is uniformly convex if and only if there exists a constant $\mu > 0$ so that for all $x, y \in X$ it holds

$$\nabla f(x)^\top (y - x) + \mu \|y - x\|^2 \leq f(y) - f(x).$$

Proof. (a) "⇒": Let f be convex. Then, for all $x, y \in X$ and all $0 < \lambda \leq 1$ it holds

$$\frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \leq \frac{(1 - \lambda) f(x) + \lambda f(y) - f(x)}{\lambda} = f(y) - f(x).$$

Taking the limit $\lambda \searrow 0$ then yields inequality (6.1) since

$$\nabla f(x)^\top (y - x) = \lim_{\lambda \searrow 0} \frac{f(x + \lambda(y - x)) - f(x)}{\lambda}.$$

"⇐": Assume that (6.1) holds. For arbitrary $x, y \in X$ and $0 \leq \lambda \leq 1$ we define

$$x_\lambda := (1 - \lambda)x + \lambda y.$$

We need to show that

$$(1 - \lambda) f(x) + \lambda f(y) - f(x_\lambda) \geq 0$$

holds. To this end, we use (6.1) and compute

$$\begin{aligned} &(1-\lambda)f(x)+\lambda f(y)-f(x_\lambda)\\ &=(1-\lambda)(f(x)-f(x_\lambda))+\lambda(f(y)-f(x_\lambda))\\ &\geq(1-\lambda)\nabla f(x_\lambda)^\top(x-x_\lambda)+\lambda\nabla f(x_\lambda)^\top(y-x_\lambda)\\ &=\nabla f(x_\lambda)^\top((1-\lambda)x+\lambda y-x_\lambda)=0. \end{aligned} \tag{6.2}$$

(b) "⇒": Let f be strictly convex. For $x, y \in X$ with $x \neq y$ and $z = (x+y)/2$ we then have

$$f(z) < \frac{1}{2}(f(x)+f(y)),$$

i.e.,

$$f(z)-f(x) < \frac{1}{2}(f(y)-f(x)).$$

Moreover, from Condition (a) and the definition of z it follows

$$f(z)-f(x) \geq \nabla f(x)^\top(z-x) = \frac{1}{2}\nabla f(x)^\top(y-x).$$

Together, we obtain

$$\nabla f(x)^\top(y-x) \leq 2(f(z)-f(x)) < f(y)-f(x).$$

"⇐": Follows directly by using ">" in (6.2).

(c) "⇒": Like in (a) we consider the difference quotient

$$\begin{aligned} \nabla f(x)^\top(y-x) &= \lim_{\lambda \searrow 0} \frac{f(x_\lambda)-f(x)}{\lambda}\\ &\leq \lim_{\lambda \searrow 0} \frac{(1-\lambda)f(x)+\lambda f(y)-\mu\lambda(1-\lambda)\|y-x\|^2-f(x)}{\lambda}\\ &= f(y)-f(x)-\mu\|y-x\|^2. \end{aligned}$$

"⇐": We use the identities

$$\|x-x_\lambda\| = \lambda\|y-x\|, \quad \|y-x_\lambda\| = (1-\lambda)\|y-x\|.$$

Similar to (6.2) we obtain

$$\begin{aligned}
&(1-\lambda)f(x)+\lambda f(y)-f(x_\lambda)\\
&=(1-\lambda)(f(x)-f(x_\lambda))+\lambda(f(y)-f(x_\lambda))\\
&\geq (1-\lambda)\left(\nabla f(x_\lambda)^\top(x-x_\lambda)+\mu\|x-x_\lambda\|^2\right)\\
&\quad+\lambda\left(\nabla f(x_\lambda)^\top(y-x_\lambda)+\mu\|y-x_\lambda\|^2\right)\\
&=\nabla f(x_\lambda)^\top((1-\lambda)x+\lambda y-x_\lambda)+\mu\left((1-\lambda)\|x-x_\lambda\|^2+\lambda\|y-x_\lambda\|^2\right)\\
&=\mu((1-\lambda)\lambda^2+\lambda(1-\lambda)^2)\|y-x\|^2=\mu\lambda(1-\lambda)\|y-x\|^2. \qquad \square
\end{aligned}$$

If f is even twice continuously differentiable we can also characterize convexity of f using the Hessian matrix of f.

Theorem 6.6. Let $f : X \to \mathbb{R}$ be twice continuously differentiable on the open and convex set $X \subseteq \mathbb{R}^n$. Then, it holds

(a) that the function f is convex if and only if the Hessian matrix $\nabla^2 f(x)$ is positive semi-definite for all $x \in X$, i.e., if and only if

$$d^\top \nabla^2 f(x) d \geq 0 \quad \text{for all} \quad x \in X,\ d \in \mathbb{R}^n;$$

(b) that the function f is strictly convex if the Hessian matrix $\nabla^2 f(x)$ is positive definite for all $x \in X$, i.e., if

$$d^\top \nabla^2 f(x) d > 0 \quad \text{for all} \quad x \in X,\ d \in \mathbb{R}^n \setminus \{0\};$$

(c) that the function f is uniformly convex if and only if the Hessian matrix $\nabla^2 f(x)$ is uniformly positive definite for all $x \in X$, i.e., if and only if there exists a constant $\mu > 0$ so that

$$d^\top \nabla^2 f(x) d \geq \mu \|d\|^2 \quad \text{for all} \quad x \in X,\ d \in \mathbb{R}^n$$

holds.

Proof. (a) "⇒": Let f be convex and let $x \in X$ and $d \in \mathbb{R}^n$ be arbitrary. Since X is open, there exists a constant $\tau = \tau(x, d) > 0$ with $x + td \in X$ for all $t \in [0, \tau]$. For $0 < t \leq \tau$, we obtain from Part (a) of Theorem 6.5 and the Taylor expansion that

$$0 \leq f(x+td) - f(x) - t\nabla f(x)^\top d = \frac{t^2}{2} d^\top \nabla^2 f(x) d + o(t^2)$$

holds. Multiplication with $2/t^2$ and taking the limit $t \searrow 0$ yields the claim.

"$\Leftarrow$": For arbitrary $x, y \in X$, the Taylor expansion yields the existence of a value $\sigma \in [0,1]$ with

$$\begin{aligned} & f(y) - f(x) \\ &= \nabla f(x)^\top (y-x) + \frac{1}{2}(y-x)^\top \nabla^2 f(x+\sigma(y-x))(y-x) \\ &\geq \nabla f(x)^\top (y-x). \end{aligned} \tag{6.3}$$

This yields, together with Part (a) of Theorem 6.5, the convexity of f.

(b) For $x, y \in X$ with $x \neq y$ we obtain the inequality in (6.3) with "$>$" instead of "$\geq$".

(c) "$\Rightarrow$": As in Part (a), for all $x \in X$ and $d \in \mathbb{R}^n \setminus \{0\}$ there exists a constant $\tau = \tau(x,d) > 0$ so that for all $0 < t \leq \tau$ it holds

$$\begin{aligned} 0 &\leq f(x+td) - f(x) - t\nabla f(x)^\top d - \mu \|td\|^2 \\ &= \frac{t^2}{2} d^\top \nabla f(x) d - t^2 \mu \|d\|^2 + o(t^2). \end{aligned}$$

Multiplication with $2/t^2$ and taking the limit $t \searrow 0$ yields

$$d^\top \nabla^2 f(x) d \geq 2\mu \|d\|^2.$$

"$\Leftarrow$": Using the notation from the proof of Part (a) we obtain

$$\begin{aligned} f(y) - f(x) &= \nabla f(x)^\top (y-x) + \frac{1}{2}(y-x)^\top \nabla^2 f(x+\sigma(y-x))(y-x) \\ &\geq \nabla f(x)^\top (y-x) + \frac{\mu}{2}\|y-x\|^2. \end{aligned}$$

Thus, using Part (c) of Theorem 6.5 proves that f is uniformly convex. □

Remark 6.7. The example $f(x) = x^4$ shows that the condition in Part (b) of the last theorem is not necessary for strict convexity.

Remark 6.8. If we flip the direction of the relations in the last two theorems we get the analogous characterizations for concavity instead of convexity.

The main reason for the importance of convexity in optimization is given in the following theorem. Since it does not require additional work, we directly proof this theorem for the constrained case.

Theorem 6.9. Let $f : X \to \mathbb{R}$ be a convex function defined on a convex set $X \subseteq \mathbb{R}^n$. Then, the following statements are true.

(a) Every local minimizer of f on X is also a global minimizer of f on X.

(b) If f is strictly convex, then f has at most one local minimizer on X and this local minimizer (if it exists) then also is the unique global minimizer of f on X.

(c) Let X be open, f be continuously differentiable on X, and suppose that $x^* \in X$ is a stationary point of f. Then, x^* is a global minimizer of f on X.

Proof. (a) Let x^* be a local minimizer of f on X. Assume further that there exists an $x \in X$ with $f(x) < f(x^*)$. Then, it holds

$$f(x^*+t(x-x^*)) \leq (1-t)f(x^*)+tf(x) < (1-t)f(x^*)+tf(x^*) = f(x^*)$$

for all $t \in (0, 1]$, which is a contradiction for t sufficiently close to 0 since x^* is a local minimizer of f on X.

(b) Assume that f has two different local minimizers x^* and y^* on X. Using Part (a), both are also global minimizers of f on X. Thus,

$$f(x) \geq f(x^*) = f(y^*) \quad \text{for all} \quad x \in X$$

holds. Using strict convexity of f then yields for $(x^* + y^*)/2 = x \in X$ that

$$f(x) < \frac{f(x^*) + f(y^*)}{2} = f(x^*) \leq f(x)$$

holds. This is a contradiction. Thus, the function f has at most one local minimizer on X, which then also is the unique global minimizer on X.

(c) Part (a) of Theorem 6.5 implies

$$f(x) - f(x^*) \geq \nabla f(x^*)^\top (x - x^*) = 0 \quad \text{for all} \quad x \in X.$$

Thus, x^* is a global minimizer of f on X. □

Definition 6.10 (Convex optimization problem)**.** A constrained optimization problem that satisfies the conditions of the last theorem is called a *convex optimization problem*.

7

The Gradient Method

In this section, we consider one of the most classic methods to solve unconstrained optimization problems: the gradient (or steepest descent) method.

Like for almost all optimization algorithms we try to answer the following questions for the gradient method:

(a) How does the method work? What is its main rationale? What is the intuition of the method?

(b) How can the method be formally described?

(c) What are the convergence properties of the method? Under which set of assumptions does it converge to what kind of points?

(d) What is the speed of convergence? How fast (or slow) does the method approach a solution (or a stationary point). What are the required assumptions to prove a certain rate of convergence?

The gradient method belongs to the class of descent methods. Descent methods are iterative algorithms and the idea of all descent methods is the following. Given a current iterate[1] $x^k \in \mathbb{R}^n$, determine a descent direction s^k (i.e., a direction in which the objective function has a negative slope; we will soon formalize this notion), compute (for this descent direction) a step size $0 < \sigma^k \in \mathbb{R}$, which ensures that we make sufficient progress towards a solution, and finally compute the new iterate

$$x^{k+1} = x^k + \sigma^k s^k.$$

A generic framework of such methods is given in Algorithm 1.

[1] Iteration indices, here k, are always denoted by super-indices.

Algorithm 1 A Generic Descent Method for Unconstrained Optimization

Input: An objective function $f : \mathbb{R}^n \to \mathbb{R}$ and an initial iterate $x^0 \in \mathbb{R}^n$.

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: **if** termination criterion is not satisified at x^k **then**
3: Compute a descent direction $s^k \in \mathbb{R}^n$ of f at x^k.
4: Compute a step size $\sigma^k \in \mathbb{R}$ with $\sigma^k > 0$ such that $f(x^k + \sigma^k s^k) < f(x^k)$ and that $f(x^k) - f(x^k + \sigma^k s^k)$ is sufficiently large.
5: Compute the new iterate $x^{k+1} = x^k + \sigma^k s^k$.
6: **else**
7: Stop and return x^k.
8: **end if**
9: **end for**

There are some algorithmic details that we need to specify in order to make Algorithm 1 an implementable method.

First, we need to discuss a reasonable termination (or stopping) criterion. Since the method only uses first-order information, it cannot be expected that it can reliably compute stronger points than stationary ones. Thus, a reasonable stopping criterion in Line 2 of Algorithm 1 is to check if

$$\nabla f(x^k) = 0$$

holds, which, in practice, is usually replaced by

$$\|\nabla f(x^k)\| < \varepsilon$$

with a user-specified tolerance $\varepsilon > 0$, e.g., $\varepsilon = 10^{-8}$. Thus, we proceed with the lines 3–5 if the norm of the gradient norm is not yet small enough.

Second, we need to specify what we mean by "descent direction". Informally speaking, we say that a direction $s \in \mathbb{R}^n \setminus \{0\}$ is a descent direction of f at x if the slope of f in the direction of s at x is negative. The slope of f at x in direction s is given by

$$\lim_{t \searrow 0} \frac{f(x + ts) - f(x)}{\|ts\|} = \frac{\nabla f(x)^\top s}{\|s\|}.$$

The formal definition thus reads as follows.

Definition 7.1 (Descent direction)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. The vector $s \in \mathbb{R}^n \setminus \{0\}$ is called a *descent direction* of f at x if $\nabla f(x)^\top s < 0$ holds.

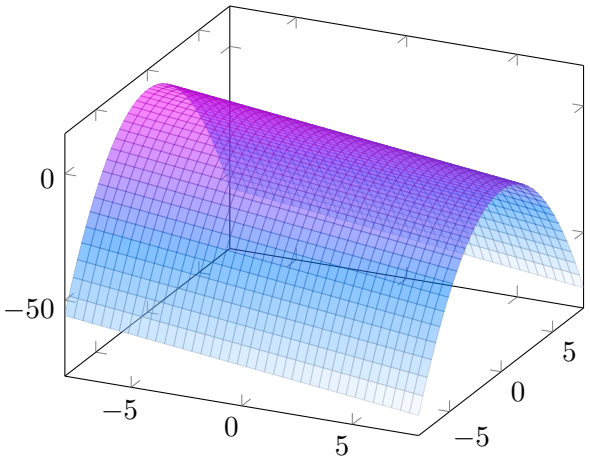


Figure 7.1: $f(x_1, x_2) = -x_1 - x_2^2$

Third, we need to discuss how step sizes σ^k can be computed that satisfy the conditions in Line 4 of Algorithm 1. This will be discussed in one of the next sections. However, before we will focus a little bit more on descent directions.

Remark 7.2. (a) Let $s \in \mathbb{R}^n$ be given and define

$$\phi : \mathbb{R} \to \mathbb{R}, \quad \phi(t) := f(x + ts).$$

It holds $\phi'(0) = \nabla f(x)^\top s$. Thus, the expression $\nabla f(x)^\top s$ exactly represents the slope of the function $\phi(t)$ at $t = 0$.

(b) For a direction s being a descent direction it is not enough that f decreases along s (or, equivalently, that $\phi(t)$ decreases at $t = 0$). The reason is that the decrease may be completely due to a negative curvature in this direction. Consider, for example, the function

$$f : \mathbb{R}^2 \to \mathbb{R}, \quad f(x) = f(x_1, x_2) = -x_1 - x_2^2,$$

which is also shown in Figure 7.1. The gradient and the Hessian matrix are given by

$$\nabla f(x) = \begin{pmatrix} -1 \\ -2x_2 \end{pmatrix}, \quad \nabla^2 f(x) = \begin{bmatrix} 0 & 0 \\ 0 & -2 \end{bmatrix}.$$

For $x = 0 \in \mathbb{R}^2$ and $s = (0, 1)^\top \in \mathbb{R}^2$, we obtain

$$\phi'(0) = \nabla f(0)^\top s = \begin{pmatrix} -1 \\ 0 \end{pmatrix}^\top \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 0.$$

This means that s is *not* a descent direction. On the other hand, f decreases along s since for all $t > 0$, it holds

$$\phi(t) = f(0 + ts) = -t^2 < 0 = \phi(0).$$

All directions $d = (d_1, d_2)^\top \in \mathbb{R}^2$ with $d_1 > 0$ are descent directions of f at $x = 0$ because

$$\nabla f(0)^\top d = \begin{pmatrix} -1 \\ 0 \end{pmatrix}^\top \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = -d_1 < 0$$

holds.

7.1 Directions of Steepest Descent

The question is still open on how to compute a specific descent direction at a given iterate. The most obvious way is to use the directions of steepest descent.

Definition 7.3 (Steepest descent directions)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function and let $x \in \mathbb{R}^n$ be chosen arbitrarily with $\nabla f(x) \neq 0$. Furthermore, let $d^* \in \mathbb{R}^n$ denote the solution of the problem

$$\min_{\|d\|=1} \nabla f(x)^\top d. \tag{7.1}$$

Then, every vector of the form $s = \lambda d^*$, $\lambda > 0$, is called a *steepest descent direction* of f at x.

The next theorem gives a closed form solution of Problem (7.1).

Theorem 7.4. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and let $x \in \mathbb{R}^n$ be chosen arbitrarily with $\nabla f(x) \neq 0$. Then, Problem (7.1) has the unique solution

$$d^* = -\frac{\nabla f(x)}{\|\nabla f(x)\|}.$$

In particular, a direction s is a steepest descent direction of f at x if and only if there exists a positive scalar $\lambda > 0$ so that s can be written as $s = -\lambda \nabla f(x)$.

Proof. The inequality of Cauchy and Schwarz states that

$$|v^\top w| \leq \|v\| \|w\|$$

holds for all $v, w \in \mathbb{R}^n$. Moreover, equality holds if and only if v and w are linearly dependent. Thus, for $d \in \mathbb{R}^n$ with $\|d\| = 1$ it holds

$$\nabla f(x)^\top d \geq -\|\nabla f(x)\|\|d\| = -\|\nabla f(x)\|$$

and equality holds if and only if

$$d = d^* = -\frac{\nabla f(x)}{\|\nabla f(x)\|}.$$

This proves the first statement. The second statement directly follows from the definition of steepest descent directions. □

The gradient method is thus obtained from Algorithm 1 if we choose the negative gradients

$$s^k = -\nabla f(x^k)$$

in Line 3, i.e., if we choose a steepest descent direction.

Remark 7.5. Note that the statement of Theorem 7.4 is only true for the Euclidean norm $\|x\| = \sqrt{x^\top x}$. If we choose a different norm in (7.1), we would also get different steepest descent directions. Consider, for instance, the case of a symmetric and positive definite matrix $A \in \mathbb{R}^{n\times n}$. In this case, $\|x\|_A := \sqrt{x^\top A x}$ also defines a norm on $\mathbb{R}^n$. The steepest descent directions of f at x with respect to the norm $\|\cdot\|_A$ are given by[2]

$$s = -\lambda A^{-1}\nabla f(x), \quad \lambda > 0.$$

7.2 The Armijo Rule

The only remaining question that needs to be answered to turn Algorithm 1 into a practical method is how to choose the step sizes σ^k in Line 4. In this lecture, we discuss the so-called Armijo rule for choosing the step sizes. For this rule, let $\beta \in (0, 1)$ and $\gamma \in (0, 1)$ be given. The Armijo rule then states to choose the largest number $\sigma^k \in \{1, \beta, \beta^2, \dots\}$ that satisfies

$$f(x^k + \sigma^k s^k) - f(x^k) \leq \sigma^k \gamma \nabla f(x^k)^\top s^k, \tag{7.2}$$

see Figure 7.2 for an illustration of this rule. In practice, frequently used parameters are $\beta = 1/2$ and $\gamma = 10^{-2}$.

The main question now is whether there always exists a step size that satisfies the Armijo rule. Fortunately, this is the case.

[2] Proving this statement is a good exercise.

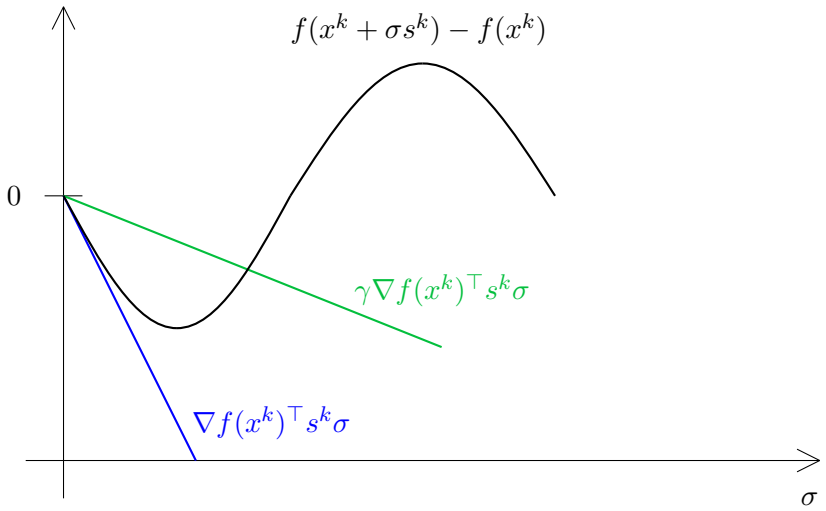


Figure 7.2: The Armijo rule. All step sizes σ are acceptable for which the graph of the black function (which is the left-hand side of the rule's inequality) is below or on the graph of the green function (the right-hand-side of the inequality). For $\gamma = 0$, the green graph would be parallel to the σ-axis, whereas $\gamma = 1$ corresponds to the blue function, which has the slope given by the directional derivative of f at the current iterate x^k.

Theorem 7.6. Let $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on the open set $U \subseteq \mathbb{R}^n$. Moreover, let $\gamma \in (0, 1)$ be given. Finally, suppose that $x \in U$ and that s is a descent direction of f at x. Then, there exists a constant $\bar{\sigma} > 0$ with

$$f(x + \sigma s) - f(x) \leq \sigma \gamma \nabla f(x)^\top s \quad \text{for all} \quad \sigma \in [0, \bar{\sigma}]. \tag{7.3}$$

Proof. Inequality (7.3) is obviously satisfied for $\sigma = 0$. Thus, let $\sigma > 0$ be sufficiently small. Then, $x + \sigma s \in U$ and

$$\begin{aligned} &\frac{f(x + \sigma s) - f(x)}{\sigma} - \gamma \nabla f(x)^\top s \\ \to\ &\nabla f(x)^\top s - \gamma \nabla f(x)^\top s = (1 - \gamma) \nabla f(x)^\top s < 0 \end{aligned}$$

holds for $\sigma \searrow 0$. Hence, we can choose $\bar{\sigma} > 0$ small enough so that

$$\frac{f(x + \sigma s) - f(x)}{\sigma} - \gamma \nabla f(x)^\top s \leq 0 \quad \text{for all} \quad \sigma \in (0, \bar{\sigma}].$$

For this $\bar{\sigma}$, (7.3) is satisfied. □

7.3 Global Convergence of the Gradient Method

Up to now, we have discussed the main principles of descent methods in general and the steepest descent method as a special case. Moreover, we formalized the general descent method.

Next, we formally specify the gradient method and prove a global convergence result for it. The method is given in Algorithm 2.

Algorithm 2 The Gradient Method for Unconstrained Optimization

Input: An objective function $f : \mathbb{R}^n \to \mathbb{R}$, an initial iterate $x^0 \in \mathbb{R}^n$, and algorithmic parameters $\beta, \gamma \in (0, 1)$.

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: **if** $\nabla f(x^k) \neq 0$ **then**
3: Set $s^k = -\nabla f(x^k)$.
4: Compute a step size $\sigma^k \in \mathbb{R}$ according to Armijo's rule (7.2).
5: Set $x^{k+1} = x^k + \sigma^k s^k$.
6: **else**
7: Stop and return x^k.
8: **end if**
9: **end for**

As mentioned above, the termination criterion is usually replaced with $\|\nabla f(x^k)\| > \varepsilon$, $\varepsilon > 0$, in practice.

Let us now prove our first convergence result for a nonlinear optimization algorithm.

Theorem 7.7. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Then, Algorithm 2 terminates after a finite number of iterations with a stationary point of f or it creates an infinite sequence $(x^k)_k$ with the following two properties:

(a) It holds $f(x^{k+1}) < f(x^k)$ for all $k \geq 0$.

(b) Every accumulation point of $(x^k)_k$ is a stationary point of f.

Proof. The only interesting situation is the one in which Algorithm 2 does not stop after a finite number of iterations. According to Theorem 7.6, Algorithm 2 then creates infinite series $(x^k)_k \subseteq \mathbb{R}^n$ and $(\sigma^k)_k \subseteq (0, 1]$ with $\nabla f(x^k) \neq 0$ and

$$f(x^{k+1}) - f(x^k) = f(x^k + \sigma^k s^k) - f(x^k) \leq -\sigma^k \gamma \|\nabla f(x^k)\|^2 < 0.$$

This shows the first property.

We still need to prove the second property. To this end, let $x^* \in \mathbb{R}^n$ be an accumulation point of $(x^k)_k$ and let $(x^{k_i})_i$ be a sub-sequence with $(x^{k_i})_i \to x^*$ for $i \to \infty$. In what follows, we denote this sub-sequence with $(x^k)_K$, i.e., $K = \{k_1, k_2, \dots\}$. The sequence $(f(x^k))_k$ is monotonically decreasing and thus has a limit $\varphi \in \mathbb{R} \cup \{-\infty\}$. In particular, we have $(f(x^{k_i})_K \to \varphi$. This implies $\varphi = f(x^*)$ and $(f(x^k))_K \to f(x^*)$ since f is continuous. Using Armijo's rule we also obtain

$$f(x^0) - f(x^*) = \sum_{k=0}^{\infty} f(x^k) - f(x^{k+1}) \geq \gamma \sum_{k=0}^{\infty} \sigma^k \|\nabla f(x^k)\|^2.$$

In particular, this implies

$$\lim_{k\to\infty} \sigma^k \|\nabla f(x^k)\|^2 = 0. \tag{7.4}$$

The remaining part of the proof is done by contradiction. Thus, assume that $\nabla f(x^*) \neq 0$ holds. Since ∇f is continuous and because $(x^{k_i})_K \to x^*$ holds, there exists an index $\ell \in K$ with

$$\|\nabla f(x^k)\| \geq \frac{\|\nabla f(x^*)\|}{2} > 0 \quad \text{for all} \quad k \in K,\ k \geq \ell.$$

Using (7.4), this implies $(\sigma^k)_k \to 0$. In particular, this means that there exists an $\ell' \in K$, $\ell' \geq \ell$, with $\sigma^k \leq \beta$ for all $k \in K$ with $k \geq \ell'$. Armijo's rule (7.2) thus yields

$$f(x^k + \beta^{-1}\sigma^k s^k) - f(x^k) > -\gamma\beta^{-1}\sigma^k \|\nabla f(x^k)\|^2 \tag{7.5}$$

for all $k \in K$, $k \geq \ell'$. We now construct a new sequence $(t^k)_K$ with $t^k = \beta^{-1}\sigma^k$. This new sequence $(t^k)_K$ converges to 0 and the mean value theorem implies the existence of a value $\tau_k \in [0, t_k]$ with

$$\begin{aligned} & \lim_{k\to\infty, k\in K} \frac{f(x^k + t^k s^k) - f(x^k)}{t^k} \\ = & \lim_{k\to\infty, k\in K} \frac{t^k \nabla f(x^k + \tau^k s^k)^\top s^k}{t^k} \\ = & -\|\nabla f(x^*)\|^2, \end{aligned}$$

and

$$\lim_{k\to\infty, k\in K} \|\nabla f(x^k)\|^2 = \|\nabla f(x^*)\|^2.$$

Together with (7.5), this implies the contradiction

$$0 < (\gamma - 1)\|\nabla f(x^*)\| \leq 0.$$

Hence, the assumption $\nabla f(x^*) \neq 0$ was wrong and the proof is finished. □

Remark 7.8. Note that the phrase "global convergence" does not mean that the algorithm converges to a global minimum. The phrase "global convergence" means that the method converges independent of the chosen initial iterate.

7.4 Speed of Convergence of the Gradient Method

We have shown that the gradient method converges to a stationary point if the objective function is continuously differentiable. This is good news. The bad news is: The speed of convergence of the gradient method may be very slow in practice. In this section, we first give a geometric intuition, why this is the case. Afterward, we consider a specific example and finally prove a convergence result for strictly convex and quadratic objective functions.

We now consider the following "exact" step size rule: Choose $\sigma^k > 0$ such that

$$f(x^k + \sigma^k s^k) = \min_{\sigma \geq 0} f(x^k + \sigma s^k).$$

One can show that the gradient method is also globally convergent for strictly convex and quadratic functions[3] if Armijo's rule is replaced with the exact rule above.

We will now try to get an geometric intuition for what happens in the gradient method. To this end, let the current iterate be x^k and the respective search direction is $s^k = -\nabla f(x^k)$. The gradient of the function (and, thus, the search direction) is orthogonal to the level set of x^k. Why is this the case? First, the level set of x^k is given by[4]

$$L_k := \{x \in \mathbb{R}^n \colon f(x) = f(x^k)\}.$$

Since f is continuously differentiable and because $\nabla f(x^k) \neq 0$ holds, L_k is a continuously differentiable hyper-surface (a curve in $\mathbb{R}^2$). We now consider an arbitrary $\mathcal{C}^1$-curve $\gamma : (-1, 1) \to L_k$ with $\gamma(0) = x^k$. It holds

$$f(\gamma(t)) = f(x^k)$$

and differentiating both sides w.r.t. t yields

$$\nabla f(\gamma(t))^\top \gamma'(t) = 0.$$

[3] ... and also for more general classes of objective functions.
[4] In $\mathbb{R}^2$, these are just the contour lines of the function.

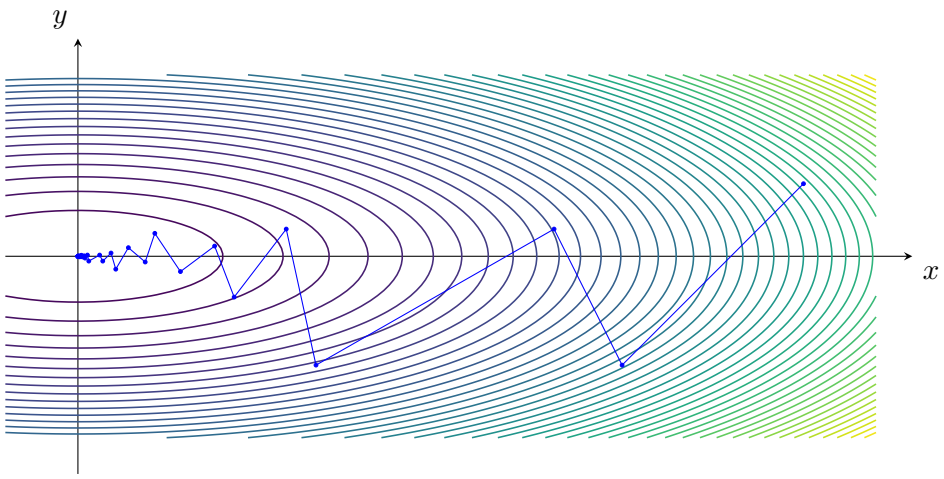


Figure 7.3: The gradient method applied to the function $f(x) = x_1^2 + 10x_2^2$ and initial iterate $x^0 = (10, 1)^\top$. Note that the iterates have been computed with the gradient method using the Armijo step size rule. This is the reason why the zig-zagging line does not have right angles. However, the qualitative behavior can be seen as well.

Thus,

$$\nabla f(\gamma(0))^\top \gamma'(0) = \nabla f(x^k)^\top \gamma'(0) = 0$$

holds. This means that the search direction $s^k = -\nabla f(x^k)$ is orthogonal to L_k. Starting at x^k, we thus follow a direction, which is orthogonal to L_k, until we reach the global minimizer $\sigma^k > 0$ of the function $\varphi(\sigma) := f(x^k + \sigma s^k)$. At this global minimizer, we have $\varphi'(\sigma^k) = 0$. Moreover, $\varphi'(\sigma) = \nabla f(x^k + \sigma s^k)^\top s^k$ implies

$$\varphi'(\sigma^k) = \nabla f(x^k + \sigma^k s^k)^\top s^k = 0.$$

Consequently, at the new point $x^{k+1} = x^k + \sigma^k s^k$, the search direction s^k is orthogonal to $\nabla f(x^{k+1})$, which is the normal of the level set L_{k+1}. In x^{k+1}, it thus holds $s^k \| L_{k+1}$. The next search direction s^{k+1} then is again orthogonal to L_{k+1}, and thus also orthogonal to s^k.

What have we seen by now? The traverse created by the method is a zig-zag line with right angles; see Figure 7.3 for an illustration of this effect. Geometrically, it is thus clear that the method may converge rather slowly if the condition number

$$\kappa(\nabla^2 f(x)) = \frac{\lambda_{\max}(\nabla^2 f(x))}{\lambda_{\min}(\nabla^2 f(x))}$$

is large.

Let us now be a little bit more specific. To this end, we consider strictly

convex and quadratic objective functions, i.e.,

$$f(x) = c^\top x + \frac{1}{2} x^\top C x, \quad c \in \mathbb{R}^n, \quad C = C^\top \in \mathbb{R}^{n \times n} \text{ positive definite.}$$

In this case, we have

$$\begin{aligned} \varphi'(\sigma) &= \nabla f(x^k + \sigma s^k)^\top s^k = (c + C(x^k + \sigma s^k))^\top s^k, \\ \varphi''(\sigma) &= (s^k)^\top \nabla^2 f(x^k + \sigma s^k) s^k = (s^k)^\top C s^k > 0 \quad \text{for all } s^k \in \mathbb{R}^n \setminus \{0\}. \end{aligned}$$

This, in particular, implies that φ is a strictly convex function. Thus, σ^k is characterized by

$$\varphi'(\sigma^k) = (c + C(x^k + \sigma^k s^k))^\top s^k = 0,$$

which implies

$$\sigma^k = -\frac{(c + Cx^k)^\top s^k}{(s^k)^\top C s^k} = -\frac{\nabla f(x^k)^\top s^k}{(s^k)^\top C s^k} = \frac{\|\nabla f(x^k)\|^2}{\nabla f(x^k)^\top C \nabla f(x^k)} = \frac{\|s^k\|^2}{(s^k)^\top C s^k}.$$

We now consider the convergence behavior of the gradient method for a specific example.

Example 7.9. We consider the minimization of the objective function

$$f : \mathbb{R}^2 \to \mathbb{R}, \quad f(x) = f(x_1, x_2) = x_1^2 + a x_2^2, \quad a > 1.$$

Thus,

$$f(x) = c^\top x + \frac{1}{2} x^\top C x$$

holds with

$$c = 0 \in \mathbb{R}^2, \quad C = \begin{bmatrix} 2 & 0 \\ 0 & 2a \end{bmatrix} \in \mathbb{R}^{2 \times 2}.$$

We start the gradient method with the exact step size rule at $x^0 = (a, 1)^\top \in \mathbb{R}^2$. The first search direction then reads

$$s^0 = -\nabla f(x^0) = -\begin{pmatrix} 2x_1^0 \\ 2a x_2^0 \end{pmatrix} = \begin{pmatrix} -2a \\ -2a \end{pmatrix}$$

and the step size is given by

$$\sigma^0 = \frac{\|s^0\|^2}{(s^0)^\top C s^0} = \frac{8a^2}{4a^2(2 + 2a)} = \frac{1}{1 + a}.$$

Thus, the next iterate can be computed:

$$x^1 = \begin{pmatrix} a \\ 1 \end{pmatrix} + \frac{1}{1+a} \begin{pmatrix} -2a \\ -2a \end{pmatrix} = \frac{a-1}{a+1} \begin{pmatrix} a \\ -1 \end{pmatrix}.$$

Let us also do the second iteration. It holds

$$\begin{aligned} s^1 &= \frac{a-1}{a+1} \begin{pmatrix} -2a \\ 2a \end{pmatrix}, \\ \sigma^1 &= \frac{\|s^1\|^2}{(s^1)^\top C s^1} = \frac{1}{1+a} = \sigma^0, \\ x^2 &= \left(\frac{a-1}{a+1} \right)^2 \begin{pmatrix} a \\ 1 \end{pmatrix} = \left(\frac{a-1}{a+1} \right)^2 x^0. \end{aligned}$$

Via induction one can now easily show that[5]

$$x^k = \left(\frac{a-1}{a+1} \right)^k \begin{pmatrix} a \\ 1 \end{pmatrix}, \quad s^k = \left(\frac{a-1}{a+1} \right)^k \begin{pmatrix} -2a \\ -2a \end{pmatrix}$$

holds for even k and that

$$x^k = \left(\frac{a-1}{a+1} \right)^k \begin{pmatrix} a \\ -1 \end{pmatrix}, \quad s^k = \left(\frac{a-1}{a+1} \right)^k \begin{pmatrix} -2a \\ 2a \end{pmatrix}$$

holds for odd k. Moreover, we have $\sigma^k = 1/(1+a)$ for all k.

We now use the above formulas to analyze the rate of the convergence of the gradient method with exact step size rule applied to this specific example of a strictly convex and quadratic function. The global minimizer of f is $x^* = 0 \in \mathbb{R}^2$ and the minimum and maximum eigenvalues of C are 2 and $2a$, respectively. Hence, the rate of convergence is given by

$$\|x^{k+1} - x^*\| = \frac{a-1}{a+1} \|x^k - x^*\| = \frac{\lambda_{\max}(C) - \lambda_{\min}(C)}{\lambda_{\max}(C) + \lambda_{\min}(C)} \|x^k - x^*\|.$$

For the objective function values, we obtain

$$f(x^k) = (x_1^k)^2 + a(x_2^k)^2 = \left(\frac{a-1}{a+1} \right)^{2k} (a^2 + a)$$

[5] This might be a good exercise!

and, consequently,

$$\begin{aligned} f(x^{k+1}) - f(x^*) &= \left(\frac{a-1}{a+1}\right)^2 (f(x^k) - f(x^*)) \\ &= \left(\frac{\lambda_{\max}(C) - \lambda_{\min}(C)}{\lambda_{\max}(C) + \lambda_{\min}(C)}\right)^2 (f(x^k) - f(x^*)). \end{aligned}$$

Thus, $\lambda_{\min}(C) \ll \lambda_{\max}(C)$ implies that the factor

$$\left(\frac{\lambda_{\max}(C) - \lambda_{\min}(C)}{\lambda_{\max}(C) + \lambda_{\min}(C)}\right)^2$$

is almost 1 and the rate of convergence is very bad.

The next theorem shows that the rate of convergence obtained in the last example is the worst possible. To prove this statement, we need another technical lemma.

Lemma 7.10 (Kantorovich's inequality)**.** Let $C \in \mathbb{R}^{n\times n}$ be a symmetric and positive definite matrix. Then,

$$\frac{\|d\|^4}{(d^\top C d)(d^\top C^{-1} d)} \geq \frac{4\lambda_{\min(C)}\lambda_{\max(C)}}{(\lambda_{\min(C)} + \lambda_{\max(C)})^2}$$

holds for all $d \in \mathbb{R}^n \setminus \{0\}$.

Proof. See, e.g., Lemma 8.5 on Page 71 in Geiger and Kanzow (2013). □

Theorem 7.11. Let $f : \mathbb{R}^n \to \mathbb{R}$ be strictly convex and quadratic. Furthermore, let $(x^k)_k$ and $(\sigma^k)_k$ be the sequences of iterates and step sizes created by the gradient method with the exact step size rule. Then,

$$f(x^{k+1}) - f(x^*) \leq \left(\frac{\lambda_{\max}(C) - \lambda_{\min}(C)}{\lambda_{\max}(C) + \lambda_{\min}(C)}\right)^2 (f(x^k) - f(x^*)), \tag{7.6a}$$

$$\|x^k - x^*\| \leq \left(\sqrt{\frac{\lambda_{\max}(C)}{\lambda_{\min}(C)}}\,\frac{\lambda_{\max}(C) - \lambda_{\min}(C)}{\lambda_{\max}(C) + \lambda_{\min}(C)}\right)^k \|x^0 - x^*\| \tag{7.6b}$$

holds, where $x^* = -C^{-1}c$ is the unique global minimizer of f and where $\lambda_{\min}(C)$ and $\lambda_{\max}(C)$ denote the smallest and largest eigenvalue of C, respectively.

Proof. Since f is quadratic, the Taylor expansion around x^* yields

$$f(x)-f(x^*) = \nabla f(x^*)^\top (x-x^*)+\frac{1}{2}(x-x^*)^\top C(x-x^*) = \frac{1}{2}(x-x^*)^\top C(x-x^*)$$

and

$$\nabla f(x) = Cx + c = Cx - Cx^* = C(x - x^*).$$

Here, we used twice that $\nabla f(x^*) = 0$ holds. The Taylor expansion around x^k leads to

$$\begin{aligned} f(x^{k+1}) &= f(x^k) + \sigma^k \nabla f(x^k)^\top s^k + \frac{(\sigma^k)^2}{2}(s^k)^\top C s^k \\ &= f(x^k) - \sigma^k \|s^k\|^2 + \frac{(\sigma^k)^2}{2}(s^k)^\top C s^k. \end{aligned}$$

Using the formula

$$\sigma^k = \frac{\|s^k\|^2}{(s^k)^\top C s^k}$$

for the exact step size, we obtain

$$\begin{aligned} f(x^{k+1}) - f(x^*) &= f(x^k) - f(x^*) - \sigma^k \|s^k\|^2 + \frac{(\sigma^k)^2}{2}(s^k)^\top C s^k \\ &= f(x^k) - f(x^*) - \frac{\|s^k\|^4}{(s^k)^\top C s^k} + \frac{1}{2}\frac{\|s^k\|^4}{(s^k)^\top C s^k} \\ &= f(x^k) - f(x^*) - \frac{1}{2}\frac{\|s^k\|^4}{(s^k)^\top C s^k}. \end{aligned}$$

This implies

$$\begin{aligned} f(x^k) - f(x^*) &= \frac{1}{2}(x^k - x^*)^\top C (x^k - x^*) \\ &= \frac{1}{2}(C(x^k - x^*))^\top C^{-1} (C(x^k - x^*)) \\ &= \frac{1}{2}(s^k)^\top C^{-1} s^k, \end{aligned}$$

which itself implies

$$\begin{aligned} &\left(1 - \frac{\|s^k\|^4}{((s^k)^\top C s^k)((s^k)^\top C^{-1} s^k)}\right)(f(x^k) - f(x^*)) \\ &= f(x^k) - f(x^*) - \frac{1}{2}\frac{\|s^k\|^4}{((s^k)^\top C s^k)((s^k)^\top C^{-1} s^k)}(s^k)^\top C^{-1} s^k \\ &= f(x^k) - f(x^*) - \frac{1}{2}\frac{\|s^k\|^4}{(s^k)^\top C s^k} \\ &= f(x^{k+1}) - f(x^*). \end{aligned}$$

The inequality of Kantorovich (Lemma 7.10) then proves

$$\begin{aligned}
& f(x^{k+1}) - f(x^*) \\
= & \left(1 - \frac{\|s^k\|^4}{((s^k)^\top C s^k)((s^k)^\top C^{-1} s^k)}\right)(f(x^k) - f(x^*)) \\
\leq & \left(1 - \frac{4\lambda_{\min(C)}\lambda_{\max(C)}}{(\lambda_{\min(C)} + \lambda_{\max(C)})^2}\right)(f(x^k) - f(x^*)) \\
= & \left(\frac{\lambda_{\max}(C) - \lambda_{\min}(C)}{\lambda_{\max}(C) + \lambda_{\min}(C)}\right)^2 (f(x^k) - f(x^*)),
\end{aligned}$$

which proves (7.6a).

To show Inequality (7.6b), we use the inequalities

$$f(x) - f(x^*) = \frac{1}{2}(x - x^*)^\top C (x - x^*) \begin{cases} \leq \frac{\lambda_{\max}(C)}{2}\|x - x^*\|^2, \\ \geq \frac{\lambda_{\min}(C)}{2}\|x - x^*\|^2. \end{cases}$$

We apply these inequalities now for $x = x^k$ as well as $x = x^{k+1}$ and obtain

$$\begin{aligned}
& \|x^{k+1} - x^*\|^2 \\
\leq & \frac{2}{\lambda_{\min}(C)}\left(f(x^{k+1}) - f(x^*)\right) \\
\leq & \frac{2}{\lambda_{\min}(C)}\left(\frac{\lambda_{\max}(C) - \lambda_{\min}(C)}{\lambda_{\max}(C) + \lambda_{\min}(C)}\right)^2 \left(f(x^k) - f(x^*)\right) \\
\leq & \frac{2}{\lambda_{\min}(C)}\frac{\lambda_{\max}(C)}{2}\left(\frac{\lambda_{\max}(C) - \lambda_{\min}(C)}{\lambda_{\max}(C) + \lambda_{\min}(C)}\right)^2 \|x^k - x^*\|^2 \\
\leq & \frac{\lambda_{\max}(C)}{\lambda_{\min}(C)}\left(\frac{\lambda_{\max}(C) - \lambda_{\min}(C)}{\lambda_{\max}(C) + \lambda_{\min}(C)}\right)^2 \|x^k - x^*\|^2
\end{aligned}$$

Taking the square root on both sides and iterating this inequality yields the claim. □

8

Newton's Method

Newton's method is one of the most important methods in numerical mathematics and optimization. It can be seen as a method for solving systems of nonlinear equations and for solving nonlinear optimization problems. We consider both variants and start with the former. To this end, we consider the nonlinear system of equations

$$F(x) = 0 \tag{8.1}$$

with a continuously differentiable function $F : \mathbb{R}^n \to \mathbb{R}^n$. The main idea of Newton's method is the following. Let $x^k \in \mathbb{R}^n$ be already computed. Obviously, (8.1) then is equivalent to

$$F(x^k + s) = 0, \tag{8.2}$$

because $s = d^k \in \mathbb{R}^n$ solves (8.2) if and only if $x = x^k + d^k$ solves (8.1). Let us now replace $F(x^k + s)$ by its first-order Taylor expansion. This leads to

$$F(x^k + s) = F(x^k) + F'(x^k)s + \rho(s),$$

where $\|\rho(s)\| = o(\|s\|)$ holds. This means, that the error term $\rho(s)$ is small for small s. In the kth iteration of Newton's method, we replace the nonlinear system of equations (8.2) with the linearized system of equations

$$F(x^k) + F'(x^k)s = 0,$$

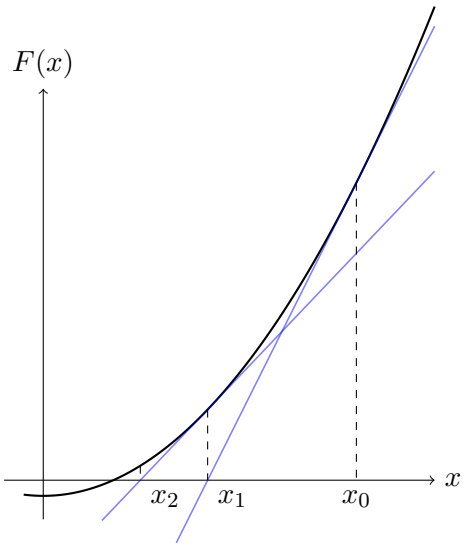


Figure 8.1: Newton's method in 1d

where F' denotes the *Jacobian*

$$F'(x) = \begin{bmatrix} \frac{\partial F_1}{\partial x_1}(x) & \dots & \frac{\partial F_1}{\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial F_n}{\partial x_1}(x) & \dots & \frac{\partial F_n}{\partial x_n}(x) \end{bmatrix}.$$

The entire method is given in Algorithm 3 and an illustration of some iterations for the one-dimensional case is given in Figure 8.1.

Algorithm 3 Newton's Method for Nonlinear Systems of Equations

Input: A function $F : \mathbb{R}^n \to \mathbb{R}^n$ and an initial iterate $x^0 \in \mathbb{R}^n$.

1: **for** $k = 0, 1, 2, \dots$ **do**
2: **if** $F(x^k) \neq 0$ **then**
3: Compute the Newton step s^k by solving the Newton equation

$$F'(x^k)s^k = -F(x^k).$$

4: Set $x^{k+1} = x^k + s^k$.
5: **else**
6: Stop and return x^k.
7: **end if**
8: **end for**

8.1 Fast Local Convergence

In the last chapter, we saw that the gradient method is globally convergent (under some mild assumptions) but that the rate of convergence may be very bad—even for strictly convex and quadratic objective functions, which is one of the nicest classes of objective functions. We will now show that Newton's method has a local convergence that is much faster. To this end, we make the following definition about the rate of convergence of sequences.

Definition 8.1 (Rates of convergence)**.** The sequence $(x^k)_k \subseteq \mathbb{R}^n$ converges

(a) *q-linearly*[1] with rate $0 < \gamma < 1$ to $x^* \in \mathbb{R}^n$, if there exists an $l \geq 0$ with

$$\|x^{k+1} - x^*\| \leq \gamma \|x^k - x^*\| \quad \text{for all} \quad k \geq l.$$

(b) *q-superlinearly* to $x^* \in \mathbb{R}^n$ if $x^k \to x^*$ and

$$\|x^{k+1} - x^*\| = o(\|x^k - x^*\|) \quad \text{for} \quad k \to \infty$$

holds. Note that the latter condition is, by definition, equivalent to

$$\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} \to 0 \quad \text{for} \quad k \to \infty.$$

(c) *q-quadratically* to $x^* \in \mathbb{R}^n$, if $x^k \to x^*$ and

$$\|x^{k+1} - x^*\| = O(\|x^k - x^*\|^2) \quad \text{for} \quad k \to \infty$$

holds. Note that the latter condition is, again by definition, equivalent to the existence of a constant $C > 0$ with

$$\|x^{k+1} - x^*\| \leq C\|x^k - x^*\|^2 \quad \text{for all} \quad k \geq 0.$$

(d) *r-linearly*[2] with rate $0 < \gamma < 1$ to $x^* \in \mathbb{R}^n$ if there exists a sequence $(\alpha^k)_k \subseteq (0, \infty)$ that converges q-linearly with rate γ to 0 so that

$$\|x^k - x^*\| \leq \alpha^k \quad \text{for} \quad k \to \infty$$

holds.

(e) *r-superlinearly* to $x^* \in \mathbb{R}^n$ if there exists a sequence $(\alpha^k)_k \subseteq (0, \infty)$

[1] The "q" stands for "quotient", since this type of convergence is defined based on the quotient of successive errors.

[2] The "r" stands for "root".

that converges q-superlinearly to 0 so that

$$\|x^k - x^*\| \leq \alpha^k \quad \text{for} \quad k \to \infty$$

holds.

(f) *r-quadratically* to $x^* \in \mathbb{R}^n$ if there exists a sequence $(\alpha^k)_k \subseteq (0, \infty)$ that converges q-quadratically to 0 so that

$$\|x^k - x^*\| \leq \alpha^k \quad \text{for} \quad k \to \infty$$

holds.

For proving the fast local convergence of Newton's method we need the following lemma.

Lemma 8.2 (Banach's Lemma)**.** The set $\mathcal{M} \subseteq \mathbb{R}^{n \times n}$ of invertible matrices is open and the mapping $\mathcal{M} \ni M \mapsto M^{-1}$ is continuous. More specifically: For $A \in \mathcal{M}$ and all $B \in \mathbb{R}^{n \times n}$ with $\|A^{-1}B\| < 1$ (and thus, in particular, if $\|A^{-1}\|\|B\| < 1$ holds), the following is true. $A + B$ is invertible and

$$\|(A+B)^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}B\|}$$

as well as

$$\|(A+B)^{-1} - A^{-1}\| \leq \frac{\|A^{-1}\|\|A^{-1}B\|}{1 - \|A^{-1}B\|}$$

holds.

Proof. Let $A \in \mathcal{M}$ be arbitrary. Furthermore, let $B \in \mathbb{R}^{n \times n}$ arbitrary with $\|A^{-1}B\| < 1$. We abbreviate $M = -A^{-1}B$. Then, the Neumann series (with $M^0 = I$) converges, i.e.,

$$S = \sum_{k=0}^{\infty} M^k,$$

since, with $S_n = \sum_{k=0}^{n} M^k$, it holds

$$\|S - S_n\| = \left\| \sum_{k=n+1}^{\infty} M^k \right\| \leq \sum_{k=n+1}^{\infty} \|M^k\| \leq \|M\|^{n+1} \sum_{k=0}^{\infty} \|M\|^k = \frac{\|M\|^{n+1}}{1 - \|M\|} \to 0$$

for $n \to \infty$. Moreover, we have

$$\begin{aligned}(I-M)S_n &= (I-M)\sum_{k=0}^{n} M^k \\ &= M^0 - M^1 + M^1 - M^2 + \cdots + M^n - M^{n+1} \\ &= I - M^{n+1}.\end{aligned}$$

Taking the limit for $n \to \infty$ leads to $(I-M)S = I$ and we have thus shown $(I-M) \in \mathcal{M}$ as well as $(I-M)^{-1} = S$. It also holds

$$A(I-M) = A - AM = A + AA^{-1}B = A + B$$

and $(A+B) \in \mathcal{M}$ with $(A+B)^{-1} = SA^{-1}$, since

$$(A+B)^{-1} = (A(I-M))^{-1} = (I-M)^{-1}A^{-1} = SA^{-1}$$

holds. Finally, we have

$$\begin{aligned}\|(A+B)^{-1}\| = \|SA^{-1}\| &\leq \|A^{-1}\|\|S\| \\ &\leq \left\|A^{-1}\right\| \left\|\sum_{k=0}^{\infty} M^k\right\| \leq \left\|A^{-1}\right\| \sum_{k=0}^{\infty} \|M\|^k \\ &= \frac{\|A^{-1}\|}{1-\|M\|}\end{aligned}$$

and

$$\begin{aligned}&\|(A+B)^{-1} - A^{-1}\| \\ &= \|SA^{-1} - A^{-1}\| \\ &= \left\|\sum_{k=0}^{\infty} M^k A^{-1} - A^{-1}\right\| \\ &\leq \|A^{-1}\| \sum_{k=1}^{\infty} \|M\|^k \\ &= \frac{\|A^{-1}\|\|M\|}{1-\|M\|}.\end{aligned}$$

□

Next, we show that a solution $x^* \in \mathbb{R}^n$ of (8.1) is isolated, if the Jacobian matrix $F'(x^*)$ is invertible.

Lemma 8.3. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable. Moreover, let $F(x^*) = 0$ and suppose that $F'(x^*)$ is invertible. Then, there exists an $\varepsilon > 0$

and a constant $\gamma > 0$ with

$$\|F(x)\| \geq \gamma \|x - x^*\| \quad \text{for all} \quad x \in B_\varepsilon(x^*).$$

This, in particular, implies that x^* is an isolated root of F.

Proof. It holds

$$\|x - x^*\| = \|F'(x^*)^{-1} F'(x^*)(x - x^*)\| \leq \|F'(x^*)^{-1}\| \|F'(x^*)(x - x^*)\|.$$

We now set

$$\gamma = \frac{1}{2\|F'(x^*)^{-1}\|},$$

which implies

$$\|F'(x^*)(x - x^*)\| \geq \frac{\|x - x^*\|}{\|F'(x^*)^{-1}\|} = 2\gamma \|x - x^*\|.$$

Using the definition of differentiability, we know that there exists an $\varepsilon > 0$ with

$$\|F(x) - F(x^*) - F'(x^*)(x - x^*)\| \leq \gamma \|x - x^*\| \quad \text{for all} \quad x \in B_\varepsilon(x^*).$$

Finally, $F(x^*) = 0$ and the triangle inequality imply

$$\begin{aligned} 2\gamma \|x - x^*\| &\leq \|F'(x^*)(x - x^*)\| \\ &= \|F(x) - (F(x) - F(x^*) - F'(x^*)(x - x^*))\| \\ &\leq \|F(x)\| + \|F(x) - F(x^*) - F'(x^*)(x - x^*)\| \\ &\leq \|F(x)\| + \gamma \|x - x^*\| \end{aligned}$$

for all $x \in B_\varepsilon(x^*)$. □

Definition 8.4 (Lipschitz continuity)**.** A function $F : \mathbb{R}^n \to \mathbb{R}^n$ is called *Lipschitz continuous* on $X \subseteq \mathbb{R}^n$ if there exists a constant (the so-called *Lipschitz constant*) $L > 0$ with

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

Before we prove the following local convergence theorem for Newton's method let us briefly recap two facts from mathematical analysis:

(a) Let $x, y \in \mathbb{R}^n$ and let $F : \mathbb{R}^n \to \mathbb{R}^n$ be sufficiently smooth. For $G(t) = F(x + t(y - x))$, i.e., $G : \mathbb{R} \to \mathbb{R}^n$, it holds $G'(t) \in \mathbb{R}^n$ with

$G'(t) = F'(x + t(y - x))(y - x)$ and

$$\int_0^1 F'(x+t(y-x))(y-x)\,\mathrm{d}t = \int_0^1 G'(t)\,\mathrm{d}t = G(1)-G(0) = F(y)-F(x).$$

(b) Let $V : [0, 1] \subset \mathbb{R} \to \mathbb{R}^n$ be continuous. Then, it holds

$$\left\| \int_0^1 V(t)\,\mathrm{d}t \right\| \leq \int_0^1 \|V(t)\|\,\mathrm{d}t.$$

The inequality can be shown by approximating the integral with Riemann sums and by using the triangle inequality.

Theorem 8.5. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable and let $x^* \in \mathbb{R}^n$ be a point with $F(x^*) = 0$. Suppose further that the Jacobian matrix $F'(x^*)$ is invertible. Then, there exist constants $\delta > 0$ and $C > 0$ so that the following statements are true:

(a) x^* is the only root of F in $B_\delta(x^*)$.

(b) $\|F'(x)^{-1}\| \leq C$ for all $x \in B_\delta(x^*)$.

(c) For all $x^0 \in B_\delta(x^*)$, Algorithm 3 terminates with $x^k = x^*$ or the algorithm creates a sequence $(x^k)_k \subseteq B_\delta(x^*)$ that converges q-superlinearly to x^*.

(d) Suppose further that F' is Lipschitz continuous on $B_\delta(x^*)$ with Lipschitz constant L. Then, the rate of convergence (if the algorithm does not terminate after a finite number of iterations) is even q-quadratic, i.e., it holds

$$\|x^{k+1} - x^*\| \leq \frac{CL}{2}\|x^k - x^*\|^2 \quad \text{for all} \quad k \geq 0.$$

Proof. (a) Lemma 8.3 implies the existence of a constant $\delta_1 > 0$ such that x^* is the only root of F in $B_{\delta_1}(x^*)$.

(b) Since F' is continuous and since $F'(x^*)$ is invertible, Banach's lemma (Lemma 8.2) implies the existence of constants $0 < \delta_2 \leq \delta_1$ and $C > 0$ with $\|F'(x)^{-1}\| \leq C$ for all $x \in B_{\delta_2}(x^*)$.

(c) For all $x, y \in \mathbb{R}^n$, Taylor's theorem implies

$$F(y) = F(x) + F'(x)(y - x) + R(x, y)$$

with error term (see the Remark (a) before the theorem)

$$R(x,y) = \int_0^1 F'(x + t(y-x))(y-x)\,\mathrm{d}t - F'(x)(y-x).$$

Moreover, from $F(x^*) = 0$ it follows for all $x^k \in B_{\delta_2}(x^*)$ that

$$\begin{aligned} x^{k+1} - x^* &= x^{k+1} - x^k + x^k - x^* \\ &= -F'(x^k)^{-1}F(x^k) + x^k - x^* \\ &= F'(x^k)^{-1}\left(-F(x^k) + F'(x^k)(x^k - x^*)\right) \\ &= F'(x^k)^{-1}\left(F(x^*) - F(x^k) - F'(x^k)(x^* - x^k)\right) \\ &= F'(x^k)^{-1}R(x^k, x^*) \end{aligned}$$

holds. Remark (b) before this theorem yields

$$\begin{aligned} \|R(x,x^*)\| &= \left\| \int_0^1 (F'(x + t(x^* - x)) - F'(x))(x^* - x)\,\mathrm{d}t \right\| \\ &\le \int_0^1 \|(F'(x + t(x^* - x)) - F'(x))(x^* - x)\|\,\mathrm{d}t \qquad (8.3) \\ &\le \int_0^1 \|F'(x + t(x^* - x)) - F'(x)\|\,\mathrm{d}t\,\|x^* - x\|. \end{aligned}$$

Moreover, the continuity of F' yields

$$\int_0^1 \|F'(x + t(x^* - x)) - F'(x)\|\,\mathrm{d}t \to 0 \quad \text{for} \quad x \to x^*. \qquad (8.4)$$

Now, let $0 < \alpha < 1$ be given arbitrarily. Due to (8.3) and (8.4), there exists a constant δ with $0 < \delta \le \delta_2$ with

$$\|R(x,x^*)\| \le \frac{\alpha}{C}\|x - x^*\| \quad \text{for all} \quad x \in B_\delta(x^*).$$

Thus, for all $x \in B_\delta(x^*)$, it holds

$$\begin{aligned} \|x^{k+1} - x^*\| &= \|F'(x^k)^{-1}R(x^k, x^*)\| \\ &\le \|F'(x^k)^{-1}\|\|R(x^k, x^*))\| \\ &\le C\frac{\alpha}{C}\|x^k - x^*\| \\ &= \alpha\|x^k - x^*\|. \end{aligned}$$

This shows the following. For $x^0 \in B_\delta(x^*)$ we have $x^1 \in B_{\alpha\delta}(x^*) \subset B_\delta(x^*)$ and, via induction, $x^k \in B_{\alpha^k\delta}(x^*) \subset B_\delta(x^*)$. If $F(x^k) = 0$, the algorithm terminates. Since $x^k \in B_\delta(x^*) \subseteq B_{\delta_1}(x^*)$ holds, this can only

be the case for $x^k = x^*$. Thus, the algorithm either terminates (after a finite number of iterations) with $x^k = x^*$ or it creates a sequence $(x^k)_k$ that converges to x^*. Using (8.3), we now get

$$\begin{aligned} &\|x^{k+1} - x^k\| \\ \leq\ &\|F'(x^k)^{-1}\| \|R(x^k, x^*))\| \\ \leq\ &C\|R(x^k, x^*))\| \\ \leq\ &C \int_0^1 \|F'(x^k + t(x^* - x^k)) - F'(x^k)\| \,\mathrm{d}t \, \|x^* - x^k\|, \end{aligned}$$

which leads to

$$\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} \leq C \int_0^1 \|F'(x^k + t(x^* - x^k)) - F'(x^k)\| \,\mathrm{d}t \to 0$$

for $k \to \infty$ due to (8.4).

(d) If F' is Lipschitz continuous in $B_\delta(x^*)$, we obtain

$$\begin{aligned} &\int_0^1 \|F'(x^k + t(x^* - x^k)) - F'(x^k)\| \,\mathrm{d}t \\ \leq\ &\int_0^1 Lt\|x^k - x^*\| \,\mathrm{d}t = \frac{L}{2}\|x^k - x^*\|. \end{aligned}$$

□

8.2 Newton's Method for Optimization Problems

Newton's method can also be used for solving unconstrained optimization problems if the objective function is twice continuous differentiable. We now consider two different derivations of this method.

8.2.1 Derivation #1

The first derivation is the easier one. The motivation is the first-order necessary condition for unconstrained optimization problems given in Theorem 5.1:

$$\nabla f(x) = 0. \tag{8.5}$$

If $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, its gradient

$$\nabla f : \mathbb{R}^n \to \mathbb{R}^n$$

is continuously differentiable and we can use Newton's method for nonlinear systems of equations (which is Algorithm 3) to solve System (8.5). This directly leads to Algorithm 4.

Algorithm 4 Newton's Method for Optimization Problems

Input: An objective function $f : \mathbb{R}^n \to \mathbb{R}$ and an initial iterate $x^0 \in \mathbb{R}^n$.

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: **if** $\nabla f(x^k) \neq 0$ **then**
3: Compute the Newton step s^k by solving the Newton equation

$$\nabla^2 f(x^k) s^k = -\nabla f(x^k).$$

4: Set $x^{k+1} = x^k + s^k$.
5: **else**
6: Stop and return x^k.
7: **end if**
8: **end for**

8.2.2 Derivation #2

The second derivation of Newton's method for unconstrained optimization problems is based on the second-order Taylor approximation of the objective function around a given iterate x^k:

$$f(x^k + s) = f(x^k) + \nabla f(x^k)^\top s + \frac{1}{2} s^\top \nabla^2 f(x^k) s + o(\|s\|^2).$$

Thus, it seems reasonable to minimize the quadratic model

$$q_k(s) = \nabla f(x^k)^\top s + \frac{1}{2} s^\top \nabla^2 f(x^k) s$$

for computing the search direction in iteration k.[3]

For what follows, we need the following technical lemma.

Lemma 8.6. Let $A \in \mathbb{R}^{n \times n}$ be symmetric and positive definite. Then, for all $\mu \in (0, \lambda_{\min}(A))$ and all symmetric matrices $B \in \mathbb{R}^{n \times n}$ with $\|B\| \leq \lambda_{\min}(A) - \mu$, it holds

$$\lambda_{\min}(A + B) \geq \mu.$$

[3] Note that we neglected the constant term $f(x^k)$ in the quadratic model.

Proof. It holds

$$\begin{aligned}
&\lambda_{\min}(A+B)\\
=\ &\min_{\|d\|=1} d^\top (A+B)d\\
\geq\ &\min_{\|d\|=1} d^\top Ad - \max_{\|d\|=1} d^\top Bd\\
\geq\ &\lambda_{\min}(A) - \|B\|\\
\geq\ &\lambda_{\min}(A) - (\lambda_{\min}(A) - \mu)\\
=\ &\mu.
\end{aligned}$$

□

Now consider a point $x^* \in \mathbb{R}^n$ that is a local minimizer of f in which the second-order sufficient conditions hold. This means that $\nabla^2 f(x^*)$ is positive definite. The last lemma then shows that there exists an $\varepsilon > 0$ so that $\nabla^2 f(x)$ is positive definite on $B_\varepsilon(x^*)$. For $x^k \in B_\varepsilon(x^*)$ this implies that the quadratic model q_k is strictly convex, which itself yields the fact that there is exactly one stationary point s^k. This stationary point is also the global minimizer of q_k and can be computed by solving

$$\nabla q_k(s^k) = \nabla f(x^k) + \nabla^2 f(x^k)s^k = 0.$$

This is exactly the Newton equation in Line 3 of Algorithm 4. Iterating the described procedure thus also leads to Algorithm 4.

We now prove the direct analogue of Theorem 8.5 for Newton's method applied to optimization problems.

Theorem 8.7. Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable and let $x^* \in \mathbb{R}^n$ be a local minimizer of f, in which the second-order sufficient conditions are satisfied. Then, there exist constants $\delta > 0$ and $\mu > 0$ so that the following statements are true:

(a) x^* is the only stationary point in $B_\delta(x^*)$.

(b) $\lambda_{\min}(\nabla^2 f(x)) \geq \mu$ for all $x \in B_\delta(x^*)$.

(c) For all $x^0 \in B_\delta(x^*)$, Algorithm 4 either terminates after a finite number of iterations with $x^k = x^*$ or it creates a sequence $(x^k)_k \subseteq B_\delta(x^*)$ that converges q-superlinearly to x^*.

(d) Suppose additionally that $\nabla^2 f$ is Lipschitz continuous on $B_\delta(x^*)$ with Lipschitz constant L, i.e.,

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\| \quad \text{for all} \quad x, y \in B_\delta(x^*)$$

holds. Then, the rate of convergence (if the algorithm does not terminate

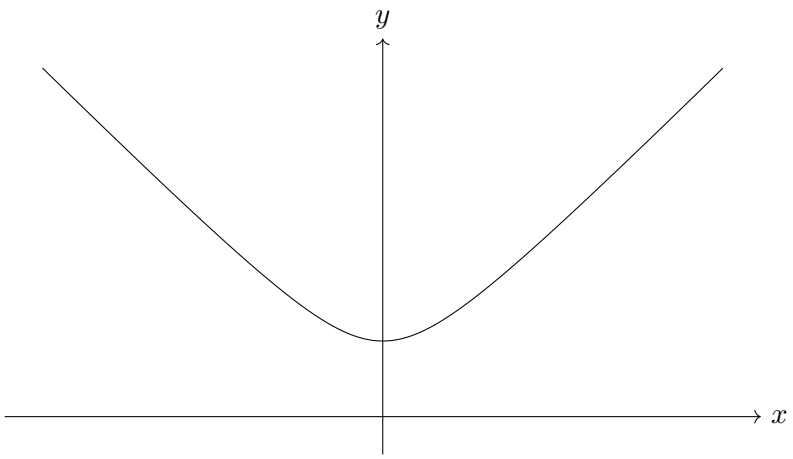


Figure 8.2: The function $f(x) = \sqrt{x^2 + 1}$

after a finite number of iterations) is q-quadratic:

$$\|x^{k+1} - x^*\| \leq \frac{L}{2\mu}\|x^k - x^*\|^2 \quad \text{for all} \quad k \geq 0.$$

Proof. (a) The second-order sufficient conditions imply that the Hessian matrix $\nabla^2 f(x^*)$ is positive definite and, thus, invertible. Hence, we can apply Theorem 8.5 with $F = \nabla f$.

(b) Follows directly from Lemma 8.6.

(c, d) For all $x \in B_\delta(x^*)$ with $\delta > 0$ so that Part (b) holds, we have

$$\|\nabla^2 f(x)^{-1}\| = \frac{1}{\lambda_{\min}(\nabla^2 f(x))} \leq \frac{1}{\mu} =: C.$$

The remaining statements follow from Part (c) and (d) of Theorem 8.5 applied to $F = \nabla f$. □

8.2.3 Global Convergence of a Damped Version

Newton's method considered so far only converges locally, i.e., its convergence is not independent of the given initial iterate.

Example 8.8 (Divergence of Newton's method)**.** Consider the objective function

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = \sqrt{x^2 + 1};$$

see Figure 8.2. Let us now consider what the Newton method is doing when applied to this function. If we choose the initial point $x^0 = 0.99$, we obtain

the iterations as given in Figure 8.3. We see that the method converges very

```
=========================================
k        f(x^k)          ||nabla f(x^k)||
=========================================
0        1.41e+00        7.04e-01
1        1.39e+00        6.96e-01
2        1.35e+00        6.74e-01
3        1.26e+00        6.06e-01
4        1.09e+00        4.05e-01
5        1.00e+00        8.66e-02
6        1.00e+00        6.58e-04
```

Figure 8.3: Fast local convergence of Newton's method

fast (in 6 iterations) and computes the correct objective function value. What happens if we choose $x^0 = 2$? The answer is given in Figure 8.4, in which the first 5 iterations are shown. Something strange happens: The objective

```
=========================================
k        f(x^k)          ||nabla f(x^k)||
=========================================
0        2.24e+00        8.94e-01
1        8.06e+00        9.92e-01
2        5.12e+02        1.00e+00
3        1.34e+08        1.00e+00
4        2.42e+24        1.00e+00
5        1.41e+73        1.00e+00
```

Figure 8.4: Divergence of Newton's method

function value seems to diverge very fast. Moreover, the gradients seem to be wrong since they do not fit together with the divergence of the objective function values. Finally, what about $x^0 = 1$? Some iterations are listed in Figure 8.5. We see that nothing happens. Both the objective function value and the gradient stay the same for all iterations.

Let us now analyze this behavior. The derivatives of the function f are given by

$$\nabla f(x) = \frac{x}{\sqrt{x^2+1}}, \quad \nabla^2 f(x) = \frac{1}{\sqrt{x^2+1}} - \frac{x^2}{(x^2+1)^{3/2}} = \frac{1}{(x^2+1)^{3/2}}.$$

The Newton equation then reads

$$\frac{1}{((x^k)^2+1)^{3/2}} s^k = -\frac{x^k}{\sqrt{(x^k)^2+1}}.$$

```
=======================================
k       f(x^k)          ||nabla f(x^k)||
=======================================
0       1.41e+00        7.07e-01
1       1.41e+00        7.07e-01
2       1.41e+00        7.07e-01
3       1.41e+00        7.07e-01
4       1.41e+00        7.07e-01
5       1.41e+00        7.07e-01
6       1.41e+00        7.07e-01
7       1.41e+00        7.07e-01
8       1.41e+00        7.07e-01
9       1.41e+00        7.07e-01
...     ...             ...
1388    1.41e+00        7.07e-01
1389    1.41e+00        7.07e-01
1390    1.41e+00        7.07e-01
...     ...             ...
41995   1.41e+00        7.07e-01
41996   1.41e+00        7.07e-01
41997   1.41e+00        7.07e-01
41998   1.41e+00        7.07e-01
41999   1.41e+00        7.07e-01
...     ...             ...
```

Figure 8.5: Cycling of Newton's method

Thus, the search direction is

$$s^k = -x^k((x^k)^2 + 1)$$

and the iterates are given by

$$x^{k+1} = x^k + s^k = -(x^k)^3.$$

This implies the following:

(a) For x^0 with $|x^0| < 1$, we have q-cubic convergence.[4]

(b) We have divergence ($|x^k| \to \infty$) for $|x^0| > 1$.

(c) For $|x^0| = 1$, we obtain

$$x^{2k} = x^0, \quad x^{2k+1} = -x^0 \quad \text{for all} \quad k \geq 0,$$

[4] Why?

i.e., the method cycles.

In order to modify Newton's method so that we obtain a globally convergent algorithm, we need to impose a step size rule. The so-called damped (or globalized) Newton's method is given in Algorithm 5.

Algorithm 5 Damped Newton's Method for Optimization Problems

Input: An objective function $f : \mathbb{R}^n \to \mathbb{R}$, an initial iterate $x^0 \in \mathbb{R}^n$, and algorithmic parameters $\beta \in (0,1)$, $\gamma \in (0,1)$, $\alpha_1, \alpha_2 > 0$, and $p > 0$.

for $k = 0, 1, 2, \ldots$ **do**
  if $\nabla f(x^k) \neq 0$ **then**
    Try to compute d^k by solving the Newton equation

$$\nabla^2 f(x^k) d^k = -\nabla f(x^k). \tag{8.6}$$

    if (8.6) is solvable and $-\nabla f(x^k)^\top d^k \geq \min\{\alpha_1, \alpha_2 \|d^k\|^p\} \|d^k\|^2$ **then**
      Set $s^k = d^k$.
    else
      Set $s^k = -\nabla f(x^k)$.
    end if
    Compute the step size $\sigma^k > 0$ using Armijo's rule (7.2).
    Set $x^{k+1} = x^k + \sigma^k s^k$.
  else
    Stop and return x^k.
  end if
end for

Theorem 8.9. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. Then, Algorithm 5 terminates after a finite number of iterations with a stationary point x^k, i.e., $\nabla f(x^k) = 0$ holds, or it creates a sequence $(x^k)_k$ so that every accumulation point of this sequence is a stationary point of f.

Proof. A proof can be found in Ulbrich and Ulbrich (2012); see Theorem 10.10 on Page 49. □

Remark 8.10. A globalization of Newton's method 3 for nonlinear systems of equations can be derived by considering Algorithm 5 applied to the optimization problem

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|F(x)\|^2.$$

Remark 8.11. It can also be shown that, under suitable assumption, there

is a transition of the damped Newton's method to Newton's method (Algorithm 4). This means that also the damped Newton's method can be shown to be a q-superlinear method.

Remark 8.12. In practice, it is often too expensive to compute the exact Hessian matrix of the objective function in every iteration of Newton's method. This leads to the question whether Newton's method also converges if one uses approximations of the Hessian matrices. The corresponding class of methods are called Newton-like or Quasi-Newton methods; for details see, e.g., Ulbrich and Ulbrich (2012) and Nocedal and Wright (2006).

Part III

Theory of Constrained Optimization Problems

9

Examples

In this part, we consider optimization problems for which the set of feasible points is not the entire $\mathbb{R}^n$ but a subset that is specified by equality and inequality constraints. Let us start with some examples.

9.1 Portfolio Optimization

Consider the situation that you are the manager of a portfolio of equities. You have a budget $B > 0$ that you want to invest in n equities so that the expected return R is at least $\rho\,\%$ and that the risk is minimized.

This situation will be modeled in the following. Let r_i be the return of equity i at the end of the considered time horizon, e.g., after one year. This quantity r_i is a random variable. Moreover, let x be the vector describing the portfolio. This means that you invest $x_i B$ in equity i. Thus, it holds

$$\sum_{i=1}^{n} x_i = 1, \quad x \geq 0.$$

The return of the entire portfolio is then given by

$$R(x) = \sum_{i=1}^{n} x_i r_i = r^\top x.$$

Furthermore, let $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ be the mean vector and the covariance matrix of r. Both are typically computed based on historical data. Then, the expected return is

$$\mathbb{E}(R(x)) = \mu^\top x$$

and the variance is given by

$$\mathrm{Var}(R(x)) = x^\top \Sigma x.$$

The latter is a measure of the risk of the portfolio. The classic portfolio optimization problem is thus given by

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & x^\top \Sigma x \\ \text{s.t.} \quad & \sum_{i=1}^{n} x_i = 1, \\ & x \geq 0, \\ & \mu^\top x \geq \rho; \end{aligned}$$

see also Markowitz (1952). The objective function is a quadratic function in x and all constraints are linear. Such an optimization problem is called a *quadratic (optimization) problem* (QP).

9.2 Optimal Placement of Microchip Components

For the design of microchips one often tries to place the (functional) modules on a microchip that are connected by signal lines as close as possible so that the signal communication times are reduced. Of course, one needs to ensure that the modules do not overlap on the microchip.

We now model such a module i, $1 \leq i \leq n$, as a circular disk with center $(x_i, y_i)^\top \in \mathbb{R}^2$ and radius $r_i > 0$. The set

$$E \subseteq \{\{i, j\} \colon 1 \leq i < j \leq n\}$$

represents the set of pairs of connected modules. Moreover, let $w_e > 0$ be a given weight factor that models the importance of a connection between two connected modules $e = \{i, j\} \in E$. Thus, a reasonable placement of modules on a microchip can be computed by solving the optimization problem

$$\begin{aligned} \min_{x,y \in \mathbb{R}^n} \quad & \sum_{e=\{i,j\} \in E} w_e \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \\ \text{s.t.} \quad & (x_i - x_j)^2 + (y_i - y_j)^2 \geq (r_i + r_j)^2, \quad 1 \leq i < j \leq n. \end{aligned}$$

The objective function models the minimization of the weighted distances between the modules and the constraints model that the modules do not overlap.

9.3 Cost Optimal Operation of a Gas Network

A very important field of applied mathematical research of the last years and decades was (and still is) the optimization of gas transport networks.[1]

A typical task of optimization for these networks is the following: The gas transport network is given and not subject to change. Moreover, at certain points of the network, the gas transporting company (often called TSO for "transmission system operator") has contracts with gas supplying and discharging customers. These contracts usually specify, among other aspects, the amount of gas that is supplied or withdrawn within the next day. The main questions for the TSO are now the following:

(a) Is there a control of the network so that the resulting physical and technical state is feasible and that all contracts are satisfied?

(b) Given that there are feasible controls, what is the cost-optimal one among these feasible controls?

The first question represents a pure feasibility task. The second question represents an optimization task.

Let us formalize this setting a little bit more. The gas transport network is modeled as a directed graph $G = (V, A)$ with node set V and arc set A. The node set is further decomposed and made up of the so-called entry nodes V_+, where gas is supplied, of so-called exit-nodes V_-, where gas is withdrawn, and of so-called inner-nodes V_0, where neither gas is supplied nor withdrawn. Thus,

$$V = V_+ \cup V_- \cup V_0$$

holds. Let us now discuss the arcs of the network. The network element that outnumbers all other types of elements are pipes. Pipes are used to transport gas. As every arc, also pipes $a \in A_{\mathrm{pi}}$ have a starting node $u \in V$ and an end node $v \in V$. That is, we may also write $a = (u, v) \in A$. If we neglect the dynamics of the problem (i.e., the dependence of the physical and technical laws with respect to time), the main physical quantities are gas mass flow q_a in a pipe $a \in A_{\mathrm{pi}}$ and gas pressure p_u, $u \in V$, that we consider to be defined on the nodes of the network. If gas flows through a pipe, turbulences at the rough inner wall of the pipe lead to friction forces that then lead to a pressure loss in the pipe. This means that if the gas mass flow q_a in pipe $a = (u, v)$ is from node u to node v, the pressure p_v at node v is smaller than the pressure p_u at node u. An often used approximation for the relation between the gas mass flow in the pipe and the pressures at the

[1] See, e.g., the research project TRR 154 "Mathematical Modelling, Simulation and Optimization Using the Example of Gas Networks": https://trr154.fau.de.

end of the pipe is the so-called *Weymouth equation*

$$p_v^2 = p_u^2 - \Lambda_a |q_a| q_a, \quad a = (u, v) \in A_{\mathrm{pi}}.$$

Here, Λ_a is a constant that encapsulates other technical parameters like the length or the inner roughness of the pipe. In other words (and if we neglect height differences): Gas flows from larger to smaller pressures. In addition, gas pressure needs to satisfy certain physical and technical bounds, i.e.,

$$p_u^- \leq p_u \leq p_u^+, \quad u \in V.$$

Thus, for being able to transport gas over large distances in the network, one needs to increase the pressure. This is done in so-called compressor machines that we model as arcs $a = (u, v) \in A_{\mathrm{cm}}$. These machines are used to increase the gas pressure, i.e., we model these machines via

$$p_v = p_u + \Delta p_a, \quad a = (u, v) \in A_{\mathrm{cm}},$$

where $\Delta p_a \geq 0$ models the controllable increase of gas pressure. In this simplified setting studied here, we do not consider other types of network elements, i.e., we have

$$A = A_{\mathrm{pi}} \cup A_{\mathrm{cm}}.$$

Finally, we need to model mass conservation at the nodes of the network. If $\delta^{\mathrm{in}}(u)$ and $\delta^{\mathrm{out}}(u)$ denote the in- and outgoing arcs of the node $u \in V$, this can be modeled by the equality constraint

$$\sum_{a \in \delta^{\mathrm{out}}(u)} q_a - \sum_{a \in \delta^{\mathrm{in}}(u)} q_a = q_u, \quad u \in V,$$

where q_u is the in- and outflow, respectively, at node u of the network. More specifically, we have

$$q_u \begin{cases} \geq 0, & u \in V_+, \\ \leq 0, & u \in V_-, \\ = 0, & u \in V_0. \end{cases}$$

With these constraints and notations at hand, a reasonable problem statement

is given by

$$\begin{aligned}
\min_{\Delta p,q,p} \quad & \sum_{a\in A_{\mathrm{cm}}} w_a \Delta p_a \\
\text{s.t.} \quad & p_v^2 = p_u^2 - \Lambda_a |q_a| q_a, \quad a = (u,v) \in A_{\mathrm{pi}}, \\
& p_u^- \le p_u \le p_u^+, \quad u \in V, \\
& p_v = p_u + \Delta p_a, \quad \Delta p_a \ge 0, \quad a = (u,v) \in A_{\mathrm{cm}}, \\
& \sum_{a\in\delta^{\mathrm{out}}(u)} q_a - \sum_{a\in\delta^{\mathrm{in}}(u)} q_a = q_u, \quad u \in V.
\end{aligned}$$

Here, the objective function models compression costs in terms of the pressure increase at all compressor stations in the network by using the cost coefficients $w_a > 0$, $a \in A_{\mathrm{cm}}$.

This is a nonlinear and nonconvex[2] optimization problem with a linear objective function.

[2] Why?

10

First-Order Optimality Conditions

In analogy to Chapter 5 we now derive and prove optimality conditions for constrained optimization problems. This will need some work and we start with the proof of a very powerful tool in optimization: the lemma of Farkas.

10.1 Farkas' Lemma

For proving the Farkas' lemma we need some technical results. First, we state a projection theorem for convex sets.

Theorem 10.1. Let $X \subseteq \mathbb{R}^n$ be a non-empty, closed, and convex set. Then, for every $y \in \mathbb{R}^n$ there exists a uniquely determined vector $z \in X$ with

$$\|y - z\| \leq \|y - x\| \quad \text{for all} \quad x \in X.$$

The vector z is called *projection* of y on X and is also denoted with $\text{Proj}_X(y)$.

Proof. Let $y \in \mathbb{R}^n$ be given. We need to show that the function

$$f(x) = \frac{1}{2}\|y - x\|^2$$

has a unique global minimum over the set X. To this end, let $w \in X$ be arbitrary. Then, $z \in X$ solves

$$\min_x \; f(x) \quad \text{s.t.} \quad x \in X$$

if and only if $z \in X$ solves the problem

$$\min_x \; f(x) \quad \text{s.t.} \quad x \in X \cap \{x \in \mathbb{R}^n \colon \|y - x\| \leq \|y - w\|\}.$$

The feasible set of the latter problem is non-empty and compact. Since the objective function is continuous, Theorem 3.3 of Weierstraß ensures the existence of a global minimizer $z \in X$. This means that

$$\|y - z\| \leq \|y - x\| \quad \text{for all} \quad x \in X$$

holds. We now show the uniqueness property by contradiction. To this end, let z^1 and z^2 be two global minimizers of the function f on X. Since X is convex, the point

$$z = \frac{1}{2}z^1 + \frac{1}{2}z^2$$

is also in X. Moreover, we have

$$\begin{aligned}
f(z) &= \frac{1}{2}\|y - z\|^2 \\
&= \frac{1}{2}\|y\|^2 - y^\top z + \frac{1}{2}\|z\|^2 \\
&= \frac{1}{2}\|y\|^2 - \frac{1}{2}y^\top z^1 - \frac{1}{2}y^\top z^2 + \frac{1}{2}\left\|\frac{1}{2}z^1 + \frac{1}{2}z^2\right\|^2 \\
&= \frac{1}{2}\|y\|^2 - \frac{1}{2}y^\top z^1 - \frac{1}{2}y^\top z^2 + \frac{1}{8}\left\|z^1\right\|^2 + \frac{1}{4}(z^1)^\top z^2 + \frac{1}{8}\left\|z^2\right\| \\
&= \frac{1}{4}\|y\|^2 - \frac{1}{2}y^\top z^1 + \frac{1}{4}\left\|z^1\right\|^2 + \frac{1}{4}\|y\|^2 - \frac{1}{2}y^\top z^2 + \frac{1}{4}\left\|z^2\right\|^2 \\
&\quad - \frac{1}{8}\left\|z^1\right\|^2 + \frac{1}{4}(z^1)^\top z^2 - \frac{1}{8}\left\|z^2\right\|^2 \\
&= \frac{1}{4}\|y - z^1\|^2 + \frac{1}{4}\left\|y - z^2\right\|^2 - \frac{1}{8}\left\|z^1 - z^2\right\|^2 \\
&= \frac{1}{2}f(z^1) + \frac{1}{2}f(z^2) - \frac{1}{8}\left\|z^1 - z^2\right\|^2.
\end{aligned}$$

Since $f(z^1) = f(z^2)$ we obtain

$$f(z) = f(z^1) - \frac{1}{8}\|z^1 - z^2\|^2.$$

As z is feasible and because z^1 is a global minimizer of f on X, the points z^1 and z^2 need to be equal. □

Theorem 10.2 (Projection theorem)**.** Let $X \subseteq \mathbb{R}^n$ be a non-empty, closed, and convex set and let $y \in \mathbb{R}^n$. Then, $z \in X$ is the projection of y on X if

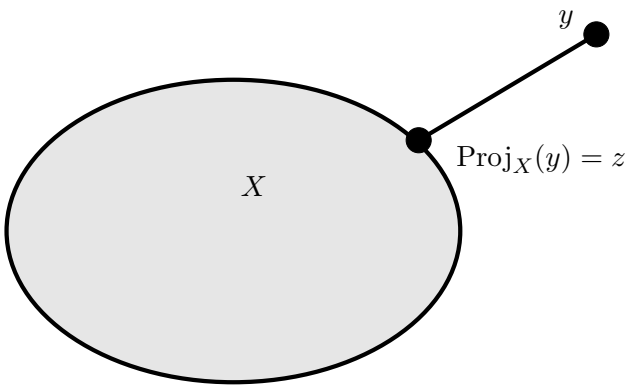


Figure 10.1: Illustration of Theorem 10.1: Projection of a vector on a non-empty, closed, and convex set

and only if

$$(z-y)^\top (x-z) \geq 0 \quad \text{for all} \quad x \in X \tag{10.1}$$

holds.

Proof. Let $y \in \mathbb{R}^n$ be given and define the function $f(x) = 1/2\|x-y\|^2$. We first assume that z is the projection of y on X. Then, $z + \lambda(x-z) \in X$ for all $x \in X$ and all $\lambda \in (0,1)$. This implies

$$f(z) \leq f(z+\lambda(x-z)) = \frac{1}{2}\|z+\lambda x - \lambda z - y\|^2 = \frac{1}{2}\|(z-y)+\lambda(x-z)\|^2,$$

which can be also written as

$$\frac{1}{2}\|z-y\|^2 \leq \frac{1}{2}\|z-y\|^2 + \lambda(z-y)^\top(x-z) + \frac{1}{2}\lambda^2\|x-z\|^2,$$

which implies

$$0 \leq \lambda(z-y)^\top(x-z) + \frac{1}{2}\lambda^2\|x-z\|^2.$$

Division by λ and taking the limit for $\lambda \searrow 0$ then yields

$$0 \leq (z-y)^\top(x-z) \quad \text{for all} \quad x \in X.$$

To prove the other implication, we assume that $z \in X$ satisfies Inequal-

ity (10.1). For an arbitrary $x \in X$ we then obtain

$$\begin{aligned} 0 &\geq (y-z)^\top (x-z) \\ &= (y-z)^\top (x-y+y-z) \\ &= \|z-y\|^2 + (y-z)^\top (x-y) \\ &\geq \|z-y\|^2 - \|y-z\|\|x-y\|, \end{aligned}$$

where we used the inequality of Cauchy and Schwarz in the last step. Moreover, the latter calculation implies

$$\|x-y\| \geq \|y-z\| \quad \text{for all} \quad x \in X,$$

i.e., z is the projection of y on X by Theorem 10.1. □

We now start to consider so-called separation theorems for which we first need the following technical lemma.

Lemma 10.3. Let $X \subseteq \mathbb{R}^n$ be a non-empty and convex set and let $\bar{x}$ be a point that is not in the interior of the set X. Then, there exists a non-zero vector $a \in \mathbb{R}^n$ such that

$$a^\top x \geq a^\top \bar{x}$$

holds for all $x \in X$.

Proof. We first consider the situation in which we impose the additional assumption that X has a non-empty interior. Note for the following that there exists a sequence $(x^k)_k$ that converges to $\bar{x}$ with $x^k \notin \mathrm{cl}(X)$ for all k. This is clear for the case that $\bar{x} \notin \mathrm{cl}(X)$ holds. If, however, $\bar{x} \in \mathrm{cl}(X)$, then $\bar{x}$ is on the boundary of X since $\bar{x} \notin \mathrm{int}(X)$. Using $\mathrm{int}(X) \neq \emptyset$, one can show that X and $\mathrm{cl}(X)$ have the same boundary and we obtain the claim. Consider now a sequence $(x^k)_k$ with the properties stated above. Let us denote with $\hat{x}^k$ the projection of x^k onto the set $\mathrm{cl}(X)$. Theorem 10.1 ensures that this projection exists since $\mathrm{cl}(X)$ is non-empty, closed, and convex. We can now use the projection theorem (Theorem 10.2) and obtain

$$(\hat{x}^k - x^k)^\top (x - \hat{x}^k) \geq 0$$

for all k and all $x \in \operatorname{cl}(X)$. This implies

$$\begin{aligned}
&(\hat{x}^k - x^k)^\top x \\
\geq\ &(\hat{x}^k - x^k)^\top \hat{x}^k \\
=\ &(\hat{x}^k - x^k)^\top (\hat{x}^k - x^k) + (\hat{x}^k - x^k)^\top x^k \\
=\ &\|\hat{x}^k - x^k\|^2 + (\hat{x}^k - x^k)^\top x^k \\
\geq\ &(\hat{x}^k - x^k)^\top x^k.
\end{aligned}$$

Using the notation

$$a_k := \frac{\hat{x}^k - x^k}{\|\hat{x}^k - x^k\|},$$

the above inequality can be re-written as[1]

$$a_k^\top x \geq a_k^\top x^k \tag{10.2}$$

for all k and all $x \in \operatorname{cl}(X)$. Since $\|a_k\| = 1$ holds, there exists a subsequence $(a_k)_{k \in K}$ that converges to a vector $a \neq 0$. Taking the limit $k \to_K \infty$, (10.2) leads to

$$a^\top x \geq a^\top \bar{x}$$

for all $x \in \operatorname{cl}(X)$. This is what we wanted to prove.

It remains to consider the case in which $\operatorname{int}(X) = \emptyset$. In this case, the arguments from above cannot be applied because X and $\operatorname{cl}(X)$ do not need to have the same boundary. If $\bar{x} \notin \operatorname{cl}(X)$, we can still use the proof from above since the argument of the coinciding boundaries is not used. Thus, it remains to consider the case $\bar{x} \in \operatorname{cl}(X)$. To this end, we consider the affine hull of X, i.e., the intersection of all affine subspaces of $\mathbb{R}^n$ that contain X. It is easy to see that the dimension of this affine hull of X is at most $n-1$ since $\operatorname{int}(X) = \emptyset$. Hence, X is a subset of an $(n-1)$-dimensional affine subspace of $\mathbb{R}^n$. In other words, X is a subset of a hyperplane

$$H = \left\{ x \in \mathbb{R}^n \colon a^\top x = \gamma \right\}$$

for some $a \in \mathbb{R}^n \setminus \{0\}$ and some $\gamma \in \mathbb{R}$. Using $\bar{x} \in \operatorname{cl}(X)$, we also obtain

$$a^\top x = \gamma = a^\top \bar{x}$$

for all $x \in X$, which completes the proof. □

With this technical lemma at hand, we can now state and proof the following separation theorem.

[1] Note that the denominator is never zero because $x^k \notin \operatorname{cl}(X)$.

Theorem 10.4 (Separation theorem)**.** Let $X_1 \subseteq \mathbb{R}^n$ and $X_2 \subseteq \mathbb{R}^n$ be two non-empty, disjoint, and convex sets. Then, there exists a non-zero vector $0 \neq a \in \mathbb{R}^n$ with

$$a^\top x_1 \leq a^\top x_2$$

for all $x_1 \in X_1$ and all $x_2 \in X_2$.

Proof. We consider the set

$$X := X_2 - X_1 := \{x \in \mathbb{R}^n \colon x = x_2 - x_1 \text{ for } x_1 \in X_1,\, x_2 \in X_2\}\,.$$

One can show that X is non-empty and convex.[2] Since the sets X_1 and X_2 are disjoint, we have $0 \notin X$. This particularly means that 0 is also not in the interior of X. Hence, Lemma 10.3 ensures the existence of a non-zero vector $a \in \mathbb{R}^n$ with $0 \leq a^\top x$ for all $x \in X$. Using the definition of X then implies

$$a^\top x_1 \leq a^\top x_2$$

for all $x_1 \in X_1$ and $x_2 \in X_2$. □

Note that, in general, the sum of two closed and convex sets does not need to be closed again. If it is required that the sum is a closed set as well, one needs one more assumption for one of the two original sets.

Lemma 10.5. Let $X_1, X_2 \subseteq \mathbb{R}^n$ be two non-empty and convex sets. Moreover, X_1 is closed and X_2 is compact. Then, the sum

$$X_1 + X_2 = \{x \in \mathbb{R}^n \colon \exists x_1 \in X_1 \text{ and } x_2 \in X_2 : x = x_1 + x_2\}$$

is a non-empty, closed, and convex set.

Proof. It is easy to see that the set $X = X_1 + X_2$ is non-empty and convex. Thus, it remains to show that the set is closed. To this end, we consider a converging sequence $(x_1^k + x_2^k)_k \subseteq X$ with $x_1^k \in X_1$ and $x_2^k \in X_2$ for all k. Since X_2 is compact, we know that the sequence $(x_2^k)_k$ is bounded. Consequently, the sequence $(x_1^k)_k$ is bounded as well because $(x_1^k + x_2^k)_k$ converges. This implies the existence of subsequences $(x_1^k)_{k\in K}$ and $(x_2^k)_{k\in K}$ that converge to some points x_1 and x_2. Thus, the sequence $(x_1^k + x_2^k)_{k\in K}$ converges to $x_1 + x_2$. The sets X_1 and X_2 are both closed, which finally leads to $x_1 + x_2 \in X_1 + X_2 = X$. Hence, X is closed. □

Theorem 10.6 (Strict separation theorem)**.** Let $X_1 \subseteq \mathbb{R}^n$ and $X_2 \subseteq \mathbb{R}^n$ two non-empty, disjoint, and convex sets. Moreover, suppose that X_1 is closed

[2] We will prove this in the exercise classes.

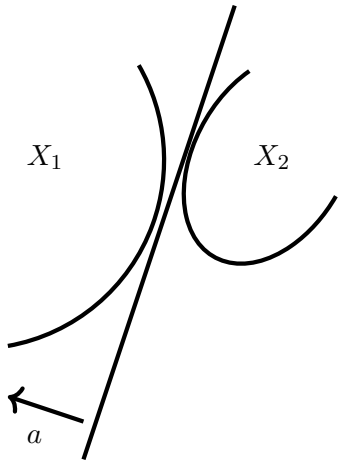


Figure 10.2: Illustration of the strict separation theorem 10.6

and that X_2 is compact. Then, there exists a non-zero vector $0 \neq a \in \mathbb{R}^n$ and a scalar $\beta \in \mathbb{R}$ with

$$a^\top x_1 < \beta < a^\top x_2 \quad \text{for all} \quad x_1 \in X_1,\ x_2 \in X_2. \tag{10.3}$$

Proof. We consider the optimization problem

$$\min_{x_1, x_2} \|x_1 - x_2\| \quad \text{s.t.} \quad x_1 \in X_1,\ x_2 \in X_2. \tag{10.4}$$

The set $X = X_1 - X_2 = \{x_1 - x_2 \colon x_1 \in X_2,\, x_2 \in X_2\}$ is non-empty and convex. Due to Lemma 10.5, the set X is closed because X_2 is compact.

Problem (10.4) is the problem to project the zero vector on the set X. Thus, Theorem 10.1 implies that Problem (10.4) is solvable. Let (x_1^*, x_2^*) now be a solution of (10.4) and define

$$a = \frac{x_2^* - x_1^*}{2}, \quad x^* = \frac{x_1^* + x_2^*}{2}, \quad \beta = a^\top x^*.$$

It holds $a \neq 0$ because $x_1^* \in X_1$, $x_2^* \in X_2$, and $X_1 \cap X_2 = \emptyset$. Moreover, one can see (by using (10.4)) that x_1^* is the projection of x^* on X_1 and that x_2^* is the projection of x^* on X_2; see also Figure 10.3. Thus, the projection theorem (Theorem 10.2) leads to

$$(x^* - x_1^*)^\top (x_1 - x_1^*) \leq 0 \quad \text{for all} \quad x_1 \in X_1.$$

From $x^* - x_1^* = a$ it then follows

$$a^\top x_1 \leq a^\top x_1^* = a^\top x^* + a^\top (x_1^* - x^*) = \beta - \|a\|^2 < \beta$$

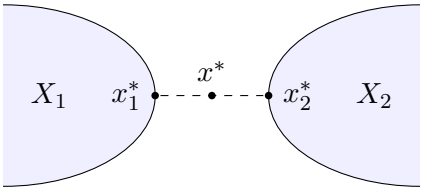


Figure 10.3: Illustration of the proof of the strict separation theorem

for all $x_1 \in X_1$. This shows the left-hand side of Inequality (10.3). The right-hand side can be shown analogously. □

We now only need one more definition and one more technical lemma before we can state and prove the famous Farkas' lemma.

Definition 10.7 (Cone)**.** A subset $X \subseteq \mathbb{R}^n$ is called a *cone* if

$$\lambda x \in X$$

for all $x \in X$ and all $\lambda > 0$.

We say that a cone X is a *closed convex cone* if X is a closed and convex subset of $\mathbb{R}^n$.

Let $a_1, \dots, a_m \in \mathbb{R}^n$ be given vectors. The set

$$\operatorname{cone}\{a_1, \dots, a_m\} = \{x_1 a_1 + \cdots + x_m a_m \colon x_1, \dots, x_m \geq 0\}$$

is a cone[3] in $\mathbb{R}^n$. It is called the *cone induced by the vectors* $a_1, \dots, a_m$.

Now, we state and prove the last technical result before we can state and prove the Farkas' lemma.

Lemma 10.8. Let $A \in \mathbb{R}^{m \times n}$ be given. Then, the set

$$X = \{y \in \mathbb{R}^n \colon y = A^\top x,\, x \geq 0\}$$

is a non-empty, closed, and convex cone.

[3]This will be shown in the exercise classes.

Proof. It is rather easy to see that X is a non-empty and convex cone. Hence, it remains to prove that X is closed. Let us denote the columns of $A^\top$ with $a_1, \dots, a_m$. With this, we can re-write X as

$$X = \{y \in \mathbb{R}^n \colon y = x_1 a_1 + \dots + x_m a_m,\, x \geq 0\} =: \operatorname{cone}\{a_1, \dots, a_m\},$$

i.e., X is the cone induced by the m vectors $a_1, \dots, a_m$. We now proceed by induction over m. For $m = 1$, we have

$$X = \{y \in \mathbb{R}^n \colon y = x_1 a_1,\, x_1 \geq 0\}.$$

This is a closed set. Thus, we now consider $m \geq 1$ and assume that every cone that is induced by less than m vectors is closed. Let now $(y^k)_k \subseteq X$ be an arbitrary sequence that converges to a point $y^* \in \mathbb{R}^n$. We have to show that $y^* \in X$ holds. Since $y^k \in X$, we obtain

$$y^k = \sum_{i=1}^{m} \lambda_i^k a_i$$

for all k and for suitably chosen $\lambda_i^k \geq 0$ with $i = 1, \dots, m$. Moreover, we know that

$$V = \left\{y \in \mathbb{R}^n \colon A^\top x,\, x \in \mathbb{R}^m\right\}$$

is a finite-dimensional and thus closed vector space with $y^k \in X \subseteq V$. Hence, the limit y^* is at least in V, which means that

$$y^* = \sum_{i=1}^{m} \alpha_i a_i$$

for suitably chosen $\alpha_i \in \mathbb{R}$ and $i = 1, \dots, m$. If all α_i are non-negative, we are done. We thus assume that there exists an index $i_0 \in \{1, \dots, m\}$ with $\alpha_{i_0} < 0$ and then define

$$\beta_k := \min\left\{\frac{\lambda_i^k}{\lambda_i^k - \alpha_i} \colon i \in \{1, \dots, m\} \text{ with } \alpha_i < 0\right\}$$

for all k. It holds $0 \leq \beta_k \leq 1$ and $\beta_k(\lambda_i^k - \alpha_i) \leq \lambda_i^k$ if $\alpha_i < 0$. This implies

$$r_{ik} := \beta_k \alpha_i + (1 - \beta_k)\lambda_i^k \geq 0 \tag{10.5}$$

for all k and all $i = 1, \dots, m$. Moreover, we define

$$\begin{aligned} z^k &:= y^k + \beta_k(y^* - y^k) \\ &= \beta_k y^* + (1 - \beta_k) y^k \\ &= \beta_k \sum_{i=1}^{m} \alpha_i a_i + (1 - \beta_k) \sum_{i=1}^{m} \lambda_i^k a_i \\ &= \sum_{i=1}^{m} \beta_k \alpha_i a_i + \sum_{i=1}^{m} (1 - \beta_k) \lambda_i^k a_i \\ &= \sum_{i=1}^{m} \left(\beta_k \alpha_i + (1 - \beta_k) \lambda_i^k \right) a_i \\ &= \sum_{i=1}^{m} r_{ik} a_i \in [y^*, y^k]. \end{aligned}$$

It holds $z^k \to y^*$ due to $y^k \to y^*$. Moreover, (10.5) implies that $z^k \in X$ holds. Let now i_k for all k be an index for which the minimum is attained in the definition of β_k, i.e., we have

$$\beta_k = \frac{\lambda_{i_k}^k}{\lambda_{i_k}^k - \alpha_{i_k}}.$$

Hence,

$$\begin{aligned} r_{i_k,k} &= \beta_k \alpha_{i_k} + (1 - \beta_k) \lambda_{i_k}^k \\ &= \frac{\lambda_{i_k}^k \alpha_{i_k}}{\lambda_{i_k}^k - \alpha_{i_k}} + \left(1 - \frac{\lambda_{i_k}^k}{\lambda_{i_k}^k - \alpha_{i_k}} \right) \lambda_{i_k}^k \\ &= \frac{\lambda_{i_k}^k \alpha_{i_k} + \left(\lambda_{i_k}^k - \alpha_{i_k} \right) \lambda_{i_k}^k - (\lambda_{i_k}^k)^2}{\lambda_{i_k}^k - \alpha_{i_k}} \\ &= 0, \end{aligned}$$

which shows that z^k is contained in the cone that is induced by the $m - 1$ vectors $\{a_1, \dots, a_m\} \setminus \{a_{i_k}\}$. By considering subsequences it can, w.l.o.g., be assumed that there is an index j (that is independent of k) so that z^k is contained in the cone induced by $\{a_1, \dots, a_m\} \setminus \{a_j\}$. This cone is closed by the induction hypothesis. Hence, the accumulation point y^* of the sequence $(z^k)_k$ is contained in the cone induced by $\{a_1, \dots, a_m\} \setminus \{a_j\}$ as well. Finally, this implies that y^* is part of the cone $X = \operatorname{cone}\{a_1, \dots, a_m\}$. □

Now, we state and prove one of the most famous results in optimization.

Lemma 10.9 (Farkas' lemma)**.** Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^n$ be given. Then, the following two statements are equivalent:

(a) The system $A^\top x = b$, $x \geq 0$ has a solution.

(b) The inequality $b^\top d \geq 0$ holds for all $d \in \mathbb{R}^n$ with $Ad \geq 0$.

Proof. Assume that the first statement is true. This means that there exists an $0 \leq x \in \mathbb{R}^m$ with

$$b^\top d = (A^\top x)^\top d = x^\top Ad \geq 0$$

for all $d \in \mathbb{R}^n$ with $Ad \geq 0$. Thus, the second statement is true as well.

We now prove the other implication by contraposition. Thus, assume that the first statement is false, which means that b is not in the set

$$X = \{y \in \mathbb{R}^n : y = A^\top x,\, x \geq 0\}.$$

The last lemma implies that X is a non-empty, closed, and convex cone. Using the strict separation theorem (Theorem 10.6), it follows the existence of a non-zero vector $a \in \mathbb{R}^n$ and a scalar $\beta \in \mathbb{R}$ with

$$a^\top y > \beta > a^\top b \tag{10.6}$$

for all $y \in X$. This implies

$$a^\top y \geq 0 > a^\top b.$$

The second (strict) inequality follows from (10.6) and $0 \in X$. The first inequality follows from also from (10.6) by additionally using $\lambda y \in X$ for $\lambda \to \infty$ and $y \in X$.[4] If we now choose $x = e_i$ for $i \in \{1, \dots, m\}$ in the definition of the cone X, we obtain

$$a^\top a_i \geq 0 > a^\top b,$$

where $a_1, \dots, a_m$ are the column vectors of the matrix $A^\top$. Thus, we obtain $Aa \geq 0$ but also $a^\top b < 0$. This means that the second statement is false. □

10.2 The Tangential Cone and the Linearized Tangential Cone

Let us start with the definition of the so-called tangential cone.

[4] Why?

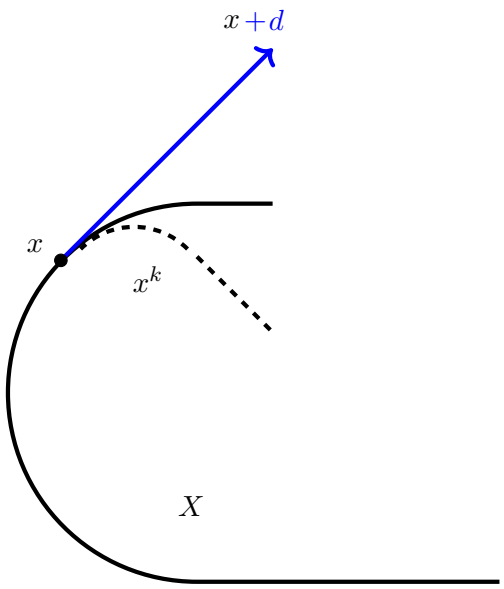


Figure 10.4: Illustration of a tangential direction

Definition 10.10 (Tangential cone)**.** Let $X \subseteq \mathbb{R}^n$ be a non-empty set. A vector $d \in \mathbb{R}^n$ is called *tangential* to X in the point $x \in X$, if sequences $(x^k)_k \subseteq X$ and $(t^k)_k \subseteq \mathbb{R}$ exist with

$$x^k \to x, \quad t^k \searrow 0, \quad \text{and} \quad \frac{x^k - x}{t^k} \to d$$

for $k \to \infty$. The set of all such directions is called the *tangential cone*[5] of X at the point $x \in X$ and is denoted with $T_X(x)$, i.e.

$$T_X(x) = \{d \in \mathbb{R}^n : \exists (x^k)_k \subseteq X, (t^k)_k \subseteq \mathbb{R} : t^k \searrow 0, x^k \to x, (x^k - x)/t^k \to d\}.$$

Lemma 10.11. Let $X \subseteq \mathbb{R}^n$ be a non-empty set. Then, the tangential cone is a cone.

Proof. This will be shown in the exercises. □

Using the tangential cone, we can state and prove our first necessary condition for constrained optimization problems.

Theorem 10.12. Let $X \subseteq \mathbb{R}^n$ be non-empty and suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Moreover, let $x^* \in X$ be a local minimizer of

[5] Sometimes this tangential cone is also called the *Bouligand tangential cone*.

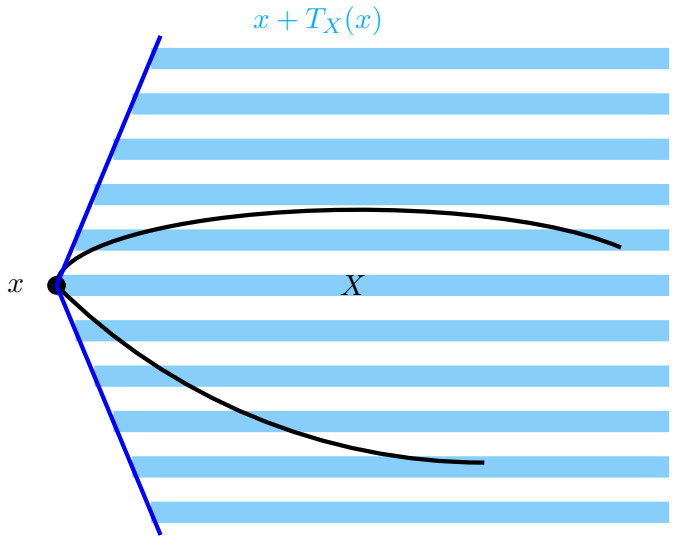


Figure 10.5: Illustration of the tangential cone (moved to the point x)

the optimization problem

$$\min_{x \in \mathbb{R}^n} \; f(x) \quad \text{s.t.} \quad x \in X. \tag{10.7}$$

Then,

$$\nabla f(x^*)^\top d \geq 0 \quad \text{for all} \quad d \in T_X(x^*) \tag{10.8}$$

holds.

Proof. Let $d \in T_X(x^*)$ be arbitrarily given. This means, by definition, that there exist sequences $(t^k)_k \searrow 0$ and $(x^k)_k \to x^*$ with

$$d = \lim_{k \to \infty} \frac{x^k - x^*}{t^k}.$$

Since f is continuously differentiable, the mean value theorem of differential calculus implies

$$f(x^k) - f(x^*) = \nabla f(\xi^k)^\top (x^k - x^*)$$

for all k, where ξ^k is a point on the line segment between x^k and x^*. In particular, $\xi^k \to x^*$ holds for $k \to \infty$ because $x^k \to x^*$ holds. For k sufficiently large, we have

$$f(x^k) - f(x^*) \geq 0$$

since x^* is a local minimizer. Thus, we also obtain

$$\nabla f(\xi^k)^\top (x^k - x^*) \geq 0$$

for sufficiently large k. Division by t^k and taking the limit $k \to \infty$ thus leads to

$$\nabla f(x^*)^\top d \geq 0. \qquad \square$$

Definition 10.13 (Stationary point)**.** A point $x^* \in \mathbb{R}^n$ that is feasible for Problem (10.7) is called *stationary* if it satisfies Condition (10.8).[6]

Note that the latter definition coincides with Definition 1.6 in the unconstrained case. This means that the two definitions are the same for $X = \mathbb{R}^n$.

Let us make one more comment regarding the stationarity condition (10.8). In other words, this condition says that for a local minimizer $x^* \in \mathbb{R}^n$ of Problem (10.7), there is no direction $d \in \mathbb{R}^n$, which is tangential to X in x^*, so that

$$\nabla f(x^*)^\top d < 0$$

holds. Thus, there is no tangential descent direction.

From now on, we consider the situation in which the feasible set X is given by a finite number of equality and inequality constraints, i.e., we consider the problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) && \text{(10.9a)} \\ \text{s.t.} \quad & g_i(x) \geq 0, \quad i \in I = \{1, \dots, m\}, && \text{(10.9b)} \\ & h_j(x) = 0, \quad j \in J = \{1, \dots, p\}. && \text{(10.9c)} \end{aligned}$$

Moreover, we assume that the objective function f as well as the constraint functions g_i, $i \in I$, and h_j, $j \in J$, are continuously differentiable.

Theorem 10.12 already states a first-order necessary condition for Problem (10.9). Unfortunately, the tangential cone is typically hard to deal with in practice. This is why we define another cone, which is easier to handle. To this end, we first need to introduce the so-called active set.

Definition 10.14 (Active inequality constraints)**.** Let $x \in X$ be a feasible point of Problem (10.9). Then, the set

$$I(x) := \{i \in I \colon g_i(x) = 0\}$$

is called the *set of active inequality constraints* at the point x.

Definition 10.15 (Linearized tangential cone)**.** Let $x \in X$ be a feasible

[6] Sometimes these stationary points are also called *B-* or *Bouligand stationary* points.

point of Problem (10.9). Then,

$$T_{\text{lin}}(x) := \{d \in \mathbb{R}^n \colon \nabla g_i(x)^\top d \geq 0, i \in I(x),\ \nabla h_j(x)^\top d = 0, j \in J\}$$

is called the *linearized tangential cone* of X at the point x.

Remark 10.16. Note that the tangential cone only depends on the geometry of the feasible set, whereas the linearized tangential cone depends on the specific algebraic description of the feasible set. The latter is not unique. For example, $x = 0$ and $x^2 = 0$ describe the same set of points but obviously differ in their algebraic description.

A key aspect for the derivation of optimality conditions for constrained optimization is the relation between the tangential cone and the linearized tangential cone. The first relation is given by the next lemma.

Lemma 10.17. Let $x \in X$ be a feasible point of Problem (10.9). Then, $T_X(x) \subseteq T_{\text{lin}}(x)$ holds.

Proof. Let $d \in T_X(x)$ be given arbitrarily. Thus, by definition, there exist sequences $t^k \searrow 0$ and $x^k \to x$ with

$$d = \lim_{k \to \infty} \frac{x^k - x}{t^k}.$$

We first prove that $\nabla g_i(x^k)^\top d \geq 0$ holds for all $i \in I(x)$. Using the mean value theorem of differential calculus, we obtain

$$0 \leq g_i(x^k) = g_i(x) + \nabla g_i(\xi^k)^\top (x^k - x) = \nabla g_i(\xi^k)^\top (x^k - x),$$

where ξ^k is a point on the line segment between x^k and x. We have that $\xi^k \to x$ for $k \to \infty$ because $x^k \to x$ holds. Division by t^k and taking the limit $k \to \infty$ leads to

$$0 \leq \nabla g_i(x)^\top d.$$

The second part, i.e., $\nabla h_j(x)^\top d = 0$ for $j \in J$, can be shown analogously. □

10.3 Constraint Qualifications and KKT Theorems

In this section, we consider so-called *constraint qualifications* that ensure that the tangential cone and the linearized tangential cone coincide in a given feasible point of the problem. Let us start with the most basic constraint qualification.

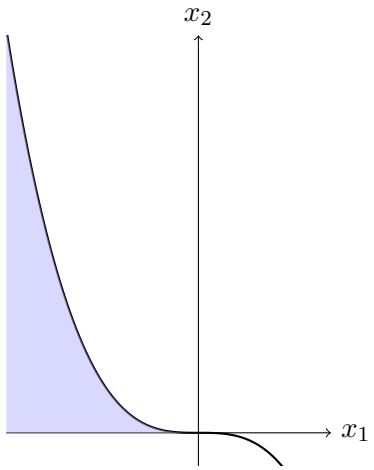


Figure 10.6: The feasible set for the constraints $-x_2 - x_1^3 \geq 0$, $x_2 \geq 0$

Definition 10.18 (Abadie constraint qualification)**.** We say that a feasible point x of Problem (10.9) satisfies the *Abadie constraint qualification (ACQ)* if $T_X(x) = T_{\text{lin}}(x)$ holds.

We will later derive conditions under which the ACQ holds at a considered point. The next example, however, shows that this is not always the case.

Example 10.19. Consider the problem

$$\min_{x \in \mathbb{R}^2} \; -x_1 \quad \text{s.t.} \quad -x_2 - x_1^3 \geq 0, \; x_2 \geq 0.$$

The feasible set is illustrated in Figure 10.6. It can be easily seen that $x^* = (0,0)^\top \in \mathbb{R}^2$ is the uniquely determined minimizer. We now define

$$g_1(x) := -x_2 - x_1^3, \quad g_2(x) := x_2$$

and see that both inequalities are active at x^*, i.e. $I(x^*) = \{1,2\}$. The linearized tangential cone is thus given by

$$T_{\text{lin}}(x^*) = \{d = (d_1, d_2)^\top \in \mathbb{R}^2 \colon d_2 = 0\}.$$

However, from the geometry of the feasible set it is clear that the tangential cone is given by

$$T_X(x^*) = \{d = (d_1, d_2)^\top \in \mathbb{R}^2 \colon d_2 = 0, \, d_1 \leq 0\}.$$

Thus, the tangential cone is a proper subset of the linearized tangential cone in this example.

For what follows, the so-called Lagrangian function plays a very important role.

Definition 10.20 (Lagrangian function)**.** The function

$$\mathcal{L}(x, \lambda, \mu) := f(x) - \sum_{i=1}^{m} \lambda_i g_i(x) - \sum_{j=1}^{p} \mu_j h_j(x)$$

is called *Lagrangian function* of Problem (10.9).

Using the Lagrangian function we can now define the Karush–Kuhn–Tucker (KKT) conditions.

Definition 10.21 (KKT conditions, KKT point, Lagrangian multipliers)**.** We consider Problem (10.9) with continuously differentiable functions f, g, and h.[7]

(a) The conditions

$$\begin{aligned} \nabla_x \mathcal{L}(x, \lambda, \mu) &= 0, \\ h(x) &= 0, \\ \lambda \geq 0,\ g(x) \geq 0,\ \lambda^\top g(x) &= 0 \end{aligned}$$

are called *Karush–Kuhn–Tucker* (or *KKT*) *conditions* of Problem (10.9). Here and in what follows,

$$\nabla_x \mathcal{L}(x, \lambda, \mu) = \nabla f(x) - \sum_{i=1}^{m} \lambda_i \nabla g_i(x) - \sum_{j=1}^{p} \mu_j \nabla h_j(x)$$

is the gradient of the Lagrangian function with respect to the variables x.

(b) Every vector $((x^*)^\top, (\lambda^*)^\top, (\mu^*)^\top)^\top \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ that satisfies the KKT conditions is called a *KKT point* of Problem (10.9). The components of λ^* and μ^* are called *Lagrangian multipliers*.

Remark 10.22. (a) Note that the KKT conditions reduce to $\nabla f(x) = 0$ in the case of an unconstrained optimization problem; see Theorem 5.1. Thus, it seems natural that the KKT conditions will also serve as first-order necessary optimality conditions in the case of constrained optimization problems.

[7] A quantity without the index usually stands for the vector; e.g., $h(x) = (h_j(x))_{j=1,\dots,p}$.

(b) The conditions

$$\lambda \geq 0,\ g(x) \geq 0,\ \lambda^\top g(x) = 0$$

are obviously equivalent to

$$\lambda_i \geq 0,\ g_i(x) \geq 0,\ \lambda_i g_i(x) = 0 \quad \text{for all} \quad i = 1, \dots, m.$$

Thus, for every KKT point $((x^*)^\top, (\lambda^*)^\top, (\mu^*)^\top)^\top \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ it holds

$$\lambda_i^* = 0 \quad \text{or} \quad g_i(x^*) = 0 \quad \text{for all} \quad i = 1, \dots, m.$$

If, in addition,

$$\lambda_i^* + g_i(x^*) > 0 \quad \text{for all} \quad i = 1, \dots, m$$

holds, i.e., if no index $i_0 \in \{1, \dots, m\}$ exists with

$$\lambda_{i_0}^* = 0 \quad \text{and} \quad g_{i_0}(x^*) = 0,$$

we say that the KKT point satisfies the *strict complementarity condition*.

(c) The KKT conditions also have a very nice geometric interpretation. Consider the problem

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad g_i(x) \geq 0,$$

which only has inequality constraints. Moreover, let $((x^*)^\top, (\lambda^*)^\top)^\top$ be a KKT point and let $I(x^*)$ be the set of active inequalities at x^*. Then, it holds

$$\nabla f(x^*) = \sum_{i \in I(x^*)} \lambda_i^* \nabla g_i(x^*)$$

because $\lambda_i^* = 0$ for all $i \notin I(x^*)$. Thus, the gradient of the objective function at a KKT point is in the cone induced by the gradients of the active inequality constraints, i.e.,

$$\nabla f(x^*) \in \operatorname{cone}\{\nabla g_i(x^*) \colon i \in I(x^*)\};$$

see also Figure 10.7.

Now, we are able to state the relation between local minimizers of a constrained optimization problem and KKT points.

Theorem 10.23 (KKT conditions under the Abadie CQ)**.** Let $x^* \in \mathbb{R}^n$ be a local minimizer of Problem (10.9). Moreover, suppose that the Abadie CQ

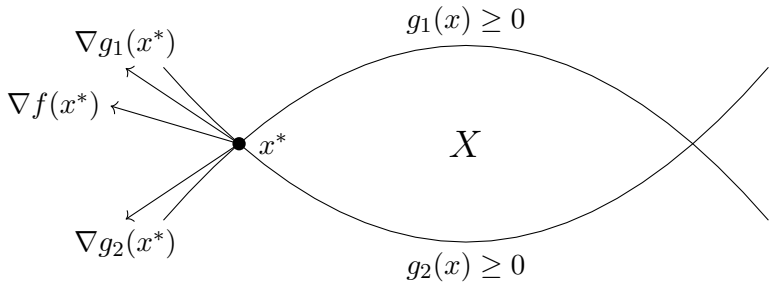


Figure 10.7: Geometric illustration of the KKT conditions

holds at x^*. Then, there exist Lagrangian multipliers $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ so that $((x^*)^\top, (\lambda^*)^\top, (\mu^*)^\top)^\top$ is a KKT point of Problem (10.9).

Proof. Since x^* is a local minimizer of Problem (10.9), we have

$$\nabla f(x^*)^\top d \geq 0 \quad \text{for all} \quad d \in T_X(x^*)$$

by Theorem 10.12. Using the Abadie CQ, this means that

$$\nabla f(x^*)^\top d \geq 0$$

holds for all vectors $d \in \mathbb{R}^n$ with $Ad \geq 0$, where A is the matrix with row vectors

$$\begin{aligned} \nabla g_i(x^*)^\top \quad &\text{for all} \quad i \in I(x^*), \\ \nabla h_j(x^*)^\top \quad &\text{for all} \quad j = 1, \ldots, p, \\ -\nabla h_j(x^*)^\top \quad &\text{for all} \quad j = 1, \ldots, p. \end{aligned}$$

Using the Farkas' lemma (Lemma 10.9 with $b = \nabla f(x^*)$) then implies that the system

$$A^\top y = \nabla f(x^*), \quad y \geq 0$$

has a solution y^*. We now denote the components of the vector y with λ_i^* for all $i \in I(x^*)$ and μ_j^+ as well as μ_j^- for all $j = 1, \ldots, p$. Then, setting $\mu_j^* := \mu_j^+ - \mu_j^-$ and $\lambda_i^* := 0$ for all $i \notin I(x^*)$ yields a KKT point $((x^*)^\top, (\lambda^*)^\top, (\mu^*)^\top))^\top$. □

In the remainder of this section we will show that the KKT conditions are also necessary optimality conditions for general, i.e., possibly nonlinear and nonconvex, optimization optimization problems under other constraint qualifications that are easier to handle than the Abadie CQ. To this end, we need the following technical result.

Lemma 10.24. Let $x^* \in \mathbb{R}^n$ be a feasible point of Problem (10.9) and let $I(x^*)$ be the set of active inequality constraints at x^*. Moreover, suppose that the gradients $\nabla h_j(x^*)$, $j = 1, \dots, p$, of all equality constraints are linearly independent and that there exists a vector $d \in \mathbb{R}^n$ with $\nabla h_j(x^*)^\top d = 0$ for all $j = 1, \dots, p$ and $\nabla g_i(x^*)^\top d > 0$ for all $i \in I(x^*)$. Then, there exists an $\varepsilon > 0$ and a curve $x : (-\varepsilon, +\varepsilon) \to \mathbb{R}^n$ with the following properties:

(a) x is continuously differentiable on $(-\varepsilon, +\varepsilon)$.

(b) $x(t) \in X$ for all $t \in [0, +\varepsilon)$, i.e., the curve is in the feasible set X of Problem (10.9) for all $t \geq 0$.

(c) $x(0) = x^*$, i.e., the curve "starts" in the given feasible point x^*.

(d) $x'(0) = d$, i.e., the curve is tangential to the vector d at $t = 0$.

Proof. First, we show that there exists an $\varepsilon > 0$ and a continuously differentiable curve $x : (-\varepsilon, +\varepsilon) \to \mathbb{R}^n$ with $x(0) = x^*$, $x'(0) = d$, and $h_j(x(t)) = 0$ for all $j = 1, \dots, p$ and all $t \in (-\varepsilon, +\varepsilon)$. To this end, we define the mapping $H : \mathbb{R}^{p+1} \to \mathbb{R}^p$ via

$$H_j(y, t) := h_j(x^* + td + h'(x^*)^\top y)$$

for all $j = 1, \dots, p$. Here, $h'(x^*) \in \mathbb{R}^{p \times n}$ is the Jacobian matrix of h at x^*. The nonlinear system of equations

$$H(y, t) = 0 \tag{10.10}$$

has the solution $(y^*, t^*) = (0, 0)$. The Jacobian at this point is given by

$$H'_y(0, 0) = h'(x^*)h'(x^*)^\top \in \mathbb{R}^{p \times p}.$$

By assumption, we know that $h'(x^*)$ has full rank. Hence, the square matrix $H'_y(0, 0)$ is regular and we can apply the implicit function theorem to the system in (10.10) at the point $(y^*, t^*) = (0, 0)$. Thus, there exists an $\varepsilon > 0$ as well as a continuously differentiable function $y : (-\varepsilon, +\varepsilon) \to \mathbb{R}^p$ with $y(0) = 0$ and $H(y(t), t) = 0$ for all $t \in (-\varepsilon, +\varepsilon)$. The implicit function theorem further implies

$$y'(t) = -(H'_y(y(t), t))^{-1} H'_t(y(t), t)$$

for all $t \in (-\varepsilon, +\varepsilon)$. Thus, we get

$$y'(0) = -(H'_y(0, 0))^{-1} H'_t(0, 0) = -(H'_y(0, 0))^{-1} h'(x^*)d = 0$$

due to $\nabla h_j(x^*)^\top d = 0$ for all $j = 1, \dots, p$. We now define the curve

$x : (-\varepsilon, +\varepsilon) \to \mathbb{R}^n$ via

$$x(t) := x^* + td + h'(x^*)^\top y(t)$$

and show that this curve has all the desired properties. First of all, it holds $x(0) = x^*$, $x'(0) = d + h'(x^*)^\top y'(0) = d$, and from $H(y(t), t) = 0$ for all $t \in (-\varepsilon, +\varepsilon)$ we get $h_j(x(t)) = 0$ for all $j = 1, \ldots, p$ and all $t \in (-\varepsilon, +\varepsilon)$.

Finally, we show that the curve is also feasible w.r.t. the inequality constraints. By continuity arguments, we have $g_i(x(t)) > 0$ for all $i \notin I(x^*)$ and all t sufficiently close to 0. We thus consider an index $i \in I(x^*)$ and define $\phi(t) := g_i(x(t))$. The chain rule implies $\phi'(t) = \nabla g_i(x(t))^\top x'(t)$. The already known properties of $x(t)$ and the assumptions regarding d thus imply $\phi'(0) = \nabla g_i(x^*)^\top d > 0$, which again implies $\phi(t) = g_i(x(t)) > 0$ for all sufficiently small $t > 0$. Hence, $x(t)$ is a continuously differentiable curve that also stays feasible w.r.t. the inequality constraints for all sufficiently small $t > 0$. □

The assumptions of the last lemma now directly lead to another constraint qualification.

Definition 10.25 (Mangasarian–Fromowitz CQ)**.** Let $x \in \mathbb{R}^n$ be a feasible point of Problem (10.9) and let $I(x)$ be the set of active inequality constraints. We say that the *Mangasarian–Fromowitz CQ (MFCQ)* is satisfied at the point x if the following two conditions hold:

(a) The gradients

$$\nabla h_j(x) \quad \text{for all} \quad j = 1, \ldots, p$$

are linearly independent.

(b) There exists a vector $d \in \mathbb{R}^n$ with

$$\begin{aligned} \nabla g_i(x)^\top d &> 0 \quad \text{for all} \quad i \in I(x), \\ \nabla h_j(x)^\top d &= 0 \quad \text{for all} \quad j = 1, \ldots, p. \end{aligned}$$

Remark 10.26. The existence of a continuously differentiable curve with the properties as stated in Lemma 10.24 sometimes is called the *constraint qualification of Kuhn–Tucker*. Thus, the lemma states that the MFCQ implies the Kuhn–Tucker CQ. Next, we show that the KKT conditions are also necessary optimality conditions under the MFCQ.

However, we first need one more property of the tangential cone.

Lemma 10.27. Let $X \subseteq \mathbb{R}^n$ be a non-empty set and let $x \in X$. Then, the tangential cone $T_X(x)$ is closed.

Proof. Let $(d^k)_k \subseteq T_X(x)$ be a sequence that converges to $d \in \mathbb{R}^n$. We need to show that $d \in T_X(x)$ holds. Then, for every k, there exists sub-sequences $(x^{k_l})_l \subseteq X$ and $(t^{k_l})_l \subseteq \mathbb{R}$ with

$$x^{k_l} \to x, \quad t^{k_l} \searrow 0, \quad \text{and} \quad \frac{x^{k_l} - x}{t^{k_l}} \to d^k$$

for $l \to \infty$. Thus, for every $k \in \mathbb{N}$, there exists an index $l(k)$ with

$$\|x^{k_{l(k)}} - x\| \leq \frac{1}{k}, \quad t^{k_{l(k)}} \leq \frac{1}{k}, \quad \text{and} \quad \left\| \frac{x^{k_{l(k)}} - x}{t^{k_{l(k)}}} - d^k \right\| \leq \frac{1}{k}.$$

For $k \to \infty$, we thus obtain sequences $(x^{k_{l(k)}})_k$ and $(t^{k_{l(k)}})_k$ that verify that $d \in T_X(x^*)$ holds. Hence, $T_X(x^*)$ is closed. □

Theorem 10.28 (KKT conditions under the MFCQ)**.** Let x^* be a local minimizer of Problem (10.9) that satisfies the MFCQ. Then, there exist Lagrangian multipliers $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ so that (x^*, λ^*, μ^*) is a KKT point of Problem (10.9).

Proof. By Theorem 10.23 it is enough to show that the MFCQ implies the ACQ. That is, we need to show $T_X(x^*) = T_{\text{lin}}(x^*)$. We already know from Lemma 10.17 that $T_X(x^*) \subseteq T_{\text{lin}}(x^*)$ holds. Thus, we need to prove $T_{\text{lin}}(x^*) \subseteq T_X(x^*)$. To this end, let $d \in T_{\text{lin}}(x^*)$ be given arbitrarily. To show that d is also part of the tangential cone, we first slightly perturb d and set

$$d(\delta) := d + \delta \hat{d}, \quad \delta > 0,$$

where $\hat{d} \in \mathbb{R}^n$ denotes the vector of the definition of the MFCQ. This means that $\hat{d}$ satisfies the conditions

$$\begin{aligned} \nabla g_i(x^*)^\top \hat{d} &> 0 \quad \text{for all} \quad i \in I(x^*), \\ \nabla h_j(x^*)^\top \hat{d} &= 0 \quad \text{for all} \quad j = 1, \ldots, p. \end{aligned}$$

Thus,

$$\begin{aligned} \nabla g_i(x^*)^\top d(\delta) &> 0 \quad \text{for all} \quad i \in I(x^*), \\ \nabla h_j(x^*)^\top d(\delta) &= 0 \quad \text{for all} \quad j = 1, \ldots, p \end{aligned}$$

holds for all $\delta > 0$. Next, we show that $d(\delta) \in T_X(x^*)$ holds for all arbitrary but fixed $\delta > 0$. Since the gradients $\nabla h_j(x^*)$, $j = 1, \ldots, p$, are linearly

independent (due to the MFCQ), Lemma 10.24 implies the existence of an $\varepsilon > 0$ and of a continuously differentiable curve $x : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ so that the points on the curve are feasible for Problem (10.9) for all $t \in [0, +\varepsilon)$. Moreover, $x(0) = x^*$ and $x'(0) = d(\delta)$ holds.[8] Consider now a sequence $(t^k)_k \subseteq \mathbb{R}$ with $t^k \searrow 0$ and set $x^k := x(t^k)$. This defines a feasible sequence $(x^k)_k$ with $x^k \to x(0) = x^*$ and

$$d(\delta) = x'(0) = \lim_{k\to\infty} \frac{x(t^k) - x(0)}{t^k} = \lim_{k\to\infty} \frac{x^k - x^*}{t^k}.$$

Thus, $d(\delta) = d + \delta\hat{d} \in T_X(x^*)$ for every fixed $\delta > 0$. For $\delta^k \searrow 0$, we obtain

$$d = \lim_{k\to\infty} d(\delta^k) \in T_X(x^*),$$

because the tangential cone $T_X(x^*)$ is closed by Lemma 10.27. □

Remark 10.29. Note that the proof of the last theorem revealed that the MFCQ implies the ACQ. The other implication is, in general, not true. To see this, consider the problem

$$\min_{x_1,x_2} \; x_1^2 + (x_2 + 1)^2 \quad \text{s.t.} \quad x_1^2 - x_2 \geq 0, \; x_2 \geq 0.$$

We consider one more constraint qualification, which is very often used in practice.

Definition 10.30 (Linear independence constraint qualification)**.** Let $x \in \mathbb{R}^n$ be a feasible point of Problem (10.9) and let $I(x)$ be the set of active inequality constraints at x. We say that the *linear independence constraint qualification (LICQ)* is satisfied in x if the gradients

$$\begin{aligned} &\nabla g_i(x) \quad \text{for all} \quad i \in I(x), \\ &\nabla h_j(x) \quad \text{for all} \quad j = 1, \ldots, p \end{aligned}$$

are linearly independent.

Theorem 10.31 (KKT conditions under the LICQ)**.** Let $x^* \in \mathbb{R}^n$ be a local minimizer of Problem (10.9) that satisfies the LICQ. Then, there exist Lagrangian multipliers $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ so that (x^*, λ^*, μ^*) is a KKT point of Problem (10.9).

Proof. We need to show that the LICQ implies the MFCQ. To this end,

[8] Note that both the curve x and ε depend on δ.

assume that the LICQ holds at x^*. Then, the first part in the definition of the MFCQ is obviously satisfied. Thus, we need to show that there exists a vector $d \in \mathbb{R}^n$ that satisfies the second part of the definition of the MFCQ. Let $I(x^*)$ be the set of active inequality constraints in x^* and set $m^* := |I(x^*)|$. We now construct a matrix $A \in \mathbb{R}^{n \times n}$ as follows. The first m^* rows are the transposed gradients $\nabla g_i(x^*)^\top$ for $i \in I(x^*)$. The next p rows are the transposed gradients $\nabla h_j(x^*)^\top$ for $j = 1, \dots, p$. Finally, the remaining $n - m^* - p$ rows can be chosen arbitrarily so that the resulting matrix has full row rank. Note that this is possible due to the LICQ. Now, we construct a vector $b \in \mathbb{R}^n$ as follows. The first m^* entries of b are 42 (see Adams (1979)), the next p components of b are 0, and the remaining entries can be chosen arbitrarily. Since A is invertible (by construction), the linear system of equations

$$Ad = b$$

has a unique solution $d \in \mathbb{R}^n$. The definitions of A and b imply that the solution d satisfies the conditions of the second part of the definition of the MFCQ. Thus, we have shown that the LICQ implies the MFCQ. The theorem then follows by applying Theorem 10.28. □

Remark 10.32. Note that the proof of the last theorem shows that the LICQ implies the MFCQ. In summary, we now know that

$$\text{LICQ} \implies \text{MFCQ} \implies \text{ACQ}.$$

Note further that the opposite implication between LICQ and MFCQ does not hold in general. To see this, consider the problem

$$\min_{x_1, x_2} \; x_1^2 + (x_2 + 1)^2 \quad \text{s.t.} \quad x_1^3 + x_2 \geq 0, \; x_2 \geq 0.$$

Lemma 10.33. Suppose that the assumptions of Theorem 10.31 hold. Then, the Lagrangian multipliers λ^* and μ^* of the KKT point (x^*, λ^*, μ^*) are unique.

Proof. The KKT conditions imply that $\lambda_i^* = 0$ for all $i \notin I(x^*)$. The uniqueness of λ_i^* for all $i \in I(x^*)$ and μ_j^* for all $j = 1, \dots, p$ then directly follows from the KKT condition

$$\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = \nabla f(x^*) - \sum_{i \in I(x^*)} \lambda_i^* \nabla g_i(x^*) - \sum_{j=1}^{p} \mu_j^* \nabla h_j(x^*) = 0$$

and the LICQ. □

10.4 The Special Case of Linear Constraints

We now consider the special case of Problem (10.9) in which all equality and inequality constraints are (affine-)linear.[9] Thus, we consider the problem

$$\min_{x\in\mathbb{R}^n} \quad f(x) \tag{10.11a}$$

$$\text{s.t.} \quad a_i^\top x \geq \alpha_i, \quad i = 1,\dots,m, \tag{10.11b}$$

$$b_j^\top x = \beta_j, \quad j = 1,\dots,p. \tag{10.11c}$$

Here, $f : \mathbb{R}^n \to \mathbb{R}$ still is a continuously differentiable function, $a_i \in \mathbb{R}^n$, $i = 1,\dots,m$, as well as $b_j \in \mathbb{R}^n$, $j = 1,\dots,p$, are vectors and $\alpha_i \in \mathbb{R}$, $i = 1,\dots,m$, and $\beta_j \in \mathbb{R}$, $j = 1,\dots,p$, are scalars.

The next theorem shows that the KKT conditions are necessary first-order optimality conditions for Problem (10.11) without any further constraint qualifications. This means that the fact that all constraints are linear itself can be seen as a constraint qualification.

Theorem 10.34 (KKT conditions under linear constraints)**.** Let $x^* \in \mathbb{R}^n$ be a local minimizer of Problem (10.11). Then, there exist Lagrangian multipliers $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ so that (x^*, λ^*, μ^*) satisfies the KKT conditions

$$\begin{aligned}
\nabla f(x^*) - \sum_{i=1}^m \lambda_i^* a_i - \sum_{j=1}^p \mu_j^* b_j = 0,& \\
a_i^\top x^* \geq \alpha_i \quad &\text{for all} \quad i = 1,\dots,m,\\
b_j^\top x^* = \beta_j \quad &\text{for all} \quad j = 1,\dots,p,\\
\lambda_i^*(a_i^\top x^* - \alpha_i) = 0 \quad &\text{for all} \quad i = 1,\dots,m,\\
\lambda_i^* \geq 0 \quad &\text{for all} \quad i = 1,\dots,m
\end{aligned}$$

of Problem (10.11).

Proof. As in the proof of Theorem 10.28 we need to show that the Abadie CQ is implied by the fact that all constraints are linear. This means, we need to show that $T_X(x^*) = T_{\text{lin}}(x^*)$ holds. Lemma 10.17 states that $T_X(x^*) \subseteq T_{\text{lin}}(x^*)$ is always true. Thus, we need to prove $T_{\text{lin}}(x^*) \subseteq T_X(x^*)$. To this end, we consider an arbitrary direction $d \in T_{\text{lin}}(x^*)$, which implies, by definition,

$$a_i^\top d \geq 0,\ i \in I(x^*), \quad \text{and} \quad b_j^\top d = 0,\ j = 1,\dots,p.$$

Let $(t^k)_k \subseteq \mathbb{R}$ be an arbitrary but non-negative sequence with $t^k \searrow 0$ for

[9] For better reading, we refer to these constraints as linear constraints in the following.

$k \to \infty$ and set

$$x^k := x^* + t^k d.$$

Then,

$$\begin{aligned} a_i^\top x^k = a_i^\top (x^* + t^k d) = \alpha_i + t^k a_i^\top d \geq \alpha_i \quad &\text{for all} \quad i \in I(x^*), \\ a_i^\top x^k = a_i^\top (x^* + t^k d) = a_i^\top x^* + t^k a_i^\top d > \alpha_i \quad &\text{for all} \quad i \notin I(x^*), \\ b_j^\top x^k = b_j^\top (x^* + t^k d) = b_j^\top x^* + t^k b_j^\top d = \beta_j \quad &\text{for all} \quad j = 1, \ldots, p \end{aligned}$$

holds. Hence, the sequence $(x^k)_k \subseteq \mathbb{R}^n$ is feasible and converges to x^*. Moreover, it holds

$$\frac{x^k - x^*}{t^k} = \frac{x^* + t^k d - x^*}{t^k} = d \to d$$

for $k \to \infty$, which shows that $d \in T_X(x^*)$. □

Remark 10.35. Note that the latter proof shows that the situation that all constraints are linear implies the Abadie CQ. Thus, if all constraints are linear, the tangential cone coincides with the linearized tangential cone. From an intuitive and geometric point of view, this should be clear, because linearizing a linear constraint should not change anything.

10.5 The Special Case of Convex Problems

We now consider convex optimization problems. We have already shown in Theorem 6.9 that every local minimizer is a global minimizer in this case. In this section, we now study the meaning of the KKT conditions for convex optimization problems. To this end, we consider the problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) && \text{(10.12a)} \\ \text{s.t.} \quad & g_i(x) \geq 0, \quad i = 1, \ldots, m, && \text{(10.12b)} \\ & b_j^\top x = \beta_j, \quad j = 1, \ldots, p, && \text{(10.12c)} \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$, are continuously differentiable, $b_j \in \mathbb{R}^n$, $j = 1, \ldots, p$, are vectors and $\beta_j \in \mathbb{R}$, $j = 1, \ldots, p$, are scalars. Moreover, f is supposed to be convex and the g_i, $i = 1, \ldots, m$, are supposed to be concave. Thus, we consider a convex objective function over a convex feasible set; cf. Definition 6.10.

As for general nonlinear problems, we also need a constraint qualification for convex problems for being able to show that the KKT conditions are meaningful necessary optimality conditions.

Definition 10.36 (Slater's constraint qualification)**.** We say that the convex problem (10.12) satisfies the *constraint qualification of Slater* if there exists a vector $\hat{x} \in \mathbb{R}^n$ so that

$$\begin{aligned} g_i(\hat{x}) &> 0 \quad \text{for all} \quad i = 1, \dots, m, \\ b_j^\top \hat{x} &= \beta_j \quad \text{for all} \quad j = 1, \dots, p \end{aligned}$$

holds. This means that $\hat{x}$ is strictly feasible w.r.t. the inequality constraints and feasible w.r.t. the equality constraints.

In order to get a geometric intuition regarding Slater's constraint qualification, one should consider the situation without equality constraints, i.e., $p = 0$. In this case, the conditions in Definition 10.36 reduce to the existence of a point $\hat{x}$ with $g_i(\hat{x}) > 0$ for all $i = 1, \dots, m$. This is nothing but the claim that the feasible set has an interior point.

Theorem 10.37 (KKT conditions for convex problems under Slater's CQ)**.** Let $x^* \in \mathbb{R}^n$ be a local (and thus global) minimizer of the convex problem (10.12). Moreover, suppose that Slater's CQ is satisfied. Then, there exist Lagrangian multipliers $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ so that (x^*, λ^*, μ^*) satisfies the KKT conditions

$$\begin{aligned} \nabla f(x^*) - \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) - \sum_{j=1}^{p} \mu_j^* b_j = 0, & \\ b_j^\top x^* = \beta_j, \quad j = 1, \dots, p, & \\ g_i(x^*) \geq 0, \quad i = 1, \dots, m, & \\ \lambda_i^* g_i(x^*) = 0, \quad i = 1, \dots, m, & \\ \lambda_i^* \geq 0, \quad i = 1, \dots, m, & \end{aligned}$$

of Problem (10.12).

Proof. As in the proofs of Theorem 10.28 and 10.31 it is enough to show that the inclusion $T_{\text{lin}}(x^*) \subseteq T_X(x^*)$ holds, where X is the feasible set of the convex problem (10.12).

To this end, remember the definition of the linearized tangential cone applied to the currently considered setting:

$$T_{\text{lin}}(x^*) := \{d \in \mathbb{R}^n \colon \nabla g_i(x^*)^\top d \geq 0, i \in I(x^*),\ b_j^\top d = 0, j = 1, \dots, p\}.$$

Next, we define

$$T_{\text{strict}}(x^*) := \{d \in \mathbb{R}^n \colon \nabla g_i(x^*)^\top d > 0, i \in I(x^*),\ b_j^\top d = 0, j = 1, \dots, p\}.$$

It is rather easy to prove that $T_{\text{strict}}(x^*) \subseteq T_X(x^*)$ holds.[10] Since, by Lemma 10.27, the tangential cone is closed, we also have

$$\text{cl}(T_{\text{strict}}(x^*)) \subseteq T_X(x^*),$$

where $\text{cl}(A)$ denotes the closure of the set A. We now show that

$$T_{\text{lin}}(x^*) \subseteq \text{cl}(T_{\text{strict}}(x^*))$$

holds, which then proves the theorem. To this end, let $d \in T_{\text{lin}}(x^*)$ be given arbitrarily. Moreover, let $\hat{x} \in \mathbb{R}^n$ be a point that is strictly feasible w.r.t. the inequality constraints of the convex problem (10.12), which exists due to Slater's CQ. We define

$$\hat{d} := \hat{x} - x^*.$$

Since the g_i are concave functions for all $i = 1, \ldots, m$, we know from Theorem 6.5 that

$$\nabla g_i(x^*)^\top \hat{d} \geq g_i(\hat{x}) - g_i(x^*) = g_i(\hat{x}) > 0 \quad \text{for all} \quad i \in I(x^*)$$

holds. Moreover, it holds

$$\nabla h_j(x^*)^\top \hat{d} = h_j(\hat{x}) - h_j(x^*) = 0 \quad \text{for all} \quad j = 1, \ldots, p$$

since the functions $h_j(x) := b_j^\top x - \beta_j$ are affine-linear. If we now set

$$d(\delta) := d + \delta \hat{d}$$

for $\delta > 0$, it follows

$$\nabla g_i(x^*)^\top d(\delta) = \nabla g_i(x^*)^\top d + \delta \nabla g_i(x^*)^\top \hat{d} > 0 \quad \text{for all} \quad i \in I(x^*)$$

and

$$\nabla h_j(x^*)^\top d(\delta) = \nabla h_j(x^*)^\top d + \delta \nabla h_j(x^*)^\top \hat{d} = 0 \quad \text{for all} \quad j = 1, \ldots, p.$$

Thus, $d(\delta) \in T_{\text{strict}}(x^*)$ for every $\delta > 0$ and for $\delta \searrow 0$ we obtain $d \in \text{cl}(T_{\text{strict}}(x^*))$. □

Up to now, we have shown that the KKT conditions are also necessary first-order optimality conditions for convex problems under Slater's CQ. For general nonlinear problems, the KKT conditions (under a reasonable CQ) are not sufficient conditions. However, for convex problems, the KKT conditions are also sufficient conditions.

[10] You will prove this in the exercise classes.

Theorem 10.38. Let $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ be a KKT point of the convex problem (10.12). Then, x^* is a local (and thus global) minimizer of (10.12).

Proof. Let $x \in \mathbb{R}^n$ be a feasible point of Problem (10.12). Then, the KKT conditions and the convexity of f imply

$$\begin{aligned} f(x) &\geq f(x^*) + \nabla f(x^*)^\top (x - x^*) \\ &= f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*)^\top (x - x^*) + \sum_{j=1}^{p} \mu_j^* b_j^\top (x - x^*) \\ &= f(x^*) + \sum_{i \in I(x^*)} \lambda_i^* \nabla g_i(x^*)^\top (x - x^*) \\ &\geq f(x^*) \end{aligned}$$

because $\lambda_i^* \geq 0$ and

$$\nabla g_i(x^*)^\top (x - x^*) \geq g_i(x) - g_i(x^*) = g_i(x) \geq 0$$

holds for all $i \in I(x^*)$. Note that the last inequality follows from the concavity of g_i. Thus, we have shown that x^* is a global minimizer of Problem (10.12). □

Note that the last theorem does not require any constraint qualification.

If we now put the results on linearly constraint and convex problems together, we obtain the following result.

Corollary 10.39. Consider the convex problem (10.12) and suppose that all constraints are given by affine-linear functions. Then, $x^* \in \mathbb{R}^n$ is a local (and thus global) minimizer of (10.12) if and only if there exists Lagrangian multipliers $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ so that (x^*, λ^*, μ^*) is a KKT point of Problem (10.12).

Lastly, we consider so-called saddle point conditions that are often used as optimality conditions for convex problems.

Definition 10.40 (Saddle point)**.** A vector $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ with $\lambda^* \geq 0$ is called a *saddle point* of the Lagrangian function $\mathcal{L}$ if

$$\mathcal{L}(x^*, \lambda, \mu) \leq \mathcal{L}(x^*, \lambda^*, \mu^*) \leq \mathcal{L}(x, \lambda^*, \mu^*) \tag{10.13}$$

holds for all $(x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ with $\lambda \geq 0$.

The right-hand side inequality in (10.13) states that $x^* \in \mathbb{R}^n$ is a global minimizer of the function $\mathcal{L}(\cdot, \lambda^*, \mu^*) : \mathbb{R}^n \to \mathbb{R}$, whereas the left-hand side inequality states that $(\lambda^*, \mu^*) \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p$ is a global maximizer of the function $\mathcal{L}(x^*, \cdot, \cdot) : \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p \to \mathbb{R}$. This also explains why such points are called saddle points of the Lagrangian function.

Theorem 10.41 (Saddle point theorem)**.** Consider the convex problem (10.12). Then, $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ is a saddle point of the Lagrangian function $\mathcal{L}$ if and only if (x^*, λ^*, μ^*) is a KKT point of (10.12).

Proof. First, assume that (x^*, λ^*, μ^*) is a saddle point of the Lagrangian function $\mathcal{L}$. The right-hand side inequality in (10.13) then implies that x^* is a global minimizer of the function $\mathcal{L}(\cdot, \lambda^*, \mu^*)$. In particular, x^* is also a stationary point of this function, which implies $\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0$. Moreover, the left-hand side inequality in (10.13) implies

$$\sum_{i=1}^{m} \lambda_i^* g_i(x^*) + \sum_{j=1}^{p} \mu_j^* h_j(x^*) \leq \sum_{i=1}^{m} \lambda_i g_i(x^*) + \sum_{j=1}^{p} \mu_j h_j(x^*) \tag{10.14}$$

for all $0 \leq \lambda \in \mathbb{R}^m$ and all $\mu \in \mathbb{R}^p$. Note that we again have $h_j(x) := b_j^\top x - \beta_j$. This implies $g(x^*) \geq 0$ and $h(x^*) = 0$ since, otherwise, (10.14) could be violated by considering sufficiently large λ and μ. Consider now $\lambda = 0$ and $\mu = \mu^*$ in (10.14). This implies

$$\sum_{i=1}^{m} \lambda_i^* g_i(x^*) \leq 0$$

and thus $\lambda_i g_i(x^*) = 0$ for all $i = 1, \ldots, m$ because $\lambda_i^* \geq 0$ and $g_i(x^*) \geq 0$ holds. In summary, we have shown that (x^*, λ^*, μ^*) is a KKT point of Problem (10.12).

We now prove the other implication. To this end, let (x^*, λ^*, μ^*) be a KKT point of Problem (10.12). This, in particular, implies that $\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0$ holds, i.e., x^* is a stationary point of the function $\mathcal{L}(\cdot, \lambda^*, \mu^*)$. Due to our assumptions, this function is convex. Thus, Theorem 6.9 implies that x^* is a global minimizer of this function, i.e., we obtain

$$\mathcal{L}(x^*, \lambda^*, \mu^*) \leq \mathcal{L}(x, \lambda^*, \mu^*)$$

for all $x \in \mathbb{R}^n$. Lastly, we use $g_i(x^*) \geq 0$, $h(x^*) = 0$, and $\lambda_i^* g_i(x^*) = 0$ to

obtain

$$\begin{aligned}\mathcal{L}(x^*, \lambda^*, \mu^*) &= f(x^*) - \sum_{i=1}^{m} \lambda_i^* g_i(x^*) - \sum_{j=1}^{p} \mu_j^* h_j(x^*) \\ &= f(x^*) \\ &\geq f(x^*) - \sum_{i=1}^{m} \lambda_i g_i(x^*) - \sum_{j=1}^{p} \mu_j h_j(x^*) \\ &= \mathcal{L}(x^*, \lambda, \mu)\end{aligned}$$

for all $\lambda \geq 0$ and all $\mu \in \mathbb{R}^p$. Thus, (x^*, λ^*, μ^*) is a saddle point of the Lagrangian function. □

Finally, we collect all results regarding convex problems.

Corollary 10.42. Consider the convex problem (10.12). Then, the following statements are true:

(a) Suppose that $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ is a saddle point of the Lagrangian function $\mathcal{L}$. Then, x^* is a global minimizer of the problem (10.12).

(b) Suppose that x^* is local (and thus global) minimizer of Problem (10.12) and that Slater's CQ is satisfied. Then, there exist Lagrangian multipliers $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ so that (x^*, λ^*, μ^*) is a saddle point of the Lagrangian function $\mathcal{L}$.

(c) Suppose that all constraints of Problem (10.12) are given by affine-linear functions. Then, $x^* \in \mathbb{R}^n$ is a local (and thus global) minimizer of Problem (10.12) if and only if there exists vectors $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ so that (x^*, λ^*, μ^*) is a saddle point of the Lagrangian function $\mathcal{L}$.

Remark 10.43. Note that the definition of a saddle point does not need the differentiability of the functions f and g_i, which is needed for the formulation of the KKT conditions. Thus, the approach using saddle points is more suitable for nonsmooth but convex optimization problems.

10.6 Fritz-John Conditions

In the last sections, we have seen that the KKT conditions (under suitable constraint qualifications) are indeed necessary first-order optimality conditions. For convex problems, we have seen that these conditions are even sufficient. In this last section on first-order conditions for constrained optimization problems, we consider so-called *Fritz-John (FJ) conditions*. These conditions

are strongly related to the KKT conditions. In contrast to the KKT conditions, the FJ conditions do not need any constraint qualification to be satisfied at a local minimizer. Consequently, the result is weaker.

For the remainder of this section let us again assume that the functions $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$, and $h : \mathbb{R}^n \to \mathbb{R}^p$ are continuously differentiable.

We start with a rather technical result.

Lemma 10.44. Let $x^* \in \mathbb{R}^n$ be a local minimizer of the optimization problem (10.9). Suppose further that the gradients $\nabla h_j(x^*)$ are linearly independent for $j = 1, \dots, p$. Moreover, let $I(x^*)$ be the set of active inequality constraints. Then, $\nabla f(x^*)^\top d \geq 0$ holds for all vectors $d \in \mathbb{R}^n$ with $\nabla h_j(x^*)^\top d = 0$ for all $j = 1, \dots, p$ and $\nabla g_i(x^*)^\top d > 0$ for all $i \in I(x^*)$.

Proof. Suppose that there exists a vector $d \in \mathbb{R}^n$ with $\nabla h_j(x^*)^\top d = 0$ for all $j = 1, \dots, p$, $\nabla g_i(x^*)^\top d > 0$ for all $i \in I(x^*)$, and $\nabla f(x^*)^\top d < 0$. Applying Lemma 10.24 implies the existence of a curve $x(t)$ with the properties as stated in the lemma. With a little bit of work (which is similar to the proof of Lemma 10.24) one can then show that $f(x(t)) < f(x^*)$ holds for sufficiently small t. From the feasibility of $x(t)$ and $x(0) = x^*$ we can then derive a contradiction to the assumption that x^* is a local minimizer. □

Definition 10.45 (Fritz-John conditions, Fritz-John point)**.** Consider the optimization problem (10.9) with continuously differentiable functions f, g, and h.

(a) The conditions

$$\begin{aligned} r\nabla f(x) - \sum_{i=1}^{m} \lambda_i \nabla g_i(x) - \sum_{j=1}^{p} \mu_j \nabla h_j(x) &= 0, \\ h(x) &= 0, \\ \lambda \geq 0,\ g(x) \geq 0,\ \lambda^\top g(x) &= 0, \\ r &\geq 0 \end{aligned}$$

are called *Fritz-John (FJ) conditions* of Problem (10.9).

(b) Every vector $(r^*, x^*, \lambda^*, \mu^*) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ that satisfies the FJ conditions is called a *Fritz-John (FJ) point* of Problem (10.9).

Remark 10.46. The only difference between the KKT and the FJ conditions is the additional scalar $r^* \geq 0$ in the FJ conditions. If for a FJ point $r^* = 1$ holds, this FJ point is also a KKT point. Moreover, it is easy to see that $(tr^*, x^*, t\lambda^*, t\mu^*)$ is a FJ point for all $t > 0$ if $(r^*, x^*, \lambda^*, \mu^*)$ is a FJ point.

Thus, in the case $r^* > 0$ we can assume $r^* = 1$ without loss of generality. Otherwise, we can simply multiply with $t = 1/r^*$. In other words: If there is a FJ point $(r^*, x^*, \lambda^*, \mu^*)$ with $r^* > 0$, then there is also a KKT point.

We now state and prove the Fritz-John analogue of the KKT theorems of the previous sections.

Theorem 10.47 (Fritz-John conditions)**.** Let $x^* \in \mathbb{R}^n$ be a local minimizer of Problem (10.9). Then, there exists a non-zero vector $(r^*, \lambda^*, \mu^*) \in \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^p$ so that $(r^*, x^*, \lambda^*, \mu^*)$ is a FJ point of Problem (10.9).

Interestingly, the theorem states the existence of a FJ point at a local minimizer without any further assumptions. However, note that the theorem does not state the existence of a FJ point with $r^* \neq 0$. It only states the existence of a non-zero vector (r^*, λ^*, μ^*) so that $(r^*, x^*, \lambda^*, \mu^*)$ is a FJ point. In this context, note that $(0, x^*, 0, 0)$ is a FJ point if and only if x^* is feasible for Problem (10.9). Thus, a FJ point with $(r^*, \lambda^*, \mu^*) = (0, 0, 0)$ is not really interesting.

Proof. We first consider the case in which the gradients $\nabla h_j(x^*)$, $j = 1, \dots, p$, are linearly dependent. In this case, there exists a non-zero multiplier $0 \neq \mu^* \in \mathbb{R}^p$ with

$$\sum_{j=1}^{p} \mu_j^* \nabla h_j(x^*) = 0.$$

If we set $r^* = 0$ and $\lambda^* = 0$, we directly see that $(r^*, x^*, \lambda^*, \mu^*)$ is a FJ point of Problem (10.9).

Thus, we need to consider the case in which the gradients $\nabla h_j(x^*)$, $j = 1, \dots, p$, are linearly independent. Moreover, let $I(x^*)$ be the set of $m^* = |I(x^*)|$ active inequalities at x^*, let $A_1 \in \mathbb{R}^{(1+m^*) \times n}$ be the matrix with rows $\nabla f(x^*)^\top$ and $-\nabla g_i(x^*)^\top$ for $i \in I(x^*)$, and let $A_2 \in \mathbb{R}^{p \times n}$ be the matrix with rows $-\nabla h_j(x^*)^\top$ for $j = 1, \dots, p$. Lemma 10.44 states that the system

$$A_1 d < 0, \quad A_2 d = 0$$

has no solution. We now consider the sets

$$C_1 := \{(x_1, x_2) \in \mathbb{R}^{1+m^*} \times \mathbb{R}^p \colon x_1 < 0,\ x_2 = 0\},$$
$$C_2 := \{(y_1, y_2) \in \mathbb{R}^{1+m^*} \times \mathbb{R}^p \colon \exists d \in \mathbb{R}^n \text{ with } y_1 = A_1 d,\ y_2 = A_2 d\}.$$

One can easily check that both C_1 and C_2 are non-empty and convex sets. The previous discussion also shows that $C_1 \cap C_2 = \emptyset$ holds. Using the separation theorem (Theorem 10.4), we know that there exists a non-zero

vector $0 \neq a = (a_1^\top, a_2^\top)^\top$ with

$$a_1^\top x_1 + a_2^\top 0 = a_1^\top x_1 \leq a_1^\top (A_1 d) + a_2^\top (A_2 d) \tag{10.15}$$

for all $x_1 < 0$ and all $d \in \mathbb{R}^n$. The inequality also holds for all $x_1 \leq 0$ and all $d \in \mathbb{R}^n$. We now fix $d = 0 \in \mathbb{R}^n$ and observe that the components of x_1 in (10.15) can be chosen arbitrarily small. This implies $a_1 \geq 0$. On the other hand, for $x_1 = 0$ we obtain

$$0 \leq (a_1^\top A_1 + a_2^\top A_2) d$$

for all $d \in \mathbb{R}^n$. In particular, for $d := -(A_1^\top a_1 + A_2^\top a_2)$ we get

$$\|A_1^\top a_1 + A_2^\top a_2\|^2 \leq 0,$$

which implies

$$A_1^\top a_1 + A_2^\top a_2 = 0.$$

We now denote the components of the vector $a_1 \in \mathbb{R}^{1+m^*}$ with r^* and λ_i^* for $i \in I(x^*)$ as well as the components of the vector $a_2 \in \mathbb{R}^p$ with μ_j^* for $j = 1, \ldots, p$. Moreover, we set $\lambda_i^* = 0$ for all $i \notin I(x^*)$. Then, (r^*, λ^*, μ^*) is a non-zero vector so that $(r^*, x^*, \lambda^*, \mu^*)$ satisfies the FJ conditions. □

Note that the theorem does not state the existence of a FJ point with $r^* > 0$. We will show in the exercises that there are problems for which no FJ point with $r^* > 0$ exist. Obviously, this needs to be related to the failure of the Abadie CQ.

11

Second-Order Optimality Conditions

In this section, we consider second-order optimality conditions for constrained optimization problems. To this end, we still consider the problem

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) && \text{(11.1a)}\\ \text{s.t.} \quad & g_i(x) \geq 0, \quad i \in I = \{1,\dots,m\}, && \text{(11.1b)}\\ & h_j(x) = 0, \quad j \in J = \{1,\dots,p\}. && \text{(11.1c)} \end{aligned}$$

However, in this section we assume that the functions $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$, and $h : \mathbb{R}^n \to \mathbb{R}^p$ are twice continuously differentiable.

Let (x^*, λ^*, μ^*) be a KKT point of Problem (11.1). As before, we define

$$I(x^*) := \{i \in \{1,\dots,m\} \colon g_i(x^*) = 0\}$$

to be the index set of active inequality constraints. We now further partition this index set by

$$\begin{aligned} I_0(x^*) &:= \{i \in I(x^*) \colon \lambda_i^* = 0\},\\ I_>(x^*) &:= \{i \in I(x^*) \colon \lambda_i^* > 0\}. \end{aligned}$$

Note that the sets $I_0(x^*)$ and $I_>(x^*)$ both depend on x^* and on the Lagrangian multiplier λ^*.

With these index sets, we further define

$$\begin{aligned}
\mathcal{T}_1(x^*) &:= \{d \in \mathbb{R}^n : \nabla g_i(x^*)^\top d = 0,\ i \in I(x^*),\\
&\qquad\qquad \nabla h_j(x^*)^\top d = 0,\ j = 1, \dots, p\}\\
\mathcal{T}_2(x^*) &:= \{d \in \mathbb{R}^n : \nabla g_i(x^*)^\top d = 0,\ i \in I_>(x^*),\\
&\qquad\qquad \nabla g_i(x^*)^\top d \geq 0,\ i \in I_0(x^*),\\
&\qquad\qquad \nabla h_j(x^*)^\top d = 0,\ j = 1, \dots, p\},\\
\mathcal{T}_3(x^*) &:= \{d \in \mathbb{R}^n : \nabla g_i(x^*)^\top d = 0,\ i \in I_>(x^*),\\
&\qquad\qquad \nabla h_j(x^*)^\top d = 0,\ j = 1, \dots, p\}.
\end{aligned}$$

The cone $\mathcal{T}_2(x^*)$ is also called *critical cone*. Since $I_>(x^*) \subseteq I(x^*)$, it obviously holds

$$\mathcal{T}_1(x^*) \subseteq \mathcal{T}_2(x^*) \subseteq \mathcal{T}_3(x^*).$$

If the KKT point (x^*, λ^*, μ^*) even satisfies strict complementarity, i.e.,

$$\lambda_i^* + g_i(x^*) \neq 0 \quad \text{for all} \quad i = 1, \dots, m,$$

then

$$\mathcal{T}_1(x^*) = \mathcal{T}_2(x^*) = \mathcal{T}_3(x^*)$$

holds.

With these notations at hand, we are now able to state and prove the first second-order optimality criterion.

Theorem 11.1 (Necessary second-order optimality condition)**.** Let $x^* \in X$ be a local minimizer of Problem (11.1) and suppose that the LICQ is satisfied in x^*. Then,

$$d^\top \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*, \mu^*) d \geq 0 \quad \text{for all} \quad d \in \mathcal{T}_2(x^*).$$

Here, λ^* and μ^* are the uniquely determined (by Lemma 10.33) Lagrangian multipliers with respect to x^*.

Proof. Let $d \in \mathcal{T}_2(x^*)$ be given and, w.l.o.g., assume that $d \neq 0$. We further decompose the index set $I_0(x^*)$ into

$$I_0^>(x^*) := \{i \in I_0(x^*) \colon \nabla g_i(x^*)^\top d > 0\}$$

and

$$I_0^=(x^*) := \{i \in I_0(x^*) \colon \nabla g_i(x^*)^\top d = 0\}.$$

Note that this decomposition depends on the specifically chosen vector

$d \in \mathcal{T}_2(x^*)$. Since the LICQ holds at x^*, the vectors

$$\nabla g_i(x^*) \quad \text{for all} \quad i \in I_>(x^*) \cup I_0^=(x^*)$$

and

$$\nabla h_j(x^*) \quad \text{for all} \quad j = 1, \ldots, p$$

are linearly independent. As in the proof of Lemma 10.24[1], we can now establish[2] the existence of an $\varepsilon > 0$ and a twice differentiable curve $x : (-\varepsilon, +\varepsilon) \to \mathbb{R}^n$ with

(a) $x(0) = x^*$,

(b) $x'(0) = d$,

(c) $g_i(x(t)) = 0$ for all $i \in I_>(x^*) \cup I_0^=(x^*)$ and $t \in (-\varepsilon, +\varepsilon)$

(d) $h_j(x(t)) = 0$ for all $j = 1, \ldots, p$ and $t \in (-\varepsilon, +\varepsilon)$,

(e) $x(t) \in X$ for all $t \in [0, +\varepsilon)$.

We now set

$$\varphi(t) := \mathcal{L}(x(t), \lambda^*, \mu^*) \quad \text{for} \quad t \in (-\varepsilon, +\varepsilon).$$

Then, φ is also twice differentiable and has the derivatives[3]

$$\varphi'(t) = x'(t)^\top \nabla_x \mathcal{L}(x(t), \lambda^*, \mu^*)$$

as well as

$$\varphi''(t) = x''(t)^\top \nabla_x \mathcal{L}(x(t), \lambda^*, \mu^*) + x'(t)^\top \nabla_{xx}^2 \mathcal{L}(x(t), \lambda^*, \mu^*) x'(t).$$

Since (x^*, λ^*, μ^*) is a KKT point of Problem (11.1), the properties of the curve $x(\cdot)$ imply

$$\varphi'(0) = x'(0)^\top \nabla_x \mathcal{L}(x(0), \lambda^*, \mu^*) = d^\top \nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0$$

and

$$\varphi''(0) = d^\top \nabla_{xx}^2 \mathcal{L}(x(0), \lambda^*, \mu^*) d = d^\top \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*, \mu^*) d$$

because of $\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0$.

Assume now that $\varphi''(0) = d^\top \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*, \mu^*) d < 0$ holds. Due to the continuity of φ'' it then also holds $\varphi''(t) < 0$ for all sufficiently small $t \in$

[1] See Lemma 2.37 in Geiger and Kanzow (2002).

[2] This, however, would need a little bit of effort.

[3] The first derivative follows by using the chain rule and the second derivative follows from the product and the chain rule.

$(-\varepsilon, +\varepsilon)$. The Taylor expansion of φ around $t = 0$ yields

$$\varphi(t) = \varphi(0) + t\varphi'(0) + \frac{t^2}{2}\varphi''(\xi_t)$$

for all $t \in (-\varepsilon, +\varepsilon)$ and a point ξ_t that depends on t. From $\varphi'(0) = 0$ and $\varphi''(\xi_t) < 0$ for sufficiently small $t \in (-\varepsilon, +\varepsilon)$ it thus follows

$$\varphi(t) < \varphi(0)$$

for these $t \in (-\varepsilon, +\varepsilon)$. Moreover, it holds

$$\varphi(0) = \mathcal{L}(x^*, \lambda^*, \mu^*) = f(x^*) - \sum_{i=1}^{m} \lambda_i^* g_i(x^*) - \sum_{j=1}^{p} \mu_j^* h_j(x^*) = f(x^*)$$

and

$$\varphi(t) = \mathcal{L}(x(t), \lambda^*, \mu^*) = f(x(t)) - \sum_{i=1}^{m} \lambda_i^* g_i(x(t)) - \sum_{j=1}^{p} \mu_j^* h_j(x(t)) = f(x(t))$$

due to the properties of the curve x and the KKT conditions of Problem (11.1). In particular, $\lambda_i^* = 0$ holds for $i \notin I_0(x^*)$. Thus, we obtain

$$f(x(t)) < f(x^*)$$

for all t sufficiently close to 0. Since the curve x is feasible for all $t \geq 0$, this is a contradiction to x^* being a local minimizer. □

Remark 11.2. For the unconstrained case, i.e.,

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad x \in X = \mathbb{R}^n,$$

we have $\mathcal{T}_2(x^*) = \mathbb{R}^n$. Hence, the last theorem reduces to the known necessary second-order optimality condition for unconstrained problems, i.e., that the Hessian matrix $\nabla^2 f(x^*)$ is positive semi-definite; cf. Theorem 5.5.

Next, we state a sufficient second-order optimality condition for constrained optimization problems.

Theorem 11.3 (Sufficient second-order optimality condition)**.** Let (x^*, λ^*, μ^*) be a KKT point of Problem (11.1) with

$$d^\top \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*, \mu^*) d > 0 \quad \text{for all} \quad d \in \mathcal{T}_2(x^*) \setminus \{0\}.$$

Then, x^* is a strict local minimizer of Problem (11.1).

Proof. Suppose that x^* is not a strict local minimizer of Problem (11.1). Then, there exists a feasible sequence $(x^k)_k \subseteq X$ with $x^k \to x^*$, $x^k \neq x^*$, and $f(x^k) \leq f(x^*)$ for all $k \in \mathbb{N}$. The direction

$$d^k := \frac{x^k - x^*}{\|x^k - x^*\|}$$

is well-defined because of $x^k \neq x^*$ for all k. It holds $\|d^k\| = 1$ for all k and, hence, the sequence is bounded. Thus, there exists a convergent subsequence. Without loss of generality, let $d^k \to d^*$ for some $d^* \in \mathbb{R}^n$. Obviously, $\|d^*\| = 1$ holds. Every h_j, $j = 1, \ldots, p$, is continuously differentiable. Hence, the mean value theorem of differential calculus can be applied and ensures a point ξ^k on the line segment between x^k and x^* such that

$$h_j(x^k) = h_j(x^*) + \nabla h_j(\xi^k)^\top (x^k - x^*)$$

holds.[4] Since x^k and x^* are both feasible, it holds $h_j(x^k) = h_j(x^*) = 0$. This implies

$$\nabla h_j(\xi^k)^\top (x^k - x^*) = 0.$$

Division by $\|x^k - x^*\|$ and taking the limit for $k \to \infty$ yields

$$\nabla h_j(x^*)^\top d^* = 0$$

because $x^k \to x^*$ also implies $\xi^k \to x^*$. Since j was arbitrary it thus follows

$$\nabla h_j(x^*)^\top d^* = 0 \quad \text{for all} \quad j = 1, \ldots, p. \tag{11.2}$$

Additionally,

$$\nabla g_i(x^*)^\top d^* \geq 0 \quad \text{for all} \quad i \in I(x^*) \tag{11.3}$$

can be shown analogously because $g_i(x^*) = 0$ and $g_i(x^k) \geq 0$ holds for all $i \in I(x^*)$ and all k. Furthermore,

$$\nabla f(x^*)^\top d^* \leq 0 \tag{11.4}$$

can be shown analogously as well by using $f(x^k) \leq f(x^*)$ for all $k \in \mathbb{N}$.

In the remainder of the proof, we consider two cases that both yield a contradiction. This will prove the claim.

Case 1 The inequalities in (11.3) are satisfied with equality for all $i \in I_>(x^*)$. Using (11.2) and (11.3) then shows $d^* \in \mathcal{T}_2(x^*)$. Further using

[4] Note that ξ^k also depends on the index j.

$h_j(x^k) = 0$, $g_i(x^k) \geq 0$, and $\lambda_i^* \geq 0$ for $i = 1, \ldots, m$ leads to

$$f(x^*) \geq f(x^k) \geq f(x^k) - \sum_{i=1}^{m} \lambda_i^* g_i(x^k) - \sum_{j=1}^{p} \mu_j^* h_j(x^k) = \ell(x^k)$$

with

$$\ell(x) := \mathcal{L}(x, \lambda^*, \mu^*) = f(x) - \sum_{i=1}^{m} \lambda_i^* g_i(x) - \sum_{j=1}^{p} \mu_j^* h_j(x).$$

Together with a suitably chosen ζ^k, the Taylor expansion of ℓ around x^* yields

$$\begin{aligned} f(x^*) &\geq \ell(x^k) \\ &= \ell(x^*) + \nabla \ell(x^*)^\top (x^k - x^*) + \frac{1}{2}(x^k - x^*)^\top \nabla^2 \ell(\zeta^k)(x^k - x^*) \\ &= f(x^*) + \frac{1}{2}(x^k - x^*)^\top \nabla^2_{xx} \mathcal{L}(\zeta^k, \lambda^*, \mu^*)(x^k - x^*) \end{aligned}$$

because of $\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0$, $h_j(x^*) = 0$ for all $j = 1, \ldots, p$, and $\lambda_i^* g_i(x^*) = 0$ for all $i = 1, \ldots, m$. Division by $\|x^k - x^*\|^2$ and taking the limit $k \to \infty$ gives

$$(d^*)^\top \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*, \mu^*) d^* \leq 0$$

since $\zeta^k \to x^*$ for $k \to \infty$. As $d^* \neq 0$ and $d^* \in \mathcal{T}_2(x^*)$, this is a contradiction to the assumptions in the theorem.

Case 2 Suppose that there exists $i_0 \in I_>(x^*)$ in (11.3) with $\nabla g_{i_0}(x^*)^\top d^* > 0$. Using $\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0$, $\lambda_i^* = 0$ for all $i \in I_0(x^*)$, as well as (11.2)–(11.4) leads to

$$\begin{aligned} 0 &\geq \nabla f(x^*)^\top d^* \\ &= \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*)^\top d^* + \sum_{j=1}^{p} \mu_j^* \nabla h_j(x^*)^\top d^* \\ &= \sum_{i \in I_>(x^*)} \lambda_i^* \nabla g_i(x^*)^\top d^* \\ &\geq \lambda_{i_0}^* \nabla g_{i_0}(x^*)^\top d^* \\ &> 0, \end{aligned}$$

which obviously is a contradiction. □

Remark 11.4. We again consider the special case of an unconstrained

optimization problem. In this case, the statement of the last theorem reduces to the condition that the Hessian matrix $\nabla^2 f(x^*)$ is positive definite. This is the same condition as in Theorem 5.7.

12

Sensitivity Analysis for Equality Constrained Problems

Up to now, we have seen that Lagrangian multipliers are crucial for formulating first- as well as second-order optimality conditions for constrained optimization problems.

In practice, these multipliers are also important because they almost always have a very specific interpretation. For example, in economic applications, Lagrangian multipliers often correspond to prices. In other applications, Lagrangian multipliers have a special physical or mechanical interpretation.

From the point of view of mathematical optimization, it is often interesting to know how much the optimal objective function value will change if one slightly alters a constraint. This is called the *sensitivity* of the optimization problem. Again, the answer is given by Lagrangian multipliers.

We start by analyzing the equality constrained optimization problem

$$\min_{x \in \mathbb{R}^n} \quad \text{s.t.} \quad f(x) \tag{12.1a}$$

$$h_j(x) = 0, \quad j \in J = \{1, \ldots, p\}, \tag{12.1b}$$

in which no inequality constraints are present.

Theorem 12.1 (Sensitivity Theorem for Equality Constrained Problems)**.** Consider Problem (12.1) and let x^* and μ^* be a local minimizer and a corresponding Lagrangian multiplier so that (x^*, μ^*) is a KKT point of Problem (12.1). Moreover, suppose that the second-order sufficient condition of Theorem 11.3 is satisfied and that the LICQ is fulfilled at x^*.

Consider now the perturbed problem

$$\min_{x\in\mathbb{R}^n} \quad f(x) \tag{12.2a}$$

$$\text{s.t.} \quad h_j(x) = u_j, \quad j \in J = \{1,\dots,p\}, \tag{12.2b}$$

which is parameterized by the right-hand side vector $u \in \mathbb{R}^p$. Then, there exists an open sphere $S \subseteq \mathbb{R}^p$ centered at $u = 0$ such that for every $u \in S$, there is an $x(u) \in \mathbb{R}^n$ and a multiplier $\mu(u) \in \mathbb{R}^p$ so that $x(u)$ is a local minimizer of the perturbed Problem (12.2) and so that $(x(u), \mu(u))$ is a KKT point of (12.2). Furthermore, $x(\cdot)$ and $\mu(\cdot)$ are continuously differentiable functions within S and we have

$$x(0) = x^* \quad \text{and} \quad \mu(0) = \mu^*.$$

Additionally, for all $u \in S$ we have

$$\nabla p(u) = \mu(u),$$

where $p(u)$ is the optimal cost parameterized by u, i.e.,

$$p(u) = f(x(u)).$$

Before we prove the theorem, we first illustrate it using a problem with a single linear constraint.

Example 12.2. We consider the problem

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & a^\top x = b. \end{aligned}$$

Let $x^* \in \mathbb{R}^n$ be a local minimizer and let $\mu^* \in \mathbb{R}$ be a corresponding Lagrangian multiplier so that (x^*, μ^*) is a KKT point of the problem. If we now change the right-hand side of the single linear constraint from b to $b + \Delta b$, the local minimizer will change as well. We denote the new local minimizer with $x^* + \Delta x$. Since $x^* + \Delta x$ is feasible for the perturbed problem, it holds

$$b + \Delta b = a^\top (x^* + \Delta x) = a^\top x^* + a^\top \Delta x = b + a^\top \Delta x.$$

Thus, the variations Δx and Δb are related via

$$a^\top \Delta x = \Delta b.$$

Moreover, we know that for the optimal Lagrangian multiplier μ^* of the

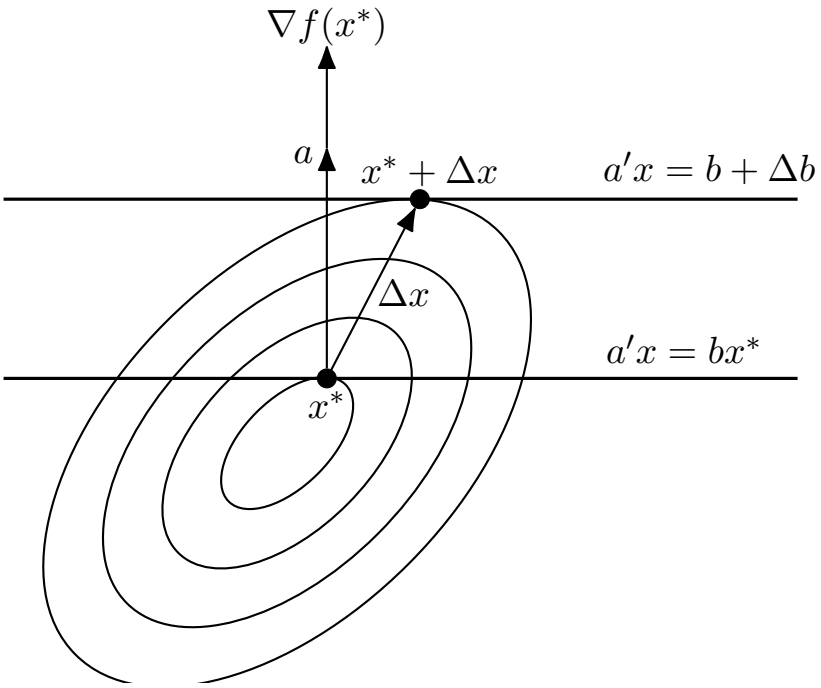


Figure 12.1: Illustration of Example 12.2

unperturbed problem, it holds

$$\nabla f(x^*) = \mu^* a.$$

With this at hand, we can deduce that the corresponding change in the objective function value is given by

$$f(x^* + \Delta x) - f(x^*) = \nabla f(x^*)^\top \Delta x + o(\|\Delta x\|) = \mu^* a^\top \Delta x + o(\|\Delta x\|).$$

Combining the two results, we obtain

$$f(x^* + \Delta x) - f(x^*) = \mu^* \Delta b + o(\|\Delta x\|).$$

Thus, up to first order, we obtain

$$\mu^* = \frac{f(x^* + \Delta x) - f(x^*)}{\Delta b}$$

and the Lagrangian multiplier μ^* gives the rate of objective function value change in dependence of the change of the right-hand side of the constraint.

Remark 12.3. Note that for equality-constrained problems, the second-order

sufficient condition of Theorem 11.3 is given by

$$d^\top \nabla^2_{xx} \mathcal{L}(x^*, \mu^*) d > 0 \quad \text{for all} \quad d \in \mathcal{T}_2(x^*) \setminus \{0\},$$

with

$$\mathcal{T}_2(x^*) = \{d \in \mathbb{R}^n \colon \nabla h_j(x^*)^\top d = 0 \text{ for all } j = 1, \dots, p\}.$$

This means that the Hessian of the Lagrangian needs to be positive definite on the null space of the gradients of the equality constraints.

Proof of Theorem 12.1. The KKT conditions of the perturbed problem are given by the system of equations

$$\nabla f(x) - \nabla h(x)\mu = 0, \quad h(x) = u. \tag{12.3}$$

Here, $\nabla h(x) \in \mathbb{R}^{n \times p}$ is the transpose of the Jacobian of the constraint function $h : \mathbb{R}^n \to \mathbb{R}^p$, i.e., the columns of $\nabla h(x)$ are the gradients $\nabla h_i(x) \in \mathbb{R}^n$, $i = 1, \dots, p$. For a fixed u, this system contains $n + p$ equations and $n + p$ variables ($x \in \mathbb{R}^n$ and $\mu \in \mathbb{R}^p$). For $u = 0$, this system has the solution (x^*, μ^*). The corresponding Jacobian matrix of System (12.3) is the $(n + p) \times (n + p)$ matrix

$$J = \begin{bmatrix} \nabla^2_{xx} \mathcal{L}(x^*, \mu^*) & -\nabla h(x^*) \\ \nabla h(x^*)^\top & 0 \end{bmatrix}.$$

We now prove that J is invertible. To this end, assume that J is not invertible. Then, there exists non-zero vector $(y^\top, z^\top)^\top$ in the null space of J, i.e.,

$$\nabla^2_{xx} \mathcal{L}(x^*, \mu^*) y - \nabla h(x^*) z = 0, \tag{12.4a}$$

$$\nabla h(x^*)^\top y = 0. \tag{12.4b}$$

By multiplying (12.4a) with $y^\top$ from the left and using (12.4b), we obtain

$$y^\top \nabla^2_{xx} \mathcal{L}(x^*, \mu^*) y = 0.$$

Thus, $y = 0$ holds because otherwise, the second-order sufficient conditions would be violated. Substituting $y = 0$ into (12.4a) then yields $z = 0$ due to the LICQ. Thus, we have shown $y = 0$ and $z = 0$, which is a contradiction. Hence, J is invertible.

We now consider System (12.3) again. Since J is invertible, the implicit function theorem (see Proposition A.25 in Appendix A of Bertsekas (2016)) can be applied and ensures for all u in some open sphere S, which is centered at $u = 0$, that there exists continuously differentiable functions $x(u)$ and $\mu(u)$

such that $x(0) = x^*$, $\mu(0) = \mu^*$, and

$$\nabla f(x(u)) - \nabla h(x(u))\mu(u) = 0, \tag{12.5a}$$
$$h(x(u)) = u. \tag{12.5b}$$

For u sufficiently close to 0, by continuity, the vectors $x(u)$ and $\mu(u)$ satisfy the second-order sufficient conditions for the perturbed problem (12.2), since they satisfy them by assumption for $u = 0$. Hence, $x(u)$ and $\mu(u)$ form a KKT point for Problem (12.2).

It remains to prove that

$$\nabla p(u) = \nabla_u f(x(u)) = \mu(u)$$

holds. To this end, we multiply Equation (12.5a) by $\nabla x(u) \in \mathbb{R}^{p \times n}$ and obtain

$$\nabla x(u)\nabla f(x(u)) - \nabla x(u)\nabla h(x(u))\mu(u) = 0.$$

Moreover, we differentiate (12.5b), i.e.,

$$h(x(u)) = u,$$

which leads to

$$I = \nabla_u h(x(u)) = \nabla x(u)\nabla h(x(u)).$$

Here, I is the $p \times p$ identity matrix. Finally, applying the chain rule yields

$$\nabla p(u) = \nabla_u f(x(u)) = \nabla x(u)\nabla f(x(u)).$$

Combing the last results, we thus obtain

$$\begin{aligned}\nabla p(u) &= \nabla x(u)\nabla f(x(u)) \\ &= \nabla x(u)\nabla h(x(u))\mu(u) \\ &= I\mu(u) = \mu(u).\end{aligned}$$

This completes the proof. □

There is also an analogous sensitivity result for equality and inequality constrained problems. However, since the main concept of sensitivity analysis using Lagrangian multipliers is well covered by the equality constrained case, we do not state and proof the respective result here. If you are interested in the more general result, you can find it in Section 4.3.2 of the book by Bertsekas (2016).

Part IV

Algorithms for Constrained Optimization Problems

13

Quadratic Programming

Optimization problems with a quadratic objective function and affine-linear constraints are called *quadratic programs*. They appear in many applications like in portfolio optimization (see Section 9.1) and also play a very important role in many algorithms for solving general constrained optimization problems, where they often appear as subproblems.

In general, a quadratic program (QP) can be stated as

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \frac{1}{2}x^\top Gx + c^\top x && \text{(13.1a)}\\ \text{s.t.} \quad & a_i^\top x \geq b_i, \quad i \in I = \{1,\dots,m\}, && \text{(13.1b)}\\ & a_j^\top x = b_j, \quad j \in J = \{1,\dots,p\}. && \text{(13.1c)} \end{aligned}$$

Here and in what follows, $G \in \mathbb{R}^{n\times n}$ is a symmetric matrix, c, a_i, $i = 1,\dots,m$, a_j, $j = 1,\dots,p$, are n-dimensional vectors, and the b_i and b_j are real scalars for $i = 1,\dots,m$ and $j = 1,\dots,p$. We call the QP (13.1) convex if the Hessian matrix G of the objective function is positive semi-definite. If G is indefinite, we call the QP nonconvex.

We now discuss equality constrained QPs, i.e., $m = 0$. In this case, we state the problem as

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & q(x) := \frac{1}{2}x^\top Gx + c^\top x && \text{(13.2a)}\\ \text{s.t.} \quad & Ax = b, && \text{(13.2b)} \end{aligned}$$

where $A \in \mathbb{R}^{p\times n}$, $p \leq n$, is the Jacobian of the constraints (with rows $a_j^\top$) and $b \in \mathbb{R}^p$ is the right-hand side vector. We assume that the matrix A has full row rank, which is equivalent to that the LICQ holds for every feasible point. The KKT conditions for a local minimizer x^* of this problem state that there exists a vector $\mu^* \in \mathbb{R}^p$ so that the following linear system of equations is

satisfied:

$$\begin{bmatrix} G & -A^\top \\ A & 0 \end{bmatrix} \begin{pmatrix} x^* \\ \mu^* \end{pmatrix} = \begin{pmatrix} -c \\ b \end{pmatrix}. \tag{13.3}$$

For practical purposes it is often useful the rewrite this system in a slightly different form in which we represent x^* as $x^* = x + d$, where x is a current estimate of the solution and d is the desired step. In this case, the system from above reads

$$\begin{bmatrix} G & A^\top \\ A & 0 \end{bmatrix} \begin{pmatrix} -d \\ \mu^* \end{pmatrix} = \begin{pmatrix} \varphi \\ \xi \end{pmatrix} \tag{13.4}$$

with

$$\xi = Ax - b, \quad \varphi = c + Gx, \quad d = x^* - x.$$

The matrix in (13.4) is called *KKT matrix*. We are interested in conditions under which this matrix is invertible. To this end, let Z be the $n \times (n-p)$ matrix whose $n-p$ columns are a basis of the null space of A. Thus, Z has full (column) rank and it holds $AZ = 0$.

Lemma 13.1. Suppose that $A \in \mathbb{R}^{p \times n}$ has full row rank and assume that the *reduced Hessian matrix* $Z^\top GZ \in \mathbb{R}^{(n-p)\times(n-p)}$ is positive definite. Then, the KKT matrix

$$K = \begin{bmatrix} G & A^\top \\ A & 0 \end{bmatrix} \in \mathbb{R}^{(n+p)\times(n+p)}$$

is invertible. Thus, the system in (13.3) has a unique solution (x^*, μ^*).

Proof. Let w and v be vectors such that

$$\begin{bmatrix} G & A^\top \\ A & 0 \end{bmatrix} \begin{pmatrix} w \\ v \end{pmatrix} = 0$$

holds. Using the second block row $Aw = 0$ and multiplying the last system from the left with $(w^\top, v^\top)$ leads to

$$0 = \begin{pmatrix} w \\ v \end{pmatrix}^\top \begin{bmatrix} G & A^\top \\ A & 0 \end{bmatrix} \begin{pmatrix} w \\ v \end{pmatrix} = w^\top Gw.$$

Since w is in the null space of A, it can be written as $w = Zu$ for some $u \in \mathbb{R}^{n-p}$. Thus, we have

$$0 = w^\top Gw = u^\top Z^\top GZu.$$

Since $Z^\top GZ$ is positive definite, this implies $u = 0$ and, hence, $w = 0$. Consequently, it holds $Gw + A^\top v = A^\top v = 0$. Because A has full row rank, this implies $v = 0$. Thus, the KKT matrix is invertible. □

Under the assumptions of the last theorem, we can show even more.

Theorem 13.2. Suppose that $A \in \mathbb{R}^{p\times n}$ has full row rank and assume that the reduced Hessian matrix $Z^\top GZ \in \mathbb{R}^{(n-p)\times(n-p)}$ is positive definite. Then, the vector x^* satisfying (13.3) is the unique global solution of Problem (13.2).

Proof. Let x be a feasible point, i.e., $Ax = b$ holds, and as before, define $d = x^* - x$. It holds $Ax^* = Ax = b$ and we thus have $Ad = 0$. Substituting $d = x^* - x$ into the objective function we obtain

$$\begin{aligned} q(x) &= \frac{1}{2}(x^* - d)^\top G(x^* - d) + c^\top(x^* - d) \\ &= \frac{1}{2}d^\top Gd - d^\top Gx^* - c^\top d + q(x^*). \end{aligned} \tag{13.5}$$

From (13.3) it follows $Gx^* = -c + A^\top \mu^*$ and also using $Ad = 0$, we obtain

$$d^\top Gx^* = d^\top(-c + A^\top \mu^*) = -d^\top c.$$

Using this relation in (13.5) yields

$$q(x) = \frac{1}{2}d^\top Gd + q(x^*).$$

Since d is in the null space of A, we can write $d = Zu$ for some vector $u \in \mathbb{R}^{n-p}$. Hence,

$$q(x) = \frac{1}{2}u^\top Z^\top GZu + q(x^*)$$

holds. Using the positive definiteness of $Z^\top GZ$ we have $q(x) > q(x^*)$ except when $u = 0$, which implies $d = 0$ and $x^* = x$. Thus, x^* is the unique global solution of Problem (13.2). □

13.1 Direct Solution of the KKT System

One can prove that the KKT matrix in (13.4) is always indefinite if $p \geq 1$ holds. For a symmetric matrix K we define its *inertia* as the triple (n_+, n_-, n_0), where n_+ is the number of positive, n_- is the number of negative, and n_0 is the number of zero eigenvalues of K, i.e.,

$$\text{inertia}(K) = (n_+, n_-, n_0).$$

It holds the following theorem. The proof can be found, e.g., in Forsgren et al. (2002).

Theorem 13.3. Let

$$K = \begin{bmatrix} G & A^\top \\ A & 0 \end{bmatrix} \in \mathbb{R}^{(n+p)\times(n+p)}$$

and suppose that $A \in \mathbb{R}^{p\times n}$ has rank p. Then,

$$\text{inertia}(K) = \text{inertia}(Z^\top G Z) + (p, p, 0)$$

holds. Thus, if $Z^\top G Z \in \mathbb{R}^{(n-p)\times(n-p)}$ is positive definite, it holds $\text{inertia}(K) = (n, p, 0)$.

Having the theorems of the last section at hand and knowing that the KKT matrix is indefinite for $p \geq 1$, we can now derive solution methods for the equality constrained QP (13.2) by solving the KKT system (13.4) in a direct way. One way to do so is to factorize the full $(n+p)\times(n+p)$ KKT matrix, which can be done using the so-called *symmetric indefinite factorization*; see, e.g., Chapter 3 in the Appendix of the book by Nocedal and Wright (2006).

However, there are also other possibilities to solve the KKT system (13.4) that only use the factorization of smaller matrices.

13.2 Schur Complement Method

Assume that G is positive definite. Then G^{-1} exists and we can multiply the first block row of (13.4) with AG^{-1}, which leads to

$$(AG^{-1}A^\top)\mu^* = AG^{-1}\varphi + Ad = AG^{-1}\varphi - \xi. \tag{13.6}$$

We now solve this symmetric and positive definite system for μ^* and then compute d (again using the first block row) by solving

$$Gd = A^\top \mu^* - \varphi. \tag{13.7}$$

In this approach, we perform operations with G^{-1} and we need to compute the factorization of the $p \times p$ matrix $AG^{-1}A^\top$.

Why is the method called the "Schur complement method"? The reason is the following. If we apply a block Gaussian elimination (using G as the block pivot element) to the KKT matrix

$$K = \begin{bmatrix} G & A^\top \\ A & 0 \end{bmatrix},$$

we obtain

$$\begin{bmatrix} G & A^\top \\ 0 & -AG^{-1}A^\top \end{bmatrix}.$$

In linear algebra, the $(2,2)$ matrix block $AG^{-1}A^\top$ is called the *Schur complement of G in the matrix K*. Doing the backsolves with this block upper triangular system we also obtain the steps in (13.6) and (13.7).

13.3 Null-Space Method

For the Schur complement method, we needed the invertibility of the matrix G. The null-space method that we now discuss does not need this assumption. Instead, we require that the constraint matrix A has full row rank and that $Z^\top GZ$ is positive definite. Similar to the Schur complement method, the null-space method also uses the specific block structure of the KKT system for decoupling it into two smaller systems.

First, we decompose the vector d into

$$d = Yd_Y + Zd_Z,$$

where $Z \in \mathbb{R}^{n\times(n-p)}$ is the null-space matrix and $Y \in \mathbb{R}^{n\times p}$ is chosen so that the matrix $[Y|Z]$ is invertible. Hence, $d_Y \in \mathbb{R}^p$ and $d_Z \in \mathbb{R}^{n-p}$. Using this decomposition in the second block row of the KKT system yields

$$(AY)d_Y = -\xi. \tag{13.8}$$

Since A has rank p and $[Y|Z] \in \mathbb{R}^{n\times n}$ is invertible, the product $A[Y|Z] = [AY|0]$ has rank p as well. Thus, $AY \in \mathbb{R}^{p\times p}$ is invertible and d_Y as the solution of (13.8) is well-defined.

Next, we can substitute the decomposition of d into the first block row of the KKT system and obtain

$$-GYd_Y - GZd_Z + A^\top\mu^* = \varphi.$$

Multiplying this equation with $Z^\top$ from the left leads to

$$(Z^\top GZ)d_Z = -Z^\top GYd_Y - Z^\top\varphi.$$

This system can be solved by performing the Cholesky factorization of the positive definite reduced Hessian matrix $Z^\top GZ$. Afterward, we can compute the entire step $d = Yd_Y + Zd_Z$. Finally, we need to determine the Lagrangian multiplier μ^*. To this end, we multiply the first block row of the KKT system with $Y^\top$ from the left and obtain

$$(AY)^\top\mu^* = Y^\top(\varphi + Gd),$$

which can be solved for μ^*.

14

A Primer on Algorithms for Constrained Optimization Problems: Penalty Methods

There are many different types of algorithms for solving general constrained optimization problems. In this lecture, we only give a primer on these methods by discussing some special instantiations of so-called penalty methods. A more detailed discussion of different methods for solving constrained optimization methods is given in the lecture "Numerical Optimization" by Prof. Schulz or in the books by Ulbrich and Ulbrich (2012), Nocedal and Wright (2006), Bertsekas (2016), and Geiger and Kanzow (2002).

The main idea of penalty methods is to move the constraints to so-called penalty terms in an extended objective function. These penalty terms are designed as to penalize the violation of the constraints. The constrained problem is then solved by solving a series of unconstrained penalty problems. Different penalty methods vary in the details on how to penalize the constraint violation in the extended objective function.

14.1 The Quadratic Penalty Method

The most intuitive way of penalizing the constraint violation is by adding a multiple of the sum of the squares of the constraint violation to the original objective function. The resulting penalty function is thus called *quadratic penalty function*.

For the ease of presentation we first present the method using a quadratic

penalty function for equality constrained problems of the form

$$\min_{x\in\mathbb{R}^n} \quad f(x) \tag{14.1a}$$
$$\text{s.t.} \quad h_j(x) = 0, \quad j \in J = \{1,\dots,p\}. \tag{14.1b}$$

Definition 14.1 (Quadratic Penalty Function, Penalty Parameter)**.** The function

$$Q(x;\pi) := f(x) + \frac{\pi}{2}\sum_{j=1}^{p} h_j^2(x), \quad \pi > 0, \tag{14.2}$$

is called the *quadratic penalty function* of Problem (14.1). Its parameter π is called the *penalty parameter*.

The rationale of the quadratic penalty function is that we penalize constraint violations with increasing severity if we drive π to ∞. Thus, the main idea of penalty methods using the quadratic penalty function is to solve a sequence of unconstrained optimization problems with objective function $Q(x;\pi^k)$ for a sequence of penalty parameters $(\pi^k)_k$ with $\pi^k \to \infty$. Since $Q(\cdot,\pi)$ is smooth if f and the h_j, $j = 1,\dots,p$, are smooth, we can apply the methods discussed in Part II to solve the penalty problem in every iteration.

Remark 14.2. One needs to be very careful when choosing the penalty parameters. This can be seen in the following example. Consider the problem

$$\min_{x_1,x_2} \quad -5x_1^2 + x_2^2 \quad \text{s.t.} \quad x_1 = 1.$$

This problem has the unique local and global optimal solution $(1,0)^\top \in \mathbb{R}^2$. However, the corresponding quadratic penalty function

$$Q(x;\pi) = -5x_1^2 + x_2^2 + \frac{\pi}{2}(x_1-1)^2$$

is unbounded for all penalty parameters $\pi < 10$.

The extension of the quadratic penalty function (14.2) to both equality and inequality constrained problems

$$\min_{x\in\mathbb{R}^n} \quad f(x) \tag{14.3a}$$
$$\text{s.t.} \quad g_i(x) \geq 0, \quad i \in I = \{1,\dots,m\}, \tag{14.3b}$$
$$h_j(x) = 0, \quad j \in J = \{1,\dots,p\}, \tag{14.3c}$$

is straightforward. In this case, a reasonable quadratic penalty function is

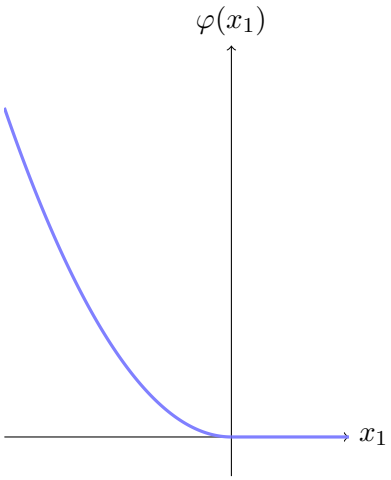


Figure 14.1: The function $\varphi(x_1) := \min\{0, x_1\}^2$.

given by

$$Q(x;\pi) := f(x) + \frac{\pi}{2}\sum_{i=1}^{m}\left([g_i(x)]^-\right)^2 + \frac{\pi}{2}\sum_{j=1}^{p} h_j^2(x), \quad \pi > 0,$$

with

$$[\cdot]^- : \mathbb{R} \to \mathbb{R}, \quad [\alpha]^- := \max\{-\alpha, 0\}.$$

Note that, in this case, Q might be less smooth than the original objective function f and the constraint functions g, h. For example, the constraint $x_1 \geq 0$ (which is perfectly smooth) leads to the penalty term $\max\{-x_1, 0\}^2 = \min\{0, x_1\}^2 =: \varphi(x_1)$, which has the derivatives

$$\varphi'(x_1) = \begin{cases} 0, & x_1 > 0, \\ 2x_1, & x_1 < 0, \end{cases}$$

and

$$\varphi''(x_1) = \begin{cases} 0, & x_1 > 0, \\ 2, & x_1 < 0. \end{cases}$$

Thus, φ is once but not twice continuously differentiable; see also Figure 14.1.

We now state a general framework for algorithms based on the quadratic penalty function (14.2) in Algorithm 6.

Remark 14.3. (a) It is not specified in Algorithm 6 how the penalty

Algorithm 6 Quadratic Penalty Method

Input: Problem (14.1), an initial penalty parameter π^0, a nonnegative sequence $(\tau^k)_k$ with $\tau^k \to 0$, and an initial starting point x_0^s.

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: Find an approximate minimizer x^k of $Q(\cdot; \pi^k)$, starting at x_k^s, and terminating when $\|\nabla_x Q(x; \pi^k)\| \leq \tau^k$.
3: **if** final convergence test is satisfied **then**
4: **return** approximate solution x^k.
5: **end if**
6: Choose new penalty parameter $\pi^{k+1} > \pi^k$.
7: Choose new starting point x_{k+1}^s.
8: **end for**

parameters should be updated. In practice, one usually follows an adaptive strategy. For example, one uses a moderate increase like $\pi^{k+1} = 1.5\pi^k$ if the penalty problem for $Q(\cdot; \pi^k)$ was hard to solve. If this problem was easy to solve, one instead may try a more aggressive update like $\pi^{k+1} = 10\pi^k$.

(b) It is also not stated explicitly how the sequence $(\tau^k)_k$ should be chosen. The only condition is $\tau^k \to 0$ for $k \to \infty$, which ensures that the minimization of the quadratic penalty problems is carried out more accurately in later iterations.

(c) As it is shown in Remark 14.2, it might be the case that a penalty problem is unbounded (and thus unsolvable) for penalty parameters that are not large enough. Practical implementations of quadratic penalty methods try to detect these situations and increase the penalty parameter in such cases.

(d) The quadratic penalty method may get numerically unstable for very large penalty parameters π. A detailed discussion of this issue and possible remedies are discussed in Chapter 17 of the book by Nocedal and Wright (2006).

14.1.1 Convergence of the Quadratic Penalty Method

We now study the convergence properties of the quadratic penalty method for equality constrained problems. In this case, the quadratic penalty function is given in (14.2).

Theorem 14.4. Suppose that each x^k is an exact global minimizer of $Q(\cdot; \pi^k)$ defined by (14.2) in Algorithm 6. Moreover, $\pi^k \to \infty$ holds for $k \to \infty$. Then, every limit point of the sequence $(x^k)_k$ is a global minimizer of Problem (14.1).

Proof. Let $\bar{x}$ be a global minimizer of Problem (14.1). Thus,

$$f(\bar{x}) \leq f(x)$$

holds for all x with $h_j(x) = 0$ for all $j = 1, \ldots, p$. Since x^k is a global minimizer of $Q(\cdot, \pi^k)$ for every k, it also holds

$$Q(x^k; \pi^k) \leq Q(\bar{x}; \pi^k),$$

which is equivalent to the inequality

$$f(x^k) + \frac{\pi^k}{2} \sum_{j=1}^{p} h_j^2(x^k) \leq f(\bar{x}) + \frac{\pi^k}{2} \sum_{j=1}^{p} h_j^2(\bar{x}) = f(\bar{x}). \tag{14.4}$$

Rearranging this inequality leads to

$$\sum_{j=1}^{p} h_j^2(x^k) \leq \frac{2}{\pi^k} \left(f(\bar{x}) - f(x^k) \right). \tag{14.5}$$

Suppose now that x^* is a limit point of $(x^k)_k$. Thus, there is a subsequence (indexed with $\mathcal{K}$) with

$$\lim_{k \in \mathcal{K}} x^k = x^*.$$

Taking the limit $k \to \infty$, $k \in \mathcal{K}$, on both sides of (14.5), we obtain

$$\lim_{k \in \mathcal{K}} \sum_{j=1}^{p} h_j^2(x^k) = \sum_{j=1}^{p} h_j^2(x^*) \leq \lim_{k \in \mathcal{K}} \frac{2}{\pi^k} \left(f(\bar{x}) - f(x^k) \right) = 0,$$

where the last equality follows from $\pi^k \to \infty$. Consequently, $h_j(x^*) = 0$ holds for all $j = 1, \ldots, p$ and x^* is feasible.

We now take the limit $k \to \infty$, $k \in \mathcal{K}$, in (14.4) and use the nonnegativity of π^k and $h_j^2(x^k)$. This leads to

$$f(x^*) \leq f(x^*) + \lim_{k \in \mathcal{K}} \frac{\pi^k}{2} \sum_{j=1}^{p} h_j^2(x^k) \leq f(\bar{x}).$$

Thus, x^* is a feasible point that has an objective function value that is not larger than the one of the global minimizer $\bar{x}$. This implies that x^* is a global minimizer as well. □

Remark 14.5. The assumptions of the last theorem require that we solve every penalty subproblem to global optimality. This desirable property, in

general, cannot be attained in practice—especially in cases in which the penalty functions $Q(\cdot;\pi^k)$ are nonconvex.

In the next theorem, we hence consider the case in which we solve penalty subproblems to stationarity approximately (which is a by far weaker assumption than solving the subproblems to global optimality).

Theorem 14.6. Suppose that the tolerances and the penalty parameters in Algorithm 6 satisfy $\tau^k \to 0$ and $\pi^k \to \infty$. Then, if a limit point x^* of the sequence $(x^k)_k$ is infeasible, it is a stationary point of the function $\|h(\cdot)\|^2$. On the other hand, if a limit point x^* is feasible and if the LICQ is satisfied in x^*, then x^* is a KKT point of Problem (14.1). For such an x^* and any infinite subsequence $\mathcal{K}$ with $\lim_{k\in\mathcal{K}} x^k = x^*$, it holds

$$\lim_{k\in\mathcal{K}} -\pi^k h_j(x^k) = \mu_j^* \quad \text{for all} \quad j = 1,\dots,p, \tag{14.6}$$

where $\mu^* \in \mathbb{R}^p$ is the vector of Lagrangian multipliers so that (x^*,μ^*) satisfies the KKT conditions.

Proof. The gradient of the quadratic penalty function is given by

$$\nabla_x Q(x^k;\pi^k) = \nabla f(x^k) + \sum_{j=1}^{p} \pi^k h_j(x^k)\nabla h_j(x^k)$$

and the termination criterion of Algorithm 6 thus yields

$$\left\| \nabla f(x^k) + \sum_{j=1}^{p} \pi^k h_j(x^k)\nabla h_j(x^k) \right\| \le \tau^k. \tag{14.7}$$

Rearranging this expression and using the inequality $\|a\| - \|b\| \le \|a+b\|$ leads to

$$\left\| \sum_{j=1}^{p} h_j(x^k)\nabla h_j(x^k) \right\| \le \frac{1}{\pi^k}\left(\tau^k + \|\nabla f(x^k)\|\right). \tag{14.8}$$

Let now x^* be a limit point of the sequence of iterates generated by Algorithm 6. Then, there exists a subsequence $\mathcal{K}$ with $\lim_{k\in\mathcal{K}} x^k = x^*$. If we now take the limit $k \to \infty$ for $k \in \mathcal{K}$, the bracketed term on the right-hand side of (14.8) tends to $\|\nabla f(x^*)\|$. Due to $\pi^k \to \infty$, the right-hand side approaches 0. Thus, for the limit of the left-hand side we obtain

$$\sum_{j=1}^{p} h_j(x^*)\nabla h_j(x^*) = 0. \tag{14.9}$$

If the constraint gradients $\nabla h_j(x^*)$ are linearly dependent, we can have $h_j(x^*) \neq 0$ for some $j = 1, \dots, p$. In this case, the limit point x^* is not feasible but is a stationary point of $\|h(\cdot)\|^2$.

If, on the other hand, the LICQ is satisfied at x^*, all constraint gradients are linearly independent and (14.9) implies $h_j(x^*) = 0$ for all $j = 1, \dots, p$. Thus, x^* is feasible. It remains to prove that x^* is also a KKT point and that (14.6) holds. Let

$$A(x)^\top := \nabla h(x) = [\nabla h_j(x)]_{j=1}^p \in \mathbb{R}^{n \times p}$$

denote the matrix of the constraint gradients, i.e., the Jacobian, and set $\mu^k := -\pi^k h(x^k)$. Using this notation, we have

$$A(x^k)^\top \mu^k = \nabla f(x^k) - \nabla_x Q(x^k; \pi^k) \quad \text{and} \quad \|\nabla_x Q(x^k; \pi^k)\| \leq \tau^k. \tag{14.10}$$

If $k \in \mathcal{K}$ is sufficiently large, the matrix $A(x^k)$ has full row rank and, thus, $A(x^k)A(x^k)^\top$ is invertible. Multiplying (14.10) with $A(x^k)$ from the left yields

$$\mu^k = \left(A(x^k)A(x^k)^\top\right)^{-1} A(x^k) \left(\nabla f(x^k) - \nabla_x Q(x^k; \pi^k)\right).$$

Consequently, taking the limit $k \to \infty$, $k \in \mathcal{K}$, leads to

$$\lim_{k \in \mathcal{K}} \mu^k = \mu^* = \left(A(x^*)A(x^*)^\top\right)^{-1} A(x^*) \nabla f(x^*).$$

Taking the limit in (14.7) as well leads to

$$\nabla f(x^*) - A(x^*)^\top \mu^* = 0.$$

Thus, (x^*, μ^*) is a KKT point of Problem (14.1). □

Remark 14.7. In contrast to the result in Theorem 14.4, the weaker assumptions of Theorem 14.6 do not exclude the situation that a limit point of the sequence of iterates may be infeasible. However, in this unlucky situation we, at least, obtain a stationary point of a reasonable measure of infeasibility. Moreover, we only obtain KKT points instead of global minimizers in the case that the limit point is feasible and LICQ is satisfied.

14.2 Exactness and the ℓ_1 Penalty Method

In the context of penalty methods, the notion of exact penalty functions is of special importance. This property implies that, for a certain value of the penalty parameter π, a single minimization w.r.t. x can yield the

exact solution of the original nonlinear optimization problem. The quadratic penalty function of the last section is not exact.

In what follows, we discuss nonsmooth and exact penalty functions. A very popular nonsmooth penalty function is the ℓ_1 penalty function, which is defined by

$$\phi_1(x;\pi) := f(x) + \pi \sum_{i=1}^{m} [g_i(x)]^- + \pi \sum_{j=1}^{p} |h_j(x)|,$$

where, again, $[\alpha]^- := \max\{0, -\alpha\}$. The name of this penalty function is motivated by the fact that the penalty term is π multiplied with the ℓ_1 norm of the constraint violation. Note that ϕ_1 is not differentiable at certain points because of the absolute value function $|\cdot|$ and the $[\cdot]^-$ function.

The following result states the exactness of the ℓ_1 penalty function.

Theorem 14.8. Suppose that x^* is a strict local solution of Problem (14.3) and that the KKT conditions are satisfied for x^* and respective Lagrangian multipliers λ^*, μ^*. Then, x^* is a local minimizer of $\phi_1(x;\pi)$ for all $\pi > \pi^*$ with

$$\pi^* = \max\{\|\lambda^*\|_\infty, \|\mu^*\|_\infty\}.$$

If, in addition, the second-order sufficient conditions of Theorem 11.3 and $\pi > \pi^*$ hold, then x^* is a strict local minimizer of $\phi_1(x;\pi)$.

Proof. The proof can be found in Han and Mangasarian (1979). □

Example 14.9. Consider the problem

$$\min_{x\in\mathbb{R}} \quad x \quad \text{s.t.} \quad x \geq 1,$$

which has the solution $x^* = 1$. It obviously holds

$$\phi_1(x;\pi) = x + \pi[x-1]^- = \begin{cases} (1-\pi)x + \pi, & x \leq 1, \\ x, & x > 1. \end{cases}$$

As it can be seen in Figure 14.2, the penalty function has a minimizer at $x^* = 1$ for $\pi > 1$ but is monotonically increasing for $\pi < 1$.

Common to all penalty methods is that they minimize the penalty function. Thus, we need to characterize stationary points of ϕ_1. However, ϕ_1 is not differentiable and we can thus not apply our standard definition of stationary

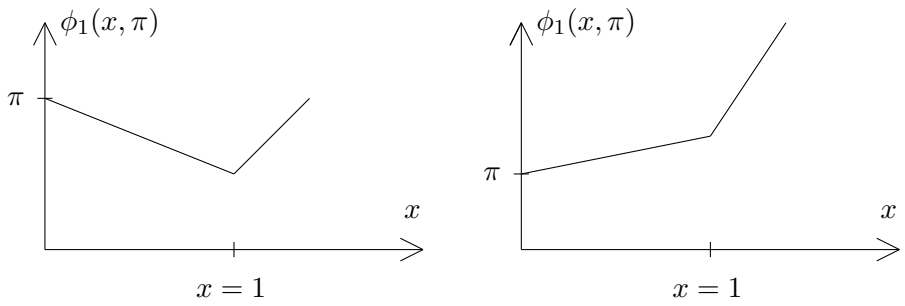


Figure 14.2: The ℓ_1 penalty function of Example 14.9 for $\pi > 1$ (left) and $\pi < 1$ (right)

points to ϕ_1 directly. Fortunately, the directional derivative exists for every direction d.

Definition 14.10 (Directional Derivative)**.** The *directional derivative* of a function $f : \mathbb{R}^n \to \mathbb{R}$ in the direction d is given by

$$D(f(x); d) := \lim_{\varepsilon \to 0} \frac{f(x + \varepsilon d) - f(x)}{\varepsilon}.$$

With this definition at hand, we can now define stationary points of the penalty function ϕ_1.

Definition 14.11 (Stationary Points of the Penalty Function ϕ_1)**.** A point $x^* \in \mathbb{R}^n$ is a *stationary point* of the penalty function ϕ_1 if

$$D(\phi_1(x^*; \pi); d) \geq 0 \quad \text{for all} \quad d \in \mathbb{R}^n.$$

In analogy, x^* is a stationary point of the ℓ_1 measure of infeasibility

$$\chi(x) := \sum_{i=1}^{m} [g_i(x)]^- + \sum_{j=1}^{p} |h_j(x)|$$

if $D(\chi(x^*); d) \geq 0$ for all $d \in \mathbb{R}^n$. If a point is infeasible for Problem (14.3) and stationary w.r.t. the infeasibility measure χ, we call this point an *infeasible stationary point*.

Example 14.12. For the function in Example 14.9, it holds

$$D(\phi_1(x^*; \pi); d) = \begin{cases} d, & \text{if } d \geq 0, \\ (1 - \pi)d, & \text{if } d < 0, \end{cases}$$

for $x^* = 1$. Thus, if $\pi > 1$ holds, we have $D(\phi_1(x^*;\pi);d) \geq 0$ for all $d \in \mathbb{R}$.

The next result shows that (under certain assumptions) stationary points of the ℓ_1 penalty function ϕ_1 correspond to KKT points of the original problem (14.3).

Theorem 14.13. Suppose that x^* is a stationary point of the penalty function $\phi_1(x;\pi)$ for all π greater than a certain threshold $\hat{\pi} > 0$. Then, if x^* is feasible for the original problem (14.3), it satisfies the KKT conditions of Problem (14.3). If x^* is not feasible for Problem (14.3), it is an infeasible stationary point.

Proof. We start with the case in which x^* is feasible. From Definition 14.10 and the definition of ϕ_1, it follows[1]

$$D(\phi_1(x^*;\pi);d) = \nabla f(x^*)^\top d + \pi \sum_{i \in I(x^*)} \left[\nabla g_i(x^*)^\top d\right]^- + \pi \sum_{j=1}^{p} \left|\nabla h_j(x^*)^\top d\right|.$$

We now consider any direction d in the linearized tangential cone $T_{\text{lin}}(x^*)$ and obtain

$$\sum_{i \in I(x^*)} \left[\nabla g_i(x^*)^\top d\right]^- + \sum_{j=1}^{p} \left|\nabla h_j(x^*)^\top d\right| = 0.$$

Thus, since x^* is a stationary point of ϕ_1 by assumption, we have

$$0 \leq D(\phi_1(x^*;\pi);d) = \nabla f(x^*)^\top d \quad \text{for all} \quad d \in T_{\text{lin}}(x^*).$$

Using Farkas' Lemma 10.9 now yields

$$\nabla f(x^*) = \sum_{i \in I(x^*)} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{p} \mu_i^* \nabla h_j(x^*)$$

for some coefficients $\lambda_i^* \geq 0$ for all $i \in I(x^*)$ as well as μ_j^* for $j = 1, \ldots, p$.

The part of the proof in which x^* is infeasible is left as an exercise. □

Remark 14.14. The ℓ_1 penalty problem

$$\min_{x \in \mathbb{R}^n} \ \phi_1(x;\pi) = f(x) + \pi \sum_{i=1}^{m} [g_i(x)]^- + \pi \sum_{j=1}^{p} |h_j(x)|$$

[1] Why?

of the equality and inequality constrained problem

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \geq 0, \quad i \in I = \{1,\dots,m\}, \\ & h_j(x) = 0, \quad j \in J = \{1,\dots,p\}, \end{aligned}$$

can also be reformulated in a smooth way if we accept to solve a constrained optimization problem in every iteration. The penalty problem then reads

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) + \pi \sum_{i\in I} s_i + \pi \sum_{j\in J} \left(s_j^+ + s_j^- \right) \\ \text{s.t.} \quad & g_i(x) + s_i \geq 0, \quad i \in I = \{1,\dots,m\}, \\ & h_j(x) + s_j^+ - s_j^- = 0, \quad j \in J = \{1,\dots,p\}, \\ & s_i \geq 0,\, i \in I, \quad s_j^+, s_j^- \geq 0,\, j \in J. \end{aligned}$$

Part V

Miscellaneous

15

What you should know now!

1. Different variants of stating a general optimization problem
2. Minimization vs. maximization: is there a real difference?
3. What is a feasible point?
4. What is a (strict) local minimizer?
5. What is a (strict) global minimizer?
6. What is the gradient?
7. What is the Hessian matrix?
8. What is a stationary point?
9. How can you classify an optimization problem?
10. What is an infimum?
11. When do we call an optimization problem solvable?
12. Can you characterize the solvability of an optimization problem?
13. How many and what reasons exist for unsolvability?
14. What is the statement of the theorem of Weierstraß?
15. What is the lower level set?
16. What is the relation of the lower level set with the set of all global minimizers?
17. What is the claim of the improved version of the theorem of Weierstraß?
18. What is the least-squares problem?
19. What is the statement of the necessary first-order optimality condition for unconstrained problems?

20. What is a saddle point?
21. What is the statement of the necessary second-order optimality condition for unconstrained problems?
22. What is the statement of the sufficient second-order optimality condition for unconstrained problems?
23. What is a convex set?
24. What is a (strictly, uniformly) convex function?
25. How can we characterize a convex function using its gradient?
26. How can we characterize a convex function using its Hessian matrix?
27. How many minimizers exist for a strictly convex function?
28. What do you know about a local minimizer of a convex function?
29. What do you know about a stationary point of a convex function?
30. What is the main idea of descent methods in general and of the gradient method in particular?
31. Can you write down a generic descent method with abstract conditions on the search direction and on the step sizes?
32. How are termination criteria realized in practice and why?
33. What is a direction of descent?
34. Is it enough for a direction being a descent direction that the function decreases along this direction? If not, why not?
35. What is a steepest descent direction?
36. How can we compute a steepest descent direction? Can you prove that this is a steepest descent direction? What is the important inequality used in the proof?
37. What is the Armijo rule? What is the role of γ? Can you draw a picture that illustrates the rule?
38. What is required to prove in order to show that the Armijo rule is "well defined"?
39. What is the statement of the global convergence theorem for the gradient method?
40. What is the geometric interpretation of the gradient method? What do you know about the relation of successive search directions?

41. What do you know about the speed of convergence of the gradient method?
42. What is the relation between the condition number of the Hessian of the objective function and the speed of convergence of the gradient method applied to strictly convex and quadratic objective functions? What is the geometric intuition behind this relationship?
43. What is the rate of convergence of the gradient method for strictly convex and quadratic functions?
44. What is the idea behind the iterative scheme of the Newton method and how does it relate to Taylor's theorem?
45. Can you formally state Newton's method?
46. What different types of convergence rates of sequences do you know? Can you formally define them? Which convergence rate implies another convergence rate?
47. What is the statement of Banach's lemma?
48. What do you know about a root of a function if the Jacobian at this point is invertible?
49. What is a Lipschitz continuous function?
50. What do you know about local convergence of Newton's method for systems of equations?
51. How is Newton's method for optimization problems derived? How many derivations do you know?
52. What do you know about local convergence of Newton's method for unconstrained optimization problems?
53. Do you know a counterexample for global convergence?
54. How does the damped method of Newton look like and what do you know about its convergence properties?
55. Can you explain the definition of the projection on a non-empty, closed, and convex set with a picture?
56. What is the statement of the projection theorem?
57. Can you explain the strict separation theorem with a picture?
58. What is cone? Do you know an example for a convex and a nonconvex cone?
59. What is the statement of Farkas' lemma?

60. What is the tangential cone? Explain it using pictures.
61. Can you prove that the tangential cone is indeed a cone?
62. How can the tangential cone be used to state a first-order necessary condition for constrained optimization problems?
63. What is a stationary point of a constrained optimization problem?
64. What is the set of active inequalities? Why is it important?
65. What is the linearized tangential cone? What is the motivation for introducing this cone?
66. Is the tangential cone always a subset of the linearized tangential cone or is it the other way around?
67. What is a constraint qualification?
68. Which constraint qualifications do you know?
69. Is the ACQ always satisfied?
70. What is the Lagrangian function of a constrained optimization problem?
71. Write down the KKT conditions! What is a KKT point? Can you draw a picture that can be used to motivate the KKT conditions geometrically?
72. What versions of the KKT theorem do you know? What is the difference between these versions? How can we prove these theorems?
73. What is the MFCQ?
74. Is the tangential cone open or closed?
75. What is the relation between the ACQ and the MFCQ?
76. What is the LICQ? What is the motivation for considering this CQ?
77. Can you prove the KKT theorem under the LICQ if you are allowed to use the KKT theorem under the MFCQ?
78. What is the relation between the MFCQ and the LICQ?
79. What do you know about the Lagrangian multipliers of a local minimizer in which the MFCQ (the LICQ) holds.
80. What do you know about the KKT conditions of problems with linear constraints?
81. Do you know a CQ that is important for convex problems? Why is it important?
82. What do you know about KKT points of convex problems?

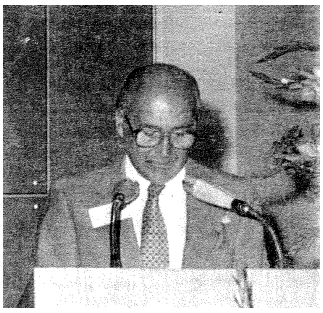
83. What is a saddle point? Why is it interesting to consider saddle points?
84. What do you know about convex problems with linear constraints?
85. What are Fritz–John conditions? What is the relation between the KKT and the FJ conditions?
86. Why is a FJ point with $r^* > 0$ of special importance? Is there a theorem without CQs that ensures the existence of a FJ point with $r^* > 0$?
87. For deriving second-order conditions for constrained optimization, we further partitioned the index set of active inequality constraints. How and why?
88. How do we define $\mathcal{T}_i(x^*)$, $i = 1, 2, 3$, and why?
89. What is the relation between $\mathcal{T}_i(x^*)$, $i = 1, 2, 3$?
90. What is the statement of the necessary second-order optimality conditions?
91. What is the statement of the sufficient second-order optimality conditions?
92. What is sensitivity analysis?
93. What is the main sensitivity result for equality constrained problems using a simple example?
94. Can you explain the main result of sensitivity analysis for equality constrained problems using a simple example?
95. What is the KKT matrix/system? How can this matrix be derived?
96. What is the reduced Hessian matrix?
97. What do you know about the solutions of the KKT system? What about uniqueness and/or global optimality?
98. What do you know about the eigenvalues of the KKT matrix?
99. What is the inertia of a matrix?
100. What do you know about the inertia of the KKT matrix? What is the assumption required for the corresponding theorem?
101. Can you explain the Schur complement method? What are the several steps of this method? How large are the linear systems that have to be solved?
102. Why does it have this name?
103. Why is the null-space method called null-space method?

104. What are the steps in the null-space method? What are the linear systems to be solved?
105. What are the main differences between the Schur complement method and the null-space method?
106. What is the main idea of penalty methods?
107. How do we define the quadratic penalty function for equality constrained problems?
108. How do we define the quadratic penalty function for equality and inequality constrained problems?
109. Is the quadratic penalty problem always well-defined?
110. Is the quadratic penalty function for equality and inequality constrained problems always smooth?
111. Can you state and explain the main framework for quadratic penalty methods?
112. What do you know about the convergence of the quadratic penalty method? We proved two theorems. What is the difference between these two theorems?
113. What is the crucial role of the LICQ in the second theorem?
114. What is the motivation to study ℓ_1 penalty functions?
115. How is the ℓ_1 penalty function defined?
116. What does exactness of a penalty function mean?
117. What is the directional derivative?
118. What is a stationary point of the ℓ_1 penalty function?
119. What do you know about stationary points of the ℓ_1 penalty function if the penalty function parameter is large enough?

16

The Mathematicians Behind This Lecture

Jean M. Abadie



Life

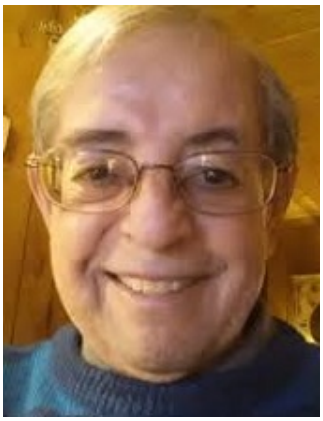
- ⋆ October 19, 1919 in Mirande, France
- † November 9, 2014 in Paris, France
- 1948 Diploma of Advanced Studies
- 1950 Training as Professeur Stagiaire at the University of Paris
- from 1950 Statistician at the Water & Forestry Research Center, Nancy, France
- from 1955 Head of Mathematics group of Électricité de France
- from 1959 Professor of Statistics at the University of Paris
- 1978-1983 French representative to special commission on systems science at NATO

- Member of the American Mathematics Society (AMS) and Société mathématique de France (SMF)
- 1980-1983 Chairman of the Mathematics Programming Society

Scientific Achievements

- Abadie Constraint Qualification

Larry Gandara Armijo



Portrait is not verified.

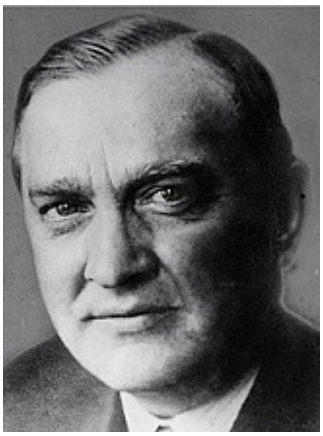
Life

- ⋆ May 25, 1955 in New York City
  (Date of birth in all likelihood incorrect)
- † September 3, 1993 in Oxnard, California

Scientific Achievements

- 1962 PhD thesis "Generalizations of Convexity for Functions of One Variable"
- In his three-page article "Minimization of functions having Lipschitz continuous first partial derivatives" (Pacific Journal of Mathematics, 1966), he first used the step size strategy for unconstrained optimization problems today known as the Armijo rule.

Stefan Banach



Life

- ⋆ March 30, 1892 in Krakow
- † August 31, 1945 in Lviv
- 1911–1913 Studies of mechanics at the Polytechnic National University in Lviv, Pre-diploma, did not complete his studies
- First World War: supervisor at road construction
- Auto-didactic studies of mathematics (and partly at the University of Krakow)
- 1922 Doctorate in mathematics at the University of Lviv with the thesis "Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales", without having obtained any degree in mathematics beforehand
- The doctoral thesis founded functional analysis
- also 1922 Habilitation with a thesis on measure theory, associate professor in Lviv
- 1924/25 Research stay in Paris, work in the "Scottish Café"
- 1927 Tenured professor in Lviv
- 1934 Corresponding member of the Polish Academy of Sciences and Fine Arts in Krakow
- 1939—1941 Dean of the faculty of Mathematics and Physics in Lviv

Scientific Achievements

- Banach space
- Banach algebra
- Hahn-Banach theorem
- Banach-Alaoglu theorem
- Banach fixed-point theorem
- Banach's lemma

Bernard Bolzano

Life

- ⋆ October 5, 1781 in Prague
- † December 18, 1848 in Prague
- Catholic priest, philosopher, and mathematician
- Studies of philosophy, mathematics, and physics at the Charles University in Prague
- 1804/1805 Doctorate in theology and ordination to the priesthood
- 1806 Tenured professor at the chair of philosophy of religion at the Charles University in Prague
- 1815 Member of the Royal Bohemian Society of Sciences
- 1818 Dean of the Philosophical Faculty at the University of Prague and director of the Department of Natural Sciences at the Royal Bohemian Society of Sciences

Scientific Achievements

- Bolzano-Weierstrass theorem
- Bolzano function
- Introduced Cauchy sequences four years before Cauchy
- Intermediate value theorem

Augustin-Louis Cauchy

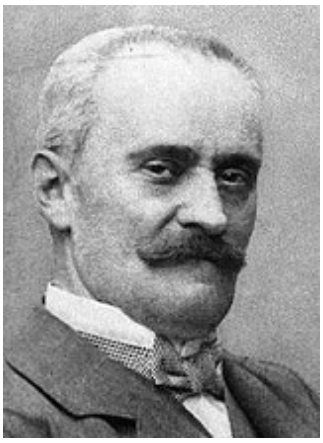
Life

- ⋆ August 21, 1789 in Paris
- † May 23, 1857 in Sceaux
- 1805 Studies of engineering at the École Polytechnique
- 1810 Participation in the construction of Port Napoléon
- 1815 Grand Prix of the French Academy of Sciences for his work on wave propagation, Assistant professor at the École Polytechnique
- 1816 Member of the Académie des Sciences
- A total of nearly 800 publications

Scientific Achievements

- Co-founder of modern analysis
- Founder of the theory of elasticity
- Cauchy sequence
- Cauchy criterion
- Cauchy-Riemann differential equations
- Integral formula of Cauchy
- Cauchy problem
- Cauchy-Schwarz inequality

Gyula (Julius) Farkas



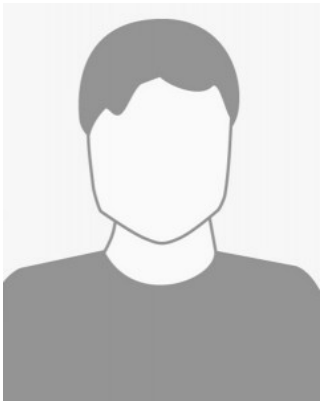
Life

- ⋆ March 28, 1847 in Sárosd
- † December 27, 1930 in Pestszentlőrinc
- Abandoned his studies of music and law in Pest
- Studies of physics and chemistry in Budapest
- 1874 Tutoring the children of Géza Batthyány, Count of Polgárdi
- Devoted himself to research in mathematics and physics during this time
- 1880 Appointment to the University of Budapest as a lecturer in Function Theory
- 1887 Professor at the University of Kolozsvár (nowadays Cluj-Napoca)
- 1898 Member of the Hungarian Academy of Sciences

Scientific Achievements

- Farkas' lemma

Stanley Fromowitz



Life

- ⋆ 1936 in Poland
- † between March 13 and April 1, 2017 in Maryland
- Studies at the University of Toronto until 1960
- Doctorate at Stanford University in the field of statistics
- Worked for the Shell Development Corporation
- Taught from 1971 until 2001 at the University of Maryland

Scientific Achievements

- Mangasarian–Fromowitz constraint qualification

Ludwig Otto Hesse

Life

- ⋆ April 22, 1811 in Königsberg
- † August 4, 1874 in Munich
- 1833–1837 Studies at the Albertus-University Königsberg (among others under Carl Gustav Jacob Jacobi)
- from 1832 Active in Corps Masovia
- 1837 Senior teacher exam in mathematics and physics
- 1840 Doctorate under Jacobi
- 1856 Corresponding member of the Göttingen Academy of Sciences
- 1859 Corresponding member of the Prussian Academy of Sciences
- 1869 Associate member of the Bavarian Academy of Sciences

Scientific Achievements

- Hessian matrix

Carl Gustav Jacob Jacobi

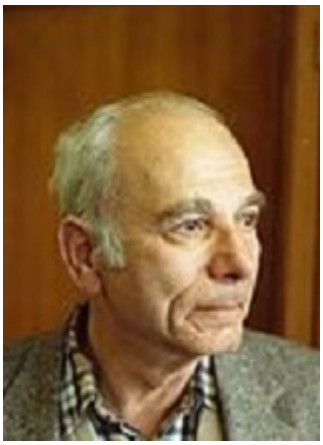
Life

- ⋆ December 10, 1804 in Potsdam
- † February 18, 1851 in Berlin
- 1821 Studies of mathematics, philosophy, and philology at the University of Berlin
- 1824 Senior teacher exam in Latin, Greek, and mathematics
- 1825 Doctorate
- 1825/26 Habilitation with inaugural lecture on differential geometry
- 1826-1843 Professorship at the University Königsberg
- Member of the Prussian Academy of Sciences

Scientific Achievements

- Jacobian matrix
- Jacobi method
- Theory of elliptic functions

Fritz John



Life

- ⋆ June 14, 1910 in Berlin
- † February 10, 1994 in New Rochelle, New York
- 1929–1933 Studies in Göttingen
- 1933 Emigrated to England after Hitler came to power
- 1934 Doctorate at the University of Göttingen under Courant
- 1935 Emigrated to the USA, Assistant professor at the University of Kentucky
- 1941 US citizenship
- 1946 Associate Professor at New York University
- 1978 Courant Chair at the Courant Institute of Mathematical Sciences at New York University
- 1981 Emeritus

Scientific Achievements

- Fritz-John conditions

Leonid Witaljevich Kantorovich

Life

- ⋆ January 19, 1912 in Saint Petersburg
- † April 7, 1986 in Moscow
- 1926–1930 Studies of mathematics at the Leningrad State University
- 1930 Doctorate
- 1934 Professor of mathematics at the Leningrad State University
- 1935 Soviet doctoral degree (equivalent to a habilitation)
- 1975 Nobel Prize in Economics (shared with Tjalling Koopman)

Scientific Achievements

- Founder of linear programming
- Kantorovich inequality

William Karush

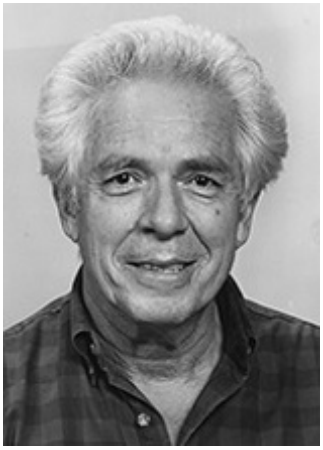
Life

- ⋆ March 1, 1917 in Chicago
- † February 22, 1997 in California
- Studies of mathematics at the University of Chicago
- 1942 Doctorate at the University of Chicago under Magnus Hestenes
- Participation in the Manhattan Project
- Signatory of the Szilard petition
- 1945 Instructor and Associate Professor at the University of Chicago
- Later worked in the field of operations research of the industrial complex
- 1967 Professor at the California State University
- 1987 Emeritus

Scientific Achievements

- 1939 Master's thesis "Minima of Functions of Several Variables with Inequalities as Side Constraints"; Karush is the first "K" in "KKT".

Harold William Kuhn



Life

- ⋆ July 29, 1925, in Santa Monica
- † July 2, 2014, in New York City
- 1950 PhD at Princeton University
- 1950/51 Fulbright Scholar at the University of Paris
- 1952–1959 Employed at Bryn Mawr College
- From 1959 Professor of Mathematics and Economics in Princeton until, he became an emeritus professor in 1995

Scientific Achievements

- Kuhn is the second “K” in “KKT”
- Hungarian method

Joseph-Louis de Lagrange

Life

- ⋆ January 25, 1736, in Turin
- † April 10, 1813, in Paris
- Within a year he acquired all the knowledge of a fully trained mathematician of his time
- At the age of 19 he was hired as a professor in mathematics at the Royal Artillery School in Turin
- 1757 Co-founder of the Turin Academy
- 1766 Director of the Royal Prussian Academy of Sciences in Berlin (as successor to Euler)
- 1787 Retired from the Académie des sciences in Paris
- 1801 Foreign member of the Göttingen Academy of Sciences
- 1808 Foreign member of the Bavarian Academy of Sciences

Scientific Achievements

- Lagrangian polynomial
- Lagrange multipliers
- Lagrangian formalism in mechanics

Rudolf Otto Sigismund Lipschitz

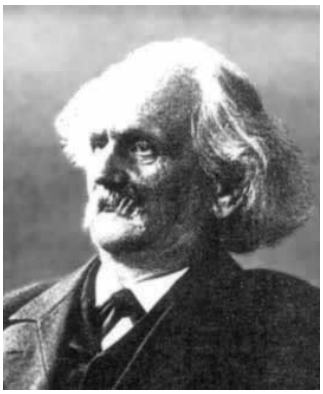
Life

- ⋆ May 14, 1832, in Koenigsberg
- † October 7, 1903, in Bonn
- 1847 Studied mathematics at the University of Königsberg
- 1853 PhD at the University of Berlin
- 1857 Privatdocent in Berlin
- 1862 Extraordinary professor at the University of Breslau
- 1874/75 Rector of the University of Bonn
- 1883 Member of the Leopoldina

Scientific Achievements

- Lipschitz continuity
- Lipschitz constant

Carl Gottfried Neumann



Life

- ⋆ May 7, 1832, in Königsberg (Prussia)
- † March 27, 1925, in Leipzig
- 1850–1855 studies in Koenigsberg
- 1856 Promotion to Dr. phil. in mathematics at the university Königsberg with the work "De problemate quodam mechanico, quod ad primam integralium ultraellipticorum classem revocatur"
- 1858 Habilitation in mathematics at the university Halle-Wittenberg with the thesis "Explicare tentatur, quomodo fiat, ut lucis planum polarizationis per vires electricas vel magnetic as declinetur"
- 1858–1863 Privatdocent in mathematics at the United Friedrichs University Halle-Wittenberg
- 1863–1863 Associate Professor of Mathematics at the United Friedrichs University Halle-Wittenberg
- 1863–1865 Professor of Mathematics at the University of Basel
- 1865–1868 Professor of Mathematics at the University of Tubingen
- 1868–1911 Professor of Mathematics at the Faculty of Philosophy at the University of Leipzig

Scientific Achievements

- Neumann series
- Neumann boundary condition

Olvi Mangasarian

Life

- ⋆ January 12, 1934, in Baghdad
- Degree in Electrical Engineering from Princeton University
- 1959 PhD at Harvard University
- Employment at Shell Development Corporation
- 1965–1967 Lecturer at the University of California
- From 1969 Professor at the University of Wisconsin–Madison

Scientific Achievements

- Mangasarian–Fromowitz constraint qualification (MFCQ)

Harry M. Markowitz

Life

- ⋆ August 24, 1927 in Chicago, IL
- 1947 Bachelor's degree in philosophy at the University of Chicago
- 1950 Master's degree in economics at the University of Chicago (student, among others, of Milton Friedman)
- 1952 Worked at the RAND Corporation, where he met George Dantzig
- 1954 Ph.D. in economics at the University of Chicago
- 1955–1956 Worked at the Cowles Foundation
- 1962 Co-founded CACI International with Herb Karr
- from 1968 Professor at University of California-Los Angeles, Wharton University, and Rutgers University, as well as visiting professor at the University of Tokyo, London School of Economics, and London Business School
- 1989 John von Neumann Theory Prize
- 1990 Nobel Memorial Prize in Economic Sciences
- From 2007 on: Adjunct professor at the Rady School of Management at University of California-San Diego
- Fellow of the Institute for Operations Research and the Management Sciences, the American Academy of Arts and Sciences, and the Econometric Society

Scientific Achievements

- Founder of modern portfolio theory
- Sparse matrix methods

- Simulation language programming SIMSCRIPT

Isaac Newton

Life

- ⋆ January 4, 1643, at Woolsthorpe-by-Colsterworth
- † March 31, 1727, in London
- Attended Trinity College, Cambridge, at the age of 18
- This was temporarily closed during the Great Plague, whereupon he returned to his parent's house, where he worked on problems of Optics, algebra and mechanics
- 1667 Fellow of Trinity College
- 1672 Ground-breaking paper on the refraction of light through a prism
- 1699 Master of the Royal Mint in London
- 1703 President of the Royal Society
- 1705 Knighted

Scientific Achievements

- Newton's method

Georg Friedrich Bernhard Riemann

Life

- ⋆ September 17, 1826, in Breselenz
- † July 20, 1866, in Selasca
- Studied mathematics in Göttingen and Berlin, among others with Carl Friedrich Gauss
- 1851 PhD under Gauss
- 1854 Habilitation
- 1857 Associate professorship

Scientific Achievements

- Riemann integral
- Riemann subtotals
- Riemann problem
- Cauchy-Riemann differential equations

Issai Schur

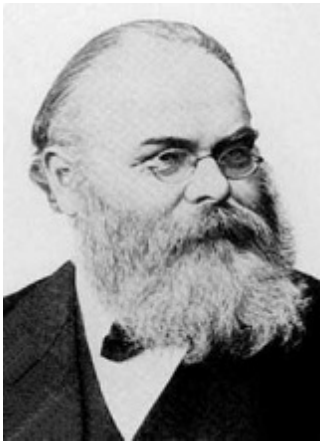
Life

- ⋆ January 10, 1875, in Mogilev, Belarus
- † January 10, 1941, in Tel Aviv, Israel
- 1894 Studied mathematics and physics at the University of Berlin
- 1901 PhD with Ferdinand Georg Frobenius and Lazarus Immanuel Fuchs
- 1903 Habilitation and privatdocent at the University of Berlin
- 1913 Professor and successor of Felix Hausdorff in Bonn
- 1919 Member of the Leopoldina
- 1922 Member of the Prussian Academy of Sciences
- 1933 Takeover by the National Socialists, which led to the dismissal of Schur from the university system

Scientific Achievements

- Schur complement

Hermann Amandus Schwarz



Life

- ⋆ January 25, 1843, in Hermsdorf
- † November 30, 1921, in Berlin
- 1864 PhD in Berlin
- 1866 Habilitation and privatdocent
- 1867–1869 Associate professor in Halle
- 1869 Associate professor at ETH Zurich
- 1875 Full professor at the University of Göttingen
- 1892 Full professor at the Berlin Friedrich-Wilhelms-University and full member of Prussian Academy of Sciences
- 1885 Member of the Leopoldina
- 1897 Corresponding member of the Russian Academy of Sciences in Saint Petersburg
- 1912 Corresponding member of the Bavarian Academy of sciences

Scientific Achievements

- Schwarz theorem

Morton Lincoln Slater

Life

- ⋆ March 14, 1921, in New York, NY, United States
- † May 24, 2002, Fort Worth, TX, United States
- Studies at the University of Wisconsin
- 1949 PhD at Harvard University
- Professor at Texas Christian University
- 1975 Chair of the Admissions Committee at Sophie Davis School
- 1986 Foundation of Gateway to Higher Education Program

Scientific Achievements

- Slater CQ

Brook Taylor

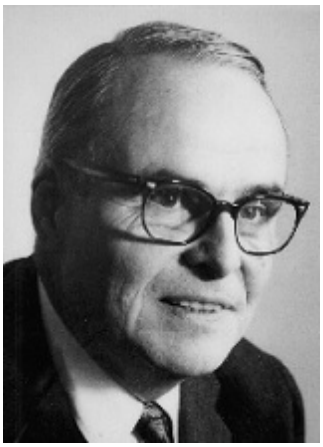
Life

- ⋆ August 18, 1685, in Edmonton
- † December 29, 1731, in London
- Mathematics studies at Cambridge
- 1715 Publication of his major work "Methodus Incrementorum Directa et Inversa", which deals with singular Solutions of differential equations and the vibrating string based on mechanical principles

Scientific Achievements

- Taylor's theorem
- Taylor series

Albert William Tucker



Life

- ⋆ November 28, 1905, in Oshawa
- † January 25, 1995, in New Jersey
- Studied mathematics at the University of Toronto
- 1932 PhD at Princeton University
- 1932-1933 National Research Fellow at Harvard University and at the University of Chicago
- 1934 Assistant Professor at Princeton University
- 1938 Associate Professor at Princeton University
- 1946–1974 Full Professor at Princeton University

Scientific Achievements

- The "T" in "KKT"
- PhD supervisor of John Nash

Karl Theodor Wilhelm Weierstrass

Life

- ⋆ October 31, 1815, in Ostenfelde
- † February 19, 1897, in Berlin
- 1834–1838 Studied law and finance at the Rheinische Friedrich-Wilhelms-University in Bonn
- 1838–1840 Studied mathematics and physics at the Academy in Muenster
- From 1841/42 various jobs as a teacher; next to mathematics also for botany and gymnastics
- First mathematical publication that really received attention in 1854 Crelles Journal: “ Zur Theorie der Abelschen Functionen”
- 1856 Honorary doctor’s degree from the Albertus University of Königsberg for that publication only
- From 1856 professor at the Friedrich-Wilhelms-University Berlin

Scientific Achievements

- Theorem of Bozen-Weierstrass
- Weierstrass theorem
- Weierstraß convergence theorem
- Weierstrass majorant criterion

Index

Bibliography

Abbott, Stephen (2015). *Understanding Analysis*. 2nd ed. Springer-Verlag New York. DOI: 10.1007/978-1-4939-2712-8.

Adams, Douglas (1979). *The Hitchhiker's Guide to the Galaxy*.

Ben-Tal, A., L. El Ghaoui, and A. Nemirovski (2009). *Robust Optimization*. Princeton University Press. ISBN: 9781400831050.

Bertsekas, Dimitri P. (2016). *Nonlinear Programming*. Athena scientific Belmont. ISBN: 978-1-886529-05-2.

Birge, John R. and Francois Louveaux (2011). *Introduction to Stochastic Programming*. Springer Science & Business Media. DOI: 10.1007/978-1-4614-0237-4.

Chvátal, Vasek (1983). *Linear Programming*. A Series of books in the mathematical sciences. New York (N. Y.): Freeman. ISBN: 0-7167-1195-8.

Forsgren, A., P. Gill, and M. Wright (2002). "Interior Methods for Nonlinear Optimization." In: *SIAM Review* 44.4, pp. 525–597. DOI: 10.1137/S0036144502414942.

Geiger, Carl and Christian Kanzow (2002). *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer. ISBN: 9783540427902.

– (2013). *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*. Springer-Verlag.

Han, S.-P. and O. L. Mangasarian (1979). "Exact penalty functions in nonlinear programming." In: *Mathematical Programming* 17.1, pp. 251–269. ISSN: 1436-4646. DOI: 10.1007/BF01588250.

Heuser, Harro (1990). *Lehrbuch der Analysis. Teil 1*. B. G. Teubner Stuttgart.

– (1993). *Lehrbuch der Analysis. Teil 2*. B. G. Teubner Stuttgart.

Horst, Reiner and Hoang Tuy (2013). *Global optimization: Deterministic approaches*. Springer Science & Business Media.

Lemarechal, C. and R Mifflin (1978). *Nonsmooth optimization*. Vol. 3.

Liberzon, Daniel (2011). *Calculus of variations and optimal control theory: a concise introduction*. Princeton University Press.

Markowitz, Harry (1952). "Portfolio Selection." In: *The Journal of Finance* 7.1, pp. 77–91. DOI: 10.1111/j.1540-6261.1952.tb01525.x.

Nocedal, Jorge and Stephen J. Wright (2006). *Numerical Optimization*. 2nd. Berlin: Springer. DOI: 10.1007/978-0-387-40065-5.

Schewe, Lars and Martin Schmidt (2019). *Optimierung von Versorgungsnetzen. Mathematische Modellierung und Lösungstechniken*. Springer Spektrum, Berlin, Heidelberg. ISBN: 978-3-662-58538-2. DOI: 10.1007/978-3-662-58539-9.

Schrijver, Alexander (1998). *Theory of linear and integer programming*. John Wiley & Sons.

Stein, Oliver (2018). *Grundzüge der Globalen Optimierung*. Springer Spektrum, Berlin, Heidelberg. DOI: 10.1007/978-3-662-55360-2.

Ulbrich, Michael and Stefan Ulbrich (2012). *Nichtlineare Optimierung*. Birkhäuser Basel. DOI: 10.1007/978-3-0346-0654-7.